GENETIC ARCHITECTURE OF COMPLEX DISEASE IN HUMANS:

A CROSS-POPULATION EXPLORATION

Urko Martínez Marigorta

DOCTORAL THESIS UPF / 2012

THESIS DIRECTOR

Dr. Arcadi Navarro

Department of Experimental and Health Sciences

UNIVERSITAT POMPEU FABRA

A los que me hicieron,

y a los que me ayudaron a hacer

Keep calm and carry on

Government of the United Kingdom, 1939

Acknowledgements

Vull donar les gràcies a tothom que m'he trobat i m'ha ajudat aquests anys. M'he sentit molt apreciat, i he pogut gaudir de la companyia de gent que m'ha donat moltes ensenyances. N'he après molt. No oblidaré tampoc les lliçons que he pogut extraure de vertaders models científics. Us he intentat copiar molt: la capacitat intel·lectual, la resistència mental, l'empenta i rauxa, la integritat científica, el seny, procurar admiració en l'enveja i la humilitat en el no saber. En especial, vull citar expressament la visió, empenta i il·lusió de l'Arcadi.

Agrair també l'amistat a tota la gent de, tant de Barcelona (euskaldunak!) com de Girona. En sigui aquesta feina que he intentat fer de valent una petita penyora.

Pocas palabras bastan para cada uno de vosotros. Todo el amor, el carácter y los esfuerzos económicos. Vosotros y yo sabemos porqué no ha sido fácil. Esté cada uno donde esté, espero orgullo de este trabajo, aunque no sea más que una pequeña parte de una lucha muy antigua. En el recuerdo.

Errexena Zu. Gaurko eta betiko. Egunero maite detena. Egunero gozatzen detena. Ikasi besterik ez det egin, eta ikastea da gehien gustuko detena. Egin dezun guztiaz, benetan harro. Noiz agertu zinan ezin gogoratu. Zu zara indarra eta bereziki, zu zara laguna bidean…

Abstract

The aetiology of common diseases is shaped by the effects of genetic and environmental factors. Big efforts have been devoted to unravel the genetic basis of disease with the hope that it will help to develop new therapeutic treatments and to achieve personalized medicine. With the development of high-throughput genotyping technologies, hundreds of association studies have described many loci associated to disease. However, the depiction of disease architecture remains incomplete. The aim of this work is to perform exhaustive comparisons across human populations to evaluate pressing questions. Our results provide new insights in the allele frequency of risk variants, their sharing across populations and the likely architecture of disease.

Resumen

La etiología de las enfermedades comunes está formada por factores genéticos y ambientales. Se ha puesto mucho empeño en describir sus bases genéticas. Este conocimiento será útil para desarrollar nuevas terapias y la medicina personalizada. Gracias a las técnicas de genotipado masivo, centenares de estudios de asociación han descrito una infinidad de genes asociados a enfermedad. Pese a ello, la arquitectura genética de las enfermedades no ha sido totalmente descrita. Esta tesis pretende llevar a cabo exhaustivas comparaciones entre poblaciones para responder diversas preguntas candentes. Nuestros resultados dan pistas sobre la frecuencia de los alelos de riesgo, su presencia entre poblaciones y la probable arquitectura de las enfermedades.

Preface

The central dogma of molecular biology describes that genetic information in biological systems is mostly transferred in a single direction that goes from DNA to RNA and proteins. It was established in the 60s (20th century!) but still prompted me and many of my lab fellows to make research in biology. It will be so forever. Yet, young people do not join anymore because of an easy dogma but the beauty of an ever filling puzzle.

Complexity is not the only trend in biology. The development of high throughput technologies such as those used in genomics is producing increasing amounts of biological data. Since a while ago, not data but making sense of it is the main priority for biologists.

One of the deepest reasons behind the abovementioned revolutions is related to a special kind of phenotype: pathology in humans. It is hoped that unveiling the factors that shape disease will permit to develop new disease treatments and large amounts of money are devoted, for instance, to medical genetics.

This thesis represents a small work within the efforts to make sense of publicly available data to understand disease complexity. No new data are provided. Nonetheless, a few inferences on the genetic architecture of disease are made. While the answers we provide will certainly age fast, they might prove helpful to anyone interested in the current debates in the field of genetics of disease.

Index

1 Introduction

Cæteris paribus…
All other things being equal...

1.1. Epidemiology and the genetics of disease

The work in this thesis focuses on the genetic architecture of complex disease across populations. Concepts from such disparate fields as epidemiology and quantitative genetics should be introduced. I prioritized an introduction about disease in populations, the role of genetics in disease, the types of risk variants and an introduction to Mendelian disease.

1.1.1. Epidemiology and Genetic Epidemiology

Epidemiology studies the occurrence and determinants of disease in populations (Rothman 2002). Usually, causation is multifactorial and variable across individuals. Epidemiology goes beyond the description of causes and measures their strength in populations. These calculations allow comparing disease in different groups; thereby giving perspectives in public health (Figure 1).

A key step in any epidemiological study is the definition of the population at risk, the group of individuals from which conclusions on disease occurrence can be extracted. The two main descriptive measures of disease occurrence are incidence (rate at which new cases occur during a specified period) and prevalence (the proportion of a population that are cases at a point in time).

Two main types of study are used in epidemiology: cohort and case-control studies. Individuals from a cohort are initially disease-free and ascertained because of a shared characteristic or risk exposure (e.g. the 1958 British Birth cohort). These individuals are followed over a period of time, and the estimates of relative risk (RR) can be drawn. Relative risk measures the ratio of disease probabilities in exposed over unexposed persons. These measures serve to weight the attributable risks in populations.

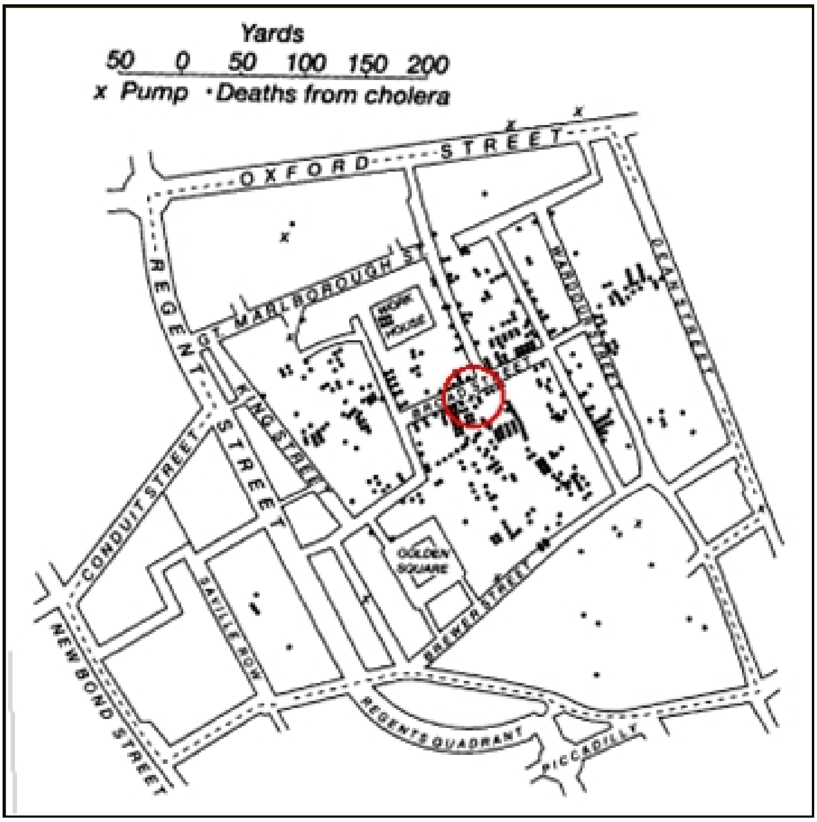


Figure 1. Map by John Snow showing London cholera cases in 1854. The clustering of cases around Broad Street's pump (in current Soho) served to avoid new cases and helped to eradicate the epidemics of cholera in London.

Although the gold standard in the field, cohort studies are lengthy in time because only a fraction of individuals will develop disease. The case-control approach is an immediate alternative, in particular for rare conditions (Rothman 2002). In this design, disease patients are identified and their suspected exposures are compared with those of healthy controls. Ideally, control individuals must be unbiased representatives of the same population that gives rises to the cases; special efforts are devoted to statistical adjustment to avoid confounding. From a contingency table, the association of exposures and case-control status is measured by means of the odds ratio (OR), the ratio of the odds of exposure in cases to the odds in the controls (Rothman 2002).

Given their observational nature, epidemiological studies must deal with all kind of possible biases. For instance, selection bias may occur if the selected individuals are not good representatives of the population at risk. Any difference in variables other than the exposure among the source under study may lead to a confounding. Epidemiologists devote especial efforts to avoid any such confounding (i.e. matching samples for known confounders).

Genetic epidemiology focuses in the study of the role of genetic and environmental factors in disease prevalence (Thomas 2004; Ziegler et al. 2010). Its primary focus lies in inherited variation through comparisons among relatives. Genetic epidemiology aims to study if a given disease “runs in families” and the role of genetics in its inheritance patterns. Together with cohort and case-control studies, family-based designs are the cornerstones of the field.

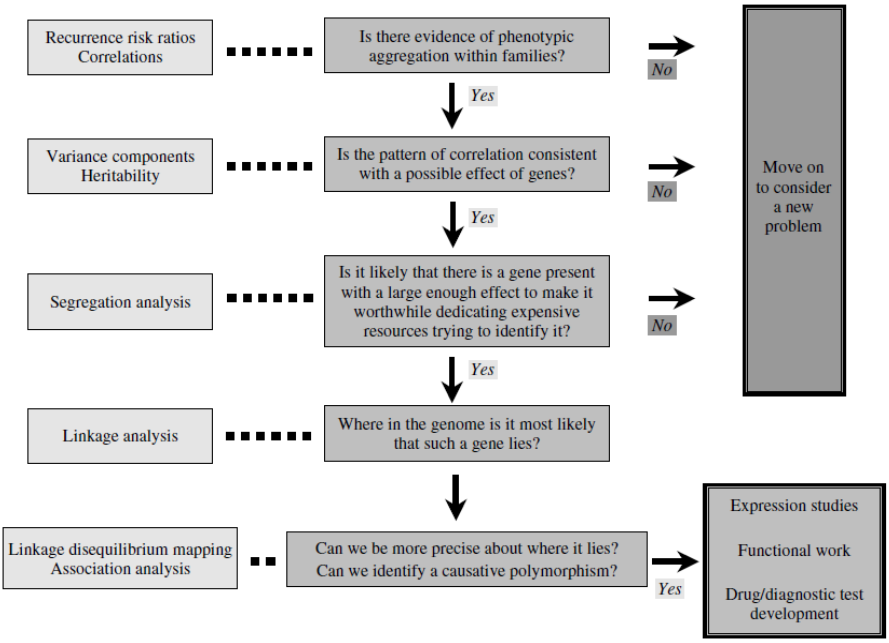

Recurrence risk ratios
Correlations
Is there evidence of phenotypic aggregation within families?
No
Yes
Variance components
Heritability
Is the pattern of correlation consistent with a possible effect of genes?
No
Yes
Segregation analysis
Is it likely that there is a gene present with a large enough effect to make it worthwhile dedicating expensive resources trying to identify it?
No
Yes
Move on to consider a new problem
Linkage analysis
Where in the genome is it most likely that such a gene lies?
Linkage disequilibrium mapping
Association analysis
Can we be more precise about where it lies?
Can we identify a causative polymorphism?
Yes
Expression studies
Functional work
Drug/diagnostic test development

Figure 2. Progression steps in genetic epidemiology.
Summary of main tasks done in genetic epidemiology, from the initial hypothesis on a role for genetics underlying familial aggregation to fine mapping (Thomas 2004).

The process of genetic epidemiology has been equalled to a progressive task whose final aim is the unravelling of the genetic basis of disease (Figure 2). Eventually, the full description of the genetic architecture of a given disease consists in the mapping of all genetic variants associated to it, along with their effect sizes and interactions with other genetic and environmental factors. Such a description would complement other evidences (e.g. migrant studies) to understand the role of genetics in the differential prevalence of disease across populations and would help in the study of the evolutionary history of disease.

1.1.2. Familial resemblance and heritability

The study of patterns of familial aggregation constitutes the first step to ensure that genetic variants shape disease aetiology. Familial resemblance refers to the increase in phenotypic correlation (e.g. disease prevalence) of relatives compared to unrelated individuals (Ziegler et al. 2010). A comparison of familial resemblances estimated from different relationships among relatives (e.g. siblings, cousins, 2nd cousins) permits an initial exploration of the range of possible genetic architectures of disease. Nevertheless, recurrence ratios among relatives can arise due to either genetic or environmental factors and thus further methods are needed to ensure that genetics plays a role in disease.

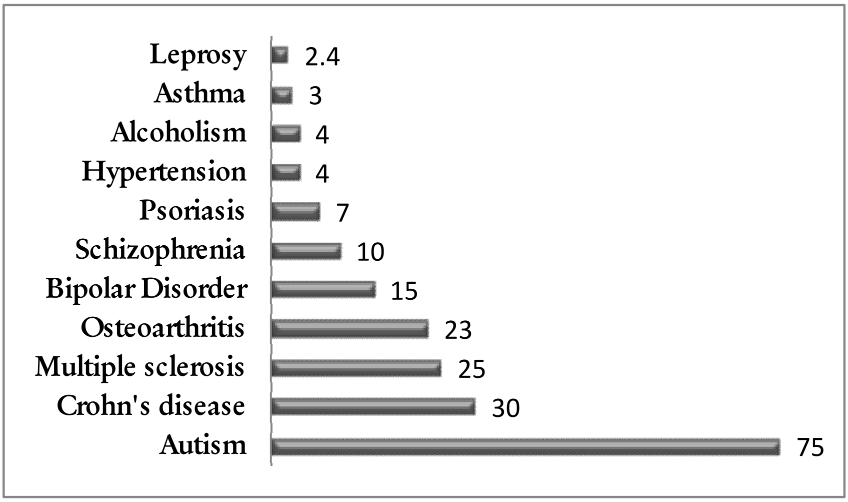


Figure 3. Sibling risk ratios for several disease traits (Ziegler et al. 2010). Recurrence risk ratios measure the prevalence of disease in relatives of cases compared to that in general population. The ranges of values risk ratios can take depend on disease prevalence.

Heritability is a population parameter that permits to weight the extent to which variation in a trait is due to genetic factors (Visscher et al. 2008). Specifically, variation observed across individuals is partitioned into unobserved genetic and environmental categories. Each of these partitions captures part of the variance, and the ratio of variances weights the influence of genetic variability in the observed sum of variances (H^2, broad-sense heritability). Further partition of the genetic category into additive, dominant and interaction effects permits to measure h^2 (narrow-sense heritability): the fraction of variation attributable to additive genetic factors that serves to predict similarities between parents and offspring. However, the incorporation of complex phenomena such as gene-by-environment interactions to the statistical model difficult these calculations.

To estimate heritability in humans, the correlation among relatives in families can be used to compare observed vs. expected resemblance (Visscher et al. 2008). The most commonly used classical methods are the regression of the offspring on the parental phenotypes, or the difference in correlation between monozygotic (MZ) and dizygotic (DZ) twins. Yet, estimations can be biased if non-genetic factors (e.g. increased environmental correlations in MZ twins) are not accurately modelled (Boomsma et al. 2002). Notably, the availability of genetic markers and re-sequencing data permits to refine the estimates from pedigrees with those calculated by means of distantly related individuals (Yang et al. 2010).

The interpretation of heritability can suffer from several misconceptions (Visscher et al. 2008). It is easily forgotten that estimates of heritability are population-specific and can vary across sex and age. Additionally, they are point estimates in time: traits with large heritability can be heavily affected by changes in the environment (e.g. secular rise in human height in modern societies). Hence, differences in environment must be ruled out to conclude that differences among populations are genetic in origin (Feldman and Lewontin 1975). Moreover, the comparison of heritability estimates across diseases and populations can be heavily affected by differences in incidence. Thus, heritability for categorical disease traits is better estimated through the threshold liability model (Visscher et al. 2008). Specifically, an underlying distribution of risk factors is assumed to assign a genetic score of risk to each individual, and prevalence is then used to set the threshold score of disease status.

Positive estimates of heritability predict a role of genetic variants in disease prevalence. Indeed, how easily large-effect alleles are found increases with heritability (phenotype predicts genotype). However, large estimates of heritability do not inform about disease architecture and are not at odds with the observation that abundant cases are sporadic and have no diseased relatives.

1.1.3. Types of genetic variation in disease

Disease forms a category of abnormal phenotypes characterized by pathology that appear from the expression of DNA under the influence of environmental factors. A consideration of the types and forces shaping polymorphism is needed to deepen our knowledge of the genetics of disease.

Permanent changes in the DNA molecule arise by mutation (Balding et al. 2007; Crow 2000; Eyre-Walker 2010). Broadly speaking, two kinds of variation can be distinguished in our genomes according to their size: point mutations and structural variation (Frazer et al. 2009). The former are substitutions of a single base known as single nucleotide polymorphisms (SNPs) that form the most prevalent change in human genomes (>53 million SNPs are deposited in the dbSNP repository, release 137). Human genomes carry 10 to 200 *de novo* single-base mutations and the two haploid genomes in individuals harbour 3×10^6 differences, or 1 in 1,000 nucleotides (Conrad et al. 2011; Reich et al. 2002).

Each form of a SNP is called allele. Most SNPs are bi-allelic (Slatkin 2008) and their frequency permits to distinguish between *major* and *minor* alleles, the frequency of the latter being the *minor allele frequency* (MAF). The sequence in origin of alleles permits to distinguish between 'ancestral' and 'derived' allele, the latter being those that have arisen more recently. Moreover, the position of SNPs in the genome permits a classification based on functionality (Cargill et al. 1999). Given the low density of genes in the genome, most SNPs are *intergenic* (Sachidanandam et al. 2001). This feature does not preclude a functional effect, as intergenic SNPs may have regulatory roles (Cooper and Shendure 2011). *Genic* SNPs are further classified into *non-coding* (*intronic* and *5′* and *3′ UTR*) and *coding* (*synonymous, missense* and *nonsense*). The action of natural selection varies across functional categories of SNPs and shapes their allele frequency in human populations (Barreiro et al. 2008).

Structural variants (SVs) are the second class of genetic variants. SVs range from small *indels* to large *chromosomal rearrangements* (Frazer et al. 2009). Each type has distinctive rates of mutation and variable genome dynamics (Zhang et al. 2009). Examples of SV role in disease are abundant (Eichler et al. 2010; Zhang et al. 2009).

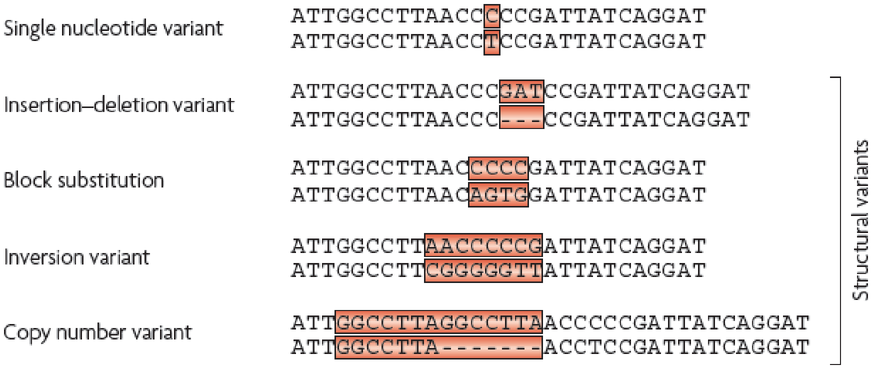

Single nucleotide variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCTCCGATTATCAGGAT
Insertion–deletion variant
ATTGGCCTTAACCCGATCCGATTATCAGGAT
ATTGGCCTTAACCC---CCGATTATCAGGAT
Block substitution
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACAGTGGATTATCAGGAT
Inversion variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTCGGGGGTTATTATCAGGAT
Copy number variant
ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTA-------ACCTCCGATTATCAGGAT
Structural variants

Figure 4. Nomenclature and representation of human genetic variation. Single nucleotide variants are DNA polymorphisms in which a single base is altered. Insertion-deletion (*Indels*) variants occur when one or more base pairs are present in some genomes but absent in others. Block substitutions describe cases in which a string of nucleotides varies between two genomes. An inversion variant is one in which the order of the base pairs is reversed in a defined section of a chromosome. Copy number variants occur when nearly identical sequences are repeated in some chromosomes but not others. Adapted from (Frazer et al. 2009)

Population geneticists study the forces that affect the evolution of polymorphism in natural populations (Hartl and Clark 2007).

Besides mutation, recombination also increases genetic variability (Slatkin 2008). Instead of creating new variants, it does so by placing different variants in the same chromosome. The homologous chromosomes of maternal and paternal origin align and exchange segments during meiosis. Thus, recombination creates new combinations of alleles via germ-line mosaics that are passed to the next generation (haplotypes). Recombination rates vary across the genome and hotspots of 1-2 kb length with >100-fold increased rates can be distinguished (Myers et al. 2005).

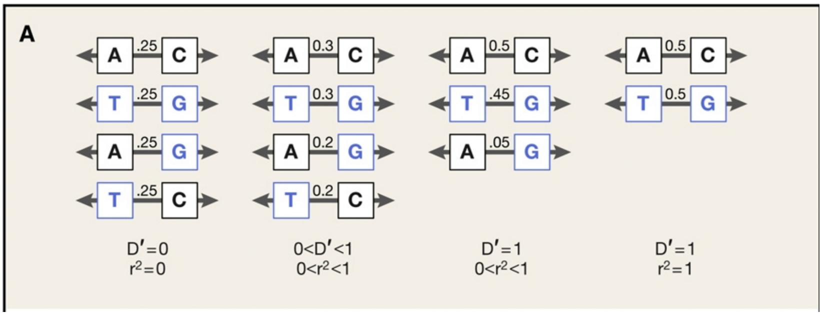


Figure 5. Differences between metrics of Linkage Disequilibrium.
Left: Two markers with 0.5 allele frequency are not linked and each resulting haplotype has a frequency of 0.25. Middle left: Alleles at one marker correlate partially with alleles at the other marker. Middle right: The two alleles are tightly linked. Right: An allele at one marker predicts perfectly the allele at the other marker (only in this situation r^2=1). Adapted from (Raychaudhuri 2011)

Haplotypes are directly related to Linkage Disequilibrium (LD). In populations, LD occurs when combinations of alleles are found as haplotypes more often than expected by their allele frequencies (Balding et al. 2007). The appearance of new alleles through mutation creates LD, and genetic drift and natural selection help in its maintenance. In contrast, the reshuffling of present haplotypes by gametic recombination forces the decay of LD. The evolutionary history of populations and the presence of recombination hotspots determine the structure and size of segments that are in LD (*haplotype blocks*) and its variation across populations (Reich et al. 2001).

Genetic drift is another force that governs the fate of genetic diversity (Hartl and Clark 2007). Living populations are of finite size, and individuals have different number of offspring. Hence, each generation represents a sampling of the allele frequencies in the previous generation. The succession of random changes in allele frequencies diminishes genetic variation, as alleles are either lost or fixed. The effects of genetic drift depend on population size, being more extensive in small populations (Hartl and Clark 2007). Importantly, the estimation of effective population size permits to enclose the effects of genetic drift in populations (Fisher 1930; Wright 1931). Genetic drift plays a major role driving differentiation across populations (Nagylaki 1985). Populations are hierarchically structured because individuals breed with partners from close habitats (Balding et al. 2007). After divergence, allele frequencies fluctuate and thus drift decreases diversity within populations whilst differentiating them.

The presence of genetic variants that alter the survival and reproduction of individuals (*fitness*) opens room for the action of natural selection. Different modes of selection are distinguished (Hartl and Clark 2007). Purifying selection decreases the frequency of variants that diminish the fitness of the individual, whilst positive selection increases the frequency of those improving fitness. Both types of selection result in a decrease of genetic diversity. Finally, balancing selection acts prioritizing the maintenance of different alleles at the same locus and leads to increased levels of heterozygosity.

1.1.4. Mendelian disease

Mendelian diseases constitute the simplest category of genetic disease and are usually monogenic (there is a single gene harbouring deleterious mutations that cause pathology). These disorders run in families through classical Mendelian inheritance. The phenotypic analysis of affected individuals (probands) permits an exploration of the likely architecture of disease (e.g. recessive - dominant segregation (Weiss 1999)).

Genetic mutations that cause Mendelian disorders evolve under strong purifying selection due to their severe effects on the fitness of individuals and are maintained at low frequency in populations (Reich and Lander 2001). Their prevalence is usually maintained by a mutation-selection balance; thereby new disease-causing mutations are culled by natural selection (Di Rienzo and Hudson 2005). This equilibrium model leads to high allelic heterogeneity, whereby mutations with different evolutionary histories in the same gene produce the same phenotype (e.g. 302 different mutation events were found in 424 families with Haemophilia B from UK). Importantly, the effective mutation rate for Mendelian disease correlates with gene length (Weiss 1999).

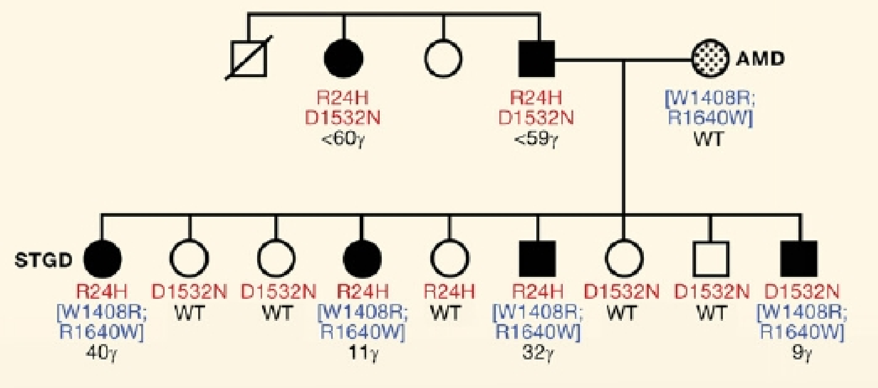


Figure 6. Phenotypic heterogeneity in Mendelian disease.
Example from a pedigree in which different combinations of ABCA4 alleles determine the age-of-onset and severity of Stargardt macular dystrophy (Lupski et al. 2011).

In last 30 years, linkage analysis has succeeded in unveiling the genetic causes of hundreds of monogenic diseases. Nonetheless, the analysis of Mendelian diseases is usually not so straightforward (Figure 6). Instead, several complications such as imprinting, age-dependant penetrance and phenotypic heterogeneity can obscure the phenotype-genotype map and alter their inheritance patterns (Ziegler et al. 2010). These phenomena blur the distinction between Mendelian and Complex diseases.

There are known knowns and known unknowns.
But there are also unknown unknowns:
things we do not know that we don't know…
Donald Rumsfeld, 2002 (adapted)

1.2. Complex diseases

Complex diseases are caused by the effects of genetic and environmental factors. The modern increase in prevalence of these diseases fuels the research to unravel their genetic architecture.

1.2.1. Introduction and prevalence

The word *complex* fits perfect to label diseases such as diabetes or schizophrenia. These diseases arise through intricate and variable interplay between environmental and genetic factors (Botstein and Risch 2003). Complex diseases show familial clustering, but follow non-Mendelian inheritance patterns (Ziegler et al. 2010). Complex diseases are usually acquired at late ages, but the age of onset is highly variable (Wright et al. 2003). Complex diseases present shared morbidities that can be classified in medical handbooks, but their manifestations differ across individuals (Weiss 1999). These inherent complexities difficult the task of uncovering risk factors and the understanding of the ultimate reasons that transform healthy individuals into patients.

Several goals underlie the interest to study the genetics of disease. Two of the main aims consist, first, in gathering biological knowledge about the pathways involved in aetiology that can lead to targets for drug design (Visscher et al. 2012) and, second, using this knowledge to develop personalized medicine and identify healthy individuals at increased risk of disease (Wray et al. 2010).

The prevalence of most complex diseases has exploded in the last two centuries (Gibson 2009). This major shift in human pathology has coincided with the decrease in prevalence of communicable diseases and the ongoing lengthening in life expectancy in developed societies (Di Rienzo and Hudson 2005; Wright et al. 2003). As shown in Figure 7, Several of the diseases linked to "affluence", such as coronary disease, have substituted trauma and infectious disease as the first causes of death (Pollard 2008).

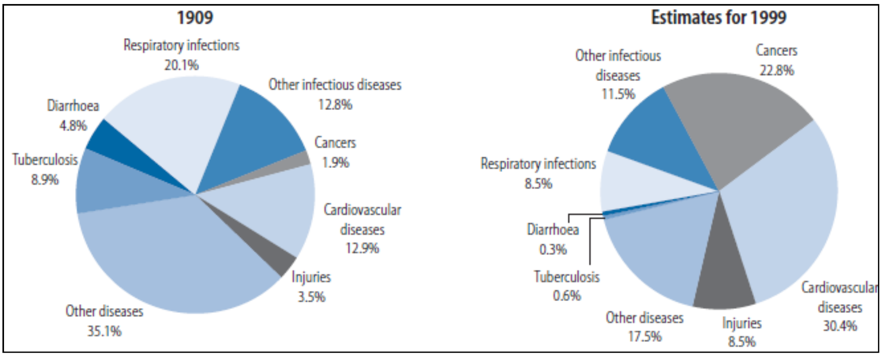


Figure 7. Change in death cause profile in Chile between 1909 and 1999. A century ago, almost half of deaths were caused by infectious disease and the number of deaths caused by cancer or cardiovascular disease constituted a minority. This situation has reversed. From (Pollard 2008).

Over the last decades, this epidemiological transition has extended into developing countries, emerging worldwide. For instance, type 2 diabetes presents epidemic prevalence in populations that are undergoing rapid lifestyle shifts such as in urban India (Diamond

2003). Several evolutionary hypotheses have been put forward to explain current prevalence of complex disease and the differences observed across human populations. Indeed, an evolutionary perspective of disease may help to choose appropriate methods to unravel their genetic architecture (Di Rienzo and Hudson 2005).

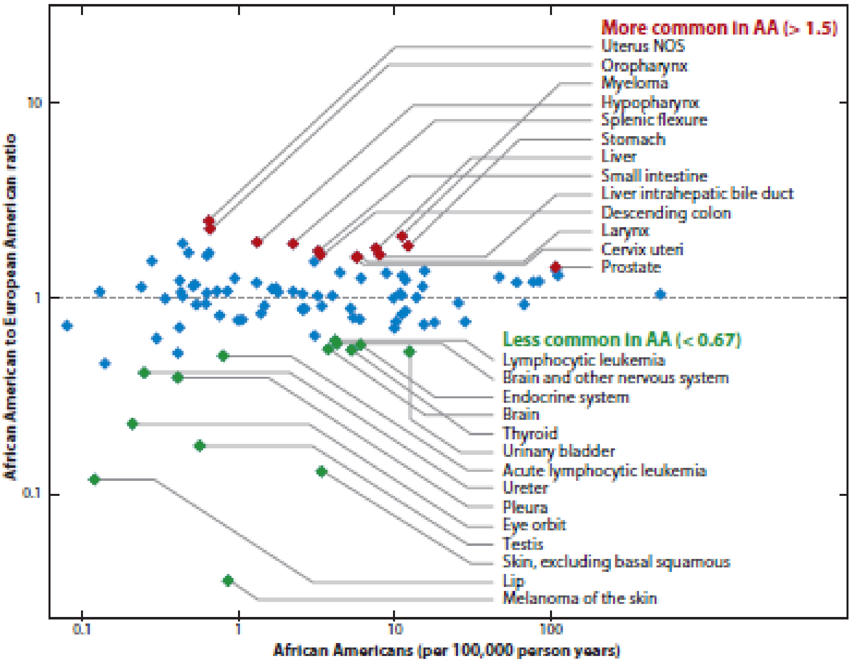


Figure 8. Disease prevalence varies across genetic ancestries. Relative frequencies of cancers in African and European Americans are shown. Cancer types that present significant differences in prevalence (marked in red and green) can be due to genetic, environmental or gene-by-environment variability. From (Winkler et al. 2010).

1.2.2. Early models and inferences of disease architecture

Genetic architecture refers to the underlying basis of a phenotypic trait, composed by the distribution of causal variants, allele frequencies, effect sizes and the patterns of pleiotropy, dominance and epistasis they maintain (Visscher et al. 2010). These parameters are modelled by the evolutionary processes that act on the phenotype. Importantly, current efforts in modelling disease evolution focus on the fitness effects of susceptibility variants in ancestral environments rather than on current clinical parameters (Di Rienzo and Hudson 2005). Nevertheless, factors such as the late age of onset, the incomplete penetrance of causal variants and the myriad of environmental contributions difficult the modelling of fitness and disease evolution (Di Rienzo and Hudson 2005).

First experimental studies in early 50s and 60s found strong associations (e.g. OR>100) between common variations in ABO and HLA and several types of cancer and autoimmune diseases (Bodmer and Bonilla 2008). Interestingly, these alleles present both high allele frequencies and large effect sizes (e.g. HLA-B27 allele in Ankylosing spondylitis).

The high frequency of these variants hinted at a role of fluctuating selection between Palaeolithic and Neolithic times (Pritchard and Cox 2002). Initial hypotheses put forward to explain prevalence focused in the large difference between current lifestyles and those in the times when modern humans evolved. James V. Neel proposed the ancestral *thrifty genotype hypothesis* model for type 2 diabetes (Neel 1962). This author aimed to explain the paradox of diabetes prevalence given its detrimental effects on fitness. This hypothesis suggests that alleles predisposing to rapid releases of insulin in humans would be adaptive under ancient cycles of feast and famine but maladaptive in modern societies. The recent environmental shift and the increase in life expectancy would have unmasked the latent genetic susceptibilities to disease (Di Rienzo and Hudson 2005). Evolutionary visions have been proposed to explain the differential prevalence of diabetes and obesity across populations ("New world syndrome"), the higher prevalence of hypertension in African Americans ("sodium retention hypothesis") or of inflammatory and allergic diseases in urban peoples ("hygiene hypothesis"). Several of these hypotheses are disaccredited, but the tension between ancestral and current environments frames the necessity of non-equilibrium models for complex disease.

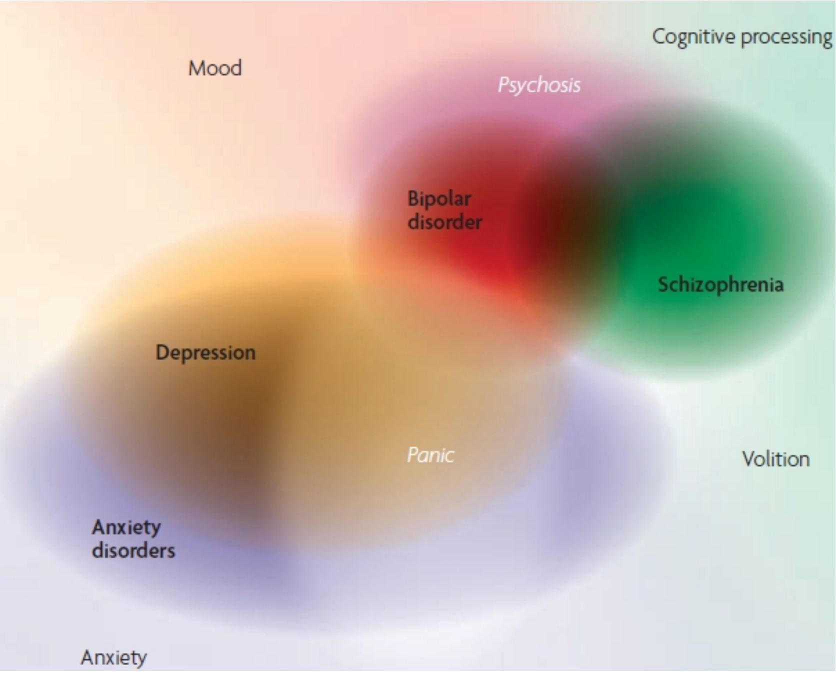


Figure 9. The nebulous architecture of complex disease.
Complex diseases arise due to interactions of genetic and environmental factors that control the underlying causal traits and emerge from the range of normal variation in healthy individuals. The figure (Burmeister et al. 2008) for psychiatric diseases serves to illustrate this point.

The study of complex disease genetics with positional cloning methods exploded in the 80s thanks to the availability of genomic maps through restriction fragment length polymorphism (Botstein et al. 1980). It is striking that most successful findings were later revealed as false positives (Botstein and Risch 2003), with the main successes being restricted to familial forms of disease (e.g. BRCA1 for breast cancer in young cases). Overall, a scant number of common variants (MAF>10%) such as APOE-ε4 for Alzheimer's disease and PPARγ for type 2 diabetes were available by mid-90s.

1.2.3. The "Common Disease/Common Variant" hypothesis

The somewhat failure of linkage methods, that had been successful for Mendelian diseases, prompted the debate on the methodology necessary to discover variants for complex disease. The few such variants available at the turn of the century evidenced the absence of high-risk alleles and suggested that different variants must account for disease aggregation in relatives.

In the late 90s, several landmark studies explored the methodological challenge to unravel disease-associated variants of low risk (Chakravarti 1999; Lander 1996; Reich and Lander 2001; Risch and Merikangas 1996). For instance, Risch and Merikangas (Risch and Merikangas 1996) compared linkage to association methods and showed the latter would be powered to find low-risk alleles in the absence of allelic heterogeneity. Importantly, the authors acknowledged the strict thresholds necessary to avoid false positives inflation due to the large number of polymorphism that ought to be tested.

Meanwhile, a variety of papers explored disease architecture from a simulation perspective (Pritchard 2001; Pritchard and Cox 2002; Reich and Lander 2001). These studies explored ranges of selective effects for risk variants to answer questions on the frequency and number of alleles expected in disease loci. Overall, only a few of the loci eventually associated to disease would harbour disease mutations. With the caveat of the difficulty to guess in advance which of the disease loci are polymorphic for these variants, the predicted values of allele identity were encouraging for the successful completion of association studies.

These works showed that a regime of weak purifying selection would prevent the fixation of disease variants, leaving them at intermediate frequencies. Given the correlation between genetic variance and heterozygosity (Visscher et al. 2012), these intermediate-frequency alleles would contribute most to the genetic variance of disease, i.e. heritability. Finally, simulation studies suggested that it was possible, at least in principle, that a relatively reduced set of common variants could account for most of the individuals with disease mutations in populations.

The works mentioned above, amongst others, contributed to the establishment of the “Common Disease/Common Variant” (CD/CV) hypothesis. This paradigm states that common variants in susceptibility genes account for complex disease genetic risk. The slight deleterious effects of these variants in individual fitness explain their high allele frequencies. Their frequency would counterweight the low relative risks of these variants, thus explaining the large prevalence of complex disease. In spite of criticisms (Terwilliger and Weiss 2003; Weiss and Terwilliger 2000), the CD/CV hypothesis was established as the paradigm in human genetics and paved the way for the wave for LD-based association studies.

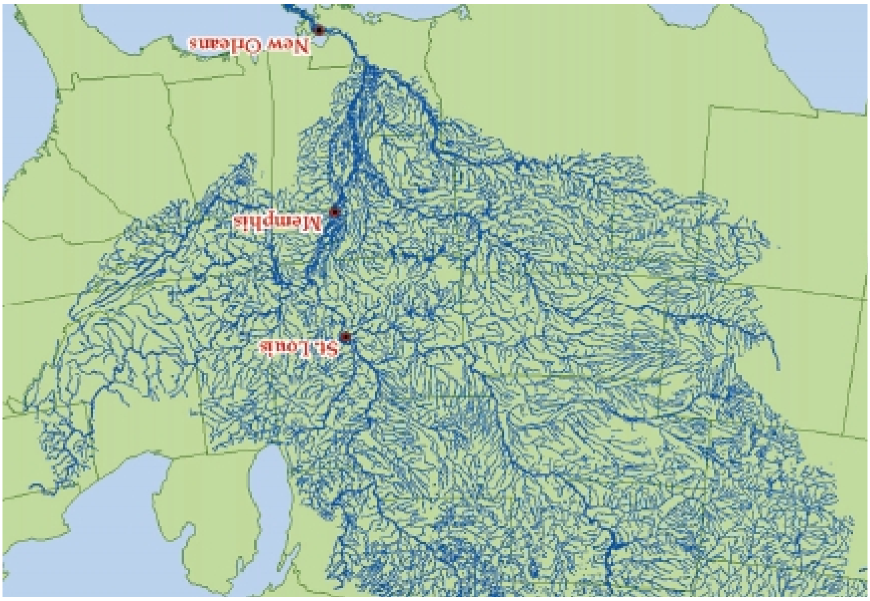


Figure 10. The difficult quest for the genetic causality of complex traits. An inverse map of the Mississippi river serves as a fluvial equivalent of complex disease: cases are the end-point of a complex network of causal factors. The contributory streams form a buffered structure that can change from flood to flood. This complexity makes difficult the assignment of marginal effects to each stream and thus the ascertainment of the major contributing rivers (Weiss 2008).

I have nothing to offer but blood, toil, tears and sweat.
Winston Churchill

1.3. LD-based association mapping

The establishment of the CD/CV hypothesis fuelled the characterization of human common variation and the patterns of LD through the HapMap project. These data served to establish association mapping as the choice method to unravel the genetic bases of complex disease. The HapMap project, an introduction to methodology and pitfalls of candidate gene association studies and the current wave of genome-wide studies are described below.

1.3.1. The HapMap project and the patterns of LD

SNPs revealed as the marker of choice in population-based association studies due to their abundance, but a proper understanding of the strength of LD in human genomes was necessary for the proper design of studies (Wright and Hastie 2007). A seminal simulation study estimated that LD would not extend beyond 3 kb (Kruglyak 1999). Conversely, posterior studies based on real data established the existence of discrete haplotype blocks that extend for several tens of kilobases (Daly et al. 2001; Jeffreys et al. 2001; Patil et al. 2001). The possibility of characterizing haplotype blocks by the genotyping of a low number of SNPs that may be shared across populations (Gabriel et al. 2002) prompted the launching of the HapMap project in October 2002 (Figure 11).

The primary object behind the HapMap Project was to develop a “haplotype map” of the human genome to study common patterns of human DNA sequence variation (The International HapMap Project 2003). The unravelling of the haplotype structure across populations was aimed to provide the basis for SNP selection in LD-based association studies (de Bakker et al. 2005; The International HapMap Consortium 2005). Four populations were chosen for high density genotyping: 30 parent-offspring trios of Yoruba ancestry (Ibadan, Nigeria); 30 trios of northern and western European ancestry from Utah (USA); 45 unrelated individuals from Tokyo (Japan); and 45 unrelated individuals from Beijing (China). After the second phase, genotype data for >3 million SNPs were characterized in 270 individuals (Frazer et al. 2007; The International HapMap Consortium 2005). The latest release “HapMap III” provided 1.6 million SNPs genotyped in >1,100 individuals from 11 populations.

HapMap data served to develop a catalogue of human common SNP variation and the patterns of correlation among variants. The latter is very relevant to ascertain the tagSNPs that best cover genetic variation within LD blocks (de Bakker et al. 2005).

Of note, the design of the SNP discovery process had important consequences on the allele frequency spectrum of SNPs genotyped by HapMap. Markers were identified by direct sequencing of a small panel of individuals and subsequently genotyped in the larger panel of individuals. This procedure introduced an ascertainment bias in SNP discovery in which rare variants are easily messed and common variants in Europeans are prioritized.

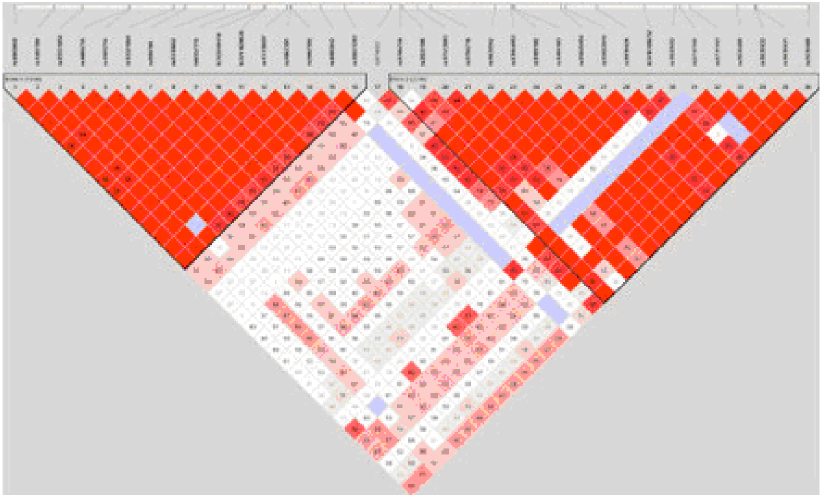


Figure 11. Haploview plot showing the LD relationships of SNPs. The strength of LD between pairs of SNPs is denoted by colouring, from white (weak) to red (maximum). Two LD blocks are distinguishable (Canzian et al. 2009).

1.3.2. Candidate gene association studies

The widespread availability of SNP markers after HapMap expedited the publication of candidate gene association studies using case-control designs (Hirschhorn et al. 2002). Genetic association aims to establish statistically supported correlations between genetic markers and the phenotype of interest (Wright and Hastie 2007). In contrast to linkage, association mapping looks for the transmission of disease status with alleles instead of with loci. Genetic associations are detected when specific alleles are more frequent in groups of affected than of non-affected individuals (Ziegler et al. 2010). Association mapping does not require pedigree sampling as it can take profit from unrelated individuals gathered from the general population.

Associations between alleles and disease arise both when the tested marker is either causal (*direct association*) or when it is in strong LD with the marker that in turn causes disease (*indirect association*). In consequence, association studies present two requirements regarding genetic variation. First, prior biological knowledge is necessary to select loci that "make biological sense" (Thomas 2004) and contain clues of participation in disease aetiology (e.g. signals in previous linkage studies). Second, dense spacing of markers is necessary to cover extensively the ascertained candidate loci.

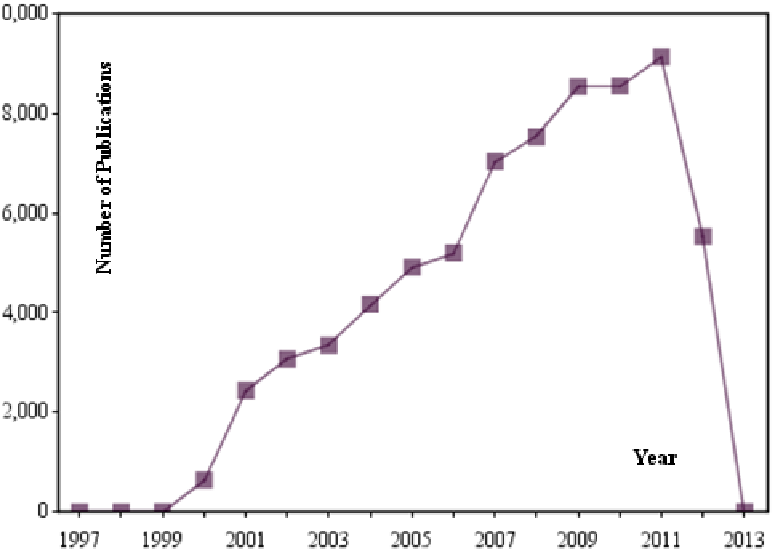


Figure 12. Explosive growth of published genetic association studies. The availability of SNP markers expedited the publication of candidate gene association studies (HuGE Navigator database (Yu et al. 2008)).

In the case-control design, frequencies of variants at the ascertained SNPs are compared in populations of cases and controls. A fundamental assumption of this method is that the ascertained individuals must provide unbiased estimates of the true allele frequency of the populations of interest (Clarke et al. 2011). In case-control studies, the penetrance and relative risks are approximated by calculating the strength of association (Odds Ratio) under different models. Conditions that are better approximated as quantitative traits (e.g. high blood pressure for hypertension) are usually assessed through linear regression. Alternative analysis of haplotypic and epistatic effects permit to test more complex patterns of association (Clarke et al. 2011; de Bakker et al. 2005).

1.3.3. Lack of replication of the candidate gene approach

The identification of several associations through candidate gene approaches created great hopes in the community and considerable hype in the media (e.g. CTLA4 and type 1 diabetes, NOS2 and Crohn's disease or ADAM33 and asthma). However, the non-replication of significant findings has constituted a major challenge of candidate gene studies. A comprehensive review of >600 studies found that only 6 out of 166 putative associations had been consistently replicated (Hirschhorn et al. 2002). Other meta-analysis described large heterogeneity in the reported effect sizes and found clues of publication bias (Ioannidis et al. 2001; Lohmueller et al. 2003).

Typically, as shown in Figure 13, an initial positive report is followed by replication studies reporting little or absence of association (Ioannidis et al. 2001). A wealth of statistical, biological and sociological reasons has been put forward to explain replication failures (Hirschhorn and Altshuler 2002).

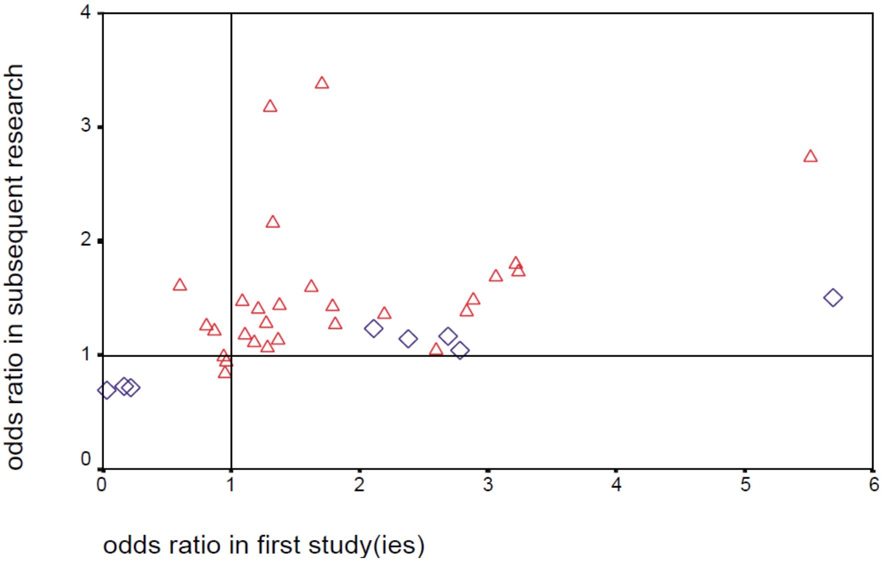


Figure 13. First studies tend to report unsupported strong effect sizes. Odds Ratio found in first and following studies for 36 associations (Ioannidis et al. 2001). Blue diamonds denote significant discrepancies.

The initial report "discovering" the association between disease and particular alleles can constitute a false-positive finding. Indeed, several quality-control aspects can result in false reports when poorly addressed (Wright and Hastie 2007). Population stratification, an artefact that arises when cases and controls do not match as regards of genetic ancestry, is one of such challenges. The confounding may appear when the underlying populations that are represented in the studied cohorts differ in allele frequencies. Even subtle correlations of genetic ancestry and disease status can inflate association test statistics. Several methods have been developed to avoid type I errors due to stratification (Devlin and

Roeder 1999; Pritchard et al. 2000a; Pritchard et al. 2000b). Statistical errors can also lead to false positives (Risch and Merikangas 1996). The significance level to reject the null hypothesis of no association (usually α=0.05) has to account properly for the number of independent tests (e.g. number of SNPs) to avoid "multiple testing". Finally, circumstances that enhance the probability of being reported depending on the outcome of the study result in publication bias (Pan et al. 2005). Ironically, an excessive zeal to harden significance thresholds together with the presence of widespread publication bias does not cancel but inflates false positive rates (Ellis 2010; Ioannidis 2005).

True findings can also fail to replicate due to lack of statistical power in replication studies. The typical sample sizes used in candidate-gene association studies (in the hundreds of samples) may not render enough power to statistically distinguish the low effect sizes of variants associated to complex disease (Chanock et al. 2007). Winner's curse, the inflated effect size typical of discovery studies, can also affect the ability of replication trials by the over-estimation of statistical power.

Finally, several factors associated to true biological heterogeneity are involved in the abundant lack of replicability problem. Heterogeneity at the phenotypic (e.g. clinical severity) and genetic levels (i.e. variation in LD patterns), as well as differences in gene-by-environment interactions (i.e. FTO effects in diabetes (Timpson et al. 2009)) explain replication failures.

The plethora of questionable associations could have ruined the prospects of ever discovering the genetic bases of disease. This perspective prompted the constitution of the NCI-NHGRI Working group on replicability (Chanock et al. 2007). Among other criteria, the consensus list included that replication studies should (i) have enough sample size, (ii) test the same phenotype, (iii) use independent samples from the same or similar population, and (iv) find a similar effect using the same genetic model. If associations are true, the combination of results across studies by means of meta-analysis should lead to a better p-value (Chanock et al. 2007).

Interestingly, several meta-analyses of candidate gene associations checked for the role of ethnicity in disease. In general, most risk variants show consistent patterns in their effects in Europeans and East Asians, albeit a few number of associations present significant differences in effect size (Ioannidis et al. 2004). Interestingly, disease-associated variants harbour levels of population differentiation that do not depart from the genomic average, but show substantial variation in allele frequencies that might help accounting for the differences in disease prevalence (Adeyemo and Rotimi 2009; Lohmueller et al. 2006; Myles et al. 2008).

1.3.4. Genome-wide association studies

The development of genotyping techniques improved the availability of polymorphism to a density of up to several SNP per kilobase (Hinds et al. 2005; Sachidanandam et al. 2001; The International HapMap Consortium 2005). Additionally, the unravelling of LD patterns allowed, in principle, to capture ~80% of the predicted >10 million common SNPs with a scattered selection of 0.5 to 1 million SNPs (Visscher et al. 2012). Eventually, HapMap permitted the development of commercial genotyping arrays that did capture >95% and 80% of common variation in Eurasian and African populations, respectively (The International HapMap 3 Consortium 2009). Thus, technological development of commercial chips for high-throughput genotyping made it feasible to look for common risk variants by means of genome-wide association studies (GWAS).

The first GWAS were published in 2005 and 2006 (Dewan et al. 2006; Klein et al. 2005). Even if using few markers and samples (~100K SNPs and <200 individuals), both studies managed to find common variants associated to age-related macular degeneration due to their large effect sizes (OR>2). In 2007, the Wellcome Trust Case Control Consortium published a GWAS for 7 different diseases (The Wellcome Trust Case Control Consortium 2007). The WTCCC paper became a landmark due to the large number of samples used, the use of shared control cohorts across disease, the confirmation of the low stratification present in Europeans, the setting of significance thresholds and the replication of previous signals. Dropping costs of commercial arrays helped in the increase of published GWAS (Clarke et al. 2011; Hindorff et al.).

GWAS present two key differences with respect of candidate gene studies. First, there is an inherent issue related to numbers.

Previously, a few tens of markers and, at most, a few hundreds of individuals were analyzed. In contrast, GWAS studies routinely test millions of markers (after imputation) in thousands of individuals gathered from diverse cohorts. Second, GWAS are said to be "hypothesis-free": they certainly look for common risk effects, but they do so looking at SNPs scattered across the genome and without any a priori list of candidate loci (Ziegler et al. 2010).

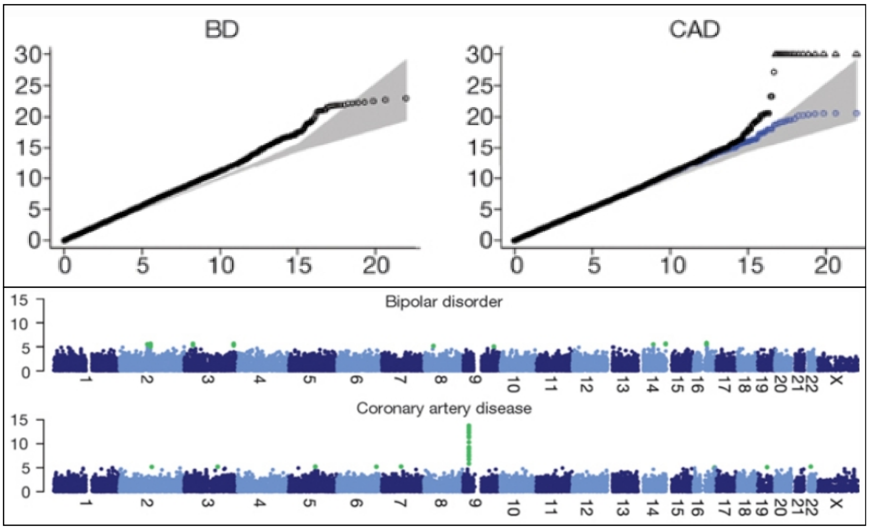


Figure 14. Display of WTCCC genome-wide results for two diseases. Top: Q-Q plots showing the genome-wide distribution of association statistic (y-axis) and that expected under the null hypothesis (x-axis and grey band). Bottom: Manhattan plots show the –log of p-values for all SNPs sorted by position. CAD shows a clear peak in 9p21 that accounts for the deviation in the Q-Q plot. The differences can be due to quality control problems (e.g. phenotype heterogeneity in bipolar disorder) or, alternatively, point at true differences in genetic architecture.

The standard approach in GWAS consists in testing the association of every single SNP to the phenotype of interest. The large number of markers difficult the examination of results and the distribution of association statistics must be inspected visually through

Manhattan and Q-Q plots (Figure 14). These tools permit to check the presence of genome-wide enrichments of low p-values and of clusters of p-values in regions that merit further interest (towers in Manhattan plots). The adjustment for multiple testing constitutes an important decision in GWAS. The two most used cut-offs are $5x10^{-7}$ (WTCCC) and $5x10^{-8}$ (strict Bonferroni), but there are several available methods available to select proper significance thresholds (McCarthy et al. 2008; Pe'er et al. 2008).

There is large variation in GWAS design as regards to the methodological approach selected to accumulate enough evidence of association (Skol et al. 2006). One of the most common practices consists in the design of multi-stage GWAS in which the results from the initial genome-wide stage are followed up through the genotyping of a few selected SNPs in a much larger sample of individuals (Visscher et al. 2010; Ziegler et al. 2010). This approach saves costs by the use of lower number of individuals in the initial stage and the enrichment of SNP coverage only in interesting regions. This procedure also permits to estimate the effect size of associated SNPs without the inflation due to the winner's curse phenomenon (Ziegler et al. 2010).

At the time of writing, more than 1,380 published studies are recorded in the *catalog* of GWAS maintained by the *NIH Office of Population Genomics* (Hindorff et al.). Similar to the situation faced in candidate gene studies, most of GWAS use exclusively individuals of European genetic ancestry. In 2010, a survey of GWAS publication patterns found that >80% of studies did not use any cohort of non-European ancestry (Rosenberg et al. 2010). Yet, the bias towards Europeans has relaxed in the very last two years (up to 84 GWAS on East Asians by May 2011 (Fu et al. 2011)).

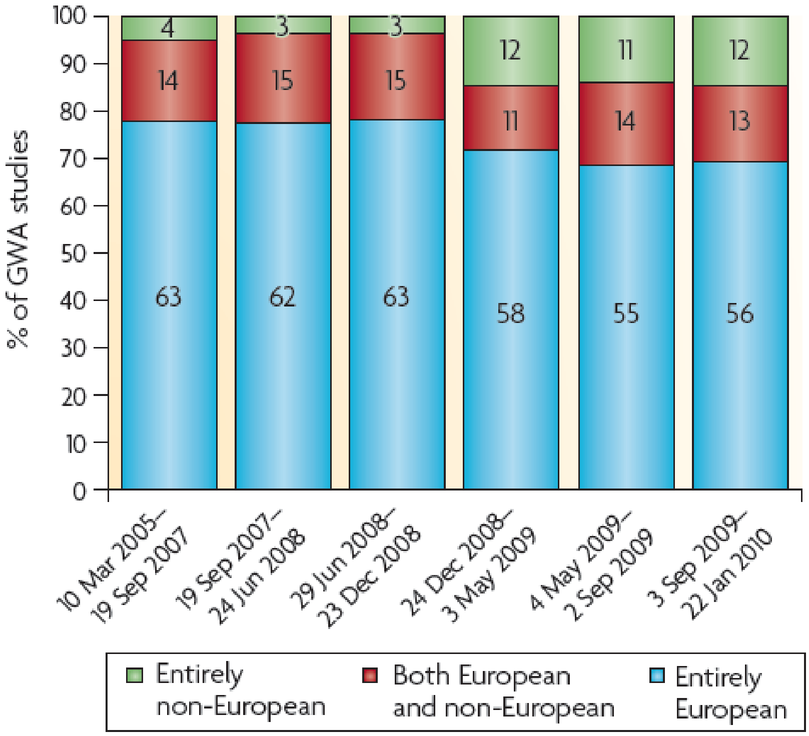


Figure 15. Evolution of populations studied in published GWAS. Most GWAS use exclusively individuals of European ancestry, but there is a slight trend over time favouring the study of non-European populations. Numbers within columns indicate the absolute number of GWAS publications per period (Rosenberg et al. 2010).

Non-European GWAS present challenges related to imputation ability, genomic coverage (tagSNP portability) and statistical power (due to SNPs ascertainment in arrays). The ability to detect disease variants can vary if they have different effect size or present disparate allele frequencies across populations (Adeyemo and Rotimi 2009). However, several reasons fuel the case for GWAS generalization across populations to better achieve the objectives of complex disease mapping (Rosenberg et al. 2010). To mention only one, the use of diverse populations permits to take profit of the variation in LD across populations and thus help in the fine mapping to narrow down the location of causal variants (Visscher et al. 2012; Zaitlen et al. 2010). In any case, preliminary comparisons of GWAS replicability observed high rates of concordance across populations (Shriner et al. 2009; Waters et al. 2010).

Crisis? What crisis?

The Sun, 10 January 1979, Winter of Discontent

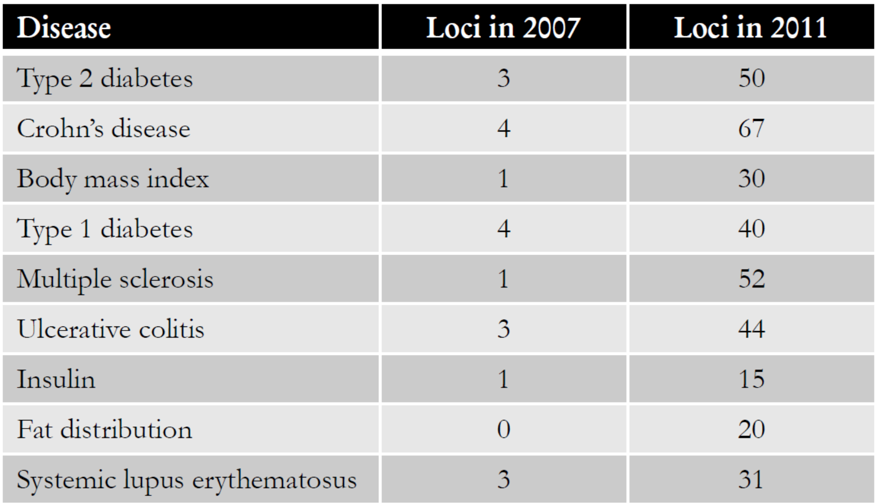
1.4. Beyond GWAS

After "five years of GWAS" (Visscher et al. 2012), geneticists work on several hundreds of robust associations between variants and disease. This wealth of data informs about functional aspects of disease and may have immediate clinical impacts. However, most of the heritability remains unfound and the degree in which causality is shared across populations should be illuminated for disease testing. A discussion of these aspects is presented.

1.4.1. Knowledge gained through GWAS

There are over 2,000 loci robustly associated to disease (Visscher et al. 2012). Interesting facts can be extracted from the analysis of the regions and variants unravelled by GWAS. The number of loci identified for each disease has increased exponentially if compared to associations discovered and replicated through the candidate gene approach (Figure 16). This observation emphasizes the limitations of an approach based on biological candidates compared to the "hypothesis-free" GWAS. There is also variation in the number of loci discovered per disease, but it appears to be correlated with study sample size (Visscher et al. 2012).

Pathway analysis of the discovered loci shows unsuspected insights into disease mechanisms (Visscher et al. 2012). New understanding has been obtained for a wide range of problems, from specific diseases (e.g. the role of IL23R in ankylosing spondylitis), to shared aetiology across disease (e.g. loci associated to disparate autoimmune diseases) and to new mechanistic connections across disease (e.g. cancer and diabetes). Interestingly, the enrichment of "druggable" hits provides targets for the translational application of GWAS (Collins 2011; Lander 2011).



Disease	Loci in 2007	Loci in 2011
Type 2 diabetes	3	50
Crohn's disease	4	67
Body mass index	1	30
Type 1 diabetes	4	40
Multiple sclerosis	1	52
Ulcerative colitis	3	44
Insulin	1	15
Fat distribution	0	20
Systemic lupus erythematosus	3	31

Figure 16. New disease loci discovered by GWAS for several autoimmune and metabolic conditions (Visscher et al. 2012).

The analysis of SNPs associated to disease also shows three interesting patterns related to functionality and the explained heritability. First, the vast majority of risk SNPs locate outside transcriptional units (~43% are intergenic (Hindorff et al. 2009)). In some cases, association signals map into gene-poor regions (e.g. 8q24 associations for several cancers). These observations confirm the role of non-coding variants in complex disease, but the exact numbers of these proportions are difficult to calculate (Hindorff et al. 2009). Second, disease-associated SNPs are enriched for eQTLs and clues of pleiotropy have been described (Nicolae et al. 2010; Sivakumaran et al. 2011). Finally, and in spite of contradictory evidence (Lachance 2010), the distribution of allele frequencies of GWAS variants contain an excess of common variants (MAF>20%) when compared to genotyped SNPs. Indeed, these variants explain substantially large proportions of disease risk when compared to SNPs at intermediate frequencies (Park et al. 2011).

After the wave of published GWAS, we have gained strong knowledge on the functional aspects of risk variants and valuable insights into the genetic architecture of disease. However, the disappointingly low amount of explained heritability constitutes the most debated issue in the GWAS era.

1.4.2. "Missing heritability" and alternative models

The integration of statistical power in the distribution of effect sizes from GWAS has allowed estimating the number of discoveries that may be expected in future studies. That is, the number of risk variants that ought to be discovered if the study sample size is known in advance (Lango Allen et al. 2010; Park et al. 2010). This exercise lends support to the hypothesis that a large number of risk variants wait to be discovered (Park et al. 2011; Park et al. 2010).

However, the bulk of genetic risk variance remains unexplained after the "low-hanging fruits" found by GWAS have already been reported. For height, the ~50 trumpeted variants described after having studied more than 30,000 people in 2008 account for less than 10% of heritability (Visscher 2008). In spite of the knowledge gained about new disease loci, GWAS results add very little to the prediction power necessary for personalized medicine. This phenomenon was coined as the "case of the missing heritability" (Maher 2008).

Several hypotheses have been put forward to explain the missing heritability problem (Eichler et al. 2010; Goldstein 2009; Maher 2008; Manolio et al. 2009). The stringent correction for multiple testing necessary to avoid false positives could swamp the signal of alleles with very small effect sizes. Alternatively, most heritability could be explained by rare variants of large effect that are not captured in commercial genotyping arrays (Figure 17).

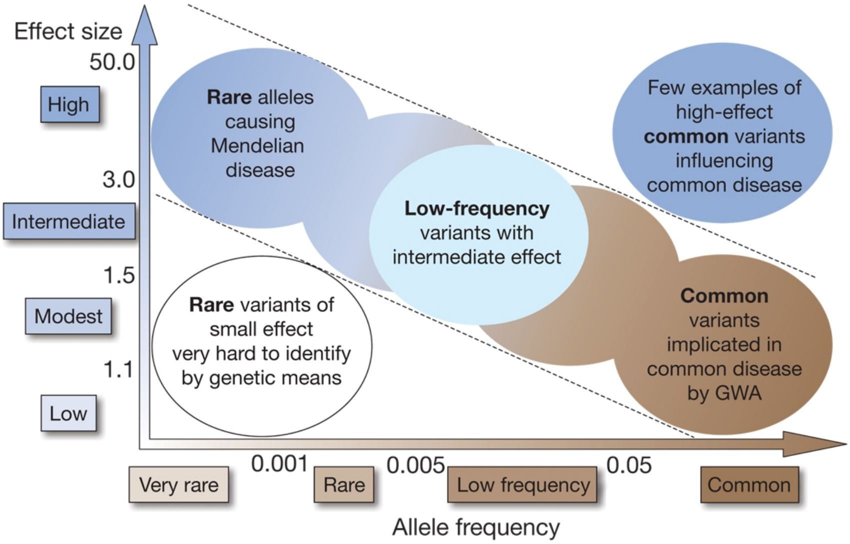

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

Figure 17. Feasibility of identifying genetic variants by risk allele frequency and effect size (odds ratio).
Interest lies in associations within the dotted lines (Manolio et al. 2009)

Other explanations champion the unexplored role of epistatic interactions (GWAS usually assume additive effects), the hidden effects of gene-by-environment interactions, the responsibility of Copy Number Variants and the presence of parent-of-origin effects (that could explain up to 14% of type 2 diabetes heritability (Eichler et al. 2010)). Inherent phenotypic heterogeneity among patients such as that faced in psychiatric disorders could recipe for GWAS failure (Burmeister et al. 2008; Terwilliger and Weiss 2003). Finally, the possibility that the heritability explained by known variants is much larger after accounting for epistatic interactions has been recently proposed (Zuk et al. 2012). The "infinitesimal model" and the "rare allele model" are the two most cited alternatives to the refuted CD/CV model of heritability (Gibson 2012).

The infinitesimal model is the classical quantitative genetics idea posing that a myriad (hundreds or thousands) of genome-wide scattered common variants account for disease risk in populations, each explaining very small bits of genetic variance (Gibson 2012). For instance, the fitting of an exponential distribution using the effect size of the ~50 height alleles known in 2008 predicts that 93,000 SNPs would be required to explain 80% of the heritability (Goldstein 2009). A model of very low-risk effects could account for the low sibling risk explained by GWAS variants (Gibson 2012; Hemminki et al. 2008). Thus, due to the stringent correction for multiple testing, most of the low-risk susceptibility variants would remain buried along with non-associated alleles. The early departure from chance and the excess of low p-values observed in

Q-Q plots for most diseases supports the infinitesimal model (Park et al. 2011). The GWAS published by the International Schizophrenia Consortium also supports this model (Purcell et al. 2009) by showing that the relaxation of p-value thresholds to include SNPs increases the predictive ability of individual risk scores to distinguish between cases and controls (Figure 18).

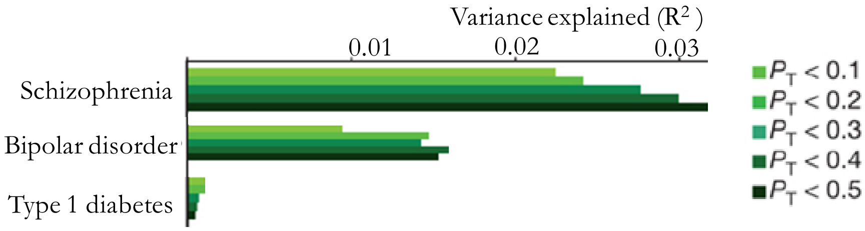


Figure 18. Variance explained in regression of case-control status. Variance explained by models to distinguish case status in target samples. The scores are build using SNPs associated to schizophrenia with five liberal thresholds (p-vakue<0.1 to p-value<0.5). SNPs discovered for schizophrenia show significant predictive power for case-control status in GWAS of bipolar disorder (Purcell et al. 2009).

Further evidence pointing at the presence of thousands of very low effect variants comes from looking at the joint effects of all genotyped SNPs instead of at individual variants. Liner models built using all genotyped SNPs are able to explain 45% of height heritability (Yang et al. 2010). Moreover, and as shown for several disease-associated traits, the percentage of variance explained by each chromosome correlates perfectly with chromosome length and gene content (Smith et al. 2011; Yang et al. 2011). Under the infinitesimal model, heritability would not be “missing” but “hidden” by the strict significance cut-offs used in GWAS. Interestingly, the discrepancy between height heritability (80%)

and that captured by SNPs from commercial arrays (45%) can be accounted by incomplete LD between genotyped common SNPs and causal variants with lower allele frequency.

The major alternative to the infinitesimal model, the rare allele model, posits that most of the variance for complex disease is due to rare variants (MAF<1%) of large effect. Evolutionary theory predicts that variants that reduce the fitness of individuals are maintained at low-allele frequencies by purifying selection. This model would fit with recent human demographic history. First, the explosive growth in census size has resulted in an excess of rare variants in human genomes (Gravel et al. 2011; Keinan and Clark 2012). Second, the analysis of allele frequency distribution for different classes of variation confirms that purifying selection maintains deleterious alleles at low frequencies (Kryukov et al. 2007) that differ across human populations (Lohmueller et al. 2008).

Simulations of mutations with pleiotropic effects on both complex traits and fitness also predict that most of genetic variance for complex traits is contributed by derived non-synonymous alleles of large effect (Eyre-Walker 2010). Additionally, the prediction that variable selective pressures have shaped the participation of rare variants could account for the variation in “hidden heritability” estimates across disease (Yang et al. 2011). The presence of susceptibility rare variants has been a constant observation in re-sequencing studies of disease loci (some are listed in (Schork et al. 2009)). The burden of rare copy number variants has also been shown to accumulate in cases of neurological disorders (Levy et al. 2011; Pinto et al. 2010). Moreover, GWAS loci overlap significantly with loci associated to Mendelian disease (Siontis et al. 2010).

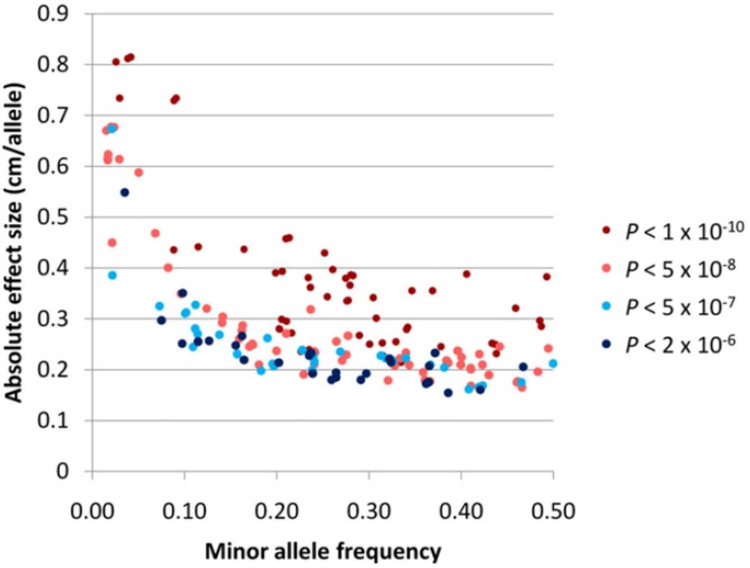


Figure 19. The complex frequency spectrum of SNPs for complex traits. The effect size of height associated SNPs identified in a gene-centric GWAS as a function of Minor allele frequency. The presence of large-effect rare variants and low-effect non-significant common alleles is shown. From (Lanktree et al. 2011)

The further possibility that rare, instead of common, variants explain the results from GWAS has been put forward (Dickson et al. 2010). Simulation data show that the accumulation of rare variants in certain haplotypes can give signal through common SNPs in LD with such variants. Under the presence of "synthetic associations", the effect size of susceptibility variants might be much larger than the reported associations. The "synthetic" model has been confirmed for several associations (e.g. NOD2 for Crohn's disease) but empirical and population genetics data refute any widespread participation of rare variants in extant GWAS results (Anderson et al. 2011; Orozco et al. 2010; Wray et al. 2011).

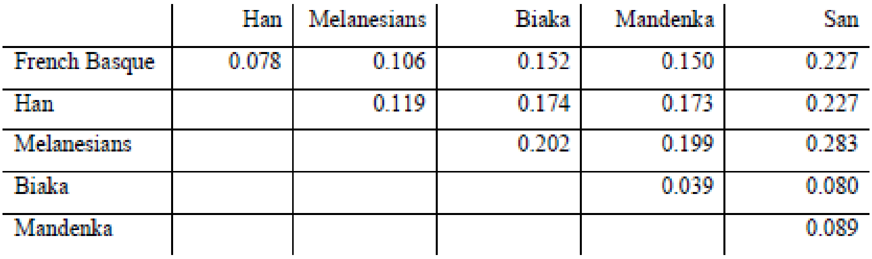
The prospect that rare variants explain the "missing heritability" was one factor encouraging the 1,000 Genomes Project (sequencing >2,500 individuals from 27 populations). The goal lies in the description of all variation with at least 1% of allele frequency and its sharing across human populations (Gravel et al. 2011).

1.4.3. Disease architecture across populations

The availability of genetic data has shed light on human origins. The out-of-Africa (OOA) model predicts diversity being highest in African populations and structured at increasing distances from Africa (Goldstein and Chikhi 2002). Two early results validated this model; namely (i) the common ancestor of human mtDNA dates at 200 kilo years ago (kya) and locates in Africa and (ii) non-Africans present subsets of African diversity (Cann et al. 1987). Microsatellites and autosomal markers have validated the single origin hypothesis (Li et al. 2008; Rosenberg et al. 2002). Nonetheless, the publication of two extinct hominin genomes confirmed that archaic genes segregate in modern humans (Green et al. 2010; Plagnol and Wall 2006; Reich et al. 2010).

Humans appear to have settled Eurasia and Oceania around 60 kya (Goldstein and Chikhi 2002). The Eurasian divergence is estimated around 17 kya, and recent gene flow occurred between Africans and Europeans (Keinan et al. 2007; Moorjani et al. 2011). America and Pacific archipelagos were colonized around 20 and 4 kya. Recently, most populations have undergone explosive expansions in Neolithic times (Coventry et al. 2010; Keinan and Clark 2012). As a result, humans live in an extraordinary diverse range of habitats and present wide phenotypic and cultural variation.

The complex demographic history of human populations has shaped the presence and frequencies of genetic variants. Most common variants predate human expansion across major landmasses and are shared across populations (The International HapMap 3 Consortium 2009). Nonetheless, the effects of genetic drift within each population has created major differences in allele frequencies at common variants (Bamshad et al. 2004). On the other hand, the explosive expansion in census size has resulted in the vast majority of human polymorphism being rare (MAF<0.05) and confined to continental populations (Gravel et al. 2011).



	Han	Melanesians	Biaka	Mandenka	San
French Basque	0.078	0.106	0.152	0.150	0.227
Han		0.119	0.174	0.173	0.227
Melanesians			0.202	0.199	0.283
Biaka				0.039	0.080
Mandenka					0.089

Figure 20. Average genetic distances among several human populations. Population structure can be studied through the fixation indexes (Wright 1922, 1969). *F-statistics* describe the departure from expected heterozygosity in panmixia due to inbreeding. The average genetic differentiation shown in the figure was calculated by means of the F_{ST} from the polymorphic sites identified in a survey of 20 autosomal regions (Wall et al. 2008). A null F_{ST} indicates there are no differences in allele frequencies among populations, and the maximum value of one is reached when differential alleles have fixed in each population.

The architecture and adaptive significance of several traits such as skin pigmentation and body size are being unravelled. The extent to which demographic and selective events have shaped the frequency and prevalence of causal variants and disease across populations remains unknown.

1.4.4. The prospects for personalized genetic medicine

The hype about the recent publication of the ENCODE results (Bernstein et al. 2012) reveals the need to deepen in our knowledge about genome functional organization. We are in a similar position regarding the knowledge about the functional basis of disease associations. For instance, the biological underpinning of 9p21 alleles to cardiovascular disease and diabetes remains as one of the few functional validations of GWAS (Harismendy et al. 2011).

The immediate application of GWAS results to the clinical setting lies in the development of genetic profiles based on risk markers that has already begun through the availability of commercial consumer testing (Jakobsdottir et al. 2009; Kraft et al. 2009; Lee et al. 2008). Potentially, genetic profiles based on risk markers may distinguish between high-risk and low-risk groups of individuals. However, the most commonly reported features of GWAS associations (i.e. p-values and OR) do not translate to clinical utility (Jakobsdottir et al. 2009). Instead, measures of profile accuracy are necessary to develop tests with clinical validity (Kraft et al. 2009). The two most important parameters of clinical utility are sensitivity and specificity (Ziegler et al. 2010). The former measures the percentage of detected individuals that will truly develop disease. This proportion sizes the total number of individuals that will benefit from early intervention after diagnosis. Specificity measures the proportion of individuals that are correctly classified as not being at risk to develop the disease. This quantity is of enormous importance as it evaluates the amount of money and personal suffering that is saved through correct ascertainment of those individuals that do not need any clinical intervention (Kraft et al. 2009).

The performance of genetic tests can be evaluated through the receiver operating characteristic (ROC). This curve serves to visually inspect sensitivity vs. specificity. Overall, the classical measure to quantify ROC performance is the area under the curve (AUC) statistic (Ziegler et al. 2010). However, additive models constructed from GWAS variants present AUC values (~75%) that are not enough to distinguish individuals that will develop disease (Jakobsdottir et al. 2009). With few variants per disease, the achieved high specificity implies low sensitivity and thus few individuals truly benefit from the early diagnosis (Wray and Visscher 2010; Wray et al. 2010). The calculation of scores from training sets using liberal thresholds somewhat improves the picture (Evans et al. 2009). However, discriminative accuracy presents great variation across disease. Indeed, the correlation of genetic predictors and phenotypes has been shown to have an upper limit that depends on the heritability of the trait (Wray et al. 2010). For instance, the discovery of 50% of the genetic variance for schizophrenia would translate to an AUC of 90% due to its large heritability. Thus, models estimated from genome-wide markers can improve the prediction of phenotypes based on phenotypes of close relatives (Lee et al. 2008). A promising aspect of GWAS results lies in the fact that, even if not causal, associated SNPs can be useful for clinical prediction if correlated with the causative alleles. Nevertheless, caution is needed to transport clinical predictors to other populations because causal markers might not be shared or might present different LD with associated SNPs (Visscher et al. 2011).

Current protocols include estimating genetic risk by means of studying family history, but a large proportion of complex cases subjects do not have close diseased relatives (Wray and Visscher 2010). Indeed, 50% of genetic variance occurs within families and thus genetic risk of disease changes across relatives with the same family history. Current estimations show that genome-wide SNP data from unrelated people have enough precision to predict phenotypes when risk estimates are calculated with >100K individuals (Goddard 2009). However, the current performance of predictions for individual genomes remains unclear (Burga and Lehner 2012; Jelier et al. 2011) and only the combination of larger samples sizes and improved genomic coverage will enhance the hopes for personalized genetic testing.

2 Objectives

The central objective of this work is to push forward our knowledge about the worldwide distribution of genetic variants associated to complex disease. I intend to examine inter-continental patterns of replication to study what they tell us about the underlying genetic architecture of disease.

This work aims to:

1. Describe the patterns of replicability of genetic associations across human populations, for both candidate gene and genome-wide association studies

2. Determine the role of factors such as allele frequency, genetic differentiation, linkage disequilibrium and statistical power in the replication of genetic associations

3. Obtain general inferences about the genetic architecture of complex disease. Specifically:

    a. Gain insight about the role of common/rare variants
    b. Quantify the sharing of risk variants across Eurasians
    c. Determine the role of synthetic associations in GWAS
    d. Model the frequency spectrum of variants discovered in the immediate future through larger association studies

3 Results

3.1.

Recent human evolution has shaped geographical differences in susceptibility to disease

Urko M. Marigorta, Oscar Lao, Ferran Casals, Francesc Calafell, Carlos Morcillo-Suárez, Rui Faria, Elena Bosch, François Serra, Jaume Bertranpetit, Hernán Dopazo, Arcadi Navarro

BMC Genomics 2011, **12**:55

3.2.

High trans-ethnic replicability of GWAS results implies common causal variants

Urko M. Marigorta and Arcadi Navarro

Manuscript submitted to PLoS Genetics

Title: High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants

Running head: GWAS variants are common and shared across Eurasia

Urko M. Marigorta[1] and Arcadi Navarro[1,2,3, §]

[1] Institute of Evolutionary Biology (UPF-CSIC), PRBB, Doctor Aiguader 88, 08003, Barcelona, Catalonia, Spain

[2] National Institute for Bioinformatics, Universitat Pompeu Fabra, Barcelona, Spain

[3] Institució Catalana de Recerca i Estudis Avançats (ICREA). Barcelona, Catalonia, Spain

[§] *corresponding author*

Email addresses:

UMM: urko.martinez@upf.edu
AN: arcadi.navarro@upf.edu

ABSTRACT

Background

Genome-wide association studies (GWAS) have illuminated the biology of disease. However, they tend to explain small fractions of risk, raising doubts about issues such as the relative roles of rare versus common variants in the genetic architecture of complex diseases and how findings in one population translate to the rest of the world. Light on these problems can be shed by studying the degree of sharing of disease-associated variants across populations.

Main Findings

We present a comprehensive survey of GWAS replicability across 28 diseases. Most loci and SNPs discovered in Europeans have been extensively replicated using peoples of European and East Asian ancestry. We found a strong and significant correlation of Odds Ratios across continents, indicating that underlying causal variants are common and shared between European and East Asian populations. As expected if that were the case, SNPs discovered in Europeans that failed to replicate in East Asians map into genomic regions with larger between-population differences in patterns of Linkage Disequilibrium. Finally, we observed that GWAS with larger sample sizes have detected variants with weaker effects rather than with lower frequencies.

*Significanc*e

Our results settle the issue of the spurious origin of GWAS associations and confirm that the vast majority of GWAS results are due to common variants. In addition, the sharing of disease alleles and the high correlation in their effect sizes show that the underlying causal variants are shared between Europeans and East Asians and that they must map close to the

associated marker SNPs. Thus, our results indicate that trans-ethnic fine mapping of causal variants is feasible.

AUTHOR SUMMARY

Describing and identifying the genetic variants that increase risk for complex diseases remains a central focus of human genetics and is fundamental for the emergent field of personalized medicine. Over the last five years, GWAS have revolutionized the field, discovering hundreds of disease loci. However, with only a handful of exceptions, the causal variants that generate the associations unveiled by GWAS have not been identified, and their frequency and degree of sharing across populations remains unknown. Here, we present the largest and most comprehensive comparison of GWAS to date. By examining the results of GWAS for 28 diseases that have been performed with peoples of European and East Asian ancestries, we show that the vast majority of associations are caused by common variants that are shared between continents and map close to the associated markers. These results indicate that the major contributors to disease risk are shared across continents and imply that trans-ethnic fine-mapping of causal variants is feasible.

INTRODUCTION

Genome-wide association studies (GWAS) have detected hundreds of risk alleles [1], generating novel biological knowledge and widening the range of diagnostic and treatment tools [2]. However, the reported effect sizes of these variants are small and their impact in individual risk prediction remains modest, raising doubts about the relevance of GWAS results [1,3-6]. Some of the most hotly debated topics are how to account for the unexplained risk [4]; what may be the role of rare variants as producers of artefactual GWAS results [7-10]; and up to what extent GWAS results are portable between populations [11-15]. Answering to these questions is pressing for two reasons. First, the description of the genetic architecture of disease is crucial for personalized medicine and, in particular, finding predictors of individual disease risk that could be applicable to different ancestries would be a major step forward [1]. Second, if sharing across populations of risk alleles were common, it would confirm trans-ethnic mapping as a powered tool that would take profit of population heterogeneity in LD and allele frequencies to identify the causal variants underlying disease susceptibility [1,15].

The available reports of the distribution of the allele frequencies of GWAS risk variants point at an excess of common variants [16] that, at least for some particular diseases [17], present consistent effects across populations. If repeated, these observations constitute empirical evidence against rare alleles as a source of synthetic associations and would point at common variants that are in LD with the associated tagSNPs in all populations. However, such studies have not been generalized across different diseases and, currently, most evidence comes from either re-sequencing efforts aimed to capture rare variants [18] or multi-ethnic replication efforts for a few risk variants [13,15,17,19]. Most meta-analysis of GWAS data, that could shed light on these issues, either have ignored population heterogeneity [2,20] or have focused on a limited set of traits [21] and GWAS [22].

By compiling data from 275 GWAS for 28 different diseases, we build the largest-to-date database of discovery and replication patterns of SNPs associated to complex diseases. We evaluate the extent to which risk variants discovered in Europeans replicate in posterior studies performed on individuals of European or East Asian ancestry and compare the risk effect sizes found in both populations. We also examine the extent up to which statistical power and differences in Linkage Disequilibrium among populations explain replication failures. Our results serve to establish the patterns of replicability of GWAS across diseases and populations to evaluate how transportable these results are and to study the relative roles of rare and common variants in explaining current GWAS results.

RESULTS AND DISCUSSION

We started by downloading all the associations in the GWAS Catalog [23] (last accessed in February 2012, see **Materials and Methods**), which represents a total of 7,145 associations with $P<10^{-5}$ reported in 1,171 papers. We focused on diseases (avoiding quantitative traits, such as height) with at least two GWAS performed with different ethnic groups. This renders a final dataset of 275 GWAS papers reporting 413 associations to 28 diseases and including peoples from European and East Asian ancestry (204 and 71 GWAS, **Tables S1 and S2**). Out of these, we ascertained 182 SNPs initially reported as genome-wide significant ($P<5\text{x}10^{-7}$) in European GWAS and for which one or more replication attempts had been performed in subsequent European and/or East Asian GWAS (177 and 225 attempts, respectively, **Tables S3 and S4**). We studied patterns of replication across studies, using the criterion that a replication was successful if the same risk allele achieved $P<0.05$. To obtain that information we examined every individual paper, since the GWAS Catalog records only $P<10^{-5}$.

Replicability rates and sharing across Europeans and East Asians

Replicability rates are high within Europeans, with 150 successful out of 177 attempts (84.6%), when only 8.8 positive replications (~5%) would be expected under the null hypothesis of no association (binomial test, $P<10^{-16}$). This excess was robust to the significance threshold (i.e. 113 observed *vs*. 0.18 expected for $P<0.001$ and 51 *vs*. $1.8\text{x}10^{-5}$ for $P<10^{-7}$, **Table S4**). This is expected, since most GWAS already contain an internal replication phase [1,24]. Interestingly, all diseases presented the same high replicability patterns, with no traces of heterogeneity in replicability (**Table S5**). These results were consistent with previous partial reports of replication for individual diseases [17,19] and confirmed that the subset of 182 genome-wide significant SNPs map in loci truly associated to disease in peoples with European ancestry.

Next, we considered the replication attempts in East Asian populations. Out of 225 replication attempts, 103 were successful at $P<0.05$ (45.8%). This replicability departs from the null expectation (103 *vs*. 11.3 expected, $P<10^{-16}$) and is robust to replication thresholds (i.e. 49 observed *vs*. 0.23 expected for p-value <0.001 and 19 *vs*. $2.3\text{x}10^{-5}$ for p-value $<10^{-7}$). Nevertheless, that figure is smaller than for Europeans, which can be expected since East Asian GWAS tend to have smaller sample sizes [15]. We tested this hypothesis by focusing in the 81 attempts with ≥80% power to replicate the Odds Ratio (OR) found in Europeans (**Table S4 and Materials and Methods**). For that subset, replicability increases dramatically to 76.5% (62 of 81 attempts at $P<0.05$). Again, we found no heterogeneity across diseases (**Table S6**).

Trans-ethnic replication indicates that risk loci are shared between Europeans and East Asians. As to the sharing and frequency of risk variants themselves, it can be explored even if they remain undiscovered. First, since most rare variants appeared after the split of Europeans and East Asians [4,12,25-27], they would have accumulated randomly in the genealogy of each allele of the marker SNPs used in GWAS. Therefore, risk alleles would not be necessary shared even if discovered through the same tagSNP. Strikingly, when considering the direction of effects instead of their significance, we observed the same risk allele than in Europeans in 85.9% of East Asian replication attempts (100% and 73.6% of attempts replicated and not replicated at $P<0.05$, respectively). This proportion departs from the 50% expectation in a scenario of independent rare variants ($P<10^{-16}$, binomial test). Secondly, the idea that the same causal variants underlie association in the two continents is reinforced by a strong correlation between ORs in Europeans and East Asians (Spearman's $r=0.736$, $P<10^{-16}$, **Figure 1**). This correlation [28], which also holds even when considering only failed replication attempts in East Asia ($r=0.53$, $P<6\cdot10^{-9}$), is unexpected for population-specific rare causal variants, as their effect size and Linkage Disequilibrium (LD) with the associated SNP would be different in each population.

Assessing the potential effect of publication bias

Publication bias could have inflated our replicability estimates [29,30]. Due to the large number of SNPs that are tested in a GWAS, the usual practice has been to report any new associated SNPs discovered in each GWAS, plus the replication status of previously associated SNPs. Therefore, our ability to gather replication attempts depends on how many of them are explicitly reported, which presents enormous variability among papers. This opens the possibility of a reporting bias, in which GWAS authors could prioritize mentioning successful replication attempts. If so, our chance of gathering a replication attempt may be heavily biased towards positive results and might inflate our estimates of replicability [30].

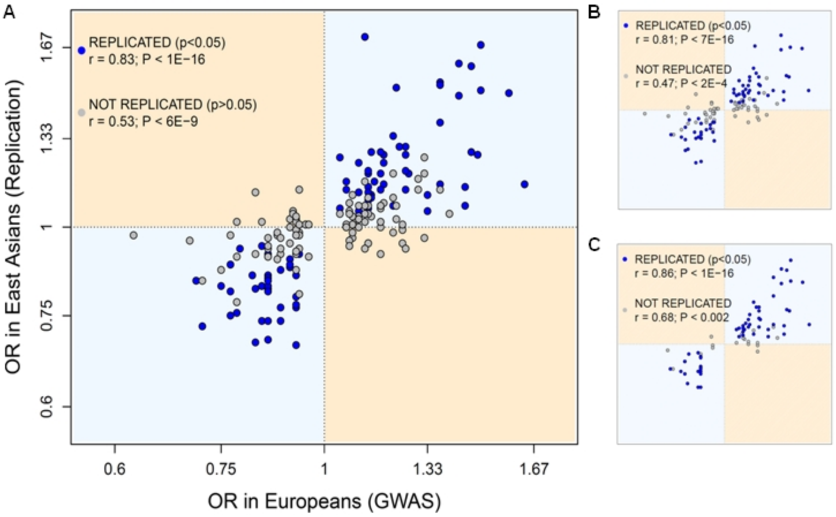


Figure 1. East Asian GWAS find the same risk allele and similar OR than European discovery GWAS.

X axis: ORs for the replication stage of the discovery European GWAS. *Y* axis: ORs for the initial stage of East Asian GWAS (**Materials and Methods**). Dots in blue indicate significant (*P*<0.05) replication attempts in East Asia; dots in grey indicate non-significant replication attempts. (A) Using all replication attempts; (B) Using only the largest replication attempt per SNP; (C) Using replication attempts with ≥80% power to replicate the OR found in Europeans.

In the most extreme version of this scenario, the 103 replications finding the same risk allele at *P*<0.05 in East Asians would be the result of Type I error with a P=0.05 threshold. In that case, the 103 positive replications would be just the 2.5% (=5% type I error x 50% probability of the same risk allele) of a 40 times larger pool of 4,120 replication attempts in East Asians (95% C.I. =3,418–4,959, assuming a Poisson distribution). In other words, 4,017 (=4,120-103) associations failing to find the same risk allele or at *P*>0.05 would have remained unreported.

To assess the potential size of that bias, we estimated the maximum number of potentially failed (*P*>0.05) but unreported replication attempts [30]. Specifically, and for each GWAS performed in East Asians, we counted the number of SNPs recorded in our database as previously discovered in European GWAS, but for which the East Asian GWAS did not explicitly mention neither a p-value nor any other information (in the main text or in the supplemental information). In total, the 416 such instances we found constitute the maximum number of cases of reporting bias in our database (**Table S7**). Most of them may not constitute reporting bias at all, since the SNPs in question may not be included in the array used for the East Asian GWAS, may be monomorphic in the studied population, may have been filtered out during QC and so on. Therefore, a systematic reporting bias cannot account for our results.

Replicability and differences in Linkage Disequilibrium and Heterozygosity

A clear prediction can be made if, as our results suggest, most associations reported by GWAS are due to the same common causal variants with similar effect sizes in the two ancestral groups: LD patterns and levels of heterozygosity should be more similar between populations in the genomic regions that contain successfully replicated SNPs than in the genomic regions with European-associated SNPs that have not reached significance in East Asians. To test this prediction, we compared the inter-continental similitude of LD and heterozygosity patterns in genomic regions harboring two different groups of disease-associated SNPs: the 47 SNPs discovered in Europeans that have been successfully replicated in every attempt with East Asians and the 65 SNPs that have never been positively replicated. We used the VarLD score [31] to measure, for each SNP, between-population LD differences in a 300-SNP region around it. We used sliding windows of 50 consecutive SNPs. As predicted, differences in LD were significantly larger for the windows centered in non-replicated (VarLD = 17.64 vs. 12.66, *P*<0.002). Moreover, these differences are only significant in the immediate vicinity of the associated SNP, and they quickly cancel out as the distance for the associated allele increases (**Figure 2** and **Table S8**). The same result was obtained using only attempts with ≥80% statistical power and contrasting 13 replicated and 38 not replicated SNPs (VarLD = 20.42 vs. 12.49, *P* = 0.045).

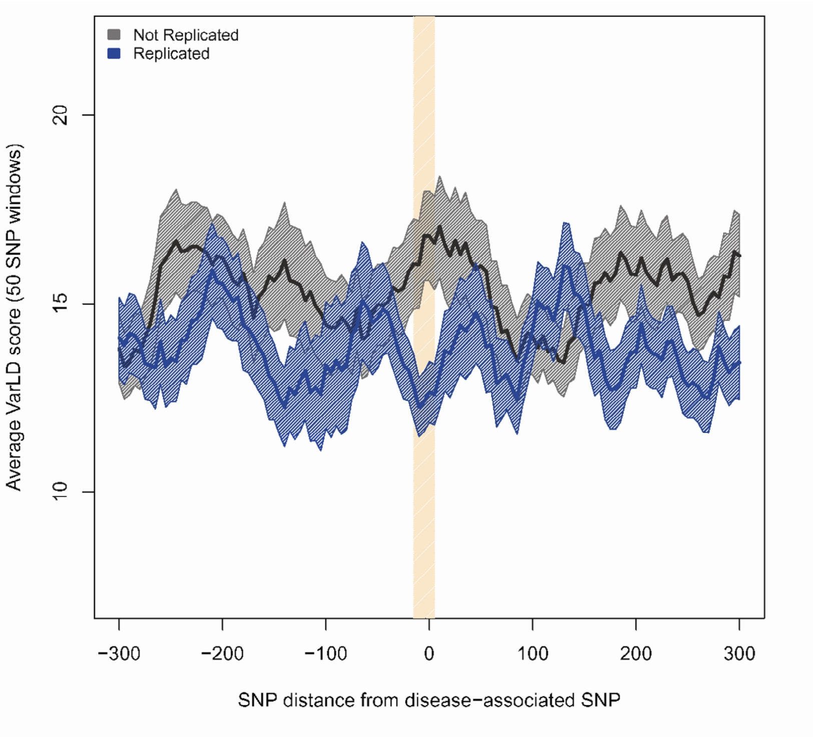


Figure 2. Regions harboring not replicated SNPs present larger differences in LD between Europeans and East Asians.

Measures of difference in LD (VarLD scores) for sliding windows of 50 SNPs with a 5-SNP step. Measures for replicated and not replicated SNPs are given as blue and black lines. Shadowed areas represent the standard error of the mean. Vertical red band indicates that all significant windows (P<0.01) locate near to the associated SNPs.

Similar patterns are observed for the windows comparing the differences in average heterozygosity between Europeans and East Asians (**Figure**

S1). Windows centered on non-replicated SNPs presented significantly larger differences in average heterozygosity across populations (0.048 vs. 0.019, *P*<0.009). Similarly to the analysis with LD, these differences accumulated in the region nearby to the associated SNPs (Figure S1), and maintained when using replication attempts with ≥80% statistical power.

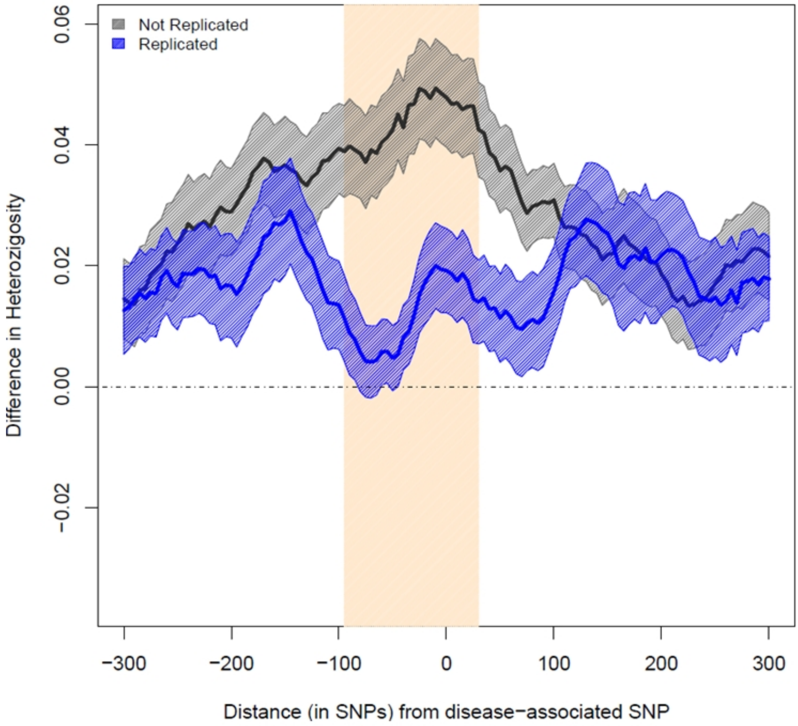


Figure S1. Difference in Heterozygosity between Europeans and East Asians.
The x axis represents the distance of each 50-SNP window from the associated SNPs. The y axis shows the difference in mean heterozygosity, namely the average for Europeans less that of East Asians (SEM indicated in shadow). The band in bisque indicates the windows with significant differences (P<0.01).

Shared LD regions and trans-ethnic mapping

Since GWAS results are highly consistent across continents, causal variants should map in regions of shared high LD with the marker SNP [1,24]. To provide a rough estimate of the average size of the genomic region harboring causal SNPs, we computed regions of similarly high LD levels that are shared between Europeans and East Asians. Specifically, and for each replicated SNP, we selected all HapMap SNPs, measured their r^2 with the associated SNP in both Europeans and East Asians and calculated the maximum range of overlapping LD in both populations. Considering $r^2 = 0.8$, we estimate that the average shared windows wherein true risk alleles lie sizes 39.5 Kb (range = 4.3–144.8, **Figure S2**). The average overlapping region not extending beyond 40 kb confirms the feasibility of trans-ethnic fine mapping [32-34].

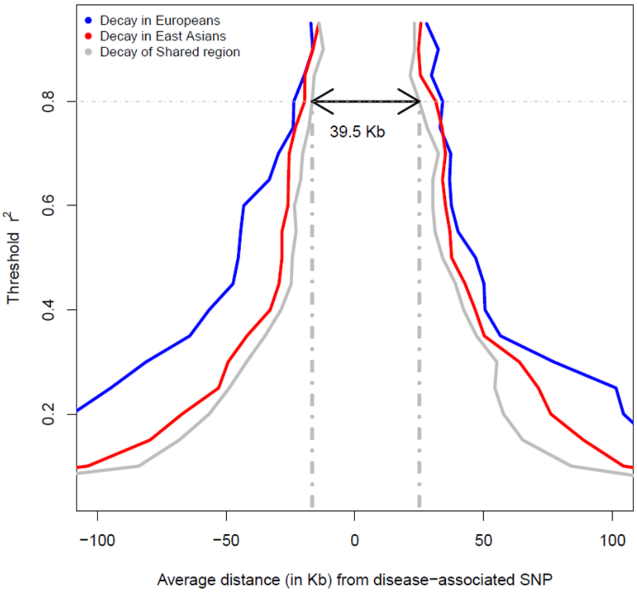


Figure S2. Average size of shared window with the same LD level.

The x axis represents the average size of the overlapping window in both populations at different levels of LD (measured by r^2). The average of $r^2 = 0.8$ is highlighted in black.

Comparison with previous results

Our results indicate that causal variants underling GWAS results are common and shared between continents, extending the observation of reports that focused in single traits [17,19]. This would seem to contradict results by us and others that highlighted heterogeneity in the genetic etiology of disease across human populations [14,21,22]. This observation contrasts with the large replicability and large correlation in OR that we observe, as well as with the role of differences in LD in explaining non-replicated associations. The apparent contradiction between the present and previous papers can be explained by two facts. First, our previous results focused on candidate-gene studies, which have been largely dominated by false positives [14]; and, second, studies that considered GWAS data had different questions, used different approaches and could gather only a limited set of traits [21] and/or associations [22].

An examination of previous datasets confirms a general trend to consistency of GWAS results across continents and emphasizes the benefits of incorporating as many associations as possible. Fu et al.[21] focused on associated SNPs discovered in East Asian GWAS. Although they used only four traits and 47 SNPs (43 loci), they demonstrated the challenges of multi-ethnic studies, and provided a framework to cope with these difficulties. As discussed by the authors, caution is warranted as to whether the disease loci and/or causal variants are population-specific. For instance, they suggested that the signals for type 2 diabetes located in PTPRD (rs17584499) and SRR (rs391300) could be population-specific, as they fail to replicate in a well-powered study in Europeans. However, we gathered several replication attempts of these signals in posterior East Asian GWAS (**Table S3**), and, out of 8 replication attempts, we observed only 1 at $P<0.05$ (when 7.44 would be expected by power) and only 4 with the same risk allele. In addition, the inclusion of more recent studies (**Table S4**) shows that an apparently European-specific variant tagged by rs12779790 (CAMK1D) could be associated also in East Asians [35]. These results make it clear that assessing a limited number of GWAS may have affected the report by Fu et al.[21].

Ntzani et al [22] examined differences in effect sizes, rather than replicability or the role of rare variants in GWAS results for 12 diseases and 4 anthropometric traits. They focused on the relatively short (~20) list of GWAS that either use samples with different ancestries in the replication stage or compare their own results with previous papers using different ancestries. In contrast, we gathered attempts from multiple GWAS on the same diseases and were able to construct a much larger and powered database. They found overall consistency in effect direction across ancestries (~82%, similar to ours of 85%), but with modest correlations in effect sizes, *rho* ≈ 0.33, that would seem contradictory with the *rho* = 0.75 we observe. Nevertheless, an almost identical correlation in OR would have been observed if limiting the study to the 22 SNPs that are shared between Ntzani et al. [22] and our dataset (rho = 0.58 and 0.53, respectively). Barring possible difference due to the different nature of the anthropometric traits analyzed Ntzani et al. [22], the previous result stresses the importance of continuously updating the list of replication attempts to guarantee powered inferences.

Effective replicability rates of larger GWAS hints at weaker but common causal variants

Of course, the finding of shared variants underlying GWAS results holds only for associations that have been published so far. Ongoing efforts to join cohorts into large consortia [36] ensure steady progress in the field and guarantee the discovery of new genetic associations to complex diseases [6,37]. It is tempting to make inferences about what may be the results of future, much larger, association studies; particularly about the frequency and degree of trans-ethnic sharing of as yet undiscovered variants. We approximated this question by considering the patterns of replicability across time. Specifically, it is clear that if the GWAS with larger sample-sizes that have been published recently for peoples of European ancestry had discovered variants with lower frequencies (variants that should be increasingly population-specific), their results should be less likely to replicate across populations. If this observation were made, it would predict decreased replicabilities in future, even larger GWAS with increased power to discover lower-frequency risk variants.

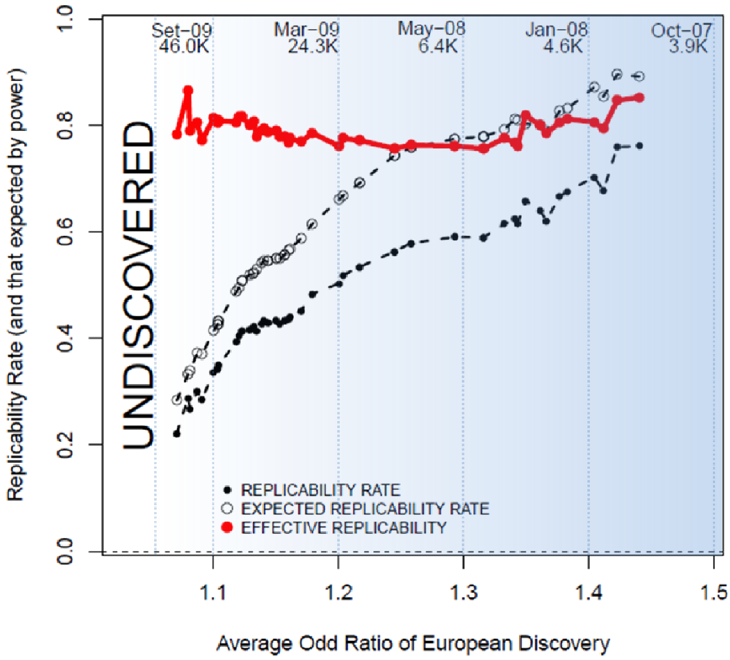


Figure 3. Replicabilities against ORs in the discovery study.

For every SNP discovered in Europeans, all the replication attempts in East Asians were considered and classified by bins of European OR. The OR of SNPs with risk alleles being major was transformed to ensure OR>1. By windows with step 0.3, the average statistical power (empty black circles), average replication success (solid black circles) and effective replicability (the ratio between observed and expected replicability, the two former quantities, red circles) are shown. Top values of the graph represent the average date of publication and sample size of discovery GWAS, for bins of 0.1 OR.

As observed in **Figure 3**, more recent GWAS have gathered larger sample sizes and unveiled associations with lower ORs. Replicability has decreased with lower ORs and tends to be lower than what would be expected out of sheer statistical power, most likely because our power calculations assume the same heterozygosities and LD patterns in different populations, which we have already showed not to be the case. An interesting inference can be made by observing effective replicability rates, the ratio between the proportion of positive replications and their statistical power. Effective replicability would be expected to decrease if the lower ORs detected by GWAS were due to lower-frequency causal variants. In contrast, we observed a constant effective replicability rate of ~80% that was independent of the OR reported in the European discovery GWAS (red line in **Figure 3**), indicating that larger GWAS detect common variants with weaker effects rather than rarer variants.

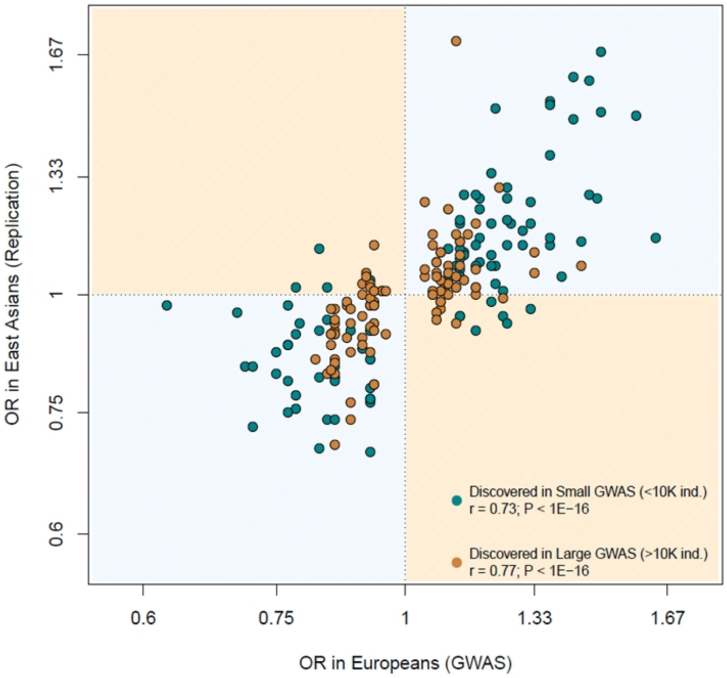


Figure S4. Similar correlation between European and East Asian OR, regardless of the discovery GWAS sample size.

The same correlation arose when using all replication attempts (as in Figure) or the filtered (n = 123) set of largest replication attempt per SNP (not shown).

This inference is confirmed by comparing the trans-continental correlation of ORs between larger and smaller GWAS. We classified the European discovery GWAS into two groups, using a threshold of 10,000 individuals to distinguish between "small" and "large" GWAS. Larger GWAS do indeed detect associations with smaller ORs (average OR 1.15 *vs*. 1.28; P<3x10^{-7}). Nevertheless, the correlation of ORs between European and Asian GWAS was the same for "small" and "large" GWAS (**Figure 4**), showing, yet again, that variants are equally shared across populations, regardless of the sample size of the discovery GWAS.

The relevance of our findings comes from the fact that, first, they settle the issue of the spurious origin of GWAS associations [7,8,10], since trans-continental replicability shows that they do correspond to true disease loci; second, they clarify the contribution of common variants to extant GWAS results, since practically all GWAS have delivered precisely what they were designed to detect [1]: associations with common variants; third, our results show that causal variants are shared across populations and that they must lie close to marker SNPs. In a context were most causal variants have not yet been found, leveraging on the increasingly varied ancestries of GWAS may help tracking them down [33,34,38]. Finally, since larger GWAS did not detect rarer variants, our findings support the infinitesimal model of the genetic architecture of disease [4]. However, it is not simple to extrapolate our results to associations that remain undiscovered so far. Whether the heritability that is not yet explained by GWAS will be partly due to risk variants in insufficient LD with common SNP markers, as suggested by some authors [6,39] or whether this heritability exists at all [40] will only be resolved by further empirical research.

MATERIALS AND METHODS

Creating a database of SNPs associated to disease

We considered the 1,171 studies indexed in the *catalog of Published Genome-Wide Association Studies* as to February 2012 (http://www.genome.gov/26525384, last accessed 14th February 2012) and classified them according to the trait under study. Each study was classified according to the genetic ancestry of the samples, considering only individuals used in the GWAS stage. Studies performed on a mixed panel were considered only if separate ancestry-specific analyses were provided and we recorded them as independent studies. We observed a strong bias towards GWAS performed with "European" (78.4%) and "East Asian" (14.9%) individuals, while much fewer studies are available for "African" (4.3%), "Hispanic" (1.2%), "Middle Eastern" (0.5%), "Native American" (0.4%) and "Oceanian" (0.3%) ancestries. Therefore, and to increase the reliability of our results, we only included GWAS performed with peoples of European and East Asian ancestry. We focused on dichotomous disease traits, avoiding anthropometric traits such as height. In order to be able to produce replicability estimates for both studied ancestries, we included in our analysis the 28 diseases for which two or more GWAS were available for Europeans and at least one in East Asians (e.g. 11 GWAS for lung cancer in Europeans and 5 in East Asians).

We built a database with 28 dichotomous disease phenotypes (**Table S1**), with data coming from 204 European and 71 East Asian GWAS. Several features of interest were recorded for each GWAS: first author, journal, year of publication, genetic ancestry, sample size in GWAS stage, total sample size in replication stage, array genotyped, genomic control factor in GWAS stage (if available), use of imputed SNPs (Y/N) and number of genomic regions achieving genome-wide significance in the initial and final stage (**Table S2**). The publications corresponding to each GWAS were downloaded from the respective journals.

For each disease, the selected studies were sorted per date of publication regardless of the population of study. Starting for the first study, we built a cumulative database of disease-associated SNPs and their replicability in successive studies. After excluding GWAS with pooled DNAs or focusing on CNVs, each GWAS publication was visually screened for two kinds of association data: the report of a new disease-associated SNPs (discovered SNPs); and the replication status of disease-associated SNPs discovered in previous GWAS (replicated SNPs). In both cases, we recorded three features from each association: (i) Odds Ratio (OR) (ii) confidence interval of the OR and (iii) the p-value.

We used several conservative criteria to include newly discovered SNPs in our database. First, to avoid the winner's curse bias, we used the OR and p-value from the replication stages of the discovery GWAS. Second, when several replication stages from the same GWAS were available, the OR from the stage with largest sample size was recorded. Only when no replication stages were available did we use the OR from the GWAS stage. Third, SNPs associated uniquely in sex-specific analyses were excluded. Fourth, ORs coming from allelic tests and additive models were prioritized over genotypic tests and other genetic models. Fifth, the genome-wide significance level for a newly discovered SNP to be included in our analysis was set at $P<5\times10^{-7}$, unless imputed SNPs were used in the GWAS, in which we toughened up the threshold to $P<5\times10^{-8}$. Sixth, for genomic regions with several genome-wide significant SNPs (SNPs less than 200 Kb from each other), we included in the study the SNP with lowest p-value. Finally, as a further conservative measure, disease-associated SNPs from the MHC region and HLA alleles were not included in the study.

To include replication attempts in our database, several conservative conditions had to be met. We only recorded attempts in which exactly the same SNP than in the discovery GWAS had been genotyped. Second, in all these cases, the p-value considered for the replication report was the one from the GWAS stage. Finally, the OR for each disease-associated SNP was referenced for the allele that had been the risk allele in the discovery study. Thus, OR < 1 means that the minor allele was found as protective in the discovery study, while OR > 1 means that the minor allele appeared as the risk allele. For SNPs with different minor alleles across populations, OR were referenced to the minor allele specific for each population. Instances of the latter are indicated in column “Shift” in **Table S4** and the shifted OR is represented in all Figures except when otherwise indicated.

A total of 413 discovered SNPs from 331 genomic regions were found to be associated to disease, 316 of those SNPs being reported for the first time in Europeans and 97 in East Asians (**Table S3**). In total, we gathered 465 replication reports, dealing with 217 out of the 405 discovered SNPs (**Table S4**). Out of the 465 replication reports, 205 and 260 corresponded, respectively, to attempts performed on Europeans and on East Asians. Since East Asian GWAS are more recent, most of the replication attempts (400 out of 465, 87%) reported the replication status of discovered SNPs that had been reported for the first time in Europeans. Therefore, we focused on the subset of 402 replication attempts gathered for 182 associated SNPs discovered in European GWAS. Out of these, a total of 177 and 225 replication attempts had been reported for Europeans and East Asians, respectively.

The 225 replication attempts in East Asians aimed to replicate a total of 131 SNPs associated to disease with genome-wide significance in Europeans, which results in an average of 1.75 replication attempts per associated SNP (range = 1–7). Thus, our estimates of replicability could be biased if replicated SNPs gathered more replication attempts per SNP, or more associated SNPs in European populations. During the analysis, and as noted in the text, we applied an additive filtering to ensure no bias in the estimates of replicability and correlations between European and East Asian OR. Specifically, we repeated the analysis selecting only the largest replication attempt per SNP, resulting in a filtered set of 123 attempts. The SNPs ascertained for the filtering are indicated in **Table S4**.

Population genetics analysis (VarLD and Heterozigosity)

Polymorphism data was downloaded from HapMap Project Phase 2 (release 24, November 2008). For each ascertained SNP, we downloaded two data sets: (i) genotypes for the associated SNP and (ii) genotypes for a 600-SNP window centered on the associated SNP. We downloaded all genotypes for all unrelated samples from the three populations of European and East Asian ancestry (CEU, JPT and CHB). JPT and CHB samples were clustered together due to their close genetic relationship.

Population differences in local patterns of Linkage Disequilibrium (LD) around disease associated SNPs were measured with the VarLD software (www.nus-cme.org.sg/software/varld.html) [31], using the targeted option for 50-SNP windows. For each population and genomic region, VarLD builds a matrix of pairwise signed r^2 values among all the SNP pairs and provides a raw score corresponding to the absolute difference in the eigen-decompositions between two matrices. This score is a summary measure of the overall LD levels in a given genomic region between two populations. We used it to measure the extent of differences in local LD between two kind of genomic regions: these containing replicated and non-replicated SNPs. To rule out the possibility that differences in LD between replicated and not-replicated SNPs are not related to the presence of the disease associated SNP, we scanned VarLD differences in consecutive windows of the same size (50 SNP), starting 300 SNPs upstream of the disease associated SNP and finishing 300 SNPs downstream, with an step of 5 SNPs. In total, we checked 121 consecutive windows around the disease associated SNP. On average, we were examining a window of 503.61 Kb centered on each associated SNP.

We used a similar sliding window approach to summarize the differences in allele frequencies between Europeans and East Asians. Again, we did it for each SNP, calculating the average heterozygosity in each window for replicated and non-replicated SNPs. Differences in heterozygosity are simply the result of subtracting the average in East Asians from that in Europeans (**Figure S1**).

Power and Statistical analyses.

As noted in the text, for some analysis we focused on the attempts that had >80% power to replicate the effect size found in Europeans. Statistical Power was calculated with the CaTS Power Calculator (www.sph.umich.edu/csg/abecasis/CaTS/) [41]. For each replication attempt we checked the power under a log-additive model to find the same effect size as in the discovery European GWAS, given the sample size of the replication GWAS and the allele frequency of the risk allele in East Asians.

Statistical analyses were performed using standard R procedures. The significance of the replicability estimates was checked by means of a binomial test, with an expected replicability rate of 0.05 under the null hypothesis of no shared associated SNPs between Europeans and East Asians. Similarly, the significance in the risk allele direction was checked by means of a binomial test, using a null expected ratio of 0.5. As indicated in the first section, differences in LD between replicated and non-replicated SNPs were checked by means of Mann-Whitney tests comparing the distributions of VarLD scores for sliding 50-SNP windows centered on the disease-associated SNPs. The same procedure was used for the average difference in heterozygosity and distributions of OR found by “small” and “large” GWAS.

ACKNOWLEDGEMENTS

We acknowledge colleagues from Navarro’s lab for technical support and helpful comments. We also thank Peter M. Visscher, Isabel Mendizabal, Rui Faria and Juan A. Rodríguez for valuable discussion and smart insight. This was supported by a PhD fellowship to U.M.M and a grant to A. N. from the Ministerio de Ciencia e Innovación (Spain, BFU2009-13409-C02-02) and by the Spanish National Institute for Bioinformatics (www.inab.org).

COMPETING INTERESTS

The authors have declared that no competing interests exist.

REFERENCES

1. Visscher PM, Brown MA, McCarthy MI, et al. (2012) Five years of GWAS discovery. Am J Hum Genet 90: 7-24.
2. Hindorff LA, Sethupathy P, Junkins HA, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.
3. Eichler EE, Flint J, Gibson G, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446-450.
4. Gibson G (2012) Rare and common variants: twenty arguments. Nature Reviews Genetics 13: 135-145.
5. Manolio TA, Collins FS, Cox NJ, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
6. Yang J, Benyamin B, McEvoy BP, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565-569.
7. Anderson CA, Soranzo N, Zeggini E, et al. (2011) Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol 9: e1000580.
8. Dickson SP, Wang K, Krantz I, et al. (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8: e1000294.
9. Orozco G, Barrett JC, Zeggini E (2010) Synthetic associations in the context of genome-wide association scan signals. Hum Mol Genet 19: R137-144.
10. Wray NR, Purcell SM, Visscher PM (2011) Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol 9: e1000579.
11. Adeyemo A, Rotimi C (2009) Genetic variants associated with complex human diseases show wide variation across multiple populations. Public Health Genomics 13: 72-79.
12. Bustamante CD, Burchard EG, De la Vega FM (2011) Genomics for the world. Nature 475: 163-165.
13. Ioannidis JP (2009) Population-wide generalizability of genome-wide discovered associations. J Natl Cancer Inst 101: 1297-1299.
14. Marigorta UM, Lao O, Casals F, et al. (2011) Recent human evolution has shaped geographical differences in susceptibility to disease. BMC Genomics 12: 55.
15. Rosenberg NA, Huang L, Jewett EM, et al. (2010) Genome-wide association studies in diverse populations. Nat Rev Genet 11: 356-366.
16. Park JH, Gail MH, Weinberg CR, et al. (2011) Distribution of allele frequencies and effect sizes and their interrelationships for

common genetic susceptibility variants. Proc Natl Acad Sci U S A 108: 18026-18031.

17. Waters KM, Stram DO, Hassanein MT, et al. (2010) Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. PLoS Genet 6.
18. Durbin RM, Abecasis GR, Altshuler DL, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
19. Waters KM, Le Marchand L, Kolonel LN, et al. (2009) Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. Cancer Epidemiol Biomarkers Prev 18: 1285-1289.
20. Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44: 243-246.
21. Fu J, Festen EA, Wijmenga C (2011) Multi-ethnic studies in complex traits. Hum Mol Genet 20: R206-213.
22. Ntzani EE, Liberopoulos G, Manolio TA, et al. (2012) Consistency of genome-wide associations across major ancestral groups. Hum Genet 131: 1057-1071.
23. Hindorff LA, MacArthur J, (European Bioinformatics Institute), Wise A, et al. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies.
24. McCarthy MI, Abecasis GR, Cardon LR, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356-369.
25. Gravel S, Henn BM, Gutenkunst RN, et al. (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108: 11983-11988.
26. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740-743.
27. Nelson MR, Wegmann D, Ehm MG, et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337: 100-104.
28. Okada Y, Terao C, Ikari K, et al. (2012) Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat Genet 44: 511-516.
29. Barsh GS, Copenhaver GP, Gibson G, et al. (2012) Guidelines for genome-wide association studies. PLoS Genet 8: e1002812.
30. Lohmueller KE, Pearce CL, Pike M, et al. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33: 177-182.

31. Ong RT, Teo YY (2010) varLD: a program for quantifying variation in linkage disequilibrium patterns between populations. Bioinformatics 26: 1269-1270.
32. Shriner D, Adeyemo A, Gerry NP, et al. (2009) Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. PLoS One 4: e8398.
33. Saxena R, Elbers CC, Guo Y, et al. (2012) Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. Am J Hum Genet 90: 410-425.
34. Zaitlen N, Pasaniuc B, Gur T, et al. (2010) Leveraging genetic variability across populations for the identification of causal variants. Am J Hum Genet 86: 23-33.
35. Cho YS, Chen CH, Hu C, et al. (2012) Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. Nat Genet 44: 67-72.
36. Sullivan PF, Daly MJ, O'Donovan M (2012) Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat Rev Genet 13: 537-551.
37. Park JH, Wacholder S, Gail MH, et al. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42: 570-575.
38. Twee-Hee Ong R, Wang X, Liu X, et al. (2012) Efficiency of trans-ethnic genome-wide meta-analysis and fine-mapping. Eur J Hum Genet.
39. Yang J, Manolio TA, Pasquale LR, et al. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43: 519-525.
40. Zuk O, Hechter E, Sunyaev HR, et al. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. PNAS 109: 1193-1198.
41. Skol AD, Scott LJ, Abecasis GR, et al. (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38: 209-213.

4 Discussion

Gentlemen, we have run out of money; now we have to think
Winston Churchill, 1945 (attributed)

Mistakes were made
Classic exonerative linguistic construct in Washington

In this section, I will discuss the findings of the two presented studies that are based on comparing the results from association studies performed across human populations. The results will be interpreted in terms of some of the current issues in the field of the genetics of complex disease. First, I will discuss the current validity of candidate gene association studies (section 4.1.1) and the role of genetic heterogeneity in replication failures (section 4.1.2). Next, I will examine the main lessons on how to address the results from genome-wide association studies (section 4.2.1) and the main inferences on the genetic architecture of disease. Finally, I will check several of the possible directions in the near future.

4.1. Inferences from candidate gene association studies

In the first study (section 3.1.) we analyzed the role of genetic heterogeneity among human populations in the replicability of genetic association studies. To address this question, we took advantage of more than 25 years of candidate gene studies and measured the degree of population differentiation (as measured by the F_{ST} statistic) in loci that have shown differential patterns of association to disease. We created two sets consisting in 890

“Global” and 37 “Continental” gene-disease associations that pointed at genetic distance being correlated with replicability. In addition, we observed that genes from highly replicated associations harbour more derived alleles present as major alleles in populations. Several risk alleles had already been shown to present differential frequencies and effect sizes across populations (Ioannidis et al. 2004; Ioannidis et al. 2001). At any rate, our work produced the first general study of candidate gene replicability in the context of population differentiation and confirmed the role of the recent evolutionary history in disease susceptibility patterns.

At the time of our analysis the predominant vision in the community was very pessimistic. Few associations showed a consistent pattern of replication and most meta-analyses pointed at the pervasive presence of publication bias and lack of statistical power (Hirschhorn and Altshuler 2002; Hirschhorn et al. 2002; Lohmueller et al. 2006; Lohmueller et al. 2003; Pan et al. 2005). At that time, the Genetic Association Database (GAD) we used informed about the features and results from over 39,000 attempts to associate genes with disease. However, we had to limit ourselves to the ~17,000 and ~7,000 attempts that, respectively, informed about the conclusions and the ancestry of the individuals tested. The figures on the abundance of candidate results contrast with the difficulty to extract sound conclusions from these efforts.

Our analysis was based on two premises. First, we hypothesized that at least some true associations had been unravelled by candidate gene methods and, hence, were available in the database. Further, part of the lack of replication problem of the true candidate associations could be attributed to genetic variability in disease architecture across populations.

4.1.1. The reliability of candidate gene results

It is difficult to evaluate our first premise. Only a handful of disease associations have been validated molecularly in the laboratory and/or from a clinical perspective. Nonetheless, the figure of true associations could be somehow approximated through a comparison to the second study based on GWAS (result 3.2). When translating the 28 diseases analyzed in the second study into the first study, they account for 462 of the 890 associations and 286 of the 403 genes ascertained for the Global set. Out of these, 19 associations have been "re-discovered" in our GWAS database. Thus, 4.1% of associations (19 out of 462) and 6.6% of disease genes (19 out of 286) from the Global set of candidate gene studies have been found again in GWAS studies. Using a conservative estimate of 20,000 genes in human genomes and considering that 289 independent genes are recorded in the GWAS database, the figure of 19 "re-discovered" genes represents a 4.6-fold enrichment compared to the four overlapping genes expected if GWAS results randomly overlapped signals from candidate gene studies.

There are few differences across diseases when looking at the number of candidate gene associations "re-discovered" by GWAS. A total of 11 out of the 28 diseases account for all the "re-discovered" loci and the number of such signals per disease strongly correlates with the total number of loci described by GWAS (rho = 0.54; *P* = 0.004). That is, the more associations described by GWAS, the highest are the chances of replicating genes from candidate gene studies.

Three diseases present the highest rate of "re-discovered" associations: alcohol dependence (2 out of 3, 67%), breast cancer (2 out of 9; 22%) and Parkinson's disease (3 out of 17; 18%). It is tempting to speculate about shared genetic basis of Mendelian forms of breast cancer and Parkinson's disease and their counterparts in the form of complex disease. If so, genes found by linkage in familial versions of disease could have successfully driven the research by candidate gene methods and, eventually, be "re-discovered" by GWAS. This scenario would fit historical practice in candidate gene studies (i.e. testing loci from linkage regions), but any comparison with Mendelian genes gathered from OMIM data would lack statistical support.

The correspondent numbers for the Continental set look similar. The 4 out of 37 "re-discovered" associations represent a larger enrichment (9-fold) of overlaps with our GAD results, but they account for only 10.8% of the signals present in the Continental set. Even if acknowledging the large number of false negatives expected in GWAS, only a tiny number of associations from the candidate gene sets appear to be confirmed in the second study.

Yet, a reassuring aspect of the analyses from our first study based on candidate gene lies in the different filters we applied to ensure the reliability of our observations. Specifically, the correlation coefficients between genetic differentiation and replicability that we report enhance progressively when filtering out (i) associations with lowly powered studies and (ii) associations with <50% replicability. Importantly, the 19 GWAS-confirmed associations present larger replicabilities than non-"re-discovered" associations (74.2% vs. 65.8%; *P*=0.052, permutation test). Indeed, none of the 19 "re-discovered" associations presents a replicability lower than 50%, while this is the case for 100 (22.6%) out of the other 443 associations (*P*=0.011; binomial test). Moreover, the candidate gene associations further validated by GWAS have been studied more times in our database (12.3 vs. 8.8 studies; *P*=0.016, permutation test). Thus, the amount of "re-discovered" candidate associations becomes progressively larger when only the most studied (and thus reliable) associations are ascertained (i.e. 13/164 and 6/57 of those with ≥8 and ≥12 studies, respectively). Future work could be done to enlighten the reasons that lie behind the poor overlap of candidate gene and GWAS results (e.g. what happens if GWAS significance thresholds are relaxed?).

The low number of "re-discovered" associations difficult the statistical validation of our results in this much shorter pool of associations. Though, the strengthened correlations between replicability and genetic differentiation when more consistent associations are selected and the strong enrichment of GWAS signals in the more reliable continental set indicate that the results we report are hardly spurious.

4.1.2. Exploring the scenarios behind candidate associations

We considered several possible explanations for the observed trends of replicability being correlated with genetic differentiation across populations, but neither statistical power nor rare variants account for the role of population heterogeneity in replicability.

A major role for statistical power was discarded because the inclusion of the average sample size in the regression analysis did not explain the differences in continental replicability. Still, rare variants could also affect replicability by giving risk exclusively in certain ethnicities. We looked for loci with an excess of rare variants in any continent, but it did not correlate with replicability. We further discarded a major role for these variants through an extensive scrutiny of the 444 papers included in the 37 continental associations. This procedure allowed us to approach the causal variants by focusing in the reported markers rather than in gene-based differentiation summaries. However, only 4 of the 54 reported markers (7.4%) presented extreme allele frequencies through populations (MAF<0.05).

The prospect that disease variants from candidate genes are common leaves room for other factors shaping their inconsistent replication patterns. Even if variants were common, and thus likely to be shared, candidate gene studies could fail to replicate if differential gene-by-environment effects or linkage disequilibrium patterns are present across populations. Additionally, our observation that highly replicated genes harbour increased amounts of major derived alleles could be explained if causal alleles are common. Specifically, phenomena such as allele surfing and the reduced ability of purifying selection to purge deleterious alleles during the out-of-Africa event shared by Europeans and East Asians might have increased the frequency of disease alleles in these populations. Only the availability of African association studies and their replicability patterns would help evaluating this possibility. In principle, we would expect a lower success in attempts to replicate "Eurasian" associations even if shared to all humankind, because the more effective purifying selection in Africans would have maintained them at lower frequencies. Indeed, genes associated to disease do harbour lower proportions of major derived alleles in Africans, but this signal could be entirely due to the ascertainment bias in SNP arrays instead of stringent purifying selection in Africans.

After tens of thousands of candidate gene studies, the literature on the genetics of common diseases remained full of uncertainties. The few triumphs drifted in a sea of failed signals. The possibility that true ethnic heterogeneity may explain replication failures is the main contribution of the first study. It truly constitutes a positive message after decades of frustration. Still, this study only serves to extract general patterns across disease and does not permit to ascertain those associations more likely to be true.

Regardless of the exact value of the thousands of published candidate gene associations, the increasing capability of genotyping technologies induced an explosive leap forward. The wave of published GWAS that started in 2007 has afforded many associations that really fill the community standards. From 2010 on, the generalization of non-European GWAS allows using the patterns of replication across populations to deepen our knowledge about the genetic architecture of disease.

4.2. Inferences from genome-wide studies

In the second study (section 3.2) we used replicability patterns of GWAS results across Europeans and East Asians to make inferences about the genetic architecture of complex disease. We created a database of 413 genome-wide significant associations described in 275 GWAS for 28 different complex diseases. This dataset built on publicly available data was used to work on several of the questions that are currently most pressing, such as the potential role of rare causal alleles in GWAS results, the transportability of disease associations across populations and the search for alternative models to explain the "missing heritability".

4.2.1. Collecting data from genome-wide association studies

We gathered 316 and 97 SNPs reported for the first time in Europeans and East Asians, respectively. Next, we carefully looked across GWAS to ascertain replication attempts for discovered associations. In total, we gathered 465 attempts describing the p-value and effect size of previously associated SNPs. Given that 87%

of attempts focused in SNPs discovered in Europeans (402 of 465), we concentrated in the 177 and 225 attempts that aimed to replicate any of the 182 SNPs discovered in Europeans.

The process of building an extensive database from a large pool of publications spanned several months. We manually checked ~25% of the literature (275 of 1,200 GWAS). This figure represented almost 100% of the literature available for our purpose of focusing on diseases studied in non-Europeans. Sadly, no studies performed on Africans were available and diseases prevalent in Africans such as malaria were not included due to the lack of counterpart studies in other populations.

We devised several conservative criteria to avoid any major bias. To name but a few, (i) we assigned the effect size from the largest replication attempt available in discovery GWAS to skip the winner's curse bias and (ii) we only considered attempts in which exactly the same SNP than in the discovery GWAS had been genotyped. The list of selection procedures is available in the Supplementary Information for study 3.2.

In addition, we followed a very stringent policy to include studies to our database. Almost 25% of eligible GWAS (67 out of 275) were discarded. For instance, we avoided any study with pooled DNA due to their proneness to false positives. Most exclusion events were based on phenotypic heterogeneity. Researchers try to enrich the fraction of genetic causality in cases through the inclusion of individuals with extreme phenotypes or with familial histories of disease. This is a sensible policy to describe new disease loci, but makes it difficult to compare effect sizes across studies without similar selection criteria.

For instance, the first GWAS published for type 2 diabetes used less than 1,300 individuals but described five different associated loci (Sladek et al. 2007). In sharp contrast, the next three GWAS on diabetes found at most four new loci even if using ~10,000 individuals (Saxena et al. 2007; Scott et al. 2007; Zeggini et al. 2007). Most probably, the success of the first study lied in the inclusion characteristics of disease subjects: cases ought to be of young onset (<45 years), non-obese (BMI<30) and with diseased relatives. It is reassuring that three out of the five described loci have been replicated again and again in ongoing GWAS. However, the first study described outstanding effect sizes that are not comparable with those from ongoing studies (i.e. rs7903146 in TCF7L2 showed an OR of 1.65 while all following European studies have described OR<1.4). Of note, exclusions were abundant for GWAS looking for risk mutations in cancer (i.e. 7 of 16 GWAS on aggressive prostate cancer were excluded due to phenotypic heterogeneity).

The case of asthma is a clear example of the drastic effects of phenotypic heterogeneity in replicability. Even if excluding several GWAS based on occupational versions of the disease, we allowed studies using childhood versions of asthma. Interestingly, asthma stands out as the only disease with a low replicability rate in

Europeans (13 in 26 attempts at P<0.05, while 21.84 would be expected out of statistical power). Thus, the adoption of strict GWAS filtration criteria was the key to ensure comparable results across studies.

One aspect of particular interest lies in the procedure followed to address the problem of reporting bias. Ensuring the gathering of all available replication attempts is necessary to guarantee proper measures of replicability. However, GWAS studies excel in numbers: hundreds of thousands of SNPs are tested and lots of false positives are expected. Thus, publications tend to report a small minority of results and in very few cases all available p-values are listed. At best, a few hundreds of best SNPs are reported in the supplementary files. Authors, publishers and readers might be interested in looking at true new loci, but such policies can heavily bias our replicability estimates if masking failed replication attempts. We devised a strategy to calculate the expected numbers of outcome reporting bias present in our database. We found that even in the extreme scenario in which all non-reported attempts were failed and had been masked by the authors, reporting bias would have negligible effects in the observed rates of replication.

4.2.2. Inferences from genome-wide association studies

In fact, risk variants associated to disease present very large replicability rates when tested in the same discovery continental population. Specifically, 85% (150 of 177) of the replication attempts performed on Europeans for SNPs discovered in that population find the same risk allele at P<0.05. In fact, intra-

European replicability rate increases to 94% when using attempts with ≥90% statistical power to replicate the discovery odds ratios (109 of 116, while 114.3 would be expected out of statistical power). Similar proportions are observed for the narrower set of replication attempts within East Asians (72%, 25 of 35). These figures confirm that stringent GWAS thresholds (i.e. $P<5*10^{-7}$) ensure that most risk alleles present in the database are true signals.

The database of GWAS replicability allowed us to investigate the effect of setting different thresholds to declare a replication attempts as significant. At *P*<0.05, we observe a 17-fold enrichment compared to the random expectation if all discovered SNPs were false positives. Interestingly, the enrichment increases progressively when more stringent thresholds are selected. Eventually, lots of positive replications are observed when almost none would be expected. Although certain SNPs could well be prone to appear as false positives because of high substructure in allele frequencies across Europeans, these results indicate that significant evidence from replication GWAS is almost conclusive to distinguish between true and false positives (Figure 21).

P-value threshold	**Observed positive replications (*P*<0.05)**	**Replications expected by chance**
0.05	150	8.8
0.01	130	1.77
0.001	113	0.177
1E-05	72	1.77E-03
1E-07	51	1.77E-05

Figure 21. Replicability observed in Europeans for different thresholds. Hardening the significance thresholds lend support to true signals.

The second result consists in confirming that most risk loci discovered through GWAS are shared between Europeans and East Asians. Replicability rates of European SNPs drop to 46% when tested in East Asians (103 of 225 attempts replicate the same allele at *P*<0.05). However, most of the observed decay can be attributed to the low statistical power of East Asian GWAS. Most SNPs discovered in Europeans are only significant (thus "discovered") in GWAS from consortia that joined efforts to increase power through larger sample size. In contrast, most GWAS in East Asians belong to the first wave of studies, and thus present much lower sample sizes (see (Rosenberg et al. 2010)). Strictly, replicability rates in East Asians rise from 46% to 72.5% (61 out of 82), 83% (50 out of 60) and 88% (48 out of 55) when using only attempts with 80%, 90% and 95% power to replicate the odds ratio found in Europeans.

These numbers emphasize that GWAS risk loci are shared across Eurasians. Further, we can explore the allele frequency of causal variants. Three pieces of evidence discard a major role for rare variants in GWAS associations and point at common and shared variants underlying the captured signals. First, we find that the direction of effects is almost universally shared across populations: the same risk alleles are found across studies, regardless of the replication p-value or effect size observed. In fact, even well-powered (>80%) but failed (*P*>0.05) East Asian replication attempts, that are strongly suggestive of European-specific disease associations, present the same risk allele in 74% occasions (14 out of 19). The concordance in risk alleles is unexpected if population-specific rare variants account for the signals of risk loci that are shared between European and East Asian GWAS. Reassuringly, similar evidence appears when looking at SNPs “discovered” in East Asians and its replication attempts performed on Europeans. Specifically, 89% (24 out of 27) of such attempts find the same risk allele as in East Asians and the figure jumps to 100% (23 out of 23) in well-powered attempts that are more likely to get the true risk allele in Europeans right, if there is any.

Further support to the presence of the same common variants in shared disease loci comes from paying attention to the odds ratios. Even if common, if risk variants responsible for the GWAS signals were different across populations, we would not expect any correlation in the estimated effects. In contrast, not only cohorts from Europe and East Asia do show the same risk allele, but do it similarly: odds ratios are correlated irrespective of the statistical significance in replication attempts.

Final evidence lies in the patterns of linkage disequilibrium and heterozygosity in replicated and not replicated associations. If the same common variants underlie most GWAS associations, the only biological reason to explain replication failures would lie in linkage disequilibrium and allele frequency differences across populations. Specifically, risk variants discovered initially in Europeans could remain undetected in East Asians if the linkage between the causal allele and the marker SNP has disappeared in the latter population. Thus, the policy we have followed to consider replication only using the same SNP would not be enough if significant SNPs are only good proxies for causal variants in certain populations. This prediction is very difficult to test directly given that very few causal variants are known. Nevertheless, we tested this prediction indirectly by measuring differences in LD patterns. The accumulation of differences in LD in not replicated SNPs lends support to the hypothesis that replication failures are due to tagSNPs being bad proxies for causal variants in East Asians.

Overall, these observations suggest a model whereby GWAS have detected associations that are: (i) common as regards to allele frequency, (ii) shared across Eurasia, (iii) not caused by synthetic associations with rare alleles and (iv) presenting very similar pathological effects across populations.

4.2.2. Immediate implications of the inferences on GWAS

Similar to the first study based on candidate gene association studies, the aforementioned analyses on GWAS replicability does not help to assign which are the causal variants but help to generate immediate useful predictions and tips for human geneticists working in complex disease:

1. Trans-ethnic mapping should help to unravel the causal variants behind GWAS associations
2. The incorporation of African individuals will be of the greatest help to narrow down the location of causal variants
3. Local re-sequencing around the most significant SNPs should suffice to map risk alleles

4. The combination of samples from different ancestries could help to overcome the limited sample size of cohorts from rare diseases (i.e. ankylosing spondylitis)
5. The effect sizes observed in Europeans are likely to be portable for genetic testing across ancestries

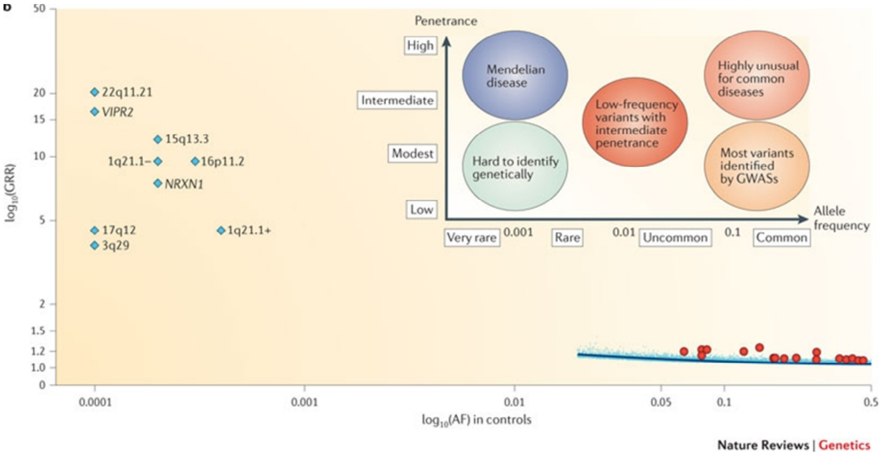


Figure 22. Known architecture of complex disease (Sullivan et al. 2012). The figure shows the allele spectrum of all common and rare variants that have been associated to schizophrenia. Three kinds of risk variants can be distinguished. First, in the left very rare alleles of large effect found through individuals with extreme phenotypes are shown. Second, the red dots in bottom right indicate the allele frequency and risk ratios of 17 SNPs described as genome-wide significant by GWAS. Finally, the blue line indicates the frequencies for the top 20,000 SNPs found in GWAS. Two predictions appear straightforward: (i) thousands of low-risk causal variants will be detected through larger GWAS and (ii) hundreds of low-frequency variants (from MAF 0.001 to 0.01) remain undiscovered.

A replicability study of extant GWAS results only informs us about the genetic architecture of disease that has been unravelled so far. However, available data can provide some hints on the “missing heritability”. We analyzed the rates of replicability across different bins of odds ratio to check whether the level of replicability holds constant or falls as we introduce low-effect risk alleles.

This exercise aims to work on current tendencies of replicability to mimic the performance of future and larger GWAS. We did so using the following reasoning: replicability should only drop when variants are not shared across continents. Importantly, the decay in statistical power as we focus on low-risk alleles (i.e. OR<1.1) must be taken into account. Otherwise, we would wrongly conclude that increasing amounts of non-shared variants are present in low-risk associations. Nevertheless, we describe a near constant effective replicability rate that is independent of the observed risk ratio.

With the caveat of the limited associations that are known so far, it does not seem that population-specific variants are being captured through low-risk associations discovered in the largest GWAS published so far. This result supports an infinitesimal model of disease thereby a large number of shared (and thus likely common) variants remain buried as false negatives in current GWAS.

4.3. A few (apparently) easy predictions and the big bet

There are three conclusions that are crystal clear to any individual that spends five years using publicly available databases to make sense of (likely biased) data on the genetics of disease. First, the use of public data will enhance your capabilities to devise projects that depart from the work in the original publication but eventually become unrealistic. Second, your happiness will (inversely) correlate with the length of the supplementary in *Nature Genetics* papers. Finally, sooner or later, you will end up showing a picture of an iceberg in every last slide shown in seminars.

Beyond such daily nuisances, any individual working in the field will experience euphoric revelations after reading interesting papers or attending funny seminars. Usually, the short episodes of excitement will happen in the middle of long periods of disappointment due to the "*technologically-driven tendencies of my field*" or "*we should go back to families*" or "*I should already know about this*". Yet, eureka moments may help predicting what is going to be next fun. I would like to cite a few predictions that may seem obvious but may also be immediately revealed as plain nonsense:

⇒ Very large studies will describe many new disease loci that will be shown to harbour lots of pathogenic rare variants

⇒ Re-analysis of already published data with pathway perspectives will be useful for drug design

⇒ Increasing number of papers will describe examples of epistasis and gene-by-environment interactions

⇒ Routine genome sequencing will discover an astonishing number of pseudo-mendelian versions of complex disease

⇒ Sequencing will serve to cure extreme mendelian phenotypes

⇒ Knowledge driven data analysis (i.e. tissue-specific expression) will be key to understand disease pathology

⇒ A molecular perspective will validate disease associations and eventually re-conquer the field, but only through a systems biology approach (not focused in DNA sequencing)

⇒ The bottleneck in the field will still lie in phenotypic data

⇒ People will fancy the increasing capabilities of genomic medicine (similar to current recreational genomics)

⇒ Most people working in the field will be employed by genomics companies still to be born

I plan to keep trying on it, but great troubles lie in the complexity of human lives: each of us experiences a singular combination of environmental and genetic variants. After the low hanging fruits delivered by the last tours de force, it might be that success lies within efforts capable of limiting the tests to be done…

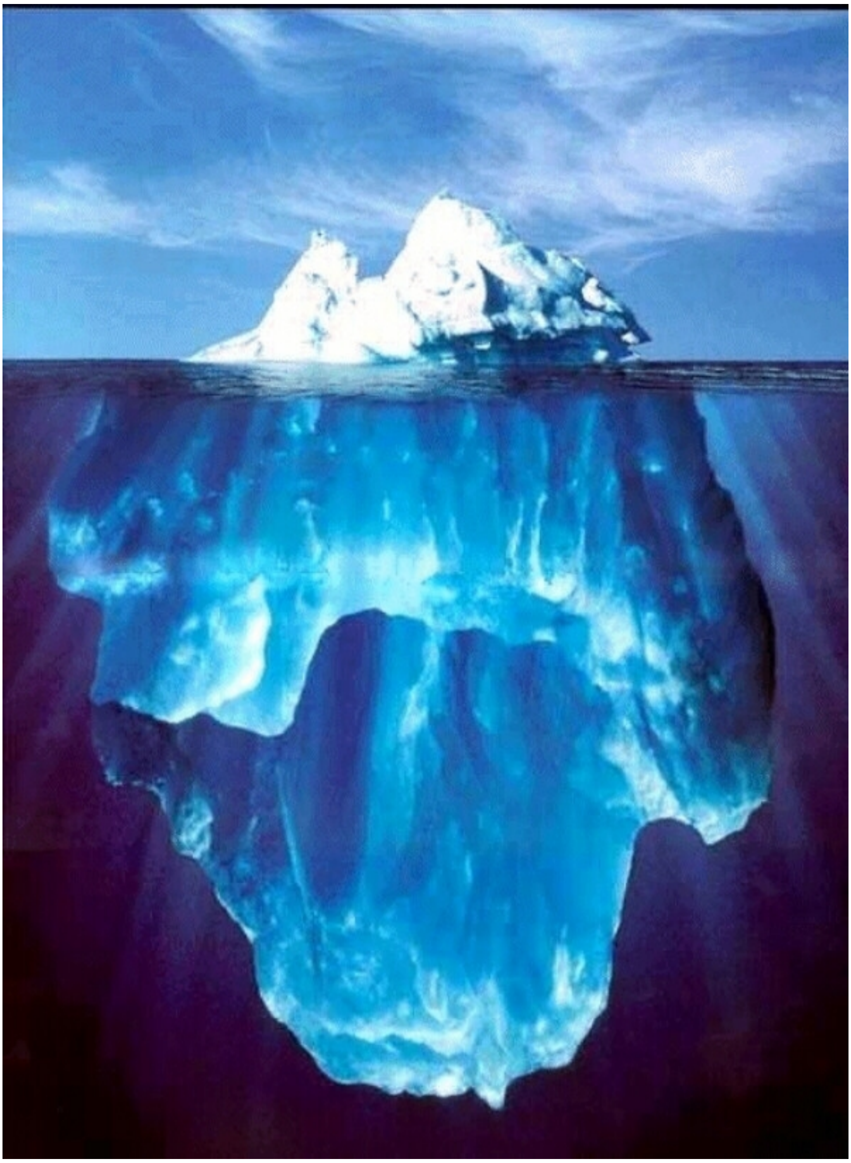
…but still the big bet is the following one by Leroy Hood in 2002:

"My prediction is that in 10-15 years, we will have identified hundreds of genes that predispose to disease. We will be able to analyze the relevant DNA sequences from these genes from a small amount of blood and use these to predict a probabilistic future health history for each individual. This is predictive medicine. Since it is an anathema in medicine to predict without being able to cure or prevent, we will use systems approaches over the next 15-25 years to place defective genes in the context of their biological systems and learn how to circumvent their limitations. This is preventive medicine. The agents for preventive medicine will include drugs, stem cell therapy, engineered proteins, genetically-engineered cells, and many others. Because each of us will have different potential disease combinations, medicine will become highly personalized. My prediction is that preventive medicine will extend the average lifespan by 10-30 years."

Leroy Hood
My Life and Adventures Integrating Biology and Technology

But there are also unknown knows…

Slavoj Žižek



Bibliography

Adeyemo A, Rotimi C (2009) Genetic variants associated with complex human diseases show wide variation across multiple populations. Public Health Genomics 13: 72-9

Anderson CA, Soranzo N, Zeggini E, Barrett JC (2011) Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol 9: e1000580

Balding DJ, Bishop M, Cannings C (2007) Handbook of Statistical Genetics, 3rd edn. John Wiley & Sons, Ltd.

Bamshad M, Wooding S, Salisbury BA, Stephens JC (2004) Deconstructing the relationship between genetics and race. Nat Rev Genet 5: 598-609

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340-5

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40: 695-701

Boomsma D, Busjahn A, Peltonen L (2002) Classical twin studies and beyond. Nat Rev Genet 3: 872-82

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl: 228-37

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314-31

Burga A, Lehner B (2012) Beyond genotype to phenotype: why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience. Febs J

Burmeister M, McInnis MG, Zollner S (2008) Psychiatric genetics: progress amid controversy. Nat Rev Genet 9: 527-40

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325: 31-6

Canzian F, Kaaks R, Cox DG, Henderson KD, Henderson BE, Berg C, Bingham S, Boeing H, Buring J, Calle EE, Chanock S, Clavel-Chapelon F, Dossus L, Feigelson HS, Haiman CA, Hankinson SE, Hoover R, Hunter DJ, Isaacs C, Lenner P, Lund E, Overvad K, Palli D, Pearce CL, Quiros JR, Riboli E, Stram DO, Thomas G, Thun MJ, Trichopoulos D, van Gils CH, Ziegler RG (2009) Genetic polymorphisms of the GNRH1 and GNRHR genes and risk of breast cancer in the National Cancer Institute Breast and Prostate Cancer Cohort Consortium (BPC3). BMC Cancer 9: 257

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231-8

Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT (2011) Basic statistical analysis in genetic case-control studies. Nat Protoc 6: 121-33

Collins FS (2011) Reengineering translational science: the time is right. Sci Transl Med 3: 90cm17

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV, Zilversmit M, Cartwright R, Rouleau GA, Daly M, Stone EA, Hurles ME, Awadalla P (2011) Variation in genome-wide mutation rates within and between human families. Nat Genet 43: 712-4

Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12: 628-40

Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR, Wheeler DA, Sabo A, Lusk C, Weiss KG, Akbar H, Cree A, Hawes AC, Newsham I, Varghese RT, Villasana D, Gross S, Joshi V, Santibanez J, Morgan M, Chang K, Iv WH, Templeton AR, Boerwinkle E, Gibbs R, Sing CF (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun 1: 131

Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet 1: 40-7

Chakravarti A (1999) Population genetics--making sense out of sequence. Nat Genet 21: 56-60

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF, Jr., Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS (2007) Replicating genotype-phenotype associations. Nature 447: 655-60

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29: 229-32

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005) Efficiency and power in genetic association studies. Nat Genet 37: 1217-23

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997-1004

Dewan A, Liu M, Hartman S, Zhang SS, Liu DT, Zhao C, Tam PO, Chan WM, Lam DS, Snyder M, Barnstable C, Pang CP, Hoh J (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. Science 314: 989-92

Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21: 596-601

Diamond J (2003) The double puzzle of diabetes. Nature 423: 599-602

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8: e1000294

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446-50

Ellis PD (2010) The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge University Press

Evans DM, Visscher PM, Wray NR (2009) Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet 18: 3525-31

Eyre-Walker A (2010) Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci U S A 107 Suppl 1: 1752-6

Feldman MW, Lewontin RC (1975) The heritability hang-up. Science 190: 1163-8

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-61

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241-51

Fu J, Festen EA, Wijmenga C (2011) Multi-ethnic studies in complex traits. Hum Mol Genet 20: R206-13

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-9

Gibson G (2009) Decanalization and the origin of complex disease. Nat Rev Genet 10: 134-40

Gibson G (2012) Rare and common variants: twenty arguments. Nat Rev Genet 13: 135-45

Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245-57

Goldstein DB (2009) Common genetic variation and human traits. N Engl J Med 360: 1696-8

Goldstein DB, Chikhi L (2002) Human migrations and population structure: what we know and why it matters. Annu Rev Genomics Hum Genet 3: 129-52

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108: 11983-8

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, Hansen NF, Durand EY, Malaspinas AS, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober B, Hoffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PL, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paabo S (2010) A draft sequence of the Neandertal genome. Science 328: 710-22

Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, Frazer KA (2011) 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature 470: 264-8

Hartl CL, Clark AG (2007) Principles of Population Genetics, 4th edn. Sinauer

Hemminki K, Forsti A, Bermejo JL (2008) The 'common disease-common variant' hypothesis and familial risks. PLoS One 3: e2504

Hindorff LA, MacArthur J, (European Bioinformatics Institute), Wise A, Junkins HA, Hall PN, Klemm AK, Manolio TA A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-7

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307: 1072-9

Hirschhorn JN, Altshuler D (2002) Once and again-issues surrounding replication in genetic association studies. J Clin Endocrinol Metab 87: 4438-41

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4: 45-61

Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2: e124

Ioannidis JP, Ntzani EE, Trikalinos TA (2004) 'Racial' differences in genetic effects for complex diseases. Nat Genet 36: 1312-8

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29: 306-9

Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 5: e1000337

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29: 217-22

Jelier R, Semple JI, Garcia-Verdugo R, Lehner B (2011) Predicting phenotypic variation in yeast from individual genome sequences. Nat Genet 43: 1270-4

Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740-3

Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet 39: 1251-5

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385-9

Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, Thomas G, Hoover R, Hunter DJ, Chanock S (2009) Beyond odds ratios--communicating disease risk based on genetic profiles. Nat Rev Genet 10: 264-9

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22: 139-44

Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80: 727-39

Lachance J (2010) Disease-associated alleles in genome-wide association studies are enriched for derived low frequency alleles relative to HapMap and neutral expectations. BMC Med Genomics 3: 57

Lander ES (1996) The new genomics: global views of biology. Science 274: 536-9

Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470: 187-97

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin

Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, Konig IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-8

Lanktree MB, Guo Y, Murtaza M, Glessner JT, Bailey SD, Onland-Moret NC, Lettre G, Ongen H, Rajagopalan R, Johnson T, Shen H, Nelson CP, Klopp N, Baumert J, Padmanabhan S, Pankratz N, Pankow JS, Shah S, Taylor K, Barnard J, Peters BJ, Maloney CM, Lobmeyer MT, Stanton A, Zafarmand MH, Romaine SP, Mehta A, van Iperen EP, Gong Y, Price TS, Smith EN, Kim CE, Li YR, Asselbergs FW, Atwood LD, Bailey KM, Bhatt D, Bauer F, Behr ER, Bhangale T, Boer JM, Boehm BO, Bradfield JP, Brown M, Braund PS, Burton PR, Carty C, Chandrupatla HR, Chen W, Connell J, Dalgeorgou C, Boer A, Drenos F, Elbers CC, Fang JC, Fox CS, Frackelton EC, Fuchs B, Furlong CE, Gibson Q, Gieger C, Goel A, Grobbee DE, Hastie C, Howard PJ, Huang GH, Johnson WC, Li Q, Kleber ME, Klein BE, Klein R, Kooperberg C, Ky B, Lacroix A, Lanken P, Lathrop M, Li M, Marshall V, Melander O, Mentch FD, Meyer NJ, Monda KL, Montpetit A, Murugesan G, Nakayama K, Nondahl D, Onipinla A, Rafelt S, Newhouse SJ, Otieno FG, Patel SR, Putt ME, Rodriguez S, Safa RN, Sawyer DB, Schreiner PJ, Simpson C, Sivapalaratnam S, Srinivasan SR, Suver C, et al. (2011) Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height. Am J Hum Genet 88: 6-18

Lee SH, van der Werf JH, Hayes BJ, Goddard ME, Visscher PM (2008) Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet 4: e1000231

Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K, Buja A, Krieger A, Yoon S, Troge J, Rodgers L, Iossifov I, Wigler M (2011) Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron 70: 886-97

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-4

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD (2008) Proportionally more deleterious genetic variation in European than in African populations. Nature 451: 994-7

Lohmueller KE, Mauney MM, Reich D, Braverman JM (2006) Variants associated with common disease are not unusually differentiated in frequency across populations. Am J Hum Genet 78: 130-6

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33: 177-82

Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA (2011) Clan genomics and the complex architecture of human disease. Cell 147: 32-43

Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18-21

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 (2009) Finding the missing heritability of complex diseases. Nature 461: 747-53

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356-69

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet 7: e1001373

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321-4

Myles S, Davison D, Barrett J, Stoneking M, Timpson N (2008) Worldwide population differentiation at disease-associated SNPs. BMC Med Genomics 1: 22

Nagylaki T (1985) Homozygosity, effective number of alleles, and interdeme differentiation in subdivided populations. Proc Natl Acad Sci U S A 82: 8611-3

Neel JV (1962) Diabetes Mellitus: A "Thrifty" Genotype Rendered Detrimental by "Progress"? Am J Hum Genet 14: 353–362

Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6: e1000888

Orozco G, Barrett JC, Zeggini E (2010) Synthetic associations in the context of genome-wide association scan signals. Hum Mol Genet 19: R137-44

Pan Z, Trikalinos TA, Kavvoura FK, Lau J, Ioannidis JP (2005) Local literature bias in genetic epidemiology: an empirical evaluation of the Chinese literature. PLoS Med 2: e334

Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Jr., Chatterjee N (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S A 108: 18026-31

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42: 570-5

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001)

Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294: 1719-23

Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32: 381-5

Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bolte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Bryson SE, Carson AR, Casallo G, Casey J, Chung BH, Cochrane L, Corsello C, Crawford EL, Crossett A, Cytrynbaum C, Dawson G, de Jonge M, Delorme R, Drmic I, Duketis E, Duque F, Estes A, Farrar P, Fernandez BA, Folstein SE, Fombonne E, Freitag CM, Gilbert J, Gillberg C, Glessner JT, Goldberg J, Green A, Green J, Guter SJ, Hakonarson H, Heron EA, Hill M, Holt R, Howe JL, Hughes G, Hus V, Igliozzi R, Kim C, Klauck SM, Kolevzon A, Korvatska O, Kustanovich V, Lajonchere CM, Lamb JA, Laskawiec M, Leboyer M, Le Couteur A, Leventhal BL, Lionel AC, Liu XQ, Lord C, Lotspeich L, Lund SC, Maestrini E, Mahoney W, Mantoulan C, Marshall CR, McConachie H, McDougle CJ, McGrath J, McMahon WM, Merikangas A, Migita O, Minshew NJ, Mirza GK, Munson J, Nelson SF, Noakes C, Noor A, Nygren G, Oliveira G, Papanikolaou K, Parr JR, Parrini B, Paton T, Pickles A, Pilorge M, et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368-72

Plagnol V, Wall JD (2006) Possible ancestral structure in human populations. PLoS Genet 2: e105

Pollard TM (2008) Western Diseases: An Evolutionary Perspective. Cambridge University Press

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124-37

Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11: 2417-23

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. Genetics 155: 945-59

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67: 170-81

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748-52

Raychaudhuri S (2011) Mapping rare and common causal alleles for complex human diseases. Cell 147: 57-69

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Paabo S (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053-60

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411: 199-204

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17: 502-10

Reich DE, Schaffner SF, Daly MJ, McVean G, Mullikin JC, Higgins JM, Richter DJ, Lander ES, Altshuler D (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. Nat Genet 32: 135-42

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-7

Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M (2010) Genome-wide association studies in diverse populations. Nat Rev Genet 11: 356-66

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298: 2381-5

Rothman KJ (2002) Epidemiology: An Introduction. Oxford University Press

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole

CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-33

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331-6

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341-5

Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19: 212-9

Shriner D, Adeyemo A, Gerry NP, Herbert A, Chen G, Doumatey A, Huang H, Zhou J, Christman MF, Rotimi CN (2009) Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. PLoS One 4: e8398

Siontis KC, Patsopoulos NA, Ioannidis JP (2010) Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. Eur J Hum Genet 18: 832-7

Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, Rudan I, McKeigue P, Wilson JF, Campbell H (2011) Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet 89: 607-18

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38: 209-13

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-5

Slatkin M (2008) Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9: 477-85

Smith EN, Koller DL, Panganiban C, Szelinger S, Zhang P, Badner JA, Barrett TB, Berrettini WH, Bloss CS, Byerley W, Coryell W, Edenberg HJ, Foroud T, Gershon ES, Greenwood TA, Guo Y, Hipolito M, Keating BJ, Lawson WB, Liu C, Mahon PB, McInnis MG, McMahon FJ, McKinney R, Murray SS, Nievergelt CM, Nurnberger JI, Jr., Nwulia EA, Potash JB, Rice J, Schulze TG, Scheftner WA, Shilling PD, Zandi PP, Zollner S, Craig DW, Schork NJ, Kelsoe JR (2011) Genome-wide association of bipolar disorder suggests an enrichment of replicable associations in regions near genes. PLoS Genet 7: e1002134

Sullivan PF, Daly MJ, O'Donovan M (2012) Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat Rev Genet 13: 537-51

Terwilliger JD, Weiss KM (2003) Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. Ann Med 35: 532-44

The International HapMap 3 Consortium (2009) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-8

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299-320

The International HapMap Project (2003) The International HapMap Project. Nature 426: 789-96

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-78

Thomas DC (2004) Statistical Methods in Genetic Epidemiology, 1st edn. Oxford University Press

Timpson NJ, Lindgren CM, Weedon MN, Randall J, Ouwehand WH, Strachan DP, Rayner NW, Walker M, Hitman GA, Doney AS, Palmer CN, Morris AD, Hattersley AT, Zeggini E, Frayling TM, McCarthy MI (2009) Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. Diabetes 58: 505-10

Visscher PM (2008) Sizing up human height variation. Nat Genet 40: 489-90

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90: 7-24

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9: 255-66

Visscher PM, McEvoy B, Yang J (2010) From Galton to GWAS: quantitative genetics of human height. Genet Res (Camb) 92: 371-9

Visscher PM, McEvoy B, Yang J (2011) From Galton to GWAS: quantitative genetics of human height. Genet Res (Camb) 92: 371-9

Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF (2008) A novel DNA sequence database for analyzing human demographic history. Genome Res 18: 1354-61

Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, Maskarinec G, Monroe KR, Kolonel LN, Altshuler D, Henderson BE, Haiman CA (2010) Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. PLoS Genet 6

Weiss KM (1999) Genetic Variation and Human Disease: Principles and Evolutionary Approaches, Reprinted edn. Cambridge University Press

Weiss KM (2008) Tilting at quixotic trait loci (QTL): an evolutionary perspective on genetic causation. Genetics 179: 1741-56

Weiss KM, Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? Nat Genet 26: 151-7

Winkler CA, Nelson GW, Smith MW (2010) Admixture mapping comes of age. Annu Rev Genomics Hum Genet 11: 65-89

Wray NR, Purcell SM, Visscher PM (2011) Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol 9: e1000579

Wray NR, Visscher PM (2010) Narrowing the boundaries of the genetic architecture of schizophrenia. Schizophr Bull 36: 14-23

Wray NR, Yang J, Goddard ME, Visscher PM (2010) The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 6: e1000864

Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H (2003) A polygenic basis for late-onset disease. Trends Genet 19: 97-106

Wright A, Hastie N (2007) Genes and Common Diseases
Cambridge University Press

Wright S (1922) Coefficients of inbreeding and relationship. American Naturalist 56: 330–338

Wright S (1931) Evolution in Mendelian populations. Genetics 16: 97-159

Wright S (1969) The Theory of Gene Frequencies. The University of Chicago Press, Chicago, IL

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565-9

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes

MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM (2011) Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet 43: 519-25

Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ (2008) A navigator for human genome epidemiology. Nat Genet 40: 124-5

Zaitlen N, Pasaniuc B, Gur T, Ziv E, Halperin E (2010) Leveraging genetic variability across populations for the identification of causal variants. Am J Hum Genet 86: 23-33

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316: 1336-41

Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10: 451-81

Ziegler A, Konig IK, Pahlke F (2010) A Statistical Approach to Genetic Epidemiology, 2nd edn. Wiley

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109: 1193-8

Appendix

A.1.

Copy number variation analysis in the great apes reveals species-specific patterns of structural variation

Elodie Gazave, Fleur Darré, Carlos Morcillo-Suárez, Natalia Petit-Marty, Angel Carreño, **Urko M. Marigorta**, Oliver A. Ryder, Antoine Blancher, Mariano Rocchi, Elena Bosch, Carl Baker, Tomàs Marquès-Bonet, Evan E. Eichler, Arcadi Navarro

Genome Research 2011, **21**: 1626-1639

Gazave E, Darré F, Morcillo-Suárez C, Petit-Marty N, Carreño A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, Baker C, Marquès-Bonet T, Eichler EE, Navarro A. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome Res. 2011; **21**: 1626-1639.

A.2.

Adaptive evolution of loci covarying with the human African Pygmy phenotype

Isabel Mendizabal, **Urko M. Marigorta**, Oscar Lao, David Comas

Human Genetics 2012, **131** (8): 1305-1317

Mendizabal I, Marigorta UM, Lao O, Comas D. Adaptive evolution of loci covarying with the African Pygmy phenotype. Hum Genet. 2012; **131** (8): 1305-1317.