HOW SERIOUSLY SHOULD WE TAKE RISK PREDICTIONS FOR MULTIFACTORIAL ILLNESSES?

Proposed for discussion by F. Clerget-Darpoux and A. Cambon-Thomsen

This text, initially written in French, has already been endorsed by a community of 11 professional societies* involved in genetic issues

The progress made by geneticists over the last three decades with regards to disease aetiology has been tremendous. In particular, aetiological heterogeneity has been shown for many common diseases with evidence for monogenic sub-entities (Alzheimer's disease, Parkinson's disease, diabetes, breast cancer, colon cancer etc.). Using molecular biology, genes involved in the monogenic forms of diseases have been identified, allowing their molecular diagnosis, better understanding of their physiopathology and research into new therapies.

However, these monogenic sub-entities most often represent very small proportions of the corresponding diseases. Thus, for the majority of patients, the primary causes of the disease are unknown and may be both multiple and heterogeneous. These other forms of the disease, known as "multifactorial", involve both genetic and environmental factors. The remarkable success of the human genome project generated great hopes. It was thought that the identification of genetic factors involved in multifactorial diseases would allow the risks for a given person to be predicted through analysis of their genome. Internet sites and numerous publications offer an evaluation of the risks of developing various diseases. Interpretation of the results presupposes that the genetic factors are numerous, each with minor, independent, cumulative effects, and that no interaction with environmental factors is involved. If this hypothesis were proven, one could imagine that genome wide studies would end up cataloguing all the genetic factors involved in each disease and would thus allow people to be classed according to their personal risk factors, as compared to a reference population.

However, in contrast to monogenic diseases or monogenic sub-entities of a disease, the primary cause(s) of a multifactorial disease may be non genetic. This is clear for some

diseases for which the environmental factor is well known. Thus, for leprosy, the genetic differences described for a predisposition to develop this disease only come into play when the Hansen bacterium is encountered. For most multifactorial diseases, the environmental factors are unknown but the exposure to environmental factors may be essential in initiating the disease. More generally, the missing information may be essential for risk prediction; this is not quantifiable unless simplistic and questionable assumptions are made. More details are provided in the text annexed to this document.

In conclusion:

- While genome wide studies provide an essential contribution to scientific knowledge of multifactorial diseases, the isolated use of information provided by them lacks any capacity to predict future onset of those diseases. It leads to an erroneous perception of the risk for the individual.
- The scientific community has a duty not to provide justification for individual risk predictions for multifactorial diseases based solely on genetic information.
- It is important to inform the public clearly and correctly on the advances in genomics and to debate the limits of their applications.

* signed by

- SFGH (Société Française de Génétique Humaine; French human genetics society)
- FFGH (Fédération Française de Génétique Humaine; French human genetics federation)
- ACLF (Association des Cytogénéticiens de Langue Française; Association of French-speaking cytogeneticists)
- ANPGM (Association Nationale des Praticiens de Génétique Moléculaire; National molecular genetics practitioners association)
- AFGCCG (Association Française de Génétique Clinique et de Conseil Génétique, French association for clinical genetics and genetic advice)
- AFCG (Association Française des Conseillers en Génétique, French genetic advisers association)
- SOFFOET (Société française de foetopathologie, French foetopathology society)
- CPEGM (Collège des Praticiens et Enseignants de Génétique Médicale, College of medical genetics practitioners and teachers)
- SFG (Société Française de Génétique; French genetics society)
- SFSP (Société Française de Santé Publique; French public health society)
- ADELF (Association des Epidémiologistes de Langue française; Association of French-speaking epidemiologists)

Annexe

Scientific arguments backing up the text co-authored by the learned societies and professional bodies for genetics and human genetics

1) MULTIFACTORIAL DISEASES AND SMALL INDEPENDENT EFFECTS

The definition of a multifactorial disease is so unclear that we often qualify it as a "complex disease" or "common illness". While the term "complex" is not wrong, it is not particularly specific to multifactorial diseases; "common" on the other hand, is inappropriate because the rates of occurrence of these diseases is extremely variable, some have rates of occurrence lower than 1‰ (autism, multiple sclerosis) and others greater than 10% (type 2 diabetes in over 65 year-olds, Alzheimer's disease in the over 80s).

Genetic analyses generally assume that a multifactorial disease is due to numerous genetic factors with minor, independent and cumulative effects (the disease appears above a certain threshold) and to environmental influences. This model, the so-called polygenic model, along with the concept of heritability was introduced by Fisher in 1918. In animal or vegetable species, for whom we can control both mating and their environment, this model is very useful for the selection of certain quantitative characteristics. This model, presented as an alternative to the monogenic model, where the principal cause of a disease is mutation of a single gene, is based on the hypothesis that genetic factors do not interact either with each other or with the environmental factors. However, for human diseases, this hypothesis is clearly false for illnesses that are only found in a specific environment, such as leprosy and coeliac disease, and is also unrealistic for the multifactorial diseases for which the environmental factors are poorly described, or are completely unknown. Taking the polygenic model for multifactorial disease back to the drawing board leads us to question the risk values and heritability concepts that are linked to it.

2) MULTIFACTORIAL DISEASES AND HERITABILITY

In a given model, heritability of a trait measures the contribution of genetic variability to total variability of the trait within a population. Its estimation assumes an absence of

interaction and of correlation between genetic and environmental factors within the population.

Idea No.1: the heritability of a disease is only valid if the hypotheses stated hold true.
For multifactorial diseases, we do not know the number of factors involved, nor the importance of their individual effects or how the genetic and environmental factors interact.
Idea No.2: Measuring heritability, supposing it is possible, is not an indicator of the comparative importance of genetic and environmental factors in the physiopathology of the disease. The confusion between contributions from variations and contributions from factors in the interpretation of heritability was decribed well before genomic studies were initiated (see Lewontin, 1974).
A strong heritability does not imply that genetic factors are important and that the environment only plays a minor role in the physiopathology of a disease. It could in fact indicate, within the population concerned, that there is little variability in exposure to environmental factors involved in the disease. When all the members of a village are exposed to the bacterium causing leprosy, the difference between those who will go on to develop the disease and those who will not, will be explained by genetic differences. Nevertheless, the determining factor remains the exposure to the bacterium. Similarly, for celiac disease (or gluten intolerance), everyone is exposed to gluten and only a small proportion of the population is affected. Genetic risk factors appear here to be essential; however, gluten is a determining causal element because its elimination from the diet is enough to cure the sufferer.
The use of erroneous interpretations of heritability is obvious on the internet sites of some companies offering genetic screening. Thus, on 23andme's site we can read that environmental factors play a more important role than genetic factors in the development of type 2 diabetes because its heritability is only 26%. Similarly, it is stated without reference, that the risks of celiac disease or age-related macular degeneration are mainly attributable to genetic factors.

Idea No.3: the limits of the notion of heritability and of its interpretation are also valid for the, very fashionable, "missing heritability" (Maher et al., Nature, 2009). According to those who promote this concept, missing heritability would reveal the "ground remaining to be covered" for the identification of all the genetic risk factors. Thus, a recent article in Nature (Manolio et al., 2009) reports a table of the proportion of explained heritability for several traits or diseases. For example, 5 genetic risk factors associated with age-related macular degeneration (AMD) would explain 50% of its heritability and the authors conclude

that work must continue to identify the genetic risk factors for the remaining 50%, despite the fact that the missing information is unmeasurable.

3) MULTIFACTORIAL DISEASES AND FAMILIAL CONCENTRATION.

For multifactorial diseases, a higher risk has been noted for relatives of persons with the disease than for a random person from the general population. The ratio of risk of disease for a sibling (brother or sister) of a person with the disease and the risk of disease in the general population (relative risk λs) is often used as an argument in favour of the existence of genetic factors and even to measure the importance of these.

- Idea No.4: The relative risk λs is very variable from one disease to another. For example, it is estimated at between 20 and 40 for multiple sclerosis (Ebers et al., 1995) and at around 1.8 for type 2 diabetes (Weijnen et al., 2002). A very small risk of developing the disease can be associated with a very large λs risk (and vice-versa). In the case of multiple sclerosis, the relative risk for a sibling is high (20 to 40) but his or her risk of developing the disease is only 2 - 3% because the rate of occurrence of the disease in the general population is between 1/1000 and 1/2000. On the contrary, for type 2 diabetes, the relative risk for a sibling is low (only 1.8) but his or her risk of developing the disease at 55 years of age is 23%, because in the general population this risk is already 13% for that age-group.

- Idea No.5: a high relative risk does not necessarily mean significant genetic contributions to the physiopathology of a disease. A part of the increase in risk for a relative, or even all of it, can be explained by family members being exposed to similar environmental factors. The increase in obesity noted for brothers and sisters of an obese child may be due to shared genetic factors, but almost certainly is also related to the fact that they share the same eating habits.

4) MULTIFACTORIAL DISEASES AND RISK CALCULATION

Genome wide association studies (GWAS) are today one of the most widely used tools for the analysis of multifactorial diseases, and it is based on their results that the internet sites and numerous publications calculate the risks of developing these diseases.

The human genome varies from one individual to the next at a very large number of positions (several tens of thousands per chromosome), these variations are also called polymorphisms,

among which the SNP (Single Nucleotide Polymorphisms) which are present in one of two forms that we will call A and a.

Association studies seek to determine the SNPs for which the frequencies of A and a are different between people presenting a disease and healthy people (controls). The significance of this difference is often measured as the ratio of the risk of developing the disease for a carrier and non carrier of the risk form (Odds Ratio or "OR"). A SNP which shows a significant difference (or an OR different from 1) indicates that in the genomic neighbourhood one or several genetic factor(s) involved in the physiopathological process is present.

- Idea No.6: The OR is a signal that only imperfectly reveals the complexity of the neighbouring genetic factors involved in the physiopatological process. Or, to put it another way, the value of the OR might be significantly higher if it was calculated directly using the genetic factors involved in the disease. This is well illustrated in coeliac disease for which the OR of the most strongly associated SNP in the HLA region is 7 (Van Heel et al., 2007) while it is 25 when the polymorphism of two genes in the HLA region is taken into account simultaneously (Margaritte-Jeannin et al., 2004).

More generally, there is a huge gap between the detection of a significant OR and the identification of the genetic variations actually involved in the pathological process. It is important to note the few advances that have been made into understanding the mechanisms underlying the associations observed over 35 years ago between HLA antigens and a large number of diseases (type 1 diabetes, multiple sclerosis).

- Idea No.7: The calculations involving a whole range of SNPs associated with a disease are based once again on the hypothesis that there are no interactions between the genetic factors. However, these obviously exist as part of complex processes which underlie, for example auto-immune or metabolic diseases.

- Idea No.8: To a high OR.can correspond a low risk of developing a disease

For example, if we consider celiac disease in Europe, the OR for carriers of functional polymorphisms of the HLA system is 25 but the risk of developing the disease is only 1%. Indeed, the risk is proportional to the frequency of occurrence of the disease – around 1/300 in Europe – but also inversely proportional to the frequency of HLA polymorphisms involved which is high - around 25% of the European population as a whole are carriers.

Idea No.9: Aetiological heterogeneity of so-called multifactorial diseases.

Genome wide association studies have revealed, for a good number of multifactorial diseases, low-intensity signals (low OR for the associated SNPs) thanks to impressive numbers (several thousands) of people with the disease. This implies that very little selection was carried out during sampling in terms of clinical and environmental homogeneity. As a consequence, it is very likely that the aetiological heterogeneity is very high in these samples. Thus the analysis of 14,586 cases of type 2 diabetes almost certainly involved both patients with no weight problems and overweight patients. Weijmens et al.'s analysis (Weijmen et al., 2002) suggests, however, that they could correspond to different aetiologies. Similarly, genome wide studies of cancers probably mix together monogenic and multifactorial forms of the same cancer type.

- Idea No.10: Importance of the missing information.

The information provided by the SNPs associated with different diseases in the genome wide studies may be very poor with regard to other types of information (family history, environmental history etc.) (Bourgain et al., 2007, Clerget-Darpoux & Elston, 2007).

In contrast with what those who measure "missing heritability" like to suggest, there is no way of measuring either the importance or the nature of the missing information. A risk may be completely changed by one new piece of information. A person carrying the highest risk forms for the type 2 diabetes-associated SNPs has, in fact, less risk of developing the disease if he or she is slim and without any family history of the disease than a person not carrying the risk forms of these SNPs but who is obese and has a family history. Similarly, on the basis of a genome analysis, someone could be declared to be at risk of developing leprosy although there is absolutely no risk if he or she is not exposed to the mycobacterium.

- **References:**

1. Fisher: The correlation between relatives on the supposition of mendelian inheritance. Transactions of the Royal Society of Edinburg, 1918, 52: 399-433.
2. Lewontin: The analysis of variance and the analysis of causes. Am J Hum Genet, 1974, 26: 400-401.
3. Bomprezzi R, Kovanen PE Martin R: New approaches to investigating heterogeneity in complex traits. J Med Genet, 2003, 40: 553-559.
4. Maher B: Personal genomes: the case of the missing heritability. Nature, 2008, 456: 18-21.

5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: Finding the missing heritability of complex diseases. Nature, 2009, 461(7265): 747-53.

6. Ebers, G. C., Sadovnick, A. D. & Risch, N. (1995) A genetic basis for familial aggregation in multiple sclerosis. Nature, 1995, 377:150-151.

7. Weijnen CF, Rich SS, Meigs JH, Krolewski AS, Warram JH: Risk of diabetes in siblings of index cases with type 2 diabetes: implications for genetic studies. Diabetes, 2002, 19:4 1-50.

8. DA van Heel, L Franke, K A Hunt, R Gwilliam, A Zhernakova, M Inouye, M C Wapenaar, M Barnardo, G Bethel, G Holmes, C Feighery, D Jewell, D Kelleher, P Kumar, S Travis, J Walters, D S Sanders, P Howdle, J Swift, R J Playford, W M McLaren, M L Mearin, C J Mulder, R McManus, R McGinnis, L R Cardon, P Deloukas, and C Wijmenga : A genome-wide association study for celiac disease identifies risk variants in the region harboring *IL2* and IL21. Nature Genetics, 2007, 39: 827 – 829.

9. P Margaritte-Jeannin, M-C Babron, M Bourgey,A Louka, F Clot, S Pecopo, I Coto, J-P Hugot, H Ascher, L Sollid, L Greco, F Clerget-Darpoux: A study of the Europen Genetic cluster on Coeliac disease Tissue antigens, 2004, 63, 562-567

10. Bourgain C, Genin E, Cox N, Clerget-Darpoux F: Are genome-wide association studies all that we need to dissect the genetic component of complex human diseases? Eur J Hum Genet, 2007,15(3): 260-263.

11. Clerget-Darpoux F, Elston RC: Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. Hum Hered, 2007, 64: 91-96.

This text is the result of the reflections and summarisations of a committee set up by the SFGH (Société française de génétique humaine; French human genetics society) consisting of:

- Alain Bernheim (INSERM research director, IGR Villejuif, SFGH vice president)
- Catherine Bourgain (INSERM researcher)
- Anne Cambon-Thomsen (INSERM research director, Toulouse)
- Françoise Clerget-Darpoux (INSERM research director)
- Pierre Darlu (CNRS research director)
- Marc Fellous (professor emeritus, université Paris VII-Diderot)
- Josué Feingold (emeritus research director INSERM)
- Gérard Huber (psychoanalyst, author, president of the « santé-solidarité_prospective 2100 » (health solidarity prospective 2100) club)
- Jean-Claude Kaplan (professor emeritus, université Paris V-Descartes)
- Jean-Louis Serre (professor, université de Versailles, SFGH president)