On Mendelian Randomisation

François Cambien, INSERM U525

According to recent statements, "*genetic association studies are undergoing a renaissance under the Banner of Mendelian randomization...Mendelian randomization offers a potentially exciting perspective on gene-disease association in the genome area "* (J Little, MJ. Khoury). In this commentary, we examine the alleged novelty of this approach and investigate the foundations of the expectations it creates

"In a correctly designed genetic association study, the laws of Mendelian genetics ensure that comparison of groups of individuals defined by genotype is equivalent to a randomized comparison". Because of this property of a genetic polymorphism, other traits, whether genetic or non-genetic (environmental, behavioral ...) should be distributed randomly across genotypes, except those that are affected by the polymorphism. As a consequence, comparisons of phenotypes across genotypes should not be biased by confounding factors and should provide insights into causal pathways.

"*To test whether a risk factor has a causal relation to disease risk, we can look for polymorphisms that affect the risk factor or the metabolic pathway on which the action of the risk factor depends, then examine the effects of these polymorphisms on disease risk in a large case-control study"* (D. Clayton and P.M. McKeigue)

the different sources of association

It may be useful to examine the various sources of association in association studies. In epidemiological studies, an association between a factor and a disease may be the result of chance (not replicable), may reflect the effect of confounding factors (replicable but disappear after adjustment on confounding factors), may be the consequence of the disease (reverse causation), or may reflect the causal role of the factor (among others) on the disease.

Situation 1 - chance. Because a chance finding is not replicable, this kind of association could be easily discarded. However non-replicability is not always easy to test and non-replication does not imply that a finding was only the consequence of chance, especially when dealing with marginal effects in the context of a network in interacting factors.

Situation 2 - confounding. is not obvious to resolve, because confounding factors are not necessarily known, or may not be measured. Randomisation of the study population in groups, differing only for the factor of interest leads to a random distribution of known and unknown confounders in the different groups, therefore removing any bias that might result in a spurious finding. This is what can be obtained in clinical trials through random assignment of treatment. There is no way in observational studies to obtain such randomisation, except in very particular situations (for example the familial setting suitable for TDT, where the expression Mendelian randomization makes sense, but certainly not in relation to genotypes as will be discussed below). In observational studies appropriate study design and statistical adjustment provide a partial protection against confounding factors. Within a study, not taking into account the possible stratification of sub-populations with different distributions of risk factors or genotypes and disease may lead to spurious associations, this particular type of confounding may be particularly important in genetic association studies.

Situation 3 - reverse causation. is especially worrying in retrospective case-control studies, but it exists also in prospective studies, because preclinical forms of disease may exist in healthy individuals, that may affect a number of phenotypes and are associated with an increased risk of disease (a typical example is atherosclerosis and myocardial infarction). Genetic studies are less prone to this bias, because the genotype is unaffected by the presence of the disease. However the difficulty reappear when the focus of an analysis is on interaction with non-genetic factors which themselves could be affected by the disease. Changes occurring as a consequence of the

disease (Biological markers) may be used as indicators of the presence and evolution of the disease. In addition these post-disease changes may play a role in the disease process (situation 4), either by counteracting it or accelerating its evolution, and the importance of these changes may depend on a number of other players including gene polymorphisms. For example: 1. we may be interested in the identification of factors that favor the apparition of cancer (situation 4) or we may be interested by factors that affect the 'host' response which may considerably influence the severity or propensity to metastasis of the cancer (both situation 3 and 4), 2. taking a drug is a response to a disease and the effect of the drug on the disease may be dependent on other factors including genetic ones (pharmacogenetics).

Situation 4 - causality. Here the identified factor is a cause of the disease or of its evolution. The ultimate goal of association studies is to identify such factors. For that purpose, studies and inference processes should be reliable, implying that they minimize situations 1-3 above.

Evidently any combination of the 4 situations is possible, for example there may be a two-ways relationship between a factor and a disease.

The use of candidate gene association studies to identify causal factors is not new

The systematic search for associations between genetic polymorphisms and complex traits constitutes an approach to disease causality that may circumvent the biases inherent to phenotype analysis. The invariance of the genotype providing a relative protection against biases that might arise from confounding factors or from retro-causation. This has been the main reason for a number of scientists to conduct association studies, well before reinvented Mendelian randomization (It may be in the nature of Mendel of always being rediscovered).

The interest for gene polymorphisms in common cardiovascular diseases research started really to increase at the end of the 1980s, when methods became available, first based on RFLPs and then on PCR, allowing to genotype large samples of subjects at an acceptable cost. In cardiovascular research, the candidate gene approach had a fairly strong support and its connection with the concept of intermediate phenotype was natural. The example of Familial Hypercholesterolemia (FH) had been very instructive, the chain of events leading to CHD in FH patients had been reconstituted, thanks in particular to the works of Goldstein and Brown.

Mutation of the LDL receptor gene --> strongly affect the fonction of the LDL receptor --> this leads to an important rise of circulating LDL cholesterol --> which in turn is associated with a major increase in the risk of CHD.

The cardiovascular epidemiologist interested in biology and genetics was aware that : 1. plasma LDL cholesterol is a major risk factor for CHD, 2. FH explains a small part of its variability in the population, 3. Familial studies show that LDL cholesterol is highly heritable, 4. ApoE e2,3,4 phenotypes (genotypes) are frequent and are significantly associated with LDL cholesterol (Sing, Davignon), 5. the different proteins involved in LDL metabolism are rather well-known, offering a good rationale for candidate gene selection.

These premisses dictated a rather straightforward research programme: 1. The variability of the candidate genes had to be assessed, 2. possible associations between identified polymorphisms and plasma LDLc and other related intermediate phenotypes had to be tested, 3. finally these polymorphisms had to be tested in relation to CHD.

This programme took momentum at the beginning of the 1990s. In parallel other biological systems of interest for CHD started to be investigated, genes involved in HDL metabolism, thrombosis, hypertrophy, vasomotricity, hypertension, glucose metabolism, inflammation progressively entered the scene, their variability was explored and they were analysed in relation to a number of intermediate phenotypes and cardiovascular disease end-points.

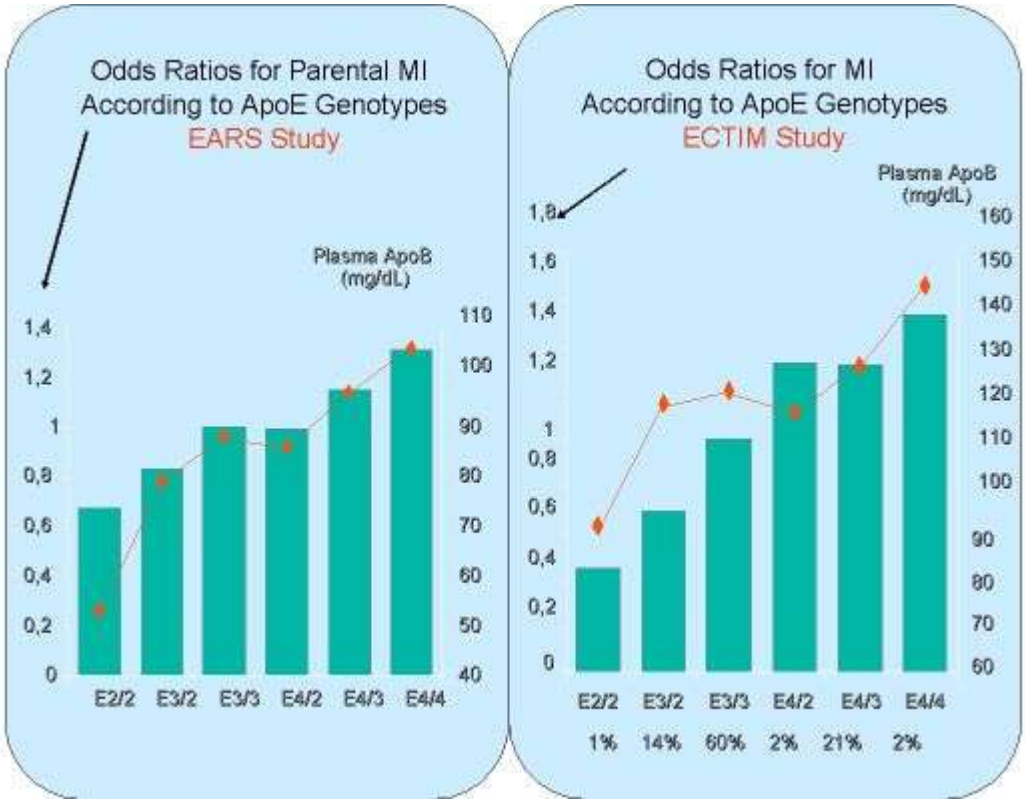

Figure 1. The ApoE polymorphisms considerably affects the recognition of ApoB/E containing lipoproteins by the ApoB/E receptor on the surface of hepatic and peripheral cells. This differential recognition influences the circulating level of LDL cholesterol and ApoB. This is true in young adults (EARS Study, Tiret et al. 1994) and in older subjects (ECTIM Study, Luc et al. 1994). It can be seen that the relationship is graded, increasing from E2/2 to E4/4 and that the relationship with MI (ECTIM) and parental history of MI (EARS) is very coherent across studies and between intermediate phenotype (ApoB) and disease.

The candidate gene approach incorporated the idea of exploring metabolic pathways or systems assumed to be involved in the pathophysiology of CHD and the concept of intermediate phenotype was tightly linked to this idea. A number of studies were designed to implement this approach. Candidate gene polymorphisms, intermediate phenotypes and diseases were frequently investigated simultaneously to assess whether associations were consistent (Cambien et al 1997). Obviously if a gene polymorphism is associated with an intermediate phenotype, for example the circulating product of this gene and both the polymorphism and the intermediate phenotype are associated with the disease endpoint in a coherent way (not in opposite directions), the results may be said to be internally consistent. This is not a proof of causality but it is better than an isolated association between the polymorphism and the disease. Actually internal consistency is one of the criteria used to support causality in epidemiological studies. This is why as much information as possible should be collected a *priori* that could be used to assess consistency, intermediate phenotypes, but also for example family history (see figure 1)

Mendelian randomization

The potentiality of genetic analysis for discovering new causal pathways and mechanisms of disease is currently being rediscoved by epidemiologists and statisticians under the name of "Mendelian randomization" which in an interesting and well documented paper is proposed to provide a new way to understand the environmental determinants of disease, ***"If polymorphisms produce phenotypic differences that mirror the biological effects of modifiable environmental exposures which in turn alter disease risk, the different polymorphisms should themselves be related to disease risk to the extent predicted by their influence on the phenotype."*** (G. Davey Smith and S. Ebrahim)

Are polymorphisms transmitted independently of one another

The Mendelian laws alluded to by Clayton and McKeigue is the random assortment of alleles during meiosis which supposedly imply that the inheritance of a particular allele is independent of the inheritance of alleles at other loci. Because it was in the pre-gene era, Mendel was speaking of differentiating characteristics, which we

might call variable traits, and his inference was correct as long as the each trait was encoded by a single gene and the traits were not genetically linked. Obviously this cannot be assumed to be so a priori for any pair of traits, linkage disequilibium (LD) exists among loci that are close to each other on the genome (resulting in non independent genotypes). LD may extend over wide regions in the genome (the best known is the HLA region on chromosome 6). It might be argued that the genome being 3 billions base pairs long and LD extending over a few tenths or hundreds Kb, and less commonly over millions Kb, genotypes on different genes of interest for a phenotype should be approximately randomly distributed. Unfortunately, genes that are relevant for a particular function or disease are not randomly distributed across the genome, clusters of genes within regions in LD are frequently not functionally independent (HLA is an obvious example, but there are many examples relating to other biological systems). This is a *fortiori* the case for polymorphisms within a single gene and particular polymorphisms may have no marginal effect (when they are examined one at a time) while they have different effects on different background haplotypes.

Non-genetic characters are not evenly distributed across genotypes

If comparison of genotypes was equivalent to a randomized comparison, we would expect that non-genetic characters are randomly distributed across genotypes except those that are affected by the genotype. The problem is that the polymorphisms we investigate and those that are in LD with them may directly or indirectly affect a large number of known and unknown factors, some of which being able to modify the relationship between the genotypes and the phenotypes of interest. For example, we expect that FH patients are more frequently treated with hypolidemic drugs than normocholesterolemic subjects and we expect that carriers of the ALDH2 variant associated with reduced alcohol tolerance will drink less alcohol than non-carriers.... We may safely assume that factors that are irrelevant for the genotype/phenotype association under study are evenly distributed across genotypes, while this may not be true for known and unknown factors that are relevant for the disease.

Davey Smith and Ebrahim acknowledge that "Mendelian randomization in genetic association studies is approximate, rather that absolute". But the degree of this approximation is impossible to specify and may considerably vary from situation to situation and the loss of information cannot be assumed to be random. As a consequence the term randomization is misleading and we are back to a more classical situation where we are reduced to perform statistical adjustment on confounding factors that we know, hoping that those we do not know are not important.

Will gene polymorphisms help us to identify biological pathways of importance for disease

The disease model assumed by the proponents of "Mendelian randomization" is simple and actually can be discussed independently of the concept of "Mendelian randomization" within the more traditional idea of consistency. The 3 components of the triangular relationship are the genotype (G), the intermediate phenotype (IP) and the disease (D). Assuming a simple causal pathway model G-->IP-->D, one wants to test whether the set of relationships [G:IP, G:D, IP:D] is consistent, i.e. whether [G:D] can be predicted from [G:IP, IP:D] from a simple statistical model. For example, if an IP mean level is 40% higher in genotype G1 than in genotype G0, and a 40% increase in IP is associated with an increased risk of Disease of 20% then we expect that the risk of disease in G1 subjects is 20% higher than in G0 subjects.

A simple G-->IP-->D model may be appropriate in a few instances, for example for monogenic disorders, but it is not in general valid even as an approximation for a complex trait (if it was, the trait would not be complex anymore since the genotype-disease relationship could be modeled as a set of additive effects). The genotype may affect several intermediate phenotypes and disease end-points (pleiotropy), The disease may influence the intermediate phenotype and even the [G:IP] relationship (retro-causation), the variability of both the intermediate phenotype and the disease may be affected by partially overlapping sets of polymorphisms and biological interactions may be assumed that simple additive or multiplicative statistical modeling ignores... In the example of

the relationship between a beta fibrinogen gene polymorphism, plasma fibrinogen and CHD, the following causal pathway G-->IP-->D-->IP is more plausible than the postulated G-->IP-->D pathway, resulting in an inconsistent set of associations under the simple causal pathway model. This finding is not unexpected because atherosclerosis generates or (for many scientists is) an inflammatory states, the expression of a large number of inflammatory factors are induced in atherosclerosis lesions and if there is a true G-->IP-->D effect involving these factors, it would likely be partially or totally masked by the variability of the IP that is a consequence of the disease state.

Inconsistency may be the rule rather than the exception

During the last 15 years, guided by the vast knowledge accumulated on the pathophysiology of the disease, polymorphisms of more than 100 candidate genes have been explored in relation to atherosclerosis and its complications. One of the main finding emerging from these studies is the inconsistency of the simple pathway model. A surprisingly large fraction of polymorphisms (G) affecting genes for which 'proximal' (Intermediate) phenotypes (IP) exist, such as RNA or protein expression, quantity or function, are strongly associated with these phenotypes. Concurrently the results of a large number of studies also indicate that associations between polymorphisms and, more 'distal', clinically relevant phenotypes and disease end-points (D), are much weaker or absent, and often inconsistent across studies. Although there may be exceptions, there is no reason to believe that polymorphisms of the still unexplored candidate genes or of unknown genes that are expected to be identified through whole genome analysis should be different from the already investigated ones with regard to the strength and complexity of their association with disease. Available data appear therefore to be inconsistent with the simple causal pathway discussed above. However they are not inconsistent with a more realistic model of biological process assuming that the variable genes are component of complex biological systems (BS) or networks in which the weaker functional consequences of genetic variability on distal than on proximal phenotypes is the consequence of the increasing complexity of the network as the number of intermediate steps and possible pathways increase.

How to approach the genetics of complex traits and what is the place of epidemiologists in this research ?

One of the major strength of epidemiology is that it provides simple risk factor models of practical interest for preventive or medical purposes. This simplicity however severely limits the domain of relevance of epidemiology, and forcing simplicity when the data are complex may not be appropriate. Genetic association studies have been envisaged as a possible way to go beyond the limitations faced by epidemiology, by providing a rational approach to the identification of new causal factors contributing to disease etiology. The large number of studies that have been conducted have shown that approaching the relationship between gene polymorphisms and complex diseases as a simple extrapolation of classical epidemiology and Mendelian genetics was untenable. Reflection around “Mendelian randomization” and its potential is based on simplifying assumptions that will be rarely appropriate. What matters is the phenotype, new phenotyping technologies are becoming available that may considerably change the perspective of epidemiological research, and hopefully will shift the focus of epidemiology from quantity to quality. Very important to understand the genetics of complex traits is the emerging science of biological systems. Biological system genetics is not focused on the analysis of single genes or proteins but on the analysis of whole biological systems in which genetic interactions among components is not ignored and the phenotypes are crucial. Approaching the genetics of complex traits requires a broad range of expertise: molecular, mathematical, epidemiological geneticists, molecular and cellular biologists, epidemiologists, computer scientists, specialists in high throughput technologies must work together. Being at the source of the data, epidemiologists occupy a key position in such collaborative framework, they should do their best to introduce new reliable and relevant measurements in their studies, this obviously will be at the expense of large numbers.

References

Cambien F., Poirier O., Mallet C., Tiret L. Coronary heart disease and genetics: an epidemiologist's view. Mol Med Today. 1997;3(5):197-203.

Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001; 358: 1356- 1360

Davey Smith G, Ebrahim S. 'Mendelian randomization' : can genetic epidemiology contribute to understanding environmental determinants of diseases ? Int J Epidemiol 2003; 32:1-22.

Little J. and Khoury MJ. Mendelian randomization: a new spin or real progress ? Lancet 2003; 9388: 930-1

Luc G, Bard JM, Arveiler D, Evans A, Cambou JP, Bingham A, Amouyel P, Schaffer P, Ruidavets JB, Cambien F, et al. Impact of apolipoprotein E polymorphism on lipoproteins and risk of myocardial infarction. The ECTIM Study. Arterioscler Thromb. 1994; 14: 1412-9.

Tiret L, de Knijff P, Menzel HJ, Ehnholm C, Nicaud V, Havekes LM. ApoE polymorphism and predisposition to coronary heart disease in youths of different European populations. The EARS Study. European Atherosclerosis Research Study.Arterioscler Thromb. 1994; 14: 1617-24.