European Journal of Human Genetics (2007) 15, 260–263
© 2007 Nature Publishing Group All rights reserved 1018-4813/07 $30.00
www.nature.com/ejhg

VIEWPOINT

Are genome-wide association studies all that we need to dissect the genetic component of complex human diseases?

Catherine Bourgain*[,1], Emmanuelle Génin[1], Nancy Cox[2,3] and Françoise Clerget-Darpoux[1]

[1]*INSERM U535, University Paris Sud, Villejuif F-94817, France;* [2]*Department of Medicine, University of Chicago, Chicago, IL, USA;* [3]*Department of Human Genetics, University of Chicago, Chicago, IL, USA*

With the availability of dense maps of anonymous and frequent SNPs spanning the whole human genome, genome-wide association studies are now becoming a reality. In this paper, we discuss the utility of these approaches to detect genetic risk variants involved in complex disease susceptibility and, in the best case scenario where a signal is detected, how helpful it will be to the understanding of the pathological process.

European Journal of Human Genetics (2007) **15,** 260–263. doi:10.1038/sj.ejhg.5201753; published online 13 December 2006

Keywords: genome-wide association study; power; modelling

Given the disappointing results obtained by the Human Genetics community through systematic linkage screenings of the genome, Risch and Merinkangas[1]argued that 'the future of the genetics of complex diseases is likely to require large scale testing by association studies'. If linkage studies have low power to detect common variants with small odds ratios (OR), they are also doing a poor job at detecting very frequent variants with high genotypic or allelic ORs, a situation where association tests might perform better. A good illustration of this point is provided by the VNTR flanking the insulin gene. The association between the locus and type I diabetes is easily detected, and the risk allele is frequent (about 70% in the general population) with an allelic OR around 3.9.[2] The distribution of alleles shared identical by descent (IBD) in sib-pairs, however, is very close to its null expectation and linkage is very difficult to detect. Even on large sib-pair samples, discordant results have been described.[3,4] An additional interesting feature of association studies is to allow a much more precise location of risk factors. Whereas the precision of location provided by linkage studies is sized in cM, it is sized in kb or even bp for association studies.

Technological developments have rapidly made large-scale association studies a reality. The HapMap project (www.hapmap.org) launched by an international consortium by the end of 2002 publicly released the first data in 2005. The first genome-wide association studies have been published recently.[5–7] Large-scale association studies using maps of anonymous and frequent SNPs are presented as the new tool that will accelerate the discovery of genes related to common diseases' (HapMap Press Release, 7 February 2005). In the present paper, we would like to debate around two questions that are central to the discussion on the power of genome-wide association studies: will most genetic variants involved in complex diseases be detectable by systematic association testing and, in the best-case scenario where a signal is observed, what will be its contribution to the understanding of the pathological process?

Will most genetic variants involved in complex diseases be detectable by systematic association testing?

The standard strategy for systematic association testing is to perform allelic tests using SNPs tagging common

*Correspondence: Dr C Bourgain, INSERM U535, University Paris XI, Genetique Epidemiologique et Structure des Populations Humaines, Hopital Paul Brousse, Batiment Leriche, B.P. 1000, Villejuif, Cedex 94817, France. Tel: +33 1 45 59 53 85; Fax: + 33 1 45 59 53 31; E-mail: bourgain@vjf.inserm.fr

Received 1 June 2006; revised 12 September 2006; accepted 27 October 2006; published online 13 December 2006

variants and/or haplotypes to limit both the multiple testing correction and the genotyping cost.

By construction, this approach has been recognized to have low power to detect the effects of rare variants, which are poorly tagged by common SNPs. Relatively rare variants are known to contribute to a variety of common, complex diseases. In Crohn's disease, the three risk alleles in the NOD2 gene have frequencies smaller than 5% in populations of European descent.[8,9] Population genetics modelling of the expected genetic variation at disease susceptibility loci has also suggested that rare variants may play an important role in complex disease susceptibility.[10,11] When selection is weak, as seems likely for many complex disease mutations, genetic drift becomes important, and the total frequency of susceptibility mutations is expected to vary widely among loci.

In contrast, almost all common SNP-based variation as well as other SNP-correlated common variants (including common deletions[12,13]) can be interrogated either directly or by surrogate SNPs or haplotypes. The issue is whether we will be able to design studies powerful enough to detect them. A number of variants with both high frequency and sizeable effects have been identified, including ApoE4 in Alzheimer's disease and cardio-vascular diseases or HLA factors in autoimmune diseases. Variants with similar sizeable effects may still have to be discovered. Some may have escaped identification by linkage studies because of a very high susceptibility allele frequency. Others may be located in regions already broadly detected by linkage as in the recent whole genome association successes.[6,7]

Aside from these large effect variants, common susceptibility alleles or haplotypes with quite modest ORs – in the range of 1.15–1.5, with the majority in the range of 1.15–1.25[14] have been shown to have reproducible effects (via meta-analysis) on various phenotypes. These variants with low marginal risks may, in fact, have strong effects on disease, but restricted to particular genetic background or environment. Indeed, complex models of interaction between genes, sex and environment have been extensively described in animal models for various complex traits.[15] Identifications of these variants – that may be far from exceptions in humans – may thus constitute important steps in the understanding of the physiopathology of the diseases, but very challenging ones.

Finally, to control for the multiple testing in genome-wide context, effects of SNPs may only be tested at the allelic[6,7] or simple haplotypic level (haplotypes made of 2 or 3 markers). But interaction between alleles or between distant SNPs within a susceptibility gene may be poorly detected at the allele level even in the presence of non-negligible effects at the genotype/haplotype level.

Whether we will be able to power studies to detect common risk alleles with $\mathrm{OR} < 1.25$ for all diseases is a critical issue. Apart from the challenge to collect huge samples without compromising on the quality of phenotypes, the crucial issue is whether human samples of sufficient sizes with homogeneous allele frequencies and pattern of linkage disequilibrium (LD) do exist. For some diseases, the inter-population variability may be too large to allow the recruitment of samples not subject to population stratification and large enough to detect variants with small effects. Similarly, once the variant is detected in a first study, it might be difficult to obtain samples to perform replication studies. First, these samples will need to be as large or even larger[16] and, second, the difference in both variant frequencies and LD pattern among populations may make the signal undetectable in other populations. The difficulty of replication is well illustrated in the recent work of Reich *et al.*[17] on multiple sclerosis, (MS) where an association on chromosome 1 in African-Americans is not replicated in another sample of Afro-Caribbeans. Consequently, it is not possible to decide whether the association described is driven by a causal variant or whether it is just another false positive result.

Altogether, we believe that these different elements invite cautiousness regarding the ability of genome-wide association studies to detect most genetic variants involved in complex diseases.

In the best-case scenario where a signal is observed, what will be its contribution to the understanding of the pathological process?

The main motivation behind looking for genetic risk factors for complex diseases is to get a better insight into the pathological process of the diseases. However, to go from an association signal to a deeper knowledge on the causal variants and their effect measurements is not that simple. The study of the HLA component in auto-immune diseases well illustrates the caveats encountered.

If the association between type I diabetes (IDDM) and the HLA class II regions has been known for more than 20 years, the causal variants of the region have still to be identified. A particular amino acid (non-'Asp57') of the DQβ chain was suspected for a while[18] as it is strongly associated with the disease: patients with an aspartic acid at position 57 in the DQβ chain are very rare. However, this hypothesis was rejected,[19] because it could neither explain the overall HLA genotype distribution nor the HLA haplotype sharing of affected sibs.[19] This example illustrates the importance of not only considering allelic association but also genotypic distributions and the importance of identifying models that can explain both the association and linkage information and not only one of them.

In celiac disease, genetic susceptibility in the HLA region is mainly explained by one DQ heterodimer whose two components (DQA1 and DQB1 alleles) can either be on the same HLA haplotype (*cis*-position) or one on each complementary haplotype (*trans*-position).[20] This functionally relevant model was found using both biological elements on the structure and function of HLA molecules[21] and the observed genotype distribution in patients. However, because it cannot clearly explain the HLA sharing among affected sibs and because there are geographical differences in the risks conferred by the DQ heterodimers,[22] this model cannot account for the whole HLA component of the disease, which suggests that other HLA factors must be involved.

In these two examples, the HLA component acts differently, through complex interactions of numerous variants. Deeper understanding could be obtained only through more detailed knowledge of gene diversity, patient genotype distributions, IBD sharing in affected siblings conditional on patient genotype, and biological understanding of the function of genes in the HLA region. Detection of an association was only the very first step.

In a recent study, Lincoln *et al.*[23] used a large-scale SNP association strategy for studying the involvement of the MHC in MS. The strongest association is observed with haplotype blocks in the HLA class II region and, conditional on this association, there is no evidence for other HLA-region risk factors for MS. These results are an additional illustration of the limits of strategies relying on only association information. Despite the use of large numbers of simplex and multiplex families with dense SNP typing, the study just ends at a result which has been known for many years.

HLA is certainly a unique region of the genome for its genetic complexity (level of polymorphism, pattern of LD and clustering of many functionally related genes). Therefore, the challenges encountered in this region are not fully representatives of the ones that we face elsewhere on the genome. Still, the HLA region is one of the most extensively studied and one exhibiting the strongest genetic factors for many complex diseases described so far. This, together with the accumulation of experiences in other regions of the genome, makes us believe that, though different, these non-HLA challenges may not be less complex.

The availability of tools (e.g. high-throughput genotyping platforms) and information (e.g. a haplotype map of the human genome) are a valued and welcome addition to the armamentarium geneticists can use to identify and characterize the genetic component to common diseases with complex transmission. But we should resist the temptation to assume that the models most favourable for the utility of these new tools must be the right models for all the variants involved in complex diseases just because our studies will work better if these most favourable models are correct. We can ill afford to go from a generation of underpowered and ultimately disappointing linkage and family studies directly to a generation of underpowered and disappointing association studies.

At the same time, the impressive diversity of causal variant models already identified favours continued diversification of strategies. Identification of biologically relevant novel pathways of disease susceptibility followed by candidate gene strategies allowing for the simultaneous consideration of these connected genes may be quite effective at identifying new variants affecting susceptibility. Linkage data contain considerable additional information capable of aiding the discrimination of causal variation from the nearby variation that is merely in LD. Consequently, linkage studies may still pay dividends even if the newer tools are as successful as we all hope, and we believe efforts to recruit family data should not be given up too quickly in favour of only large samples of cases and controls. Linkage and association studies should be considered as complementary strategies and not as concurrent.

References

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 Bell GI, Horita S, Karam JH: A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 1984; **33**: 176–183.
3 Cox NJ, Wapelhorst B, Morrison VA *et al*: Seven regions of the genome show evidence of linkage to type 1 diabetes in a consensus analysis of 767 multiplex families. *Am J Hum Genet* 2001; **69**: 820–830.
4 Concannon P, Gogolin-Ewens KJ, Hinds DA *et al*: A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nat Genet* 1998; **19**: 292–296.
5 Maraganore DM, de Andrade M, Lesnick TG *et al*: High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet* 2005; **77**: 685–693.
6 Klein RJ, Zeiss C, Chew EY *et al*: Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**: 385–389.
7 Herbert A, Gerry NP, McQueen MB *et al*: A common genetic variant is associated with adult and childhood obesity. *Science* 2006; **312**: 279–283.
8 Lesage S, Zouali H, Cezard JP *et al*: CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* 2002; **70**: 845–857.
9 Vermeire S, Wild G, Kocher K *et al*: CARD15 genetic variation in a Quebec population: prevalence, genotype-phenotype relationship, and haplotype structure. *Am J Hum Genet* 2002; **71**: 74–83.
10 Pritchard JK, Cox NJ: The allelic architecture of human disease genes: common disease-common variant or not? *Hum Mol Genet* 2002; **11**: 2417–2423.
11 Pritchard JK: Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001; **69**: 124–137.
12 McCarroll SA, Hadnott TN, Perry GH *et al*: Common deletion polymorphisms in the human genome. *Nat Genet* 2006; **38**: 86–92.
13 Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA: Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* 2006; **38**: 82–85.
14 Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003; **33**: 177–182.

15 Mackay TF: The genetic architecture of quantitative traits: lessons from Drosophila. *Curr Opin Genet Dev* 2004; **14**: 253–257.

16 Terwilliger JD, Haghighi F, Hiekkalinna TS, Goring HH: A bias-ed assessment of the use of SNPs in human complex traits. *Curr Opin Genet Dev* 2002; **12**: 726–734.

17 Reich D, Patterson N, De Jager PL *et al*: A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat Genet* 2005; **37**: 1113–1118.

18 Todd JA, Bell JI, McDevitt HO: HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* 1987; **329**: 599–604.

19 Clerget-Darpoux F, Babron MC, Deschamps I, Hors J: Complementation and maternal effect in insulin-dependent diabetes. *Am J Hum Genet* 1991; **49**: 42–48.

20 Sollid LM, Thorsby E: The primary association of celiac disease to a given HLA-DQ alpha/beta heterodimer explains the divergent HLA-DR associations observed in various Caucasian populations. *Tissue Antigens* 1990; **36**: 136–137.

21 Sollid LM, Thorsby E: HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis. *Gastroenterology* 1993; **105**: 910–922.

22 Margaritte-Jeannin P, Babron MC, Bourgey M *et al*: HLA-DQ relative risks for coeliac disease in European populations: a study of the European Genetics Cluster on Coeliac Disease. *Tissue Antigens* 2004; **63**: 562–567.

23 Lincoln MR, Montpetit A, Cader MZ *et al*: A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis. *Nat Genet* 2005; **37**: 1108–1112.