Published in final edited form as:

Eur Psychiatry. 2010 June ; 25(5): 307–309. doi:10.1016/j.eurpsy.2009.12.019.

Statistical challenges for genome-wide association studies of suicidality using family data

Jessica Lasky-Su^{1,2,3} and Christoph Lange^{1,3,4}

¹Channing Laboratory, Brigham and Women's Hospital, Boston, MA, USA

²Department of Medicine, Harvard Medical School, Boston, MA USA

³Center for Genomic Medicine, Brigham and Women's Hospital, Boston, MA, USA

⁴Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Abstract

The etiology of suicide is complex in nature with both environmental and genetic causes that are extremely diverse. This extensive heterogeneity weakens the relationship between genotype and phenotype and as a result, we face many challenges when studying the genetic etiology of suicide. We are now in the midst of a genetics revolution, where genotyping costs are decreasing and genotyping speed is increasing at a fast rate, allowing genetic association studies to genotype thousands to millions of SNPs that cover the entire human genome. As such, genome-wide association studies (GWAS) are now the norm. In this article we address several statistical challenges that occur when studying the genetic etiology of suicidality in the age of the genetics revolution. These challenges include 1) the large number of statistical tests; 2) complex phenotypes that are difficult to quantify; and 3) modest genetic effect sizes. We address these statistical issues in the context of family-based study designs. Specifically we discuss several statistical extensions of family-based association tests (FBATs) that work to alleviate these challenges. As our intention it to describe how statistical methodology may work to identify disease variants for suicidality, we avoid the mathematical details of the methodologies presented.

Keywords

Family-based association methods; genetics; suicidality; psychiatric disorders; candidate gene; genome-wide association

Family-based association studies

There are two distinctly different designs for genetic studies: the family-based study design and designs that use unrelated individuals, both of which have their strength and weaknesses with regard to genome-wide association studies (GWAS). The most basic family-based design genotypes two parents and an affected offspring. The statistical test for this design is the transmission disequilibrium test (TDT)[1], which compares the observed number of alleles that are transmitted to the expected offspring compared with the expected number of

Corresponding Author: Jessica Lasky-Su 181 Longwood Ave. Boston, MA, 02115, USA rejas@channing.harvard.edu Phone: 978-465-5398 Fax: 617-525-0958.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

transmissions under Mendel's law. The TDT has been generalized to the FBATs that allow for multiple siblings, phenotypes of any distribution, all genetic models, and adjustment for covariates.

The FBAT statistic

We assume that there are n parent-offspring trios, that each offspring has m coded genotypes, $k = 1, 2, \dots, m$. Let x_{ik} be the coded genotype of the k^{th} marker for the i^{th} proband that can be coded in an additive, dominant, or recessive manner. The variables p_{ik1} and p_{ik2} denote the parental genotypes for the parents of the i^{th} proband at the k^{th} marker, y_i denotes the trait for the i^{th} proband, and μ denotes the offset for that trait. Therefore the FBAT statistic, $FBAT_k$, tests for an association between the k^{th} marker locus and the trait of interest and is given by the following[2]:

$$FBAT_k^2 = \left\{ \sum_{i=1}^n (y_i - \mu) [x_{ik} - E(x_{ik} | p_{ik1}, p_{ik2})] \right\}^2 / \sum_{i=1}^n (y_i - \mu)^2 \text{Var}(x_{ik} | p_{ik1}, p_{ik2}) \mathbf{X}_1^2$$

The FBAT test statistic has been extended several ways to address multiple testing, the generation of complex phenotypes, and analyzing multiple markers simultaneously.

Multiple testing

GWAS analyses evaluate the genetic associations for hundreds of thousands to millions of SNPs. As such, the multiple testing from the large numbers of SNPs often makes it difficult to identify genetic associations that are genome-wide significant. This is particularly true when studying a disease like suicidality, where the genetic effect sizes of any given single nucleotide polymorphism (SNP) is likely to be small. One of the attractive features of FBATs is that the test statistic can be divided into two independent components, commonly known as the “within” and “between” components. Because the FBAT statistic only uses the within component, Lange et al.[3–5] has proposed several ways to use the between component, through the use of the conditional mean model, in order to increase the power of FBATs. Vansteen et al.[5] proposed to use the between component in an initial “screening stage” whereby an estimate of the genetic effect size is calculated at each SNP. This estimate of the genetic effect size can then be used to select a subset of SNPs that are actually tested using the FBAT test statistic. Because the between and within components are completely independent, this ranking of the data based on the genetic effect size and the selection of a subset of SNPs to test, does not bias the subsequent FBAT statistic. This screening approach, which is implemented in the PBAT program[6], offers a way to reduce the multiple tests, thereby increasing the ability to identify genome-wide significant SNPs. It has also been found to be more powerful than standard GWAS analyses with family-based data[5]. Several disease variants have been identified using this approach, including the identification of genetic variants that influence Alzheimer's disease[5,7–9] and many extensions and improvements on this methodology have been proposed[10]. This screening method is an ideal tool to apply to genetic studies of suicidality for several reasons. First, unlike the TDT, this allows one to capitalize on all of the information that is obtained from family trios. Second, it is important to note that, although suicidality has a distinct genetic component[11], this behavior is also often the result of another underlying illness. Therefore suicidality is an uncommonly heterogeneous disease and the genetic underpinnings of the disorder are likely to be even more difficult to detect. With severe multiple testing corrections, the likelihood of identifying these associations becomes even smaller. Therefore, if the number of statistical tests can be reduced using the screening method, this approach is likely to greatly improve the power of detecting disease variants for suicidality.

Complex phenotypes

The best way to identify true genetic associations is to use highly heritable phenotypes in the genetic analyses. However, selecting these phenotypes for suicidality is not trivial. Psychiatric disorders in general are often diagnosed using a complex set of criteria making it difficult to determine what phenotypes to use in genetic analyses. Often the best approach for generating heritable genetic phenotypes is to derive them statistically. The FBAT-principal components (FBAT-PC) methodology is such an approach that can be used to construct SNP-specific phenotypes that are maximally heritable through the use of several relevant quantitative phenotypes[12]. In order to use FBAT-PC, a group of related phenotypes with similar distributional characteristics are first identified. When multiple phenotypes are used, FBAT-PC uses principal components analysis to construct an overall phenotype that amplifies the trait heritability by aggregating the genetic components of all measurements into a single univariate phenotype with maximal heritability. This univariate trait with maximal heritability is constructed as follows. First, the conditional mean model proposed by Lange et al.[12] is used to generate genetic effect sizes for each phenotype/measurement. The effect size estimates are used along with information on the mode of inheritance and allele frequencies to estimate a matrix of genetic variances at that SNP. Second, the phenotypic variance matrix is calculated using the original set of phenotypes as well as any other desired covariates. Once the genetic and phenotypic variance matrices are generated, generalized principal components is used to identify the set of weights that maximize the heritability at the given SNP. As it can be shown, the remaining principal component do not contain any information about the genetic association; The FBAT-PC methodology has been successfully implemented in several association studies, both in GWAS[7,13] and candidate gene studies[14,15]. We implemented in FBAT-PC in a samples of parent offspring trios affected with attention deficit hyperactive disorder using 9 inattentive and 9 hyperactive-impulsive symptoms. Using the between component information, a set of weights are generated for each of the 18 symptoms and a composite phenotype is generated as described above to maximize the heritability at each SNP while remaining statistically independent of any subsequent association test. It is important to note that each of these symptoms were measured on the same ordinal scale, so that the implementation of FBAT-PC was appropriate. This methodology is very applicable to studying the genetics of suicidality because this phenotype is difficult to measure, until an individual actually commits suicide. In addition, it is unclear if actually using the act of committing suicide is the best, most heritable phenotype to use in genetic epidemiological studies. Therefore, to study suicidality it is likely best to use several phenotypes that measure suicidal ideation. Rather than deciding on one phenotype to use, several phenotypes for suicidal ideation could be incorporated into the FBAT-PC framework to generate a maximally heritable phenotype for each SNP being analyzed.

Another approach to increase the heritability of the phenotype is to use phenotypes that are closer in the biologic pathway to the actual genotypes being studied (Figure 1). Specifically, when selecting phenotypes that are further down the biologic pathway, away from the genotype information, the strength of the relationship between the genotype and phenotype decreases. This is particularly true when studying the genetics of suicidality, as it is often the case where committing suicide occurs after another clinical phenotype, thereby weakening the genotype-phenotype relationship further. For example, it is sometimes the case that the clinical phenotype is another psychiatric disorder, such as depression or drug dependence. In this case, committing suicide may be the result of an episode of depression or consuming illicit drugs. As a result there is even more distance in the causal pathway between the gene and the outcome of interest, which further weakens the genotype-suicide relationship. Using phenotypes closer to the genotype, such as gene expression levels, will result in a much higher heritability estimates than clinical traits. These increases in heritability also increase

the genetic effect size, thereby making the genetic effect easier to detect. Therefore, for studying the genetics of suicidality, using gene expression levels as a phenotype is one of the optimal approaches to take. Brain expression levels are difficult, if not impossible to obtain for most psychiatric disorders, however, expression levels are often easier to obtain from individuals who have committed suicide, as autopsies must be performed in the event of a suicide.

Increasing the number of SNPs to be tested

Another way to increase the power to identify disease variants is to study several genetic markers simultaneously. The cumulative genetic effect of multiple SNPs is more likely to have a higher heritability than any of the individual SNPs. Other methods, including global haplotype tests [16], regression methods [17,18], and multimarker tests [19], address multiple testing by combining multiple SNPs into a single test. Global haplotype tests enumerate the possible haplotypes and then evaluate the null hypothesis that no haplotype is associated with the disease. Regression-based approaches look at associations between a trait and all linear combination of SNPs. Multimarker methods [19] test multiple SNPs simultaneously by combining individual marker scores and pairwise correlations while avoiding the use of haplotype structure. Another statistical test, multivariate FBAT (MFBAT) [20], incorporates both multiple markers and multiple phenotypes into a single test statistic. By reducing the number of tests across two dimensions, the total number of statistical tests is reduced to one and the multiple testing problem is addressed more efficiently than other methods. The MFBAT methodology is also flexible in nature and can incorporate phenotypes that have varying distributional properties. These methodologies are particularly useful to identify disease variants for suicidality, as this increases the overall genetic effect that is being studied at one time. Because the genetics underlying suicidality are very heterogeneous, by using multiple markers in one statistical test, a heterogeneous disease, with more than one genetic effect can be combined into one statistical test and thereby increase the overall power of the test.

Summary

We have discussed several challenges for studying the genetic etiology of suicide in the context of the genetics revolution. We discuss how family-based approaches offer several approaches to address these issues that will increase the overall power we have to identify genetic associations for suicidality.

References

- [1]. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;52:506–16. [PubMed: 8447318]
- [2]. Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Heredity* 2000;50:211–23. [PubMed: 10782012]
- [3]. Lange C, Laird NM. On a general class of conditional tests for family-based association studies in genetics: The asymptotic distribution, the conditional power, and optimality considerations. *Genet Epidemiol* 2002;23:165–80.
- [4]. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. Using the noninformative families in family-based association tests: a powerful new testing strategy. *Am J Hum Genet* 2003;73:801–11. [PubMed: 14502464]
- [5]. Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, DeMeo DL, et al. Genomic screening and replication using the same data set in family-based association testing. *Nat Genet* 2005;37:683–91. [PubMed: 15937480]

- [6]. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. PBAT: tools for family-based association studies. *Am J Hum Genet* 2004;74:367–9. [PubMed: 14740322]
- [7]. Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, et al. A common genetic variant is associated with adult and childhood obesity. *Science* 2006;312:279–83. [PubMed: 16614226]
- [8]. Lasky-Su J, Lyon HN, Emilsson V, Heid IM, Molony C, Raby BA, et al. On the replication of genetic associations: timing can be everything! *Am J Hum Genet* 2008;82:849–58. [PubMed: 18387595]
- [9]. Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF, et al. Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. *Am J Hum Genet* 2008;83:623–32. [PubMed: 18976728]
- [10]. Ionita-Laza I, McQueen MB, Laird NM, Lange C. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am J Hum Genet* 2007;81:607–14. [PubMed: 17701906]
- [11]. Wasserman D, Geijer T, Sokolowski M, Rozanov V, Wasserman J. Nature and nurture in suicidal behavior, the role of genetics: some novel findings concerning personality traits and neural conduction. *Physiol Behav* 2007;92:245–9. [PubMed: 17585962]
- [12]. Lange C, van Steen K, Andrew T, Lyon H, DeMeo DL, Raby B, et al. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat Appl Genet & Mol Biol* 2004;3:1–27. 2004.
- [13]. Lasky-Su J, Neale BM, Franke B, Anney RJ, Zhou K, Maller JB, et al. Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identifies novel associations and confirms candidate gene associations. *Am J Med Genet B Neuropsychiatr Genet* 2008;147B:1345–54. [PubMed: 18821565]
- [14]. Lasky-Su J, Lange C, Biederman J, Tsuang M, Doyle AE, Smoller JW, et al. Family-based association analysis of a statistically derived quantitative traits for ADHD reveal an association in DRD4 with inattentive symptoms in ADHD individuals. *Am J Med Genet B Neuropsychiatr Genet* 2008;147B:100–6. [PubMed: 17579349]
- [15]. Lasky-Su J, Banaschewski T, Buitelaar J, Franke B, Brookes K, Sonuga-Barke E, et al. Partial replication of a DRD4 association in ADHD individuals using a statistically derived quantitative trait for ADHD in a family-based association test. *Biol Psychiatry* 2007;62:985–90. [PubMed: 17560555]
- [16]. Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM. Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet Epidemiol* 2004;26:61–9. [PubMed: 14691957]
- [17]. Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol* 2004;27:415–28. [PubMed: 15481099]
- [18]. Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 2003;56:18–31. [PubMed: 14614235]
- [19]. Rakovski CS, Xu X, Lazarus R, Blacker D, Laird NM. A new multimarker test for family-based association studies. *Genet Epidemiol* 2007;31:9–17. [PubMed: 17086514]
- [20]. Lasky-Su J, Murphy A, McQueen MB, Weiss ST, Lange C. *European Journal of Human Genetics*. resubmitted.

genotype $\Rightarrow$ gene expression $\Rightarrow$ signaling $\Rightarrow$ cellular phenotype $\Rightarrow$ organ dysfunction $\Rightarrow$ clinical phenotype

Figure 1.

Common biological measurements (i.e., intermediate phenotypes) that can be measured and occur between a genotype and the clinical phenotype. Because these intermediate phenotypes occur closer to the genotype, the heritability and therefore any potential association is easier to detect. Therefore, it is often advantageous to use these intermediate phenotypes in genetic analyses.