Human Genome Variation 2006: emerging views on structural variation and large-scale SNP analysis

Gonçalo Abecasis[1], Paul Kwong-Hang Tam[2], Carlos D Bustamante[3], Elaine A Ostrander[4], Stephen W Scherer[5], Stephen J Chanock[6], Pui-Yan Kwok[7,9] & Anthony J Brookes[8,9]

The eighth annual Human Genome Variation Meeting was held in September 2006 in the Hong Kong Special Administrative Region, China. The meeting highlighted recent advances in characterization of genetic variation, including genome-wide association studies and structural variation.

The eighth annual Human Genome Variation Meeting was held September 14–16, 2006 in Hong Kong, bringing together experts in population and evolutionary genetics, genotyping and sequencing technologies, genome informatics and gene mapping. This meeting series began almost a decade ago and was then entitled 'SNPs and Complex Genome Analysis', in response to the emergence of technologies for SNP detection and the resulting focus of the field on characterizing this type of variation. The first small gathering, 'SNP1998', included about 50 investigators who came together for workshop-style brainstorming and debate. The event was extremely successful and consequently grew into a leading international meeting that now reluctantly but wisely limits attendee numbers to under 200 to retain its dynamic 'workshop' atmosphere. The organizing committee of about six active researchers changes each year to keep the content topical, with Anthony Brookes and, in more recent years, Pui-Yan Kwok and Stephen Chanock taking the role of main organizer. The meeting has witnessed exceedingly rapid maturation of the SNP field and has lately seen this expand to include characterization of structural variation (**Box 1**). In response, and reflecting the multidisciplinary challenge that is genome variation research, the meeting title was this year changed to 'Human Genome Variation'.

After years of method development, SNP discovery, sequencing and high-throughput genotyping by the community, a notable discussion point in this year's meeting was the likely outcome of the first wave of genome-wide SNP association studies that make use of these previous advances. There was a general, but not universal, view that these studies will provide important insights not only into disease and pharmacogenetics but also into normal genetic variation. The underlying technologies will, however, undoubtedly open new vistas for related disciplines such as basic population genetics and structural variation analysis. Alongside the enthusiasm for the latest SNP genotyping technologies and the discoveries that the current crop of genome-wide association studies might bring, there was also already a clear sense that these technologies may soon be replaced by new, inexpensive sequencing methods. These can be expected to provide even higher-throughput capabilities and allow more detailed analysis of individual genomes.

Structural variation provides new challenges and opportunities

The ubiquity of structural variation, and especially copy number variation (CNV), in the human genome was illustrated in presentations by Stephen Scherer, Anthony Brookes, Alexander Urban, Ching Lau and Charles Lee. It is now clear that multi-kilobase, and even multi-megabase, instances of such variants are abundant throughout our genome, and they have the potential to affect human disease—for example, by dramatically altering gene expression. A consortium effort to identify most large and common structural variation was reported, and the resulting primary data have since been made available online (at http://projects.tcag.ca/variation/ and http://www.sanger.ac.uk/humgen/cnv/data/). But there is currently no optimal way of assessing these very interesting polymorphisms on a genomic scale, and an even bigger technology gap exists regarding short-range, low–copy number and/or rare structural variants—of which there may be many. Presentations by Yuan-Tsong Chen and Chack-Yung Yu nevertheless detailed clinical implications for CNV in disease outcomes and pharmacogenomics. Generally, there is only limited overlap in the CNV sets (including insertions and deletions) identified by different experimental methods, even when these different methods were applied to the same sample set. This all suggests that structural variation will be a key feature of the meeting series for many years to come.

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA. [2]Division of Paediatric Surgery, University of Hong Kong, Hong Kong Special Administrative Region, China. [3]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, USA. [4]Cancer Genetics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA. [5]The Centre for Applied Genomics, Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada. [6]Pediatric Oncology Branch, National Cancer Institute, Bethesda, Maryland, USA. [7]Cardiovascular Research Institute and Institute for Human Genetics, University of California, San Francisco, California, USA. [8]Department of Genetics, University of Leicester, Leicester, UK. [9]These authors contributed equally to this work.
e-mail: ajb97@leicester.ac.uk

BOX 1 PAST MEETINGS IN THIS SERIES

This meeting series has focused on characterizing genome variation but has changed in focus along with the field, following technological developments and shifts in approaches. Below are representative reports on conferences in this series over the past decade.

Anonymous. SNP attack on complex traits. *Nat. Genet.* **20**, 217–218 (1998).

Pennisi, E. A closer look at SNPs suggests difficulties. *Science* **281**, 1787–1789 (1998).

Syvanen, A.C., Landegren, U., Isaksson, A., Gyllensten, U. & Brookes A. First international SNP meeting at Skokloster, Sweden, August 1998. Enthusiasm mixed with scepticism about single-nucleotide polymorphism markers for dissecting complex disorders. *Eur. J. Hum. Genet.* **7**, 98–101 (1999).

Isaksson, A. *et al.* Discovery, scoring and utilization of human single nucleotide polymorphisms: a multidisciplinary problem. *Eur. J. Hum. Genet.* **8**, 154–156 (2000).

Nelson, D.L. SNPs, linkage disequilibrium, human genetic variation and Native American culture. *Trends Genet.* **17**, 15–16 (2001).

Brookes, A.J., Kwok, P.-Y., White, P.S., Oefner, P.J. SNP 2000: third international meeting on single nucleotide polymorphism and complex genome analysis (special issue). *Hum. Mutat.* **17**, 241–356 (2001).

White, P.S., Kwok, P.-Y., Oefner, P., Brookes, A.J. 3rd international meeting on single nucleotide polymorphism and complex genome analysis: SNPs: 'some notable progress'. *Eur. J. Hum. Genet.* **9**, 316–318 (2001).

Brookes, A.J. 4th international meeting on single nucleotide polymorphism and complex genome analysis. Various uses for DNA variations. *Eur. J. Hum. Genet.* **10**, 153–155 (2002).

Clark, A.G. The 6th international meeting on single nucleotide polymorphism and complex genome analysis. *Pharmacogenomics* **5**, 153–156 (2004).

Rocha, D. *et al.* Seventh international meeting on single nucleotide polymorphism and complex genome analysis: 'ever bigger scans and an increasingly variable genome'. *Hum. Genet.* **119**, 451–456 (2006).

From SNP genotyping to global genetic variation

High-density SNP genotyping across the genome is increasingly commonplace, and the increased data density provided by such technologies furnishes a rich resource for investigation of genome variation and population genetics. Design and interpretation of these studies will need to allow for known and unknown regions of structural variation which, as shown by Anthony Brookes, can sometimes lead to SNP genotyping errors. Describing their analysis of high-throughput SNP data sets, Molly Przeworski and Carlos Bustamante presented new strategies for detecting signals of natural selection across the genome, with a particular focus on genic regions. Jonathan Pritchard presented work extending the now generally accepted premise that tag SNPs selected from a small number of populations genotyped by the International HapMap consortium capture common variants in a very diverse set of human populations. Thus, not only are most genetic variants shared across the globe, but their patterns of linkage disequilibrium and haplotype variation are also highly similar. Constantin Polychronakos highlighted unusual strand bias in synonymous amino acid substitutions across the genome, while Peter Keightley discussed selective constraint in noncoding regions—both talks thus offering important insights for disease gene mapping. Also relevant to complex disease association analysis, Charles Rotimi highlighted the value of using the more diverse chromosomes and smaller haplotype blocks that exist in African populations and crystallized the valuable message that "under the skin, we are all Africans."

Toward genome-wide association studies

With laboratory technologies racing forward, many speakers felt that the scale of genomic data sets will challenge existing analysis tools. With this in mind, Pak Sham, Christoph Lange and Gonçalo Abecasis showcased new, efficient methods for analyzing genome-wide association study data sets and debated the most effective manner in which these data could be used to uncover variants underlying human diseases. Stephen Chanock described one such project, the Cancer Genetic Markers of Susceptibility (CGEMS; http://cgems.cancer.gov) study, which will help with analysis development by providing qualified investigators with access to raw genotype data, as well as by posting precomputed results online. Such genome-wide data sets will be a treasure trove not only for identifying disease susceptibility variants but also for many other uses (such as studying structural variants in the genome). Hopefully, this standard of open access will be matched by other similar large studies. In the area of personalized medicine, Yuan-Tsong Chen illustrated how technological advances have shortened the interval between collecting suitable patient materials and finding disease genes by reporting several alleles that are strongly associated with adverse drug responses. The strength and specificity of the associations presented suggests that the clinical implementation of pharmacogenomics may not be far off, but bridging the communications gap between the research community and clinical medical professionals will require some effort.

Applying the very latest technologies, a number of genome-wide studies are now underway. Andrew Clark and Neil Risch provided some exciting first glimpses into these large data sets, including some remarkable findings in population genetics. For instance, Andrew Clark examined global FST in different regions of the US and noted clear evidence of a gradient of allele frequencies that could partially be explained by the demographic history. He also noted some striking differences in heterozygosity and patterns of linkage disequilibrium between HapMap Caucasians and Ashkenazi subjects from an ongoing study. Thus, the deeper we look into the genome, the more surprises we see, and our experimental designs may need to be reconsidered. For example, Neil Risch argued that once genotypes for hundreds of thousands of SNPs are routinely available and most common variants are tagged, the standard association study design will probably make redundant some of the current advantages of admixture mapping.

New genetic associations for human diseases provide backdrop

Several speakers presented compelling cases for new disease associations with a wide range of conditions in study populations from across the world. Kyoyung Song discussed the immunogenetic basis of systemic lupus erythematosus in different Asian populations, while Francisco De La Vega presented new findings regarding the genetics of Crohn disease. Enhancer and promoter mutations rather than coding region defects are steadily being recognized as important in some pathologies, such as in many cases of Hirschsprung disease, as described by Paul Tam. And perhaps the genetic basis of complex traits might often entail rare variants rather than common variants that can be found by today's genome-wide strategies. In line with this, Grant Montgomery reported rare variants in a candidate gene (*GDF9*) that contribute to variation in dizygotic twinning. Taking a different strategy, Hidetoshi Inoko made the case for genome-wide association analysis with dense microsatellites. To simplify both the genetic and the phenotypic sides of the equation, Nathan Sutter presented compelling data on common genetic variation and growth in dogs, highlighting a rich resource for discovery and validation of disease associations relevant to humans.

Hopefully, positive association signals such as these will be reproducible when examined by others. In the future, risk variants detected by increasingly deep genome-wide association studies will likewise have to undergo replication tests of validity. Perhaps these signals will be stronger and more reliable since they were produced by comprehensive scans, but alter-

natively the findings from these large studies could be overwhelmed by false positives owing to the many semi-independent tests (markers) that will have been considered.

Genetic variation and gene expression profiling

The burgeoning field of germline genetic variation and gene expression profiling was the subject of another session. Expression patterns represent phenotypes that are probably more directly influenced by genetic variants than are typical disease traits, and as such the effect sizes will be larger and easier to detect. Vivian Nap-Yee Chan reported data from such analyses, and Dan Schaid presented new methods to search for SNPs that are associated with an exceptionally large number of gene expression traits, aiming to detect genomic variation in 'master regulators' of gene expression. Making sense of such findings will require an understanding of where in relation to genes one should expect to find elements that influence gene expression. On this point Ed Liu showed some surprising data on global and specific effects of p53-related genes, indicating that many control elements may be located a long way from the transcribed exons.

New technologies geared for resequencing

Technology development is the lifeblood of human genome variation research. Improved methods are always needed, especially for new areas of interest such as structural variation. Many now believe that ultra–high-throughput (re)sequencing may be key to properly unlocking the genome's secrets, and this was apparent in talks by David Bentley, Jingyue Ju and Elaine Mardis. One example application of resequencing would be the discovery of somatic mutations, an advance that would hold promise for development of new therapies for cancer and perhaps other diseases. Some commercial developments in resequencing technologies were presented, with goals of extending the reach of genome-wide association studies beyond assessment of common variants to detect rare variants. The new analytical challenges that will arise from such large-scale resequencing projects, especially to detect those functionally important but rare variants from among the background diversity, were a topic of interest in discussions. Genome-wide resequencing could even lead to deemphasizing the role of genome-wide association studies of common variants, even though this latter approach has become commonplace only relatively recently. Moving forward, we should tie together the new technologies and approaches for characterizing genome variation, as emphasized by Pui-Yan Kwok, who discussed an integrated approach toward genome-wide association studies using parallel investigation in other sectors, such as gene expression and protein profiling.

A presentation by Kenshi Hayashi suggested advantages of moving away from studying the normal diploid genome. In an elegant example of using a rare resource, he showed how complete hydatidiform moles could be used to make extensive and unambiguous haplotype maps. But before such new technologies and strategies overtake current platforms, there is still a role for thorough annotation of genetic variation across exons. To this end, Tamsin Eades described the Exon Sequencing (ExoSeq) project of the Wellcome Trust Sanger Institute, which aims to sequence all exons in DNA samples from 48 unrelated individuals of north and west European origin.

The meeting evolves and continues

The 2006 Human Genome Variation Meeting was permeated by a pioneer atmosphere that has been apparent at each gathering in this series, although the focus of the meeting has constantly changed and the scales of the reported studies have increased. One might predict that next year's meeting (tentatively scheduled for September 2007 in Spain) will include exciting updates on the first wave of genome-wide association studies in complex disease, new insights and technologies for structural variation analysis and bioinformatics and database developments that will be critical for integrating all this genotype-phenotype information. The meeting will continue to 'evolve' along with the changing research in the field, but it remains focused on a desire to understand how human genomes vary and how this makes us who and what we are.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.