Meeting report

AACR Special Conference: SNPs, haplotypes, and cancer – applications in molecular epidemiology, Key Biscayne, Florida, USA, 13–17 September 2003

David G Cox

Department of Epidemiology and Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA

Corresponding author: David G Cox (e-mail: dcox@hsph.harvard.edu)

Published: 18 December 2003

Breast Cancer Res 2004, **6**:E9 (DOI 10.1186/bcr756)

Abstract

This American Association for Cancer Research Special Conference brought together scientists with diverse expertise to address issues related to use of appropriate epidemiological, statistical, and laboratory methods to study the genetic epidemiology of cancer. Discussions focused on experiences with association studies using single nucleotide polymorphisms and haplotypes, their limitations, and what is needed to improve on the current 'state of the art'. Various studies were presented in different contexts, ranging from candidate gene studies to whole genome scans, and conducted in prospective cohorts, case–control studies, and other study designs. Common problems such as determining the probability that observed associations are false negative or false positive, the potential effects of admixture, and determining which polymorphisms to examine in which genes and in which populations were examined. Problems specific to haplotype analysis were discussed, with emphasis on haplotype block structures and on how to use haplotypes in analysis. Questions were also posed as to determining the functional relevance of single nucleotide polymorphisms in molecular epidemiology. Finally, future directions, using specific examples, were addressed.

Keywords: cancer, haplotypes, molecular epidemiology, single nucleotide polymorphisms

Introduction

In their letter to conference attendees, Drs Timothy Rebbeck, Fred Kadlubar, and Christine Ambrosone proposed three main questions as aims to be targeted in the Special Conference. First, what have we learned from association studies of single nucleotide polymorphisms (SNPs)? Second, what have we been unable to accomplish and why? Finally, what approaches are required to improve the state of the science?

The conference was aimed primarily at epidemiologists who are either performing association studies with SNPs and haplotypes, or who are planning to do so in the near future.

Keynote talks

John D Potter (Fred Hutchinson Cancer Research Institute, Seattle, WA, USA) and Charles R Cantor (Sequenom, San Diego, CA, USA) were the keynote speakers. Dr Potter gave a general introduction to SNPs and haplotypes, emphasizing that using the haplotype strategy reduces the amount of genotyping necessary and increases the power of a study. However, he proposed that most studies conducted today have limited power to detect low penetrance risk alleles, or complex gene–environment interactions. He proposed 'the last cohort', which would consist of at least half a million ethnically diverse individuals. Cell lines should be established and incident tumor samples would need to be collected. This sort of approach would be able to answer questions about association in most common diseases and to address various interactions. However, this approach may still have low power to detect associations in complex diseases with diverse subtypes.

Dr Cantor presented data from population-based association studies, in which he constructed DNA pools to assess allele frequencies in the case and control groups. The

NCI = National Cancer Institute; SNP = single nucleotide polymorphism.

pools were then genotyped in whole genome scans using 30,000 to 60,000 markers. This method allows analysis of the distribution of genotype frequencies between the two populations. Samples are very carefully separated into different groups, whether this is by case–control status or even within subsets based on molecular or environmental descriptions. Then, each pool is genotyped for a panel of SNPs, covering the entire genome. If any evidence of association is discovered, then individual samples can be genotyped to confirm the association.

Current paradigms for SNP studies

Stephen J O'Brien (National Cancer Institute [NCI], Frederick, MD, USA) presented his views on mapping by admixture linkage disequilibrium as a shortcut to multifactorial and complex disease gene discovery. By studying diseases that exhibit different incidence between ethnicities, in admixed populations linkage disquilibrium may be extensive around genes associated with the disease.

Choosing suitable genetic markers was a subject touched upon by both Sholom Wacholder (NCI, Bethesda, MD, USA) and Joel N Hirschhorn (Children's Hospital, Boston, MA, USA). Both presented material showing that good selection not only of markers but also of diseases in which to study these markers affects the likelihood that a study will yield a false-negative or false-positive finding. A common thread was that associations must be found repeatedly if they are to be believed, and it is difficult to determine when enough evidence is enough.

Genetics and population structure in cancer gene identification

Haplotypes are regarded as important tools not only in association studies but also in evolutionary biology. Daniel O Stram (University of Southern California, Los Angeles, CA, USA) presented current thinking on the use of haplotypes in association studies, including determining haplotype tagging SNPs, or those SNPs that describe all the haplotypic variation present across a region of the genome within a particular population.

Population substructure was the topic of discussion for both Jonathan Pritchard (University of Chicago, Chicago, IL, USA) and Rick A Kittles (Howard University, Washington, DC, USA). Given that certain diseases are more common in some populations than in others, it is possible that differences between cases and controls in the mix of ethnicities can lead to spurious associations. By analyzing markers that differ between ethnicities, but are not likely to be associated with disease, it is possible to detect the degree of confounding by ethnicity, or population stratification, that is present in a study.

David Balding (Imperial College, London, UK) presented the idea of using haplotypes in fine-scale mapping of disease susceptibility loci. This is accomplished by reconstructing the evolutionary history of the haplotypes in the case population. Although these methods are becoming more computationally feasible, they are still very computer intensive. Future directions include multistage designs to combine information from haplotype tagging SNP selection and the full case-control study, incorporation of block structures of haplotypes and their inherent uncertainty, allowing for population structure, and combining associations and sharing of genetic variation between phenotype.

Applying high throughput SNP technology to epidemiologic studies

SNP assays must be carefully designed because, as Dr Meredith Yeager (NCI, Gaithersburg, MD, USA) pointed out, adjacent SNPs may exist in the region of a targeted SNP that may interfere with primer or probe hybridization and cause misclassification of results. Once SNPs are selected, they must be genotyped in large numbers of samples. Dr Stephen Chanock (NCI, Gaithersburg, MD, USA) described his experience running a core genotyping facility. Robotics and laboratory information management systems are essential to limit errors in sample handling and processing, as well as in data collection and management. As a collaborating site in the HapMap project, Dr Stacey Gabriel (Whitehead Institute, Cambridge, MA, USA) applies high throughput technology in genotyping, and described genotyping platforms applicable to the large sizes of studies currently being proposed.

One method of reducing time and cost of whole genome association studies is to pool samples together, as presented by Eric Lai (GlaxoSmithKline, Research Triangle Park, NC, USA). By creating well defined pools, allele frequencies for hundreds of samples can be determined, and case prevalence compared with control prevalence. Care must be taken to have homogeneous samples within a pool, in order to limit error. Although pooling could be a good screening tool for determining candidate regions or SNPs for disease association, it is limited because of the rigidity of pools. For instance, it would be difficult to remove one or more individuals from a pool because they changed phenotype or withdrew consent to be included in a study.

Pak Sham (Social, Genetics and Developmental Psychiatry Research Centre, Institute of Psychiatry, King's College, London, UK) also pointed out that pooling samples constitutes an initial screen and can not replace individual genotyping. Also, accurate DNA quantification is essential to ensure that each individual is equally contributing allele information to the entire pool.

Strategies for selecting SNPs and haplotypes

dbSNP (http://www.ncbi.nlm.nih.gov/SNP) is one of the main repositories of SNP information on the internet.

Stephen Sherry (National Center for Biotechnology Information, Bethesda, MD, USA) presented its content. The database is searchable and can be downloaded in its entirety if so desired. Obvious candidate SNPs to study would be those that regulate gene expression. Allelic imbalance, or unequal expression of allelic transcripts, could provide evidence of this. Thomas Hudson (McGill University Health Centre, Montreal, Quebec, Canada) described how polymorphisms that effect splicing of mRNA or transcription/translation of the gene product could be responsible for such an effect, as well as those that alter mRNA or protein stability. SNPs creating allelic imbalance are of added interest, considering that they may alter the possibility that a SNP that changes some aspect of protein function is not be expressed, therefore potentially negating its effect.

One factor effecting the selection of SNPs to study is allele frequency. A common theory in genetic epidemiology is that risk for common disease is associated with common alleles. Gilles Thomas (Centre d'Etude du Polymorphisme Humain [CEPH], Paris, France), proposed that this may not always hold true, especially in instances where a large number of rare SNPs in one gene convey risk for disease.

A previously mentioned tool that could prove useful in selecting SNPs to use in studies would be to focus on those that differ in allele frequency between populations. These are SNPs that may have been under some sort of evolutionary selective pressure and therefore alter some aspect of disease risk. Anthony Brookes (Karolinska Institute, Stockholm, Sweden) has been cataloguing these SNPs. He is also finding that 'duplicons' (regions of the genome that are either exactly repeated or are highly homologous to other regions) can give rise to spurious interindividual variation. Increasing the complexity is the possibility that interindividual variation (polymorphisms) can indeed occur in such regions, but because of duplication an individual could have other than the normal one or two copies of a particular sequence. Carefully designed assays can determine the cause of variation seen in duplicons once they are known to be present.

Determining the functional relevance of human SNPs in molecular epidemiology

Only a small fraction of the many tens of thousands of genes already discovered in the human genome have a known function. Considering this, determining the effect on gene function or regulation for all of the SNPs in those genes have is a truly daunting task. Tim Hubbard (Welcome Trust Sanger Institute, Cambridge, UK) described the Ensembl database, which is designed to draw together information from many sources and automatically annotate gene function and the variation present in each gene.

Although determining the functional significance of SNPs in proteins in general is not straightforward, it can be considerably easier for those genes involved in drug metabolism. Genetic variation in drug metabolism was known before the genes actually encoding the responsible enzymes were discovered. The field of pharmacogenomics addresses the issues of such variation and hopes to describe the interindividual differences in response to therapeutic drugs. Additionally, genetic differences between tumors can affect the way in which cancer responds to therapy. Richard Weinshilboum (Mayo Medical School, Mayo Clinic, Rochester, MN, USA) described the history of genetic polymorphisms in genes related to drug metabolism, specifically the sulfotransferase enzymes. William E Evans (St. Jude Children's Hospital, Memphis, TN, USA) discussed differences in response to chemotherapy in acute lymphoblastic leukemia. Gareth Morgan (University of Leeds, Leeds, UK) reported that personalized medicine is coming, taking into consideration germ line (inherited) or tumor (somatic) mutations that determine response to a particular treatment, and that side effects as well as end-points and survival must be taken into consideration when designing clinical studies.

Future directions for SNPs, haplotypes, and cancer

Specific examples of large epidemiological studies relating SNPs and haplotypes to cancer were presented by Nat Rothman (NCI, Bethesda, MD, USA) and David Hunter (Harvard School of Public Health, Boston, MA, USA). The Interlymph Study (non-Hodgkins lymphoma) and the NCI Breast and Prostate Cancer Cohort Consortium are two of the largest, most ambitious cohort studies ever undertaken. Both aim to tackle the problems of determining gene–environment interaction by increasing sample size. This increase is carried out by collaborations with many centers, each with comparable data sets, within the framework of prospective cohorts.

Daniela Seminara (NCI, Bethesda, MD, USA) presented the funding mechanisms at the NCI specifically for collaboration between groups under the Epidemiology and Genetics Research Program. The main purpose of the NCI in these endeavors is to not only provide monetary support but also to provide support in building the infrastructure necessary to carry out such large collaborations successfully.

DeCode Genetics Inc, (Reykjavik, Iceland) has been collecting and genotyping individuals from Iceland with various outcomes, including cancers, and comparing the data with those from healthy individuals. Because of its history of isolation and good record keeping, the genealogy of most people in Iceland is known, and detailed family trees can be drawn. Although this approach will be

useful in detecting risk alleles in the population of Iceland, the probability that any risk alleles will be due to founder effect is fairly high, and therefore they may be of limited significance outside Iceland.

Conclusion

The conference ended with a great feeling of success. Those who had little or no experience and exposure to use of SNPs and haplotypes in epidemiology came away with a better understanding of how to go about setting up such studies, and how to interpret studies in the literature. The substantial methodologic difficulties existing in this area were outlined and guidelines were laid out, fueled in part by recent collaborations such as Interlymph and the NCI Cohort Breast and Prostate Cancer Consortium, on how to design, conduct, and analyze data from large, collaborative efforts.

Competing interests

None declared.

Correspondence

David G Cox, Department of Epidemiology and Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. Tel: +1 617 432 2262; fax: +1 617 432 1722; e-mail: dcox@hsph.harvard.edu