How important are rare variants in common disease?

Aude Saint Pierre and Emmanuelle Génin

Advance Access publication date 8 July 2014

Abstract

Genome-wide association studies have uncovered hundreds of common genetic variants involved in complex diseases. However, for most complex diseases, these common genetic variants only marginally contribute to disease susceptibility. It is now argued that rare variants located in different genes could in fact play a more important role in disease susceptibility than common variants. These rare genetic variants were not captured by genome-wide association studies using single nucleotide polymorphism-chips but with the advent of next-generation sequencing technologies, they have become detectable. It is now possible to study their contribution to common disease by resequencing samples of cases and controls or by using new genotyping exome arrays that cover rare alleles. In this review, we address the question of the contribution of rare variants in common disease by taking the examples of different diseases for which some resequencing studies have already been performed, and by summarizing the results of simulation studies conducted so far to investigate the genetic architecture of complex traits in human. So far, empirical data have not allowed the exclusion of many models except the most extreme ones involving only a small number of rare variants with large effects contributing to complex disease. To unravel the genetic architecture of complex disease, case-control data will not be sufficient, and alternative study designs need to be proposed together with methodological developments.

Keywords: *rare variants; common disease; common variants; population genetics; mutation-selection balance; next-generation sequencing*

INTRODUCTION

The genetic basis of common human diseases has been the subject of many studies. Genetic factors have been identified in most of the common diseases, but how many more genetic factors are to be found is still an open question. Technological advances in molecular biology and next-generation sequencing approaches have made it possible to explore the entire genome of individuals and not only some predefined loci where common variants have been detected. This opens up some new opportunities in the genetic research on common diseases but also raises some new challenges. Indeed, to exploit these data in the most efficient way, it is important to estimate what sort of genetic factors we should be searching for.

For several years, the paradigm that prevailed in complex disease genetic studies was the so-called 'common disease-common variants' (CDCV) paradigm, and it was believed that the genetic susceptibility to common disease was because of the genetic variants that are relatively frequent in general populations and have low penetrances [1]. It was at the opposite spectrum from rare Mendelian diseases due to very rare variants (pathogenic mutations) with very high penetrances that could only be found in affected individuals and not in the general population. To identify these common variants, genome-wide association studies have been conducted, where single nucleotide polymorphism (SNP)-chips containing probes to genotype hundred thousands of SNPs were used on very large samples of cases and

Corresponding author. Emmanuelle Génin, Inserm U1078, Génétique, Génomique fonctionnelle et Biotechnologies, 46 rue Félix Le Dantec, CS 51819, 29 218 Brest Cedex 2, France. Tel.: +33 02 98 22 34 08; +33 2 98 467 910; E-mail: Emmanuelle.genin@inserm.fr

Aude Saint Pierre is a post-doctoral fellow in statistical genetics at Inserm in Brest working on population stratification in different regions of France and its impact on rare variant association tests.

Emmanuelle Génin is a research director at Inserm in Brest. Her research interests are the development and the evaluation of methods derived from population genetics to evidence genes involved in complex diseases.

controls [2]. In several diseases, Genome-Wide Association Studies (GWAS) were successful at discovering numerous novel associations in genomic regions that were not suspected to be involved and at demonstrating the role of novel metabolic pathways in these pathologies. However, most of the common variants identified are difficult to link with the disease. Several of the associations concern SNPs that are located in noncoding regions of the genome and even when the signals detected are located in or close to genes, it is usually very difficult to identify the causal variants tagged by the associated SNPs. These common variants have limited utility in clinical practice now, as we are still awaiting for the development and testing of risk models incorporating multiple genetic and other risk factors; however, the expectation is that eventually disease risk prediction could be improved by using common variation as well [3]. For most common diseases, taken together, all the common variants found associated only explain a very limited proportion of the disease heritability. A good example is type 2 diabetes where, in 2011, 44 common susceptibility variants were known that have almost all been identified using GWASs and meta-analyses of GWASs. These 44 loci, however, only explain 10% of the familial clustering suggesting that additional genetic risk factors were likely to play a role in the disease [4]. The same observation was made for almost all common diseases and has led to the missing heritability problem [5]. Rather than missing, it was then suggested that the heritability might in fact be hidden because a much more substantial amount of heritability can be explained when taking into account all SNPs and not only the significant ones. This was first illustrated for height where the proportion of variance explained by the 180 statistically significant SNPs was only 10% but increases to 45% when considering the entire set of genotyped SNPs [6]. More recently, a similar analysis performed in type 2 diabetes found that between 49% and 63.9% of the liability-scale variance was explained by common variants across the genome [7], suggesting that the single-SNP analyses conducted so far have not identified all the common genetic variants contributing to the disease and that even larger studies will always identify new loci. Still, there is a significant proportion of the genetic variance that is unexplained by common variants and a need to look for other sources. Among the other possible culprits, rare genetic variants have been particularly highlighted in the last few years because progress made in sequencing technologies have made them more easily accessible at the scale of the entire genome and not only in candidate regions.

In this review, we will discuss the contribution of rare variants to common disease in the light of the large-scale resequencing efforts made in the last few years to understand the whole pattern of genetic variation found in human populations and how it compares with population genetic theories.

WHAT IS A RARE VARIANT?

There is no consensus in the literature on what is called a rare variant. First, the term variant is sometimes used to represent a locus and sometimes to represent an allele at a locus. For example, in Frazer *et al.* [8], the authors define a rare variant as a genetic variant with a minor allele frequency (MAF) of $<1\%$, and so they consider that the variant is the locus and that the rarity refers to the frequency of the less frequent (or minor) allele at the locus. However, in the first papers where the CDCV hypothesis was presented, the word variant was used to refer to the allele. Indeed, in his seminal paper, Lander [9] wrote 'If the genes are the human elements, the common variants are the abundant isotopes'. In the following, we will use the term variant to represent an allele at a locus.

Second, the meaning of 'rare' is also different from one study to the other. Some authors differentiate rare and common variants on the basis of a frequency threshold. For example, in Frazer *et al.* [8], a threshold of 1% is suggested with rare defining alleles with a frequency $<1\%$, whereas in Gorlov *et al.* [10], the threshold is 5%. Bodmer and Bonilla [11] also use an upper limit of 1% but suggest a lower limit of $\sim 0.1\%$ to distinguish rare variants from a third category of variants that include what they call 'clearly deleterious mutations'. The authors, however, recognized that these frequency boundaries are not absolute and that there might be some overlaps between 'low-frequency common variants' and 'high-frequency rare variants'. In their review paper, Cirulli and Goldstein [12] define four categories of variants based on their frequency: 'very common' variants with a frequency between 5 and 50%, 'less common' with a frequency between 1 and 5%, 'rare (but not private)' with a frequency of $<1\%$ but that are still polymorphic in one or more major human populations and 'private' that are restricted to

probands and immediate relatives, the clan as defined by Lupski *et al.* [13]. Cirulli and Golstein [12] also insist on the fact that this categorization is not only an academic one but also has major implications for the analysis, as different strategies will be required to evidence these different categories of variants. The 'very common' ones can be identified using SNP-chips and current GWAS, the 'less common' ones using the catalog of variants identified in the 1000 Genomes Project, the 'rare but not private' variants will be amenable to a framework of extreme phenotype resequencing and co-segregation in families and the 'private' will be very difficult to identify except perhaps through co-segregation in families for some of them. More recently, Zuk *et al.* [14] defined common variants as variants frequent enough to allow their individual testing in cases and controls and suggested that, given the actual sample sizes, common variants can include variants with frequencies down to 0.5%; i.e. variants that can be seen once in 100 individuals. Rare variants are then defined in contrast to common variants as those variants that cannot be tested individually but need to be aggregated in association tests. Rather than defining rare variants based on some arbitrary and relative frequency thresholds that would not have the same meaning in terms of power of detection based on the sample sizes, they might be defined, in contrast to the common variants, by the fact that they cannot be detected using usual association tests as conducted in GWAS. Note that in this case, the important factor is the actual count of alleles in the sample rather than the frequency as previously acknowledged by Joyce and Tavaré [15].

DIFFERENT TYPES OF RARE GENETIC VARIANTS

Genetic variants can broadly be categorized into two different classes depending on their nucleotide composition: single nucleotide variants and structural variants [8]. Single nucleotide variants are generated by changes at a single nucleotide position on the DNA sequence. This DNA position where there exist different single nucleotide variants can also be called an SNP, although some people restrict the use of SNP to DNA positions where the two alleles are frequent, typically with an MAF >1%. In most cases, single nucleotide variants are created by the replacement during replication of a nucleotide that carries a given base (Adenine, Thymine, Guanine or Cytosine) to a nucleotide that carries a different base. They therefore generate modification of neither the DNA sequence length nor the nucleotide order. This replacement is what is called a point mutation and is often unique at a given DNA position so that only two variants could be observed at any position. There are, however, reports of DNA sequence positions where more than two alleles exist, i.e. multiallelic SNPs [16]. It is difficult to estimate the fraction of SNPs that are multiallelic, but a recent query of dbSNPs limited to autosomes (query performed on 28 February 2014) found 286 623 SNPs with more than two alleles (272 888 with three alleles and 13 735 with four alleles) of 49 252 236 autosomal SNPs (0.58%). Some of these could of course be errors due, for example, to insertion in the sequence or incorrect allele calls. But, what this large number of multiallelic SNPs indicates is the fact that the 'infinite site model' that is at the basis of population genetics theories and that assumes that two different mutations are not possible at a single nucleotide position might suffer from exceptions that are not so rare.

Structural variants are basically all other kinds of DNA variations. This is thus a rather heterogeneous class of variants that includes insertion–deletion (indels) polymorphisms where one or a few nucleotides (the limit in length is not really clear in the literature) are missing in some DNA sequences, block substitutions where several adjacent nucleotides are changed, inversions where the order of the nucleotides in a given genomic region is changed and copy-number variants (CNV) that are DNA sequences of $\geq$1 kb present in variable numbers in comparison with a reference genome [17]. Note that in this definition, variant refers to the locus rather than the allele that would, in this case, be the number of copies. The same authors also make a distinction between CNV and copy-number polymorphisms by defining a copy-number polymorphism as a CNV that occurs in >1% of the population. Nevertheless, structural variants represent an important fraction of genetic variants and, like single nucleotide variants, they can be rare in the population (only shared by a few individuals and even private) or, at the opposite of the frequency spectrum, common and shared by many individuals from different populations over the world [18]. CNVs can be identified using SNP-chip data by looking at the signal intensity at the different probes. This CNV calling is done individual by individual, and thus in

theory, there is no limit on the frequency of the CNVs that are detectable, and even CNVs that are private to individuals can be detected. However, in practice, there are important limits on the CNV detection power of SNP-chip signal intensity [19], and only a small fraction of novel CNVs are in fact detectable from these kinds of data.

THE GENETIC ARCHITECTURE OF COMMON DISEASES

There are basically two opposite models of genetic architecture that have been proposed in the literature to explain the common disease susceptibility. These models differ by the number of distinct alleles involved at a given disease locus and by the frequency of these alleles. Under the CDCV model, it is assumed that the major contributors to common disease risk are genetic variants with relatively high frequency in the population and low penetrance. Whereas under the 'common disease-rare variant model' (CDRV), it is assumed that there exist plenty of rare genetic variants each with high penetrance that can be involved in common disease susceptibility. Under this latter model, there is thus an important level of heterogeneity, but the genetic variants are expected to have clearer functional impact by impairing protein function or production (see Schork *et al.* [20] for a review). These two different models involved different evolutionary scenarios. If risk alleles are frequent as under the CDCV model, then it could be either that (i) they are not evolutionarily deleterious, which would make sense for late-onset diseases that occur after reproduction age, (ii) they are favored through balancing selection because they might confer some advantage at the heterozygous state (heterozygous advantage) or through their effect on some other traits (antagonist pleiotropy) or (iii) there have been some changes in the direction of selection and, although they were neutral in the past (or even favorable) in a different environment, they have recently become medically detrimental (see, for example, the 'thrifty gene hypothesis' for obesity-related traits [21]). If, on the contrary, risk is conferred by an accumulation of rare variants, this would require that mutation rates are high enough to compensate their loss through selection and drift. Moreover, if these rare alleles have relatively high penetrance, they should be deleterious enough to impair some biological process, but not too deleterious, to ensure that individuals carrying them live long enough to express the disease.

Some early attempts to explore the contribution of rare and common variants to complex diseases were made by simulations under the mutation-selection balance framework, where novel genetic variants are continuously created by mutations and then disappear or increase in frequency under the action of random drift and selection [1, 22, 23]. The results were found to be very sensitive to population genetics modeling assumptions and in particular to the mutation rate for disease alleles and to the intensity of selection. If mutation rates are high and novel variants are slightly deleterious, then a high genetic heterogeneity is expected (corresponding to the CDRV hypothesis), whereas if mutation rates are lower and/or novel variants are neutral, less heterogeneity is expected with, at each locus, a single (or a few) predominant alleles contributing to the disease that are frequent in the population (CDCV hypothesis). The population demography is also an important parameter, as the allele frequency spectrum strongly depends on population growth rates and bottleneck events. The adequacy of the mutation-selection balance framework to model common disease genetic architecture is, however, questionable. Indeed, a basic assumption of this model is that there exist some links between the fitness of the variants and their phenotypic effects. For Mendelian diseases involving highly deleterious alleles segregating in families, this simple population genetic model where disease alleles are generated through mutation and removed by purifying selection fits well. This is, however, not the case for common diseases that show complex patterns of inheritance, incomplete penetrance, late onset and complex interactions between genes and with environmental factors and thus a connection between disease susceptibility and fitness that is not as clear. Moreover, the fitness and phenotypic effects of genetic variants that are of importance today for human diseases are those that existed in the ancestral human populations, and because they depend on environmental exposures that have changed during history, they might have been very different than the ones observed in contemporary populations [24]. The impact of positive selection on ancestral alleles that could have become detrimental under changing environmental conditions is also a matter of debate. A recent study by Ayub *et al.* [25] did not detect more signals of positive selection at 65 loci associated with type 2 diabetes than in

other regions of the genome. There was also no evidence of any enrichment in ancestral status for the risk alleles as would have been expected under the thrifty gene hypothesis [21].

AN ABUNDANCE OF RARE SLIGHTLY DELETERIOUS VARIANTS IN THE HUMAN GENOME

With the growing availability of resequencing data on various human populations, it has been possible to see how the theoretical models of common disease genetic architecture fit with real data. Using data on Human Mendelian disease causing mutations and a data set of 37 genes sequenced in 1500 individual human chromosomes, Kryukov *et al.* [26] found that 20% of new missense mutations in humans result in a loss of function, whereas 27% are effectively neutral. Thus, the remaining 53% of new missense mutations have mildly deleterious effects. Focusing only on the least frequent missense alleles (those only seen in singletons among the 1500 chromosomes), the fraction of mildly deleterious variants even increases up to 70% of the alleles that show a heterozygous fitness loss in the range 0.001–0.003. These results fit well with the CDRV hypothesis and furthermore suggest that the low allele frequency of an amino acid variant can, by itself, serves as a predictor of its functional significance. Gorlov *et al.* [27] come to a similar conclusion by interrogating HapMap Phase II/Encode and SeattleSNP data to determine whether there is an enrichment in functionally relevant SNPs among the SNPs with MAF $< 5\%$ as compared with SNPs with higher MAFs.

The 1000 Genomes Project that describes the whole genome of 1092 individuals from 14 populations worldwide has confirmed, at the genome-wide scale, these observations made on candidate gene regions of the abundance of low-frequency variants and of their enrichment in potentially functional mutations. It has also evidenced their geographic differentiation and the impact of demography with rapid population growth leading to an excess of rare variants compared with what would be expected under a model of constant population size [28]. This excess of rare variants and the shape of the allele frequency spectrum can then be used to estimate population demographic history and the rates of population expansion [29, 30]. However, recent deep resequencing efforts on very large samples of individuals suggest that population growth rates obtained on small samples are probably underestimated and that there has been an accelerating population growth in the last 2000–3000 years that led to an increased burden of rare variants [31–35]. These recent rare variants that are private to individuals or groups of closely related individuals could constitute an important reservoir of disease risk alleles as suggested by Lupski *et al.* [13] under the concept of clan genomics.

EMPIRICAL EVALUATION OF THE ROLE OF RARE VARIANTS IN COMMON DISEASES

There are many examples of rare and low-frequency variants associated with complex traits and our aim here is not to provide an exhaustive list of them (for a review of some of the relevant studies, see Table 1 in Schork *et al.* [20]). Rather, we will discuss here the results of some recently published resequencing studies on cases and controls that have addressed the issue of the contribution of rare variants in complex traits.

The first study is a large resequencing effort of 25 GWAS-identified genes for autoimmune diseases in 24 892 subjects with six autoimmune diseases and 17 019 controls [36]. A total of 2990 variants in protein-coding regions of these genes were identified among which 73.6% were novel (never reported in public databases), 97.1% had a frequency $<0.5\%$ in the controls (the frequency threshold considered by the authors for rare variants) and 68.9% were only seen in one or two individuals. These numbers are similar to what was also observed in 202 drug target genes sequenced in 14 002 individuals [34]. On these data, the authors first performed single-locus analysis with each of the seven studied phenotypes in an attempt to detect possible rare variants with strong effect that could be shared among several cases. They found some signals with some low-frequency variants, but those were in fact explained by common variants that were already identified by GWAS. Next, they consider the possibility of heterogeneity and that the susceptibility could be due to multiple rare variants within the same gene. They thus applied burden association tests that look for a difference in the load of variants both rare and predicted to be of functional impact (they considered variants with a frequency $<0.5\%$ in controls that

were annotated as nonsynonymous, premature stop or splice-site altering). They used different types of tests to combine the information on these different variants and failed to detect any significant association with any of the phenotypes. They concluded that, in the seven tested autoimmune diseases, there was little support for a significant impact of rare coding-region variants in the known risk genes. The GWAS signals previously detected in these diseases were not explained by rare variants with strong effects through synthetic associations [37]. However, it cannot be excluded that, for these diseases, rare variants located outside these GWAS regions contribute significantly to the risk. Moreover, the frequency threshold used to filter out common variants could also have an impact on the results. It would be interesting to reevaluate the association using tests that consider the overall spectrum of variation within each gene to test for a cumulative effect of rare and common variants [38]. Moreover, it is also possible that a polygenic burden of rare coding variants located in different genes is involved, and effects are not detectable at the individual gene level as recently found in schizophrenia exomes [39].

Another relevant study is a whole-exome sequencing study of 1000 cases affected by type 2 diabetes and 1000 controls from Denmark, where the authors tested for association using a wide-range of strategies from single-marker tests to gene set analysis with different allele frequency thresholds to include variants in these tests [40]. They failed to detect any significant signal after correction for multiple testing but performed some simulations based on the observed patterns of genetic variation in their data to assess the power of the gene-based association tests under different scenarios. Their simulation approach is original in that it does not fix effect sizes or allele frequencies but fixes the heritability and considers that it is equally divided among different numbers of coding variants. They found that the power to detect a particular gene effect was limited, but their study was powerful enough to detect at least one gene if rare variations in <20 genes were involved in type 2 diabetes risk. Because they were not able to evidence any such association, they concluded that low-frequency variants in a small number of genes do not explain a large amount of type 2 diabetes heritability. Their simulation model is simplistic in that it assumes that each variant equally contributes to the heritability of the trait, however, the approach calibrated on real data is interesting.

A similar idea was also recently highlighted by Agarwala *et al.* [41], who proposed an integrated simulation framework to test the fit of different disease architecture models with real data. They went even further in their approach in that they consider empirical data from different genetic studies (linkage, GWAS, polygenic score and sequencing studies). They simulated genetic variation at the population scale consistent with empirical sequencing data and considered a wide range of disease genetic models to generate phenotypes of individuals. Then, they sampled in this phenotyped population to mimic different types of genetic study (epidemiological estimates of sibling recurrence risk, linkage scans in affected sibpairs, GWAS and replication in large case-control samples and polygene score logistic regression studies). They compared the results of these *in silico* studies against the results of these different types of studies conducted so far on type 2 diabetes. More than 50 different disease models were considered that were defined by two parameters: the mutation target size T and the coupling τ between the effects of the variant on fitness and on disease. The mutational target size T is the number of nucleotides that could influence disease risk if mutated and they considered values of T ranging from 75 kb to 3.75 Mb corresponding to situations where there could be between 1 and 1500 genes involved. The coupling parameter τ was varied between tight coupling ($\tau = 1$), where variants with large effects on fitness were assumed to have large effects on disease risk and no coupling ($\tau = 0$), where these two effects were assumed to be independent. Comparing the *in silico* genetic studies produced under these different models against the empirical data for type 2 diabetes, they found that only the most extreme models could be excluded and a wide range of models was compatible from models with moderate τ, where a large part of the heritability is explained by rare alleles (CDRV hypothesis) to models with weak τ, where the contribution of rare variants is more limited (CDCV hypothesis). The results of this complex simulation study could leave the feeling that it is much ado about nothing as, at the end, we are not able to quantify the respective contribution of rare and common variants in the genetic architecture of type 2 diabetes. However, the interest of the study is to provide tools to integrate the information coming from both population genetics studies and genetic epidemiological studies. This is indeed a necessary step toward a comprehensive understanding of the genetic bases of common diseases [24].

The failures of common disease exome sequencing studies to evidence genes differentially enriched in rare coding variants among cases and controls could be due to their limited sample sizes. A recent study by Zuk *et al.* [14] suggest that samples at least as large as those that were genotyped on SNP-chips for GWAS will need to be resequenced to gain enough power to test for association with rare variants in an agnostic manner, testing all genes using gene-based tests and/or gene-set analyses. They computed the power of rare variant association tests under various scenarios of mutation rates, selection coefficients and effect sizes of alleles within genes and reached the conclusion that at least 25 000 cases would be needed in the discovery phase to reach a 90% power to detect genes that contain missense mutations associated with an increased disease risk at least 3-fold. These numbers are much larger than expected and between 10 and 25 orders of magnitude from the sizes of the samples studied by exome sequencing in the papers highlighted above. They raised the question of whether alternative strategies that rely on formal genetics should be preferred [42]. This is well illustrated by the study by Cruchaga *et al.* [43], where the role in Alzheimer's disease of rare coding variants in the phospholipase D3 gene could be evidenced by sequencing no more than 40 exomes in the exploratory phase (29 affected and 11 unaffected individuals). The key point here was the selection of the cases for the exome study. They were sampled from 14 large families containing at least four cases with late-onset Alzheimer's disease. Such multiplex families have been collected for many common diseases and have been underexploited compared with case-control data in the GWAS era [44]. This is really a pity, as they contain important information useful to gain insights into the importance of rare and common variants in complex diseases, allowing modeling of their respective contribution rather than just detection of effects difficult to link to phenotypes [45].

CONCLUSION

The question of the importance of rare variants in common diseases is central for study designs and had been raised by many investigators from different fields. We have tried in this article to summarize some of the most relevant works, and at the end, we must admit there is no definite answer to this question. We are tempted to just answer 'that depends'. It sounds a bit like the question of the required sample sizes often asked to statisticians and to which the answer could be 'the more the better' or 'tell me what you want to find and I'll tell you the minimum sample size you need'. Here, coming back to the facts, it is clear that rare variants are more abundant in the human genome than common variants. It is among them that most of the deleterious mutations would be found, and thus they are certainly worth continuing to search for. Technological advances have made them accessible, and this is good news. However, we should be careful that technology remains a tool provider for science and not a driver of science. When power computation shows that samples of 10 000 of individuals would need to be sequenced to test for association with rare variants, we should ask the question of whether there are not clever and more effective approaches that could be used.

Among the different approaches to unravel genetic variants contributing to complex traits, population studies at a fine geographic scale could be an interesting strategy. It should allow the detection of low-frequency genetic variants, which have arisen recently and are still concentrated in limited geographical areas. These variants are difficult to investigate in studies combining data from cases and controls sampled in heterogeneous populations. They may, however, play a substantial role in disease susceptibility. Efforts to detect them are currently ongoing in some regions of western France such as Brittany or Vendee where there has been limited migration in recent history. Spatial genetic epidemiology studies will be conducted to correlate clusters of genetic variants with clusters of diseases. Insights into the genetic structure of populations and the spatial distribution of disease will be gained that will allow the design of more powerful genetic epidemiology studies taking heterogeneity into account more effectively.

Key points

- Advances in sequencing technologies make it possible to resequence the whole genome or exome of large samples of individuals.
- An abundance of rare genetic variants only present in one or a few individuals are found.
- These rare variants are enriched in functional variants and mildly deleterious variants.
- They can constitute an important reservoir of disease risk alleles.
- Case-control designs might not be the most efficient design and methodological developments are required to propose alternative strategies.

FUNDING

This work was supported by a grant from the Britanny Region (dispositif SAD stratégie d'attractibilité durable-projet STATEX) and from Association Gaetan Saleun.

References

1. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001;**17**:502–10.
2. Panoutsopoulou K, Zeggini E. Finding common susceptibility variants for complex disease: past, present: future. *Brief Funct Genomics Proteomics* 2009;**8**:345–52.
3. Manolio TA. Genomewide association studies and assessment of the risk of disease. *New Eng J Med* 2010;**363**:166–76.
4. Wheeler E, Barroso I. Genome-wide association studies and type 2 diabetes. *Brief Funct Genomics* 2011;**10**:52–60.
5. Manolio TA, Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
6. Yang J, Benyamin B, McEvoy BP, *et al*. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010;**42**:565–9.
7. Morris AP, Voight BF, Teslovich TM, *et al*. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 2012;**44**:981–90.
8. Frazer KA, Murray SS, Schork NJ, *et al*. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 2009;**10**:241–51.
9. Lander ES. The new genomics: global views of biology. *Science* 1996;**274**:536–9.
10. Gorlov IP, Gorlova OY, Frazier ML, *et al*. Evolutionary evidence of the effect of rare variants on disease etiology. *Clin Genet* 2011;**79**:199–206.
11. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008; **40**:695–701.
12. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010;**11**:415–25.
13. Lupski JR, Belmont JW, Boerwinkle E, *et al*. Clan genomics and the complex architecture of human disease. *Cell* 2011; **147**:32–43.
14. Zuk O, Schaffner SF, Samocha K, *et al*. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci USA* 2014;**111**:E455–64.
15. Joyce P, Tavare S. The distribution of rare alleles. *J Math Biol* 1995;**33**:602–18.
16. Huebner C, Petermann I, Browning BL, *et al*. Triallelic single nucleotide polymorphisms and genotyping error in genetic epidemiology studies: MDR1 (ABCB1) G2677/T/A as an example. *Cancer Epidemiol Biomarkers Prev* 2007;**16**: 1185–92.
17. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* 2006;**7**:85–97.
18. Redon R, Ishikawa S, Fitch KR, *et al*. Global variation in copy number in the human genome. *Nature* 2006;**444**: 444–54.
19. Marenne G, Rodriguez-Santiago B, Closas MG, *et al*. Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish Bladder Cancer/EPICURO study. *Hum Mutat* 2011;**32**:240–8.
20. Schork NJ, Murray SS, Frazer KA, *et al*. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 2009;**19**:212–9.
21. Neel JV. Diabetes mellitus: a ''thrifty'' genotype rendered detrimental by ''progress''? *Am J Hum Genet* 1962;**14**: 353–62.
22. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001;**69**: 124–37.
23. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant . . . or not? *Hum Mol Genet* 2002;**11**:2417–23.
24. Di Rienzo A. Population genetics models of common diseases. *Curr Opin Genet Dev* 2006;**16**:630–6.
25. Ayub Q, Moutsianas L, Chen Y, *et al*. Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. *Am J Hum Genet* 2014;**94**:176–85.
26. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 2007;**80**:727–39.
27. Gorlov IP, Gorlova OY, Sunyaev SR, *et al*. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* 2008;**82**:100–12.
28. Abecasis GR, Auton A, Brooks LD, *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.
29. Gravel S, Henn BM, Gutenkunst RN, *et al*. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA* 2011;**108**:11983–8.
30. Gutenkunst RN, Hernandez RD, Williamson SH, *et al*. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 2009;**5**:e1000695.
31. Coventry A, Bull-Otterson LM, Liu X, *et al*. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* 2010;**1**:131.
32. Gazave E, Ma L, Chang D, *et al*. Neutral genomic regions refine models of recent rapid human population growth. *Proc Natl Acad Sci USA* 2014;**111**:757–62.
33. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 2012;**336**:740–3.
34. Nelson MR, Wegmann D, Ehm MG, *et al*. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 2012;**337**:100–4.
35. Tennessen JA, Bigham AW, O'Connor TD, *et al*. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 2012; **337**:64–9.
36. Hunt KA, Mistry V, Bockett NA, *et al*. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* 2013;**498**:232–5.
37. Dickson SP, Wang K, Krantz I, *et al*. Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010;**8**: e1000294.

38. Ionita-Laza I, Lee S, Makarov V, *et al*. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* 2013;**92**:841–53.
39. Purcell SM, Moran JL, Fromer M, *et al*. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 2014; **506**:185–90.
40. Lohmueller KE, Sparso T, Li Q, *et al*. Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *Am J Hum Genet* 2013;**93**:1072–86.
41. Agarwala V, Flannick J, Sunyaev S, *et al*. Evaluating empirical bounds on complex disease genetic architecture. *Nat Genet* 2013;**45**:1418–27.
42. Clerget-Darpoux F, Elston RC. Will formal genetics become dispensable? *Hum Hered* 2013;**76**:47–52.
43. Cruchaga C, Karch CM, Jin SC, *et al*. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature* 2014;**505**:550–4.
44. Bourgain C, Genin E, Cox N, *et al*. Are genome-wide association studies all that we need to dissect the genetic component of complex human diseases? *Euro J Hum Genet* 2007;**15**:260–3.
45. Clerget-Darpoux F, Elston RC. Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Hum Hered* 2007;**64**: 91–6.