Research on Rare Variants for Complex Diseases

Chee Seng Ku, National University of Singapore, Singapore
Patrik KE Magnusson, Karolinska Institutet, Stockholm, Sweden
Kee Seng Chia, National University of Singapore, Singapore
Yudi Pawitan, Karolinska Institutet, Stockholm, Sweden

Despite the notable success of genome-wide association studies (GWAS) in revealing numerous new disease-associated genes and loci, all the identified single nucleotide polymorphisms (SNPs) collectively only account for a small proportion of the heritability of complex diseases. The failure of GWAS-SNPs to account most heritability of complex diseases has invoked some considerable discussion on the reasons for the missing heritability. One of the factors likely to contribute to the missing heritability is the rare variants, which have not been well examined for disease association in GWAS. However, currently the tools and approaches are readily available to investigate rare variants.

Introduction

Over the past five years, immense success have been achieved in dissecting the genetic basis of complex diseases by genome-wide association studies (GWAS) (A Catalogue of Published Genome-Wide Association Studies http://www.genome.gov/26525384). However, although nearly 3000 single nucleotide polymorphisms (SNPs) have been reported to be associated with various human complex diseases and traits, they are most likely the surrogate markers which are in strong linkage disequilibrium with disease variants (Hindorff et al., 2009). The disease variants in most GWAS-loci remain to be uncovered and the surrogate markers need not necessarily be tagging other SNPs, as the disease variants could also be in the form of indels, copy number variants (CNVs) or other non-SNP variants. This was well demonstrated in the discovery of the 20-kb deletion located immediately upstream of the IRGM (immunity-related GTPase family, M) gene for Crohn disease (McCarroll et al., 2008). In addition, studies have also found evidence of correlations of about 30 GWAS-SNPs with CNVs at correlation coefficient ($r^2 > 0.5$) suggesting the possible associations of CNVs with various human complex diseases and traits (Conrad et al., 2010).

The genome-wide study performed by Wellcome Trust Case Control Consortium (WTCCC) investigating the association between $\sim$3400 CNVs and 8 common diseases in 19,000 samples, however, did not find novel discoveries (Wellcome Trust Case Control Consortium, 2010). See also: Copy Number Variation in the Human Genome; Genetic Variation: Human; Relevance of Copy Number Variation to Human Genetic Disease

Despite the notable success of GWAS in revealing numerous new disease-associated genes and loci, the results have been disappointing in at least one aspect, all the GWAS-SNPs collectively only account for a small proportion of the heritability of complex diseases (Maher, 2008). This is because most GWAS-SNPs confer small effect sizes (odds ratio, OR $< 1.5$). As a result, most heritability of complex diseases is still unexplained, even though several diseases are claimed to be well-studied by GWAS and meta-analysis of sufficiently large sample sizes. For example, all the 18 SNPs that were identified for type-2 diabetes cumulatively only account for $\sim 6\%$ of its heritability, and for Crohn disease that is only about 20%, even though more than 30 SNPs have been found (Manolio et al., 2009). This unexplained heritability is commonly known as 'missing heritability' and it has been described as the ‘dark matter’ in the genetics of complex diseases. Currently, the validity of common disease common variant (CD/CV) hypothesis has been challenged because of the missing heritability. But it is important to bear in mind that other common non-SNP variants like indels and CNVs have not been fully examined for disease associations, therefore, how much the heritability of complex diseases can be explained by common variants is still an unanswered question.

ELS subject area: Genetics and Disease

How to cite:
Ku, Chee Seng; Magnusson, Patrik KE; Chia, Kee Seng; and Pawitan, Yudi (September 2010) Research on Rare Variants for Complex Diseases. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0022658
question. See also: Genome-Wide Association Studies; Genome-wide Association Studies: The Success, Failure and Future

The failure of GWAS-SNPs to account most heritability of complex diseases has invoked some considerable discussion on the reasons for the missing heritability and the strategies for finding the remaining disease variants (Eichler et al., 2010). In fact, one of the factors likely to contribute to the missing heritability is the rare variants, which have not been well examined for disease association in GWAS. Additionally, we have also generally ignored gene–gene and gene–environmental interactions. They are likely to contribute some fraction of the heritability, although they are much harder to detect than main effects. The roles of rare variants in complex diseases have been increasingly recognised (Bodmer and Bonilla, 2008). The rare variants refer to the ‘rarer’ or less common SNPs (minor allele frequency, MAF < 5%) and the rare single nucleotide substitutions that are not classified as SNPs (MAF < 1%). The rare variants are not limited to SNPs but also encompass other non-SNP genetic variants like rare CNVs. Here we focus on the discussion of the strategies to find rarer SNPs and the issues and problems in the analysis for disease associations. We also give an overview of the development of technologies which have enabled research on rare variants and the emerging data supporting their associations with complex diseases.

Total Number of Disease Variants

The genetic architecture refers to the total number of genetic variants accounting for all the heritability of a complex disease and their allele frequencies and effect sizes. The number of disease variants that is needed to explain the totality of heritability depends on their allele frequencies and effect sizes. For example, the common SNP in complement factor H (CFH) with an odds ratio 4.6 (per risk allele) has explained a significant proportion of the heritability of age-related macular degeneration (Klein et al., 2005). And so far, the five confirmed loci for the eye disease have already accounted 50% of its heritability. In contrast, for about the same number of confirmed loci for systemic lupus erythematosus, only 15% of its heritability was accounted (Manolio et al., 2009). As such, this clearly shows that different diseases have different profiles and patterns in their genetic architecture (Evans et al., 2009).

The relationship between the number of disease variants, allele frequencies and effect sizes has been demonstrated in our simulation study to predict the number of genetic variants that remain to be discovered. We used type-2 diabetes as an example and assumed a similar allele frequencies and effect sizes of the confirmed SNPs that have been identified for the disease, at the time of our analysis. Our simulation analysis estimated that a total of 812 ‘common-low effect’ variants will be needed to account for all heritability of the disease. However, only about 80 variants are required if they are ‘rare-high effect’ variants. The MAF of ‘common-low’ variants is defined from 7.3% to 49.9% and the odds ratio is within the range 1.05–1.15, whereas for ‘rare-high’ variants, their MAF is defined from 0.73% to 4.99% and the range for odds ratio is 1.63–4.05. Here, the ‘disease variant’ is not specifically referred to SNPs; it could also be other forms of genetic variants (Pawitan et al., 2009).

Currently, empirical data show both the common and rare variants are contributing to the heritability of diseases (Schork et al., 2009). Since common SNPs have already been well-studied in GWAS but still unable to account most heritability of complex diseases, so it is logical that the focus of research is now directed towards rarer SNPs with potentially larger effect sizes. The number of rarer variants conferring larger effect sizes that is needed to explain most heritability of complex diseases is more manageable for functional studies to investigate their biological implications. Furthermore, the number also looks more practical for diagnostic test and disease risk prediction model.

Meta-analysis of GWAS Data

Now it is clear that, from the results of GWAS, the number of common SNPs having large effect sizes is limited, and the common SNPs conferring small to moderate effect sizes (OR < 1.5) have been shown to explain only a small proportion of heritability of complex diseases. Thus it is believed that rare variants with larger effect sizes are potentially contributing to the missing heritability. However, the rare variants are not easy to be detected and studied using current genotyping arrays which mainly focus on common SNPs. In addition, the statistical power of a single GWAS is also inadequate to capture them. Knowing these constraints, thus it is obvious that the subsequent step is to expand the content of genotyping arrays to cover rarer SNPs and to enhance the statistical powers of the studies to detect them. The timely arrival of next generation genotyping arrays leveraging the SNPs data from the 1000 Genomes Project has made the interrogation of less common SNPs technically feasible. This is discussed in a more detail in the subsequent section.

However, in terms of increasing the statistical power to capture less common SNPs, a rather straightforward way is to combine several GWAS datasets investigating the same diseases into a meta-analysis. Merging the SNP data of several GWAS that employed different genotyping arrays have been possible using genotype imputation methods (Li et al., 2009). Imputation uncertainty has been a problem in meta-analysis, but a recent study has addressed how this factor can be incorporated into the association analysis and subsequent meta-analysis (de Bakker et al., 2010).

Meta-analysis is a promising strategy to detect associations of less common SNPs with larger effect sizes. This has been well demonstrated in several studies. For example, in the meta-analysis of GWAS for multiple sclerosis, a less common SNP (risk allele frequency of 0.02 in controls) in TNFRSF1A (tumor necrosis factor receptor super family,
member 1A) with an odds ratio of 1.6 was found (De Jager et al., 2009). Similarly, Barrett et al. (2008) also found a rarer SNP in the region containing LRKK2 (leucine-rich repeat kinase 2) and MUC19 (mucin 19, oligomeric) for Crohn disease, with an odds ratio ~1.5 and the risk allele frequency in controls as 0.017. Other studies including the finding of two less common SNPs in 10q21.2 (ANK3 (ankyrin 3, node of Ranvier (ankyrin G)) gene) and 14q13.1 carrying an odds ratio >1.4 for bipolar disorder (Ferreira et al., 2008). So, these studies have confirmed that the less common disease-associated SNPs can also be discovered if they are included in the meta-analysis achieving adequate statistical power. Therefore, for diseases that have already been studied by a number of GWAS, meta-analysis is the next natural step to proceed in an attempt to capture rarer SNPs.

**Next Generation Genotyping Arrays**

GWAS have almost reached their limit of discovery for common SNPs using current genotyping arrays, which mainly focus on SNPs with MAF >5%. Therefore, expanding the content to cover less common SNPs, indels and CNVs is urgently needed to make new discoveries in future GWAS. In fact, a series of next generation genotyping arrays (HumanOmni2.5 and Omni5.0) and supplementary arrays (HumanOmni1S and Omni2.5S) will be launched by Illumina in 2010 (please see Illumina Whole-Genome Genotyping Product Roadmap, http://www.illumina.com/applications/gwas/ilmn); these arrays genotype from 2.5 to 5.0 million SNPs. The HumanOmni2.5 arrays have already been commercially marketed. The additional SNPs in the next generation genotyping arrays are chosen from the 1000 Genomes Project. The Omni2.5 array will mainly target for the SNPs that fall within the range 2.5–5%, whereas the Omni5.0 array will increase the coverage of SNPs down to 1%. These tools will certainly increase the ability of researchers to open the ‘black-box’ containing less common SNPs. Although the next generation genotyping arrays have expanded their content targeting SNPs down to 1%, single nucleotide variants (<1%) are still beyond their detection range. Since the next generation genotyping arrays have expanded the coverage of SNPs with MAF of 1–5%, so the missing heritability is expected to be increasingly accounted. See also: High-Throughput Single Nucleotide Polymorphisms Genotyping Technologies; Next Generation Sequencing Technologies and Their Applications; Whole Genome Resequencing and 1000 Genomes Project

However, several considerations have to be kept in mind. The SNPs from the lower spectrum of allele frequency are considered as ‘rarer SNPs’, thus to what extent they are shared between populations is not yet clear. The rarer variants are usually of recent origin, as they are likely to have occurred in populations after the human migration out of Africa. As a result, rarer variants are more likely to be population or geographical-specific, and this has been supported by the data from 1000 Genomes Project. It was found that most (~8 million) of the ~9 million new SNPs identified in the pilot phase of 1000 Genomes Project are specific to one HapMap population (Via et al., 2010). Therefore, it is expected that a considerable fraction of the added SNP content in the Illumina® 2.5- and 5.0-million arrays might be nonpolymorphic in some populations; as such these SNPs are not informative in GWAS analysis. Probably this problem can be overcome by customising the SNP content in genotyping arrays for each population, or it should be partially solved with the development of semi-custom arrays. This will allow researchers to choose additional rarer SNPs specific in their populations using the data produced by the 1000 Genomes Project. Since the next generation genotyping arrays have not been widely evaluated in diverse populations (beyond the HapMap populations), so to what extent that the SNPs are not polymorphic in other populations is unclear at this point.

The success of finding the associations of rarer SNPs using next generation genotyping arrays is also dependent on the adequate statistical power. The next phase of GWAS will require a much larger sample size than the current standard to ensure that the less common SNPs are observed in sufficient number of cases and controls. Furthermore, an adequate number of the minor alleles (in homozygous and heterozygous states) for rarer SNPs are also needed for genotype calling, which is based on the three genotype-intensity clusters. More importantly, sample sizes have to be large enough to achieve the required statistical power for multiple-testing correction of 2.5–5 million SNPs.

**Target Sequencing**

The next generation of genotyping array is a promising tool to interrogate less common SNPs in GWAS, but several issues exist, such as the concern of population-specific rarer SNPs and genotype calling if the sample size constraint does not give sufficient number of homozygotes and heterozygotes of the minor allele for clustering. Thus, it is obvious that a more robust method to these problems will be needed. Sequencing approach will overcome the aforementioned problems faced by the genotyping arrays. This is because sequencing is a more efficient method to detect less common SNPs or single nucleotide variants (<1%) regardless of population.

Although sequencing technologies are advancing rapidly and the cost is declining, the whole genome sequencing approach is still prohibitively expensive to be applied in a large sample set for genetic association studies (Mardis, 2008). In contrast, targeted sequencing strategy is more feasible and affordable, and it is being advocated in genetics community. The targeted regions for sequencing are usually the exome (all exons in the human genome) (Ng et al., 2009), the loci or genes identified in GWAS (Nejentsev et al., 2009) and selections of biologically plausible genes and pathways for the diseases (Ahituv et al., 2007).
Sequencing of the biological plausible genes and disease pathways is probably more challenging, because it is driven by a priori hypotheses, whereas the patho-physiology of most complex diseases is still poorly understood. Nevertheless, there have been some studies that sequence genes associated with monogenic forms of obesity in humans and mice, as well as genes involved in body-weight-related pathways (Ahituv et al., 2007). A candidate gene-based sequencing study also found a rare nonsynonymous SNP in ANGPTL4 (angiopoietin-like 4) (with 1.3% frequency in European Americans) associated with low plasma levels of triglycerides (Romeo et al., 2007). These early studies have shown that sequencing of biological plausible candidate genes is a promising approach for rare variant discoveries. These studies used Sanger sequencing and the low-throughput of that method is inadequate for large-scale studies like exome sequencing.

Nowadays, large-scale targeted sequencing studies have become technically feasible with the development of several high-throughput sequence capture or enrichment methods, which enabled the isolation of tens of megabase genomic regions in one experiment (Turner et al., 2009). These sequence-capture methods are commercially available, for example, the NimbleGen Sequence Capture Arrays and SureSelect Target Enrichment Systems by Roche® and Agilent Technologies® respectively. The high-throughput sequence-capture methods coupled with next generation sequencing technologies have also made the large-scale targeted sequencing studies affordable. Furthermore, barcoding of the samples allows hundreds of samples to be sequenced per instrument run in next generation sequencers. In addition to the higher sample throughput, this barcoding system also avoids over-sequencing the samples that only require sequencing of specific regions that comprised of several hundred megabases instead of the whole genome. Facilitated by these technological developments, the number of targeted sequencing studies is expected to increase considerably in the coming years.

Sequencing of GWAS-loci

GWAS have identified thousands of genetic loci containing common SNPs for complex diseases, but one intriguing question is to what extent these loci harbour rarer variants that are independently associated with the diseases. It has been assumed that the common SNPs identified by GWAS should be tagging for common disease variants because they are in strong linkage disequilibrium (high $r^2$ value). However, this is not necessarily the case; empirical data have shown that several rare variants can be responsible for GWAS association signals. In addition, the study also found that rare variants are likely to account for many association signals identified by GWAS because they can easily induce synthetic associations that are credited to common SNPs (Dickson et al., 2010). Thus targeted sequencing to further interrogate the GWAS-loci for rare variants is highly recommended and the success of this strategy has been demonstrated.

One notable example is the finding of several rarer SNPs (allele frequency <3%) in IFIH1 (interferon induced with helicase C domain 1) lowering the risk of type-1 diabetes (Nejentsev et al., 2009). In the study, they resequenced 144 target regions covering exons and regulatory sequences of 10 genes. Some of the genes contain common type-1 diabetes-associated polymorphisms. These less common SNPs are associated with the disease independent of each other and of the common SNPs. However, it is also noteworthy that the study did not find associations of rarer SNPs on the other genes containing common type-1 diabetes-associated SNPs. In any case, this study has demonstrated the ability of sequencing approaches to harvest those rarer SNPs which were omitted from the GWAS screening. The research on resequencing of GWAS-loci is still limited, thus more studies will be needed in the near future to investigate to what extent the usefulness of this strategy is enhancing the studies of rare variants.

Exome resequencing

Sequencing of the protein coding regions is an appealing idea because most rare missense alleles are deleterious in human (Kryukov et al., 2007). In addition, the MAF distribution of possibly damaging nonsynonymous SNPs was found to be shifted towards rare SNPs in comparison to the benign and synonymous SNPs that are not likely to be functional. There is also an inverse relationship between MAF and the proportion of nonsynonymous SNPs predicted to be protein disturbing (Gorlov et al., 2008). This suggests that rare variants are more likely to be functional than common variants. As such, these studies have directly provided strong evidence supporting the exome sequencing approach to detect rarer SNPs of functional importance. This will also expedite the research in identifying causal SNPs.

Proper selection of samples for exome sequencing plays a key role in enhancing the power to detect disease- or trait-associated variants, for example, choosing samples from the extreme ends (such as from the top and bottom 10 percentiles) of a quantitative trait such as body mass index, blood pressure and other continuous traits. For example, one study has sequenced the coding exons and splice junctions of 58 genes in several hundreds of extremely obese individuals, compared to lean subjects and found multiple rare alleles contribute to obesity in the population (Ahituv et al., 2007). Similarly, careful selection of cases with early onset of disease or multiple-affected family members will enrich for genetic risk factors to be detected.

So far, exome sequencing studies on hundreds or thousands of samples are not yet done for complex diseases, but a number of such studies have been performed for Mendelian disorders in a small sample set. For example, a sequencing study of 12 human exomes has demonstrated the sensitivity and specificity of the exome sequencing approach in identifying rare and common variants in the coding sequences (Ng et al., 2009). In addition, they also showed that the candidate genes for Mendelian disorders can be identified by exome sequencing in a small number of
unrelated cases. In fact, the number of exome sequencing studies to uncover the genetic abnormalities and mutations of Mendelian disorders has been increasingly published (Lalonde et al., 2010; Ng et al., 2010). Similarly, the International Cancer Genome Consortium (ICGC) also adopted the exome sequencing approach to decipher the cancer genome.

It is likely that there will be a transient era for targeted sequencing studies instead of a direct leap into whole genome sequencing. In terms of cost, exome sequencing is also more affordable than whole genome sequencing. Undeniably whole genome sequencing will become a highly used technique, but this will not happen until the price of sequencing is cost-effective enough, and the strategies and analysis protocols and tools are in place to process the huge volume of sequencing data for genetic variants detection and subsequent association studies.

Issues and Problems in Analysis of Rare Variants

The strategy of sequencing of a subset of samples and then genotyping the discovered rare variants in additional cases and controls is commonly adopted in targeted sequencing studies. For example, Knight and colleagues resequenced the exons of ABCA13 (ATP-binding cassette, sub-family A (ABC1), member 13) gene in 100 schizophrenia cases and 100 controls and multiple rare coding variants were identified. These rare variants were then genotyped in a larger sample set of several hundred cases and controls (Knight et al., 2009). Currently there is an increased trend in targeted sequencing studies for other complex diseases in addition to cancers (Haller et al., 2009; Prickett et al., 2009). Therefore, the studies need to be carefully designed to find genuine associations of rarer SNPs. In addition, several critical issues in the analysis of rare variants have to be carefully addressed.

Published data have shown that if only cases are sequenced and then the identified rare variants are genotyped in controls or the vice versa, this can lead to an increase in type I error. Similarly, the type I error may also be inflated where different proportions of cases and controls are sequenced and the remaining samples are genotyped. However, the inflation of type I error can be overcome if the samples that are used for rare variant discovery are not included in the subsequent association study (Li and Leal, 2009). So, this implies that a different set of samples should be used for rare variant discovery and association study. Additionally, other factors such as the initial sample size to be sequenced, variant frequencies and the number of variants within the genomic region also determine whether an inflation of type I error will occur. Therefore, many factors have to be considered carefully to avoid excessive false-positive associations, and proper

Figure 1 Strategies for finding disease-associated SNPs or variants.
design of sequencing studies is a critical factor determining the success of finding the disease-associated rare variants (Li and Leal, 2009). In addition, several new methods for detecting associations with rare variants using sequencing data have also been developed (Li and Leal, 2008; Morris and Zeggini, 2010).

**Summary and Future Perspectives**

The failure of GWAS-SNPs to account for most of the heritability of complex diseases has diverted the effort to investigate rare variants. With the arrival of next generation genotyping arrays targeting less common SNPs down to 1% allele frequency, the time is now ripe to open the ‘black-box’ containing such variants. At the same time, the targeted sequencing approaches have also become technically feasible and affordable for hundreds to thousands of samples. This is credited to the development of sequence-capture methods and high-throughput sequencing technologies. The methods of analysis are being developed, and the issues and problems in analysis of rare variants using sequencing data have also been increasingly recognised and addressed. In addition to SNPs, the current genotyping arrays have already been shown to be powerful for studying rare CNVs in complex diseases. These studies have produced exciting results for a number of diseases like autism, schizophrenia and obesity (Sebat et al., 2007; Walsh et al., 2008; Walters et al., 2010). Hence, the increased marker density and resolution in next generation genotyping arrays certainly will also enhance the research on rare CNVs. Thanks to the developments of cutting-edge technologies researchers are now entering into a new era of GWAS.

The various strategies (Figure 1) for research on rare variants that we have outlined and discussed can be executed at the same time. For diseases where multiple GWAS have been performed, meta-analysis should be carried out to enhance the statistical power to detect less common SNPs. Further genotyping on the samples that have already been interrogated by first generation genotyping arrays is worth considering. This step will increase the coverage of less common SNPs and this additional genotyping work can be easily completed using the supplementary Human Omni1S and Omni2.5S arrays. In addition, targeted sequencing should also be performed on the GWAS-loci to interrogate to what extent the GWAS association signals can be accounted by multiple rare variants. Although exome sequencing studies have not been done yet for complex diseases, sequencing of the coding regions will enable us to detect rare deleterious nonsynonymous SNPs and causal variants.

**References**


Walsh T, McClellan JM, McCarthy SE et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320: 539–543.


Further Reading

