Variation in the human genome and risk to common disease

Keystone Symposium on 'Human Genome Sequence Variation and the Inherited Basis of Common Diseases'
January 8–13, 2004, Breckenridge, Colorado, USA

Paul IW de Bakker†, Richa Saxena & Robert R Graham
†Author for correspondence
Massachusetts General Hospital, Department of Molecular Biology, Boston, MA 02114, USA
Tel: +1 617 726 5942; Fax: +1 617 726 5937;
E-mail: debakker@molbio.mgh.harvard.edu

Common diseases, such as Type 2 diabetes, affect millions of people across the globe, but many remain poorly understood and often lack effective treatments. Pinning down the genetic risk factors of common diseases could spur the development of better therapies. The completion of the Human Genome Project has produced resources and sparked technological advances, providing an unprecedented glimpse into the organization and function of the genome. The goal of the recent Keystone Symposium entitled 'Human Genome Sequence Variation and the Inherited Basis of Common Diseases' was to provide an overview of sequence variation patterns in the human genome, give insight into the genetic architecture of common disease, and discuss methods to identify variants and pathways that underlie common disease.

Another goal of the meeting was to bring researchers in various disciplines together to share insights, in the hope that diverse perspectives will catalyze progress on this complex problem. Along these lines, the meeting was held concurrently with the 'Natural Variation and Quantitative Genetic in Model Organisms' symposium, sharing several joint sessions as well as numerous poster sessions and social activities.

The HapMap and ENCODE projects

Most sites in the genome that vary in the general human population are rare, while the bulk of genetic diversity in any given individual is due to common variants. The latter are likely to exercise modest yet real effects on disease risk, and studies using large samples will convincingly expose such variants [1]. In recent years, human geneticists have proposed to conduct comprehensive surveys of the common genetic variation to identify risk factors that contribute to common disease.

The public databases of single nucleotide polymorphisms (SNPs) contain at present a significant fraction of the ~ 10 million common SNPs that are thought to exist in the entire human population. Initially, performing whole-genome scans of individuals was thought to be impractical as an enormous number of SNPs would have to be screened. It seemed prohibitively expensive considering the large number of people that are typically required to reach adequate statistical power. However, that pessimistic picture changed a few years ago when a number of studies [2] reported an extensive correlation structure (linkage disequilibrium [LD]) in the human genome which would allow efficient tagging on a genome-wide scale. To this end, the International HapMap Project's mandate is to chart the correlation structure at an unprecedented level of precision for different populations [3].

So what does the data at present look like? M Daly (Whitehead Institute, USA) presented early data from the NHGRI-funded ENCyclopedia Of DNA Elements (ENCODE) project. This pilot project is part of the HapMap and aims to extensively characterize a number of 500-kb regions of varying gene density and non-exonic conservation to give immediate insight into their detailed structure. Lessons learned from these regions can then be directly applied to the HapMap, most importantly to determine appropriate marker densities to ensure optimal coverage of the variation. From five ENCODE regions genotyped at high marker densities, Daly showed that LD across the regions is significant and that much of the data has a block-like structure. The finding that a limited number of markers (tags) can capture most of the variation present is of critical importance. If this holds true for the remainder of the genome, then the HapMap will turn out to be the prudent strategy for improving the efficiency and power of large-scale (genome-wide) association studies.

Daly briefly discussed the concept of haplotype blocks and emphasized that they can be used as convenient descriptors of local patterns of LD. But there is really a lot more to LD than just discrete blocks, due to the presence of higher order correlations. Adjacent blocks are often tightly linked, suggesting that a simple block-by-block tagging approach is unlikely to be optimal. In fact, the observed LD features of the ENCODE regions strongly suggest that the ability to choose tags optimally (when HapMap is complete) results in a two- to fourfold gain in efficiency (i.e., requiring fewer markers) compared to random marker selection. It is estimated that as few as 500,000 markers will be able to adequately capture all the common variation in the genome.

Tag single nucleotide polymorphisms

While the above projections inspire enthusiasm for the HapMap Project, its ultimate utility will be judged by its capability to provide tags that assay even unobserved variable sites in unexamined individuals, given that the HapMap is only representative of a limited, and therefore incomplete, sample. This general issue of external validation was addressed directly by D Goldstein (University College London, UK), who suggested that rarer polymorphisms may not be amenable to an effective tagging approach, which might limit the scope of the HapMap. Using a logistic regression method, Goldstein compared the internal and external performance of selected tags and found that sites with a < 7% minor allele frequency in an independent subset of individuals were predicted rather poorly. It is not clear how this result applies to tagging approaches in general (or the HapMap in particular), especially in light of the small sample size of his analysis, with only 64 chromosomes divided into training and test sets of equal size. Spurious correlations in the training set may cause overfitting and thus result in 'greedy' tagging with poor performance in the actual test set. These findings call for the assignment of significance levels for hidden site prediction and validation in independent data sets. Several presenters also expressed the view that analytical improvements to existing methods could be developed to take advantage of the longer range LD surrounding rarer polymorphisms. These issues do not detract from the principal justification for a tagging approach based on the HapMap. It is certainly encouraging to see that the field is moving towards evaluating concretely the trade-off between the number of markers and statistical power, as well as the rigorous assessment of the fraction of common variation in the genome that can be captured effectively via tagging.

Recombination hot spots and haplotype blocks

Early sperm studies have identified specific hot spots in which crossing-over events cluster [4]. These so-called recombination hot spots influence the patterns of LD in the genome with important implications for disease gene mapping. L Cardon (Wellcome Trust Centre for Human Genetics, UK) demonstrated that a physical versus genetic correspondence map reveals a non-uniform distribution of recombination rates across the genome, reinforcing the notion of recombination hot spots. Supporting these findings, R Durbin (Wellcome Trust Sanger Institute, UK) directly observed recombination events in pedigrees and demonstrated that haplotype block boundaries correlate well with recombination events. P Donnelly (University of Oxford, UK) presented a 'composite-likelihood' method for estimating recombination rates from genotype data and concluded that ~ 80% of all recombination events occur within ~ 20% of the human genome sequence. R Hudson (University of Chicago, USA) explained how a composite-likelihood approach can be applied to fine-scale mapping by estimating genotype probabilities. These studies all help explain the nature of the population genetic forces that have shaped sequence diversity in the human population over time.

Genetic architecture of human disease

One goal of the meeting was to describe progress in dissecting the genetics of common disease. Most speakers agreed that the value of human genetic analysis is primarily to illuminate and validate molecular pathways that contribute to disease in order to treat disease in the human population more effectively, rather than to predict disease in the individual. Different approaches to the study of human diseases were presented and highlighted the challenges inherent to complex disease genetics.

A prevailing theme was that genetic association studies should be designed to detect both rare and common variation, as both can be informative in complex diseases. In his keynote address, A Chakravarti (Johns Hopkins University, USA) provided a historical perspective on disease genetics and illustrated the power of association studies in understanding the molecular basis of the rare Mendelian diseases β-thalassemia and cystic fibrosis, and of Hirschsprung disease. For the latter, complex disease, Chakravarti described how whole genome association scans identified multiple loci that influence the complex inheritance patterns, chief among them a genetic interaction between a common haplotype in the *RET* oncogene and mutations in the endothelin receptor B gene *EDNRB* [5]. A very common risk haplotype in *RET*, often homozygous in affected individuals, explains the rare cases where the mutation in Hirschsprung disease maps to *EDNRB* but does not associate with *RET*. This work is a remarkable demonstration that rare diseases can also be influenced by common variation.

A different approach to dissecting pathways underlying common disease was presented by R Lifton (Yale University, USA), who argued that the study of rare diseases may illuminate genes and pathways involved in common disease, and that medications targeting these pathways can be used to treat common forms of a disease. Families with rare Mendelian syndromes of high and low blood pressure were studied to gain understanding of the molecular pathways involved in hypertension. Most genes identified using linkage analysis in these families were found to be members of the renin–angiotensin pathway. Drugs that target this renal salt-handling pathway have been tremendously successful for the effective treatment of hypertension, even though mutations in these genes are extremely rare. The recent elucidation of with-no-lysine kinase 4 (WNK4) as a molecular switch controlling renal sodium and potassium excretion is another example of how a rare syndrome can inform about the causes of more common forms of the disease [6].

D Altshuler (Massachusetts General Hospital, USA) underscored the theme that the true value in human genetics lies in the understanding of the molecular basis of disease. To illustrate that causal alleles need not always be rare

and selected against, he showed that for puberty-onset Type 1 diabetes, which is lethal if untreated, three risk alleles (in the human leukocyte antigen [HLA], insulin [*INS*] and cytotoxic T lymphocyte antigen 4 [*CTLA4*] genes) are, in fact, common, and together explain more than half of the inherited risk. This example and others persuasively demonstrate that the balance between evolutionary selection and the frequency spectra of mutations/variants influencing disease is extremely complicated. This prompted Altshuler to propose that the prior probability distribution of variants being associated to disease is generally much flatter (with regard to allele frequency) than has been argued before. He also suggests that an association study should state explicitly the prior probability of finding an association in the candidate locus to disease. This will be largely dependent on a number of factors: number of variants and their allele frequency spectra; genomic annotation (whether a variant is coding or regulatory); effect (risk) on disease; and correlation amongst variants. Prior probabilities are typically extremely low (on the order of 10^{-6} for a SNP and 10^{-4} for a gene), so that much more substantial evidence would be required before an association can be declared as real. Consideration of gene–gene or gene–environment interactions would lower the priors even more. Other evidence may include biological plausibility, linkage in a previous study, work in model organisms, biochemical or other functional data, and gene expression analyses. In the context of recent work on Type 2 diabetes [7], two further issues were highlighted: the need for large, independent replication studies to confirm reported associations, and the limitation of LD mapping to resolve a causal allele or SNP in a haplotype-based test or when multiple variants are tightly linked.

D Hunter (Harvard School of Public Health, USA) assessed gene–environment interactions for known polymorphisms associated with several common diseases. As power is decreased in stratification of patients by environmental influences, large sample sizes are key, and prospective cohorts may be less biased than case/control groups ascertained for a particular disease. Of three nested case/control studies derived from large cohorts, he specifically addressed the evidence for the influence of the PPAR-γ Pro12Ala variant on physiological response to fat intake [8], a substantial decline in cognitive ability in apolipoprotein E4 (ApoE4)-carrying women with untreated hypertension, and the genetic variation explaining the regular use of aspirin and a decreased risk of colon cancer.

General requirements for genetic association studies were summarized by S O'Brien (National Cancer Institute, USA) in the context of the 15 AIDS restriction genes that modify susceptibility to HIV infection or AIDS progression. These requirements are worth repeating here: large sample sizes with detailed phenotype definitions; strong epidemiological influence in selection of effects to examine; selection of candidate genes with plausible functional rationalization; assessment of association of single variants and combination of variants (haplotypes); and independent replication.

Pharmacogenetics

There is tremendous heterogeneity in the way patients respond to drugs, in terms of both unwanted side effects (toxicity) and treatment efficacy (pharmacokinetics and dynamics). Although the overall pharmacological effects of drugs are typically not monogenic traits, D Goldstein argues that drug response is a 'simpler' complex trait than are common diseases in general, as some of the important pharmacogenetic variation resides in only few candidate genes for any given drug [9]. These candidate genes tend to be rather obvious; for example, drug-metabolism enzymes, transporters, and receptors. This was illustrated by a literature survey of the genetic determinants of drug response. Most reported associations are in the drug target itself (or in the broader pathway), whereas nearly one-third are in drug-metabolizing enzymes. He further commented that pharmacogenetic testing has not been successfully integrated into clinical practice due to intrinsic limitations of the studies performed hitherto, mostly pointing the finger at small sample sizes, limited genetic information (testing only candidate polymorphisms instead of examining all common variation at a locus), neglect of gene–gene interactions, and population stratification. It was suggested by others that the reluctance of medical professionals, insurance providers and regulatory hurdles are also to blame.

Genetic association studies may help drug discovery by pointing out potential drug targets. One can imagine that when a coding variant that changes protein structure is associated with disease risk, the protein in question may represent a key step in the disease pathway, and therefore constitute an attractive drug target. For example, variants in the genes encoding PPAR-γ and sulfonylurea receptor 1 (SUR1)/K_{ir}6.2 have been shown to be associated with risk to Type 2 diabetes. PPAR-γ is a nuclear hormone receptor regulating adipogenesis and is the target of the thiazolidinedione class of oral hypoglycemic agents (most notably, rosiglitazone), while SUR1 is the target of sulfonylurea drugs (e.g., glyburide and glipizide).

Comparative genomics and the regulation of gene expression

Even though the sequence of ~ 3 billion base pairs of the human genome is known, the biological function of the vast majority remains unclear. Variants that alter protein sequence (and structure) or influence gene expression patterns can cause disease but only ~ 1% of the genome is encompassed in coding sequence and the bases that regulate gene expression have proven more difficult to identify. Comparative genomics and studies of regulation of gene expression are likely to further our understanding of genome structure, and could make disease gene mapping studies more efficient and successful.

Comparative genomics offers a powerful tool to identify the base pairs that

are evolutionarily conserved across species and provide insight into their biological role. E Lander (Whitehead Institute, USA) presented results of such studies in humans and yeast [10]. The complete sequence of multiple closely related yeast strains allowed a significant refinement of the annotation of the yeast genome. An algorithm to identify coding sequence was developed that takes advantage of the observation that coding regions tend to be highly conserved across species. The relative absence of mutations, gaps and frameshift mutations across species coupled with standard gene parameters, proved sensitive enough to identify exons of > 120 base pairs. Using this algorithm, ~ 500 sequences that were initially annotated as genes were found to be non-functional, reducing the current gene count in yeast to ~ 5700. Comparative genomics techniques were used to identify regulatory regions in the yeast genome. Across the yeast species, intergenic regions were strongly conserved, and a comparative analysis identified novel regulatory motifs. An analysis of the binding motifs of known transcription factors suggested a pattern of combinatorial control, where multiple transcription factors interact at a single locus to regulate a gene.

The lessons learned and tools produced to study the yeast genomes should prove invaluable for future comparative genomics projects, especially those in the human genome. Lander went on to propose that the future availability of the complete sequences of 16 mammalian genomes would enable us to perform similar comprehensive analyses on the human genome.

Elucidating the regulation of gene expression may also provide insights into the function of the genome. The availability of genome sequences and the development of sensitive genome-wide expression analyses have allowed researchers to address this complex problem for the first time. L Kruglyak (Fred Hutchison Cancer Research Center, USA) described a model in yeast that allows the genetic analysis of expression patterns [11], and proposed that transcript levels could be used as a tool to dissect regulatory networks. An expression map of the yeast genome was described that compares expression level differences between two parental yeast species and their progeny. About half of the yeast genome was differentially expressed between the two strains, yet only 10% of genes account for nearly half of the variance. Kruglyak estimated that the differences in most loci were the result of interactions with other genes. In fact, a small number of loci (15) were linked to changes in the expression of thousands of genes. K White (Yale University, USA) and S Monks (University of Washington, USA) reported on their efforts to elucidate the regulation of gene expression in the fruit fly and mouse, respectively.

Expert opinion

This conference provided a unique opportunity for interaction and exchange of ideas between scientists in traditionally disparate fields of genetics. Certainly, much progress has been made in understanding the detailed patterns of LD within the human genome. Empirical data and theoretical calculations have revealed patterns of non-uniform recombination rates across the genome, which is a significant advance. Preliminary analyses of the HapMap look very encouraging as SNP-tagging approaches will improve the efficiency and power of large-scale association studies.

The second major theme of the conference was the genetic architecture of complex disease, methods to improve the design of association studies, and more rigorous criteria to evaluate statistical significance. As is evident from studies carried out by various geneticists, many different approaches have led to advances in dissecting genes and pathways associated with common diseases. But it is clear that there are guiding principles that every geneticist should follow in genetic association studies: clear phenotype definition and assignment; removal of confounding bias or population stratification by case/control matching; and proper correction for multiple hypothesis testing.

Highlights

- Genome-wide patterns of sequence variation (LD) and the key forces that shape them, including recombination hot spots, are now coming into better focus.
- More efficient and powerful association studies can be designed using SNP-tagging methods to identify markers that capture the variation at a locus.
- Genetic associations must be evaluated using strict criteria and appropriately corrected for multiple hypotheses tested to ensure adequate statistical power – independent replication in a large sample remains the only reliable way to confirm a putative association.

Outlook

To elucidate the role of variants that underlie common disease, genetic association studies are the method of choice, but their success critically depends on the ability to collect large samples of patients (and controls). As a potential solution, F Collins (NIH, USA) revealed tentative plans to form a large prospective cohort (involving hundreds of thousands of individuals) as an open-access resource for studying complex diseases, within which nested case/control studies can be performed. Although this initiative aims to provide sufficient power to examine gene–gene and gene–environment interactions, it remains to be seen whether this plan will ever be realized, given the enormous investment needed, infrastructure challenges and ethical issues.

A cautionary note from industry was offered by K Lindpaintner (F Hoffmann-La Roche, Switzerland) who declared that personalized medicine or pharmacogenetics will not deliver the 'magic bullet' to treat common diseases any time soon. With the heritability of common diseases ranging from 0.2 to 0.5, he emphasized the overall difficulty in developing effective drugs given the large array of etiological risk factors. Indeed, R&D expenditure is steadily rising, drug development pipelines have not become any shorter, and New Chemical Entity failure

rates remain high. Recent successes in drug discovery notwithstanding, Lindpaintner argues that projected advances in genetics and genomics have been widely overstated. Rather, these exciting developments 'will provide help but no patent solutions' [12].

While many of us in the 'postgenome era' are concerned with delivering public health benefits that have been promised by genome projects and the like, it is indeed a somewhat depressing fact that only few pharmacogenetic tests are routinely employed in the clinic. Hopefully, further advances in analytical tools and genotyping technologies will eventually aid the clinician in deciding upon the optimal pharmacotherapy for the individual patient. The developments presented at this meeting offer optimistic prospects for unravelling the genetic components that predispose individuals to disease and help pave the way for finding better treatments.

Bibliography

1. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* 33, 177-182 (2003).
2. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229-232 (2001).
3. The International HapMap Consortium: The International HapMap Project. *Nature* 426, 789-796 (2003).
4. Jeffreys AJ, Kauppi L, Neumann R: Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* 29, 217-222 (2001).
5. Carrasquillo MM, McCallion AS, Puffenberger EG, Kashuk CS, Nouri N, Chakravarti A: Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat. Genet.* 32, 237-244 (2002).
6. Kahle KT, Wilson FH, Leng Q *et al.*: WNK4 regulates the balance between renal NaCl reabsorption and K^+ secretion. *Nat. Genet.* 35, 372-376 (2003).
7. Florez JC, Hirschhorn J, Altshuler D: The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu. Rev. Genomics Hum. Genet.* 4, 257-291 (2003).
8. Memisoglu A, Hu FB, Hankinson SE *et al.*: Prospective study of the association between the proline to alanine codon 12 polymorphism in the PPARγ gene and Type 2 diabetes. *Diabetes Care* 26, 2915-2917 (2003).
9. Goldstein DB, Tate SK, Sisodiya SM: Pharmacogenetics goes genomic. *Nat. Rev. Genet.* 4(12), 937-947 (2003).
10. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241-254 (2003).
11. Yvert G, Brem RB, Whittle J *et al.*: Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.* 35, 57-64 (2003).
12. Lindpaintner K: The impact of pharmacogenetics and pharmacogenomics on drug discovery. *Nat. Rev. Drug Discov.* 1, 463-469 (2002).