Commentary

Human Heredity

Hum Hered 2015;79:1–4
DOI: 10.1159/000370327

Published online: February 6, 2015

The Missing Heritability Paradigm: A Dramatic Resurgence of the GIGO Syndrome in Genetics

Emmanuelle Génin[a] Françoise Clerget-Darpoux[b]

[a]INSERM UMR1078, CHU Brest, Université Bretagne Occidentale, Brest, and [b]INSERM UMR781, Institut Imagine, Université Paris Descartes, Paris, France

In the search for the genetic factors underlying multifactorial diseases, the way is paved by epidemics of the GIGO (Garbage-In Garbage-Out) syndrome. Indeed, the use of simplistic models for these diseases leads to erroneous conclusions. The first epidemic of the GIGO syndrome occurred in the 1980s with the publication of the first comprehensive genetic maps of markers that enabled linkage analysis at the genome-wide level. During this decade, the lod score method, previously developed by Morton [1] to find genes involved in monogenic diseases, was used in the study of multifactorial diseases. The method requires the specification of a genetic model for the correspondence between the genotypes at the disease locus and the phenotype [2]. For most multifactorial diseases that involve multiple genetic and environmental risk factors, this correspondence is not known. The simple model used for monogenic diseases that assumes a rare, highly penetrant, dominant or recessive gene mutation is irrelevant for multifactorial diseases. This did not prevent researchers from using this model, and numerous tables of meaningless lod scores were published at that time, leading to invalid conclusions [3].

Since the first report of the complete human genome sequence in 2008 [4], we have witnessed a dramatic resurgence of the GIGO syndrome, with numerous publications reporting heritability estimates for different multifactorial diseases and claiming the existence of some so-called missing heritability [5]. In these computations of heritability (table 1), the assumed model for multifactorial diseases is no longer a monogenic model but a polygenic additive one with a liability threshold [6–9]. An infinite number of factors (both genetic and environmental) with weak, independent and additive effects are assumed to contribute to an underlying liability that is normally distributed, and a threshold is defined beyond which an individual is affected. Rather than the heritability of a multifactorial disease, it is thus the heritability of the liability of the disease that is estimated. It is also under this model that a novel method has been developed that allows the computation of heritability estimates from genome-wide association study data of unrelated individuals [10, 11].

Is this model suitable for multifactorial diseases? Some argue that observations are consistent with this model [12]: the relative risks estimated at individual SNPs are weak, there is no evidence for statistical interaction between associated SNPs, and the variance explained by each chromosome is proportional to its length. However, consistency with a model does not mean that the model is true. It may simply result from

0001–5652/15/0791–0001$39.50/0

E-Mail karger@karger.com
www.karger.com/hhe

Françoise Clerget-Darpoux
INSERM UMR781, Institut Imagine, Université Paris Descartes
24 Boulevard du Montparnasse
FR–75015 Paris (France)
E-Mail francoise.clerget@inserm.fr

the poor information available and a lack of power to reject the model. In particular, information on individual SNPs is useful to detect susceptibility loci, as evidenced by multiple replicated genome-wide association study hits, but it is very inefficient for measuring gene effects. The contrast between patients and controls concerning biallelic markers does not allow a good estimation of the genotypic risks associated with functional variations that are likely to be more complex than a single nucleotide change. This is the case for several disease susceptibility genes for which the information obtained from SNPs leads to an underestimation of the true genotypic risks and an incorrect classification of individuals in risk categories. Examples include the insulin gene in type 1 diabetes [13, 14], PTPN22 in rheumatoid arthritis [15, 16], IL2RA in multiple sclerosis [17, 18] and HLA in celiac disease [19–22] (table 2). Assuming that a multifactorial disease is triggered by the small and independent effects of many factors is overly simplistic.

Nonetheless, many studies of multifactorial diseases keep evaluating the heritability under the additive polygenic model, whereas genetic variance and consequently heritability estimates strongly depend on the genetic model assumed for disease transmission. Heritability estimates for the liability to schizophrenia vary from about 10–15% under a monogenic model to >75% under an additive polygenic model [23], and half of the latter value under a more complex model involving interactions between pathways [24]. Even though all these models fit the recurrence risks of schizophrenia in relatives, they lead to very different heritability estimates. Which of these estimates is correct? It is not possible to say, and they are probably all wrong, since for schizophrenia – as well as for the majority of multifactorial diseases – the underlying model is not known. Multifactorial diseases most likely have a very complex and heterogeneous etiology. This is well illustrated by the evolution of heritability estimates for the liability to diabetes before (heritability 0.75) [25] and after the differentiation between various forms of diabetes: monogenic (MODY), type 1 diabetes (heritability 0.88) [26] and type 2 diabetes (heritability 0.26) [27]. The genetic homogeneity of the liability to type 2 diabetes is itself questionable [28]. Most of the assumed models are too simplistic to allow any useful prediction.

Not only do heritability estimates depend on the underlying genetic model, but they also depend on the environmental variance. If there is no environmental variance, as for example in the case of tuberculosis in a population homogeneously exposed to Koch bacillus, the

Table 1. Definition of Heritability

The heritability of a trait is defined as the proportion of phenotypic variance attributable to genetic differences. The variance of the phenotype (Var P) is partitioned into a sum of genetic (Var G) and environmental variances (Var E), assuming no interaction and no correlation between the genetic and environmental factors involved in the phenotype.

$$H^2 = \frac{\text{Var G}}{\text{Var P}} = \frac{\text{Var G}}{\text{Var G} + \text{Var E}}$$

Estimating heritability implies assumptions on genetic and environmental variances.

Table 2. Examples of suceptibility genes the effect of which is not well captured by a single SNP

Insulin gene and type 1 diabetes
A genome-wide association study performed in 2007 [13] reported an association between type 1 diabetes and a SNP in the insulin gene. Under the assumption of an additive effect of the risk allele, the relative risk (i.e. OR) for a risk-allele carrier versus a noncarrier was estimated at 1.25. However, in 1984, based on data reporting an association between type 1 diabetes and a variable number of tandem repeats in the insulin gene promoter [14], the OR was 3.2.

PTPN22 in rheumatoid arthritis
The OR for SNP rs2476601 reported as to be associated with rheumatoid arthritis [16] is estimated at 1.6, whereas taking into account the information on 3 interactive SNPs of PTPN22, the relative risks for rheumatoid arthritis vary from 1 to 4.7 [15]. In addition, the use of only the information on SNP rs2476601 leads to a strong misclassification of individuals in terms of relative risks.

IL2RA and multiple sclerosis
The OR for SNP rs2104286 associated with multiple sclerosis in a genome-wide study [17] is estimated at 1.19, whereas the differential risk between the least and most at-risk genotype is 4 for the joint information provided by the 2 SNPs rs2256774 and rs3118470 [18].

HLA and celiac disease
A genome-wide association study reported an association between a SNP (rs2187668) of the HLA region and celiac disease with an OR of 7 [19], whereas the OR estimate is 8.4 for the HLA B8 antigen – known for more than 30 years to be associated with celiac disease [20] – and it is 27 for the HLA DQ2 heterodimer directly involved in the disease process [21, 22]. This heterodimer is encoded by 2 HLA genes interacting together, and its effect could not be summarized by just a biallelic polymorphism. An important proportion of heterodimer carriers (those encoding the heterodimer in the trans position) will be classified as having low risk when considering only the information provided by SNP rs2187668.

heritability is 1. For most multifactorial diseases or human traits, the environmental risk factors are unknown, and contrary to animal or plant species, their variance is impossible to control. This variance may strongly vary between different populations and between different generations.

If heritability estimates of human traits are so unreliable, then one may wonder why their use in human genetics is so widespread. It may stem from the strange game proposed to geneticists to search for the 'missing heritability of complex diseases' [29]. A clear winner in this game is the impact factor of *Nature*, with the number of citations of that single article ([29]) being >3,000 (and an additional one in this Commentary). Unfortunately, genetics is a clear looser, with a lack of diversity in the strategies used for the study of multifactorial diseases. It is particularly disturbing to observe that information available on marker-disease transmission in families – albeit essential for genetic modelling – is nowadays so often ignored by geneticists [30].

GIGO epidemics seem to occur after major revolutions in genetics. Thanks to molecular biology, the progress on monogenic diseases during the last two decades of the 20th century has been incredible. In this new century, with the sequencing of the genome, we may also expect a huge advance in the understanding of human diseases, in particular with the identification of many monogenic subentities. It seems, however, that in their enthusiasm, geneticists have forgotten that most human diseases are very heterogeneous and complex. It took a long time to escape from the monogenic paradigm; it is now urgent to escape from the polygenic one! We support Nelson et al. [31] in their call for a paradigm change in quantitative genetics. The issue is even more crucial for multifactorial diseases as the quantitative liability variable is an unobserved abstract construction [32]. Preventing the GIGO syndrome requires a move away from simplistic models and the development of novel strategies, combining different sources of information to progress in the understanding of disease etiology [33].

References

1 Morton NE: Sequential tests for the detection of linkage. Am J Hum Genet 1955;7:277–318.
2 Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J: Effects of misspecifying genetic parameters in lod score analysis. Biometrics 1986;42:393–399.
3 Clerget-Darpoux F, Bonaiti-Pellie C: An exclusion map covering the whole genome: a new challenge for genetic epidemiologists? Am J Hum Genet 1993;52:442–443.
4 Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452:872–876.
5 Maher B: Personal genomes: the case of the missing heritability. Nature 2008;456:18–21.
6 Fisher R: The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb Earth Sci 1918;52:399–433.
7 Falconer DS: An Introduction to Quantitative Genetics. New York, Ronald, 1960.
8 Dempster ER, Lerner IM: Heritability of threshold characters. Genetics 1950;35:212–236.
9 Gottesman II, Shields J: A polygenic theory of schizophrenia. Proc Natl Acad Sci USA 1967;58:199–205.
10 Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: Common SNPs explain a large proportion of the heritability for human height. Nat Genet 2010;42:565–569.
11 Yang J, Lee SH, Goddard ME, Visscher PM: GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88:76–82.
12 Hill WG, Goddard ME, Visscher PM: Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 2008;4:e1000008.
13 Todd JA, Walker NM, Cooper JD, et al: Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007;39:857–864.
14 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.
15 Begovich AB, Carlton VE, Honigberg LA, et al: A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am J Hum Genet 2004;75:330–337.
16 Bourgey M, Perdry H, Clerget-Darpoux F: Modeling the effect of PTPN22 in rheumatoid arthritis. BMC Proc 2007;1(suppl 1):S37.
17 Multiple Sclerosis Genetics Consortium; Hafler DA, Compston A, Sawcer S, et al: Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 2007;357:851–862.
18 Babron MC, Perdry H, Handel AE, Ramagopalan SV, Damotte V, Fontaine B, Muller-Myhsok B, Ebers GC, Clerget-Darpoux F: Determination of the real effect of genes identified in GWAS: the example of IL2RA in multiple sclerosis. Eur J Hum Genet 2012;20:321–325.
19 van Heel DA, Franke L, Hunt KA, et al: A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 2007;39:827–829.
20 Dausset J, Svejgaard A (eds): HLA and Disease. Copenhagen, Munksgaard, 1977.
21 Sollid LM, Markussen G, Ek J, Gjerde H, Vartdal F, Thorsby E: Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. J Exp Med 1989;169:345–350.
22 Margaritte-Jeannin P, Babron M-C, Bourgey M, Louka AS, Clot F, Percopo S, Coto I, Hugot J-P, Ascher H, Sollid LM, Greco L, Clerget-Darpoux F: 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 Elston RC: Discussion: Methodologies in human behavior genetics. Soc Biol 1973;20:276–279.
24 Zuk O, Hechter E, Sunyaev SR, Lander ES: The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci USA 2012;109:1193–1198.
25 Harvald B, Hauge M: Hereditary factors elucidated by twin studies (edited by Neel J, Shaw M, Schull W in 1965). Symposium on Contribution of Genetics to Epidemiological Studies of Chronic Diseases, Ann Arbor, 1963.
26 Hyttinen V, Kaprio J, Kinnunen L, Koskenvuo M, Tuomilehto J: Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. Diabetes 2003;52:1052–1055.
27 Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H: Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance – a population-based twin study. Diabetologia 1999;42:139–145.
28 Weijnen CF, Rich SS, Meigs JB, Krolewski AS, Warram JH: Risk of diabetes in siblings of index cases with type 2 diabetes: implications for genetic studies. Diabet Med 2002;19:41–50.
29 Manolio TA, Collins FS, Cox NJ, et al: Finding the missing heritability of complex diseases. Nature 2009;461:747–753.
30 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.
31 Nelson RM, Pettersson ME, Carlborg O: A century after Fisher: time for a new paradigm in quantitative genetics. Trends Genet 2013;29:669–676.
32 Benchek PH, Morris NJ: How meaningful are heritability estimates of liability? Hum Genet 2013;132:1351–1360.
33 Bourgain C, Genin E, Cox N, Clerget-Darpoux F: What do we really need to dissect the genetic component of complex human diseases? Eur J Hum Genet 2007;15:260–263.