Title

Missing heritability of complex diseases: Enlightenment by genetic variants from intermediate phenotypes

Journal

BioEssays, 38(7)

ISSN

0265-9247

Authors

Blanco-Gómez, Adrián

Castillo-Lluva, Sonia

Del Mar Sáez-Freire, María

et al.

Publication Date

2016-07-01

DOI

10.1002/bies.201600084

Peer reviewed

Missing heritability of complex diseases: Enlightenment by genetic variants from intermediate phenotypes

*Adrián Blanco-Gómez[1)2)**], Sonia Castillo-Lluva[1)2)**], María del Mar Sáez-Freire[1)2)**], Lourdes Hontecillas-Prieto[1)2)], Jian Hua Mao[3)], Andrés Castellanos-Martín[1)2)*,***] and Jesus Pérez-Losada[1)2)*,***]*

Diseases of complex origin have a component of quantitative genetics that contributes to their susceptibility and phenotypic variability. However, after several studies, a major part of the genetic component of complex phenotypes has still not been found, a situation known as "missing heritability." Although there have been many hypotheses put forward to explain the reasons for the missing heritability, its definitive causes remain unknown. Complex diseases are caused by multiple intermediate phenotypes involved in their pathogenesis and, very often, each one of these intermediate phenotypes also has a component of quantitative inheritance. Here we propose that at least part of the missing heritability can be explained by the genetic component of intermediate phenotypes that is not detectable at the level of the main complex trait. At the same time, the identification of the genetic component of intermediate phenotypes provides an opportunity to identify part of the missing heritability of complex diseases.

Keywords:

complex diseases; heritability; intermediate phenotype or endophenotype; missing heritability

Introduction: The problem of missing heritability

The variability of presentation of complex phenotypes, such as those involved in complex diseases, has a component of quantitative inheritance, consisting of the sum of effects of different allelic forms that interact with each other and with the environment [1–3]. The quantitative inheritance not only contributes to the heterogeneity of complex traits, but also contributes, to a greater or lesser extent, to the phenotypic heterogeneity of traits with Mendelian inheritance. In these cases, the effect of the main gene, principally responsible for the phenotype, is modified by the function of low penetrance genes, also called modifier genes. These modifier genes contribute to the expression of any given phenotype, even

DOI 10.1002/bies.201600084

[1)] Instituto de Biología Molecular y Celular del Cáncer (CIC-IBMCC), Universidad de Salamanca/CSIC, Salamanca, Spain
[2)] Instituto de Investigación Biomédica de Salamanca (IBSAL), Salamanca, Spain
[3)] Life Sciences Division, Lawrence Berkeley National Laboratory (LBNL), University of California, Berkeley, CA, USA

[**]These authors contribute equally to this work as co-first authors and are listed in alphabetical order.
[***]These authors contribute equally as senior authors.
Current address of Sonia Castillo-Lluva: Departamento de Bioquímica y Biología Molecular I, Facultad de Biología, Universidad Complutense de Madrid, Madrid, Spain.
Current address of Lourdes Hontecillas-Prieto: Departamento de Patología Molecular, Instituto de Biomedicina de Sevilla (IBiS), Sevilla, Spain.
Current address of Andrés Castellanos-Martín: Institute for Research in Biomedicine (IRB), Barcelona, Spain.

***Corresponding authors:**
Andrés Castellanos Martín
E-mail: andres.castellanos@irbbarcelona.org
Jesús Pérez Losada
E-mail: jperezlosada@usal.es

Abbreviations:
DSV, DNA sequence variant; **GWAS,** genome-wide association studies; **QTLs,** quantitative trait loci.

acquired ones, as in the susceptibility and evolution of infectious diseases [4].

The heterogeneity of complex phenotypes in a population is the result of phenotypic variance owing to environmental influence and to the associated genetic components. *Heritability in a broad sense* is defined as the proportion of the phenotype variance due to genetic components. *Heritability in a strict or narrow sense* refers to the proportion of phenotypic variance due to the additive genetic components between different allelic variants or DNA sequence variants (DSVs) [5]. It is interesting to note that 88% of DSVs are found in both intronic (45%) and intergenic (43%) non-coding regions [6]. In linkage studies, genomic regions associated with phenotype variability, are named quantitative trait loci (QTLs). The heritability of a complex phenotype also depends on the estimated contribution of environmental factors; thus the more phenotypic variance explained by the environment, the less phenotype variance explained by genetic factors and vice versa. Hence, the estimation of phenotypic variance can vary considerably among populations subjected to different environmental factors; for example, for susceptibility to lung cancer development in a smoking population versus in a population of non-smokers.

Many features, such as body size, have a very high heritability. Thus, it is estimated that 80–90% of variance in body size is explained by inheritance [7]. However, although many alleles associated with this phenotype have been identified, they only explain around 5–10% of the phenotype variance. This is also the case of many other complex phenotypes and, in general, there is a high level of discrepancy between the proportion of phenotypic variance expected to be explained by genetic influences, known as the "expected heritability," and the heritability really explained by the DSVs identified so far. This difference is known as the "missing heritability" [8, 9].

Genome-wide association studies (GWAS) allow for the assessment of between half a million to two million SNPs along the human genome that represent the most common haplotypes. Thus, GWAS were initially designed to identify the allelic forms responsible for quantitative inheritance of complex phenotypes, such as sporadic cancer, Alzheimer's disease, diabetes, among others. Minor allele frequency (MAF) refers to the frequency at which the least common allele occurs in a given population. According to this definition, common or frequent variants (MAF > 0.05 or 5%), low-frequency variants (MAF = 0.01–0.05 or between 1 and 5%) and rare variants (MAF < 0.01 of less than 1%) can be distinguished. It is assumed that the genetic component of complex diseases is the sum of the effects of common or frequent genetic variants, which is known as the *common disease/common variant* hypothesis. Therefore, with the arrival of GWAS it was thought that it would be easier to identify the different DSVs that contribute to the pathogenesis of various complex traits, and certainly there have been many QTLs and DSVs identified associated with the variability of a number of complex phenotypes. This collection can be accessed via the website: www.genome.gov/gwastudies/.

In general, the effect of DSVs on total phenotypic variability has been relatively small, as most of the DSVs identified are separately/individually associated with less than 1–2% of the phenotypic variance, and all together explain less than 30% of the variance observed [10]. Therefore, the usefulness of GWAS in identifying genetic determinants responsible for the variance of complex phenotypes has been questioned. In conclusion, GWAS seem to be insufficient to solve the problem of the "missing heritability" [9, 11–14]. For this reason, it has also been suggested that at least part of the genetic component of complex phenotypes may be due to rare variants [15]. This is the *common disease/rare variant* hypothesis [16]. These rare variants are relatively difficult to detect by GWAS because in order to be detected, they should be present in the population in a proportion between at least 1 and 10%. Since these rare allelic forms are less frequently present, to detect them, it would be necessary to significantly increase the sample size or use meta-analysis studies [17–19]. However, even under these conditions, it would probably not be sufficient enough to detect them if we consider that part of the difficulty in detecting association with rare variants in GWAS could be attributable to the fact that these variants, until recently, have been underrepresented on SNP microarrays. Hence, rare variants are in low linkage disequilibrium with common SNPs in microarrays, and have been imputed to limited success [20]. The new massive sequencing techniques, both whole-exome sequencing and whole-genome sequencing, could be more useful in the identification of rare variants. In particular whole-genome sequencing would be ideal for finding both exon and intron DSVs. Whole-genome sequencing has already been used successfully in identifying causative mutations for rare diseases caused by a mutation in a single gene [21–27]. Similarly, these massive sequencing strategies are being increasingly used to identify rare allelic forms that contribute to the pathogenesis of common complex phenotypes [15, 28–30] and there is also recent empirical evidence that low-frequency and rare variants may be connected to complex diseases [31–34]. Certainly, these studies have identified rare variants associated with the variability of complex phenotypes, but both the individual effect of each variant and the effect of the sum of all of them together on phenotype variance, are still too small to explain a significant proportion of the "missing heritability." In addition, in spite of the relatively larger effect sizes of rare variants, one could expect that their effect on population risk would be small, merely because of their low frequency in the population [35].

It has also been proposed that part of the missing heritability could be explained by genetic interactions or epistasis [36]. The estimated heritability in the strict sense is based on the assumption that there are no interactions between alleles and that their effects are only additive. But this assumption is not completely true because the genetic interactions also affect the calculation of broad-sense heritability [37]. Therefore, it has been proposed that a significant part of the missing heritability may not be due to the genetic variants that still remain to be identified, but to the epistatic genetic interactions between the genetic variants already discovered [36]. Thus, even though all allelic forms responsible for the variability of a particular complex phenotype were identified, they could not explain all the phenotype variance by an additive effect. Consequently, the proportion of explained heritability would always be less than the total heritability [36]. In any case, the demonstration of epistasis in human population studies is very difficult and would require huge samples. Thus, the magnitude of the contribution of epistasis to the missing heritability of complex phenotypes still remains to be determined. It has also been proposed that epigenetic changes that contribute to variable gene expression between individuals may also contribute to complex phenotype variability and its heritability [38]. Nevertheless, none of these explanations have so far satisfactorily explained the large proportion of missing heritability [39].

Complex phenotypes are a consequence of multiple intermediate phenotypes located at different levels

To understand the possible cause of missing heritability, we should revise the concept of complex phenotype. Many major diseases, as well as physiological and pathophysiological processes are considered complex phenotypes, such as aging, sporadic cancer, diabetes mellitus, ischemic heart disease, or autoimmune diseases. These diseases are the consequence of another series of second order or intermediate phenotypes that are causally related to them and involved in their pathogenesis (Fig. 1). For example, the susceptibility to ischemic heart disease is influenced by other intermediate phenotypes such as blood pressure, hypercholesterolemia, or susceptibility to pro-atherogenic agents present in tobacco, among others [40]. Another example is that of aging susceptibility, which could be the result of different intermediate phenotypes related to biological mechanisms such as the ability to repair DNA damage, telomere length attrition, or others that control susceptibility to oxidative damage [41].

In addition, the variability in the phenotypic presentation of a complex trait that may be considered as a first order phenotype or the outcome phenotype depends on the phenotypic variability of each one of those intermediate phenotypes or second order phenotypes involved in its pathogenesis. Thus, the grade of susceptibility to a complex disease depends on the grade of expression of those second order phenotypes involved in its pathogenesis. As an example, in the case of ischemic heart disease, the grade of susceptibility to the disease depends on the grade of susceptibility to hypertension, hypercholesterolemia, etc. On the other hand, these intermediate phenotypes of second order that contribute to the pathogenesis of the main complex trait are very often complex phenotypes that in turn depend on other intermediate phenotypes involved in their pathogenesis, or third order phenotypes in respect to the main one. For example, susceptibility to hypertension is associated with intermediate phenotypes that control the arterial tone such as the activity of calcium channels, the activity of the renin-angiotensin-aldosterone system, the renal function, or the autonomous nervous system activity, among others [42]. In short, the expression of a main complex phenotype would be determined by a series of complex intermediate phenotypes of second order and these, in turn, by others of third order and so on, reaching an increasing level of complexity (Fig. 1).

In the global structure of a complex phenotype or complex disease, we can consider the existence of a *surface level* where the main features that define the complex disease would be located (Fig. 2). For example, in the case of the ischemic heart disease, this surface level would include chest pain, specific electrocardiographic changes, elevated cardiac enzymes due to the myocardial necrosis, the obstruction of the coronary artery by the plaque of atherosclerosis, etc. Furthermore, intermediate phenotypes, those that could modify the manifestations of a complex disease, would be located at different levels within this hierarchy. In relation to this, *the organic or systemic level*, including the physiological and pathophysiological processes that influence the manifestations of the complex disease, would be located just below the main surface level. This would include intermediate phenotypes related to the endocrine-metabolic system, function of specific organs such as respiratory, digestive, kidney, etc. Following the already exhibited example of ischemic heart disease, the *organic or systemic level* would include processes such as hypertension, hypercholesterolemia, etc. Below this, at *the tissue level*, the interactions between different cell types from the tissue responsible for correct function, tissue growth and repair would take place. The interactions between tissue cells would lead to differentiation, proliferation, and apoptosis of those cells from the tissue, among others. In ischemic heart disease, this would include the cellular processes that contribute to the formation of the atherogenesis plaque, such as local infiltration of macrophages, endothelial damage, proliferation of smooth muscle cells in the arterial wall, etc. At *the intrinsic-cellular or intracellular level*, just below the *tissue level*, processes that determine proliferation, apoptosis, and cell differentiation, would take place. These processes include aspects such as traffic between organelles, vesicle trafficking processes in the endoplasmic reticulum, in the Golgi apparatus, etc. *The molecular level* would include intracellular and extracellular signaling pathways, transcription factors, membrane receptors, etc., responsible for cellular and extracellular processes and would comprise the synthesis of different molecular components, posttranslational modifications, and their degradation by ubiquitination, etc. *The transcriptional level* would include both the RNA and the epigenetic modifications that determine the expression of certain genes, influenced by the functional requirements at each moment. Finally, at the bottom of the hierarchy, *the genomic level* would include genes and intergenic and intragenic regulatory regions (Fig. 2).

The different grade of function at each level results in its own intrinsic phenotypic variability and contributes to the phenotypic variability of the upper levels, particularly to the first one. For example, the different functioning of macrophages may contribute to a higher or lower susceptibility to form the plaque of atheroma, and thus contribute to the phenotype of ischemic heart disease located at the upper level. Obviously, the function of each level is influenced by the lowest one, the genomic level, and its interaction with the environment, but the grade of function and phenotype expression of each level is also influenced by other variables and not only depends on the genomic influence. For example, at the molecular level, protein function is not only influenced by the DNA that determines their amino acid sequence, but also by the intracellular protein compartmentalization at a given time, by the degree of protein activation due to posttranslational modifications in response to signaling

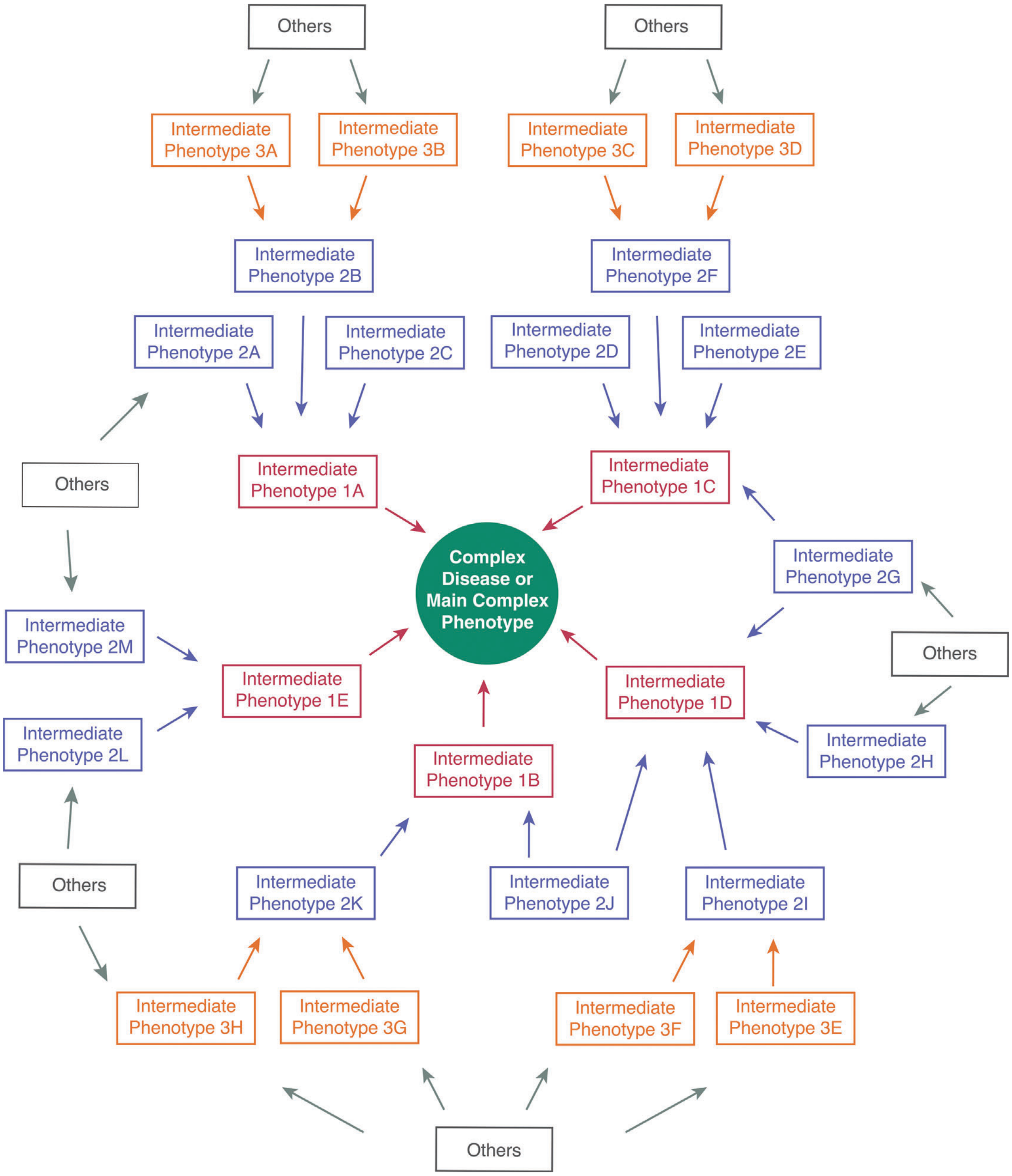
Others
Others
Intermediate Phenotype 3A
Intermediate Phenotype 3B
Intermediate Phenotype 3C
Intermediate Phenotype 3D
Intermediate Phenotype 2B
Intermediate Phenotype 2F
Intermediate Phenotype 2A
Intermediate Phenotype 2C
Intermediate Phenotype 2D
Intermediate Phenotype 2E
Others
Intermediate Phenotype 1A
Intermediate Phenotype 1C
Complex Disease or Main Complex Phenotype
Intermediate Phenotype 2G
Intermediate Phenotype 2M
Others
Intermediate Phenotype 1E
Intermediate Phenotype 1D
Intermediate Phenotype 2L
Intermediate Phenotype 2H
Intermediate Phenotype 1B
Others
Intermediate Phenotype 2K
Intermediate Phenotype 2J
Intermediate Phenotype 2I
Intermediate Phenotype 3H
Intermediate Phenotype 3G
Intermediate Phenotype 3F
Intermediate Phenotype 3E
Others

Figure 1. The grade of susceptibility and phenotypic variation in the presentation of complex diseases or complex traits depends on a number of intermediate phenotypes of second order that influence their pathogenesis; many of which are themselves complex phenotypes, which are in turn influenced by other intermediate phenotypes of third order. All of this creates a series of complex interactions between phenotypes. It must also be considered that the molecular and genetic determinants below each phenotype all acquire a structure of systems biology.

pathways, by its degradation by the ubiquitin system, or by organic intracellular conditions that maintain their tertiary structure, among others [43].

There are many functional interactions between components within each level, and between levels, that respond to feedback allowing cells, tissues, and ultimately the whole organism to respond to its functional requirements at every moment. These intra- and inter-level interactions have a structure of systems biology that can be represented as networks, where the effects of an alteration in a given level produce an impact far beyond its immediate target and level. Thus, the alteration of a gene has implications beyond its immediate influence due to its related associations. For example, a gene that influences obesity, will contribute to generating a systemic low-grade inflammation that promotes the development of the metabolic syndrome and type II diabetes and certain types of cancer. Adipose tissue is also involved in hormone synthesis, such as estrogens, where the production of fatty tissue may contribute to breast cancer susceptibility. In addition, estrogens are implicated in many interactions at different levels, such as with the coagulation system, the immune system, and autoimmune diseases which

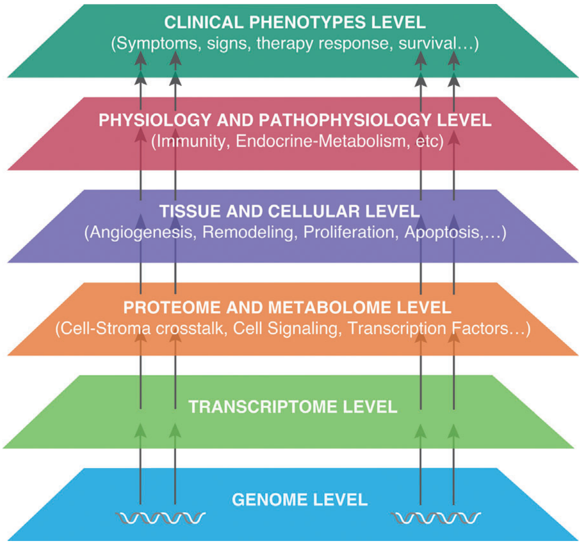
CLINICAL PHENOTYPES LEVEL
(Symptoms, signs, therapy response, survival…)
PHYSIOLOGY AND PATHOPHYSIOLOGY LEVEL
(Immunity, Endocrine-Metabolism, etc)
TISSUE AND CELLULAR LEVEL
(Angiogenesis, Remodeling, Proliferation, Apoptosis,…)
PROTEOME AND METABOLOME LEVEL
(Cell-Stroma crosstalk, Cell Signaling, Transcription Factors…)
TRANSCRIPTOME LEVEL
GENOME LEVEL

Figure 2. Considering that phenotypic traits of the genesis of a complex disease are within the first or surface level, the intermediate phenotypes that contribute to its pathogenesis may reside at different levels: systemic, organic/tissue, cellular, and molecular. There are multiple interactions within each level and between levels in order to meet the needs of the whole organism at every moment. All of this is strongly influenced by the lowest or genomic level and the interactions with the environment.

are most common in women; obesity is also a stimulus for bone formation, but also for osteoarthritis, etc. [44, 45].

Moreover, in many cases, pathophysiological, cellular, and molecular components from one or more of these levels will be involved in the pathogenesis of several complex traits at the same time, and this will influence the expression of several of these complex traits simultaneously. This would explain the epidemiological association between complex diseases because of the existence of common pathogenic processes. As an example, this would be the case of the association of autoimmune processes with cancer, or cancer and obesity; inflammation, obesity, and thromboembolic disease or between the latter, cancer and aging. In a broader sense, and according to the pathogenic processes that present a commonality, an attempt has been made to group various diseases by families, thus redefining the taxonomy of disease, known as diseasome [46, 47].

Using intermediate phenotypes to identify genetic determinants of diseases of complex origin

The association between different complex diseases sometimes is not only due to the common pathogenic mechanisms, but also because they share the influence of the same genetic variants, known as *crossphenotype associations*. This is the case of Crohn's disease and ulcerative colitis [48], other autoimmune diseases [49, 50] or psychiatric disorders [51, 52]. This is connected to the concept of *biological pleiotropy* or true pleiotropy, when one or more genetic variants are associated with two phenotypes that are not related. This means that one of them is not causally related with the other, or, in other words, that one of them is not an intermediate phenotype of the other [53, 54]. This concept is different to what was coined by Solovieff et al., known as *mediated pleiotropy*, when a genetic variant is linked to a complex phenotype because the genetic variant is associated with an intermediated phenotype that in turn is causally related with the outcome complex phenotype [55]. For example, Voight et al., found genetic variants associated with both low-density lipoprotein (LDL) serum levels and the risk of myocardial infarction [56]. On the other hand, for diseases of complex origin, such as ischemic heart disease, autoimmune diseases, etc., it is quite likely that every single DSV identified is influencing the complex disease through intermediated phenotypes from different levels. In fact, most of time, when we identify DSVs associated with complex phenotypes, the difficult part is to identify the mechanisms of the intermediated phenotypes by which these genetic variants are influencing the complex phenotype.

In recent years, new statistical strategies have been developed to identify associations between genetic loci or genetic variants and several traits in order to identify pleiotropy [55, 57]. Moreover, statistical strategies have been developed to distinguish between biological pleiotropy and mediated pleiotropy, such as *Mendelian randomization*, to identify causal association between an intermediate phenotype and the outcome one [56, 58, 59]. In a similar way, strategies to identify genetic variants associated with a complex phenotype through their association with intermediate phenotypes have also been carried out [60, 61].

The use of intermediate phenotypes to identify the genetic component of diseases has been used mainly in the field of Psychiatry [60]. In 1965, Douglass Falconer introduced the idea of normally distributed quantitative traits as liability for genetically determined disorders, known as the multifactorial threshold model, which has been applied to non-Mendelizing common diseases. This idea was adapted to schizophrenia by Gottesman and Shield in 1967 [62]. Psychiatric disorders are complex diseases whose classification between entities is problematic because their definition and diagnosis are based on behavioral characteristics that are difficult to quantify. This is the reason why the difficulty of finding genetic determinants in psychiatric disorders, both by linkage and association studies, is widely acknowledged [60]. Therefore, an effort was made to identify other intermediate phenotypes that were associated with the disease and that were easier to quantify, employing imaging tools such as computed tomography (CT) scans, magnetic resonance imaging (MRI), or positron emission tomography (PET), levels of neurotransmitters, etc., in order to be to find the genetic determinants associated with the main disease [60].

In Psychiatric Genetics, the term endophenotype is mainly used which is equivalent to intermediate phenotype. It was initially described by Gottesman and Shields [63] and adapted from the zoology field [64]. This term initially referred to the endogenous phenotypes identified by biochemical or microbiological determinations, and was well-adapted in Psychiatric Genetics to fill the gap between the lowly quantifiable descriptions of mental illness and Genetics. The rationale for using an intermediate phenotype for the identification of the genetic component of a complex phenotype is that if the intermediate phenotype associated with the disease is very directly related to it, and represents a more elementary phenotype, the number of genes required to produce variations in these intermediate phenotypes would be less than the number of genes involved in the production of the main disorder.

The intermediate phenotype used to identify part of the genetic characteristics of a complex phenotype must have certain characteristics; it can even be a phenotype from the surface level that contributes to the definition of the disease. If the intermediate phenotype used is from a lower level, it must contribute to the pathogenesis of the disease, and not just be a mere biological marker without a common genetic basis with the main phenotype. Thus, initially four criteria have been proposed to define an intermediate phenotype or endophenotype [60] based on the previous criteria used to define a marker in psychiatric genetics [61]: (i) the intermediate phenotype is associated with illness in the population; (ii) the intermediate phenotype is heritable; (iii) the intermediate phenotype is primarily state-independent (manifests in an individual whether or not illness is active); and (iv) within families the intermediate phenotype and illness co-segregate. Subsequently, an additional criterion that may be useful for identifying intermediate phenotype of diseases that display complex inheritance patterns was suggested by Leboyer et al. [65]: The intermediate phenotype found in affected family members is found in unaffected family members at a higher rate than in the general population. The use of intermediate phenotypes to identify genetic determinants of complex diseases has been used successfully not only in psychiatric disorders but also in other non-psychiatric disorders, such as the long QT syndrome [66, 67] and various ECG traits associated with cardiovascular morbidity [68], the idiopathic hemochromatosis [69], the juvenile myoclonic epilepsy [70], and the familial adenomatous polyposis [71] or systemic lupus erythematosus [72, 73], among others.

Missing heritability of complex traits could be partly explained by genetic determinants of intermediate phenotypes

The study of phenotypic variability in human complex diseases is difficult to carry out when considering the influence caused by intermediate phenotypes. The enormous genomic variability between different human populations and their complex interaction with the environment contribute to this difficulty. However, these types of studies are more plausible in simplified models such as those generated by crosses between mouse strains genetically homogeneous and with relatively homogeneous phenotypes. In these inbred strains, all mice are genetically identical and have every allele in homozygosis. Furthermore, the interaction with the environment is simplified as they are housed in stable conditions in animal facilities free of specific pathogens (SPF) [74, 75]. These cohorts of mice generated by intercrosses or backcrosses between inbred strains with divergent phenotype behavior have been used frequently for linkage studies and identification of QTLs, and also offer a unique opportunity to carry out studies through multiple levels of intermediate phenotypes associated with complex traits and diseases, such as cancer, aging, heterogeneous response to therapy, etc.

Recently, we dissected the phenotypic heterogeneity of ERBB2+ breast cancer in some of the previously described levels in a simplified model generated by a backcross between two mouse strains with divergent susceptibility to breast cancer [76]. We dissected the disease in different pathophenotypes, such as latency, the time of disease, the grade of local tumor growth, the number of metastases, etc. In addition, we studied the associations of those pathophenotypes with different intermediate phenotypes including different components of cell signaling such as AKT/mTOR and MAPK/ERK pathways in both tumor and livers, and a range of metabolites in serum from disease free mice. We identified QTLs associated with the heterogeneous behavior of the different pathophenotypes of the disease, and QTLs associated with the variability of the different intermediate phenotypes studied. After carrying out this study, we observed that most of the tumor QTLs associated with different pathophenotypes did not overlap with the QTLs associated with intermediate phenotypes, even with those that could be expected to be strongly associated with tumor variability,

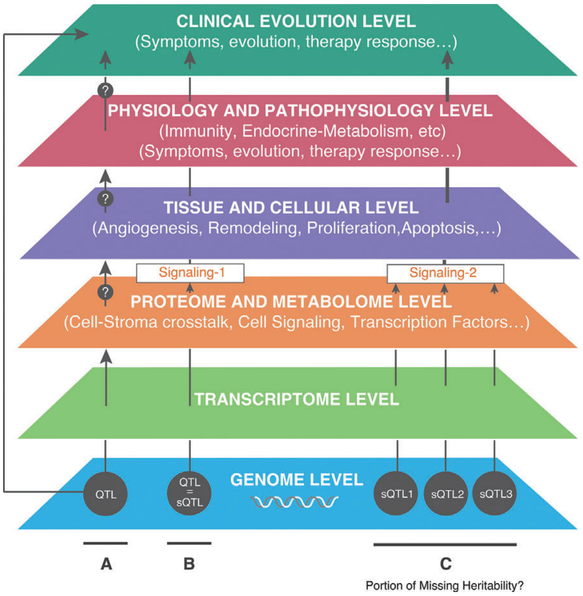
CLINICAL EVOLUTION LEVEL
(Symptoms, evolution, therapy response…)
?
PHYSIOLOGY AND PATHOPHYSIOLOGY LEVEL
(Immunity, Endocrine-Metabolism, etc)
(Symptoms, evolution, therapy response…)
?
TISSUE AND CELLULAR LEVEL
(Angiogenesis, Remodeling, Proliferation,Apoptosis,…)
Signaling-1
Signaling-2
?
PROTEOME AND METABOLOME LEVEL
(Cell-Stroma crosstalk, Cell Signaling, Transcription Factors…)
TRANSCRIPTOME LEVEL
GENOME LEVEL
QTL
QTL = sQTL
sQTL1
sQTL2
sQTL3
A
B
C
Portion of Missing Heritability?

Figure 3. The influence of a quantitative trait locus (QTL) or DNA sequence variant (DSV) over the susceptibility and/or variable presentation of a complex disease would be exerted by some intermediate phenotype located within any of the levels previously indicated (**A**). For example, it is possible that a QTL or DSV will exert its influence on a complex disease through a particular signaling pathway (**B**). It is feasible that another signaling pathway influences the variability of the complex disease in a very significant manner, but this signaling pathway in turn may be itself a complex phenotype influenced by multiple QTL. It is possible that these QTLs are, individually, powerful enough to affect a significant proportion of the variability of the signaling pathway. However, they would not induce a powerful enough variation in the signal to affect the main phenotype, and therefore be detected as genetic modifiers of the complex disease. These QTL influencing intermediate phenotypes, undetectable at the level of the main phenotype, would form part of the “missing heritability” (**C**).

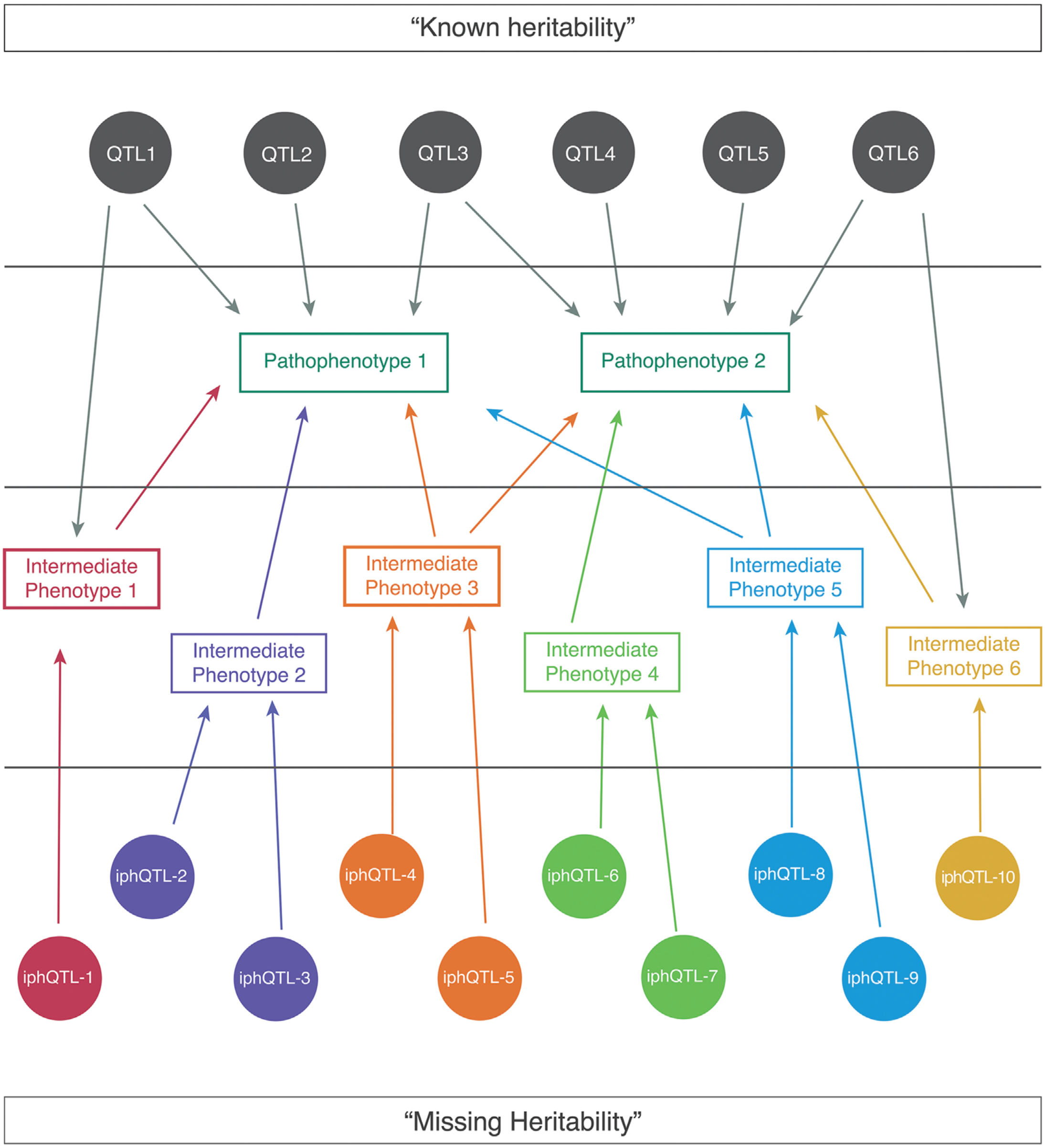

Figure 4. It is possible to dissect a complex disease in different pathophenotypes, and to identify a number of genetic determinants associated to the variability of these pathophenotypes in a population, that constitute the known proportion of the heritability of the disease. There are a number of intermediate phenotypes that contribute to the pathogenesis of the disease. The identification of genetic determinants that contribute to the variability of these intermediate phenotypes can be a global strategy to identify part of the missing heritability of complex diseases. iphQTL = intermediate phenotype QTL.

such as those related with tumor cell signaling [76]. Thus, an intermediate phenotype associated with the pathogenesis of a complex trait, in turn presents its own variability along the population, and is also controlled by a number of QTLs that seem to be different.

As already mentioned, complex phenotypes such as complex diseases result from different intermediate phenotypes involved in its pathogenesis that, to a greater or lesser extent, have a component of complex traits with polygenic influence. Thus, when a DSV is associated with a complex phenotype, its effect is exercised through some intermediate phenotype. Following the example of the first part of this review, it is possible that a DSV that influences susceptibility to ischemic heart disease does so through its influence on susceptibility to hypertension. However, it is not uncommon that all DSVs that contribute to the susceptibility to hypertension are not detected as being responsible for susceptibility to ischemic heart disease. Thus, these DSVs would be strong enough to contribute to the phenotypic variability of hypertension, but the proportion of hypertension variability affected by those DSVs is not significant enough to be detected as a modifier of the susceptibility to ischemic heart disease. This does not mean that the whole hypertension variability does not contribute to the variable susceptibility to ischemic heart disease, but part of the polygenic component of the hypertension cannot be detected at the main complex phenotype level, which in this case is ischemic heart disease.

Therefore, for a QTL associated with phenotype variability to be detected, it has to contribute to a minimum detectable level of the variability of that phenotype (Fig. 3). This allows for the detection of a QTL when it reaches a LOD score by linkage analysis. Similarly, when an intermediate phenotype contributes to the variability of a complex phenotype, the intermediate phenotype should contribute to a minimum significant variability of this complex phenotype in order to be detected as significantly associated with the complex phenotype. The QTL associated with an intermediate phenotype that is associated with the pathogenesis of a main complex phenotype, could not be detected as a QTL of the main phenotype. This could be explained if the QTL is not responsible for a sufficient fraction of the intermediate phenotype variability. So that the latter is not sufficient to individually contribute to a detectable level of the variability of the complex phenotype.

Given the amount of secondary, tertiary, and other intermediate phenotypes which can contribute to the pathogenesis and variability of a complex phenotype, basically with all of them possessing a component of quantitative inheritance, it is not surprising that many of the QTLs that contribute to the variability of all these intermediate phenotypes cannot be detected as genetic determinants of the main phenotype, and would constitute much of the missing heritability [76]. For the same reason, and due to the huge amount of interactions between intermediate phenotypes and the enormous amount of QTLs associated with them, it seems that identification of the whole scenario of missing heritability of a complex phenotype probably could not be carried out. As a conclusion, we believe that the allelic forms that contribute to the variability of intermediate phenotypes that are not strong enough to be detected at the main phenotype constitute much of the missing heritability. Therefore, at least part of the missing heritability that affects a complex phenotype could be identified considering the heritability of intermediate phenotypes that participate directly in the pathogenesis of the complex phenotype (Fig. 4).

Future perspectives

Certainly, it is possible to generate multivariate models with intermediate phenotypes as variables from these different levels that are able to predict main phenotype variability more adequately than if we only consider the genetic elements [76]. In terms of heritability in the narrow sense, much of the missing heritability could be detected considering the heritability of a complex phenotype as the sum of the heritability of each of the intermediate phenotypes that contribute to the pathogenesis of the complex phenotype expressed in a model of multivariate analysis, where the value of each intermediate phenotype would be corrected by a different coefficient whose value would depend on the contribution of each intermediate phenotype to the variance of the complex phenotype. In this sense, Li et al. proposed a modified inverse-variance weighted meta-analysis method to combine disease status and quantitative intermediate phenotypes information, and showed that it was a powerful tool for detecting genetic variants in complex disease association studies, especially when the effects of the susceptibility loci are minor. They showed that these statistical tests combining both disease status and quantitative risk factors including intermediate phenotypes are more powerful than case-control studies [77].

We propose that the integration of the heritability from different intermediate phenotypes in multivariate models will allow the identification of an important part of the missing heritability of complex diseases. This in turn will allow for the better understanding of the pathogenesis of these diseases, and the identification of new genetic variants that may influence the risk, prognosis, and the response of disease to treatment; hence, moving toward the possibility of more personalized medicine [78].

Acknowledgments

JPL was partially supported by FEDER and MICINN (PLE2009-119, SAF2014-56989-R), Instituto de Salud Carlos III (PI07/0057, PI10/00328, PIE14/00066), Junta de Castilla y León (CSI034U13, BIO/SA31/15), "Proyectos integrados IBSAL 2015" (IBY15/00003), the "Eugenio Rodríguez Pascual", the "Fundación Inbiomed" (Instituto Oncológico Obra Social de la Caja Guipozcoa-San Sebastian, Kutxa), and the "Fundación Sandra Ibarra de Solidaridad frente al Cáncer". CR-C is funded by Q3718001E (2009–2010) y GRS 612/A/11 (2011–2012) and "the Fundación Eugenio Rodríguez Pascual". AC was supported by FIS (PI07/0057) and MICINN (PLE2009-119). SCLL was funded by a JAEdoc Fellowship (CSIC)/FSE. JHM was supported by the National Institutes of Health, a National Cancer Institute grant (R01 CA116481), and the Low-Dose Scientific Focus Area, Office of Biological & Environmental Research, US Department of Energy (DE-AC02-05CH11231). We thank Dr. Elena de la Casa Esperón for their useful comments on the manuscript, Emma Keck from the University of Salamanca for her help in English editing and Rocío Puras Pardo for figure construction. We apologize to the many colleagues whose work could not be cited due to space restrictions.

The authors have declared no conflicts of interest.

References

1. **Collins FS, Guyer MS, Charkravarti A.** 1997. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**: 1580–1.
2. **Reich DE, Lander ES.** 2001. On the allelic spectrum of human disease. *Trends Genet* **17**: 502–10.
3. **Pritchard JK.** 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**: 124–37.
4. **Boon AC, deBeauchamp J, Hollmann A, Luke J**, et al. 2009. Host genetic variation affects resistance to infection with a highly pathogenic H5N1 influenza A virus in mice. *J Virol* **83**: 10417–26.
5. **Visscher PM, Hill WG, Wray NR.** 2008. Heritability in the genomics era-concepts and misconceptions. *Nat Rev Genet* **9**: 255–66.
6. **Hindorff LA, Sethupathy P, Junkins HA, Ramos EM**, et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* **106**: 9362–7.
7. **Visscher PM.** 2008. Sizing up human height variation. *Nat Genet* **40**: 489–90.
8. **Maher B.** 2008. Personal genomes: the case of the missing heritability. *Nature* **456**: 18–21.
9. **Manolio TA, Collins FS, Cox NJ, Goldstein DB**, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–53.

10. **Lander ES.** 2011. Initial impact of the sequencing of the human genome. *Nature* **470**: 187–97.
11. **McClellan J, King MC.** 2010. Genomic analysis of mental illness: a changing landscape. *Jama* **303**: 2523–4.
12. **Cirulli ET, Goldstein DB.** 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**: 415–25.
13. **Antonarakis SE, Chakravarti A, Cohen JC, Hardy J.** 2010. Mendelian disorders and multifactorial traits: the big divide or one for all? *Nat Rev Genet* **11**: 380–4.
14. **Marian AJ, Belmont J.** 2011. Strategic approaches to unraveling genetic causes of cardiovascular diseases. *Circ Res* **108**: 1252–69.
15. **Abecasis GR, Altshuler D, Auton A, Brooks LD**, et al. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–73.
16. **Bodmer W, Bonilla C.** 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40**: 695–701.
17. **Newton-Cheh C, Johnson T, Gateva V, Tobin MD**, et al. 2009. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* **41**: 666–76.
18. **Sotoodehnia N, Isaacs A, de Bakker PI, Dorr M**, et al. 2010. Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nat Genet* **42**: 1068–76.
19. **Teslovich TM, Musunuru K, Smith AV, Edmondson AC**, et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**: 707–13.
20. **Auer PL, Lettre G.** 2015. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* **7**: 16.
21. **Regalado ES, Guo DC, Villamizar C, Avidan N**, et al. 2011. Exome sequencing identifies SMAD3 mutations as a cause of familial thoracic aortic aneurysm and dissection with intracranial and other arterial aneurysms. *Circ Res* **109**: 680–6.
22. **Choi M, Scholl UI, Ji W, Liu T**, et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* **106**: 19096–101.
23. **Ng SB, Bigham AW, Buckingham KJ, Hannibal MC**, et al. 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42**: 790–3.
24. **Ng SB, Buckingham KJ, Lee C, Bigham AW**, et al. 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**: 30–5.
25. **Choi M, Scholl UI, Yue P, Bjorklund P**, et al. 2011. K+ channel mutations in adrenal aldosterone-producing adenomas and hereditary hypertension. *Science* **331**: 768–72.
26. **Bamshad MJ, Ng SB, Bigham AW, Tabor HK**, et al. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**: 745–55.
27. **Comino-Mendez I, Gracia-Aznarez FJ, Schiavi F, Landa I**, et al. 2011. Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nat Genet* **43**: 663–7.
28. **Levy S, Sutton G, Ng PC, Feuk L**, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5**: e254.
29. **Ng PC, Levy S, Huang J, Stockwell TB**, et al. 2008. Genetic variation in an individual human exome. *PLoS Genet* **4**: e1000160.
30. **Marian AJ.** 2009. Nature's genetic gradients and the clinical phenotype. *Circ Cardiovasc Genet* **2**: 537–9.
31. **Rivas MA, Beaudoin M, Gardet A, Stevens C**, et al. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* **43**: 1066–73.
32. **Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G**, et al. 2012. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat Genet* **44**: 1326–9.
33. **Jonsson T, Atwal JK, Steinberg S, Snaedal J**, et al. 2012. A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**: 96–9.
34. **Fritsche LG, Igl W, Bailey JN, Grassmann F**, et al. 2016. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet* **48**: 134–43.
35. **Pritchard JK, Cox NJ.** 2002. The allelic architecture of human disease genes: common disease-common variant. . .or not? *Hum Mol Genet* **11**: 2417–23.
36. **Zuk O, Hechter E, Sunyaev SR, Lander ES.** 2012. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* **109**: 1193–8.
37. **Song YS, Wang F, Slatkin M.** 2010. General epistatic models of the risk of complex diseases. *Genetics* **186**: 1467–73.
38. **Furrow RE, Christiansen FB, Feldman MW.** 2011. Environment-sensitive epigenetics and the heritability of complex diseases. *Genetics* **189**: 1377–87.
39. **Marian AJ.** 2012. Elements of 'missing heritability'. *Curr Opin Cardiol* **27**: 197–201.
40. **Libby P.** 2013. Mechanisms of acute coronary syndromes and their implications for therapy. *N Engl J Med* **368**: 2004–13.
41. **Lopez-Otin C, Blasco MA, Partridge L, Serrano M**, et al. 2013. The hallmarks of aging. *Cell* **153**: 1194–217.
42. **Adrogue HJ, Madias NE.** 2007. Sodium and potassium in the pathogenesis of hypertension. *N Engl J Med* **356**: 1966–78.
43. **Deribe YL, Pawson T, Dikic I.** 2010. Post-translational modifications in signal integration. *Nat Struct Mol Biol* **17**: 666–72.
44. **McCarthy MI.** 2010. Genomics, type 2 diabetes, and obesity. *N Engl J Med* **363**: 2339–50.
45. **Shulman GI.** 2014. Ectopic fat in insulin resistance, dyslipidemia, and cardiometabolic disease. *N Engl J Med* **371**: 1131–41.
46. **Lee DS, Park J, Kay KA, Christakis NA**, et al. 2008. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci USA* **105**: 9880–5.
47. **Barabasi AL, Gulbahce N, Loscalzo J.** 2011. Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12**: 56–68.
48. **Jostins L, Ripke S, Weersma RK, Duerr RH**, et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**: 119–24.
49. **Sirota M, Schaub MA, Batzoglou S, Robinson WH**, et al. 2009. Autoimmune disease classification by inverse association with SNP alleles *PLoS Genet* **5**: e1000792.
50. **Cotsapas C, Voight BF, Rossin E, Lage K**, et al. 2011. Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet* **7**: e1002254.
51. **Purcell SM, Wray NR, Stone JL, Visscher PM**, et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**: 748–52.
52. **Lichtenstein P, Yip BH, Bjork C, Pawitan Y**, et al. 2009. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* **373**: 234–9.
53. **Stearns FW.** 2010. One hundred years of pleiotropy: a retrospective. *Genetics* **186**: 767–73.
54. **Wagner GP, Zhang J.** 2011. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat Rev Genet* **12**: 204–13.
55. **Solovieff N, Cotsapas C, Lee PH, Purcell SM**, et al. 2013. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**: 483–95.
56. **Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R**, et al. 2012. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**: 572–80.
57. **Shriner D.** 2012. Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Front Genet* **3**: 1.
58. **Lawlor DA, Harbord RM, Sterne JA, Timpson N**, et al. 2008. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**: 1133–63.
59. **Glymour MM, Tchetgen Tchetgen EJ, Robins JM.** 2012. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions *Am J Epidemiol* **175**: 332–9.
60. **Gottesman, II, Gould TD.** 2003. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* **160**: 636–45.
61. **Gershon ES, Goldin LR.** 1986. Clinical methods in psychiatric genetics. I. Robustness of genetic marker investigative strategies. *Acta Psychiatr Scand* **74**: 113–8.
62. **Gottesman II, Shields J.** 1967. A polygenic theory of schizophrenia. *Proc Natl Acad Sci USA* **58**: 199–205.
63. **Gottesman II, Shields J.** 1973. Genetic theorizing and schizophrenia. *Br J Psychiatry* **122**: 15–30.
64. **John B, Lewis KR.** 1966. Chromosome variability and geographic distribution in insects. *Science* **152**: 711–21.
65. **Leboyer M, Bellivier F, Nosten-Bertrand M, Jouvent R**, et al. 1998. Psychiatric genetics: search for phenotypes. *Trends Neurosci* **21**: 102–5.
66. **Keating M, Atkinson D, Dunn C, Timothy K**, et al. 1991. Linkage of a cardiac arrhythmia, the long QT syndrome, and the Harvey ras-1 gene. *Science* **252**: 704–6.
67. **Vincent GM, Timothy KW, Leppert M, Keating M.** 1992. The spectrum of symptoms and QT intervals in carriers of the gene for the long-QT syndrome. *N Engl J Med* **327**: 846–52.
68. **Silva CT, Kors JA, Amin N, Dehghan A**, et al. 2015. Heritabilities, proportions of heritabilities explained by GWAS findings, and implications of cross-phenotype effects on PR interval. *Hum Genet* **134**: 1211–9.

69. **Lalouel JM, Le Mignon L, Simon M, Fauchet R**, et al. 1985. Genetic analysis of idiopathic hemochromatosis using both qualitative (disease status) and quantitative (serum iron) information. *Am J Hum Genet* **37**: 700–18.
70. **Greenberg DA, Delgado-Escueta AV, Widelitz H, Sparkes RS**, et al. 1988. Juvenile myoclonic epilepsy (JME) may be linked to the BF and HLA loci on human chromosome 6. *Am J Med Genet* **31**: 185–92.
71. **Leppert M, Burt R, Hughes JP, Samowitz W**, et al. 1990. Genetic analysis of an inherited predisposition to colon cancer in a family with a variable number of adenomatous polyps. *N Engl J Med* **322**: 904–8.
72. **Kariuki SN, Franek BS, Kumar AA, Arrington J**, et al. 2010. Trait-stratified genome-wide association study identifies novel and diverse genetic associations with serologic and cytokine phenotypes in systemic lupus erythematosus. *Arthritis Res Ther* **12**: R151.
73. **Kariuki SN, Ghodke-Puranik Y, Dorschner JM, Chrabot BS**, et al. 2015. Genetic analysis of the pathogenic molecular sub-phenotype interferon-alpha identifies multiple novel loci involved in systemic lupus erythematosus. *Genes Immun* **16**: 15–23.
74. **Balmain A.** 2002. Cancer as a complex genetic trait: tumor susceptibility in humans and mouse models. *Cell* **108**: 145–52.
75. **Hunter KW, Crawford NP.** 2008. The future of mouse QTL mapping to diagnose disease in mice in the age of whole-genome association studies. *Annu Rev Genet* **42**: 131–41.
76. **Castellanos-Martin A, Castillo-Lluva S, Saez-Freire Mdel M, Blanco-Gomez A**, et al. 2015. Unraveling heterogeneous susceptibility and the evolution of breast cancer using a systems biology approach. *Genome Biol* **16**: 40.
77. **Li Y, Huang J, Amos CI.** 2012. Genetic association analysis of complex diseases incorporating intermediate phenotype information *PLoS ONe* **7**: e46612.
78. **Pharoah PD, Antoniou AC, Easton DF, Ponder BA.** 2008. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med* **358**: 2796–803.