A guide on gene prioritization in studies of psychiatric disorders

SVEN STRINGER,^{1,2*} KIM C. CERRONE,^{1*} WIM VAN DEN BRINK,¹ JULIA F. VAN DEN BERG,^{3,4}
DAMIAAN DENYS,¹ RENE S. KAHN² & ESKE M. DERKS¹

1 Department of Psychiatry, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

2 Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

3 Parnassia Psychiatric Institute, The Hague, The Netherlands

4 Department of Clinical Psychology, Leiden University, Leiden, The Netherlands

Key words

guide, gene prioritization,
psychiatry, genetics

Correspondence

Eske M. Derks, Professor of
Genetics in Psychiatry, AMC-
APC, room PA1-179,
Meibergdreef 5, 1105 AZ,
Amsterdam, The Netherlands.
Email: e.m.derks@amc.uva.nl

*These authors contributed
equally to this work

Received 10 April 2015;
revised 23 June 2015;
accepted 2 July 2015

Abstract

There has been an increasing interest in the identification of genetic variants causing individual differences in human behavior. Psychiatrists have contributed to the genetics field by defining the most important behavioral characteristics and by studying the association between genetic variants and behavioral differences within phenotypically well-characterized samples in which detailed assessments have been collected (e.g. neuroimaging). These samples are typically limited in size and are therefore not suitable for a genome-wide association analysis. Instead, gene association studies conducted in such samples typically focus on a few genes of interest, allowing smaller sample sizes. However, the selection of high-priority genes is not always straightforward and psychiatrists will usually have a limited background in genetics. We aim to fill this gap by (i) providing a basic introduction to genetics; (ii) showing how the selection of genes of interest can be optimized by the use of two web tools: Polysearch and Gene Prospector; (iii) illustrating how statistical power analyses can be performed and discussing the importance of sufficiently powered studies. This guide can help psychiatrists with limited experience in genetics in designing genetic studies that allow identification of specific behavioral, cognitive, or neural correlates of genetic risk variants, while avoiding common pitfalls. Copyright © 2015 John Wiley & Sons, Ltd.

General introduction

Since the early 1990s, there has been a rapidly increasing interest in the identification of genetic variants causing individual differences in human behavior. Genetic research has moved forward quickly with initiatives such as the international HapMap project ([http://hapmap.ncbi.nlm.nih.](http://hapmap.ncbi.nlm.nih.gov/)

<http://hapmap.ncbi.nlm.nih.gov/>) (Gibbs *et al.*, 2003) describing the common patterns of human DNA sequence variation and the 1000 genomes project (<http://www.1000genomes.org/>) (1000 Genomes Project Consortium, 2012) providing a map of both common and rare genetic variations. The costs of genotyping have decreased substantially over the years, making genetic research accessible for many researchers, including those

for whom genetics is not their primary expertise. For example, psychiatrists have aimed to investigate which genetic variants are associated with complex traits, including drug and alcohol dependence, schizophrenia, and depression. This type of research is important to obtain more insight into the etiological nature of behavioral differences across subjects and to understand why some individuals are more vulnerable to suffer from mental disorders than others. The specific behavioral characteristics of an individual, caused by the expression of genes, the influence of environmental factors, and the interactions between the two, is referred to as the individual's phenotype. Even though recent studies have led to the identification of significantly associated genetic loci, the contribution of psychiatrists to the genetics field will be important to increase the understanding of the precise nature of the specific behavioral, cognitive, or neural correlates of identified genetic risk variants.

This review provides a guideline to researchers who wish to incorporate genetics into their studies but do not have a formal background in this topic. We will first introduce some basic concepts used in genetics (next section). As explained more extensively in the third section, we will focus specifically on studies in which relatively few genes are being investigated. Genetic studies in the psychiatric field are usually small scale, whereas large sample sizes are required for genome-wide association studies (GWAS). In the fourth section we will illustrate how the vast amount of information that is available in public databases can be used to prioritize genes. Text mining software has been developed to facilitate searching through the enormous literature. Other tools provide access to findings from previously conducted genetic studies or prioritize genes based on known protein–protein interactions. However, these tools are usually presented in great technical detail in dedicated bioinformatics journals and may therefore be less accessible to non-geneticists. Furthermore, because of the high speed with which novel tools are introduced, it is a challenge to select tools that are useful for a specific purpose. We will therefore provide a practical guide and some illustrations to facilitate the use of such tools in the Supporting Information of this article. Designing a genetic study also requires considerations about sample size in order to have sufficient statistical power. Therefore, we explain in the fifth section how genetic power calculations can be performed; this section includes illustrations and practical examples. When one or multiple genes of interest have been identified, the next step is the selection of genetic variants (i.e. polymorphisms) as will be discussed in the sixth section. In the final section we further discuss the

implications of performing genetic studies in the post GWAS era. To improve the readability of this article, practical guidelines for using the web tools discussed in this article will be provided in Supporting Information available online.

DNA: the basis of heredity

Deoxyribonucleic acid (DNA) is our genetic material. The complete set of DNA is called the genome and consists of 23 pairs of chromosomes inside the nuclei of somatic cells of the human body (Figure 1). DNA is built up out of four different molecules, called nucleotides, composed of the bases adenine (A), cytosine (C), thymine (T) and guanine (G) (Brown, 2002). DNA is structured in a double helix. The two strands are paired in such a way that adenine always pairs with thymine (A-T) and guanine always pairs with cytosine (G-C), forming base pairs. The two strands are complementary to each other and thus contain essentially the same information. Consequently, when determining a person's genotype it is only necessary to investigate the base-pair sequence at a single strand. Furthermore, genes are regions of DNA that encode information needed for the production of proteins. To decode this genetic information into proteins, DNA is first transcribed into messenger ribonucleic acid (mRNA), which is subsequently translated into proteins with the help of transfer RNA (tRNA) (Figure 2). Proteins consist of amino acids and are the working forces of the cells that create and break down the molecules in an organism. The nucleotides in the DNA direct the synthesis of proteins by dictating the sequence of amino acids of which the protein is composed (Brown, 2002). Three nucleotides code for one specific amino acid. Only 1% of the genome is translated into protein; this is called the exome. The remaining part of the genome does not directly code for proteins, but is likely to be relevant in regulating the transcription of the coding parts of DNA into RNA and protein molecules (ENCODE Project Consortium, 2012).

Genetic polymorphisms

Genetic studies typically aim to identify genetic variants that contribute to phenotypic differences between individuals. Therefore, these studies focus on locations on the DNA at which the sequence of base-pairs is different across individuals. These genetic variants are called polymorphisms and the possible variants of a polymorphism are referred to as alleles. See Figure 3 for an example of a single nucleotide polymorphism (SNP). The frequency at which alleles occur in a population is often represented as a function of the frequency of the least common, or minor allele (minor allele frequency; MAF), which is

Figure 1. Organization of deoxyribonucleic acid (DNA) in the nucleus of the cell. Reprinted with permission from Darryl Leja (2010a).

one of the primary characteristics of polymorphisms. In the literature, polymorphisms are usually distinguished based on their MAF and are referred to as rare ($MAF < 0.01$), low-frequency ($0.01 < MAF < 0.05$) or common ($MAF > 0.05$) variants.

Genetic markers and the concept of linkage disequilibrium (LD)

Two alleles (i.e. variants of polymorphisms) which are located at different positions at the same chromosome, are in linkage disequilibrium (LD) if they are not inherited independently but are often transmitted together. Generally, alleles which are located close together at the same chromosome will show stronger LD. LD allows us to obtain information about the role of causal variants across

the whole genome by genotyping a relatively small subset of the polymorphisms, called genetic markers (Neale *et al.*, 2008). A marker is a genetic polymorphism that is known to have multiple alleles in the population, but is not necessarily directly associated with the expression of a phenotype. Despite the lack of a direct association, it may be indirectly associated with a phenotype if the marker is transmitted together with a causal genetic variant, thus being in LD. Therefore, these markers can be thought of as little flags on the genome that mark the spots at which a causal genetic variant (i.e. a variant that is directly associated with the expression of a phenotype) can be present. Genetic markers are crucial in genetic linkage and association studies, since they allow for the detection of causal genetic variants (i.e. trait loci) which are in LD with the marker (Neale *et al.*, 2008). In GWAS the

Figure 2. The processes of transcription and translation. This figure shows the process of transcription, from deoxyribonucleic acid (DNA) to mature messenger ribonucleic acid (mRNA). And outside the nuclear membrane of the cell in the cytoplasm, the process of translation occurs where proteins are built from amino acids with the help of transfer RNA (tRNA) molecules that are guided by ribosomes. Reprinted with permission from Darryl Leja (2010b).

association between millions of single nucleotide polymorphisms and the phenotype of interest is investigated. Because these SNPs are in LD with nearby genetic variants that are not captured by the GWAS array, these unobserved genetic variants may still be detected through statistical association. Therefore, a GWAS tests for the involvement of the vast majority of the low-frequency and common genetic variants throughout the genome.

Population stratification

Population stratification is the presence of a systematic difference in allele frequencies between population subgroups with different ethnic backgrounds. To avoid spurious associations it is important to carefully match cases and controls based on ethnicity. Ethnic diversity may be large (e.g. Caucasian versus Asian), but more subtle

Figure 3. Schematic representation of a single nucleotide polymorphism (SNP) on two pieces of deoxyribonucleic acid (DNA).

differences exist and population stratification has also shown to be present within Caucasian populations in Europe (Lao *et al.*, 2008). In GWAS information from millions of SNPs are available. It is therefore relatively straightforward to correct for population stratification by studying allele differences between subpopulations. In genetic studies in which few genetic variants have been genotyped, it is more difficult to control for ethnic differences. Therefore ethnic background should be carefully assessed, for example through registration of ethnicity of participants and their grandparents and analyses should be based on an ethnically homogeneous group.

Gene prioritization studies in the genome-wide association era

Genome-wide association is currently the most common approach for the search of novel genetic variants, although other genome-wide approaches, including whole exome sequencing (i.e. genotyping all bases in the coding regions of the genome) and whole genome sequencing (i.e. genotyping all bases of the entire genome) are becoming increasingly popular. Genome-wide association is a study design in which associations between a disease of interest and millions of common variants across the whole genome are tested. The GWAS catalog (<http://www.genome.gov/gwastudies/>) archives the results from GWAS that include at least 100,000 SNPs which are associated with a disorder at a significance level of $p < 1.0 \times 10^{-5}$ (Welter *et al.*, 2014).

Despite the popularity of GWAS, we argue that studies that are designed to investigate the role of a more limited number of genetic variants, are still important. Such studies are not only useful for the identification of hypothesis-driven genetic associations, but can also be applied to find specific phenotypic characteristics associated with novel genetic variants identified in GWAS. The primary strength of more focused studies is that they can be conducted in relatively small samples. Because of the strict correction for multiple testing that is required to avoid false-positive findings, GWAS are not the preferred design for small-scaled studies. To correctly account for the large number of statistical tests in genome-wide studies, a significance level of 5×10^{-8} was shown to effectively control type-I error rate (Dudbridge and Gusnanto, 2008; Hoggart *et al.*, 2008). Therefore multiple testing correction results in low statistical power to detect true effects. Using a hypothesis-driven approach by focusing on a few genes of interest limits the number of statistical tests and allows for a less stringent multiple testing correction. Furthermore, even though GWAS have resulted in the detection

of genetic variants that are statistically associated with a disease, the biological role of these novel variants is often unknown. Genetic risk factors are often located in intergenic regions and may be associated with a disease through regulation of gene expression levels (Niculescu, 2013). It is therefore not trivial to discover the functional role of a genetic variant that is statistically associated with a disorder. Genetic studies conducted in small-scaled samples in which extensive phenotypic information has been collected can contribute to elucidating the underlying biological processes. In the following sections we provide guidelines for gene prioritization and will show how to avoid potential pitfalls when conducting a genetic study.

Gene prioritization

As previously mentioned, selecting genes of interest may be guided by the use of web tools that combine information from GWAS, reviews, meta-analyses, and other genetic studies testing the association between the trait of interest and specific genetic polymorphisms. The following section provides a brief description of two web tools that facilitate the selection of genes of interest. Supporting Information Appendix 1 provides a tutorial, explaining in detail how to use these tools.

Choice of web tools

Based on the overview provided by Tranchevent *et al.* (2011) "A guide to web tools to prioritize candidate genes", we have chosen to use the two web tools Gene Prospector (<http://www.hugenavigator.net/HuGENavigator/geneProspectorStartPage.do>) and PolySearch (<http://wishart.biology.ualberta.ca/polysearch/>). These tools are both user friendly and suitable for finding promising genes of interest (Tranchevent *et al.*, 2011). The choice of these specific tools is based on the type of web resources used in their search (literature and text mining), the fact that the output genes are prioritized, and because they allow for searching the genome based on a keyword such as a disease name or a set of keywords related to a disease. Furthermore, both tools allow for genome-wide searching. Gene Prospector searches for human genes in relation to a disease phenotype or other related phenotypes. It selects and ranks genes based on a literature search of genetic association studies and thus relies exclusively on text mining (Tranchevent *et al.*, 2011; Yu *et al.*, 2008). PolySearch selects and analyses relationships between diseases, genes and other pathways using several biomedical databases (e.g. Entrez SNP, OMIM, HapMap) (Cheng *et al.*, 2008; Gibbs *et al.*, 2003; Hamosh *et al.*, 2005). Both tools complement each other

and it is therefore advised to use them together and to combine the information for obtaining a more comprehensive overview.

Example

To illustrate the use of these web tools we used “alcohol dependence” as an example input phenotype. The web tools then provide a table with all genes associated with alcohol dependence in the literature. Since both web tools use different search algorithms (see Supporting Information Appendix 1), combining the output of both tools will lead to a more comprehensive list of possible associations. When comparing the top 20 hits from both tables, in this example, 12 genes overlap (see Table 1). Since text mining tools can easily produce false positive associations, the findings need to be manually inspected to verify whether these are likely to represent true associations. False positive associations can occur for example when a gene is listed as important solely because it is mentioned in the abstract of a GWAS, while significant findings with that gene are not necessarily present. The UCSC Genome Browser (<https://genome.ucsc.edu/>) and links to relevant meta-analyses provided by Gene Prospector can be used for the manual curation of top results (see Supporting Information Appendix 1,

Figure S3 and Table S2 for a more detailed description of this follow-up search).

Statistical power

In the previous section we illustrated how to prioritize genes with alcohol dependence as an example phenotype. Suppose we have identified *ADH1B* as one of the most interesting genes for a genetic study in which we aim to identify genetic variants that influence low alcohol response in adolescents with a family history of alcohol use disorder. In practice, we would test polymorphisms within or close to *ADH1B*. Although it may be tempting to test all polymorphisms within this gene, small sample sizes in candidate gene studies force us to limit the number of variants to achieve sufficient statistical power. It is therefore important to determine in advance how much power we can expect to have given our sample size, the number of genetic variants included in the study, the MAF, and other study characteristics.

In this section we will focus on the impact of effect size, sample size, and the number of tests on statistical power. Although not explicitly investigated here, power depends on other factors as well, such as the MAF of a genetic variant. Rare causal genetic variants are much more difficult to detect than common causal variants, since the statistical power to detect significant association with such variants is low, unless the effect size is very large (Sham and Purcell, 2014). Similarly, for case-control samples it is not only the total sample size that matters, but also the relative number of cases and controls. Because an equal number of cases and controls results in optimal power, increasing the smallest subgroup (i.e. cases or controls) in size will have the biggest impact on power.

The Genetic Power Calculator by Purcell *et al.* (2003) (<http://pngu.mgh.harvard.edu/~purcell/gpc/>) is a useful online tool to perform genetic power calculations based on specific study characteristics. Supporting Information Appendix 2 describes in detail how to perform a power analysis with the Genetic Power Calculator. Here, we will illustrate the importance of performing a power analysis for both quantitative trait studies with continuous variables and case-control studies with dichotomous variables. For simplicity, we assume that analyzed SNPs are independent causal variants with additive genetic effects and no other causal variants in LD. First, assuming we want to test whether *ADH1B* is associated with a quantitative trait such as ethanol metabolism, how many independent genetic variants in *ADH1B* can we realistically test? As an effect size measure, we define the heritability explained by a genetic variant as the variance in phenotype (i.e. ethanol

Table 1. Candidate genes for alcohol dependence according to Gene Prospector and Polysearch

Gene Prospector	Polysearch	Overlap
<i>ALDH2</i>	<i>OPRM1</i>	<i>ADH1B</i>
<i>ADH1B</i>	<i>GABRA2</i>	<i>DRD2</i>
<i>DRD2</i>	<i>5HTT</i>	<i>OPRM1</i>
<i>SLC6A4</i>	<i>ADH1B</i>	<i>GABRA2</i>
<i>ADH1C</i>	<i>DRD2</i>	<i>SLC6A3</i>
<i>OPRM1</i>	<i>BDNF</i>	<i>ALDH2</i>
<i>GABRA2</i>	<i>HTR1B</i>	<i>HTR2A</i>
<i>SLC6A3</i>	<i>SLC6A3</i>	<i>COMT</i>
<i>CYP2E1</i>	<i>ANKK1</i>	<i>HTR1B</i>
<i>HTR2A</i>	<i>CHRM2</i>	<i>ADH4</i>
<i>COMT</i>	<i>ADH1C</i>	<i>ANKK1</i>
<i>MAOA</i>	<i>ADH4</i>	<i>BDNF</i>
<i>HTR1B</i>	<i>PDYN</i>	
<i>DRD4</i>	<i>ALDH2</i>	
<i>ADH4</i>	<i>ADH</i>	
<i>ANKK1</i>	<i>NPY</i>	
<i>TPH1</i>	<i>COMT</i>	
<i>HTR1A</i>	<i>HTR2A</i>	
<i>BDNF</i>	<i>GATA4</i>	
<i>MTHFR</i>	<i>SNCA</i>	

metabolism) explained by that particular variant. Figure 4 illustrates the effect of three factors that determine statistical power: heritability explained by a genetic variant, sample size, and the number of genetic variants included in the study. The heritability explained by single genetic variants associated with quantitative phenotypes is typically smaller than $< 1\%$, even if the heritability of the phenotype is large. Therefore the maximum explained variation in this example is set at 2% . Figure 4 shows the power to detect associations for different scenarios for genetic variants with a MAF of 0.25. Even if a genetic variant has a heritability of 1% and the sample size is 400, the power to detect an association is only 50% . If we choose to test 10 independent genetic variants with similar effect sizes and correct for multiple testing, the power drops to 20% .

Second, for phenotypes that are not quantitative but dichotomous, a similar picture emerges. Suppose we want to test whether *ADH1B* is associated with a dichotomous trait such as binge drinking before the age of 14 years, assuming a population prevalence of 1% , MAF of 0.25 and a sample case-control ratio of one. Figure 5 shows the power to detect genetic associations in such a case-control study. In this example, the effect size measure is the odds ratio. If the odds ratio were 1.5, which is a relatively large effect size, a case-control sample with 200 cases and 200 controls would achieve a power of 70% for testing a single variant. However,

when testing 10 independent variants genetic power decreases to 50% due to multiple testing correction. As Figure 5 shows, statistical power is somewhat lower at higher population prevalence, but is mostly affected by the effect size, sample size, and number of genetic variants to test.

The earlier mentioned examples illustrated the results of power analyses for relatively large effect sizes. Although it is difficult to know in advance which effect size to expect, results from GWAS suggest effect sizes are often small. Significant odds ratios in case-control studies typically range between 1.1 and 1.4 and are most often found in the lower end of this range (Ioannidis *et al.*, 2010; Park *et al.*, 2011). For quantitative traits, genetic variants with a heritability of 1% are extremely rare, and even a heritability of 0.5% can be considered large (Park *et al.*, 2011). As the power plots show, small samples will typically not yield sufficient power of 80% for genetic testing, especially if many independent genetic variants are being tested. In practice, testing the involvement of a large number of genetic variants with small effect sizes, requires sample sizes in the order of thousands to achieve sufficient power. Sample sizes in the order of hundreds are only acceptable if few, relatively frequent genetic variants are tested for which strong associations with similar phenotypes of interest have been reported in previous studies.

Figure 4. Power in quantitative trait studies. Power is shown as a function of sample size, number of single nucleotide polymorphisms (SNPs) tested, and heritability (h^2) of the genetic variant. Minor allele frequency is fixed at 0.25. Tested SNPs are assumed to be independent causal variants with no other causal variants in linkage disequilibrium.

Figure 5. Power in case-control studies. Power is shown as a function of sample size, number of single nucleotide polymorphisms (SNPs) tested, prevalence (prev), and odds ratio (OR). Minor allele frequency is fixed at 0.25 and case-control ratio is one. Tested SNPs are assumed to be independent causal variants with no other causal variants in linkage disequilibrium.

From gene to polymorphism

In the previous fourth section, we have discussed several tools which can be used to guide the selection of a set of candidate genes. This is not the endpoint of designing a genetic study, since we are now faced with the next step: the selection of genetic polymorphisms. Understanding which genes contribute to a disorder requires genotyping and investigating those genetic variants that explain phenotypic variation in the population. In the fifth section

we have discussed the importance of determining in advance how many polymorphisms can be analyzed. In this section we discuss which polymorphisms to select. Selecting polymorphisms within a specific gene of interest is not always as straightforward as it may seem. We will discuss two scenarios and will describe some of the important aspects on the selection of polymorphisms. However, this is a complex topic, and it is beyond the scope of this article to provide a comprehensive guide. For a more

thorough discussion of this topic, we refer the reader to a review by Patnala *et al.* (2013).

Scenario 1: Selecting genetic polymorphisms based on known genetic associations

First, we will discuss the selection of polymorphisms for genes which are selected because genetic variants in or near these genes have previously been found to be associated with the disorder of interest. For example, GWAS have shown that the SNP rs6943555 in the autism susceptibility candidate 2 (*AUTS2*) gene is associated with alcohol consumption (Schumann *et al.*, 2011). We may therefore be interested in the role of *AUTS2* in the brain reward system and aim to investigate this gene in a genetic study in which functional neuroimaging data is available. In this example, the choice for a particular polymorphism can be based on the previous results of GWAS; we can simply select the genetic variant that was reported to be most strongly associated with the disorder in the original publication. If multiple genetic variants have been reported in the original paper, it depends on the amount of LD whether all of these variants or a specific subset should be selected for further investigation.

Scenario 2: Selection genetic polymorphisms based on indirect evidence

Much more complex is the second scenario, in which a gene of interest was selected not because of the presence of significantly associated variants in this gene, but as a result of indirect evidence. For example, protein–protein interaction analyses may show that the product (protein) of a gene physically interacts with the protein of a well-known candidate gene for the trait of interest and we may therefore hypothesize that the novel gene is also associated with the trait. In this scenario, there are no specific polymorphisms which are highlighted in this gene, but instead, the entire gene is the unit of interest. Nevertheless, when designing a genetic study, we need to select one or multiple genetic polymorphisms. We will make recommendations on how to proceed, although no simple guidelines are yet available, and consulting a geneticist may be crucial.

One possibility is to sequence the entire coding region of the gene of interest. Targeted sequencing of candidate genes is technically feasible and the costs will be relatively low, in the order of 10 to 20 dollars per gene. However, the number of statistical tests will be large, due to the large number of variants that will need to be tested. A second option is to select genetic polymorphisms which influence the expression levels of the gene of interest: expression

quantitative trait loci (eQTL). For this purpose, publicly available eQTL data (Lonsdale *et al.*, 2013; Xia *et al.*, 2012) can be used for the selection of relevant polymorphisms. Third, one could specifically focus on SNPs that directly alter the protein product, such as mis-sense SNPs (i.e. SNPs which result in a codon that codes for a different amino acid) or SNPs associated with alternative splicing, a mechanism by which different forms of mRNAs are generated from the same genetic code. The impact of mis-sense SNPs on protein structure can be estimated through PolyPhen-2 (<http://genetics.bwh.harvard.edu/pph2/>) (Adzhubei *et al.*, 2013). SNPs associated with alternative splicing can be identified with the R package sQTLseeker (Monlong *et al.*, 2014).

Relevant characteristics of genetic polymorphisms

Here, we will briefly describe some of the characteristics that are important to consider when selecting genetic polymorphisms, including the frequency of the alleles in the general population and ethnic variability of the SNP. Information on these characteristics can be obtained from dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>) (Sherry *et al.*, 2001). As we discussed earlier in the fifth section, the statistical power to detect significant associations will be lower for causal genetic variants with a low MAF. For this reason and also considering the fact that rare genetic variants have been found to be more prone to genotyping errors (Wall *et al.*, 2014), selection of variants with relatively high (e.g. > 0.2) MAF is generally preferred. Furthermore, the MAF of genetic polymorphisms may differ across ethnic populations (i.e. population stratification) and this may lead to false-positive associations (Marchini *et al.*, 2004). For example, a polymorphism (5-HTTLPR) in the serotonin transporter gene (*SLC6A4*), an extensively studied genetic variant, shows large allele frequency differences across European and Asian populations (Hu *et al.*, 2006). For such variants, ethnic differences may lead to spurious findings if ethnicity is correlated with the phenotype of interest. Generally, it is advised to include only participants from the same ethnic background, but more subtle ethnic differences are difficult to rule out. Therefore, ethnic variability in MAF is something to be aware of when selecting a genetic polymorphism.

Finally, publicly available data of the ENCODE study (ENCODE Project Consortium, 2012) contains information on the existence of functionally relevant markers at a specific genomic location. The different types of functional variation may be useful for evaluating the relevance of a specific genetic polymorphism. For example, if

transcription factors bind to a specific genomic location, it is more likely that this locus has a functional role.

Discussion

In this article we have provided practical guidelines on how to select genetic variants for a genetic association study. We have shown how gene prioritization can be informed by the use of two bioinformatics tools: PolySearch and Gene Prospector. Power analyses of binary and continuous traits were conducted for a range of effect sizes and sample sizes. Finally, we have discussed some of the relevant aspects of the selection of specific genetic polymorphisms for a specific gene.

PolySearch and Gene Prospector are bioinformatics tools which combine information from different resources to obtain a ranking among possible genes for a specific trait of interest. In this article, we have illustrated the application of these tools using alcohol dependence as an example trait of interest. A high overlap was observed; 12 genes are included in the top 20 of both tools. However, discrepancies are also clearly present, which illustrates the importance of using multiple tools for gene prioritization. Manual evaluation of the top results is essential, and is greatly facilitated by these tools since it is rather straightforward to obtain detailed insight in the supportive evidence (e.g. links to the results of meta-analyses) and to evaluate the validity of these results.

The statistical power analyses showed that for the range of effect sizes most commonly reported in genetic studies of complex traits, sample sizes of a few hundred subjects are not sufficient in case-control designs or population sampling of continuous phenotypes. In GWAS sample size has been scaled up from thousands to tens of thousands of subjects, resulting in the first genome-wide significant associations (see review of Visscher *et al.*, 2012). We show here that the same is true for candidate-gene studies, in which only a few genetic variants are tested. Instead of sample sizes of ~50 to ~200 cases, we expect that candidate gene studies should include at least 500 subjects, unless few genetic variants with relatively large effect sizes will be investigated. We emphasize that performing an underpowered study is similar to a speculative investment. We might be fortunate and find a significant result, but a research grant can probably be spent better. While a non-significant result in a well-powered study (power > 80%) at least indicates that the true effect is probably smaller than you expected, the same result in an underpowered study does not provide information and is a waste of money. We therefore advise to specify a reasonable expected effect size in advance, perform a power

analysis, and, if necessary, to limit the number risk variants tested to achieve sufficient power.

In addition to limiting the number of genetic polymorphisms in the study, there are several other ways to improve the statistical power. First, the statistical power to detect genetic variants with low frequency in the population can be increased by using linkage analysis within a family design (Ott *et al.*, 2011). In linkage analysis, one or more extended families are analyzed to identify genetic regions that segregate with the phenotype to the next generation. However, the genetic regions that are identified with linkage analysis are relatively large, compared to the LD regions identified in GWAS. Therefore, additional analyses will be required to identify the causal genetic variants. Second, statistical power may be increased by testing the association with multiple statistically associated SNPs simultaneously by performing haplotype based tests (Browning and Browning, 2007). Third, statistical power can be increased by the use of polygenic risk score analysis (Salvatore *et al.*, 2014; Purcell *et al.*, 2009; Belsky *et al.*, 2013). In this type of analysis, multiple genetic variants are combined to create a single polygenic risk score based on the genetic effect sizes which have been estimated in a discovery sample. Association between this risk score and the phenotype is then tested in a target sample that should be completely independent of the discovery sample. With this approach, instead of identifying the individual contribution of each genetic variant, a polygenic risk score analysis identifies the cumulative contribution of multiple genetic variants to the phenotype of interest. Fourth, although Bonferroni correction is the most widely used approach for controlling type-I error rate, this correction is overly conservative when testing the association with multiple correlated SNPs. Instead, permutation testing can be used to estimate permutation-based *p*-values by constructing an empirical frequency distribution of the *p*-values under the null hypothesis of no association (Sham and Purcell, 2014).

Fifth, statistical power largely depends on effect size and some argue that the statistical association between genetic variants and endophenotypes might be stronger than those seen for the association with clinical diagnosis (Gottesman and Gould, 2003). Endophenotypes are biological phenotypes, such as functional neuroimaging measures, which are more directly regulated by gene expression compared to behavioral phenotypes. For example, Mier *et al.* (2009) have conducted a meta-analysis testing the association between the Val158Met (rs4680) in the *COMT* gene (rs4680) and prefrontal activation and reported a large effect size ($d = 0.73$; corresponding to 12% explained variance). However, Flint and Munafò

(2007) argue that the genetic architecture of endophenotypes might not be easier to dissect than that of traditional phenotypes. Finally, meta-analysis can be an excellent tool both in genome-wide association studies and candidate gene studies to further increase power and reduce false-positive findings by statistically synthesizing the findings of multiple independent studies to calculate combined effect sizes (Evangelou and Ioannidis, 2013). Using funnel plots, meta-analysis can also provide insight into the extent of publication bias, caused by the preferential publication of positive associations compared to null findings (Richard *et al.*, 2009).

In conclusion, genetic association studies which focus on few genetic variants of high interest are particularly well suited to detect associations between genetic variants and refined phenotypes that cannot be assessed in large sample sizes. An important challenge is to prioritize and select those genes which are most likely associated with the outcome measure of interest. In this article we have provided guidelines for prioritizing genes using bioinformatics tools that search the currently available data and literature.

Furthermore, we have illustrated how genetic power calculations can be performed and have conducted power analyses for a range of effect sizes and sample sizes. Caveats of candidate gene studies have been discussed, including the possibility of false-positive associations due to ethnic heterogeneity within the samples, and possible ways to circumvent these problems have been explained. Awareness and practical guidelines of the primary tools and techniques that are available in the field of genetics will hopefully contribute to an increased understanding of genetics for psychiatrists. That would be a great step forward, since this would allow the genetics field to benefit from the expertise of psychiatrists in describing those phenotypic characteristics that are most strongly associated with genetic variation.

Acknowledgements

This work was supported by the Volksbond Rotterdam Foundation.

References

- Adzhubei I., Jordan D.M., Sunyaev S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, Chapter 7, Unit 7.
- Belsky D.W., Moffitt T.E., Baker T.B., Biddle A.K., Evans J.P., Harrington H., Houts R., Meier M., Sugden K., Williams B. (2013) Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study. *JAMA Psychiatry*, **70**, 534–542.
- Brown T.A. (2002) *Genomes*. Oxford: Wiley-Liss.
- Browning B.L., Browning S.R. (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology*, **31**, 365–375.
- Cheng D., Knox C., Young N., Stothard P., Damaraju S., Wishart D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research*, **36**, W399–W405.
- Darryl Leja, NGHRI. (2010a) Chromosome (Advanced). NGHRI Digital Media Database. <http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85282>
- Darryl Leja, NGHRI. (2010b) Translation. NGHRI Digital Media Database. <http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85252>
- Dudbridge F., Gusnanto A. (2008) Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, **32**, 227–234.
- ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
- Evangelou E., Ioannidis J.P. (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews. Genetics*, **14**, 379–389.
- Flint J., Munafò M.R. (2007) The endophenotype concept in psychiatric genetics. *Psychological Medicine*, **37**, 163–180.
- Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
- Gibbs R.A., Belmont J.W., Hardenbol P., Willis T.D., Yu F., Yang H., Ch'ang L. Y., Huang W., Liu B., Shen Y. (2003) The international HapMap project. *Nature*, **426**, 789–796.
- Gottesman I.I., Gould T.D. (2003) The endophenotype concept in psychiatry: etymology and strategic intentions. *American Journal of Psychiatry*, **160**, 636–645.
- Hamosh A., Scott A.F., Amberger J.S., Bocchini C.A., McKusick V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**, D514–D517.
- Hoggart C.J., Clark T.G., De Iorio M., Whittaker J. C., Balding D.J. (2008) Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology*, **32**, 179–185.
- Hu X.Z., Lipsky R.H., Zhu G., Akhtar L.A., Taubman J., Greenberg B.D., Xu K., Arnold P.D., Richter M.A., Kennedy J.L. (2006) Serotonin transporter promoter gain-of-function genotypes are linked to obsessive-compulsive disorder. *The American Journal of Human Genetics*, **78**, 815–826.
- Ioannidis J.P., Castaldi P., Evangelou E. (2010) A compendium of genome-wide associations for cancer: critical synopsis and reappraisal. *Journal of the National Cancer Institute*, **102**(12), 846–858.
- Lao O., Lu T.T., Nothnagel M., Junge O., Freitag-Wolf S., Caliebe A., Balasakova M., Bertranpetit J., Bindoff L.A., Comas D. (2008) Correlation between genetic and geographic structure in Europe. *Current Biology*, **18**, 1241–1248.
- Lonsdale J., Thomas J., Salvatore M., Phillips R., Lo E., Shad S., Hasz R., Walters G., Garcia F., Young N. (2013) The genotype-tissue expression (GTEx) project. *Nature Genetics*, **45**, 580–585.
- Marchini J., Cardon L.R., Phillips M.S., Donnelly P. (2004) The effects of human population structure on large genetic association studies. *Nature Genetics*, **36**, 512–517.

- Mier D., Kirsch P., Meyer-Lindenberg A. (2009) Neural substrates of pleiotropic action of genetic variation in COMT: a meta-analysis. *Molecular Psychiatry*, **15**, 918–927.
- Monlong J., Calvo M., Ferreira P.G., Guigo R. (2014) Identification of genetic variants associated with alternative splicing using sQTLseeker. *Nature Communications*, **5**, 4698.
- Neale B.M., Ferreira M.A.R., Medland S.E., Posthuma D. (2008) Statistical Genetics: Gene Mapping through Linkage and Association. New York: Taylor & Francis.
- Niculescu A.B. (2013) Convergent functional genomics of psychiatric disorders. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics*, **162**, 587–594.
- Ott J., Kamatani Y., Lathrop M. (2011) Family-based designs for genome-wide association studies. *Nature Reviews. Genetics*, **12**, 465–474.
- Park J.H., Gail M.H., Weinberg C.R., Carroll R.J., Chung C.C., Wang Z., Chanock S.J., Fraumeni J.F., Chatterjee N. (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, **108**, 18026–18031.
- Patnala R., Clements J., Batra J. (2013) Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genetics*, **14**, 39.
- Purcell S.M., Cherny S.S., Sham P.C. (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, **19**, 149–150.
- Purcell S.M., Wray N.R., Stone J.L., Visscher P.M., O'Donovan M.C., Sullivan P.F., Sklar P. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.
- Richard J., Pillemer D.B., Light R.J. (2009) Summing Up: The Science of Reviewing Research. Cambridge, MA: Harvard University Press.
- Salvatore J.E., Aliev F., Edwards A.C., Evans D.M., Macleod J., Hickman M., Lewis G., Kendler K.S., Loukola A., Korhonen T. (2014) Polygenic scores predict alcohol problems in an independent sample and show moderation by the environment. *Genes*, **5**, 330–346.
- Schumann G., Coin L. J., Lourdasamy A., Charoen P., Berger K.H., Stacey D., Desrivieres S., Aliev F.A., Khan A. A., Amin N., Aulchenko Y. S., Bakalkin G., Bakker S.J., Balkau B., Beulens J.W., Bilbao A., de Boer R.A., Beury D., Bots M.L., Breetvelt E.J., Cauchi S., Cavalcanti-Proenca C., Chambers J.C., Clarke T.K., Dahmen N., de Geus E.J., Dick D., Ducci F., Easton A., Edenberg H.J., Esko T., Fernandez-Medarde A., Foroud T., Freimer N.B., Girault J.A., Grobbee D.E., Guarrera S., Gudbjartsson D.F., Hartikainen A.L., Heath A.C., Hesselbrock V., Hofman A., Hottenga J. J., Isohanni M.K., Kaprio J., Khaw K.T., Kuehnel B., Laitinen J., Lobbens S., Luan J., Mangino M., Maroteaux M., Matullo G., McCarthy M.I., Mueller C., Navis G., Numans M.E., Nunez A., Nyholt D.R., Onland-Moret C.N., Oostra B.A., O'Reilly P.F., Palkovits M., Penninx B.W., Polidoro S., Pouta A., Prokopenko I., Ricceri F., Santos E., Smit J.H., Soranzo N., Song K., Sovio U., Stumvoll M., Surak I., Thorgerisson T.E., Thorsteinsdottir U., Troakes C., Tyrfinngsson T., Tonjes A., Uitterwaal C.S., Uitterlinden A.G., van der H.P., van der Schouw Y. T., Staehlin O., Vogelzangs N., Vollenweider P., Waechter G., Wareham N.J., Waterworth D.M., Whitfield J.B., Wichmann E.H., Willemsen G., Witteman J. C., Yuan X., Zhai G., Zhao J.H., Zhang W., Martin N.G., Metspalu A., Doering A., Scott J., Spector T.D., Loos R.J., Boomsma D.I., Mooser V., Peltonen L., Stefansson K., van Duijn C.M., Vineis P., Sommer W.H., Kooner J.S., Spanagel R., Heberlein U.A., Jarvelin M.R., Elliott P. (2011) Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 7119–7124.
- Sham P.C., Purcell S.M. (2014) Statistical power and significance testing in large-scale genetic studies. *Nature Reviews. Genetics*, **15**, 335–346.
- Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**, 308–311.
- Tranchevent L.C., Capdevila F.B., Nitsch D., De Moor B., De Causmaecker P., Moreau Y. (2011) A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, **12**, 22–32.
- Visscher P.M., Brown M.A., McCarthy M.I., Yang J. (2012) Five years of GWAS discovery. *The American Journal of Human Genetics*, **90**, 7–24.
- Wall J.D., Tang L.F., Zerbe B., Kvale M.N., Kwok P.Y., Schaefer C., Risch N. (2014) Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Research*, **24**, 1734–1739.
- Welter D., MacArthur J., Morales J., Burdett T., Hall P., Junkins H., Klemm A., Flicek P., Manolio T., Hindorf L. (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, **42**, D1001–D1006.
- Xia K., Shabalin A.A., Huang S., Madar V., Zhou Y.H., Wang W., Zou F., Sun W., Sullivan P.F., Wright F.A. (2012) seeQTL: a searchable database for human eQTLs. *Bioinformatics*, **28**, 451–452.
- Yu W., Wulf A., Liu T., Khoury M.J., Gwinn M. (2008) Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics*, **9**, 528.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.