

## Technophiles seek genomic imperfections with the Greek gods at Atlantis

Hamish S Scott

**The 8<sup>th</sup> International Symposium on Mutations in the Genome covered a broad range of technological developments and their applications to detecting genomic sequence variation.**

This year, the Human Genome Organization (HUGO) sponsored a "Mutation Detection" meeting from 31 May to 4 June on the Cycladic island of Santorini, Greece. The Cyclades owe their name to the circle they form around the sacred isle of Delos, the birth-place of the Greek gods Apollo and Artemis. This circle form is due to a series of volcanic eruptions, several of which were centered at the now scenic and peaceful Santorini. Many believe that one such eruption was responsible for the myth of Atlantis, and that Santorini is in fact Atlantis.

The 220 registrants came from 35 countries and five continents; only Antarctica and South America were not represented. The conference, organized by Richard Cotton (Australia), Aglaia Athanassiadou (Greece), Johan den Dunnen (The Netherlands) and Ann-Christine Syvanen (Sweden), was booked 3 months in advance, thanks to not only the location but also the program. The meeting's multidisciplinary nature was reflected in the presentations by human, plant and animal geneticists; clinicians; diagnostic laboratory scientists; chemists; bioinformaticians; and engineers. The technology-based meeting also had a strong industry presence.

Meeting participants reported on the application and evolution of existing DNA technologies, as well as the development of new DNA technologies, for the detection, analysis and documentation of genetic variation. All current pathways to identify genes contributing to mendelian and complex traits eventually converge

on sequence variant testing in candidate genes. But the rapid advances and diversification in DNA technologies and genome sciences, along with the heterogeneity of the problems being studied, mean that there is no unique pathway to success and there are a number of intellectual and physical dichotomies among which the investigator must choose. Though faced with even more options, at least conference participants were left well-informed. Here, I discuss some of the main themes presented.

### Mutation detection

For those searching for new sequence variants, the conference satchel included a pleasant surprise: an Applied Biosystems application note

on conformation sensitive capillary electrophoresis (CSCE) to be used as a sequencing prescreen. Many have awaited this information because now, using a commercially available semidenaturing polymer, this protocol can be transported up to the 96-capillary 3730xls, markedly increasing throughput. The downside is that a capillary array dedicated to CSCE is required. Chris Mattocks (National Genetics Reference Laboratory, UK) showed how this technique is being set up in a diagnostic laboratory, to be used on oncogenes such as *BRCA1*. To increase flexibility and decrease the cost of fluorescent primers, his lab uses 'universal fluorescent primers'. Matt Hayden (Adelaide, Australia) and others showed that



Happy conference-goers enjoy the magnificent scenery of Santorini at the conference dinner. Photo courtesy of Matt Trau (University of Queensland, Australia).

*Hamish S. Scott is in the Genetics and Bioinformatics Division of The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, 3050 Victoria, Australia. e-mail: hscott@wehi.edu.au*

similar universal priming strategies also work for microsatellites and denaturing high-performance liquid chromatography<sup>1</sup>.

Richard Wooster (Cancer Genome Project, Sanger Institute, UK) described the high-throughput sequencing and mutation detection now being used in place of CSCE. This Project included bidirectional sequencing of 518 kinase genes from breast cancer cell lines and tumors. Experimental design, sample tracking and data analysis for a project of this scale required considerable automation to generate a pipeline<sup>2</sup>.

The sequencing capacity of the Cancer Genome Project is well beyond the requirements and abilities of most diagnostic and research laboratories, not least because of the difficulties in obtaining adequate and appropriate samples. As Tania Tabone (Genomic Disorders Research Centre, Australia) described, there is still a need for a mutation detection method that is simple, sensitive and cost-efficient to carry out in most diagnostic and research laboratories. She presented the 'mismatch oxidation assay', which measures changes in absorbance caused by oxidation of mismatched nucleotides with the old chemical cleavage favorite, potassium permanganate (KMnO<sub>4</sub>). To date, however, the assay has only ~70% sensitivity and has been done only on short amplicons.

Possibly fitting into this simple, sensitive and cost-efficient niche for detection of new sequence variants is high-resolution melting, presented by Jason McKinney (Idaho Technology, USA) and Michael Hoffmann (Roche Applied Science, Germany). High-resolution melting analysis relies on Idaho Technology's proprietary DNA binding dye called LCGreen, which is thought to saturate heteroduplexes. Even partial denaturation must displace some dye, resulting in a shift in fluorescence. A thermodynamically precise instrument, capable of taking many readings during the denaturation, is also required. Idaho Technology has produced a 96- or 384-block instrument, called a LightScanner, that has all the hardware and software solutions for dedicated high-resolution melting analyses. A pre-sequencing scan of 384 heteroduplexed samples is now possible in ~5 min, with no post-PCR purification or manipulation. Roche Applied Science has also produced a machine with a similar capacity for high-resolution melting, the LightCycler 480.

### Complex disorders

Depending on one's position on the 'common disease—common variant' hypothesis, analyses of complex disorders may differ entirely from mutation analysis or detection of new sequence variants. Many common functional SNPs may already be described by the HapMap project.

Notwithstanding this, there are choices to be made, for example, between genome-wide or candidate analysis for association. Affymetrix SNP chips will soon have ~500,000 SNPs. This should increase the chances of assaying functional SNPs. ParAllele offers a combination of genome-wide and candidate analyses with their MegAllele Genotyping Human 10K cSNP Panel containing ~10,000 nonsynonymous public SNPs that may code for functional changes.

But much of the human genome is organized into a series of regions with high linkage disequilibrium (LD). For linkage and association, the statistical problems that LD already brings at the 10K level with Affymetrix SNP chips are challenging and require corrections to avoid introducing false positive lod scores, as shown by Garry Hannan (CSIRO, Australia) in an analysis of families with prostate cancer. Additionally, statistical and computation issues will increase exponentially with analyses of large data sets.

Assaying a selected number of SNPs, called tagging SNPs, in regions of high LD may be sufficient to detect linkage and association. But the LD regions may differ between populations. Elin Lõhmussaar (Tartu, Estonia) compared LD patterns of selected regions of the genome in several relatively isolated European populations to that of the HapMap and also assessed the transferability of LD information among populations. The results suggest that, although future denser HapMap data sets will allow selection of tagging SNPs in most LD regions for use in central European populations, for an unknown proportion of genes, the HapMap reference data will need to be augmented with tagging SNPs defined in local populations, especially for isolated and peripheral populations<sup>3</sup>.

Part of the basis of the HapMap project is that carrying out genetic analysis with haplotypes maximizes the power of the study. But many studies don't include family information, and computer algorithms used to predict extended haplotypes have a relatively high error rate. If the haplotypes of a small number of markers are known, the accuracy of the computer algorithms is greatly improved. Therefore, Pui-Yan Kwok and Ming Xiao (San Francisco, USA) created an instrument that they envision will carry out automated data analysis and haplotype calling of 1,000 haplotypes per week. Long PCR products are labeled with padlock probes and forced through a type of maze to linearize the single molecules before they pass under the five-color detector.

With a similar goal to the engineering feat described above, James Wetmur (Mount Sinai, USA) presented molecular haplotyping by linking emulsion PCR. The PCR is done in an emulsion generated simply by vortexing a mixture of

oils with the PCR mix. The emulsion ensures that, normally, only one genome is present in each aqueous drop and thus becomes the target of a PCR reaction. The use of 'linking' primers near the SNPs of interest allows formation of minichromosomes, thereby preserving the phase information or haplotype of two polymorphic loci. The haplotypes are then scored by standard allele-specific PCR<sup>4</sup>.

Presentations by Lili Milani and Elin Grundberg (both from Uppsala, Sweden) looked at the allelic imbalance of expression using SNPs in mRNAs. This approach aims to refine a list of candidate genes for a complex disease by detecting an alteration from the expected 1:1 ratio of expression of the alleles. Though looking at different diseases and genes and using different techniques, allelic imbalance of expression was detected in both studies but no association or proven regulatory SNPs have been identified to date.

### Bypass everything and sequence?

Two presentations from industry described how close we are to the \$1,000 genome, or at least the \$100,000 genome. Harold Swerdlow (Solexa, UK) described the company's cloned single molecule arrays for genome-wide resequencing. Randomly fragmented genomic DNA is attached to a chip, and a four-color sequencing strategy then generates reads of ~25 bases from each of the millions of fragments. These are then aligned back to a reference sequence and can be used to determine sequence differences. Approximately 70–80% of the 25-base fragments can be mapped back to a mammalian genome.

Kenton Lohman (454 Life Sciences, USA) described their DNA sequencing system, recently used to assemble the ~580,000-nucleotide complete genome of *Mycoplasmagenitalium*<sup>5</sup>. Again, genomic DNA is randomly fragmented, and each fragment is clonally amplified while attached to beads in emulsion, to ensure one fragment per bead. The beads are then deposited into PicoTiter plates, one bead per well. Simultaneous pyrosequencing of each of the ~1.6 million wells results in reads with an average length in excess of 100 bases and more than 20 Mbp of sequence per 4-h instrument run.

Both methods should have greater than a 100× throughput advantage over traditional fluorescent sequencing, and other planned applications include expression profiling and karyotyping using these technologies. Both techniques should be adaptable to in-depth sequencing of multiple amplicons. The most obvious application in human may be in cancer, as haplotype information is lost for complex disorders.

Kalim Mir (Oxford, UK) presented a less well-developed new approach to genome sequencing that shares many similarities to Solexa's. Instead of randomly distributing the sheared DNA on a chip, hybridization to a microarray is used to reorder the genome before sequencing. Single molecules in a hybridized spot on the microarray are then analyzed. Additionally, by forcing a stretched linear display of the captured DNA, this method is hoped to retain long-range information such as haplotypes. Matt Trau (Queensland, Australia) demonstrated the potential power of 'Nano-Balls' for storing and retrieving biological information. One of the reasons for their potential power is the increased surface area; M. Trau questioned why biologists often rely on flat surfaces in two dimensions. Although 454 Life Sciences' technology partially uses this capacity, one wonders whether a totally fluidic three-dimensional sequencing technology might be around the corner.

### Miniaturization and throughput

Automation, scale and miniaturization are always drivers of high-throughput technologies. Theodore Christopoulos (Patras, Greece) presented microfluidic devices, one including integrated heating and detectors, for both reverse transcription and PCR<sup>6,7</sup>. PCRs of 2  $\mu$ l could be done in only 60 s. He showed that, even at such low reaction volumes, detection of genotyping results was possible with a disposable dipstick using oligonucleotide probes labeled with gold nanoparticles<sup>8</sup>. As no instrumentation was required for detection, or for some of the PCR analyses, one can imagine a field deployment of this type of system for pathogen detection and genetic diagnoses in third-world countries, for example. Clarissa Consolandi (Milan, Italy) presented a similar scenario as a 'Lab on a Chip' that resembled a standard microscope slide with an embedded microfluidic section and control module. After adding the sample through an inlet port, the slide can be placed in a housing tray in a standard desktop PC containing the application software to control the PCR. Cycle times had been reduced to only 22 min, and future plans include an inbuilt DNA detector.

Raphael Sandaltzopoulos (Thrace, Greece) presented impressive progress in the development of nanoarrays. Proteins or nucleic acids are immobilized in monolayers with nanometer features and are still free to interact, for example, by hybridization. Single molecules can be detected using a standard CD photodiode like those found in computers or stereos. With the number of features that could be placed on a standard CD, and the already widespread availability of the detection system, this group envisions these nanoarrays being analyzed at family doctors' offices.

### Large-scale genomic rearrangements

Jan Schouten (MRC, Holland) presented an update on the flexibility of multiplex ligation-dependent probe amplification (MLPA). Judging from both posters and oral presentations, MLPA has become one of the most popular research and diagnostic tools for detection of both small and large genomic rearrangements involving copy number changes. Schouten also demonstrated the use of MLPA for detection of CpG methylation changes and genotyping of single-nucleotide variants.

A number of investigators used array technologies to look for larger scale genomic rearrangements. Nigel Carter (Sanger Institute, UK) used array comparative genomic hybridization (CGH) to detect clinically relevant chromosomal deletions and duplications in individuals with mental retardation. Hau Ren (MCRI, Australia) used the 100K Affymetrix SNP chips for high-resolution karyotyping, as a type of CGH that can also detect uniparental disomy. One of the Sanger array CGH profiles, presenting a control experiment of anonymous male DNA versus pooled male DNA, showed presumably polymorphic copy number variation of a sizable genomic fragment. Anthony Brookes (Leicester, UK) presented data from a quantitative SNP scoring technique showing that a large number of duplicons (duplicated genomic segments, with 90–100% similarity, >1 kb length) not only contain a large number of SNPs but also are polymorphic in copy number themselves<sup>9</sup>. There are already demonstrations of phenotypic consequences for some polymorphic duplicons. This could cause havoc with karyotypic analyses and also some results of linkage and association studies.

### Databases and bioinformatics

Given the large amount of data generated by genetics and genomics studies, sophisticated informatics and bioinformatics are needed to drive and analyze these technologies. For example, what will be needed in terms of computer hardware and software to resequence a complex genome? Software solutions to speed up data analyses from standard sequencing traces received attention from Jonathan Liu with the popular Mutation Surveyor (Softgenetics, USA) and Jurgen Del Favero (Antwerp, Belgium) with novoSNP<sup>10</sup>. George Patrinos (Erasmus, Netherlands) and Dick Cotton (Genomic Disorders Research Centre, Australia) highlighted the difficulties in collecting and 'databasing' all laboratory-based evidence of sequence variation and function with relevant clinical information<sup>11,12</sup>. There were suggestions and solutions offered, but it is difficult to see them coming to fruition, although variant databases are proven to be useful<sup>13</sup>.

### What will it mean?

Summing up the meeting, Uta Francke (Stanford, USA), wearing her clinical geneticist's hat, asked what this will mean for the people. We are still a long way from having 'one test for all in the clinic', a test that is cheap and widely available for clinical diagnoses, as well as pharmacogenetic and pharmacogenomic analyses. It is still difficult to imagine the day when a person's full genomic sequence, including haplotype information, will affect their health care treatment, management and reproductive choices. This is particularly true for complex disorders. With risk assessment from genetic epidemiology studies, however, personal genomic sequences may bring preventative medicine to the fore. Ed Southern (Oxford, UK) commented that it is often the simplest techniques that get widespread use when applied intelligently rather than the ultra-complicated. But what appears complicated to us today may be child's play for the next generation.

1. Guipponi, M. *et al.* Universal fluorescent labeling of PCR products for DHPLC analysis: reducing cost and increasing sample throughput. *Biotechniques* **39**, 34, 36, 38, 40 (2005).
2. Stephens, P. *et al.* A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.* **37**, 590–592 (2005).
3. Mueller, J.C. *et al.* Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am. J. Hum. Genet.* **76**, 387–398 (2005).
4. Wetmur, J.G. *et al.* Molecular haplotyping by linking emulsion PCR: analysis of paraoxonase 1 haplotypes and phenotypes. *Nucleic Acids Res.* **33**, 2615–2619 (2005).
5. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, advance online publication 31 July 2005 (doi:10.1038/nature03959).
6. Obeid, P.J., Christopoulos, T.K., Crabtree, H.J. & Backhouse, C.J. Microfabricated device for DNA and RNA amplification by continuous-flow polymerase chain reaction and reverse transcription-polymerase chain reaction with cycle number selection. *Anal. Chem.* **75**, 288–295 (2003).
7. Obeid, P.J., Christopoulos, T.K. & Ioannou, P.C. Rapid analysis of genetically modified organisms by in-house developed capillary electrophoresis chip and laser-induced fluorescence system. *Electrophoresis* **25**, 922–930 (2004).
8. Kalogianni, D.P., Koraki, T., Christopoulos, T.K. & Ioannou, P.C. Nanoparticle-based DNA biosensor for visual detection of genetically modified organisms. *Biosens. Bioelectron.*, published online 31 May 2005 (doi:10.1016/j.bios.2005.04.016).
9. Fredman, D. *et al.* Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet.* **36**, 861–866 (2004).
10. Weckx, S. *et al.* novoSNP, a novel computational tool for sequence variation discovery. *Genome Res.* **15**, 436–442 (2005).
11. Patrinos, G.P. & Brookes, A.J. DNA, diseases and databases: disastrously deficient. *Trends Genet.* **21**, 333–338 (2005).
12. Cotton, R.G., Auerbach, A.D. & Oetting, W.S. A call for mutations. *Genet. Med.* **7**, 370 (2005).
13. Vogt, G. *et al.* Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations. *Nat. Genet.* **37**, 692–700 (2005).