



Published in final edited form as:

Prev Sci. 2018 January ; 19(1): 58–67. doi:10.1007/s11121-017-0785-1.

Critical Issues in the Inclusion of Genetic and Epigenetic Information in Prevention and Intervention Trials

Shawn J. Latendresse¹, Rashed Musci², and Brion S. Maher²

¹Department of Psychology and Neuroscience, Baylor University, Waco, TX

²Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Abstract

Human genetic research in the past decade has generated a wealth of data from the genome-wide association scan era, much of which is catalogued and freely available. These data will typically test the relationship between a single nucleotide variant or polymorphism (SNP) and some outcome, disease or trait. Ongoing investigations will yield a similar wealth of data regarding epigenetic phenomena. These data will typically test the relationship between DNA methylation at a single genomic location/region some outcome. Most of these findings will be the result of cross sectional investigations typically using ascertained cases and controls. Consequently, most methodological consideration focuses on methods appropriate for simple case control comparisons. It is expected that a growing number of investigators with longitudinal experimental prevention or intervention cohorts will also measure genetic and epigenetic indicators as part of their investigations, harvesting the wealth of information generated by the GWAS era to allow for targeted hypothesis testing in the next generation of prevention and intervention trials. Herein, we discuss appropriate quality control and statistical modelling of genetic, polygenic and epigenetic measures in longitudinal models. We specifically discuss quality control, population stratification, genotype imputation, pathway approaches, and proper modelling of GxE interaction.

Keywords

Prevention; Genetic; Polygenic Risk; Methylation; GWAS

The availability of relatively inexpensive and accessible genomic arrays has led to a growth in the inclusion of genetic and epigenetic data in prevention and intervention trials (Brody, Yu, Chen, Beach, & Miller, 2015; Musci et al., 2015; Vandenberg et al., 2016). Researchers carefully plan and design subject ascertainment, randomization, interventions and outcomes measurement. A similar level of attention must be paid to the inclusion of genomic data to maximize the ability of the field to find true associations. We present a series of

Corresponding authors: Brion Maher, PhD, 624 N. Broadway Ave., Baltimore, MD 21205, brion@jhu.edu. Shawn Latendresse, PhD, One Bear Place #97334, Waco, TX 76798, Shawn_Latendresse@baylor.edu.

Compliance with Ethical Standards

Claims of potential conflict of interest. Drs. Latendresse, Musci and Maher have no potential conflicts of interest to report.

Ethical approval. For this type of study ethical approval is not required.

Formal consent: For this type of study formal consent is not required.

considerations and solutions to dealing with large scale genomic data that need to be addressed before, during or after actual genetic or epigenetic association testing aimed at limiting the genetic association replication crisis in the context of intervention and prevention studies largely predicated on the notion that randomized trials, which will be relatively underpowered when compared to large, genome-wide association scan consortia, can glean information from those GWAS (i.e., Post-GWAS) to limit testing to those loci (genetic, epigenetic) that have been previously associated with similar outcomes. (Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis, 2001).

In an ideal situation, human geneticists elucidate the genetic basis of a disorder by identifying a missing or aberrant protein, or a candidate gene, which can be associated with the trait. Although this approach has been successful for many single-gene human diseases, it has not been very effective in the study of more complex outcomes such as psychiatric/behavioral disorders, which are the result of small contributions of hundreds or thousands of genetic variants. Prior to the development of molecular genetic markers, other methods were utilized to provide evidence that genetic factors are important for the variation in behavioral and psychiatric outcomes. Those methods included twin, adoption and family studies, each testing specific hypotheses regarding the relationship between genetic and phenotypic similarity, that is, disorders or traits with a genetic component should be more concordant among those who are more closely genetically-related. While providing evidence for a genetic, or heritable, contribution to a trait or disease, these approaches were not designed to identify the contribution of specific genes or polymorphisms (a variation in the genetic sequence present in >1% of the population).

The development of technology to measure genetic polymorphisms shifted the approach to linkage analysis, relying primarily on repeat polymorphisms, such as microsatellites, to test the cosegregation of chromosomal regions with disease in families. Concurrently, there was a growing focus on association mapping, where a polymorphism in or near a gene putatively underlying the pathophysiology of the disease, a “candidate gene”, is tested for allele frequency differences between cases and controls. Unfortunately, the candidate gene era, even in cases where the neurobiology underlying specific behaviors or psychiatric disorders was thought to be understood, was largely unsuccessful at identifying replicating variants influencing a phenotype or outcome (Farrell et al., 2015). However, in whole, these lines of investigation have contributed greatly to our understanding of the nature of a host of human traits and diseases and are comprehensively reviewed elsewhere (Zandi, Wilcox, Dong, Chon, & Maher, 2012)

Subsequent progress in genetic technology, the development of a dense set of single nucleotide polymorphisms (SNPs) that capture a substantial proportion of common genetic variation across the genome and the assumption that common alleles with moderate effect sizes were largely responsible for observed heritability, created conditions for genetic studies of disorders of complex etiologic architecture, such as behavioral and psychiatric outcomes. The naiveté that flourished at the onset of the era of the genome-wide association study (GWAS), the primary approach to assessing the impact of genetic variation on common diseases or phenotypes in which hundreds of thousands or millions of single nucleotide polymorphisms (SNPs) spaced throughout the human genome are tested without any

theoretical justification, has largely waned. A major drawback to this approach is that the signals expected for complex diseases are unlikely to meet strict thresholds for multiple test correction that are necessary when hundreds of thousands or millions of tests are performed, and true signals are likely to be blended with false signals (Zaykin & Zhivotovsky, 2005). Though initially deemed “unsuccessful”, larger sample sizes, needed to overcome the multiple testing burden consequent to conducting 10^6 hypothesis tests, have yielded many successes and GWAS has proven useful in identifying some regions influencing variation in psychiatric/behavioral traits, a trend that is expected to continue. Additionally, polygenic approaches, which index tens, hundreds, or thousands of SNPs to create composite indices of genetic risk, have been developed (Maher, 2015).

A more recent focus is on epigenetics. For the purpose of discussion here, epigenetics is defined as genomic influences, other than the actual base (A,C,T,G) sequence, which impact on gene expression and ultimately on disease risk. While there are several classes of epigenetic changes discussed in the literature, the focus here is on DNA methylation. Briefly, DNA methylation refers to the addition of a methyl group to cytosine in areas of the genome that are enriched with C and G nucleotides. While the technical details of methylation are beyond the scope of this manuscript, the impact of DNA methylation may be central to understanding how gene expression is regulated. Important regions of many genes are potential targets of methylation, including the promoter region. Thus, methylation is a mechanism by which the expression of a gene can be reduced or silenced, including in the differentiation of tissues and specialization of cells. For decades, the role of methylation in important biological processes, such as X-chromosome inactivation, has been well known. Recently, the role of methylation in response to the environment coupled with the development of new array-based technologies allows for the assessment of variation in methylation at sites throughout the genome. Epidemiologic studies have identified specific environmental risk factors. However, biologic pathways along which environments “get under the skin” and influence mental health have only begun to be investigated.

Understanding the interplay between genes and environment remains at the forefront of research in the behavioral sciences into many developmental and disease etiology processes. Over the past few years several informative critical reviews have been published highlighting the successes, challenges and hype surrounding the investigation of gene-environment interaction (Dick et al., 2015; Duncan & Keller, 2011). Some of the issues are the relatively straightforward problems that plague many areas of research, such as sample size and publication bias. Other issues like hype, interpretation, replication and impact are not unique to GxE testing but shared across much of genetic association testing (Ioannidis et al., 2001). The replication problem is best exemplified by a case where multiple groups dealing with the same set of available data reached conflicting conclusions regarding the mediation of the relationship between stress and depression by 5-HTTLPR, a finding (Caspi et al., 2003) that received unparalleled attention and nearly fostered a subfield aimed at its replication (Munafò, Durrant, Lewis, & Flint, 2009; Risch et al., 2009).

Herein, we present general guidelines for performing high quality genetic and epigenetic analyses while avoiding the pitfalls that commonly occur. These issues include genotype quality control, correction for population stratification, and genotype imputation. We also

discuss single marker modelling of genetic and epigenetic data. Lastly, we discuss approaches frequently applied after single marker testing including polygenic and pathway approaches, and the inclusion of functional data in association tests. We also highlight specific topics that are unique to studies of intervention or prevention.

Quality Control

Maximizing genotype accuracy is a key step in increasing the power to detect true genotype-phenotype relationships. Marker and subject-level checks are performed to ensure data precision. These steps are more thoroughly described elsewhere (Anderson et al., 2010) but reviewed here briefly. First, on a per-marker level, Hardy-Weinberg equilibrium (HWE) is tested and markers that exhibit large deviations from the expected distribution of genotypes, given the observed allele frequencies, are removed. The rationale for this step is to eliminate any markers that may exhibit evidence of systematic genotype error (e.g., excess failure of a particular allele). It is important to recognize that minor deviations from HWE will be observed by chance when testing such a large number of markers. Thus, it is typical to use a stringent criterion for dropping markers exhibiting Hardy-Weinberg disequilibrium ($p < 0.001$). In addition, markers exhibiting sample-wide genotype call rates of less than 95% can also be eliminated. A high frequency of missing data is generally interpreted as an indicator of a poor quality marker. If cases and controls are present in a dataset then differences in missing rates between cases and controls can be used as a criterion for marker elimination. Lastly, although this step has become less common with an increasing interest in rare variants, SNPs with a minor allele frequency below a threshold are eliminated in the interest of power. On the per subject level, individuals for whom greater than 5% of the total markers assessed fail to be genotyped (i.e., cannot be called or are otherwise missing) are also eliminated, as this is likely due to poor DNA quality. In addition, subjects with a genetic sex (e.g., heterozygous X chromosome markers in a subject purported to be male) inconsistent with the reported sex are eliminated. Lastly, the availability of large-scale genetic data allows for the estimation of identity-by-state between individuals in a dataset. Identity-by-state can be used as a proxy for identity-by-descent, or the degree of allele sharing due to common ancestry, a common measure of relatedness. Obviously, in samples where families are ascertained, it is expected that relatives will be present in the dataset and appropriate methods for accounting for dependence between subjects will be used in subsequent association testing. However, when unexpected relatedness among individuals is discovered (e.g., IBD/IBS $> .1875$, or midway between the expected genetic similarity of second- and third-degree relatives; Anderson et al., 2010) it is the usual practice to randomly delete all but one of the correlated observations.

Population Stratification

In the context of genetic association testing, population stratification refers to systematic genetic differences between subpopulations. This is especially problematic in instances where the investigators are blind to the presence of stratification and/or the population substrata also differ phenotypically. In these instances, association analyses will be prone to generate spurious genotype-phenotype relationships and it is especially important in prevention trials to ensure that, even after the fact, randomization properly accounted for

population stratification. Based on early results in the candidate gene era, this was thought to be a serious issue, and cause of the lack of replication in many candidate gene studies (Tabor, Risch, & Myers, 2002). Contradictory opinions notwithstanding (Hutchison, Stallings, McGeary, & Bryan, 2004), the availability of genome-wide SNP data made it very apparent that population structure or stratification was a potential source of spurious false positive results, even in samples thought to be homogeneous (Burton et al., 2007; Freedman et al., 2004). Several methods, including genomic control (Devlin & Roeder, 1999) and STRUCTURE (Pritchard & Rosenberg, 1999; Pritchard & Donnelly, 2001) are available to deal with this issue by either correcting the test statistic for the average level of stratification or a priori grouping of population subsets. At the onset of the GWAS era, the wealth of genome-wide data gave rise to additional approaches that rely on the correlation structure of genetic information to identify cryptic population structure. EIGENSTRAT (Price et al., 2006) is an example of an approach that uses genome-wide data to infer principal components of population membership. Subjects are assigned a score for each of these principal components representing their membership in a given population cluster. These scores can then be used in all subsequent analyses to account for population structure. This process is automated in PLINK 2 and the cluster PC scores are included as covariates in subsequent analyses to account for population stratification (Chang et al., 2015). In practice, to reduce the computational intensity and avoid using redundant information from SNPs in linkage disequilibrium (i.e., non-randomly associated) investigators frequently select a subset of SNPs randomly across the genome to estimate stratification using the PCA approach. Although these may or may not be a priori identified ancestry information markers, it has been shown that “randomly” selected SNPs perform equally well (Montana & Pritchard, 2004). Several investigators (Choudhry et al., 2006; Sankararaman, Sridhar, Kimmel, & Halperin, 2008) have noted that differences in global (genome-wide) versus local (at a gene or LD block) ancestry exist, especially in admixed populations (e.g., African-American). Consequently, methods have been developed to estimate local ancestry, or the proportion of ancestry at each SNP attributable to known reference populations. This approach can be applied in a genome-wide context (WinPOP/LAMP; (Pasaniuc, Sankararaman, Kimmel, & Halperin, 2009; Pasaniuc et al., 2011; Sankararaman et al., 2008)) to estimate proportions of local ancestry at each region and those estimates used to account for stratification based on those mixing proportions in subsequent tests in specific candidate regions. Importantly, Keller (Keller, 2014), suggests that although including these marker-based variables in linear models examining moderation of specific genetic associations (e.g., gene-by-intervention effects) effectively removes the confounding influence of population stratification, this is only true with respect to main effects. Thus, fully accounting for the effect of population stratification on an interaction requires the inclusion of all stratification related product terms (e.g., gene-by-stratification and intervention-by-stratification effects).

Genotype Imputation

There are numerous approaches to direct imputation of genotypes. Initially, the goal of such approaches was to exploit linkage disequilibrium to allow the testing of ungenotyped markers (Clark & Li, 2007). For example, a known, functional SNP may be highly

correlated with multiple nearby genotyped SNPs. Given an accurate reference panel as the source of correlation between markers, the un-genotyped marker can be imputed and tested for association (Clark & Li, 2007). As GWAS data became widely available, the use grew with the goal of creating common marker sets to allow meta-analysis of datasets genotyped on different GWAS panels. These approaches rely on assigning genotypes above a specified level of certainty and subsequent analysis, using standard approaches, of the resultant inferred genotypes (Lin & Huang, 2007). Genome-wide SNP imputation is commonly performed using Impute2 (Howie, Donnelly, & Marchini, 2009) or MACH (Li, Willer, Ding, Scheet, & Abecasis, 2010) on data that are pre-phased using SHAPEIT (Delaneau, Zagury, & Marchini, 2013). Pre-phasing refers to the computational process of constructing haplotypes, or linear combinations of alleles along a chromosome, prior to imputation. The primary advantage of pre-phasing is a dramatic improvement in the speed of imputation. The 1000 Genomes Project sample is commonly used as a reference (1000 Genomes Project Consortium et al., 2010). In samples where diverse genetic backgrounds are present, it is common to use multiple or all reference samples within the 1000 Genomes Project. Importantly, each approach to imputation generates a measure of imputation confidence which can be used in subsequent tests to account for uncertainty. A typical approach is to recode the genotype to capture uncertainty. For example, two alleles imputed with 99% certainty would be coded as .99 and summed to create a genotype 1.98, representing two nearly certain alleles comprising a homozygous genotype.

Statistical Model: Main Effects

Individual SNP or epigenetic marker association analyses can be performed by modeling genotype counts or methylation signature (or change) in linear or logistic models. In some models including genotype data there may be an a priori motivation for collapsing genotype groups to compare two, instead of three, genotype groups. In its simplest form this test, commonly termed the measured genotype analysis (Boerwinkle, Chakraborty, & Sing, 1986), models the influence of genotypic variation at a given locus on variation in the quantitative trait, essentially a one-way ANOVA. The extension to logistic and linear regression for genetic association testing for discrete and continuous outcome variables is common and appropriate. It is typical to code the genotype as a trichotomous indicator (0,1,2). This approach provides a distinct advantage in that it can easily be extended to incorporate measured ancestry covariates and environmental moderators or mediators. When imputed SNPs are used, it is common to replace the genotype predictor with uncertainty-adjusted dosage. The use of linear mixed models (LMM) for genetic association analysis has grown rapidly in recent years to allow for valid tests of the relationship between a measured SNP and phenotypic outcome and are especially useful in prevention trials, where longitudinal data are collected (Eu-Ahsunthornwattana et al., 2014). Simpler and more complicated models can be estimated using, for example, **lme4** in R (Bates, Maechler, Bolker, & Walker, 2014) with genotype predicting longitudinal methylation outcomes and/or methylation predicting subsequent behavioral outcomes. An additional advantage of using linear/logistic models is the easy extension to explore mechanisms (e.g., mediation or moderation) of discovered associations. For example, by employing the **mediation** package in R, specific indirect pathways wherein methylation serves as a mediator of environmental

influence(s) on subsequent behavior(s) could also be tested (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014).

Incorporating biological information

A wealth of extant biological information, including previous GWAS of similar traits, knowledge of polymorphism function, gene regulation, tissue specific expression patterns and biological organization of genes into pathways can be leveraged to improve our etiological understanding of disease and behavior. For example, an investigator may want to prioritize sites that are epigenetically modified during development in examining the potential mediators of the impact of an intervention or to focus on the genes likely to be involved in a disorder that arises during a particular developmental period (Birnbaum, Jaffe, Hyde, Kleinman, & Weinberger, 2014).

Prior information can be harnessed in multiple ways to increase the power of genetic association testing. While there are robust developments in methodology that allow for the inclusion of functional information, we highlight an existing approach that incorporates functional information while correcting for multiple testing. The weighted FDR (False Discovery Rate) allows for the inclusion of a prior value in the FDR (Roeder, Bacanu, Wasserman, & Devlin, 2006). This prior allows additional information (previous GWAS findings, biological plausibility, etc) to be included as a weight in the calculation of the FDR potentially influencing the relative ranks of p-values of a new set of analyses. For instance, investigators can up-weight SNPs with evidence a significant impact on expression or methylation variation in the brain. The approach was recently used, with success, in an ‘informed GWAS’, or iGWAS, to include prior information from previous genome-wide scans in a novel analysis (Fortney et al., 2015).

Great advances are occurring in evaluating functional significance of genomic regions and variants. Variations found by genome sequencing can be evaluated based on functional predictions, relevance to developmental mechanisms, frequency in population and disease databases, previous GWAS evidence, or known effect on expression. Noncoding variants can be evaluated based on epigenetic signatures derived from specific resources including ENCODE and Braincloud. Existing expression and methylation data to allow the discovery eQTLs and meQTLs at specific developmental periods in brain is available in BrainCloud (Colantuoni et al., 2011; Jaffe et al., 2014; Numata et al., 2012). Relevant functional data, frequently freely available, will allow for more prudent hypothesis testing than is currently routine and wFDR allows one to place a higher prior probability on genetic variants with prior evidence of function relevant to the trait of interest.

Polygenic/Pathway Approaches

Due to the individually small effect sizes of contributing loci and stringent statistical significance criterion of GWAS most modest effect size polymorphisms will not be deemed “significant” and “replicated” in many GWAS (Purcell et al., 2009). Multiple approaches to aggregating those effects have been developed, thereby increasing the power to detect a polygenic signal and aiding in understanding the nature of complex traits. Purcell and

colleagues (Purcell et al., 2009) used an approach in which two stage GWAS data were used to select a set of “independent” SNPs in linkage equilibrium that generated p-values below some arbitrary threshold (P_T) in one sample as a discovery stage. Those SNPs were then used to create polygenic sum scores, with each allele weighted by the logarithm of the odds ratio from the discovery sample, to be tested in a second sample. The terms “polygenic scores” (PGS), “genetic risk scores” (GRS) and “polygenic risk scores” (PRS) are now used interchangeably to describe metrics comprising a large number of SNPs pooled together to represent a measured set of variants underlying a particular trait or disease. Dudbridge (Dudbridge, 2013) examined the power and predictive accuracy of polygenic scores for discrete and continuous traits and found that large discovery samples, which best separate the true from null effects at the tail of the p-value distribution, yield the most precise polygenic scores. In the post-GWAS era, the use of existing GWAS results, such as those from the Psychiatric Genomics Consortium, can be used as “discovery” samples from which to generate polygenic scores for a novel study. Tools for automated creation of polygenic scores are available in the Plink 2 software package (Chang et al., 2015).

In addition to using polygenic scores as predictors in statistical models testing the impact of many genetic influences on a particular outcome, approaches have been developed to estimate what could accurately be termed “marker-based” or “molecular” heritability. The general approach of this method uses genome-wide markers to examine deviation from expected genetic-phenotypic similarity (Visscher et al., 2006). Methods such as genome-wide complex trait analysis (GCTA) rely on population-based samples with available GWAS data (Yang et al., 2011). In GCTA, a genetic relationship matrix is derived from all available SNPs and used to estimate the proportion of phenotypic variation accounted for by the genome-wide genetic differences. Importantly, while this method can generate an estimate of the phenotypic variation accounted for by genetic differences, it does not identify specifically which variants or pathways account for it. Extensions of the method claim to improve accuracy by correcting for linkage disequilibrium, as opposed to pruning out correlated (but possibly true independent effect) SNPs (Vilhjalmsson et al., 2015) or by relying on HaploSNPs (Bhatia et al., 2015), large shared segments in high LD that can be recoded into a regional genotype. Overall, GCTA and its extensions would be useful in describing and testing the polygenic architecture of a trait, for example response to treatment or intervention, but have limited utility in defining the specific genes or pathways involved.

A more useful approach for prevention, intervention, and treatment genetic studies are those that jointly test multiple genetic predictors that are grouped in the same biological pathways. The basic motivation for pathway-based analyses is the high likelihood that genetic associations with an outcome will co-occur in SNPs grouped within the same biological pathway. There are two broad classes of pathway analysis: those that test whether an excess of statistically significant results occur in SNPs in a pathway and those that test whether the top signals are more closely related biologically than by chance. Commonly used approaches include DAPPLE (Rossin et al., 2011), Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), ALIGATOR (Holmans et al., 2009), MAGENTA (Segrè et al., 2010), FORGE (Pedroso et al., 2012) and INRICH (Lee, O’Dushlaine, Thomas, & Purcell, 2012). Although varying in methodology, these approaches rely on gene sets or pathways defined in specific databases (e.g., KEGG, Gene Ontology) to organize SNPs for excess

statistical significance within a pathway, while accounting for potential confounders like gene size and LD pattern within those genes. In addition to differences among the methodologies, it is also important to note that an inherent limitation of these approaches is the accuracy with which genes are grouped into pathways. Consequently, given the relative strengths and weaknesses of some existing pathway-based approaches and pathway databases, a combined rank approach has recently been developed for aggregating data across multiple approaches (Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium, 2015). Importantly, these methods can be applied to specific candidate pathways, for example “stress response”, if a particular set of genes is thought to impact an outcome of interest.

Methylation

Epidemiologic studies have identified specific childhood environmental exposures as substantial risk factors for subsequent behavioral disorders. Epigenetic processes mediate the impact of environmental influences (e.g., life experiences) on risk of illness through regulation of gene expression and function. The role of stressors in epigenetic variation in animal models is well established but extant human epigenetic studies are relatively small and typically rely on simple phenotypes (Tsankova, Renthal, Kumar, & Nestler, 2007). However, preliminary studies reporting associations between methylation and aspects of addiction (Hopf & Bonci, 2010), prenatal stress (Oberlander et al., 2008), childhood abuse (McGowan et al., 2009), PTSD (Uddin et al., 2010), and depression (Uddin et al., 2011) provide motivation for the inclusion of epigenetic measures in exploration of the relationship between intervention and later behavioral outcomes. Although acquired adverse epigenetic changes were once thought to be permanent, new evidence suggests they are plastic and potentially reversible (Kelly, De Carvalho, & Jones, 2010), opening the possibility for the impact of targeted interventions.

The role of methylation as a relatively stable marker of promoter-mediated regulation of gene expression makes it a logical target for understanding the mechanisms of environmental exposure. Moreover, methylation is known to vary both between and within individuals assessed at multiple time-points (Langevin et al., 2011). Evidence suggests several important driving forces behind differential methylation including, underlying genetic variation (meQTLs) (Smith et al., 2014), life experiences, differential tissue and cell-types, and chronological age (Horvath, 2013). Variation in DNA methylation has been investigated as a mediator of the physiologic responses to acute and chronic exposures, including environmental adversity or response to intervention/prevention. Longitudinal data allows for the examination of change in epigenetic profiles potentially reflecting response to an intervention which potentially confer liability to outcome phenotypes. Identification of these mechanisms will inform prevention strategies at the most critical point(s) in the developmental risk trajectory.

Accurate measurement of DNA methylation provides the potential for a molecular record of response to the environment. Development of novel array technology for surveying CpG methylation across the genome has led to an explosion in the study of the relationship between the epigenome and behavior. While these arrays can be used with any tissue type,

the most common use is in whole blood samples. Consequently, our discussion here assumes use of whole blood samples. Epigenomic microarray data present a set of additional challenges beyond those posed by genetic microarray data. Comprehensive discussion is presented elsewhere but, in short, quality assessment, scaling and normalization, and removal of batch and other technical artifacts must be performed prior to analysis using packages such as minfi (Aryee et al., 2014). A particularly important aspect of methylation analysis is the removal confounding due to cellular heterogeneity. Importantly, even within a sample of lymphocytes, each cell type (e.g., B cell, T cell, NK cell) will have a unique epigenomic profile and heterogeneity in the relative cell proportions between individuals can lead to spurious results. Cellular heterogeneity is frequently corrected using the approach of Houseman and colleagues (Houseman et al., 2012). After correction and normalization, a quantitative metric representing percent methylation can be used as a dependent or independent variable in subsequent analysis.

Gene-Environment interplay

Integration of the behavioral sciences with genetics necessitates the simultaneous testing of the impact of genes and environment on phenotypic outcome. A usual approach is to explicitly test for an interaction between a specific genetic variant and an environment variable (GxE), whether or not a specific a priori hypothesis exists. While generating a host of apparently significant and often high profile findings, this approach has been met with non-replication and a great deal of criticism. While these issues are discussed in depth elsewhere, we will briefly mention several pitfalls to avoid. As with all statistical model fitting, it is important to be aware of distributional assumptions. However, given awareness of the potential impact of violations of distributional assumptions (e.g., Gaussian) in environmental measures on subsequent testing of interaction effects often leads investigators to impose artificial thresholds (e.g., median splits) on environmental indicators, consequently reducing statistical power. We recommend the use of quantile normalization as an alternative approach (Irizarry et al., 2003). Some investigators attempt to rely on methods that omit the main effects of the gene and environment and model only the interaction. Discussed elsewhere (Keller, 2014), this approach will yield inaccurate results. It is essential to include the main effect of any variable in a model that tests for its interaction with another variable to avoid an increase in Type I error. This increase in Type I error is magnified in models where gene-environment correlation is present (rGE). Recognizing the potential for spurious results, many statistical packages (including R), prohibit users from directly modelling interactions alone. A workaround used by some investigators is to pre-compute the interaction term (literally $G \times E$) and test it as the lone independent variable in the model. This leads to highly spurious results when there are marginal effects of the gene or environment but no effect of GxE. Caution is advised when using this approach. It is well known that mean-centering of (quasi)continuous variables is requisite for removing any collinearity between an interaction (i.e., product) term and its component predictors, and avoids spurious results (Aiken and West 1991). Given proper treatment of genetic and environmental indicators, GxE interaction can be formally tested in many statistical models, including linear mixed models.

Interpretation of a statistically significant GxE interaction is of particular interest. In the context of behavioural outcomes, two competing models exist for explaining gene-environment interplay. The diathesis stress model (Monroe & Simons, 1991) and differential susceptibility model (Belsky, 1997). Others have referred to these as fan-shaped vs cross-over interactions, respectively, due to their plotted appearance (Roisman et al., 2012). In the differential susceptibility model a particular genotype will perform significantly better under one environmental condition versus other genotypes and will also perform significantly worse than other genotype under an alternative environmental condition (a cross-over). In the diathesis-stress model, genotypes differ under only one environmental condition (a fan-shape). Frequently investigators plot interactions with the environment on the x-axis and phenotype on the y-axis, with separate lines per genotype, and draw conclusions based on visual inspection. However, formal statistical procedures exist for determining regions of significance, values of the environmental exposure under which genotypes differ significantly on phenotypic measures. In the instance of discrete environmental measures, this approach is simple since it involves comparing genotypes on phenotypic measure at each level of the environment to determine where they differ significantly. For continuous measures of exposure this can be reformulated as the “pick-a-point” or simple slopes approach, where significant differences in mean or slope can be tested at specific values of the continuous environment (Rogosa, 1980). The obvious weakness of this approach is its reliance on selection of arbitrary environmental values for testing. A preferred approach relies on the calculation of regions of significance or the range of values of the environment where the genotype groups differ significantly on phenotype values. As reviewed in Preacher et al (Preacher, Curran, & Bauer, 2006), this can be accomplished using the Johnson-Neyman (Johnson & Neyman, 1936) approach or the use of confidence bands to identify the range of environmental measures on which the values differ on slope.

An important consideration in the context of gene-by-intervention analyses is that the differential susceptibility model is not specifically testable, as there is no “negative” treatment condition under which the crossover can be observed. More comprehensive treatments of the topic are available elsewhere (for example, (Manuck & McCaffery, 2014))

Genetics within Randomized Control Trials

Prevention trials provide a unique opportunity for researchers to explore gene-environment interplay, particularly within the context of differential susceptibility (Bakermans-Kranenburg & van IJzendoorn, 2015). The randomization of participants to trial arms allows for careful evaluation of the role of genetics as well as the role of the environment. This is specifically relevant because of the careful manner in which the environment is controlled, thus limiting potential gene-environment correlation. Further, these trials generally provide rich phenotypic information for more careful modelling. There are some limitations, the largest being the limited sample sizes. Researchers can improve the power associated with these statistical tests by limiting the number of genetic regions explored through use of Post-GWAS candidate selection, polygenic scoring or pathway analysis. Previous work has explored differential susceptibility utilizing meta-analytic methods, and demonstrated the remarkable increase in power for detecting GxE in intervention trials (Bakermans-Kranenburg & Van IJzendoorn, 2015).

Conclusion

Inclusion of genetic and epigenetic measures in longitudinal prevention studies will allow for the elucidation of the mechanism(s) by which preventive measures operate, including the detection of biological moderators and/or mediators of effectiveness. Multiple steps must be taken to ensure valid results. Appropriate quality control steps, at the SNP and individual level, must be taken to avoid inclusion of problematic genetic markers and contaminated or swapped individual samples. Careful imputation of markers using a valid reference panel will allow for the testing of nearly all genomic variation, while accounting for the uncertainty inherent to imputation. Proper approaches to accounting for genetic ancestry avoids spurious association due to population stratification. Ancestry principal components can be used as covariates in subsequent models testing genotype-phenotype relationships to prevent false positive results due to stratification. Polygenic scoring is a powerful approach to capturing composite genetic risk and may be useful for examining the impact of prevention or intervention efforts across the biological risk spectrum. Inclusion of polygenic risk as the lone genetic predictor, or in addition to a single SNP, in association testing is a valid strategy for jointly testing association. The growth of large scale curated databases of gene function, tissue specific and developmental timing of expression, methylation sensitive sites, and genomic functional annotations provide relevant prior information that can be used in weighted hypothesis testing approaches such as the weighted FDR. The increased availability of arrays to provide genome-wide indicators of baseline and change in site-specific methylation is a boon to those interested in exploring the mechanisms by which intervention may serve to impact the genome. As with genotype data, great care must be taken with methylation array data to ensure that technical artefacts are removed and data are properly normalized before conducting association tests. Since methylation signals can change over time, a new set of mechanistic hypotheses, including whether or not a particular site mediates the impact of an intervention, are possible. Lastly, GxE testing is at the core of inclusion of genetic measures in intervention and prevention studies (i.e., does modifying the environment impact outcome differently by genotype). Consequently, safeguards must be taken to avoid the generation of spurious results.

Acknowledgments

Funding. This work was supported by National Institute on Drug Abuse (NIDA) Grants R01DA036525 and R01DA039408, and National Institute on Alcoholism and Alcohol Abuse Grant K01AA020333.

References

- Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, ... McVean GA. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467(7319):1061–1073. DOI: 10.1038/nature09534 [PubMed: 20981092]
- Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nature Protocols*. 2010; 5(9):1564–1573. [PubMed: 21085122]
- Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics (Oxford, England)*. 2014; 30(10):1363–1369. DOI: 10.1093/bioinformatics/btu049

- Bakermans-Kranenburg MJ, van IJzendoorn MH. The hidden efficacy of interventions: Gene× environment experiments from a differential susceptibility perspective. *Annual Review of Psychology*. 2015; 66:381–409.
- Bates D, Maechler M, Bolker B, Walker S. *lme4: Linear mixed-effects models using eigen and S4*. R Package Version. 2014; 1(7)
- Belsky J. Variation in susceptibility to environmental influence: An evolutionary argument. *Psychological Inquiry*. 1997; 8(3):182–186.
- Bhatia G, Gusev A, Loh P, Vilhjálmsón BJ, Ripke S, Purcell S, ... Kendler KS. Haplotypes of common SNPs can explain missing heritability of complex diseases. *bioRxiv*. 2015 022418.
- Birnbaum R, Jaffe AE, Hyde TM, Kleinman JE, Weinberger DR. Prenatal expression patterns of genes associated with neuropsychiatric disorders. *American Journal of Psychiatry*. 2014
- Boerwinkle E, Chakraborty R, Sing C. The use of measured genotype information in the analysis of quantitative phenotypes in man. *Annals of Human Genetics*. 1986; 50(2):181–194. [PubMed: 3435047]
- Brody GH, Yu T, Chen E, Beach SR, Miller GE. Family-centered prevention ameliorates the longitudinal association between risky family processes and epigenetic aging. *Journal of Child Psychology and Psychiatry*. 2015
- Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, ... Samani NJ. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447(7145):661–678. [PubMed: 17554300]
- Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, ... Poulton R. Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science (New York, NY)*. 2003; 301(5631):386–389. DOI: 10.1126/science.1083968
- Chang CC, Chow CC, Tellier L, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*. 2015; 4(7)
- Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL. ... Genetics of Asthma in Latino Americans GALA Study. Population stratification confounds genetic association studies among latinos. *Human Genetics*. 2006; 118(5):652–664. DOI: 10.1007/s00439-005-0071-3 [PubMed: 16283388]
- Clark AG, Li J. Conjuring SNPs to detect associations. *Nature Genetics*. 2007; 39(7):815–816. [PubMed: 17597769]
- Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, ... Kleinman JE. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature*. 2011; 478(7370):519–523. DOI: 10.1038/nature10524 [PubMed: 22031444]
- Delaneau O, Zagury J, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*. 2013; 10(1):5–6. [PubMed: 23269371]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999; 55(4):997–1004. [PubMed: 11315092]
- Dick DM, Agrawal A, Keller MC, Adkins A, Aliev F, Monroe S, ... Sher KJ. Candidate gene-environment interaction research: Reflections and recommendations. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*. 2015; 10(1):37–59. DOI: 10.1177/1745691614556682 [PubMed: 25620996]
- Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet*. 2013; 9(3):e1003348. [PubMed: 23555274]
- Duncan LE, Keller MC. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *American Journal of Psychiatry*. 2011
- Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SM, Blackwell JM, Cordell HJ. Wellcome Trust Case Control Consortium 2. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet*. 2014; 10(7):e1004445. [PubMed: 25033443]
- Farrell M, Werge T, Sklar P, Owen M, Ophoff R, O'donovan M, ... Sullivan PF. Evaluating historical candidate genes for schizophrenia. *Molecular Psychiatry*. 2015; 20(5):555–562. [PubMed: 25754081]

- Fortney K, Dobriban E, Garagnani P, Pirazzini C, Monti D, Mari D, ... Owen AB. Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS Genet.* 2015; 11(12):e1005728. [PubMed: 26677855]
- Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, ... Pato CN. Assessing the impact of population stratification on genetic association studies. *Nature Genetics.* 2004; 36(4):388–393. [PubMed: 15052270]
- Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, ... Craddock N. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *American Journal of Human Genetics.* 2009; 85(1):13–24. DOI: 10.1016/j.ajhg.2009.05.011 [PubMed: 19539887]
- Hopf FW, Bonci A. Dnmt3a: Addiction's molecular forget-me-not? *Nature Neuroscience.* 2010; 13(9): 1041–1043. DOI: 10.1038/nn0910-1041 [PubMed: 20740031]
- Horvath S. DNA methylation age of human tissues and cell types. *Genome Biology.* 2013; 14(10): 3156.
- Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, ... Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 2012; 13:86-2105-13-86.doi: 10.1186/1471-2105-13-86 [PubMed: 22568884]
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009; 5(6):e1000529. [PubMed: 19543373]
- Hutchison KE, Stallings M, McGeary J, Bryan A. Population stratification in the candidate gene study: Fatal threat or red herring? *Psychological Bulletin.* 2004; 130(1):66. [PubMed: 14717650]
- Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nature Genetics.* 2001; 29(3):306–309. [PubMed: 11600885]
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England).* 2003; 4(2):249–264. DOI: 10.1093/biostatistics/4.2.249
- Jaffe AE, Gao Y, Tao R, Hyde TM, Weinberger DR, Kleinman JE. The methylome of the human frontal cortex across development. *bioRxiv.* 2014; doi: 10.1101/005504
- Johnson PO, Neyman J. Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs.* 1936
- Keller MC. Gen \times environment interaction studies have not properly controlled for potential confounders: The problem and the (simple) solution. *Biological Psychiatry.* 2014; 75(1):18–24. [PubMed: 24135711]
- Kelly TK, De Carvalho DD, Jones PA. Epigenetic modifications as therapeutic targets. *Nature Biotechnology.* 2010; 28(10):1069–1078. DOI: 10.1038/nbt.1678
- Langevin SM, Houseman EA, Christensen BC, Wiencke JK, Nelson HH, Karagas MR, ... Kelsey KT. The influence of aging, environmental exposures and local sequence features on the variation of DNA methylation in blood. *Epigenetics : Official Journal of the DNA Methylation Society.* 2011; 6(7):908–919.
- Lee PH, O'Dushlaine C, Thomas B, Purcell SM. INRICH: Interval-based enrichment analysis for genome-wide association studies. *Bioinformatics (Oxford, England).* 2012; 28(13):1797–1799. DOI: 10.1093/bioinformatics/bts191
- Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology.* 2010; 34(8):816–834. [PubMed: 21058334]
- Lin DY, Huang BE. The use of inferred haplotypes in downstream analyses. *American Journal of Human Genetics.* 2007; 80(3):577–579. [PubMed: 17380613]
- Maher BS. Polygenic scores in epidemiology: Risk prediction, etiology, and clinical utility. *Current Epidemiology Reports.* 2015; 2(4):239–244. [PubMed: 26664818]
- Manuck SB, McCaffery JM. Gene-environment interaction. *Annual Review of Psychology.* 2014; 65:41–70.

- McGowan PO, Sasaki A, D'Alessio AC, Dymov S, Labonte B, Szyf M, ... Meaney MJ. Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nature Neuroscience*. 2009; 12(3):342–348. DOI: 10.1038/nn.2270 [PubMed: 19234457]
- Monroe SM, Simons AD. Diathesis-stress theories in the context of life stress research: Implications for the depressive disorders. *Psychological Bulletin*. 1991; 110(3):406. [PubMed: 1758917]
- Montana G, Pritchard JK. Statistical tests for admixture mapping with case-control and cases-only data. *American Journal of Human Genetics*. 2004; 75(5):771–789. [PubMed: 15386213]
- Munafò MR, Durrant C, Lewis G, Flint J. Gene× environment interactions at the serotonin transporter locus. *Biological Psychiatry*. 2009; 65(3):211–219. [PubMed: 18691701]
- Musci RJ, Masyn KE, Uhl G, Maher B, Kellam SG, Ialongo NS. Polygenic score× intervention moderation: An application of discrete-time survival analysis to modeling the timing of first tobacco use among urban youth. *Development and Psychopathology*. 2015; 27(01):111–122. [PubMed: 25640834]
- Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience*. 2015; 18(2):199–209. DOI: 10.1038/nn.3922 [PubMed: 25599223]
- Numata S, Ye T, Hyde TM, Guitart-Navarro X, Tao R, Wininger M, ... Lipska BK. DNA methylation signatures in development and aging of the human prefrontal cortex. *American Journal of Human Genetics*. 2012; 90(2):260–272. DOI: 10.1016/j.ajhg.2011.12.020 [PubMed: 22305529]
- Oberlander TF, Weinberg J, Papsdorf M, Grunau R, Misri S, Devlin AM. Prenatal exposure to maternal depression, neonatal methylation of human glucocorticoid receptor gene (NR3C1) and infant cortisol stress responses. *Epigenetics : Official Journal of the DNA Methylation Society*. 2008; 3(2):97–106.
- Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. *Bioinformatics (Oxford, England)*. 2009; 25(12):i213–21. DOI: 10.1093/bioinformatics/btp197
- Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Kao WH, ... Price AL. Enhanced statistical tests for GWAS in admixed populations: Assessment using african americans from CARE and a breast cancer consortium. *PLoS Genetics*. 2011; 7(4):e1001371.doi: 10.1371/journal.pgen.1001371 [PubMed: 21541012]
- Pedroso I, Lourdasamy A, Rietschel M, Nöthen MM, Cichon S, McGuffin P, ... Breen G. Common genetic variants and gene-expression changes associated with bipolar disorder are over-represented in brain signaling pathway genes. *Biological Psychiatry*. 2012; 72(4):311–317. [PubMed: 22502986]
- Preacher KJ, Curran PJ, Bauer DJ. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*. 2006; 31(4):437–448.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*. 2006; 38(8):904–909. [PubMed: 16862161]
- Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theoretical Population Biology*. 2001; 60(3):227–237. [PubMed: 11855957]
- Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics*. 1999; 65(1):220–228. [PubMed: 10364535]
- Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009; 460(7256):748–752. [PubMed: 19571811]
- Risch N, Herrell R, Lehner T, Liang K, Eaves L, Hoh J, ... Merikangas KR. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: A meta-analysis. *Jama*. 2009; 301(23):2462–2471. [PubMed: 19531786]
- Roeder K, Bacanu SA, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. *American Journal of Human Genetics*. 2006; 78(2):243–252. [PubMed: 16400608]

- Rogosa D. Comparing nonparallel regression lines. *Psychological Bulletin*. 1980; 88(2):307.
- Roisman GI, Newman DA, Fraley RC, Haltigan JD, Groh AM, Haydon KC. Distinguishing differential susceptibility from diathesis–stress: Recommendations for evaluating interaction effects. *Development and Psychopathology*. 2012; 24(02):389–409. [PubMed: 22559121]
- Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, ... Daly MJ. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genetics*. 2011; 7(1):e1001273.doi: 10.1371/journal.pgen.1001273 [PubMed: 21249183]
- Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. *American Journal of Human Genetics*. 2008; 82(2):290–303. DOI: 10.1016/j.ajhg.2007.09.022 [PubMed: 18252211]
- Segrè AV, Groop L, Mootha VK, Daly MJ, Altshuler D. Diagram Consortium & Magic Investigators. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet*. 2010; 6(8):e1001058. [PubMed: 20714348]
- Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, ... Conneely KN. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*. 2014; 15 145-2164-15-145. doi: 10.1186/1471-2164-15-145
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, ... Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(43):15545–15550. 0506580102 [pii]. [PubMed: 16199517]
- Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: Practical considerations. *Nature Reviews Genetics*. 2002; 3(5):391–397.
- Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. Mediation: R package for causal mediation analysis. 2014
- Tsankova N, Renthal W, Kumar A, Nestler EJ. Epigenetic regulation in psychiatric disorders. *Nature Reviews Neuroscience*. 2007; 8(5):355–367. DOI: 10.1038/nrn2132 [PubMed: 17453016]
- Uddin M, Aiello AE, Wildman DE, Koenen KC, Pawelec G, de Los Santos R, ... Galea S. Epigenetic and immune function profiles associated with posttraumatic stress disorder. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107(20):9470–9475. DOI: 10.1073/pnas.0910794107 [PubMed: 20439746]
- Uddin M, Koenen KC, Aiello AE, Wildman DE, de los Santos R, Galea S. Epigenetic and inflammatory marker profiles associated with depression in a community-based epidemiologic sample. *Psychological Medicine*. 2011; 41(5):997–1007. DOI: 10.1017/S0033291710001674 [PubMed: 20836906]
- Vandenbergh DJ, Schlomer GL, Cleveland HH, Schink AE, Hair KL, Feinberg ME, ... Redmond C. An adolescent substance prevention model blocks the effect of CHRNA5 genotype on smoking during high school. *Nicotine & Tobacco Research : Official Journal of the Society for Research on Nicotine and Tobacco*. 2016; 18(2):212–220. DOI: 10.1093/ntr/ntv095 [PubMed: 25941207]
- Vilhjalmsson B, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, ... Do R. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *bioRxiv*. 2015 015859.
- Visscher PM, Medland SE, Ferreira M, Morley KI, Zhu G, Cornes BK, ... Martin NG. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet*. 2006; 2(3):e41. [PubMed: 16565746]
- Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, ... Visscher PM. Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*. 2011; 43(6):519–525. DOI: 10.1038/ng.823 [PubMed: 21552263]
- Zandi PP, Wilcox HC, Dong L, Chon S, Maher B. Genes as a source of risk for mental disorders. *Public Mental Health*. 2012:201.
- Zaykin DV, Zhivotovsky LA. Ranks of genuine associations in whole-genome scans. *Genetics*. 2005; 171(2):813–823. [PubMed: 16020784]