MEETING REPORT   **Open Access**

'Sifting the significance from the data' - the impact of high-throughput genomic technologies on human genetics and health care

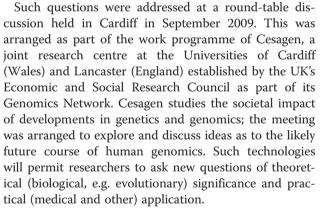Angus J Clarke[1*], David N Cooper[1], Michael Krawczak[2], Chris Tyler-Smith[3], Helen M Wallace[4], Andrew O M Wilkie[5], Frances Lucy Raymond[6], Ruth Chadwick[7], Nick Craddock[8], Ros John[9], John Gallacher[10] and Mathias Chiano[11]

Abstract

This report is of a round-table discussion held in Cardiff in September 2009 for Cesagen, a research centre within the Genomics Network of the UK's Economic and Social Research Council. The meeting was arranged to explore ideas as to the likely future course of human genomics. The achievements of genomics research were reviewed, and the likely constraints on the pace of future progress were explored. New knowledge is transforming biology and our understanding of evolution and human disease. The difficulties we face now concern the interpretation rather than the generation of new sequence data. Our understanding of gene-environment interaction is held back by our current primitive tools for measuring environmental factors, and in addition, there may be fundamental constraints on what can be known about these complex interactions.

Introduction

The advent of 'high throughput' or 'next generation' genomic sequencing technologies has raised expectations of what laboratory genetics has, and will have, to offer to both the clinician and the patient. Our greatly increased ability to generate nucleic acid sequence data raises the question of how this rapidly accumulating mass of new genomic information will be interpreted and when this will become feasible in general clinical practise. How much sense will have been made of the new data within the next 10 to 15 years? What new questions will it be possible for us to pose once these new technologies are readily available to provide the information upon which biologically meaningful answers to these questions can be based? The extended lag time between generating a basic understanding of the pathogenesis of many single gene (Mendelian) disorders and devising effective remedies has long been acknowledged. In the context of complex disorders and the many known quantitative (non-disease) traits, will the lag time between data collection and its interpretation be shortened?

Such questions were addressed at a round-table discussion held in Cardiff in September 2009. This was arranged as part of the work programme of Cesagen, a joint research centre at the Universities of Cardiff (Wales) and Lancaster (England) established by the UK's Economic and Social Research Council as part of its Genomics Network. Cesagen studies the societal impact of developments in genetics and genomics; the meeting was arranged to explore and discuss ideas as to the likely future course of human genomics. Such technologies will permit researchers to ask new questions of theoretical (biological, e.g. evolutionary) significance and practical (medical and other) application.

The starting point for the discussion was the recent leap in DNA sequencing capability developed by several different commercial enterprises. There was no need to dwell upon the specifics of the technologies; rather, the point was to address the potential benefits for the sciences and clinical medicine of having large volumes of genomic sequence data available, effectively without major constraints of time or cost and alongside an emerging capacity to determine the CpG methylation status of the corresponding nucleotide sequences, the 'methylome', and other modifications of chromatin [1]. Several short presentations were made during the round-table

* Correspondence: clarkeaj@cardiff.ac.uk
[1]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, Wales CF14 4XN, UK
Full list of author information is available at the end of the article

sessions, but the emphasis of the meeting was on open discussion among the 20 participants, who have backgrounds in human and evolutionary genetics, clinical medicine, social science and philosophy. The perspectives developed at the meeting have been refined 'virtually' since then through multiple cycles of e-mail exchanges.

Experience of exomics

Useful insights had already been obtained through two large studies: the Genetics of Learning Disability (GOLD) study with complete X chromosomal 'exome' sequences from >200 patients with sex-linked cognitive impairment [2] and the description of the full exomes of 12 selected individuals [3]. Frances Lucy Raymond (Cambridge, England) was able to draw a number of general lessons from the GOLD study:

- Sample quality is critical for obtaining reliable DNA sequence data; poor sample quality generates multiple sequence variants per sample, making it difficult to distinguish real sequence from experimental artefact; the assumption that poor quality samples would not amplify was false.
- Truncating sequence variants - termed nulls - are found in approximately 1% of genes on the X chromosome but are nevertheless compatible with normal life in the hemizygous male.
- Missense variants are common, with an average of four unique single nucleotide non-synonymous variants per family with X-linked intellectual disability.
- Data access remains a sensitive issue; there is a need to publish variants and allele frequencies while preserving the anonymity of research participants. This would generally entail publishing aggregate data and new variants rather than individual-specific haplotype information.

These findings emphasise what is already well known, that a disrupted gene need not necessarily lead to a clinically overt disease state, i.e. many genes are dispensable, and even for inactivated essential genes, penetrance will often be incomplete [4,5]. The difficulties inherent in this analysis with regard to the interpretation of previously unreported variants - mutations that may be of pathogenic significance - will be even more substantial when we consider autosomal loci, of which (almost) everyone will have two copies [6]. Additional new approaches to determine the functional significance of genomic variation will be required if any coherent interpretation of the approaching deluge of data is to be feasible [7].

Epigenetics

Epigenetic influences together introduce an additional layer of complexity over and above the functionality of genomic DNA sequences [4]. The inter-relationship of DNA sequence and DNA sequence modification has not yet been adequately addressed in either theory or experiment; it should, however, be noted that this relationship is likely to be bi-directional with certain polymorphic variants influencing methylation status and the methylation status in turn influencing mutability. Ros John (Cardiff, Wales) introduced this topic and its relationship to known imprinting phenomena, referring to both epidemiological and experimental work, *viz.* records of the Dutch famine at the end of World War II [8] and contemporary work with the agouti mouse [9]. The term 'epigenetics' includes the study of both conventional gene-gene (GxG) interactions, long recognised in principle but in practise exceedingly difficult to measure in humans, and the still under-researched area of 'predictive adaptive responses'. The latter has been elaborated theoretically [10], triggering very interesting - but so far perhaps more tantalising than illuminating - laboratory work on the possible influence of early life experiences (before or soon after birth) on patterns of DNA methylation and subsequent disease susceptibility in the face of dietary challenges and other environmental circumstances.

Our understanding of the intricate role of the epigenome in directing differentiation, communicating cell fate and mediating adaptive response will be informed by comprehensive surveys of epigenetic marks across space (cell type/state) and time (development and ageing), and in response to environmental challenges (diet, stress). To this end, a number of ingenious genome-wide, high-throughput technologies have been developed. Some are based on the ability to selectively capture methylated fragments of the genome using methyl-binding proteins and then sequencing these fragments. MeDIP-seq, MBD-seq and MethylCap-seq have all been applied successfully, although with differences in selectivity for specific sequences [11,12]. A readout of the whole genome methylation status can also be obtained by applying a bisulfite-sequencing technique, whereby genomic DNA is exposed to HSO_3^-, which deaminates only the unmethylated (and, hence, unprotected) cytosine residues resulting in a change in the DNA sequence. This technique may provide greater coverage of the genome [13]. The genome-wide assessment of histone modifications is a more complex task since these modifications are numerous, and activating and silencing marks are not necessarily mutually exclusive [14]. Chromatin immunoprecipitation followed by high-throughput sequencing is the preferred technique for determining the genome-wide locations of specific histone modifications [15,16]. Although there are still issues, in particular, the biased representation of certain sequences, computational analyses and further technical refinements will undoubtedly increase the sensitivity of these techniques and further reduce the costs.

One potentially very important aspect of GxG and gene-environment (GxE) interactions that may limit their importance, or at least our ability to detect them, is the small effective population size (N_e in population genetic notation) of humans [17]. Another important question is the extent to which the assessment of CpG methylation can serve as an adequate biomarker for the full range of epigenetic influences. A systems biology approach to the recognition of biological networks may be one way of addressing GxG and GxE interactions in humans [18,19], given that the types of breeding experiments pioneered in *Drosophila* species are clearly ruled out in humans, not only on ethical grounds but also by the restrictions of time-scale available to researchers. The collection of biological samples and data on the large Helmholtz cohort in Germany is intended to address some of these issues.

Behaviour and psychiatry

When one turns to consider human behaviour and psychiatric disorders, rather than the (relatively) simple phenotypes associated with human growth or with metabolic and developmental disorders, additional layers of complexity become evident in the GxE interactions. As Anita Thapar (Cardiff, Wales) made clear, there is a pressing need for good measures of both the environmental factors and the behavioural phenotypes of potential interest. The GxE interactions can be looked at as purely statistical phenomena or, more interestingly and perhaps more importantly, as phenomena of potential biological meaning. Here more than anywhere, and in deference to the great David Hume, it is essential to avoid the all-too-easy slide from the observation of an association into unwarranted assumptions about causation. Thus, if one observed an association between attention deficit hyperactivity disorder in children and maternal cigarette smoking, what causal influences might have been at work, - e.g. an environmental effect of smoking in pregnancy or a common genetic basis to both behaviours? One opportunity to clarify such questions, although difficult to establish, would be longitudinal studies of children born following conception by IVF. Some of these questions are profound, touching our very humanity; for instance, how (through what process) does the early maltreatment of children result in later antisocial behaviour? How may social interventions be designed effectively to interrupt such pathogenetic linkages, for the long-term benefit of society as a whole? The Bradford-Hill criteria [20] address this problem of causation but have yet to be fully tested in the genomic era.

Improved methods of measuring environmental factors and their behavioural correlates are clearly required for the full benefits of the progress in genome sequencing to be attained; these measurements should be made without lapsing into implicitly deterministic assumptions.

Common disease genetics

Michael Krawczak (Kiel, Germany) challenged the assumptions underlying much 'common disease genetics' - especially the limitations of its most influential models. He expanded upon three critical underlying assumptions made by Reich and Lander [21,22]:

- The allelic spectrum of common disease genes is simple.
- There was a single, sudden expansion event of the human population (at least, the out-of-Africa population) from 10,000 members to its modern size.
- The part played by selection in the spread of alleles predisposing to common diseases was relatively small.

This model has led to the neglect of new mutants arising from existing variants during the turnover of the allelic spectrum of common disease. Krawczak explained why these assumptions are highly implausible and what effect this is likely to have had on the current allelic spectrum of genetic predisposition to complex disease. In particular, as is apparent both when other modelling approaches, such as coalescence theory, are used and from the data currently accumulating:

- The allelic spectrum of complex disorders, like that of Mendelian disorders, is likely to be diverse.
- There may be only a weak correlation between the risk of disease and the population frequency of risk alleles.
- The functional effects of predisposing alleles may be weakly or even inversely related to the associated disease risk (some predisposing alleles may even, counter-intuitively and due to stochastic processes, be more frequent in unaffected than affected individuals).
- The mutational load in unaffected individuals may be particularly high for common diseases.

Mapping the 'polygenes'

Krawczak also discussed the limitations of the available strategies for mapping the loci which contribute to the common, complex disorders. The power of the affected sib pair (ASP) linkage approach has been regarded as low [21] in comparison to association studies employing the transmission disequilibrium test (TDT), not a case–control design, but that view also had shortcomings:

- The paper assumed knowledge of the risk allele when assessing the TDT but not when assessing the ASP method.
- The relative risks invoked were unrealistically high.
- The association approach was found to be advantageous for studies of candidate genes (or

regions) but not necessarily for genome-wide methods exploiting linkage disequilibrium, such as genome-wide association studies (GWAS).

There are good grounds for using both family studies and GWAS to utilise all the available evidence [23]. Family-based approaches that could be used with whole-genome sequencing (potential scenarios) include:

- sequencing the index patients and evaluating the inheritance pattern of 'interesting' genes or gene regions in their families
- sequencing the parents of the patient, comparing the results to population controls, and using TDT on 'promising' alleles
- assessing the co-inheritance pattern of unlinked genes in families to derive candidate regions for GxG interactions
- defining and assessing pre-disease or sub-clinical disease states in genetically-defined high-risk individuals identified in families.

Andrew O M Wilkie (Oxford, England) mentioned that the supposed 'triumphs' of the common disease/common variants (CD/CV) model conveniently ignore the fact that the relative risks conferred by susceptibility alleles cluster much more closely around 1.0 than anyone envisaged two decades ago; relative risks above 2.0 are decidedly unusual. Hence, the majority of heritability remains unexplained, and useful personalised prediction in healthy individuals (largely based on family history) has hardly improved over this time.

Additional challenges to our understanding include the greater heritability of diseases (e.g. schizophrenia, as presented by Nick Craddock of Cardiff, Wales) than would be expected given the CD/CV model of disease causation and the greater difficulty of interpreting any estimates of heritability when the disease phenotype results from the joint action of several predisposing genes (as Helen M Wallace (Buxton, UK) reminded us). Whether these insights into pathogenesis will eventually convert to promised improvements in human health remains to be seen.

There was discussion about the possible 'replication' of the results of GWAS, led by Mathias Chiano (Stevenage, England). Even the question of what exactly 'replication' would entail needs to be clarified, as when specifying the population(s) from which study participants are drawn. If an attempt is made to replicate a given study in a different population, the meaning of a negative result will be unclear because the relevant genetic factors may differ between this and the original population; there are no easy answers in relation to population differences.

Selection

Chris Tyler-Smith (Hinxton, England) continued with the theme of human history and the contemporary traces of past selection. Selection can take many forms but the presence of high levels of polymorphism suggests either an ancient origin for neutral mutations or the maintenance of polymorphism by fluctuating or balancing selection. Recognising past selection is possible by (i) counting offspring, (ii) looking for patterns of DNA variation in populations that depart from neutrality (such as long haplotypes, skewed allele frequency spectra, large differences between populations), and (iii) functional studies. Recent genomic studies comparing populations have drawn upon the second category of evidence that detects the net effects of positive selection over millennia, but these have mostly been based on simple models of selection and are less able to detect phenomena such as heterozygote advantage, frequency-dependent selection, disruptive selection - selection for both homozygotes - or antagonistic but stabilising selection with contrary effects at different stages of the life cycle in the two sexes or in different environments.

While there are grounds for thinking that selective 'sweeps' have been uncommon in recent human evolution [24], the role of selection in maintaining substantial levels of polymorphism remains less clear. Although a biological (natural selection-based) account for many genetic observations consistent with selection remains elusive, this could be the result of the challenging nature of such studies. The persistence of polymorphism without a clear genomic signature may result from changes in the strength and direction of selection. Despite these complexities, there are widely accepted examples of positive selection, including disease-associated variants and those influencing visible traits such as skin colour or hair, where sexual selection through gene-culture interaction may have been relevant [25,26].

Human genetic variation: the 1000 Genomes Project

Results from the 1000 Genomes Project [27] and other studies indicate that, despite the low level of genetic variation in humans compared with other apes, there are still enormous inter-individual differences in genome sequence, which can now be identified in a comprehensive way at the level of CNVs, indels and SNPs. Notably, there are on average 250 to 300 loss-of-function variants such as 'nonsense' or splicing SNPs in the genome of the average 'healthy' person, 50 to 100 previously implicated, disease-associated variants and approximately 80 genes that vary in copy number. As suggested by the GOLD study (*op cit*), this greatly complicates the clinical interpretation of an individual's genome sequence.

The first goal of the 1000 Genomes Project has been to capture important data about background human genetic variation which will be essential in studies of both inherited and acquired disease. The samples being studied:

- were initially drawn from urban populations in four countries (HapMap Nigerian, Japanese, Chinese, and families with European ancestry from the USA) but additional samples will be drawn from elsewhere, including South Asia
- were from anonymous individuals with no phenotypic information provided (except sex, population of origin and that they are adults competent to provide consent), but do have consent for free web release of the sequence data, and are available as cell-lines to researchers
- because full × 30-depth sequencing to detect heterozygous variants in a single individual effectively would be too costly, low coverage sequencing of samples from many individuals is being used to capture the variants shared between individuals; the full project aims to detect essentially all polymorphic variants (i.e. those at >1% frequency) in the geographical areas investigated that lie in the accessible part of the genome.

1000 Genomes pilot studies

- Samples (179) at depth × 2-4 should identify most variants present at >5% - so far, 17.2 million SNPs have been found, of which >50% were novel.
- Two trios sequenced at × 30-60 revealed many 'new' (*de novo*) mutation events (see below)
- Approximately 1,000 genes were sequenced deeply in samples from 700 volunteers.

The next task: it will be necessary to draw upon large population studies to seek evidence of variable mutation penetrance.

Mosaicism and foetal DNA in the maternal circulation

High-throughput sequencing is also opening up the detection of mosaicism, which is likely to have many biological consequences in addition to neoplasia [28]. With a read depth of × 30, it is difficult to detect mosaicism at or below the level of 5% because of the 1% error in typing each nucleotide position. This could have applications in the study of cancers, e.g. the evolution of individual tumours as reflected in tumour-specific DNA [29], and in prenatal diagnosis, where free foetal DNA is currently used to look for alleles not present in the mother, such as Rhesus or Y chromosome sequences, but is likely soon to be applicable to population screening, especially for trisomy 21 [30-

32]. Greater read depths will be required for clinically important applications reliant upon the detection of low levels of mosaicism [33].

Phenotypes and taxonomy

Craddock led a discussion on disease phenotypes and their relation to disease taxonomy, especially in relation to psychiatric disorders. The definition of a phenotype is crucial as it may be influenced by numerous factors including 'culture' and tradition, established or desired patterns of health service utilisation, interacting environmental variables and previously accepted models of disease causation.

There has been a growing awareness that the specific single genes in which mutations cause a few rare disorders may also be (more loosely) associated with disease predispositions apparent in GWAS. Thus, rare mutations in the *CACNA1C* gene can result in a multisystem disorder, manifesting as cardiac dysrhythmias, epilepsy, autism, cognitive impairment and abnormal physical features, whereas variation in the same gene is more weakly associated with bipolar disease, schizophrenia and unipolar depression, as evidenced by GWAS. The benefits of identifying specific genes in which variation contributes to disease susceptibility include understanding the pathogenesis, establishing a diagnosis, devising new therapies and selecting the most appropriate treatment for the individual patient.

Following the course of human genetics research over the past 15 years, one may perceive a shift in focus from rare to common diseases and from genes of large effect in which mutations cause disease to genes of small effect in which variants merely modify the risk of disease. Mutations in genes of large effect are usually accepted unequivocally as important causal factors; the modifying genes of lesser effect can, however, also provide insights into pathogenesis when several loci in the same developmental or functional pathway interact to contribute to a specific disease. It is also clear that defining the phenotype is, itself, an iterative process, taking the researcher back and forth between the phenotype and the genotype. The researcher can use different values of a given parameter in order to establish which yields the clearest discrimination between individuals with and without disease.

The finding that sequence variation at some of the same loci influence both schizophrenia and bipolar disease has cast doubt on the validity of the Kraepelinian dichotomy of the psychoses into these categories [34,35]. The exclusion of 'intermediate' categories of patients from much previous research had in effect prevented this dogma from being challenged in the past - it became a closed 'view of the world' (a prophecy that was both dogmatic and self-fulfilling).

More recently still, it has become clear that new mutations at many loci contribute very substantially to schizophrenia, confirming the indications from studies

of *de novo* CNVs that *de novo* rare mutations contribute importantly to this relatively common disease [36,37] and demonstrating that the reservations of Krawczak and others about the CD/CV model were warranted.

Mutation and selection at single loci

There is still much to be learned from the detailed study of single loci, in addition to genome-wide research that does not assume the applicability of the CD/CV model of complex disorders. As an example, Wilkie presented information from his long term studies of the fibroblast growth factor receptor (*FGFR*) genes.

As background to the *FGFR* studies, germline mutation rates have been estimated to be about 2.5×10^{-8} [38] or 1.8×10^{-8} [39] - equivalent to approximately 120 nucleotide substitutions per birth. The fitness of amino acid substitutions has been estimated [40,41], and purifying selection is thought to act on 2.5% to 5% of the genome (of which 1% will be coding sequence). This will lead to 1 to 3 harmful amino acid substitutions per birth, plus a similar number from each previous generation. Recent data from the 1000 Genomes Project suggest a slightly lower rate of new mutations, closer to 10^{-8} per base pair per generation, and that the influence of purifying selection can be detected through a reduction in genetic variation at distances of up to 100 kb from genes, implying that few parts of the genome are entirely unaffected by such selection. Similarly low rates of mutation have been found by others [42].

Wilkie described how he set out to explain the apparently very high levels of mutation in specific genes using the *FGFR* genes as a model. The most frequent transition mutation in the human germline is the *FGFR3* mutation that causes achondroplasia, but Wilkie chose to study the Apert syndrome mutation 755 C > G in *FGFR2*, which is the most common transversion mutation in humans. The mutations originate exclusively from the healthy fathers of the affected individuals, who tend to be older than average for the population (paternal age effect). From the analysis of normal sperm and testes, it can be deduced that the positions that mutate in *FGFR* genes are not true mutation hotspots; rather, there is positive selection for spermatogonia carrying the mutation because the mutation confers gain-of-function properties upon the encoded protein [43-45]. Wilkie looked for these mutations in an uncommon testicular tumour - spermatocytic seminomas - and a comparable spectrum of mutations was found here as in congenital disorders and in bladder tumours: they all act on the Ras/MAPK pathway, which plays a central role in regulating proliferation and other critical cellular processes. The testis may be viewed as a 'bioreactor' for selfish mutations that promote the clonal growth of spermatogonia harbouring one of these growth-promoting mutations [46]. The effects of each mutation vary with the degree to which it is activating:

1) highly activating mutations: lethal disorders or testicular tumours
2) moderately activating mutations: congenital malformation syndromes exhibiting a paternal age effect
3) weakly activating mutations: rare sequence variants that may predispose to a wide spectrum of disorders, involving for example neurodevelopment and cancer predisposition

It should be noted that the latter category of mutations will be undetectable by GWAS, and so may contribute to missing heritability [47].

The paternal age effect for schizophrenia may occur through a similar mechanism. It should be remembered that more than 1% individuals with schizophrenia have a *de novo* CNV, often a rare but recurrent CNV [48], while CNVs have been shown to be much more common in cases of schizophrenia than in controls (5% of controls, 15% to 20% of cases) [49] and, interestingly, to overlap with the CNVs found in attention deficit hyperactivity disorder [50]. However, the frequency of *de novo* CNVs is high enough in controls that such *de novo* events cannot be taken as *ipse facto* proof of pathogenicity.

Pharmacogenetics

Chiano led a discussion on pharmacogenetics and medical practice. Genetics is just one of the factors that influence the safety and efficacy of particular drugs. These factors include dosage, 'the environment', compliance, other drugs, diet, age and the co-occurrence of other diseases. Efficacy can vary from 80% for COX-2 inhibitors to 25% for cancer chemotherapy. The abacavir-associated hypersensitivity reaction can occur in 5% to 8%; there is a major susceptibility locus in the HLA region [51,52]. Proponents have been disappointed at the slow pace of introduction of pharmacogenetic testing into clinical practice. Meanwhile, somatic genetic tests are proving to be helpful in guiding the treatment of some malignancies but have been slower to enter regular clinical practice for a wider set of indications despite the longstanding recognition of the relevant pharmacogenetic phenomena: additional criteria of utility and cost-effectiveness also need to be fulfilled [53].

Privacy and consent

Ruth Chadwick (Cardiff, Wales) led a discussion on issues of privacy and consent in genomic research. It is becoming impossible for medical researchers to guarantee privacy to the research participants they recruit - especially with the pressure from funding agencies who insist upon

open-access archiving of genomic sequence data, as these data inevitably contain potentially identifying information. Indeed, it would now be misleading to promise privacy of personal genome information to research participants in exchange for consent to donate samples.

There are difficulties at the level of the technology (sample collection and storage, and data generation, storage and access) and the interpretation of information collected. It has been argued that the very concept of privacy will have to be renegotiated in the context of the rationale(s) for data sharing, especially at international level. Indeed, there has been a move from 'traditional' notions of informed consent to one of a broad consent and then to open consent. One must either strengthen the traditional practises of consent - ensuring that participants understand the difficulties of ensuring privacy before giving consent (while at the same time pursuing the data protection strategies that are possible) - or rethink the whole notion of privacy [54]. In clinical practice, there may be recurrent difficulties when molecular studies of a malignancy, designed to guide therapeutics, thereby, also yield unwelcome information about prognosis or implications for close relatives [55]. In the context of research, however, new forms of consent (e.g. broad consent taken online) may well be readily accepted [56]. One approach is for the researcher to move from guaranteeing (and so protecting) privacy to practising veracity, explaining to potential participants that their data will be accessed by, and shared with, others; those consenting to research have to be open to this [57].

Those participating in 'personal genomics' would be deluding themselves if they thought that privacy was still possible once their genome sequences became accessible online [58,59]). This is the age of bioinformatics, surveillance, Facebook and Twitter. There is inevitably a risk of personal consequences for research participants when their data are released. These will take different forms under different circumstances. Access to health care may become more difficult in countries without state health care or universal and compulsory health insurance, as when health care is made available through commercial insurance schemes. In addition, if information about other family members becomes available, then biological family relationships might be shown to differ from the pattern of social relationships presented in public (e.g. paternity may have been misattributed).

One perspective reveals a clash of research cultures: molecular scientists (who seek instant open access to all data to maximise the scientific - and perhaps commercial - exploitation of the data) and the clinicians (who are often more insistent upon protecting their patients and more modest in their assessment of the likely short-term benefits to the participants, and other patients, of the research itself).

One key notion is that of the acceptable uses for research data. Participants tend to trust 'the system' to use their data in a 'good' way. The system of research governance is designed (in part) to ensure that such trust is warranted - although the system of governance also serves other, more institutional purposes. The prospect of internet-based marketing corporations using access to research data and to electronic health records as an opportunity to market more products seems both manipulative and cynical.

The concept of 'consent' is also problematic in ways unrelated to the concept of privacy. As Raymond explained, families participating in research studies could allow their genuine altruism to lead them to sign the research consent forms because their thoroughly worthy impulses blinded them to the potential for problems to arise from participation. Such problems might include results of uncertain significance, which may remain difficult to interpret for many years or whose significance might only be clarified through the performance of family studies that could be confusing or distressing; even with such family studies, the significance of the results may remain opaque. If the reasons underlying the importance and relevance of consent to research were explained, then many families might decline participation.

One response to these possibilities has been the suggestion that a solution should be found through information technology itself, with the development of a system of 'data enclaves', within which researchers can utilise data in a predefined analysis but from which they cannot export any data about individuals - only aggregate data and the derived results of analysis.

Limited clinical utility of genetic association

Wallace set out her answer to the important question, 'What are the (biological) limits to the predictive value and clinical utility of the 'predict and prevent' strategy?' An early approach to this led to the conclusion that environmental interventions targeted by genotype might significantly reduce the incidence of some diseases [60], dependent on the magnitude of gene-environment interactions (in the statistical sense of their influence on risk at a population level). However, there was an error in the measure of clinical utility used in this paper, and in addition, it took no account of the proportion of the population in the high-risk group (thus, erroneously suggesting that a genetic test which identified the whole population as at high risk would have high clinical utility). Further, the aetiology of a given disease will place constraints on the potential magnitude of both the genetic component of a disease and of any gene-environment interaction. Wallace set out her approach to assessing clinical utility in the context of genetic testing [61].

Utility, U_{GE} = population impact – gamma, where gamma is the fraction of the population classified as at high genotypic risk, and the population impact is the avoidable proportion of disease occurring in this fraction of the population, compared to the population as a whole.

Utility is larger if gamma is small, and targeting is only effective in the presence of GxE interactions. If there is no GxE interaction, genotyping performs no better than random selection in terms of reducing the incidence of the disease in a population. If high-risk individuals have less to gain from intervention than the low-risk group, then targeting has negative utility. There is a range of solutions for any given disease, which can be mapped out if data on relative risks from both twin and family studies, plus environmental data, is known. Each solution depends on the model that is assumed for gene-gene and gene-environment interactions. Because of the number of alternative models, the sample sizes required to evaluate or validate risk predictions may be too high, and it could be impossibly difficult to generate good association (GWAS) data and integrate it with environmental information [62]. The model also confirms earlier findings that heritability estimates can be high even in the complete absence of any genetic component to a disease [63] and that gene-gene and gene-environment interactions, as well as the equal environments assumption, may account for some (or even all) of the 'missing heritability' of complex diseases.

It should be noted that recent modelling of heritability and GWAS data suggest that a much greater fraction of the heritability of quantitative traits can be accounted for by GWAS data than had been found previously, once allowance is made for the incomplete linkage disequilibrium between the genotyped SNPs and the 'true' causal factors [64-66]. However, such modelling is still unable to identify the factors involved and cannot deliver improved capacity to predict disease. Furthermore, only a small proportion of the calculated heritability has yet been identified for any common disease [67].

Conclusion

The space provided by this workshop and the subsequent e-mail exchanges has served to clarify our thoughts and allowed several modest conclusions to emerge, although unanimity was not achieved on all issues. Four points, in particular, became progressively clearer during these discussions.

First, the new insights derived from the application of high throughput sequencing are likely to transform our understanding of human biology, especially in the context of human genetic disease and evolution. Clinical applications of human genome sequencing are emerging in oncology and reproductive genetics and are beginning to impact upon other clinical areas and disciplines. The old debates about nature vs. nurture are likely to re-emerge in new guises, and great care will be needed in case fruitless debate consumes too much energy or misconceived arguments lead to the inappropriate application of theory in policy and practise across various walks of life, most especially psychiatry, education and the law.

Second, and despite a full acceptance of this first point, major problems remain in moving from the rapidly accumulating raw sequence data and catalogues of genetic variation to warranted conclusions in their interpretation. These problems include questions relating to the clinical utility of conclusions drawn on the basis of genetic association studies that lack a plausible pathogenic mechanism underlying the observed association. There will inevitably be uncertainties in the face of a morass of interacting variables, but some of the difficulties should be resolved over time - although perhaps on a timescale of decades rather than months. In addition, our current measures of environmental factors are primitive, and our ability to record the variation in such factors over a lifetime of many years is still crude.

Finally, and despite the continuing advances in information technologies, the accumulation of both genomic and environmental data may outstrip the capacity of our information systems to store and analyse the data [68]. These limits to computability may not be merely practical in nature - to be overcome once technology has moved on - but may represent constraints as to what can be known or calculated even in principle, given limits to global population size, the nonrandom distribution of genotypes and the exponentially increasing number of statistical comparisons that can be made; how can one correct adequately for such astronomical numbers of comparisons? Deductive, hypothesis-driven research on the basis of such data will be challenging!

Acquiring the necessary genome sequence data appears now to be almost the easy step by comparison with the quantification of environmental factors and of human phenotypes, the physical limits to data storage and constraints in principle to computability and interpretation.

Hedgecoe (Cesagen, Cardiff, Wales), Ruth McNally (Cesagen, Lancaster, England), Anita Thapar (Psychological Medicine, Cardiff University, Wales).

Author details
[1]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, Wales CF14 4XN, UK. [2]Institute of Medical Informatics and Statistics, Christian-Albrechts University Kiel, Brunswiker Straße 10, Kiel 24105, Germany. [3]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [4]GeneWatch UK, 60 Lightwood Rd, Buxton SK17 7BB, UK. [5]Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Headington, Oxford OX3 9DS, UK. [6]Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK. [7]Cesagen, 6 Museum Place, Cardiff University, Cardiff, Wales CF10 3BG, UK. [8]Department of Psychological Medicine, Cardiff University School of Medicine, Heath Park, Cardiff CF14 4XN, UK. [9]Cardiff School of Biosciences, Cardiff University, Cardiff, Wales CF10 3AX, UK. [10]Clinical Epidemiology, School of Medicine, Cardiff University, Cardiff, Wales CF14 4XN, UK. [11]GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, UK.

References

1. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473**:43–49.
2. Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, O'Meara S, Latimer C, Dicks E, Menzies A, Stephens P, Blow M, Greenman C, Xue Y, Tyler-Smith C, Thompson D, Gray K, Andrews J, Barthorpe S, Buck G, Cole J, Dunmore R, Jones D, Maddison M, Mironenko T, Turner R, Turrell K, Varian J, West S, Widaa S, *et al*: **A systematic large-scale resequencing screen of X-chromosome coding exons in mental retardation.** *Nat Genet* 2009, **41**:535–543.
3. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature* 2009, **461**:272–276.
4. Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, Chuzhanova N, Krawczak M, Kehrer-Sawatzki H, Stenson PD: **Genes mutations, and human inherited disease at the dawn of the age of personalized genomics.** *Hum Mutat* 2010, **31**:631–655.
5. MacArthur DG, Tyler-Smith C: **Loss-of-function variants in the genomes of healthy humans.** *Hum Mol Genet* 2010, **19**:R125–R130.
6. Raymond FL, Whibley A, Stratton MR, Gecz J: **Lessons learnt from large-scale exon re-sequencing of the X chromosome.** *Hum Mol Genet* 2009, **18**:R60–R64.
7. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, James M, Liu P, Tichelaar JW, Vikis HG, You M, Mills IG: **Principles for the post-GWAS functional characterization of cancer risk loci.** *Nat Genet* 2011, **43**:513–518.
8. Schulz LC: **The Dutch Hunger Winter and the developmental origins of health and disease.** *Proc Natl Acad of Sci USA* 2010, **107**:16757–16758.
9. Morgan HD, Sutherland HG, Martin DI, Whitelaw E: **Epigenetic inheritance at the agouti locus in the mouse.** *Nat Genet* 1999, **23**:314–318.
10. Gluckman P, Hanson M: *The Fetal Matrix. Evolution, Development and Disease.* Cambridge: Cambridge University Press; 2005.
11. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavaré S, Beck S: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.** *Nat Biotechnol* 2008, **26**:779–785.
12. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, Olshen A, Ballinger T, Zhou X, Forsberg KJ, Gu J, Echipare L, O'Geen H, Lister R, Pelizzola M, Xi Y, Epstein CB, Bernstein BE, Hawkins RD, Ren B, Chung WY, Gu H, Bock C, Gnirke A, Zhang MQ, Haussler D, *et al*: **Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications.** *Nat Biotechnol* 2010, **28**:1097–1105.
13. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, Jäger N, Gnirke A, Stunnenberg HG, Meissner A: **Quantitative comparison of genome-wide DNA methylation mapping technologies.** *Nat Biotechnol* 2010, **28**:1106–1114.
14. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells.** *Cell* 2006, **125**:315–326.
15. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129**:823–837.
16. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448**:553–560.
17. Charlesworth B: **Effective population size and patterns of molecular evolution and variation.** *Nat Rev Genet* 2009, **10**:195–205.
18. Flint J, Mott R: **Applying mouse complex-trait resources to behavioural genetics.** *Nature* 2008, **456**:724–727.
19. Emily M, Mailund T, Hein J, Schauser L, Schierup MH: **Using biological networks to search for interacting loci in genome-wide association studies.** *Eur J Hum Genet* 2009, **17**:1231–1234.
20. Bradford-Hill A: **The environment and disease: association or causation?** *Proc R Soc Med* 1965, **58**:295–300.
21. Reich DE, Lander ES: **On the allelic spectrum of human disease.** *Trends in Genetics* 2001, **17**:502–510.
22. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273**:1516–1517.
23. Clerget-Darpoux F, Elston RC: **Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association.** *Hum Hered* 2007, **64**:91–96.
24. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, 1000 Genomes Project, Sella G, Przeworski M: **Classic selective sweeps were rare in recent human evolution.** *Science* 2011, **331**:920–924.
25. Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, Macarthur DG, Yngvadottir B, Nica AC, Woodwark C, Chen Y, Conrad DF, Ayub Q, Mehdi SQ, Li P, Tyler-Smith C: **Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation.** *Genetics* 2009, **183**:1065–1077.
26. Laland KN, Odling-Smee J, Myles S: **How culture shaped the human genome: bringing genetics and the human sciences together.** *Nat Rev Genet* 2010, **11**:137–148.
27. The 1000 Genomes Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061–1073.
28. De S: **Somatic mosaicism in healthy human tissues.** *Trends in Genetics* 2011, **27**:217–223.
29. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K, Turashvili G, Varhol R, Warren RL, Watson P, Zhao Y, Caldas C, Huntsman D, Hirst M, Marra MA, Aparicio S: **Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution.** *Nature* 2009, **461**:809–813.
30. Sehnert AJ, Rhees B, Comstock D, de Feo E, Heilek G, Burke J, Rava RP: **Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood.** *Clin Chem* 2011, **57**:1042–1049.
31. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B: **Detection and quantification of rare mutations with massively parallel sequencing.** *Proc Natl Acad Sci USA* 2011, **108**:9530–9535.
32. Chiu RWK, Akolekar R, Zheng YW, Leung TY, Sun H, Chan KC, Lun FM, Go AT, Lau ET, To WW, Leung WC, Tang RY, Au-Yeung SK, Lam H, Kung YY, Zhang X, van Vugt JM, Minekawa R, Tang MH, Wang J, Oudejans CB, Lau TK, Nicolaides KH, Lo YM: **Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study.** *British Medical Journal* 2011, **342**. doi:10.1136/bmj.c7401.
33. Lo YM, Chan KC, Sun H, Chen EZ, Jiang P, Lun FM, Zheng YW, Leung TY, Lau TK, Cantor CR, Chiu RW: **Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus.** *Science Translational Medicine* 2010, **2**:61–91.
34. Williams HJ, Craddock N, Russo G, Hamshere ML, Moskvina V, Dwyer S, Smith RL, Green E, Grozeva D, Holmans P, Owen MJ, O'Donovan MC: **Most genome-wide significant susceptibility loci for schizophrenia and bipolar disorder reported to date cross-traditional diagnostic boundaries.** *Hum Mol Genet* 2011, **20**:387–391.

35. Green EK, Grozeva D, Jones I, Jones L, Kirov G, Caesar S, Gordon-Smith K, Fraser C, Forty L, Russell E, Hamshere ML, Moskvina V, Nikolov I, Farmer A, McGuffin P, Wellcome Trust Case Control Consortium, Holmans PA, Owen MJ, O'Donovan MC, Craddock N: **The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia.** *Molecular Psychiatry* 2010, **15**:1016–1022.
36. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, Dionne-Laporte A, Spiegelman D, Henrion E, Diallo O, Thibodeau P, Bachand I, Bao JY, Tong AH, Lin CH, Millet B, Jaafari N, Joober R, Dion PA, Lok S, Krebs MO, Rouleau GA: **Increased exonic de novo mutation rate in individuals with schizophrenia.** *Nat Genet* 2011, **43**:860–863.
37. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, Gogos JA, Karayiorgou M: **Exome sequencing supports a de novo mutational paradigm for schizophrenia.** *Nat Genet* 2011, **43**:864–868.
38. Nachman MW, Crowell SL: **Estimate of the mutation rate per nucleotide in humans.** *Genetics* 2000, **156**:297–304.
39. Kondrashov AS: **Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases.** *Hum Mutat* 2003, **21**:12–27.
40. Ponting CP, Lunter G: **Signatures of adaptive evolution within human non-coding sequence.** *Hum Mol Genet* 2006, **15**:R170–175.
41. Eyre-Walker A, Keightley PD: **The distribution of fitness effects of new mutations.** *Nat Rev Genet* 2007, **8**:610–618.
42. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ: **Analysis of genetic inheritance in a family quartet by whole-genome sequencing.** *Science* 2010, **328**:636–639.
43. Goriely A, McVean GA, Röjmyr M, Ingemarsson B, Wilkie AO: **Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line.** *Science* 2003, **301**:643–646.
44. Goriely A, McVean GA, van Pelt AM, O'Rourke AW, Wall SA, de Rooij DG, Wilkie AO: **Gain-of-function amino acid substitutions drive positive selection of FGFR2 mutations in human spermatogonia.** *Proc Natl Acad Sci USA* 2005, **102**:6051–6056.
45. Qin J, Calabrese P, Tiemann-Boege I, Shinde DN, Yoon SR, Gelfand D, Bauer K, Arnheim N: **The molecular anatomy of spontaneous germline mutations in human testes.** *Public Library of Science Biology* 2007, **5**:e224.
46. Goriely A, Hansen RM, Taylor IB, Olesen IA, Jacobsen GK, McGowan SJ, Pfeifer SP, McVean GA, Rajpert-De Meyts E, Wilkie AO: **Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors.** *Nat Genet* 2009, **41**:1247–1252.
47. Goriely A, Wilkie AOM: **Missing heritability: paternal age effect mutations and selfish spermatogonia.** *Nat Rev Genet* 2010, **11**:589.
48. Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller HJ, Hartmann A, *et al*: **Large recurrent microdeletions associated with schizophrenia.** *Nature* 2008, **455**:232–236.
49. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, *et al*: **Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia.** *Science* 2008, **320**:539–543.
50. Williams NM, Zaharieva I, Martin A, Langley K, Mantripragada K, Fossdal R, Stefansson H, Stefansson K, Magnusson P, Gudmundsson OO, Gustafsson O, Holmans P, Owen MJ, O'Donovan M, Thapar A: **Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis.** *Lancet* 2010, **376**:1401–1408.
51. Mallal S, Nolan D, Witt C, Masel G, Martin AM, Moore C, Sayer D, Castley A, Mamotte C, Maxwell D, James I, Christiansen FT: **Association between presence of HLA-B*5701 HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir.** *Lancet* 2002, **359**:727–732.
52. Mallal S, Phillips E, Carosi G, Molina JM, Workman C, Tomazic J, Jägel-Guedes E, Rugina S, Kozyrev O, Cid JF, Hay P, Nolan D, Hughes S, Hughes A, Ryan S, Fitch N, Thorborn D, Benbow A, PREDICT-1 Study Team: **HLA-B*5701 screening for hypersensitivity to abacavir.** *N Engl J Med* 2008, **358**:568–578.
53. Furness P, Zimmern R, Wright C, Adams M: *The Evaluation of Diagnostic Laboratory Tests and Complex Biomarkers.* Cambridge and The Royal College of Pathologists, London: Public Health Genetics Foundation; 2008.
54. Manson N, O'Neill O: *Rethinking Informed Consent in Bioethics.* Cambridge, England: Cambridge University Press; 2007.
55. Lobb EA, Barlow-Stewart K, Suthers G, Hallowell N: **Treatment-focused DNA testing for newly diagnosed breast cancer patients: some implications for clinical practice.** *Clin Genet* 2010, **77**:350–354.
56. Wood F, Kowalczuk J, Elwyn G, Mitchell C, Gallacher J: **Achieving online consent to participation in large-scale gene-environment studies: a tangible destination.** *Journal of Medical Ethics* 2011, **37**:487–492.
57. Lunshof J, Chadwick R, Vorhaus D, Church G: **From genetic privacy to open consent.** *Nat Rev Genet* 2008, **9**:406–411.
58. Knoppers BM: **Consent to 'personal' genomics and privacy.** *European Molecular Biology Reports* 2010, **11**:416–419.
59. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M, Sagreiya H, Whaley R, Knowles JW, Chou MF, Thakuria JV, Rosenbaum AM, Zaranek AW, Church GM, Greely HT, Quake SR, *et al*: **Clinical assessment incorporating a personal genome.** *Lancet* 2010, **375**:1525–1535.
60. Khoury MJ, Wagener DK: **Epidemiological evaluation of the use of genetics to improve the predictive value of disease risk factors.** *Am J Hum Genet* 1995, **56**:835–844.
61. Wallace HM: **Model of gene-gene and gene-environment interactions and its implications for targeting environmental interventions by genotype.** *Theoretical Biology and Medical Modelling* 2006, **3**:35.
62. Terwilliger JD, Weiss KM: **Confounding ascertainment bias, and the blind quest for a genetic 'fountain of youth'.** *Annals of Medicine* 2003, **35**:532–544.
63. Guo S: **The behaviors of some heritability estimators in the complete absence of genetic factors.** *Hum Hered* 1999, **49**:215–228.
64. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: **Common SNPs explain a large proportion of the heritability for human height.** *Nat Genet* 2010, **42**:565–569.
65. Visscher PM, Yang J, Goddard ME: **A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al (2010).** *Twin Research and Human Genetics* 2010, **13**:517–24.
66. Lanktree MB, Guo Y, Murtaza M, Glessner JT, Bailey SD, Onland-Moret NC, Lettre G, Ongen H, Rajagopalan R, Johnson T, Shen H, Nelson CP, Klopp N, Baumert J, Padmanabhan S, Pankratz N, Pankow JS, Shah S, Taylor K, Barnard J, Peters BJ, Maloney CM, Lobmeyer MT, Stanton A, Zafarmand MH, Romaine SP, Mehta A, van Iperen EP, Gong Y, Price TS, *et al*: **Meta-analysis of Dense Genecentric Association Studies reveals common and uncommon variants associated with height.** *Am J Hum Genet* 2011, **88**:6–18.
67. So HC, Li M, Sham PC: **Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study.** *Genetic Epidemiology* 2011, **35**:447–456.
68. Kahn SD: **On the future of genomic data.** *Science* 2011, **331**:728–729.