

**Molecular Epidemiology, Cluster
Analysis, and Drug Resistance Prediction
of *Mycobacterium tuberculosis* Complex in
Ireland using Conventional Methods and
Whole Genome Sequencing**

A thesis submitted to Trinity College, Dublin for the
Degree of Doctor of Philosophy
2017

Emma Roycroft MSc

Under the supervision of Prof. Thomas Rogers, Dr. Margaret
Fitzgibbon and Dr. Ronan O'Toole



Declaration

I, Emma Roycroft, declare that this thesis has not been submitted as an exercise for a degree at this, or any other, university and is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository, or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Emma Roycroft

Summary

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* Complex (MTBC), is the joint leading cause of death worldwide due to a single infectious agent, with HIV/AIDS, and remains a major challenge to public health in both low- and high-prevalence countries. It is estimated to have killed 1.5 million people in 2014 (1.1 million co-infected with HIV/AIDS). Latent TB infection is estimated to affect approximately one third of the world's population. TB has been statutorily notifiable in Ireland since 1947. Cases of TB in Ireland have declined from a high of 230 cases per 100,000 in the early 1950s to 6.9 per 100,000 in 2015. However, TB incidence has plateaued at this level in recent years and prevalence remains high in urban centres. Ireland is a European island nation that has experienced mass emigration and immigration over the last century. Immigration of people from outside the European Union, as well as free movement policies within the European Union, has been associated with the spread of TB, especially multi- and extensively-drug resistant TB (MDR/XDR-TB). MDR-TB is defined as resistance to isoniazid and rifampicin, while XDR-TB is defined as resistance to the above plus a fluoroquinolone and aminoglycoside. Drug-resistance-associated mutations have been discovered that can predict phenotypic resistance in MTBC. Disruption of transmission chains is a key factor in controlling the spread of TB. Seven global lineages of MTBC have been elucidated. Mycobacterial Interspersed Repetitive Units – Variable Number Tandem Repeat (MIRU-VNTR) genotyping is the established genotyping method. With the advent of 'sequencing by synthesis' Whole Genome, or Next Generation, Sequencing (WGS/NGS) has become more accessible, faster, and less costly, to the diagnostic laboratory. It is the ultimate genotyping tool, and its uses could be extended much further than genotyping alone.

The primary aim of this study was to examine MTBC genotyping data collected at the Irish Mycobacteria Reference Laboratory (IMRL) from 2010-14 in order to assess the distribution of lineages present and to build on work already published in this regard (n=1,305 strains). It was clear from the results that clusters of MTBC were present in Ireland, and that this was worth further investigation by exploiting the higher resolution achieved by WGS (n=11 informative clusters chosen). It was hypothesised that MIRU-VNTR genotyping had over-estimated TB transmission events, that WGS could resolve these clusters with greater resolution, and that WGS would be a valuable addition to conventional genotyping currently in place in the IMRL. The MIRU-VNTR genotyping results also revealed that MDR/XDR-TB cases had increased in number over the period of the study. A comprehensive survey of MDR/XDR-TB (collected 2001-14, n=42 isolates from 41 patients) was consequently undertaken in order to characterise the strains circulating in Ireland. A number of different WGS analysis platforms were evaluated for their use in drug resistance prediction compared to phenotypic drug susceptibility testing (DST) – an algorithm designed by

Walker and Kohl *et al*, online web-tools PhyResSe and TB Profiler, and ReseqTB data-sharing platform. The IMRL also took part in an international collaborative pilot study that aimed to introduce WGS techniques for mycobacterial identification, MTBC drug resistance prediction, and outbreak detection. This was the most comprehensive molecular characterisation of MTBC strains found in Ireland that has been performed to date.

MIRU-VNTR genotyping is an excellent first-line tool for surveillance of the molecular epidemiology of MTBC in Ireland. Higher diversity of lineages was found within strains collected from 2010-14, than in previous Irish studies, due to the presence of six global lineages, including West African lineages 5 and 6. Euro-American lineage 4 remains predominant. MDR/XDR-TB is present in low, but ultimately increasing, numbers of cases, although mono-resistance remains below 5%. Median cluster size was 2. Within Euro-American lineage 4, clusters of 2 cases constituted over 50%, which compared to previous studies. However, 2.7% clusters were greater than 25 cases, which was not seen previously.

Cluster analysis results showed that WGS could both rule-in and rule-out outbreaks with greater discrimination than MIRU-VNTR genotyping and could confirm recent transmission events, making it a valuable tool in the fight against TB, especially in cases where MIRU-VNTR genotypes are identical. From this and other studies, it is clear that WGS alone cannot infer all epidemiological information. Epidemiological data, contact tracing and genotyping all remain essential tools for MTBC cluster and outbreak investigation.

The molecular characterisation of MDR/XDR-TB strains in Ireland from 2001-14 has proven that these strains are not being readily transmitted within the Irish population, that the drug resistant strains are similar to those found circulating in Europe, and that despite their high diversity, the drug-resistance-associated mutations they harboured were largely similar. WGS genotypic drug resistance prediction matched phenotypic DST in most cases. Statistical sensitivity varied widely, depending on the drug, and analysis platform used, while specificity ranged from 86-100%. When rifampicin and isoniazid alone were analysed, sensitivities ranged from 90-100%, better than rapid molecular tests already available (83-93%). For fluoroquinolones, WGS analysis resulted in sensitivity of 71%. Sensitivity for aminoglycosides and other drugs was lower, although specificity remained high. TB Profiler and PhyReSe resulted in the highest overall sensitivity compared to phenotypic DST. However, a high number of false positives was seen with TB Profiler (n=13). While isoniazid, rifampicin and fluoroquinolone genotypic results were reliable, others require more evidence of phenotypic-genotypic correlation if they are to be used diagnostically. WGS analysis, although currently not at a stage where it could replace phenotypic DST completely, would be an invaluable additional tool within the laboratory for rapid drug resistance prediction of MTBC.

The international collaborative pilot study provided proof of principle that WGS could be employed for drug resistance prediction, nearest neighbour relatedness analysis that could flag outbreaks, and mycobacterial identification, improving laboratory turnaround times significantly, and even decreasing costs by 7% annually according to UK site figures.

Despite its limitations, WGS represents a ‘game-changing’ technology for MTBC and many other microbiological applications.

Acknowledgements

I would like to acknowledge the invaluable guidance and support I have received from my Supervisors during the course of my Ph.D.: Prof. Thomas Rogers, Dr. Margaret Fitzgibbon and Dr. Ronan O'Toole. Each of you brought different expertise to the table, all of which benefited me throughout the process. Thank you for giving me the opportunity to carry out the research, to collaborate internationally, to attend conferences and courses far and wide, and to learn the many skills required in order to complete a PhD.

I also wish to acknowledge support and funding received from the Clinical Microbiology Dept., Trinity College, and the IMRL, Labmed Directorate, St. James' Hospital, in order to carry out the study. The Laboratory Manager John Gibbons, IMRL Chief Medical Scientist Dr. Margaret Fitzgibbon, Former Microbiology Chief Medical Scientist, Helen Barry, and Senior Medical Scientists with responsibility for rosters (in particular Brenda Moloney, Sarah Lalor and Lorraine Montgomery) facilitated my reduced hours for three years and for that I am very grateful.

Many people had an impact on this research. Thank you to:

- The patients who contributed to the study by consenting to have their samples tested for TB. They may not be aware of the study, but the knowledge learned from it will hopefully help people like them in the future
- Prof. Derrick Crook, Prof. Tim Peto, Dr. Tim Walker, Dr. Louise Pankhurst, Dr. Antonina Votintseva, Dr. Ana Gibertoni Cruz, and Carlos Del Ojo Elias, for involving the IMRL in the MGIT Pilot Study and for support and assistance with HiSeq cluster WGS and analysis. I learned so much from the time I spent in Oxford.
- Micheál Mac Aogáin, who is a rare mix of Microbiologist and Computer Scientist, for his helpful assistance with bioinformatics, especially for his collaboration on the two papers that have been written, and life in general
- Prof. Stephen Gordon, Dr. Damien Farrell, and Kevin Conlon, University College, Dublin, and Dr. Javier Nunez, Animal Health and Veterinary Laboratories Agency, UK, for assistance in sequencing and analysing fourteen MTBC strains from an institutional outbreak
- Philomena Raftery and Dr. Margaret Fitzgibbon who performed much of the MIRU-VNTR genotyping in this study.
- The remaining core members of the IMRL (Auveen Griffin, Lorraine Montgomery, Stacey O'Gorman, Martina Kelly, Siobhán Crilly, Syro Hickey (especially for the coffee!), Maeve

Keane, Ronan Gibbons and Simone Mok), every one of the IMRL team has helped me at one point or another and I really appreciate it.

- Treasured colleagues in the SJH Microbiology Laboratory (too many to mention by name, you know who you are!) who have given me encouragement and support and listened to my rants on rare occasions
- The external laboratories that contributed both clinical information and isolates when requested, especially Senior Medical Scientists Mary Lynch Healy in Cork University Hospital and Christine Yearsley in the Mater University Hospital
- St. James' Hospital Surveillance Scientist, Mary Kelleher, and members of the HSE-East Public Health team, especially Dr. Mary O'Meara, Surveillance Scientist Philomena Downes and Dr. Andrew Morgan, for their help with collecting clinical data on clusters of interest and drug-resistant cases
- Health Protection Surveillance Centre colleagues, especially Sarah Jackson and Dr. Joan O'Donnell, for their continued collaboration
- Dr. Brendan Crowley for help with HIV data collection and sage advice and encouragement
- Dr. Geraldine Moloney, Dr. Katie Dunne, Prof. Stephen Smith and Dr. Helen Miajlovic, for feedback, advice, encouragement, ice-cream and cakes along the way
- Prof. Joe Keane and Dr. Helen Miajlovic for their direct and constructive feedback following my continuation report. Their points helped to focus and greatly improve the research.
- Noel Gibbons, for giving me a love of all things TB-related
- My family; my wonderful mum, Kay Roycroft, sister and best friend, Helen, unflappable brother, Bryan, brother-in-law Conor, sister-in-law Liz, niece Kate, all the Fitzgeralds and Hartes (my Dublin family), and those that have inspired and influenced me who are now on the other side. Thank you for your unwavering support, unconditional love, and pride in my efforts and achievements.

Thank you to my husband, Ian Fitzgerald, for lunches, dinners, resilience, strength, and so much love.

Presentations and Publications

Oral Presentations

'Molecular Epidemiology and Drug Resistance in *M.tb.* In Ireland' – Trinity College, Dublin, Research Away Day – November 2012

'Molecular genotyping of *M. tb.*; Know your enemy!' – presentation to Public Health HSE-East, Dr. Steeven's Hospital, Dublin - March 2013

'MGIT Pilot Study – Preliminary Data from the IMRL' - Presentation given to the TB Multi-disciplinary Team in John Houston Ward, SJH, Dublin – February 2014

'TB: Current Concepts and Future Challenges' – SeroSep Microbiology Seminar, Red Cow Hotel, Dublin – November 2015

'Augmenting Public Health TB Surveillance with Whole Genome Sequencing: the IMRL Experience' – Public Health Winter Scientific Meeting – RCPI, Kildare Street, Dublin – December 2015

'Augmenting Public Health TB Surveillance with Genotyping: the IMRL experience' – Infection and Immunity Translational Research Group Research Day, Queen's University, Belfast – Jan 2016

'Extracting Useful Clinical Drug Resistance Data from the Genome of *Mycobacterium tuberculosis*: Next Generation Sequencing (NGS) for the Diagnostic Laboratory' – Irish Next Generation Sequencing Conference, Trinity Centre for Biomedical Science, Dublin – June 2016

Poster Presentations

'A "snapshot" of genetic lineages of *M. tb.* in the Republic of Ireland, 2010-11' – presented at European Society of Mycobacteriology Congress, Brasov, Romania, July 2012, Molecular Medicine Ireland Annual Scientific Meeting, March 2013, and SJH Centre for Learning and Development Clinical Audit and Research Seminar, April 2013

'Mining the Whole Genome of Mycobacteria - HICF Collaboration Data' - presented at Molecular Medicine Ireland Annual Education and Training Scientific Meeting 2014, Science Gallery, and Biomedica, RDS, Dublin – April 2014

'Molecular Characterisation of the first Extensively Drug Resistant (XDR) TB in Ireland' - European Society of Mycobacteriology Congress 2014, Vienna, Austria – June 2014

'Augmenting Public Health TB Surveillance with Whole Genome Sequencing: the IMRL Experience' – Public Health Winter Scientific Meeting – RCPI, Kildare Street, Dublin, December 2015, and American Society of Microbiology Microbe 2016, Boston, Massachusetts, USA, June 2016

'Evaluation of online web tools for the prediction of drug resistance and genotyping in *Mycobacterium tuberculosis* Complex in Ireland, a low prevalence country, from a User's perspective' - European Society of Mycobacteriology Congress, Catania, Sicily - July 2016

Publications

A snapshot of genetic lineages of *Mycobacterium tuberculosis* in Ireland over a two-year period, 2010 and 2011. Fitzgibbon MM, Gibbons N, Roycroft E, Jackson S, O'Donnell J, O'Flanagan D, Rogers TR. Euro Surveillance 2013;18(3) (Appendix 1)

Draft Genome Sequence of the First Isolate of Extensively Drug-Resistant *Mycobacterium tuberculosis* in Ireland. Roycroft E, Mac Aogáin M, O'Toole RF, Fitzgibbon M, Rogers TR. Genome Announcements 2014; 2(5) (Appendix 2)

Rapid, Comprehensive, and Affordable Mycobacterial Diagnosis with Whole-genome Sequencing: a Prospective Study. Pankhurst J, Del Ojo Elias C, Votintseva A, Walker T, Cole K, Davies J, Fermont J, Gascoyne-Binzi D, Kohl T, Kong C, Lemaitre N, Niemann S, Paul J, Rogers T, Roycroft E *et al.* *The Lancet Respiratory Medicine*, 4(1), pp.49-58. (Appendix 3)

List of Abbreviations

A	Adenine
A	Alanine
Ac/Ac2PIM2	Tri-/tetra-acylated phosphatidyl-myo-inositol-dimannoside
Ac/Ac2PIM6	Tri-/tetra-acylated phosphatidyl-myo-inositol-hexamannoside
AFB	Acid Fast Bacilli
AG	Arabinogalactan
AGP	Arabinogalactanpeptidoglycan complex
AIDS	Acquired Immune Deficiency Syndrome
AMK	amikacin
AMNCH	Adelaide and Meath Hospital incorporating the National Children's Hospital, Tallaght, Dublin 22
APC	Antigen Presenting Cell
AS	All other mycobacterial species
ATYP	ZN morphology atypical or NTM-like
AWS	Amazon Web Services
BACTEC	Becton Dickinson Bactec liquid culture system
BAL	Broncho-alveolar Lavage
BAM	Binary Alignment/Map format
BCF	Binary Calling File
BCG	Bacille Calmette Guerin
BLAST	Basic Local Alignment Search Tool
BMGF	Bill and Melinda Gates Foundation
BWA	Burrows Wheeler Aligner
BX	Biopsy
C	Cytosine
C	Cysteine
CAP	capreomycin
CF	Cystic Fibrosis
CFP	Culture Filtrate Protein
CFZ	clofazimine
CI	Confidence Interval
CIDR	Computerised Infectious Disease Reporting
CIP	ciprofloxacin
CL3	Containment Level 3
CLA	clarithromycin

CM	Common Mycobacteria
COMPASS-TB	Complete Pathogen Sequencing Solution - Tuberculosis
C-Path	Critical Path Institute
CryPTIC	Comprehensive Resistance Prediction for Tuberculosis – International Consortium
CSF	Cerebro-spinal Fluid
CSO	Central Statistics Office
CTAB	cetyl trimethylammonium bromide
CYC	cycloserine
D	Aspartic Acid
DAT	Diacyltrehalose
Delhi/CAS	Delhi/Central Asian Strain
DNA	Deoxyribonucleic Acid
dNTP	deoxynucleotide triphosphates
DOT	Directly Observed Therapy
DPG	Diphosphatidylglycerol
DR	Direct Repeat
dsDNA HS	Double-stranded DNA High Sensitivity assay
DST	Drug susceptibility testing
DST	Drug susceptibility testing
E	Glutamic Acid
E/EMB	ethambutol
EAI	East African Indian lineage
ECDC	European Centre for Disease Control
EPR	Electronic Patient Record
ER	Emma Roycroft (TCD and IMRL)
ESAT-6	Early Secretory Antigenic Target
ETI	ethionamide
EU	European Union
F	Phenylalanine
FIND	Foundation for Innovative and New Diagnostics
G	Guanine
G	Glycine
GalNH₂	Galactosamine residue
GC	Glycine-Cytosine
GNU	GNU not Unix
GTR	Generalised Time Reversible
GUI	Graphical User Inter-face

H	Histidine
HCW	Healthcare Worker
HIV	Human Immunodeficiency Virus
HKY	Hasegawa, Kishino and Yano
HPA	Health Protection Agency (now PHE)
HPSC	Health Protection Surveillance Centre
I	Isoleucine
I/INH	isoniazid
IFNγ	Interferon Gamma
IGV	Integrated Genome Browser
IL-2	Interleukin
IMRL	Irish Mycobacteria Reference Laboratory
IMS	industrial methylated spirit
IQR	Inter-Quartile Range
ISO15189	International Standards Organisation
IT	Information Technology
IVDU	Intra-venous Drug User
K	Lysine
k	keto
KAN	kanamycin
L	Leucine
LAM	Latin American Mediterranean lineage
LAM	Lipoarabinomannan
LIS	Laboratory Information System
LJ	Lowenstein Jensen
LM	Lipomannan
LPA	line-probe assay
LSP	Long Sequence PCR
LTBI	Latent Tuberculosis Infection
LTS	Long Term Support
LZD	linezolid
M	Methionine
m	methoxy
MA	Mycolic acids
MALDI-TOF	Matrix Assisted Laser Desorption/Ionisation –Time of Flight
MDR-TB	Multi-Drug Resistant Tuberculosis
MEM	Maximal Exact Matches
MF	Margaret Fitzgibbon (IMRL)

MGIT	Mycobacteria Growth Indicator Tube
MIC	Minimum inhibitory concentration
MIM	Mycobacterial inner membrane
MIRU-VNTR	Mycobacterial Interspersed Repetitive Units – Variable Number Tandem Repeats
MLST	Multi-Locus Sequence Typing
MLVA	Multiple Locus Variant Analysis
MMM	Modernising Medical Microbiology
MOM	Mycobacterial outer membrane
MRSA	Methicillin-resistant <i>Staphylococcus aureus</i>
MSC	Microbiological safety cabinet
MSP	Minimum Spanning Tree
MTBC/MtbC	Mycobacterium tuberculosis Complex
M.tb	Mycobacterium tuberculosis
MXF	moxifloxacin
N	Asparagine
N/A	Not Applicable
NaOH	Sodium Hydroxide
NCBI	National Centre for Biotechnology Information
NDWG	New Diagnostics Working Group
NGS	Next Generation Sequencing
NI	Northern Ireland
NICE	National Institute for Health and Care Excellence
NIH	National Institutes of Health
NPF	Naso-Pharyngeal Flora
NTC	Non-Template Control
NTM	non-tuberculous mycobacteria
OFX	ofloxacin
OS	Operating System
P	Proline
PE	Phosphatidylethanolamine
P/PZA	pyrazinamide
PANTA	polymixin B, amphotericin B, naladixic acid, trimethoprim, azlocillin
PAS	para-aminosalicylate sodium
PAT	Polyacyltrehalose
PDIM	Phthiocerol dimycocerosate
PG	Peptidoglycan

PI	Phosphatidyl-myo-inositol
PIM	Phosphotidylmyo-inositol mannoside
PCR	Polymerase Chain Reaction
PE/PPE	Proline-Glutamate / Proline-Proline-Glutamate - novel regions of the MTBC genome encoded by <i>pe/ppe</i> genes, that make up 10% coding region
PE	paired end
PGRS	Polymorphic GC-rich sequence
PHE	Public Health England
PhyML	Phylogenetics based on Maximum Likelihood
PhyResSe	Phylo-Resistance-Search-Engine
PPD	Purified Protein Derivative
PPE	Personal protective equipment
PR	Philomena Raftery (IMRL)
PRO	prothionamide
Q	Glutamine
QC	Quality Control
R	Arginine
R	Resistant
R/RIF	rifampicin
RFB	rifabutin
RFLP	Restriction Fragment Length Polymorphism
RLL	Right Lower Lobe
RMB	Right Mid Lobe
RNA	Ribo-nucleic Acid
RoD	Regions of Difference
RT	Room Temperature
RUL	Right Upper Lobe
S	Serine
S	Susceptible
S/SM	streptomycin
SaaS	Software as a Service
SAMtools	Sequence Alignment/Map tools
SBS	Sequencing by Synthesis
SGL	Sulfoglycolipid
SIRE	streptomycin, isoniazid, rifampicin, ethambutol
SJH	St. James' Hospital, Dublin 8
SLV	Single Locus Variation

SM	streptomycin
SMRL	Scottish Mycobacteria Reference Laboratory
SNV	Single Nucleotide Variation
SRA	Short Read Archive
T	Thymine
T	Threonine
TB	Tuberculosis
TCD	Trinity College, Dublin
TDR	Total Drug Resistance (TB)
TNF	Tumour Necrosis Factor
TST	Tuberculosis Skin Test
TTP	Time to Positivity
TYP	ZN morphology typical of MTBC
UCD	University College, Dublin
UK	United Kingdom
UK	United Kingdom
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
US CDC	United States Centres for Disease Control
USA	United States of America
UV	ultra-violet light
V	Valine
VCF	Variant Calling File
W	Tryptophan
WGS	Whole Genome Sequencing
WHO	World Health Organisation
XDR-TB	Extensively-Drug Resistant Tuberculosis
Y	Tyrosine
ZN	Ziehl Neelsen

Table of Contents

Summary.....	i
Acknowledgements	iv
Presentations and Publications	vi
Oral Presentations.....	vi
Poster Presentations.....	vi
Publications	vii
List of Abbreviations.....	viii
Table of Contents	xiv
List of Figures.....	xxi
List of Tables.....	xxiv
1 General Introduction and Hypotheses	1
1.1 Background.....	1
1.2 Mycobacterium Species.....	1
1.2.1 Mycobacterium tuberculosis Complex (MTBC).....	1
1.3 History and Expansion of Tuberculosis.....	2
1.4 Global Epidemiology of TB	2
1.5 Epidemiology of TB in Ireland.....	2
1.6 Reasons TB has not been eradicated	3
1.7 Pathogenesis of TB and latency.....	3
1.7.1 Innate Immune Response.....	3
1.7.2 Adaptive Immune Response	4
1.7.3 Granuloma formation	4
1.8 Symptoms and Clinical Diagnosis.....	5
1.8.1 Tuberculin Skin Test and Interferon Gamma Release Assays	5
1.9 TB Treatment and Vaccination.....	5
1.10 Drug Resistance.....	6
1.10.1 Genes involved in resistance	7

1.11	Irish Mycobacteria Reference Laboratory (IMRL)	8
1.11.1	Microscopy and Morphology	8
1.11.2	Culture and Drug Susceptibility Testing	8
1.11.3	Rapid Molecular Tests	8
1.11.4	MIRU-VNTR Genotyping	9
1.12	Bio-informatic Analysis Software.....	9
1.12.1	Linux Operating System and the Command Line.....	9
1.12.2	Open-source Bio-informatics software	10
1.12.3	Commercial Bio-informatics software	10
1.13	Phylogenetic Analysis.....	10
1.13.1	Multiple Sequence Alignment.....	11
1.13.2	Neighbour Joining Tree.....	11
1.13.3	Maximum Likelihood Tree	11
1.13.4	Models of Evolution.....	12
1.13.5	UPGMA tree	12
1.13.6	Minimum Spanning Tree	12
1.14	Whole Genome Sequencing (WGS)	12
1.14.1	Progression from Sanger to Next Generation Sequencing	12
1.14.2	Bridge Amplification	13
1.14.3	Sequencing by Synthesis.....	13
1.15	Impetus for this PhD Research.....	14
1.16	Hypotheses and Aims of the Study	15
2	Materials and Methods.....	17
2.1	Study Ethics, Safety, and Storage of Isolates.....	17
2.2	Working in a Containment Level 3 (CL3) facility with a Category 3 Pathogen.....	17
2.3	Storage of Mycobacterial Isolates at the IMRL	17
2.4	Sample Selection.....	17
2.4.1	Sample Selection for Molecular Epidemiology of MTBC in Ireland	18
2.4.2	Sample Selection for In-depth whole genome sequencing analysis of MIRU-VNTR Genotype clusters of interest.....	18

2.4.3	Sample Selection for Molecular Characterisation of MDR/XDR-TB in Ireland	19
2.4.4	Sample Selection for BD Bactec™ MGIT™ 960 Early Positive Culture Pilot Study	19
2.5	Mycobacterial Culture using the BD Bactec™ 960 MGIT™ Liquid Culture System and Lowenstein Jensen solid medium and Identification using Hain GenoType Line Probe Assays.	20
2.6	Ziehl Neelsen Stain (ZN).....	20
2.7	Anti-tuberculous Drug Susceptibility Testing (DST) using the BD Bactec™ MGIT™ 960 Liquid Culture System.....	21
2.8	Heat Inactivation of Isolates	21
2.9	Hain Lifescience GenoType Line Probe Hybridisation Assays for rapid molecular detection of drug resistance in tuberculosis.....	22
2.10	Crude DNA Extraction from ZN positive liquid cultures.....	22
2.10.1	Crude Extraction.....	22
2.10.2	Hain Lifescience GenoLyse® Extraction	23
2.11	Nucleic Acid Amplification.....	23
2.12	PCR Product Line Probe Assay Reverse Hybridisation and Analysis	24
2.13	Automated MIRU-VNTR Genotyping of <i>Mycobacterium tuberculosis</i> Complex (MTBC)	24
2.13.1	MIRU-VNTR locus amplification.....	24
2.13.2	MIRU-VNTR Fragment Analysis	24
2.13.3	Analysis and Allele Assignment using GeneMapper® Software.....	25
2.13.4	MLVA Compare Software and the MIRU-VNTR ^{plus} online database.....	25
2.14	Whole Genome Sequencing using Illumina® Next Generation Sequencing Technology (MiSeq® and HiSeq®)	26
2.14.1	Whole genome DNA Extraction.....	26
2.14.2	AMPure XP beads ‘Clean-Up’	27
2.14.3	DNA Extract and/or Library Quantification.....	28
2.14.4	Illumina® Nextera XT Library Preparation and Quantification.....	28
2.14.5	Illumina® MiSeq® and BaseSpace Cloud Computing	30
2.14.6	Illumina® HiSeq®	30
2.15	Whole Genome Sequencing Analysis of output fastq files	31

2.15.1	Illumina® Basespace cloud platform	31
2.15.2	Linux Operating System	31
2.15.3	FastQC software, quality control, and trimming.....	31
2.15.4	BWA-MEM alignment.....	31
2.15.5	SAMtools and Bcftools for Variant Calling.....	32
2.15.6	Whole Genome Cluster Analysis	33
2.15.7	Whole Genome Variant Analysis for Resistance Mutation Detection.....	37
2.15.8	Workflow analysis of Draft Genome Sequence of the first XDR-TB strain in Ireland using Linux Command Line and open-source software.....	38
2.15.9	Algorithm developed by Walker and Kohl et al for resistance mutation detection in 23 candidate genes	40
2.15.10	TB Profiler freely available online web-tool for analysis of MTBC genomes	41
2.15.11	PhyResSe freely available online web-tool for analysis of MTBC genomes	41
2.15.12	Comparison of Online Web-tools for the Prediction of TB Drug Resistance	42
2.15.13	ReseqTB data-sharing platform mutation database.....	42
2.16	Comparison of WGS versus conventional DST for anti-tuberculous drug resistance detection in an MDR/XDR-TB cohort.....	42
2.17	MGIT™ Pilot Study pipeline for identification of all mycobacteria, and resistance profiling and Nearest-Neighbour Relatedness Analysis of MTBC.....	42
3	Molecular Epidemiology of MTBC in Ireland, 2010-14	45
3.1	Introduction	45
3.2	Results.....	47
3.2.1	Samples included in the Study	47
3.2.2	Mixed MTBC Infection.....	47
3.2.3	Health Protection Surveillance Centre TB Data, 2010-14	47
3.2.4	MTBC Lineage Distribution in Isolates collected in Ireland, 2010-14	48
3.2.5	MTBC Clusters identified	48
3.2.6	Anti-TB Drug Susceptibility Patterns Observed.....	48
3.3	Discussion	50
4	Whole Genome Phylogenetic Analysis of MTBC Isolates from Eleven Informative MIRU-VNTR genotyping Clusters.....	54

4.1	Introduction	54
4.2	Results	57
4.2.1	IMRL Cluster 1 – community-based substance abuse, Haarlem.....	57
4.2.2	IMRL Cluster 2 – community-based substance abuse, LAM.....	59
4.2.3	IMRL Cluster 3 – community-based drug-susceptible, H37RV (Dublin).....	60
4.2.4	IMRL Cluster 4 – community-based drug-susceptible, H37RV (countrywide).....	61
4.2.5	IMRL Cluster 5 – regionally-dispersed ethnic group, Delhi/CAS	61
4.2.6	IMRL Cluster 6 – regionally-dispersed ethnic group, EAI	62
4.2.7	IMRL Cluster 7 – community-based drug-resistant, Beijing	62
4.2.10	IMRL Cluster 10 – residential institution, Beijing	64
4.2.11	IMRL Cluster 11 – non-residential institution, Euro-American.....	65
4.3	Discussion.....	66
5	Molecular Characterisation of MDR-/XDR-TB in Ireland, collected 2001-2014, and Evaluation of WGS Analysis Techniques compared to Conventional Phenotypic Methods for Drug Resistance Prediction.....	70
5.1	Introduction	70
5.2	Results	73
5.2.1	Sample Selection and Patient Demographics	73
5.2.2	Molecular Epidemiology of the MDR/XDR-TB Cohort – MIRU-VNTR Genotyping 74	
5.2.3	Molecular Epidemiology of the MDR/XDR-TB Cohort – Whole Genome Sequencing compared to Conventional Genotyping	75
5.2.4	Phenotypic Drug Susceptibility Testing.....	77
5.2.5	Rapid Molecular Line Probe Assays <i>MTBDRplus</i> and <i>MTBDRsl</i>	77
5.2.6	Drug Resistance Prediction using Whole Genome Sequencing	80
5.2.7	Sequential Isolate Study I: Progression from Susceptible to MDR-TB	87
5.2.8	Sequential Isolate Study II: Progression from MDR-TB to XDR-TB.....	88
5.2.9	Sequential Isolate Study III	89
5.2.10	Draft Genome Sequence of the First XDR-TB Isolate in Ireland.....	89
5.3	Discussion.....	91
5.3.1	Molecular Epidemiology of MDR/XDR-TB in Ireland and compared with Europe	91

5.3.2	Multi- and Extensive Drug Resistance in Ireland and compared worldwide.....	92
5.3.3	Draft genome of the first XDR in Ireland	94
5.3.4	Sequential isolates, progression to MDR-TB, progression to XDR-TB, re-infection 94	
5.3.5	Drug Resistance Prediction using rapid genotypic tools (LPAs and WGS) compared to phenotypic DST	96
5.3.6	Comparison of online web-tools for drug resistance prediction in MTBC	98
5.3.7	Limitations of the Study	98
5.3.8	Health Outcomes.....	99
5.4	Conclusions	100
6	Prospective Pilot study to identify Mycobacteria, and Detect Drug Resistance and Nearest- Neighbour Relatedness of MTBC, using Whole Genome Sequencing of Early Positive Liquid Cultures	101
6.1	Introduction	101
6.2	Results.....	103
6.2.1	COMPASS-TB Study Group	103
6.2.2	Samples Submitted.....	103
6.2.3	Sample and Illumina MiSeq Run Details.....	103
6.2.4	Identification using MMM WGS workflow compared to Conventional Methods	103
6.2.5	Anti-TB Drug Resistance Prediction using MMM WGS workflow compared to Conventional DST.....	105
6.2.6	Nearest Neighbour Relatedness using MMM WGS workflow.....	106
6.2.7	Reporting of Results by the MMM WGS workflow	107
6.2.8	WGS Contamination with Human and Nasopharyngeal Flora DNA.....	107
6.2.9	Reporting Times and Costs associated with WGS compared to Conventional Methods	107
6.3	Discussion	109
7	General Discussion, Conclusions, and Future Directions	115
7.1	MIRU-VNTR genotyping	116
7.2	Cluster analysis	117
7.3	MDR/XDR-TB.....	119

7.4	MGIT Pilot study.....	121
7.5	Critical Evaluation of Whole Genome Sequencing.....	121
7.5.1	Advantages over Conventional Methods.....	121
7.5.2	Limitations and challenges	122
7.6	Systems biology approach	124
7.7	Conclusions	125
7.8	Future directions and hypotheses generated.....	126
	Bibliography	127

List of Figures

Figure 1. Phylogenetic tree of Actinobacteria based on 1,500 nucleotides of 16S rRNA	1
Figure 2. The Mycobacterial Cell Wall	1
Figure 3. The complete genome of Mycobacterium tuberculosis	1
Figure 4. Proposed out-of-Africa expansion of human MTBC	2
Figure 5. Latency and Reactivation	4
Figure 6. Granuloma formation by the tubercle bacillus	4
Figure 7. IMRL Algorithm for Culture and DST of Mycobacteria species versus Algorithm proposed by the MGIT Pilot Study	8
Figure 8. Image of a Ziehl Neelsen Stain of Mycobacteria, taken at the IMRL	8
Figure 9. Visual representation of the principle of MIRU-VNTR genotyping	9
Figure 10. Next Generation Sequencing-by-Synthesis	13
Figure 11. Proposed Evolution of the seven Global Lineages of modern MTBC strains	14
Figure 12. Illumina Miseq Next Generation Sequencing Platform	18
Figure 13. Examples of Hain GenoType MTBDRplus Line Probe Assay reverse hybridised strips	22
Figure 14. Hybridisation strips for version 1 and 2 Hain Genotype MTBDRplus Line Probe Assay	22
Figure 15. Algorithm for phylogenetic lineage-calling for 24-locus MIRU-VNTR genotypes	26
Figure 16. Quickgene Mini-80 Instrument used for Adapted Quickgene extraction protocol for MDR/XDR-TB cohort, and isolates for whole genome cluster analysis	27
Figure 17. Illumina MiSeq Run Detail for MGIT Pilot Study Run 2	30
Figure 18. Per Base Sequence Quality for a fastq file (read 1, IEXDR1) measured with FastQC software	31
Figure 19. Representations of the workflow proposed and implemented by the COMPASS-TB group as part of the MGIT Pilot Study	42
Figure 20. Example of the report format (with QC report) envisaged by the MGIT Pilot Study collaborators (IMRL15).	44
Figure 21. Categorical Radial Neighbour Joining dendrogram of 24-locus MIRU-VNTR genotypes collected in Ireland from January 2010 to December 2014 (n = 1305)	48
Figure 22. Overview of the distribution of all Beijing strains (2010-14) visualised via a minimum-spanning tree built using MLVA Compare software	48
Figure 23. Zoomed area of the Beijing Minimum Spanning Tree showing the largest Beijing outbreak within the 2010-14 cohort, Cluster 10 (n=20), using MLVA Compare software	48
Figure 24. Zoomed area of the LAM lineage distribution (2010-14) visualised in a Minimum Spanning Tree showing the largest LAM outbreaks within the cohort, using MLVA Compare software.	48

Figure 25. Zoomed area of the Euro-American sub-lineage Haarlem distribution (2010-14, Cluster 1) visualised in a Minimum Spanning Tree, built using MLVA Compare software.	48
Figure 26. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 1 isolates	57
Figure 27. Genetic distances within the three largest clusters found; Clusters 1, 2, and 10	58
Figure 28. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 2a and 2b isolates	59
Figure 29. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 3 isolates	60
Figure 30. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 4a and 4b isolates	61
Figure 31. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 5 isolates and sequences from a UK Midlands Study	61
Figure 32. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 6 isolates	62
Figure 33. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 7 isolates	62
Figure 34. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 8 isolates	63
Figure 35. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 9a, 9b, and 9c isolates	63
Figure 36. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 10 isolates	64
Figure 37. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 11 isolates	65
Figure 38. MIRU-VNTRplus database neighbour joining phylogeny constructed with MIRU-VNTR genotyping data from MDR/XDR-TB isolates collected in Ireland from 2001-2014.	74
Figure 39. Polar radial WGS Maximum likelihood phylogeny constructed using variant calling files from the whole genomes of MDR/XDR-TB strains isolated in Ireland from 2001-2014.	75
Figure 40. Transformed cladogram constructed using the whole genome data from MDR/XDR-TB isolates collected in Ireland from 2001-2014.	75
Figure 41. Distribution of mycobacterial species found over the course of the study and correlation of species identification (by NGS versus conventional methods) for IMRL isolates (n=36).	

Figure 42. The bigger picture; Distribution of mycobacterial species found over the course of the study and correlation of species identification (by NGS versus conventional methods) for all participating sites (n=356).	105
Figure 43. Percentage of DNA contamination (human and nasopharyngeal flora) present compared to auramine microscopy result in IMRL MGIT Study isolates	107
Figure 44. Percentage of DNA contamination (human and nasopharyngeal flora) present compared to Sample Type in IMRL MGIT Study isolates	107
Figure 45. Percentage of DNA contamination (human and nasopharyngeal flora) present compared to Time to Positivity (TTP) in IMRL MGIT Study isolates	107
Figure 46. The Cycle of Poverty and Tuberculosis	115
Figure 47. Systems Biology Approach to the development of new Clinical Tools, Therapeutics and Vaccines	124

List of Tables

Table 1 Global lineages of M.tuberculosis Complex and their Sub-lineages	2
Table 2. Summary of Programs, Websites, Software and Scripts used throughout the Study.	31
Table 3. Distribution of global and sub-lineages lineages among Mycobacterium tuberculosis Complex isolates in Ireland, 2010-11 (n=361) compared to 2010-2014 (n=1,306).	46
Table 4. HPSC outbreak data reported from 2010 to 2014 inclusive.	47
Table 5. Total number, and size break-down, of MIRU-VNTR genotyping clusters present in Ireland from 2010-2014	48
Table 6. Mono-resistance seen in MTBC isolates collected between January 2010 and December 2014.	48
Table 7. Compiled details of isolates chosen for in-depth WGS analysis of MIRU-VNTR genotyping informative clusters.	57
Table 8. MDR/XDR-TB Patient demographics.	73
Table 9. Lineage assignation according to the whole genomes of the MDR/XDR-TB cohort, collected in Ireland 2001-14., assigned by TB Profiler and PhyResSe.	75
Table 10. MDR/XDR-TB conventional phenotypic drug susceptibility testing (DST) results, 2001-2014.	77
Table 11. Sensitivity and specificity of WGS Methods for Drug Resistance Prediction in MTBC (genotypic) compared to phenotypic DST for various anti-tuberculous drugs	77
Table 12. Single Nucleotide Variations (SNVs) that have been associated with resistance to isoniazid, rifampicin and ethambutol, found in MDR/XDR-TB isolates collected in Ireland from 2001-2014.	80
Table 13. Single Nucleotide Variations (SNVs) that have been associated with resistance to pyrazinamide, streptomycin, fluoroquinolones and aminoglycosides, found in MDR/XDR-TB isolates collected in Ireland from 2001-2014.	80
Table 14. Single Nucleotide Variations (SNVs) found in the MDR/XDR-TB cohort, collected 2001-14, when 43 further genes were analysed.	80
Table 15. Summary of comparison between Walker/Kohl et al WGS analysis algorithm (genotypic) and conventional isoniazid, rifampicin, ethambutol and pyrazinamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014	80
Table 16. Summary of comparison between Walker/Kohl et al WGS analysis algorithm (genotypic) and conventional streptomycin, fluoroquinolone and aminoglycoside DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014	80
Table 17. Summary of correlation between ReseqTB resistance mutation catalogue (genotypic) compared to isoniazid, rifampicin, ethambutol and pyrazinamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014	80

Table 18. Summary of correlation between ReseqTB resistance mutation catalogue (genotypic) compared to fluoroquinolone, aminoglycoside and ethionamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014	80
Table 19. Summary of comparison between PhyResSe TB NGS analysis web-tool (genotypic) and isoniazid, rifampicin, ethambutol and pyrazinamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014	80
Table 20. Summary of comparison between PhyResSe TB NGS analysis web-tool (genotypic) and streptomycin, fluoroquinolone and aminoglycoside DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014	80
Table 21. Summary of comparison between TB Profiler NGS analysis web-tool (genotypic) and isoniazid, rifampicin, ethambutol and pyrazinamide DST (phenotypic) for the MDR/XDR-TB cohort from 2001-2014	80
Table 22. Summary of comparison between TB Profiler NGS analysis web-tool (genotypic) and streptomycin, fluoroquinolones, aminoglycosides and ethionamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014.	80
Table 23. Discrepancies found for various TB drugs when analysed using phenotypic DST compared to Walker/Kohl algorithm, PhyResSe, TB Profiler and ReseqTB catalogue, where DST available. CAP not reported by PhyResSe, ETI not reported by Walker/Kohl algorithm, PAS, LZD and CFZ not reported by Walker/Kohl algorithm, PhyResSe or ReseqTB.	80
Table 23. Comparison of open-source TB NGS analysis web-tools PhyResSe and TB Profiler	80
Table 25. Diagnostic susceptibility testing (DST) performed by different laboratories on IEXDR1 3	90
Table 26. SNVs found in IEXDR1	90
Table 27. MIRU-VNTR profile of IEXDR1	90
Table 28. Results of COMPASS-TB MGIT Pilot Study: IMRL isolates	103
Table 29. Table comparing identification of isolates using NGS (COMPASS-TB pipeline) and Hain GenoType LPAs CM, AS and MTBC for IMRL isolates	104
Table 30. Table comparing Drug Resistance Prediction results achieved for M. tuberculosis Complex isolates using NGS (COMPASS-TB pipeline) vs. phenotypic DST	105

Chapter 1.

General Introduction and Hypotheses

1 General Introduction and Hypotheses

1.1 Background

Tuberculosis (TB) is the joint leading cause of death worldwide due to a single infectious agent, along with HIV/AIDS, and remains a major challenge to public health. It is estimated to have killed 1.5 million people in 2014 (1.1 million co-infected with HIV/AIDS) [1]. It is also estimated that today approximately one third of the world's population is latently infected with *M. tuberculosis* and it remains the leading killer among those infected with HIV as an AIDS defining illness [2, 3]. 95% of TB deaths occur in the developing world. In 2012, over 10 million children were orphaned due to TB deaths [4].

1.2 Mycobacterium Species

Evolutionary studies have found that *Mycobacterium* species are members of the *Actinomycetales* within the Actinobacteria family (Figure 1) [5]. *Mycobacterium* species can be sub-divided into non-tuberculous mycobacteria (NTM) and *Mycobacterium tuberculosis* Complex (MTBC). NTM can be further sub-divided into slow- and rapid-growers. Examples of the more common slow-growers include *Mycobacterium avium*, *M. intracellulare*, *M. kansasii*, *M. xenopi*, *M. lentiflavum*, *M. marinum*, *M. ulcerans*, and *M. chimaera*. Examples of rapid-growers include *Mycobacterium fortuitum*, *M. chelonae*, *M. abscessus*, and *M. goodii*. NTMs are opportunistic pathogens which tend to cause diseases that are less severe than tuberculosis, but can be chronic in nature and affect immuno-compromised patients such as those with HIV or cystic fibrosis [6, 7]. To date, 171 species of NTM have been identified.

1.2.1 Mycobacterium tuberculosis Complex (MTBC)

Tuberculosis is caused by infection with an acid-fast bacillus of the *Mycobacterium tuberculosis* Complex (MTBC), which can include *M. tuberculosis*, *M. africanum*, *M. bovis*, *M. bovis BCG*, *M. microti*, *M. canettii*, *M. caprae* or *M. pinnipedii*. Most cases are caused by *M. tuberculosis* and affect the respiratory system, although TB may affect any part of the body. MTBC are fastidious slow-growers (divide approximately every 15-20 hours). They appear 'rough, tough, tacky and buff' on solid medium, primarily because of the presence of mycolic acids in their bacterial cell walls (Figure 2). They grow aerobically at 35-37°C in Middlebrook 7H9 liquid culture medium or Lowenstein Jensen solid culture medium. The *M. tuberculosis* genome was sequenced in 1998 (Figure 3) [8].

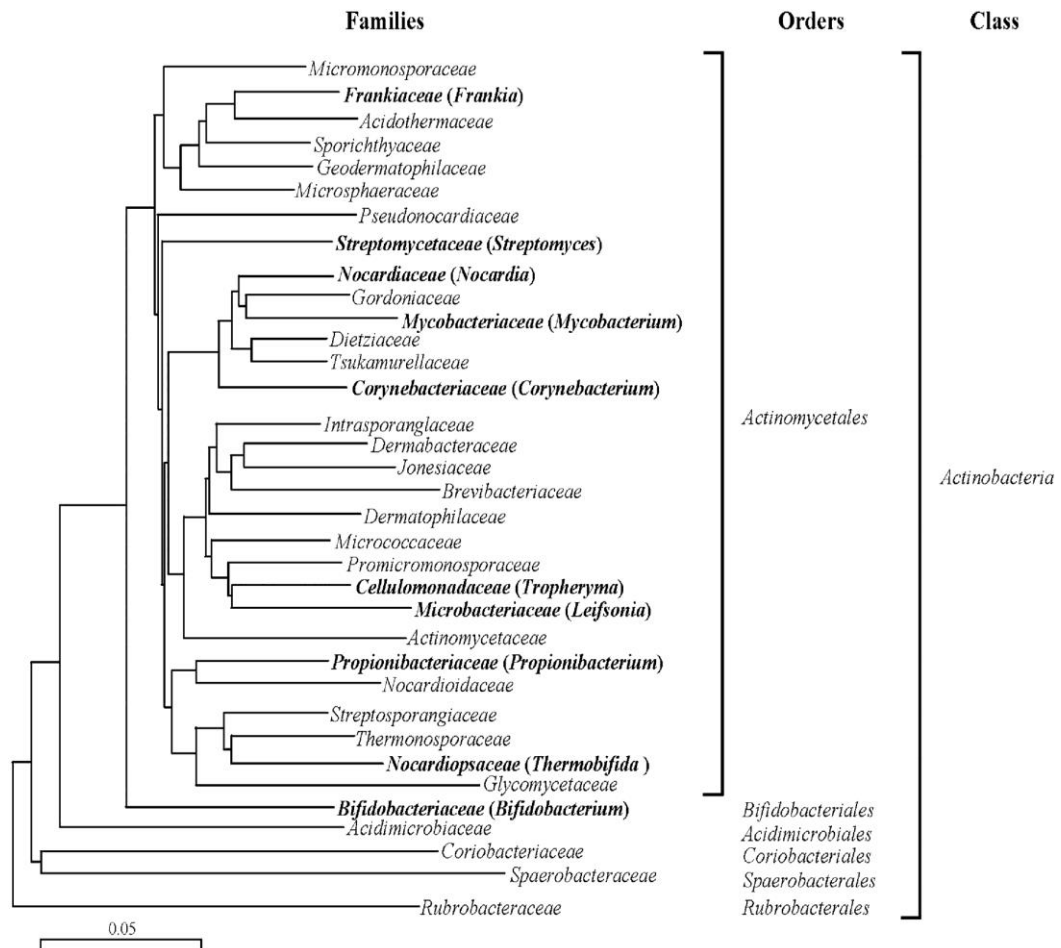


Figure 1. Phylogenetic tree of Actinobacteria based on 1,500 nucleotides of 16S rRNA

Mycobacteriaceae are members of the *Actinomycetales* [5]

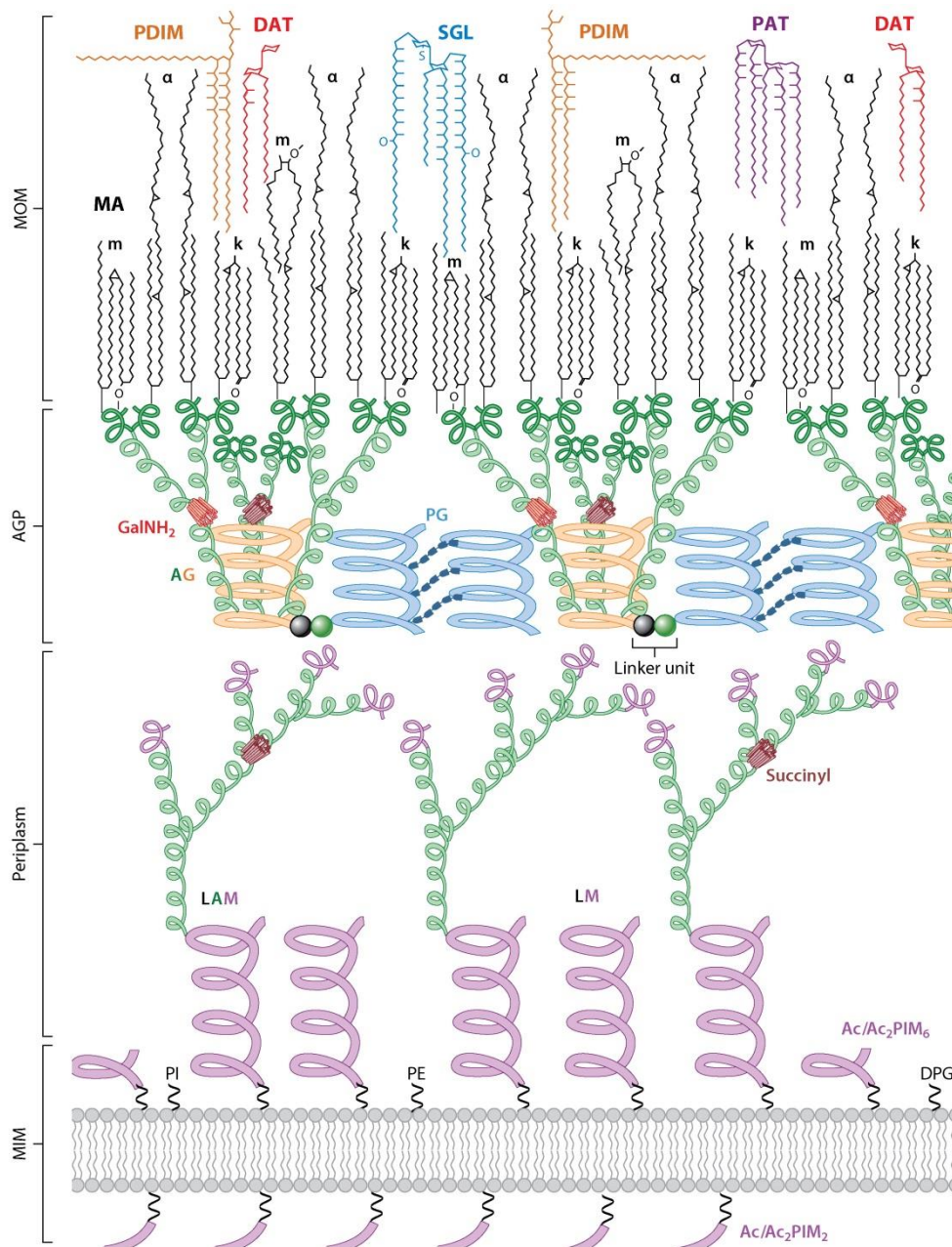


Figure 2. The Mycobacterial Cell Wall

A figure depicting the mycobacterial cell wall and its components [248]. MTBC cell walls are thicker, more waxy and hydrophobic than other bacterial cell walls. They consist of an outer mycolic acid membrane, an arabinogalactan-peptidoglycan layer, a periplasm, and an inner membrane. Ac/Ac₂PIM₂, tri-/tetra-acylated phosphatidyl-myco-inositol-dimannoside; Ac/Ac₂PIM₆, tri-/tetra-acylated phosphatidyl-myco-inositol-hexamannoside; AG, arabinogalactan; AGP, arabinogalactan-peptidoglycan complex; DAT, diacyltrehalose; DPG, diphosphatidylglycerol; GalNH₂, galactosamine residue; k, keto; LAM, lipoarabinomannan; LM, lipomannan; m, methoxy; MA, mycolic acids; MIM, mycobacterial inner membrane; MOM, mycobacterial outer membrane; PAT, polyacyltrehalose; PDIM, phthiocerol dimycocerosate; PE, phosphatidylethanolamine; PG, peptidoglycan; PI, phosphatidyl-myco-inositol; SGL, sulfoglycolipid.

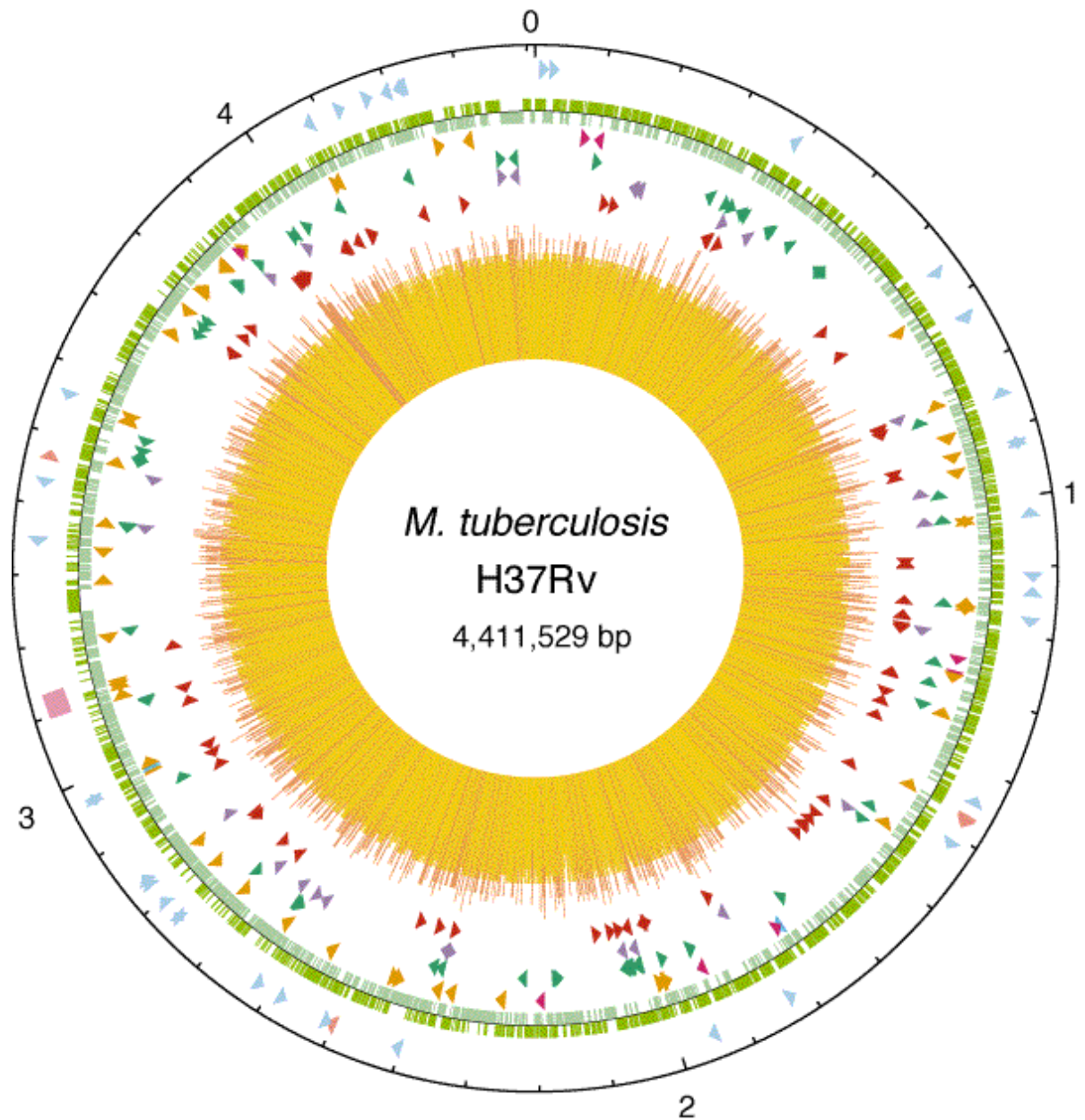


Figure 3. The complete genome of *Mycobacterium tuberculosis*

The complete genome of *Mycobacterium tuberculosis* was sequenced in 1998 [8]. A histogram in the centre of the ring shows GC content (<65% GC content in yellow, >65% in red). The next rings represent PGRS regions in dark red, the next PE family members in purple (excluding PGRS), the next the PPE family members in green, followed by regions of repetitive DNA (insertion sequences orange, 13E12 family members in dark pink and prophage in blue). The two green rings show the coding sequence by strand (clockwise dark green, anticlockwise light green). RNA genes (tRNA in blue, others in pink) and the direct repeat region (pink block) form the penultimate ring. The outer ring represents the scale in Mb (4.4 million bp long), where 0 is the origin of replication.

1.3 History and Expansion of Tuberculosis

The tubercle bacillus was discovered by Robert Koch in 1882 to be the causative agent of tuberculosis [9]. Tuberculosis is thought to have evolved with humans originating from the Horn of Africa [10]. The earliest paleo-pathological evidence of human tuberculosis was found in ancient Syria (8,800 – 7,600 BC) and a Neolithic settlement in the Eastern Mediterranean (7,000 BC) [11, 12]. Ancient MTBC DNA studies have been performed using varying methods, such as sequencing of insertion sequence IS6110, restriction fragment length polymorphism (RFLP), sequencing of the 65kDa heat-shock protein gene *hsp*, *rpoB*, and whole genome sequencing [13]. Ancient TB DNA studies have found MTBC in Egyptian mummies and in archaeological skeletons from the Bronze Age [14, 15]. Coalescent modelling of the whole genomes of MTBC, by contrast, point to its much earlier emergence approximately 70,000 years ago, from where it accompanied humans out of Africa, expanding along with human population density in the Neolithic period (Figure 4) [16]. Some strains have been found to expand along with great wars and mass human migrations [17]. Tuberculosis caused 20% of deaths between the 17th and 19th centuries [18]. Many famous historical figures such as Emily Bronte, George Orwell and Frederic Chopin died of consumption, or the ‘white death’ [19]. From ancient TB DNA studies of 18th century genomes from well-documented mummified bodies in Hungary, it seems that Lineage 4 may have featured widely, and that mixed infection seems to have been a feature of tuberculosis at the peak of its prevalence in Europe [12].

1.4 Global Epidemiology of TB

The World Health Organisation (WHO) estimated that the global prevalence rate of TB has fallen by 42% since 1990 [1]. This only represents the number of cases that were officially notified. The true number is thought to be much higher than this. Globally, the highest TB burden countries include India, Indonesia and China [1]. European prevalence of TB is a tale of two halves. Western Europe (e.g Germany, France, Spain, Scandinavia etc.) is considered a low prevalence setting. Eastern Europe, and especially countries of the former Soviet Union, is deemed high prevalence [20]. The collapse of communist public health systems in these former Soviet States is thought to have allowed the evolution and spread of multi-drug resistant TB (MDR-TB) [21]. There are seven global lineages of MTBC, that are broadly geographically based (Table 1.).

1.5 Epidemiology of TB in Ireland

Tuberculosis has been a statutorily notifiable disease in Ireland since the 1947 Health Act was written into legislation. Cases of TB in Ireland have declined from a high of approximately 7,000 per year in the early 1950s (230 cases per 100,000 population) to just 6.9 per 100,000 in 2015 [19, 22, 23]. However, TB incidence has plateaued at this level in recent years. Its prevalence remains

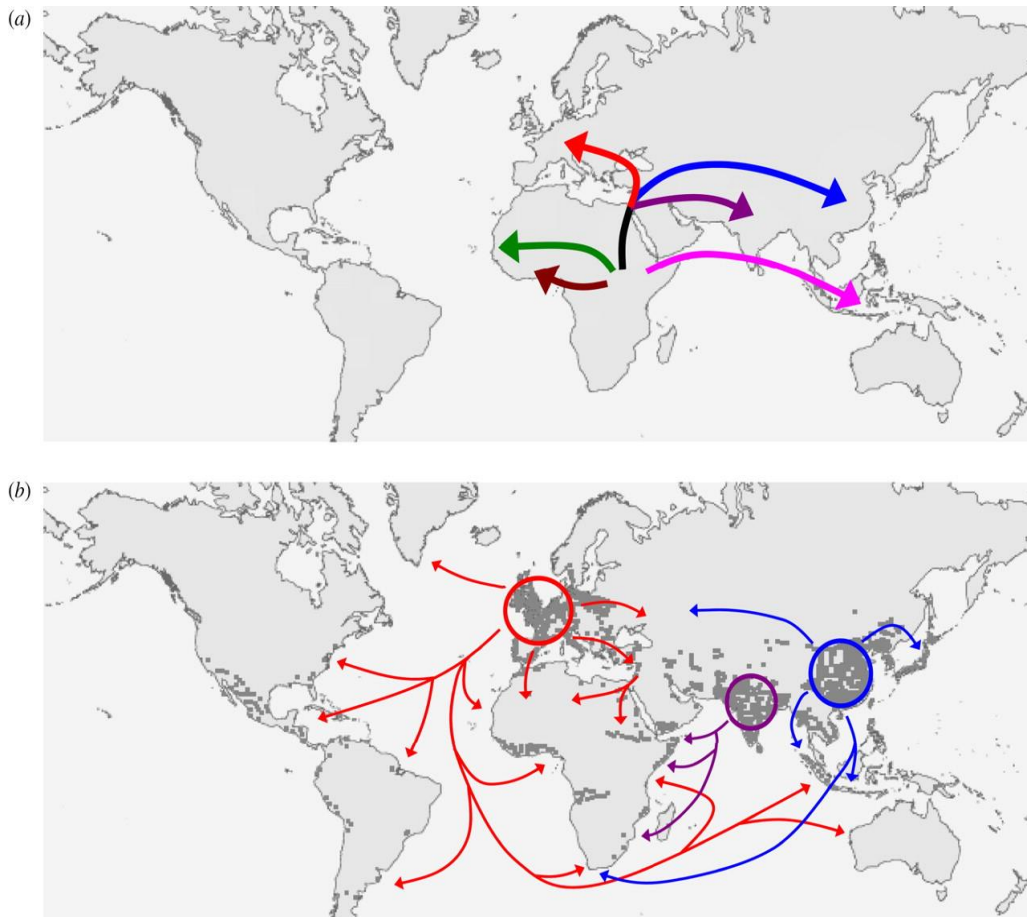


Figure 4. Proposed out-of-Africa expansion of human MTBC

(a) MTBC originated in Africa and some lineages accompanied the Out-of-Africa migrations of modern humans [249]

(b) The three evolutionarily 'modern' MTBC lineages seeded Europe, India and China, respectively, and expanded as a consequence of the increases in the human populations in these regions centuries ago. These lineages then spread throughout the world via exploration, trade and conquest.

Red lines correspond to Euro-American Lineage 4, purple to East African Indian Lineage 3, blue to East Asian Lineage 2, pink to Indo-Oceanic, green and brown to West African Lineages 5 and 6.

Lineage no.	Global Lineage	Sub-Lineage
1	Indo-Oceanic	EAI
2	East Asian	Beijing
3	East African Indian	Delhi-CAS
4	Euro-American	Haarlem, LAM, H37Rv, Cameroon, Ghana, S, TUR, X, Uganda I & II, New-1, URAL
5	West African-1	West African 1
6	West African-2	West African 2
-	Bovis	Bovis
7	Ethiopian	-

Table 1 Global lineages of *M.tuberculosis* Complex and their Sub-lineages

Broadly geographically based. LAM – Latin American Mediterranean, CAS – Central Asian Strain, EAI – East African Indian

high in urban centres. Forty-five percent of all TB cases in Ireland in 2015 were reported in urban centres [22]. This is thought to be due to deprivation associated with addiction and homelessness in these areas [23]. Ireland is a European island nation that has experienced mass emigration and immigration over the last century [24, 25]. Immigration of people from outside the European Union, as well as free movement policies within the European Union, has been associated with the spread of TB, especially multi- and extensively-drug resistant TB (MDR/XDR-TB) [26, 27].

1.6 Reasons TB has not been eradicated

Tuberculosis continues to be a threat to public health in both developing and industrialised countries. The reasons it has not been eradicated include the fact that latent tuberculosis provides a constant reservoir of infection. HIV/AIDS, poverty, poor public health systems and TB control programs, development of drug resistance, lack of access to treatment in some countries, and movement of people from areas of high to low prevalence are all plausible reasons why TB has not been controlled. Even higher-income, low-prevalence countries like Ireland struggle to bring the prevalence down to zero.

1.7 Pathogenesis of TB and latency

M. tuberculosis has no known reservoir outside of humans, although other members of the MTBC are zoonotic in nature [28]. Airborne transmission in aerosolised droplets makes it highly contagious. Prolonged close contact with an infectious individual, especially in confined, poorly-ventilated spaces is required to be infected [29]. It mainly affects the lungs, but can infect any organ [30]. Either latent or active infection can occur. MTBC is a skilled evader of the immune response, even though it does not seem to harbour classical bacterial virulence factors nor does it seem to gain them via horizontal gene transfer, except in the early stages of its evolution [31, 32].

1.7.1 Innate Immune Response

On infection, bacteria are recognised and phagocytosed by myeloid dendritic cells and alveolar macrophages in the lung via recognition molecular patterns (MAMPs) and pattern recognition receptors (PRRs) (such as toll-like receptors TLRs 2 and 4), which induces a pro-inflammatory response [33, 34]. MTBC can prevent phagosome-lysosome mediated killing and modulate the innate immune response in a number of ways [35, 36]. LAM scavenges Reactive Oxygen Intermediates (ROIs), and mycolic acids may help to resist hydrogen peroxide [37, 38]. MTBC can also enter the cytosol. This is thought to be dependent on two proteins secreted by the bacteria (Type VII secretion system) – early secretory antigenic target 6 (ESAT-6) and culture filtrate protein (CFP-10) [39, 40].

1.7.2 Adaptive Immune Response

The adaptive immune response is also activated by alveolar macrophages and other antigen-presenting cells (APCs), which drain into the lymphatic system, presenting TB antigens to CD4+ and CD8+ T cells, which proliferate in the presence of IL-12 and IFN γ to produce a Th-1 response (usually 2-6 weeks following infection with MTBC) [41, 42]. The Th-1 effector T cells migrate back to the lungs to react again with the infected macrophages, producing more IFN γ .

1.7.3 Granuloma formation

Cytokines and chemokines (such as RANTES, MIP1 α , MIP1 β , MCP-1, -3, -5 and IP10) are released which mediate the recruitment of inactivated monocytes, lymphocytes, natural killer cells, and neutrophils, which form a cellular mass, or granuloma, around the bacilli, which can be seen on chest radiographs as lesions suggestive of TB infection (Figures 5 and 6) [41, 43, 44]. Chemokine receptors, such as CCR-5 (a receptor for RANTES, MIP1 α and MIP1 β) also mediate granuloma formation [45]. The innate response forms the primary granuloma, which is then amplified by the adaptive response to form a solid granuloma (latency) or the liquified central area releases live tubercle bacilli into the lung to be expelled (active infection). The granuloma can be seen as both a host defence against the bacilli, and a cocoon where dormant bacilli can linger.

The granuloma itself consists of many different cells, including necrotic infected macrophages at the centre, surrounded by foam cells and giant cells, dendritic cells, epitheloid macrophages (both apoptotic infected and regular), macrophages (both apoptotic infected and regular), fibroblasts, neutrophils, natural killer cells and T and B cells, all surrounded by epithelial cells [46].

Latent tuberculosis infection (LTBI), although still not fully understood, is now defined as a dynamic model (or spectrum, depending on the host) of endogenous reactivation and damage response occur constantly within the 'healthy' host [47]. From latency, active infection can occur at any time, depending on the host cell-mediated immune response. TB tends to reactivate when the immune system is lowered for any reason (old age, HIV infection, immunosuppressive medication, smoking, lowered nutrition or social deprivation) [48]. It is estimated that 90% of people who are infected with latent TB will never reactivate. However 10% will, and most will within one year of contact [49]. Latency and reactivation are determined at the level of the granuloma. In some cases the granuloma is eliminated or mineralised, in some cases caseation and necrosis occur prior to reactivation [43]. Particularly because of latency, prevention is better than cure in the case of TB.

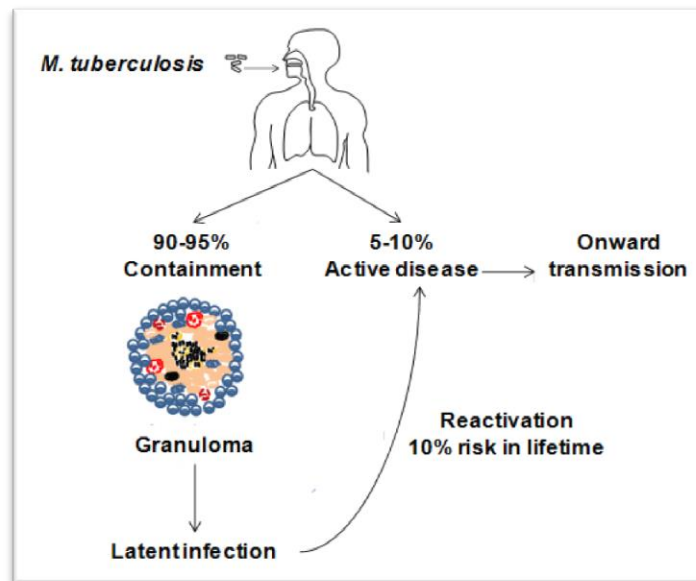


Figure 5. Latency and Reactivation

A figure depicting the inhalation of tubercle bacilli in aerosolised droplets, followed by granuloma formation and subsequent reactivation where active infection is manifest and could lead to further transmission (figure courtesy of bitesized.immunology.org)

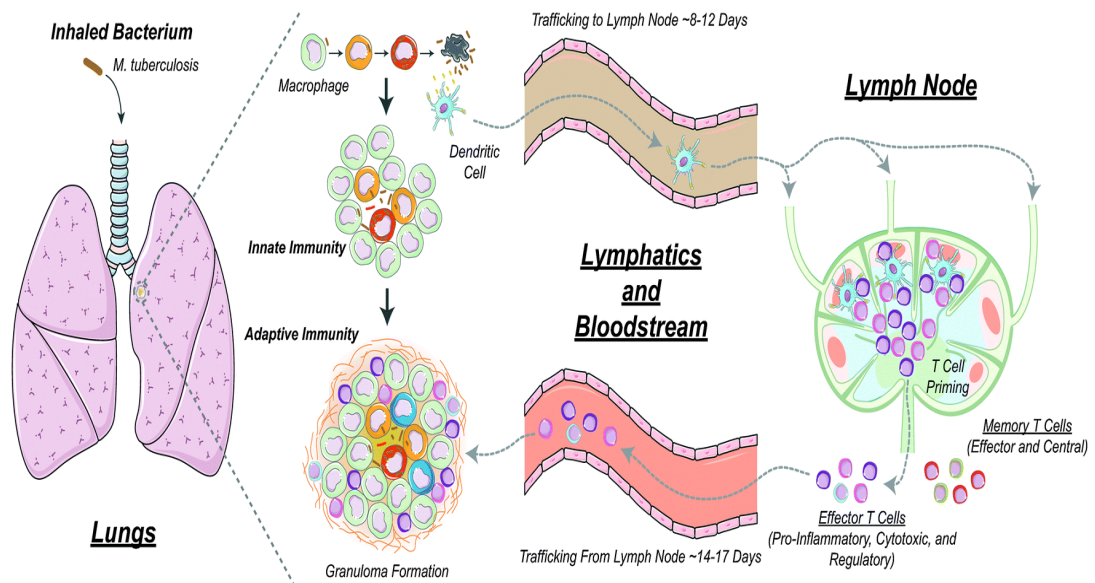


Figure 6. Granuloma formation by the tubercle bacillus

Overview of the immune response to *M. tuberculosis* infection [41]. It replicates within macrophages. Some bacteria are killed via the innate immune response. Dendritic cells present antigen to naive T cells in the lymph node, generating effector T cells (CD4+ and CD8+) that travel back to the site of infection to kill bacteria via the adaptive immune response. Granulomas are formed and strengthened by both types of immunity.

1.8 Symptoms and Clinical Diagnosis

Primary active respiratory TB symptoms include dramatic weight loss, drenching night sweats, haemoptysis, chronic cough, shortness of breath, and pyrexia of unknown origin, while cavitating lesions can be seen in the lungs on chest radiograph. Over 80% of TB cases occur within the lungs and approximately 19% are extra-pulmonary in nature [50]. Primary MTBC infection is defined as the first exposure to TB, followed by either latency or active infection. Following initial infection and treatment, patients may go on to reactivate old lesions, or be re-infected with a new strain of MTBC.

1.8.1 Tuberculin Skin Test and Interferon Gamma Release Assays

If an individual is suspected of TB exposure, they will be tested for LTBI. The Mantoux test, or Tuberculin Skin Test (TST) was developed by Robert Koch in 1890, and made safe for use by Florence Seibert in the 1930s. In principle, the inoculated purified protein derivative (PPD) of tuberculin causes the release of Interferon-gamma ($IFN\gamma$), Interleukin-8 (IL-8) and Tumour Necrosis Factor ($TNF\alpha$) if the individual has been previously exposed to TB. In practice, the TST can be confounded by different factors, such as the ‘boosting effect’ thought to be caused by serial testing, and must be interpreted with clinical caution. Interferon-gamma release assays (e.g. Quantiferon-TB Gold and T-Spot TB) are seen as a more sensitive alternative, however these too are fraught with interpretation issues [51]. They work on the principle that a person infected with TB will have circulating T-cells that have been exposed to, and sensitised by, antigens specific to TB (ESAT-6/CFP-10). If T-cells encounter TB again, they produce $IFN\gamma$, which can be measured in the patient’s blood. Both tests have been associated with difficulties in interpretation and reproducibility, especially in risk groups such as those with HIV, although they have been considered useful for confirming a positive result in one another (i.e. TST confirms IGRA and *vice versa*) [51, 52].

1.9 TB Treatment and Vaccination

Without treatment, the death rate would be much higher; about 70% of those with active disease would die within 10 years [53]. Combination therapy is essential for TB therapy in order to attack both dormant and active organisms. Anti-tuberculous drugs were first shown to be effective in the 1940s [1]. Treatment includes first-, second- and increasingly third-line drugs, as outlined in the table below.

TB therapy	Combinations of drugs included in current regimens	Use and duration
First line drugs	isoniazid, rifampicin, ethambutol, pyrazinamide	Mainly used in drug susceptible cases, 6-9 months
Second line drugs	Fluoroquinolones (ciprofloxacin, levofloxacin, moxifloxacin), aminoglycosides (kanamycin and amikacin - injectable, others include streptomycin) cycloserine, polypeptides (capreomycin), thioamides (ethionamide)	Used in drug resistant cases, up to 24 months
Third line drugs	linezolid, clarithromycin	Used in more extreme drug resistant cases in conjunction with 1 st and 2 nd line drugs

Between 2000 and 2014, TB treatment saved 35 million lives among HIV-negative people [1].

Isoniazid inhibits the synthesis of mycolic acids in the cell wall by activation of *katG* catalase-peroxidase. Rifampicin inhibits bacterial DNA-dependent RNA synthesis. Pyrazinamide acts as a prodrug. The MTBC enzyme pyrazinamidase converts it to its active form, pyrazinoic acid, which stops mycobacterial growth. Ethambutol inhibits arabinosyl transferase which is involved in cell wall synthesis. Streptomycin inhibits protein synthesis. The newest drugs available are bedaquiline and delamanid. Bedaquiline blocks ATP synthase which is used by MTBC to generate energy. Delamanid is a nitroimidazole.

The Bacille-Calmette Guerin (BCG) vaccine, originally developed in the early 1900s, is the only TB vaccine available despite many attempts to discover new vaccines. It is thought to be effective in preventing TB meningitis in the first year of life but its efficacy is debated over the course of a lifetime. Its efficacy is thought to be affected by prior exposure to MTBC infection or sensitisation with environmental mycobacteria [54]. The BCG vaccine is administered universally to children in Ireland, although the necessity to do this is currently under review (personal communication, Dr. Mary O'Meara, Public Health).

1.10 Drug Resistance

Genetic drug resistance (mono-, multi- and extreme-drug resistance) in MTBC is thought to be mainly due to the accumulation of unlinked chromosomal mutations which occur spontaneously at low frequency in various genes, and not due to the acquisition of mobile genetic elements such as

plasmids or transposons, as occurs in most other bacterial species [32]. Other phenotypic mechanisms of resistance include efflux pumps, which have been studied more in recent times [55]. MTBC contains a relatively large number of putative drug efflux pumps [55]. Efflux pump inhibitor drugs have even been found to improve anti-tubercular drug therapy when used in conjunction with traditional regimens [56]. Therapy with three first-line anti-tuberculous drugs was thought to overcome any possible resistance [57]. However, functional ‘mono-therapy’ due to irregular drug concentrations *in vivo* (caused by various host factors), inappropriate prescribing of drugs and patient non-compliance have amplified these mutations and given rise to multi-drug resistance [58]. The resistant isolate can then be transmitted to the next patient (acquired resistance) [59].

MDR TB is defined as resistance to at least isoniazid and rifampicin. XDR TB is defined as additional resistance to fluoroquinolones and one of the second-line injectable drugs. The first case of totally drug resistant (TDR) TB was reported in Italy in 2007 [60]. Resistant cases could take up to two years to treat, with complicated and extremely expensive therapeutic regimens that, in many cases, still fail. MDR and XDR TB represent a serious challenge for tuberculosis control in countries of the former Soviet Union, Asia and Africa. Survival of patients is significantly reduced by drug resistance [59].

1.10.1 Genes involved in resistance

The main genes (along with their promoter regions, 100bp upstream of each gene) that have been found to be involved in resistance are *rpoB*, *katG*, *gyrA*, *inhA*, *rrs*, and *embB* [61]. The gene *rpoB* encodes a DNA-directed RNA polymerase and is believed to be involved in rifampicin resistance. It inhibits RNA synthesis [62]. The gene *katG* encodes catalase-peroxidase-peroxynitrate and is associated with isoniazid resistance. Another isoniazid (low level) resistance-associated gene, *inhA*, encodes meromycolic acid enoyl reductase and works with *fabG1* encoded meromycolic acid 3-oxo-acyl reductase to inhibit mycolic acid synthesis [62]. A 16S ribosomal RNA is encoded by *rrs*, which is believed to be involved in the inhibition of mycobacterial protein synthesis in aminoglycoside resistance [62]. The gene *gyrA* encodes DNA gyrase sub-unit A and is thought to be associated with fluoroquinolone resistance, where it inhibits DNA synthesis. Finally, the *embB* gene encodes arabinogalactan arabinosyl transferase and has been associated with ethambutol resistance, where it inhibits mycobacterial arabinogalactan synthesis [62]. Of course, there are many more genes that have also been associated with drug resistance in MTBC, but to varying extents. Even the most newly discovered drugs, bedaquiline and delamanid, mutations in genes such as *ddn* (nitroreductase) and *mmpL3* (membrane transporter) have been found to cause resistance by inhibition of mycolic acid synthesis and production of reactive nitrogen species [62].

1.11 Irish Mycobacteria Reference Laboratory (IMRL)

The IMRL was established in 2001 and is an ISO15189 accredited national reference laboratory that processes approximately 6,000 specimens and 400 cultures per year for 18 external users, and plays a central role as part of a multi-disciplinary tuberculosis care team at St. James's Hospital, Dublin. Services offered include auramine and Ziehl Neelsen (ZN) microscopy, mycobacterial culture and MTBC drug susceptibility testing (DST), molecular testing for identification of, and drug resistance detection in, mycobacteria species (both rapid and conventional) and a national genotyping service (Figure 7).

1.11.1 Microscopy and Morphology

Mycobacterial bacilli are acid-fast, which can be exploited by different staining methods for microscopic visualisation. Auramine utilises fluorescence to visualise fluorescing bacilli against a dark background. Ziehl Neelsen (ZN) microscopy stains acid-fast bacilli pink with carbol fuchsin while malachite green counterstains the background (Figure 8). A positive smear microscopy signifies that a patient is infectious and should be isolated to prevent further transmission. A negative smear microscopy is required to enable them to leave isolation while on TB treatment.

1.11.2 Culture and Drug Susceptibility Testing

The BACTEC[®] MGIT[®] 960 instrument is a fully automated system that exploits the fluorescence of an oxygen sensor to detect growth of mycobacteria in culture. A ruthenium pentahydrate oxygen sensor embedded in silicon at the bottom of a tube containing 7 ml of modified Middlebrook 7H9 broth fluoresces following the oxygen reduction induced by aerobically metabolizing bacteria within the medium. Both detection and DST can be performed using this liquid culture-based method [63, 64]. Cultures are incubated at 35°C for 42 days in order to deem them negative for the presence of mycobacteria.

1.11.3 Rapid Molecular Tests

Cepheid GeneXpert MTB/RIF can detect MTBC and rifampicin resistance-associated mutations, using real-time PCR, within 2 hours of sample receipt, directly on a respiratory specimen [65]. BD MGIT TBc test is a rapid immune-chromatographic assay that detects the MPT64 antigen in growing MTBC cultures, and takes 15 min to complete. Line probe assays are used for rapid genotypic resistance detection of mutations associated with rifampicin, isoniazid, ethambutol, fluoroquinolones and aminoglycosides [66]. Line probe assays use reverse hybridisation of dedicated DNA strips impregnated with probes specific for both wild-type and drug-resistance-associated mutations to detect drug resistance within two days of culture positivity.

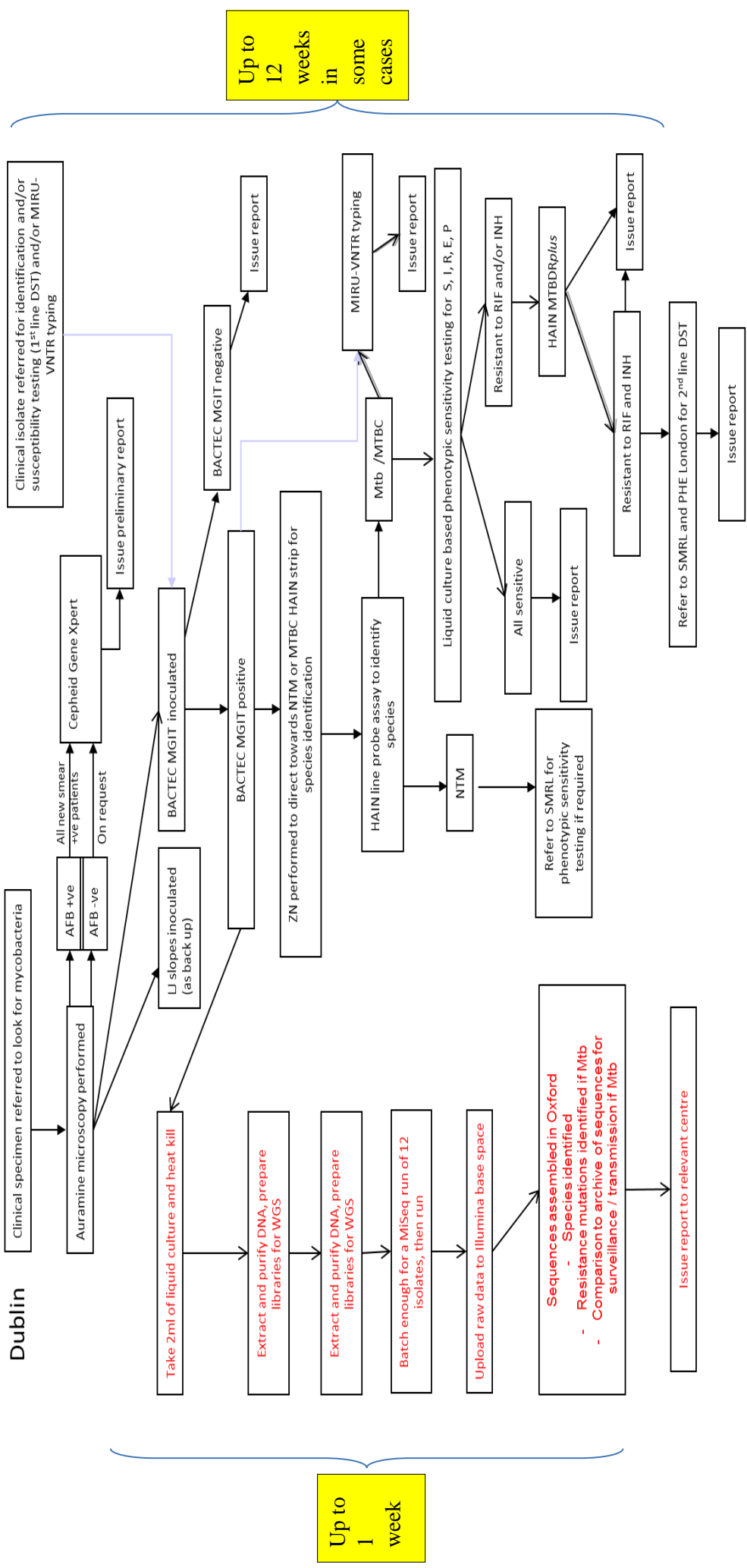


Figure 7. IMRL Algorithm for Culture and DST of Mycobacteria species versus Algorithm proposed by the MGIT Pilot Study

Algorithm currently in place in the IMRL for clinical specimens referred for mycobacterial culture (in black). Alternative algorithm in place for the collaborative MGIT Pilot Study in which the IMRL participated (in red). General time-frames are included in yellow. The proposed whole genome sequencing algorithm could save a significant amount of time for the diagnostic laboratory.

AFB - Acid Fast Bacilli, WGS - whole genome sequencing, *M.tb*/MTBC - *Mycobacterium tuberculosis* Complex, DST – Drug susceptibility testing, MIRU-VNTR - mycobacterial interspersed repetitive units variable number tandem repeats, S - streptomycin, I/INH - isoniazid, R/RIF - rifampicin, E - ethambutol, P - pyrazinamide, MGIT – mycobacteria growth indicator tube, SMRL – Scottish Mycobacteria Reference Laboratory, PHE – Public Health England, LJ – Lowenstein Jensen, NTM – non-tuberculous mycobacteria.

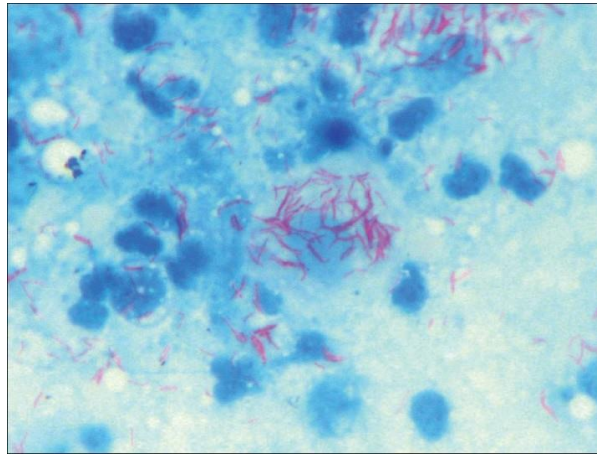


Figure 8. Image of a Ziehl Neelsen Stain of Mycobacteria, taken at the IMRL

'Acid-fast' bacilli stain pink due to their uptake of carbol fuchsin, acid-alcohol fails to remove the carbol fuchsin through cell walls that are impermeable to acid. Smears are counter-stained with malachite green.

1.11.4 MIRU-VNTR Genotyping

Variable Number Tandem Repeats (VNTRs) of micro- or mini-satellite DNA, also called Mycobacterial Interspersed Repetitive Units (MIRUs), have emerged as valuable markers for genotyping of MTBC (Figure 9). It has been shown that MIRU-VNTR typing can provide unique molecular insights into the population structure of the MTBC and provides clear criteria for the identification of the different MTBC lineages and sub-lineages [67]. It employs PCR amplification using primers specific for the flanking regions of the VNTRs and determines the sizes of the amplicons using electrophoretic migration. As the length of the repeat units is known, these sizes reflect the numbers of the amplified VNTR copies. The final result is a numerical code, corresponding to the repeat number in each VNTR locus. Such numerical genotypes are intrinsically portable and are thus particularly convenient for both intra- and inter-laboratory comparative studies. An optimised set of 24 MIRU-VNTR markers was proposed by an international consortium including ten European and American laboratories in 2006 [68]. The MIRU-VNTR*plus* database is a web-tool available for analysing sequences, naming strains and phylogenetic studies [69].

1.12 Bio-informatic Analysis Software

Computational molecular evolution and whole genome sequencing (WGS) analysis programs have been developed, and are designed to be used, with operating systems other than Windows. This can present a challenge to those unfamiliar with these operating systems.

1.12.1 Linux Operating System and the Command Line

Linux is an open-source operating system kernel, designed by Linus Torvalds, released under a GNU General Public License. It was originally developed by computer programmers for computer programmers. There is a culture of sharing programs freely in the programming world, where a program will be shared and improved by many for the greater good. The Linux Command Line is where the user interacts with the computer directly, instead of a graphical user interface (GUI) being used as a translator between the user and computer. The operator must learn to work in different programming languages in order to be able to install programs, run programs, save, copy and edit files. While this can be challenging to the new user, it opens up a world of possibility for working with large data-sets like whole genome sequencing data. There is a wide-ranging knowledgeable online community that can be accessed through technical forums for help and information if required.

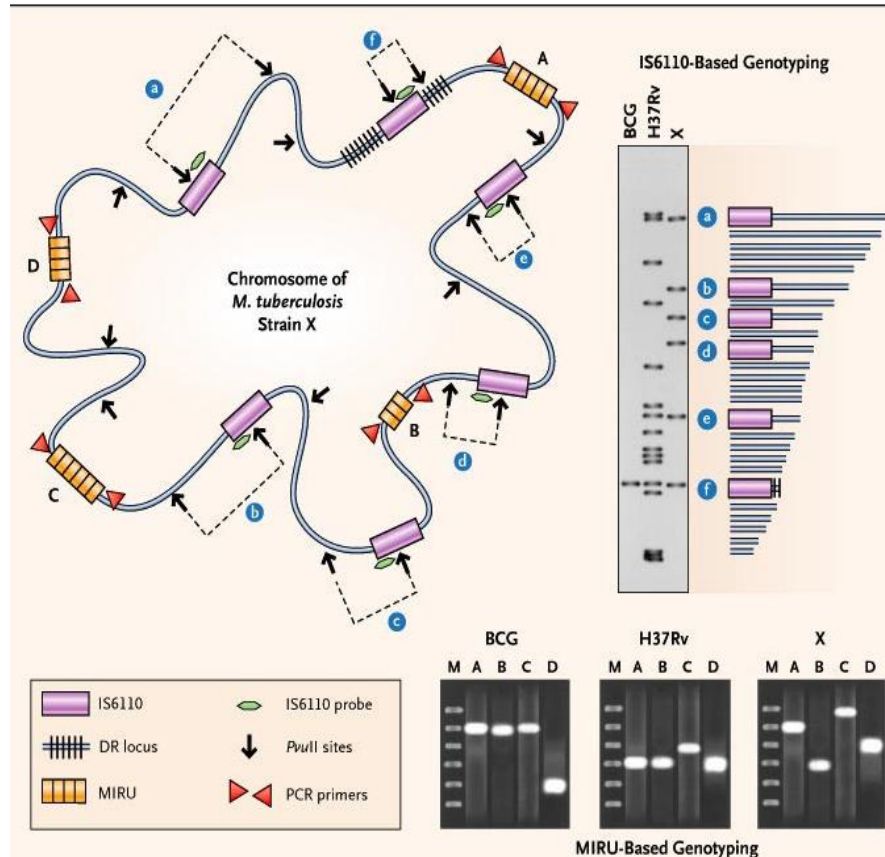


Figure 9. Visual representation of the principle of MIRU-VNTR genotyping

This figure shows the different genotyping methods available for MTBC: IS6110 RFLP analysis, spoligotyping and MIRU-VNTR genotyping [250]

IS6110 IS6110 is an insertion sequence unique to MTBC that has been used for restriction fragment length polymorphism (RFLP) analysis. Specific lineages display unique RFLP patterns. PvuII is the restriction enzyme used.

DR locus Spoligotyping is based on the variability of 43 spacers in the DR (direct repeat) locus.

MIRU-VNTR MIRU-VNTR genotyping is based on the different numbers of repeats of microsatellite DNA at informative loci across the MTBC genome that are amplified using PCR. Each of the 24 fragments is analysed, a number is assigned to each locus according to the number of repeats found, and the result is a 24-digit portable genotype.

1.12.2 Open-source Bio-informatics software

Examples of open-source software include mapping programs like Burrows Wheeler Aligner (BWA) or Bowtie [70]. An open-source variant analysis package called SAMtools is used to analyse variants in WGS data [71]. Many phylogenetic packages are open-source, e.g. PhyML, Seaview [72, 73]. Galaxy.org is a freely available web-based tool for analysing WGS data that incorporates many of the tools used for biological data analysis, and is accessible from a Windows environment [74]. Biolinux 8 is a Linux operating system (OS) which has many of these biological software programs already installed for convenience [75]. TB Profiler and PhyResSe are SaaS (software as a service) online web-tools that are designed to perform the above analysis for the user for MTBC genomes, using raw input data files [76, 77].

1.12.3 Commercial Bio-informatics software

Geneious R9 is an example of commercially available user-friendly software that can be run on Windows OS. It will perform read mapping, assembly, neighbour-joining phylogenetics on consensus sequences mapped to reference genomes, visualisation of the resulting tree and distance matrices, and can also be used as an integrated genome viewer [78]. DNASTAR Lasergene and Bionumerics are two other commercial platforms.

1.13 Phylogenetic Analysis

Despite opinions to the contrary, evidence suggests that every organism alive today, and all those that lived before them, have a shared heritage that dates back to the origin of life about 3.8 billion years ago [79]. Charles Darwin sketched his first evolutionary tree in 1837, defining evolutionary biology that is still used today. Phylogenetics is defined as ‘the evaluation of hypotheses about historical patterns of evolutionary descent in the form of evolutionary trees’ [79]. ‘Phyle’ means tribe, and ‘genesis’ means birth. It is a guessing game. One can only construct a phylogenetic tree from what is known, however, there could be a vast amount of unknown information missing from that tree. This must always be taken into account.

For the purposes of the current study, a phylogenetic tree is a mathematical representation of the evolutionary relationships that exist between a set of DNA sequences that may have diverged over a short period of time, compared to the billions of years the data with which some phylogenetic trees are constructed. Complex mathematical algorithms are used for construction. Methods are broadly distance-based or character-based. Character-based methods can be Bayesian or Maximum Likelihood. Bayesian methods are generally used for constructing phylogenies over much longer periods of evolutionary time and where much of the ancient ancestry may not be known, for instance in plant and animal computational molecular evolution. Although Bayesian methods are

becoming more commonly used for infectious disease molecular epidemiology, for the purposes of the current study, maximum likelihood methods were sufficient since it is the method most commonly used in the literature to date. Constructing a phylogenetic tree begins with an alignment of sequences. The shape of the tree, and therefore the pattern of branching that it hypothesises are known as ‘topology’. An analogous group is one that includes an ancestor and all of its descendants, and is called a ‘clade’. A phylogram visualises branch lengths that correspond with the distance (in evolutionary time) between two species. A cladogram transforms branch lengths to represent the clustering of species more readily, and cannot be relied on for any more than branching pattern [79]. Bootstrapping is a statistical method whereby reliability of branching on a phylogenetic tree is tested by replicating the construction of the tree 100 times, or as many times as is required, depending on time available.

1.13.1 Multiple Sequence Alignment

Sequence alignment is a hypothesis about homology of sequences. Phylogenetic trees rely on the differences between the aligned sequences. Alignments can be challenging to achieve with confidence since there may be repeats in the sequence, insertions or deletions in some but not all species, gaps where no sequencing data is present for some species and not others. This becomes even more challenging when whole genomes of millions of base pairs are being aligned. It can be sub-divided into pair-wise, global, and local alignment. Pair-wise is the simplest and least computationally demanding. There are numerous alignment theories, algorithms and software programs from which to choose. One must choose the most appropriate alignment method for the task at hand based on the computational resources available [80].

1.13.2 Neighbour Joining Tree

An example of a distance based method would be a neighbour-joining tree, which incorporates a step-wise clustering method for its construction. It joins the most closely-related sequences first, therefore minimising branch length, and is regarded as the most representative of true evolution. Neighbour-joining is the most commonly used distance-based method and is considered a fast and useful preliminary analysis on which to base further analysis [81].

1.13.3 Maximum Likelihood Tree

Maximum-likelihood is a character-based method that calculates the probability of expecting each possible nucleotide (character) at every node for each sequence, when given a data-set and a model of evolution. The tree with the highest probability is chosen. Despite being computationally demanding, maximum-likelihood methods are useful as they are statistically-based and, therefore, comparisons of different trees, parameters, and models can be tested [81].

1.13.4 Models of Evolution

Several models of evolution have been proposed to infer genetic distances between sequences. At least 56 models have been developed. Each model uses different parameters for the frequency of each base, and the probability of each base substitution. For instance, the Jukes-Cantor model assumes all four bases (A, C, T, G) and all types of substitution occur at equal frequency [81]. Another model, GTR (Generalised Time Reversible) model, was described in 1986 by Simon Tavaré. Time reversibility assumes that, although two states may occur with different frequencies, the amount of change from state x to y is equal to the amount of change from y to x .

1.13.5 UPGMA tree

The Unweighted Pair Group Method with Arithmetic Mean tree (UPGMA) is a simple hierarchical clustering method [82]. It constructs a rooted tree that reflects the structure present in a pairwise similarity matrix. At each step, two clusters are combined into a higher-level cluster, and so on, until the tree is constructed. This method results in a cladogram, and is used to visualise MIRU-VNTR genotyping clusters along with a neighbour joining algorithm.

1.13.6 Minimum Spanning Tree

Minimum spanning trees are frequently used in molecular epidemiology research to estimate relationships among individual strains or isolates. They represent a set of edges (or connections) between nodes (or individuals) that are connected to each other by the shortest distance possible. Rather than a definitive representation of epidemiological direction of transmission, these trees are a highly visual impression of the bigger picture of how an outbreak might be progressing [83].

1.14 Whole Genome Sequencing (WGS)

Whole-genome sequencing has been found to trace the spread of a specific strain more precisely than other typing methods, and solve cryptic transmission chains [17]. It offers the highest available level of resolution of pathogen genomes and is, therefore, considered the ultimate genotyping tool.

1.14.1 Progression from Sanger to Next Generation Sequencing

In 1953, the structure of DNA was discovered by Watson, Crick and Rosalind Franklin [84]. In 1973, the first DNA sequence was published on the *lac* operon by Gilbert and Maxam [85]. Fred Sanger published his automated sequencing method in 1977, which led the way for a sequencing revolution [86]. The entire MTBC genome was first elucidated in 1998 (Figure 3) [8], and the entire human genome was sequenced in 2001 [87]. Sanger sequencing was advanced and

manipulated over the years. Massively parallel sequencing by synthesis became the game-changer with regard to whole genome sequencing, making it accessible to many more scientists worldwide for many different applications [88].

1.14.2 Bridge Amplification

Short-read massively-parallel NGS was made possible by the invention of bridge amplification and sequencing-by-synthesis technology (Figure 10). Sequencing-by-synthesis happens on a flow-cell. A flow-cell is a glass slide lawned with two types of oligonucleotide, one of which is complementary to the adaptor on the previously-prepared DNA library of fragments. The library fragments of random size are washed over the flow-cell where they can attach via the above adaptors. DNA polymerase makes a complementary strand. Denaturation follows and the original strand is washed away. Bridge amplification happens next. This is where the strands fold over and hybridise to the flow-cell with their second adaptor oligonucleotide. A complementary strand is formed by DNA polymerases. The now double-strand is denatured, resulting in two single copies that are tethered to the flow cell (forward and reverse strands). This process is repeated many times, resulting in many clusters of DNA strands attached in various spots across the glass flow-cell. Reverse strands are cleaved and washed away, leaving only the forward strands. In order to prevent unwanted priming, 3'-ends are blocked and sequencing-by-synthesis can begin.

1.14.3 Sequencing by Synthesis

Sequencing-by-synthesis begins with attachment of the first sequencing primer to produce the first NGS read (Read 1). Each cycle, the flow-cell is washed with a mixture of nucleotides that compete for the next spot on the complementary strand. Only one nucleotide is added per cycle. The number of cycles determines the length of the read. Fifty, 300, 500 and 600 cycle kits are currently available. When the nucleotide is added it fluoresces at a characteristic wavelength, depending on whether it is an adenine (A), cytosine (C), thymine (T) or guanine (G). Then the nucleotide mix is washed away, another cycle begins, and so on. For a given cluster on the flow-cell, all added nucleotides are excited simultaneously.

After completion of the first cycle read, read products are washed away. An Index 1 read primer is hybridised to the template DNA strand and this goes through the same steps as for Read 1. Following cleavage, the 3'-end of the strand is un-blocked, and an Index read 2 primer is attached and reads the remaining multiplexing information from the strand during another round similar to that of Read 1. Neither of these cycles is included in the results but is necessary to associate that particular read back to its original sample within a multiplexed reaction. Following this, bridge amplification creates a complementary strand to the current template. The original strand is cleaved

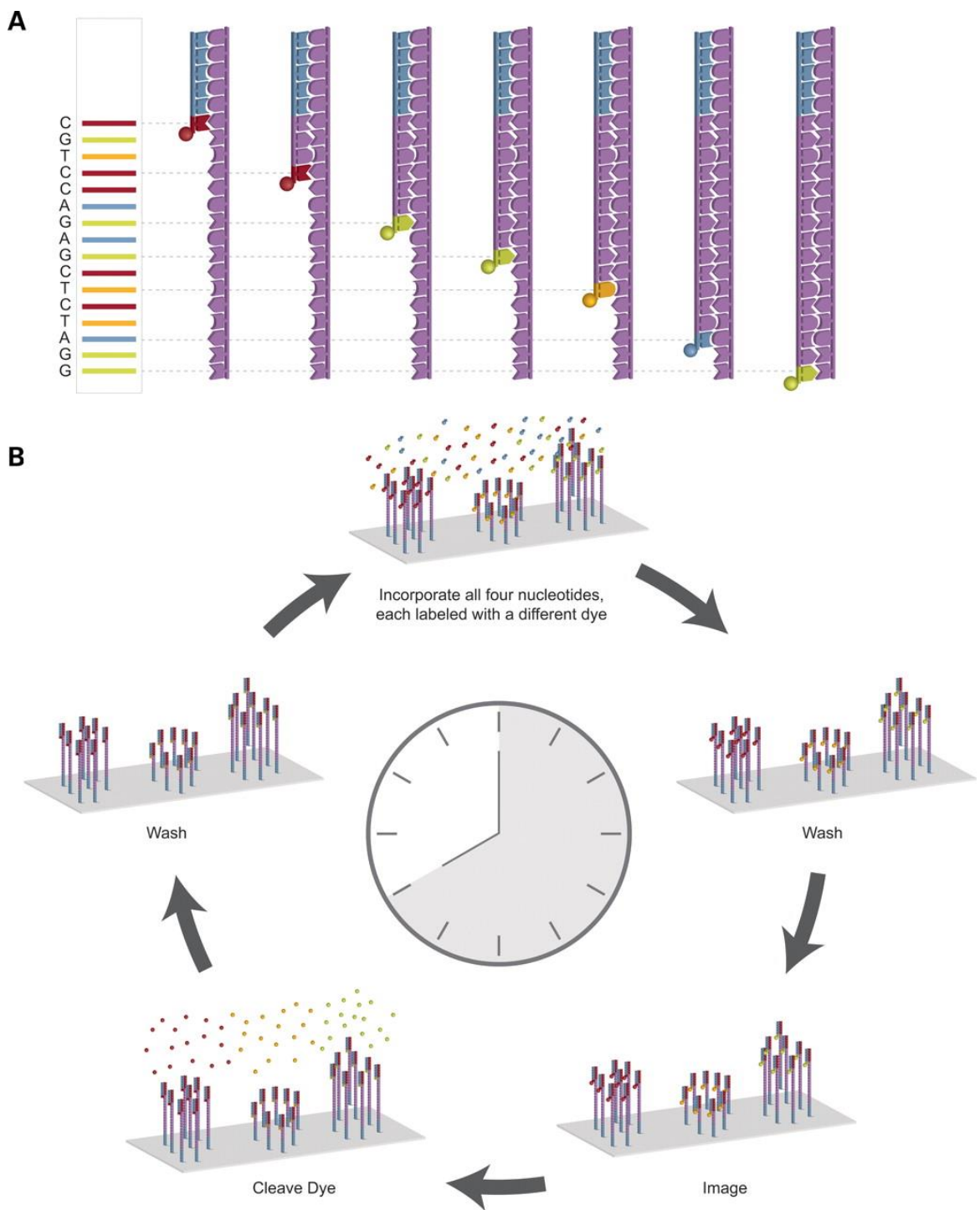


Figure 10. Next Generation Sequencing-by-Synthesis

(a) Shows how sequencing by synthesis is recorded using fluorescent snapshots of all clusters on a flow cell, followed by washing and initiation of the next cycle. (b) shows the cycle as it progresses on the flow cell. (Figure courtesy of Illumina)

and washed away, leaving the original sequence of the reverse strand. Sequencing-by-synthesis is repeated as above, leaving millions of reads within clusters that represent all random library fragments. Following completion of cycles, the resulting reads are de-multiplexed back to their original samples and exported as fastq files, which are sequencing read files in a format that is very similar to fasta format but also includes the sequencing quality of the reads.

1.15 Impetus for this PhD Research

Drug resistance in MTBC is not a new phenomenon. Soon after the introduction of streptomycin in 1944, streptomycin-resistant strains of MTBC began to emerge. Available drugs are old by now (e.g. isoniazid and pyrazinamide were discovered in 1952, ethambutol in 1961) [89]. The only vaccine on the market (BCG) was developed in the early 1900s. Because MTBC is a slow-growing fastidious organism, phenotypic susceptibility results take a relatively long time to elucidate (min. 7-10 days). New drugs, vaccines and rapid molecular tests are urgently needed [89]. Understanding the molecular mechanisms and nuances of an organism will eventually help in controlling its spread [32]. Molecular characterisation of MTBC plays an important role in understanding the mechanisms of its origin and transmission in a population, as well as how it develops drug resistance. It may also play a role in the development of prophylaxis (e.g. vaccine targets), or prevention of transmission of the disease (e.g. drug targets). Understanding the mechanisms of resistance could also enable development of rapid commercial molecular diagnostic tools [90, 91]. To date, genotyping of tuberculosis has manifested seven major global lineages (Figure 11). Many countries have published data on MTBC molecular epidemiology, helping to build a picture of TB evolution worldwide [92-104]. Ireland has two publications to date that give a preliminary insight into its molecular epidemiology [105, 106]. However, a more comprehensive nationwide survey of MTBC would be useful in order to discover national clustering patterns and drug resistance. WGS could reveal even more about the underlying mutations leading to resistance in MDR/XDR-TB cohort found in Ireland, as well as further information on the patterns of outbreaks that are occurring among us.

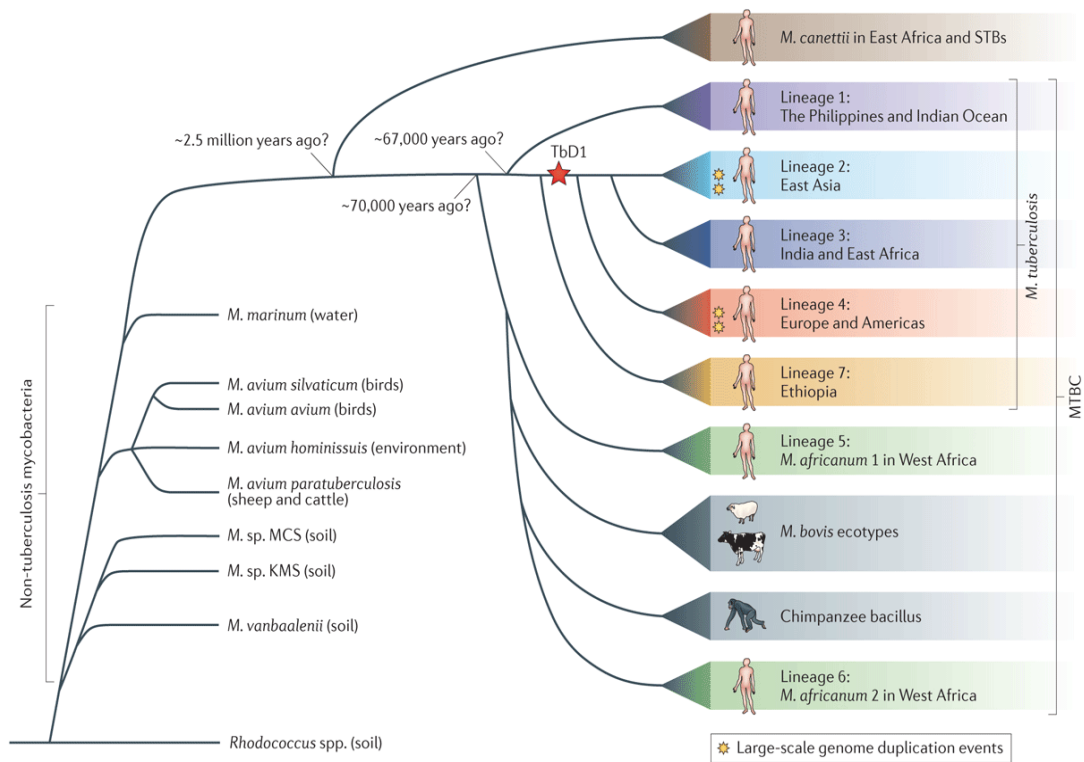


Figure 11. Proposed Evolution of the seven Global Lineages of modern MTBC strains

There are 7 global lineages which are roughly geographic in origin [124]. TbD1 indicates the deletion event specific for *M. tuberculosis* lineages 2, 3 and 4. Evolutionary distances are not to scale.

1.16 Hypotheses and Aims of the Study

The overall translational aim of this project was to improve the quality of the IMRL service through research and collaboration (with Trinity College and others) that could directly benefit patients in the long-term. Hypotheses are in italics below.

The primary aim was to examine five years of MTBC MIRU-VNTR genotyping data collected at the IMRL from 2010-14 in order to assess the distribution of lineages present and to build on work already published in this regard.

Prospective MIRU-VNTR genotyping is essential for primary surveillance of MTBC in Ireland and flagging of national and local outbreaks

The benefit of MIRU-VNTR genotyping increases over time, ie, outbreaks are more readily identified when there is a prospective genotyping method in place

From the molecular epidemiology results extracted for the period 2010-14, it was clear that clusters of TB were present in Ireland, and that this was worth further investigation with a higher resolution, i.e. whole genome sequencing.

Whole genome sequencing improves upon the MIRU-VNTR genotyping method currently in place in the IMRL, further delineating clusters of identical MIRU-VNTR genotypes, especially where epidemiological information is difficult to collect

MIRU-VNTR genotyping over-clusters isolates in certain cases, which could be misleading, were it to be relied on completely

Public Health surveillance, identification of super-spreaders, and contact tracing, would be greatly augmented by the use and interpretation of whole genome sequencing data in the IMRL

The MIRU-VNTR genotyping results also revealed that MDR/XDR-TB had increased over the period of the study. A comprehensive survey of MDR/XDR-TB (collected 2001-14) was consequently undertaken in order to characterise the strains circulating in Ireland, and to assess the drug-resistance-associated mutations they harboured using whole genome sequencing compared to phenotypic susceptibility testing.

MDR/XDR-TB is not being readily transmitted in the Irish-born population

MDR/XDR-TB is not being spread from the immigrant population to the Irish-born population

The drug-resistant MTBC strains present in Ireland are identical to those currently circulating in Europe

*Whole genome sequencing can accurately predict the phenotypic resistance profile of *M. tuberculosis* in a cohort of multi-drug resistant strains found in Ireland*

WGS would improve MDR/XDR-TB detection and accelerate drug resistance prediction if it was introduced in the diagnostic laboratory

Collaborative analysis of whole genome clusters was performed in Oxford with the Modernising Medical Microbiology Group whose translational aim was to establish a routine diagnostic technique for rapid mycobacterial species identification, epidemiological typing and drug susceptibility testing of TB in a routine diagnostic laboratory. The IMRL was invited to collaborate in this study with the aim of introducing WGS techniques, which had been validated on an international scale, to the IMRL.

Whole genome sequencing could replace traditional methods of identification, susceptibility testing and genotyping in the diagnostic laboratory, due to its more rapid turnaround times, accuracy, and decreasing costs

This is the most comprehensive work done to date on MIRU-VNTR Genotyping of MTBC in Ireland, and the first documented work using WGS on MTBC strains found in Ireland.

Ethics

Ethics approval was received for this project from the St. James' Hospital/Adelaide and Meath Hospital, incorporating the National Children's Hospital, Dublin (SJH/AMNCH) Research Ethics Committee (2013/05/01).

Chapter 2.

Materials and Methods

2 Materials and Methods

Unless otherwise stated, all of experimental procedures and analysis carried out as part of this study was performed by Emma Roycroft.

2.1 Study Ethics, Safety, and Storage of Isolates

Ethical approval was received for this study from the SJH/AMNCH Ethics committee; number 2013/05/01.

Thorough Personnel and Project Risk Assessments were carried out prior to commencement of the study.

The XDR-TB WGS project has been deposited in the European Nucleotide Archive (ENA) under the accession no. CCJS01000000, as will the remainder of isolates in this project. Reads from the MGIT Pilot Study were deposited in the National Centre for Biotechnology Information (NCBI) Short Read Archive (SRA) (Bio-Project PRJNA268101 and PRJNA302362).

2.2 Working in a Containment Level 3 (CL3) facility with a Category 3 Pathogen

All cultures were initially processed at the IMRL CL3 laboratory, since *M. tuberculosis* is a Category 3 pathogen. Strict safety protocols were adhered to when processing cultures in the CL3 laboratory. Appropriate personal protective equipment (PPE) was worn at all times, and training and competency were maintained over the course of the study. Only when the cultures were heat-inactivated could they be removed for further processing (see section 2.8 Heat Inactivation of Isolates).

2.3 Storage of Mycobacterial Isolates at the IMRL

The IMRL retains all TB culture isolates in duplicate in 7H9 Middlebrook medium (BD, New Jersey, USA) at -80°C indefinitely. This constitutes a national archive which contains isolates from various parts of Ireland, some of which date back to 1998.

2.4 Sample Selection

2.4.1 Sample Selection for Molecular Epidemiology of MTBC in Ireland

One isolate per patient per body site is genotyped in the IMRL every year. Isolates collected between January 2010 and December 2014, and received in the IMRL, were included in this

study. MIRU-VNTR genotyping data from the Northern Ireland Mycobacteria Reference Laboratory isolates (n=67) collected by colleagues in Belfast during this time-period were included in order to examine whether transmission may have occurred between jurisdictions.

2.4.2 Sample Selection for In-depth whole genome sequencing analysis of MIRU-VNTR Genotype clusters of interest

The IMRL Microsoft (Redmond, Washington, USA) Access master database (where all MIRU-VNTR genotyping data is recorded) was used to find duplicate MIRU-VNTR genotypes (selecting 24/24 exact-duplicate clusters, and clusters that had 1/24 SLV difference). This was enhanced and checked by using the MIRU-VNTR*plus* database [69]. Nine clusters of interest representing Lineages 1-4 (Indo-Oceanic, East Asian, East African Indian and Euro-American, n=103) were chosen, whole genome DNA was extracted and quantified, and sent via DX courier service (Iver, UK), to the Wellcome Trust Genomics Centre, Oxford for high-throughput sequencing on the Illumina[®] HiSeq[®], in collaboration with Prof. Derrick Crook and Modernising Medical Microbiology (MMM) group.

An institutional outbreak, already under investigation, was also chosen for this study (Cluster 10). Fourteen isolates had been prepared (DNA extraction and library preparation performed in University College Dublin, UCD, Belfield) and sent for sequencing, prior to the start-date of the current study, to the Beijing Genomics Institute, in collaboration with Prof. Stephen Gordon, UCD, and colleagues. A further 10 isolates were found to be part of the outbreak following this, one of which was community-acquired. These were prepared in-house and sequenced on the Illumina[®] MiSeq[®] in the TrinSeq Laboratory, Institute of Molecular Medicine, Dublin (Figure 12). Remainder isolates, and a further cluster (Euro-American lineage 4), were also sequenced in-house (n=16).

Public Health colleagues interrogated the (Computerised Infectious Disease Reporting) CIDR database for epidemiological information on patients within the clusters, as far as was possible, back to 2011, and a Microsoft Access database was investigated for cases preceding that time. Clinical details were accessed from patient notes and request forms and the Electronic Patient Record (EPR) in St. James' Hospital (SJH) in order to build a picture of the clusters to compare with the phylogenetic analysis. HPSC data was also used where possible.

2.4.3 Sample Selection for Molecular Characterisation of MDR/XDR-TB in Ireland

Isolates were selected on the basis of their drug resistance (DST) profile. Using the SJH Laboratory Information System (LIS) and Microsoft Access MIRU-VNTR Genotyping master database, a list of isolates that were at least phenotypically resistant to isoniazid and rifampicin



Figure 12. Illumina MiSeq Next Generation Sequencing Platform

The instrument has a small bench-top footprint (68.6 x 52.3 x 56.5 cm). Consumable reagents used are placed in front of the instrument in this figure. The flow cell is placed in the drawer on the left-hand-side of the instrument. The user interacts with the instrument via the display screen. (www.illumina.com)

was generated for the MDR-/XDR-TB molecular characterisation study. The electronic inventory of isolates was interrogated in order to find the selected MDR-/XDR-TB patient isolates. These were grown for at least four weeks using the BD Bactec™ MGIT™ 960 liquid culture system prior to whole genome DNA extraction and next generation sequencing (NGS).

Since tuberculosis is a notifiable disease in Ireland, the (Health Protection Surveillance Centre) HPSC collects data on it, including data on multi- and extensively-drug resistant cases. From 1998-2014, the HPSC recorded 42 cases of MDR. Due to this data being anonymous, and the fact that the date of notification could be different to the date the isolate was received in the IMRL, cases could not be correlated exactly. Despite this, 42 MDR-/XDR-TB isolates from 41 separate patients were retrieved from the bio-bank. Two different MDR-TB isolates were collected from the same patient at different time points. Thirty nine were whole genome sequenced using NGS; three isolates failed to grow for sequencing, however did have DST and rapid molecular test data performed (IEMDR37-39 inclusive).

2.4.4 Sample Selection for BD Bactec™ MGIT™ 960 Early Positive Culture Pilot Study

For this study (see Chapter 6), all mycobacterial culture-positive isolates (not including repeat isolates from the same patient), received in the IMRL in the month of October 2013, were included. A follow-on set of isolates, from the month of February 2014, was also included. A test run was performed before the ‘live’ runs began. Eight isolates were included; four mycobacterial culture-positive isolates received in the IMRL in September 2013, and four random culture-positive isolates (not taken at Day 0).

Specimens received in the IMRL were cultured using the BD Bactec™ MGIT™ 960 liquid culture system as per IMRL algorithm (Figure 7). When the cultures flagged positive, this was considered ‘Day 0’. A ZN stain was performed as per IMRL protocol. One ml was taken from all ZN-positive isolates, on either Day 0 or Day 1, heat-inactivated, sonicated, and frozen. The remaining culture (6 ml) was used by the IMRL for identification and DST using traditional methods as per protocol. Taking this aliquot did not affect the follow-on work. Isolates were prepared when sufficient numbers had been collected for an Illumina® MiSeq® run (11 isolates and one positive control per run) (Figure 12).

2.5 Mycobacterial Culture using the BD Bactec™ 960 MGIT™ Liquid Culture System and Lowenstein Jensen solid medium and Identification using Hain GenoType Line Probe Assays

The IMRL uses the BD Bactec™ MGIT™ 960 liquid culture system to process all specimens and referred cultures for detection, identification, and drug susceptibility testing (DST) if MTBC identified (see IMRL algorithm, Figure 7, and Section 1.11.2 for more details of the principal of the method). All specimens were decontaminated (from bacterial or yeast contamination) with 2% NaOH, followed by reconstitution with neutralising buffer. They were then centrifuged for 15 min at 4,200 rpm, supernatant removed, and 500 µl of deposit inoculated into a BBL™ MGIT™ vial containing Middlebrook 7H9 medium with growth supplement (BD BBL™ MGIT™ SIRE™) and a mixture of antibiotics which inhibit bacteria other than mycobacteria (BD BBL™ MGIT™ PANTA™, New Jersey, USA). BBL™ MGIT™ vials (7 ml) were incubated in the MGIT™ instrument for a maximum of 42 days at 35°C (± 2°C). When 'Δ Growth / Δ Time' reached a pre-defined threshold, the MGIT™ culture was deemed positive. Microscopy using a ZN stain (Section 1.11.1) was used to confirm this result, and to determine the morphology of the isolate. The culture was also inoculated onto Columbia blood agar with horse blood (E&O Laboratories, Bonnybridge, Scotland) for 48 hours to out rule bacterial contamination (i.e. a false positive).

Culture-positive isolates were identified using various molecular tests; BD Bactec™ MGIT™ TBC (identifies MTBC), Hain GenoType CM (identifies common mycobacteria and/or mixtures, e.g. MTBC, *M. kansasii*, *M. intracellulare*, *M. avium* etc.), Hain GenoType MTBC (differentiates members of the MTBC, e.g. *M. bovis*, *M. africanum*, *M. microti*, *M. bovis* BCG and others) and Hain GenoType AS (identifies less common mycobacteria such as *M. haemophilum*, *M. marinum* etc.). Figure 7 displays the algorithm currently in place in the IMRL for clinical specimens. Once identification and/or DST was completed, each isolate was frozen in duplicate at -80°C in the IMRL bio-bank. For the MGIT™ Pilot Study, 1 ml of culture was taken directly from each 'fresh' BBL™ MGIT™ vial. In all other cases, isolates were taken from the bio-bank and either grown in BBL™ MGIT™ liquid medium (plus growth supplement and antibiotic mixture) or directly on Lowenstein Jensen (LJ) medium (E&O Laboratories) until a confluent growth was seen.

2.6 Ziehl Neelsen Stain (ZN)

Mycobacteria were found to be 'acid-fast' by Robert Koch, which means they are impermeable to acid. This can be exploited to identify mycobacteria within a clinical specimen. Carbol fuchsin is used to stain the bacilli. They are then incubated with acid-alcohol and counter-stained with malachite green. Since the 'acid-fast' bacilli are impermeable to acid, they retain the pink carbol fuchsin, but not the green counter-stain (Figure 8).

A smear was made on a clean glass slide using one drop of thiomersal (Sigma Aldrich, St. Louis, Missouri, USA) and one drop of culture and dried on a hot plate at 65°C. The slide was flooded with a mixture of 5% phenol-in-ethanol for 5 min to fix. Carbol Fuchsin (PanReac Applichem, Darmstadt, Germany) was applied, and the slide was flamed until evaporation was seen emerging from the smear, followed by 5 min incubation at RT. The slide was washed with tap water, followed by application of a 25% acid-in-alcohol mixture (sulphuric acid in industrial methylated spirit, Sigma-Aldrich) for 30s and left for 2 min. This mixture was washed off with water and Malachite green (Merck, Darmstadt, Germany) was applied for 1 min. The slide was washed a final time and dried on a hot plate before the microscopy could be performed. If mycobacteria were present, bacilli would appear pink on a green background.

2.7 Anti-tuberculous Drug Susceptibility Testing (DST) using the BD BactecTM MGITTM 960 Liquid Culture System

Isolates were grown in MGITTM 960 liquid culture (BD, New Jersey, USA) and, when the required threshold was reached, DST was performed as previously described using the reference standard method [63, 64, 107]. Neat concentrations of isolates (at 1-2 days after reaching their growth threshold) were grown in the presence of a critical concentration of the desired drug compared to a 1:100 dilution of each neat isolate (growth control). Drugs tested included: rifampicin (1.0 µg/ml), isoniazid (0.1 and 0.4 µg/ml), pyrazinamide (100 µg/ml), streptomycin (1.0 and 4.0 µg/ml), ethambutol (5.0 µg/ml), kanamycin (2.5 µg/ml), moxifloxacin (0.25 and 1.0 µg/ml), ofloxacin (2.0 µg/ml), amikacin (1.0 µg/ml) and capreomycin (2.5 µg/ml) (BD). Different methods and drug regimens (both radiometric and non-radiometric) have been used over the years for DST of MTBC [108, 109]. Isolates for second line susceptibility testing were sent to supra-national reference laboratories in the UK. Information on the older methods used has not been included here. The reference standard (WHO-endorsed) MGITTM DST was used in the majority of cases. Other drugs tested, but not necessarily on all isolates, include clofazimine (4.0 µg/ml), linezolid (1.0 µg/ml), cycloserine (40 µg/ml), PAS (2.0 µg/ml), ethionamide (5.0 µg/ml), prothionamide (2.5 µg/ml), ciprofloxacin (1.0 µg/ml), rifabutin (0.5 µg/ml) and clarithromycin (0.5 µg/ml). Where conventional DST data was not available for a drug, it could not be included for comparison with whole genome sequencing resistance analysis.

2.8 Heat Inactivation of Isolates

All cultures were heat-inactivated by complete submersion in a water-bath at 95°C. For the adapted QuickGene Mini-80 protocol, the duration was 2 hours. When the MGITTM Pilot study and crude extraction were undertaken, the duration was 30 min. For MIRU-VNTR genotyping extraction (Hain GenoLyse[®] kit), because a lysis step was included, the duration was 5 min. Regardless of the extraction protocol, a heat-inactivation control sample, grown with the test cultures, was heat

inactivated alongside the test cultures, and a pre-inactivation aliquot and post-inactivation aliquot were cultured for up to 6 weeks in the BD MGIT™ 960 liquid culture system to prove that heat inactivation was sufficient to kill all bacilli present. The pre-inactivation controls grew within 2 weeks; the post-inactivation controls were negative after 42 days culture.

2.9 Hain Lifescience GenoType Line Probe Hybridisation Assays for rapid molecular detection of drug resistance in tuberculosis

GenoType MTBDR*plus* v2.0 and MTBDR*sl* v1.0 and v2.0 line-probe assays (Hain Lifescience GmbH, Nehren, Germany) were performed on isolates according to the manufacturer's instructions, either *in vitro* or *in silico* [110, 111]. Please refer to Section 1.11.3 for more details of the principal of the method. Heat-inactivated DNA extracts were amplified using specific primers for informative regions of the genes *rpoB*, *inhA*, *katG*, *rrs*, *gyrA*, *gyrB*, *embB* and *eis*. Reverse hybridisation was then performed on dedicated DNA strips which had been impregnated with specific probes for MTBDR*plus* v2.0 or MTBDR*sl* v1.0 and v2.0. A characteristic pattern was visualised and interpreted to indicate the presence of MTBC, and the presence or absence of drug-resistance-associated mutations to rifampicin, isoniazid, aminoglycosides, fluoroquinolones, ethambutol and low-level kanamycin. Conjugate, amplification and genus controls were included on the strip. A positive, negative and non-template control (NTC) were included with each batch. Figure 13 and 14 show examples of hybridisation strips that are used in the MTBDR*plus* v2.0 and MTBDR*sl* assays v1.0 and v2.0.

2.10 Crude DNA Extraction from ZN positive liquid cultures

Different extraction methods were used for different purposes. Mycobacterial DNA is not as easy to extract in a pure form, as other types of bacterial DNA. Line probe assays and MIRU-VNTR genotyping were performed using crude or GenoLyse® extracts but whole genome sequencing needed more DNA and further manipulation.

2.10.1 Crude Extraction

This method was performed in a microbiological safety cabinet (MSC) in the St. James' Hospital (SJH) IMRL CL3 laboratory, wearing appropriate PPE. Growth on solid media such as LJ slopes, or Middlebrook 7H10 plates could be tested. One µl of culture was taken with a disposable loop and re-suspended in 300 µl of molecular grade water, without disturbing the solid media. Growth in BBL™ MGIT™ culture tubes could also be crudely extracted. In the MSC, 1.0 ml of positive culture in 2.0 ml micro-centrifuge tubes was centrifuge at 13,000 rpm for 15 min (16,249 x g). Supernatant was carefully removed and the deposit re-suspended in 300 µl molecular grade water, followed by resuspension of pellet for 10s. Extracts were heat-inactivated in a 95°C water bath for

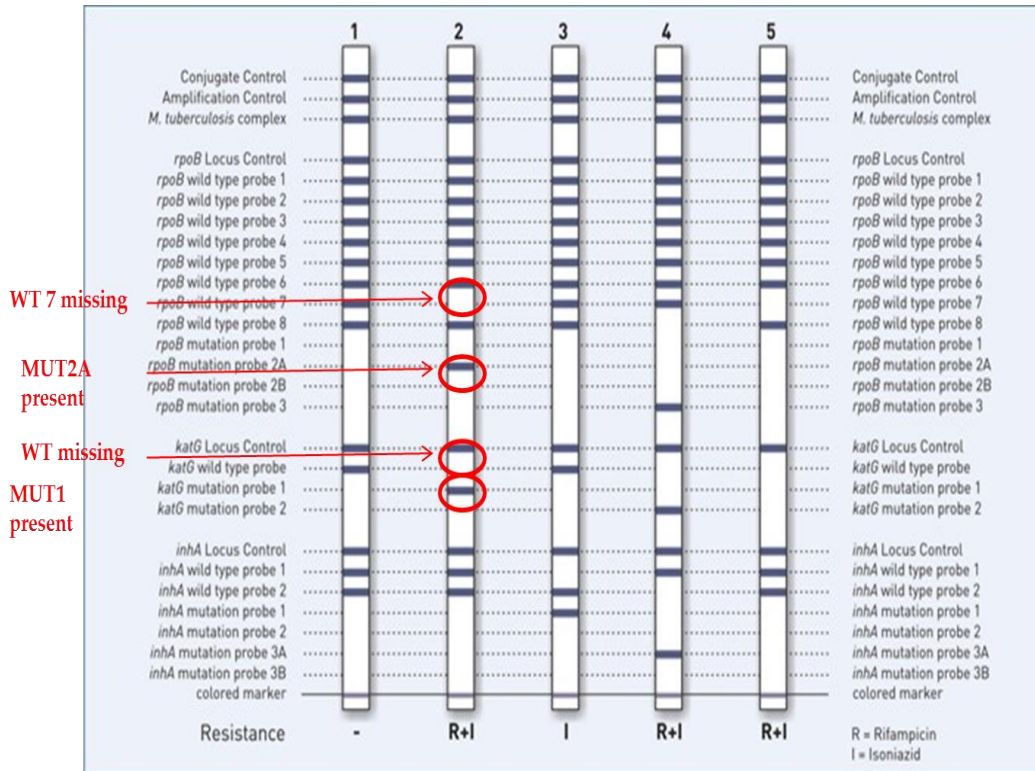


Figure 13. Examples of Hain GenoType MTBDR_{plus} Line Probe Assay reverse hybridised strips

Strips are hybridised with oligonucleotide probes specific for wildtype and mutations in genes associated with rifampicin (*rpoB*) and isoniazid (*katG* and *inhA*) resistance in MTBC. If wildtype probes are missing, resistance is indicated. If a mutation probe is hybridised, this will correspond to a specific mutation indicated by the kit. Example 2 shows an amplified extract of MTBC that is resistant to rifampicin (*rpoB* wildtype 7 missing, MUT2A present) and isoniazid (*katG* wildtype missing, MUT1 present). (Figure courtesy of Hain Lifescience)

	GenoType MTBDRs/ VER 1.0	GenoType MTBDRs/ VER 2.0
	<p>Conjugate Control (CC) Amplification Control (AC) <i>M. tuberculosis</i> complex (TUB)</p> <p><i>gyrA</i> Locus Control (<i>gyrA</i>) <i>gyrA</i> wild type probe 1 (<i>gyrA</i> WT1) <i>gyrA</i> wild type probe 2 (<i>gyrA</i> WT2) <i>gyrA</i> wild type probe 3 (<i>gyrA</i> WT3) <i>gyrA</i> mutation probe 1 (<i>gyrA</i> MUT1) <i>gyrA</i> mutation probe 2 (<i>gyrA</i> MUT2) <i>gyrA</i> mutation probe 3A (<i>gyrA</i> MUT3A) <i>gyrA</i> mutation probe 3B (<i>gyrA</i> MUT3B) <i>gyrA</i> mutation probe 3C (<i>gyrA</i> MUT3C) <i>gyrA</i> mutation probe 3D (<i>gyrA</i> MUT3D)</p> <p><i>rrs</i> Locus Control (<i>rrs</i>) <i>rrs</i> wild type probe 1 (<i>rrs</i> WT1) <i>rrs</i> wild type probe 2 (<i>rrs</i> WT2) <i>rrs</i> mutation probe 1 (<i>rrs</i> MUT1) <i>rrs</i> mutation probe 2 (<i>rrs</i> MUT2)</p> <p><i>embB</i> Locus Control (<i>embB</i>) <i>embB</i> wild type probe 1 (<i>embB</i> WT1) <i>embB</i> mutation probe 1A (<i>embB</i> MUT1A) <i>embB</i> mutation probe 1B (<i>embB</i> MUT1B)</p> <p>colored marker</p> <p>Differences between the two versions are marked in red</p>	<p>Conjugate Control (CC) Amplification Control (AC) <i>M. tuberculosis</i> complex (TUB)</p> <p><i>gyrA</i> Locus Control (<i>gyrA</i>) <i>gyrA</i> wild type probe 1 (<i>gyrA</i> WT1) <i>gyrA</i> wild type probe 2 (<i>gyrA</i> WT2) <i>gyrA</i> wild type probe 3 (<i>gyrA</i> WT3) <i>gyrA</i> mutation probe 1 (<i>gyrA</i> MUT1) <i>gyrA</i> mutation probe 2 (<i>gyrA</i> MUT2) <i>gyrA</i> mutation probe 3A (<i>gyrA</i> MUT3A) <i>gyrA</i> mutation probe 3B (<i>gyrA</i> MUT3B) <i>gyrA</i> mutation probe 3C (<i>gyrA</i> MUT3C) <i>gyrA</i> mutation probe 3D (<i>gyrA</i> MUT3D)</p> <p><i>gyrB</i> Locus Control (<i>gyrB</i>) <i>gyrB</i> wild type probe 1 (<i>gyrB</i> WT1) <i>gyrB</i> mutation probe 1 (<i>gyrB</i> MUT1) <i>gyrB</i> mutation probe 2 (<i>gyrB</i> MUT2)</p> <p><i>rrs</i> Locus Control (<i>rrs</i>) <i>rrs</i> wild type probe 1 (<i>rrs</i> WT1) <i>rrs</i> wild type probe 2 (<i>rrs</i> WT2) <i>rrs</i> mutation probe 1 (<i>rrs</i> MUT1) <i>rrs</i> mutation probe 2 (<i>rrs</i> MUT2)</p> <p><i>eis</i> Locus Control (<i>eis</i>) <i>eis</i> wild type probe 1 (<i>eis</i> WT1) <i>eis</i> wild type probe 2 (<i>eis</i> WT2) <i>eis</i> wild type probe 3 (<i>eis</i> WT3) <i>eis</i> mutation probe 1 (<i>eis</i> MUT1)</p> <p>colored marker</p>
Detection of	<i>M. tuberculosis</i> complex and its resistances to fluoroquinolones, aminoglycosides/cyclic peptides and ethambutol	<i>M. tuberculosis</i> complex and its resistances to fluoroquinolones and aminoglycosides/cyclic peptides
Sample Material	smear-positive pulmonary samples, cultivated samples	smear-positive and -negative pulmonary samples, cultivated samples
Ready-to-use amplification mix	-	✓
Ethambutol	Mutations in the <i>embB</i> gene that are involved in ethambutol resistance	
	✓	-
Fluoroquinolone	Mutations in the <i>gyrB</i> gene that are involved in fluoroquinolone resistance	
	-	✓
Kanamycin	Mutations in the <i>eis</i> gene that are involved in kanamycin resistance	
	-	✓

Figure 14. Hybridisation strips for version 1 and 2 Hain GenoType MTBDRplus Line Probe Assay

Strips are hybridised with oligonucleotide probes specific for wildtype and mutations in genes associated with fluoroquinolone (*gyrA*, *gyrB*), aminoglycoside (*rrs*), low-level kanamycin (*eis*), and ethambutol (*embB*) resistance in MTBC. If wildtype probes are missing, resistance is indicated. If a mutation probe is hybridised, this will correspond to a specific mutation indicated by the kit. (Figure courtesy of Hain Lifescience)

30 min, followed by centrifugation for 5 min to pellet cell debris. The supernatant containing DNA was transferred to a pre-labelled 0.5 ml PCR tube and stored at -20°C.

2.10.2 Hain Lifescience GenoLyse[®] Extraction

The protocol for crude DNA extraction (section 2.10.1) was followed, however, instead of resuspension in molecular grade water, 100 µl Hain GenoLyse[®] Lysing Reagent was used, and extracts were heat-inactivated at 95°C for 5 min. Once heat-inactivated, extracts were centrifuged briefly and 100 µl Hain GenoLyse[®] Neutralising Reagent was added to each tube, followed by resuspension of pellet. Extracts were centrifuged at 13,000 rpm for 5 min to pellet the cell debris, followed by transfer of the supernatant DNA extract to 0.5 ml PCR tubes.

2.11 Nucleic Acid Amplification

PCR amplification was performed on all DNA extracts requiring resistance mutation detection using line probe assays (LPAs) and/or MIRU-VNTR genotyping. The process was uni-directional, and performed in separate areas, according to good molecular laboratory practice.

Following DNA extraction, heat-inactivated DNA extracts were removed from the CL3 laboratory to the Molecular Laboratory, where two PCR workstations were designated for separate reagent preparation and DNA manipulation. Both workstations and dedicated pipettes were cleaned thoroughly prior to use with Microsol 3+ decontaminant (Anachem, Bedfordshire, UK), followed by molecular grade water, followed by 70% industrial methylated spirit (IMS). UV light was then applied to ensure no DNA contamination was present. PCR reagents were brought to RT from -20°C, mixed and centrifuged prior to use. Master-mixes, containing forward and reverse primers, dNTPs, buffer, molecular grade water and Taq polymerase, were prepared according to optimised methods for MTBDR^{plus} v2.0, MTBDR^{sl} and MIRU-VNTR genotyping. Master-mixes were aliquoted into each well of a 96-well PCR plate (Applied Biosystems, Foster City, California, USA, California, USA), or 0.5 ml PCR tube. 5 µl DNA extract were added when performing LPAs, 2 µl DNA extract were added when performing MIRU-VNTR genotyping.

For one MTBDR assay, 35 µl master-mix Am-A, 10 µl master-mix Am-B, and 5 µl DNA, were added. PCR tubes were closed securely and PCR was run on a thermal cycler (Veriti, Applied Biosystems, Foster City, California, USA, California, USA) using an optimised set of parameters, according to the manufacturer's instructions, in a dedicated PCR Product room. A positive (H37Rv) and negative (*Staphylococcus epidermidis*) quality control and NTC (molecular grade water) were run with every batch, from extraction through to amplification.

For MIRU-VNTR genotyping, 6 quadruplex PCRs, which contained four primer sets each, were required. For each test, 8 µl of each quadruplex mastermix were added to the 96-well plate, followed by 2 µl of appropriate DNA; 6 wells of the plate were associated with one isolate. An adhesive film (AB gene, Portsmouth, New Hampshire, USA) was used to tightly seal the 96-well plate to avoid evaporation during amplification, and the plate was centrifuged briefly prior to PCR amplification in a thermal cycler (Veriti, Applied Biosystems) using optimised parameters in a dedicated PCR Product room. A positive control (archived external quality assurance scheme MTBC DNA extract) and NTC (molecular grade water) were run with every plate.

2.12 PCR Product Line Probe Assay Reverse Hybridisation and Analysis

Following PCR, the products were run on the Hain GenoType protocol on an automated instrument (GT-Blot 48, Hain Lifescience GmbH, Nehren, Germany) according to the manufacturer's instructions. Following a Streptavidin-Biotin substrate-conjugate reaction, banding patterns on each hybridisation strip were visualised. Figure 13 and 14 show examples of what banding patterns should look like.

2.13 Automated MIRU-VNTR Genotyping of *Mycobacterium tuberculosis* Complex (MTBC)

Genotyping was performed by members of the IMRL (Dr. Margaret Fitzgibbon [n=398], Philomena Raftery [n=337] and Emma Roycroft [n=491]). Please refer to Section 1.11.4 for details of the principal of the method.

2.13.1 MIRU-VNTR locus amplification

Twenty-four informative loci, in 6 quadruplexes, were amplified according to the manufacturer's instructions (Genoscreen, Lille, France). Also, see Section 2.11 Nucleic Acid Amplification.

2.13.2 MIRU-VNTR Fragment Analysis

In the PCR product room, following amplification, the 96-well PCR plate, containing amplified locus fragments, was centrifuged briefly and 2 µl from each well added to the corresponding well on a new 96-well plate which contained a mixture of GeneScan 1200 LIZ dye Size Standard and Hi-Di formamide (both Applied Biosystems, Foster City, California, USA). This was followed by denaturation for 5 min at 95°C, centrifugation and fragment analysis.

The PCR fragments, labelled with the four different fluorescent dyes within each multiplex, were combined with the above internal size standard (labelled by a fifth dye, LIZ) and then analysed in

individual capillaries on the ABI 3130 Genetic Analyser (Applied Biosystems) for size determination (36 hours for 96-well plate).

2.13.3 Analysis and Allele Assignment using GeneMapper® Software

Every locus contains a certain number of repetitive units (MIRU/VNTR). The more repeats, the larger the fragment. The less repeats, the smaller the fragment. The fragments were run, using an electric current, through a capillary array containing an electrophoretic gel polymer at different rates due to their size, and peaks were created where the fragments pass a detector. GeneMapper® v 5.0 (ThermoFisher Scientific) was calibrated to recognise each possible fragment size at each of the 24 loci using the resulting electropherogram peaks. ‘Bins’ were created in order to visualise where the fragments best fit. For instance, a locus-X fragment, which has Y repeats, will travel through the capillary array at a certain rate and create a peak at a certain point. This peak will fall into a corresponding ‘Bin’ on the GeneMapper® software so that the User can visualise the pattern.

Fragment analysis results retrieved from the ABI 3130 Genetic Analyser were analysed using GeneMapper® software and alleles were assigned accordingly. Stutter, pull-up, and spurious peaks were examined in order to call the correct peak, and therefore report the correct fragment size for each locus. Raw data was analysed where necessary. Peaks were manually assigned where GeneMapper® failed to assign an allele. Loci were repeated wherever necessary. Double alleles, where isolates containing two variants of a same clone or two independent strains, where two peaks emerged above stutter peak background. Genotypes were exported to a Microsoft Access database for further reporting, data storage and analysis.

2.13.4 MLVA Compare Software and the MIRU-VNTR^{plus} online database

Lineage calling was performed using MLVA Compare and the MIRU-VNTR^{plus} database, using similarity search, tree-based, and manual identification of strains based on their 24-locus MIRU-VNTR genotypes. MLVA Compare software (Ridom, GmbH, Munster, Germany) and MIRU-VNTR^{plus} database are interfaced [69]. Each new MIRU-VNTR genotype was uploaded to a Microsoft Excel file, with loci in the order that is required by MLVA Compare software and MIRU-VNTR^{plus} database (<http://www.miru-vnrplus.org/MIRU/index.faces>). The file was imported into the software, from where the Multi-locus Variant Analysis (MLVA) MtbC15-9 genotype could be interpreted, followed by lineage calling based on 24-locus MIRU-VNTR patterns, either by similarity search or by phylogenetic tree compared to reference strains from the online database. A neighbour joining or Unweighted Pair Group Method with Arithmetic Mean (UPGMA) tree could be built using reference strains and test strains. This could be viewed as a dendrogram or radial tree. MLVA Compare can also visualise the strains using a minimum-

spanning tree. Results were exported to the master database from MLVA Compare. An algorithm designed by Allix-Beguec *et al* was used to designate strain lineages from phylogenetic trees (Figure 15) [69]. The master database was interrogated for duplicates of each new genotype.

2.14 Whole Genome Sequencing using Illumina[®] Next Generation Sequencing Technology (MiSeq[®] and HiSeq[®])

2.14.1 Whole genome DNA Extraction

Illumina[®] HiSeq[®] requires more input DNA than Illumina[®] MiSeq[®] (100 µl of 6ng/µl, compared to 5 µl of 0.2ng/µl). Please refer to Section 1.14 for more details of the principal of the method.

2.14.1.1 Adapted QuickGene Mini-80 Protocol, originally developed by Walker et al

This method was used for the MDR/XDR-TB cohort, and cluster analysis isolates that were sent to the Wellcome Trust Genomics Centre Illumina[®] HiSeq[®] for sequencing. Positive and negative controls were included with every extraction and library preparation, and in certain cases, for WGS (H37Rv MTBC reference strain and non-template control).

Whole genome DNA extraction of the MDR/XDR-TB cohort was performed on four-week-old MGIT[™] cultures (on liquid medium and solid LJ medium) which had been frozen in Middlebrook 7H9 medium (BD) at -80°C, using a previously published adapted Autogen QuickGene (Kurabo Bio-medical, Osaka, Japan) protocol [112]. Isolates for Illumina[®] HiSeq[®] sequencing were grown for at least 8 weeks, also on solid medium (LJ slopes) in some cases, in order to achieve the required concentration. Heat-inactivation and positive and negative controls were included with each batch.

Three ml of each culture were centrifuged for 15 min at 13,000 rpm (1 ml centrifuged, supernatant removed and reconstituted x 3 times) in a 2 ml micro-tube, followed by addition of 400 µl 0.9% sterile saline, and pellet thoroughly resuspended. Cultures growing on LJ slopes were harvested along with the liquid cultures in order to increase the yield. Isolates were heat inactivated at 95°C for 2 hours, sonicated for 15 min, and frozen overnight. Twenty µl EDT (Proteinase K) and 300 µl LDT lysis buffer from the QuickGene Tissue DNA Extraction kit (Kurabo Bio-medical, Osaka, Japan) were added to each defrosted isolate, resuspended, and transferred to Lysing Matrix B tubes (MP Bio, Santa Ana, California, USA). Extracts were mechanically disrupted twice (40s at 6000 m/s) using a MagNA Lyser (Roche, Basel, Switzerland) and centrifuged at 13,000 rpm for 10 min, after which the supernatant was transferred to a fresh labelled tube, and incubated for 10 min submerged in a water-bath at 70°C. Molecular grade ethanol (240 µl, 96-99% v/v) was added,

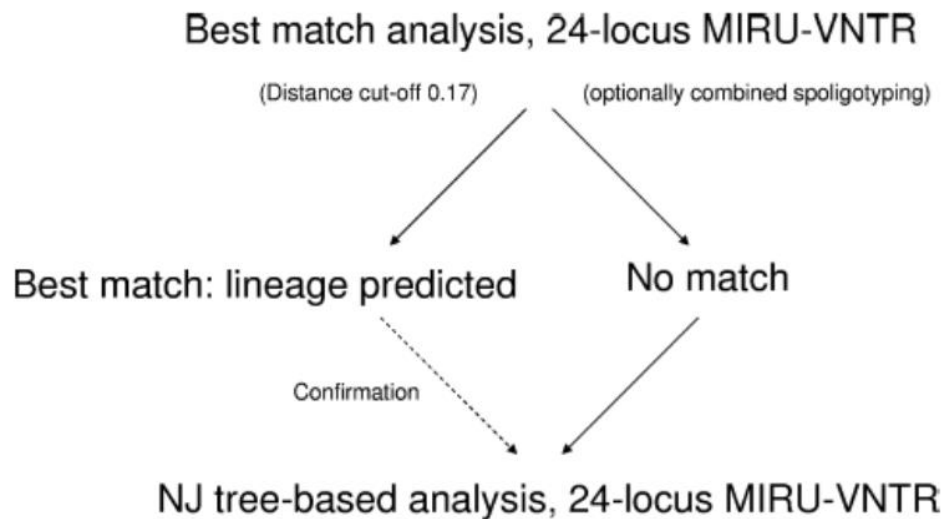


Figure 15. Algorithm for phylogenetic lineage-calling for 24-locus MIRU-VNTR genotypes

This figure shows an algorithm for use with the MIRU-VNTR*plus* online database when assigning lineage to newly-tested MIRU-VNTR genotypes [69]. A similarity search is the first-line method, where the ‘best match’ lineage is predicted or there is no match. Confirmation of the best match, or lineage assignation, is then performed using a neighbor joining phylogeny constructed from 180 reference strains and the queried new strains.

followed by resuspension of the pellet (15s). The QuickGene Mini-80 compact benchtop extraction instrument was set up (8 isolates processed per run). Extracts were applied to each tube (1-8), ensuring the rack was placed over the discard tubes (Figure 16), and the handle turned to filter the extracts. WDT (ethanol wash buffer) from the kit (750 µl) was added, and the handle turned. This was repeated twice more, after which the rack was moved back so that the purified DNA could be eluted into fresh labelled 0.5 ml tubes (eluted with 100 µl CDT elution buffer, incubated for 2 min). The DNA was then quantified (section 2.14.3). For Illumina[®] HiSeq[®], at least 100 µl of 6ng/µl input DNA was required. For Illumina[®] MiSeq[®], at least 5 µl of 0.2ng/µl input DNA was sufficient.

2.14.1.2 MGIT[™] Pilot Study Protocol for Extraction of Early-positive Mycobacterial Cultures [113]

Once BD Bactec[™] MGIT[™] 960 liquid cultures reached the positivity threshold, they were removed from the instrument. Either on that day (Day '0'), or on the following day (Day '1'), 1 ml culture was removed, sonicated for 15 min and heat-inactivated for 30 min at 95°C. Isolates were centrifuged for 15 min at 13,000 rpm, supernatant was removed and 1 ml 0.9% sterile saline was added, followed by re-suspension and a repeat centrifuging step. Supernatant was removed and the pellet re-suspended in 700 µl molecular-grade water. The entire extracts were transferred to Lysing Matrix B tubes (MP Bio, Santa Ana, California, USA) and mechanically disrupted 3 times (40s at 6000 m/s, with 5 min rest between each round) using a MagNA Lyser (Roche, Basel, Switzerland) and centrifuged at 13,000 rpm for 10 min, after which 450 µl supernatant was transferred to a new labelled tube. A 1/10 volume (45 µl) of 3M sodium acetate (Sigma Aldrich, St. Louis, Missouri, USA) was added, along with 2 volumes (1 ml) of ice-cold molecular grade ethanol (Sigma Aldrich), were added followed by resuspension of pellet and incubation at -20°C for 1 hour. The extracts were centrifuged for 15 min at 13,000 rpm. Supernatant was removed and the pellet was washed with 1 ml 70% molecular grade ethanol for one min, followed again by removal of supernatant. The pellets were air-dried at RT for 10-15 min and then re-suspended in 50 µl molecular grade water by heating at 55°C for 10 min in a water-bath, re-suspending the pellet 1-2 times during incubation. Forty-five µl supernatant was added to a new 0.5 ml PCR tube, followed by 'clean-up' using AMPure XP beads (Agencourt, Beverley, Massachusetts, USA).

2.14.2 AMPure XP beads 'Clean-Up'

This method was used in DNA extraction and DNA library preparation protocols.

Beads were mixed well and 1.8x volume added to the DNA in a 96-well 0.8 ml plate (AB Gene, Portsmouth, New Hampshire, USA), vortexed for 20s and incubated for 10 min at RT. Samples were placed on a magnetic stand (Ambion, Life Technologies, Carlsbad, California, USA) for 3



Figure 16. Quickgene Mini-80 Instrument used for Adapted Quickgene extraction protocol for MDR/XDR-TB cohort, and isolates for whole genome cluster analysis

Whole genome extraction can be accelerated using this instrument. Extraction methods include the toxic CTAB method and other manual, labour-intensive methods. This instrument has been found to achieve similar results when adapted into an otherwise manual extraction method [112] (www.autogen.com)

min followed by careful removal of supernatant. Leaving the plate on the stand, samples were washed twice with 80% molecular grade ethanol and air-dried for 10-15 min, followed by removal from the stand and thorough re-suspension in 26 μ l molecular grade water. Samples were replaced on the magnetic stand for 3 min, after which 25 μ l supernatant (without beads) was removed to new labelled tubes, followed by DNA quantification (section 2.14.3).

2.14.3 DNA Extract and/or Library Quantification

The Quant-iT Qubit™ dsDNA HS Assay Kits for use with the Qubit® Fluorimeter (Invitrogen, Carlsbad, California, USA) are designed for DNA quantitation of double-stranded DNA (dsDNA) for initial sample concentrations from 10 pg/ μ L to 100 ng/ μ L. The kit provides concentrated assay reagent, dilution buffer, and pre-diluted DNA standards.

Once all reagents were at RT, thoroughly mixed and centrifuged briefly, Qubit tubes were set up (two for standard 1 and 2, and one for each sample to be measured). Quant-iT working solution was made by diluting the Quant-iT reagent 1:200 in Quant-iT buffer (200 μ l of working solution was required for each sample and standard). 190 μ l was added to each standard tube, and 198 μ l was added to each test tube. Ten μ l of standard 1 and 2, and 2 μ l of each sample were added to the appropriate tubes and vortexed for 2-3s, followed by incubation at RT for 2 min. The fluorimeter was calibrated using standards 1 and 2, according to the instructions on the instrument for ‘Quant-iT High Sensitivity assay’, followed by reading of each test. The first reading measured how much DNA was in 200 μ l (the sample size, diluted). In order to find out how much DNA is in 2 μ l, a calculation was performed on the instrument. Results were displayed in ug/ml, which is equivalent to ng/ μ l.

2.14.4 Illumina® Nextera XT Library Preparation and Quantification

MGIT™ Pilot Study adapted library preparation protocol [114]

2.14.4.1 Tagmentation of DNA

DNA was ‘tagmented’ with Illumina Adaptor sequences. Once reagents were at RT, mixed and centrifuged briefly, 10 μ l of Tagment DNA buffer from the Nextera XT Library Preparation kit (Illumina®, San Diego, California, USA) was added to each well to be used, followed by 5 μ l of input DNA at 0.2 ng/ μ l (1ng total) and 5 μ l of Amplicon Tagment Mix, followed by gentle mixing. The plate was sealed, centrifuged briefly, followed by PCR at 55°C for 5 min. The reaction was neutralised immediately once samples had returned to 10°C with 5 μ l Neutralise Tagment buffer, gently mixed and centrifuged briefly before incubating for 5 min at RT.

2.14.4.2 Index PCR Setup and Clean-up

A template was set up for each run, which included a pair of indices which would identify each library within the multiplexed assay, and with which Illumina MiSeq software could de-multiplex the reads following sequencing. Once thawed to RT, mixed and centrifuged briefly, 15 µl of Nextera Primer Mix was added to each well containing clean tagmented DNA, followed by addition of 5 µl each of appropriate index 1 (i7) and index 2 (i5). The plate was sealed and centrifuged briefly followed by PCR for 15 cycles (72°C for 3 min, followed by 15 cycles of 95°C for 30s, 95°C for 10s, 55°C for 30s and 72°C for 30s, followed by 72°C for 5 min). Index PCR clean-up was performed using AMPure XP beads (section 2.14.2), followed by quantification using Quant-iT Qubit™ Assay (section 2.14.3)

2.14.4.3 Library Normalisation

Using Qubit™ measurements (ng/µl), libraries were normalised to 1.6 ng/µl, or 4 ng/µl, to achieve 4nM or 10nM libraries respectively (or in rare cases, 0.8 ng/µl, 2nM), using Tris-Cl 10 mM pH 8.5 with 0.1% Tween 20 (Sigma Aldrich, St.Louis, Missouri, USA). (Conversion factor of 1ng/µl = 2.5 nM)

2.14.4.4 Library Pooling and Denaturation

Normalised libraries were mixed and 5 µl of each were transferred to a new 2 ml micro-tube, followed by thorough mixing of the pool. One ml of fresh 0.2 N NaOH (Sigma Aldrich) was prepared and 15 µl was added to a new micro-tube. Fifteen µl of the pool were added, mixed thoroughly and left to denature to single stranded DNA for 5 min at RT. To obtain 50 pM of denatured pool using a 4 nM pool, 25 µl of denatured DNA were added to 975 µl of pre-chilled HT1 (Hybridisation buffer). To obtain 50 pM using a 10 nM pool, 10 µl of denatured DNA were added to 990 µl of pre-chilled HT1. To obtain 20 pM using a 2 nM pool, 20 µl of denatured DNA were added to 980 µl of pre-chilled HT1. To obtain 600 µl of desired input concentration, the pooled, denatured, diluted DNA was diluted further using the table below:

Final pool concentration	16 pM	20 pM	Final pool conc.	15 pM
50 pM denatured DNA	192 µl	240 µl	20 pM denatured DNA	750 µl
Pre-chilled HT1	408 µl	360 µl	Pre-chilled HT1	250 µl

The final aliquot was vortexed, centrifuged briefly, and kept on ice before loading into the Illumina® MiSeq® reagent cartridge.

2.14.5 Illumina[®] MiSeq[®] and BaseSpace Cloud Computing

Following a post-run wash, the MiSeq[®] was ready for use. Steps indicated on the instrument display screen were followed. The wash reagents were removed from the instrument. The flow-cell, where clusters are formed and sequencing-by-synthesis takes place, was removed from its transport medium, dried with lint-free tissue, washed extremely gently with molecular grade water, followed by 70% molecular grade ethanol, thoroughly and gently dried once more and placed carefully in its seating on the instrument when prompted. The MiSeq[®] run was set up to send all output data to Illumina[®] BaseSpace, a cloud computing environment, in which the sequencing run could be viewed in real-time from any personal computer with access to the internet, and from where the sequences may be downloaded following completion of the run (in fastq file format). The .csv template file (containing sample and index information), that was uploaded prior to initialising the sequencing run, contains vital information which enables the MiSeq[®] to map the correct sequence data back to its corresponding sample. The reagent cartridge (containing the final normalised library pool) and buffer were loaded onto the instrument when prompted, and the run could begin once all checks were complete. A 300-cycle run took 27 hours to complete, after which a post-run wash was performed using Tween 20 (Sigma Aldrich), according to the manufacturer's instructions.

After cycle 25 was complete, metrics could be remotely monitored in order to assess the performance of the MiSeq[®] run. For a 300-cycle run (2 x 150), cluster density should be 800-1200 K/mm², data quality scores should be >75% Q30, clusters passing filter should be a high percentage, estimated yield should be > 1 Gb, data-by-cycle intensities should be > 200 (they may drop over the course of the run but should not drop below 200). Figure 17 shows the MiSeq[®] run detail that can be seen on BaseSpace for MGIT[™] Pilot Study Run 2.

2.14.6 Illumina[®] HiSeq[®]

Once sufficient DNA concentration was reached (at least 100 µl of 6 ng/µl input DNA required), DNA extracts were sent to the Modernising Medical Microbiology group in Oxford via an approved courier (DX, Iver, UK), from where they were further processed for sequencing on the Illumina[®] HiSeq[®] located at the Wellcome Trust Genomics Centre, Oxford. The main difference between the HiSeq[®] and MiSeq[®] is that the HiSeq[®] could perform NGS on 196 MTBC genomes at once (~600 Gb data output), while the MiSeq[®] could process approximately 12 MTBC genomes at once (300 cycle kit) (~5-12 Gb data output). The cost per sample is less when using the HiSeq[®] but the run must be at capacity in order to be cost-effective.

28/1/2014

Run Detail Db_MGIT_Run02 - BaseSpace

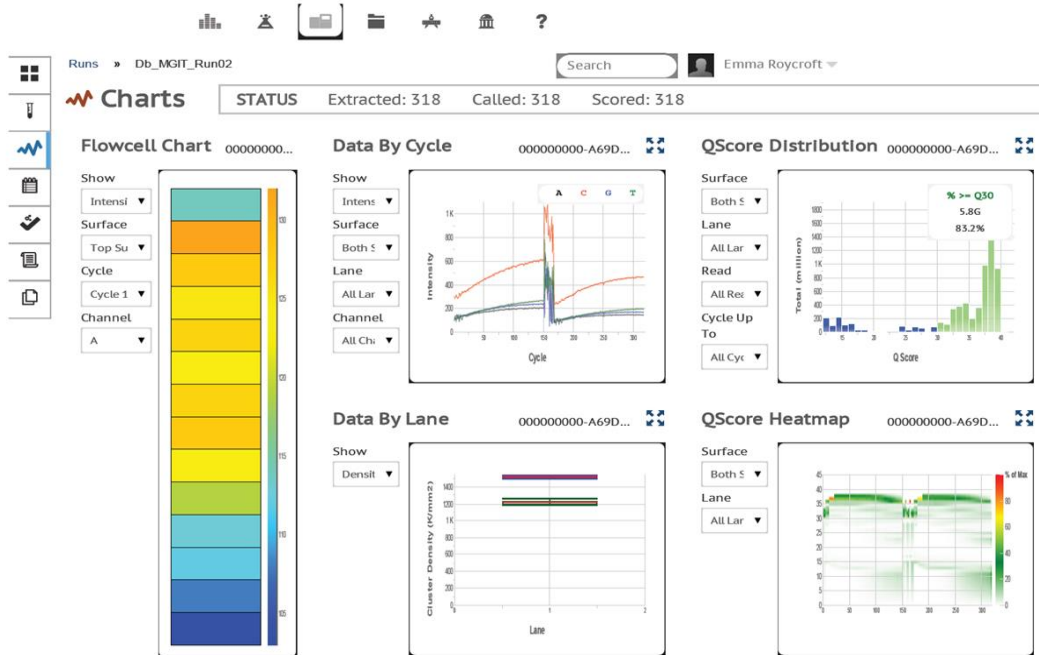


Figure 17. Illumina MiSeq Run Detail for MGIT Pilot Study Run 2

Information seen in real-time on BaseSpace as the MiSeq run progresses. The screenshot includes a flowcell chart with intensity of clustering shown as a heatmap, Data by Cycle, Q Score distribution, which represents the quality of the sequencing data emerging by histogram and heatmap, and Data by Lane that shows the cluster density and clusters passing filter. A good MiSeq run will have 800-1200 K/mm² cluster density with most of those passing filter, and Q scores of >30.

2.15 Whole Genome Sequencing Analysis of output fastq files

Table 2 contains a summary of all programs, websites, software and scripts used throughout the study.

2.15.1 Illumina[®] Basespace cloud platform

Illumina[®] MiSeq[®] output data was de-multiplexed and uploaded, in the form of paired end fastq files (forward read 1, R1, and reverse read 2, R2), automatically to BaseSpace cloud computing storage and analysis platform, from where it could be stored, downloaded for further in-house analysis, or shared with the MMM group in Oxford for further analysis using bespoke workflows.

2.15.2 Linux Operating System

Many open-source software programs have been created to analyse fastq data. Some of them incorporate a graphical user interface (GUI), but some do not. The ones that do not incorporate a GUI can be written in many computing languages, and must be operated using a command line terminal. Bio-linux 8 was used in this study (based on Ubuntu Linux 14.04 LTS), since it already incorporates over 250 bioinformatics analysis tools [75]. Please refer to Section 1.12 for more details on bioinformatics.

2.15.3 FastQC software, quality control, and trimming

FastQC was used to assess the quality of the output fastq files. Where fastq files were below desired quality, trimming could be done using Trimmomatic based on quality, or read length by assessing at what read length quality began to decline [115, 116]. Import of data from Binary Alignment Map (BAM), (Sequence Alignment Map) SAM or fastq files is possible with FastQC. The parameters measured were: Basic statistics, Per base sequence quality, Per sequence quality scores, Per base sequence content, Per base GC content, Per sequence GC content, Per base N content, Sequence length distribution, Sequence duplication levels, Over-represented sequences and Kmer content. FastQC could be performed again on the trimmed fastq files, and metrics should have improved. Figure 18 displays the Per Base Sequence Quality achieved for Read 1 fastq file of IEXDR1, the first XDR-TB isolated in Ireland. Adapter trimming, set up as part of the initial MiSeq[®] run parameters, increased quality control since sequences where adapters had inadvertently been sequenced were removed prior to analysis.

2.15.4 BWA-MEM alignment

BWA (Burrows Wheeler Alignment, version 0.7.12) is a software program that aligns (or maps) low-divergent sequences against reference genomes (local alignment) [117]. BWA-MEM (Maximal Exact Matches) was the algorithm used as Illumina[®] fastq reads were over 70bp in

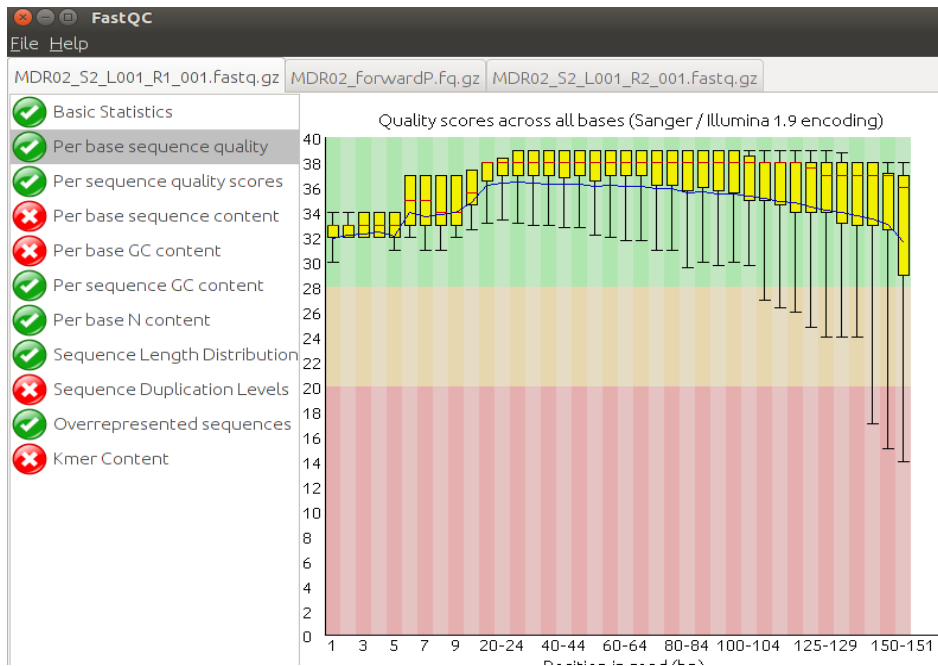


Figure 18. Per Base Sequence Quality for a fastq file (read 1, IEXDR1) measured with FastQC software

This figure is a screenshot from the FASTQC program that assesses the quality of sequencing data from raw fastq files [115]. The example above shows that this isolate sequenced with very high quality. Even though Per Base Sequence Content, Per Base GC content, Sequence Duplication Levels and Kmer Content have ‘failed’, the sequence has passed the most important parameters and can be used for sequencing. When de novo assembly is required, kmer content may interfere, however when mapping to a reference genome, these parameters are less important.

Table 2. Summary of Programs, Websites, Software and Scripts used throughout the Study

The names of the programs/software packages/scripts are included, along with a brief explanation and reference where available. NK – not known, N/A – not applicable, ER – Emma Roycroft

Program	Version	Purpose	Reference
BWA-MEM	0.7.12	alignment	[117]
SAMtools	0.1.7 - 0.1.9	variant calling	[251]
Trimmomatic	0.32	trimming	[116]
Maq	0.7.1	changes the format of fastq files	[252]
Smalt	NK	mapping	[252]
Bowtie	1.1.1	alignment	[70]
vcftools	0.1.5	variant calling	[253]
FastQC	0.11.3	quality control	[115]
TBProfiler	NK	freely-available online web-tool for the analysis of MTBC whole genomes	[77]
PhyResSe	27.0	freely-available online web-tool for the analysis of MTBC whole genomes	[76]
Seaview	4.0	phylogenetics visualisation	[72]
Artemis	16.0.0	genome browser	[254]
MEGA	NK	alignment	[252]
Geneious	R9	commercial genome browser, variant calling, phylogenetics	[74]
PhyML	3.0	maximum likelihood phylogenetics	[73]
MIRU-VNTRplus database	N/A	lineage calling and phylogenetics for MIRU-VNTR genotyping data	[69]
Figtree	1.4.2	phylogenetics visualisation	[255]
MLVA Compare	1.02	lineage calling and phylogenetics for MIRU-VNTR genotyping data	[256]
GeneMapper	5.0	fragment analysis software for MIRU-VNTR genotyping	N/A
Illumina BaseSpace	1.0	cloud-computing platform that stores sequencing output	N/A
Biolinux	8.0	bioinformatics operating system, based in Ubuntu 14.04 LTS, contains many bioinformatics programs already installed	[252]
Stampy	1.0.13	mapping	[118]
Galaxy.org	various	freely-available online tool-box where the User can use otherwise Linux-based programs in a Windows environment	[74]
IGV	NK	Integrated Genome Viewer, provided by the Broad Institute	N/A
Stata	13.1	Statistical package	N/A
Squirrel Walk	N/A	custom script	[114]
Genefinder	N/A	custom macro script for Microsoft Excel	ER
Fasta2Phylip.pl	N/A	custom script	[112]
Gp_seq_2012.py	N/A	custom script	[112]
pad_mask_Mltree.py	N/A	custom script	[112]

length. The algorithm seeded alignments with the most exact matches with the reference genome, and then extended seeds with the affine-gap Smith-Waterman algorithm. As the name suggests, gap penalties are utilised in alignment algorithms to calculate which gaps are more detrimental to the alignment than others. Affine gap penalties compare just two gap scenarios, which makes the algorithm faster than performing general gap penalty calculations. An index was created for the reference genome in the fasta format, H37Rv (Genbank accession AL123456.3, NCBI reference sequence number NC_000962.3), followed by processing of BWA-MEM alignment algorithm. A SAM file was produced for use with SAMtools. A synopsis of the commands used is as follows:

```
bwa index ref.fa
```

```
### H37Rv reference is indexed
```

```
bwa mem ref.fa read1.fq read2.fq > aln-pe.sam
```

```
### fastq files for read 1 and 2 are pair-end-aligned
```

2.15.5 SAMtools and Bcftools for Variant Calling

SAM (Sequence Alignment Map) files are used to store large nucleotide sequence alignments in a compact format. BAM is the compressed binary equivalent of these files. SAMtools (version 0.1.9) work with these files, manipulating alignments via commands such as ‘view’, ‘sort’, ‘merge’, ‘index’, by removing PCR duplicates, and creating ‘per-position’ VCF, BCF or pileup files using ‘mpileup’ [71]. Commands can be combined for faster processing of files, using Unix pipes. A synopsis of the commands used in this study is as follows:

```
samtools view [options] in.sam/in.bam
```

```
### view the input sam or bam file
```

```
samtools sort -T /tmp/aln.sorted -o aln.sorted.bam aln.bam
```

```
### sort the bam file and place in a temporary directory called aln.sorted, the output to be called aln.sorted.bam and aln.bam
```

```
samtools index aln.sorted.bam
```

```
### index the aligned sorted bam file
```

```
samtools merge aln.sorted1.bam aln.sorted2.bam -o aln.sorted12merged.bam
```

```
### merge read 1 and 2 aligned sorted bam files, the output should be called aln.sorted12merged.bam
```

samtools faidx ref.fasta

samtools index the H37Rv reference to a fasta file

samtools tview aln.sorted.bam ref.fasta

view the aligned sorted bam file against the H37Rv reference fasta file

samtools mpileup

perform the alignment

samtools fastq input.bam > output.fastq

where a fastq output is required, this command is used

samtools fasta input.bam > output.fasta

where a fasta output is required, this command is used

2.15.6 Whole Genome Cluster Analysis

Please refer to Section 1.13 for more details on phylogenetic analysis.

2.15.6.1 Geneious R9 for Whole Genome Cluster Analysis

This commercial software is designed to bring together many molecular analytical tools in one package for use by biologists who are not as familiar with computer science [78]. It uses many open-source software programs, but combines them with a GUI instead of Linux command line.

Fastq files were imported into Geneious R9 software, where both 5' and 3' ends were trimmed based on a more than 5% chance of base call error. Paired-end reads were then mapped to the H37Rv (GenBank AL123456.3, NCBI NC000962.3) reference genome. For variant analysis, minimum variant frequency was set at 0.25, while maximum variant P-value was set at 10⁻⁶ and minimum strand bias at 10⁻⁵ when exceeding 65% bias. A consensus sequence of each strain of interest (which had already been mapped to H37Rv) was set followed by multiple alignment with H37Rv. Geneious Tree Builder was used to build a consensus phylogenetic tree from the multiple alignment (Bootstrap 100 re-sampled replicates, branch support threshold > 50%), using the Jukes Cantor genetic distance model and H37Rv as the outgroup. Other models of evolution possible were HKY or Tamuri-Nei. A neighbour-joining or UPGMA tree could be used for phylogenetic reconstructions.

2.15.6.2 Phylogenetic Analysis of Whole Genome Clusters using method developed by Oxford MMM (masked maximum-likelihood tree building)

MIRU-VNTR genotyped clusters were whole genome extracted and sent to MMM collaborators in Oxford in November 2014 and January 2015. Library preparation and NGS on Illumina[®] HiSeq[®] instruments were co-ordinated by MMM with the Wellcome Trust Genomics Centre, Oxford. Phylogenetic reconstructions of each cluster (1-9) were carried out in collaboration with this group, who had developed a workflow for MTBC cluster analysis for a previous publication [112]. The HiSeq[®] paired-end reads were mapped with Stampy (version 1.0.13) to H37Rv, without BWA pre-mapping, with an expected substitution rate of 0.01. Repetitive regions (make up approximately 7% of the genome) were located using a self-self BLAST method, followed by masking, using a custom script, prior to further analysis. Variant calling was performed using SAMtools mpileup (homozygous in a diploid model). Only variants with variant frequency of > 75% and read depth of at least 5 reads were accepted. Each variant had to occur in each direction and could not be located within 12 base pairs (bp) of another variant or indel [118]. Maximum likelihood (ML) trees were constructed using concatenated variable sites across clustered genomes with PhyML 3.0, and visualised using Seaview [72, 73]. The clusters were analysed while taking into account a previously calculated threshold for recent transmission. Less than 5 Single Nucleotide Variants (SNVs) between samples indicated recent transmission (i.e. an outbreak), 5-12 SNVs indicated possible transmission depending on the time-frame of the outbreak, > 12 - < 20 SNVs represented a grey-zone, and > 20 SNVs out-ruled recent transmission based on the estimated mutation rate of 0.1 SNV per genome per year of *M. tuberculosis*.

For each cluster, a .txt file, which incorporated a list of the filename paths was created. Custom python and perl scripts, new folders 'Pad' and 'Tree', and a copy of PhyML 3.0 for Linux, were imported into the working directory (Linux OS, command line):

- Fasta2Phylip.pl – changed fasta file format to Phylip format, compatible with PhyML
- Gp_seq_2012.py – generated a fasta SNV file, a distance matrix (.dat file), and a .txt file containing the genomic positions of the SNVs
- pad_mask_MLtree.py – created the maximum likelihood phylogeny

2.15.6.3 Phylogenetic Analysis of Whole Genome Clusters using method developed by Dr. Javier Nunez at the Animal Health and Veterinary Laboratories Agency, UK

Data files for one isolate were named “filename_1.fq.gz” for one end of the fragment and “filename_2.fq.gz” for the other end of the fragment (they were paired-end sequenced). Quality of the data sets was assessed with FASTQC [115]. The files were unzipped:

```
gunzip -c filename_1.fq.gz > filename_temp_1.fastq
gunzip -c filename_2.fq.gz > filename_temp_2.fastq
### fastq files were unzipped using gunzip program
```

Some alignment tools can handle several different formats, some others cannot. Sanger format has become the standard. Files were transformed to Sanger format, followed by alignment and mapping to H37Rv with MAQ and SMALT, BWA, or Bowtie, which are installed within Biolinux 8 [75].

```
maq sol2sanger filename_temp_2.fastq filename_1.fastq
maq sol2sanger filename_temp_2.fastq filename_2.fastq
### maq program was used to convert the fastq files from Solexa to Sanger format (older
sequencing run outputs can be in Solexa format)
```

AND

```
smalt index -k 13 -s 6 H37Rv.fna H37Rv
smalt map -n 7 -f samsoft -o filename.sam H37Rv filename_1.fastq filename_2.fastq
### smalt program used to index the H37Rv reference, followed by mapping of fastq files to the
reference strain
```

OR

```
bwa index -p H37Rv.fna H37Rv
bwa aln H37Rv filename_1.fastq > filename_1.sai
bwa aln H37Rv filename_2.fastq > filename_2.sai
bwa sampe $2 filename_1.sai filename_2.sai filename_1.fastq filename_2.fastq > filename.sam
### as above BWA was used to index the H37Rv reference, fastq reads 1 and 2 were then indexed
and aligned to H37Rv, the resulting files were merged to a sam file
```

OR

```
bowtie-build H37Rv.fna H37Rv
bowtie H37Rv -1 filename_1.fastq -2 filename_2.fastq filename.sam -S -p 7
### bowtie is another program that can be used to index the H37Rv reference and map the fastq
files to the reference, resulting in a sam file
```

The resulting SAM file was manipulated with SAMtools and bcftools to generate the consensus sequence and the variant calling file (VCF).

```
samtools view -Shu filename.sam > filename.bam
```

```
### convert sam into bam, a binary version of bam format
```

```
samtools sort filename.bam filename.non-uni.sorted
```

```
### sort the data within the bam file
```

```
samtools rmdup -S filename.non-uni.sorted.bam filename.sorted.bam
```

```
### delete duplicated reads
```

```
samtools index filename.sorted.bam
```

```
### make an index file of the bam file
```

```
samtools faidx H37Rv
```

```
### make an index file of H37Rv
```

```
samtools mpileup -uf H37Rv filename.sorted.bam" > filename.pileup.bcf
```

```
### create a pileup file from the bam file
```

```
##### consensus sequence
```

```
bcftools view -cg filename.pileup.bcf > filename.pileup.vcf
```

```
### create a text version of the bcf file
```

```
perl vcfutils.pl vcf2fq filename.pileup.vcf > filename.fq
```

```
### calculate the consensus sequence or genome for your isolate
```

```
##### variant calling
```

```
bcftools view -vcg filename.pileup.bcf > filename.snp
```

```
### calculate what position contain SNVs.
```

From this point, the software used to combine the SNVs files from several isolates into a single table was custom built. MEGA was used to construct phylogenetic trees, and IGV was used to visualise alignments. Areas with low coverage, for instance PE-PPE repeat regions, were not included in the phylogenetic constructions.

2.15.7 Whole Genome Variant Analysis for Resistance Mutation Detection

2.15.7.1 Geneious R9 for Whole Genome Variant Analysis

Fastq files were imported into Geneious R9 software, where both 5' and 3' ends were trimmed based on a more than 5% chance of base call error. Paired-end reads were then mapped to the H37Rv reference genome [78]. For variant analysis, minimum variant frequency was set at 0.25, while maximum variant P-value was set at 10^{-6} and minimum strand bias at 10^{-5} when exceeding 65% bias. All variants (inside and outside coding regions, synonymous and non-synonymous) were extracted at a minimum read-depth of 5x, to a Microsoft Excel file from where they could be investigated further, both manually and using custom macros to search for candidate genes (n=23 for the Walker/Kohl *et al* algorithm, and a further 43 genes from databases such as TBDreamDB, the Broad Institute, and recent published literature) (Section 2.15.9). The SNVs were filtered for those that were non-synonymous with variant frequency over 90%. Putative novel mutations were subjected to a thorough online search of the literature.

2.15.7.2 Galaxy.org

Galaxy.org is a freely-available, open-source, web-based platform for data-intensive bioinformatics research. The Galaxy Team is a part of the Centre for Comparative Genomics and Bioinformatics at Pennsylvania State University, and the Department of Biology and at Johns Hopkins University. Similar to Geneious R9, it brings together many frequently-used software solutions for genomic analysis, and makes it more user-friendly to those biologists who are not familiar with computer science. If the web server is not working quickly enough, or the data being used is too large for the web-based Galaxy.org, then Amazon Web Services (AWS) Cloud computing can be used; 'instances' can be created in the cloud where analysis can then take place, subverting the need for a faster web-server or increased PC memory capacity.

Fastq files were imported using 'Get Data, Upload file'. Sequence quality was examined using FastQC (section 2.15.3). Sequences were trimmed using 'NGS: QC and manipulation', followed by repeat FastQC. 'FastQ joiner' was used to associate forward and reverse paired-end reads. Reads were mapped to H37Rv with Bowtie or BWA for Illumina® [70]. The output SAM file was analysed with SAMtools for variant calling.

2.15.8 Workflow analysis of Draft Genome Sequence of the first XDR-TB strain in Ireland using Linux Command Line and open-source software

This workflow was performed in September 2014 (Appendix 2).

Filenames and details:

IEXDR1_S2_L001_R1_001.fastq.gz (687.8 MB) 4889178 reads

IEXDR1_S2_L001_R2_001.fastq.gz (659.4 MB) 4889178 reads

IEXDR1 failed a number of FastQC metrics, most likely due to the presence of adaptor/index sequences, since adaptor trimming had not been set up as part of the original Miseq® run. Trimmomatic was used to trim any adapter/index sequences from the real sequence. All reads that matched Illumina® Nextera indices and adapters were removed by providing a file containing these sequences (NexteraPE-PE.fa).

The command line argument used was:

```
sudo      java      -jar      /usr/local/Trimmomatic-0.32/trimmomatic-0.32.jar      PE
IEXDR1_S2_L001_R1_001.fastq.gz      IEXDR1_S2_L001_R2_001.fastq.gz
IEXDR1_forwardP.fq.gz      IEXDR1_forwardUP.fq.gz      IEXDR1_reverseP.fq.gz
IEXDR1_reverseUP.fq.gz      ILLUMINA®CLIP:NexteraPE-PE.fa:2:30:10      LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15
```

Output filenames and details:

IEXDR1_forwardP.fq.gz (387.4 MB) 3482497 reads

IEXDR1_reverseP.fq.gz (390.1 MB) 3482497 reads

FastQC was performed on these smaller files which contained trimmed reads, and quality had improved, although sequence duplication and Per Base GC/Sequence content still 'failed'. Also, Sequence length distribution does not look as consistent as it did previously because of the trimming.

Reads were mapped to H37Rv using the following custom shell script which runs through all the BWA commands (explained in section 2.15.4 and 2.15.5):

```
#!/bin/bash
```

```
#read in values from command line
```

```
fastq1=$1
```

```
fastq2=$2
```

```
ref=$3
```



```

output=$4

#index the ref file
bwa index $ref

#make dir for mapping
mkdir $output.mapping

#map
bwa aln $ref $fastq1 > $output.mapping/F.sai
bwa aln $ref $fastq2 > $output.mapping/R.sai
bwa sampe -a 300 $ref $output.mapping/F.sai $output.mapping/R.sai $fastq1 $fastq2 >
$output.mapping/$output.sam

#create a sorted and indexed bam file
samtools view -b -S $output.mapping/$output.sam > $output.mapping/$output.tmp.bam
samtools sort $output.mapping/$output.tmp.bam $output.mapping/$output
samtools index $output.mapping/$output.bam

```

The output BAM file described the alignment of sequence reads to the reference, which was then input into SAMtools ‘mpileup’ (section 2.15.5) with a series of input options (e.g. –Q20 minimum base quality must have a phred score of 20) in a single custom command line argument which runs SAMtools mpileup on the BAM file, followed by conversion to a consensus sequence based on the base calls at each nucleotide position (vcfutils.pl perl script), followed by ‘awk’ command which changed the file header ‘@gi|444893469|emb|AL123456.3| *Mycobacterium tuberculosis* H37Rv complete genome/’ to ‘>gi|444893469|emb|AL123456.3| *Mycobacterium tuberculosis* H37Rv complete genome’ which changed the fastq file format to fasta:

```

samtools mpileup -DSugBf ../H37rv.fasta -Q20 -q30 -o40 -e20 -h100 -m2 -C50 -D100
IEXDR1.bam | bcftools view -cg - | perl /usr/share/samtools/vcfutils.pl vcf2fq | awk
'/@gi|444893469|emb|AL123456.3| Mycobacterium tuberculosis H37Rv complete genome/,/^+$/|
perl -pe "s/@/>/;s/\+//>" > IEXDR1cns.fasta

```

This resulted in a final file IEXDR1cns.fasta. Original reads were mapped to this consensus in order to confirm that they map successfully. Potential errors in mapping could be noted at this stage, mostly present around repetitive regions. It also enabled manual visualisation of ambiguous calls in the consensus sequence. Manual visualisation in a genome browser, although laborious, ensured confidence in the consensus sequence. Ambiguous regions were annotated or deleted.

ABACUS was used to ensure the correct order of contigs relative to the H37Rv genome and a program from the PAGIT suite – IMAGE – was used to close as many contig gaps as possible [119, 120].

Excerpt from XDR-TB paper (Appendix 2):

‘Whole-genome sequencing was performed to provide further molecular confirmation of IEXDR1 (lineage 2, East Asian, or Beijing strain). Genomic DNA was sequenced using an Illumina[®] Miseq[®]. Paired-end reads were mapped to the *M. tuberculosis* H37Rv reference genome (AL123456.3) by Burrows-Wheeler Alignment [117]. This yielded a mapped-read-depth of 196-fold, covering 97.6% of the H37Rv genome. A consensus sequence was called using the SAMtools mpileup command [71]. The IMAGE algorithm was employed to extend contigs and close gaps in the assembly producing a final draft assembly [119].’

2.15.9 Algorithm developed by Walker and Kohl et al for resistance mutation detection in 23 candidate genes

Geneious R9 software was used to produce Microsoft Excel files containing all variants present for each isolate (as in Section 2.15.7.1 above). A custom macro was designed to extract all variants from 23 candidate genes (*ahpC, fabG1, inhA, katG, ndh, rpoB, embA, embB, embC, embR, iniA, iniC, manB, rmlD, pncA, rpsA, gyrA, gyrB, rpsL, gidB, rrs, tlyA and eis*) associated with resistance (including single nucleotide variations, indels, and substitutions). The details of the macro script are as follows:

```
### find all genes present in the variant file of each isolate using a custom list, and transfer them to a new worksheet for further analysis
```

```
Sub GeneFinder()
```

```
Dim srchLen, gName, nxtRw As Integer
```

```
Dim g As Range
```

```
'Clear Sheet 2 and Copy Column Headings
```

```
Sheets(2).Cells.ClearContents
```

```
Sheets(1).Rows(1).Copy Destination:=Sheets(2).Rows(1)
```

```
'Determine length of Search Column from Sheet3
```

```
srchLen = Sheets(3).Range("A" & Rows.Count).End(xlUp).Row
```

```
'Loop through list in Sheet3, Column A. As each value is
```

```
'found in Sheet1, Column I, copy it to the next row in Sheet2
```

```
With Sheets(1).Columns("I")
```

```
For gName = 2 To srchLen
```

```
Set g = .Find(Sheets(3).Range("A" & gName), lookat:=xlWhole)
```

```

    If Not g Is Nothing Then
firstAddress = g.Address
    Do
g.Value = .Find(Sheets(3).Range("A" & gName), lookat:=xlWhole)
Set g = .FindNext(g)
nxtRw = Sheets(2).Range("E" & Rows.Count).End(xlUp).Row + 1
    g.EntireRow.Copy Destination:=Sheets(2).Range("A" & nxtRw)
    Loop While Not g Is Nothing And g.Address <> firstAddress

    End If
Next
End With
End Sub

```

Mutation types included benign, phylogenetic, uncharacterised, homoplasic, previously seen on the literature, resistance-determinants. Filtering was performed manually according to an algorithm designed by Walker and Kohl *et al* (232 resistance-determinants) and mutations collated [61].

2.15.10 TB Profiler freely available online web-tool for analysis of MTBC genomes

Raw fastq files from the MDR/XDR-TB cohort were uploaded to TB Profiler web-tool and results downloaded and collated [77]. One isolate (R1 and R2 fastq) could be uploaded at one time. An internet connection was required to connect to <http://tbdr.lshtm.ac.uk/>. A report detailing the lineage, main spoligotype and Region of Difference genotyping results, 14 drugs and drug-resistance-associated mutations (if found within the gene or its promoter region), and a list of other mutations in candidate genes, was produced, which could be printed. An MDR/XDR prediction was also made. A Microsoft Excel file was used to collate the mutation data for the MDR/XDR-TB cohort.

2.15.11 PhyResSe freely available online web-tool for analysis of MTBC genomes

Raw fastq files from the MDR/XDR-TB cohort were uploaded to PhyResSe web-tool and results downloaded and collated [76]. A batch of fastq files could be uploaded at once. An internet connection was required to connect to <http://bioinf.fz-borstel.de/mchips/phyresse/>. It was possible to back up the session with a secure 'key'. A number of open-source programs were used (FastQC, Qualimap, SAMtools) along with custom scripts which could be visualised by the end user if desired. Results detailing the read quality, mapping quality, lineage and variant calling file with resistance mutations associated with each drug highlighted, could be downloaded by the user, in

different file formats, depending on user needs. A Microsoft Excel file was used to collate the mutation data for the MDR/XDR-TB cohort. The resistance catalogue is regularly updated as new information becomes available. Version 27 was used for this study.

2.15.12 Comparison of Online Web-tools for the Prediction of TB Drug Resistance

PhyResSe and TB Profiler were compared under numerous headings such as user-friendliness, speed, drugs included, genes investigated, genotypic drug resistance prediction and lineage calling, and results tabulated.

2.15.13 ReseqTB data-sharing platform mutation database

The same Geneious R9 variant files that were used to analyse the Walker and Kohl *et al* algorithm (section 2.15.9) were used to collate drug-resistance-associated mutations in the MDR/XDR-TB cohort according to the ReSeqTB resistance mutation catalogue [121]. The ReseqTB platform was not yet operational, therefore the catalogue was downloaded and mutations searched for manually. A Microsoft Excel file was used to collate the mutation data for the MDR/XDR-TB cohort.

2.16 Comparison of WGS versus conventional DST for anti-tuberculous drug resistance detection in an MDR/XDR-TB cohort

Genotypic resistance prediction, performed using PhyResSe, TB Profiler, ReSeqTB, Walker/Kohl *et al* algorithm, and Hain GenoType LPAs, was compared to phenotypic DST results and discrepancies analysed [61, 76, 77, 121-123]. Even though study numbers were low, and therefore confidence intervals were inevitably high, it was seen as a good comparison exercise to use sensitivity and specificity for each drug, or class of drug, using each tool, compared to the reference standard DST.

2.17 MGITTM Pilot Study pipeline for identification of all mycobacteria, and resistance profiling and Nearest-Neighbour Relatedness Analysis of MTBC

Figure 19 displays a simplified version of the pilot study design, and a more in-depth version of the bio-informatics workflow designed by the COMPASS-TB group. Illumina[®] MiSeq[®] sequencing runs were performed in the TrinSeq laboratory, Institute of Molecular Medicine, Trinity College, Dublin, and results shared via BaseSpace with the Oxford MMM group for analysis. The data was downloaded and matched to the submitted plate generator file, from where it could automatically enter the analysis pipeline.

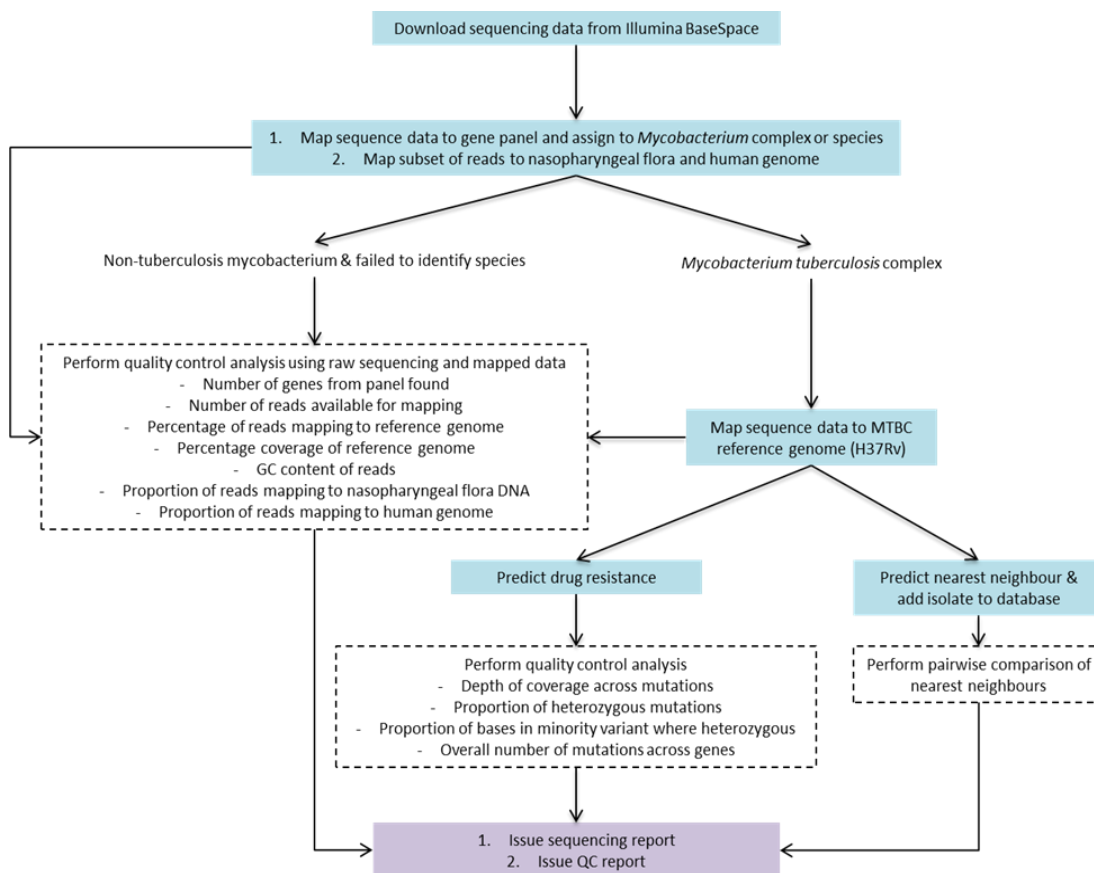
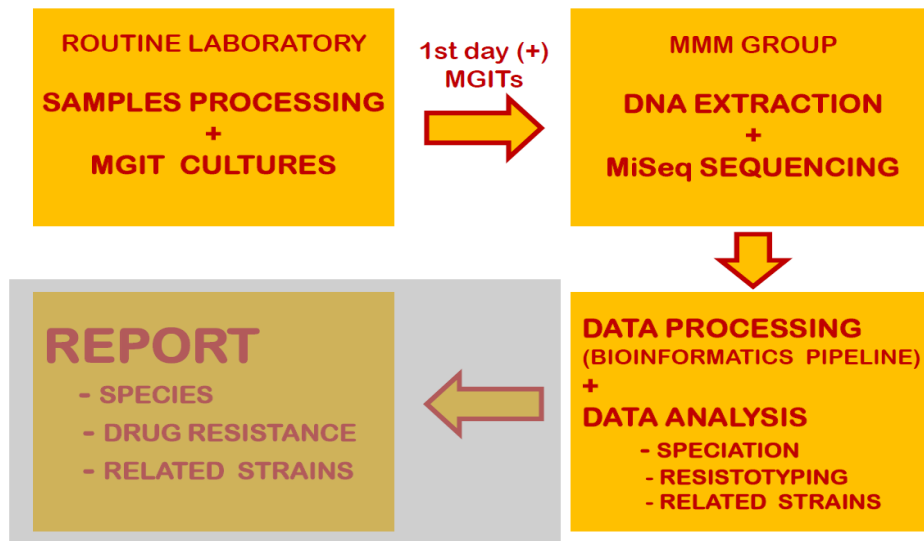


Figure 19. Representations of the workflow proposed and implemented by the COMPASS-TB group as part of the MGIT Pilot Study

Two flow diagrams of the COMPASS-TB WGS data processing pipeline. The top diagram is a simplified workflow. The lower diagram details the bio-informatics workflow in more detail. Light blue areas - WGS processing, purple - reporting, no shading - manual processing steps. QC = Quality control

In order to assess DNA contamination, the first 50,000 reads of every isolate were mapped with Bowtie (v 2.2.0) to the human genome (GRCh37/hg19) and nasal and mouth flora from the National Institutes of Health (NIH) Human Microbiome Project. Raw read and mapping data were used to assess data quality: percentage GC content (~65% in mycobacteria), median insert size, reads available for mapping, number of reads successfully mapped to H37Rv, percentage of reads mapped to H37Rv, percentage coverage of the TB reference genome and percentage mapped to the human genome. For instance, if GC content and mapping to the reference genome were both low, this could indicate lower quality sequencing data due to higher levels of contamination with human and/or nasopharyngeal flora (NPF). If GC content was 65%, and other indicators were satisfactory but the percentage reads mapped to H37Rv were low, a non-tuberculous mycobacterium species was probably present. If GC content was acceptable, but reads available for mapping were < 1 million, and the percentage of the genome covered was too low, the isolate should be repeated. Repetition of extraction was not possible for this study, since only one aliquot was taken on Day 0/1 for sequencing, and the remainder used for conventional culture and susceptibility. However, repetition from the library preparation stage was possible.

A gene presence/absence algorithm was used to identify mycobacterial species or MTBC, using 'normal' parameters or 'relaxed' parameters (where sequencing read quality was low). A group of bioinformaticians and biologists in the Oxford MMM group designed this algorithm using clustered informative genes derived from a large catalogue of commercially-available reference strains they had whole genome sequenced, or downloaded from NCBI (n=169). For example, 100 genes identified MTBC, 33 genes identified *M. avium* Complex and 100 genes were associated with *M. goodnae*. These clustered genes were used to construct a 'mycobacterium pangenome', against which raw reads could be mapped.

If MTBC was identified, the reads were mapped to H37Rv with Stampy [118]. Once mapped, the MTBC genome was interrogated for the presence of a catalogue of drug-resistance-associated mutations which had been derived from a search of the available literature and consultation with experts in the field [76], and Hain GenoType MTBDR*plus* v2.0 v2.0 and MTBDR*sl* mutations (138 resistance determinants altogether, for isoniazid, rifampicin, ethambutol, pyrazinamide, streptomycin, fluoroquinolones and aminoglycosides) [114]. A read-depth of at least 5 was required for a sensitive or resistant prediction to be made. Where a mixed call was made (heterozygosity), at least 5 reads in both directions were required, and the minority base call had to be present in at least 10% of calls. Where excessive numbers of mutations were seen in a gene, resistance prediction could not be reliably performed for the corresponding drug.

The MMM Group has collected a large database of MTBC genomes from previous studies (2,191), which constitutes a set of reference strains, along with those which are publicly available [112].

These genomes were each aligned to the reference genome (H37Rv, AL123456.3, NC000962.3) and variant files extracted and compiled. Maximum likelihood phylogenetic trees could be built from their variant files. This process is computationally intensive and time-consuming. It would not be feasible to build this tree every time a new genome became available. Instead, bioinformaticians from the MMM Group designed a rapid nearest-neighbour algorithm, 'Squirrel Walk', which could be used to interrogate each new genome that was sequenced, in real-time. The compiled variant file from the 'reference' maximum likelihood tree was used to extract a database of SNV differences between each node of the tree. The algorithm reported all matches within 20 SNVs of the queried isolate, or the single closest match if all reference isolates were > 20 SNVs apart. If recent, or possible, transmission was suspected (i.e. < 5 SNV or 6-12 SNV difference between strains, respectively), follow-on whole genome cluster analysis was performed (section 2.15.6.2). New variants between the queried isolate and its adjacent node were stored in a 'bucket' at the node, enabling new isolate data to be iteratively added to the Squirrel Walk database. This avoided further computationally intensive phylogenetics, although sub-clades of the tree would have to be re-built if the 'buckets' became over-populated (this was not necessary over the course of the pilot study). The report included the origin of the nearest-neighbour reference strain. Identification, drug susceptibility, and nearest-neighbour relatedness results were reported to collaborators as it would be envisaged if the pilot study were to be rolled out (Figure 20).

Once the pilot study time-frame was complete, data was collated and analysed by the MMM Group using Stata 13.1 (StatCorp, USA), using routine methods as the reference standard. Specimens were anonymised. Data submitted from the IMRL on each isolate included: sample collection date, date MGIT™ vial reached its positivity threshold, heat-inactivation date, identification date, DST completion date, drugs tested and susceptibility profile and MIRU-VNTR genotyping completion date. WGS data collected included: sample extraction date, NGS performed date, date results were shared via BaseSpace, identification date, drug resistance prediction date, nearest-neighbour matching date and report complete date. A costing analysis was performed with selected UK sites only, which included completion of a comprehensive micro-costing questionnaire.

Whole Genome Sequencing Report from MGIT Positive Samples

Not for diagnostic use


Sample Details			
Sequencing Location	Dublin	Date Received in Lab	1 st Nov 2013
Local LIMS Specimen ID	IMRL15	Run Date	7 th Jan 2014
GUID			

Sample/Sequencing Quality	
Comments	

Organism Identification
Mycobacterium tuberculosis

Resistotype				
Drug	Prediction	HAIN Mutation	Extended Catalogue	Ambiguous
Isoniazid	Resistant	katG_S315T		
Rifampicin	Resistant	rpoB_S450L		
Ethambutol	Resistant	embB_M306V		
Pyrazinamide	Sensitive			
Streptomycin	Sensitive			
Moxifloxacin	Sensitive			
Amikacin	Sensitive			

Relatedness			
Nearest neighbour(s)			Genealogy
GUID	No. of SNPs Apart	Centre	
C00018007	84	Birmingham	

Authorised	
Signature: 	Print name: Timothy Walker
Position:	Date: 13 January 2014

QC Report

Sample Details			
Sequencing Location	Dublin	Date Received in Lab	1 st Nov 2013
Local LIMS Specimen ID	IMRL15	Run Date	7 th Jan 2014
GUID			

Sequencing run Statistics	
GC Content	60%
Median Insert Size	223
Reads For Mapping	5934172
Reads Mapped to TB Reference	4569883
Percentage Mapped to TB Reference	77%
Coverage of TB Reference	91.8%
Percentage Mapped to Human Genome	23%
QC comment	

Figure 20. Example of the report format (with QC report) envisaged by the MGIT Pilot Study collaborators (IMRL15).

For IMRL15, WGS analysis identified Mycobacterium tuberculosis that is resistant to isoniazid, rifampicin and ethambutol (MDR-TB) due to resistance-associated mutations found in 23 candidate genes from a catalogue of high- and low-confidence mutations. These all happen to be mutations included in Hain GenoType LPAs. The genome of this isolate is 84 SNVs from an isolate previously found in Birmingham, so no outbreak was flagged here. GC content was 60%, 91.8% of the reference genome was covered, and there was 23% human DNA contamination found.

Chapter 3.

Molecular Epidemiology in Ireland, 2010-14

3 Molecular Epidemiology of MTBC in Ireland, 2010-14

3.1 Introduction

Tuberculosis prevalence in Ireland, although low, has remained between 7.9 and 11.0 per 100,000 inhabitants over the last decade [22]. This has, to some extent, been influenced by inward human migration from higher prevalence countries [22]. Disruption of transmission chains is key to controlling tuberculosis both at a national and international level. ‘Know your enemy’ and you will defeat it¹. Genotyping is a powerful tool in the fight against TB. *Mycobacterium tuberculosis* is a relatively stable clonal organism that is well-suited to molecular genotyping methods. Modern strains are thought to have evolved with humans about 2.5 million years ago (Figure 11) [124]. Seven global lineages have been elucidated, which are broadly geographically-based: Indo-Oceanic Lineage 1, East Asian lineage 2, East African Indian lineage 3, Euro-American Lineage 4, West African I and II (Lineage 5 and 6), and Ethiopian Lineage 7 (Table 1 and Figure 11). *Mycobacterium bovis* spp. fits with West African clades [125]. Genotyping can shed light on whether an outbreak has occurred, whether a patient has relapsed or is newly re-infected, or whether laboratory cross-contamination has occurred and the isolate is falsely positive. Geographical distribution of strains can be tracked using genotyping and correlation of the impact of genotype on resistance, fitness, and virulence is also possible. Numerous MTBC genotyping methods have been employed in different settings to date: Insertion sequence IS6110 fingerprinting using Restriction Fragment Length Polymorphism (RFLP), Spacer Oligonucleotide typing of the Direct Repeat DR region (spoligotyping), deletion mapping for Regions of Difference (RoD) or Long-Sequence PCR (LSP), Single Nucleotide Polymorphism Multi-locus Sequence Typing (MLST using a set of SNVs) and whole genome MLST. Each method has its own strengths and weaknesses [126]. Insufficient resolution, technical difficulties and challenges with standardisation have been experienced using some of these methods [126, 127]. Automated 24-locus MIRU-VNTR genotyping, while not without intra-laboratory variation, is seen as the prospective surveillance method of choice [126, 128].

In November 2009, MIRU-VNTR genotyping was introduced at the Irish Mycobacteria Reference Laboratory (IMRL). This typing method relies on the principle that different lineages of MTBC contain characteristic numbers of tandem repeat regions at various loci in the genome [129]. Twenty-four informative genomic loci are amplified and repeats at each locus enumerated (see Figure 9). This results in a portable 24-digit genotype that can be used for further phylogenetic analysis and worldwide comparison. Genotyping has been found to complement and augment Public Health surveillance and contact tracing by recognising potential outbreaks (clusters), sometimes before epidemiological links have been established [130, 131]. In Ireland, lineage details are uploaded to the national Computerised Infectious Disease Reporting database (CIDR),

¹ Sun Tzu, The Art of War

which is then linked to the patient epidemiological data collected by public health specialists. Anonymised national CIDR data is collated and reported by the Health Protection Surveillance Centre (HPSC) quarterly.

Enhanced surveillance of TB began in Ireland in 1998. An amendment to the Infectious Disease Regulations 1981, in January 2004, made ‘outbreaks, unusual clusters, or changing patterns of illness statutorily notifiable by medical practitioners and clinical directors of laboratories’ from that time onwards [22]. The national TB clustering rate is seen as a measure of the TB control program. The lower the rate of clustering, and the smaller the cluster size, the less transmission is occurring, therefore better TB control is evident. The IMRL definition of a cluster is two or more isolates with identical 24-locus MIRU-VNTR genotypes. The HPSC definition of an outbreak is the occurrence of cases of active TB disease above the expected level over a given time period (6 months) in a geographical area, facility, or within a specific population group [57].

A molecular epidemiology study performed in the southwest of the country was undertaken in 2010, looking at 171 strains of MTBC (collected 2004-06) focussed on the urban centre of Cork, using spoligotyping and 24-locus MIRU-VNTR genotyping [105]. The most common spoligotypes detected were ST0137 (X2, 16.9%) and ST0351 (U, 15.8%). Spoligotype ST351 (which corresponds to Euro-American lineage 4) was the most common spoligotype found among Irish-born patients. With a combination of the two genotyping methods, they found fifteen clusters containing 47 isolates (27.5% clustering). They observed that patients were more likely to be linked with a cluster if they were Irish-born and under the age of 55 years. Genotyping did not match epidemiological data in all cases.

In a recent publication from the IMRL, the first analysis of the population structure of tuberculosis in the Ireland (2010-11) was reported (Appendix 1) [106]. The analysis yielded four global lineages of MTBC (see Table 3). The majority (63%) belonged to Euro-American lineage 4; while Indo-Oceanic lineage 1, East Asian lineage 2, and East African Indian lineage 3 represented 12%, 12%, and 13% of isolates, respectively. Sub-lineages H37Rv (20%), Haarlem (21%) and LAM (22%) were most prevalent among Euro-American lineage 4 strains. One large East Asian lineage 2 cluster was identified whilst smaller clusters of isolates were identified amongst Euro-American lineage 4. Four multi-drug resistant tuberculosis (MDR-TB) cases, representing East Asian lineage 2 (Beijing) and Euro-American lineage 4, were identified during this period. Rates of mono-resistance to either isoniazid or streptomycin were low at 5%.

The original study provided a “snapshot” of the genetic diversity of *M. tuberculosis* in Ireland. The present study aimed to build on this work, proving that prospective genotyping is essential for TB

GLOBAL LINEAGE	SUB-LINEAGE	NO. OF ISOLATES 2010-11	ISOLATES %	NO. OF ISOLATES 2010-2014	ISOLATES %	
1	INDO-OCEANIC	EAST AFRICAN INDIAN	42	12	145	11.1
2	EAST ASIAN	BEIJING	45	12	117	9
3	EAST AFRICAN INDIAN	DELHI/CENTRAL ASIAN	45	13	128	9.8
4	EURO-AMERICAN	LINEAGE 4 TOTAL	229	63	863	66.1
		SUB-LINEAGE UNSPECIFIED	11 (4.8%)		378 (43.8%)	
		LATIN AMERICAN MEDITERRANEAN	52 (22.7%)		155 (18%)	
		HAARLEM	47 (20.5%)		207 (24%)	
		H37RV	44 (19.2%)		7 (0.8%)	
		HAARLEM/X	29 (12.7%)		N/A	
		CAMEROON	13 (5.8%)		39 (4.5%)	
		S	6 (2.6%)		18 (2.1%)	
		TUR	8 (3.5%)		9 (1%)	
		X	5 (2.2%)		26 (3%)	
		GHANA	3 (1.25%)		0	
		URAL	3 (1.25%)		10 (1.2%)	
		UGANDA I+II	6 (2.6%)		6 (0.7%)	
		NEW-1	2 (0.9%)		8 (0.9%)	
5 + 6	WEST AFRICAN I+II	WEST AFRICAN I+II	0		11	0.9
-	BOVIS	BOVIS	N/A		41	3.1
7	ETHIOPIAN	-	0		0	0
TOTAL			361	100	1305	100

Table 3. Distribution of global and sub-lineages lineages among *Mycobacterium tuberculosis* Complex isolates in Ireland, 2010-11 (n=361) compared to 2010-2014 (n=1,306).

Even though there are almost four times more isolates from 2010-14 as there were from 2010-11, the distribution remains largely similar. Euro-American Lineage 4 still predominates. Higher diversity of strains has, however, been found within the larger cohort.

surveillance and flagging of outbreaks, by analysing MTBC MIRU-VNTR genotyping data over a longer time-frame, from 2010-2014 inclusive.

3.2 Results

3.2.1 Samples included in the Study

1,305 isolates, collected between January 2010 and December 2014, and received in the IMRL, were included in the study (5 years in total). Twelve isolates were genotyped by the Scottish Mycobacteria Reference Laboratory (SMRL). MIRU-VNTR data received from Northern Ireland for the time-period was also included (n=67, received in the Northern Ireland Mycobacteria Reference Laboratory, Belfast, January 2010 – December 2013). Disregarding Belfast isolates, for which no clinical details were available, the median age, at date of collection, was 39 years (IQR 29-54). Appendix 1 includes the published article on the molecular epidemiology of MTBC in Ireland, 2010-2011.

3.2.2 Mixed MTBC Infection

No mixed infection was observed in this cohort, although MIRU-VNTR double alleles were present. Two double alleles were detected in 3 isolates (2 Euro-American and 1 Bovis strain) and one double allele in 9 isolates (7 Euro-American, 1 Bovis and 1 EAI strain).

3.2.3 Health Protection Surveillance Centre TB Data, 2010-14

1,904 TB cases were reported by the HPSC from 2010 to 2014 inclusive. Of these, 1,340 were reported culture positive. This indicates that 97.4% of culture-positive cases were represented in the current study. The crude prevalence rate ranged from 7.1 to 9.2 per 100,000. The percentage of foreign-born cases ranged from 40.7% in 2010 at its lowest to 46.7% at its highest, in 2011. India, Pakistan, Nigeria and the Philippines were the countries from where most non-Irish patients originated. The average male to female infection ratio was approximately 3:2. On average, 61.6% involved pulmonary disease alone, 33.1% extra-pulmonary disease alone, and 20.5% had both pulmonary and extra-pulmonary components. When extra-pulmonary disease was involved, it was most commonly in lymphatic, extra-thoracic, and pleural body sites. Thirty-six outbreaks were reported by the HPSC; 5 in the community, 9 in non-residential institutions, 2 in residential institutions, 8 within private family homes, 9 across extended families and 3 across more than one location (Table 4). Non-residential institutions included schools/universities/colleges, workplaces and public houses. Residential institutions included prisons and healthcare facilities.

Year	No. of TB Outbreaks	Active Cases	LTBI Cases	Community	Non-Residential Institution	Residential Institution	Family Private House	Family Extended	Across >1 Location
2010	7	41	60	1	4 (3 Schools, 1 Workplace)	-	-	2	-
2011	5	42	15	-	2 (1 School, 1 Public House)	1 Prison	-	2	-
2012	7	24	4	1	-	-	4	1	1
2013	12	41	155	2	2 (1 University/ College, 1 Workplace)	1 Healthcare Facility	2	3	2
2014	5	16	10	1	1 Public House	-	2	1	-

Table 4. HPSC outbreak data reported from 2010 to 2014 inclusive.

Details include how many latent TB cases were associated with the outbreak, and whether the outbreak occurred in the community, non-residential institutions such as public houses or universities, or residential settings like prisons or healthcare facilities. Most outbreaks occurred in 2013 (n=12) where there were 155 LTBI cases also found.

3.2.4 MTBC Lineage Distribution in Isolates collected in Ireland, 2010-14

Lineages present included 66.1% Euro-American lineage 4 (n=863), 11.1% Indo-Oceanic lineage 1 (n=145), 9.8% East African Indian lineage 3 (n=128), 9% East Asian lineage 2 (n=117), 3.1% Bovis (n=41) and 0.9% West African I lineage 5 (n=6) and II lineage 6 (n=5). No Ethiopian Lineage 7 isolates were found. The lineage distribution of all strains is visualised in a neighbour joining phylogenetic tree in Figure 21. When Euro-American lineage 4 was categorised into sub-lineage, 43.8% Euro-American (sub-lineage not defined, n=378), 24% Haarlem (n=208), 18% LAM (n=155), 4.5% Cameroon (n=39), 3% X (n=26), 2.1% S (n=18), 1.2% Ural (n=10), 1% TUR (n=9), 0.9% New-1 (n=8), 0.8% H37Rv (n=7), 0.7% Uganda I and II (n=6) were found. The lineage distribution comparison between the 'snapshot' time period (2010-11) and the current study period (January 2010-December 2014 inclusive) is described in Table 3.

3.2.5 MTBC Clusters identified

Between 2010 and 2014, 152 clusters were identified (n=551) in the IMRL, while the remainder were unique. The overall clustering rate, therefore, was 42.2%. The break-down and size of these clusters is shown in Table 5. The largest number of unique clusters occurred within the Euro-American lineage 4 (n=421; Euro-American, sub-lineage un-defined [n=47], followed by Haarlem [n=25] and LAM [n=16]). Clusters consisted most commonly of two isolates (n=85 clusters), followed by three (n=29) and four isolates (n=15) (median cluster size = 2, IQR 2-3.25). The five largest clusters occurred in Euro-American lineage 4, sub-lineage Haarlem (n=33), followed by LAM (n=28 and n=25), East Asian lineage 2, sub-lineage Beijing (n=20), and Indo-Oceanic lineage 1, sub-lineage EAI (n=11). Figures 22-25 visualise the four largest outbreak clusters on a minimum-spanning phylogenetic tree (MSP). Clusters that contained both IMRL and Northern Ireland isolates were found within EAI, Delhi/CAS, Euro-American (sub-lineage un-defined), Haarlem and H37Rv sub-lineages. Isolates from a previous publication based on UK MIRU-VNTR genotyping (n=264, collected 2010-2013) were compared with the current cohort in order to investigate if transmission was evident across the Irish sea [112]. Three isolates with identical MIRU-VNTR genotypes to a UK cluster, related to an ethnic group, were found. No other links were found to UK genotypes.

3.2.6 Anti-TB Drug Susceptibility Patterns Observed

Susceptibility data was available for 1,192 (89%) isolates. Susceptibility data was not available for the Belfast isolates (n=67) and some isolates from external laboratories, who sent MTBC isolates to the IMRL solely for MIRU-VNTR genotyping (n=43). Three isolates failed to grow for DST, or were mixed or contaminated. Mono-resistance found is detailed in Table 6. Percentages ranged from 0.08% in rifampicin to 4.9% in isoniazid.

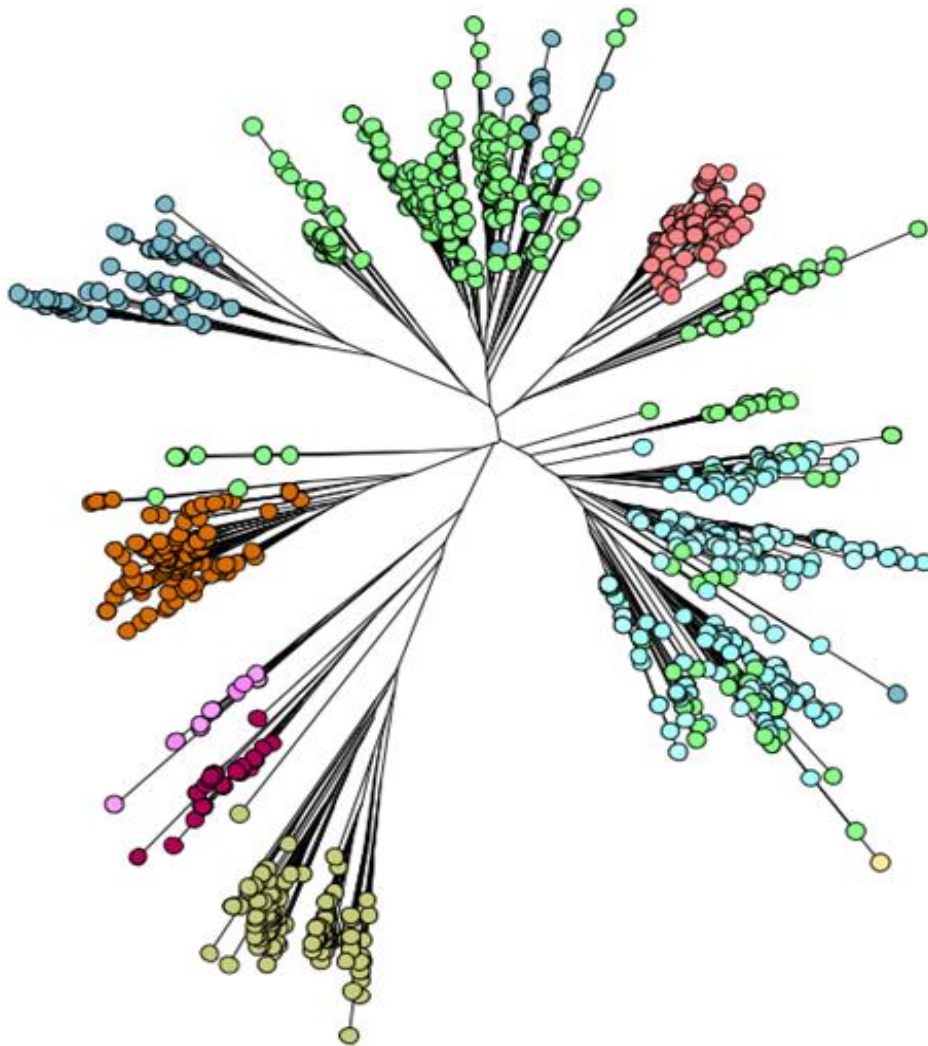


Figure 21. Categorical Radial Neighbour Joining dendrogram of 24-locus MIRU-VNTR genotypes collected in Ireland from January 2010 to December 2014 (n = 1305)

Created using MLVA Compare software which links to the MIRU-VNTRplus online database. Single yellow isolate, bottom right of figure, *Mycobacterium canetti* (similar to Euro-American Lineage 4).

Euro-American lineage 4 predominates within these isolates. 6/7 global lineages are present.

- Bovis
- Indo-Oceanic Lineage 1, sub-lineage EAI
- West African I and II
- Euro-American Lineage 4, sub-lineage Haarlem
- Euro-American Lineage 4, sub-lineage unspecified
- East Asian Lineage 2 sub-lineage Beijing
- Euro-American Lineage 4, sub-lineage LAM
- East African Indian Lineage 3, sub-lineage Delhi/CAS

MIRU-VNTR Genotyping Cluster Size (n=)	Euro-American	Haarlem	LAM	Beijing	Delhi/CAS	EAI	Bovis	Cameroon	X	S	New-1	H37RV	TUR	Ural	TOTAL CLUSTERS OF EACH SIZE
2	25	12	10	5	8	5*	5	3	4	2	2	2	1	1	85
3	8	5	2	3	-	2	1	3*	-	2	1	1*	1	-	29
4	8	2	-	1*	2*	1*	-	-	-	-	-	-	-	1	15
5	2	2	2	-	1	-	-	-	-	-	-	-	-	-	7
6	1*	-	-	-	-	-	-	-	-	-	-	-	-	-	1
7	2	1	-	2	-	-	-	-	1	-	-	-	-	-	6
8	1*	1	-	-	-	-	-	1*	-	-	-	-	-	-	3
10	-	1	-	-	-	-	-	-	-	-	-	-	-	-	1
11	-	-	-	-	-	-	1	-	-	-	-	-	-	-	1
20	-	-	-	1*	-	-	-	-	-	-	-	-	-	-	1
25	-	-	1*	-	-	-	-	-	-	-	-	-	-	-	1
28	-	-	1*	-	-	-	-	-	-	-	-	-	-	-	1
33	-	1*	-	-	-	-	-	-	-	-	-	-	-	-	1
TOTAL NO. OF CLUSTERS	47	25	16	12	11	8	7	7	5	4	3	3	2	2	152

Table 5. Total number, and size break-down, of MIRU-VNTR genotyping clusters present in Ireland from 2010-2014

* isolates from these clusters were whole genome sequenced. The most common cluster size was n=2, most clusters occurred within Euro-American lineage 4 (which includes Euro-American, Haarlem, LAM Latin American Mediterranean, Cameroon, X, S, New-1, H37RV, TUR and Ural). The largest clusters were seen in Haarlem, LAM and Beijing strains.

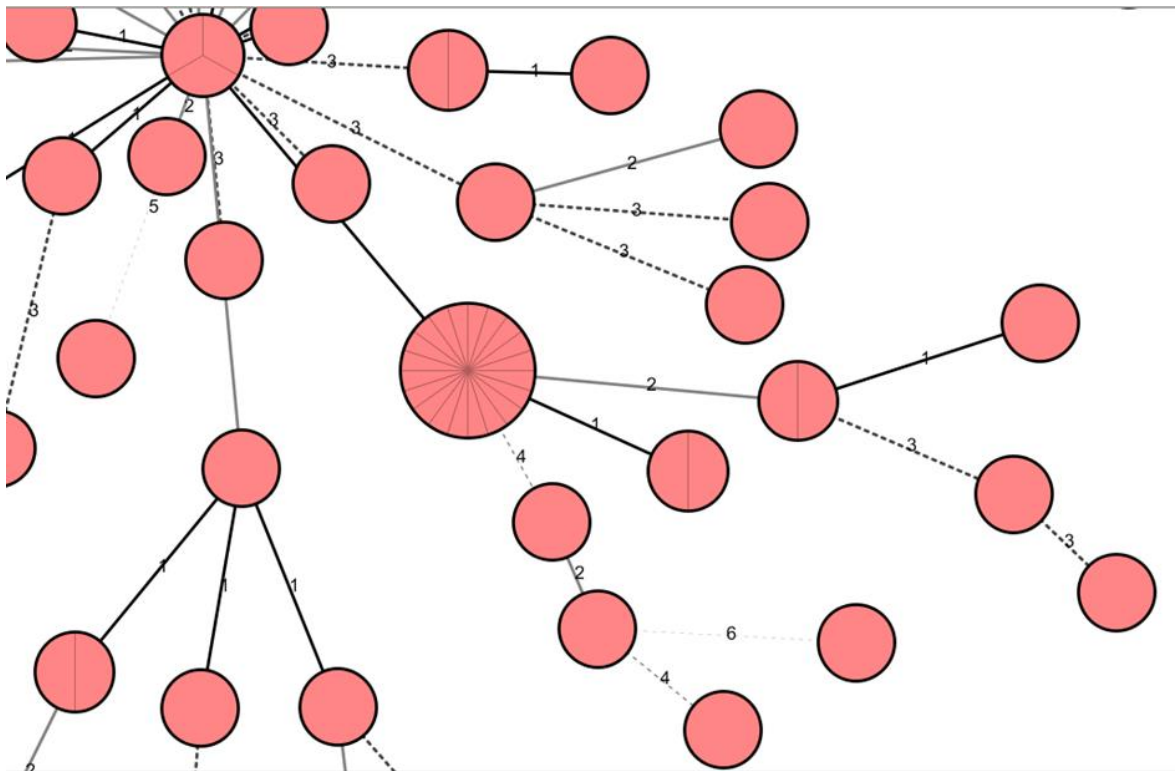


Figure 23. Zoomed area of the Beijing Minimum Spanning Tree showing the largest Beijing outbreak within the 2010-14 cohort, Cluster 10 (n=20), using MLVA Compare software

Circles represent isolates with a certain MIRU-VNTR genotype. The larger the circle, the more isolates with that genotype are found. Coloured segments represent individual cases with identical 24-locus MIRU-VNTR genotypes. Full black lines represent one SLV difference between connected cases. Lighter black lines represent two SLV differences. Broken lines represent 3 and 4 further SNV differences and so on.

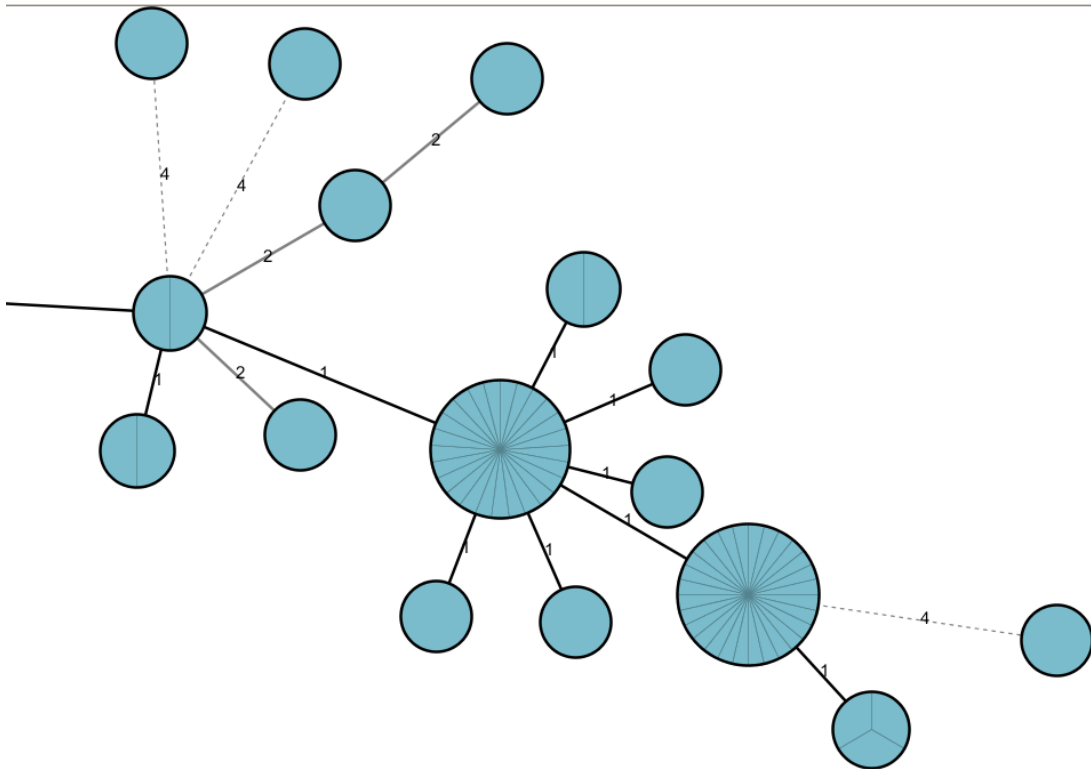


Figure 24. Zoomed area of the LAM lineage distribution (2010-14) visualised in a Minimum Spanning Tree showing the largest LAM outbreaks within the cohort, using MLVA Compare software.

Circles represent isolates with a certain MIRU-VNTR genotype. The larger the circle, the more isolates with that genotype are found. Coloured segments represent individual cases with identical 24-locus MIRU-VNTR genotypes. Full black lines represent one SLV difference between connected cases. Lighter black lines represent two SLV differences. Broken lines represent 3 and 4 further SNV differences and so on. Two large clusters (n=25, n=28) which are 1 SLV apart are clearly visible (Cluster 2a and 2b).

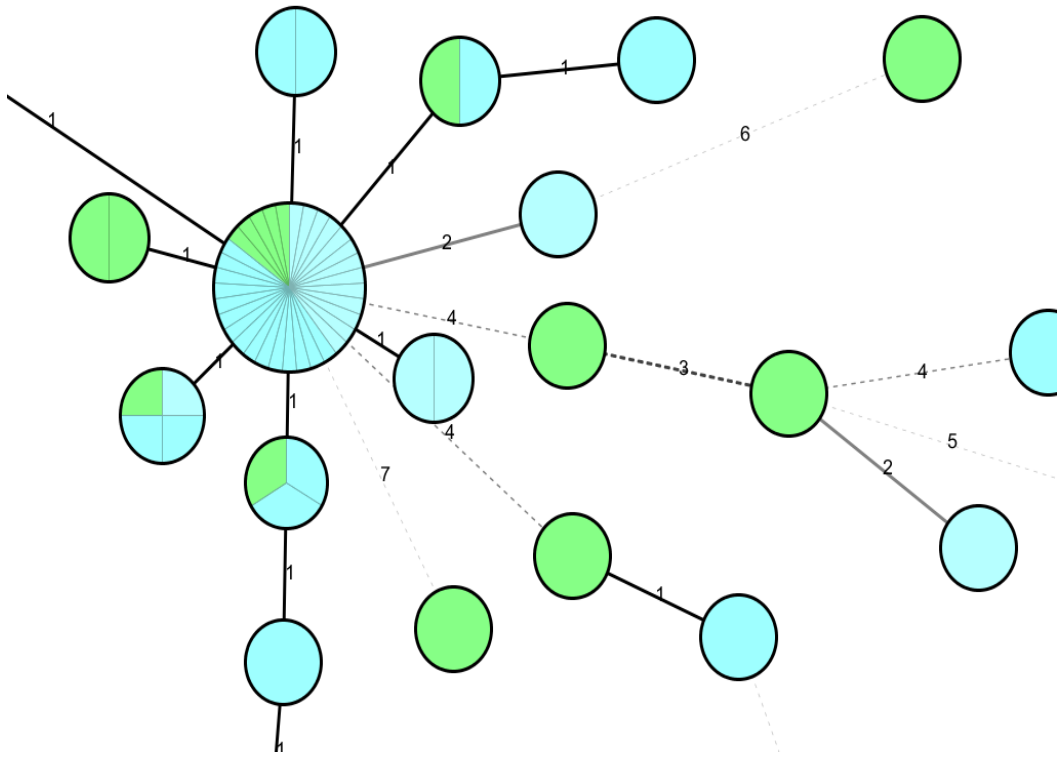


Figure 25. Zoomed area of the Euro-American sub-lineage Haarlem distribution (2010-14, Cluster 1) visualised in a Minimum Spanning Tree, built using MLVA Compare software.

Circles represent isolates with a certain MIRU-VNTR genotype. The larger the circle, the more isolates with that genotype are found. Coloured segments represent individual cases with identical 24-locus MIRU-VNTR genotypes. Full black lines represent one SLV difference between connected cases. Lighter black lines represent two SLV differences. Broken lines represent 3 and 4 further SNV differences and so on. One large cluster is visible (n=33).

DRUG	NO. MTBC ISOLATES	NO. MONO-RESISTANT	MONO-RESISTANT %	95% CI RANGE
SM	1153	36	3.1	2.1 - 4.1
INH	1192	58	4.9	3.7 - 6.1
RIF	1192	1	0.08	0 - 0.2
EMB	1190	3	0.3	0 - 0.6
PZA	660	19	2.9	1.6 - 4.2

Table 6. Mono-resistance seen in MTBC isolates collected between January 2010 and December 2014.

SM – streptomycin, INH – isoniazid, RIF, rifampicin, EMB – ethambutol, PZA – pyrazinamide, CI confidence interval. Mono-resistance to any drug did not exceed 5%.

Eighteen MDR-TB strains were isolated in the IMRL, collected between 2010 and 2014, two of which were further identified as XDR-TB. One further isolate, from an Irish-born patient, was rifampicin mono-resistant but clinically suspected of MDR-TB (Haarlem lineage). From 2010-14, the HPSC reported sixteen MDR-TB cases, all of whom were born outside Ireland. MDR/XDR-TB isolates were found within the following lineages: East Asian lineage 2 (Beijing, n=6), Euro-American lineage 4 (LAM [n=5], Cameroon [n=1], Ural [n=2], Euro-American un-defined lineage [n=2]) and East African Indian lineage 3 (Delhi/CAS, n=2). Three MDR-TB, and two MDR/XDR-TB clusters were identified: MLVA MtbC 15-9 genotypes 94-32 (Beijing), 8958-32 (Delhi/CAS), 163-15 (Ural) were MDR clusters, while 100-32 (Beijing) and 843-52 (LAM) were mixtures of MDR- and XDR-TB strains. Clusters were not greater than two isolates in size. One was a household cluster (Delhi/CAS 8958-32) and the others occurred within the community. The remainder (n=8) were unique MDR-TB strains. MDR/XDR-TB will be discussed in more detail in Chapter 5.

3.3 Discussion

The current study aimed to build on the previous publication examining the genetic diversity of *M. tuberculosis* in Ireland, by analysing more isolates over a longer time period, i.e. 5 years [106]. This was the most comprehensive MTBC molecular epidemiology study undertaken to date in Ireland. Previous population analyses outlined the relatively high diversity of strains present in Ireland [105, 106]. The presence of West African I and II lineages (lineages 5 and 6) in the current study indicates even greater diversity than previously found, i.e. the presence of all known global lineages except for Ethiopian Lineage 7. Global lineage proportions do not seem to have changed significantly over time (see Table 3). Immigration from areas of high TB prevalence is an ongoing challenge to TB control in many low-prevalence countries [26, 27, 132]. The diversity found could reflect the fact that much of the active TB disease found in Ireland develops in those born outside of Ireland, from diverse regions of the world [22]. This diversity suggests that Ireland's European Union (EU) status, rather than its island status, influences MTBC distribution. The HPSC reports that India, Pakistan, Nigeria and the Philippines are countries from where TB presents most often, but several other countries of origin are reported representing every global continent (n=41 different countries in 2014 alone). Euro-American Lineage 4, however, is still the predominant lineage found, which correlates geographically with Western European lineages, and previous publications [105, 106]. From the results, it is probable that some cross-border transmission has occurred between Northern Ireland and Ireland (7 clusters found, n=67 isolates), although the direction of transmission cannot be determined. However, TB genotypes do not seem to be as transmissible from, or to, the UK Midlands (1 cluster found, n=364 isolates). Cross-border travel and communication may be more common on the island of Ireland than between the UK Midlands and Ireland. Another hypothesis is that immigrants from areas of high TB prevalence to the UK, who settle in the Midlands, do not tend to travel to Ireland subsequently to any great degree. MDR/XDR-TB genotypes that have been associated with EU 'cross-border' clusters were found in this cohort (MtbC15-9 100-32, 94-32 and 843-52), suggesting that movement of people does have a role to play. If drug resistant strains are present in the country, it might take some time, but it may be inevitable that transmission within the population will occur at some point in the future, if it has not already happened.

Rates of mono-resistance were low in the previous studies, as they were in the current study (< 5%). This could suggest that non-compliance with TB treatment regimens, and associated drug-resistant organism selection pressure, is not a major problem in Ireland. Four MDR strains were reported by Fitzgibbon *et al* [106]. MDR/XDR-TB numbers had risen to 19 at the end of 2014. Although the numbers varied over the years (from a peak in 2012 [n=6] to a trough in 2014 [n=2]), this represents an increasing trend overall and a possible threat to TB control in Ireland. MDR/XDR-TB treatment is much more complex (up to two years treatment), high-risk (approximately 50% success rate in 2014) and costly (50-200 times the cost) than regular TB

treatment regimens [1, 4]. However, the clusters were no bigger than two cases and no Irish-born individuals were infected, indicating that there has been no evidence of transmission into the Irish-born population between year 2010 and 2014. Of course, it is still possible that some Irish-born patients are latently infected with resistant strains, but have not yet presented.

The WHO and The Global Fund estimate that the TB funding gap in Europe is 200 million US dollars [133]. Since resources for TB surveillance and treatment in Ireland are decreasing due to its perceived low prevalence, they should be targeted at high-risk groups, multi-drug resistance, and directly-observed therapy (DOT). The HPSC reports that one of the biggest risk factors for developing MDR/XDR-TB is to be from a high-prevalence country, and that many immigrants reactivate latent TB within 2 years of arriving into the country [57]. A recommendation made in the Guidelines on the Prevention and Control of Tuberculosis in Ireland 2010, amended in 2014, was that every person from an area of > 40 per 100,000 TB prevalence, and staying in the country for more than 3 months, should be provided with the opportunity for TB screening [57].

Mixed infection was not observed in the MTBC cohort. However, the presence of double alleles could suggest the presence of sub-populations of these strains *in vivo*. Double alleles were found in Euro-American lineage 4 strains more than other strains, but not significantly so.

Culture-positive cases represented 63.6% in the original IMRL study. In the current study, this number has risen to 97.4%. The European Centre for Disease Control (ECDC) target percentage for culture-confirmation of new pulmonary TB cases is 80%. This indicates that the IMRL is performing well. The original study was, as the name suggests, a ‘snapshot’ in time, and therefore may not have reflected the true percentage of culture-positive cases reported by the IMRL for the years 2010-11.

Cluster size fell from a mean of 3.5 TB cases to a mean of 2 TB cases when five years of data were taken into account (median also 2, IQR 2-3.25). Therefore the original study over-estimated the ‘true’ mean cluster size. When only Euro-American lineage 4 cases were taken into account, the original IMRL study found a rate of 36.6%, whereas the current study found a higher clustering rate of 48.8%. These rates were not stratified by size of cluster. The rate may be higher, but this was because of the larger clusters present in the current study (no clusters with more than 12 cases in the previous study, 2.7% clusters contained more than 25 cases in the current study). Also, clustering rate and size determination may not be entirely accurate with MIRU-VNTR genotyping, and this will be discussed further in Chapter 4. Nonetheless, this shows that a hypothesis put forward in the original paper, i.e. that the study period was too short in length to draw clear conclusions regarding clustering, was valid. Transmission chains can be tracked more readily when prospective genotyping is in place. These chains can then be flagged in a timely manner, and

possibly broken. More clusters appear to be present, but for the most part, their numbers are low. Eighteen clusters were found within Euro-American lineage 4 in the original IMRL paper, involving 63 TB cases; 114 clusters involving 421 TB cases were found within Lineage 4 isolates when the 5-year time period was taken into account. The largest cluster among isolates in the 2004-06 study were from spoligotyping clades X and U, which correspond with Euro-American lineage 4 [105]. The largest cluster in the original study was found within the LAM sub-lineage (n=12 cases). The largest clusters in the current study were within sub-lineages LAM (n = 25, 24) and Haarlem (n = 22), followed by East Asian lineage 2 (Beijing, n = 20). These clusters have clearly seen a large expansion over just a few years, indicating possible transmission within the population. The Haarlem cluster had just four documented cases in 2010-11, but now has 33. The LAM clusters consisted of 8 and 12 cases in the original study and have grown to 25 and 28 in the current study. These clusters are consistent with HPSC outbreak surveillance data from the period, which is detailed in Table 4. The Beijing cluster was associated with a residential institutional outbreak, which began in 2011, among both non-Irish-born and Irish-born patients. The others were community outbreaks associated with more than one location. Minimum-spanning trees visualise the outbreak strains, however the direction of transmission cannot be estimated with just 24 categorical values for comparison (Figure 22-25). The HPSC reported 164 active TB cases from 36 outbreaks. However, 244 further latent TB cases were detected through contact tracing. These 244 individuals could re-activate at any time in their lives. Latent TB is a confounder of all molecular genotyping methods since transmission links between LTBI patients, and between LTBI and active TB cases, cannot be proven. If latent TB DNA could be genotyped in some way, this would represent a leap forward for MTBC surveillance.

MIRU-VNTR_{plus} online web-tool contains a set of 186 reference strains that can be incorporated into a phylogenetic tree, from where new lineages can be assigned. If a large number of genotypes is being queried, each of those queried genotypes now becomes a 'reference strain' on the tree, changing the topology of the tree, and in certain cases, the clade into which that genotype of interest may fit. Extensive experience with this tool has changed the interpretation somewhat since the original study was undertaken. For instance, Haarlem/X is now assigned by its global lineage Euro-American lineage 4. Strains are currently not assigned a sub-lineage unless they fit convincingly within a sub-lineage clade on the phylogenetic tree. The number of MIRU-VNTR_{plus} reference strains is insufficient to legislate for every new genotype that is queried. Therefore Euro-American lineage 4 strains are often not assigned a sub-lineage, especially if there is uncertainty concerning the clade with which they may best align. Lineages are reported to CIDR along with 24-locus genotype, since it was decided that the lineage is a more accessible way to cluster cases than a 24-digit code. However, perhaps the MLVA MtbC15-9, or simply the global lineage, may be a more accurate reporting method in the future. It may also be useful to assign a cluster number, to

any isolate that has clustered, when reporting to the CIDR database. This could help to harmonise the outbreak data between the laboratory, the HPSC, and Public Health.

Correlating anonymous HPSC data on TB cases and outbreaks with IMRL data was challenging. CIDR database was used by all Health Service Executive (HSE) areas from 2011 onwards, therefore some data may have been missing prior to this time. Public Health epidemiological information represents the missing link between the laboratory and the HPSC. The HPSC would define an outbreak as two or more apparently related cases of TB, or, within a higher prevalence setting, a trend above the baseline prevalence. In the original study, the HPSC confirmed five of eleven IMRL clusters as outbreaks under their definition. In the current study period, the IMRL reported 152 clusters, however, the HPSC reported just 36 outbreaks (Table 4). The IMRL lacks the epidemiological evidence to truly link these cases and determine the direction of transmission. MIRU-VNTR genotyping relies on this data to prove transmission. It is possible that at least some of these clusters are truly linked, but as yet have not been epidemiologically proven. However, MIRU-VNTR genotyping has been found to over-estimate genotyping clusters in both low- and high-prevalence settings [134, 135]. Identical MIRU-VNTR genotypes may not represent recent transmission. In some cases, even though MIRU-VNTR lineages may have come from a recent common ancestor originally, their genomes may have evolved differently over time, or may have been subject to convergent evolution [136]. The estimated mutation rate of *M. tuberculosis* is 0.5 SNVs per genome per year [112]. The 24 MIRU-VNTR loci may remain identical, but other regions of the genome (which consists of 4.4 million base pairs) could have changed. It was hypothesised in the original study that whole genome sequencing represents a higher resolution genotyping method, and this will be explored in Chapter 4.

In summary, the study showed that there is a high level of diversity in MTBC strains present in Ireland. Euro-American lineage 4 remains the predominant lineage. MDR/XDR-TB is present in variable amounts but ultimately increasing, although mono-resistance remains below 5%. Median cluster size was two. Within Euro-American lineage 4, clusters of two cases constituted over 50%, which compared to previous studies. However, 2.7% of clusters were greater than 25 cases, which was not seen previously. MIRU-VNTR genotyping is an excellent first-line tool for prospective genotyping of MTBC in Ireland.

Chapter 4.

Whole Genome Phylogenetic Analysis of MTBC Isolates from 11 Informative MIRU-VNTR Genotyping Clusters

4 Whole Genome Phylogenetic Analysis of MTBC Isolates from Eleven Informative MIRU-VNTR genotyping Clusters

4.1 Introduction

Outbreaks of tuberculosis infection can severely impact communities and institutions alike, causing illness in many cases and anxiety in many more latently-infected contacts. TB genotyping aims to disrupt transmission chains by flagging potential outbreaks. Breaking the chain of transmission is the ultimate goal. Contact tracing is integral to this. Household TB contacts tend to develop active infection within one year of latent infection [137]. The intervention yield of household TB contact investigations, especially in the case of MDR/XDR-TB, has been found to be relatively high [137]. The more virulent, or drug-resistant, the strain, the more effort should be put into stopping its transmission. Coscolla *et al* concluded from their studies that the most geographically widespread lineages (East Asian, 2, and Euro-American, 4) are more virulent than geographically restricted strains [138]. East Asian lineage 2, sub-lineage Beijing, strains are thought to have spread globally in several waves, roughly coinciding with historical events such as the industrial revolution, the First World War and HIV epidemics [139]. Multi-drug resistant Beijing clones emerged following the fall of the Soviet Union [139]. The movement of people, from outside the EU and within, enables the movement of these strains from areas of high prevalence to areas of low prevalence. Contact tracing can be extremely challenging; immigrants may not receive TB screening, either pre- or post-entry to Ireland, TB patients (both Irish and non-Irish born) and their contacts sometimes live chaotic lives and may not comply with treatment, LTBI testing is not always easy to interpret, extensive resources are needed to contact trace, and it is difficult to know how far from the perceived index case to trace. LTBI makes outbreak investigation even more difficult for TB, since the individual who first presents may not be the index case, and contacts have a 5-15% chance of presenting at any stage in their lives [1]. Resources for TB contact-tracing in low-income countries may be almost non-existent, however even in high-income countries, resources may be hard to come by when the majority of people believe that TB is no longer a significant problem [140].

As discussed in Chapter 3, following an amendment to regulations in 2004, outbreaks of TB became statutorily notifiable and are reported annually by the HPSC (Table 4). Genotyping was introduced in the IMRL in 2009 in order to augment public health contact tracing and surveillance on a national scale. While 24-locus MIRU-VNTR genotyping has been recognised as a very useful technique, in certain cases, its resolution is not sufficient to distinguish “outbreak” strains where epidemiological links have not been established or are difficult to obtain [141]. MIRU-VNTR genotyping has been found to over-estimate clustering, especially in low-incidence TB settings [135, 142, 143]. In Switzerland, for instance, Stucki *et al* found 35 MIRU-VNTR clusters (n=90

cases). WGS confirmed just 17 of those [135]. They showed that foreign-born MIRU-VNTR genotypes tend to be over-estimated more than Swiss-born patients' genotypes.

WGS has the potential to strongly impact on diagnostic and public health microbiology for many different organisms, including *E. coli*, MRSA and MTBC, and viruses like HIV [144]. Several studies have been undertaken to prove the hypothesis that WGS resolves TB clusters more thoroughly than classical genotyping methods [145]. Without an estimation of the mutation rate of MTBC, it would be difficult to infer patient-to-patient transmission or relatedness. The TB molecular clock was investigated by three separate groups, and estimated to be 0.3-0.5 SNVs per genome per year [112, 143, 146]. Gardy *et al* delineated a Canadian outbreak in 2011 with greater resolution using WGS analysis and social networking. What was originally thought to be a clonal outbreak with missing epidemiological links, turned out to be the simultaneous expansion of two different strains of MTBC that had evolved from a common ancestor regularly seen in that region. Specific members of a high-risk social group, associated with substance misuse, who were highly infectious, were hypothesised to have contributed to the large size of the outbreak. These individuals were termed 'super-spreaders' [142]. In another study, published by Roetzer *et al* in 2013, genomic clustering based on WGS matched contact tracing and geographical distribution of strains better than classical genotyping methods like spoligotyping, RFLP IS6110 typing, and MIRU-VNTR genotyping [143]. WGS subdivided the outbreak (n=86 isolates) into seven clusters and 36 unique strains. One of those clusters (the 'Hamburg' clone) emerged between 1993 and 1997, and continued for at least 14 years. Within that time-frame, mutation of just 3 SNVs was seen [143]. Kato-Maeda *et al* suggested that current tools did not give enough information about the direction and sequence of transmission of short-term outbreaks. They uncovered an outbreak whose microevolution could be traced by SNV acquisition over time in sequential patients (n=9 patients); they saw 0-2 SNVs per transmission event [147]. The results matched epidemiological evidence in 8 out of 9 cases. Even when epidemiological evidence has not suggested an outbreak, WGS has been able to prove otherwise. Two individuals attending the same school, but with no other links and no knowledge of one another, contracted genomically identical MTBC (sub-lineage Beijing). Most of the other students were not in the country to be screened for LTBI, but it is hypothesised that the true index case was among them [148]. Without WGS, this transmission could not have been confirmed since both students were of Asian origin where the Beijing lineage is endemic. Research by Walker *et al* has contributed significantly to our knowledge of TB outbreak WGS analysis and interpretation. They investigated rate of mutation in cross-sectional and longitudinal isolates from individuals and families respectively (0.5 SNVs per genome per year) and estimated, from characterisation of household and community clusters, that < 5 SNVs is representative of recent transmission but that > 20 SNVs rules out transmission within a MIRU-VNTR cluster. They delineated outbreaks with greater resolution than MIRU-VNTR genotyping, inferred direction of transmission, and identified super-spreaders within the cohort [112]. In a

follow-up observational study on transmission of TB in the Oxfordshire area of the UK from 2007-12, they found that most non-UK-born patients from high-incidence countries reactivate LTBI in the UK but are not involved in onward transmission of TB. Nonetheless, the incidence in this area due to the number of non-UK TB cases was 31 times the average for a low-incidence country [149].

Tuberculosis prevalence in prisons worldwide is consistently higher than that of the general population, especially in in-mates who inject drugs and those who are infected with viral hepatitis and/or HIV [150]. Brazil has the 4th largest incarcerated population in the world, with TB incidence within those prisons 20 times that of the average population [151]. A study undertaken in Brazil assessed the relationship between urban TB cases and prison-related cases and found that 54% could be linked to the prison outbreak [151]. Russian gulags have been implicated in the spread of MDR/XDR-TB. In Russian prisons in 2004, Medecins San Frontier estimated that 22% of new cases, and 40% of retreatment cases were MDR-TB [152]. This is a significant number when one takes into account that an estimated 74,000 prisoners in Russia were found to have active TB disease in 2003 [153]. Ireland is not exempt. The HPSC documented a TB outbreak in a prison that began in 2011, which continues, with cases still emerging in 2016. Even though Ireland is a low prevalence country, there are pockets of high prevalence, especially in the capital city, Dublin, where it can reach 40 cases per 100,000, compared to the countrywide average of 6.9 cases per 100,000 [22]. This is possibly due to more individuals that are at high-risk of contracting TB dwelling in urban areas, e.g. those experiencing homelessness, addiction and/or deprivation. There have also been documented outbreaks across the country in residential institutions, non-residential institutions, private families, extended families and communities (Table 4). High-risk groups, such as the Irish Traveller population, also experience a higher crude TB incidence rate than the Irish-born population, as well as outbreaks [154]. Chapter 3 indicated that Ireland has experienced clusters of MIRU-VNTR genotyping over the course of 5 years. These clusters have not been investigated to date in order to find if perhaps some of them may have been over-estimated.

In this study, clusters of interest were chosen for WGS analysis in order to determine whether MIRU-VNTR genotyping has over-estimated TB transmission in Ireland, that WGS can resolve these clusters with greater resolution, and that WGS would be a valuable addition to conventional genotyping currently in place in the IMRL.

4.2 Results

Illumina Next-Generation-Sequencing was used to further investigate MIRU-VNTR clusters. Eleven clusters were chosen from the IMRL national collection over the period 2006 to 2014 (n =138 isolates, Table 7). The largest of these was Cluster 10 from an institutional outbreak (n=24) followed by Cluster 1 (n=23), associated with community-based substance abuse. Cluster 2, also related to substance abuse, contained two community-based sub-clusters 2a and 2b (2a, n=21 and, 2b, n=20) that differed by one MIRU-VNTR single locus variation (SLV). Eight smaller clusters were also included, ranging in size from 3 to 12 cases. The majority of patients within these clusters were of Irish origin. Based on previous publications, a single nucleotide variation (SNV) difference of less than 5 (< 5) was associated with recent transmission, more than 5 but less than 12 (> 5 - < 12) SNVs with possible transmission, more than 12 but less than 20 (> 12 - < 20) SNVs was considered a grey-zone, and more than 20 (> 20) SNVs were not associated with transmission, based on the estimated mutation rate of 0.5 SNVs per genome per year for MTBC [112].

4.2.1 IMRL Cluster 1 – community-based substance abuse, Haarlem

Euro-American lineage 4, sub-lineage Haarlem, cluster 1 consisted of identical 24-locus MIRU-VNTR genotypes (Table 7). Whole genome sequenced isolates were collected between 2006 and 2014, the longest time-span of all the clusters chosen. Twenty-three isolates were sequenced, however according to the IMRL MIRU-VNTR genotyping database, 51 identical isolates have been found as of August 2016, indicating that this clone continues to expand. Four possible family/household outbreaks occurred within this cluster, consisting of 9 cases.

When the maximum likelihood phylogenetic tree was constructed for Cluster 1, isolates grouped broadly into 5 clades (Figure 26). IMRLH32, 35 and 37 were family members living at the same address. While IMRLH35 and 37 (mother and daughter) were indistinguishable, IMRLH32 (father) was 141 SNVs apart. MIRU-VNTR genotyping was repeated on this isolate in order to confirm that it was the correct patient. The MIRU-VNTR genotype of the isolate represents a Haarlem strain unrelated to the other two cases. IMRLH94 was more similar to IMRLH35 and 37 (> 5 - < 20 SNVs apart) than IMRLH32, and the individual had an address in the same area of Dublin. However, no other epidemiological link was evident. IMRLH37, a daughter of IMRLH35 and 32, presented in December 2006, so could possibly be the index case. Her mother presented in October 2009. IMRLH94 presented 4 years later, so it could be possible that transmission occurred here since a timeframe of 4 years would allow for mutation of at least 2 SNVs to occur.

IMRLH05, 10, 11, 31 and 39 also form a clade with IMRLH94 in which no isolate is more than 19 SNVs apart, which represents a grey-zone result (> 12 - < 20). No striking epidemiological evidence accompanies this clade. IMRLH31 was the first to present, in April 2011. However this

Cluster No.	MIRU-VNTR Type	No. of isolates	WGS isolate time-span	Sub-Lineage	M/F	MEDIAN AGE	GEOGRAPHICAL AREA	RATIO OF IRISH/NON-IRISH BORN	PULMONARY/EXTRA-PULMONARY/NK	MAIN RISK FACTOR
1	223225342334425143323_32	23	2006-14	Haarlem	7:3	45	DUBLIN,KILDARE,CORK, LIMERICK	61% IRISH,9%NON-IRISH (INDIA+SOMALIA),30%NK	70%/13%,17%	ALCOHOL MISUSE, IMMUNOSUPPRESSION
2a	142244332224126143322622	21	2010-13	LAM	29:12	47	DUBLIN,MEATH,GALWAY, WICKLOW,	68% IRISH,2%MAURITIAN,30%NK	71%/5%/24%	FROM A COUNTRY WITH HIGH TB ENDEMICITY
2b	142244332224126153322622	20	2010-13				WEXFORD,CLARE			
3	245243122334225143335522	3	2010-12	H37Rv	100% female	30	DUBLIN AREA	66.7% IRISH,33.3%ROMANIAN	66.7%/33.3%	NK
4a	224213322534226153332422	6	2011-13	H37Rv	6:4	24	CAVAN,MONAGHAN, TIPPERARY, DUBLIN	12.5% IRISH,12.5% ZIMBABWE,75% NK	87.5%/12.5%	FROM A COUNTRY WITH HIGH TB ENDEMICITY
4b	224213322534226153335422	2	2011							
5	242247432244225113342543	4	2010-11	Delhi/CAS	3:1	20	DUBLIN,CORK,LIMERICK	50% SOMALIAN,50% NK	25%/50%/25%	FROM A COUNTRY WITH HIGH TB ENDEMICITY
6a	2145243a2843266223342713	4	2009-13	EAI	5:1	37	DUBLIN,MEATH, PHILIPPINES	50% FILIPINO,50% NK	33.3%/66.7%	FROM A COUNTRY WITH HIGH TB ENDEMICITY
6b	2145243a2943266223342713	2	2011-12							
7	244233352644425153353623	4	2009-14	Beijing	1:1	47	LIMERICK,MONAGHAN, LAOIS, DUBLIN	25% AZERBAIJAN,75% NK	75%/25%	NK
8	22421333154422515333_522	3	2010-12	Cameroon	2:1	42	MEATH,LOUTH,DUBLIN	33.3%CAMEROON,66.7% NK	33.3%/66.7%	FROM A COUNTRY WITH HIGH TB ENDEMICITY
9a	224213331644225153332522	2	2010	Cameroon	5:12	24	DUBLIN,LIMERICK,CLARE, GALWAY	29% IRISH,6% UK,65% NK	82%/12%/6%	FROM A COUNTRY WITH HIGH TB ENDEMICITY
9b	22421433164422515333_522	9	2006-13	Cameroon	5:12	24	DUBLIN,LIMERICK,CLARE, GALWAY		82%/12%/6%	
9c	224214331644225153332522	3	2007-11	Cameroon	5:12	24	DUBLIN,LIMERICK,CLARE, GALWAY		82%/12%/6%	
10	244233352644424173353823	24	2011-13	Beijing	11:2	33	DUBLIN, GALWAY,LAOIS, MAINLY CLOVERHILL	ESTIMATED 62.5% IRISH	88%/8%/4%	IMMUNOSUPPRESSION, TB CONTACT
11	244215342234425153334542	8	2013-2015	Euro-American	7:1	60	CO.CORK	ESTIMATED 100% IRISH	83.3%/16.7%	NK

Table 7. Compiled details of isolates chosen for in-depth WGS analysis of MIRU-VNTR genotyping informative clusters.

Included are risk factors associated with members of the cluster, where isolates were collected from, where the patients involved originated from, median age, ratio of male to females, number of isolates, MIRU-VNTR genotype that was identical, sub-lineage, and time-frame of the cluster. NK – not known

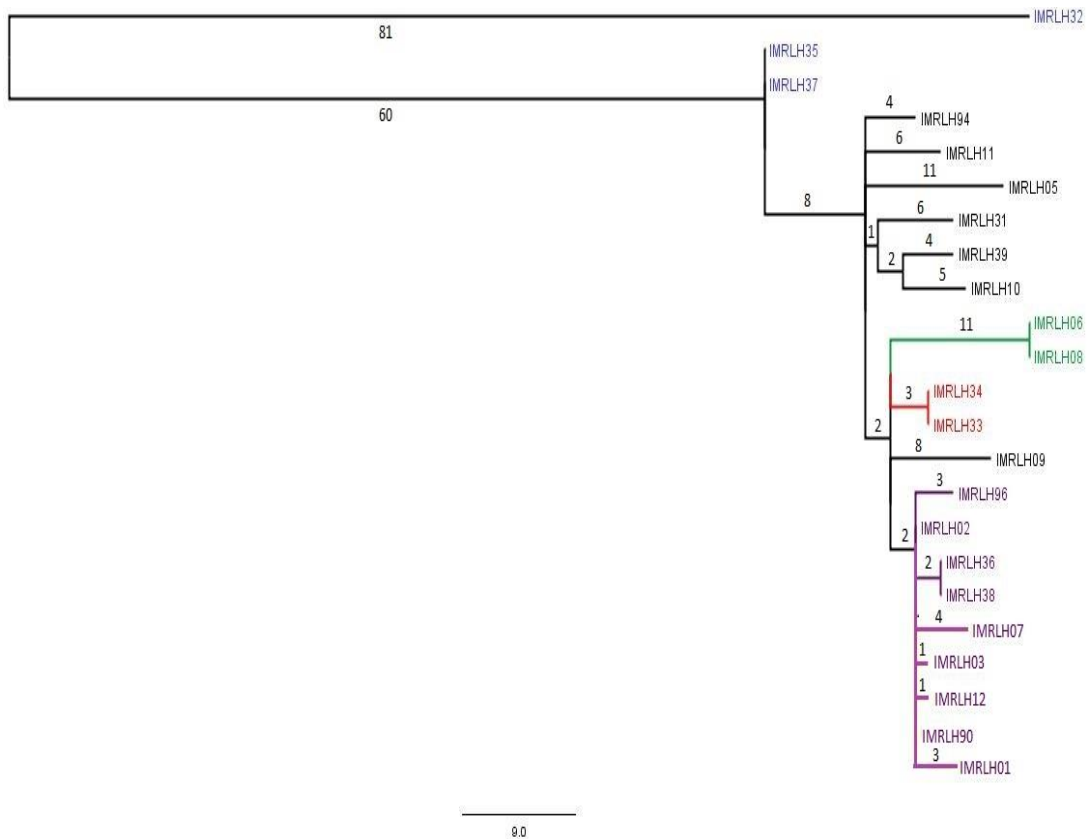


Figure 26. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 1 isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. Coloured isolates (blue, green, red, and purple) refer to separate clades. WGS separates the identical MIRU-VNTR genotypes broadly into five clades and several unique strains (black). IMRLH32, 35 and 37 (blue) represent a household outbreak, as do IMRLH36 and 38 (purple).

patient was immunosuppressed and presented with a neck lump, which would not suggest that he was infectious.

The second case to present was IMRLH39, which was a pulmonary case.

Individuals born outside Ireland (IMRLH33, Somalia, and 34, India) were found to be identical, indicating recent transmission between them, most likely within Ireland. The most closely-related cases to these were IMRLH90, 03 and 12. IMRLH03 had TB meningitis. IMRLH12 had pulmonary TB in January 2009, before the non-Irish born cases presented (in August and November of the same year), so transmission is a possibility here. IMRLH90 did not present until 2013.

IMRLH06 and IMRLH08 had identical surnames and originated from the same geographical area. Although unconfirmed, recent transmission looks likely between these patients (<5 SNVs difference).

IMRLH01, 02, 03, 07, 09, 12, 36, 38, 90 and 96 represent a grey-zone clade separated by no more than 20 SNVs. IMRLH36 and 38 occurred as part of a family outbreak at one address, which was linked to contact with a TB case. IMRLH01, 03, 07, 12, and 90, all no more than 7 SNVs apart. The common risk factor that the majority of these cases shared was alcohol misuse, and anecdotally, some of these individuals may have shared a homeless hostel on a number of occasions, which could have presented the opportunity for transmission to occur. IMRLH01, 02, 12, and 90 are < 5 SNVs apart, which provides more compelling evidence to suggest transmission between these patients. IMRLH12 presented in 2009, while the others presented between December 2011 and July 2013, therefore could have been the index case. IMRLH02 and IMRLH12 could be the link between the more distant cases within this clade, since these patients were associated with substance misuse and the homeless hostel, IMRLH12 presenting in 2009, followed by IMRLH02 in 2011, followed in quick succession by the remainder. This could possibly suggest IMRLH02 was a super-spreader. IMRLH07 is the only patient who does not fit the general picture of this clade, since he was a retired professional with diabetes. IMRLH96 does not fit this picture either, since it is associated with an 80 year old retired professional, although contact (between IMRLH07 and/or 96) with any of the other people within the clade has not been ruled out.

Results suggest that MIRU-VNTR cluster 1 was not a single clone but, in fact, a series of smaller clusters. Figure 27 visualises genetic distances within this cryptic cluster. MIRU-VNTR genotyping has over-estimated transmission here.

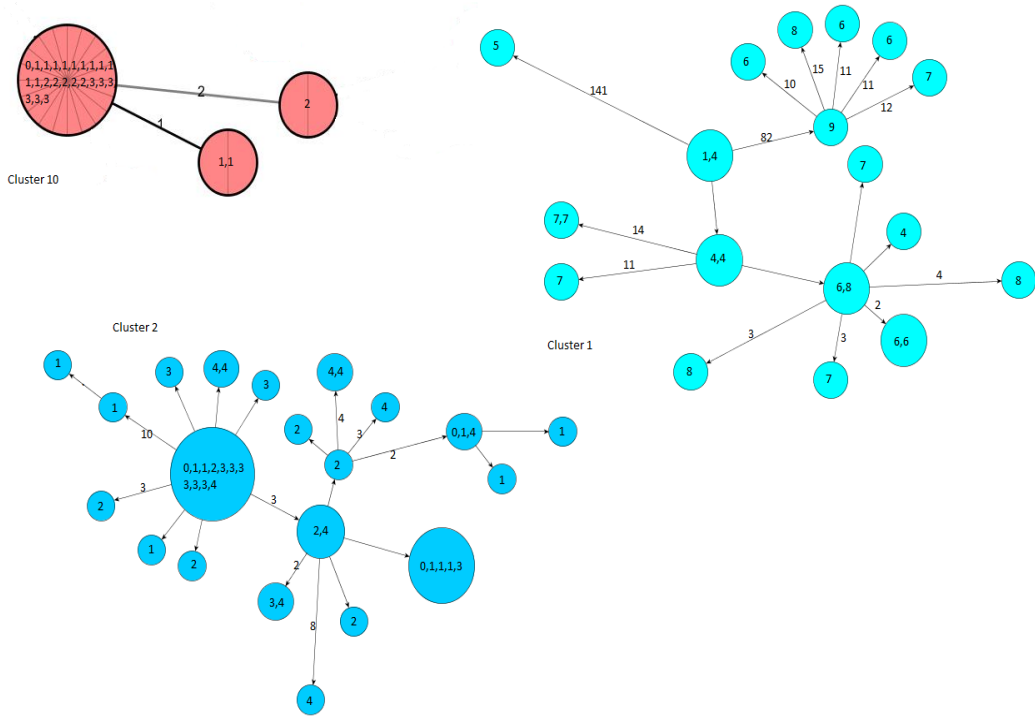


Figure 27. Genetic distances within the three largest clusters found; Clusters 1, 2, and 10

Genetic distances were taken from maximum likelihood phylogenies constructed from the genomes of each cohort. Each circle represents a node of people who were infected with isolates separated by no SNVs. Each number within a circle represents a case, and the number refers to the year of the outbreak in which they first presented (the first-infected represented by 0 in the larger sub-clusters). Lines between circles without a label represent 1 SNV difference, lines with number labels represent that number of SNVs between circles (not to scale). Direction of transmission is not included but it is predicted that each cluster would expand to the right.

4.2.2 IMRL Cluster 2 – community-based substance abuse, LAM

Euro-American lineage 4, sub-lineage LAM, cluster 2 consisted of two sub-lineages which differed by one MIRU-VNTR SNV at locus 2996 (Table 7). Cluster 2a contained 4 repeats at this locus, while cluster 2b contained 5 (Table 7). Twenty-one and twenty isolates were chosen for WGS analysis, however there are now 37 and 38 identical isolates found within these clusters, respectively, in the IMRL database as of August 2016. Galway-based isolates only clustered with genotype 2b. Five possible family/household outbreaks had occurred within the cluster, consisting of 10 individuals. Contact with another case of TB and substance misuse, were the most common risk factors found among cases.

A maximum likelihood tree was constructed to visualise the relationship between Cluster 2 TB genomes (Figure 28). Genetic distances are also represented in Figure 27. The first clade observed was between IMRLH48 and 55. No epidemiological evidence links these cases that presented in July and August 2010. A large cluster, consisting of IMRLH13-23, 47, 49, and 50-54 were all found to be within 5 SNVs of each other (n=18). Eleven of these isolates' whole genomes were indistinguishable. This suggests that there was a possible super-spreader who transmitted infection within a short time frame to many others. All isolates within this clade were from MIRU-VNTR cluster 2b. IMRLH17, 22 and 52 shared the same address, as did IMRLH19 and 20. IMRLH50 and 53 were neighbours of IMRLH19 and 20, living on the same road in Dublin. IMRLH13 and 23 were family contacts. IMRLH53 presented in January 2010 with pulmonary TB, before the remainder of cluster b (February 2010 – March 2013), therefore could have been the index case. Only two MIRU-VNTR cluster b isolates were observed to be different from this group (IMRLH48 and 55 above), and these too were within the grey-zone (<20 SNVs apart).

IMRLH64, 71 and 72 were < 5 SNVs apart. These three cases occurred in females (one Mauritian-born, the others Irish-born) with addresses in the Dublin area, one of whom had been previously treated for TB and presented in 2011. No other epidemiological information was available to link these patients. IMRLH62, 67 and 68 formed another Irish-born clade associated with family contacts. IMRLH24, 42, 45, 46 and 66 shared common TB risk factors of immunosuppression and substance misuse and constitute another clade. IMRLH45 and 46 were family members within this clade. Another family group (father and son) were indistinguishable when their genomes were compared (IMRLH43 and 70). The son was the presumed index case who presented in September 2012, followed by his father in October 2013. IMRLH69, although no more than 16 SNVs away from the furthest isolate within cluster 2a, is not as closely linked genomically. IMRLH40, 41, 44, 60, 61, 63 and 65 form the final clade within cluster 2a. All cases presented between October 2010 and May 2012 with the most common TB risk factors cited being alcohol misuse and previous history of TB infection. Two of the cases were family members (IMRLH40 and 60) and these group together as indistinguishable from the genomes of three others (IMRLH44, 61 and 65). Of

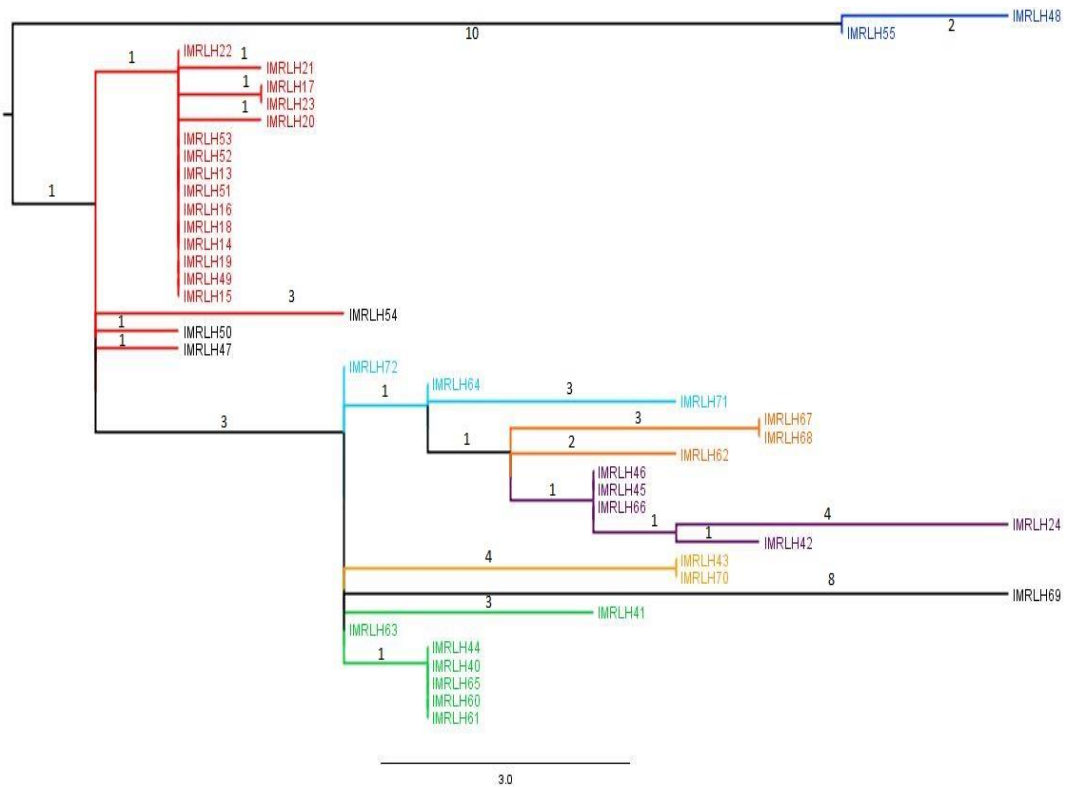


Figure 28. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 2a and 2b isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. Cluster 2a and 2b are separated by 3 SNVs. Coloured branches refer to 7 separate outbreaks (blue, light blue, red, purple, orange, yellow and green). The cluster in red indicates the presence of a super-spreader, as does the cluster in green. IMRLH17, 22 and 52 represent a household outbreak, as do IMRLH19 and 20, and IMRLH13 and 23 (all present within the red cluster). IMRLH62, 67 and 68 were also household contacts (orange), as were IMRLH45 and 46 (purple), as were IMRLH43 and 70 (yellow), and IMRLH40 and 60 (green).

these, IMRLH61 was the first to present with pulmonary TB, in October 2010, so could potentially be a super-spreader of the disease.

Cluster 2a and 2b were just 3 SNVs apart. There were just 24 SNVs between the most distant isolates within the entire cluster. Since this number was higher than the estimated threshold for a possible outbreak (ie > 20 SNVs), the overall cluster does not indicate recent transmission itself but rather 7 separate outbreaks. IMRLH63 and 72 were the two closest cluster 2a isolates to cluster 2b. IMRLH63 cited contact with a TB case as a risk factor. It is possible that this is the link between Cluster 2a and 2b. Public health colleagues suspected a north Dublin inner-city public house as the original source of this outbreak, however it is unclear exactly how it spread countrywide, even as far as an island on the western seaboard.

MIRU-VNTR genotyping over-estimated cluster 2a, however this still represents one of the largest 'true' clusters resolved by WGS (n=18).

4.2.3 IMRL Cluster 3 – community-based drug-susceptible, H37RV (Dublin)

Euro-American lineage 4, sub-lineage H37Rv cluster 3 was chosen since it was a smaller cluster with a less common genotype (Table 7). Three cases were sequenced (collected 2010-12, from the Dublin area) and four isolates match this MIRU-VNTR genotype in total in the IMRL database as of August 2016.

Figure 29 details the phylogenetic tree constructed from the whole genomes of cluster 3 isolates. Two Irish cases (IMRLH29 and 30) differed by a single SNV (ie < 5 SNVs apart), therefore suggest recent transmission. However, this result did not necessarily match the clinical details of these patients. IMRLH29 presented in May 2010 with extra-pulmonary TB. IMRLH30, although Irish-born, had documented residence in a country of high endemicity and presented with pulmonary TB in December 2012. IMRLH28 was from a country of high endemicity (Romania) and presented with pulmonary TB in January 2012. Since the isolate was 11 SNVs distant from IMRLH29 and 30, this would represent a possible, but less likely, relationship (> 5 - < 12 SNVs apart).

WGS and MIRU-VNTR genotyping agree for one sub-cluster, although no epidemiological evidence links the transmission. It resolves a further identical MIRU-VNTR genotype strain.

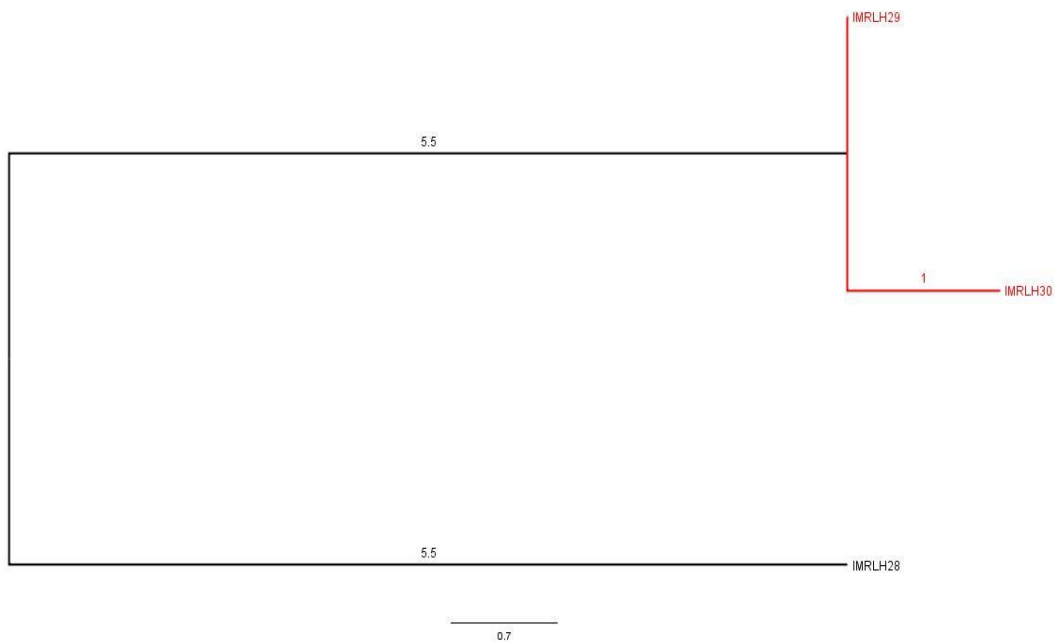


Figure 29. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 3 isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. Coloured branches refer to a separate outbreak (red).

4.2.4 IMRL Cluster 4 – community-based drug-susceptible, H37RV (countrywide)

Euro-American lineage 4, sub-lineage H37Rv cluster 4 consisted of two pan-susceptible MTBC sub-clusters (4a, collected 2011-13, n=6 and 4b, collected 2011, n=2) that differed by one MIRU-VNTR SLV at locus 3690 (cluster 4a consisted of 2 repeats at this locus, cluster 4b consisted of 5) (Table 7).

Results can be seen in Figure 30, a phylogenetic tree constructed with variant files from the whole genomes of this cohort. Isolate IMRLH80 did not produce enough DNA at the extraction stage, and therefore could not be whole genome sequenced. MIRU-VNTR cluster 4b (n=2) consisted of two cases (IMRLH81 and 82), one of which was a Zimbabwean patient. From an internet search, the other patient's surname has also been associated with Zimbabwean origin. These two isolates were indistinguishable, and were at least 164 SNVs distant from cluster 4a. Since > 20 SNVs separate the clades, this is a distinct cluster, which is consistent with recent transmission. The cases presented in March and August 2011, IMRLH81 being the possible index case. IMRLH78 does not fit within either of the two clades found. This patient presented in 2011 with TB in the sacrum. IMRLH73, 74, 76 and 79 formed the second clade (cluster 4a). All cases presented between December 2012 and June 2013 (6 months). Three of the cases, ranging from 21-24 years of age, were associated with a non-residential institutional outbreak in a university, but the fourth case, which was the first to be isolated, was 81 years of age, where TB was discovered at post-mortem (IMRLH79). An epidemiological link could not be found at the time. WGS suggests that these cases represent recent transmission as part of an outbreak (ie < 2 SNVs apart).

WGS broadly agrees with the two separate MIRU-VNTR genotyping sub-clusters, with one isolate seemingly un-related.

4.2.5 IMRL Cluster 5 – regionally-dispersed ethnic group, Delhi/CAS

East African Indian lineage 3, sub-lineage Delhi/CAS cluster 5 consisted of three isolates, collected from 2010 to 2011 (Table 7). This number has since grown to five isolates with identical MIRU-VNTR genotypes in the IMRL database. This cluster was the only MIRU-VNTR genotype found in the database to match a genotype found in a group of patients sequenced as part of an outbreak associated with substance misuse in a previously-published UK Midlands study [149].

The phylogenetic tree built with these genomes, and the genomes from the UK Midlands outbreak (Figure 31), suggest that there was, in fact, no link between them, even though their MIRU-VNTR genotypes were identical (at least 99 SNVs different). Furthermore, the 3 isolates found in Ireland (IMRLH25, 26, and 27) were not related either (at least 55 SNVs different, ie > 20 SNVs). MIRU-VNTR over-estimated transmission here. It is likely that these patients were reactivating latent

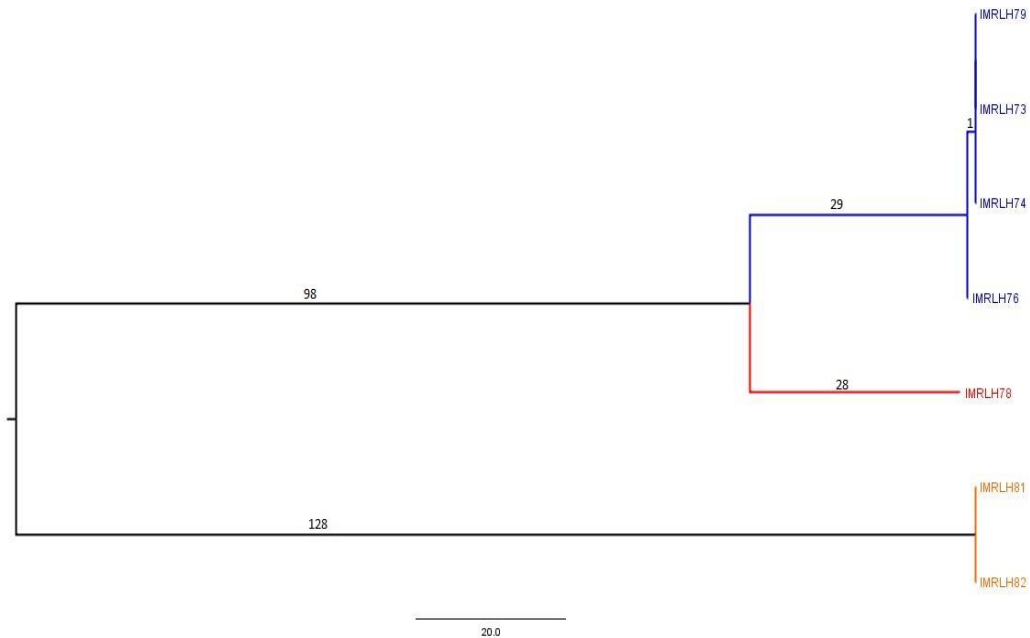


Figure 30. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 4a and 4b isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. Coloured branches refer to separate clades (red, purple and orange). IMRLH73, 74 and 76 represent a non-residential institutional outbreak (Cluster 4A), while IMRLH81 and 82 represent two seemingly un-related Zimbabwean patients.

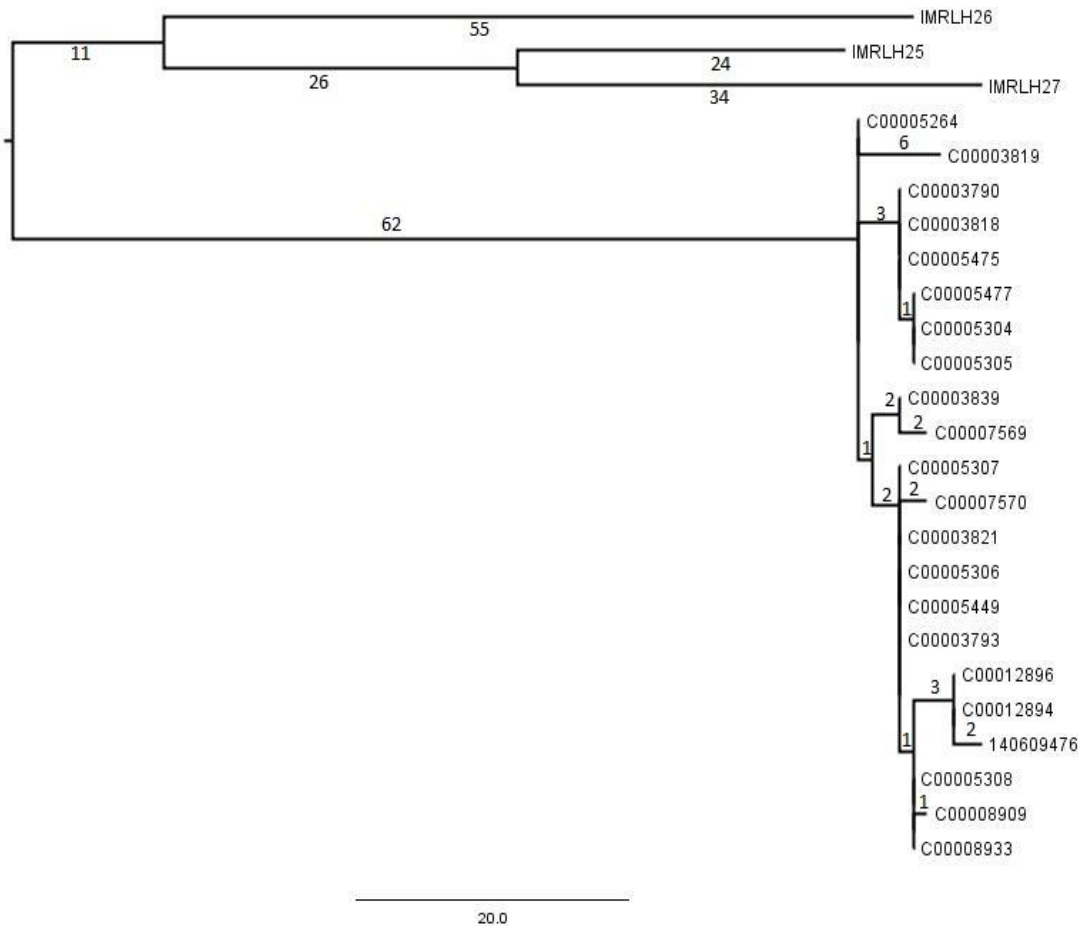


Figure 31. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 5 isolates and sequences from a UK Midlands Study

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. Included also are isolates from an outbreak associated with substance misuse from a previous publication based in the UK (those labelled ‘C0000...’), which seems to be separated by at least 133 SNVs, therefore un-related [9]. The cases found in Ireland are also un-related according to WGS.

infection that they had contracted elsewhere. This hypothesis is supported further by the fact that two of the cases were extra-pulmonary in nature.

WGS delineated this identical MIRU-VNTR cluster into separate un-related strains. MIRU-VNTR genotyping has over-estimated transmission here.

4.2.6 IMRL Cluster 6 – regionally-dispersed ethnic group, EAI

Indo-Oceanic lineage 1, sub-lineage EAI, cluster 6 consisted of two sub-clusters separated by one MIRU-VNTR SLV at locus 2163b (cluster 6a, collected 2009-13, n=2 and cluster 6b, collected 2011-12, n=4) (Table 7). Fifty per cent of cases were originally from the Philippines and cited being from a country of high endemicity as a TB risk factor. Isolates were pan-susceptible with first-line DST except for one isolate which was resistant to isoniazid, ethambutol and streptomycin (IMRLH85).

Figure 32 visualises the maximum likelihood tree constructed using Cluster 6 whole genomes. Every isolate was separated by at least 68 SNVs (ie >20 SNVs) therefore recent transmission was ruled out, even though each isolate within cluster 6a (and 6b) had identical 24-locus MIRU-VNTR genotypes and were living in the same geographical region. Although not every case country of origin was recorded, public health colleagues confirmed anecdotally that all of these patients were originally from the Phillipines. Only two of the cases were pulmonary in nature, the others presented with extra-pulmonary TB. This pattern suggests reactivation of TB rather than recent transmission. MIRU-VNTR over-estimated transmission in this case.

WGS further resolved this identical MIRU-VNTR genotyping cluster into separate un-related strains. Conventional genotyping has over-estimated transmission.

4.2.7 IMRL Cluster 7 – community-based drug-resistant, Beijing

East Asian lineage 2, sub-lineage Beijing cluster 7 was chosen on the basis of its lineage, and the fact that one isolate was an MDR-TB strain (IMRLH57), and one isolate was resistant to isoniazid and streptomycin (IMRLH59), while the remaining cases were susceptible (n=2, Table 7, Figure 33). Although small in size when sequencing was performed (isolates collected 2009-14), this MIRU-VNTR cluster has grown in size from 4 to 10 as of August 2016.

The maximum likelihood phylogenetic tree constructed with these isolate genomes indicates that, while their MIRU-VNTR genotypes were identical, and they may have originated from a common ancestor, their whole genomes are sufficiently different to rule out recent transmission (ie an

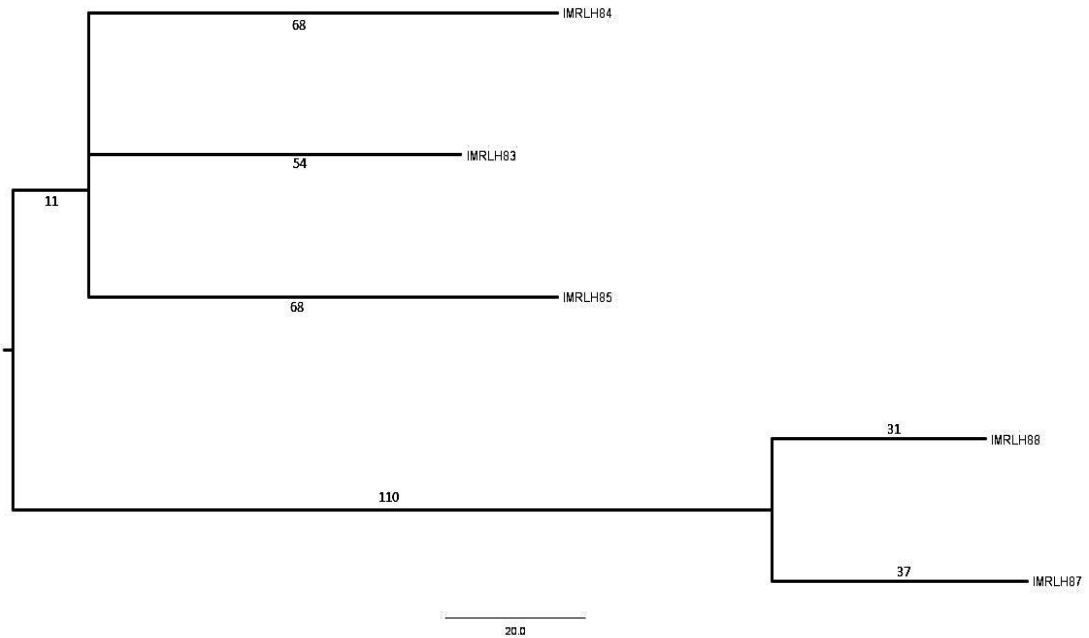


Figure 32. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 6 isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. No isolate was linked to any other according to WGS, even though MIRU-VNTR genotyping clustered them all as identical.

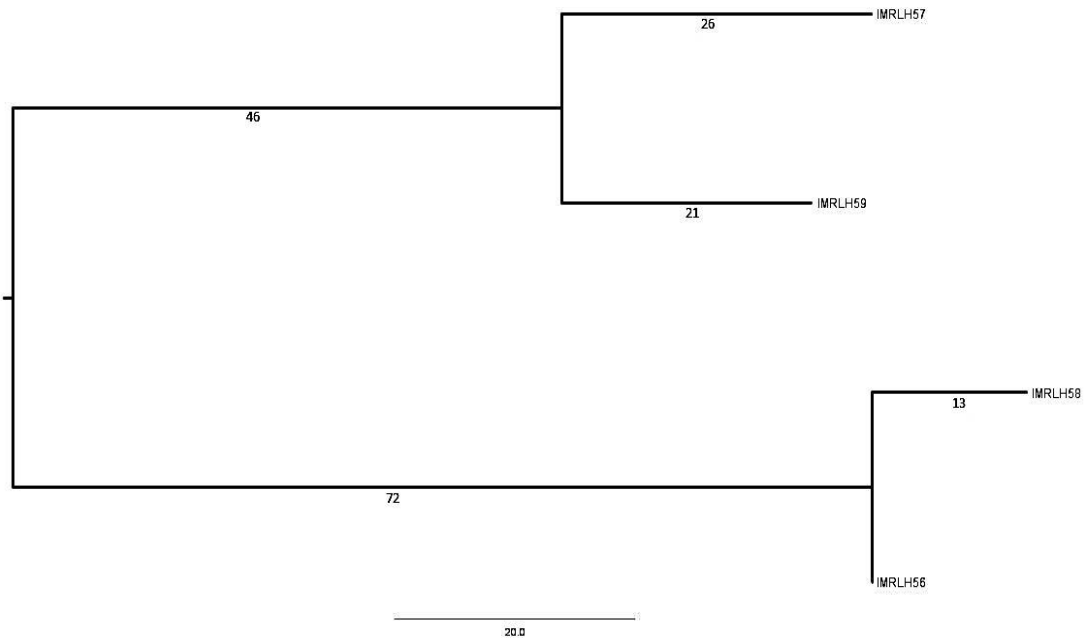


Figure 33. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 7 isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. No isolate was linked to any other according to WGS, even though MIRU-VNTR genotyping clustered them all as identical.

outbreak). The closest cases were 13 SNVs apart (IMRLH56 and 58). One of these patients (IMRLH58) presented with renal TB in 2012, and the other (IMRLH56) had pulmonary TB in 2014, so there is no obvious epidemiological link either. IMRLH57 and 59 form a separate clade 118 SNVs distant. These were the drug resistant isolates. It does not look likely that this drug resistance was transmitted from one patient to the other, nor to any other patient, even though they fit broadly into the same clade.

WGS further delineated this identical MIRU-VNTR genotyping cluster into separate un-related strains, therefore transmission has been over-estimated.

4.2.8 IMRL Cluster 8 – community-based MDR-TB and drug-susceptible, Cameroon

Euro-American lineage 4, sub-lineage Cameroon cluster 8 was chosen on the basis of a mixture of identical strains with differing drug susceptibilities (Table 7). The cluster consisted of 3 isolates (collected 2010-12), one of which was an MDR-TB, from a patient with pulmonary TB. The other two patients had extra-pulmonary TB.

Analysis showed that these isolates all differed by at least 22 SNVs suggesting that they derived from a common ancestor but were not the result of recent transmission (Figure 34). MDR-TB was not found to have been transmitted within this cohort, but rather has been reactivated within the patient, or transmitted from an individual outside the scope of the study.

WGS resolved this cluster into separate un-related strains. MIRU-VNTR genotyping, therefore, has over-estimated TB transmission in this case.

4.2.9 IMRL Cluster 9 – community-based low level INH resistance, Cameroon

Euro-American lineage 4, sub-lineage Cameroon cluster 9 (n=12) can be sub-divided into three sub-clusters (9a, collected 2010, n=2, 9b, collected 2006-13, n=7, and 9c, collected 2007-11, n=3) (Table 7). Clusters 9b and 9c were 23/24 MIRU-VNTR loci identical, and were, in turn, different to cluster 9a by one SLV. Cluster 9b differed from cluster 9c within locus 4052, where cluster 9b contained 2 repeats, and cluster 9c failed to amplify the locus. Cluster 9a differed from 9b and 9c at locus 960, where the former contained 3 repeats and the latter contained 4. Cluster 9b and 9c isolates all displayed low level isoniazid resistance only (all other first-line drugs were susceptible), the remainder were susceptible to all drugs. Hain Genotype MTBDR_{plus} line probe assays confirmed the presence of low-level isoniazid resistance. Two possible family outbreaks were observed within this cluster, involving 5 cases.

A phylogenetic tree of variant files from the above whole genomes was constructed using PhyML (Figure 35). IMRLH101, 102, 107 and 111 formed a clade (< 5 SNVs difference) which

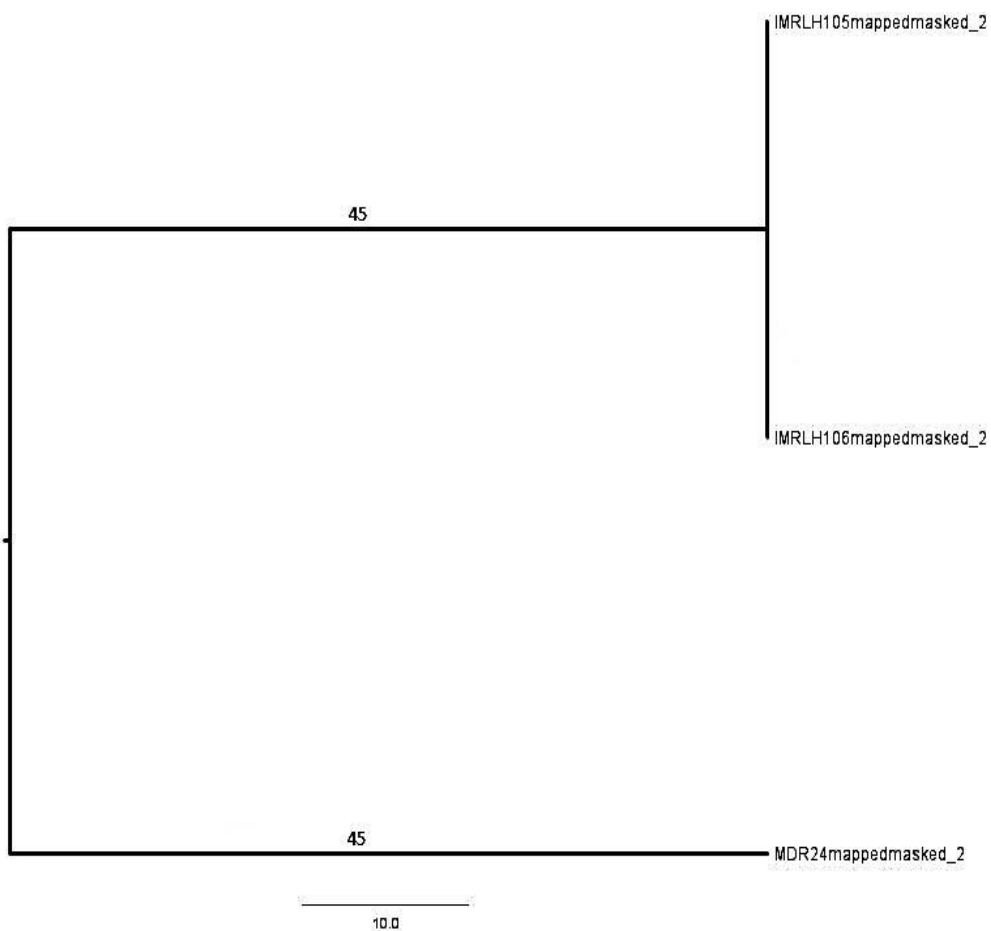


Figure 34. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 8 isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. WGS showed that these isolates were separated by at least 22 SNVs, and therefore un-related, even though their MIRU-VNTR genotypes were identical.

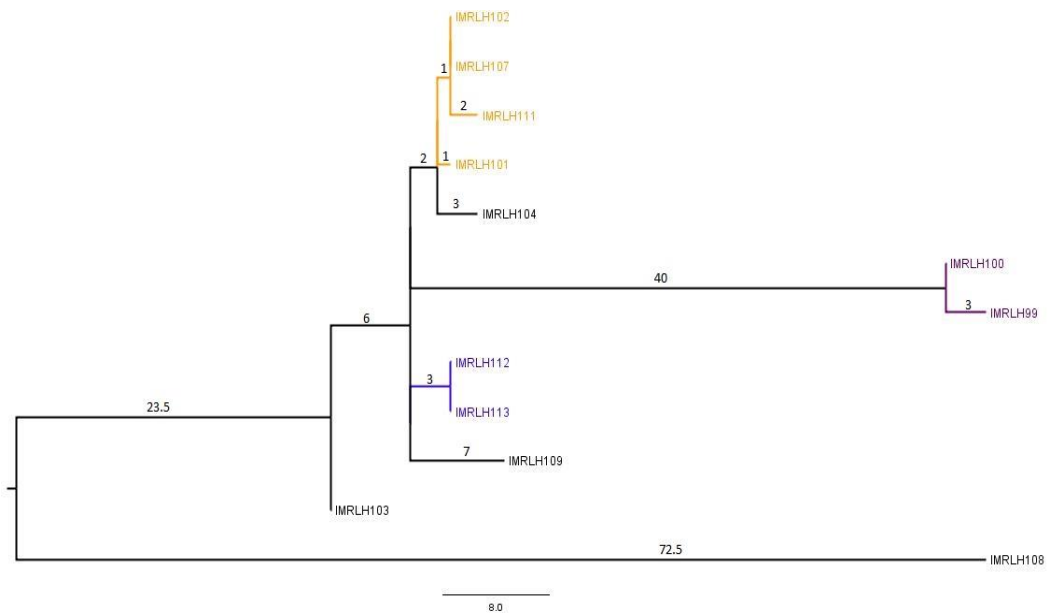


Figure 35. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 9a, 9b, and 9c isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. The tree is separated broadly into 4 clades. WGS agreed with MIRU-VNTR genotyping in most cases except for a major discrepancy (IMRLH108 is the mother of 4-month old IMRLH103, but they are un-related according to WGS). IMRLH102 and 107 represent another household outbreak.

represented recent transmission. IMRLH102 and 107 were family members living at the same address, the latter patient having been the first to present in 2007. IMRLH101 had an address in a different public health area. After the possible index case, the other cases within the clade did not present until 2011, over which time 2 SNVs could have naturally mutated. All but one of the above isolates were within MIRU cluster 9b. IMRLH111 genotyped with cluster 9c. The discrepant MIRU locus failed to amplify, so it is possible that, had it amplified, it would have matched cluster 9b. IMRLH99 and 100 were the two cluster 9a cases, and they were at least 40 SNVs apart from any of the other isolates, which would correlate with the slightly differing MIRU-VNTR genotype. IMRLH100 was a healthcare worker, which could be seen as a TB risk factor. IMRLH112 and 113 were identical cases, one patient from the UK and one from Ireland, that were possibly linked by substance misuse. IMRLH103 and 108 were patients with the same surname and address, however, their isolates differed by 96 SNVs, which would rule out recent transmission. These patients (a mother and 4-month-old daughter) presented in the same month, May 2013. IMRLH109 does not fit well into any of the clades. This patient had a geographically distinct address.

WGS agreed with MIRU-VNTR genotyping in all cases except for one major discrepancy, which is more likely to be related to WGS since an epidemiological link exists between the two cases.

4.2.10 IMRL Cluster 10 – residential institution, Beijing

East Asian lineage 2, sub-lineage Beijing, cluster 10 consisted of 24 isolates collected from 2011-13 from an outbreak recognised originally by public health colleagues in a residential institution (Table 7). This prison outbreak has been publicly documented by the HPSC [22]. This pan-susceptible TB cluster has increased to 29 in the IMRL database as of August 2016. Eight isolates were associated with the institution. Other isolates were collected from counties Dublin, Galway and Laois. The outbreak has subsequently been discovered in the community. The ratio of male to female cases was 11:2 and the median age was 33 years (IQR 27-40). Country of origin was not recorded, however, 62.5% of cases were estimated to be of Irish origin from a survey of patient surnames. The majority of cases were associated with pulmonary TB (88%), however extra-pulmonary cases were also involved from body sites such as pleural fluid. One case of laryngeal TB was also recorded. One possible household outbreak was observed involving two patients. Immunosuppression and contact with a TB case were the most commonly cited TB risk factors.

A maximum likelihood tree constructed using the variant calling files of this cohort of whole genomes (Figure 36) indicates that < 5 SNVs separate them. This is also clearly when genetic distances are observed (Figure 27). The infecting TB strain did not mutate by more than 2 SNVs, which is somewhat in line with the timing of the outbreak (March 2011 to December 2013). This pattern is strongly suggestive of a highly infectious super-spreader. Epidemiologically, the index

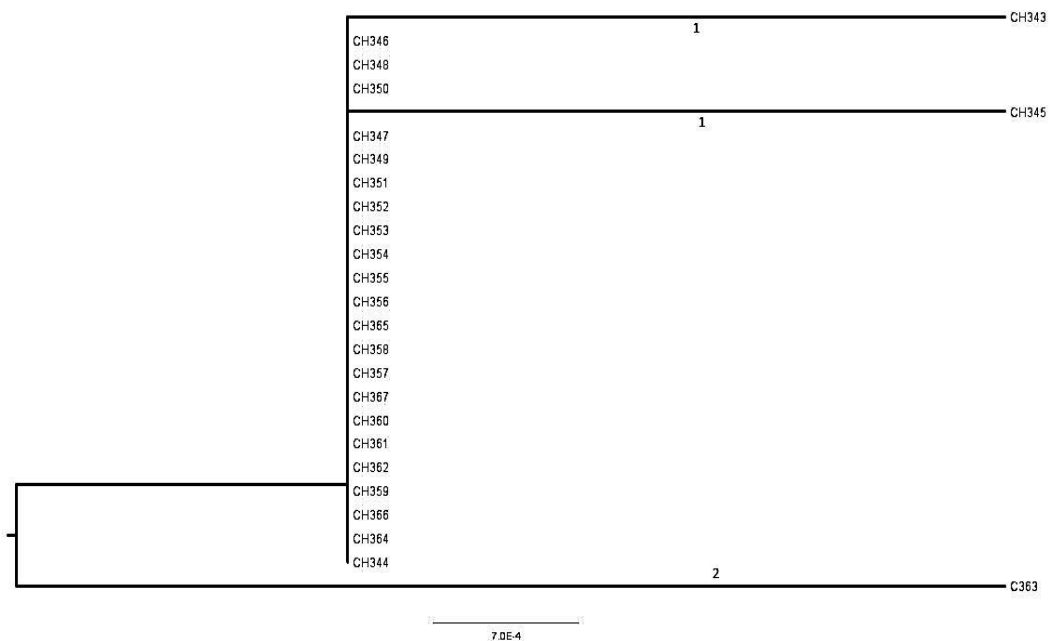


Figure 36. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 10 isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. WGS and MIRU-VNTR genotyping correlate. This entire cluster represents recent transmission both within an institution and in the community. A super-spreader is highly likely to have been involved. CH347 was diagnosed with laryngeal TB, which might lead one to suspect he was the super-spreader.

case was thought to have laryngeal TB (CH347), which is considered a more infectious form of the disease than pulmonary TB [155].

WGS confirmed the MIRU-VNTR genotyping result.

4.2.11 IMRL Cluster 11 – non-residential institution, Euro-American

Euro-American lineage 4 cluster 11 (sub-lineage un-defined) consisted of 8 cases collected from 2013 to 2015 in county Cork (Table 7). This outbreak was associated with a non-residential institution that these patients frequented and was originally flagged by public health colleagues.

Figure 37 shows the PhyML maximum likelihood phylogenetic tree constructed with variant nucleotides from the whole genomes of this cohort. Nine SNVs separated the most distant isolates (ie > 5 - < 12, most likely related). Two separate clades were seen. The first (IMRLH121, 123-125 and 127) were < 5 SNVs apart with addresses within a very small geographical area. IMRLH124 was first to present in February 2013. Variants seem to have been acquired between this case and subsequent cases. IMRLH124 also seems to have been the closest case to the second clade, which included IMRLH120, 122 and 126. IMRLH126 shared an address with members of the first clade, which was another possible route of transmission. No obvious super-spreader was seen.

WGS resolved this outbreak further and raised the possibility of a link between the two separate clades found. MIRU-VNTR has likely over-estimated transmission here, although there was an epidemiological link between some of the cases.

Overall, MIRU-VNTR over-estimated TB transmission in the majority of cases, with WGS delineating these clusters further. However, MIRU-VNTR genotyping correctly proved transmission in two large outbreaks, and WGS was discrepant in two cases of family household contact.

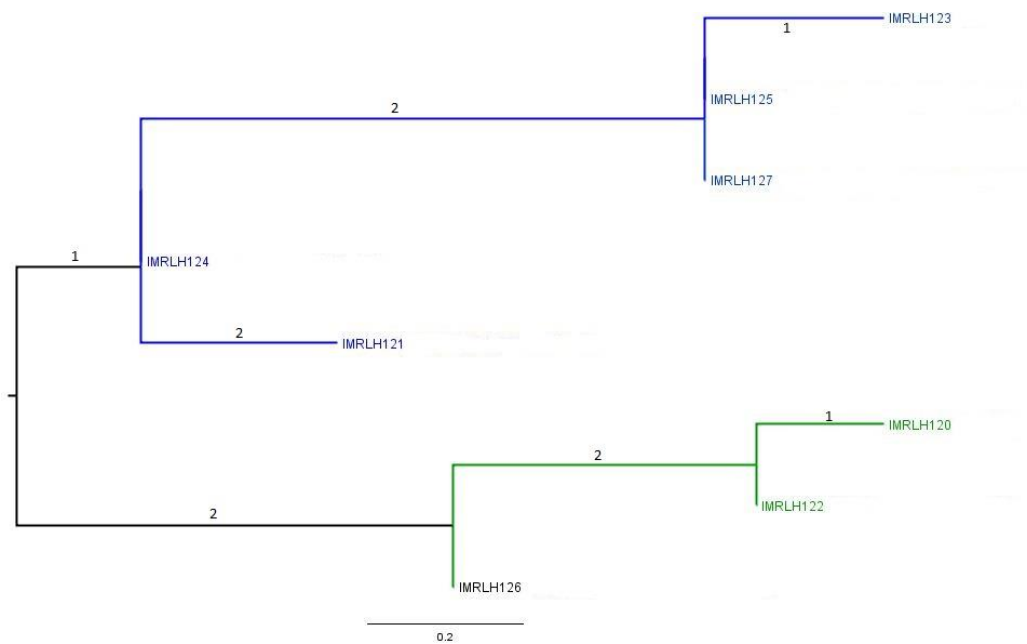


Figure 37. PhyML Maximum likelihood phylogenetic tree constructed using the variant calling files from the whole genomes of Cluster 11 isolates

Sequences were mapped to a H37Rv reference genome. Repeat regions were masked. Branch lengths represent the number of SNVs between each isolate. WGS separates these isolates into two clades, with a possible link case (IMRLH124), however the overall cluster likely constitutes transmission since no isolate is further than 9 SNVs. Substance misuse and geographical region links these cases.

4.3 Discussion

WGS is the ultimate genotyping tool, since it compares the entire genome of each isolate. Eleven MIRU-VNTR clusters of interest were analysed in this study. WGS analysis further resolved and added valuable information in all cases. WGS confirmed MIRU-VNTR genotyping results for Clusters 10 and 11. MIRU-VNTR genotyping was proven to over-estimate clustering in Clusters 1 – 8 inclusive. Despite this, there were many sub-clusters found that did confirm classical genotyping results. Possible super-spreaders were identified within Clusters 2 and 10. It could be hypothesised that the only direction of transmission to be inferred from Cluster 10 is from the prison out into the community. Direction of transmission within the other clusters was challenging to predict.

Ideally, for a study of this nature, all isolates over a certain time-frame should be sequenced, otherwise the extent to which classical genotyping has mis-represented clustering cannot be fully assessed. Resources would not allow for this, however it was hypothesised that performing WGS on clusters of interest would reveal enough information to be useful, and that future research could be based upon the findings. Full access to public health records and enhanced surveillance information on every patient in each cluster would also have been ideal. However, due to data protection and resource constraints, this was not possible, and only certain information was available for the study.

Public health departments in Ireland are divided into four areas. Since it is a relatively small country, and movement of people cannot be monitored, this separation could be seen as detrimental to TB contact tracing unless communication between areas is robust. Contact tracing can be challenging with TB and other infectious diseases. Pertinent tracing information may not always be shared with public health by patients. They may do this for personal reasons, or because they do not think the information is relevant, or they may have forgotten. The CIDR database attempts to monitor infectious diseases across the entire country and depends on the timely notification of new cases. Nevertheless, outbreaks could be missed in this setting. Another aspect of this, and other MTBC WGS studies, is that LTBI confounds the analysis. Infected contacts may present with active infection soon after contact, or may not reactivate until much later in life, or may never become ill with TB, making chains of transmission more complicated to interpret. Gardy *et al* originally investigated an outbreak using WGS and social networking (incidentally the strain was not found in the IMRL database). They are now working on ways to declare a TB outbreak over with genetic epidemiology, in order to concentrate public health resources to the outbreaks where they would be most effective [156].

Cluster 10 posed a particularly challenging outbreak for public health colleagues, since the prison in which it took place was a remand prison where in-mates are regularly moved to other prisons or

back out to the community within relatively short time-frames. Contact tracing, treatment and compliance contributed to a complicated public health workload. This was the first documented case of a prison outbreak Ireland. Genotyping helped public health colleagues to make the case that this was a true outbreak and that resources should be made available to prevent it from being transmitted to the community. Any doubt that this Beijing strain, a strain which has been documented to be over-clustered by MIRU-VNTR genotyping, was not an outbreak was put to rest with these definitive WGS analysis findings. Despite best efforts, this genotype has been found in the community. A super-spreader was almost certainly involved in this outbreak. If super-spreaders could be identified, resources could be focussed on treating and isolating those particular patients until they were proven smear-negative, preventing further transmission; the ultimate goal of genotyping.

NGS technology has its limitations like any other technology. MTBC contains many repeat regions that cannot be reliably sequenced using this short read technology. Every repetitive unit will have the same sequence. Each of those sequences will map to just one of the repeat units of the reference genome repeat region and will not disperse over the entire repetitive area, leading to mapping errors. As a result, repeat regions were masked for the purposes of cluster analysis. While only a small percentage (~5%) of the genome was not covered, some variations could have been missed. However, masking repeat regions is the consensus method for WGS cluster analysis, so results are comparative to other studies.

Anomalies also occurred where epidemiological links were confirmed by classical genotyping, but refuted by WGS analysis. IMRLH32 (Cluster 1, Figure 26) was 141 SNVs apart from his family contacts who lived at the same address. MIRU-VNTR genotyping linked them as identical. Genotyping was repeated on the whole genome DNA extract with the same result. Mapping, variant calling and maximum likelihood phylogenetic reconstructions were also repeated with the same result. WGS could not, however, be repeated. Haarlem has been found to be prevalent in Ireland (Chapter 3), therefore it is possible that this case was contracted elsewhere at some other time and emerged independently of the other cases. His daughter was positive in 2006, his wife in 2009 and he presented in 2010. IMRLH108 (Cluster 9, Figure 35) experienced a similar discrepancy with epidemiological data. IMRLH108 (mother) was 96 SNVs distant from IMRLH103 (4 month old daughter). Perhaps the infant was infected via another route and the index case was not included for analysis here. The WGS result does not seem plausible for either of these cases. Sequencing errors are not always apparent. WGS analysis is based on mathematics and not on intuitive biology. Phylogenetic trees analysed with different parameters, different models of evolution, different analysis software, can produce slightly different results. One must be satisfied that the analysis method used is reliable and reproducible. Nevertheless anomalies will probably

occur. No method is completely error-proof and critical thinking and clinical judgement are required when interpreting this data.

Cluster 2 represented two larger MIRU-VNTR genotype clusters that were separated by one SLV. In cases where there is an epidemiological link, one can be approximately 95% confident that these isolates are related, especially if the difference in the number of repeats at that locus is just one, which it was here (Supply, P., personal communication) [68, 157]. In this case, however, there was no clear epidemiological evidence to work with. Cluster 2a did contain a 'true' cluster with a possible super-spreader, although it is not possible to guess which case that may have been since their genomes are almost all identical. Neither was it possible to determine the direction of transmission. Determining the direction of transmission depends on higher diversity among strains.

For Clusters 5, 6 and 8, the hypothesis that foreign-born patients do not tend to transmit TB within the country they have migrated to, but instead reactivate latent TB, seems to hold [158]. Many of these cases were non-Irish born, and were extra-pulmonary cases, which would suggest that they had contracted TB previously, had reactivated, and were not infectious. Cluster 4a and 4b differed by more than one repeat at a single locus (2 repeats versus 5 repeats), and formed two distant clades with WGS. This upholds the above hypothesis that MIRU-VNTR genotyping can rule out transmission reliably when SLVs differ. The main issue with MIRU-VNTR genotyping occurs when genotypes are identical.

Cluster 1 duration was the longest, 2006-2014, and this may have impacted somewhat on the results, ie that the cluster was actually a number of separate outbreaks. Over this period, if a true clonal expansion was happening, one would have expected just 4 SNVs to have naturally mutated. In reality, the clades are a lot more than 4 SNVs apart (Figures 26 and 27). Conventional genotyping has clearly over-clustered here, however, outbreaks do not tend to last 8 years or more, although it is possible (for example the 'Hamburg' clone) [143]. Figure 27 shows how cryptic outbreaks like Cluster 1 and Cluster 2, where epidemiological evidence may be difficult to obtain, can be visualised more clearly with WGS. Direction of transmission can only be estimated, but would be likely to expand to the right of the clusters shown.

When there were > 5 and < 20 SNVs present, transmission prediction was more difficult to infer. One case could have remained latent for many years while developing SNVs through some unknown mechanism which were then transferred. Epidemiological links would be essential to prove these links. Genetic distances could be inferred with maximum likelihood, however, direction of transmission was not possible to infer from the phylogenetic trees alone, and without epidemiological data available, could not be reliably estimated. Future studies should ideally be planned and coordinated with public health colleagues from the outset. The evidence that there is

not as much clustering in Ireland as classical genotyping would suggest is welcome news from a public health perspective, as is the evidence that multi-drug resistant isolates have not been transmitted among individuals in this cohort.

Results show that WGS can both rule-in and rule-out outbreaks with greater discrimination than MIRU-VNTR genotyping and can confirm recent transmission events, making it a valuable tool in the fight against TB. While MIRU-VNTR serves as an excellent first-line method to rule out transmission, WGS analysis is necessary in cases where MIRU-VNTR genotypes are identical in order to truly confirm or deny recent transmission. Decreasing costs associated with WGS will allow this invaluable technology to be used in the diagnostic setting. This issue will be discussed in the following chapters. From this and other studies, it is clear that WGS alone cannot infer all epidemiological information [145]. There is still much to be learned about MTBC genomic diversity and transmission before that would be possible. Epidemiological data, contact tracing and genotyping all remain essential tools for MTBC cluster and outbreak investigation. However, WGS could be an attractive method for best directing Public Health resources, i.e. if a cluster is confirmed by WGS, and a possible super-spreader has been found, resources would best be directed towards that patient and their close contacts, whereas if WGS rules out recent transmission, further resources could be saved in that instance, and re-directed to where they are required.

Chapter 5.

**Molecular Characterisation of MDR/XDR-TB in
Ireland, collected 2001-14,**

and

**Evaluation of WGS Analysis Techniques compared
to Conventional Phenotypic Methods for Drug
Resistance Prediction**

5 Molecular Characterisation of MDR-/XDR-TB in Ireland, collected 2001-2014, and Evaluation of WGS Analysis Techniques compared to Conventional Phenotypic Methods for Drug Resistance Prediction

5.1 Introduction

Multi- and extensive- anti-tuberculous drug resistance threatens the global management of tuberculosis [58, 159]. One hundred and five countries in the world, including low-prevalence countries, like Ireland, have reported XDR-TB to date [1, 160-166]. MDR-TB occurs when an isolate displays resistance to rifampicin and isoniazid. Extensively-drug-resistant TB displays resistance to the above plus a fluoroquinolone and an injectable aminoglycoside. Both MDR- and XDR-TB require complicated, lengthy treatment (approximately 2 years) which can include many side-effects and may not be successful (50% global success rate reported in 2015) [1]. Treatment was estimated to be upwards of €10,000 for susceptible TB, €57,000 in the case MDR-TB, and over €170,000 in the case of XDR-TB [167]. The WHO reports that 111,000 people were commenced on MDR-TB therapy in 2014, however it is estimated that 300,000 people developed MDR-TB but were never adequately tested [1]. MDR/XDR-TB poses an increasing threat to public health and TB control in Ireland. From 1998-2014, the HPSC reported 7,162 cases of TB, of which MDR/XDR-TB represented 0.6% (n=42). Even though numbers are relatively low, MDR-TB cases increased year on year from 2003 to a high of seven cases in 2007, and have been present in variable numbers since then [22].

Due to the slow and fastidious growth patterns of TB, drug resistance can take many weeks to confirm in the diagnostic laboratory [114]. Phenotypic drug susceptibility testing is not always reliable, especially in the case of pyrazinamide [168, 169]. Various rapid molecular tests have been developed to determine drug resistance (e.g. line probe assays Hain GenoType MTBDR*plus* and MTBDR*sl* for direct respiratory specimens or cultures, and Cepheid GenXpert MTB/RIF for direct specimens) [122, 123, 170]. GenoType MTBDR*plus* detects mutations in the *rpoB* gene which confer resistance to rifampicin, and *inhA* and *katG* genes which confer resistance to low- and high-level isoniazid respectively [171, 172]. GenoType MTBDR*sl* version 1.0 detects mutations in the genes *gyrA* and *embB* which are predictive of phenotypic resistance to fluoroquinolones and ethambutol respectively, and mutations in the *rrs* region which are thought to be involved in aminoglycoside resistance [173-177]. GenoType MTBDR*sl* version 2.0 replaces *embB* with probes for the genes *gyrB* and *eis*, involved in fluoroquinolone and kanamycin resistance [178-180]. While extremely useful, rapid molecular assays cannot cover all possible mutations, and do not take into account the development of novel mutations across the entire genome.

WGS is becoming more accessible for the diagnostic laboratory (as discussed in other chapters), although data analysis remains challenging [114, 181, 182]. Studies, including the international

collaboration that will be discussed in Chapter 6, have shown that WGS using Illumina Next Generation Sequencing (NGS) is a 'rapid' and 'comprehensive' method for drug resistance profiling and epidemiology of TB [114]. The TB genome is relatively stable, therefore lends itself well to WGS [8, 32, 183]. Several international studies have been undertaken to determine the correlation between WGS for genotypic detection of drug-resistance associated mutations present in TB genomes and phenotypic drug susceptibility testing (DST) results [61, 76, 77, 121], to the benefit of the wider scientific community working with TB. The better the correlation between genotypic resistance prediction and phenotypic DST, the more confidence can be placed in the variant, either as a resistance determinant, a benign polymorphism, or a phylogenetic (or lineage-defining) variation.

Walker and Kohl *et al* designed an algorithm to predict anti-tuberculous drug resistance in a retrospective cohort study which involved collation of a resistance mutation catalogue, 2,099 MTBC isolates used as a training set on which to test that catalogue, and a validation set of 1,552 further genomes on which the final set of drug-resistance-associated could be verified [184]. With their final set of resistance determinants, they were able to predict phenotypic resistance (for all drugs that had phenotypic DST available) in the validation set with sensitivity ranging from 45.5% (95% CI 16.7-76.6) for ofloxacin to 96.8% (CI 94.1-98.5) for rifampicin, and specificity ranging from 94.2% (CI 92.7-95.4) for ethambutol to 100% (99.4-100) for ofloxacin.

Coll *et al* also designed a drug resistance algorithm and catalogue of validated mutations for 11 anti-TB drugs using 792 MTBC strains from six countries. They developed a freely-available, online, drug resistance prediction research tool, TB Profiler, which accepts raw output data from NGS platforms (fastq files), and rapidly reports lineage (using phylogenetic SNVs) and drug resistance using the validated mutations [77]. Compared to phenotypic DST, sensitivity ranged from 60% (CI 29.6-90.4) for moxifloxacin to 96.2% (CI 93.9-98.5) for rifampicin. Specificity ranged from 68.7% (CI 52.6-84.8) for moxifloxacin to 100% (CI 100-100) for isoniazid.

Feuerriegel *et al* almost simultaneously published another freely-available online research tool to predict resistance and lineage in MTBC with an independent resistance mutation catalogue, Phylo-Resistance Search Engine (PhyResSe) [76]. Raw fastq data can be uploaded, where it undergoes vigorous quality control at the raw data, read mapping and SNV calling levels, followed by prediction of resistance and lineage calling using data from a well-characterised set of in-house isolates from around the world. Results were compared with Sanger sequencing results, rather than phenotypic DST, and concordance of 97.83% to 100% were achieved.

Genomic sequencing analysis represents a new scientific paradigm. Many scientists are unfamiliar with programming languages and using the Linux command line to write scripts or design custom

analysis pipelines that can handle large amounts of data output, and may therefore struggle with NGS analysis. PhyResSe and TB Profiler are freely-available web-tools (each incorporates a catalogue of drug-resistance-associated mutations and promoter regions) that aim to make this process more accessible to the user [76, 77]. ReSeqTb is a data-sharing platform which seeks to consolidate information on mutations that have strong links to phenotypic drug resistance [121].

A study on TB drug resistance in an Irish hospital population from 1991 to 2001 found eight cases of MDR-TB [185]. This represented an increase on a previous survey, which had included TB cases from 1982-85 [186]. They found that a previous history of TB and being a foreign national were the most important risk factors associated with development of MDR-TB. Researchers stated that TB disease decreased dramatically in Ireland due to the fact that the 1947 Health Act made it part of the national public health program, which meant it was free to all. They underlined that the decrease had led to detrimental gaps in the knowledge around TB patient management, and the increase of iatrogenic resistance because of treatment mismanagement and failure. This was cited as a more important issue than non-Irish-nationals emigrating from the newer European Union (EU) accession countries at the time.

Ireland has experienced a small, variable, but steady growth in MDR/XDR-TB from 2001-2014 (the HPSC reported no cases in 2002 to 7 cases in 2007) [22]. This provided the impetus to perform a survey of the isolates collected from that period. The survey included conventional techniques such as MIRU-VNTR genotyping, rapid molecular line probe assays, and phenotypic DST. In order to characterise the mutations that are present in the MDR/XDR-TB cohort in Ireland, WGS analysis was performed on each isolate and analysed using the afore-mentioned platforms that were designed to detect drug-resistance-associated mutations. The isolates could then be compared to those from other jurisdictions in terms of their molecular epidemiology and drug resistance. Correlation between WGS resistance prediction and conventional DST was investigated, and comparison of methods used for genomic analysis, such as PhyResSe, TB Profiler, ReseqTB resistance mutation catalogue and Walker and Kohl *et al* algorithm was performed.

It was hypothesised that (a), MDR/XDR-TB strains are not being readily transmitted in the Irish population, (b), that the drug resistant strains would most likely be those found currently circulating in Europe, (c), that WGS for drug resistance detection would match phenotypic DST in the cohort being tested and could be extremely useful for rapid diagnosis of MDR/XDR-TB cases in the routine diagnostic laboratory.

5.2 Results

From 1998-2014 there were 7,162 cases of TB reported in Ireland by the HPSC, of which 42 were MDR/XDR-TB, representing 0.6% of cases. The first HPSC-recorded case of MDR-TB was in 1998, although drug resistance was documented prior to this, and the first XDR-TB was reported in 2005 [12, 40, 41].

5.2.1 Sample Selection and Patient Demographics

Forty two MDR/XDR-TB isolates from 41 individual patients were collected. Patient demographics are detailed in Table 8. One patient was infected with two different MDR strains at different time points (2004 and 2007). Three isolates were not whole genome sequenced. One isolate included was clinically suspected of MDR-TB although never proven diagnostically (IEMDR26). Seventeen isolates had been processed as specimens at the IMRL, while the remainder were referred from external laboratories. When IMRL reports were examined, the average amount of time taken from receipt to final report (which would have included referral to a supra-national reference laboratory for second-line susceptibilities) was 170 days (range 38-190).

Patient demographics are recorded in Table 8. Isolates were collected nationally. One isolate was collected from a prison and one was from a direct provision centre for asylum seekers. Cases were 58.5% male, 39% female, and 2.5% of unknown gender. Median age at presentation was 34 years (IQR 29-40). Seventy-three per cent (n=30) of cases were pulmonary and 5% (n=2) were extra-pulmonary in nature, while the remainder were unknown. Twenty-seven cases (66%) were confirmed non-Irish born and 3 cases were confirmed Irish-born (7%). The remainder were unknown (n=11, 27%), however, from an internet survey of the origins of their surnames, a further 5 were estimated to have been born in Ireland, 5 were estimated to have been born elsewhere, and one case remained unknown. Countries of origin (both confirmed and estimated) of MDR isolates were extremely diverse: 41.5% Russia or former Soviet State, 19.5% Western Europe (Irish-born), 17.1% Africa, 17.1% Asia, 2.4% South America and 2.4% unknown. Being from a high TB burden country was the most common risk factor among cases. Eight patients had documented previous history of TB and treatment (19.5%), 3 of whom were associated with non-compliance with treatment, while two have no documented history. The others remain unknown. Twenty four out of forty-one (59%) patients had confirmed HIV negative status, 5/41 (12%) of patients had confirmed HIV co-infection, the remainder (12/41, 29%) had unknown HIV status or were not tested for HIV. Three further patients were recorded as being immune-compromised due to hepatitis C virus infection.

Study No.	M/F	Address	Country of Origin	Age	TB Risk Factors	Specimen Type	Collected	Lineage	MIRU-VNTR Genotype	MbC 15-9
MDR01	M	TIPPERARY	IRELAND	37	NK	SPUTUM	2004	GHANA	223263342334425143233613	67-25
MDR01E/EARLY	M	TIPPERARY	IRELAND	34	NK	SPUTUM	2001	GHANA	223263342334425143233613	67-25
IE/EXDR1	F	DUBLIN	LITHUANIA	26	I	SPUTUM	2005	BEIJING	244233352644425173353723	100-32
MDR03	M	DUBLIN	IRELAND	49	NK	SPUTUM	2004	EAI	215834379493266223342713	1740-44
MDR04*	M	DUBLIN	LITHUANIA	34	1,2	SPUTUM	2004	BEIJING	244233352644425173353723	100-32
MDR05	M	DUBLIN	INDIA	27	1,3	NK	2005	EAI	224334382363144223374613	1741-212
MDR06	F	DUBLIN	NIGERIA	31	I	SPUTUM	2007	S	23334312434225143233822	1742-17
MDR07	M	SLIGO	?RUSSIA	24	NK	SPUTUM	2007	BEIJING	244233352644425173353723	100-32
MDR08	F	DUBLIN	GEORGIA	33	1,4,5	SPUTUM	2007	LAM	132254332224125153322622	843-52
MDR09*	M	DUBLIN	LITHUANIA	37	1,2,4	SPUTUM	2007	LAM	132244332224125153322622	121-52
MDR10	M	MEATH	LITHUANIA	37	I	NK	2007	BEIJING	244233352644425173353723	100-32
MDR11	F	NK	ZIMBABWE	29	1,6	URINE	2008	BEIJING	24421323242411614352102	1743-54
MDR12	F	DUBLIN	AZERBAIJAN	29	1,4,5	SPUTUM	2009	BEIJING	244233352644425153353623	1065-32
MDR13	M	DUBLIN	ZIMBABWE	42	1,6	NK	2009	S	23334312334225143233a22	1772-17
MDR14	F	NK	MONGOLIA	28	I	NK	2006	BEIJING	244233362544425173353823	1773-32
MDR15	F	DUBLIN	?ALBANIA	27	NK	NK	2003	BEIJING	244233352644425173353723	100-32
MDR16	M	NK	NK	19	NK	SPUTUM	2005	LAM	132275332224126153322_22	?-51
MDR17	M	DUBLIN	ROMANIA	19	I	SPUTUM	2010	EURO-AMERICAN	222253122434225143335922	1655-15
MDR18	M	WESTMEATH	CHINA	26	1,6	BAL	2010	BEIJING	244233342_4425173353323	?-32
MDR19	M	DUBLIN	LATVIA	39	1,6	SPUTUM	2011	URAL	23523723244425113323632	163-15
MDR20	M	DUBLIN	UKRAINE	30	1,4,5,6,7	SPUTUM	2011	BEIJING	244233352644425173353723	100-32
MDR21	F	DUBLIN	SOUTH AFRICA	34	1,6	NK	2011	LAM	134254332224121143322722	8075-482
MDR22	M	GALWAY	INDIA	16	I	NK	2012	DELHI/CAS	222236452244225173353623	8958-32
MDR23	F	DUBLIN	MONGOLIA	29	I	SPUTUM	2012	BEIJING	244233362544425173353823	1773-32
MDR24	M	MEATH	NIGERIA	37	I	SPUTUM	2012	LAM	22421333154422515333_522	?-2,6
MDR25	F	GALWAY	INDIA	38	I	PLEURAL BX	2012	DELHI/CAS	222236452244225173353623	8958-32
MDR26	F	NK	IRELAND	45	NK	NK	2010	HAARLEM	213226332434425153333732	?-15
MDR27	M	DUBLIN	LATVIA	40	1,2,4,8	SPUTUM	2014	LAM	132254332224125153322622	843-52
MDR27/LATER	M	DUBLIN	LATVIA	40	1,2,4,8	SPUTUM	2014	LAM	132244332224125153322622	843-52
MDR28	F	MEATH	MOLDOVA	29	1,4,6	SPUTUM	2011	LAM	132253(4)3322241251533223(6)22	?-52
MDR29	F	DUBLIN	SOMALIA	24	I	BAL	2013	EURO-AMERICAN	214225132134425113333a22	12428-15
MDR30	M	LAOIS	LITHUANIA	44	1,4,9	SPUTUM	2013	URAL	235237232244425113323632	163-15
MDR31	F	DUBLIN	ZIMBABWE	32	1,4,6	SPUTUM	2013	LAM	244214132324116152442822	12880-443
MDR32	F	DUBLIN	RUSSIA	37	1,6	SPUTUM	2013	BEIJING	244233352644425173353723	100-32
MDR33	M	NK	?LITHUANIA	50	NK	LUNG	2012	BEIJING	244233352644425153353823	94-32
MDR34	M	CORK	RUSSIA	49	I	SPUTUM	2012	BEIJING	244233352644425153353823	94-32
MDR35	M	ROSCOMMON	LATVIA	53	6,8	SPUTUM	2014	LAM	132254332224125153322622	843-52
MDR36	M	NK	?IRELAND	67	NK	SPUTUM	2005	EURO-AMERICAN	214213322434226153334122	?-62
MDR37 (WGS NP)	F	NK	?IRELAND	80	NK	SPUTUM	2005	X	224224342234425153332832	?-15
MDR38 (WGS NP)	M	KILDARE	?IRELAND	40	NK	SPUTUM	2005	DELHI/CAS	232236442244225143353743	9045-32
MDR39 (WGS NP)	M	NK	?IRELAND	76	NK	SPUTUM	2005	EURO-AMERICAN	22424312242422515335522	1286-15
MDR40	M	DUBLIN	?IRELAND	53	NK	SPUTUM	2005	HAARLEM	224224332334425153333732	?-15
MDR41	M	LOUTH	IVORY COAST	33	NK	SPUTUM	2004	GHANA	223263342334425143233613	67-25
MDR42	NK	NK	?SOUTH AMERICA	18	NK	SPUTUM	2001	BEIJING	244233352644425153353823	94-32

Table 8. MDR/XDR-TB Patient demographics.

Demographics included were gender, address, country of origin, age at presentation, risk factors for TB, specimen type, year of collection, sub-lineage, MIRU-VNTR genotype and MtbC15-9 code. * indicates two different MDR strains separately isolated from the same patient. TB risk factors: 1 -From a high TB burden country, 2-smoker, 3-occupation (HCW), 4-previous history of TB, 5-non-compliance with treatment, 6-immunocompromised, 7-intra-venous drug user (IVDU), 8-alcohol misuse and 9-incarcerated. Age corresponds to age in years at time of presentation. WGS NP – whole genome sequencing not performed, BAL – bronchoalveolar lavage, ? indicates that the country of origin has not been confirmed, NK – not known. Sequential isolates were included (IEMDR01 and IEMDR01EARLY, MDR27 and MDR27LATER).

5.2.2 Molecular Epidemiology of the MDR/XDR-TB Cohort – MIRU-VNTR Genotyping

Forty two MDR/XDR-TB isolates were MIRU-VNTR genotyped (detailed in Table 8). The distribution of lineages was as follows: 54.7% Euro-American Lineage 4 (n=23), 33.3% East Asian Lineage 2 (n=14), 7.2% East African Indian Lineage 3 (n=3) and 4.8% Indo Oceanic Lineage 1 (n=2). One patient had two different strains of MTBC; a Beijing strain in 2004 and a LAM strain in 2007.

A neighbour-joining phylogeny was constructed using the categorical 24-locus MIRU-VNTR genotypes, which details the country of origin and MtbC15-9 code of each isolate (Figure 38). Seven MLVA MtbC15-9 clusters were identified through 24-locus MIRU-VNTR genotyping: East Asian Lineage 1, sub-lineage Beijing 100-32 (n=7), 94-32 (n=3) and 1773-32 (n=2), Euro-American Lineage 4, sub-lineage Ghana 67-25 (n=2), sub-lineage LAM 843-52 (n=3) and sub-lineage Ural 163-15 (n=2), and East African Indian Lineage 3, sub-lineage Delhi/CAS 8958-32 (n=2). The remainder were unique (n=20) (Table 8).

The Ghana cluster (163-15) was observed in an Irish-born patient and a non-Irish-born patient (an asylum seeker from the Ivory Coast, 7 months in Ireland, presented in December 1999) with unknown epidemiological links. The Beijing cluster (100-32) was collected from patients born in the former Soviet Union, two of whom were immune-compromised, and two of whom were confirmed as XDR-TB. No known links were found within Beijing cluster 94-32. The LAM (843-52) cluster included a patient who progressed from MDR- to XDR-TB over the course of treatment (further details in section 5.2.8). Each of the cases originated from former Soviet Union countries. The Ural cluster found (163-15) included patients from Latvia and Lithuania. One of the patients was incarcerated and had shared a cell with two others. It was unclear if the other individual, who was hepatitis C positive, had spent time in prison. This strain is consistent with nosocomial outbreaks in the Republic of Moldova associated with long-term in-patient care [187]. The Delhi/CAS cluster (8958-32) involved a family household outbreak in a mother and son who were originally from India. The final cluster (Beijing, 1773-32) involved two individuals with no documented link except for the fact that they were Mongolian-born. The first patient presented in 2006, and the second did not arrive in Ireland until 2009 and presented in 2012.

All of these clustered genotypes have been found to be identical to genotypes that have been implicated in cross-border clustering by the European Centre for Disease Control (ECDC) from 2003 to 2014 [188].

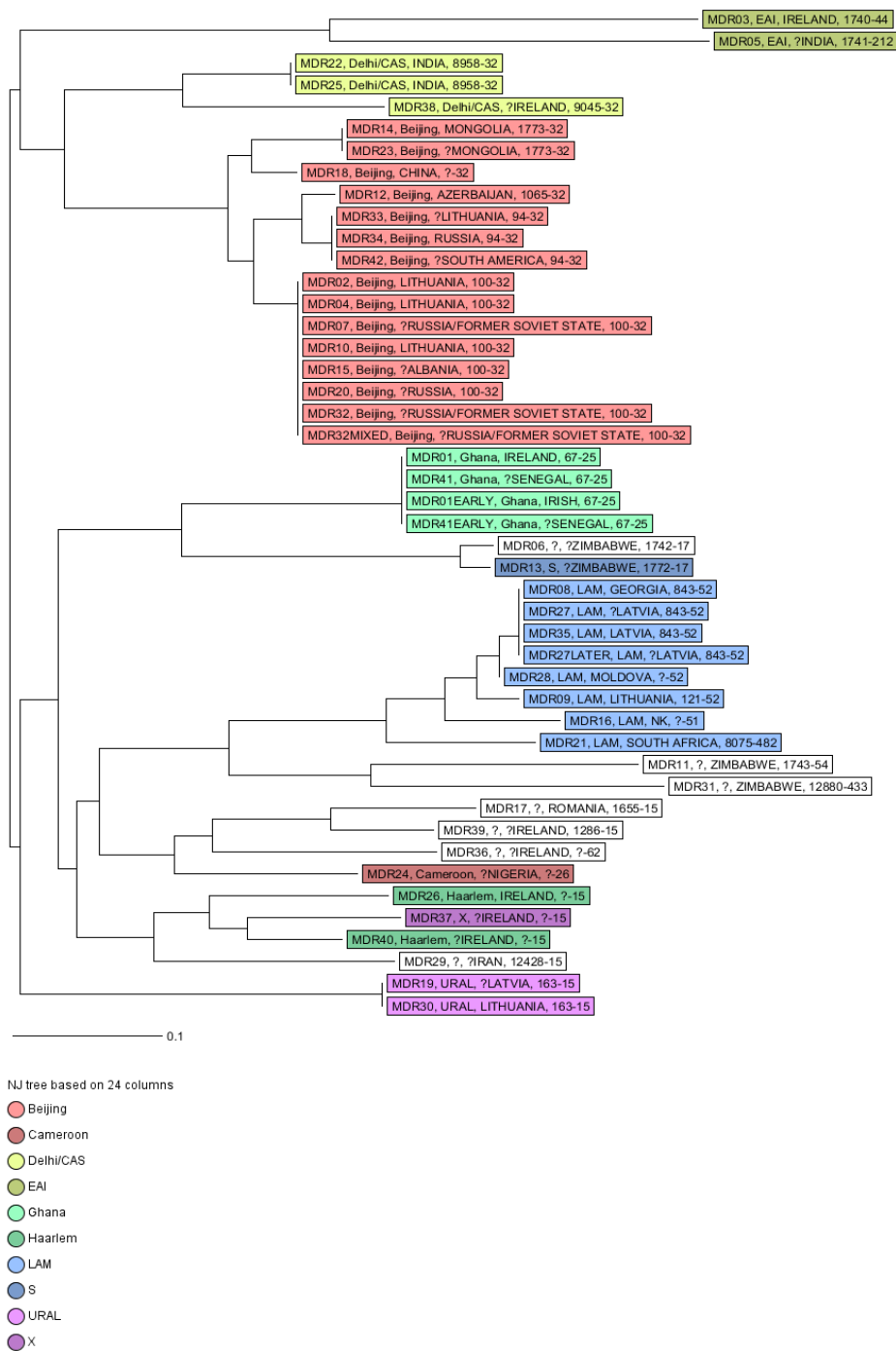


Figure 38. MIRU-VNTRplus database neighbour joining phylogeny constructed with MIRU-VNTR genotyping data from MDR/XDR-TB isolates collected in Ireland from 2001-2014.

Colours indicate the lineage assigned according to MIRU-VNTR genotyping pattern detected. No colour indicates that no lineage had been assigned to that isolate. Mtbc 15-9 codes are included for each isolate, as are countries of origin. ? indicates that the individual's country of origin was estimated.

5.2.3 Molecular Epidemiology of the MDR/XDR-TB Cohort – Whole Genome Sequencing compared to Conventional Genotyping

5.2.3.1 Global Lineage Assignment

Web-tools TB Profiler and PhyResSe were used to compare lineage-calling with WGS analysis against conventional MIRU-VNTR genotyping. Of thirty-nine MDR/XDR-TB isolates whole genome sequenced, WGS analysis matched the global lineages called by MIRU-VNTR genotype in all cases (Table 8 and 9). Minor differences were found when sub-lineage was investigated.

5.2.3.2 Cluster Analysis

5.2.3.2.1 Where Whole Genome Sequencing and MIRU-VNTR Genotyping correlated

A maximum likelihood phylogenetic tree, which incorporates MtbC15-9 within its taxa, was constructed, with the PhyResSe online web-tool, using the whole genomes of the MDR/XDR-TB cohort (Figure 39). A confirmatory maximum likelihood tree was created with Seaview software [72], which is visualised as a transformed cladogram in Figure 40. WGS analysis differed with MtbC15-9 clusters in some cases. Similar parameters were used to indicate recent transmission between cases as that used for outbreak analysis in Chapter 4 (ie < 5 SNVs representative of recent transmission).

Conventional genotyping grouped IEMDR02, 04, 07, 10, 15, 20 and 32 together, denoted by MLVA MtbC15-9 100-32 (Figure 38). When the whole genomes of these isolates were analysed, 14 SNVs was the most that distinguished any of them, which indicates that these lie in the grey-zone (ie >12 - <20 SNVs). Epidemiological information would be required to link the furthest cases. A sub-cluster which included IEMDR02 (collected 2005), IEMDR10 (collected 2007), and IEMDR32 (collected 2013) were within 5 SNVs of each other, which would suggest transmission.

MtbC15-9 8958-32 clustered IEMDR22 and 25 together. WGS phylogenies showed that they were separated by just 2 SNVs (<5 SNVs), confirming the MIRU-VNTR genotyping result within this household outbreak setting.

MtbC15-9 67-25 clustered IEMDR01 and IEMDR41 as identical (Ghana sub-lineage). WGS confirms probable transmission between these two cases, which could represent the first confirmed transmission of MDR-TB in Ireland. Both patients had second isolates that were confirmed identical MIRU-VNTR, and IEMDR01EARLY represents an earlier isolate of IEMDR01 (collected 2001) that also matched IEMDR41 genomically (ie < 5 SNVs apart, Figure 38).

Table 9. Lineage assignation according to the whole genomes of the MDR/XDR-TB cohort, collected in Ireland 2001-14., assigned by TB Profiler and PhyResSe.

Whole Genome global lineage can be compared to MIRU-VNTR genotyping lineage in Table 7. This table also represents a comparison between web-tools TB Profiler and PhyResSe. TB Profiler has assigned spoligotype and region of difference from the whole genome as well as the global lineage. PhyResSe has called the global lineage alone [76, 77].

STUDY NO.	TB PROFILER Lineage	TB PROFILER Main spoligotype	TB PROFILER RoD	PHYRESSE Lineage
IEMDR01	Euro-american 4, 4.1	LAM; T; S; X; H	RD105, RD105; RD207, RD105;RD207;RD181	Ghana
IEMDR02	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD239	Beijing
IEMDR03	Indo-oceanic 1, 1.2.1	EAI, EAI2	RD105, RD105; RD207, RD105;RD207;RD181	EAI 'Manila'
IEMDR04	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD239	Beijing
IEMDR05	Indo-oceanic 1, 1.1, 1.2	EAI, EAI3; EAI4; EAI5; EAI6; EAI13; EAI15	RD239	EAI
IEMDR06	Euro-american 4, 4.4, 4.4.1, 4.4.1.1	LAM; T; S; X; H; S; T; S; Orphans	RD105, RD105; RD207, RD105;RD207;RD181	'S-type'
IEMDR07	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD115	Beijing
IEMDR08	Euro-american 4, 4.3, 4.3.4, 9	LAM; T; S; X; H; LAM; T; T1	RD115	LAM
IEMDR09	Euro-american 4, 4.3, 4.3.3	LAM; T; S; X; H; mainly LAM, LAM; T	RD115	LAM
IEMDR10	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD105, RD105; RD207, RD105;RD207;RD181	Beijing
IEMDR11	Euro-american 4, 4.3, 4.3.4, 4.3.4.2	LAM; T; S; X; H; mainly LAM, LAM, LAM1; LAM4; LAM11	RD174	LAM
IEMDR12	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD105, RD105; RD207, RD105;RD207;RD181	Beijing
IEMDR13	Euro-american 4, 4.4, 4.4.1, 4.4.1.1	LAM; T; S; X; H; S; T; S; Orphans	RD105, RD105; RD207, RD105;RD207;RD181	'S-type'
IEMDR14	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD105, RD105; RD207, RD105;RD207;RD181	Beijing
IEMDR15	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD105, RD105; RD207, RD105;RD207;RD181	Beijing
IEMDR16	Euro-american 4, 4.3, 4.3.3	LAM; T; S; X; H; mainly LAM, LAM; T	RD115	LAM
IEMDR17	Euro-american 4, 4.8	LAM; T; S; X; H; T1; T2; T3; T5	RD219	Euro-American Super-lineage
IEMDR18	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD105, RD105; RD207, RD105;RD207;RD181	Beijing
IEMDR19	Euro-american 4, 4.2, 4.2.1	LAM; T; S; X; H; H; T; LAM, H3; H4	RD105, RD105; RD207, RD105;RD207;RD181	Beijing
IEMDR20	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD105, RD105; RD207, RD105;RD207;RD181	Ural
IEMDR21	Euro-american 4, 4.3, 4.3.3	LAM; T; S; X; H; mainly LAM, LAM; T	RD115	Beijing
IEMDR22	East African Indian 3, 3.1.2	CAS, CAS; CAS2	RD750	LAM
IEMDR23	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD105, RD105; RD207, RD105;RD207;RD181	Delhi-CAS
IEMDR24	Euro-american 4, 4.6, 4.6.2, 4.6.2.2	Beijing, Beijing RD207, Beijing RD181	RD726	Beijing
IEMDR25	East African Indian 3, 3.1.2	LAM; T; S; X; H; T; LAM, LAM10-CAM	RD750	Cameroon
IEMDR26	Euro-american 4, 4.1, 4.1.2, 4.1.2.1	CAS, CAS; CAS2	RD182	Delhi-CAS
IEMDR27	Euro-american 4, 4.3, 4.3.3	LAM; T; S; X; H; T; X; H; T; H; T1; H1	RD115	Haarlem
IEMDR28	Euro-american 4, 4.3, 4.3.3	LAM; T; S; X; H; mainly LAM, LAM; T	RD115	LAM
IEMDR29	Euro-american 4, 4.2, 4.2.2	LAM; T; S; X; H; H; T; LAM, T; LAM7-TUR	RD115	Euro-American Super-lineage
IEMDR30	Euro-american 4, 4.2, 4.2.1	LAM; T; S; X; H; H; T; LAM, H3; H4	RD174	Euro-American Super-lineage
IEMDR31	Euro-american 4, 4.3, 4.3.4, 4.3.4.2, 4.3.4.2.1	LAM; T; S; X; H; mainly LAM, LAM, LAM1; LAM4; LAM11, LAM11	RD105, RD105; RD207, RD105;RD207;RD182	Ural
IEMDR32	East Asian 2, 2.2, 2.2.2	Beijing, Beijing RD207, Beijing RD182	RD105, RD105; RD207, RD105;RD207;RD183	LAM
IEMDR33	East Asian 2, 2.2, 2.2.3	Beijing, Beijing RD207, Beijing RD183	RD105, RD105; RD207, RD105;RD207;RD183	Beijing
IEMDR34	East Asian 2, 2.2, 2.2.4	Beijing, Beijing RD207, Beijing RD184	RD105, RD105; RD207, RD105;RD207;RD184	Beijing
IEMDR35	Euro-american 4, 4.3, 4.3.3	LAM; T; S; X; H; mainly LAM, LAM; T	RD115	LAM
IEMDR36	Euro-american 4, 4.9	LAM; T; S; X; H; T1	RD183	Euro-American Super-lineage
IEMDR40	Euro-american 4, 4.1, 4.1.1, 4.1.1.1	LAM; T; S; X; H; T; X; H; X1; X2; X3; X2	RD105, RD105; RD207, RD105;RD207;RD181	Euro-American Super-lineage
IEMDR41	Euro-american 4, 4.1	LAM; T; S; X; H; T; X; H	RD105, RD105; RD207, RD105;RD207;RD181	Ghana
IEMDR42	East Asian 2, 2.2, 2.2.1	Beijing, Beijing RD207, Beijing RD181	RD105, RD105; RD207, RD105;RD207;RD181	Beijing

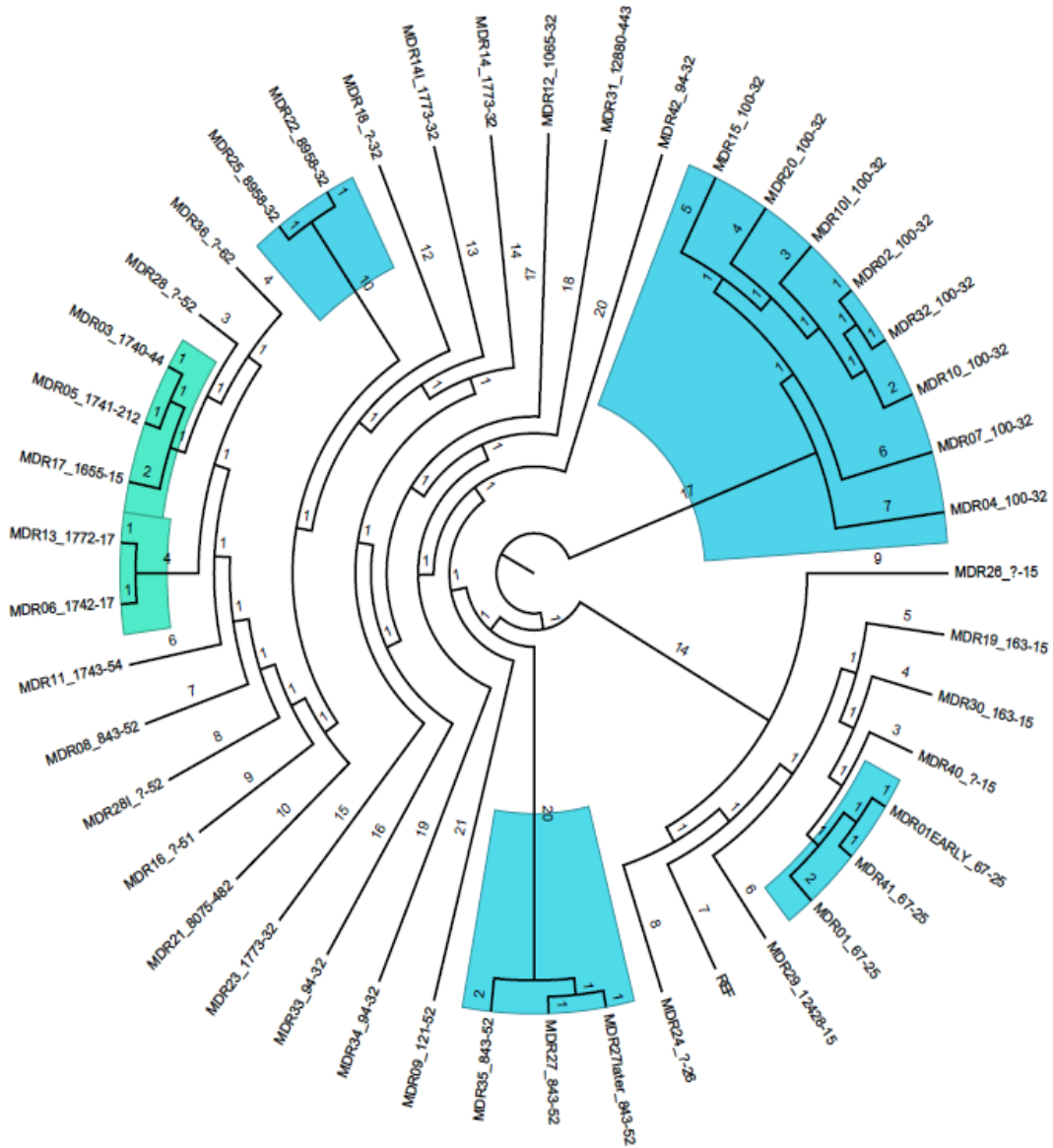
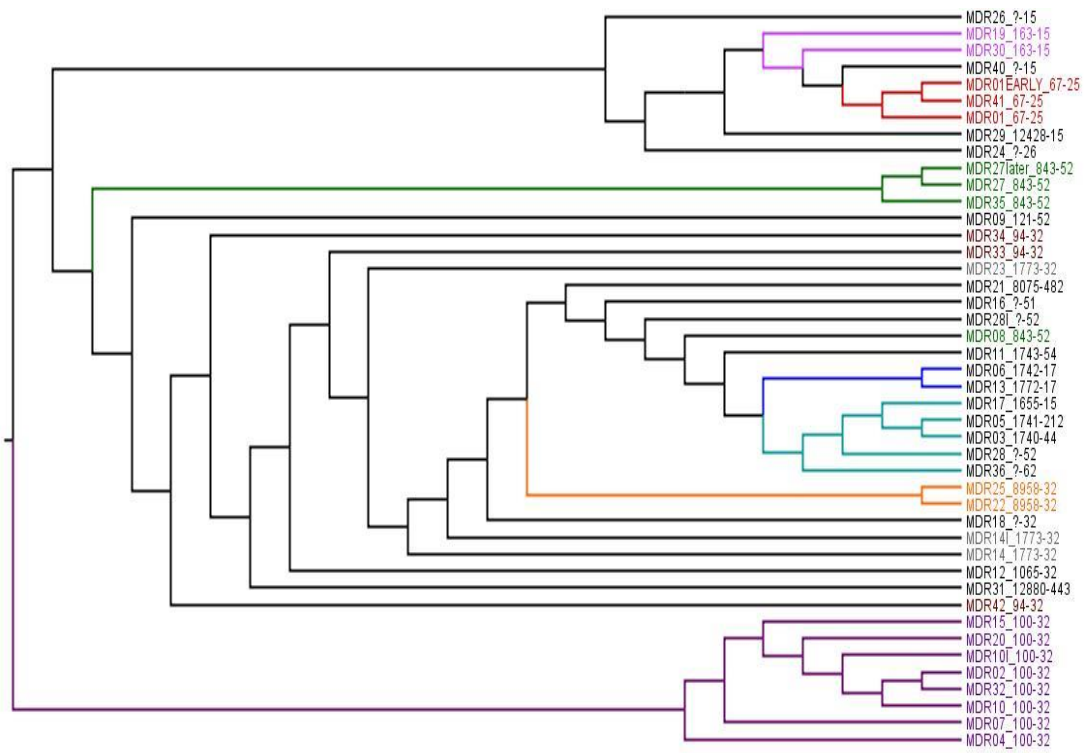


Figure 39. Polar radial WGS Maximum likelihood phylogeny constructed using variant calling files from the whole genomes of MDR/XDR-TB strains isolated in Ireland from 2001-2014.

(PhyResSe software used for phylogenetic reconstruction). MtbC15-9 codes are incorporated into the taxa labels. Clades in light blue represent those where WGS and MtbC15-9 matched. Clades in green represent WGS clusters where MtbC15-9 genotypes differ.



3.0

Figure 40. Transformed cladogram constructed using the whole genome data from MDR/XDR-TB isolates collected in Ireland from 2001-2014.

This is the same information as Figure 36 represented in a different way. Where branches and taxa are the same colour, MtbC 15-9 and WGS correlated. Where branches are coloured but taxa are not, WGS has designated these isolates a clade but the MtbC 15-9 codes have not. Where taxa are coloured but branches are not, WGS has not placed these isolates within a clade as genetically identical even though their MtbC 15-9 codes would suggest they are identical.

IEMDR41 was collected in 2004 from an asylum seeker from the Ivory Coast. IEMDR01 (Irish) was a long-term patient in a healthcare facility in Dublin. On further investigation, it was discovered that IEMDR41 also spent time at this facility. Since the Irish patient presented first, this could be evidence of transmission to the non-Irish born individual. Since no other MIRU-VNTR genotypes matched these two in the IMRL database, there was no other individual that may have been involved in the chain of transmission. Lineage could not be used to indicate direction of transmission since Ghana strains have been seen in Irish TB cases previously.

5.2.3.2.2 Where Whole Genome Sequencing and MIRU-VNTR Genotyping differed

MtbC15-9 843-52 clustered IEMDR08, 27 and 35 together. Figures 39 and 40, constructed with their whole genomes, refute this. There was probable transmission between IEMDR27 and 35 (ie < 5 SNVs apart), but IEMDR08 was separated by 42 SNVs and was, therefore, unrelated according to WGS analysis.

MtbC15-9 94-32 grouped MIRU-VNTR genotypes IEMDR33, 34, and 42 as identical. When WGS analysis was performed, these isolates were found broadly within the same clade, however 38 SNVs separated IEMDR33 and 34, and 40 SNVs separated IEMDR34 and 42, therefore no transmission was evident.

While MtbC15-9 1773-32 isolates IEMDR14 and 23 grouped together broadly on the WGS phylogenetic trees, they were 31 SNVs distant and therefore genomically unrelated, again refuting any transmission between these cases.

MtbC15-9 163-15 isolates IEMDR19 (collected 2011) and 30 (collected 2013) were identical according to MIRU-VNTR genotyping (Figure 38), however, 11 SNVs separated them in the WGS phylogeny. Since this represents a grey-zone result, a link was unlikely between these cases based on a mutation rate of 0.5 SNVs per genome per year. One isolate was collected two years later than the other, which would have enabled 1-2 SNVs at most to have been acquired.

IEMDR06 and 13 clustered together in the same clade on the WGS phylogenetic tree (Figure 39). They were also within the same clade of the MIRU-VNTR genotyping phylogeny (Figure 38) but their MtbC15-9 codes differed (1742-17 and 1772-17 respectively). This was considered a minor discrepancy.

WGS analysis grouped IEMDR03, 05 and 17 together, with no more than 5 SNVs separating the furthest isolates. Their MtbC15-9 codes differed significantly (1740-44 EAI, 1741-212 EAI and 1655-15 Euro-American, respectively). These isolates represent recent transmission according to

their whole genomes. The patients all had addresses in the Dublin area, but no other epidemiological link. One Irish patient showed evidence of progression from susceptible to MDR-TB (IEMDR03, Section 5.2.7), and was associated with non-compliance. The other cases were non-Irish-born. IEMDR05 was the first to present, in January 2005, while Irish-born IEMDR03 presented with MDR-TB in March 2005, and non-Irish-born IEMDR17 did not present until 2010.

5.2.4 Phenotypic Drug Susceptibility Testing

Table 10 describes phenotypic drug susceptibility results for the MDR/XDR-TB cohort. All isolates were resistant to rifampicin and isoniazid except one (IEMDR26, isoniazid susceptible) which was identified as an MDR-TB in 2010 on the basis of rifampicin resistance. DST availability varied depending on tests available at the time of diagnosis. Outside of rifampicin and isoniazid, rifabutin (19/21, 90.5%), streptomycin (28/42, 67%), ethambutol (22/42 intermediate or resistant, 52.4%) and pyrazinamide (18/42, 42%) showed the most resistance. Linezolid and cycloserine were the only drugs that were fully susceptible (0/10 and 0/6 resistant, respectively). Of the aminoglycosides, kanamycin showed the most resistance (9/22, 40.9%) compared to amikacin and capreomycin (both 5/34, 14.7%). Relatively low resistance to fluoroquinolones was seen (ciprofloxacin 3/16, 18.8%, low level moxifloxacin 5/24, 20.8%, high level moxifloxacin 3/24, 12.5% and ofloxacin 3/22, 13.6%). IEXDR1[164], IEMDR27 and IEMDR32 displayed XDR-TB resistance patterns. Others exhibited pre-XDR-TB resistance patterns. IEMDR12 was resistant to fluoroquinolones, but not aminoglycosides. IEMDR16, IEMDR29 and IEMDR33 were resistant to aminoglycosides, but not fluoroquinolones. IEMDR26, while only rifampicin resistant (i.e. not an MDR-TB phenotype), did display resistance to kanamycin, capreomycin, and low-level moxifloxacin. IEMDR15, IEMDR20, IEMDR34, IEMDR40 were resistant to kanamycin, but not any of the other second-line agents. IEMDR09, IEMDR35 and IEMDR36 also showed possible pre-XDR patterns.

5.2.5 Rapid Molecular Line Probe Assays *MTBDRplus* and *MTBDRsl*

Sensitivity and specificity of Hain Genotype *MTBDRplus* and *MTBDRsl* LPAs (n=42 isolates) with respect to the reference standard, phenotypic DST, are shown in Table 11. Overall, sensitivity of 78% (95% CI 70-85) and specificity of 98% (CI 82-100) were seen with this genotypic method.

5.2.5.1 Hain GenoType *MTBDRplus*

DST correlated with *MTBDRplus* LPA in 34 out of 42 cases (81%). Seven discrepancies were found where isoniazid was involved; four cases where isoniazid was phenotypically resistant, but no resistance mutation was found, and three cases where isoniazid was phenotypically resistant but only low-level resistance was detected by the LPA. Three discrepancies were found where

Table 10. MDR/XDR-TB conventional phenotypic drug susceptibility testing (DST) results, 2001-2014.

IEXDR1 was previously published in 2014 [21]. S susceptible, R resistant, INH isoniazid, RIF rifampicin, EMB ethambutol, PZA pyrazinamide, SM streptomycin, AMK amikacin, KAN kanamycin, CAP capreomycin, ETI ethionamide, PAS para-amino-salicylic acid, LZD linezolid, CFZ clofazimine, CYC cycloserine, PRO prothionamide, CIP ciprofloxacin, MXF moxifloxacin, OFX ofloxacin, RFB rifabutin, CLA clarithromycin.

MDR no.	SM	SM	RIF	INH	INH	EMB	PZA	CYC	PAS	CFZ	ETH	PRO	AMK	KAN	CAP	CIP	MXF	MXF	OFX	RFB	CLA	LZD
CRITICAL CONCENTRATION ug/ml	1.0	4.0	1.0	0.1	0.4	5.0	100	40.0	2.0	4.0	5.0	2.5	1.0	2.5	2.5	1.0	0.25	2.0	2.0	0.5	0.5	1.0
IEMDR01	R	R	R	R	R	R	R	S	S/R	R	R	S	R	R	S	S/R				R		
IEMDR03	S	R	R	R	S	S	S		R			S	S	S	S	S	S	S				
IEMDR04	R	R	R	R	R	R	S	S	S	S	R	S										
IEMDR05	R	R	R	R	R	R	S	S	S	S	S											
IEMDR06	R	R	R	R	R	R	R					S								S		
IEMDR07	R	R	R	R	R	R	R					S	S	S	S	S				R	R	
IEMDR08	R	R	R	R	R	R	S					R	S									
IEMDR09	R	R	R	R	R	R	R					S								R	R	
IEMDR10	R	R	R	R	R	R	R					S	S	S	S	S				R	R	
IEMDR11	S	R	R	R	R	R	R					S	S							R	R	
IEMDR12	S	R	R	R	R	R	S		S	S		R	S	S	R	R	R	R	R	R	S	
IEMDR13	R	R	R	R	I	S	S		S	R		S	S							R	S	
IEMDR14	R	R	R	R	R	R	R					R	S							R	S	
IEMDR15	R	R	R	R	I	R	R					S	R	S	S	S	S	S				
IEMDR16	R	R	R	R	S	R	R					R	R	S	S	S	S	S				
IEMDR17	S	R	R	R	R	S	S					S								R	R	
IEMDR18	R	R	R	R	R	R	R		S	S		S	S	S	S	S	S	S		R	R	S
IEMDR19	R	R	R	R	R	R	S					S	S							R	S	
IEMDR20	R	R	R	R	R	R	R					R	S	R	S	S	S	S	R	R	S	
IEMDR21	R	R	R	R	R	R	S					S	S							R		
IEMDR22	S	R	R	R	R	S	S					S	S	S	S	S	S	S				
IEMDR23	R	R	R	R	R	S	R					R	S	S	S	S	S	S				
IEMDR24	S	R	R	R	R	S	S		R			S	S	S	S	S	S	S	R	R	S	
IEMDR25	S	R	R	R	R	S	S					S	S	S	S	S	S	S				
IEMDR26	S	R	R	S	S	S	S					S	R	R	R	R	R	S				
IEMDR27	R	R	R	R	R	R	R	S	R	S		R	S	S	R	R	R	R	R	R		
IEMDR28	R	R	R	R	R	S	S		S	S		R	S						S	S		S
IEMDR29	R	R	R	R	R	S	R		S			S/R	R	R	R	S	S	S	R	R	S	
IEMDR30	R	R	R	R	R	R	S					S	S	S	S	S	S	S				S
IEMDR31	S	S	R	R	R	S/R	S					S	S	S	S	S	S	S	R	R	S	
IEMDR32	R	R	R	R	R	R	R	S	S	R	S	S	R	R	R	R	R	R	R	R		S
IEMDR33	R	R	R	R	R	R	S						R	R	S	S	S	S				
IEMDR34	R	R	R	R	R	S	R					R	S	R	S	S	S	S				
IEMDR35	R	R	R	R	R	R	R					R	S	R	R	R	R	R	R	R		S
IEMDR36	S	R	R	R	R	S	S						S	S	S	R	S	S				
IEMDR37	S	R	R	R	R	S	S															
IEMDR38	S	R	R	R	R	S																
IEMDR39	S	R	R	R	R	S	S															
IEMDR40	S	R	R	R	R	S	S						S	R	S	S	S	S				
IEMDR41	R	R	R	R	R	S	R						S	S	S	S	S	S				
IEMDR42	R	R	R	R	R	R	S						S	S	S	S	S	S				

DRUG	WALKER/KOHL ET AL ALGORITHM									
	PHENOTYPICALLY RESISTANT			PHENOTYPICALLY SUSCEPTIBLE			ALL			
	GENOTYPE		TOTAL	GENOTYPE		TOTAL	SENSITIVITY %	95% CI	SPECIFICITY %	95% CI
	R	S		R	S					
INH	34	4	38	0	1	1	90	75-97	100	2.5-100
RIF	35	4	39	0	0	0	90	76-97	NA	NA
EMB	19	4	23	2	14	16	83	61-99	88	62-99
PZA	1	19	20	0	19	19	5	0-25	100	82-100
SM	24	4	28	1	10	11	86	67-96	91	59-100
FQ	5	2	7	0	27	27	71	29-96	100	87-100
AMK	4	1	5	4	25	29	80	28-100	86	68-96
KAN	3	6	9	2	11	13	33	8-70	85	55-98
CAP	5	1	6	3	25	28	83	36-100	89	72-98
TOTAL	130	45	175	12	132	144	75	67-81	92	86-96
TB PROFILER WEB-TOOL										
	PHENOTYPICALLY RESISTANT			PHENOTYPICALLY SUSCEPTIBLE			ALL			
	GENOTYPE		TOTAL	GENOTYPE		TOTAL	SENSITIVITY %	95% CI	SPECIFICITY %	95% CI
	R	S		R	S					
INH	38	0	38	0	1	1	100	91-100	100	2.5-100
RIF	38	1	39	0	0	0	100	91-100	NA	NA
EMB	23	0	23	4	12	16	85	66-96	100	74-100
PZA	11	9	20	2	17	19	55	32-77	89	67-99
SM	25	3	28	1	10	11	89	72-98	91	59-100
FQ	5	2	7	0	27	27	71	29-96	100	87-100
AMK	4	1	5	4	25	29	50	16-84	96	80-100
KAN	7	2	9	1	12	13	70	35-93	92	62-100
CAP	3	3	6	0	28	28	50	12-88	100	88-100
ETI	1	1	2	1	4	5	50	1-99	80	28-100
PAS	0	4	4	0	8	8	NA	NA	100	63-100
LZD	0	0	0	0	10	10	NA	NA	100	69-100
CFZ	0	1	1	0	15	15	NA	NA	100	78-100
TOTAL	155	27	182	13	169	182	85	79-90	93	88-96
PHYRESSE WEB-TOOL										
	PHENOTYPICALLY RESISTANT			PHENOTYPICALLY SUSCEPTIBLE			ALL			
	GENOTYPE		TOTAL	GENOTYPE		TOTAL	SENSITIVITY %	95% CI	SPECIFICITY %	95% CI
	R	S		R	S					
INH	36	2	38	0	1	1	95	82-99	100	2.5-100
RIF	39	0	39	0	0	0	100	91-100	NA	NA
EMB	22	1	23	3	13	16	96	78-100	81	54-96
PZA	12	8	20	1	18	19	60	36-91	95	74-100
SM	24	4	28	1	10	11	89	71-98	92	62-100
FQ	5	2	7	0	27	27	71	29-96	100	87-100
AMK	3	2	5	0	29	29	60	15-95	100	88-100
KAN	5	4	9	0	13	13	56	21-86	100	75-100
TOTAL	146	23	169	5	111	116	86	80-91	96	90-99
RESEQTB RESISTANCE MUTATION CATALOGUE										
	PHENOTYPICALLY RESISTANT			PHENOTYPICALLY SUSCEPTIBLE			ALL			
	GENOTYPE		TOTAL	GENOTYPE		TOTAL	SENSITIVITY %	95% CI	SPECIFICITY %	95% CI
	R	S		R	S					
INH	34	4	38	0	1	1	90	75-97	100	2.5-100
RIF	38	1	39	0	0	0	97	87-100	NA	NA
EMB	19	4	23	2	14	16	83	61-95	100	77-100
SM	20	8	28	0	11	11	71	51-87	100	72-100
FQ	5	2	7	0	27	27	71	29-96	100	87-100
AMK	3	2	5	0	29	29	60	15-95	100	88-100
KAN	2	7	9	0	13	13	22	2.8-60	100	75-100
CAP	3	3	6	0	28	28	50	12-88	100	87-100
ETI	0	2	2	0	5	5	0	0-84	100	48-100
TOTAL	124	33	157	2	128	130	79	72-85	99	95-100
HAIN GENOTYPE LINE PROBE ASSAYS MTBDRPLUS AND MTBDRSL										
	PHENOTYPICALLY RESISTANT			PHENOTYPICALLY SUSCEPTIBLE			ALL			
	GENOTYPE		TOTAL	GENOTYPE		TOTAL	SENSITIVITY %	95% CI	SPECIFICITY %	95% CI
	R	S		R	S					
INH	34	7	41	0	1	1	83	68-93	100	2.5-100
RIF	39	3	42	0	0	0	93	81-99	NA	NA
EMB	16	7	23	2	14	16	70	47-87	87.5	62-99
FQ	5	2	7	0	27	27	83	36-100	100	88-100
AG	3	4	7	0	28	28	38	9-76	100	88-100
KAN (L)	4	5	9	0	13	13	44	14-79	100	77-100
TOTAL	101	28	129	2	83	85	78	70-85	98	92-100

Table 11. Sensitivity and specificity of WGS Methods for Drug Resistance Prediction in MTBC (genotypic) compared to phenotypic DST for various anti-tuberculous drugs

Included for comparison were Walker/Kohl *et al* algorithm, ReseqTB resistance mutation catalogue, PhyResSe and TB Profiler web-tools. Although numbers were low, and therefore confidence intervals wide, it was seen as a good comparison exercise to record sensitivity and specificity for these methods.

rifampicin was phenotypically resistant but no drug-resistance-associated mutations were found using the assay.

5.2.5.1.1 Discrepancies between MTBDRplus and phenotypic DST, and how they may be resolved with WGS analysis

Isoniazid

IEMDR03 was phenotypically resistant to both isoniazid and rifampicin, however MTBDRplus only detected S531L (Ser450Leu) in *rpoB* while *inhA* and *katG* displayed a wildtype banding pattern. This patient represents an example of progression from susceptible to MDR-TB over the course of treatment and is dealt with in more detail in Section 5.2.7 below.

While IEMDR14 harboured MTBDRplus mutations for *inhA* (C-15T) and *rpoB* (S450L, or S531L *E.coli* numbering, plus possible mutations in codon 513-519), it was not representative of a true MDR-TB isolate since no *katG* mutations were apparent. High-level isoniazid resistance could possibly be due to an SNV found in the *fabG1* promoter region (T-15C) during WGS analysis, which is outside the scope of MTBDRplus.

For IEMDR17, where isoniazid phenotypic resistance was not matched by the discovery of mutations in *katG* or *inhA*, resistance could possibly be due to *ahpC* promoter region mutation (C-52T) and/or *katG* (Gln295Pro) which are both outside the scope of MTBDRplus, but found using WGS.

Only low-level isoniazid resistance was found by the LPA in IEMDR23, however the isolate displayed phenotypic resistance. This could possibly be explained by the discovery of *fabG1* promoter region mutation (T-15C) by WGS analysis, which is, again, outside the scope of MTBDRplus.

Rifampicin

IEMDR06 was phenotypically resistant to both rifampicin and isoniazid but only a *katG* mutation was found with MTBDRplus. Rifampicin resistance could possibly be due to *rpoB* (I491F, or I572F with *E.coli* numbering), found using WGS, which is outside the scope of MTBDRplus.

5.2.5.2 Hain GenoType MTBDRsl

Version 1 and 2 correlated well with each other (100% agreement for *gyrA*, *gyrB* and *rrs*). When evaluating this assay, phenotypic DST results were not available on every isolate; fluoroquinolones (n=5 not available) and aminoglycosides (n=17 kanamycin not available, n = 5 amikacin/capreomycin not available). *In vitro* and *in silico* results were analysed.

5.2.5.2.1 Discrepancies between MTBDRsl and phenotypic DST, and how they may be resolved with WGS analysis

Ethambutol

Taking into account both *in vitro* and *in silico* results, when ethambutol resistance (*embB* gene) was analysed, MTBDRsl correlated with DST in 31/39 (79.5%) cases. There were eight discrepancies; six where the phenotypic DST was resistant but no mutations were found by the LPA, and two where the DST was susceptible but a mutation was detected. For one isolate (IEMDR12), *embB* D354A (D354A) was detected, which may explain the phenotypic DST result. In another isolate (IEMDR27), WGS analysis found Q497R mutation in *embB* which could have caused ethambutol resistance. In the remaining resistant DST cases (n=4), no other relevant mutations were detected with WGS analysis that might explain the ethambutol resistance.

In the cases where DST was susceptible but a mutation was present, the mutations were *embB* M306V and M306I, both of which have also been seen in ethambutol susceptible isolates that were resistant to other drugs [189, 190].

Fluoroquinolones

When fluoroquinolone resistance was analysed (*gyrA* and *gyrB* genes), MTBDRsl correlated with DST in 32/34 (94.1%) cases. One discrepancy (IEXDR1) was most probably due to the DST available at the time (2005). Ciprofloxacin was found to be susceptible, and resistant, in two separate laboratories. Other fluoroquinolones were not tested. Both *gyrA* and *gyrB* correlated well for the other fluoroquinolones (ie moxifloxacin and ofloxacin).

The second discrepancy occurred in IEMDR36, where phenotypic moxifloxacin DST displayed low-level resistance, although no mutations could be detected in *gyrA* or *gyrB*. WGS analysis failed to detect any other mutation that may explain this resistance.

Aminoglycosides

When aminoglycoside resistance was analysed (*rrs* gene), MTBDRsl correlated with phenotypic DST in 24/34 (70.5%) of cases. Among ten discrepancies, there was one case where a mutation was detected but capreomycin was susceptible (although amikacin was resistant).

In nine cases, no mutation was detected in *rrs* by the LPA, although phenotypic aminoglycoside resistance was seen (kanamycin-only resistant, n=5, amikacin and kanamycin resistant, n=2, and capreomycin-only resistant, n=2). Other genes may have been involved in aminoglycoside resistance in these instances.

Low Level Kanamycin

When low-level kanamycin resistance (*eis* gene) was analysed independently, MTBDR_{sl} correlated with phenotypic DST in 17/22 (77.3%) of cases. In all of the discrepant cases, phenotypic resistance was seen, while no mutation was detected by the LPA. Three of those cases were kanamycin-only resistant, i.e. aminoglycoside resistant but not explained by mutations in either *rrs* or *eis*. WGS analysis failed to detect other plausible mutations to explain this phenotypic resistance. In two isolates (IEMDR29 and IEMDR32), *rrs* mutation A1401G predicted phenotypic resistance to capreomycin, amikacin and kanamycin, therefore an *eis* mutation may not have been required to cause kanamycin resistance.

5.2.6 Drug Resistance Prediction using Whole Genome Sequencing

WGS resistance prediction was performed manually with Geneious R9 software and resistance mutation catalogues from Walker/Kohl *et al* and ReseqTB data sharing platform, and online using PhyResSe and TB Profiler WGS analysis web-tools and their respective resistance mutation catalogues (25 genes and their promoter regions – 100bp upstream of each gene). Results were compared to phenotypic DST. Sensitivity and specificity results for these, and Hain LPA comparison with phenotypic DST, are described in Table 11. A comparison of the above web-tools was also undertaken. A further 43 genes, that were found in the literature to have an association with drug resistance in MTBC, were also analysed manually (Tables 12-24).

Fastq files from the MDR/XDR-TB isolates were mapped to the H37Rv reference genome (AL123456, NC_000962.3) with a mean coverage of 157, a mapping quality of 59 according to Qualimap, and with a mean mapped read percentage of 95.9% [44].

5.2.6.1 Drug-resistance-associated mutations present in MDR/XDR-TB Isolates collected in Ireland from 2001 to 2014

Drug-resistance-associated mutations found within twenty-five candidate genes (and their promoter regions, i.e. 100bp upstream of each gene) in the MDR/XDR-TB cohort are shown in Tables 12 and 13.

The *katG* S315T mutation was by far the most common SNV found which was associated with isoniazid resistance (n=34/37, 92%), followed by *fabG1* C-15T (n=9/37, 24%) which was seen in conjunction with *katG* S315T in seven cases.

STUDY NO.	ISONIAZID	RIFAMPICIN	ETHAMBUTOL
Rv	Rv2428, 1483, 1484, 1908c, 1854c, 2245	Rv0667, 0668	Rv3794, 3795, 3793, 1267c, 0342, 0343, 3264c, 3266c
GENE	<i>ahpC</i> , <i>fabG1</i> (<i>mabA</i>), <i>inhA</i> , <i>katG</i> , <i>ndh</i> , <i>kasA</i>	<i>rpoB</i> , <i>rpoC</i>	<i>embA</i> , <i>embB</i> , <i>embC</i> , <i>embR</i> , <i>iniA</i> , <i>iniC</i> , <i>manB</i> , <i>rmlD</i>
IEMDR01	<i>katG S315T</i> , <i>fabG1 T-8C</i>	<i>rpoB H445Y</i> (H526Y <i>E.coli</i>)	<i>embB M306V</i> , D328Y
IEMDR01 EARLY	<i>katG S315T</i> , <i>fabG1 T-8C</i>	<i>rpoB H445Y</i> (H526Y <i>E.coli</i>)	<i>embB M306V</i> , M306I
IEXDR1	<i>katG S315T</i>	<i>rpoB H445Y</i> (H526Y <i>E.coli</i>)	<i>embB M306V</i>
IEMDR03	–	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	–
IEMDR04	<i>katG S315T</i>	<i>rpoB H445Y</i> (H526Y <i>E.coli</i>)	<i>embA C-8T</i>
IEMDR05	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB D328Y</i>
IEMDR06	<i>katG S315T</i>	<i>rpoB I491F</i> (I572F <i>E.coli</i>)	<i>embB M306I</i>
IEMDR07	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB M306V</i>
IEMDR08	<i>katG S315T</i> , <i>fabG1 C-15T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>), <i>rpoC F452S</i>	–
IEMDR09	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB M306I</i> , D328Y
IEMDR10	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>), <i>rpoC F452S</i>	<i>embB M306V</i>
IEMDR11	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB M306I</i> , Q497R
IEMDR12	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB D354A</i>
IEMDR13	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB M306I</i>
IEMDR14	<i>fabG1 C-15T</i> , <i>inhA S94A</i>	<i>rpoB S450L</i> , D435Y (S531L, D516Y <i>E.coli</i>)	<i>embB M306V</i>
IEMDR15	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>), <i>rpoC F452S</i>	<i>embB M306V</i>
IEMDR16	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>), <i>rpoC F452S</i> , G332R	<i>embB M306I</i>
IEMDR17	<i>ahpC C-52T</i> , <i>katG Q295P</i>	<i>rpoB H445D</i> (H526D <i>E.coli</i>)	–
IEMDR18	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB M306I</i> , <i>embA C-12T</i>
IEMDR19	<i>katG S315T</i> , <i>fabG1 C-15T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB S297A</i>
IEMDR20	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB M306I</i>
IEMDR21	<i>katG S315T</i>	<i>rpoB H445Y</i> (H526Y <i>E.coli</i>)	<i>embB M306V</i>
IEMDR22	<i>katG S315T</i>	<i>rpoB D435Y</i> (D516Y <i>E.coli</i>)	–
IEMDR23	<i>fabG1 C-15T</i> , <i>inhA S94A</i>	<i>rpoB S450L</i> , D435Y (S531L, D516Y <i>E.coli</i>)	<i>embB M306V</i>
IEMDR24	<i>katG S315T</i> , P241P	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	–
IEMDR25	<i>katG S315T</i>	<i>rpoB D435Y</i> (D516Y <i>E.coli</i>)	–
IEMDR26	–	<i>rpoB H445Y</i> (H526Y <i>E.coli</i>)	–
IEMDR27	<i>katG S315T</i> , <i>fabG1 C-15T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB Q497R</i>
IEMDR27 LATER	<i>katG S315T</i> , <i>fabG1 C-15T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB Q497R</i>
IEMDR28	<i>katG S315T</i> , <i>fabG1 C-15T</i>	<i>rpoB H445L</i> (H526L <i>E.coli</i>)	–
IEMDR29	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB Y319S</i>
IEMDR30	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB M306V</i>
IEMDR31	<i>katG S315T</i> , <i>ahpC G-48A</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB M306V</i>
IEMDR32	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB G406A</i>
IEMDR33	<i>katG S315T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB M306V</i>
IEMDR34	<i>katG S315T</i> , <i>fabG1 C-15T</i>	<i>rpoB D435V</i> , D435A (D516V, D516A <i>E.coli</i>)	–
IEMDR35	<i>katG S315T</i> , <i>fabG1 C-15T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embB Q497R</i>
IEMDR36	<i>katG S315T</i>	<i>rpoB D435V</i> (D516V <i>E.coli</i>)	–
IEMDR40	<i>katG S315T</i>	<i>rpoB H445R</i> (H526R <i>E.coli</i>)	–
IEMDR41	<i>katG S315T</i> , <i>fabG1 T-8C</i>	<i>rpoB H445Y</i> (H526Y <i>E.coli</i>)	–
IEMDR42	<i>katG S315T</i> , <i>fabG1 C-15T</i>	<i>rpoB S450L</i> (S531L <i>E.coli</i>)	<i>embA C-12T</i> , <i>embB Y334H</i>

Table 12. Single Nucleotide Variations (SNVs) that have been associated with resistance to isoniazid, rifampicin and ethambutol, found in MDR/XDR-TB isolates collected in Ireland from 2001-2014.

Twenty five genes were analysed using different resistance mutation catalogues as well as manual detection.

STUDY NO.	PYRAZINAMIDE	STREPTOMYCIN	FLUOROQUINOLONES	AMINOGLYCOSIDES
Rv	Rv1630, 2043c	Rv1630, 3919, 1694, Rvnr01/MTB000019	Rv0006,0005	Rv3919, Rvnr01/MTB000019, Rv1694, 2416c,
GENE	<i>rpsA, pncA</i>	<i>rpsL, gidB, tlyA, rrs</i>	<i>gyrA, gyrB</i>	<i>gidB, rrs, tlyA, eis</i>
IEMDR01	<i>pncA</i> C14Stop	-	-	-
IEMDR01 EARLY	<i>pncA</i> C14Stop	-	-	-
IEXDR1	<i>pncA</i> G132C	<i>rpsL</i> K43R, <i>rrs</i> A1401G	<i>gyrA</i> D94A	<i>rrs</i> A1401G
IEMDR03	-	-	-	-
IEMDR04	<i>pncA</i> D12N	<i>rpsL</i> K43R	<i>gyrB</i> N499T	<i>eis</i> C-10T
IEMDR05	<i>pncA</i> C14G	<i>rrs</i> A514C	-	<i>rrs</i> A514C
IEMDR06	-	<i>rpsL</i> K43R	-	-
IEMDR07	<i>pncA</i> D136N	<i>rpsL</i> K43R	-	<i>eis</i> C-10T
IEMDR08	-	-	-	-
IEMDR09	-	<i>rpsL</i> K43R	<i>gyrA</i> A90V	<i>eis</i> C-10T
IEMDR10	-	<i>rpsL</i> K43R	-	<i>eis</i> C-10T
IEMDR11	-	-	-	-
IEMDR12	<i>pncA</i> A-12G	<i>rrs</i> C517T	<i>gyrA</i> S91P	<i>rrs</i> C517T
IEMDR13	-	<i>rpsL</i> K43R	-	-
IEMDR14	<i>pncA</i> D12E	<i>rpsL</i> K43R	-	-
IEMDR15	-	<i>rpsL</i> K43R	-	<i>eis</i> C-10T
IEMDR16	-	<i>rpsL</i> K43R	-	<i>eis</i> C-10T
IEMDR17	-	-	-	-
IEMDR18	<i>pncA</i> Q10P	<i>rpsL</i> K43R	-	-
IEMDR19	-	<i>rpsL</i> K43R	-	<i>eis</i> C-10T
IEMDR20	<i>pncA</i> M175V	<i>rpsL</i> K43R	-	-
IEMDR21	-	<i>rpsL</i> K88M	-	-
IEMDR22	-	-	-	-
IEMDR23	<i>pncA</i> D12E	<i>rpsL</i> K43R	-	-
IEMDR24	-	-	-	-
IEMDR25	-	-	-	-
IEMDR26	-	-	-	-
IEMDR27	<i>pncA</i> H51Y	<i>rrs</i> A514C	-	<i>rrs</i> A514C
IEMDR27 LATER	<i>pncA</i> H51Y	<i>rrs</i> A514C	<i>gyrA</i> D94Y	<i>rrs</i> A514C
IEMDR28	-	<i>rrs</i> A514C	-	<i>rrs</i> A514C
IEMDR29	<i>pncA</i> G108R	<i>rpsL</i> K43R, <i>rrs</i> A1401G	-	<i>rrs</i> A1401G
IEMDR30	-	<i>rpsL</i> K88R	-	-
IEMDR31	-	-	-	-
IEMDR32	<i>pncA</i> I31S	<i>rpsL</i> K43R, <i>rrs</i> A1401G	<i>gyrA</i> A90V	<i>rrs</i> A1401G
IEMDR33	-	<i>rpsL</i> K43R, <i>rrs</i> C517T	-	<i>rrs</i> C517T, <i>eis</i> G-14A
IEMDR34	<i>pncA</i> Y103H	<i>rpsL</i> K43R	-	<i>eis</i> C-10T
IEMDR35	<i>pncA</i> H51Y	<i>rrs</i> A514C	-	<i>rrs</i> A514C
IEMDR36	-	-	-	-
IEMDR40	-	-	-	-
IEMDR41	-	-	-	-
IEMDR42	<i>pncA</i> C14Stop	<i>rpsL</i> K88R	-	-

Table 13. Single Nucleotide Variations (SNVs) that have been associated with resistance to pyrazinamide, streptomycin, fluoroquinolones and aminoglycosides, found in MDR/XDR-TB isolates collected in Ireland from 2001-2014.

Twenty five genes were analysed using different resistance mutation catalogues as well as manual detection.

GENE	SNV	NO. OF ISOLATES WITH SNV	LITERATURE SEARCH
<i>accD6/Rv2247</i>	D229G	14	[201]
<i>cycA/Rv1704c</i>	R93L	39	[192]
	D238N	2	not found
	T13P	5	not found
	H16P	7	not found
	R477G	1	not found
	D20A	4	not found
<i>ddlA/Rv2981c</i>	T365A	37	[192]
	W210G	1	not found
<i>efpA/Rv2846c</i>	I73T	1	[196]
<i>ethA/Rv3854c</i>	Y211C	1	not found
	C403R	1	not found
	P334A	1	not found
	H281P	1	not found
	P284T	1	not found
<i>fabD/Rv2243</i>	T115A	8	not found
	S275N	2	[196]
<i>fadE24/Rv3139</i>	I430L	2	not found
<i>fbpC/Rv0129c</i>	G158S	2	Phylogenetic
<i>iniB/Rv0341</i>	G113D	2	not found
<i>oxyR'/Rv2427a</i>	L13F	3	not found
Rv1592c	F417L	2	not found
	I322V	1	phylogenetic
Rv2242	M323T	2	not found
Rv3124/moaR1	R145S	2	not found
	A116S	1	not found
	P54S	2	[199]
Rv3125c/PPE49	L327W	1	not found
<i>thyA/Rv2764c</i>	Q97R	1	[200]
	T202A	9	phylogenetic
Rv3728	T289A	2	not found
	D563N	2	not found
	V488I	1	not found
	H921N	1	not found
Rv3790/dprE1	A356T	1	not found
Rv0407/fgd1	K270M	1	phylogenetic
Rv1173/fbiC	W678G	3	not found
Rv3361c/mfpA	L178V	1	not found
Rv2911/dacB2/dacB	A286T	1	not found
	R2Q	7	not found
Rv0486/mshA	A187V	13	phylogenetic
	N111S	1	phylogenetic

Table 14. Single Nucleotide Variations (SNVs) found in the MDR/XDR-TB cohort, collected 2001-14, when 43 further genes were analysed.

No mutations were found within 23 genes (Rv3423, 1305, 2566c, 3601c, 0701, Rvnr02/MTB000020, Rv1772, 3126, 0206c, 1988, 3262, 3547, 0016c, 2068c, 0116c, 2518c, 0664, 1909c, 0340, 1258, 3197a, 2245, and 3855). Mutations found are included in the table above. If references were found for the mutation, they were included. If no reference was found, the table records 'not found'. If the mutation was found, but related to lineage, it was designated 'phylogenetic'.

Table 15. Summary of comparison between Walker/Kohl *et al* WGS analysis algorithm (genotypic) and conventional isoniazid, rifampicin, ethambutol and pyrazinamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014

Discrepancies are highlighted in purple. R – resistant, S= sensitive, L – low, MUT – mutation, see list of abbreviations for amino acid and nucleotide abbreviations. Promoter region refers to -100bp upstream of the gene.

	ISONIAZID			RIFAMPICIN			ETHAMBUTOL			PYRAZINAMIDE		
GENES + PROMOTER REGIONS	<i>ahpC, fabG1(mabA), inhA, katG</i>			<i>rpoB</i>			<i>embA, embB, embC, embR, iniA, iniC, manB, rmlD</i>			<i>pncA, rpsA</i>		
Rv	Rv2428, Rv1483, Rv1484, Rv1908c			Rv0667			Rv3794, Rv3795, Rv3793, Rv1267c, Rv0342, Rv0343, Rv3264c, Rv3266c			Rv2043c, Rv1630		
STUDY NO.	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT
IEMDR01	R	R	<i>katG</i> S315T, <i>fabG1</i> T-8C	R	R	<i>rpoB</i> H445Y	R	R	<i>embB</i> M306V, M306I	R	S	-
IEXDR1	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> H445Y	R	R	<i>embB</i> M306V	R	S	-
IEMDR03	R	S	-	R	R	<i>rpoB</i> S450L	S	S	-	S	S	-
IEMDR04	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> H445Y	S	R	<i>embA</i> C-8T	S	S	-
IEMDR05	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	S	-	S	S	-
IEMDR06	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> I491F	R	R	<i>embB</i> M306I	R	S	-
IEMDR07	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306V	R	S	-
IEMDR08	R	R	<i>katG</i> S315T, <i>fabG1</i> C-15T	R	R	<i>rpoB</i> S450L	S	S	-	S	S	-
IEMDR09	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	S	<i>embB</i> M306I	R	S	-
IEMDR10	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306V	R	S	-
IEMDR11	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306I, <i>embB</i> Q497R	R	S	-
IEMDR12	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> D354A	R	S	-
IEMDR13	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306I	S	S	-
IEMDR14	R	R (L)	<i>fabG1</i> C-15T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306V	R	S	-
IEMDR15	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306V	R	S	-
IEMDR16	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	S	R	<i>embB</i> M306I	R	S	-
IEMDR17	R	S	-	R	R	<i>rpoB</i> H445D	S	S	-	S	S	-
IEMDR18	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306I	R	R	<i>pncA</i> Q10P
IEMDR19	R	R	<i>katG</i> S315T, <i>fabG1</i> C-15T	R	R	<i>rpoB</i> S450L	R	S	-	S	S	-
IEMDR20	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306I	R	S	-
IEMDR21	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> H445Y	R	R	<i>embB</i> M306V	S	S	-
IEMDR22	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> D435Y	S	S	-	S	S	-
IEMDR23	R	R (L)	<i>fabG1</i> C-15T	R	R	<i>rpoB</i> S450L	S	R	<i>embB</i> M306V	R	S	-
IEMDR24	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	S	S	-	S	S	-
IEMDR25	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> D435Y	S	S	-	S	S	-
IEMDR26	S	S	-	R	R	<i>rpoB</i> H445Y	S	S	-	S	S	-
IEMDR27	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> Q497R	R	S	-
IEMDR28	R	R	<i>katG</i> S315T, <i>fabG1</i> C-15T	R	S	-	S	S	-	S	S	-
IEMDR29	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	S	S	-	R	S	-
IEMDR30	R	R	<i>katG</i> S315T, <i>ahpC</i> G-48A	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306V	S	S	-
IEMDR31	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	S/R	R	<i>embB</i> M306V	S	S	-
IEMDR32	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> G406A	R	S	-
IEMDR33	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> M306V	S	S	-
IEMDR34	R	R	<i>katG</i> S315T, <i>fabG1</i> C-15T	R	R	<i>rpoB</i> D435V, D435A	S	S	-	R	S	-
IEMDR35	R	R	<i>katG</i> S315T, <i>fabG1</i> C-15T	R	R	<i>rpoB</i> S450L	R	R	<i>embB</i> Q497R	R	S	-
IEMDR36	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> D435V	S	S	-	S	S	-
IEMDR40	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> H445R	S	S	-	S	S	-
IEMDR41	R	R	<i>katG</i> S315T, <i>fabG1</i> C-15T	R	R	<i>rpoB</i> H445Y	S	S	-	R	S	-
IEMDR42	R	R	<i>katG</i> S315T	R	R	<i>rpoB</i> S450L	R	S	-	S	S	-

Table 16. Summary of comparison between Walker/Kohl *et al* WGS analysis algorithm (genotypic) and conventional streptomycin, fluoroquinolone and aminoglycoside DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014

Discrepancies are highlighted in purple. R – resistant, S= sensitive, L – low, MUT – mutation, see list of abbreviations for further abbreviations. Promoter region refers to -100bp upstream of the gene.

	STREPTOMYCIN			FLUOROQUINOLONES					AMIKACIN			KANAMYCIN			CAPREOMYCIN		
GENES + PROMOTER REGIONS	<i>rpsL, gidB, tlyA, rrs</i>			<i>gyrA, gyrB</i>					<i>gidB, rrs, tlyA</i>			<i>eis, gidB, rrs, tlyA</i>			<i>gidB, rrs, tlyA</i>		
Rv	Rv1630, Rv3919, Rv1694, Rvnr01/MTB000019			Rv0006, Rv0005					Rv3919, Rvnr01/MTB000019, Rv1694			Rv2416c, Rv3919, Rvnr01/MTB000019, Rv1694			Rv3919, Rvnr01/MTB000019, Rv1694		
STUDY NO.	DST	WGS	MUT	OFX DST	MOX DST	CIP DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT
IEMDR01	R	S					S			S			S			S	
IEMDR1	R	R	<i>rpsL K43R, rrs A1401G</i>			S/R		<i>gyrA D94A</i>	R	R	<i>rrs A1401G</i>		R	<i>rrs A1401G</i>	R	R	<i>rrs A1401G</i>
IEMDR03	S	S		S	S		S		S	S		S	S		S	S	
IEMDR04	R	R	<i>rpsL K43R</i>				S			S			S			S	
IEMDR05	R	R	<i>rrs A514C</i>				S			R	<i>rrs A514C</i>		R	<i>rrs A514C</i>		R	<i>rrs A514C</i>
IEMDR06	R	R	<i>rpsL K43R</i>				S			S			S			S	
IEMDR07	R	R	<i>rpsL K43R</i>			S	S		S	S			S		S	S	
IEMDR08	R	S					S		S	S			S		S	S	
IEMDR09	R	R	<i>rpsL K43R</i>			R	R	<i>gyrA A90V</i>		S			S		S	S	
IEMDR10	R	R	<i>rpsL K43R</i>			S	S		S	S			S		S	S	
IEMDR11	S	S				S	S		S	S			S		S	S	
IEMDR12	S	R	<i>rrs C517T</i>	R	R	R	R	<i>gyrA S91P</i>	S	R	<i>rrs C517T</i>	S	R	<i>rrs C517T</i>	S	R	<i>rrs C517T</i>
IEMDR13	R	R	<i>rpsL K43R</i>			S	S		S	S			S		S	S	
IEMDR14	R	R	<i>rpsL K43R</i>			S	S		S	S			S		S	S	
IEMDR15	R	R	<i>rpsL K43R</i>	S	S		S		S	S		R	S		S	S	
IEMDR16	R	R	<i>rpsL K43R</i>	S	S		S		R	S		R	S		S	S	
IEMDR17	S	S				S	S		S	S			S		S	S	
IEMDR18	R	R	<i>rpsL K43R</i>		S	S	S		S	S		S	S		S	S	
IEMDR19	R	R	<i>rpsL K88R</i>			S	S		S	S			S		S	S	
IEMDR20	R	R	<i>rpsL K43R</i>	S	S	S	S		S	S		R	S		S	S	
IEMDR21	R	S			S	S	S		S	S			S		S	S	
IEMDR22	S	S		S	S		S		S	S		S	S		S	S	
IEMDR23	R	R	<i>rpsL K43R</i>	S	S		S		S	S		S	S		S	S	
IEMDR24	S	S		S	S	S	S		S	S		S	S		S	S	
IEMDR25	S	S		S	S		S		S	S		S	S		S	S	
IEMDR26	S	S		S	R (L)		S		S	S		R	S		R	S	
IEMDR27	R	R	<i>rrs A514C</i>	R	R		R	<i>gyrA D94Y</i>	S	R	<i>rrs A514C</i>	S	R	<i>rrs A514C</i>	R	R	<i>rrs A514C</i>
IEMDR28	R	R	<i>rrs A514C</i>			S	S		S	R	<i>rrs A514C</i>		R	<i>rrs A514C</i>	S	R	<i>rrs A514C</i>
IEMDR29	R	R	<i>rpsL K43R, rrs A1401G</i>	S	S	S	S		R	R	<i>rrs A1401G</i>	R	R	<i>rrs A1401G</i>	R	R	<i>rrs A1401G</i>
IEMDR30	R	R	<i>rpsL K88R</i>	S	S		S		S	S		S	S		S	S	
IEMDR31	S	S		S	S		S		S	S		S	S		S	S	
IEMDR32	R	R	<i>rpsL K43R, rrs A1401G</i>	R	R		R	<i>gyrA A90V</i>	R	R	<i>rrs A1401G</i>	R	R	<i>rrs A1401G</i>	R	R	<i>rrs A1401G</i>
IEMDR33	R	R	<i>rpsL K43R, rrs C517T</i>	S	S		S		R	R	<i>rrs C517T</i>	R	R	<i>rrs C517T</i>	S	R	<i>rrs C517T</i>
IEMDR34	R	R	<i>rpsL K43R</i>	S	S		S		S	S		R	S		S	S	
IEMDR35	R	R	<i>rrs A514C</i>	S	S		S		S	R	<i>rrs A514C</i>		R	<i>rrs A514C</i>	R	R	<i>rrs A514C</i>
IEMDR36	S	S		S	R (L)		S		S	S		S	S		S	S	
IEMDR40	S	S		S	S		S		S	S		R	S		S	S	
IEMDR41	R	S		S	S		S		S	S		S	S		S	S	
IEMDR42	R	R	<i>rpsL K88R</i>	S	S		S		S	S		S	S		S	S	

Table 17. Summary of correlation between ReseqTB resistance mutation catalogue (genotypic) compared to isoniazid, rifampicin, ethambutol and pyrazinamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014

Discrepancies are highlighted in purple. R – resistant, S= sensitive, L – low, MUT – mutation, see list of abbreviations for further abbreviations. Promoter region refers to -100bp upstream of the gene.

GENES + PROMOTERS		ISONIAZID			RIFAMPICIN			ETHAMBUTOL			STREPTOMYCIN		
		<i>inhA, katG</i>			<i>rpoB</i>			<i>embB</i>			<i>rpsL</i>		
Rv		Rv484, Rv1908c			Rv0667			Rv3795			Rv0682		
STUDY NO.	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT	
IEMDR01	R	R	<i>fabG1 T-8C, katG S315T</i>	R	R	<i>rpoB H445Y</i>	R	R	<i>embB M306V, M306I</i>	R	S		
IEXDR1	R	R	<i>KatG S315T</i>	R	R	<i>rpoB H445Y</i>	R	R	<i>embB M306V</i>	R	R	<i>rpsL K43R</i>	
IEMDR03	R	S		R	R	<i>rpoB S450L</i>	R	S		S	S		
IEMDR04	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445Y</i>	R	R		R	R	<i>rpsL K43R</i>	
IEMDR05	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	S		R	S		
IEMDR06	R	R	<i>katG S315T</i>	R	R	<i>rpoB I491F</i>	R	R	<i>embB M306I</i>	R	R	<i>rpsL K43R</i>	
IEMDR07	R	R	<i>KatG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	R	R	<i>rpsL K43R</i>	
IEMDR08	R	R	<i>fabG1 C-15T, katG S315T</i>	R	R	<i>rpoB S450L</i>	R	S		R	S		
IEMDR09	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I</i>	R	R	<i>rpsL K43R</i>	
IEMDR10	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	R	R	<i>rpsL K43R</i>	
IEMDR11	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I, Q497R</i>	S	S		
IEMDR12	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	S		S	S		
IEMDR13	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I</i>	R	R	<i>rpsL K43R</i>	
IEMDR14	R	(L)	<i>inhA S94A, fabG1 C-15T</i>	R	R	<i>rpoB S450L, D435Y</i>	R	R	<i>embB M306V</i>	R	R	<i>rpsL K43R</i>	
IEMDR15	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	R	R	<i>rpsL K43R</i>	
IEMDR16	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I</i>	R	R	<i>rpsL K43R</i>	
IEMDR17	R	S		R	R	<i>rpoB H445D</i>	S	S		S	S		
IEMDR18	R	R	<i>KatG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I</i>	R	R	<i>rpsL K43R</i>	
IEMDR19	R	R	<i>fabG1 C-15T, katG S315T</i>	R	R	<i>rpoB S450L</i>	R	S		R	R	<i>rpsL K88R</i>	
IEMDR20	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I</i>	R	R	<i>rpsL K43R</i>	
IEMDR21	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445Y</i>	R	R	<i>embB M306V</i>	R	S		
IEMDR22	R	R	<i>katG S315T</i>	R	R	<i>rpoB D435Y</i>	S	S		S	S		
IEMDR23	R	(L)	<i>inhA S94A, fabG1 C-15T</i>	R	R	<i>rpoB S450L, D435Y</i>	S	R	<i>embB M306V</i>	R	R	<i>rpsL K43R</i>	
IEMDR24	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S	S		S	S		
IEMDR25	R	R	<i>katG S315T</i>	R	R	<i>rpoB D435Y</i>	S	S		S	S		
IEMDR26	S	S		R	R	<i>rpoB H445Y</i>	S	S		S	S		
IEMDR27	R	R	<i>katG S315T, fabG1 C-15T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB Q497R</i>	R	S		
IEMDR28	R	R	<i>katG S315T, fabG1 C-15T</i>	R	S		S	S		R	S		
IEMDR29	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S	S		R	R	<i>rpsL K43R</i>	
IEMDR30	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	R	R	<i>rpsL K88R</i>	
IEMDR31	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S/R	R	<i>embB M306V</i>	S	S		
IEMDR32	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>EmbB G406A</i>	R	R	<i>rpsL K43R</i>	
IEMDR33	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	R	R	<i>rpsL K43R</i>	
IEMDR34	R	R	<i>katG S315T, fabG1 C-15T</i>	R	R	<i>rpoB D435Y</i>	S	S		R	R	<i>rpsL K43R</i>	
IEMDR35	R	R	<i>katG S315T, fabG1 C-15T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB Q497R</i>	R	S		
IEMDR36	R	R	<i>katG S315T</i>	R	R	<i>rpoB D435Y</i>	S	S		S	S		
IEMDR40	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445R</i>	S	S		S	S		
IEMDR41	R	R	<i>fabG1 T-8C, katG S315T</i>	R	R	<i>rpoB H445Y</i>	S	S		R	S		
IEMDR42	R	R	<i>katG S315T, fabG1 C-15T</i>	R	R	<i>rpoB S450L</i>	R	S		R	R	<i>rpsL K88R</i>	

Table 18. Summary of correlation between ReseqTB resistance mutation catalogue (genotypic) compared to fluoroquinolone, aminoglycoside and ethionamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014

Discrepancies are highlighted in purple. R – resistant, S= sensitive, L – low, MUT – mutation, see list of abbreviations for further abbreviations. Promoter region refers to -100bp upstream of the gene.

	FLUOROQUINOLONES						AMIKACIN			KANAMYCIN			CAPREOMYCIN			ETHIONAMIDE		
GENES + PROMOTERS	<i>gyrA</i>						<i>rrs</i>			<i>rrs</i>			<i>rrs</i>			<i>inhA</i>		
Rv	Rv0006						Rvnr01/ MTB000019			Rvnr01/ MTB000019			Rvnr01/ MTB000019			Rv1484		
STUDY NO.	OFX DST	MXF DST	CIP DST	WGS	MXF MUT	OFX MUT	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT
IEMDR01	-	-	-	S	-	-	-	S	-	-	S	-	-	S	-	R	S	-
IEXDR1	-	-	S/R	R	<i>gyrA</i> <i>D94A</i>	<i>GyrA</i> <i>D94A</i>	R	R	<i>rrs</i> <i>A140I</i> <i>G</i>	-	R	<i>rrs</i> <i>A140I</i> <i>G</i>	R	R	<i>rrs</i> <i>A140I</i> <i>G</i>	S	S	-
IEMDR03	S	S	-	S	-	-	S	S	-	S	S	-	S	S	-	-	S	-
IEMDR04	-	-	-	S	-	-	-	S	-	-	S	-	-	S	-	R	S	-
IEMDR05	-	-	-	S	-	-	-	S	-	-	S	-	-	S	-	S	S	-
IEMDR06	-	-	-	S	-	-	-	S	-	-	S	-	-	S	-	-	S	-
IEMDR07	-	-	S	S	-	-	S	S	-	-	S	-	S	S	-	-	S	-
IEMDR08	-	-	-	S	-	-	S	S	-	-	S	-	S	S	-	-	S	-
IEMDR09	-	-	R	R	<i>gyrA</i> <i>A90V</i>	<i>gyrA</i> <i>A90V</i>	-	S	-	-	S	-	S	S	-	-	S	-
IEMDR10	-	-	S	S	-	-	S	S	-	-	S	-	S	S	-	-	S	-
IEMDR11	-	-	S	S	-	-	S	S	-	-	S	-	S	S	-	-	S	-
IEMDR12	R	R	R	R	<i>gyrA</i> <i>S91P</i>	<i>gyrA</i> <i>S91P</i>	S	S	-	S	S	-	S	S	-	-	S	-
IEMDR13	-	-	S	S	-	-	S	S	-	-	S	-	S	S	-	-	S	-
IEMDR14	-	-	S	S	-	-	S	S	-	-	S	-	S	S	-	R	<i>inhA</i> <i>S94A</i>	-
IEMDR15	S	S	-	S	-	-	S	S	-	R	S	-	S	S	-	-	S	-
IEMDR16	S	S	-	S	-	-	R	S	-	R	S	-	S	S	-	-	S	-
IEMDR17	-	-	S	S	-	-	S	S	-	-	S	-	S	S	-	-	S	-
IEMDR18	-	S	S	S	-	-	S	S	-	S	S	-	S	S	-	-	S	-
IEMDR19	-	-	S	S	-	-	S	S	-	-	S	-	S	S	-	-	S	-
IEMDR20	S	S	S	S	-	-	S	S	-	R	S	-	S	S	-	-	S	-
IEMDR21	-	S	S	S	-	-	S	S	-	-	S	-	S	S	-	-	S	-
IEMDR22	S	S	-	S	-	-	S	S	-	S	S	-	S	S	-	-	S	-
IEMDR23	S	S	-	S	-	-	S	S	-	S	S	-	S	S	-	R	<i>inhA</i> <i>S94A</i>	-
IEMDR24	S	S	S	S	-	-	S	S	-	S	S	-	S	S	-	-	S	-
IEMDR25	S	S	-	S	-	-	S	S	-	S	S	-	S	S	-	-	S	-
IEMDR26	S	R (L)	-	S	-	-	S	S	-	R	S	-	R	S	-	-	S	-
IEMDR27	R	R	-	R (OF X)	-	<i>gyrA</i> <i>D94Y</i>	S	S	-	S	S	-	R	S	-	S	S	-
IEMDR28	-	-	S	S	-	-	S	S	-	-	S	-	S	S	-	-	S	-
IEMDR29	S	S	S	S	-	-	R	R	<i>rrs</i> <i>A140I</i> <i>G</i>	R	R	<i>rrs</i> <i>A140I</i> <i>G</i>	R	R	<i>rrs</i> <i>A140I</i> <i>G</i>	-	S	-
IEMDR30	S	S	-	S	-	-	S	S	-	S	S	-	S	S	-	-	S	-
IEMDR31	S	S	-	S	-	-	S	S	-	S	S	-	S	S	-	S	S	-
IEMDR32	R	R	-	R (OF X)	-	<i>gyrA</i> <i>A90V</i>	R	R	<i>rrs</i> <i>A140I</i> <i>G</i>	R	R	<i>rrs</i> <i>A140I</i> <i>G</i>	R	R	<i>rrs</i> <i>A140I</i> <i>G</i>	S	S	-
IEMDR33	S	S	-	S	-	-	R	S	-	R	S	-	S	S	-	-	S	-
IEMDR34	S	S	-	S	-	-	S	S	-	R	S	-	S	S	-	-	S	-
IEMDR35	S	S	-	S	-	-	S	S	-	-	S	-	R	S	-	-	S	-
IEMDR36	S	R (L)	-	S	-	-	S	S	-	S	S	-	S	S	-	-	S	-
IEMDR40	S	S	-	S	-	-	S	S	-	R	S	-	S	S	-	-	S	-
IEMDR41	S	S	-	S	-	-	S	S	-	S	S	-	S	S	-	-	S	-
IEMDR42	S	S	-	S	-	-	S	S	-	S	S	-	S	S	-	-	S	-

Table 19. Summary of comparison between PhyResSe TB NGS analysis web-tool (genotypic) and isoniazid, rifampicin, ethambutol and pyrazinamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014

Discrepancies are highlighted in purple. R – resistant, S= sensitive, L – low, MUT – mutation, see list of abbreviations for further abbreviations. Promoter region refers to -100bp upstream of the gene.

	ISONIAZID			RIFAMPICIN			ETHAMBUTOL			PYRAZINAMIDE		
GENES + PROMOTERS	<i>inhA, katG, fabG1, ndh, ahpC</i>			<i>rpoB</i>			<i>embC, embA, embB</i>			<i>rpsA, pncA</i>		
Rv	Rv1484, Rv1908c, Rv1483, Rv1854c, Rv2428			Rv0667			Rv3793, Rv3794, Rv3795			Rv1630, Rv2043c		
STUDY NO.	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT
IEMDR01	R	R	<i>fabG1 T-8C, katG S315T</i>	R	R	<i>rpoB H445Y</i>	R	R	<i>embB M306V, M306I</i>	R	R	<i>pncA C14Stop</i>
IEXDR1	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445Y</i>	R	R	<i>embB M306V</i>	R	R	<i>pncA G132C</i>
IEMDR03	R	S	-	R	R	<i>rpoB S450L</i>	S	S	-	S	S	-
IEMDR04	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445Y</i>	S	S	-	S	R	<i>pncA D12N</i>
IEMDR05	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB D328Y</i>	S	S	-
IEMDR06	R	R	<i>katG S315T</i>	R	R	<i>rpoB I491F</i>	R	R	<i>embB M306I</i>	R	S	-
IEMDR07	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	R	R	<i>pncA D136N</i>
IEMDR08	R	R	<i>fabG1 C-15T, katG S315T</i>	R	R	<i>rpoB S450L</i>	S	S	-	S	S	-
IEMDR09	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I, D328Y</i>	R	S	-
IEMDR10	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	R	S	-
IEMDR11	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I, Q497R</i>	R	S	-
IEMDR12	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	S	-	R	S	-
IEMDR13	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I</i>	S	S	-
IEMDR14	R	R	<i>inhA S94A, fabG1 C-15T</i>	R	R	<i>rpoB S450L, D435Y</i>	R	R	<i>embB M306V</i>	R	R	<i>pncA D12E</i>
IEMDR15	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	R	S	-
IEMDR16	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S	R	<i>embB M306I</i>	R	S	-
IEMDR17	R	S	-	R	R	<i>rpoB H445D</i>	S	S	-	S	S	-
IEMDR18	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I, embA C-12T</i>	R	R	<i>pncA Q10P</i>
IEMDR19	R	R	<i>fabG1 C-15T, katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB S297A</i>	S	S	-
IEMDR20	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I</i>	R	R	<i>pncA M175V</i>
IEMDR21	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445Y</i>	R	R	<i>embB M306V</i>	S	S	-
IEMDR22	R	R	<i>katG S315T</i>	R	R	<i>rpoB D435Y</i>	S	S	-	S	S	-
IEMDR23	R	R	<i>inhA S94A, fabG1 C-15T</i>	R	R	<i>rpoB S450L, D435Y</i>	S	R	<i>embB M306V</i>	R	R	<i>pncA D12E</i>
IEMDR24	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S	S	-	S	S	-
IEMDR25	R	R	<i>katG S315T</i>	R	R	<i>rpoB D435Y</i>	S	S	-	S	S	-
IEMDR26	S	S	-	R	R	<i>rpoB H445Y</i>	S	S	-	S	S	-
IEMDR27	R	R	<i>katG S315T, fabG1 C-15T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB Q497R</i>	R	R	<i>pncA H51Y</i>
IEMDR28	R	R	<i>katG S315T, fabG1 C-15T</i>	R	R	<i>rpoB H445L</i>	S	S	-	S	S	-
IEMDR29	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S	R	<i>embB Y319S</i>	R	R	<i>pncA G108R</i>
IEMDR30	R	R	<i>katG S315T, ahpC G-48A</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	S	S	-
IEMDR31	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S/R	R	<i>embB M306V</i>	S	S	-
IEMDR32	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB G406A</i>	R	R	<i>pncA I31S</i>
IEMDR33	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	S	S	-
IEMDR34	R	R	<i>katG S315T, fabG1 C-15T</i>	R	R	<i>rpoB D435V</i>	S	S	-	R	S	-
IEMDR35	R	R	<i>katG S315T, fabG1 C-15T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB Q497R</i>	R	R	<i>pncA H51Y</i>
IEMDR36	R	R	<i>katG S315T</i>	R	R	<i>rpoB D435V</i>	S	S	-	S	S	-
IEMDR40	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445R</i>	S	S	-	S	S	-
IEMDR41	R	R	<i>fabG1 T-8C, katG S315T</i>	R	R	<i>rpoB H445Y</i>	S	S	-	R	R	<i>pncA C14Stop</i>
IEMDR42	R	R	<i>katG S315T, fabG1 C-15T</i>	R	R	<i>RpoB S450L</i>	R	R	<i>embA C-12T, embB Y334H</i>	S	S	-

GENES + PROMOTERS	STREPTOMYCIN			FLUOROQUINOLONES					AMIKACIN			KANAMYCIN		
	<i>rrs,gidB</i>			<i>gyrA,gyrB</i>					<i>rrs</i>			<i>eis</i>		
Rv	Rvnr01/MTB000019, Rv3919			Rv0006,Rv0005					Rvnr01/MTB000019			Rv2416c		
STUDY NO.	DST	WGS	MUT	OFX DST	MOX DST	CIP DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT
IEMDR01	R	S					S			S			S	
IEXDR1	R	R	<i>rpsL K43R</i>			S/R	R	<i>gyrA D94A</i>	R	R	<i>rrs A1401G</i>		S	
IEMDR03	S	S		S	S		S		S	S		S	S	
IEMDR04	R	R	<i>rpsL K43R</i>				R	<i>gyrB N499T</i>		S			R	<i>eis C-10T</i>
IEMDR05	R	R	<i>rrs A 514C</i>				S			S			S	
IEMDR06	R	R	<i>rpsL K43R</i>				S			S			S	
IEMDR07	R	R	<i>rpsL K43R</i>			S	S		S	S			R	<i>eis C-10T</i>
IEMDR08	R	S					S		S	S			S	
IEMDR09	R	R	<i>rpsL K43R</i>			R	R	<i>gyrA A90V</i>		S			R	<i>eis C-10T</i>
IEMDR10	R	R	<i>rpsL K43R</i>			S	S		S	S			R	<i>eis C-10T</i>
IEMDR11	S	S				S	S		S	S			S	
IEMDR12	S	R	<i>rrs C517T</i>	R	R	R	R	<i>gyrA S91P</i>	S	S		S	S	
IEMDR13	R	R	<i>rpsL K43R</i>			S	S		S	S			S	
IEMDR14	R	R	<i>rpsL K43R</i>			S	S		S	S			S	
IEMDR15	R	R	<i>rpsL K43R</i>	S	S		S		S	S		R	R	<i>eis C-10T</i>
IEMDR16	R	R	<i>rpsL K43R</i>	S	S		S		R	S		R	R	<i>eis C-10T</i>
IEMDR17	S	S				S	S		S	S			S	
IEMDR18	R	R	<i>RpsL K43R</i>		S	S	S		S	S		S	S	
IEMDR19	R	R	<i>rpsL K88R</i>			S	S		S	S			S	
IEMDR20	R	R	<i>rpsL K43R</i>	S	S	S	S		S	S		R	R	<i>eis C-10T</i>
IEMDR21	R	S			S	S	S		S	S			S	
IEMDR22	S	S		S	S		S		S	S		S	S	
IEMDR23	R	R	<i>RpsL K43R</i>	S	S		S		S	S		S	S	
IEMDR24	S	S		S	S	S	S		S	S		S	S	
IEMDR25	S	S		S	S		S		S	S		S	S	
IEMDR26	S	S		S	R (L)		S		S	S		R	S	
IEMDR27	R	R	<i>rrs A514C</i>	R	R		R	<i>gyrA D94Y</i>	S	S		S	S	
IEMDR28	R	R	<i>rrs A514C</i>			S	S		S	S			S	
IEMDR29	R	R	<i>rpsL K43R</i>	S	S	S	S		R	R	<i>rrs A1401G</i>	R	S	
IEMDR30	R	R	<i>rpsL K88R</i>	S	S		S		S	S		S	S	
IEMDR31	S	S		S	S		S		S	S		S	S	
IEMDR32	R	R	<i>rpsL K43R</i>	R	R		R	<i>gyrA A90V</i>	R	R	<i>rrs A1401G</i>	R	S	
IEMDR33	R	R	<i>rpsL K43R</i> <i>rrs C517T</i>	S	S		S		R	S		R	R	<i>eis G-14A</i>
IEMDR34	R	R	<i>rpsL K43R</i>	S	S		S		S	S		R	R	<i>eis C-10T</i>
IEMDR35	R	R	<i>rrs A514C</i>	S	S		S		S	S			S	
IEMDR36	S	S		S	R (L)		S		S	S		S	S	
IEMDR40	S	S		S	S		S		S	S		R	S	
IEMDR41	R	S		S	S		S		S	S		S	S	
IEMDR42	R	R	<i>rpsL K88R</i>	S	S		S		S	S		S	S	

Table 20. Summary of comparison between PhyResSe TB NGS analysis web-tool (genotypic) and streptomycin, fluoroquinolone and aminoglycoside DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014

Discrepancies are highlighted in purple. R – resistant, S= sensitive, L – low, MUT – mutation, see list of abbreviations for further abbreviations. Promoter region refers to -100bp upstream of the gene.

	ISONIAZID			RIFAMPICIN			ETHAMBUTOL			PYRAZINAMIDE		
GENES + PROMOTERS	<i>inhA, katG, ahpC, kasA, fabG1 (mabA)</i>			<i>rpoB, rpoC</i>			<i>embC, embA, embB, embR</i>			<i>rpsA, pncA</i>		
Rv	Rv1484, Rv1908c, Rv2428, Rv2245, Rv1483			Rv0667, Rv0668			Rv3793, Rv3794, Rv3795, Rv1267c			Rv1630, Rv2043c		
STUDY NO.	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT	DST	WGS	MUT
IEMDR01	R	R	<i>fabG1 T-8C, katG S315T</i>	R	R	<i>rpoB H445Y</i>	R	R	<i>embB M306V, M306I</i>	R	R	<i>pncA C14Stop</i>
IEEXDR1	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445Y</i>	R	R	<i>embB M306V</i>	R	S	-
IEMDR03	R	R	<i>kasA G312S</i>	R	R	<i>rpoB S450L</i>	S	S	-	S	S	-
IEMDR04	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445Y</i>	S	S	-	S	R	<i>pncA D12N</i>
IEMDR05	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB D328Y</i>	S	R	<i>pncA C14Stop</i>
IEMDR06	R	R	<i>katG S315T</i>	R	R	<i>rpoB I491F</i>	R	R	<i>embB M306I</i>	R	S	-
IEMDR07	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L, rpoC F452S</i>	R	R	<i>embB M306V</i>	R	R	<i>pncA D136N</i>
IEMDR08	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S	S	-	S	S	-
IEMDR09	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I, D328G</i>	R	S	-
IEMDR10	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L, rpoC F452S</i>	R	R	<i>embB M306V</i>	R	S	-
IEMDR11	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I, Q497R</i>	R	S	-
IEMDR12	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB D354A</i>	R	R	<i>pncA A-12G</i>
IEMDR13	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306I</i>	S	S	-
IEMDR14	R	R	<i>fabG1 C-15T, inhA S94A</i>	R	R	<i>rpoB D435Y, rpoB S450L</i>	R	R	<i>embB M306V</i>	R	S	-
IEMDR15	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L, rpoC F452S</i>	R	R	<i>embB M306V</i>	R	S	-
IEMDR16	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L, rpoC G332R, F452S</i>	S	R	<i>embB M306I</i>	R	S	-
IEMDR17	R	R	<i>ahpC C-52T, katG Q295P</i>	R	R	<i>rpoB H445D</i>	S	S	-	S	S	-
IEMDR18	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embA C-12T, embB M306I</i>	R	R	<i>pncA Q10P</i>
IEMDR19	R	R	<i>fabG1C-15T, katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embA C-16T, embB S297A</i>	S	S	-
IEMDR20	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB N296H, M306I</i>	R	R	<i>pncA M175V</i>
IEMDR21	R	R	<i>katG 315T</i>	R	R	<i>rpoB H445Y</i>	R	R	<i>embB M306V</i>	S	S	-
IEMDR22	R	R	<i>katG S315T</i>	R	R	<i>rpoB D435Y</i>	S	S	-	S	S	-
IEMDR23	R	R	<i>fabG1 C-15T, inhA S94A</i>	R	R	<i>rpoB D435Y, rpoB S450L</i>	S	R	<i>embB M306V</i>	R	S	-
IEMDR24	R	R	<i>katG S315T, katG P241P</i>	R	R	<i>rpoB S450L</i>	S	S	-	S	S	-
IEMDR25	R	R	<i>katG S315T</i>	R	R	<i>rpoB D435Y</i>	S	S	-	S	S	-
IEMDR26	S	S	-	R	R	<i>rpoB H445Y</i>	S	S	-	S	S	-
IEMDR27	R	R	<i>fabG1C-15T, katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB Q497R</i>	R	R	<i>pncA H51Y</i>
IEMDR28	R	R	<i>fabG1 C-15T, katG S315T</i>	R	R	<i>rpoB H445L</i>	S	R	<i>embA C-16T</i>	S	S	-
IEMDR29	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S	R	<i>embB Y319S</i>	R	R	<i>pncA G108R</i>
IEMDR30	R	R	<i>ahpC G-48A, katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	S	S	-
IEMDR31	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	S/R	R	<i>embB M306V</i>	S	S	-
IEMDR32	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB G406A</i>	R	R	<i>pncA I31S</i>
IEMDR33	R	R	<i>katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB M306V</i>	S	S	-
IEMDR34	R	R	<i>fabG1 C-15T, katG S315T</i>	R	S	-	S	S	-	R	R	<i>pncA Y103H</i>
IEMDR35	R	R	<i>fabG1 C-15T, katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embB Q497R</i>	R	R	<i>pncA H51Y</i>
IEMDR36	R	R	<i>katG S315T</i>	R	R	<i>rpoB D435V</i>	S	S	-	S	S	-
IEMDR40	R	R	<i>katG S315T</i>	R	R	<i>rpoB H445R</i>	S	S	-	S	S	-
IEMDR41	R	R	<i>fabG1T-8C, katG S315T</i>	R	R	<i>rpoB H445Y</i>	S	S	-	R	R	<i>pncA C14Stop</i>
IEMDR42	R	R	<i>fabG1 C-15T, katG S315T</i>	R	R	<i>rpoB S450L</i>	R	R	<i>embA C-12T, embB Y334H</i>	S	S	-

Table 21. Summary of comparison between TB Profiler NGS analysis web-tool (genotypic) and isoniazid, rifampicin, ethambutol and pyrazinamide DST (phenotypic) for the MDR/XDR-TB cohort from 2001-2014

Discrepancies are highlighted in purple. R – resistant, S= sensitive, L – low, MUT – mutation, see list of abbreviations for further abbreviations. Promoter region refers to -100bp upstream of the gene.

GENES + PROMOTERS Rv	STREPTOMYCIN				FLUOROQUINOLONES				AMIKACIN				KANAMYCIN				CAPREOMYCIN				ETHIONAMIDE											
	DST	WGS	MUT		gyrA, gyrB	OFX	DST	MOX	DST	CIP	DST	WGS	MUT		rrs	DST	WGS	MUT		rrs, flvA	DST	WGS	MUT		ethA, ethR, inhA	DST	WGS	MUT				
Rv0682	R	S			Rv0006, Rv0005					R	S			R	S					R	S				R	S						
IEMDR01	R	S								R	S			R	S					R	S				R	S						
IEMDR02	R	R								R	S			R	S					R	S				R	S						
IEMDR03	S	S								R	S			R	S					R	S				R	S						
IEMDR04	R	R								R	S			R	S					R	S				R	S						
IEMDR05	R	R								R	S			R	S					R	S				R	S						
IEMDR06	R	R								R	S			R	S					R	S				R	S						
IEMDR07	R	R								R	S			R	S					R	S				R	S						
IEMDR08	R	S								R	S			R	S					R	S				R	S						
IEMDR09	R	R								R	S			R	S					R	S				R	S						
IEMDR10	R	R								R	S			R	S					R	S				R	S						
IEMDR11	S	S								R	S			R	S					R	S				R	S						
IEMDR12	S	R								R	S			R	S					R	S				R	S						
IEMDR13	R	R								R	S			R	S					R	S				R	S						
IEMDR14	R	R								R	S			R	S					R	S				R	S						
IEMDR15	R	R								R	S			R	S					R	S				R	S						
IEMDR16	R	R								R	S			R	S					R	S				R	S						
IEMDR17	S	S								R	S			R	S					R	S				R	S						
IEMDR18	R	R								R	S			R	S					R	S				R	S						
IEMDR19	R	R								R	S			R	S					R	S				R	S						
IEMDR20	R	R								R	S			R	S					R	S				R	S						
IEMDR21	R	R								R	S			R	S					R	S				R	S						
IEMDR22	S	S								R	S			R	S					R	S				R	S						
IEMDR23	R	R								R	S			R	S					R	S				R	S						
IEMDR24	S	S								R	S			R	S					R	S				R	S						
IEMDR25	S	S								R	S			R	S					R	S				R	S						
IEMDR26	S	S								R	S			R	S					R	S				R	S						
IEMDR27	R	R								R	S			R	S					R	S				R	S						
IEMDR28	R	R								R	S			R	S					R	S				R	S						
IEMDR29	R	R								R	S			R	S					R	S				R	S						
IEMDR30	R	R								R	S			R	S					R	S				R	S						
IEMDR31	S	S								R	S			R	S					R	S				R	S						
IEMDR32	R	R								R	S			R	S					R	S				R	S						
IEMDR33	R	R								R	S			R	S					R	S				R	S						
IEMDR34	R	R								R	S			R	S					R	S				R	S						
IEMDR35	R	R								R	S			R	S					R	S				R	S						
IEMDR36	S	S								R	S			R	S					R	S				R	S						
IEMDR40	S	S								R	S			R	S					R	S				R	S						
IEMDR41	R	R								R	S			R	S					R	S				R	S						
IEMDR42	R	R								R	S			R	S					R	S				R	S						

Table 22. Summary of comparison between TB Profiler NGS analysis web-tool (genotypic) and streptomycin, fluoroquinolones, aminoglycosides and ethionamide DST (phenotypic) for MDR/XDR-TB isolates collected in Ireland from 2001-2014.

Discrepancies are highlighted in purple. R – resistant, S= sensitive, L – low, MUT – mutation, see list of abbreviations for further abbreviations. Promoter region refers to -100bp upstream of the gene.

TB DRUG	DST AVAILABLE	DISCREPANT ISOLATES	DST	WALKER/KOHL ALGORITHM	PHYRESSE	TB PROFILER	RESEQ TB	MUTATIONS INVOLVED/NOTES
INH	N=39	IEMDR03	R	S	S	R	S	<i>kasA</i> G312S – PHYLOGENETIC SNV
		IEMDR14, IEMDR23	R (H)	R (L)	R (L)	R (L)	R (L)	<i>fabG1</i> T-15C, <i>inhA</i> S94A
		IEMDR17	R	S	S	R	S	<i>ahpC</i> C-52T, <i>kaiG</i> Q295P
RIF	N=39	IEMDR28	R	S	R	R	S	<i>rpoB</i> H445L
		IEMDR34	R	R	R	S	R	<i>rpoB</i> D435V, <i>rpoB</i> D435A
		IEMDR12	R	R	S	R	S	<i>embB</i> D354A
EMB	N=39	IEMDR19	R	S	R	R	S	<i>embA</i> C-16T, <i>embB</i> S297A
		IEMDR29	S	S	R	R	S	<i>embB</i> Y319S
PZA	N=21	VARIOUS DISCREPANCIES INVOLVING VARIOUS MUTATIONS						
STR	N=39	IEMDR01, IEMDR08	R	S	S	S	S	–
		IEMDR05, IE MDR27, IEMDR28, IEMDR35	R	R	R	R	S	<i>rrs</i> A514C
		IEMDR12	S	R	R	R	S	<i>rrs</i> C517T
FQ	N=34	IEMDR21	R	S	S	R	S	<i>rpsL</i> K88M
		IEMDR04 (NO DST AVAILABLE)	–	S	R	R	S	<i>gyrB</i> N499T
		IEMDR26, IEMDR36	R MXF (L)	S	S	S	S	–
AMK	N=34	IEMDR12, IEMDR27, IEMDR28, IEMDR35	S	R	S	R	S	<i>rrs</i> A514C, <i>rrs</i> C517T
		IEMDR33	R	R	S	R	S	<i>rrs</i> C517T
		IEMDR16	R	S	S	S	S	–
KAN	N=22	IEMDR12	S	R	S	R	S	<i>rrs</i> C517T, <i>eis</i> C-10T
		IEMDR15, IEMDR16, IEMDR20, IEMDR34	R	S	R	R	S	<i>eis</i> C-10T
		IEMDR26, IEMDR40	R	S	S	S	S	–
CAP	N=33	IEMDR29, IEMDR32	R	R	S	R	S	<i>rrs</i> A514C
		IEMDR33	R	R	R	R	S	<i>rrs</i> A1401G
		IEMDR12, IE MDR33	S	R	–	S	S	<i>rrs</i> C517T, <i>eis</i> C-10T
ETI	N=8	IEMDR26	R	S	S	S	S	<i>rrs</i> C517T
		IEMDR27, IEMDR35	R	R	R	S	S	–
		IEMDR28	S	R	S	S	S	<i>rrs</i> A514C
LZD	N=10	IEMDR01	R	–	–	R	S	<i>fabG1</i> T-8C
		IEXDR1, IEMDR24, IEMDR27	R	–	–	R	S	<i>fabG1</i> T-15C

Table 23. Discrepancies found for various TB drugs when analysed using phenotypic DST compared to Walker/Kohl algorithm, PhyResSe, TB Profiler and ReseqTB catalogue, where DST available. CAP not reported by PhyResSe, ETI not reported by Walker/Kohl algorithm, PAS, LZD and CFZ not reported by Walker/Kohl algorithm, PhyResSe or ReseqTB.

Antituberculous drugs - INH isoniazid, RIF rifampicin, EMB ethambutol, PZA pyrazinamide. R resistant, S susceptible (or no mutations found in the case of the WGS algorithm), S/R found to be close to the breakpoint of the drug being tested, DST phenotypic drug susceptibility testing result, SNV single nucleotide variation, Y yes, N no, Y/N correlates in some cases. Amino acids – S serine, T threonine, H histidine, Y tyrosine, L leucine, I isoleucine, F phenylalanine, D aspartic acid, V valine, A alanine, M methionine, Q glutamine, R arginine (*fabG1* promoter region mutations – T thymine, C cytosine nucleotides).

CRITERIA	TB Profiler	PhyResSE
User-friendliness (out of 10)	9	9
User interface (out of 10)	7	8
Speed	10 minutes per sample , median run time 5min (range 2-10)	>24 hours for 42 samples
Ability to access result files easily	Download report only	Many files accessible e.g. FastQC, VCF, Fasta
Report available	Yes, one report type available	report in four parts, and overall, also, collated report for ALL uploaded strains
Report style	clear, concise	clear, more detail
Table of results	Yes	yes
Creators	Coll <i>et al</i> [77]	Feuerriegel <i>et al</i> [76]
First public release date	April 2015	April 2015
Open source	yes	yes
Cost	free	free
URL	http://tldr.lshim.ac.uk/	https://bioinf.fz-borstel.de/mchips/phyresse/
Level of User Support	Article cited only, no other user support	every step explained in detail for users
File formats accepted	fastq	fastq
Data Security	not obvious, data seems to be publicly available although pseudonymised by user	private sessions with private 32-character key' sent once SSL encryption has been established, kept by client as a Cookie, jQuery used to upload files
User feedback	Insufficient user control	FastQC should have html file available
Reference strain used	H37Rv Genbank accession number NC_000962.3	H37Rv Genbank accession number NC_000962.3
Mapping software	Snap algorithm	BWA-MEM followed by FastQC and Qualimap of resulting BAM file
Variant calling software	SAMtools, VCFtools, KvarQ	SAMtools, Picardtools, GATK
Programming language	Perl/PHP	CGI, Session Perl module
No. of drug resistance mutations	1,325 variants at 992 nucleotide positions from 31 loci, six promoters, 25 coding regions	614 variants in coding, ribosomal and intergenic regions
List of mutations curated	Not documented	yes, version updated at intervals
Sources of drug resistance mutation discovery	TBDreadMDB, MUBIL-TB-DB, phylogenetic SNPs at drug resistant loci removed	Published literature, experimental data, publicly available data
Drugs included	AMK, CAP, EMB, ETH, INH, KAN, MOX, OFX, PZA, RMP, STR (PAS, LZD, CFZ, BDQ sensitivity and specificity not determined)	AMK, CAP, EMB, ETH, INH, KAN, FQ, PZA, RMP, STR, PAS, LZD
Lineage specific mutations included for identification	yes	yes
Novel polymorphisms detected	yes	yes
Compared against	Xpert MTB RIF, MTBDRplus, MTBDRsl, phenotypic DST	DST, Sanger sequencing of resistance genes
Limitations	DST availability, especially for EMB and PZA, poor specificity to CAP and EMB, MOX had few DST results	DST availability
Quality control	Q30 (1 error per 1000bp) variants only allowed	fastqValidator for fastq files, FASTQC and Qualimap for BAM files
Upload	1 file at a time	as many files as required at once

Table 24. Comparison of open-source TB NGS analysis web-tools PhyResSe and TB Profiler

Web-tools were compared in areas such as quality control, user-friendliness, report format, limitations, resistance mutation catalogue, drugs included, software used, level of user support, and others.

RpoB S450L (S531L with *E.coli* numbering) was the most common SNV associated with rifampicin resistance (n=25/39, 64%), followed by SNVs at *rpoB* codon 445 (position 526 with *E.coli* numbering) (H445Y, n=6, H445D, n=1, H445L, n=1, and H445R, n=1, overall 23%).

Ethambutol resistance was dominated by SNVs at *embB* codon 306 (M306V, n=11/27, 41%, M306I, n=8/27, 30%). Pyrazinamide mutations were scattered across the *pncA* gene. Streptomycin resistance was mostly associated with *rpsL* K43R (18/26, 69%). Fluoroquinolone (FQ) drug-resistance-associated mutations were seen less than any of the other drugs (n=6/39, 15% isolates with drug-resistance-associated mutations). Aminoglycoside resistance was mainly associated with kanamycin resistance (*eis* gene mutations constituted 9/17, or 53% of mutations seen).

5.2.6.2 Novel Mutations detected with NGS

Two novel mutations were found within the 25 candidate genes (not included in Tables 12 and 13).

IEMDR04 had a non-synonymous SNV Ala209Val (coverage 234, variant frequency 93.2%, average base call quality 34) in *ndh/Rv1854c*, a gene thought to be involved in isoniazid resistance. IEMDR04 was phenotypically isoniazid resistant.

The second occurred in IEXDR1, which had the non-synonymous SNV Val371Ala (coverage 330, variant frequency 99.1%, average base call quality 35) in *embC/Rv3793*, a gene thought to be associated with ethambutol resistance. IEXDR1 was phenotypically resistant to ethambutol. These mutations were not found in the literature previously.

Since the isolates also harboured mutations that were previously found to be associated with isoniazid and ethambutol resistance, these mutations did not increase the sensitivity of the analysis in these cases, however they may have an effect where the other mutations are not present.

5.2.6.3 Mutations found in other candidate genes

A total of 43 candidate genes (excluding the 25 detailed in Tables 12 and 13), where any association with resistance had been previously cited, were analysed manually within the MDR/XDR-TB cohort. This list of genes was derived from the TBDream database, the Broad Institute TB database (no longer available) and a number of publications [191-194]. Non-synonymous SNVs with over 90% variant frequency were searched for. No mutations were seen in 23 of these genes. Table 14 details the mutations found within the remaining 21.

None of the drug-resistance-associated mutations associated with bedaquiline and delamanid cross resistance were detected in the cohort [195]. Mutations in Rv0486 *mshA* (A187V and N111S),

Rv0407 *fgdI* (K270M), Rv2764c *thyA* (T202A) and Rv0129c *fpbC* (G158S), were found to be related to lineage rather than resistance (phylogenetic) [76]. Others were found in both susceptible and resistant strains (Rv2846c *efpA* [I73T], Rv3139 *fadE24* [I430L] and Rv1592c [I322V]) [196-198]. Other mutations were found elsewhere but not proven with high confidence to be associated with resistance, for example Rv3124 *moaR1* P54S [199], and Rv2764c *thyA* Q97R [200]. Two mutations were also found by Koser *et al* during a study carried out on a mixed population of XDR-TB within the same patient (Rv1704c *cycA* [R93L] and Rv2981c *ddlA* [T365A], both related to cycloserine resistance). However, almost all of the isolates in this cohort harboured these two mutations. Either all of those isolates are resistant to cycloserine, or these mutations are coincidental. It has previously been hypothesised that a mutation in Rv2247 *accD6* (D229G) is involved in isoniazid resistance, and is present within 14 isolates in the current cohort [201]. The remainder were not found in the published literature.

5.2.6.4 NGS analysis using the Walker/Kohl *et al* Algorithm compared to Phenotypic DST

A summary of the results of comparison between genotypic and phenotypic drug resistance prediction using drug-resistance-associated mutations derived by Walker/Kohl *et al* can be seen in Tables 15 and 16. Discrepancies are high-lighted. Sensitivity and specificity results for drugs included in the Walker/Kohl *et al* algorithm compared to DST can be seen in Table 11. Overall, sensitivity of 75% (95% CI 67-81) and specificity of 92% (CI 86-96) was seen with this WGS analysis method.

Mixed calls were found in IEMDR01 for ethambutol (*embB* M306V 54.5% and M306I 25.3%) and IEMDR34 for rifampicin (*rpoB* D435V 53.5% and D435A 45.6%, *E.coli* numbering D572V and D572A).

IEMDR08 had the lowest coverage when mapped to H37Rv (10.55 +/- 5.72, mapped reads 10.92%) which could have contributed to the discrepancy found for streptomycin. All other isolates had at least over 65-fold coverage over at least 77% of the genome, so do not explain the remaining discrepancies.

By far the most discrepancies occurred where pyrazinamide was found to be phenotypically resistant, but no resistance-associated mutation was found (n=19). Many of the remaining discrepancies were false negatives, where a resistance-associated mutation was not found to match phenotypic resistance.

False positives occurred in fewer cases. IEMDR12 harboured *rrs* C517T (which has been associated with aminoglycoside resistance) but kanamycin, capreomycin, streptomycin and amikacin were phenotypically susceptible. IEMDR27 harboured *rrs* A514C (also associated with aminoglycoside resistance) but was phenotypically susceptible to kanamycin and amikacin (although resistant to capreomycin). IEMDR28 also harboured *rrs* A514C but was phenotypically susceptible to amikacin and capreomycin.

5.2.6.4.1 ‘Uncharacterised’ Mutations according to Walker/Kohl algorithm

Walker/Kohl *et al* published uncharacterised mutations found in their study that were only found in resistant phenotypes [184]. Two of these mutations were found among the current cohort.

One (IEMDR04) was a promoter region SNV in *embA* (C-8T at nucleotide position 4243225, see Table 15) that was only found in ethambutol resistant strains in the aforementioned study. Here, ethambutol was found to be susceptible.

The second uncharacterised mutation found was in a promoter region of *ahpC* gene in IEMDR30 (G-48A at nucleotide position 2726145, 99% variant frequency). This was found, as predicted by Walker/Kohl *et al*, to be in an isoniazid resistant isolate, and co-exists with the high confidence mutation *katG* S315T. Mutations in this *oxyR*'-*ahpC* region have previously been associated with a fitness compensatory role [202]. When TB Profiler and PhyResSe were subsequently used to analyse the isolates, this mutation was also found.

5.2.6.5 NGS analysis with ReSeqTB drug resistance mutation catalogue compared to Phenotypic DST

A summary of the results of comparison between genotypic and phenotypic drug resistance prediction using drug-resistance-associated mutations from the ReSeqTB data-sharing platform can be seen in Tables 17 and 18. Discrepancies are high-lighted. WGS analysis results for sensitivity and specificity of drugs included in the ReSeqTB catalogue of drug-resistance-associated mutations compared to DST can be seen in Table 11. Overall, sensitivity of 79% (CI 72-85) and specificity of 99% (CI 95-100) was seen with this WGS analysis method.

Two false positives were found where ethambutol was phenotypically susceptible but resistance-associated mutations were found using WGS (IEMDR16 *embB* M306I and IEMDR23 *embB* M306V). The remaining discrepancies were false negatives where phenotypic resistance was seen but no mutation was found, the majority of which were associated with ethambutol, pyrazinamide and kanamycin.

5.2.6.6 *PhyResSe* NGS analysis compared to Phenotypic DST

A summary of the results of comparison between genotypic and phenotypic drug resistance prediction using drug-resistance-associated mutations from the ReseqTB data-sharing platform can be seen in Tables 19 and 20. Discrepancies are high-lighted. WGS analysis results for sensitivity and specificity of drugs using PhyResSe compared to DST can be seen in Table 11. Overall, sensitivity of 86% (CI 80-91) and specificity of 96% (CI 90-99) was seen with this WGS analysis method.

Five false positives were found. Three isolates contained resistance-associated mutations for ethambutol but were phenotypically susceptible – IEMDR16 *embB* M306I, IEMDR23 *embB* M306V and *embB* Y319S. IEMDR04 harboured a *pncA* mutation (associated with pyrazinamide resistance), D12N, but was phenotypically susceptible. IEMDR12 harboured *rrs* C517T, however streptomycin was susceptible (this mutation is not associated with cross-resistance to other aminoglycosides by the PhyResSe platform). False negatives were found most commonly with ethambutol and streptomycin.

5.2.6.7 *TB Profiler* NGS analysis compared to Phenotypic DST

A summary of the results of comparison between genotypic and phenotypic drug resistance prediction using drug-resistance-associated mutations from the ReseqTB data-sharing platform can be seen in Tables 21 and 22. Discrepancies are high-lighted. WGS analysis results for sensitivity and specificity of drugs analysed using TB Profiler compared to phenotypic DST can be seen in Table 11. Overall, sensitivity of 85% (CI 79-90) and specificity of 93% (CI 88-96) was seen with this WGS analysis method.

Thirteen false positives were found. IEMDR12 harboured *rrs* C517T (which has been associated with aminoglycoside resistance) but kanamycin, capreomycin, streptomycin and amikacin were phenotypically susceptible. The mutation *eis* C-10T (associated with kanamycin resistance) was also present in IEMDR12 even though kanamycin was susceptible. IEMDR27 harboured *rrs* A514C (associated with aminoglycoside resistance) but was phenotypically susceptible to kanamycin and amikacin (although resistant to capreomycin). The same mutation was found in IEMDR28, however the isolate was phenotypically susceptible to amikacin and capreomycin. This *rrs* mutation was also found in IEMDR35, however it was amikacin susceptible. Four false positives were seen involving ethambutol susceptible isolates – IEMDR16 (*embB* M306I), IEMDR23 (*embB* M306V), IEMDR28 (*embA* C-16T) and IEMDR29 (*embB* Y319S). Pyrazinamide susceptible isolates IEMDR04 and IEMDR05 harboured mutations *pncA* D12N and C14Stop respectively. IEMDR27 contained an ethionamide resistance-associated mutation *fabG1* (C-15T)

although it was phenotypically susceptible. Most false negative results were seen in relation to pyrazinamide (n=11).

5.2.6.8 Discrepancies that resulted when all platforms were compared

Table 23 visualises the discrepancies between phenotypic DST and genotypic drug resistance prediction when all platforms were taken into account.

Isoniazid

IEMDR03 was found resistant by DST and TB Profiler, but susceptible by Walker and Kohl *et al* algorithm, ReseqTB or PhyResSe. The mutation predicted by TB Profiler was *kasA* G312S, which has been shown to be a phylogenetic SNV [203].

IEMDR14 and IEMDR23 were found high-level resistant phenotypically but this was not predicted genotypically by any method, even though mutations *fabG1* C-15T and *inhA* S94A (related to low-level resistance) were found.

IEMDR17 was found resistant with DST and TB Profiler only. Two drug-resistance-associated mutations were found by TB Profiler – *ahpC* C-52T and *katG* Q295P. *AhpC* C-52T has been found in XDR-TB isolates previously [204]. *KatG* Q295P has also been reported previously, but associated with low-level isoniazid resistance [205].

An additional isoniazid mutation was found by TB Profiler in IEMDR24 – *katG* P241P. An extra mutation was found by TB Profiler and PhyResSe in IEMDR30 – *ahpC* G-48A, which has been seen previously [206]. This had also been found as an ‘uncharacterised’ mutation by Walker/Kohl *et al* that had only been seen in resistant isolates. These were not discrepancies as such, but instead, could have some compensatory role in isoniazid resistance.

Rifampicin

TB Profiler and PhyResSe predicted *rpoB* H445L (*E.coli* numbering H526L) as the SNV causing phenotypic resistance in IEMDR28, but the other methods did not report this form of the mutation as being associated with rifampicin resistance.

Walker and Kohl *et al* showed a mixed call for IEMDR34 (*rpoB* D435V 53.5% and D435A 45.6%, *E.coli* numbering D572V and D572A), however TB Profiler did not report either mutation, and PhyResSe and ReseqTB reported just one mutation *rpoB* D435V (*E.coli* numbering D572V).

Additional rifampicin mutations were found in six isolates (see Tables 19-22). TB Profiler reported *rpoC* F452S alongside *rpoB* S450L (*E.coli* numbering S531L) in IEMDR06, IEMDR10 and IEMDR15, which could be a compensatory mutation in the Beijing lineage [207]. TB Profiler found two extra mutations in IEMDR16 - *rpoC* F452S and *rpoC* G332R, also a possible compensatory mutation [208]. TB Profiler and PhyResSe found an extra mutation in IEMDR14 D435Y (*E.coli* numbering D572Y) alongside *rpoB* S450L (*E.coli* numbering S531L).

Ethambutol

Walker and Kohl *et al* and TB Profiler predicted that *embB* S354A was the SNV causing phenotypic resistance in IEMDR12, however PhyResSe and ReseqTB did not report resistance. In IEMDR19, TB Profiler and PhyResSe found *embB* S297A to be causing resistance, but ReseqTB and Walker and Kohl *et al* did not report this. TB Profiler found an extra mutation in IEMDR19 – *embA* C-16T. IEMDR29 was phenotypically susceptible. ReseqTB and Walker and Kohl *et al* correlated with this result, however both TB Profiler and PhyResSe predicted resistance (*embB* Y319S).

Pyrazinamide

Various discrepancies across the platforms were seen. The highest diversity of mutations was also seen with this drug.

Streptomycin

Across all genotypic methods, no mutation was seen that could explain phenotypic streptomycin resistance in IEMDR01 or 08. Four isolates were found to be resistant by DST and all genotypic methods but ReseqTB, which does not include *rrs* A514C in its catalogue. ReseqTB did not include another *rrs* mutation (C517T) in its catalogue either, but in the case of IEMDR12, it was the only method to correlate with the susceptible DST. TB Profiler was the only genotypic method to correlate with phenotypic resistance found in IEMDR21 (mutation *rpsL* K88M).

Fluoroquinolones

Although phenotypic DST was not available for IEMDR04, the mutation found by TB Profiler, *gyrB* N499T, caused a discrepancy between it and the other platforms. Also, no mutation was found among any of the platforms that would explain low level moxifloxacin resistance found phenotypically in IEMDR26 and 36.

Aminoglycosides

The *rrs* mutations C517T and A514C caused most discrepancies. Neither PhyResSe nor ReseqTB recognise these mutations as causing aminoglycoside cross-resistance, while the other platforms do. Phenotypic kanamycin resistance in IEMDR26 and 40 could not be explained by a mutation.

PhyResSe does not consider *rrs* A1401G a kanamycin-associated mutation, whereas the other methods do. The other methods correlated with phenotypic DST in this instance (IEMDR29 and 32).

Ethionamide

ReseqTB does not recognise *fabG1* T-8C as being associated with ethionamide resistance, whereas TB Profiler does. In the case of IEMDR01, the TB Profiler correlated with phenotypic DST. Similarly, ReseqTB does not include *fabG1* C-15T as a mutation associated with ethionamide resistance, while the other platforms do. ReseqTB correlates with phenotypic DST for IEMDR27.

Linezolid

Linezolid was not reported by PhyResSe, ReseqTB or the Walker/Kohl algorithm due to the lack of phenotypic DST information available. TB Profiler failed to find a mutation that could correlate with phenotypic resistance to linezolid in IEXDR1, IEMDR24, and 27.

5.2.6.9 Web-tool Comparison: TB Profiler compared to PhyResSe NGS analysis

A web-tool comparison of PhyResSe and TB Profiler was also performed. Tables 9 and 24 display the results of the comparison. Both methods successfully identified the isolates from raw fastq files, and provided drug resistance profiles and genotyping for each isolate. TB Profiler performed the analysis in less time than PhyResSe (10 min per isolate compared to over 24 hours for 42 isolates respectively), however PhyResSe performed significant amounts of quality control, which takes longer to complete. TB Profiler allows download of a drug resistance profile and lineage report. Both had high sensitivity and specificity for isoniazid and rifampicin. TB Profiler and PhyResSe displayed 100% correlation when global TB lineage calling was compared, however they differed based on sub-lineage in 36% of cases (14/39) (Table 24).

5.2.7 Sequential Isolate Study I: Progression from Susceptible to MDR-TB

A pan-susceptible pulmonary isolate was recovered from a sample taken in February 2004 from a male, 49 year old, Irish patient with no known risk factors for TB (Table 8). Wild-type patterns were seen for *rpoB*, *katG* and *inhA* on MTBDR*plus* at that time. A second pulmonary isolate from the same patient, recovered in September 2004, was phenotypically resistant to isoniazid while showing the same MTBDR*plus* result as before. In March 2005, a third pulmonary isolate (IEMDR03, Table 8, Figures 38-40) displayed phenotypic resistance to both isoniazid and rifampicin. An *rpoB* mutation (S450L, or S531L with *E.coli* numbering) was seen on MTBDR*plus* hybridisation strip, confirming the presence of rifampicin resistance. However, isoniazid resistance could not be confirmed with the LPA. Clinically, the patient was confirmed as having developed MDR-TB.

WGS analysis was also performed on the sequential isolates from 2004. Only non-synonymous SNVs with read depth above 20 and variant frequency above 95% were considered. No compensatory mutations in *rpoC* were identified [208, 209]. TB Profiler platform was the only one to correlate with the isoniazid DST, citing that *kasA* (Gly312Ser) was present in both. However, this mutation has been proven to be associated with EAI Lineage 3 and not involved in resistance [203]. On further examination of the differences between the sequential isolates (70 possible resistance-associated genes searched), the second isolate had acquired a mutation on *katG* Q439R (coverage 137, variant frequency 96.4%), which was not found in the first. However, the third isolate did not harbour that particular mutation, but instead had acquired a different *katG* mutation D381A (coverage 177, variant frequency 99.4%), as well as a high confidence *rpoB* mutation S450L (coverage 243, variant frequency 100%), which correlated with the MTBDR*plus* assay result. An SNV in Rv3728 was also found in the latter isolate, V488I (coverage 151, variant frequency 91.4%). The former *katG* Q439R mutation has been seen before, however is not high confidence for drug resistance as such, and seems to have been lost prior to the development of classical MDR [210]. Neither *katG* D381A or Rv3728 V488I was found in the literature. However, any novel *katG* gene could play a role in resistance to isoniazid, it has been found [201].

5.2.8 Sequential Isolate Study II: Progression from MDR-TB to XDR-TB

In February 2014, an isolate (IEMDR27) was received in the IMRL from a male, 40 year old, Latvian patient with clinical symptoms of pulmonary TB, who was a smoker, associated with alcohol misuse, from a high TB burden country (Table 8, Figures 38-40). The isolate was resistant to all first-line agents, as well as rifabutin, prothionamide, PAS and capreomycin, and susceptible to amikacin, cycloserine, ethionamide, kanamycin, linezolid, moxifloxacin and ofloxacin at that time (Table 10). The patient was diagnosed with MDR-TB. A second pulmonary isolate was recovered 9 months later (November 2014, IEMDR27LATER) which was found to be resistant to the above plus fluoroquinolones (moxifloxacin and ofloxacin). This represents the first case of transition from MDR- to XDR-TB while on treatment in Ireland.

Sixty-nine genes putatively involved in resistance were analysed in the whole genomes of the earlier and later isolates. Only non-synonymous SNVs with read depth above 20 and variant frequency above 95% were considered. No compensatory mutations in *rpoC* were identified [208, 209]. Both isolates shared high confidence mutations such as Rv0667 *rpoB* S450L (S531L *E.coli* numbering) which is related to rifampicin resistance, Rv1908c *katG* S315T, related to isoniazid resistance, Rv3795 *embB* Q497R, related to ethambutol resistance, and Rv2043c *pncA* H51Y, related to pyrazinamide resistance. One significant mutation was acquired between February and November. A mutation in Rv0006 *gyrA* gene, D94Y, which has been associated strongly with

fluoroquinolone resistance (coverage 153, variant frequency, 98.7%) was found. A high confidence mutation associated with aminoglycoside resistance was not found, however an XDR-TB phenotype was confirmed due to the presence of capreomycin resistance.

5.2.9 Sequential Isolate Study III

IEMDR01EARLY was collected in 2001 from a male 34 year old Irish patient with no known TB risk factors, who was treated over a long period of time. The patient remained culture positive at least up to 2004, when another isolate was received (IEMDR01) (Table 8, 10, Figures 38-40). This was anecdotally associated with non-compliance. Phenotypic DST was performed elsewhere on the first isolate, and was recorded as at least isoniazid and rifampicin resistant. In 2004, the IMRL reported all first-line drugs resistant, as well as PAS and ethionamide.

WGS was performed on both isolates in 2014. Only non-synonymous SNVs with read depth above 20 and variant frequency above 95% were considered. No compensatory mutations in *rpoC* were identified [208, 209]. The isolates shared high confidence mutations for isoniazid (Rv1908c *katG* S315T) and rifampicin (Rv0667 *rpoB* H445Y, or H526Y with *E.coli* numbering). IEMDR01EARLY harboured high confidence mutations associated with phenotypic ethambutol resistance Rv3795 *embB* M306V and M306I, however only M306V was present by the time the second isolate was recovered. IEMDR01 also harboured another ethambutol associated mutation, *embB* D328Y. Two phylogenetic, and 6 other, SNVs were observed in IEMDR01EARLY that were not present in the later IEMDR01. IEMDR01 contained 9 phylogenetic and 11 other SNVs that were not originally present in the 2001 isolate. None of the non-phylogenetic mutations were included in the resistance catalogues of any of the WGS analysis platforms used in the current study [76, 77, 121, 184].

5.2.10 Draft Genome Sequence of the First XDR-TB Isolate in Ireland

As part of the analysis, a paper was published in Genome Announcements which characterised the whole genome sequence of the first XDR-TB strain in Ireland. Below are details extracted from the paper, see Appendix 2 [164]:

The first Irish XDR-TB case was isolated in the IMRL in 2005 from a 26 year old female Lithuanian patient (IEXDR1, Table 8, 10, Figures 38-40) [211, 212]. First-line DST was completed within three weeks (streptomycin, isoniazid, rifampicin, ethambutol and pyrazinamide resistant), second- and third-line within five weeks (amikacin, clarithromycin, ciprofloxacin, rifabutin resistant, capreomycin, clofazimine and prothionamide susceptible) with some intra-laboratory

discrepancies, and fourteen weeks (PAS highly resistant, ethionamide and cycloserine susceptible) (Table 25).

In March 2014, WGS was performed to provide further molecular confirmation of IEXDR1 (lineage 2, East Asian, or Beijing strain). The analysis yielded a mapped-read-depth of 196-fold, covering 97.6% of the H37Rv genome. A final draft assembly of 4,340,174 bp consisting of 109 contigs (largest contig, 217,725bp) was achieved [119]. Variant analysis recovered 1,492 SNVs in the assembled genome with respect to H37Rv, of which 810 were non-synonymous (depth of coverage ≥ 20 -fold, variant frequency $\geq 95\%$) (Table 26).

Non-synonymous mutations were identified in genes Rv0667 (*rpoB*) [H526Y] and Rv1908c (*katG*) [S315T]. High-confidence SNVs were also found for second-line fluoroquinolones in gene Rv0006 (*gyrA*) [D94A] and aminoglycosides in MTB000019 (*rrs*) [A1401G] [193]. This was consistent with the XDR phenotype of IEXDR1. NGS data also correlated with the DST resistance profile of IEXDR1 for ethambutol (Rv3795 (*embB*) [M306V] and Rv3793 (*embC*) [V371A]), pyrazinamide (Rv2043c (*pncA*) [G132C]), and streptomycin (Rv0682 (*rpsL*) [K43R]). Other SNVs, which may confer resistance to antibiotics such as ethionamide and thioacetazone (Rv3854c (*ethA*) [Y211C] and Rv0644c (*mmaA2*) [E213D]), cycloserine (Rv1704c (*cycA*) [R93L]) and cotrimoxazole (Rv3764c (*thyA*) [Q97R] were also identified in the IEXDR1 genome, although their specificities are not as well defined [192].

Previously-described phylogenetically-informative polymorphisms (*katG* [R463L], Rv2629 [D64A], *embA* [C76C, TGC/TGT] and [Q38Q, CAA/CAG], *rpsA* [R212R, CGA/CGC], *gidB* [E92D] and *mshA* [A187V] confirm the presence of a Beijing strain (MtbC15-9 code, 100-32, Table 27) [213].

Anti-tuberculous Agent	DST Results (2005)			
	IMRL	SMRL	HPA	Lithuania
Streptomycin	R	R	–	R
Isoniazid (0.1)	R	R	–	R
Isoniazid (High Level) (0.4)	R	R	–	R
Rifampicin	R	R	–	R
Ethambutol	R	R	–	R
Pyrazinamide	R	R	–	R
Amikacin	–	R	highly R	R
Clarithromycin	–	R	S	–
Clofazimine	–	S	–	–
Ciprofloxacin	–	R	S	–
Rifabutin	–	R	–	–
Capreomycin	–	S	S	S
Prothionamide	–	S	–	–
Cycloserine	–	–	S	S
Ethionamide	–	–	S	S
PAS	–	–	highly R	R
Ofloxacin	–	–	–	R
Kanamycin	–	–	–	R

Table 25. Diagnostic susceptibility testing (DST) performed by different laboratories on IEXDR1

Irish Mycobacteria Reference Laboratory (IMRL), Scottish Mycobacteria Reference Laboratory (SMRL), Health Protection Agency (HPA, now Public Health England) and a Lithuanian Doctor's DST report received at the time [164]. Highlighted in purple are intra-laboratory discrepancies.

Table 26. SNVs found in IEXDR1

Some of the SNVs found had already been described in the literature, and the anti-tuberculous drug resistance with which those SNVs have been associated was recorded in the table. Phylogenetic SNVs were also found.

Gene	Locus tag	Mutation position in IEXDR01 (vs H37Rv reference AL123456)	Variant Frequency	Coverage	Possibly linked to resistance to... [23]
<i>gyrA</i>	Rv0006	G668D	96.90%	325	Phylogenetic
<i>gyrA</i>	Rv0006	S95T	97.20%	290	fluoroquinolones
<i>gyrA</i>	Rv0006	D94A	96.10%	283	fluoroquinolones
<i>gyrA</i>	Rv0006	E21Q	97.10%	314	Phylogenetic
<i>iniB</i>	Rv0341	A347S	30.00%	10	-
<i>rpoB</i>	Rv0667	1075 (SYN)	95.80%	286	?rifampicin
<i>rpoB</i>	Rv0667	H445Y (H526Y, E.coli)	96.40%	309	rifampicin
<i>rpoB</i>	Rv0667	L378R	96.70%	239	?rifampicin
<i>rpsL</i>	Rv0682	K43R	95.90%	340	?streptomycin
<i>tlyA</i>	Rv1694	11 (SYN)	97.50%	280	-
<i>katG</i>	Rv1908c	S315T	97.80%	278	isoniazid
<i>katG</i>	Rv1908c	R463L	98.40%	314	Phylogenetic
<i>pncA</i>	Rv2043c	G132C	98.50%	407	?pyrazinamide
<i>thyA</i>	Rv2764c	Q97R	98.20%	285	?cotrimoxazole
<i>ddlA</i>	Rv2981c	T365A	98.10%	266	?cycloserine
<i>embC</i>	Rv3793	927 (SYN)	98.40%	192	?ethambutol
<i>embC</i>	Rv3793	V371A	98.20%	337	?ethambutol
<i>embA</i>	Rv3794	76 (SYN)	98.90%	279	Phylogenetic
<i>embA</i>	Rv3794	38 (SYN)	96.40%	279	Phylogenetic
<i>embB</i>	Rv3795	M306V	98.60%	348	ethambutol
<i>ethA</i>	Rv3854c	Y211C	97.60%	340	?thioacetazone
<i>gidB</i>	Rv3919c	E92D	99.20%	381	Phylogenetic
<i>gid</i>	Rv3919c	205 (SYN)	97.60%	328	?aminoglycosides
<i>fabD</i>	Rv2243	201 (SYN)	98.40%	183	-
<i>fabD</i>	Rv2243	T115A	96.50%	115	-
<i>accD6</i>	Rv2247	D229G	98.40%	248	-
<i>accD6</i>	Rv2247	200 (SYN)	98.80%	248	-
-	Rv3728	325 (SYN)	97.00%	199	?aminoglycosides
<i>mmpL3</i>	Rv0206c	122 (SYN)	97.80%	278	-
<i>dprE1</i>	Rv3790	153 (SYN)	96.90%	223	-
<i>cycA</i>	Rv1704c	R93L	97.40%	227	?cycloserine
<i>fgd1</i>	Rv0407	320 (SYN)	97.20%	282	-
<i>mshA</i>	Rv0486	A187V	97.90%	328	Phylogenetic
<i>mmaA2</i>	Rv0644c	E213D	97.10%	314	?thioacetazone
<i>rrs</i>	-	A1401G	-	-	aminoglycosides

Number of tandem repeats	Genome position (H37Rv) number
2	580
7	2996
3	802
3	960
3	1644
5	3192
4	424
4	577
4	2165
4	2401
3	3690
2	4156
6	2163b
5	1955
7	4052
2	154
5	2531
3	4348
2	2059
1	2687
3	3007
4	2347
2	2461
3	3171

Table 27. MIRU-VNTR profile of IEXDR1

This signature is representative of an East Asian Lineage 2 Beijing strain [164].

5.3 Discussion

This study characterised the molecular epidemiology and drug resistance present in MDR/XDR-TB isolates collected in Ireland between 2001 and 2014. A detailed picture of the MDR/XDR-TB present in Ireland could be useful in the fight to control its spread. It is clear that most of the cases arose in those born outside Ireland. This correlates with previous studies on drug resistance in Ireland [185]. Cases came from a diverse range of countries. The 2011 census of Ireland recorded a 143% increase in non-Irish nationals in 9 years. Over the course of these years, the largest increase was among Polish, Lithuanian, Romanian, Indian and Latvian populations. The majority of non-Irish cases occurred in patients originally from former Soviet Union states, which would correlate with the Central Statistics Office information. Poland and Romania were not in the Soviet Union but neighboured countries who were, and had ties with the Soviet Union. Latvia and Lithuania were former Soviet states. The emergence and spread of MDR/XDR-TB following the fall of the Soviet Union, and its public health system, has been documented [17].

Movement of people within the EU, and from outside the EU, is considered a threat to the global management of MDR/XDR-TB, but it is difficult to see how this could ever be curtailed without impinging on people's human rights [26, 27]. Instead, other ways of managing the risk must be sought. For example, immigrant screening is offered to individuals on entry to Ireland, however, it is not compulsory. People may fear the consequences of testing positive for LTBI or active TB infection, or may not be aware that there is no charge for TB treatment in most EU states. Despite this, compulsory screening, at point of entry, of immigrants from high prevalence regions, would almost certainly find more LTBI and active infection, which could be dealt with immediately to the benefit of both the individual and the wider community [26, 27, 132, 214-216]. The immigrants may have come from war-torn countries, may have travelled long distances under severe stress and may have contracted TB on their journey. Timely screening and thorough follow-up, of documented and un-documented migrants, as well as dealing with the related social issues (stigma, fear, poverty), could have a high intervention yield and should be a priority for TB control in Ireland [217].

5.3.1 Molecular Epidemiology of MDR/XDR-TB in Ireland and compared with Europe

The diversity of MDR/XDR-TB strains correlates with the overall diversity of susceptible MTBC strains found in Ireland, investigated with MIRU-VNTR genotyping (Chapter 2 and Table 3 and 8). Lineage distribution for the MDR-TB strains consisted of 54.7% Euro-American, 33.3% East Asian, 7.2% East African Indian, and 4.8% Indo-Oceanic. Lineage distribution of susceptible strains consisted of 66.1% Euro-American, 9% East Asian, 9.8% East African Indian and 11.1% Indo-Oceanic. For the Beijing strain, the relative risk of isolating an MDR/XDR-TB strain is 3.8,

which is considered significant (p-value <0.0001). Association between this lineage and drug resistance has been previously observed [17, 218]. None of the other lineages were found to be associated with drug resistance.

Comparing the Irish MDR/XDR-TB cohort to European clusters reported by the ECDC on 2014 data, we share 7 'cross-border cluster' genotypes (MtbC15-9), which provides the first molecular evidence that supports the above hypothesis that, like other countries in Europe, movement of people within and from outside the EU has encouraged the spread of MDR/XDR-TB towards Ireland [188]. Its island status has not protected it from its EU status. For the most part, WGS phylogenetics agreed with MIRU-VNTR genotyping clusters, which strengthens the theory further. This evidence could provide the impetus for improved TB screening and follow-up care at points of entry to Ireland.

WGS and MIRU-VNTR genotyping did not correlate for every cluster however. Three MIRU-VNTR clusters were refuted by WGS, and another group of distinct MIRU-VNTR genotypes clustered together when their whole genomes were analysed. As in chapter 4, which compared WGS analysis of outbreaks with MIRU-VNTR genotyping, there are occasional discrepancies with MIRU-VNTR genotyping due to its resolution (24 genomic loci compared to 4.4 million genomic loci with WGS).

5.3.2 Multi- and Extensive Drug Resistance in Ireland and compared worldwide

On analysing the results of this cohort, it could, in fact be deemed a multi-national study of MDR/XDR-TB since countries from every continent were included. When compared to worldwide studies on drug resistance in tuberculosis, the current study cohort is extremely similar in its mutation make-up to worldwide strains. One study examined rifampicin (*rpoB*), isoniazid (*katG*, *inhA*), fluoroquinolone (*gyrA*, *gyrB*), and aminoglycoside (*rrs*, *eis*) resistance in 417 isolates from India, Moldova, the Philippines and South Africa [193]. The most common mutations found were also found in the current study. They noted little regional variation among strains, which is borne out by the current study. One variation they did see was with *rrs* and *eis* promoter mutations associated with kanamycin resistance. In India and South Africa, kanamycin resistance was mainly caused by *rrs* mutations, while in Moldova, it was mainly associated with *eis* mutations. Isolates in the current cohort who originated in India or South Africa were not resistant to kanamycin, or were not tested for kanamycin. The only Moldovan isolate did have an *rrs* mutation but was not phenotypically tested for kanamycin. Perhaps these variations are to do with empirical drug regimen and consequent selection pressure in those regions.

Another study analysed rifampicin (*rpoB*), isoniazid (*katG*, *mabA-inhA* promoter region), and fluoroquinolones (*gyrA*) in MDR/XDR-TB isolates in Belarus, China, Iran/Iraq, Honduras, Romania and Uganda (n=117) [194]. Researchers did find a large regional difference in mutations here, seeing a narrower set of mutations and more fluoroquinolone resistance in the higher TB prevalence countries. The main *rpoB* mutation found in the current cohort (S450L, or S531L *E. coli* numbering) was also found across all sites in the multi-national study, followed by H445Y (or H526Y *E. coli* numbering) which was mainly found in Romanian isolates in the multi-national study. Similar to the current study, *katG* S315T was found in 71% of cases. They found that combination of the mutations *katG* S315T and *inhA* C-15T were most commonly seen in Romanian isolates. This combination was found in our study in 7 isolates, but was not found in the one Romanian isolate. It was, instead, found in isolates from South America, Russia, Latvia, Moldova and Georgia (all high TB prevalence regions). The one isolate from Romania had two uncommon mutations. Perhaps variation and combination of isoniazid mutations is characteristic of this geographical region, or more common among former Soviet Union states and their neighbouring states.

A study on MTBC in Ireland in 1987 asked if tuberculosis drug resistance was a problem [186]. Even though there has been confirmed evidence that cross-border clusters of MTBC genotypes have reached Ireland, there has been no confirmed transmission of MDR/XDR-TB to the native population. There has been transmission within immigrant families, and possibly between individuals living together who were from the same geographical area, but no transmission to an Irish individual, although reactivation of LTBI could prove otherwise in years to come. The only possible transmission events that could have occurred seem to be from Irish to non-Irish patients, which may be a consequence of non-compliance or incorrect treatment regimen, proving the hypothesis of the previous Irish study, that iatrogenic resistance could pose a greater threat to MDR-TB control and prevention in Ireland than the import of MDR-TB from areas of high prevalence [185]. Treatment failure could be due to decreased medical knowledge around TB treatment due to its low prevalence, penetration failure of the particular drug regimen, incomplete course of treatment due to side-effects or non-compliance, or failure to identify the need to place the patient on directly observed therapy (or the lack of resources to do so).

Drug regimens, and non-compliance with treatment, as with many other types of bacteria, have been part of the reason that drug resistance has developed, due to selection pressure when certain regimens are used empirically. IEMDR03 is an example of micro-evolution from susceptible to drug-resistant over 15 months. Although it does not look like most of the MDR/XDR-TB isolates developed their resistance within Ireland, it is worth bearing in mind for the Irish cohort. Micro-evolution within an already-circulating strain, like those of the Haarlem sub-lineage, could be a more significant threat to TB control in Ireland than import of MDR-TB from areas of high

prevalence. Even IEMDR27, who was not an Irish patient, developed XDR-TB while on what experts deemed to be the correct regimen. Each sequential isolate study has shown that micro-evolution can, and has, occurred *in vivo*; the first evidence for this in Ireland. However, from the study results, it can be seen that the Irish-born MDR-TB cases do not differ significantly from the non-Irish born cases in the drug-resistance-associated mutations that they harbour.

It is imperative to get the balance right between over-estimating and under-estimating drug resistance. The cost (monetary, time, and health) of toxic second- and third-line TB drug regimens is high. Phenotypic culture is tried and tested and still has a place in TB diagnostics. New drugs such as bedaquiline and delamanid have come to the market in recent years. Phenotypic DST is now becoming available for these, which will add to the knowledge base around resistance. Even though culture will remain at the heart of TB diagnostics at least for the near future, WGS could decrease the amount of time that laboratory scientists are exposed to high-risk MDR/XDR-TB isolates. For example, DST may not have to be repeated once a resistant isolate is detected since WGS performed on the heat-inactivated culture could confirm this instead.

5.3.3 Draft genome of the first XDR in Ireland

This was the first XDR-TB isolated in Ireland. The first case of MDR-TB was not documented previously in the literature. The WGS confirms the presence of XDR-TB through the presence of high-confidence mutations for isoniazid, rifampicin, fluoroquinolones and aminoglycosides. At the time of testing, there was an intra-laboratory discrepancy for clarithromycin and ciprofloxacin. This could have been because the MIC of the isolate was close to the breakpoint of the antibiotic, or due to two different DST methods used. The MIC is more useful than a breakpoint method, especially in the case of emerging resistance to a particular drug. However, it is also more time-consuming and labour-intensive to perform. Theoretically, had WGS been in use at the time, the resistance profile would have been available within one week. Furthermore, there would not have been any discrepancy regarding clarithromycin or ciprofloxacin.

5.3.4 Sequential isolates, progression to MDR-TB, progression to XDR-TB, re-infection

The results from these three studies show that micro-evolution of MTBC has occurred within Ireland. Two of the patients were of Irish origin.

One Irish patient (IEMDR01EARLY and IEMDR01) had drug-resistance-associated mutations for rifampicin and isoniazid to begin with. However ethambutol mutations converted from mixed *embB* M306I and M306V to M306V only over time. These mutations have been associated with both ethambutol susceptible and resistant strains ('flip-flop' phenomenon). Perhaps as the isolate

reaches an MIC close to the drug's critical concentration, the polymorphism first occurs to change methionine to isoleucine, followed by a further change to valine in order to lead to complete ethambutol resistance. This patient was anecdotally associated with non-compliance. Perhaps selection pressure of varying drug concentration encouraged the genetic change. There seemed to be high variation between the strains. This was the same strain that could have been involved in transmission with IEMDR41. IEMDR41 shares the same drug-resistance-associated mutations as IEMDR01, except for the above ethambutol mutations. This could be seen as an indication of the direction of transmission, i.e. that the strain acquires resistance as time passes and that therefore the non-Irish-born patient transmitted to the Irish-born. This hypothesis is strengthened by the fact that the strain was designated Ghana lineage, associated with the Ivory Coast geographical region. However epidemiological evidence suggests that the Irish-born patient presented first which would refute that hypothesis.

The second Irish patient (IEMDR03) was also associated with non-compliance. The pattern was clearer here, from susceptible to isoniazid resistant after 9 months, to MDR-TB 6 months later. Interestingly, this patient grouped very closely with another patient (non-Irish healthcare worker, presented in 2005). They were of the same lineage, EAI, however they had different MtbC15-9 genotypes. With this lineage, i.e. a lineage that is more commonly found in Asia, one might assume that the non-Irish patient transmitted to the Irish patient, however the Irish patient was, once again, the first to present. It is possible they were both infected by a common source, but it is also possible that the Irish patient attended a healthcare facility over the course of his illness where he came into contact with the Indian healthcare worker and transmitted the MDR-TB he had developed *in vivo* to her. The main evidence to refute their contact and transmission is the lack of epidemiological links and the mutations their genomes harbour. IEMDR05 contains *katG* S315T and other mutations that make it a candidate for pre-XDR-TB. It is therefore more likely that this patient reactivated TB she contracted in her country of origin, which has a very high MDR/XDR-TB burden. Their genomes are similar on a mathematically computed tree. This tree must always be interpreted with caution.

The last sequential case concerned the development of XDR-TB during treatment in Ireland. This patient was hospitalised, so non-compliance could not have been an issue. It just takes one mutation to rule out an entire class of TB drugs, which are already scarce to come by. Clinicians are faced with even more difficult decisions in these cases, where it is a matter of weighing up the side-effects and toxicity of the drugs against the efficacy of the treatment.

One Lithuanian patient had contracted MDR-TB twice over the course of 3 years. The first was a Beijing strain (2004) and the second a LAM strain (2007). Fourteen out of twenty-four MIRU-VNTR loci were distinct and their MtbC15-9 codes differed significantly (100-32 and 121-52,

respectively). The earlier isolate was susceptible to ethambutol, pyrazinamide, and clarithromycin, while the later isolate was resistant. The mutations they harbour are different for fluoroquinolones, ethambutol, pyrazinamide and rifampicin. When their whole genomes were analysed, they were at least 48 SNVs apart, which would also indicate that they were un-related. This patient had visited his home country within this timeframe and it is hypothesised that he contracted both strains in that region, which is known to have a high MDR-TB burden [1]. The two isolates do not seem to have co-existed or transferred resistance from one to the other. Nevertheless, it must be considered a possibility, when analysing MTBC, that there could be mixed infection, co-infection and/or hetero-resistance at play.

If sequential isolates were found to develop resistance in a particular pattern, results could be used to monitor therapeutic response to drug regimens, and if, or when, to change those regimens. Future studies should focus on the genomic micro-evolution of these sequential isolates in order to find out how their MTBC genomes change over time. Future work should focus on sequential isolates and the beneficial information they can provide.

5.3.5 Drug Resistance Prediction using rapid genotypic tools (LPAs and WGS) compared to phenotypic DST

Overall PhyResSe and TB Profiler performed with the greatest sensitivity. However it is also true that TB Profiler had the largest amount of false positives. False positives could be viewed as very major errors since they could have a significant clinical impact on the drug regimen available to the patient if the mutation rules out using an essential drug incorrectly. False negatives could be seen as major errors since the method is not correctly identifying resistance in the MTBC strain, however, in combination with other drugs in the regimen, this drug may still be somewhat effective. When the Walker/Kohl *et al* algorithm was originally validated, it performed with sensitivity of 81.6% and specificity of 98% compared to the LPAs tested in that study that performed with sensitivity of 81.6% and specificity of 98.2% [61]. In the current study, the Walker/Kohl *et al* algorithm did not perform as well (sensitivity 75%, specificity 92%) but remained almost equivalent to the LPAs tested (sensitivity 78%, specificity 98%). While the numbers in this cohort were low, it can be clearly seen that WGS performed at least as well as the rapid molecular tests currently available. LPAs are excellent tools. However, it is the added value that WGS can bring that is at stake here. WGS will not only tell us about the rifampicin and isoniazid mutations that are well-characterised, it will also tell us the emerging novel mutations that might be present, and it will provide that information for any amount of genes required. Sixty-nine genes were surveyed in all in this study. Also, the data is not discarded following the test. A database of genomes can be built up over time, which can be accessed for any reason in the future. If new mutations are found and published, these genomes can be re-analysed with this information

in mind. This is not to mention the further value that WGS can offer, as can be seen in chapters 4 and 6, which could include species identification and outbreak analysis.

Since there was no commercially-available software that looks for candidate genes of resistance in *M. tuberculosis*, parts of this sequencing analysis were done manually, one isolate at a time, using separate software programs, custom macros, and manual searching. Ideally, a script or program could be designed which would analyse all candidate genes for resistance in a short period of time. Many genomes could be analysed in parallel. This is the type of NGS workflow that was developed by the MMM and COMPASS-TB Study Groups [61, 112, 114]. That work will be discussed further in chapter 6. The other promising solutions are the online tools PhyResSe and TB Profiler [76, 77]. These workflows have not been robustly tested in clinical trials to date, but would need to be in order to satisfy the needs of clinical users and their patients.

The major impediment to these workflows would be that, as new information emerges, the parameters would have to change. This would have to be managed carefully and could be costly. For the diagnostic laboratory, this could be a significant limitation. Conversely, it would be a huge advantage to the diagnostic laboratory if an application could be regularly updated with the most recent, cutting-edge information. Currently, updates have to be incorporated regularly into databases for Minimum Inhibitory Concentration (MIC) testing, and (Matrix-assisted Laser Desorption/ionisation – time of flight) MALDI-TOF analysis, in the bacteriology diagnostic laboratory. These updates, although sometimes inconvenient, are a part of the working laboratory, just as software updates are for personal computers or mobile devices. Currently, however, the bottleneck that exists for post-sequencing analysis remains a challenge.

ReseqTB data-sharing platform seeks to consolidate the known high-confidence mutations across the entire TB research community, bringing together collaborators such as the Bill and Melinda Gates Foundation (BMGF), Critical Path Institute (C-Path), Foundation for Innovative and New Diagnostics (FIND), WHO, New Diagnostics Working Group (NDWG) and others. Collaborators would be able to access a ‘one-stop source of curated, aggregated, clinically relevant genetic and associated meta-data for global MTBC strains’ [121].

Sharing of relevant resistance data, both genotypic and phenotypic, as well as patient demographics and health outcome data, is the key factor in developing a genotypic solution for MTBC whole genome resistance detection. This solution could not only benefit high-income countries, but also the developing world. The most common resistance determinants have already been formulated into one simple point-of-care assay; the Cepheid GeneXpert MTB/RIF, which can detect MTBC, and rifampicin resistance, in under two hours. The rapid molecular tests discussed in this chapter also utilise informative mutations discovered through research (Hain GenoType MTBDR*plus* and

MTBDRs/), and they are upgrading as new research emerges. However, these tests are expensive. Point-of-care technology is rapidly changing in order to meet developing world needs and costs. Once the most informative mutations are elucidated, a quick and simple test could be developed around their detection.

5.3.6 Comparison of online web-tools for drug resistance prediction in MTBC

Both TB Profiler and PhyResSe successfully identified the isolates from raw fastq data, and provided drug resistance profiles and genotyping for each isolate. PhyResSe will also call the sub-lineage based on a set of phylogenetic SNVs [219]. TB Profiler will call the sub-lineage in some cases, but also reports on the spoligotype and Region of Difference (RoD) results. There have always been slight differences between methods for sub-lineage-calling, therefore it is not surprising that there were minor discrepancies.

While it is impressive that these tools can extract so much lineage information from whole genomes, they seem to be relying on historical nomenclature for the moment. If there is a move towards some form of WGS genotyping, a new nomenclature may be needed which might need to incorporate more sub-lineages based on consensus sets of lineage-related SNVs [220].

Between PhyResSe and TB Profiler, while both had 100% specificity for isoniazid and rifampicin, TB Profiler had higher sensitivity for isoniazid (100% vs. 94.7%), while PhyResSe sensitivity to rifampicin was higher (100% vs. 97.4%). TB Profiler performed the analysis in less time than PhyResSe (10 min per isolate compared to over 24 hours for 42 isolates respectively), however PhyResSe performed significant amounts of quality control, which takes longer to complete. TB Profiler allows download of a drug resistance profile and lineage report. PhyResSe allows download of data at every step in numerous file formats.

Both TB Profiler and PhyResSe are excellent online web-tools that allow users to input raw fastq data for rapid and reliable drug resistance profiling. TB Profiler may be more suited to settings where there is less bio-informatics expertise, whereas PhyResSe allows more computational transparency. These web-tools are for 'research use only' at the moment, but with more controlled studies of their specificity and sensitivity, and clinical trials, they could be the tools of the future.

5.3.7 Limitations of the Study

There were some limitations to the study. One challenging issue was collection of patient demographics and details under strict data protection and confidentiality conditions. It was difficult to know when the MDR/XDR-TB patients arrived in Ireland, and even more difficult to find out if,

and/or when, they had visited their home country, or for how long they stayed. Other information that could have been useful would be whether the patient had addiction issues, contact information, or previous history of TB. While this information was collected for some patients, it was not for others.

Phenotypic DST is the reference standard. This is a limitation in itself since it is not always 100% reliable, especially for drugs like ethambutol and pyrazinamide [168, 221]. There were also many gaps in the data for second- and third-line susceptibilities (not available for 15/42 isolates), mainly due to the isolates coming from external hospitals, or the differences over time in drug testing policies in supra-national reference laboratories, which ranged from testing two drugs to eleven in the case of an XDR-TB case.

Ideally, isolates should all have been tested with micro-titre MIC in a controlled environment in order to accurately determine the correlation between phenotypic DST and genotypic drug resistance prediction. A large international alliance is currently studying this correlation as part of the CRYPTIC (Comprehensive Resistance Prediction for Tuberculosis: an international Consortium) Project.

5.3.8 Health Outcomes

The ideal health outcome is that the patient is treated at an earlier stage with the correct regimen for the TB they have contracted. Particularly in the case of MDR/XDR-TB, conventional phenotypic testing can take excessively long (up to 170 days for some of the cases in this cohort). This is generally due to slow growth of the organism, repeat testing to confirm results, and referral to supra-national reference laboratories where the isolate must be grown again in order to perform the required second- and third-line DST. Any method that could decrease this time-frame would be welcomed. WGS has the potential to change the diagnostic and public health microbiology laboratory, and these changes could have a marked impact on the management of tuberculosis [144, 222, 223]. Public Health England are planning to introduce this technology for mycobacterial reference services in the UK [224]. Although, the outputs from the current study may not have impacted on a real-time TB case, other studies have shown that WGS can make a difference to health outcomes. Pankhurst *et al* showed that WGS prediction of drug resistance and outbreak discovery prevented further transmission of a case of MDR-TB within the community from the time of discovery, before the phenotypic DST results were available [114]. This will be discussed further in chapter 6, which details the results of an international collaborative study that aimed to prove the usefulness of WGS analysis for TB diagnostics.

5.4 Conclusions

The molecular characterisation of MDR/XDR-TB strains in Ireland from 2001-14 has proven that these strains are not being readily transmitted within the Irish population, that the drug resistant strains are similar to those found circulating in Europe, and that despite their high diversity, their drug-resistance-associated mutations are largely similar even though they differ elsewhere in their genomes. This reflects the stability of the MTBC genome. WGS matched phenotypic DST in most cases, although discrepancies were found. While isoniazid, rifampicin and fluoroquinolone genotypic results were particularly reliable, others require more phenotypic-genotypic correlation evidence if they are to be used diagnostically. That being said, WGS could transform drug-resistance prediction for the diagnostic laboratory by dramatically reducing the time to detection of resistance which could impact on patient treatment and health outcomes in the future.

Chapter 6.

Prospective Pilot Study to Identify Mycobacteria, and Detect Drug Resistance and Nearest- Neighbour Relatedness in MTBC, using WGS of Early Positive Liquid Cultures

6 Prospective Pilot study to identify Mycobacteria, and Detect Drug Resistance and Nearest-Neighbour Relatedness of MTBC, using Whole Genome Sequencing of Early Positive Liquid Cultures

6.1 Introduction

For optimum recovery of mycobacteria, specimens are cultured for up to six weeks using a liquid-based culture-system [63]. Culture-positive isolates are identified using various molecular tests (2 days) and if MTBC is detected, susceptibility-testing is performed to first-line anti-tuberculous agents (up to 2 weeks) [64]. If MDR-TB is suspected, the empiric treatment regimen is changed and the isolate is subjected to testing with second- and third-line drugs, which include fluoroquinolones and aminoglycosides [107]. Figure 7 displays the algorithm currently in place in the IMRL for clinical specimens. Due to the fastidious nature of MTBC *in vitro*, results may not be available for many weeks, which can have knock-on effects for patient treatment and transmission of MDR/XDR-TB [211]. Rapid molecular technologies are becoming increasingly useful for bypassing these time-intensive processes and for guiding appropriate treatment regimens [144, 192, 193].

Genotypic molecular assays are already available for rapid detection and resistance prediction directly on specimens. Xpert[®] MTB/RIF (Cepheid[®], Sunnyvale, California, USA) detects MTBC and rifampicin resistance within 2 hours, and Hain GenoType MTBDR*plus* and MTBDR*sl* (Hain Lifescience, Nehren, Germany) detect MTBC and first- and second-line drug resistance respectively, within two days. While extremely useful, they are susceptible to PCR inhibitors, they do not detect all possible resistance, and, in practice, they supplement culture, rather than replacing it. As discussed in Chapter 5, WGS variant analysis has been shown to accurately predict phenotypic resistance in many bacteria, in particular TB [61, 76, 77, 225, 226].

MIRU-VNTR genotyping for TB surveillance and outbreak detection is performed prospectively in Europe, the United States and Canada. While this is the established first-line genotyping method of choice, it has been shown to over-cluster at times, and is time-consuming and labour-intensive to perform (approximately 4-5 days from extraction to analysis for 14 isolates) [135]. WGS has been shown to provide higher resolution for TB cluster analysis [142].

Modernising Medical Microbiology (MMM) is a group based in Oxford (Nuffield Department of Medicine, John Radcliffe Hospital), directed by Prof. Derrick Crook and Prof. Tim Peto, who have pioneered the use of WGS to delineate TB outbreaks and transmission in the UK [112, 149, 158]. The IMRL participated in an international collaboration led by this group. The aim of the pilot study was to prove that WGS, using Illumina[®] MiSeq[®] NGS, could be performed directly on newly-positive mycobacterial cultures, within a short time-frame, and that the sequencing data

could be used for species identification of all mycobacteria, followed by resistance prediction and nearest-neighbour relatedness analysis (using 2,191 reference isolates from the MMM database) whenever MTBC was identified. Figures 7 and 19 show the WGS workflow proposed by the collaborative group, which could consolidate conventional techniques into one rapid method that would allow numerous outputs to be measured simultaneously. This was the first prospective study to attempt to test the hypothesis that this rapid WGS workflow could be possible for the diagnostic laboratory, especially in reference laboratories in higher income settings. Collaboration in this study enabled a large international sample set that included Irish isolates to be used for validation of NGS for routine diagnostics in the IMRL.

6.2 Results

6.2.1 COMPASS-TB Study Group

The IMRL collected isolates for this international collaborative pilot study, with the Oxford MMM group at the hub, for the month of October 2013, and again from January - March 2014. Laboratories from France (Lille), Germany (Borstel), Canada (Vancouver), Ireland (IMRL) and various UK sites (Birmingham, Brighton, Oxford and Leeds) contributed to the study and collaborators formed the umbrella COMPASS-TB Study Group (Complete Pathogen Sequencing Solution). Appendix 3 includes the published article from this study.

6.2.2 Samples Submitted

The entire study recorded 356 isolates submitted, 23 of which were from the IMRL. In practice, thirty-six isolates from the IMRL were submitted, 35 of which were sequenced on four Illumina[®] MiSeq[®] 300-cycle runs. This comprised 23 individual patient isolates (IMRL3 and IMRL5 were from the same patient), 8 isolates from a test run (included for analysis here), 2 positive controls (H37Rv), one negative control, one isolate failed extraction (IMRL26) and one isolate was disregarded (IMRL18). IMRL26 whole genome extraction did not achieve sufficient input DNA concentration, and could not be repeated as the remaining culture was required by the diagnostic laboratory.

6.2.3 Sample and Illumina MiSeq Run Details

Table 28 describes the sample type, microscopy results, time-to-positivity, GC content, and how the sequencing run performed for each isolate (i.e. MiSeq[®] cluster density, number of reads generated, percentage of reference genome covered). Original sample types were either respiratory, urinary or from lymph node biopsies. DNA extract concentrations ranged from 0.07-21.7 ng/μl, library preparation concentrations from 1.3-54 nM, cluster densities from 302 to 1520 K/mm², and reads for mapping from 314,586-10,378,278. For MTBC, 70.4 - 92.5% of the reference genome was covered. GC content for mycobacteria should be approximately 65%. This was maintained in most cases (47-69%).

6.2.4 Identification using MMM WGS workflow compared to Conventional Methods

Twenty six isolates from the IMRL (26/31, 84%, 95% CI 71-97) were identified correctly by the NGS pipeline (see Table 29). WGS successfully identified 93% (CI 90-96) of species.

Table 28. Results of COMPASS-TB MGIT Pilot Study: IMRL isolates

23 isolates were included in the final study. Sample type and auramine positivity, Illumina[®] Miseq[®] 300 cycle run performance and nearest-neighbor relatedness are included. ATYP - atypical ZN morphology - probable non-tuberculous mycobacterium, TYP - typical ZN morphology - probable MTBC, TTP time to positivity, BAL - Broncho-alveolar lavage, RLL - right lower lobe, RUL - right upper lobe, RMB - right main bronchus, CSF - cerebrospinal fluid, CF - cystic fibrosis, BX - biopsy

Sample	Sample Type	Microscopy (P/N)	TTP (days)	ZN morphology	Miseq Cluster Density (K/mm ²)	GC content	No. of Reads	% TB Ref genome covered	Origin of closest neighbour	Pairwise Nearest Neighbour genomic distance (SNP)
TEST 1	sputum	N	15	TYP	1311	0.50	9000518	77.8	Database	190
TEST 2	sputum	N	13	ATYP	1311	0.47	9006662	4.6	N/A	N/A
TEST 3	BAL lingua	N	9	ATYP	1311	0.53	7981992	10.4	N/A	N/A
TEST 4	BAL RUL	P	9	ATYP	1311	0.68	7620646	45.0	N/A	N/A
TEST 5	CSF	N/A	N/A	TYP	1311	0.43	774294	0.1	N/A	N/A
TEST 6	sputum	N/A	N/A	TYP	1311	0.63	868204	77.2	Database	117
TEST 7	sputum	N/A	N/A	TYP	1311	0.61	1017270	70.4	Database	98
TEST 8	pleural fluid	N/A	N/A	TYP	1311	0.63	3646936	92.2	Database	143
IMRL1	sputum	P	28	TYP	302	0.63	1038994	91.5	Oxford	173
IMRL2	BAL RMB	P	12	ATYP	302	0.69	945158	41.0	N/A	N/A
IMRL3	sputum	P	4	TYP	302	0.59	1654366	91.8	IMRL5 (Dublin)	0
IMRL4	sputum	P	7	TYP	302	0.60	711106	89.9	Birmingham	147
IMRL5	sputum	P	6	TYP	302	0.63	314586	78.8	IMRL3 (Dublin)	0
IMRL6	CF sputum	N	7	ATYP	302	0.58	790442	10.0	N/A	N/A
IMRL7	ileal biopsy	P	14	TYP	302	0.66	1066756	91.5	Vancouver	149
IMRL8	sputum	N	14	TYP	302	0.64	599010	88.7	Dublin	128
IMRL9	sputum	P	91	TYP	302	0.54	967066	89.3	Oxford	267
IMRL10	lymph node bx	N	18	TYP	302	0.59	1309348	91.2	Birmingham	192
IMRL11	sputum	unknown	12	TYP	302	0.62	737884	90.8	Birmingham	246
IMRL12	CF sputum	unknown	13	ATYP	1520	0.69	1136532	49.2	N/A	N/A
IMRL13	BAL RLL	N	17	ATYP	1520	0.66	7391174	40.6	N/A	N/A
IMRL14	sputum	P	7	TYP	1520	0.64	4100286	92.0	Birmingham	140
IMRL15	sputum	P	2	TYP	1520	0.60	5934172	91.8	Birmingham	84
IMRL16	tissue	N	17	TYP	1520	0.66	3020444	90.3	OxfordBCGCL1	2
IMRL17	sputum	P	5	TYP + ATYP	1520	0.63	2271564	91.6	Eastern Europe	157
IMRL18	disregard	disregard	N/A	disregard	1520	0.65	1804436	91.5	Birmingham	92
IMRL19	Reference Strain	N/A	N/A	TYP	1520	0.66	7512060	92.5	H37Rv ref strain	0
IMRL20	Negative	N	N/A	Negative	1520	N/A	10378278	N/A	Negative control	N/A
IMRL21	BAL	N	17	ATYP	1099	0.67	3379366	56.6	N/A	N/A
IMRL22	sputum	N	20	ATYP	1099	0.65	5291352	72.6	N/A	N/A
IMRL23	BAL	N	15	ATYP	1099	0.67	5409622	44.4	N/A	N/A
IMRL24	urine	P	11	TYP	1099	0.65	7022768	92.4	Birmingham	41
IMRL25	sputum	N	20	ATYP	1099	0.65	7812186	45.4	N/A	N/A
IMRL26	sputum	P	13	TYP	1099	N/A	N/A	N/A	N/A	N/A
IMRL27	BAL	P	8	ATYP	1099	0.68	4525708	66.2	N/A	N/A
IMRL28	Reference Strain	N/A	N/A	TYP	1099	0.65	2632584	92.4	IMRL19H37	0

One identification failed (Test5). One isolate (IMRL17) was identified as a mixture of *M. tuberculosis* and *M. avium* by traditional methods (Hain GenoType CM, clearly visible on ZN microscopy on Day 0). However, the NGS failed to identify the *M. avium*. No *M. avium* genes (0/33) were found in the isolate by the NGS pipeline. In order to investigate this discrepancy, Hain GenoType CM and MTBC were performed on the whole genome extract. Only *M. tuberculosis* was detected. The crude extract was repeated and ZN microscopy re-checked. Evidence of a mixed culture was confirmed. This compared to 1,562,589 reads that mapped to H37Rv with an average coverage of 56.6%. The patient also had clinical symptoms of, and risk factors for, *M. avium* infection. The patient did appear to clear this infection, however, and the next isolate received did not harbour a mixture. Resistance predictions were confirmed by whole genome sequencing this pure strain, although it was not part of the MGIT Pilot Study *per se*. Reads from both mixed and pure isolates were mapped in-house using Geneious R9 software to *M. avium* reference genome (NC_008595.1). 445,982 out of 2,271,564 (19.6%) reads from the pilot study isolate mapped with an average coverage of 12.3, covering 33.7% of the genome. 1,654,724 out of 7,450,798 (22.2%) reads from the subsequent pure isolate mapped with an average coverage of 45.7, covering 34.2% of the *M. avium* genome.

Conversely, two isolates (IMRL21 and IMRL27) were identified by the IMRL as *M. avium* only, and by the NGS pipeline as a mixture of *M. tuberculosis* and *M. avium*. On further investigation, using Hain GenoType CM performed on NGS extracts, *M. avium* only was detected, indicating that the isolates may have been contaminated at the DNA library preparation stage. However, internal quality control passed for extraction and library preparation for all four MiSeq[®] runs. DNA library preparation was repeated and the NGS pipeline identified *M. avium* only, therefore these results were resolved for the purposes of the study.

One isolate (IMRL22) was identified using Hain GenoType CM LPA as *M. intracellulare* and by the NGS pipeline as a mixture of *M. tuberculosis* and *M. chimaera* (*M. avium* Complex). On further investigation, the isolate was not contaminated at the library preparation stage, but may have been contaminated at the DNA extraction stage. As the remaining culture was required by the diagnostic laboratory, it was not possible to re-extract the DNA. Hain GenoType CM was performed on the NGS extract which resulted in the detection of *M. intracellulare* only. Hain GenoType MTBC was performed on the same extract, as it was hypothesised that the *M. intracellulare* DNA might have out-numbered any *M. tuberculosis* DNA that may have been present for the Hain GenoType CM assay (which should have detected both species, if present), and the result was a weak *M. tuberculosis* pattern. The patient was not suspected to have a mixed infection. Proof of this was found when Hain GenoType MTBC was performed on the subsequent isolate from this patient; only *M. intracellulare* was found using Hain GenoType CM. If IMRL21, 22 and 27 had not been

ISOLATE	Oxford ID	IMRL ID
TEST01	<i>M. tuberculosis</i> Complex (MTBC)	<i>M.tuberculosis</i>
TEST02	<i>M. avium</i> Complex	<i>M. avium</i>
TEST03	<i>M. fortuitum</i>	<i>M.fortuitum</i>
TEST04	<i>M. avium</i> Complex	<i>M. avium</i>
TEST05	Failed	<i>M.tuberculosis</i>
TEST06	<i>M. tuberculosis</i> Complex (MTBC)	<i>M.tuberculosis</i>
TEST07	<i>M. tuberculosis</i> Complex (MTBC)	<i>M.tuberculosis</i>
TEST08	<i>M. tuberculosis</i> Complex (MTBC)	<i>M.tuberculosis</i>
IMRL01	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL02	<i>M. avium</i> Complex	<i>M. avium</i>
IMRL03	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL04	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL05	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL06	<i>M. abscessus</i> Complex	<i>M. Abscessus</i>
IMRL07	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL08	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL09	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL10	<i>M. tuberculosis</i>	<i>M. africanum</i>
IMRL11	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL12	<i>M. avium</i> Complex	<i>M. avium</i>
IMRL13	<i>M. celatum</i>	<i>M. celatum</i>
IMRL14	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL15	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL16	<i>M. bovis</i> BCG	<i>M. bovis</i> BCG
IMRL17	<i>M. tuberculosis</i>	<i>M. tuberculosis</i> and <i>M. avium</i>
IMRL18	Disregarded	Disregarded
IMRL19	H37Rv positive control	H37Rv positive control
IMRL20	Negative control	Negative control
IMRL21	<i>M. avium</i> Complex/ <i>M. tuberculosis</i> Complex (repeat library prep = <i>M. avium</i> only, contaminated at library prep stage)	<i>M. avium</i>
IMRL22	<i>M. chimaera</i> (<i>M. avium</i> Complex)/ <i>M. tuberculosis</i> Complex (repeat library prep = mixture, contaminated at the extraction stage)	<i>M. intracellulare</i>
IMRL23	<i>M. avium</i> Complex	<i>M. avium</i>
IMRL24	<i>M. tuberculosis</i>	<i>M. tuberculosis</i>
IMRL25	<i>M. lentiflavum</i>	<i>M. lentiflavum</i>
IMRL26	Extraction failed	Extraction failed
IMRL27	<i>M. avium</i> Complex/ <i>M. tuberculosis</i> Complex (repeat library prep = <i>M. avium</i> only, contaminated at the library prep stage)	<i>M. avium</i>
IMRL28	H37Rv positive control	H37Rv positive control

Table 29. Table comparing identification of isolates using NGS (COMPASS-TB pipeline) and Hain GenoType LPAs CM, AS and MTBC for IMRL isolates

Discrepancies are highlighted in purple. LPAs were performed in the IMRL.

contaminated, the percentage identified correctly by the NGS pipeline would be 94% (CI 86-100), which correlates well with the study as a whole.

The distribution of IMRL mycobacterial species can be seen in Figure 41. A diverse range of species was uncovered over the course of the study. MTBC accounts for the largest proportion of isolates (55%, CI 38-73), followed by *M. avium* Complex (16%, CI 3-29), while the remainder were unique. Figure 42 displays the distribution of mycobacterial species for all sites (n = 356) [114]. MTBC also accounts for the largest proportion of isolates (47.3%, CI 43-53) within the larger group, followed by *M. avium* Complex (13.5%, CI 10-17).

6.2.5 Anti-TB Drug Resistance Prediction using MMM WGS workflow compared to Conventional DST

Fourteen IMRL isolates (from 13 individual patients) were identified as MTBC. In almost all cases, the phenotypic DST results matched NGS predictions for isoniazid, rifampicin, ethambutol, streptomycin and pyrazinamide (5 discrepancies out of a possible 120, 96% accuracy, CI 92-99) (see Table 30). Taking the entire study into account, the accuracy was 93% (44 discrepancies out of a possible 672, CI91-95).

More than five reads were required by the NGS pipeline in order to confirm a mutation at a particular location in the genome. For IMRL05, the coverage where the *rpoB* mutation was located was low. There were four C (cytosine) reads at *rpoB* reference position 445 (H37Rv wildtype), and one T (thymine) which would represent a resistance-associated SNV, resulting in an ambiguous read that could not be interpreted. Since four reads were wildtype and one a mutant, this most probably suggested a susceptible prediction, which would correlate with the DST. Furthermore, IMRL03 was the same patient (performed in error), and the coverage for *rpoB* was sufficient to call it susceptible.

Moxifloxacin DST was available for two isolates, both of which correlated with NGS. Six isolates (out of a total of 14) had low coverage and/or ambiguous reads in the *rrs* gene (associated with resistance to amikacin). Phenotypic amikacin DST results were only available for two IMRL isolates. One correlated with the NGS prediction but the other did not.

IMRL16 was an *M. bovis* BCG vaccine strain. Phenotypic DST results from the IMRL to date show that this strain is intrinsically resistant to low-level isoniazid and pyrazinamide (unpublished data). However, other sites with the same BCG vaccine strain do not report this low level isoniazid resistance (personal communication, Dr. Tim Walker). It does not seem to be caused by the classical *inhA* mutation. Mutations on the *ndh* gene have been implicated in mycobacterial cross-resistance to isoniazid and ethionamide [227]. This gene was investigated in IMRL16 and a

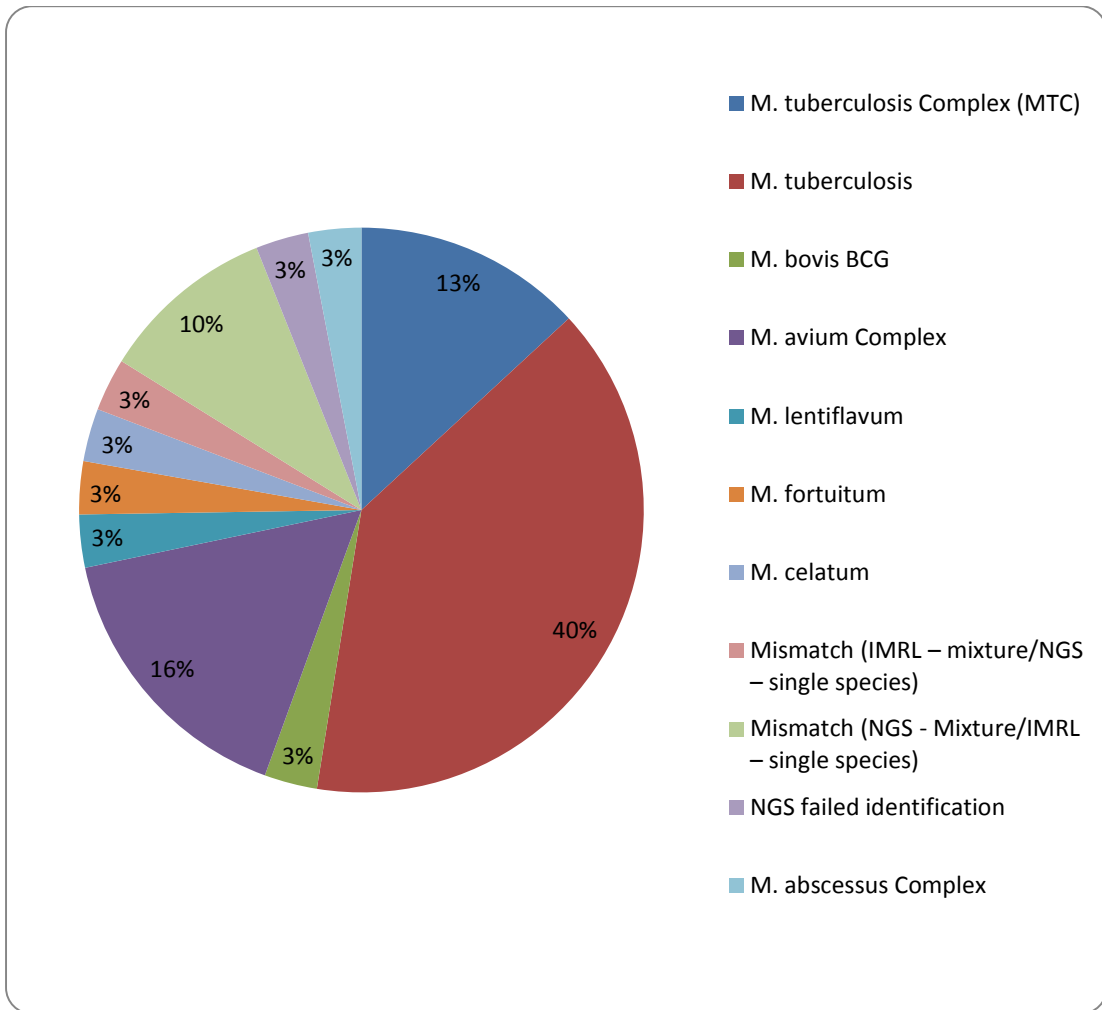


Figure 41. Distribution of mycobacterial species found over the course of the study and correlation of species identification (by NGS versus conventional methods) for IMRL isolates (n=36).

A diverse range of isolates was recovered over the course of the study, suggesting that the pilot was robustly tested. MTBC (which includes *M. tuberculosis*, *M. bovis*, *M. bovis BCG*, *M. microtii*, *M. africanum*) was the most commonly isolated organism (55% in total), followed by *M. avium* Complex (16%). Approximately 16% resulted in a partial mismatch or identification failure prior to investigation of discrepancies.

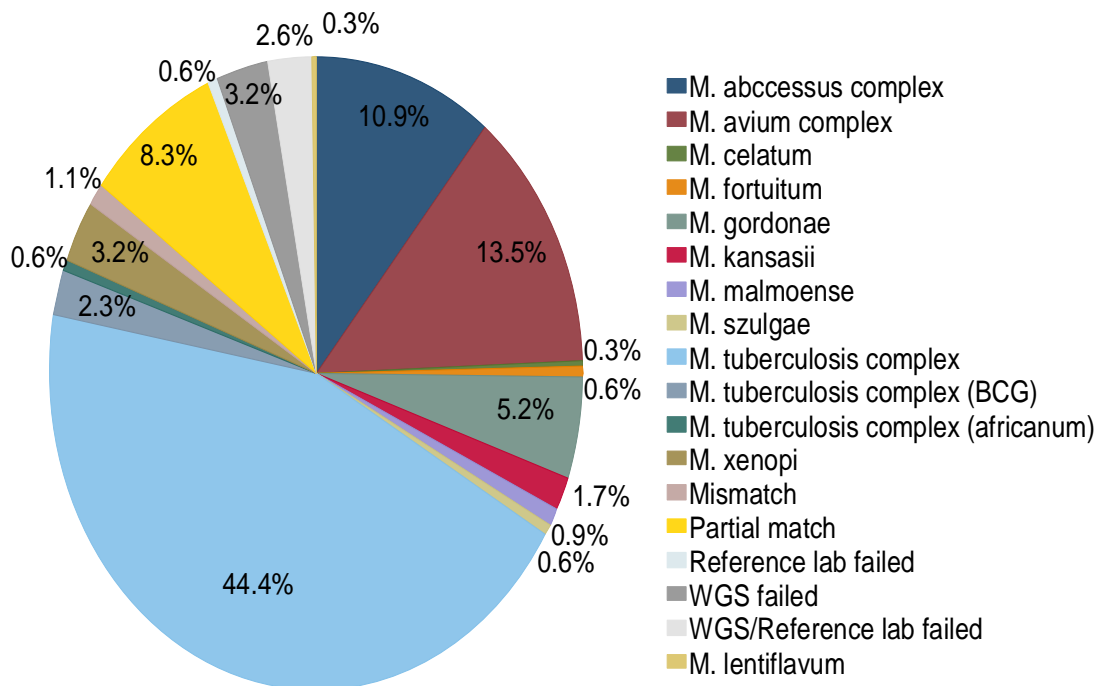


Figure 42. The bigger picture; Distribution of mycobacterial species found over the course of the study and correlation of species identification (by NGS versus conventional methods) for all participating sites (n=356).

This figure demonstrates that MTBC (which includes *M. tuberculosis*, *M. bovis*, *M. bovis BCG*, *M. microtii*, *M. africanum*) was the most commonly isolated organism in this study (47%). Approximately 10% resulted in a mismatch, partial mismatch, or failed identification, for various reasons, including DNA extraction failure and laboratory cross contamination.

Table 30. Table comparing Drug Resistance Prediction results achieved for *M. tuberculosis* Complex isolates using NGS (COMPASS-TB pipeline) vs. phenotypic DST

SMRL = Scottish Mycobacteria Reference Laboratory. S = susceptible, R = resistant. Susceptibility to moxifloxacin and amikacin is not currently tested in the IMRL. * IMRL03 and IMRL05 represent two isolates from the same patient. Discrepancies are highlighted.

mutation was found (*ndh* G313R, coverage 105, variant frequency 98.1%) which has been seen in *M. tuberculosis* isoniazid resistant and susceptible strains previously [196].

IMRL15 contains an ethambutol mutation which has been known to correlate with both resistant and susceptible phenotypes (a ‘flip flop’ effect) [61]. The isolate had been found phenotypically resistant, but was reported as susceptible by the IMRL following repeat testing that did not confirm the resistant result.

IMRL17, as mentioned above, was found to be a mixture of *M. avium* and *M. tuberculosis*, however the amount of *M. avium* present in the 1 ml that was taken for NGS, seems to have been below WGS detection limits. There was contamination visible across some of the genes (e.g. *rrs*). Table 30 highlights discrepancies in IMRL17 between the NGS pipeline and IMRL DST for ethambutol and pyrazinamide also. The NGS pipeline did not uncover any known SNVs associated with resistance to either antibiotic, and yet the isolates were phenotypically resistant. These DST results were confirmed in the IMRL, and also by an external reference laboratory. As mentioned above, a ‘true’ pure *M. tuberculosis* isolate was subsequently sequenced. Comparison of the two genomic resistance profiles yielded identical results, except for the presence of an *rrs* mutation (A1401G) that has been associated with phenotypic resistance to aminoglycosides, found in the pure isolate only. A resistance mutation (found in both mixed and pure isolates) that was not originally included in the resistance mutation catalogue (*embB* G406A), nor in the Hain mutation list, was subsequently believed to be involved in ethambutol resistance in this, and one other, isolate [199, 228, 229]. A *pncA* mutation (I31S) was found in both the mixed and pure isolates, which was not included in the mutation catalogue at the time, but is now considered to confer resistance to pyrazinamide in MTBC.

6.2.6 Nearest Neighbour Relatedness using MMM WGS workflow

Table 28 also includes the nearest-neighbour relatedness analysis performed using the MMM pipeline (‘Squirrel Walk’). Each isolate was compared to the bank of MTBC genomes collected to date by the MMM Group (2,191 isolates at the time of the study). An SNV difference of < 12 would suggest some relatedness, prompting further, more in-depth analysis. For instance, there is a 2 SNV difference between IMRL16 and the BCG vaccine strain sequenced by the pipeline. IMRL16 is, in fact, a BCG vaccine strain. H37Rv reference strain was identical (zero SNV difference) when sequenced a second time, confirmed by nearest-neighbour relatedness analysis. IMRL3 and IMRL5 were isolates from the same patient, sequenced in error. The nearest-neighbour relatedness analysis confirmed this (zero SNV difference). No other IMRL isolates had <12 SNVs distance from any strains in the MMM database, therefore no clusters were identified. However, the larger study did link 15 of 91 UK isolates to an outbreak.

6.2.7 Reporting of Results by the MMM WGS workflow

Figure 20 displays the report form envisaged by the COMPASS-TB Study Group. For example, the NGS report for IMRL15 contains the species identification (*M. tuberculosis*), its resistance profile (resistance to isoniazid, rifampicin, and ethambutol predicted) and associated mutations, and its relatedness to isolates in the MMM database using the Squirrel Walk algorithm (84 SNVs apart from a strain from Birmingham). The QC report suggests that GC content was 60% (H37Rv genome GC content is 65%). There were 5,934,172 reads available for mapping, 4,569,883 (77%) of which mapped to the TB genome, and the remainder of which was human DNA contamination (23%). Despite the contamination, 91.8% of the TB genome was covered. The SNVs detected were *katG* S315T, which is related to isoniazid resistance, *rpoB* S450L which is related to rifampicin resistance, and *embB* M306V which has been associated with both resistant and susceptible phenotypes, and therefore represents a ‘flip-flop’ mutation as mentioned previously. All mutations were checked manually using in-house Geneious R7 software. The mutations found by the MMM group matched those found with Geneious software for IMRL15. Hain GenoType MTBDR*plus* and MTBDR*sl* also correlated.

6.2.8 WGS Contamination with Human and Nasopharyngeal Flora DNA

Both human DNA and nasopharyngeal flora (NPF) DNA contamination were recorded during the Pilot Study. When IMRL isolates alone were examined (Figure 43-45), more human DNA contamination was present when the microscopy result was negative. There was no significant relationship between NPF DNA contamination and microscopy (see Figure 43). Sputa exhibited more NPF DNA contamination than the other sample types. Sample type did not seem to have a significant effect on the amount of human DNA present (see Figure 44). Time to positivity (TTP) did not seem to have a significant impact on either human DNA or NPF DNA contamination over time (see Figure 43). Contamination rates over the entire cohort of isolates were not made available for comparison. The study was published in 2016 [114].

6.2.9 Reporting Times and Costs associated with WGS compared to Conventional Methods

When the entire study was taken into account, median time from MGIT[®] positivity to DST report was 25 days (IQR 14-32). If MIRU-VNTR genotyping was included, i.e. final reports, this rose to 31 days (IQR 21-60). WGS final reports also took 31 days (IQR 21-44). This was believed to be due to batching delays, and delays in sharing sequencing data, and working on a 5-day week rather than the 7-day week in place in diagnostic laboratories. When the comparison was balanced out to take these issues into account (i.e. just timings recorded), reference laboratory DST reports were generated a median of 15 days (IQR 9-25) behind WGS reports (median 24 days [IQR20-33]

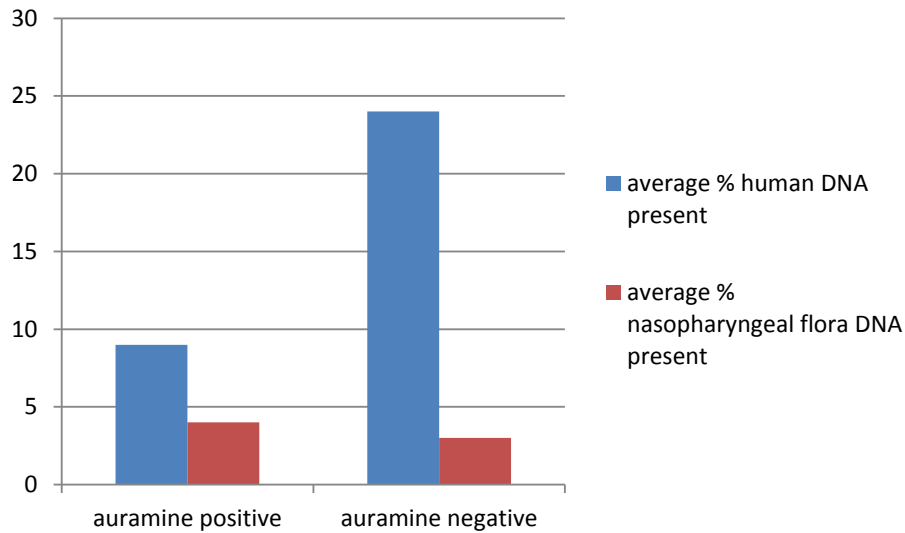


Figure 43. Percentage of DNA contamination (human and nasopharyngeal flora) present compared to auramine microscopy result in IMRL MGIT Study isolates

More human DNA was present when the auramine microscopy was negative. There was no significant difference in nasopharyngeal flora DNA levels in either positive or negative microscopy.

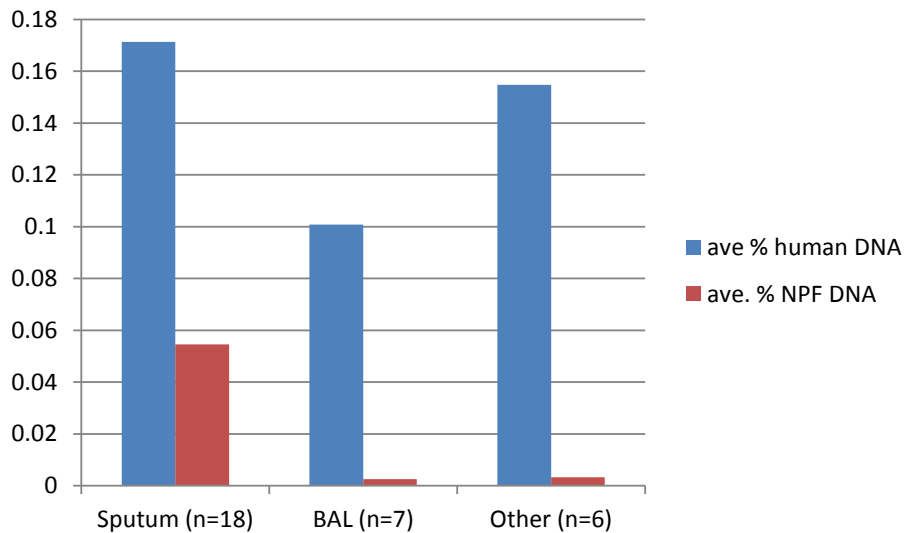


Figure 44. Percentage of DNA contamination (human and nasopharyngeal flora) present compared to Sample Type in IMRL MGIT Study isolates

Sputa exhibited more nasopharyngeal flora DNA contamination than the other sample type, however, human DNA contamination did not seem to differ according to sample type.

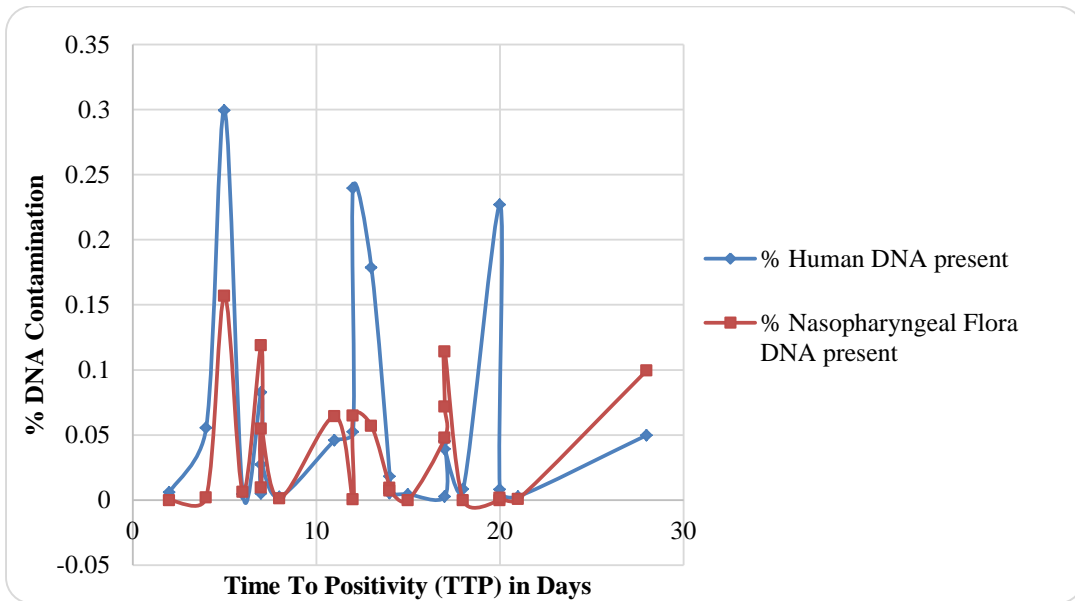


Figure 45. Percentage of DNA contamination (human and nasopharyngeal flora) present compared to Time to Positivity (TTP) in IMRL MGIT Study isolates

Time to positivity did not significantly impact on the amount of human or nasopharyngeal flora present over time.

versus 8 days [IQR 6-9]). Relatedness reporting was reported in 9 days (IQR 6-10) using WGS versus 32 days (IQR 22-42) for reference laboratory genotyping [114]. Once WGS data was shared, a full diagnostic report was issued in a median of 5 days (IQR 3-7).

Costs were measured by UK sites only. When overall costs were measured in a reference laboratory for culture, identification, DST, molecular assays and MIRU-VNTR genotyping, an annual saving of 7% was estimated [114].

Reads from this study were deposited in the National Center for Biotechnology Information (NCBI) Short Read Archive (Bioproject PRJNA268101 and PRJNA302362).

6.3 Discussion

The MMM Group and others are working on ways to analyse the whole genomes of bacteria, by creating a network of bioinformaticians and scientists who collaborate to design sequencing analysis workflows. Because the two disciplines are symbiotic to this process, the MMM Group is well-placed to successfully produce reliable analyses. Prior to this study, they had designed workflows for outbreak investigation using WGS which were discussed more thoroughly in Chapter 4 [112, 149, 158]. The MGIT pilot study design goes much further than cluster analysis, and includes workflows for species identification and resistance-associated gene variant analysis, and a nearest-neighbour relatedness algorithm (Squirrel Walk) that saves computational power and time, while still using an enhanced reference database of over 2,000 genomes to interrogate each new isolate. These custom workflows allow analysis on a large scale in a shorter time-frame. The Pilot Study has shown that WGS can be used diagnostically on an international scale. All diagnostic results for clinical and public health action could be simultaneously available in as little as 2 days from completing sequencing. Bioinformatics can be a significant bottle-neck when introducing WGS techniques to the diagnostic laboratory [230], as discussed in Chapter 5. The MMM Group have designed a workflow that could theoretically be used by any laboratory that can share data electronically. The workflow, from DNA extraction to WGS analysis was standardised and operating procedures and work-sheets made available for collaborators in a user-friendly format [113].

The WGS analysis performed very well considering the wide range of extraction concentrations, library preparation concentrations, and cluster densities with which it had to contend. Read depth had to be set quite low in order to cover the TB genome. For other WGS applications, a higher read depth would be desirable. These cultures were barely positive. They were not dissimilar to specimens. Specimen whole genome extraction has been challenging to date [231]. The MMM Group are currently working on whole genome extraction for clinical samples, which would make the workflow even more rapid by by-passing mycobacterial culture time (time-to-positivity ranged from 4 – 28 days, one isolate was a referred culture that took a long time to grow in the IMRL, 91 days from date of collection). DNA extraction is currently the rate-limiting step for WGS, since it can take longer to perform than sequencing in some cases.

Species identification using the gene presence/absence WGS algorithm preformed impressively. IMRL17 identification, however, constitutes a significant discrepancy, i.e. a mixture of *M. tuberculosis* and *M. avium* was found by the IMRL, and not by the WGS workflow. Re-testing using traditional methods supported the WGS identification. As the WGS extract was subsequently found not to have *M. avium* present, the nature and homogeneity of the sample comes into question. The portion that was used to prepare the ZN stain, on the same day the WGS aliquot was taken (Day 0), was clearly mixed. It is possible that the 1 ml portion removed for sequencing

contained *M. tuberculosis* DNA which overwhelmed the numbers of *M. avium* present. The patient was treated and the *M. avium* infection was resolved rapidly. The sample was taken at a very early stage of infection and may not have had sufficient numbers to be detected. This also happened with two other mixed isolates in the study, however the WGS analysis found only *M. avium* in these cases. In one case, the patient had confirmed co-infection but the sequencing quality was very poor (88% human DNA contamination), and in the other case the WGS could not be repeated. Not finding mixtures, even when present in small numbers, could be a limitation of the technology. In order to investigate further, a second isolate from the same patient, which did not harbour *M. avium*, was whole genome sequenced, and results compared. Both sequences were mapped to *M. avium* in-house and both mapped similarly (< 25% reference covered). This, and the fact that the non-tuberculous mycobacteria in the study mapped to H37Rv within a range of 4.6-49.2%, highlights the fact that this technology is mathematical, and not intuitive. It works on a series of algorithms. Occasionally, short reads will align to the incorrect part of the genome, without noting an error. Sequencing longer reads could circumvent this.

When anti-TB drug resistance prediction was analysed against standard DST, WGS performed with 93% accuracy across the entire study group (n=168 MTBC isolates, 628 out of 672 correct predictions). Some isolates were duplicates, so after de-duplication, 467 out of 508 predictions correlated. Another 434 predictions could have been made had phenotypic DST been available. If an up-dated catalogue were to be used today, just a year later, for example version 27 PhyResSe catalogue, many more prediction would likely correlate.

Some discrepancies were found. IMRL16 was a BCG vaccine strain that was just 2 SNVs apart from the UK vaccine strain in the database. Phenotypic DST indicated that IMRL16 was low-level isoniazid and pyrazinamide resistant (breakpoint = 0.1 µg/ml); all vaccine strains found to date in the IMRL have displayed this phenotype (unpublished data). A known pyrazinamide resistance mutation (*pncA* H57D) was found using WGS. However, the whole genome did not display a mutation in *inhA* or the *fabG1* promoter region, which would usually be present in low-level isoniazid resistant *M. tuberculosis*, nor do laboratories in the UK find this phenotypic resistance in their vaccine strains (personal communication). The *ndh* gene was also investigated with no conclusive mutation found. This warranted investigation. The live attenuated BCG vaccine was developed in the 1920s at the Pasteur Institute in Lille, France, and is the only TB vaccine currently available. More recently, BCG has been shown to be an effective immunotherapy for bladder cancers. Post-vaccination cutaneous rashes and disseminated infections are rare, but possible, complications, especially in immune-compromised patients [232, 233]. Different vaccine strains are in use around the world to immunise children. The vaccine used in Ireland (Danish strain 1331) has been in use since 2001. Prior to that, Copenhagen 1077 strain was used. The UK also uses Danish strain 1331, so they should at least have had the same phenotypic DST profile, particularly

since their genomes were so closely related. It was hypothesised that mutations may have been picked up through numerous sub-culturing steps during vaccine production originally. However, due to ‘seed lot’ preparations which have been available since the 1960s, the sequences and susceptibilities within each BCG strain should not differ significantly.

A literature search revealed that different strains of BCG have differing susceptibility profiles to anti-tuberculous drugs [234]. BCG vaccine strains can also differ slightly in their MIRU-VNTR genotypes [235]. The Danish strain has documented low-level isoniazid resistance at a breakpoint of 0.1 mg/l, as has the Connaught strain [234]. However, MICs of 0.2 and 0.4 mg/l were found in different susceptibility studies [233, 236]. This indicated that different DST methods could potentially call this resistant or susceptible. The IMRL would err on the side of caution and call this low-level resistant. A concern was raised following this investigation, that this strain may be used in BCG immunotherapy for bladder cancer. Arend *et al* argue that the Danish strain is no more than a ‘sheep in wolf’s clothing’, since it is no more virulent than the susceptible strains [237]. However, inoculation of a patient with any resistant organism, even though attenuated, could be considered unethical, due to the possibility of disseminated infection with that resistant organism, especially in an immuno-compromised patient [232, 233]. Further investigation confirmed that the Pasteur 1173P2 strain is used for bladder cancer immunotherapy in Ireland, which has been found to have an MIC of 0.1-0.2 in two different studies, i.e. less than half the Danish strain [233, 236].

Ethambutol DST has been found to be problematic due to its bacteriostatic nature [175]. The SNV M306V found in *embB* in IMRL15 is a commonly-found mutation, associated with ethambutol resistance, which has been known to exhibit a ‘flip flop’ phenomenon, i.e. could be found resistant or susceptible [228, 238]. DST was performed in the IMRL a total of four times. First, the ethambutol exhibited a resistant phenotype. It was resistant on repeat, but contaminated with other bacteria. The third repeat was susceptible, as was the fourth. The IMRL reported the ethambutol as susceptible since the resistant result failed to be confirmed. This mutation could be a good indicator of emerging resistance to ethambutol, i.e. the ‘flip flop’ is possibly due to the MIC being close to the breakpoint for ethambutol (5µg/ml). A small in-house study on ethambutol was carried out in the IMRL (not included here). As part of this study, IMRL15 was tested with ethambutol for a prolonged period of time. It did indeed exhibit resistance to ethambutol at the prolonged time-period. Hain Genotype MTBDR_{sl} assay also subsequently confirmed the presence of the *embB* M306V mutation. If WGS was used in the diagnostic laboratory, especially in the case of MDR-TB like IMRL15, the clinicians could have been alerted to this mutation earlier, and ethambutol use may have been avoided altogether, or at least used with caution. Ethambutol was included in the patient’s drug regimen in this case, and the outcome was successful despite the ‘discrepant’ result. Perhaps the *in vitro* breakpoint does not reflect true resistance *in vivo*, or the combination of anti-

tuberculous drugs was successful, rather than the individual drugs. When a drug is bacteriostatic, perhaps it can still be of some use even if the MIC is increased.

Discrepancies found when analysing the resistance profile of IMRL17 (Table 30) yielded more questions. The NGS pipeline analysed candidate genes known to be involved in resistance. No mutations were found to ethambutol, pyrazinamide or amikacin. It was possible that novel mutations were at play in this case, or that the DST was incorrect, although they had been confirmed in the IMRL and an external reference laboratory. A potential problem with the reliability of the results in the case of a mixture between *M. tuberculosis* and *M. avium* is that the same genes may be present in both strains. For instance, all four genes that were found to contain drug-resistance-associated mutations in this mixture (*katG*, *rpoB*, *rpsL*, *gyrA*) are found in both *M. tuberculosis* and *M. avium*. Hain GenoType *MTBDRplus* and *MTBDRsl* were also performed, and assays correlated with the WGS results, and each other, except for one difference. A mutation in the *rrs* gene (A1401G) was discovered in the subsequently sequenced pure isolate, but not in the mixed. The WGS confirmed this result. Mutations in this gene are associated with resistance to aminoglycosides [239]. *M. avium* also contains this gene. Perhaps some *M. avium rrs* gene reads were present in sufficient amounts to mask the *M. tuberculosis rrs* reads, and produce false negative results. This may explain the aminoglycoside discrepancy, but did not provide an explanation for the ethambutol and pyrazinamide. A resistance mutation in *embB* G406A was subsequently believed to be involved in ethambutol resistance in this, and one other, isolate [199, 228, 229].

Pyrazinamide was included in the study, but compared to the many mutations that could be present in the *pncA* gene (Rv2043c), only 6 mutations were found with sufficient confidence to include in the catalogue, originally designed by Feuerriegel *et al* to create the PhyResSe web-tool for TB genomic analysis, discussed in Chapter 5 [76]. Since then, the PhyResSe catalogue has grown significantly (now version 27). On interrogation of the newest version, a *pncA* mutation (I31S) was found in both the mixed and pure isolates, which is considered to confer resistance to pyrazinamide in MTBC. These two discrepancies highlight the fact that no core resistance mutation catalogue has been agreed upon worldwide to date. Even when this may be defined, it will have to be thoroughly clinically tested and robust. It will also be necessary to update it regularly with any new information that comes to light. This is a limitation for a diagnostic laboratory that may have to purchase new software updates regularly at a significant financial cost. Another software limitation could be storage of large amounts of isolate data and security of cloud-computing environments. While it may seem reasonable to a research scientist to transfer anonymised data electronically for pipeline analysis, it may be more challenging to convince hospital IT departments, and those responsible for data protection, that this is a feasible way to use ‘patient’ data.

Nearest neighbour relatedness was analysed using WGS in order to be able to predict, track, and break transmission chains of TB outbreaks in a timely and reliable manner. Of the IMRL strains submitted, except for IMRL3 and IMRL5 (same patient), the BCG strain (IMRL16) and H37Rv, pairwise nearest-neighbour was >41 SNVs (range 41-267, table 28), which indicates that no IMRL isolates were found to be related to any of the 2,191 reference strains in the database. MIRU-VNTR genotype clusters from a previous MMM study (n=264, [112], supplementary data) were analysed against genotypes present in Ireland from January 2010 - February 2013, and only one cluster was identified (Chapters 3 and 4), so this result underpins the above indication that there does not seem to be any significant inter-country transmission from the UK to Ireland, or vice versa. Within the UK study cohort, 15 out of 91 isolates were associated with 9 clusters in the database. Eight of the 9 clusters were already familiar to public health colleagues, and confirmed by MIRU-VNTR genotyping. WGS diagnosed two patients as MDR-TB and alerted Public Health to the ninth cluster found. The first patient's isolate was sequenced and analysed within a week after culture-positivity, and therefore pre-empted the phenotypic DST and genotyping. The laboratory performed Hain GenoType MTBDR*plus* and MTBDR*sl* urgently and the patient was commenced on second-line treatment more quickly than it would have been possible had WGS not been performed. This patient, following diagnosis, was on first-line therapy in the community, awaiting DST results. MDR-TB could have been transmitted within this time. Rapid WGS has the potential to decrease TB burden by breaking transmission chains.

A limitation to WGS analysis was low read-depth associated with contamination by non-mycobacterial DNA. For instance, Test 5 identification failed. This was due to low coverage of the reference genome, most likely affected by the amount of human DNA contamination present (61%). Following the 'Test' runs performed at each site, it was decided to change the DNA extraction protocol to include a saline wash prior to ethanol precipitation. The saline seemed to make a significant difference to the amount of human DNA contamination present. If the analysis maps to just reference genomes of interest, the contaminants can be filtered out. However, the less contamination, the more reads of interest obtained, the greater the coverage and quality of the sequence.

More Human DNA contamination was present when the smear microscopy result was negative. This is not surprising considering the microscopy is a measure of the mycobacterial load in the sample. If the microscopy is negative, then the proportion of mycobacteria will probably be less compared to the proportion of human DNA present. There was no significant relationship between NPF DNA contamination and smear microscopy.

Sputa exhibited more NPF DNA contamination than the other sample types. This is probably due to the nature of the sample. Sample type does not seem to have a significant effect on the amount of

human DNA present. Time to positivity (TTP) does not seem to have an effect on the amount of human DNA or NPF DNA contamination present. In the case of human DNA contamination, this is somewhat surprising since one would expect the proportion of human DNA to decrease in relation to the amount of mycobacterial DNA as the mycobacteria grow over time. These limitations could possibly be addressed by increasing the volume of MGIT sample taken for DNA extraction and ensuring effective decontamination of sample prior to MGIT inoculation.

Inadvertent DNA cross-contamination seems to have occurred on one MiSeq® run, both at the DNA extraction stage and the library preparation stage. Positive and negative controls were included at each stage. However, they did not exhibit any evidence of contamination (DNA quantification was zero for the negative control). Because the analysis is so sensitive, the contamination was discovered by the NGS pipeline. However, as the negative control did not reveal the contamination, these isolates (IMRL21, 22 and 27) could have been inadvertently reported as mixed species had the conventional molecular methods not been employed in parallel. There could, potentially, be many different types of DNA contamination present (from both living and dead organisms) when using a MGIT culture on Day 0. Some contamination could be from mycobacteria species on the same run and this would be difficult to distinguish if no other methods were present with which to compare the results. Contamination becomes an even greater issue when considering DNA sequencing directly from a clinical sample. DNA contamination remains a challenge for rapid WGS.

Handling a category 3 organism in a CL3 facility is a hazardous task. IMRL safety training, competency of staff, emergency procedure drills and maintenance of laboratory negative pressure levels, are all required to be up-to-date and faultless. Never-the-less, staff are still at a significantly higher risk than those handling category 2 organisms. Any method that will reduce the amount of handling of the organism is a progressive step. WGS could radically reduce the amount of handling time, while producing equivalent results, in a shorter time-frame. Any TB clinician or public health specialist would be encouraged to hear that reliable DST reports could be generated 15 days faster, and genotyping results could be available 13 days faster than at present. The high cost of WGS is often seen as a barrier to implementation in the diagnostic laboratory [222, 231]. However, an overall saving of 7% shows that it is a 'rapid, comprehensive and affordable' alternative [114]. The published article can be seen in Appendix 3.

This international collaborative prospective pilot study has provided proof of principle that WGS could be used in the diagnostic laboratory for mycobacterial diagnosis, genotypic surveillance and susceptibility testing, especially in cases where drug-resistance is suspected, to at least augment the service already provided, if not to replace it in the long-term.

Chapter 7.

General Discussion, Conclusions, and Future Directions

7 General Discussion, Conclusions, and Future Directions

Over the course of the study, it became clear that tuberculosis is still a disease of deprivation. One published study looked at a TB outbreak in crack cocaine users in Canada [142]. Another studied the prevalence of TB in the Traveller population in Ireland [154]. Six out of eleven clusters of another study were associated with substance misuse [112]. Prisoners who inject drugs were another cohort of TB patients studied [150]. In the 1900s, the TB problems in Ireland were associated with over-crowding and malnutrition in urban centres, especially tenement buildings in Dublin City Centre. Now, the outbreaks are still associated with the lower echelons of society. TB does not get media attention unless there is an outbreak. Although TB outbreaks within the Traveller community and in a prison were reported in the national news, they did not get prime time news headlines that a disease outbreak with more rapid progression might get. Instead they were largely ignored. No Ebola virus case was ever found in Ireland and yet it was in the headlines regularly. From an internet search, 5 newspaper articles and one radio broadcast were found referring to the above prison outbreak, from June 2011 to May 2013, which could have far-reaching consequences for the wider community. An internet search for Ebola uncovers such headlines as 'Ebola terror hits Ireland' (Irish Mirror), more than 5 articles within 3 months, as well as Health Service Executive publications on how to be ready for its spread. TB seems to be a forgotten illness, as are the people who it infects.

It has been hypothesised that we should not look at TB as a biomedical problem, but as a bio-social problem. Ortblad *et al* suggest that 'TB is a litmus test for society' (Figure 46) [240]. Increased levels of TB infection mean that people cannot go to work. They may be isolated in hospital, or at home, or very ill from the infection itself or side effects of treatment. This, in turn, leads to increased problems with poverty and social development. Poverty alone can lead to malnutrition, poor social development, smoking and alcohol-related dependencies, and poor living conditions, which, in turn, lead to decreased immunity and to increased susceptibility to TB and other infections. And so the cycle continues. Poverty is a causal determinant of TB, just as TB is a causal determinant of poverty. This cycle could be associated with South Africa, which has one of the highest TB prevalence rates in the world, or Ireland [241]. The authors point out that TB was on the decrease even before anti-tuberculous drugs were accessible due to improved standards of living. While standards of living may be on the increase for sections of our society, they remain dire for some.

The overall aim of this study was to improve and expand the research performed at the IMRL for the eventual benefit of the patient. The reference laboratory regularly performs publishable research, however it is done on an *ad hoc* basis, and is rarely published due to time constraints and lack of resources. Trinity College Clinical Microbiology Department is on the same campus as the

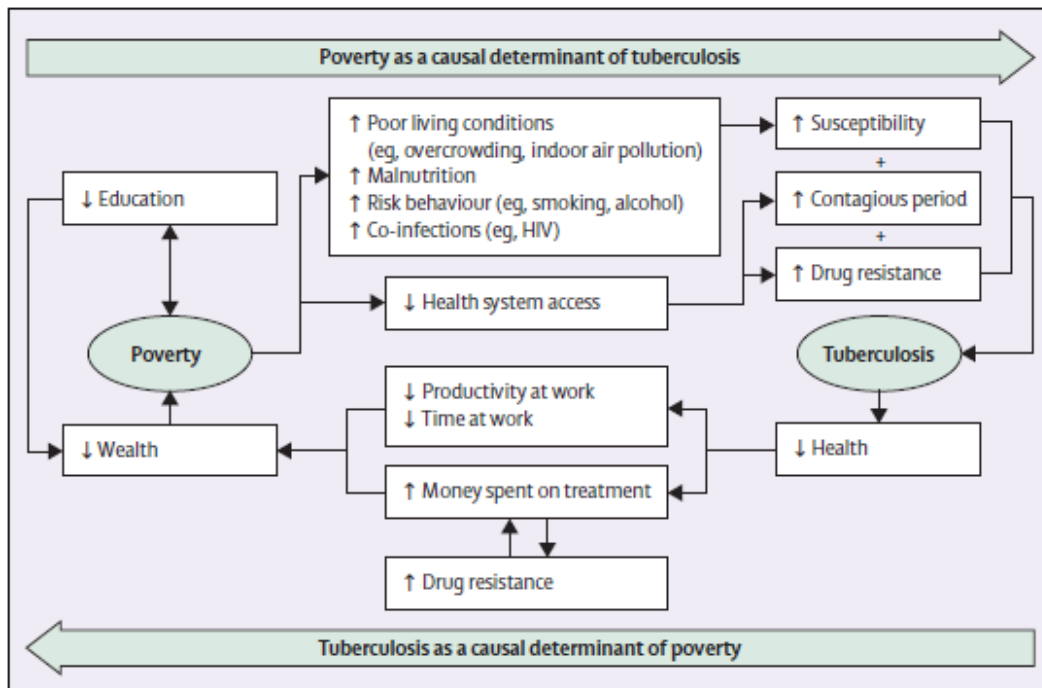


Figure 46. The Cycle of Poverty and Tuberculosis

Low socio-economic conditions and lifestyles lead to increased susceptibility to TB, which may not be diagnosed promptly, leading to increased transmission [240]. While ill or on treatment, the patient may not be able to attend work, which could lead to loss of employment or pay, which leads to poverty which leads back to complete to cycle of malnutrition, poor living conditions and risk behavior, which are all risk factors for development of TB. TB is a bio-social problem, and not as much of a bio-medical problem as it is treated.

IMRL and both parties could exploit this unique opportunity to collaborate further. During the current study, the IMRL has collaborated with Trinity College, and on an international scale [114]. The research has enabled publications that represent the IMRL widely so, in that regard, the aim has been fulfilled [106, 114, 164]. The MGIT Pilot Study has shown that WGS is capable of improving health outcomes for TB patients alongside the processes that are in place. However, it may be challenging to put a business case forward for an expensive platform to an already-underfunded hospital management when it would not be replacing many other tests.

7.1 MIRU-VNTR genotyping

MIRU-VNTR genotyping data was analysed over a five-year period. The study was in agreement with previous published literature which observed high diversity within MTBC strains in Ireland previously [105, 106]. The lineage distribution within the larger group studied was equivalent to previous studies. The predominant EU-wide lineage, Euro-American lineage 4, remains the most common genotype isolated.

The high diversity reflects the multi-cultural population present in Ireland, which is, in part, due to its membership of the EU and the free movement of people between member states. This evidence is further strengthened by the fact that at least seven cross-EU-border genotypes have been found in Ireland (susceptible cross-border genotypes not available for the analysis) [188].

Although the Ireland-Northern-Ireland study was not comprehensive (n=67 NI isolates), identical genotypes were found between strains collected in Northern Ireland (NI) and Ireland. According to the WHO, where Northern Ireland is included in UK data, the UK as a whole had a TB rate of 10 cases per 100,000 in 2014 (Ireland is recorded as 6.4 cases per 100,000) [1]. According to the Northern Ireland Public Health Agency, it has a lower TB prevalence rate than Ireland (5.2 cases per 100,000 in 2014 versus 7.1 cases per 100,000 in Ireland, reported by the HPSC) [22, 242]. This begs the question why there is such a large discrepancy between Northern Ireland and the UK as a whole, and also why, on the same island of Ireland, there is a difference in prevalence between the two regions. In documented record, Ireland, while low prevalence, has never reached as low as 5.2 cases per 100,000 [22]. By mid-2014, the Northern Ireland Statistics and Research Agency reported 13,093 immigrants within the resident population. The CSO reported 60,600 immigrants present in Ireland by mid-2014. The non-UK-born population in the UK as a whole in 2014 was 8.3 million according to the Office for National Statistics. From these figures, it is clear that Northern Ireland does not see as much immigration as Ireland or the rest of the UK, which could explain the differences. However, Northern Ireland has had a lower incidence of TB over the last century than Ireland, so this may not explain the discrepancy fully. Also, these crude incidence rates do not represent the overall incidence, which is much higher in urban centres, like London, Belfast and

Dublin [22, 242]. Rural TB incidence tends to be much lower than urban rates, and comparable within Ireland and Britain.

Irish people are also likely to contribute to the diversity of strains present. Since nationality is not always recorded on TB presentation, it was not possible to record country of birth for every patient. However, anecdotally, when all Euro-American lineage 4 strains, as well as the outbreak clusters that were analysed with WGS, were removed, and the remainder analysed for surnames and forenames commonly related to Irish origin, Irish patients presented with strains such as West African 2 lineage 6 (n=2), EAI lineage 1 (n=8), Delhi/CAS lineage 3 (n=4) and Beijing lineage 2 (n=12). Irish people have emigrated worldwide in large numbers since famine times [25]. Many have travelled home from high TB burden countries, either to visit or to stay, and could have transmitted TB from diverse geographical regions within that time.

Even though the evidence to suggest that some strains are more virulent than others may not be overwhelming, some clones are being expanding across Europe more readily than others (MtbC15-9 100-32 and 94-32 Beijing lineage 2 genotypes are examples) [188]. Even if transmission of imported strains has not occurred in great numbers to date, it is only a matter of time before it could happen. Surveillance gives us the tools to be able to track and trend TB molecular epidemiology in Ireland, which could alert to that transmission in real-time. Prospective surveillance gives us a detailed and accurate view of the situation and how it has changed over time. The longer time-frame indicated that the cluster size was not as high as assumed from previous studies, which could be seen as an indication that TB control is working successfully. However, the overall clustering sizes and rates do not take into account the larger clusters that seem to be expanding in the population. This could mean many more LTBI contacts within the community. For instance, in 2013, the HPSC recorded 155 LTBI cases from just 12 outbreaks (Table 4) [22]. The larger community clusters, although challenging for public health, would be the best clusters to focus resources on in the future in order to curtail their further spread.

7.2 Cluster analysis

This study was the first documented work using WGS to analyse MTBC outbreaks found in Ireland. Clusters of interest from the MIRU-VNTR genotyping study were analysed in more detail using WGS, based on previous publications where WGS had delineated MIRU-VNTR genotyping clusters with greater resolution [112, 142, 143]. The original hypothesis, that MIRU-VNTR genotyping over-clusters isolates in certain cases, which could be mis-leading were it to be relied on completely, was proven (Clusters 1,2a, and 5-8 inclusive), although conventional genotyping performed well in many cases (mainly correlated in clusters 3, 4 and 9), and correlated completely in two cases (Cluster10 and 11).

A second hypothesis, that WGS improves on MIRU-VNTR genotyping by adding value and extra depth to the analysis in cases where epidemiological data is difficult to collect, was upheld in general, with some exceptions. Cluster 2a and 2b represent an epidemiologically cryptic cluster. WGS confirmed two separate outbreaks (differing by one MIRU-VNTR SLV) (Figure 28). Furthermore, the presence of two possible super-spreaders was indicated by the data. Two discrepancies were observed, where household contacts failed to be confirmed as transmission by WGS, even though MIRU-VNTR genotyping found they were identical. Although unlikely, it is possible that a sequencing error occurred, or the incorrect isolates were tested, or that there was an anomaly in the phylogenetic tree (which predicted other genotypes correctly).

Super-spreaders were also indicated in Clusters 1 and 10 (Figures 26 and 36). The hypothesis, that public health surveillance, identification of super-spreaders, and contact tracing, would be greatly augmented by the use and interpretation of WGS data in the IMRL, was proven by the study. The caveat remains that no method is 100% reliable and that clinical judgement and epidemiological evidence cannot be replaced by WGS.

NICE guidelines suggest the principles of TB control are as follows:

- Early diagnosis, case finding, and contact tracing following diagnosis of new active cases
- Directly Observed Therapy (DOT) and/or adherence to treatment
- Drug resistance and awareness of drug interactions and side-effects
- Free treatment for all
- Dealing with social and cultural barriers associated with TB
- Role of allied health professionals in identifying and referring cases appropriately
- Preventing circulation of mis-information that causes fear about TB, for instance within households
- Education of a vast array of different risk groups about the above topics [243].

In order to follow these guidelines, a multi-disciplinary approach would be the most beneficial way of tackling TB clustering and outbreaks in Ireland [243]. Ideally, a co-ordinated team made up of clinicians, nurses, pharmacists, clinical microbiologists, public health specialists, HPSC experts, and the IMRL, would work in tandem on prospective outbreak surveillance, detection, contact tracing, and prevention. At present, Ireland lacks a countrywide TB Controller that could co-ordinate a team like this. A national strategy and funding are also needed for LTBI testing and treatment and immigrant screening. However, it is difficult to secure funding for a tuberculosis strategy in a low-prevalence setting.

7.3 MDR/XDR-TB

This study was the first documented in-depth analysis of drug resistance in MDR/XDR-TB using WGS in Ireland [164]. The hypothesis that MDR/XDR-TB is not being readily transmitted in the Irish-born population can be accepted due to the evidence produced in the study. The most closely-related strains were between non-Irish born cases, or between both Irish-born and non-Irish-born cases. The second hypothesis, i.e. that MDR/XDR-TB is not being spread from the immigrant population to the Irish-born population, was upheld for all but 5 cases (n=2 clusters). In one case, the Irish patient was associated with non-compliance and presented first, therefore it could be assumed he was the index case and no transmission from a non-Irish-born individual to an Irish individual had occurred. However, since the strains are so similar, it would be impossible to confirm this without more epidemiological evidence than the fact that they attended the same healthcare facility.

The second case, where the two most closely genomically-linked cases were an Indian nurse and an Irish-born gentleman who was associated with non-compliance, had no known epidemiological link either. They presented within 3 months of each other with MDR-TB. Their MtbC15-9 codes were significantly different. Twelve MIRU-VNTR loci differed, although they were both EAI lineage 1 and branched together phylogenetically (Figure 38). Their drug-resistance-associated mutations differed significantly. IEMDR03 represented a patient who had developed MDR-TB while on anti-TB therapy, therefore had actually presented much earlier, in 2004. This is an example of where phylogenetic trees cannot be completely relied upon since their construction is based on mathematics and not biology or epidemiology. Because of perceived problems with repeat regions and drug-resistance-associated mutations, those regions are often removed for phylogenetic analysis, which could cause discrepancies in cases where isolates are very similar. As with WGS cluster analysis of the 11 susceptible clusters, WGS throws up some questions when weighed against epidemiological evidence. This study has shown no evidence of transmission of MDR/XDR-TB from the non-Irish population to the native Irish population.

Other research questions asked whether WGS can accurately predict the phenotypic resistance profile of *M. tuberculosis* in a cohort of multi-drug resistant strains found in Ireland, and whether it could improve MDR/XDR-TB detection and accelerate drug resistance prediction if it was introduced in the diagnostic laboratory. WGS has proven to be an excellent tool for predicting resistance of MTBC to some drugs, i.e. rifampicin, isoniazid, and fluoroquinolones. However, its performance surrounding other drugs has not been as good. This is, for the most part, probably due to phenotypic DST being the 'gold' standard. Phenotypic DST has never been 100% reliable. If a less than reliable reference standard is being used, discrepancies are bound to occur. The other downside of phenotypic DST is that the amount of phenotypic evidence for some drugs (rifampicin and isoniazid) far out-weighs the results available for others. In the IMRL, second- and third-line

DST has always been performed, based on clinical need, in a supra-national reference laboratory (now performed in-house). Furthermore, the cohort of drugs tested has changed to augment clinical need, which led to gaps in the DST data (Table 10). For example, DST was not performed on linezolid up until 2010. DST on clarithromycin has not been performed since 2011.

The drug-resistance-associated mutations that are currently being investigated by various research groups and collaborators could eventually be used to create user-friendly, reliable, point-of-care tests that could be used in developing countries with high MDR/XDR-TB prevalence. The research being performed in wealthier countries will benefit poorer nations in the long-run.

When discussing drug-resistance mutations, one must always bear in mind that there are other mechanisms of drug resistance than the accumulation of chromosomal mutations over time, or due to selection pressure. Efflux pumps are a prolific area of research currently [55, 56]. WGS could aid this research area by identifying new efflux pumps that could be involved in resistance within MTBC, or genetic mutations within efflux pumps that may cause a pathogen to be more or less susceptible to particular drugs. Efflux pumps were not included in the current study due to time and funding constraints, but would be an excellent direction for future research.

While a global consensus resistance mutation catalogue has not yet been developed, there is already sufficient evidence to suggest that WGS could augment and accelerate the methods already in place for drug-resistance prediction in the IMRL. Ideally, pending WGS being performed successfully on clinical samples, suspect specimens would be tested with a rapid molecular assay, such as Cepheid GeneXpert MTB/RIF, on receipt. If positive, once MTBC cultures flagged positive, whole genome sequencing could be performed as soon as possible on the suspect culture. Meanwhile, DST in the form of comprehensive micro-titre plate MICs would be performed in parallel. WGS would be available prior to DST on a preliminary basis, and could give the clinician a rapid first-line drug resistance profile that they could use to formulate a drug regimen tailored to the patient. Mutations would be labelled as ‘high- or low-confidence’, or with the MIC with which they correlated, according to how much evidence was available regarding their association with resistance in MTBC. If, or when, the research surrounding them was sufficiently robust in nature, they could be upgraded to ‘high-confidence’. The MIC profile would follow. Constant communication between laboratory and respiratory team, as part of a multi-disciplinary approach, would enable any genotype-phenotype discrepancies to be flagged as soon as possible. This may not be the solution for every laboratory, but would be especially beneficial in a reference laboratory environment.

7.4 MGIT Pilot study

The IMRL participated in an international collaborative study to prove the hypothesis that WGS could replace traditional methods of identification, susceptibility testing and genotyping in the diagnostic laboratory, due to its more rapid turnaround times, accuracy, and decreasing costs. NICE guidelines (2016) recommend research that will answer the following question:

‘In people with suspected TB, what is the relative clinical and cost effectiveness of universal and risk-based use of rapid nucleic amplification tests?’ [243].

The MGIT pilot study carried out cost effectiveness studies in some centres, however it was not done in the IMRL. This shows that the study was timely and the research appropriate. Cost-effectiveness studies would have to be carried out in the IMRL if WGS were to be introduced in targeted situations. It is unlikely that WGS would be performed universally in the IMRL, at least in the near future.

Following the publication of the international pilot study, Public Health England plan to introduce the workflow into reference laboratories across the country, replacing current phenotypic DST methods with custom-designed micro-titre plate MICs (personal communication, Prof. Derrick Crook) [114, 224]. They seem confident in its success, but it may be too soon for this kind of drastic change. The worldwide collaborative CryPTIC project is currently assessing the newly-designed micro-titre plate MICs. Using the MIC results, they plan to consolidate a resistance mutation catalogue, with thousands of MTBC isolates and their genomes, with which to move forward. This is envisaged to take five years. Perhaps the PHE would do well to wait until then to move completely to the new system.

In this study, proof of principle was provided that WGS could replace conventional methods for identification, outbreak detection, and drug resistance prediction, however, it could not be done in every mycobacterial laboratory, nor could it be done without more controlled prospective clinical trials having proven it a better alternative. However, as already stated, it could augment methods already in place, and replace some rapid molecular tests.

7.5 Critical Evaluation of Whole Genome Sequencing

7.5.1 Advantages over Conventional Methods

WGS is the ultimate genotyping tool. It can provide added value in abundance by consolidating other molecular tests into one assay. The information can be stored and used as reference for future testing. A reference database of varied, wide-ranging, genomic MTBC data enables better quality

phylogenetic analysis. At any time, the database can be interrogated on the basis of newly-emerging research. WGS is a safe technique since hazardous MTBC cultures are heat-inactivated prior to further whole genome extraction and processing. WGS platforms can be shared across multi-disciplinary platforms, such as the TrinSeq NGS laboratory at Trinity College, since the common denominator is DNA whether one is working with humans, micro-organism, animals, or plants.

WGS could be of benefit to Clinical Microbiology, not only for MTBC, but for other mycobacteria and bacteria also. Projects are currently ongoing in Trinity College, Dublin, and St. James' Hospital that are utilising NGS technology for *Mycobacterium chimaera*, *Mycobacterium abscessus*, and *Neisseria gonorrhoea*. Any fastidious pathogen that contains sufficient DNA can be utilised. Even if sufficient DNA cannot be achieved (for instance with some viruses), PCR techniques can be employed to amplify the target DNA, and then use NGS technology to find resistance mutations, or perform phylogenetic analysis, or elucidate the pathogen present. A multi-disciplinary approach would benefit the entire Pathology Laboratory, since human DNA can also be analysed using NGS technology.

MDR/XDR-TB isolates could be incubating for many weeks in the CL3 laboratory, whether it is being grown prior to detection, or for phenotypic DST, or for repeat DST once resistance is detected. WGS is safer than growing a high bacterial load of a hazardous, highly contagious, multi-drug resistant organism. Once the MTBC DNA is heat-inactivated, it poses no further threat of transmission.

Turnaround times for clinical results could be considerably decreased by using WGS in the mycobacteria laboratory. The MGIT Pilot Study has proven this by decreasing time to reporting of species identification and DST by a median of 15 days, and nearest-neighbour relatedness by 23 days. These decreases are significant in a TB setting where every day an infectious person is on the incorrect treatment regimen is a day where transmission could occur. In the near future, it will be possible to sequence directly from clinical samples, and this could decrease turnaround times even further.

7.5.2 Limitations and challenges

Of course, there are limitations with the technology, like any other method that has been discovered. While it is, scientifically speaking, extremely beneficial to have a large database of reference strains, the dilemma of where and how to store large amounts of patient data remains an issue, especially in a hospital environment, where data protection is of utmost importance. Storing pseudonymised data in the cloud would be the best solution, although cloud computing is seen as

insecure in a data protection capacity by the information technology department in St. James' Hospital. WGS could be seen as offering too much information when simple questions are being asked. Perhaps there is no need to store all this data, when it may never be required. Despite WGS having decreased in cost, it is still an expensive process. Costs for the MGIT Pilot Study were based on having a sequencing instrument in place, and replacing DST. The instruments themselves cost over €100,000 to purchase, which would be a large outlay for an already under-funded reference laboratory, and since the susceptible strains isolated would still require DST, the IMRL could not replace DST. It might remain out of reach for the moment, even for many higher-income countries.

Sequencing instruments must run at capacity in order to be worth their high cost. Batching of isolates may be necessary, which could negate the decreased turnaround time that WGS offers. The solution to this could be to use sequencing instruments as part of a multi-disciplinary organisation. An example of this would be TrinSeq, a company based on the SJH campus with links to Trinity College, who charge for the use of their Illumina MiSeq[®] on a first-come first-served basis. Users from many different facets of research use the service, for sequencing bacterial, viral, human and animal DNA. Consequently, users do not have to bear the high cost of instrument purchase and maintenance. On the other hand, molecular tests are currently batched and performed weekly, and this seems to have been sufficient for users to date.

Takiff *et al* state, and others agree, that 'with advances in technology, WGS of clinical specimens could become routine in high-income countries; however, its relevance will probably depend on easy-to-use software to efficiently process the sequences produced and accessible genomic databases that can be mined for future studies' [141, 223, 244, 245]. When designing a WGS analysis workflow, many decisions have to be made, such as which software to use for each task. Individual research groups are using myriad different programs integrated with custom scripts. In a constantly changing computational environment, it is challenging to know if the software being used is the optimum. Bio-informatics expertise is essential, at least when setting up an analysis pipeline. Most scientists will not have extensive bio-informatics experience, and it could take years to learn. In the diagnostic laboratory, it is not feasible to expect everyone to be a bio-informatics expert. The workflow must be targeted at the member of the team with the least knowledge.

As with any molecular method, while finding a mutation may predict phenotypic resistance, not finding a mutation is not a definite indicator of susceptibility. Any report format envisaged would have to incorporate that caveat. Furthermore, as with all *in vitro* molecular tests, what happens in the laboratory may not reflect what is happening *in vivo*. Every new method that becomes available in microbiology has its advantages and challenges. The challenges, especially those of 'big data' analysis, will eventually be solved for the diagnostic laboratory.

Contamination with other types of DNA, whether human, nasopharyngeal flora, other mycobacteria, or otherwise, constitutes the main obstacle to directly sequencing clinical samples. Even the ‘newly-positive’ mycobacterial cultures included in the MGIT Pilot Study contained varying amounts of contamination, which is not surprising since the MGIT vials they were taken from contained centrifuged deposits of clinical samples. Positive and negative controls, as used in the classical sense, did not seem to flag the mycobacterial contamination that was present. ZN morphology clearly demonstrated the mixture found. It could be challenging to prove whether a mixture is truly present or not with WGS, especially where culture was no longer in use.

7.6 Systems biology approach

Despite the WHO and others’ best efforts to eradicate it (e.g. STOP-TB Partnership, END-TB Strategy, TB Alliance), TB is still the joint-leading cause of death due to an infectious agent worldwide [1]. Unfortunately, not many new drugs, and no effective vaccine, have been discovered to date. WGS research has not been able to explain differences in transmissibility between strains or lineages, or why some strains are more virulent than others, or more likely to develop drug resistance [244].

The theme for WHO World TB Day 2016 was ‘Unite to End TB’. New drugs, vaccines, and clinical tools are urgently required. Researchers and contributors from all aspects of TB research need to work together to make that happen, through a systems biology approach (Figure 47) [246]. ReseqTB data-sharing platform is a good example of collaborative work between groups with different interests in TB (WHO, US CDC, FIND, C-Path, NDWG) [121].

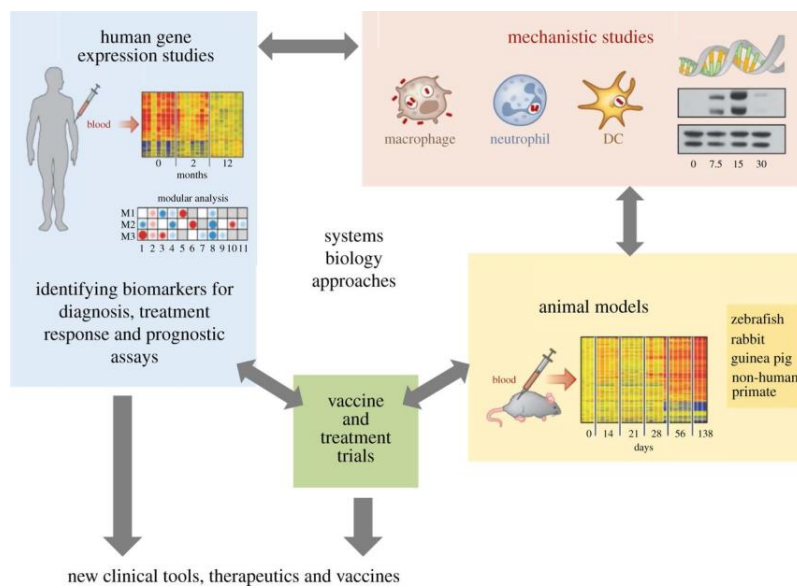


Figure 47. Systems Biology Approach to the development of new Clinical Tools, Therapeutics and Vaccines

Only through a sustained multi-disciplinary approach will the essential new clinical tools, vaccines and drugs be developed [246]. Molecular drug resistance mutation prediction feed into this approach since new tests could be designed around the resistance mutations found.

7.7 Conclusions

MIRU-VNTR genotyping is an excellent first-line tool for surveillance of the molecular epidemiology of MTBC in Ireland. A high diversity of lineages was found within strains collected between 2010 and 2014 inclusive, which correlates with previous Irish studies. A high clustering rate was associated with a cluster size of two in over 50% of cases. A small number of clones had seen large expansion since the previous study was undertaken in 2010-11.

WGS cluster analysis of eleven clusters of interest resulted in MIRU-VNTR genotyping having over-estimated clustering in 6 out of 11 clusters. Some clusters contained sub-clades that agreed with MIRU-VNTR genotyping (n=3). MIRU-VNTR genotyping was confirmed using WGS for two clusters. Furthermore, possible super-spreaders were indicated in the data of 3 clusters, where public health resources could be focussed to prevent further TB transmission. Two major discrepancies were noted where WGS did not correlate with household contacts known to public health. Despite this, WGS would be a useful addition to the diagnostic laboratory in cases where identical MIRU-VNTR genotypes were found.

Drug resistance prediction using WGS compared to phenotypic DST resulted in widely varying sensitivity, depending on the drug, and analysis platform used, while specificity ranged from 86-100%. When rifampicin and isoniazid alone were analysed, sensitivities ranged from 90-100%, better than rapid molecular tests already available (83-93%). When fluoroquinolones were analysed, WGS analysis resulted in sensitivity of 71%, less than the currently available LPA (83%). Sensitivity for aminoglycosides and other drugs was much lower, although specificity remained high. TB Profiler and PhyReSe resulted in the highest overall sensitivity compared to DST (higher than LPAs also). However, a high rate of false positivity was seen with TB Profiler (n=13). It appears there is a trade-off required to find a level where false positivity is minimised and sensitivity is optimised. The sensitivity and specificity ranges above were utilised for comparison. A larger sample size, and more phenotypic DST data, would be required in order to improve their confidence intervals.

WGS analysis, although currently not at a stage where it could replace phenotypic DST completely, would be an invaluable additional tool within the laboratory for rapid drug resistance prediction of MTBC. The MGIT Pilot Study provided proof of principle that WGS could not only be employed for drug resistance prediction, but also for nearest neighbour relatedness analysis that could flag new or expanding outbreaks, and mycobacterial identification in significantly less time than conventional methods.

Despite its limitations, WGS represents a 'game-changing' technology for MTBC and many other microbiological applications.

7.8 Future directions and hypotheses generated

Whole genome sequencing could be performed directly from sputum, further decreasing turnaround times for diagnosis, susceptibility testing and genotyping.

There is much work still to be done in order to whole genome sequence successfully using a clinical sample that may have more contaminating DNA present than mycobacterial DNA. The MMM Group is currently working on an extraction method for this purpose.

MIRU-VNTR genotyping could be performed directly on clinical samples as a rapid first-line indicator of TB transmission.

It has been shown that MIRU-VNTR genotyping directly on clinical specimens is possible, however, it has never been attempted in a study in Ireland. This could accelerate the flagging of potential outbreaks, augmenting public health without impacting largely on work practices or costs in the IMRL [247].

Whole genome sequencing on newly positive cultures could be introduced to the routine diagnostic laboratory in cases of suspected drug resistance and outbreaks.

A blinded in-house prospective WGS study could assist in increasing the level of confidence surrounding genomic results for clinicians. Sequencing analysis needs to be standardised prior to introduction of WGS in any form.

Sequential isolates will give many more insights into TB transmission and micro-evolution.

Plans to sequence sequential isolates in a collaborative study are already underway at the IMRL.

Bibliography

Bibliography

1. WHO, *World Health Organisation Global Tuberculosis Report 2015*. 2015, World Health Organisation: Online. p. 1-192.
2. (CDC), C.f.D.C.a.P., *Tuberculosis genotyping--United States, 2004-2010*. MMWR Morb Mortal Wkly Rep, 2012. **61**(36): p. 723-5.
3. Sudre, P., G. ten Dam, and A. Kochi, *Tuberculosis: a global overview of the situation today*. Bull World Health Organ, 1992. **70**(2): p. 149-59.
4. Kanabus, A. *TB Facts; Information about tuberculosis*. 2016; Available from: www.tbfacts.org/treatment-of-drug-resistant-tb/.
5. Ventura, M., et al., *Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum*. Microbiology and Molecular Biology Reviews, 2007. **71**(3): p. 495-548.
6. O'Driscoll, C., et al., *Molecular epidemiology of Mycobacterium abscessus complex isolates in Ireland*. Journal of Cystic Fibrosis, 2016. **15**(2): p. 179-185.
7. Mohamed Buhary, T., S.L. Gayed, and I. Hafeez, *Pericardial effusion with Mycobacterium avium complex in HIV-infected patients*. BMJ Case Rep, 2016. **2016**.
8. Cole, S.T., et al., *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence*. Nature, 1998. **393**(6685): p. 537-44.
9. Koch, R., [*The etiology of tuberculosis by Dr. Robert Koch. From the Berliner Klinische Wochenschrift, Volume 19 (1882)*]. Zentralbl Bakteriol Mikrobiol Hyg A, 1882. **251**(3): p. 287-96.
10. Blouin, Y., et al., *Significance of the identification in the Horn of Africa of an exceptionally deep branching Mycobacterium tuberculosis clade*. PLoS One, 2012. **7**(12): p. e52841.
11. Baker, O., et al., *Human tuberculosis predates domestication in ancient Syria*. Tuberculosis (Edinb), 2015. **95 Suppl 1**: p. S4-S12.
12. Hershkovitz, I., et al., *Detection and molecular characterization of 9,000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean*. PLoS One, 2008. **3**(10): p. e3426.
13. Witas, H.W., et al., *Molecular studies on ancient M. tuberculosis and M. leprae: methods of pathogen and host DNA analysis*. European Journal of Clinical Microbiology & Infectious Diseases, 2015. **34**(9): p. 1733-1749.
14. Zink, A.R., et al., *Characterization of Mycobacterium tuberculosis complex DNAs from Egyptian mummies by spoligotyping*. J Clin Microbiol, 2003. **41**(1): p. 359-67.
15. Suzuki, T., H. Fujita, and J.G. Choi, *Brief communication: new evidence of tuberculosis from prehistoric Korea--Population movement and early evidence of tuberculosis in far East Asia*. Am J Phys Anthropol, 2008. **136**(3): p. 357-60.
16. Comas, I., et al., *Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans*. Nat Genet, 2013. **45**(10): p. 1176-82.
17. Merker, M., et al., *Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage*. Nat Genet, 2015. **advance online publication**.
18. Wilson, L.G., *Commentary: Medicine, population, and tuberculosis*. Int J Epidemiol, 2005. **34**(3): p. 521-4.
19. Dormandy, T., *The White Death: A History of Tuberculosis*. 1999, The Hambledon Press, London and Rio Grande.
20. Team, E.E., *ECDC and WHO/Europe joint report on tuberculosis surveillance and monitoring in Europe*. Euro Surveill, 2014. **19**(11).
21. Mokrousov, I., *Molecular structure of Mycobacterium tuberculosis population in Russia and its interaction with neighboring countries*. International Journal of Mycobacteriology, (0).
22. HPSC. *Reports on the Epidemiology of TB in Ireland*. [Report (Online), available at: <http://www.hpsc.ie/A-Z/VaccinePreventable/TuberculosisTB/Epidemiology/AnnualReports/>] 1998-2014; Reports on the Epidemiology of TB in Ireland from 1998 to 2014]. Available from: <http://www.hpsc.ie/A-Z/VaccinePreventable/TuberculosisTB/Epidemiology/AnnualReports/>.
23. Fair, E., et al., *Molecular epidemiologic investigation of tuberculosis in an area of increasing incidence in inner-city Dublin*. Ir Med J, 2006. **99**(3): p. 87-90.

24. Office, C.S., *Population and Migration Estimates to April 2015*, C.S.O. Ireland, Editor. 2015, Central Statistics Office, Ireland: Online.
25. O'Rourke, K., *Emigration and living standards in Ireland since the Famine*. J Popul Econ, 1995. **8**(4): p. 407-21.
26. Hollo, V., et al., *The effect of migration within the European Union/European Economic Area on the distribution of tuberculosis, 2007 to 2013*. Euro Surveill, 2016. **21**(12).
27. Ködmön, C., P. Zucs, and M.J. van der Werf, *Migration-related tuberculosis: epidemiology and characteristics of tuberculosis cases originating outside the European Union and European Economic Area, 2007 to 2013*. Euro Surveill, 2016. **21**(12).
28. Brites, D. and S. Gagneux, *Co-evolution of Mycobacterium tuberculosis and Homo sapiens*. Immunol Rev, 2015. **264**(1): p. 6-24.
29. Beggs, C.B., et al., *The transmission of tuberculosis in confined spaces: an analytical review of alternative epidemiological models*. Int J Tuberc Lung Dis, 2003. **7**(11): p. 1015-26.
30. Esmail, H., et al., *The ongoing challenge of latent tuberculosis*. Philos Trans R Soc Lond B Biol Sci, 2014. **369**(1645): p. 20130437.
31. Reva, O., I. Korotetskiy, and A. Ilin, *Role of the horizontal gene exchange in evolution of pathogenic Mycobacteria*. BMC Evol Biol, 2015. **15 Suppl 1**: p. S2.
32. Eldholm, V. and F. Balloux, *Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out*. Trends Microbiol, 2016.
33. Wolf, A.J., et al., *Mycobacterium tuberculosis infects dendritic cells with high frequency and impairs their function in vivo*. J Immunol, 2007. **179**(4): p. 2509-19.
34. Kawai, T. and S. Akira, *The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors*. Nat Immunol, 2010. **11**(5): p. 373-84.
35. McDonough, K.A., Y. Kress, and B.R. Bloom, *The interaction of Mycobacterium tuberculosis with macrophages: a study of phagolysosome fusion*. Infect Agents Dis, 1993. **2**(4): p. 232-5.
36. McDonough, K.A., Y. Kress, and B.R. Bloom, *Pathogenesis of tuberculosis: interaction of Mycobacterium tuberculosis with macrophages*. Infect Immun, 1993. **61**(7): p. 2763-73.
37. Chan, J., et al., *Lipoarabinomannan, a possible virulence factor involved in persistence of Mycobacterium tuberculosis within macrophages*. Infect Immun, 1991. **59**(5): p. 1755-61.
38. Yuan, Y., et al., *Identification of a gene involved in the biosynthesis of cyclopropanated mycolic acids in Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A, 1995. **92**(14): p. 6630-4.
39. van der Wel, N., et al., *M. tuberculosis and M. leprae translocate from the phagolysosome to the cytosol in myeloid cells*. Cell, 2007. **129**(7): p. 1287-98.
40. de Jonge, M.I., et al., *ESAT-6 from Mycobacterium tuberculosis dissociates from its putative chaperone CFP-10 under acidic conditions and exhibits membrane-lysing activity*. J Bacteriol, 2007. **189**(16): p. 6028-34.
41. Linderman, J.J., et al., *A multi-scale approach to designing therapeutics for tuberculosis*. Integr Biol (Camb), 2015. **7**(5): p. 591-609.
42. Bodnar, K.A., N.V. Serbina, and J.L. Flynn, *Fate of Mycobacterium tuberculosis within murine dendritic cells*. Infect Immun, 2001. **69**(2): p. 800-9.
43. Russell, D.G., *Who puts the tubercle in tuberculosis?* Nat Rev Microbiol, 2007. **5**(1): p. 39-47.
44. Orme, I.M. and A.M. Cooper, *Cytokine/chemokine cascades in immunity to tuberculosis*. Immunol Today, 1999. **20**(7): p. 307-12.
45. Fraziano, M., et al., *Expression of CCR5 is increased in human monocyte-derived macrophages and alveolar macrophages in the course of in vivo and in vitro Mycobacterium tuberculosis infection*. AIDS Res Hum Retroviruses, 1999. **15**(10): p. 869-74.
46. Ramakrishnan, L., *Revisiting the role of the granuloma in tuberculosis*. Nat Rev Immunol, 2012. **12**(5): p. 352-66.
47. Ahmad, S., *Pathogenesis, immunology, and diagnosis of latent Mycobacterium tuberculosis infection*. Clin Dev Immunol, 2011. **2011**: p. 814943.

48. O'Leary, S.M., et al., *Cigarette Smoking Impairs Human Pulmonary Immunity to Mycobacterium tuberculosis*. American Journal of Respiratory and Critical Care Medicine, 2014.
49. North, R.J. and Y.J. Jung, *Immunity to tuberculosis*. Annu Rev Immunol, 2004. **22**: p. 599-623.
50. Sandgren, A., V. Hollo, and M.J. van der Werf, *Extrapulmonary tuberculosis in the European Union and European Economic Area, 2002 to 2011*. Euro Surveill, 2013. **18**(12).
51. Cheallaigh, C.N., et al., *Interferon gamma release assays for the diagnosis of latent TB infection in HIV-infected individuals in a low TB burden country*. PLoS One, 2013. **8**(1): p. e53330.
52. Fitzgerald, I.J., *Specific Immune-based Detection of Latent Tuberculosis Infection*, in *Clinical Microbiology, Trinity College, Dublin*. 2010, Trinity College, Dublin: Trinity College, Dublin.
53. Tiemersma, E.W., et al., *Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: a systematic review*. PLoS One, 2011. **6**(4): p. e17601.
54. Mangtani, P., et al., *Protection by BCG vaccine against tuberculosis: a systematic review of randomized controlled trials*. Clin Infect Dis, 2014. **58**(4): p. 470-80.
55. da Silva, P.E.A., et al., *Efflux as a mechanism for drug resistance in Mycobacterium tuberculosis*. Fems Immunology and Medical Microbiology, 2011. **63**(1): p. 1-9.
56. Coelho, T., et al., *Enhancement of antibiotic activity by efflux inhibitors against multidrug resistant Mycobacterium tuberculosis clinical isolates from Brazil*. Front Microbiol, 2015. **6**: p. 330.
57. Committee, N.T.A., *Guidelines on the Prevention and Control of Tuberculosis in Ireland 2010, amended 2014*, I. Health Protection Surveillance Centre and Health Service Executive, Editor. 2010, Health Protection Surveillance Centre, Ireland: Online.
58. Dheda, K., et al., *Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis*. Lancet Respir Med, 2014. **2**(4): p. 321-38.
59. Zhang, Y. and W.W. Yew, *Mechanisms of drug resistance in Mycobacterium tuberculosis*. Int J Tuberc Lung Dis, 2009. **13**(11): p. 1320-30.
60. Migliori, G.B., et al., *First tuberculosis cases in Italy resistant to all tested drugs*. Euro Surveill, 2007. **12**(5): p. E070517.1.
61. Walker, T.M., et al., *Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study*. Lancet Infect Dis, 2015. **15**(10): p. 1193-202.
62. Zhang, Y. and W.W. Yew, *Mechanisms of drug resistance in Mycobacterium tuberculosis: update 2015*. Int J Tuberc Lung Dis, 2015. **19**(11): p. 1276-89.
63. Tortoli, E., et al., *Use of BACTEC MGIT 960 for recovery of mycobacteria from clinical specimens: multicenter study*. J Clin Microbiol, 1999. **37**(11): p. 3578-82.
64. Tortoli, E., et al., *Evaluation of automated BACTEC MGIT 960 system for testing susceptibility of Mycobacterium tuberculosis to four major antituberculous drugs: comparison with the radiometric BACTEC 460TB method and the agar plate method of proportion*. J Clin Microbiol, 2002. **40**(2): p. 607-10.
65. Evans, C.A., *GeneXpert--a game-changer for tuberculosis control?* PLoS Med, 2011. **8**(7): p. e1001064.
66. Tomasicchio, M., et al., *The diagnostic accuracy of the MTBDRplus and MTBDRsl assays for drug-resistant TB detection when performed on sputum and culture isolates*. Sci Rep, 2016. **6**: p. 17850.
67. Allix-Beguec, C., M. Fauville-Dufaux, and P. Supply, *Three-year population-based evaluation of standardized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of Mycobacterium tuberculosis*. J Clin Microbiol, 2008. **46**(4): p. 1398-406.
68. Supply, P., et al., *Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of Mycobacterium tuberculosis*. Journal of Clinical Microbiology, 2006. **44**(12): p. 4498-4510.
69. Allix-Beguec, C., et al., *Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic*

- identification of *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol*, 2008. **46**(8): p. 2692-9.
70. Langmead, B., *Aligning short sequencing reads with Bowtie*. *Curr Protoc Bioinformatics*, 2010. **Chapter 11**: p. Unit 11.7.
 71. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. *Bioinformatics*, 2009. **25**(16): p. 2078-9.
 72. Gouy, M., S. Guindon, and O. Gascuel, *SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building*. *Mol Biol Evol*, 2010. **27**(2): p. 221-4.
 73. Guindon, S., et al., *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0*. *Syst Biol*, 2010. **59**(3): p. 307-21.
 74. Afgan, E., et al., *The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update*. *Nucleic Acids Res*, 2016. **44**(W1): p. W3-W10.
 75. Field, D., et al., *Open software for biologists: from famine to feast*. *Nat Biotechnol*, 2006. **24**(7): p. 801-3.
 76. Feuerriegel, S., et al., *PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data*. *J Clin Microbiol*, 2015. **53**(6): p. 1908-14.
 77. Coll, F., et al., *Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences*. *Genome Med*, 2015. **7**(1): p. 51.
 78. Kearse, M., et al., *Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data*. *Bioinformatics*, 2012. **28**(12): p. 1647-9.
 79. Gregory, T.R., *Understanding Evolutionary Trees*. *Evolution: Education and Outreach*, 2008. **1**(2): p. 121-137.
 80. Lemey, P., *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. 2009: Cambridge University Press.
 81. Salemi, M. and A.-M. Vandamme, *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*. 2003: Cambridge University Press.
 82. Sokal, R.R. and C.D. Michener, *A statistical method for evaluating systematic relationships*. *University of Kansas Scientific Bulletin*, 1958. **28**: p. 1409-1438.
 83. Salipante, S.J. and B.G. Hall, *Inadequacies of minimum spanning trees in molecular epidemiology*. *J Clin Microbiol*, 2011. **49**(10): p. 3568-75.
 84. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids*. *Nature*, 1953. **171**(4356): p. 737-738.
 85. Gilbert, W. and A. Maxam, *The nucleotide sequence of the lac operator*. *Proc Natl Acad Sci U S A*, 1973. **70**(12): p. 3581-4.
 86. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*. *Proc Natl Acad Sci U S A*, 1977. **74**(12): p. 5463-7.
 87. Venter, J.C., et al., *The sequence of the human genome*. *science*, 2001. **291**(5507): p. 1304-1351.
 88. Shendure, J., et al., *Advanced sequencing technologies: methods and goals*. *Nat Rev Genet*, 2004. **5**(5): p. 335-44.
 89. Ma, Z., et al., *Global tuberculosis drug development pipeline: the need and the reality*. *Lancet*, 2010. **375**(9731): p. 2100-9.
 90. Nicol, M.P. and R.J. Wilkinson, *The clinical consequences of strain diversity in Mycobacterium tuberculosis*. *Trans R Soc Trop Med Hyg*, 2008. **102**(10): p. 955-65.
 91. Mathema, B., et al., *Molecular epidemiology of tuberculosis: current insights*. *Clin Microbiol Rev*, 2006. **19**(4): p. 658-85.
 92. Jagielski, T., et al., *Methodological and Clinical Aspects of the Molecular Epidemiology of Mycobacterium tuberculosis and Other Mycobacteria*. *Clin Microbiol Rev*, 2016. **29**(2): p. 239-90.
 93. Hoza, A.S., et al., *Molecular characterization of Mycobacterium tuberculosis isolates from Tanga, Tanzania: First insight of MIRU-VNTR and microarray-based spoligotyping in a high burden country*. *Tuberculosis (Edinb)*, 2016. **98**: p. 116-24.
 94. Li, Y., et al., *Characterization of Mycobacterium tuberculosis isolates from Hebei, China: genotypes and drug susceptibility phenotypes*. *BMC Infect Dis*, 2016. **16**: p. 107.

95. Chaidir, L., et al., *Predominance of modern Mycobacterium tuberculosis strains and active transmission of Beijing sublineage in Jayapura, Indonesia Papua*. Infect Genet Evol, 2016. **39**: p. 187-93.
96. Yimer, S.A., et al., *Mycobacterium tuberculosis lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in Amhara Region, Ethiopia*. J Clin Microbiol, 2015. **53**(4): p. 1301-9.
97. Chen, Y.Y., et al., *Molecular epidemiology of tuberculosis in Kaohsiung City located at southern Taiwan, 2000-2008*. PLoS One, 2015. **10**(1): p. e0117061.
98. Mokrousov, I., et al., *Molecular snapshot of Mycobacterium tuberculosis population structure and drug-resistance in Kyrgyzstan*. Tuberculosis (Edinb), 2013. **93**(5): p. 501-7.
99. Aleksic, E., et al., *First molecular epidemiology study of Mycobacterium tuberculosis in Kiribati*. PLoS One, 2013. **8**(1): p. e55423.
100. Lim, L.K., et al., *Molecular epidemiology of Mycobacterium tuberculosis complex in Singapore, 2006-2012*. PLoS One, 2013. **8**(12): p. e84487.
101. Varghese, B., et al., *Tuberculosis transmission among immigrants and autochthonous populations of the eastern province of Saudi Arabia*. PLoS One, 2013. **8**(10): p. e77635.
102. Cerezo, I., et al., *A first insight on the population structure of Mycobacterium tuberculosis complex as studied by spoligotyping and MIRU-VNTRs in Bogota, Colombia*. Infect Genet Evol, 2012. **12**(4): p. 657-63.
103. Krawczyk, M., et al., *Epidemiological analysis of Mycobacterium tuberculosis strains isolated in Lodz, Poland*. Int J Tuberc Lung Dis, 2011. **15**(9): p. 1252-8, i.
104. Christianson, S., et al., *Evaluation of 24 locus MIRU-VNTR genotyping of Mycobacterium tuberculosis isolates in Canada*. Tuberculosis (Edinb), 2010. **90**(1): p. 31-8.
105. Ojo, O.O., et al., *Molecular epidemiology of Mycobacterium tuberculosis clinical isolates in Southwest Ireland*. Infection, Genetics and Evolution, 2010. **10**(7): p. 1110-1116.
106. Fitzgibbon, M.M., et al., *A snapshot of genetic lineages of Mycobacterium tuberculosis in Ireland over a two-year period, 2010 and 2011*. Euro Surveill, 2013. **18**(3).
107. Rüscher-Gerdes, S., et al., *Multicenter laboratory validation of the BACTEC MGIT 960 technique for testing susceptibilities of Mycobacterium tuberculosis to classical second-line drugs and newer antimicrobials*. J Clin Microbiol, 2006. **44**(3): p. 688-92.
108. Huang, T.S., et al., *Antimicrobial susceptibility testing of Mycobacterium tuberculosis to first-line drugs: comparisons of the MGIT 960 and BACTEC 460 systems*. Ann Clin Lab Sci, 2002. **32**(2): p. 142-7.
109. Pfyffer, G.E., et al., *Multicenter laboratory validation of susceptibility testing of Mycobacterium tuberculosis against classical second-line and newer antimicrobial drugs by using the radiometric BACTEC 460 technique and the proportion method with solid media*. J Clin Microbiol, 1999. **37**(10): p. 3179-86.
110. Brossier, F., et al., *Performance of the genotype MTBDR line probe assay for detection of resistance to rifampin and isoniazid in strains of Mycobacterium tuberculosis with low- and high-level resistance*. J Clin Microbiol, 2006. **44**(10): p. 3659-64.
111. Kiet, V.S., et al., *Evaluation of the MTBDRsl test for detection of second-line-drug resistance in Mycobacterium tuberculosis*. J Clin Microbiol, 2010. **48**(8): p. 2934-9.
112. Walker, T.M., et al., *Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study*. Lancet Infect Dis, 2013. **13**(2): p. 137-46.
113. Votintseva, A.A., et al., *Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures*. J Clin Microbiol, 2015.
114. Pankhurst, L.J., et al., *Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study*. Lancet Respir Med, 2016. **4**(1): p. 49-58.
115. Andrews, S. *FastQC A Quality Control tool for High Throughput Sequence Data*. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
116. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics, 2014. **30**(15): p. 2114-20.
117. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics, 2009. **25**(14): p. 1754-60.
118. Lunter, G. and M. Goodson, *Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads*. Genome Res, 2011. **21**(6): p. 936-9.

119. Swain, M.T., et al., *A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs*. Nat Protoc, 2012. **7**(7): p. 1260-84.
120. Di Camillo, B., et al., *ABACUS: an entropy-based cumulative bivariate statistic robust to rare variants and different direction of genotype effect*. Bioinformatics, 2014. **30**(3): p. 384-91.
121. Schito, M. and D.L. Dolinger, *A Collaborative Approach for 'ReSeq-ing' Mycobacterium tuberculosis Drug Resistance: Convergence for Drug and Diagnostic Developers*. EBioMedicine, 2015. **2**(10): p. 1262-5.
122. Hillemann, D., et al., *Use of the genotype MTBDR assay for rapid detection of rifampin and isoniazid resistance in Mycobacterium tuberculosis complex isolates*. J Clin Microbiol, 2005. **43**(8): p. 3699-703.
123. Hillemann, D., S. Rüsç-Gerdes, and E. Richter, *Feasibility of the GenoType MTBDRsl assay for fluoroquinolone, amikacin-capreomycin, and ethambutol resistance testing of Mycobacterium tuberculosis strains and clinical specimens*. J Clin Microbiol, 2009. **47**(6): p. 1767-72.
124. Galagan, J.E., *Genomic insights into tuberculosis*. Nat Rev Genet, 2014. **15**(5): p. 307-20.
125. Gagneux, S., et al., *Variable host-pathogen compatibility in Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A, 2006. **103**(8): p. 2869-73.
126. Jagielski, T., et al., *Current methods in the molecular typing of Mycobacterium tuberculosis and other mycobacteria*. Biomed Res Int, 2014. **2014**: p. 645802.
127. Cannas, A., et al., *Molecular Typing of Mycobacterium Tuberculosis Strains: A Fundamental Tool for Tuberculosis Control and Elimination*. Infect Dis Rep, 2016. **8**(2): p. 6567.
128. de Beer, J.L., et al., *Second worldwide proficiency study on variable number of tandem repeats typing of Mycobacterium tuberculosis complex*. Int J Tuberc Lung Dis, 2014. **18**(5): p. 594-600.
129. Allix-Beguec, C., et al., *Standardised PCR-based molecular epidemiology of tuberculosis*. Eur Respir J, 2008. **31**(5): p. 1077-84.
130. Borgdorff, M.W. and D. van Soolingen, *The re-emergence of tuberculosis: what have we learnt from molecular epidemiology?* Clin Microbiol Infect, 2013. **19**(10): p. 889-901.
131. Alonso-Rodriguez, N., et al., *Prospective universal application of mycobacterial interspersed repetitive-unit-variable-number tandem-repeat genotyping to characterize Mycobacterium tuberculosis isolates for fast identification of clustered and orphan cases*. J Clin Microbiol, 2009. **47**(7): p. 2026-32.
132. Kronfol, N.M. and Z. Mansour, *Tuberculosis and migration: a review*. East Mediterr Health J, 2013. **19**(8): p. 739-48.
133. World Health Organisation, T.G.F., *Tuberculosis Financing and Funding Gaps in 118 Countries Eligible for Global Fund Support*, in Online. 2016, Stop TB Partnership: Online.
134. Jamieson, F.B., et al., *Whole-genome sequencing of the Mycobacterium tuberculosis Manila sublineage results in less clustering and better resolution than mycobacterial interspersed repetitive-unit-variable-number tandem-repeat (MIRU-VNTR) typing and spoligotyping*. J Clin Microbiol, 2014. **52**(10): p. 3795-8.
135. Stucki, D., et al., *Standard Genotyping Overestimates Transmission of Mycobacterium tuberculosis among Immigrants in a Low-Incidence Country*. J Clin Microbiol, 2016. **54**(7): p. 1862-70.
136. Comas, I., et al., *Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies*. PLoS One, 2009. **4**(11): p. e7815.
137. Shah, N.S., et al., *Yield of Contact Investigations in Households of Patients With Drug-Resistant Tuberculosis: Systematic Review and Meta-Analysis*. Clinical Infectious Diseases, 2014. **58**(3): p. 381-391.
138. Coscolla, M. and S. Gagneux, *Consequences of genomic diversity in Mycobacterium tuberculosis*. Semin Immunol, 2014. **26**(6): p. 431-44.
139. Merker, M., et al., *Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage*. Nat Genet, 2015.
140. Pringle, D., *The resurgence of tuberculosis in the Republic of Ireland: perceptions and reality*. Soc Sci Med, 2009. **68**(4): p. 620-4.

141. Wyres, K., et al., *Whole Genome Sequencing Analysis and Interpretation in Clinical and Public Health Microbiology Laboratories: What Are the Requirements and How Do Existing Tools Compare?* Pathogens, 2014. **3**(2): p. 437-458.
142. Gardy, J.L., et al., *Whole-genome sequencing and social-network analysis of a tuberculosis outbreak.* N Engl J Med, 2011. **364**(8): p. 730-9.
143. Roetzer, A., et al., *Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study.* PLoS Med, 2013. **10**(2): p. e1001387.
144. Koser, C.U., et al., *Routine use of microbial whole genome sequencing in diagnostic and public health microbiology.* PLoS Pathog, 2012. **8**(8): p. e1002824.
145. Hatherell, H.A., et al., *Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review.* BMC Med, 2016. **14**: p. 21.
146. Bryant, J.M., et al., *Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data.* BMC Infect Dis, 2013. **13**: p. 110.
147. Kato-Maeda, M., et al., *Use of whole genome sequencing to determine the microevolution of Mycobacterium tuberculosis during an outbreak.* PLoS One, 2013. **8**(3): p. e58235.
148. Torok, M.E., et al., *Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak.* J Clin Microbiol, 2013. **51**(2): p. 611-4.
149. Walker, T.M., et al., *Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study.* Lancet Respir Med, 2014. **2**(4): p. 285-92.
150. Dolan, K., et al., *Global burden of HIV, viral hepatitis, and tuberculosis in prisoners and detainees.* The Lancet.
151. Sacchi, F.P., et al., *Prisons as reservoir for community transmission of tuberculosis, Brazil.* Emerg Infect Dis, 2015. **21**(3): p. 452-5.
152. Lafontaine, D., et al., *Treatment of multidrug-resistant tuberculosis in Russian prisons.* Lancet, 2004. **363**(9404): p. 246-7.
153. Aris, B., *Russia's health crisis fuels 20-year cut in lifespan estimates.* Lancet, 2003. **362**(9395): p. 1557.
154. O'Toole, R.F., et al., *Tuberculosis incidence in the Irish Traveller population in Ireland from 2002 to 2013.* Epidemiol Infect, 2015: p. 1-7.
155. Inoue, T., *[Difference in transmissibility between bronchial and laryngeal tuberculosis--a retrospective epidemiological study of TB patients newly registered in recent 19 years in Aichi Prefecture, Japan].* Kekkaku, 2006. **81**(6): p. 419-24.
156. Hatherell, H.-A., et al., *Declaring a tuberculosis outbreak over with genomic epidemiology.* Microbial Genomics, 2016. **2**(5).
157. Caminero, J.A., et al., *Epidemiological evidence of the spread of a Mycobacterium tuberculosis strain of the Beijing genotype on Gran Canaria Island.* Am J Respir Crit Care Med, 2001. **164**(7): p. 1165-70.
158. Walker, T.M., et al., *Contact investigations for outbreaks of Mycobacterium tuberculosis: advances through whole genome sequencing.* Clin Microbiol Infect, 2013. **19**(9): p. 796-802.
159. Couvin, D. and N. Rastogi, *Tuberculosis - A global emergency: Tools and methods to monitor, understand, and control the epidemic with specific example of the Beijing lineage.* Tuberculosis (Edinb), 2015. **95 Suppl 1**: p. S177-89.
160. Singhal, P., et al., *A study on pre-XDR & XDR tuberculosis & their prevalent genotypes in clinical isolates of Mycobacterium tuberculosis in north India.* Indian J Med Res, 2016. **143**(3): p. 341-7.
161. Zhao, L.L., et al., *Molecular characterisation of extensively drug-resistant Mycobacterium tuberculosis isolates in China.* Int J Antimicrob Agents, 2015. **45**(2): p. 137-43.
162. Caceres, O., et al., *Characterization of the Genetic Diversity of Extensively-Drug Resistant Mycobacterium tuberculosis Clinical Isolates from Pulmonary Tuberculosis Patients in Peru.* PLoS One, 2014. **9**(12): p. e112789.
163. O'Toole, R.F., et al., *Draft Genome Sequence of the First Isolate of Extensively Drug-Resistant Mycobacterium tuberculosis in New Zealand.* Genome Announc, 2014. **2**(3).
164. Roycroft, E., et al., *Draft Genome Sequence of the First Isolate of Extensively Drug-Resistant Mycobacterium tuberculosis in Ireland.* Genome Announc, 2014. **2**(5).

165. Silva, C., et al., *Mycobacterial interspersed repetitive unit typing and mutational profile for multidrug-resistant and extensively drug-resistant tuberculosis surveillance in Portugal: a 3-year period overview*. Int J Antimicrob Agents, 2014. **44**(6): p. 546-51.
166. Poudel, A., et al., *Characterization of extensively drug-resistant Mycobacterium tuberculosis in Nepal*. Tuberculosis (Edinb), 2013. **93**(1): p. 84-8.
167. Diel, R., et al., *Costs of tuberculosis disease in the European Union: a systematic analysis and cost calculation*. Eur Respir J, 2014. **43**(2): p. 554-65.
168. Chedore, P., et al., *Potential for erroneous results indicating resistance when using the Bactec MGIT 960 system for testing susceptibility of Mycobacterium tuberculosis to pyrazinamide*. J Clin Microbiol, 2010. **48**(1): p. 300-1.
169. Zhang, Y., S. Permar, and Z. Sun, *Conditions that may affect the results of susceptibility testing of Mycobacterium tuberculosis to pyrazinamide*. J Med Microbiol, 2002. **51**(1): p. 42-9.
170. Boehme, C.C., et al., *Rapid molecular detection of tuberculosis and rifampin resistance*. N Engl J Med, 2010. **363**(11): p. 1005-15.
171. Miller, L.P., J.T. Crawford, and T.M. Shinnick, *The rpoB gene of Mycobacterium tuberculosis*. Antimicrob Agents Chemother, 1994. **38**(4): p. 805-11.
172. Rouse, D.A., et al., *Characterization of the katG and inhA genes of isoniazid-resistant clinical isolates of Mycobacterium tuberculosis*. Antimicrob Agents Chemother, 1995. **39**(11): p. 2472-7.
173. Cambau, E., W. Sougakoff, and V. Jarlier, *Amplification and nucleotide sequence of the quinolone resistance-determining region in the gyrA gene of mycobacteria*. FEMS Microbiol Lett, 1994. **116**(1): p. 49-54.
174. Alcaide, F., G.E. Pfyffer, and A. Telenti, *Role of embB in natural and acquired resistance to ethambutol in mycobacteria*. Antimicrob Agents Chemother, 1997. **41**(10): p. 2270-3.
175. Sreevatsan, S., et al., *Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations*. Antimicrob Agents Chemother, 1997. **41**(8): p. 1677-81.
176. Sreevatsan, S., et al., *Characterization of rpsL and rrs mutations in streptomycin-resistant Mycobacterium tuberculosis isolates from diverse geographic localities*. Antimicrob Agents Chemother, 1996. **40**(4): p. 1024-6.
177. Ramaswamy, S. and J.M. Musser, *Molecular genetic basis of antimicrobial agent resistance in Mycobacterium tuberculosis: 1998 update*. Tuber Lung Dis, 1998. **79**(1): p. 3-29.
178. Brossier, F., et al., *Performance of the New Version (v2.0) of the GenoType MTBDRsl Test for Detection of Resistance to Second-Line Drugs in Multidrug-Resistant Mycobacterium tuberculosis Complex Strains*. J Clin Microbiol, 2016. **54**(6): p. 1573-80.
179. Takiff, H.E., et al., *Cloning and nucleotide sequence of Mycobacterium tuberculosis gyrA and gyrB genes and detection of quinolone resistance mutations*. Antimicrob Agents Chemother, 1994. **38**(4): p. 773-80.
180. Zaunbrecher, M.A., et al., *Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A, 2009. **106**(47): p. 20004-9.
181. Didelot, X., et al., *Transforming clinical microbiology with bacterial genome sequencing*. Nat Rev Genet, 2012. **13**(9): p. 601-12.
182. Reuter, S., et al., *Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology*. JAMA Intern Med, 2013. **173**(15): p. 1397-404.
183. Cole, S.T. and B.G. Barrell, *Analysis of the genome of Mycobacterium tuberculosis H37Rv*. Novartis Found Symp, 1998. **217**: p. 160-72; discussion 172-7.
184. Walker, T.M., et al., *Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study*. The Lancet Infectious Diseases, (0).
185. Min, S.M., et al., *Antibiotic resistant tuberculosis and bovine tuberculosis in an Irish hospital population (1991 to 2001)*. Ir Med J, 2005. **98**(2): p. 38-40.
186. Collins, C., et al., *Is bovine, atypical or resistant tuberculosis a problem?* Ir Med J, 1987. **80**(2): p. 66-7.
187. Crudu, V., et al., *Nosocomial transmission of multidrug-resistant tuberculosis*. Int J Tuberc Lung Dis, 2015. **19**(12): p. 1520-3.

188. ECDC, *Molecular Typing for Surveillance of Multidrug-resistant tuberculosis in the EU/EEA*. January 2016, European Centre for Disease Control: Online.
189. Mokrousov, I., et al., *Detection of embB306 mutations in ethambutol-susceptible clinical isolates of Mycobacterium tuberculosis from Northwestern Russia: implications for genotypic resistance testing*. J Clin Microbiol, 2002. **40**(10): p. 3810-3.
190. Tracevska, T., et al., *Characterisation of rpsL, rrs and embB mutations associated with streptomycin and ethambutol resistance in Mycobacterium tuberculosis*. Res Microbiol, 2004. **155**(10): p. 830-4.
191. Sandgren, A., et al., *Tuberculosis drug resistance mutation database*. PLoS Med, 2009. **6**(2): p. e2.
192. Koser, C.U., et al., *Whole-genome sequencing for rapid susceptibility testing of M. tuberculosis*. N Engl J Med, 2013. **369**(3): p. 290-2.
193. Rodwell, T.C., et al., *Predicting extensively drug-resistant Mycobacterium tuberculosis phenotypes with genetic mutations*. J Clin Microbiol, 2014. **52**(3): p. 781-9.
194. Rosales-Klintz, S., et al., *Drug resistance-related mutations in multidrug-resistant Mycobacterium tuberculosis isolates from diverse geographical regions*. Int J Mycobacteriol, 2012. **1**(3): p. 124-30.
195. *Acquired Resistance to Bedaquiline and Delamanid in Therapy for Tuberculosis*. N Engl J Med, 2015. **373**(25): p. e29.
196. Vilcheze, C. and W.R. Jacobs, Jr., *Resistance to Isoniazid and Ethionamide in Mycobacterium tuberculosis: Genes, Mutations, and Causalities*. Microbiol Spectr, 2014. **2**(4): p. MGM2-0014-2013.
197. Zhang, Q., et al., *Whole genome analysis of an MDR Beijing/W strain of Mycobacterium tuberculosis with large genomic deletions associated with resistance to isoniazid*. Gene, 2016. **582**(2): p. 128-36.
198. Regmi, S.M., et al., *Polymorphisms in drug-resistant-related genes shared among drug-resistant and pan-susceptible strains of sequence type 10, Beijing family of Mycobacterium tuberculosis*. Int J Mycobacteriol, 2015. **4**(1): p. 67-72.
199. Ramaswamy SV, Amin AG, Göksel S, et al. *Molecular Genetic Analysis of Nucleotide Polymorphisms Associated with Ethambutol Resistance in Human Isolates of Mycobacterium tuberculosis*. Antimicrobial Agents and Chemotherapy. 2000;44(2):326-336.
200. Phelan, J., et al., *Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance*. BMC Med, 2016. **14**: p. 31.
201. Torres, J.N., et al., *Novel katG mutations causing isoniazid resistance in clinical M. tuberculosis isolates*. Emerg Microbes Infect, 2015. **4**(7): p. e42.
202. Jagielski, T., et al., *Detection of mutations associated with isoniazid resistance in multidrug-resistant Mycobacterium tuberculosis clinical isolates*. J Antimicrob Chemother, 2014. **69**(9): p. 2369-75.
203. Sun, Y.J., et al., *Analysis of the role of Mycobacterium tuberculosis kasA gene mutations in isoniazid resistance*. Clin Microbiol Infect, 2007. **13**(8): p. 833-5.
204. Lin, N., et al., *Draft genome sequences of two super-extensively drug-resistant isolates of Mycobacterium tuberculosis from China*. FEMS Microbiol Lett, 2013. **347**(2): p. 93-6.
205. Ravibalan, T., et al., *Characterization of katG and rpoB gene mutations in multi drug resistant Mycobacterium tuberculosis clinical isolates*. Int. J. Curr. Microbiol. Appl. Sci, 2014. **3**: p. 1072-1080.
206. Dalla Costa, E.R., et al., *Correlations of mutations in katG, oxyR-ahpC and inhA genes and in vitro susceptibility in Mycobacterium tuberculosis clinical strains segregated by spoligotype families from tuberculosis prevalent countries in South America*. BMC Microbiol, 2009. **9**: p. 39.
207. de Beer, J.L., et al., *A putative compensatory mutation and a specific marker exclusively detected in isolates of the European Beijing MDR-TB outbreak strain*.
208. de Vos, M., et al., *Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission*. Antimicrob Agents Chemother, 2013. **57**(2): p. 827-32.

209. Comas, I., et al., *Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes*. Nat Genet, 2012. **44**(1): p. 106-10.
210. Jeeves, R.E., et al., *Mycobacterium tuberculosis Is Resistant to Isoniazid at a Slow Growth Rate by Single Nucleotide Polymorphisms in katG Codon Ser315*. PLoS One, 2015. **10**(9): p. e0138253.
211. Mc Laughlin, A.M., et al., *Extensively drug-resistant tuberculosis (XDR-TB) - a potential threat in Ireland*. Open Respir Med J, 2007. **1**: p. 7-9.
212. Kennedy, B., et al., *Extensively drug-resistant tuberculosis: first report of a case in Ireland*. Euro Surveill, 2008. **13**(30).
213. Feuerriegel, S., C.U. Koser, and S. Niemann, *Phylogenetic polymorphisms in antibiotic resistance genes of the Mycobacterium tuberculosis complex*. J Antimicrob Chemother, 2014. **69**(5): p. 1205-10.
214. Hardy, A.B., et al., *Cost-effectiveness of the NICE guidelines for screening for latent tuberculosis infection: the QuantiFERON-TB Gold IGRA alone is more cost-effective for immigrants from high burden countries*. Thorax, 2010. **65**(2): p. 178-80.
215. Burgess, D., *Immigrant Health in Toronto, Canada: Addressing Food Insecurity as a Social Determinant of Tuberculosis*. Soc Work Public Health, 2016: p. 1-9.
216. Haukaas, F.S., et al., *Immigrant screening for latent tuberculosis in Norway: a cost-effectiveness analysis*. Eur J Health Econ, 2016.
217. van der Werf, M.J. and P. Kramarz, *Tackling tuberculosis in migrants*. Lancet Infect Dis, 2016. **16**(8): p. 877-8.
218. Glynn, J.R., et al., *Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review*. Emerg Infect Dis, 2002. **8**(8): p. 843-9.
219. Homolka, S., et al., *High resolution discrimination of clinical Mycobacterium tuberculosis complex strains based on single nucleotide polymorphisms*. PLoS One, 2012. **7**(7): p. e39855.
220. Maiden, M.C., et al., *MLST revisited: the gene-by-gene approach to bacterial genomics*. Nat Rev Microbiol, 2013. **11**(10): p. 728-36.
221. Christianson, S., et al., *Re-evaluation of the critical concentration for ethambutol antimicrobial sensitivity testing on the MGIT 960*. PLoS One, 2014. **9**(9): p. e108911.
222. Hasnain, S.E., et al., *Whole genome sequencing: A new paradigm in the surveillance and control of human tuberculosis*. Tuberculosis (Edinb), 2014.
223. Witney, A.A., et al., *Clinical use of whole genome sequencing for Mycobacterium tuberculosis*. BMC Med, 2016. **14**: p. 46.
224. PHE, P.H.E., *Annual TB Update 2015*. 2015, Public Health England: Online.
225. Witney, A.A., et al., *Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases*. J Clin Microbiol, 2015. **53**(5): p. 1473-83.
226. Koser, C.U., M.J. Ellington, and S.J. Peacock, *Whole-genome sequencing to control antimicrobial resistance*. Trends Genet, 2014. **30**(9): p. 401-7.
227. Vilcheze, C., et al., *Altered NADH/NAD+ ratio mediates coresistance to isoniazid and ethionamide in mycobacteria*. Antimicrob Agents Chemother, 2005. **49**(2): p. 708-20.
228. Starks, A.M., et al., *Mutations at embB codon 306 are an important molecular indicator of ethambutol resistance in Mycobacterium tuberculosis*. Antimicrob Agents Chemother, 2009. **53**(3): p. 1061-6.
229. Srivastava, S., et al., *emb nucleotide polymorphisms and the role of embB306 mutations in Mycobacterium tuberculosis resistance to ethambutol*. Int J Med Microbiol, 2009. **299**(4): p. 269-80.
230. Hasman, H., et al., *Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples*. J Clin Microbiol, 2014. **52**(1): p. 139-46.
231. Brown, A.C., et al., *Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples*. J Clin Microbiol, 2015. **53**(7): p. 2230-7.
232. Watts, M.R., et al., *Editor's choice: Implications of isoniazid resistance in Mycobacterium bovis Bacillus Calmette-Guérin used for immunotherapy in bladder cancer*. Clin Infect Dis, 2011. **52**(1): p. 86-8.

233. Kolibab, K., S.C. Derrick, and S.L. Morris, *Sensitivity to isoniazid of Mycobacterium bovis BCG strains and BCG disseminated disease isolates*. J Clin Microbiol, 2011. **49**(6): p. 2380-1.
234. Ritz, N., et al., *Susceptibility of Mycobacterium bovis BCG vaccine strains to antituberculous antibiotics*. Antimicrob Agents Chemother, 2009. **53**(1): p. 316-8.
235. Al-Hajoj, S., et al., *Molecular confirmation of Bacillus Calmette Guerin vaccine related adverse events among Saudi Arabian children*. PLoS One, 2014. **9**(11): p. e113472.
236. Durek, C., et al., *Sensitivity of BCG to modern antibiotics*. Eur Urol, 2000. **37 Suppl 1**: p. 21-5.
237. Arend, S.M. and D. van Soolingen, *Editor's choice: Editorial commentary: Low level INH-resistant BCG: a sheep in wolf's clothing?* Clin Infect Dis, 2011. **52**(1): p. 89-93.
238. Johnson, R., et al., *Ethambutol resistance testing by mutation detection*. Int J Tuberc Lung Dis, 2006. **10**(1): p. 68-73.
239. Georghiou, S.B., et al., *Evaluation of genetic mutations associated with Mycobacterium tuberculosis resistance to amikacin, kanamycin and capreomycin: a systematic review*. PLoS One, 2012. **7**(3): p. e33275.
240. Ortblad, K.F., et al., *Stopping tuberculosis: a biosocial model for sustainable development*. Lancet, 2015. **386**(10010): p. 2354-62.
241. WHO. *World Health Organisation Global Tuberculosis Report 2014*; Available from: http://www.who.int/tb/publications/global_report/en/.
242. Public Health Agency, N.I., *Epidemiology of tuberculosis in Northern Ireland: Annual surveillance report 2014*. 2016, HSC Public Health Agency, NI: Online.
243. NICE, *Tuberculosis, NICE Guideline, NG33*, in *NICE Guidelines*. 2016, National Institute for Health and Care Excellence: Online.
244. Takiff, H.E. and O. Feo, *Clinical value of whole-genome sequencing of Mycobacterium tuberculosis*. Lancet Infect Dis, 2015. **15**(9): p. 1077-90.
245. Robinson, E.R., T.M. Walker, and M.J. Pallen, *Genomics and outbreak investigation: from sequence to consequence*. Genome Med, 2013. **5**(4): p. 36.
246. Blankley, S., et al., *The application of transcriptional blood signatures to enhance our understanding of the host response to infection: the example of tuberculosis*. Philos Trans R Soc Lond B Biol Sci, 2014. **369**(1645): p. 20130427.
247. Bidovec-Stojkovič, U., et al., *Prospective genotyping of Mycobacterium tuberculosis from fresh clinical samples*. PLoS One, 2014. **9**(10): p. e109547.
248. Jankute, M., et al., *Assembly of the Mycobacterial Cell Wall*. Annu Rev Microbiol, 2015. **69**: p. 405-23.
249. Hershberg, R., et al., *High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography*. PLoS Biol, 2008. **6**(12): p. e311.
250. Barnes, P.F. and M.D. Cave, *Molecular epidemiology of tuberculosis*. N Engl J Med, 2003. **349**(12): p. 1149-56.
251. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
252. Krampis, K., et al., *Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community*. BMC Bioinformatics, 2012. **13**: p. 42.
253. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
254. Rutherford, K., et al., *Artemis: sequence visualization and annotation*. Bioinformatics, 2000. **16**(10): p. 944-945.
255. Drummond, A.J. and A. Rambaut, *BEAST: Bayesian evolutionary analysis by sampling trees*. BMC Evol Biol, 2007. **7**: p. 214.
256. Allix-Béguec, C., et al., *Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of Mycobacterium tuberculosis complex isolates*. J Clin Microbiol, 2008. **46**(8): p. 2692-9.

Appendices

Appendix 1.

Euro-Surveillance Paper, published 2013

A snapshot of genetic lineages of *Mycobacterium tuberculosis* in Ireland over a two-year period, 2010 and 2011

M M Fitzgibbon (MFitzgibbon@STJAMES.IE)¹, N Gibbons¹, E Roycroft^{1,2}, S Jackson³, J O'Donnell³, D O'Flanagan³, T R Rogers^{1,2}

1. Irish Mycobacteria Reference Laboratory, St. James' Hospital, Dublin, Ireland

2. Department of Clinical Microbiology, Trinity College, Dublin, Ireland

3. Health Protection Surveillance Centre, Dublin, Ireland

Citation style for this article:

Fitzgibbon MM, Gibbons N, Roycroft E, Jackson S, O'Donnell J, O'Flanagan D, Rogers TR. A snapshot of genetic lineages of *Mycobacterium tuberculosis* in Ireland over a two-year period, 2010 and 2011. *Euro Surveill.* 2013;18(3):pii=20367. Available online: <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20367>

Article submitted on 09 July 2012 / published on 17 January 2013

Mycobacterial interspersed repetitive-unit-variable-number tandem repeat typing alone was used to investigate the genetic lineages among 361 *Mycobacterium tuberculosis* strains circulating in Ireland over a two-year period, 2010 and 2011. The majority of isolates, 63% (229/361), belonged to lineage 4 (Euro-American), while lineages 1 (Indo-Oceanic), 2 (East-Asian) and 3 (East-African–Indian) represented 12% of isolates each (42/361, 45/361, and 45/361, respectively). Sub-lineages Beijing (lineage 2), East-African–Indian (lineage 1) and Delhi/central-Asian (lineage 3) predominated among foreign-born cases, while a higher proportion of Euro-American lineages were identified among cases born in Ireland. Eighteen molecular clusters involving 63 tuberculosis (TB) cases were identified across four sub-lineages of lineage 4. While the mean cluster size was 3.5 TB cases, the largest cluster (involving 12 Irish-born cases) was identified in the Latin American–Mediterranean sub-lineage. Clustering of isolates was higher among Irish-born TB cases (47 of 63 clustered cases), whereas only one cluster (3/63) involved solely foreign-born individuals. Four multidrug-resistant cases identified during this period represented lineages 2 and 4. This study provides the first insight into the structure of the *M. tuberculosis* population in Ireland.

Introduction

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* remains a serious challenge to public health worldwide. Despite an overall decline in case notification rates for TB across Europe, rates vary significantly, with the highest rates reported from eastern Europe and the Baltic States [1]. Multidrug-resistant (MDR) TB continues to be a major problem and an added burden in high-incidence countries such as Romania (108.2/100,000) and the Baltic states Lithuania (62.1/100,000), Latvia (43.2/100,000) and Estonia (30.7/100,000) [1]. Ireland has a low incidence, with notification rates that ranged between 9.7 and 11.3 per 100,000 population between 2001 and 2010 although to some extent this may have

been influenced by migrants arriving from high-burden countries [1-2]. In Ireland, TB is a statutorily notifiable disease, and in a recent report on the epidemiology of TB in the country, the proportion of culture-confirmed TB cases was 71.2% in 2009 and 63.2% in 2010 (data from 2010 were not finalised at the time of submission) [3]. These proportions are similar to those reported for previous years [3].

Disruption of transmission chains is a key factor in controlling TB both at a national and international level [4]. In recent years, there have been significant advances in developing the molecular tools required for rapid diagnosis of TB [4]. Analysis of variable-number tandem repeat (VNTR) sequences at mycobacterial interspersed repetitive units (MIRU) has emerged as a valuable marker for genotyping strains of the *M. tuberculosis* complex [5]. In large population-based studies, MIRU-VNTR typing has been shown to have similar discriminatory power when compared to IS6110 restriction fragment length polymorphism (RFLP) typing [5-7]. An optimised set of 24 MIRU-VNTR markers has become the gold standard for genotyping *M. tuberculosis* complex strains worldwide [5,7]. MIRU-VNTR typing is a PCR-based method that yields rapid, reproducible results that are expressed as a 24-digit numerical code which allows for easy exchange of data [5-8]. This method can be applied to early mycobacterial cultures and more recently has been successfully applied directly to smear-positive specimens [5,9]. Previous studies have shown MIRU-VNTR typing to be useful in comparing strains (i) at national and international level, (ii) among household contacts, (iii) associated with drug resistance and (iv) to determine the evolutionary pathway of TB [5-6,10-16]. At European level, MIRU-VNTR typing has been adopted for the molecular surveillance of the international transmission of MDR-TB and extensively drug-resistant TB [7].

In November 2009, molecular genotyping in the form of 24-locus MIRU-VNTR typing was introduced at the

Irish Mycobacteria Reference Laboratory (IMRL) which both cultures and receives isolates from microbiology laboratories around the country. To allow for rapid, high-throughput genotyping of *M. tuberculosis*, the commercial MIRU-VNTR typing kit (GenoScreen, Lille, France) was introduced in 2010 [5,7]. As 24-locus MIRU-VNTR typing is considered the gold standard genotyping method, all *M. tuberculosis* isolates identified at the IMRL are currently typed prospectively with this method, and it is envisaged that all *M. tuberculosis* isolates recovered since 2000 will be typed on a retrospective basis.

Here we report the first analysis of the structure of the *M. tuberculosis* population in Ireland for isolates recovered during 2010 and 2011 following the introduction of 24-locus MIRU-VNTR typing to the diagnostic laboratory. It needs to be noted that at the time there was an under-representation of isolates from the southern region of Ireland.

Methods

MIRU-VNTR typing

M. tuberculosis isolates (n=361) recovered in or referred to the IMRL over a two-year period (2010–11) were typed with the MIRU-VNTR typing kit (GenoScreen) [5]. Validation of the MIRU-VNTR technique was performed using the MIRU-VNTR Calibration Kit (GenoScreen). PCR products were subjected to electrophoresis using a 3130 genetic analyser (Applied Biosystems). Sizing of fragments and MIRU-VNTR allele assignment was performed using GeneMapper software (Applied Biosystems). Phylogenetic lineages were assigned to each isolate using the MIRU-VNTR*plus* online tool [17–18].

The 24-locus MIRU-VNTR panel comprised the following loci: MIRU 02, VNTR 42, VNTR 43, MIRU 04, MIRU 40, MIRU 10, MIRU 16, VNTR 1955, MIRU 20, QUB 11b, ETR A, VNTR 46, VNTR 47, VNTR 48, MIRU 23, MIRU 24, MIRU 26, MIRU 27, VNTR 49, MIRU 31, VNTR 52, QUB 26, VNTR 53, and MIRU 39. The MIRU-VNTR profiles are reported as a series of 24 numbers that correspond to the number of alleles at each of the loci described above.

Clusters of isolates were defined as two or more isolates with indistinguishable MIRU-VNTR patterns. The strain clustering rate was calculated as $(n_c - c)/n$, where n_c was the total number of strain-clustered cases, c was the number of clusters and n was the total number of isolates [5].

Epidemiological analysis

Enhanced surveillance of TB was implemented in Ireland in 1998. Enhanced TB notification forms are completed by public health doctors, summarising all available clinical, microbiological, histological and epidemiological information. These data are collated in the regional public health departments. Anonymised

data are then submitted electronically to the Health Protection Surveillance Centre (HPSC) for the production of reports on a weekly, quarterly and annual basis.

Since January 2011, cases of TB have been reported through the Computerised Infectious Disease Reporting system (CIDR). CIDR is a web-based system developed to integrate case-based clinical and laboratory data in order to manage the surveillance and control of notifiable infectious diseases in Ireland. Prior to using CIDR for TB surveillance, MIRU-VNTR typing results were not linked to case-based epidemiological data. In addition to recording sporadic case-based data, CIDR also facilitates the reporting of clustered cases, according to Irish outbreak case definitions [2]. Clustered cases can be reported via a summary aggregate outbreak data module to which the relevant disaggregate case-based surveillance data can also be linked.

Data analysis

The TB enhanced surveillance data for 2011 (epidemiological and linked laboratory data) used in this publication were extracted from CIDR on 17 April 2012 using Business Objects XI software and were analysed using Microsoft Excel. Data for 2011 were provisional at the time of extraction and subject to ongoing validation and revision.

Results

Results are presented in two separate sections. In the first part, genotyping results for isolates recovered in 2010–11 are presented. As epidemiological data linking was available from 2011 onwards, the second part of the results section (enhanced surveillance) refers to genotyping results linked to epidemiological data for 2011 isolates only.

Mycobacterium tuberculosis genotyping, 2010–11

Some 361 *M. tuberculosis* isolates were recovered in or referred to the IMRL during 2010–11, representing 63.6% of culture-positive cases identified through the national TB surveillance system in that period. Genotyping of *M. tuberculosis* isolates recovered during the study period yielded four global lineages (Table 1). The majority (63%) belonged to lineage 4 (Euro-American), while lineages 1 (Indo-Oceanic), 2 (East-Asian) and 3 (East-African–Indian) represented 12% each. Among the 229 Euro-American strains, sub-lineages Latin American–Mediterranean (LAM) (23%), Haarlem (21%), H37Rv (19%) and Haarlem/X (13%) were most prevalent (Table 1).

Within lineage 4, 18 clusters were identified involving 63 TB cases (Table 2). The strain clustering rate varied between different sub-lineages, but was highest for the LAM sub-lineage (6.9%). While the mean cluster size was 3.5 TB cases, the largest cluster (involving 12 Irish-born cases, representing 19% of all clustered isolates) was identified within the LAM sub-lineage (Table

TABLE 1

Distribution of lineages among *Mycobacterium tuberculosis* isolates, Ireland, 2010–11 (n=361)

Global lineage	Sub-lineage	No. of isolates	% of isolates
1 Indo-Oceanic	East-African–Indian	42	12
2 East-Asian	Beijing	45	12
3 East–African–Indian	Delhi/central-Asian	45	12
4 Euro-American	Lineage 4 total	229	63
	Latin American–Mediterranean	52	
	Haarlem	47	
	H37Rv	44	
	Haarlem/X	29	
	Cameroon	13	
	S	6	
	TUR	8	
	X	5	
	Ghana	3	
	URAL	3	
	Uganda I & II	6	
	NEW-1	2	
Others ^a	11		
Total		361	100

^a The category Others includes isolates for which the sub-lineages were not clearly defined

2). Four other clusters within the LAM sub-lineage contained between 4.7% (3/63) and 12.7% (8/63) of clustered cases. Among the clustered cases of Haarlem, H37Rv and Haarlem/X, cluster sizes ranged from 3.2 to 9.5%, at 3.2% and from 3.2 to 4.7% of isolates, respectively.

Only one cluster (3/63) contained exclusively foreign-born individuals, 12 clusters (47/63) involved Irish-born cases only, while five clusters (13/63) were mixed. In addition, one small cluster was observed among the isolates from lineage 2.

The four MDR-TB cases identified during this period represented lineages 2 (Beijing) and 4 (Ural, H37Rv and LAM). None of the MDR-TB cases were clustered.

Tuberculosis enhanced surveillance data for 2011 isolates (epidemiological and laboratory)

In 2011, 432 TB cases were provisionally reported on CIDR, of which approximately 166 (38%) were typed. At the time of data extraction, 136 TB cases were updated to include MIRU typing results (representing 81.9% of 166 typed isolates). Of the 136 TB cases with a MIRU typing result, 34 were clustered in 11 clusters with different MIRU types. Clusters ranged in size from eight to

two TB cases. Of the 11 MIRU type clusters, five, comprising 18 TB cases, were confirmed by public health departments as outbreaks meeting the Irish case definition.

The Beijing sub-lineage was most prevalent (15.4%) and associated with a small cluster. Sub-lineages Haarlem, LAM, and H37Rv were most prevalent among lineage 4 strains, while lineage 1 and lineage 3 represented 11.8% and 10.3% of typed isolates, respectively. Interestingly, isolates recovered from pulmonary specimens were mostly correlated with lineage 4 strains, while the majority of isolates recovered from extra-pulmonary specimens belonged to lineages 1 and 3 (Figure 1). In lineage 3, nine of 14 isolates were recovered from patients born in Pakistan, while the remaining five isolates were recovered from patients born in India (n=2), Kenya (n=1), Nepal (n=1) and Nigeria (n=1). Only one lineage 1 isolate was recovered from an Irish-born patient, while six were recovered from patients born in the Philippines. Other countries represented among lineage 1 isolates were Bangladesh, India, Mozambique, Pakistan, Somalia and Vietnam.

The distribution of lineages among Irish-born and foreign-born TB cases is shown in Figure 2. Of the 127 TB cases for whom MIRU-VNTR and country of birth were known, 51.5% were foreign-born and 41% were Irish-born. Lineages 1, 2 and 3 predominated among foreign-born TB cases, while a higher proportion of lineage 4 isolates were identified among Irish-born cases.

Discussion

This study has provided a snapshot of the genetic diversity of *M. tuberculosis* in Ireland. Due to the small numbers of isolates in our study, statistical analysis would not be significant and was not performed. Although data on sub-lineages were analysed by age and sex, the resulting frequencies were too small to draw firm conclusions from. However, when age and sex analyses were further stratified by country of birth, these data were broadly similar to the age and sex profile of the Irish TB notification data.

A large diverse group of isolates has been identified, suggesting a low degree of active transmission among TB patients. The distribution of genetic lineages is similar to other recent studies that used different typing techniques and in which lineage 4 (Euro-American) predominated among circulating *M. tuberculosis* strains [12,19–21]. In previous work conducted in the south-west region of Ireland, lineage 4 predominated, and clustering of isolates was associated with Irish nationals and lineage 4 isolates only [22]. In our study, the distribution of genetic lineages among extra-pulmonary specimens (where lineages 1 and 3 predominated) was similar to a recently published large-scale study conducted in the United States (US) investigating the relationship between genetic lineages and clinical sites of infection [23]. In the US study, the highest percentage of isolates recovered from extra-pulmonary

TABLE 2

 Clusters of *Mycobacterium tuberculosis* isolates within lineage 4 (Euro-American), Ireland, 2010–11 (n=172 isolates)

Sub-lineage	Total no. of cases	No. of clustered cases (%)	No. of clusters	Strain clustering rate (%)	No. of isolates/ cluster (% clustered cases)	MIRU-VNTR profile ^a
Latin American–Mediterranean	52	30 (8.3)	5	6.9	3 (4.7)	124244332224126153332832
					8 (12.7)	142244332224126153322622
					12 (19)	142244332224126143322622
					3 (4.7)	132244332224125153322222
					4 (6.3)	132244332224126133322622
Haarlem	47	18 (5)	6	3.3	2 (3.2)	22322534233442514332332
					6 (9.5)	223225342334425153323_32
					2 (3.2)	12323533263442514342332
					4 (6.3)	223225342334425143323_32
					2 (3.2)	223235331532423153333632
H37Rv	44	8 (2.2)	4	1.1	2 (3.2)	224243122234225153234422
					2 (3.2)	224243122434225153335512
					2 (3.2)	224213222534226153334522
					2 (3.2)	224213222334226153335522
Haarlem/X	29	7 (1.9)	3	1.1	2 (3.2)	224234342334425154135832
					3 (4.7)	243244332434425153343832
					2 (3.2)	243234332234425143331832
Total clustered lineage 4 cases	172	63	18	-	-	-

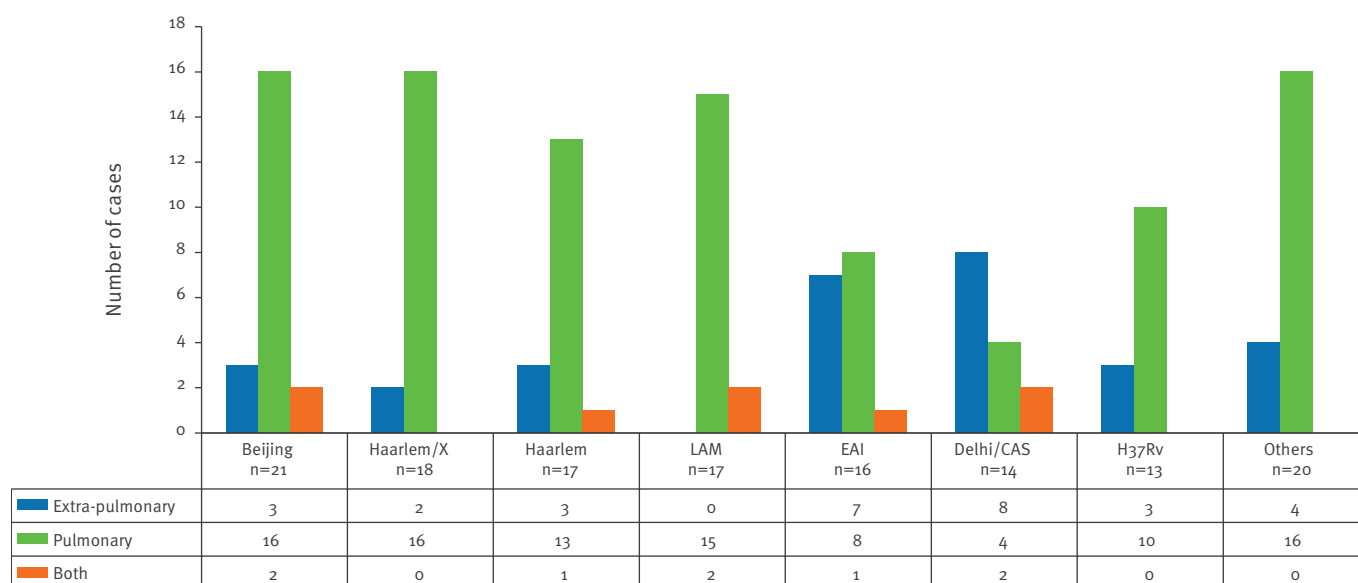
^a The numbers in this 24-digit profile correspond to the number of alleles at each of the following loci: MIRU 02, VNTR 42, VNTR 43, MIRU 04, MIRU 40, MIRU 10, MIRU 16, VNTR 1955, MIRU 20, QUB 11b, ETR A, VNTR 46, VNTR 47, VNTR 48, MIRU 23, MIRU 24, MIRU 26, MIRU 27, VNTR 49, MIRU 31, VNTR 52, QUB 26, VNTR 53, MIRU 39.

specimens was from lineages 1 (22.6%) and 3 (34.3%) [23]. However, due to the small numbers of exclusive extra-pulmonary specimens (n=30) and limited epidemiological data, statistical analysis of the relationship between lineage and clinical site of infection was not possible in our report.

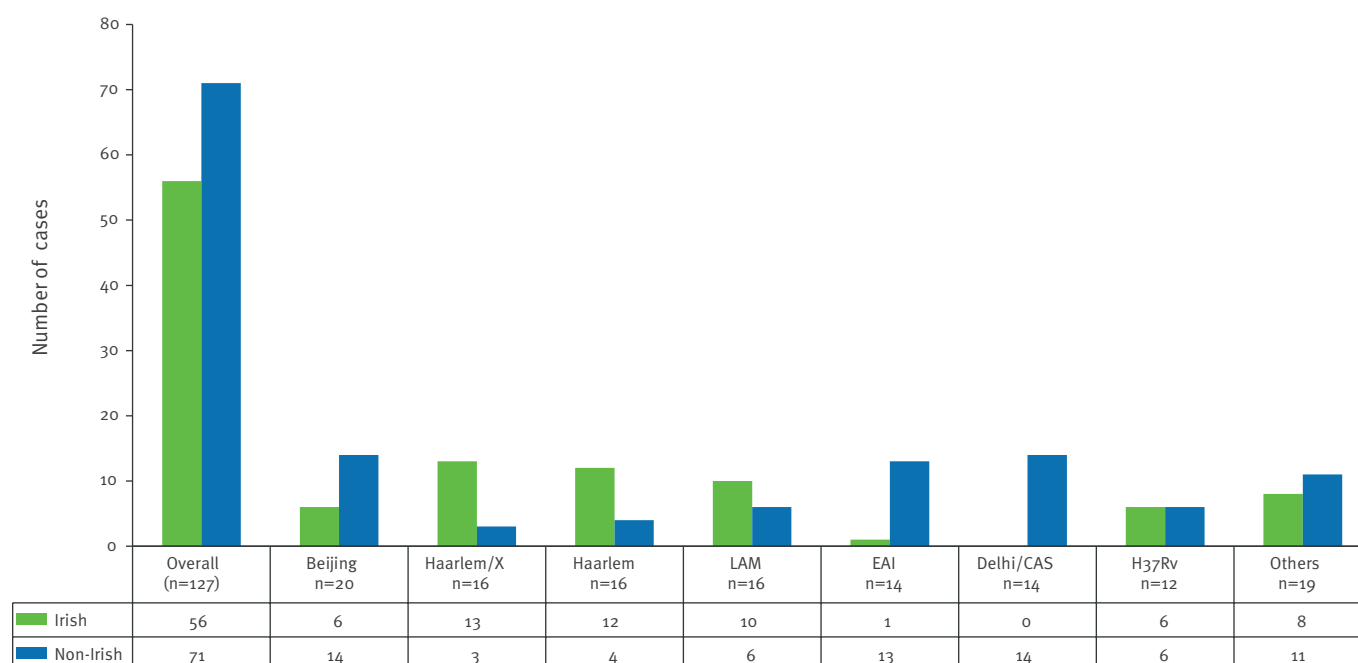
Molecular clustering of isolates in our study was more common among Irish-born individuals. These findings were similar to a previous Irish study conducted by Ojo et al. in the south-west region of Ireland, but unlike a recent study conducted in Switzerland [12,22]. We identified 18 clusters in lineage 4, and the mean cluster size was 3.5 TB cases. The largest cluster, involving 12 TB cases, belonged to the LAM lineage and spanned a period of 18 months. A second cluster identified in the LAM lineage differed by a single locus variant (SLV) at locus 2996. Similarly, in the Haarlem lineage, the two largest clusters differed by a SLV at locus 2996 also. MIRU 26 (or locus 2996) has yielded stable comparable results in a large-scale study investigating 824 *M. tuberculosis* isolates conducted at the Institut Pasteur de Lille, France, in 2006 [5]. Molecular typing played a key role in identifying a dominant *M. tuberculosis* strain (known as the Mercian strain) circulating in the West Midlands region in the United Kingdom (UK) over a five-year period, highlighting the importance of

cluster analysis [14]. Prospective molecular typing can identify rapidly expanding clusters of *M. tuberculosis* before they spread further into the community. A single dominant MIRU-VNTR type was not observed in this study, however, this could be due to the study period being short. In contrast, prospective molecular typing of *M. tuberculosis* by RFLP, performed since 1993 in the Netherlands, has proven to be effective. DNA fingerprinting data has been shown to be a powerful tool in defining epidemiological links and guiding TB control programmes in the Netherlands [24–25].

Another limitation of this study is that only one typing method was used to investigate the *M. tuberculosis* population structure in Ireland. Previous studies have shown that a combination of MIRU-VNTR typing and spoligotyping can differentiate more readily between *M. tuberculosis* strains [26–27]. However, in a previous Irish study using both spoligotyping and MIRU-VNTR typing, MIRU-VNTR typing identified clusters among spoligotype groups, thus providing supporting evidence that MIRU-VNTR typing is a more discriminatory typing method [22]. The discriminatory power of the 24-locus MIRU-VNTR panel used in this study has shown to be similar to IS6110 RFLP analysis [5]. However, the discriminatory power of 24-locus MIRU-VNTR typing differs among genetic lineages, and the inclusion of

FIGURE 1Distribution of *Mycobacterium tuberculosis* lineages by site of infection, Ireland, 2011 (n=136)

CAS: central-Asian; EAI: east-African–Indian; LAM: Latin American–Mediterranean.

FIGURE 2Distribution of lineages among Irish and non-Irish typed *Mycobacterium tuberculosis* cases, Ireland, 2011 (n=127)

CAS: central-Asian; EAI: east-African–Indian; LAM: Latin American–Mediterranean.

additional hypervariable loci may be required to differentiate among strains of lineages 2 (Beijing) and 3 (Delhi/central-Asian). For enhanced cluster or outbreak analysis, whole-genome sequencing has been shown to differentiate among strains with identical 24-locus MIRU-VNTR patterns [28-29]. The role of whole-genome sequencing in investigating community outbreaks in the UK was reported recently [29]. Walker et al. estimated that the rate of genetic changes was 0.5 single nucleotide polymorphisms (SNPs) per genome per year. Furthermore, the maximum number of genetic changes over three years would be five SNPs and 10 SNPs over 10 years [29]. It has also been proposed that clustering of isolates increases over longer periods as transmission chains are more efficiently analysed and reported [30]. But the *M. tuberculosis* genotype involved in the cluster must be considered as for example the Beijing lineage has increased ability to spread and cause disease. While clustering was limited in our study, the study period was too short to draw clear conclusions.

Although the reproducibility of MIRU-VNTR typing has been well documented, results from the first worldwide proficiency study on this method were surprising [7]. Intra- and inter-laboratory reproducibility varied depending on the typing methods employed in each laboratory. In our setting, when the commercial MIRU-VNTR typing kit was used to analyse the quality control panel, 100% concordance was achieved with the reference data (30/30 tested strains) and 100% intra-laboratory reproducibility was achieved. These findings are important to consider when typing data is exchanged between laboratories.

Although six of the 11 MIRU typing clusters identified during 2011 were not confirmed as outbreaks by public health departments, it is possible that the reason why four of these clusters did not meet the Irish TB outbreak case definitions was the small number (n=2) of involved cases [2].

In summary, this study has provided the first insights into the structure of the *M. tuberculosis* population in Ireland. Although the incidence of TB has remained static in Ireland over the last decade, there has been mass immigration to this island nation. Not surprisingly, lineage 4 predominated among circulating strains of *M. tuberculosis* in the present study. But the degree of diversity among *M. tuberculosis* was unexpected. Future studies in the IMRL involving retrospective genotyping analysis of *M. tuberculosis* isolates collected since 2000 may provide an interesting epidemiological picture. Continued molecular surveillance is important as it has been suggested that the transmissibility profile of *M. tuberculosis* strains may be influenced by their genetic and evolutionary background. This understanding of the dynamics of *M. tuberculosis* strains will provide novel insights into the *M. tuberculosis* population structure and how it relates to the epidemiology of TB in Europe and beyond.

Acknowledgments

The authors would like to thank all the departments of public health, clinicians and laboratories for providing the surveillance data on these TB cases.

References

1. European Centre for Disease Prevention and Control (ECDC). Annual Epidemiological Report 2011. Reporting on 2009 surveillance data and 2010 epidemic intelligence data. Stockholm: ECDC; 2011. Available from: http://www.ecdc.europa.eu/en/publications/Publications/Forms/ECDC_DispatchForm.aspx?ID=767
2. Health Protection Surveillance Centre (HPSC). Guidelines on the Prevention and Control of Tuberculosis in Ireland 2010.. Dublin: HPSC; 2010. Available from: <http://www.hpsc.ie/hpsc/AboutHPSC/ScientificCommittees/Publications/File,4349,en.pdf>
3. Health Protection Surveillance Centre (HPSC). Report on the Epidemiology of TB in Ireland, 2009. Dublin: HPSC, 2012. [Accessed 17 Jan 2013]. Available from: <http://www.ndsc.ie/hpsc/A-Z/VaccinePreventable/TuberculosisTB/Epidemiology/SurveillanceReports/2009/File,13261,en.pdf>
4. Barnes PF, Cave MD. Molecular epidemiology of tuberculosis. *New Engl J Med*. 2003;349(12):1149-56.
5. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable number tandem repeat typing of Mycobacterium tuberculosis. *J Clin Microbiol*. 2006;44(12):4498-510.
6. Bidovec-Stojkovic U, Zolnir-Dovc M, Supply P. One year nationwide evaluation of 24-locus MIRU-VNTR genotyping on Slovenian Mycobacterium tuberculosis isolates. *Respir Med*. 2011;105 Suppl 1:S67-73.
7. de Beer J, Kremer K, Ködmön C, Supply P, van Soolingen D, Global Network for the Molecular Surveillance of Tuberculosis 2009. First worldwide proficiency study on variable-number tandem-repeat typing of Mycobacterium tuberculosis complex strains. *J Clin Microbiol*. 2012;50(3):662-9.
8. Kanduma E, McHugh TD, Gillespie SH. Molecular methods for Mycobacterium tuberculosis strain typing: a user's guide. *J Appl Microbiol*. 2003;94(5):781-91.
9. Alonso M, Herranz M, Martinez Lirola M, Gonzalez-Rivera M, Bouza E, Garcia de Viedma D. Real-time molecular epidemiology of tuberculosis by direct genotyping of smear-positive clinical specimens. *J Clin Microbiol*. 2012;50(5):1755-7.
10. Augustynowicz-Kopeć E, Jagielski T, Kozłowska M, Kremer K, Van Soolingen D, Bielecki J, et al. Transmission of tuberculosis within family-households. *J Infect*. 2012;64(6):596-608.
11. Prodinger WM, Polanecký V, Kozáková B, Müllerová M, Mezenský L, Kaustová J, et al. Molecular epidemiology of tuberculosis in the Czech Republic, 2004: analysis of *M. tuberculosis* complex isolates originating from the city of Prague, south Moravia and the Moravian-Silesian region. *Cent Eur J Public Health*. 2006;14(4):168-74.
12. Fenner L, Gagneux S, Helbling P, Battagay M, Rieder HL, Pfyffer GE, et al. Mycobacterium tuberculosis transmission in a country with low incidence: role of immigration and HIV infection. *J Clin Microbiol*. 2012;50(2):388-95.
13. Maguire H, Brailsford S, Carless J, Yates M, Altass L, Yates S, et al. Large outbreak of isoniazid-mono-resistant tuberculosis in London, 1995 to 2006: case-control study and recommendations. *Euro Surveill*. 2011;16(13):pii=19830. Available from: <http://www.eurosurveillance.org/Viewarticle.aspx?ArticleId=19830>
14. Evans JT, Serafino Wani RL, Anderson L, Gibson AL, Grace Smith E, Wood A, et al. A geographically-restricted but prevalent Mycobacterium tuberculosis strain identified in the West Midlands region of the UK between 1995 and 2008. *PLoS One*. 2011; 6(3):e17930.
15. Wirth T, Hildebrand F, Allix-Béguec C, Wölbelling F, Kubica T, Kremer K, et al. Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS Pathog*. 2008;4(9):e1000160.
16. Sails AD, Barrett A, Sarginson S, Magee JG, Maynard P, Hafeez I, et al. Molecular epidemiology of Mycobacterium tuberculosis in East Lancashire 2001-2009. *Thorax*. 2011;66(8):709-13.
17. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria. *Nucleic Acids Res*. 2010;38:W326-31.
18. Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. Evaluation and strategy for use of MIRU-VNTRplus, a

- multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol.* 2008;46(8):2692-9.
19. Roetzer A, Sieglinde S, Diel R, Gasau F, Ubben T, di Nauta A, et al. Evaluation of *Mycobacterium tuberculosis* typing methods in a 4-year study in Schleswig-Holstein, Northern Germany. *J Clin Microbiol.* 2011;49(12):4173-8.
 20. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajj SA, et al. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol.* 2006;6:23.
 21. Homolka S, Post E, Oberhauser B, George AG, Westman L, Dafaie D, et al. High genetic diversity among *Mycobacterium tuberculosis* complex strains from Sierra Leone. *BMC Microbiol.* 2008;8:103.
 22. Ojo OO, Sheehan S, Corcoran D, Nikolayevsky V, Brown T, O'Sullivan M, et al. Molecular epidemiology of *Mycobacterium tuberculosis* isolates in Southwest Ireland. *Infect Genet Evol.* 2010;10(7):1110-6.
 23. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. Relationship between *Mycobacterium tuberculosis* phylogenetic lineage and clinical site of tuberculosis. *Clin Infect Dis.* 2012;54(2):211-9.
 24. Borgdorff MW, Nagelkerke NJD, de Haas PEW, van Soolingen D. Transmission of *Mycobacterium tuberculosis* depending on the age and sex of source cases. *Am J Epidemiol.* 2001;154(10):934-43.
 25. Lambregts-van Weezenbeek CS, Sebek MM, van Gerven PJ, de Vries G, Verver S, et al. Tuberculosis contact investigation and DNA fingerprint surveillance in The Netherlands: 6 years' experience with nation-wide cluster feedback and cluster monitoring. *Int J Tuberc Lung Dis.* 2003;7(12 Suppl 3):S463-70.
 26. Oelemann MC, Diel R, Vatin V, Haas W, Rüsche-Gerdes S, Locht C, et al. Assessment of an optimized mycobacterial interspersed repetitive-unit-variable number of tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. *J Clin Microbiol.* 2007;45(3):691-7.
 27. Valcheva V, Mokrousov I, Narvskaya O, Rastogi N, Markova N. Utility of new 24-locus variable-number tandem-repeat typing for discriminating *Mycobacterium tuberculosis* clinical isolates collected in Bulgaria. *J Clin Microbiol.* 2008;46(9):3005-11.
 28. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodtkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med.* 2011;364(8):730-9.
 29. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dediccoat MJ et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis.* 2012. pii:S1473-3099(12)70277-3. doi: 10.1016/S1473-3099(12)70277-3.
 30. Glynn JR, Bauer J, de Boer AS, Borgdorff MW, Fine PE, Godfrey-Faussett P, et al. Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. European concerted action on molecular epidemiology and control of tuberculosis. *Int J Tuberc Lung Dis.* 1999;3(12):1055-60.

Appendix 2.

Genome Announcements Paper, published 2014

Draft Genome Sequence of the First Isolate of Extensively Drug-Resistant *Mycobacterium tuberculosis* in Ireland

Emma Roycroft,^{a,b} Micheál Mac Aogáin,^a Ronan F. O'Toole,^{a*} Margaret Fitzgibbon,^b Thomas R. Rogers^{a,b}

Department of Clinical Microbiology, Trinity College Dublin, St. James's Hospital, Dublin, Ireland^a; Irish Mycobacteria Reference Laboratory, Labmed Directorate, St. James's Hospital, Dublin, Ireland^b

* Present address: Ronan F. O'Toole, Breathe Well Centre of Research Excellence, School of Medicine, University of Tasmania, Tasmania, Australia.

E.R. and M.M.A. contributed equally to this work.

Extensive drug resistance is an emerging threat to the control of tuberculosis (TB) worldwide, even in countries with low TB incidence. We report the draft whole-genome sequence of the first reported extensively drug-resistant TB (XDR-TB) strain isolated in Ireland (a low-incidence setting) and describe a number of single-nucleotide variations that correlate with its XDR phenotype.

Received 27 August 2014 Accepted 2 September 2014 Published 9 October 2014

Citation Roycroft E, Mac Aogáin M, O'Toole RF, Fitzgibbon M, Rogers TR. 2014. Draft genome sequence of the first isolate of extensively drug-resistant *Mycobacterium tuberculosis* in Ireland. *Genome Announc.* 2(5):e01002-14. doi:10.1128/genomeA.01002-14.

Copyright © 2014 Roycroft et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 3.0 Unported license](http://creativecommons.org/licenses/by/3.0/).

Address correspondence to Emma Roycroft, roycrofe@tcd.ie.

Multidrug resistance (MDR) in tuberculosis (TB) threatens the global management of the disease, which is already a leading cause of infectious mortality worldwide, with an estimated 450,000 MDR-TB cases reported in 2012 (1). Approximately 10% of MDR-TB cases (those resistant to rifampin and isoniazid) are further defined as extensively drug resistant (XDR)-TB, due to their resistance to second-line drugs, fluoroquinolones and injectable aminoglycosides (2). Long turnaround times (2 to 4 weeks) for phenotypic drug susceptibility testing (DST) (due to the fastidious nature of the organism) can hamper the appropriate treatment of XDR-TB by delaying access to antibiotic susceptibility data (3). Next-generation sequencing (NGS) can highlight resistance in a timely manner in order to effectively manage treatment and minimize further transmission of resistant strains (4–6).

The first Irish XDR-TB strain was isolated in the Irish Mycobacteria Reference Laboratory (IMRL) in 2005 (IEXDR1) (7, 8). First-line DST was completed within 3 weeks (found to be streptomycin, isoniazid, rifampin, ethambutol, and pyrazinamide resistant), second-line DST within 5 weeks (found to be amikacin, clarithromycin, ciprofloxacin, and rifabutin resistant, as well as capreomycin, clofazimine, and prothionamide susceptible), and the remainder within 14 weeks (found to be *para*-aminosalicylate sodium [PAS] resistant and ethionamide and cycloserine susceptible).

In 2014, NGS was performed to provide further molecular characterization of IEXDR1 (lineage 2 or Beijing strain). Genomic DNA was sequenced using an Illumina MiSeq. Paired-end reads were mapped to the *Mycobacterium tuberculosis* H37Rv reference genome (accession no. AL123456.3) using the Burrows-Wheeler Aligner (9). This yielded a mapped-read depth of 196-fold, covering 97.6% of the H37Rv genome. A consensus sequence was called using the SAMtools mpileup command (10). The IMAGE algorithm was employed to extend contigs and close gaps in the assembly, producing a final draft assembly of 4,340,174 bp, consisting of

109 contigs (11). Single-nucleotide polymorphism (SNP) analysis was performed using Geneious R7 (version 7.1.5; Biomatters); 1,492 SNPs were detected in the assembled genome with respect to the genome of H37Rv, of which 810 were nonsynonymous (depth of coverage, ≥ 20 -fold [average, 276]; variant frequency, $\geq 95\%$).

Nonsynonymous mutations were identified in genes Rv0667/*rpoB* (H526Y) and Rv1908c/*katG* (S315T). There is strong correlation between substitutions in *rpoB* (H526Y) and *katG* (S315T) and phenotypic resistance to rifampin and isoniazid, respectively (4, 12). High-confidence SNPs were also found for fluoroquinolone resistance in gene Rv0006 (*gyrA*) (D94A) and aminoglycoside resistance in MTB000019/*rrs* (a1401g) (12). This is consistent with the XDR phenotype of IEXDR1. Other high-confidence mutations found in IEXDR1 for ethambutol (Rv3795/*embB* [M306V]) and streptomycin (Rv0682/*rpsL* [K43R]) correlate with its drug resistance profile (13, 14). SNPs that may confer resistance to pyrazinamide (Rv2043c/*pncA* [G132C]) and PAS (Rv3764c/*thyA* [Q97R]) were also identified, although their specificities and sensitivities are not as well defined (http://www.broadinstitute.org/annotation/genome/mtb_drug_resistance.1/DirectedSequencingHome.html).

Previously described phylogenetically informative polymorphisms (Rv1908c/*katG* [R463L], Rv2629 [D64A], Rv3794/*embA* [C76C, TGC/TGT] and [Q38Q, CAA/CAG], Rv1630/*rpsA* [R212R, CGA/CGC], Rv3919c/*gidB* [E92D], and Rv0486/*mshA* [A187V]) confirm the presence of a Beijing strain (15).

In summary, using NGS, this isolate was confirmed to be XDR-TB in a considerably shorter turnaround time than that for conventional DST. This underlines the potential of NGS in the diagnostic laboratory, especially for MDR- and XDR-TB cases.

Nucleotide sequence accession number. This whole-genome sequencing project has been deposited in the European Nucleotide Archive under the accession no. CCJS00000000.

ACKNOWLEDGMENTS

We acknowledge support and funding received from the Clinical Microbiology Department, Trinity College, Dublin, and the Irish Mycobacteria Reference Laboratory and Microbiology Department, Labmed Directorate, St. James' Hospital, Dublin.

REFERENCES

1. World Health Organization. 2013. Global tuberculosis report. World Health Organization, Geneva, Switzerland.
2. CDC. 2006. Revised definition of extensively drug-resistant tuberculosis. *MMWR Morb. Mortal. Wkly. Rep.* 55:1176.
3. Campbell PJ, Morlock GP, Sikes RD, Dalton TL, Metchock B, Starks AM, Hooks DP, Cowan LS, Plikaytis BB, Posey JE. 2011. Molecular detection of mutations associated with first- and second-line drug resistance compared with conventional drug susceptibility testing of *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* 55:2032–2041. <http://dx.doi.org/10.1128/AAC.01550-10>.
4. Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, Carmichael AJ, Parkhill J, Smith GP, Peacock SJ. 2013. Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. *N. Engl. J. Med.* 369:290–292. <http://dx.doi.org/10.1056/NEJMc1215305>.
5. Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.* 8:e1002824. <http://dx.doi.org/10.1371/journal.ppat.1002824>.
6. Wyres K, Conway T, Garg S, Queiroz C, Reumann M, Holt K, Rusu L. 2014. WGS analysis and interpretation in clinical and public health microbiology laboratories: what are the requirements and how do existing tools compare? *Pathogens* 3:437–458. <http://dx.doi.org/10.3390/pathogens3020437>.
7. Mc Laughlin AM, O'Donnell RA, Gibbons N, Scully M, O'Flanagan D, Keane J. 2007. Extensively drug-resistant tuberculosis (XDR-TB)—a potential threat in Ireland. *Open Respir. Med. J.* 1:7–9. <http://dx.doi.org/10.2174/1874306400701010007>.
8. Kennedy B, Lyons O, McLoughlin AM, Gibbons N, O'Flanagan D, Keane J. 2008. Extensively drug-resistant tuberculosis: first report of a case in Ireland. *Euro Surveill.* 13:pil=18935. <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=18935>.
9. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760. <http://dx.doi.org/10.1093/bioinformatics/btp324>.
10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. <http://dx.doi.org/10.1093/bioinformatics/btp352>.
11. Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD. 2012. A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat. Protoc.* 7:1260–1284. <http://dx.doi.org/10.1038/nprot.2012.068>.
12. Rodwell TC, Valafar F, Douglas J, Qian L, Garfein RS, Chawla A, Torres J, Zadorozhny V, Kim MS, Hoshide M, Catanzaro D, Jackson L, Lin G, Desmond E, Rodrigues C, Eisenach K, Victor TC, Ismail N, Crudu V, Gler MT, Catanzaro A. 2014. Predicting extensively drug-resistant *Mycobacterium tuberculosis* phenotypes with genetic mutations. *J. Clin. Microbiol.* 52:781–789. <http://dx.doi.org/10.1128/JCM.02701-13>.
13. Sreevatsan S, Stockbauer KE, Pan X, Kreiswirth BN, Moghazeh SL, Jacobs WR, Jr, Telenti A, Musser JM. 1997. Ethambutol resistance in *Mycobacterium tuberculosis*: critical role of *embB* mutations. *Antimicrob. Agents Chemother.* 41:1677–1681.
14. Nair J, Rouse DA, Bai GH, Morris SL. 1993. The *rpsL* gene and streptomycin resistance in single and multiple drug-resistant strains of *Mycobacterium tuberculosis*. *Mol. Microbiol.* 10:521–527. <http://dx.doi.org/10.1111/j.1365-2958.1993.tb00924.x>.
15. Feuerriegel S, Köser CU, Niemann S. 2014. Phylogenetic polymorphisms in antibiotic resistance genes of the *Mycobacterium tuberculosis* complex. *J. Antimicrob. Chemother.* 69:1205–1210. <http://dx.doi.org/10.1093/jac/dkt535>.

Appendix 3.

**Lancet Respiratory Medicine Paper, published
2016**

Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study



Louise J Pankhurst*, Carlos del Ojo Elias*, Antonina A Votintseva*, Timothy M Walker*, Kevin Cole, Jim Davies, Jilles M Fermont, Deborah M Gascoyne-Binzi, Thomas A Kohl, Clare Kong, Nadine Lemaitre, Stefan Niemann, John Paul, Thomas R Rogers, Emma Roycroft, E Grace Smith, Philip Supply, Patrick Tang, Mark H Wilcox, Sarah Wordsworth, David Wyllie, Li Xu, Derrick W Crook, for the COMPASS-TB Study Group†

Summary

Background Slow and cumbersome laboratory diagnostics for *Mycobacterium tuberculosis* complex (MTBC) risk delayed treatment and poor patient outcomes. Whole-genome sequencing (WGS) could potentially provide a rapid and comprehensive diagnostic solution. In this prospective study, we compare real-time WGS with routine MTBC diagnostic workflows.

Methods We compared sequencing mycobacteria from all newly positive liquid cultures with routine laboratory diagnostic workflows across eight laboratories in Europe and North America for diagnostic accuracy, processing times, and cost between Sept 6, 2013, and April 14, 2014. We sequenced specimens once using local Illumina MiSeq platforms and processed data centrally using a semi-automated bioinformatics pipeline. We identified species or complex using gene presence or absence, predicted drug susceptibilities from resistance-conferring mutations identified from reference-mapped MTBC genomes, and calculated genetic distance to previously sequenced UK MTBC isolates to detect outbreaks. WGS data processing and analysis was done by staff masked to routine reference laboratory and clinical results. We also did a microcosting analysis to assess the financial viability of WGS-based diagnostics.

Findings Compared with routine results, WGS predicted species with 93% (95% CI 90–96; 322 of 345 specimens; 356 mycobacteria specimens submitted) accuracy and drug susceptibility also with 93% (91–95; 628 of 672 specimens; 168 MTBC specimens identified) accuracy, with one sequencing attempt. WGS linked 15 (16% [95% CI 10–26]) of 91 UK patients to an outbreak. WGS diagnosed a case of multidrug-resistant tuberculosis before routine diagnosis was completed and discovered a new multidrug-resistant tuberculosis cluster. Full WGS diagnostics could be generated in a median of 9 days (IQR 6–10), a median of 21 days (IQR 14–32) faster than final reference laboratory reports were produced (median of 31 days [IQR 21–44]), at a cost of £481 per culture-positive specimen, whereas routine diagnosis costs £518, equating to a WGS-based diagnosis cost that is 7% cheaper annually than are present diagnostic workflows.

Interpretation We have shown that WGS has a scalable, rapid turnaround, and is a financially feasible method for full MTBC diagnostics. Continued improvements to mycobacterial processing, bioinformatics, and analysis will improve the accuracy, speed, and scope of WGS-based diagnosis.

Funding National Institute for Health Research, Department of Health, Wellcome Trust, British Columbia Centre for Disease Control Foundation for Population and Public Health, Department of Clinical Microbiology, Trinity College Dublin.

Copyright © Pankhurst et al. Open Access article distributed under the terms of CC BY.

Introduction

In 2013, WHO estimated that *Mycobacterium tuberculosis* complex (MTBC) caused 9 million new active infections and 1.5 million deaths worldwide.¹ Non-tuberculous mycobacteria also cause considerable morbidity and mortality.² Protracted MTBC diagnosis and phenotypic drug susceptibility testing (DST) due to slow growth in culture contribute to reported treatment initiation delays of 8–80 days from first contact with health services, risking poor clinical outcomes and transmission control.^{3–6} Although genotypic assays such as the Cepheid Xpert MTB/RIF (Cepheid, Sunnyvale, CA, USA) and Hain line-probe (Hain Lifescience, Nehren, Germany) assays can rapidly (less than a day) identify mycobacterial

species and mutations conferring MTBC drug resistance independent of culture, they do not detect all resistance-conferring mutations and are typically still used after microbial culture.^{5–9} Besides identifying species and doing DST, high-income countries also genotype MTBC using mycobacterial interspersed repetitive unit-variable-number tandem repeat (MIRU-VNTR) for outbreak detection.

Findings from retrospective studies^{6,8–19} show the potential for whole-genome sequencing (WGS) to predict drug susceptibility and simultaneously track outbreaks with high resolution. WGS could replace the entire MTBC diagnostic workflow from Mycobacteria Growth Indicator Tubes (MGITs; BACTEC MGIT; Beckton

Lancet Respir Med 2016;
4: 49–58

Published Online
December 3, 2015
[http://dx.doi.org/10.1016/S2213-2600\(15\)00466-X](http://dx.doi.org/10.1016/S2213-2600(15)00466-X)

See [Comment](#) page 6

*Contributed equally

†Members listed in the appendix

Microbiology and Infectious Diseases, Nuffield Department of Clinical Medicine, John Radcliffe Hospital (L J Pankhurst PhD, C del Ojo Elias MSc, A A Votintseva PhD, T M Walker MRCP, D W Crook FRCPATH, D Wyllie FRCPATH), Health Economics Research Centre, Nuffield Department of Population Health (J M Fermont MSc, S Wordsworth PhD), and Department of Computer Science (Prof J Davies PhD), University of Oxford, Oxford, UK; Brighton and Sussex University Hospitals NHS Trust, Brighton, UK (K Cole BSc, J Paul MD); Public Health England Regional Centre for Mycobacteriology, Birmingham Heartlands Hospital NHS Foundation Trust, Birmingham, UK (L Xu PhD, E G Smith FRCPATH); Leeds Teaching Hospitals NHS Trust, Leeds, UK (D M Gascoyne-Binzi PhD, M H Wilcox FRCPATH); Université de Lille, Centre national de la recherche scientifique Unité mixte de recherche 8204, Institut national de la santé et de la recherche médicale U1019, Centre Hospitalier Universitaire, and Center for Infection and Immunity of Lille, Institut Pasteur de Lille, Lille, France (N Lemaitre PhD, P Supply PhD); Genoscreen, Lille, France (P Supply); British Columbia Public Health Microbiology and Reference Laboratory, Vancouver, Canada (C Kong BSc, P Tang PhD); Molecular Mycobacteriology,

Forschungszentrum Borstel,
Leibniz-Zentrum für Medizin
und Biowissenschaften,
Schleswig-Holstein, Germany
(T A Kohl PhD,
S Niemann DSc ScD); German
Center for Infection Research,
Borstel, Germany (S Niemann);
Department of Clinical
Microbiology Trinity College
Dublin and Irish Mycobacteria
Reference Laboratory, St
James's Hospital, Dublin,
Ireland (E Roycroft MSc,
T R Rogers FRCPATH); and Public
Health England, Oxford, UK
(D Wyllie)

Correspondence to:
Dr Louise J Pankhurst,
Microbiology and Infectious
Diseases, Nuffield Department of
Clinical Medicine, John Radcliffe
Hospital, University of Oxford,
Oxford OX3 9DU, UK
louise.pankhurst@ndm.ox.ac.uk

See Online for appendix

Research in context

Evidence before this study

We searched PubMed for studies published before July 1, 2015, with no language restrictions, using the search terms “whole genome sequencing”, “diagnosis”, “infection”, “mycobacterium”, and “tuberculosis”. Full diagnosis of *Mycobacterium tuberculosis* consists of identification of the organism, establishment of antibiotic sensitivity profiles, and outbreak investigation. During the last 5 years, whole-genome sequencing (WGS) has been increasingly used to assist aspects of this diagnostic pathway. Its primary use has been outbreak investigations, for which WGS provides higher-resolution outbreak tracing than does traditional typing. WGS has begun to be used to elucidate drug resistance profiles for tuberculosis and discover new resistance-conferring mutations. In view of the slow (1–2 months) diagnostic time for tuberculosis with culture-based methods, introduction of molecular assays, including WGS, to replace aspects of the tuberculosis diagnostic pathway has been recognised to improve outbreak control and potentially expedite patient treatment. However, many barriers to widespread adoption of WGS have been raised. These barriers are high diagnostic costs and personnel efforts and an absence of automated sequence analysis pipelines and supporting IT infrastructure.

Added value of this study

In this study, we implement an end-to-end WGS-based diagnostic system for tuberculosis in eight laboratories in Europe and North America. Using a decentralised sequencing,

centralised analysis model and a semi-automated bioinformatics pipeline, we sequenced newly positive mycobacterium cultures and generated diagnostic reports identifying the mycobacteria present, and for tuberculosis, predicted resistance to first-line and second-line antibiotics and did outbreak analysis. Prospective assessment of the WGS-based diagnostic system allows diagnostic accuracy to be directly compared with routine clinical diagnosis, which shows how WGS is scalable and how it could provide full diagnostic information weeks faster than routine clinical diagnostics could. Through a microcosting analysis comparing routine with WGS-based diagnostics, the financial feasibility of WGS-based diagnostics in high-income countries is established.

Implications of all the available evidence

WGS is now evidently capable of replacing traditional diagnostic procedures for tuberculosis in high-income settings. It offers clear benefits compared with traditional diagnostics, allowing rapid identification and control of outbreaks and minimising empirical treatment of patients through simultaneous first-line and second-line drug susceptibility prediction. Results from this study have led to a full feasibility study of use of WGS-based tuberculosis diagnostics in the UK, representing a paradigm shift in infectious disease diagnostics. Furthermore, as WGS technology continues to develop and portable systems become available, WGS will revolutionise diagnostics in low-income settings.

Dickinson, Franklin Lakes, NJ, USA) thanks to demonstrable benefits for outbreak detection, growing knowledge bases for drug resistance-conferring mutations, and reliable DNA isolation from newly positive culture.^{6,8–20} So far, to our knowledge, no investigators have assessed this process prospectively. In this prospective study, we compare real-time WGS with routine MTBC diagnostic workflows at Illumina MiSeq (Illumina, San Diego, CA, USA)-equipped laboratories across Europe and North America. We also do a microcosting analysis to assess the financial viability of WGS-based diagnostics.

Methods

Study design

Eight participating laboratories in the UK, Ireland, Germany, France, and Canada processed all newly positive MGIT cultures from specimens submitted for mycobacterial testing on the examining physician's request between Sept 6, 2013, and April 14, 2014 (appendix p 19). We used no other selection criteria (except for the German centre where only the second positive primary culture from MTBC patients was available for processing). If patients had duplicate specimens—eg, from different body sites or at different times—we included them.

Because this study assessed service delivery methods without returning results for clinical management, research ethics committee approval was not required in the UK. Other centres obtained local ethics committee approval.

Procedures

Routine diagnostic procedures at all centres included species identification (Hain GenoType MTBC/CM/AS), and for MTBC, MIRU-VNTR and culture with isoniazid, rifampicin, ethambutol, and pyrazinamide to establish drug susceptibility. Culture with streptomycin was also done in some centres. Subsequent culture of rifampicin-resistant isolates with fluoroquinolones and aminoglycosides was done to establish additional drug susceptibilities.

Each site prepared DNA and did WGS. We heat inactivated 1–2 mL MGIT culture aliquots at 95°C for 0.5–2 h, adhering to local protocols, and always retaining sufficient MGIT culture for routine diagnostic procedures to avoid compromising patient care. We isolated DNA as previously described (appendix);²⁰ an in-house protocol was used in the Canadian centre. We prepared sequencing libraries for the MiSeq platform using a modified Nextera XT (Illumina) protocol (appendix) and sequenced pools of 11–15 MGIT samples plus *M tuberculosis* H37Rv

or BCG DNA (positive control) with MiSeq version 2 2×150 bp paired-end read cartridges. We processed each sample once, repeating sequencing only for poor overall run performance. We deposited reads in the National Center for Biotechnology Information Short Read Archive (BioProject PRJNA268101 and PRJNA302362; appendix).

Staff doing WGS processing and analysis were masked to routine reference laboratory and clinical results. We shared MiSeq runs via Illumina BaseSpace and downloaded them at the Oxford centre for semi-automated analysis by a bespoke bioinformatics pipeline (appendix). A gene presence or absence algorithm identified mycobacterial species (using a catalogue of 169 sequenced mycobacterial strains). We mapped isolates identified as MTBC to the H37Rv reference genome (GenBank NC000962.2; Stampy version 1.0.22) and examined them for mutations deemed to confer phenotypic resistance to isoniazid, rifampicin, ethambutol, pyrazinamide, streptomycin, fluoroquinolones, or aminoglycosides on the basis of a published catalogue of high-confidence resistance-determining alleles (appendix).⁶ A minimum sequencing depth of five reads was needed to identify mutations; when we found more than one base at a single site, if the minority variant consisted of at least 10% of the total base calls and had a depth of at least five reads, we predicted a mixed phenotype. We identified cases compatible with transmission from a published database of 2191 UK MTBC sequences, representing all worldwide lineages, as previously described (using a threshold of 12 or fewer single-nucleotide polymorphisms (SNPs) on the basis of maximum diversity within different body sites and over time within a patient, and within household outbreaks; appendix).¹³

We estimated sequencing quality with several methods. For all specimens, we mapped the first 50 000 reads (Bowtie version 2.2.0) to the human genome (GRCh37/hg19) and nasal and mouth flora in the National Institutes of Health Human Microbiome Project. To verify that this sample was random, we selected 1% of reads using SAMtools 1.2 (SAMtools view -s option). We mapped these randomly selected reads using the same methods and yielded the same results (data not shown). We assessed the number of reads mapping to these human, nasal, and mouth databases, guanine-cytosine (GC) content (expected to be about 65% for mycobacteria), the number of reads available for analysis, the number of reads mapping to the reference genome, and reference genome coverage as predictors of accurate species identification using multivariable fractional polynomial logistic regression (Stata mfp; backwards elimination threshold $p=0.05$) in Stata 13.1, treating each specimen as an independent observation.²¹ We reported quality control data to the sequencing centre together with mycobacterial species, and for MTBC, drug susceptibility predictions and closest genomic match (appendix).

We gathered anonymised routine diagnostic data from local clinical laboratories and regional mycobacterial reference laboratories after WGS processing. All MTBC duplicate specimens were identified by the participating centres and analysis of drug resistance and outbreak incidence done with and without removal of duplicates (appendix). We did not identify duplicates for non-MTBC specimens. If routine and WGS results differed, we did additional quality checking and routine workflow assays; we did not repeat WGS. We calculated confidence

	n
Routine methods and WGS identification failed	9
Routine methods failed	2*
Identified by routine methods	345 (100%)
Concordant	322 (93%)
<i>M tuberculosis</i> complex	157 (46%)
<i>M avium</i> complex	71 (21%)
<i>M abscessus</i> complex	39 (11%)
<i>M gordonae</i>	18 (5%)
<i>M xenopi</i>	11 (3%)
<i>M tuberculosis</i> complex (BCG)	8 (2%)
<i>M kansasii</i>	6 (2%)
<i>M malmoense</i>	3 (1%)
<i>M fortuitum</i>	2 (1%)
<i>M szulgai</i>	2 (1%)
<i>M tuberculosis</i> complex (<i>M africanum</i>)	2 (1%)
<i>M celatum</i>	1 (<1%)
<i>M lentiflavum</i>	1 (<1%)
<i>M tuberculosis</i> complex and <i>M avium</i> complex	1 (<1%)
Part concordant	10 (3%)
WGS gained one species	5 (1%)†
WGS missed one species	3 (1%)‡
WGS identified related species	1 (<1%)§
WGS identified subspecies	1 (<1%)¶
Discordant	3 (1%)
WGS failed	10 (3%)

Data in parentheses are % of specimens identified by routine methods. M=Mycobacterium. WGS=whole-genome sequencing. *WGS identified *M avium* complex (good quality WGS; 32 [97%] of 33 genes identified) and *M kumamotoense* (poor quality WGS; 58 [58%] of 100 genes identified). †One routine *M tuberculosis* complex (MTBC); WGS identified MTBC plus *M avium* complex (retesting not possible); three routine *M avium* complexes: WGS identified MTBC plus *M avium* complexes (two retesting supported routine; one supported WGS); one routine *M abscessus* complex: WGS identified *M abscessus* plus *M avium* complex (retesting supported routine); appendix. ‡One routine MTBC plus *M avium* complex: WGS identified MTBC (retesting supported WGS); two routine MTBC plus *M avium* complex: WGS *M avium* complex (one poor quality WGS; two [6%] of 33 genes identified; one unable to retest); appendix. §One undescribed mycobacterial species similar to *M avium*: WGS identified *M avium* complex (appendix). ¶One routine *M fortuitum*: WGS identified *M fortuitum*-*acetamidolyticum* (poor quality WGS; 33 [33%] of 100 genes identified); appendix. ||One routine *M kansasii*: WGS identified *M avium* complex (routine testing subsequently confirmed *M avium* complex); one routine *M avium* complex: WGS *M scrofulaceum* (poor quality WGS; one [1%] of 72 genes identified); one routine MTBC: WGS identified *M abscessus* complex (good quality WGS; 38 [95%] of 40 genes identified; unable to retest specimen); appendix.

Table 1: Concordance between single WGS and routine laboratory methods for mycobacterial speciation

For the Human Microbiome Project see <http://www.hmpdacc.org>

	DST successful: resistant				DST successful: sensitive				DST failed				DST not attempted			
	Resistant	Sensitive	Mixed*	Failed†	Resistant	Sensitive	Mixed*	Failed†	Resistant	Sensitive	Mixed*	Failed†‡	Resistant	Sensitive	Mixed*	Failed†
Total across drugs and drug classes§	40 (100%) /19 (100%)	7 (100%) /6 (100%)	1 (100%) /1 (100%)	0	1 (100%) /1 (100%)	618 (100%) /120 (100%)	5 (100%) /4 (100%)	31 (100%) /8 (100%)	0	1 (100%) /1 (100%)	0	4 (100%) /0	6 (100%) /4 (100%)	427 (100%) /118 (100%)	7 (100%) /6 (100%)	28 (100%) /10 (100%)
First-line drugs																
Isoniazid	13 (33%) /11 (56%)	2 (29%) /2 (33%)	1 (100%) /1 (100%)	0	0	143 (23%) /105 (88%)	0	7 (23%) /7 (88%)	0	1 (100%) /1 (100%)	0	1 (25%) /0	0	0	0	0
Rifampicin	5 (13%) /4 (21%)	1 (14%) /1 (17%)	0	0	0	148 (24%) /111 (93%)	4 (80%) /3 (75%)	9 (29%) /8 (100%)	0	0	0	1 (25%) /0	0	0	0	0
Ethambutol	5 (13%) /4 (21%)	1 (14%) /1 (17%)	0	0	1 (100%) /1 (100%)	153 (25%) /114 (95%)	0	7 (23%) /7 (88%)	0	0	0	1 (25%) /0	0	0	0	0
Pyrazinamide	8 (20%) /5 (26%)	1 (14%) /1 (17%)	0	0	0	149 (24%) /113 (94%)	1 (20%) /1 (25%)	8 (26%) /7 (88%)	0	0	0	1 (25%) /0	0	0	0	0
Second-line drugs																
Streptomycin	5 (13%) /3 (16%)	1 (14%) /1 (17%)	0	0	0	14 (2%) /12 (10%)	0	0	0	0	0	0	2 (33%) /1 (25%)	138 (32%) /103 (87%)	0	8 (29%) /7 (70%)
Fluoroquinolones	3 (8%) /3 (16%)	1 (14%) /1 (17%)	0	0	0	6 (1%) /5 (4%)	0	0	0	0	0	0	2 (33%) /2 (50%)	148 (35%) /109 (92%)	0	8 (29%) /7 (70%)
Aminoglycosides	1 (3%) /1 (5%)	0	0	0	0	5 (1%) /4 (3%)	0	0	0	0	0	0	2 (33%) /1 (25%)	141 (33%) /105 (89%)	7 (100%) /6 (100%)	12 (43%) /10 (100%)

Data are number of specimens (%)/number of patients (%). DST=drug susceptibility testing. *Resistant and sensitive. †Failed whole-genome sequencing prediction (insufficient sequencing data to predict drug resistance, each specimen sequenced only once). ‡There are zero patients because these samples have been removed during removal of duplicate specimens. §The numbers of patients do not add to the totals because patients can be counted in more than one category.

Table 2: Whole-genome sequencing resistance predictions for *Mycobacterium tuberculosis* complex specimens compared with phenotypic DST

intervals for WGS sensitivity, specificity, or accuracy compared with routine diagnostics in Stata 13.1 (Stata cii).

To assess the financial viability of WGS-based diagnostics, we did a microcosting analysis at a local clinical laboratory (John Radcliffe Hospital, Oxford, UK) and regional reference laboratory (Birmingham Heartlands Hospital Trusts, Birmingham, UK). We collected data using questionnaires based on standard operating procedures, expert consultations, and interviews with laboratory staff. Questionnaires were completed by clinical scientists doing mycobacterial processing and financial managers, who collected costs associated with staff time, error rates, equipment, and consumables (appendix). We obtained basic cost data (staff time, consumables, and equipment only) via interview with clinical scientists for second-line phenotypic DST (done at the National Mycobacterial Reference Laboratory, London, UK). We annualised costs using the throughput of the Birmingham regional reference laboratory for 2014.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Each participating site collected positive MGIT samples for between 9 and 158 days (appendix p 19). 27 (21%) of 127 MTBC patients provided two to six specimens. Median time from positivity to inactivation was 4 days (IQR 1–5; 96 recorded). Median read depth, based on read length and number of reads mapping to the reference genome, was 73 (IQR 36–99; see appendix for alternative read-depth metrics).

356 MGIT specimens were submitted. 345 (97%) were identified to species or complex by routine diagnostic workflows (table 1). In 326 (94%) cases, both Hain and WGS assays identified a single species, of which three (1%) were discordant. Two species were identified in nine (3%) of cases, of which eight (89%) were discordant for one of the species. WGS predictions were concordant with routine results in 322 (93% [95% CI 90–96]) of 345 specimens (including duplicate specimens). Hain and WGS assays identified MTBC in 168 (52%) of 322 concordant specimens. Of the discordant isolates, three (13%) of 23 were MTBC cases identified by the reference laboratory alone, three (13%) were MTBC cases identified by WGS alone, and two (9%) were identified in a co-infection by either WGS or the reference laboratory (but not both). In a further six (26%) MTBC cases (identified by the reference laboratory), WGS failed. Overall, MTBC was identified with 95% (95% CI 91–98) sensitivity and 98% (95–100) specificity (including duplicate specimens). Causes of discordant results were

mixed or contaminated samples and poor quality sequencing. Failure to identify MTBC increased significantly and independently if GC content fell below 50% (showing low-GC non-mycobacterial DNA contamination) or if the total number of sequencing reads fell below 1 million (appendix p 16; $p < 0.005$).

For 168 MTBC specimens identified by WGS and routine diagnosis, 628 (93% [95% CI 91–95]) of 672 WGS-based first-line drug susceptibility predictions were concordant with reference laboratory DST. After deduplication, 467 (92% [89–94]) of 508 predictions across 127 specimens were concordant. Overall, WGS resistance prediction failed on 63 occasions across 15 specimens. Eight (53%) specimens failed all predictions, four (27%) failed aminoglycosides only, two (13%) failed rifampicin only, and one (7%) failed pyrazinamide only. When WGS prediction failed, all but one DST result was sensitive (one DST also failed). WGS made an additional 434 (mainly second-line) predictions when DST was not done, including four specimens with monoresistance to second-line drugs (appendix). In 13 cases (across 11 specimens and four drugs), resistance-conferring mutations and wild-type alleles occurred as a mixture, preventing phenotypic predictions. We noted the highest number of mixtures (7 [4%] of 168 specimens) in genes conferring resistance to aminoglycosides (table 2). Eight WGS predictions (across six specimens and six drugs) were discordant with DST. Of seven phenotypically resistant specimens with no resistance-conferring mutations in the catalogue, six (86%) contained unclassified variants in the relevant genes (appendix). Of the 127 deduplicated specimens, WGS reported 22 (17%) incidences of drug-resistant MTBC, of which five (23%) were *M bovis* (monoresistant to pyrazinamide). Four (3%) of 127 patients were infected with multidrug-resistant tuberculosis.

Pairwise SNP distances between 16 UK-sourced H37Rv-positive controls were all zero; one, from different starting material, was eight SNPs or fewer from other replicates. 68 (40%) of 168 MTBC specimens were 12 SNPs or fewer from at least one other specimen in the available UK database or previously sequenced in this study; however, 33 (49%) of these specimens (15 patients) were linked only to another specimen from the same patient.¹³ In the 127 patients with MTBC, the median pairwise distance to the next nearest patient was 113 (IQR 48–173). 22 (17%) patients with MTBC (19 [86%] UK and three [14%] non-UK; 35 specimens) were linked to different patients. This number included five (23%) patients (four [80%] UK; one [20%] non-UK) with vaccine-strain BCG and two (9%) patients linked only to each other (same non-UK centre). In total, 15 (16% [95% CI 10–26]) of 91 UK patients studied were linked to nine outbreak clusters in the UK database; including two (13%) patients unexpectedly linked to an outbreak cluster (subsequently confirmed epidemiologically; pilot cluster 6; appendix) and two (13%) from different UK regions linked to an isoniazid-resistant cluster in a third

region (subsequently confirmed epidemiologically and via MIRU-VNTR; appendix). Eight (89%) of nine outbreak clusters were already being investigated by local health protection teams, with well established epidemiological links supported by MIRU-VNTR data (available for eight [53%] of 15 UK study patients linked to these pre-existing clusters). The remaining cluster involved patients infected with multidrug-resistant tuberculosis. In this case, WGS provided the first diagnosis and outbreak alert (panel).

Unadjusted median time from MGIT positivity to DST reporting was 25 days (IQR 14–32), whereas for final reports, including MIRU-VNTR genotype reporting, it was 31 days (IQR 21–44); similarly, full WGS-based reports were available in 31 days (IQR 21–60). WGS processing delays were driven by sample batching for sequencing and delays in sharing sequencing data (figure 2). Additionally, we adhered to a 5 day working week for WGS-based processing, rather than the 7 day working week in clinical laboratories. After WGS sharing, we generated full diagnostic reports in a median

Panel: Early multidrug-resistant tuberculosis diagnosis and outbreak discovery

Two patients were linked to a previously sequenced extremely drug-resistant tuberculosis case to form a new cluster (figure 1). The first case, for which sequencing and analysis was completed a week after Mycobacteria Growth Indicator Tube positivity, was seven single-nucleotide polymorphisms from a 2010 isolate from a patient residing about 300 miles away.¹³ The study specimen was predicted to be resistant to isoniazid, rifampicin, and aminoglycosides, whereas the 2010 isolate was phenotypically and genotypically resistant to the same drugs and to fluoroquinolones (with *gyrA*⁴⁹⁰ mutation). Because the study specimen had yet to be processed by the reference laboratory, whole-genome sequencing (WGS) provided the first multidrug-resistant tuberculosis alert to the participating centre. This patient, with smear-positive pulmonary tuberculosis, was receiving first-line treatment in the community while awaiting laboratory results. Prompted by WGS, urgent Hain MTBDR_{plus} and MTBDR_{sl} assays were done by the reference laboratory, confirming WGS drug resistance predictions. Mycobacterial interspersed repetitive unit-variable-number tandem repeat was identical across the 2010 and study isolates, confirming the genetic relation. The patient was admitted to hospital for appropriate treatment on the basis of reference laboratory Hain results. Epidemiological investigation showed that both patients originated from the same European country. 2 months later, a second study specimen from the same centre was zero SNPs from the first specimen, with an identical drug resistance profile, consistent with direct transmission.¹³ Mycobacterial interspersed repetitive unit-variable-number tandem repeat again confirmed the genetic relation, although no epidemiological links between these patients have yet been identified.

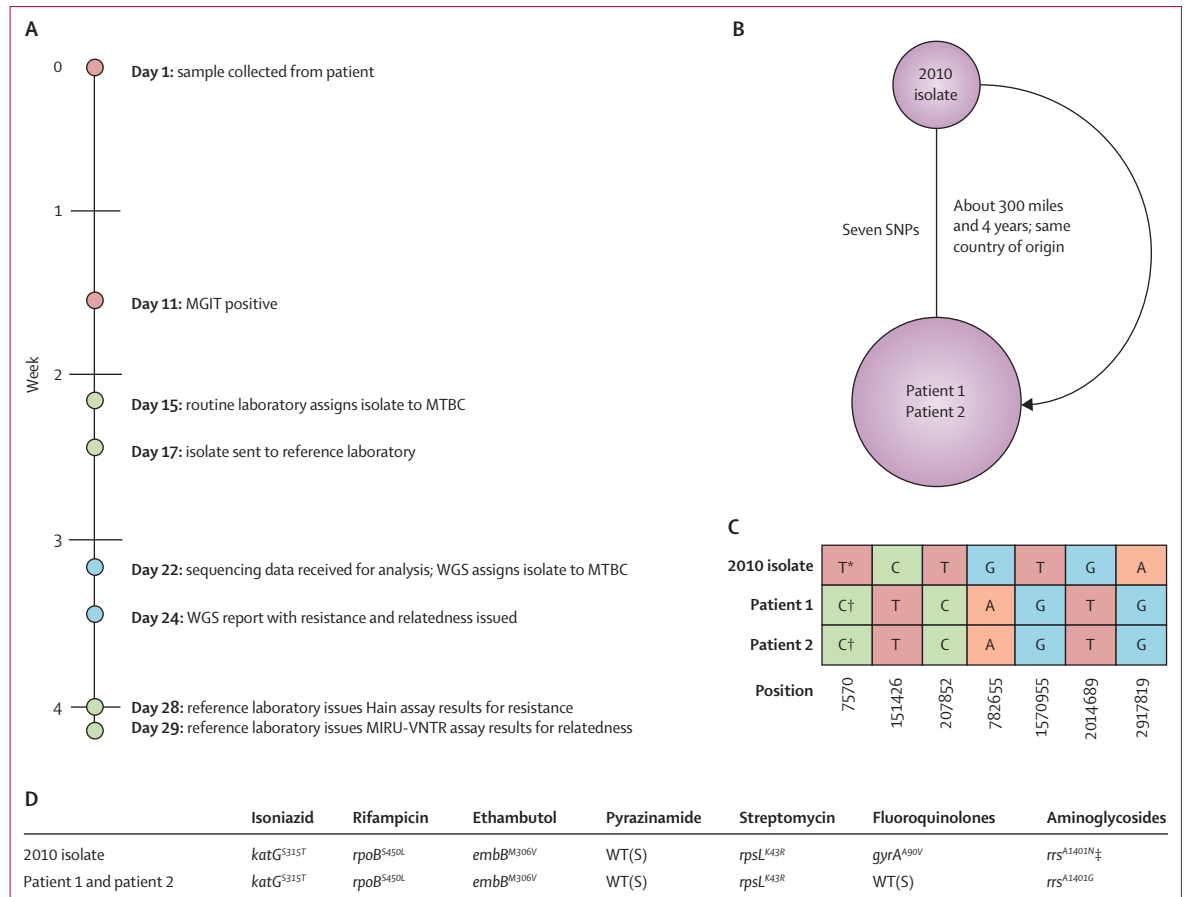


Figure 1: Details of pilot study cluster 7

(A) Timeline of the full WGS-based, routine-based, and reference laboratory-based diagnosis for patient 1 (red circles represent processes needed by both routine or reference and WGS processes, green circles represent routine or reference processes, and blue circles represent WGS processes); (B) the genomic relation between cluster isolates as established by WGS nearest neighbours; (C) detail of the seven SNP differences between cluster isolates; (D) detail of the resistance-conferring mutations identified in the 2010 isolate and the two study patients. A=adenine. C=cytosine. G=guanine. MIRU-VNTR=mycobacterial interspersed repetitive unit-variable-number tandem repeat. MGIT=Mycobacteria Growth Indicator Tube. MTBC=Mycobacterium tuberculosis complex. SNP=single-nucleotide polymorphism. T=thymine. WGS=whole-genome sequencing. WT(S)=wild-type (susceptible). *Mutation conferring resistance to fluoroquinolones. †Wild-type (sensitive). ‡Heterozygous at this position (four wild-type base calls A vs 29 resistance-conferring variant base calls G).

of 5 days (IQR 3–7). The time delay from sample batching would be minimised in high-throughput laboratories. To estimate the potential speed of WGS-based diagnosis, we compared the recorded times from 2 days before sequencing (to allow for sample preparation and 7 day working weeks) to WGS report generation with reference laboratory reporting times (using the date that specimens were sent to the reference laboratory as the starting point). We found that reference laboratory reports were generated a median of 15 days (IQR 9–25) slower than we could produce WGS reports for drug resistance (median 24 days [IQR 20–33] vs 8 days [6–9]) and 21 days (14–32) slower for relatedness (32 days [22–42] vs 9 days [6–10]; figure 2).

The cost of WGS-based diagnosis, routine diagnostic costs for a non-tuberculous mycobacteria (NTM; culture and species identification only), costs for a fully sensitive MTBC (culture, species identification, MIRU-VNTR, and

first-line DST), and costs for a drug-resistant MTBC (culture, species identification, MIRU-VNTR, first-line and second-line DST if resistant to rifampicin) are shown in table 3 (detailed breakdown in appendix). Consumables were the main cost driver for all processes other than for DST, for which staff costs dominated.

Costs (calculated with reported throughput for 2014) for routine diagnostic workflows were £518 per culture-positive specimen, consisting of MGIT culture for all samples received, species identification for culture-positive specimens, and MIRU-VNTR and DST for MTBC-positive specimens. For WGS-based diagnosis, consisting of MGIT culture for all samples received and WGS for culture-positive specimens, the per-culture-positive specimen cost would be £481, which is 7% cheaper than are routine diagnostics. To do DST as per present workflows alongside WGS would cost £540 per culture-positive specimen, which is 4% more expensive

than are routine diagnostics (table 3). Increasing sample throughput decreased costs overall (table 3, appendix). Variation in sequencing batch size, throughput, error rates, equipment, consumables, and overhead costs could alter overall WGS costs by up to 17% (appendix).

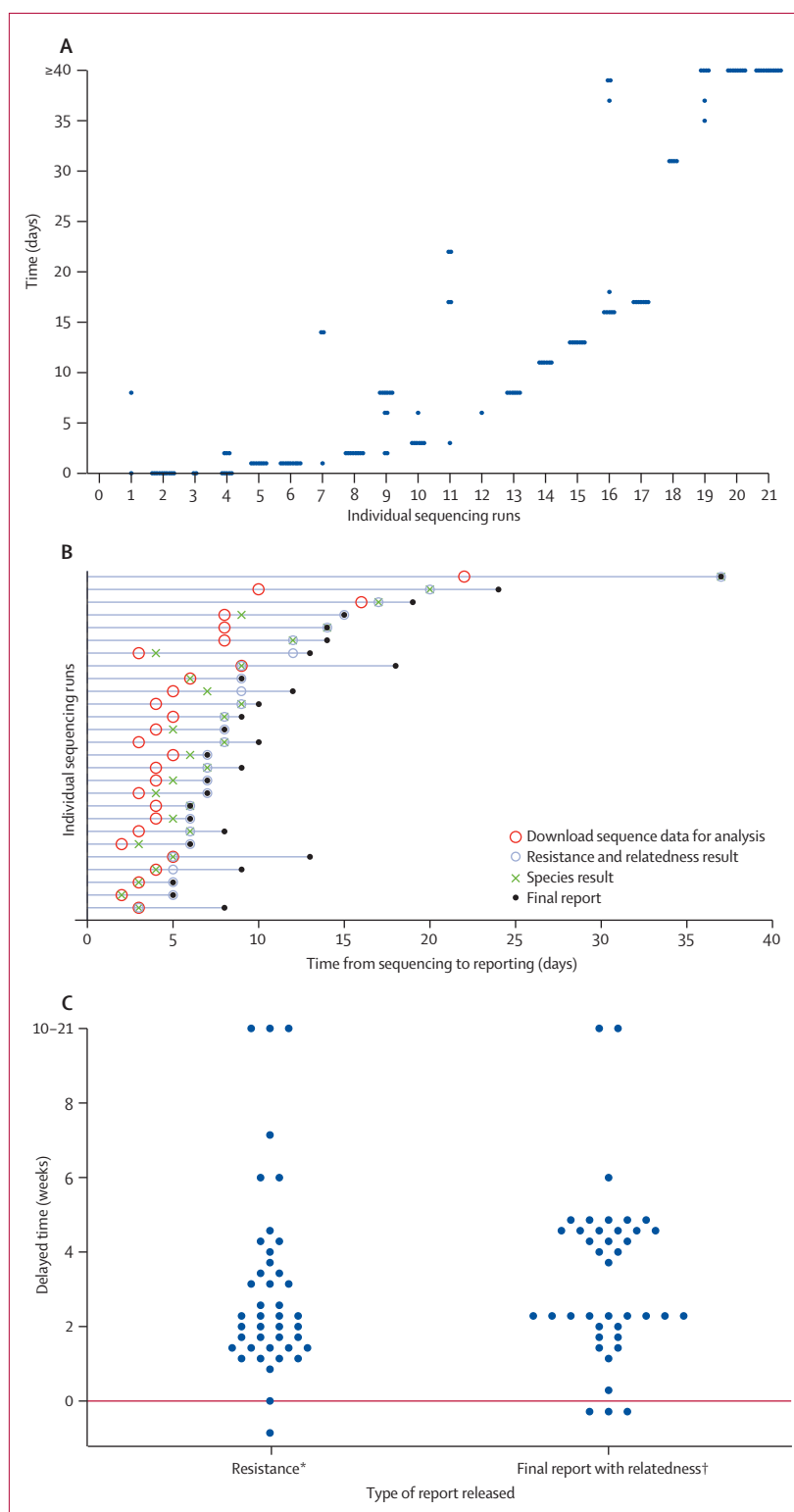
Discussion

In this prospective, multicentre, international pilot study, we assessed the real-time performance and cost of WGS for laboratory diagnosis of mycobacterial infection. With use of prototype software and only one sequencing attempt, 93% of mycobacteria were concordant with routine laboratory complex or species-level identification and 93% of susceptibility predictions for MTBC isolates were concordant with DST. The same sequencing data linked 16% of UK patients to an outbreak, identified an inter-regional cluster of isoniazid-resistant MTBC, and discovered transmission of multidrug-resistant tuberculosis, substantially hastening diagnosis and appropriate treatment of one patient. Generation of full diagnostic information by WGS was 7% cheaper than by routine methods.

This study provides proof-of-principle in high-income countries that local sample collection and sequencing with centralised analysis can be applied on an international scale. The median of 5 days between WGS data sharing and full diagnostic report availability compares favourably with present diagnostic workflows.^{5,12} The resistotyping and nearest-neighbour algorithms are rapid and scalable, retaining the resolution of whole-genome analysis. Our workflow does not depend on the specific algorithms used, which could easily be refined and improved. Consequently, Public Health England is currently assessing the suitability of WGS from early-positive MGIT cultures to replace routine clinical tuberculosis processing.²²

For species identification, Hain assays are capable of detecting 40 NTM and *M tuberculosis* species or complexes. The sequenced catalogue of 169 mycobacterial type strains used here provided broad diagnostic capacity. Although the gene presence or absence algorithm could not distinguish between different complex members, for

MTBC, this issue was resolved by phylogenetic analysis after reference genome mapping. This approach could be extended to other species' complexes.



WGS drug susceptibility predictions were highly concordant with the reference standard, simultaneously predicting first-line and second-line susceptibilities at no additional cost, despite incomplete knowledge of genotype–phenotype relations. Overall, the 17% incidence of drug resistance and 3% incidence of multidrug-resistant tuberculosis, as assessed by WGS, was well below occurrence in high-incidence settings.¹⁰ Of eight discrepant results, four had unclassified variants not in our prespecified catalogue, but with evidence of association with drug resistance in the wider scientific literature. Investigations have provided robust, evidence-based resistance prediction catalogues for use in future investigations and algorithms for the iterative addition of new or rare mutations to the prediction catalogue, with phenotypic support.^{6,8,10,14} No previously described mutations could explain phenotypic resistance for the remaining specimens; with only one sequencing attempt per isolate, these data could not be verified. Drug resistance might also be mediated by post-transcriptional or post-translational

protein modifications and efflux pump activation.^{23–25} Such resistance mechanisms remain little studied and should be investigated further as genotype–phenotype comparisons continue. However, DST methods for drugs other than isoniazid and rifampicin are also imperfect.^{26,27}

Minority variants complicating phenotypic predictions were present in 12 isolates. These variants might be due to emerging resistance or mixed infection within the patient, sample contamination with nasopharyngeal flora, or cross-contamination. We noted more mixed calls (4%) in the 16S gene (*rrs*) conferring aminoglycoside resistance than in the other resistance-conferring genes examined, potentially due to similar reads from other bacterial species mapping to the H37Rv reference. Because MGITs are inoculated with primary clinical samples, removal of other bacterial DNA, and subsequent resistance prediction where removal has been incomplete, pose challenges.²⁰ Despite this challenge, overall sensitivity was similar to that reported from other available drug susceptibility prediction tools and with use of pure-culture samples.^{6,8}

	Throughput in 2014 (n)*	Total per sample in 2014 (£)	10% fewer samples per year (£)	10% more samples per year (£)
WGS and routine clinical workflows				
MGIT culture	15 265	52.39	52.90	51.97
Cepheid Xpert MTB/RIF	617	99.66	102.35	97.44
WGS workflow only				
WGS	2207	118.55	120.16	117.26
Routine clinical workflows only				
Identification assays	2207	55.05	55.28	54.87
Hain MTBC	866
Hain CM/AS	1341
MIRU-VNTR	866	107.75	110.89	105.18
First-line DST	866	135.47	137.12	134.13
Limited second-line DST†	62	93.01	93.24	92.83
Second-line DST‡	62	101.27	104.24	98.86
WGS workflow scenarios				
MGIT culture and WGS	..	170.94	173.06	169.23
MGIT culture and WGS and first-line DST	..	306.41	310.18	303.36
MGIT culture and WGS and first-line DST and full second-line DST	..	500.68	507.66	495.05
Routine clinical workflow scenarios				
Culture and identification assays	..	107.44	108.18	106.84
Culture and identification assays and MIRU-VNTR and first-line DST	..	350.66	356.19	346.15
Culture and identification assays and MIRU-VNTR and first-line DST and full second-line DST	..	544.93	553.69	537.84
Total workflow costs				
WGS-based diagnostics	..	480.91	486.01	476.75
WGS-based diagnostics and first-line and full second-line DST	..	539.53	545.37	534.73
Routine clinical workflow-based diagnostics	..	518.31	524.00	513.64

Error rates reported in this study: 1% microscopy, 2% MGIT culture, 10% Cepheid Xpert MTB/RIF, <1% species identification (Hain ID), 13% DNA extraction for WGS, 4% WGS, 1% WGS data analysis, 10% MIRU-VNTR, and <1% DST. WGS=whole-genome sequencing. MGIT=Mycobacteria Growth Indicator Tube. MIRU-VNTR=mycobacterial interspersed repetitive unit-variable-number tandem repeat. DST=drug susceptibility testing. *Number reported in the Birmingham reference laboratory (Birmingham, UK) for 2014. Negative tests not counted for Hain GenoType MTBC (Hain Lifescience, Nehren, Germany), Hain GenoType Mycobacterium CM/AS, second-line DST, Hain GenoType MTBDRplus, or Hain GenoType MTBDRsl. †Done at the Birmingham clinical laboratory. ‡Done at a second reference laboratory (London, UK). Based on staff time, consumables, and equipment only.

Table 3: Total cost per sample by process, accounting for error rates

Throughout this investigation, WGS results were provided together with sequence quality feedback to allow laboratory staff to improve interpretation and practice. Other methods provide a small amount of data quality feedback and drug resistance identification, but do not provide species or outbreak analysis.^{6,8} Although WGS was not repeated in this study, retesting isolates when species identification, DST, or MIRU-VNTR deliver questionable results is routine in diagnostic laboratories, and such a system will be implemented for WGS. As more specimens are sequenced than at present, robust algorithms to identify isolates for resequencing will be developed to prevent inaccurate results being reported and improve WGS sensitivity and specificity.

16% of sequenced UK MTBC isolates were linked to one of nine UK outbreak clusters, including one spanning three regions. Large geographical distances separating genetically related isolates have been previously reported,¹⁰ with the authors concluding that low genetic divergence might not always represent transmission or that casual contact might be important in MTBC transmission. Application of WGS on a wide scale, as shown by the US Food and Drug Administration,²⁸ will only serve to increase the number of links detected between patients and outbreaks. Most surprisingly during this investigation, WGS diagnosed multidrug-resistant tuberculosis in one patient, only subsequently confirmed by the reference laboratory. This diagnosis directly affected the individual patient's care and reduced onward transmission risk. Identification of a second patient with genetically identical multidrug-resistant tuberculosis shows the need to rapidly identify infected patients to minimise the risk of transmission.

An often cited obstacle to clinical implementation of WGS is cost.^{29–31} For the laboratories and workflows costed here, high WGS costs for NTM diagnosis were outweighed by savings made in MTBC diagnosis, leading to an overall saving of 7% per year for a reference centre. If present DST workflows are continued alongside WGS, costs would be 4% greater per year than with present workflows alone. However, these additional costs would be mitigated by replacement of any molecular DST done alongside phenotypic DST with WGS. The cost-effectiveness of replacement of phenotypic with other rapid genotypic assays in terms of patient care has already been shown³¹ and is probably similar for WGS. The decentralised-sequencing, centralised-analysis model used in this study minimises computational and technical support costs. WGS costs could fall further when implemented diagnostically, which will need less skilled staff than at present. However, availability of skilled staff is likely to remain a key limitation of adoption of WGS in low-income settings. Fully automated analysis and reporting and strategic placement of benchtop sequencers would also reduce diagnostic delays, as reported in this study. Furthermore, progress in development of direct-from-sample and point-of-care WGS is continuing, and,

combined with the analysis algorithms shown, will revolutionise MTBC diagnosis.^{31,32}

WGS allows simultaneous prediction of mycobacterial species, first-line and second-line drug resistance, the ability to monitor emergence of new resistance mechanisms, and high-resolution outbreak monitoring on a timescale weeks faster than with traditional diagnostics. Coupled with public health interventions, WGS will transform MTBC patient care and disease control and has the potential to transform diagnosis of other infectious diseases.⁵ Furthermore, our cost estimates have shown that WGS-based diagnostics will provide value for money, demonstrating how WGS can replace mycobacterial diagnostic workflows from positive MGIT culture in high-income countries.

Contributors

DWC, SN, JP, TRR, EGS, PS, and MHW designed the study. CdOE and TMW designed and constructed the database and bioinformatics pipeline used in the study. JD designed the informatics infrastructure to support the study. LJP did the literature search. KC, DMG-B, TAK, CK, NL, LJP, ER, PT, AAV, TMW, and LX collected clinical specimens, extracted DNA, did whole-genome sequencing, and collected the clinical metadata associated with specimens. LJP, AAV, TMW, and DW did initial sequencing and quality control analysis. LJP and TMW collated and analysed sequencing and clinical data. JMF, EGS, and SW designed the health economics investigation and collected and analysed the data. LJP and JMF produced the tables and figures. DWC, LJP, JP, TRR, PS, AAV, TMW, MHW, and SW edited the report.

Declaration of interests

PS is a consultant for Genoscreen. TMW is a Medical Research Council Research Training Fellow. DWC is a National Institute for Health Research senior investigator. All other authors declare no competing interests.

Acknowledgments

All UK sites were supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre based at Oxford University Hospitals NHS Trust and University of Oxford. In addition, the research was supported by the Department of Health and Wellcome Trust through the Health Innovation Challenge Fund (T5–358) and the NIHR Health Protection Research Unit (HPRU-2012–10041). This report is independent research done for NIHR. The views expressed in this publication are those of the authors and not necessarily those of the NHS, NIHR, or the Department of Health. The British Columbia Centre for Disease Control, Vancouver, Canada, was supported by the British Columbia Centre for Disease Control Foundation for Population and Public Health. The Irish centre was supported by the Department of Clinical Microbiology, Trinity College Dublin, Ireland. The Lille and Borstel centres were supported by the EU Seventh Framework Program (278864; PathoNGenTrace project). We thank the microbiology laboratory and staff at each participating centre for assistance with sample collection. LJP thanks Phuong Quan for excellent advice about analytical and statistical queries and Timothy E A Peto for valuable assistance in producing the figures.

References

- 1 WHO. Global tuberculosis report. Geneva: World Health Organization, 2015.
- 2 van Ingen J, Bendien SA, de Lange WC, et al. Clinical relevance of non-tuberculous mycobacteria isolated in the Nijmegen-Arnhem region, the Netherlands. *Thorax* 2009; **64**: 502–06.
- 3 Paynter S, Hayward A, Wilkinson P, Lozewicz S, Coker R. Patient and health service delays in initiating treatment for patients with pulmonary tuberculosis: retrospective cohort study. *Int J Tuberc Lung Dis* 2004; **8**: 180–85.
- 4 Anderson LF, Tamme S, Brown T, et al. Transmission of multidrug-resistant tuberculosis in the UK: a cross-sectional molecular and epidemiological study of clustering and contact tracing. *Lancet Infect Dis* 2014; **14**: 406–15.

- 5 Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet* 2012; **13**: 601–12.
- 6 Feuerriegel S, Schleusener V, Beckert P, et al. PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J Clin Microbiol* 2015; **53**: 1908–14.
- 7 Parrish N, Carrol K. Importance of improved TB diagnostics in addressing the extensively drug-resistant TB crisis. *Future Microbiol* 2008; **3**: 405–13.
- 8 Coll F, McNerney R, Preston MD, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med* 2015; **7**: 51.
- 9 Witney AA, Gould KA, Arnold A, et al. Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. *J Clin Microbiol* 2015; **53**: 1473–83.
- 10 Casali N, Nikolayevskyy V, Balabanova Y, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 2014; **46**: 279–86.
- 11 Clark TG, Mallard K, Coll F, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. *PLoS One* 2013; **8**: e83012.
- 12 Koser CU, Bryant JM, Becq J, et al. Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. *N Engl J Med* 2013; **369**: 290–92.
- 13 Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* 2013; **13**: 137–46.
- 14 Walker TM, Lalor MK, Broda A, et al. Assessment of *Mycobacterium tuberculosis* transmission in Oxfordshire, UK, 2007–12, with whole pathogen genome sequences: an observational study. *Lancet Respir Med* 2014; **2**: 285–92.
- 15 Walker TM, Kohl TA, Omar SV, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis* 2015; **15**: 1193–202.
- 16 Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 2011; **364**: 730–39.
- 17 Roetzer A, Diel R, Kohl TA, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med* 2013; **10**: e1001387.
- 18 Eldholm V, Monteserin J, Rieux A, et al. Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat Commun* 2015; **6**: 7119.
- 19 Guerra-Assuncao JA, Crampin AC, Houben RM, et al. Large-scale whole genome sequencing of *M tuberculosis* provides insights into transmission in a high prevalence area. *Elife* 2015; **4**: e05166.
- 20 Votintseva AA, Pankhurst LJ, Anson LW, et al. Mycobacterial DNA extraction and whole-genome sequencing from early positive liquid (MGIT) cultures. *J Clin Microbiol* 2015; **53**: 1137–43.
- 21 Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: Wiley, 2008.
- 22 Public Health England. March, 2015. Annual TB update 2015. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/415540/Annual_TB_update_2015.pdf (accessed Nov 17, 2015).
- 23 Cain JA, Solis N, Cordwell SJ. Beyond gene expression: the impact of protein post-translational modifications in bacteria. *J Proteomics* 2014; **97**: 265–86.
- 24 Siddiqi N, Das R, Pathak N, et al. *Mycobacterium tuberculosis* isolate with a distinct genomic identity overexpresses a tap-like efflux pump. *Infection* 2004; **32**: 109–11.
- 25 Xie L, Liu W, Li Q, et al. First succinyl-proteome profiling of extensively drug-resistant *Mycobacterium tuberculosis* revealed involvement of succinylation in cellular physiology. *J Proteome Res* 2015; **14**: 107–19.
- 26 Horne DJ, Pinto LM, Arentz M, et al. Diagnostic accuracy and reproducibility of WHO-endorsed phenotypic drug susceptibility testing methods for first-line and second-line antituberculosis drugs. *J Clin Microbiol* 2013; **51**: 393–401.
- 27 Fricke WF, Rasko DA. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat Rev Genet* 2014; **15**: 49–55.
- 28 US Food and Drug Administration. March 5, 2015. Whole genome sequencing (WGS) program. <http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS> (accessed July 2, 2015).
- 29 Hasnain SE, O'Toole RF, Grover S, Ehtesham NZ. Whole genome sequencing: a new paradigm in the surveillance and control of human tuberculosis. *Tuberculosis* 2015; **95**: 91–94.
- 30 Lecuit M, Eloit M. The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. *Front Cell Infect Microbiol* 2014; **4**: 25.
- 31 Drobniewski F, Cooke M, Jordan J, et al. Systematic review, meta-analysis and economic modelling of molecular diagnostic tests for antibiotic resistance in tuberculosis. *Health Technol Assess* 2015; **19**: 1–188.
- 32 Brown AC, Bryant JM, Einer-Jensen K, et al. Rapid whole-genome sequencing of *Mycobacterium tuberculosis* isolates directly from clinical samples. *J Clin Microbiol* 2015; **53**: 2230–37.