Meeting report

SNPing variation from genomes

David A Liberles

Address: Department of Biochemistry and Biophysics and Stockholm Bioinformatics Center, Stockholm University, 10691 Stockholm, Sweden. E-mail: liberles@sbc.su.se

Published: 12 December 2001

Genome ***Biology*** 2001, **3(1)**:reports4001.1–4001.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/3/1/reports/4001

A report on the fourth International Meeting on Single Nucleotide Polymorphisms and Complex Genome Analysis, Stockholm, Sweden, 10-13 October 2001.

In the shadow of the attack on the World Trade Center in New York, the scientific community began the process of returning to normality although many American speakers canceled their presentations or delivered them by phone. The attack also led to altered content of the meeting program. For example, Bernadette Modell (Royal Free and University College Medical School, London, UK) spoke on the importance of genetic screening to prevent congenital disorders in the Islamic Middle East, where consanguineous marriages represent 15-40% of all marriages. She recommended an increase in western funding to support such screening in these countries, which are among both the wealthiest and the poorest in the world. In no talk was the effect of the September 11 attack more apparent than in that of Craig Venter (Celera Genomics, Rockville, USA). Venter returned to the new dynamic in world politics many times during his talk. He spoke of the importance of Celera's ability to sequence a bacterial genome in one day as an important anti-terrorism tool, of Celera's involvement in using sequence-based methods to help with identification of the victims of the attack, and finished with an extremely emotive photo display of the aftermath of the hijacking and World Trade Center collapse.

Despite these recent disruptions to normal work, Venter also spoke on Celera's continuing efforts in genomics, including the work on the *Anopheles gambiae* mosquito genome and the mouse genome. Starting with the unexpectedly low number of genes in the draft-sequenced human genome that, he speculated, arose from previous underestimation of the number of low-gene-density deserts when the initial extrapolation estimates of gene numbers were made. He did identify four major gene groups in which expansion of gene number has occurred - those involved in the immune system, hemostasis, signal transduction, and nucleic-acid-binding or transcription factors. Many of these large-scale gene duplications predated human existence, so a comparative approach for analysis of them is needed. A comparison of genes in the human and mouse genomes (based on Venter's presentation) can be seen in Table 1.

A second theme introduced by Venter was that of single nucleotide polymorphisms (SNPs). Although many of the SNPs within the human population do not appear to have significant biological effects, some do. For example, CCR5 polymorphisms were apparently selected during the Black Death that occurred 700 years ago in Europe, with the result that 9% of Europeans are immune to infection through CCR5 by agents including HIV, compared to a background of less than 0.1% of Africans. Venter went on to emphasize that very few polymorphisms have absolutely deterministic effects; many more exert their effects in a probabilistic way, with important influences from the environment as well.

The second keynote address, delivered by phone, came from Eric Lander (Whitehead Institute, Cambridge, USA), who presented his vision of human population structure and history. The human population is small, but growing fast, from an effective population size 3,000 generations ago of

Table 1

A comparison of the similarity of the human and mouse genomes, as presented by Craig Venter

Organism	Approximate number of high confidence genes	Percent homologs
New human sequence assembly	25,000	94.5
Mouse	23,000	96.6

The percent homologs column indicates the percentage of human genes that have homologs in the mouse genome and vice versa.

some 10,000 individuals living in Africa. Although examples of disease alleles are found at both high frequency (for example the respiratory disease cystic fibrosis) and low frequency (for example Wilson's disease, a disease of copper metabolism), Lander believes that most variations occur at high frequency because of the population expansion from a bottleneck that included only relatively few individuals and so had a simple allelic spectrum and that that is also true of most common disease alleles (common disease-common allele association model). Searching for the places that disease originated in ancestral haplotypes (using the phenomenon of linkage disequilibrium or LD, which identifies co-inherited portions of the genome) is facilitated in populations that have been isolated or have gone through recent bottlenecks, such as the Finnish population (isolated for approximately 100 generations). In general, 'old' loci have more polymorphism and less LD, while newer loci have less polymorphism and longer stretches of sequence correlation between alleles. Finally, Lander described using haplotypes to map the intestinal inflammatory condition Crohn's Disease, leading to the identification of five candidate genes that must now be explored biologically.

Many of the subsequent lectures fell into distinct categories, touching on subjects raised in the keynote addresses, including SNPs in genomes and populations, disease-association studies, technology, and evolution.

SNPs in genomes and populations

Two talks were critical of the model for disease gene association through linkage disequilibrium presented by Lander. Andrew Clark (Pennsylvania State University, University Park, USA) examined the idea of common disease-common allele correlation by examining population structure. Using the ancestral population size of 10,000 before a population expansion, one would expect 8 rather than the observed 1.1 alleles per gene proposed by Lander. Furthermore, the human population according to his model had 391,504 total generations before the expansion and 1.3×10^{12} generations after the expansion, implying that much more than 99% of all mutations have occurred since the population expansion and, by extension, most disease-causing mutations are likely to be rare.

Joseph Terwilliger (Columbia University, New York, USA) gave a talk entitled "The return of the evil mutant association study - (just when you thought it was safe...)." Starting with the question "Is haplotype mapping important?" he went on to stress that gene-phenotype relationships are more important than disease marker-gene marker (linkage) relationships. He emphasized that to ensure that the same phenotype always reflects the same genotype, it is best to examine genetically related individuals, and improved sequencing technology may make haplotype maps irrelevant. He further brought home the point that traits are not diseases and that multiple traits can affect disease, where each trait may be affected by environmental factors and by multiple genes. This talk inspired considerable audience discussion.

Many talks focused on the collections of SNPs now available. Lincoln Stein (Cold Spring Harbor Laboratory, USA), Stephen Sherry (National Center for Biotechnology Information (NCBI), Bethesda, USA), and David Fredman (Karolinska Institute, Stockholm, Sweden) described publicly available collections of SNPs (see NCBI-dbSNP [http://www.ncbi.nlm.nih.gov/SNP] and HGBASE [http://hgvbase.cgb.ki.se]). At present, 2.2 million quality-controlled SNPs are available, including 33,405 exonic SNPs, approximately one-third of which are nonsynonymous. Of these nonsynonymous SNPs, 1,333 can be mapped onto tertiary protein structures. Another 130 SNPs have been traced to mRNA splice sites. It appears that 77% of SNPs are informative in at least one population, 40% in a given population, and 25% in all three populations (Caucasians, Africans, and Asians). SNPs are also available for species other than humans, including chimpanzee. James Weber (Marshfield Medical Research Foundation, Marshfield, USA) added that 76% of polymorphisms are substitutions, while 22% were insertion or deletion (indel) events, and 2% involved other events.

Kenneth Kidd (Yale University, New Haven, USA) presented ALFRED [http://alfred.med.yale.edu/alfred/index.asp], his database of the frequencies of a large collection of SNPs and haplotypes across populations. He found that some loci show geographic distributions that mirror known populations in a cline, while others do not. From the database, very little LD was seen in sub-Saharan African populations, more in northeast Africa, and increasing amounts in Europe, Asia, and the Americas. He theorized that a significant distribution of polymorphism existed in the original African population, but only a fraction of it migrated out of Africa. Such an understanding of population structures may have medically important implications, such as in pharmacogenomics. Kidd, as well as Kristin Ardlie (Genomics Collaborative Inc, Cambridge, USA) and Gabor Marth (National Center for Biotechnology Information (NCBI), Bethesda, USA), described the mixed nature of the African-American population, which is now thought to include a 25% admixture of Caucasian haplotypes.

Leena Peltonen (University of California, Los Angeles, USA) described the structure of the Finnish population in more detail. Finland underwent rapid population expansion 900 years ago, but has seen migration to unpopulated parts of the country (including eastern populations) only since 1750. She has developed specialized Finnish microarrays to screen 2,400 Finns for 31 disease mutations that are found in Finland, finding that about one-third of Finns carry one of the mutations. SNP frequencies can vary between the different regions of Finland because of migration bottlenecks. Another interesting point that emerged in the discussion was that of the history of lactose intolerance. From gene

sequencing in the baboon, lactose intolerance appears to be the ancestral state, with lactose tolerance having arisen in the human population before migrations out of Africa and now being seen in populations worldwide.

Disease association studies

Jeffrey Long (University of Michigan, Ann Arbor, USA) analyzed the genetics of alcohol metabolism in Asian populations. Two genes involved in this pathway, alcohol dehydrogenase (*ADH*) and acetaldehyde dehydrogenase 2 (*ALDH2*) show deficiency alleles with high frequencies involving several haplotypes that do not fit with the geographical distributions and known patterns of migration of different populations. This may be evidence for selection on these alleles, possibly related to the alleles providing resistance to parasites such as malaria.

Ariel Darvasi (Hebrew University and IDgene Pharmaceuticals Ltd., Jerusalem, Israel) presented data on LD in Finnish, Sardinian, and Ashkenazi Jewish populations. Retitling his talk, "The return of the good old mutant-association study - (just when you thought it would never work...)," Darvasi showed an interesting association in Ashkenazi populations between schizophrenia and the catechol-*O*-methyltransferase (*COMT*) gene that inactivates estradiol metabolites.

Gilles Thomas (Fondation Jean Dausset CEPH, Paris, France) found an association between the candidate *CARD15* gene and Crohn's Disease through a positional cloning approach. *CARD15* activates NFκB and is expressed in monocytes, consistent with molecular and clinical studies including regression after bone marrow transplantation. David Weiner (University of California and Acadia Pharmaceuticals, San Diego, USA) analyzed genetic variation in drug-target genes. Focusing on G-protein-coupled receptors, *in vitro* analysis of SNP-encoded phenotypes showed that agonists can effect different variants of G-protein-coupled receptors dissimilarly.

Finally, Claes Wahlestadt (Karolinska Institute, Stockholm, Sweden) and Alan Schafer (Incyte Genomics Ltd., Cambridge, UK) examined human obesity and type 2 diabetes, respectively. Wahlestadt identified the neuropeptide Y gene as having a major effect in appetite control, consistent with a known role for the encoded peptide in appetite regulation. Schafer used a candidate-gene-based approach to screen polymorphisms in approximately 250 candidate genes from known functional roles, identifying many rare variants, including frame shifts. Type 2 diabetes appears to be a multigenic trait, with risk factors from obesity as well as directly causative mutations.

Technology

Advances in SNP-related technology gave a glimpse of the future of research into genome variation. Itsik Pe'er (Tel Aviv University, Israel) presented an approach for computational resequencing, to search for SNPs using DNA microarrays. This approach, which allows for detection of indel events, can be useful in analyzing tumors, new individuals in a population, or a new species, to detect mutations. In personal discussions, he indicated that the approach can be extended to interspecific mRNA expression levels using a Bayesian framework of statistical analysis based upon an explicit model of evolution.

Another impressive technology lecture was presented by Colin Barnes (Solexa Ltd., Essex, UK). He presented the use of a new type of chemistry to chip-sequence a genome using 20-30 cycles of polymerization with fluorescently labeled nucleotides on a large number of parallel single DNA molecules. This technology has the potential to significantly increase the speed, and decrease the cost, of genome sequencing and resequencing efforts of new individuals in the human population and for other species as well.

Evolution

Polymorphisms also provide insights into evolution at a molecular level, as illustrated by a number of speakers. Justin Fay (University of Chicago, USA) examined a large number (183) of genes in *Drosophila* and in humans using the McDonald-Kreitman test (comparing nonsynonymous and synonymous polymorphisms with nonsynonymous and synonymous substitutions). Here, nonsynonymous changes alter the encoded amino acid while synonymous changes do not; substitutions represent changes that have been differentially fixed between species. He found evidence for many sites under negative-selective pressure, with some under positive-selective pressure, a finding not consistent with the neutral theory of molecular evolution, where most substitutions have either a deleterious or neutral effect on the protein. Clay Stephens (presenting by phone; Genaissance Pharmaceuticals, New Haven, USA) calculated Tajima's D statistic (which compares the number of pairwise fixed differences between species with the number of intraspecific polymorphic sites) on human polymorphisms with chimpanzee as an outgroup, finding that 90% have values less than 0 (reflecting negative or conservative selective pressures), while several have very positive values (indicating selective pressures for a change of protein function).

Thomas Mitchell-Olds (Max-Planck Institute of Chemical Ecology, Jena, Germany) also used the Tajima D statistic to analyze *Arabidopsis* genes. He found some positive values against a mean negative value using this measure, which he interpreted to be reflective of population expansion. Turning to selection for resistance to herbivory through the production of insecticidal glucosinolates by the myrosinase enzyme, he examined the *man1* locus with positive Tajima's D and the *man2* locus with negative values. He hypothesized that selection pressure on myrosinase was a balance between

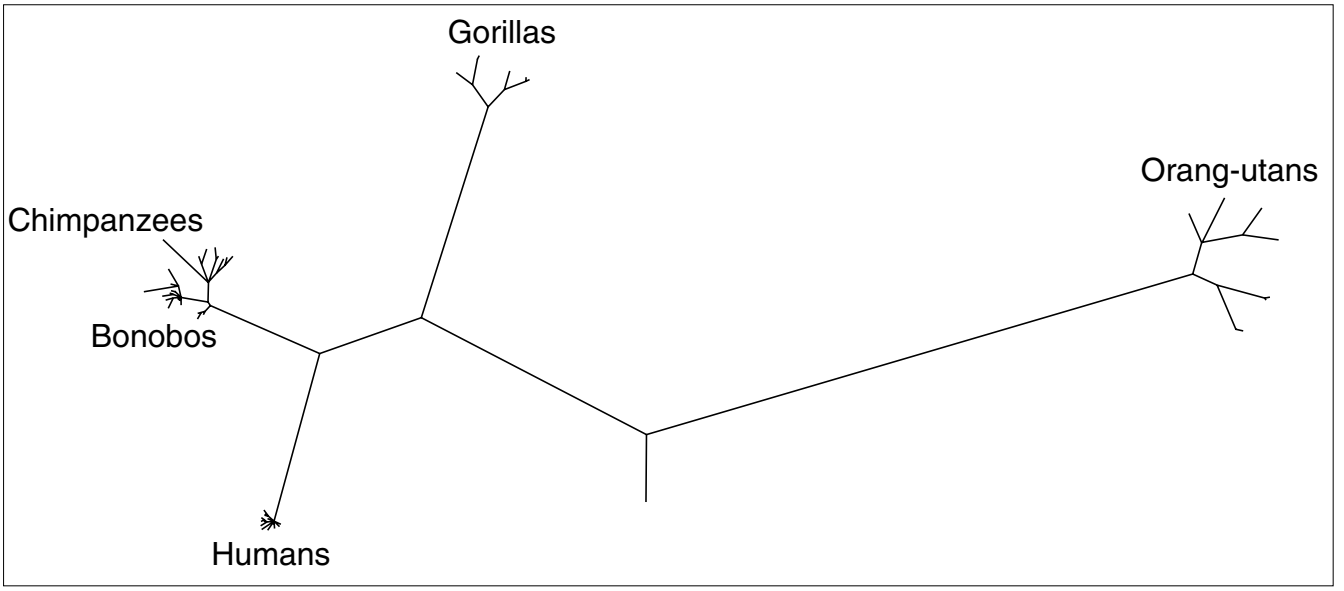

Figure 1
Maximum likelihood phylogenetic tree of human, chimpanzee, bonobo, gorilla and orang-utan Xq13.3 sequences. Humans show much less diversity than other ape species. Reproduced with permission from Kaessmann *et al.*, *Nat Genet* 2001, **27:**155-156.

defense against generalist herbivores, which are repelled by glucosinolates, and specialist herbivores, which are attracted by them.

The last talk of the conference was presented by Svante Pääbo (Max-Planck Institute for Evolutionary Anthropology, Leipzig, Germany). By comparing human and chimpanzee sequences, he found a 1.3% divergence that was not equally distributed among chromosomes, with the most on the Y and the least on the X. Many of the differences are found as transitions at CpG sites. He also noted that there is between three and four times more variation in chimpanzee populations than in human ones. Western chimpanzees have much less variation than other chimpanzee populations, but are not monophyletic. Figure 1 shows the relative divergence of great ape species. To understand the differences between humans and other apes, Pääbo compared gene expression levels for a set of genes in both the prefrontal cortex of the brain and the liver as well as coding sequence differences in specific genes, for example the gene encoding the FOXP2 putative transcription factor associated with language. After identifying many good candidate genes, the need for appropriate biological assays has emerged.

As genetic studies and evolutionary studies come together with functional studies, our understanding of the functioning (and malfunctioning) of the human genome and the selective pressures on it will be SNPed from the realms of mystery.

Acknowledgements
I am grateful to Jennifer Lee and Arno Liberles for careful reading of this review, and to Svante Pääbo for kindly providing Figure 1.