Genome **Biology**

MEETING REPORT

Human genetics in full resolution

Alicia Schiavi, Nicholas Light, Veronique Adoue, Stephan Busche and Tomi Pastinen*

Abstract

A report on the 12th International Congress of Human Genetics, joint with the 61st annual American Society of Human Genetics conference, Montreal, Quebec, 11-15 October 2011.

The 12th International Congress of Human Genetics (http://www.ichg2011.org/), along with the 61st annual American Society of Human Genetics conference, was one of the largest gatherings yet of human geneticists, with 7,500 attendees from more than 85 countries worldwide, including over 120 invited presentations and 3,800 contributed presentations.

Compared with earlier meetings, a notable difference was the full penetration of next-generation sequencing (NGS) in studies of Mendelian disease, complex traits and functional genomics. The progress in 'production scale' NGS has generated some exciting data in all fields of human genomics, and we highlight here a few of the many insightful presentations in key topic areas.

Cataloging Mendelian disease

With progress in exome sequencing and emerging whole genome sequencing studies, it is a mere matter of time before all monogenic disease variants are discovered. An exciting disease study was presented by Leslie Biesecker (National Human Genome Research Institute, National Institutes of Health, Bethesda, USA), whose group identified a mosaic activating mutation in *AKT1* as the molecular basis of Proteus syndrome using massively parallel exome sequencing of affected tissues. Proteus syndrome has been proposed to be the condition that affected the Elephant Man, Joseph Merrick. Novel follow-up studies revealed that these mutations in *Akt1* in mice manifested a phenotype of leukemia. Segolene Ayme (INSERM, Paris, France), part of the Rare Disease Task Force established by the European Commission

Public Health Directorate, discussed the expectation that through the work of a number of different consortiums the genetic cause of approximately 10,000 Mendelian diseases will be understood by 2020. Advances in establishing these projects are allowing a shift of the volume of sequencing resources onto the intricacies of complex diseases.

Unraveling the missing heritability in complex disease studies

Over 1,500 genome-wide association studies (GWASs) for complex traits have been reported thus far. Trey Ideker (University of California, San Diego, USA) described the current problems facing GWASs, which include insufficient statistical power, missing biological integration, and the effect on complex traits of the many genetic interactions between markers that GWASs do not detect. A common thread throughout dealt with the question of integrating different kinds of high-throughput data. Although many trait-associated single nucleotide polymorphisms (SNPs) are expression quantitative trait loci (eQTLs), the integration of genomic data with other types of biological data to better decipher complex traits and diseases remains a challenge.

Many presenters proposed different approaches using combined data to focus on functional sites only. Nancy Cox (University of Chicago, USA) demonstrated an approach of not looking at all genetic variants independently, but instead using functional unit analysis to improve power with respect to multiple testing bias by focusing on the level of the gene (tens of thousands) rather than the level of the SNP (millions). Use of this analysis on pharmacogenetic studies of cancer chemotherapy drugs yielded genome-wide significant genes, four associated with both cisplatin and carboplatin and 11 associated with etoposide and pemetrexed. Moreover, *HSPG2* (encoding heparan sulfate proteoglycan 2) was found to show a genome-wide significant association with many drugs. Ideker, meanwhile, proposed to combine GWAS data with protein networks to give a higher-level map of genetic interactions, which can help to find the right interactions. He also proposed future protein-network-based diagnostics.

John Stamatoyannopoulos (University of Washington, St Louis, USA) highlighted DNase I hypersensitive sites

*Correspondence: tomi.pastinen@mcgill.ca
Department of Human Genetics, McGill University and Genome Quebec Innovation Centre, 740 Dr. Penfield Avenue, Montreal, Quebec H3A 1A4, Canada

BioMed Central

(DHSs) as strong predictors of regulatory DNA across over 100 datasets from different human cells and tissues and the significant overlap between GWAS variants and DHSs. These phenotypically interesting sites also show imbalanced allelic states, reinforcing their functionality. Furthermore, DHSs represent the cumulative action of several transcription factors and show extraordinary lineage specificity. Stamatoyannopoulos showed the detection of evolutionarily conserved early developmental enhancers by looking at DHSs in adult cells. Jacob Degner (University of Chicago, USA) used a combination of DHSs with RNA sequencing data to link gene expression levels with SNPs. He presented correlation between DHS read depth and nearby variants called chromatin accessibility quantitative trait loci (caQTLs). He showed allele specificity for both DNase I sensitivity and transcription factor binding at caQTLs. Finally, Degner pointed out a link between gene expression and caQTLs, of which over a third also alter gene expression in the population (eQTLs).

Until recently, research has focused on understanding complex diseases by uncovering common variants in individuals and within populations. However, as discussed by various presenters, the advent of large-scale projects also aimed at uncovering rare variants. The reasons for studying rare variants were outlined by Goncalo Abecasis (University of Michigan, Ann Arbor, USA): this will complete the genetic architecture for complex diseases and aid in understanding the functional linkage of a locus to a trait, such as for rare segregating alleles in a population. Rare variants, compared with common variants, have a larger effect size in associations with disease. Abecasis used low density lipoprotein (LDL) levels, which are consistently associated with risk for heart disease, to better understand quantitative trait genetics. Abecasis' group used whole-genome sequencing on a Sardinian population to identify an LDL population-specific association with the β-hemoglobin (*HBB*) locus, as well as validating previously known genes. Shaun Purcell (Massachusetts General Hospital, Boston, USA) addressed the need to study rare variants to understand schizophrenia and other complex diseases through methods including gene-based rare variant tests. Rare variants with a minor allele frequency (MAF) of less than 5% may contribute up to 50% of variation. Purcell performed a large study involving exome sequencing on 1,024 samples combined with GWAS for 5,603 samples and found 240,000 coding SNPs with a MAF of 0.035%. So far, none of the coding SNPs showed experiment-wide significance.

Various presenters at the meeting discussed the study of copy number variation (CNV) to understand complex diseases, such as, but not limited to, schizophrenia, autism spectrum disorder (ASD) and intellectual disability (ID).

Stephen Scherer (Hospital for Sick Children, Toronto, Canada) investigated whether similar genes are being identified independently in ASD and ID studies. He presented a combined approach with genome scanning using high-density microarrays with NGS for CNV dissection. Scherer observed an enrichment of genetic CNV deletions (2.1-fold), compared with CNV duplications, in the combined ASD and ID loci cases versus controls. Moreover, rare CNVs tended to cluster into two major networks, cell projection/motility and kinase networks. Although the majority of CNVs so far identified are likely to be deleterious, Wenli Gu (Baylor College of Medicine, Houston, USA) described one example of a duplication CNV, Dp(11)17, that has been shown to be protective against metabolic syndrome in humans and mice.

Cancer genomics

The high intensity of cancer genomics was evident at the conference, cancer being a keyword in approximately 20% of presentations. Thomas Hudson (Ontario Institute for Cancer Research, Toronto, Canada) introduced the International Cancer Genome Consortium (ICGC), which aims to coordinate the systematic epigenomic, genomic and transcriptomic analysis of 25,000 cancer specimens derived from 50 different cancer types or subtypes. Stacey Gabriel (Broad Institute, Cambridge, USA), who represented The Cancer Genome Atlas (TCGA) program of the USA, which is involved in the analysis of 25 forms of cancer, reported new biological insights into cancer diversity, with somatic point mutations and other genomic variations shown to differ 1,000-fold between cancers. Whereas glioblastomas harbor relatively few copy number variants, ovarian carcinomas seem to be characterized by a multitude of variants, with the tumor suppressor gene *TP53* altered in almost every sample. Gabriel also revealed some newly identified targets: the tumor suppressor gene *TP63* for squamous cell lung carcinoma and *IGF2* (encoding insulin-like growth factor 2) for colon carcinoma, with the latter cancer being analyzed in depth in an upcoming publication.

For pancreatic cancer, which is historically difficult to characterize owing to specimen impurity, ultra-deep sequencing seems to be the way to go. Gabriel and colleagues carried out exome sequencing with 300x coverage to differentiate artifacts from true variants. John McPherson (Ontario Institute for Cancer Research) reported that exome and whole-genome sequencing of 113 pancreatic cancer samples carried out so far identified the oncogene *KRAS* to be mutated in most of the tumors. Youyong Lu (Beijing Institute of Cancer Research, Beijing, China), representing the stomach cancer genome project, reported the finding of a novel point mutation in *MUC17* (encoding a mucin) that correlates with the

clinical outcome of gastric cancer. Michael R Stratton (Wellcome Trust Sanger Institute, Hinxton, UK) used NGS approaches to investigate the evolution of the cancer genome. Stratton estimated that five to seven driver mutations are needed for solid cancer initiation and maybe fewer for hematological cancer initiation, and these are accompanied by tens to hundreds of thousands of passenger mutations. Stratton pointed out that the currently known human oncogenic driver mutations are located in 416 genes, and 380 of the mutations appeared only somatically. Furthermore, 59 of the mutations seem to be recessive, 357 dominant. In an elegant experiment that involved the exome sequencing of a primary breast cancer tumor and tumor-derived liver metastases of the same patient, Stratton showed that subsequent liver metastases derive from spreading of the primary metastasis rather than from independent primary tumor-derived cells, and that the comparison of individual metastasis mutation profiles enabled the generation of a metastasis pedigree.

Large-scale studies: insight into the past and future

The steadily dropping price of DNA sequencing is opening up possibilities for large-scale population studies using NGS technology that were previously impractical. Novel studies using NGS on modern populations to reveal ancient admixture, relevant to understanding human evolution and disease, were an exciting development at the conference. Jake Byrnes (Wellcome Trust Sanger Institute) presented his center's research in collaboration with the University of Puerto Rico on the first known reconstruction of a genome of a now extinct human population, the Taino, using the 1000 Genomes Project sequencing data of trios from a modern admixed population of Puerto Ricans. Another interesting talk on population admixture was delivered by Sriram Sankararaman (Broad Institute), whose laboratory used 1000 Genomes Project data to compute the average linkage disequilibrium between SNP pairs over a range of genomic lengths, to narrow down the true date of ancient admixture between modern humans and Neanderthals to 37,000 to 86,000 years before present, concurrent with when modern humans likely encountered Neanderthals in western Eurasia.

The 1000 Genomes Project itself presented the conclusion of phase 1, in conjunction with the release of an updated integrated variant set (Gil McVean, University of Oxford, UK). The consortium has currently analyzed the genomes of 1,092 individuals from 14 populations, through low coverage whole-genome sequencing and deep exome sequencing, resulting in the identification of over 40 million variants, which catalog genetic variants down to below the 1% frequency range. Phases 2 and 3 of

the project will see an expansion of the sample size to around 2,500, to further probe rare variants and increase the sample diversity. Despite the efforts of larger consortiums such as the 1000 Genomes Project to sample a range of populations, more diversity in population genetics studies at large is still needed. Sameer Soi (University of Pennsylvania, Philadelphia, USA) pointed out that although Africa is by far the most genetically diverse continent, it is among the most widely underrepresented groups when it comes to genome-wide studies. The importance of this discrepancy was made clear by Rong Chen (Stanford University, Stanford, USA) in his talk on type 2 diabetes in European, African and East-Asian cohorts, which demonstrated the extreme differentiation in disease risk alleles between human populations. Elinor Karlsson (Harvard University, Cambridge, USA) discussed her group's research on cholera susceptibility in Bangladesh, using a novel computational approach in combination with public datasets to generate significant hits from a small sample size. Karlsson's group genotyped 36 Bengali trios using 1-million-SNP arrays, with follow-up using the composite of multiple signals method, to identify 322 signals of natural selection. These signals were analysed with INRICH, a new pathway analysis tool that detects enriched gene sets, enabling the identification of a module of co-expressed genes linked to *IKBKG*, encoding a modulator of NF-κB.

Given the rapid advances in disease variant identification and the falling costs of sequencing, the utility of large-scale whole-genome sequencing studies for rare variant discovery, and even as a clinical tool, is becoming increasingly plausible. Perhaps one of the more ambitious projects discussed at the conference was presented by James Lupski (Baylor College of Medicine), who spoke briefly on the recently announced FarGen project, an initiative that will provide voluntary whole-genome sequencing to all 50,000 residents of the Faroe Islands. The data will be accessible to other researchers via an online database but, perhaps more importantly, the study could serve as a potential model for whole-genome sequencing initiatives worldwide. The possibility of future genetic screening, as in the FarGen project, using NGS in clinical settings seems to be a major milestone in the field; however, it was not without its critics.

Future challenges

Breakthroughs in NGS-based human genomics were evident throughout the meeting. How will NGS alter public health? A panel discussion on the subject exposed a range of potential ethical and practical issues, ranging from poor availability of genetic counseling services to possible misuse of genetic information. These key issues

will need to be addressed before whole-genome sequencing goes mainstream, but nevertheless the question seems to be more one of when than if.