

Genomic Variation in a Global Village: Report of the 10th Annual Human Genome Variation Meeting 2008

Anthony J. Brookes,¹ Stephen J. Chanock,² Thomas J. Hudson,³ Leena Peltonen,⁴ Gonçalo Abecasis,⁵ Pui-Yan Kwok,^{6,7} and Stephen W. Scherer^{8*}

¹Department of Genetics, University of Leicester, Leicester, United Kingdom; ²Division of Cancer Epidemiology and Genetics and Center for Cancer Research, National Cancer Institute, Bethesda, Maryland; ³Ontario Institute of Cancer Research, Toronto, Ontario, Canada; ⁴Wellcome Trust Sanger Institute, Cambridge, United Kingdom; ⁵Department of Biostatistics, University of Michigan, Ann Arbor, Michigan; ⁶Department of Dermatology, Cardiovascular Research Institute, University of California, San Francisco, San Francisco, California; ⁷Institute for Human Genetics, University of California, San Francisco, San Francisco, California; ⁸The Centre for Applied Genomics, Program in Genetics and Genomic Biology, Research Institute, The Hospital for Sick Children, Toronto, Ontario, Canada

Received 6 February 2009; accepted revised manuscript 13 February 2009.

Published online 3 March 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21008

ABSTRACT: The Centre for Applied Genomics of the Hospital for Sick Children and the University of Toronto hosted the 10th Human Genome Variation (HGV) Meeting in Toronto, Canada, in October 2008, welcoming about 240 registrants from 34 countries. During the 3 days of plenary workshops, keynote address, and poster sessions, a strong cross-disciplinary trend was evident, integrating expertise from technology and computation, through biology and medicine, to ethics and law. Single nucleotide polymorphisms (SNPs) as well as the larger copy number variants (CNVs) are recognized by ever-improving array and next-generation sequencing technologies, and the data are being incorporated into studies that are increasingly genome-wide as well as global in scope. A greater challenge is to convert data to information, through databases, and to use the information for greater understanding of human variation. In the wake of publications of the first individual genome sequences, an inaugural public forum provided the opportunity to debate whether we are ready for personalized medicine through direct-to-consumer testing. The HGV meetings foster collaboration, and fruits of the interactions from 2008 are anticipated for the 11th annual meeting in September 2009.

Hum Mutat 30:1134–1138, 2009. © 2009 Wiley-Liss, Inc.

KEY WORDS: SNP; CNV; GWAS; personalized medicine

Introduction

The 10th international Human Genome Variation (HGV) Meeting convened in October 2008, hosted by The Centre for Applied Genomics of the Hospital for Sick Children and the University of Toronto (Toronto, Canada). This annual gathering began in 1998 as a small workshop focussed on “Single Nucleotide Polymorphisms (SNPs) and Complex Genome Analysis.” Although the number of

*Correspondence to: Stephen W. Scherer, PhD, FRSC, The Centre for Applied Genomics, The Hospital for Sick Children, 14th Floor, Toronto Medical Discovery Tower/MaRS Discovery District, 101 College St., Toronto, Ontario, M5G 1L7, Canada. E-mail: steve@genet.sickkids.on.ca

delegates has expanded incrementally (to about 240 registrants representing 34 countries) and the scope of variation described has broadened appreciably, the conference’s interactive workshop flavor prevails. A tone of anticipation was set with opening remarks by Anthony Brookes and Stephen Scherer, who explained that everyone in attendance had brought relevant science with them and that titles of talks had been withheld from the program to encourage speakers to present their “hottest” data. The 3-day format included eight plenary sessions (Box 1) with 38 speakers, interspersed with breakout poster-viewing sessions, a keynote address by Professor Svante Pääbo (Max Planck Institute, Germany), and an inaugural public forum entitled “Personalized Medicine: Are We Ready?” (Box 2) Additional information about the meeting is at www.tcag.ca/hgv2008, including the full program of speakers.

In addition to the global breadth of representation, the professional expertise of delegates covered a spectrum from genetics and genomics, statistics and computational biology, through medical practice, human and vertebrate biology, epidemiology, technology development, to ethics and the law. A move to more cross-disciplinary presentations was evident and reflected the conference’s aim to merge rather than stratify approaches to the study of human genome variation. Notable themes emerging were of integration (of approaches, sources of data, and types of data) and location (in the world, and in the genome). We discuss these, as well as social implications and other trends manifest in this year’s proceedings, and add predictions of what is likely to highlight research directions in this field, anticipating the 11th meeting due to be held in Estonia in September 2009.

Integration of Data

Ten years ago, SNPs were the *raison d’être* for the first conference, and a decade of productive research has upheld the belief in the value of these genomic representatives. At the 2006 meeting, the “new kids on the block” were the relatively large structural variants, particularly copy number variants (CNVs), and that same year the meeting title changed to “Human Genome Variation” in order to acknowledge and accommodate the enhanced spectrum of genomic variation. In 2008, there was strong evidence that technologies—both arrays and next-generation sequencing—as well as analytical and annotation methods are evolving to accommodate and integrate the different forms of

Box 1. Plenary sessions

1. Genome Variation and Data Integration
2. Functional Variability: In Search of Plausibility
3. Comparative Genomics and Populations
4. Genome-wide Studies and Disease Susceptibility
5. Structural Variants, SNPs, and Phenotypes
6. Genome Variation: Applications and Ethics
7. Genome Variation Technologies, Tools, and Bioinformatics
8. Population Genetics and Genomics

variation (S. McCarroll, J. Mudge, and T. Manolio). Interspecies comparative approaches using mouse (C. Webber) and canine (K. Lindblad-Toh) disease models are also yielding increased success in characterizing some genes.

A comprehensive genetic/genomic approach revealed that some carriers of single nucleotide mutations in the *TP53* gene in families with Li-Fraumeni syndrome also have increased numbers of genomic CNVs, which may be precursors to somatic changes and tumor formation (D. Malkin). Genomic variation and expression profiles were integrated to detect loci involved in quantitative and complex traits, such as the *DISC1* gene in schizophrenia, exploring regulatory effects on its expression (L. Almasy). SNP arrays populated with probes for CNVs are now commercially available, and integrated data allowed the study of segregating CNVs in HapMap samples by virtue of linkage to SNP haplotypes (S. McCarroll). Meredith Yeager described the need to integrate fine mapping by genomic sequencing for the transition from genome-wide association studies (GWAS) to direct functional studies of candidate genes. Meta-analyses, by pooling data from related studies, provide a means to detect variants that may be rare or have small effects (J. Hirshhorn, M. McCarthy, and J. Rioux), and analytical methods to support such analyses were described, such as imputing missing data (N. Zaitlen). Esteban Parra described how admixture mapping complements GWAS for identification of loci involved in complex traits with population differences. Tuuli Lappalainen and Lars Feuk each drew attention to the need for good comparative studies of the performance of statistical approaches and algorithms for different datasets. Finally, several presenters described advances in building and integrating genomic databases (Table 1), which will be key in the future development of the field.

Location—Location—Location

Genome Variation Analysis Goes Global

This year's proceedings reflected a marked transition to more international studies, tapping global resources for an expanded perspective of human variation. Genomic samples from the HapMap and Human Genome Diversity Projects (HGDP) continue to be mined, and those from the 1000 Genomes Project and European profiles of structural and sequence VARIation of the human genome in DISease (gEUVADIS) Project are highly anticipated. Tuuli Lappalainen presented findings from GWAS of four Northern European countries, focusing on natural selection and loci with interpopulation differences, in search of causes of human disease. Sarah Tishkoff described the step up from targeted genotyping to genome arrays in her work to characterize the substructure of African populations and their relationship to the diaspora. Admixture also characterizes the Hispanic populations of the Americas, and Esteban Parra spoke of building a resource of ancestry-informative markers to facilitate admixture mapping for type 2 diabetes. The Hebrew University

Box 2. Audience response at the public forum: "Personalized Medicine: Are We Ready?"

- "I've been listening to why this might not be a good idea, but is there a real risk?"
—"under 30" law student
- "If a patient walked into my office with one of those print-outs, it would take me much more than 20 minutes to look at it, which is all I have."
—medical student
- "Is it possible to turn this into something the public can actually understand?"
—journalism student

Genetic Resource of well-characterized Ashkenazi case-control samples has been widely utilized for collaborative GWAS of common disorders. Xavier Estivill outlined an ambitious initiative in search of population-specific CNVs, using pools of HapMap and HGDP samples and arrayCGH; results of the pilot study demonstrate clear population clustering, with encouraging correlation to population differences in expression phenotypes. Developing countries have compelling reasons to invest in large-scale genotyping projects, and Billie-Jo Hardy presented case studies in Mexico, India, Thailand, and South Africa undertaken to assess motivations and challenges for studying and managing their own diversity.

Collaboration in the Global Village

Genomics is big science, and two aspects of findings from early GWAS are driving the move to collaborative investigations. One is that many variants relevant to traits of interest are either rare or of small phenotypic effect, and therefore require very large sample sizes for ascertainment. The other is that populations clearly differ genetically, and knowledge of the breadth of such difference is essential before applications such as medical interventions can be considered. The differences are also tools in themselves toward finding underlying genetic bases of human traits. Steve Scherer's opening comment was of "collaboration" and the theme prevailed.

Location in the Genome Matters

The exome has been the focus of attention for early-phase genomics, but variation exists also between exons, near genes, and in nongenic regions. De novo alterations (including somatic) may disrupt the natural environment for functional elements; as pointed out by Dalila Pinto, the number of CNVs (or SNPs for that matter) may not be as important as where they exert their influence. Elaine Mardis asked what important nonexonic changes might be missed in hypothesis-driven surveys of the genome. On this theme, several presentations considered aspects of regulation of gene function by variable elements outside of coding sequences. Eddie Rubin has exploited two characteristics to screen for human enhancer elements: conservation of these nongenic sequences among species, and their tendency to be a site for protein coactivator binding. As many as 5,000 regulatory elements are anticipated, along with the opportunity to investigate variation and its impact. Steven McCarroll described an example whereby a common deletion upstream of the *IRGM* gene, in complete linkage disequilibrium with a SNP that is strongly associated with Crohn disease, appears to be the disease-causing element through its impact on *IRGM* expression and function in autophagy. David Malkin explored the location of CNVs present in excess in *TP53* mutation carriers, and found them predominantly around other cancer-related genes, particularly DNA repair genes. Laura Almasy outlined strategies for exploring the basic biology of upstream and downstream effects by noncoding sequences on variable gene

Table 1. Databases and Catalogs

Abbreviation	Full name	Website	Described by
DECIPHER	Database of Chromosomal Imbalance and Phenotype Using Ensembl Resources	https://decipher.sanger.ac.uk/information	N. Carter
DGV	Database of Genomic Variants	http://projects.tcag.ca/variation	L. Feuk
GEN2PHEN	Genotype to Phenotype	www.gen2phen.org	A. Brookes
HGVbaseG2P	Human Genome Variation Association Database	www.hgvbaseg2p.org	A. Brookes
VEGA	Vertebrate Genome Annotation Database	http://vega.sanger.ac.uk/index.html	J. Mudge
NHGRI GWAS catalog	National Human Genome Research Institute Genome-Wide Association Studies Catalog	www.genome.gov/gwastudies	T. Manolio

expression in complex traits, with a particular focus on psychiatric endophenotypes.

Medical and Social Issues Arising

Genetic Privacy

One major issue garnering attention throughout the conference was the impact of the recent publication [Homer et al., 2008] describing statistical methods to discern the presence of a given individual sample among aggregate genotype data, followed by the action of certain funding agencies to eliminate open access to pooled data from GWAS (S. Chanock). Discussion concerned implications for other databases, such as the HGVbase database of genetic association studies (A. Brookes), and the DECIPHER database, which place individual data (albeit limited to pathological variants) in the public realm (N. Carter). David Cox considered whether either the extreme view expressed in connection with the Personal Genome Project—suggesting that privacy value concerns are outdated—or that reflected by the funders' action, represent broad public sentiment with respect to "trust," and he advocated more thoughtful discussion before further action. In anticipation of such concerns, Anthony Brookes proposed a comprehensive researcher identification system to both track and recognize researcher activities, and enable coordinated control of access to sensitive genotype–phenotype data (more information provided at www.gen2phen.org). Stephen Chanock pointed out that methods to obscure frequency data would rather defeat the purpose of their publication for research, and advocated public discussion as a first step toward a "genomic age."

Genetic Nondiscrimination

The long-awaited Genetic Information Nondiscrimination Act (GINA) was signed into U.S. law in 2008 (D. Cox). In the context of discussion on public representations of personal genomics, Tim Caulfield acknowledged that opportunity for discrimination is the biggest public concern associated with genetic disclosure, but questioned whether such concerns are justified by experience, and, therefore, whether GINA was necessary.

Populations and Personal Genomics

A recent discovery of diabetes susceptibility genes specific to East Asian samples exemplifies the importance of extending GWAS to a wider range of populations (M. McCarthy). The developing world may not benefit from personalized therapeutics designed primarily for populations in wealthier nations (B-J Hardy). Tim Caulfield cited evidence that knowledge of one's personal genotype may have limited impact on preventive behaviors. David Cox pointed out that use of individual genomic variation data (such as from whole-genome sequencing) to

recommend altered clinical care will require prospective clinical trials, and Mary Relling raised the dilemma that funding agencies expect participants in clinical trials to be treated equally, despite the aspiration for more personalized therapies.

For the first time, the conference program included an open public forum, held at Victoria College of the University of Toronto. "Personalized Medicine: Are we ready?" brought together a diverse panel of professionals from business, academic, and hospital sectors to present views on the issue of direct-to-consumer genetic testing, followed by questions from an audience of about 500. A videocast of the session is freely available at www.tcag.ca/hgv2008. Two companies that offer such testing were represented (23andme and Navigenics), advocating the opportunity to provide information to a curious public; the remaining panelists challenged with wide-ranging concerns, particularly around the concept of risk. When polled about whether they would undertake the test for themselves if it were "free," 1 out of 7 panelists said "yes." A different age demographic and attitude from among the audience was made evident by their questions such as in Box 2. When polled with the same question about taking the test, about 75% raised their hands. The moderator, Steve Scherer, noted that such forums in various host countries would reflect issues specific to the local health care system (which in Canada provides universal access) and global variation in public interests, questions, and concerns.

Noted Trends and New Developments

Ascendancy of Genome-Wide Analysis

As hinted at in the 2007 conference proceedings [Estivill et al., 2008], GWAS have yielded dramatic increases in documented associations in the past 2 years. Teri Manolio provided an overview from the NHGRI catalog of 180 publications (to September 2008), and notable examples were cited by John Rioux for inflammatory bowel disease (at 32 risk factor genes) and autoimmune disease (68 associated risk variants), Mark McCarthy for type 2 diabetes (at 18 associated loci), and Joel Hirschhorn for height (43 loci) and body mass index (8 loci). Ariel Darvasi noted that despite the impressive yield of approximately 400 novel (and often robust) SNP associations with complex traits, these still explain little of the genetic variation. Meredith Yeager asked, once we find these associations, then what?, reminding the audience that correlation does not mean causation, and fine mapping and functional studies are essential follow-ups.

Not only SNP genotyping, but sequencing of individuals has gone genome-wide, and the two first examples—Craig Venter and James Watson—were described by Pauline Ng and Michael Egholm, respectively, with additional insights from Len Pennachio. Elaine Mardis described applications of next-generation sequencing in cancer diagnostics and asked whether sequencing will soon replace microarrays. She described the first cancer genome sequence: an acute myeloid leukemia genome and its

matched normal (skin) counterpart. Edison Liu speculated that array technologies may be reaching their limits of detection, and that pair-end sequencing strategies, which detect structural and copy number variants, will be ready for association studies within two years. Several noted that as generation and accumulation of data become increasingly efficient, interpretation of such data remains rate-limiting, but essential.

Matters of Size

Large studies reveal large numbers of variants with significant but modest effect (J. Hirschhorn and M. McCarthy). Although initially discouraging, these small effects may, nonetheless, draw attention to regions and pathways of biological relevance or identify drug targets (J. Hirschhorn and J. Rioux).

Known CNVs are becoming more numerous and smaller. Genome-wide structural variation, first reported in 2004, has quickly become established in the variation spectrum. Lars Feuk described clear improvements in resolution of CNVs with tools (including sequencing) targeted for their detection, noting that enhanced resolution has created an apparent shift downward in size, and that there has been an explosion in reports of variants in the range of 100 bp to 1 kb (i.e. smaller than the operational definition of CNV). From some 30,000 structural and copy number variants already deposited in the Database of Genomic Variants (DGV), he noted that ~0.7% of the genome varies in copy number between two diploid genomes, affecting a significantly larger fraction of the genome than other types of variation (such as SNPs).

Toward Hypothesis-Free Analysis

The scope of methodologies is starting to allow more screening approaches that detect variation without prior hypotheses, although some studies advocate hybrid panels combining candidate gene regions with untargeted genome-wide probes. Three SNPs in one GWAS experiment drew attention to an otherwise low-priority candidate gene for influence on drug clearance in treatment of acute lymphoblastic leukemia, and this gene is now recognized as an important metabolic determinant with pharmacogenomic implications (M. Relling). Len Pennacchio commented that the classical genetics paradigm is shifting toward ascertainment of genetic variants first, and then the associated traits. Eventually, whole-genome sequencing will be the means to unbiased detection of variation (M. Egholm and P. Ng).

Opportunities for Individualization

The whole-genome sequence of an individual will provide the ultimate opportunity for specific and appropriate health interventions; however, the first such initiatives, as described in the past 12 months, illustrate that we are far from being able to interpret usefully the wealth of genomic variation that is uncovered. In the meantime, studies of world population variation (such as described above) are beginning to enhance opportunities for more tailored approaches to health, using knowledge of group differences. Pharmacogenomics is at the forefront of individual-targeted care according to genotype, and examples from cancer therapies were provided by Mary Relling.

Common and Rare Variants

Classical genetics first dealt most readily with the “sports of nature”—relatively rare genetic mutations of independent nature

with strong phenotypic effect. Great anticipation from the Human Genome and HapMap Projects was for the opportunity to study common variation and its impact on common disease; the mushrooming accumulation of recent data from GWAS is its outcome. Now, larger studies and meta-analyses are needed to detect the progressively rarer residual variation, but with new sequencing strategies for fine mapping, focus is returning to low-frequency and rare variants (P. Ng and M. McCarthy). David Cox noted that rare variants are more likely to yield novel therapeutics.

New Methodologies

Technical improvements abound among array technologies for genotyping, comparative hybridization, and sequencing, with analysis of data scrambling to keep pace with its generation. Keynote speaker Svante Pääbo described the exciting application of new (454) sequencing technology to ancient genomes, particularly of Neandertal. Overcoming hurdles associated with massive degradation and contamination, his team reported completion of a mitochondrial sequence from the Vindija site in August 2008, and anticipated one-fold coverage of the nuclear sequence by the end of 2008 (15-fold coverage within a few years). They will most likely not be able to address the really interesting questions about genome evolution, divergence, and selection because of small sample sizes.

The future for visualization and analysis of individual DNA molecules was foreshadowed with two related talks. Pui Kwok argued the need for better means to assay structural variation such as inversion or duplication; for example, with single molecules that retain allelic contiguity. He described efficient methods for DNA barcoding with allele-specific or nick-labeling, to create color-coded linear patterns suitable for automated analysis. Han Cao then demonstrated technology to linearize DNA fragments of any size in nanochannel arrays, with dynamic analysis of single molecules and applications such as barcode analysis, fragment sizing, sequencing, or protein interactions. There was some discussion of the recent press releases from Complete Genomics and their purported \$10,000 genome sequence, and the participants were intrigued but also wanted to see the proof.

Among the computational strategies discussed were several that related to pooling. Itsik Pe'er described resequencing of overlapping pools of individuals to efficiently discover rare variants while retaining the means to identify those who are the source of variation. In the search for population-specific CNVs, Xavier Estivill reported pooling samples from each group to remove the effects of individual variation, followed by arrayCGH between pools and principal component analysis of the CNVs. Natural pooling occurs in admixed populations, and can be exploited to map variants for complex traits that show population differences (E. Parra).

When data, rather than samples, are combined, such as for meta-analysis, strategies such as imputation may be necessary (N. Zaitlen). Rather than pooling, Gonçalo Abecasis discussed a strategy to increase efficiency of sequencing large numbers of individuals by combining data from those with some shared haplotypes in order to impute missing genotypes.

Things to Come

Steve Scherer summarized the meeting as one with many converging themes around “Genomic Variation in a Global Village.” The concept seems fitting, given the transformative impact of genetics/genomics and the Internet on society, and that the

expression "global village" was coined by the University of Toronto philosopher Marshall McLuhan. The 11th Human Genome Variation meeting will likely extend this idea, with themes such as: enumeration and influence of common CNVs; sequence data from more individuals and more of the genome; interactions among genes, variants, regulators, and environmental influences described as networks; effective ways to describe and database phenotypic variables; variation in biomolecules and their expression; as well as other topics sure to emerge in upcoming months.

Acknowledgments

We thank Dr. Janet Buchanan for help with preparation of the conference report. The meeting was supported by The Centre for Applied Genomics, the Canadian Institute for Advanced Research, the Ontario Ministry of

Research and Innovation, Genome Canada, the Canadian Institutes of Health Research, McLaughlin Centre for Molecular Medicine, Ontario Institute of Cancer Research, National Human Genome Research Institute, Affymetrix, Illumina, Agilent Technologies, Oxford Gene Technology, Applied Biosystems, and Roche Pharmaceuticals.

References

- Estivill X, Cox NJ, Chanock SJ, Kwok PY, Scherer SW, Brookes AJ. 2008. SNPs meet CNVs in genome-wide association studies: HGV2007 meeting report. *PLoS Genet* 4:e1000068.
- Homer N, Szelling S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 4:e1000167.