Proceedings of the Nutrition Society (2009), **68**, 113–121 doi:10.1017/S0029665109001037
 First published online 11 February 2009

A Meeting of the Nutrition Society, hosted by the Irish Section, was held at the O'Reilly Hall, University College Dublin, Dublin, Republic of Ireland on 18–20 June 2008

Symposium on 'The challenge of translating nutrition research into public health nutrition'

Session 2: Personalised nutrition
Genetic variation and disease risk: new advances

John Scott
School of Biochemistry and Immunology, University of Dublin, Trinity College, Dublin 2, Republic of Ireland

Variations in human DNA, most frequently single-nucleotide polymorphisms (SNPs), can have functional consequences ranging from severe to none. Variations in outcome (phenotype) can be compared, from cystic fibrosis through haemochromatosis to general familial risks in, for example, colo-rectal cancer (CRC). Cystic fibrosis and haemochromatosis have severe phenotypes with high penetrance, with signs and symptoms always or mostly present; thus, they have been easy to identify from family studies. However, the familial risks that are known to contribute markedly to CRC are unknown. The sequencing of the human genome has now made possible the identification of these and other disease variants. Knowing the DNA sequence in an idealised individual adds little unless variants that increase (or decrease) disease risk from the norm can be identified. Such variants can be expected to be very common in the general population, but have low penetrance and only change risk to a limited extent. Many patients will not have the risk variant and many 'normal' patients will have the risk variant. Thus, very large case–control cohorts are essential. These case–control cohorts can be analysed at three different levels: (1) individual SNPs; (2) individual genes; (3) genome-wide analysis (GWA). Level 1 looks for case–control differences for specific SNPs. Alternatively, new technology can be applied to examine a range of SNPs within a gene to track differences in its regulation as well as in function. Finally, the whole genome with $\geq 0{\cdot}5 \times 10^6$ SNPs could be marked. The first two approaches involve selecting 'candidate' SNPs or genes, while GWA looks for any variation in the genome that is enriched in the cases. All three approaches carry the certainty that significant associations will be found by statistical chance, for which correction must be made. This latter issue is helped by large numbers and by independent replication cohorts.

Genetic variation: Disease risk: Individual SNPs: Individual genes: Genome-wide analysis

Traditionally, considerations of genetic variation were restricted to what conventionally would have been termed 'inborn errors of metabolism'[1]. However, it was also appreciated that there was a strong association between family history and risk of almost every common disease. For example, having a parent, sibling or close relative with colon cancer would have increased an individual's risk of having this form of cancer. While some of these increased risks could be explained by probable common environmental risks, it was clear that for particular diseases a proportion of the increased risk was a result of the inheritance of genetic variants. Inborn errors showing a genetic variation are more properly termed mutations because of their low prevalence, where prevalence is the frequency with which the variant occurs in a population; an example would be classical haemophilia or cystic fibrosis. The invariable presence of the disease for those individuals with these rare variants is described as such a variant showing 'high penetrance'. Generally, the genetic variants responsible were relatively easily identified through comparison

Abbreviations: NTD, neural-tube defect; SNPs, single nucleotide polymorphisms; GWA, genome-wide analysis.

Corresponding author: Professor John Scott, fax 353 1 6772400, email jscott@tcd.ie

of the sequence of DNA between cases and unaffected relatives. However, as the technology for sequencing DNA improved and became more automated, the identification of the genetic cause of such diseases increased. Furthermore, it soon became evident that in the case of some diseases subjects with the required genetic variant, even in the homozygous state, did not always develop the disease, unlike the classical in-born errors where the variant always resulted in the disease. It therefore became clear that some genetic variants confer a disease risk that often, but not always, leads to the signs and symptoms of the disease[1]. This finding established the concept that this type of variant has intermediate penetrance, an example being hereditary haemochromatosis.

Of these two types of genetic variation, inborn errors of metabolism have been traditionally of little general public health interest or importance because, although they often have devastating consequences for the affected individual, they are very rare. A good example of such a rare high-penetrance variant would be cystic fibrosis. An example of a more-frequent variant with variable penetrance would be haemochromatosis. Such variants, which may be arbitrarily described as infrequent variants compared with, for example, cystic fibrosis, are more common and have intermediate penetrance; examples include hereditary haemochromatosis and BRCA 1 and BRCA 2 variants for breast cancer. These intermediate-frequency variants with medium to high penetrance are of some public health interest as well as being of great interest at the level of the individual. However, a third type of genetic variation is really the most important from the public health point of view. It is now clear that there are a considerable number of frequent variants with low penetrance. These variants are considered to be the genetic variations of real public health interest[2]. They are known from epidemiological evidence to be very common and therefore constitute a large burden of disease risk. They could be expected to have a heterozygous prevalence of half or even a majority of the population, with a corresponding homozygous prevalence between 13% and 29%. Thus, in the present review genetic variation has been rather arbitrarily divided into three types: rare mutations with high penetrance; infrequent variants with medium to high penetrance; frequent variants with low penetrance. This arbitrary distinction and difference will eventually merge with the identification of numerous examples of each type. However, it introduces the two most important considerations of genetic variation and disease risk, i.e. the frequency of the variation and the penetrance of the variation.

The concept of frequency is easy to understand and can be illustrated with practical examples, but the concept of penetrance is more difficult to explain and even to understand at a biological and clinical level. It is not to be confused with the clinical severity of the ultimate outcomes, usually termed the phenotype. It is rather more the concept that for many genetic variants there is a wide inter-individual variation in the ultimate consequence of the presence of the variant from the perspective of the severity of the clinical signs and symptoms and ultimately an ensuing disease and its effects. Thus, for the classical inborn errors the presence of the variant has very predictable and inevitable consequences. The consequences of such rare mutations or variants are thus often well understood both at a clinical and even biochemical level. On the other hand, with the infrequent variants, which often have medium to high penetrance, the presence of the variant does not inevitably carry clinical or disease consequences; an individual may have the variant in the complete absence of any signs or symptoms. Some of these variations in penetrance may be explicable on the basis of age or gender. However, currently, there remain many unanswered questions as to why penetrance for some genetic variants differs so much between individuals. In general terms, some of this difference may be attributable to differences in environmental exposure, although often the reason is unclear. Undoubtedly, many of these differences in penetrance are attributable to the presence or absence of other genetic variations that protect or exacerbate the biochemical, and ultimately clinical, effect of the original genetic variant. In relation to the third category of genetic variation, the frequent variants with low penetrance, the variant itself may not be clinically or even biochemically linked to a disease in any obvious way. However, the presence of the variant increases the risk of a disease in some complex and often unclear way. The sequencing of the human genome has introduced a new era for this whole area via case (patient)–control comparison of the biochemical and clinical consequences of the presence of a particular genetic variant. While the initial effort and consequent outcome in this area of research will inevitably concentrate on risk or potentially-harmful variants, such investigations will inevitably begin to identify more and more beneficial variants, which will explain why the anticipated clinical consequences of the presence of a particular harmful variant are less severe or even absent.

Rare mutations (inborn errors of metabolism) with high penetrance

The main characteristics associated with such mutations are summarised in Table 1, which also shows a few specific examples of well-known inborn errors such as cystic fibrosis. It is useful to take one such example and see how well it corresponds with the characteristics. Cystic fibrosis is a rare homozygous recessive genetic mutation, in that to get the disease requires two affected genes, one inherited from each of the parents, who are thus described as ‘carriers’. This inheritance ultimately produces two altered identical underperforming proteins. These proteins in cystic fibrosis are involved in the regulation of Na channelling through the epithelial cells that line the gut wall and also, importantly, the lungs. A decreased ability to secrete Na causes an accumulation of mucus[3].

It is suggested that the presence of one impaired copy of the protein, as would exist in those heterozygous for the variant, would decrease Na loss to a more manageable but not critical extent. During evolution the presence of such a variant in the face of a fluid-losing disease such as cholera would have been a selective advantage. However, it is clear that the dramatic reduction in the ability to extrude Na that exists in those individuals who are homozygous for cystic fibrosis has profound clinical consequences. The

Table 1. Characteristics of mutations ('inborn errors of metabolism'), infrequent variants and frequent variants with well-known examples

	Mutations 'inborn errors of metabolism'	Infrequent variants	Frequent variants
Phenotype	Severe	Severe, moderate or mild	Mild or absent
Clinical sequelae	Severe	Severe, moderate or mild	Not absolute but increased
Penetrance	High (disease usually or always present)	Severe to infrequent	Mostly low or general risk
Frequency of the variation	Rare; selected against during evolution	Relatively common; founder mutations; not selected against	Very common; probably no evolution selection
Importance	Individually important	Individually important, community important	Individually mostly unimportant
General public health	Unimportant	Often considered important	General importance
Genetic screening and counselling	Desirable or essential	Desirable if possible	Not helpful
Biochemical insights into function	High	Moderate to high	Only general
Examples			
	Cystic fibrosis PKU Homocystinuria	BRCA variants for breast cancer (high penetrance): 80% lifetime risk of disease; risk-dependent, women, risk in later life Haemochromotosis variants, for Fe overload (medium penetrance): homozygous prevalence in Ireland is one in eighty; lifetime risk of arthritis 10%, cirrosis 5%, diabetics 2%; risk dependent for: (1) men *v.* women (blood loss); (2) age, chronic disease; (3) unknown susceptibility to liver damage; (4) alcohol abuse	Common low-penetrance variants: CRC (most cases) NTD: C677T MTHFR: homozygous 10%, heterozygous 40%, wild type 50%; 15% population attributable risk G1958A MTHFD1 (mitochondria): material risk of an NTD birth
Variant in the population	If an individual has the variant, they get the phenotype or disease (high penetrance)	Those with the variants generally get the disease	Many of the general population will have the variant
	Easy to identify the genetic variant through family studies		Variants are low penetrance (many controls will have the variant and no disease): did not get it yet (e.g. age, environment); have other protective genetic variants or other environmental (diet) protection
	Variant will be in grandparents, parents, siblings etc.		Many cases will not have the variant: other causes (genetic or environmental) of the disease; other different diseases under a single heading; variant only changes risk (and other factors necessary, e.g. poor diet, allergies)

PKU, phenylketonuria; CRC, colo-rectal cancer; NTD, neural-tube defects; MTHFR, 5,10-methylenetetrahydrofolate reductase; MTHFD1, methylenetetrahydrofolate dehydrogenase/methenyltetrahydrofolate cyclohydrolase/formyltetrahydrofolate synthase.

most common clinical manifestation is an accumulation of mucus in the lungs with an increased susceptibility to infection, a poor quality of life and a premature death. Thus, for cystic fibrosis, as for most of the well-established inborn errors of metabolism, being homozygous for the variant equates with manifestation of the disease. The inevitable process of obvious and severe clinical signs and symptoms of such mutations results in them showing what is called 'high penetrance'. Thus, the disease is always or certainly usually present. Being able to identify those individuals with and without the variant was important in the identification of the gene involved, i.e. looking for a variant present in cases and absent in controls. These controls were frequently siblings or close relatives without the disease but with a large extent of overlap in their genetic make-up. Thus, the offending variant was traditionally easy to identify even though ultimately this process involved very intensive genetic sequencing, which at the time was slow and labour intensive. By such efforts the cystic fibrosis variant was identified by two groups simultaneously in 1991[4]. It subsequently became clear that approximately three-quarters of northern European diseases result from a single amino acid change[5]. The genetic and often molecular basis for other inborn errors was likewise established before the human genome was sequenced in its entirety. Inborn errors have often given biochemical insights into specific biochemical mechanisms, being in a sense nature's genetic mutations. They often have huge clinical consequences for the case involved and their families, frequently requiring a large clinical input and genetic counselling. However, while important at the level of the individual their rarity makes them of little general public health importance. In addition, their rarity makes general screening of the public for their

presence currently not cost effective, although as the technology changes, and with it the potential to generate individual DNA profiles, their routine identification may become a reality. It remains to be seen whether this knowledge is one that individuals or indeed societies will embrace.

Infrequent variants

While the presence of inborn errors and their genetic cause by specific genetic variations had been known for many years, what was also appreciated was that genetic variation could also carry with it a high risk of a particular phenotype or disease that while frequently present was not always present as in inborn errors. Such genetic diseases were often severe and clinically like inborn errors, although unlike inborn errors the signs and symptoms were not always present in those individuals who were known to have the variant. The characteristics of such infrequent variants showing this variable extent of penetrance are summarised in Table 1, together with two specific examples that illustrate the outlined characteristics, BRCA variants for breast cancer and haemochromatosis.

BRCA 1 and BRCA 2

Two variants exist termed BRCA 1 and BRCA 2, and while they carry a greatly increased risk for breast cancer, the disease is not always present[6]. The first obvious reason for this situation is the fact that only women are at significant risk; although such high cancer risk occurs in men, it is much lower because of their small amount of mammary tissue. Risk is also highly dependent on age, increasing in women as they mature, with risk gradually reducing in later life as their breast tissue decreases. However, the lifetime risk for either BRCA 1 or BRCA 2 is between 50% and 80%[7]. Thus, the reality is that generally approximately 25% and possibly $\leq$50% of women in whom the presence of the variant has been established never develop breast cancer, at least to the point where it is recognisable even at its earliest clinical stages.

Haemochromatosis

Another example of a variant that is relatively infrequent with only medium penetrance is hereditary haemochromatosis (Table 1). Almost all this disease is caused by one single nucleotide variation, G845A, which causes a change in the usually-coded cysteine residue to a tyrosine residue at amino acid 282 in the HFE protein[8]. This protein, an MHC class-I-like protein, is now known to interact with transferrin receptor 1, which in turn mediates transferrin-bound Fe uptake into almost all cell types[9]. The C282Y mutation disrupts a disulphide bridge in HFE, which is needed for its binding to β2-microglobulin, a necessary step in the stabilisation of HFE and its expression on the cell surface where it interacts with transferrin receptor 1[9]. HFE has a minor, but constant, role in the regulation of the synthesis of what is now known to be the major regulation protein for Fe absorption, i.e. hepcidin. Hepcidin synthesis is also more fundamentally governed by two other proteins, transferrin receptor 2 and haemojuvelin[10]. Genetic variants that affect haemojuvelin result in a dramatic high-penetrance form of the disease termed juvenile haemochromatosis. Such variants that directly affect the synthesis of hepcidin are therefore very high penetrance but very rare; they are thus more like true inborn errors. By contrast, the C282Y mutation interferes with only part of the control of hepcidin synthesis, producing a gradual plasma loading and a general accumulation of Fe in tissues[10]. Thus, as might be expected, patients with the C282Y mutation are free of clinical symptoms until later life, and even this deferral is mitigated downwards in women because of loss of Fe in menstrual bleeding during their child-bearing years. However, there is a large spectrum in the extent to which the presence of the variant is reflected differently in different individuals, all of whom are homozygous for the C282Y mutation. Some individuals will have a marked increase in the extent of Fe attached to circulating transferrin and a very marked increase in their circulating Fe level. Others with the same nutrition will have relatively modest increases in these two variables. Obvious explanations for this difference in outcome might be sought in inter-individual differences in dietary Fe, but such a difference is not found to be the explanation. Similarly, a difference in chronic loss of blood, and with it Fe loss, could explain the difference in cellular and plasma Fe levels, but apart from the effect on younger females this explanation is apparently not the answer. It thus seems probable that other as yet unknown compensating or exacerbating genetic variants exist in other genes that influence the overall outcome, inevitably resulting in a wide spectrum of effect when comparing individuals of quite similar age and backgrounds, all of whom are homozygous for the C282Y variant. Such a spectrum of different penetrance is of itself difficult to explain at a biological level, but what is even more perplexing is that the effects of an excess of circulating Fe and high transferrin saturation also differ widely for no apparent reason. Thus, some subjects with relatively modest increases in circulating Fe and at a relatively young age have severe clinical consequences of having the C282Y mutation, including damage to the liver (cirrhosis), the pancreas (diabetes) and the joints (arthritis). Other subjects with really massive increases in circulating Fe, as indicated by the biomarker ferritin (the diagnostic marker for Fe status), and almost total saturation of their transferrin transport protein, have no clinical signs or symptoms. Additionally, when treated to remove the excess Fe through weekly pheblotomy such patients do not seem to have any increased long-term risk of these conditions over those without the variation. It is unclear why these differences in penetrance exist. Again, like the differing levels of circulating Fe in subjects with C282Y for the variant the extent of damage does not have any obvious environmental or nutritional explanation. Thus, such variation in penetrance is usually attributed to the presence or absence of other genetic variants that increase or decrease the response to the insult that cells are confronted with as a result of having excess Fe. Such a variant might help cells by facilitating a better response to the damage caused by excess Fe deposition. Alternatively, it might be better for cells to ignore a constant untreated insult rather than to

mount a strong defence response, be that either an antioxidant one or an immune one. Such a chronic response might eventually produce undesirable effects such as increased fibrosis, leading in turn to a change in the architecture of the liver or pancreas, with resultant cirrhosis or diabetes.

The general characteristics of infrequent variants of medium to low penetrance need to be considered from the public health viewpoint (Table 1). In the specific case of haemochromatosis it seems probable that as many as one in eighty of the Irish population are homozygous for the C282Y variant. It is very difficult to get a measure of the true clinical consequence[11]. It is very probable that a substantial burden of early and potentially-treatable arthritis has its origin in the lack of diagnosis and treatment. Similarly, a not infrequent finding when patients attending diabetic clinics are screened for the variant is its enrichment over that found in the general population. Similarly, a proportion of liver damage ultimately leading to cirrhosis may have a percentage of patients with the variant as the root cause. This latter case is not made any clearer, particularly historically, by seeking the explanation for such cirrhosis in the excess use of alcohol, which of course is a completely independent cause of liver damage. Confusion between haemachromatosis and alcoholic liver disease has been exacerbated by the fact that the group at highest risk of haemachromatosis would be elderly males who traditionally also comprise a large proportion of those with alcoholic abuse.

There is also an interesting debate as to the value of screening. It is considered in most countries that, although screening seeks to identify about one in eighty in the population, the very variable penetrance probably means that the lifetime risk of serious clinical consequences for those homozygous for the C282Y variant is relatively small. Thus, in general, countries do not screen the general population. Their response is more to indicate that an estimation of ferritin is appropriate in the investigation of most elderly males and those females with any compatible symptoms. An elevated ferritin level on its own means little, since it is elevated in infection and other conditions. However, an elevated ferritin level accompanied by a high transferrin saturation is diagnostic of haemochromatosis. The genetic test may be undertaken for confirmation but a serum ferritin estimation as a means of exclusion (or inclusion) of haemochromatosis is relatively simple. It seems appropriate that it would be more widely used, even if not screened for, bearing in mind that haemochromatosis is completely treatable by pheblotomy in its early stages.

Frequent variants

Disease risk

The two forms of genetic variation already discussed are either rare or at least infrequent; thus, even though they contribute respectively to invariable and frequent clinical consequences, the overall impact on the general community is not large. Historically, the inborn errors and the other medium-to-high penetrance variants were identified progressively through family studies. However, it was always understood that a genetic component, sometimes a very important component, contributed to many chronic diseases such as cancer, CVD, osteoporosis and type 2 diabetes[12]. In fact, all diseases would have variable risk that is inherited. The key difference is that while these variants are very common, each such variant just contributes to an increase in risk to a greater or lesser extent. As summarised in Table 1, many apparently-normal individuals may not have the disease or in fact never get the disease, yet they have the risk variant; however the reason for this outcome is unclear. Clearly, some diseases only manifest themselves with advancing years. Even allowing for this age factor it may be that other environmental triggers are needed in concert with a risk variant in order to trigger the disease (Table 1). Thus, factors such as exposure to a cancerous xenobiotic, over- or under-provision of a nutrient or an infection may be necessary to convert a risk from a specific variant into an expression of a disease. Equally, for risk to become a reality may require the concomitant involvement of other common variants[13]. Such other variants may have the effect of increasing risk, e.g. two variants impacting either independently or dependently on a vital process. It seems also very probable that there may be beneficial variants that act to reduce the disease risk of having a negative variant. Two specific examples are highlighted in Table 1, neural-tube defects (NTD) and colo-rectal cancer.

Neural-tube defects

The presence of one variant associated with folic acid metabolism, i.e. C677T polymorphism for the enzyme 5,10-methylenetetrahydrofolate reductase, makes a small but important contribution to the explanation of a birth that is affected by spina bifida or another NTD, which is described as the population attributable risk[14]. The variant is very common in the Irish population and those of European origin, with a widely different prevalence worldwide. The prevalence of being homozygous or heterozygous for the variant is approximately 10% and 40% respectively[14]. Thus, over half the population have at least one flawed copy of this variant and thus in turn so does every child born in the country. Clearly, half the population are not born with an NTD. The real prevalence rate for NTD has dropped from a high of seven per 1000 births to about one per 1000 births[14]. Clearly, many of these children will be heterozygous for the variant and a substantial minority will be homozygous for the variant[15]. Case–control studies comparing affected births with unaffected births report a population attributable risk of 15% and 25% for being homozygous and heterozygous respectively. This risk value is a combination of two considerations, the numbers of births with either one or both genes affected and superimposed on these data is the actual prevalence of NTD at the time in question. It is thus clear that for any 1000 births approximately 100 babies will have both genes and a further 400 babies will have one affected gene. However, there would on average be only one affected birth per 1000 births. Thus, this variant is a very good example of a very common variant but one that

Table 2. Technological approaches used in case–control comparisons to investigate genetic variants and disease risk

Single-nucleotide polymorphism (SNPs)	Single gene	Genome-wide analysis
Case–control difference	Between five and twenty SNPs, most single genes could be covered and cover depends on the size of the gene and HapMap	$\geq 0{\cdot}5 \times 10^6$ SNPs
Difference for individual SNPs	The following are marked:	Ranged along whole genome
Biological plausibility: SNPs changes on amino acid near active site, inhibitor site	(a) Reading frame (amino acid, protein)	Case–control comparison
Area of gene conserved in nature	(b) Splice sites (between introns and exons)	Depends on HapMap
Now low cost: 10 cents per subject per SNPs; 500 cases, 500 controls, one SNPs €100; 5000 cases, 5000 controls for five SNPs €5000	(c) 5′ Regulatory region	Looks for any part of the gene (SNPs) that is increased (risk) or decreased (benefit) in cases compared with controls
	(d) 3′ mRNA region	Cost €800 000
	(e) Insertions, deletions, repeats	
	'Illumina' run*: 1500 SNPs, full analysis of 100–120 genes €200 000	

HapMap, haplotype map of the human genome, cataloguing common human genetic variants[22].
*Analysis using technology provided by Illumina, San Diego, CA, USA (see text).

has very low penetrance. For a birth to be affected, the presence of the variant constitutes only an increase in risk, and even then it is a relatively small contribution. The meaning of a population attributable risk needs to be understood[14]. It suggests that the presence of two affected genes causes a 15% risk but still leaves 85% of the risk unexplained in those homozygous babies. In the particular case of NTD there is one known large environmental or nutritional factor, i.e. folate status. Progression from a low to a high folate status, within the range of folate status considered to be adequate, reduces the risk of an affected birth by a factor of 10[16]. NTD appear to be a single condition that is approximately 75% preventable by the ingestion of additional folic acid[16].

A further example of a high-frequency low-penetrance variant is the variant of the mitochondrial form of the so-called trifunctional enzyme methylenetetrahydrofolate dehydrogenase/methenyltetrahydrofolate cyclohydrolase/formyltetrahydrofolate synthase (Table 1), which produces an amino acid change in the synthase zone of the enzyme. It was initially identified[17] and subsequently confirmed to be a maternal risk for an NTD-affected birth[18]. The function of the enzyme is in transfer of C_1 units within the mitochondria, which are then exported to the cytoplasm and are known to be the major source of C units for purine and thus DNA biosynthesis. Thus, it is understandable that an alteration in its activity or in its expression (amount made) would potentially increase risk for NTD, or indeed other risk such as the recent emergence of apparent risk for heart disease[19,20], which was identified using genome-wide analysis (GWA; see later). The comparison of how these two examples of risk have emerged demonstrates how the field of genetic variation has changed dramatically during the last two decades.

Approaches used to study genetic variants and disease risk

Candidate single-nucleotide polymorphisms

In earlier studies of inborn errors the approach was to sequence genes that were thought to be involved, which was time consuming and expensive. This approach was then largely replaced by investigating changes in the sequence of a specific base in the DNA, i.e. single-nucleotide polymorphisms (SNPs; Table 2). Early methods for investigating SNPs used PCR technology and primers of the sequence adjacent to the gene of interest. The ultimate DNA produced was then subjected to digestion with a specific endonuclease, which cleaved the single-stranded DNA at specific sites depending on whether or not there was change in nucleotide sequence of the DNA. The resultant fragments were initially identified on denaturing polyacrylamide gels, with subsequent detection using MS technology. However, more recently, demands for SNPs detection in large numbers of individuals, both cases and controls, has led to the availability of low-cost technology for this purpose. The author's group has used one technology provided by K Bioscience UK (Hoddesdon, Herts., UK), which together with similar commercial alternatives have made possible the analysis of very large numbers of SNPs in the same or indeed different genes and is now within the grasp of any research group. It might be expected that such advances would lead to the investigation of large numbers of candidate SNPs in large numbers of different genes. Certainly, this approach was followed initially, and much effort was invested in deciding which SNPs would be good candidates, driven largely by choosing SNPs that could be expected to alter function (Fig. 1). This process produced a list of priorities for choosing how attractive an alteration in a particular SNPs might be in the form of the questions: is it in the coding region, which would thus mean that it would have a direct effect in the protein produced; would it alter an amino acid in this protein and, if yes, is this amino acid situated in what looks like a functionally-important area of the protein; is this section of the gene conserved from species to species in evolution, suggesting some essential and sustained function. In retrospect, while a lot of these speculations were interesting, they were of little real value. Even if the three-dimensional structure of the protein in question was understood and thus the amino acids likely to be involved in the active site could be determined, it was clear at a biological level that specific amino acids or short sequences far removed from the active site are frequently important. They could be involved at a regulatory site, for

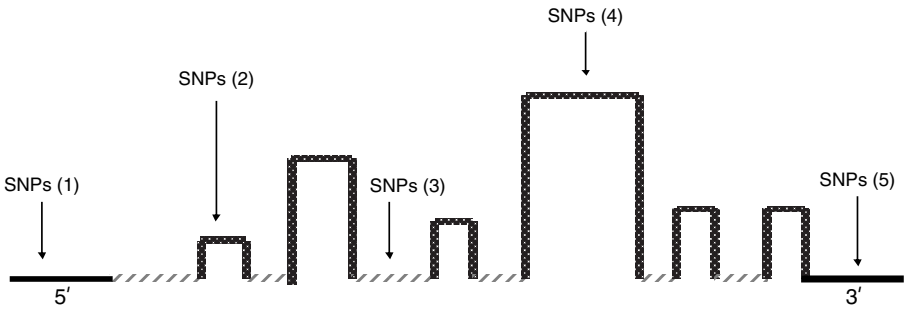

Fig. 1. Section of the human genome, illustrating single-nucleotide polymorphisms (SNPs) in (1) the 5′ regulatory region, (2, 4) introns (▩), (3) an exon (▨) and (5) the mRNA coding sequence.

the binding of an allosteric inhibitor or agonist. They could be involved in a chaperone function and thus involved in bringing the protein to its place of function. Equally, such distance sequences determine the turnover or half-life of the protein. Thus, predicting which amino acid changes would and would not matter is a risky business. What was also emerging is that the intronic and non-coding areas of genes are important (Fig. 1); although this factor has always been appreciated for SNPs at splice sites between introns and exons. It was also clear that the regulatory region that controls gene expression is in the 5′ region upstream from the exons that determine the sequence of the ultimate protein. It became clearer that such regulation sites could not always be expected to be proximate to the start site. They could be some distance away and might involve expression and regulation of other gene products. Similarly, more examples became evident in which the 3′ downstream untranslated region is clearly important. The most obvious reason is that the mRNA sequence contains a sequence at the 3′ region and beyond the last exon. This mRNA section has a large effect on the stability of the mRNA and with that the amount of protein that is actually made by the ribosome. The other complication of looking for other functions and only relating them to altered SNPs is that quite a lot of genetic change, perhaps as much as 30%, is a result of insertions, deletions or repeating sections of DNA, with the 3′ region perhaps also having a regulatory function[21].

Candidate genes

All this focused research was looking at candidate genes rather than candidate SNPs (Table 2). This approach largely ignores functional considerations in the direct sense but depends on the fact that DNA is inherited from the forefathers and parents in sections. However, the size of this intact section varies; the older the variation the greater the extent to which it becomes truly mixed over evolution. This aspect, however, has been accurately chronicled by the emergence of the so-called HapMap (a haplotype map of the human genome, cataloguing common human genetic variants)[22]. Thus, a section of the human genome can be taken and SNPs selected that will reflect any alteration in part of a gene that shows differences between a case and a control. The number of SNPs needed to completely mark a particular gene is dependent on this HapMap and, of course, the size of the gene in question. However, it is now possible to determine a selection of SNPs that will identify with confidence any other inheritance patterns. These SNPs will be selected not on some potential change in function but purely on their ability to characterise or mark the section of the genome or a particular gene that is under investigation. This approach thus moves from looking at individual SNPs that may be directly involved in a disease to characterising a specific gene with multiple SNPs, i.e. going from a candidate SNPs approach to a candidate gene approach. The number of SNPs needed for reliability, as outlined earlier, depends on the size of the gene and the HapMap. For some areas of the genome that are not highly equilibrated over evolution the HapMap may indicate that an individual SNPs can cover a large section. The practical result is that on average as few as five or as many as thirty SNPs are needed to characterise a particular gene. The great advantage of looking at individual genes rather than individual SNPs is that risk associations can be sought not just for risk of alterations leading to an altered amino acid and therefore an altered protein, but the involvement of a base change in the 5′ regulated region can also be identified. Furthermore, it will track changes in the 3′ region that may increase or decrease the half-life of the mRNA through alterations in the section of the DNA at the 3′ prime end. In addition, it is appreciated that while disease risk continues to be identified the issue is not infrequently that multiple copies of a section of DNA may be repeated, resulting in variable amounts of the enzyme or protein being synthesised. A good example is one to five repeats of a sequence that codes for thymidylate synthase[23]. Equally, a sector of the gene may be deleted or there may be variable copy numbers. Looking at sufficient SNPs can, of course, use the commercially-available options mentioned earlier. However, a point comes when looking at large numbers where a more appropriate technology is required, such as that offered by Illumina (San Diego, CA, USA)[24]. This technology uses primers of the sequence on either side of the SNPs to be tested, where the primers will not allele to the DNA if a SNPs has been inserted. The usual package would involve simultaneously looking at 1500 SNPs in hundreds or even thousands of cases and controls in a single disease. While 1500 SNPs may seem to be a large number, it is worth considering gene size and the HapMap. A single ‘Illumina’ run has the potential to analyse 100 to 150 genes at the most.

Genome-wide analysis

This approach of looking for markers as distinct from functional SNPs has its ultimate expression in GWA[24] (Table 2). Here, a few companies employing commercially-patented technologies use usually >750 000 SNPs to analyse the whole genome in one single analysis (Table 2). This approach makes substantial use of the HapMap, which is now elucidated in the literature and even adapted to take into account differences in ethnic origin. It again relies on DNA being inherited in sections through ancestors, parents etc. As with using SNPs markers for individual genes, it depends on the concept that if a particular genetic variation is involved in increasing or even decreasing the risk of a disease, the genetic variation involved will have transferred with its adjacent sections of DNA. While adjacent sections will not be involved in the disease if they are marked by a particular SNPs, an altered distribution between the cases and controls will identify a similar altered distribution in a genetic variation that is a true risk for the disease in question. What the technology will do is identify a section of the genome that differs significantly between cases and controls. This section of the genome could be quite large, depending on the number of SNPs used to mark that area, which of course will depend on the HapMap. The area identified may thus code for several or dozens of gene products. These sequences will be in the data base and will be known to code for, in some instances, known functional proteins, e.g. enzymes, receptors etc. Alternatively, while the sequence of such proteins may be known, their current function (if any) in human biology may as yet be unknown. This area of research holds out a great future for biochemists and molecular biologists. The identification of which one of these very specific proteins with known or unidentified function is involved in a particular disease will be the future. Herein lies one great strength of GWA. It does not make any *a priori* assumptions as to function. It is really quite at odds with the conventional scientific methologies that have served science so well over the years, i.e. hypothesis-driven research. The basis of virtually all experimental biologies has been to put forward a hypothesis based on known existing experimental and sometimes observational research. The next stage is to design experiments or to test this hypothesis. The design of these experiments is such that if the hypothesis is the correct explanation of what is happening at a biological level, then the ensuing results will come out in a way that can be predicted in advance. If the experimental data are robust enough and have sufficient statistical power, a negative outcome means that the hypothesis as stated is incorrect and must be either modified and retained or abandoned. The collection of data that is merely a marker or just associated with a disease has proved, with some notable exceptions, not to lead to any systematic scientific progress. This non-hypothesis approach is frequently termed in a derogatory way by experimental scientists 'a fishing trip'. While GWA is not hypothesis based, it is clearly a different dimension from simply data accumulation.

The current situation in relation to GWA has been well summarised[25], with the conclusion that to date approximately 100 loci for approximately forty common diseases and traits have been identified. These studies using the HapMap as a guide to select marker SNPs and GWA using large patient and control populations are destined to elucidate the complex inter-relationship between genetic variation and disease risk, including how disease risk is modified by environmental factors such as nutrition.

Table 3. Factors to consider when performing analyses using genome-wide analysis approaches to the study of genetic variations associated with disease risk

High number of questions asked
Statistically some results will be positive by chance
Bonforroni correction
Divide *P* value by number of SNPs examined
What SNPs will remain significant
Large numbers needed (power)
Analyse a primary and then confirmatory cohort
Depends on other replication cohorts

SNPs, single-nucleotide polymorphisms.

The future

The technology to resolve these issues is thus available. While this technology is moderately expensive, the number of GWA that are appearing month by month in the literature is a clear indication that funding agencies consider such studies are essential. However, for each such analysis to be worthwhile it must have a sufficient numbers of cases and controls to be able to withstand the very significant statistical (Bonferrori) correction required in such multiple testing (Table 3). This aspect can only be dealt with in the first instance by large numbers, certainly many hundreds, or ideally many thousands, of cases and controls. The issue of replication studies to confirm or exclude an initially positive result is also vital. First, in such replication studies the number of associations it will look for will now be a directed approach and will be small in number, thus involving only a small statistical correction. In addition, such subsequent analyses, if they adhere sufficiently to common patient and control criteria and have similar or at least overlapping protocols, can be combined in a meta-analysis. Such replication studies and meta-analyses will then transform what is a probable genetic variation for a risk into a reality. The area of the gene involved can then be investigated, giving a new insight not only into disease risk but the basic biology and aetiology of disease.

Conclusions

Historical focus has been on very-low-prevalence high-penetrance genetic variants such as cystic fibrosis. This approach gave way to more common variants that had variable extents of penetrance, which could lead to clinical signs or symptoms in some but by no means all subjects homozygous for the variant; an example being haemachromatosis. In both instances the association between the disease and the presence of a causative genetic variant was

sufficiently often present to allow its identification using family studies. However, it was always recognised that a substantial amount of risk for most common diseases, e.g. heart disease, colon cancer etc., was driven by genetic variations. To explain the risk of such diseases associated with such variants, it was clear that they would have to be common, with a homozygous prevalence of $\geq 10\%$. As indicated in the present review, although variants that influence disease risk with a lower prevalence than 10% almost certainly do exist, they would have insufficient impact on disease risk to be detectable. There the matter rested for many decades until two innovations took place. First, the human genome was sequenced, making possible the identification of genetic variation in large disease case–control cohorts and in the general population. Second, the technology for detecting SNPs in large populations became feasible and affordable. This advance led initially to larger studies excluding and including individual SNPs, which in turn gave way to marking a whole candidate gene with multiple SNPs. The ultimate expression of this development has been GWA. This approach marks the gene with between $0{\cdot}75 \times 10^6$ and 1×10^6 SNPs. It seeks to demonstrate an association between individual markers and their involvement in disease, compared with control populations. Variants can either increase or decrease case risk. The testing of such a large numbers of SNPs inevitably gives rise to 'false positives' that are the result of chance, which requires an extensive statistical correction for the numbers of SNPs tested. Some variants will survive this correction but might still be the result of chance. Deciding real from chance association requires replication of this change in further case–control populations. In these studies only a relatively small number of apparently-positive SNPs are retested, requiring a smaller statistical correction for a chance finding.

References

1. Klug WS, Cumming MR & Spencer AA (2006) *Concepts of Genetics*, 8th ed. Upper Saddle River, NJ: Pearson Prentice Hall.
2. Davey-Smith G, Ebrahim S, Lewis S *et al.* (2005) Genetic epidemiology and public health: hope, hype and future prospects. *Lancet* **366**, 1484–1498.
3. Boyle MP (2007) Adult cystic fibrosis. *JAMA* **298**, 1787–1793.
4. Welsh MJ, Tsui LC, Boat TF *et al.* (1995) Cystic fibrosis. In *The Metabolic and Molecular Basis of Inherited Disease*, 7th ed., pp. 3799–3876 [LR Scriver, AL Beaudet, WS Sly and Valle, editors]. New York: McGraw Hill.
5. Cystic Fibrosis Genetic Analysis Consortium (1994) Population variation of common cystic fibrosis mutations. *Hum Mutat* **4**, 167–171.
6. Easton DF, Deffenbaugh AM, Pruss D *et al.* (2007) A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am J Hum Genet* **81**, 873–883.
7. Ford D, Easton DF, Bishop DT *et al.* (1994) Risks of cancer in BRCA1 mutation carriers. *Lancet* **343**, 692–695.
8. Theil EC (2004) Iron, ferritin, and nutrition. *Annu Rev Nutr* **24**, 327–343.
9. Pietrangelo A (2006) Hereditary hemochromatosis. *Annu Rev Nutr* **26**, 251–270.
10. Nemeth E & Ganz T (2006) Regulation of iron metabolism by hepcidin. *Annu Rev Nutr* **26**, 323–342.
11. Pietrangelo A (2004) Heriditary hemochromatosis – a new look at an old disease. *N Engl J Med* **350**, 2383–2389.
12. Frayling TM (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet* **8**, 657–662.
13. Hunter DJ (2005) Gene-environment interactions in human diseases. *Nat Rev Genet* **6**, 287–298.
14. Kirke PN, Mills JL, Molloy AM *et al.* (2004) Impact of the MTHFR 677T polymorphism on risk of neural tube defects: case-control study. *Br Med J* **328**, 1535–1536.
15. Shields DC, Kirke PN, Mills JL *et al.* (1999) Thermolabile variant of methylenetetrahydrofolate reductase and neural tube defects: an evaluation of genetic risk and the relative importance of the genotypes of the embryo and the mother. *Am J Hum Genet* **64**, 1045–1055.
16. Daly LE, Kirke PM, Molloy A *et al.* (1995) Folate levels and neural tube defects: implications for prevention. *JAMA* **274**, 1698–1702.
17. Brody LC, Conley M, Cox C *et al.* (2002) A polymorphism, R653Q, in the trifunctional enzyme methylenetetrahydrofolate dehydrogenase/methenyltetrahydrofolate cyclohydrolase/formyltetrahydrofolatesynthase is a maternal genetic risk factor for neural tube defects: report of the Birth Defects Research Group. *Am J Hum Genet* **71**, 1207–1215.
18. Parle-McDermott A, Kirke PN, Mills JL *et al.* (2006) Confirmation of the R653Q polymorphism of the trifunctional C1-synthase enzyme as a maternal risk for neural tube defects in the Irish population. *Eur J Hum Genet* **14**, 768–772.
19. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.
20. Ziegler A, Thompson JR, Schunkert H *et al.* (2007) Genomewide association analysis of coronary artery disease. *N Engl J Med* **357**, 443–453.
21. Harrison PR (1990) Molecular mechanisms involved in the regulation of gene expression during cell differentiation and development. *Immunol Ser* **49**, 411–465.
22. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437**, 1299–1320.
23. Ulrich CM, Bigler J, Bostick R *et al.* (2002) Thymidylate synthase promoter polymorphism, interaction with folate intake, and risk of colorectal adenomas. *Cancer Res* **62**, 3361–3364.
24. Hirschlorn JN & Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95–97.
25. Manolio TA, Brooks LD & Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* **118**, 1590–1605.