Second International Meeting on Single Nucleotide Polymorphism and Complex Genome Analysis

17th - 20th September 1999, Schloß Hohenkammer, Germany

Correspondence Anthony.Brookes@cgr.ki.se

Authors / Meeting Organisers
Anthony Brookes (1), Ulf Landegren (2), Ann-Christine Syv„nen (3), Peer Bork (4), Anders Isaksson (2), Christian Stein (5), Flavio Ortigao (6).

Affiliations
(1) Center for Genomics Research, Karolinska Institute, Theorells v„g 3, S-171 77 Stockholm, Sweden.
(2) Department of Genetics and Pathology, Biomedical Center, Uppsala University, S-751 23 Uppsala, Sweden.
(3) Department of Medical Sciences, Molecular Medicine, Uppsala University Hospital, S-751 85 Uppsala, Sweden.
(4) European Molecular Biology Laboratory, Meyerhofstr. 1, Heidelberg, PF 10.2209, Germany.
(5) Fraunhofer Patent Centre, Patent and Licensing Agency for the DHGP, Leonrodstr. 68, 80636 Munich, Germany.
(6) Interactiva Biotechnologie GmbH, Sedanstr.10,Geb.8, D-89077 Ulm, Germany.

General Comments

Competitive registration procedures based upon submission of a relevant scientific abstract enabled 100 international researchers to recently attend the 2nd International Meeting on SNPs and Complex Genome Analysis in Schloß Hohenkammer, Germany. Participants included 30 invited speakers, and both industry and academia were represented at the meeting. Despite the unfortunate visitation of Hurricane Floyd to the US East coast on the 16th September, all but one of the 31 American registrants managed to find alternative flights. We sincerely thank all the US delegates for making this effort to attend! Thanks are also extended to the corporate sponsors of the meeting, namely the flagship sponsor Interactiva GmbH, as well as Amersham Pharmacia Biotech AB, Glaxo Wellcome, Hybaid Ltd, Millenium Predictive Medicine, Orchid Biocomputer Inc, Perkin-Elmer Applied Biosystems, Pharmacia & Upjohn Inc., Professional Genetics Laboratory, Pyrosequencing AB. Other support was generously provided by the German Ministry for Education and Research and the EU through a contract with HUGO Europe (EC contract BMH4-CT97-2031 to HUGO).

The scientific program was based upon three days of 25-minute oral presentations, plus two poster sessions and a two-hour open discussion session towards the close of the meeting. Subjects covered included; General Considerations, Mapping & Discovery, Technology (2 sessions), Population Genetics (2 sessions), Complex Diseases (3 sessions), Databases & Bio-Informatics (2 sessions). Less formal interactions were spirited' and were helped along by a guided tour of the Bavarian city center of Freising, dinner at an ancient brewery, a champagne reception and a Grand Bavarian Buffet.

The overall tone of the meeting suggested that SNP research is still somewhat in its infancy, though certainly more developed that at the time of the previous years meeting. There remains no solid consensus about detailed aspects of how best to proceed and this is at a time when vast sums of money are being spent in public and private SNP programs. Perhaps this is not so bad, since these activities are certainly propelling the field forward and will at least furnish a number of useful markers. But discovery alone will not tell us how we can or should utilize SNPs. Instead, this wisdom has to come from empirically determined guidelines, and studies based upon best guesses and theory using latest technologies. A large fraction of the presentations concerned a move in precisely this direction. And a smattering of critical insights and an emerging convergence of ideas were notable, particularly in presentations concerning the utilization of SNPs.

SNP Discovery

There are many, many SNPs to be discovered but which ones should we chase down first, and how? A great deal of money is available (now largely committed) for SNP discovery, but there is a continual need to re-evaluate strategies and objectives. Initial self-serving collection of SNPs for private use and patenting by industry has generally been accepted to be inefficient and unwise. Therefore, The SNP Consortium (TSC) was created, as summarized by Arthur Holden (TSC Chairman, Chicago, USA). This enterprise has established a well-structured program towards the discovery and public cataloging of 300,000 random genomic SNPs (at a 95% accuracy target) in two years, plus detailed mapping of over 170,000 of these. In complementary undertakings, Ken Buetow (National Cancer Institute, Maryland, USA), Shamil Sunyaev and Peer Bork (EMBL, Heidelberg, Germany), and Don Morris (Incyte Pharmaceuticals, Palo Alto, USA) reported on efficient discovery of SNPs by automated alignment and comparison of EST sequences. The scope of this endeavor is of course limited by the depth and number of distinct EST contigs to perhaps a few tens of thousands of SNPs, but the exercise is highly cost effective and the intra-genic location of these variants could make them more individually useful than the markers discovered by TSC.

Other reported discovery efforts were linked to specific disease research programs, and fell into two categories. Firstly, regional genomic DNA level discovery was reported as a means to provide SNPs to more precisely localize (by association analysis in populations of families and case/control materials) disease related mutations already mapped to large chromosome regions by linkage scans. Examples were reported by Michal Prochazka (National Institutes of Health, Arizona, USA) and Allen Roses (Glaxo Wellcome, North Carolina). Secondly, several groups reported SNP discovery to underpin genome-wide candidate gene approaches to disease gene identification. Thus, extensive lists of genes of highest interest are being specifically targeted for intra-genic SNP discovery. Resulting sets of mostly coding sequence SNPs were reported by Anthony Brookes (Karolinska Institute, Stockholm, Sweden) for neurodegeneration candidates, by Ken Buetow (National Cancer Institute, Maryland, USA) for cancer candidates, and by Pamela Sklar (Whitehead Institute, Boston, USA) for several diseases.

SNP Scoring and Detection Technologies

A plethora of competing SNP genotyping methods are being developed - and in a wry comment on this Michael O'Donovan (University of Wales College of Medicine, Cardiff, UK) joked that "..a man without a method is not a man". Truly useful methods will need to meet stringent requirements of both high-throughput and accuracy, as was emphasized by Joe Terwilliger (Columbia University, New York, USA). He pointed out that extremely large sample materials may be required to achieve sufficient statistical power in association studies, and that a genotyping error rate of as little as 1% could have a disastrous effect on statistical power. The cost of the SNP scoring methods was also a key consideration. Claimed genotyping costs ranged from a fraction of one dollar to many dollars per sample.

Two principal approaches to SNP scoring are being taken. The first is to develop fast and simple methods for scoring SNPs in individual reactions. Molecular beacons presented by Sanjay Tyagi (Public Health Research Institute, New York, USA) are allele-specific hybridization probes that become fluorescent upon target binding. In Dynamic Allele Specific Hybridization (DASH) (Anthony Brookes, Karolinska Institute, Stockholm, Sweden) allele specific hybridization of oligonucleotides to immobilized target molecules is monitored using a low-cost fluorescent intercalating dye over a temperature gradient, thereby significantly enhancing the specificity of the allele discrimination. DNA-polymerase-assisted allele distinction is utilized in "Pyrosquencing" (P†l Nyren, Royal Institute of Technolgy, Stockholm, Sweden) and in the template-directed dye-terminator (TDI) method (Pui-Yan Kwok, Washington university School of Medicine, St. Louis, USA). Signal detection in pyrosequencing is based upon chemiluminescence induced by pyrophosphate released during primer extension, while TDI uses fluorescence polarization - a detection principle that is also applicable to allele discrimination by oligonucleotide ligation and using the 5'-exonuclease ("TaqMan") assay. The assays described above have in common that they all depend upon PCR amplification. The "Invader" assay, presented by Shaun Lonergan (Third Wave Technologies, Inc., Wisconsin, USA), employs isothermal cyclic enzymatic cleavage of mismatched junctions between probe and target, and is independent of PCR. All these methods for individual typing of SNPs are or will soon be

available through commercialization of instruments and/or reagents.

The other principal approach towards SNP genotyping is to score multiple SNPs from each sample in a multiplexed fashion. This can be achieved by using immobilized oligonucleotides on microarrays. Single-nucleotide primer extension (minisequencing) technology on micorarrays was presented by Ann-Christine Syv„nen (Uppsala University, Sweden), and illustrated by determining the frequencies of a panel of disease causing mutations in the Finnish population (Tomi Pastinen, National Public Health Institute, Helsinki, Finland). The principle of single base extension in conjunction with generic oligonucleotide microarrays carrying sequences complementary to tag sequences on extension primers was reported by Pamela Sklar (Whitehead Institute Genome Center, Cambridge, USA). An alternative to this was described by Allen Roses (Glaxo Wellcome, North Carolina, USA). In his system, microspheres that are labeled with a variety of distinct fluorophores can thus be separated in a fluorescent cell sorter, carry sequences complementary to tags on extension primers. Tagged oligonucleotide ligation on arrays followed by detection using MALDI-TOF mass-spectrometry was described by Nick Housby (University of Oxford, UK). The requirement for PCR prior to the detection reaction is likewise the major bottleneck for multiplex SNP scoring methods. Multiplex analysis via replication of padlock probes on microarrays described by Anders Isaksson (Uppsala University, Sweden) may offer a way to circumvent PCR

The possibility of increasing throughput and decreasing the cost of SNP scoring in association studies by simultaneous analysis of pooled DNA samples was discussed by speakers such as James Weber (Center for Medical Genetics, Marshfield, USA) and Michael O'Donovan (University of Wales College of Medicine, Cardiff, UK). Sequence analysis of pooled DNA samples is also being used by TSC to discover high frequency SNPs (Arthur Holden, TSC, Chicago, USA).

Diseases and Phenotypes

The argument that genome variation contributes to phenotypic variation is both elegant and persuasive. However, it is uncertain to what extent this genetic component of disease etiology can be practically dissected. The ultimate goal of performing objective whole genome analysis entailing hundreds of thousands of SNPs and ultra-high-throughput assays is yet to be attained. Furthermore, major questions remain unanswered, such as the extent and variance in linkage disequilibrium (LD) in different chromosomal regions and in different populations, as well the real level of architectural complexity of common disease.

Given the above, SNPs and association analysis are presently being used as a means i) to home-in on disease related mutations within large regions previously identified by linkage scans, ii) to screen several variations surrounding a few or many pre-chosen candidate genes, or iii) to follow-on from extensive sequence studies of single candidate genes to determine associations between specific haplotypes and disease. A small number of successful uses of these strategies show that they can sometimes work - though we don't know how often, or how we can maximize the rate of success. But there exist many potential pitfalls. Papers presented on these subjects gave some insights into what can and cannot be achieved, in theory and practice. Allen Roses (Glaxo Wellcome, North Carolina) showed that a 4MB SNP map around the APOE locus would have enabled one to locate the increased risk that the E4 allele contributes to Alzheimer's Disease (AD). But this risk allele is in many ways a rather ideal case (10 fold increased risk in homozygotes, medium frequency allele, and similar effect in almost all populations studied). Another such case in point was presented by Michal Prochazka (National Institutes of Health, Arizona, USA) who is using SNPs to follow-up on a chromosome 1q21-q23 region linked to Type II diabetes in PIMA Indians, revealing four positively associated SNPs around the GIRK3 locus.

But what can one do if there is no initial linkage to guide ones search? Presently, the answer seems to be to make educated guesses about which genes are of likely importance. Anthony Brookes (Karolinska Institute, Stockholm, Sweden) summarized attempts to do this for AD. By spreading a wide net comprising 400 intra-genic SNPs derived from over 250 genes from four candidate systems, he is undertaking a pathway-based hypothesis testing' strategy in genetically homogenous populations enriched for genetic causation (twin materials and early onset cases). Initial data show distinct evidence for an involvement of one of the four tested pathways. Michael O'Donovan (University of Wales, Cardiff, UK) is following a similar strategy, but also including a linkage component, to analyze AD, schizophrenia, and dyslexia. Detailed single gene

analyses were presented by Ryk Ward (University of Oxford, Oxford, UK) who is evaluating ACE gene haplotype specific effects upon plasma ACE levels and indices of blood pressure, and by Margret Hoehe (Max-Delbrueck-Center, Berlin, Germany) who is searching for correlations between OPRM1 gene haplotypes and substance dependence. Both studies began from a strong prior belief in the relevance of the tested candidate gene. The data of Ryk Ward showed how intricate estimations of haplotype configurations, regression and cladistic analyses can lead one towards the precise intra-genic location of the pathogenic allele(s) and genotype combinations. This evolutionary perspective was a key take-home message.

In final cautionary notes from Joseph Terwilliger (Columbia University, New York, USA), the current emphasis of most SNP research upon tools and genotyping (finding the marker to pathogenic allele link) was contrasted with the relative lack of attention being given to careful study design and sample ascertainment (finding the pathogenic allele to disease link) comments that provided the primary focus of a subsequent two hour discussion session. Joe's astute question " if you can't find families then why are you looking for underlying genes?" crystallized his concern that many case/control association studies may be futile because the lack of even rare families segregating the disease could indicate too high genetic complexity. His advice was to work hard towards careful patient/family selection and phenotype choice, and to combine association and linkage statistics in a single analysis.

Population Genetics

Disease gene studies and population genetics efforts would benefit from knowing the typical distances up to which allelic variants can be expected to be in linkage disequilibrium (LD). Alison Dunning (University of Cambridge, Cambridge, UK) reported work to establish LD maps in four human populations that differed in the time elapsed since they were founded. She observed significant LD between most neighboring markers and sometimes between markers over 100 kb apart. Another common interest of disease and population geneticists concerns extant haplotypes. In this regard, Svante P„„bo described work to characterize haplotypes found in a 10 kb low recombination region on the human X chromosome. Sampled individuals were selected worldwide according to language, and the study encompassed other primates. He found far greater sequence diversity among chimpanzees, and identified haplotypes ancestral to mankind.

Charles Aquadro proposed using SNP analyses to look for footprints of adaptive evolution. Selective sweeps would be expected to reduce allelic heterogeneity in the vicinity of selected alleles. He suggested that analysis of ETLs (evolutionary trait loci) could serve to identify regions including allele targets of such selective fixation. These regions could include beneficial alleles that have come to dominate in the population, and would therefore be important targets of study.

Databases and Bioinformatics

Given future high throughput detection technologies it will be important to have adequate systems for the collection and integration of all this data. Jan Olsson (CyberGene AB, Sweden) discussed efforts to link local solutions and specific requirements to the outside world via the internet. Marianne Siegfried (Interactiva GmbH, Germany, and Karolinska Institute, Sweden) presented the HGBASE database which is run by a European consortium, has a human gene SNP focus and presently contains 7,000 entries with another 15,000 being processed for imminent release. Heikki Lehv„slaiho (EMBL-EBI, Cambridge, UK) presented software and database solutions to link various locus-specific mutation databases with SNP databases allowing complex queries. It was agreed that efforts should be combined to provide the community with high quality databases and that the requirements for these will definitely increase.

Currently, many groups try to mine SNPs from EST data. Don Morris (Incyte Pharmaceuticals, Palo Alto, USA) described progress on the Incyte EST dataset harvesting more than 40,000 candidates - but it is hard to measure allele frequencies in silico, and so many rare alleles might be among them. Shamil Sunyaev (EMBL, Heidelberg, Germany) presented efforts at EMBL to identify 9,000 cSNP candidates in public databases. It is difficult to compare such numbers as they depend upon prediction accuracy which is probably only 60% - 80% in most cases. This is, however, sufficiently accurate to provide candidates for particular genes that can then be tested. Shamil Sunyaev also reported attempts to map large numbers of

SNPs onto human chromosomes as well as onto three-dimensional protein structures. Such approaches will generate useful hypotheses and these will be integrated with experimental efforts to understand phenotypic differences as well as human evolution using SNP data.

Intellectual Property and Commercial Issues

Rules and strategies on the patenting of SNPs are changing. The buzz-word at this years conference was 'pre-competitive research'. The open international efforts of TSC have added a new dimension to attitudes towards intellectual property in the genomics field. The pharmaceutical industry has given an unmistakable sign that they are now less prepared to invest large sums of money into private efforts towards pre-competitive research and tool development. The European Directive on the Legal Protection of Biotechnological Inventions (98/44/EG, July 1998) and the Decision of the Administrative Council, amending the Implementing Regulations to the European Patent Convention (1st September, 1999) will help, at least partially, to clarify what is and what is not patentable as regards SNPs, ESTs and partial genomic sequences. There is a clear trend towards a situation in which a patent on a partial nucleic acid sequence or SNP will only be granted when it is possible to define a function and commercial application of this sequence or SNP.

Commercially, there are three areas where profits are expected to be generated, i) applying SNPs to pharmacogenomics by discovering functional implications, ii) genotyping individuals for particular SNPs, and iii) creating technology platforms for SNP discovery and utilization. In the first of these situations, it is hoped that proprietary rights to the right SNP' will generate profits via licensing, and that drug development will be enhanced by genetic profiling' in clinical trials. In the third situation, profits can be generated over shorter time-scales by technology supply. But meaningful business from SNP knowledge will only be realized if solid correlations between SNPs and gene function are determined. A key question is whether the most obvious and rewarding SNPs (from a cost/benefit standpoint) have already been discovered and patented by the early-birds, or if the late-comers still have a good chance of finding valuable SNPs.

The online presentation of this publication is a special feature of the Human Genome Project Information Web site.