European Journal of Human Genetics (2005) 13, 893–897
© 2005 Nature Publishing Group All rights reserved 1018-4813/05 $30.00
www.nature.com/ejhg

NEWS AND COMMENTARIES

Bioinformatics

Computers or clinicians for complex disease risk assessment?

Carolyn Hoppe

European Journal of Human Genetics (2005) **13**, 893–894. doi:10.1038/sj.ejhg.5201441; published online 11 May 2005

A new study in which Paola Sebastiani and co-workers apply a computer-assisted method to determine the combined effects of multiple candidate gene single nucleotide polymorphisms (SNPs) on stroke risk in sickle cell anemia (SCA) could be a model for how bioinformatic SNP analysis will be used for risk assessment for complex diseases in the future.

The completion of the human genome project, coupled with technological advances in large-scale genotyping and bioinformatics tools, has prompted a resurgence in the use of disease association studies to detect major susceptibility genes. The three million SNP sites identified so far exemplify the allelic complexity of the human genome and serve as markers for large association studies, so providing a valuable resource for investigating the genetic basis of many complex human diseases.

Although SCA is considered a 'monogenic' disease, arising from a single point mutation in the β-globin gene, it is characterized by considerable clinical heterogeneity. Stroke is a particularly devastating manifestation of SCA, afflicting 11% of children before the age of 20 years.[1] As only a fraction of SCA patients develop stroke, environmental and genetic modifiers beyond the sickle gene mutation must account for this phenotypic variability.

Several limited association studies that attempted to identify single SNPs within individual stroke susceptibility genes have produced inconsistent results. This inconsistency has cast doubt on the contribution of potentially relevant genetic modifiers of stroke risk in SCA.[2–7] However, because complex phenotypes such as stroke presumably arise from multiple interacting genes located throughout the genome, the optimal approach would be to search for sets of markers in different genes and analyze these markers jointly, rather than individually. Most statistical approaches typically evaluate the effects of individual SNPs one at a time, and if a significant disease association is found, the SNP is then considered to be near or within a susceptibility gene.[8] However, when large numbers of SNPs are tested simultaneously and related to a single patient phenotype, a true association may not be distinguished from a false one that chance alone has caused (ie a Type I error).[9] Furthermore, such a marker-by-marker approach ignores the multigenic nature of complex diseases, and fails to account for possible interactions between susceptibility genes.

Sebastiani and colleagues offer a viable alternative approach to disease association analyses that uses a machine-learning method derived from the field of artificial intelligence. This approach, based on Bayesian networks, allows one to inferentially explore previously undetermined relationships among genetic and clinical variables, and describe these relationships, once identified. The model integrates the relationships between multiple SNPs, clinical variables and phenotype, and so overcomes many of the limitations of current statistical approaches to disease association studies.

To identify SNPs contributing to stroke risk in SCA, the authors surveyed 108 SNPs distributed over 39 candidate genes in an unselected population of 1398 African-American adults with SCA. They used the Bayesian network algorithm to analyze genotyping results from 92 subjects with documented clinical stroke and 1306 subjects without stroke. An overall 'dependency network' was derived from the joint probabilities of the interdependent relationships between SNPs, clinical variables and stroke.

In all, 25 SNPs among 11 biologically plausible candidate genes, including those that encode BMP6, TGFBR3, SELP and CSF2, in combination with HbF (fetal hemoglobin) level, were found to directly modulate stroke risk. Interestingly, HbF level was not found to be independently associated with stroke in previous analyses of the same study cohort.[1] Another nine genes were found to be associated with stroke via interactions with direct markers.

The model was also used to predict stroke risk based on the combined effects of the genetic and clinical markers identified in the network. Although the individual contribution of each SNP was modest, the simultaneous effect of all 25 identified SNPs and their interaction with clinical variables predicted stroke risk with an accuracy of 98.5%. The results of this analysis were validated in a separate population of 114 individuals with SCA, including seven with reported stroke and 107 without stroke. The Bayesian network model correctly predicted the presence of stroke in 100%, and the absence of stroke in 98%, of the study subjects. The overall predictive accuracy of 98.2% using this model was compared to a logistic regression model that identified only five of the 25 SNPs found in the Bayesian network model and gave an overall accuracy of 88% in the same set of individuals. However, while the ability of the Bayesian network algorithm to predict stroke in this study appears impressive, the validation of this model was based on a small sample that included only seven stroke patients.

The limited availability of large numbers of well-characterized cases and controls represents another challenge in the study of genetic markers in SCA. By using previously collected samples and clinical data from a representative national cohort of individuals with SCA, this study exemplifies how biological sample repositories linked to clinical databases may be efficiently used to successfully perform large disease association studies. However, because cerebrovascular disease in SCA is heterogeneous, manifested as ischemic stroke, intracranial hemorrhage and silent infarction, rigorous phenotypic characterization of cases and controls is imperative.

The lack of available clinical and neuroimaging data needed for optimal phenotypic classification limited this study. Despite these limitations, Sebastiani and co-workers have powerfully demonstrated that multiple SNP sites from different genes over distant parts of the genome are better at identifying overt stroke in SCA than any single SNP or previously identified clinical variable alone. Their results highlight the combined influence of several candidate susceptibility genes on stroke and suggest biological pathways to be explored in future mechanistic studies.

The potential utility of the Bayesian network algorithm is illustrated by the model's ability to determine accurately the relative genetic and clinical effects on stroke risk, find the most probable combination of genetic variants leading to stroke and predict an individual's odds for developing stroke given his/her genotypic profile. As more candidate SNPs and clinical markers are identified, this predictive algorithm will undoubtedly become an invaluable tool in genetic association studies aimed at identifying disease susceptibility genes. The complex interactions modeled through this approach might ultimately translate into clinical benefit through early identification and targeted intervention in those individuals at greatest risk for a particular disease phenotype such as stroke. The computer may well replace the clinician in determining stroke risk, but it will be left to the clinician to apply this information in caring for the patient■

Carolyn Hoppe is at the Department of Hematology/Oncology, Children's Hospital Oakland, Oakland, CA, USA.
E-mail: choppe@mail.cho.org

References

1 Ohene-Frempong K, Weiner SJ, Sleeper LA *et al*: Cerebrovascular accidents in sickle cell disease: rates and risk factors. *Blood* 1998; **91**: 288–294.
2 Carroll JE, McKie V, Kutlar A: Are sickle cell disease patients with stroke genetically predisposed to the event by inheriting a tendency to high tumor necrosis factor levels? *Am J Hematol* 1998; **58**: 250.
3 Taylor JG, Tang D, Foster CB, Serjeant GR, Rodgers GP, Chanock SJ: Patterns of low-affinity immunoglobulin receptor polymorphisms in stroke and homozygous sickle cell disease. *Am J Hematol* 2002; **69**: 109–114.
4 Kahn MJ, Scher C, Rozans M, Michaels RK, Leissinger C, Krause J: Factor V Leiden is not responsible for stroke in patients with sickling disorders and is uncommon in African Americans with sickle cell disease. *Am J Hematol* 1997; **54**: 12–15.
5 Cumming AM, Olujohungbe A, Keeney S, Singh H, Hay CR, Serjeant GR: The methylenetetrahydrofolate reductase gene C677T polymorphism in patients with homozygous sickle cell disease and stroke. *Br J Haematol* 1999; **107**: 569–571.
6 Andrade F, Annichino-Bizzacchi J, Saad S, Costa F, Arruda V: Prothrombin mutant, factor V Leiden, and thermolabile variant of methylenetetrahydrofolate reductase among patients with sickle cell disease in Brazil. *Am J Hematol* 1998; **59**: 46–50.
7 Zimmerman SA, Ware RE: Inherited DNA mutations contributing to thrombotic complications in patients with sickle cell disease. *Am J Hematol* 1998; **59**: 267–272.
8 Sebastiani P, Yu YH, Ramoni MF: Bayesian machine learning and its potential applications to the genomic study of oral oncology. *Adv Dent Res* 2003; **17**: 104–108.
9 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.

Genomics

The amazing complexity of the human transcriptome

Martin C Frith, Michael Pheasant and John S Mattick

European Journal of Human Genetics (2005) **13,** 894–897. doi:10.1038/sj.ejhg.5201459; published online 22 June 2005

New work from Tom Gingeras and colleagues extends the findings of a series of recent global analyses of transcription[1–7] by revealing a much larger number of non-polyadenylated (polyA−) transcripts than expected and an extraordinary level of organizational complexity in the human transcriptome.

A variety of recent evidence indicates that the majority of sequences in eukaryotic genomes are transcribed and that the proportion of transcribed nonprotein-coding sequences increases with developmental complexity (Table 1). However, it is the novel approaches that Gingeras and colleagues employed that allowed them to add spectacularly to the findings of these previous studies. In particular, the use of tiling arrays to identify transcribed fragments ('transfrags') of the human genome gives more complete and global coverage of the transcriptome than standard cDNA cloning and sequencing approaches, although the relationship between adjacent transfrags that derive from the nearby genomic region is initially uncertain (see below).

Cheng *et al*[6] isolated mature (ie post-spliced) cytoplasmic polyA+ RNA from eight human cell lines and interrogated tiling chips covering 10 human chromosomes in triplicate. They found that the detectable transfrags in each cell line covered on average 5% of the genomic sequences on the arrays. Cumulatively, 10% of the genomic sequences were represented in the polyA+ RNA fraction of one or more cell lines, indicating that many of the observed RNAs were cell type