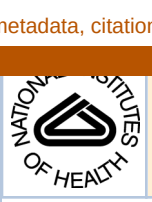

Published in final edited form as:

Twin Res Hum Genet. 2013 April ; 16(2): 505–515. doi:10.1017/thg.2013.6.

Genes, environments, and developmental GEWIS: Methods for a multi-site study of early substance abuse

E. J. Costello, Ph.D.[1], **Lindon Eaves, Ph.D.**[2], **Patrick Sullivan, M.D FRANZCP**[3], **Martin Kennedy, Ph.D.**[4], **Kevin Conway, Ph.D.**[5], **Daniel E. Adkins, Ph.D.**[6], **A. Angold, MRCPsych**[1], **Shaunna L Clark, Ph.D.**[6], **Alaattin Erkanli, Ph.D**[1], **Joseph L McClay, Ph.D.**[6], **William Copeland, Ph.D.**[1], **Hermine H. Maes, Ph.D.**[2], **Youfang Liu, Ph.D.**[7], **Ashwin A. Patkar, M.D**[1], **Judy Silberg, Ph.D.**[2], and **Edwin van den Oord, Ph.D.**[2]

[1]Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Durham, NC 27710, USA [2]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, PO Box 980003, Richmond VA 23298-0003 [3]Department of Genetics, 120 Mason Farm Road, 5097 Genetic Medicine Building, CB#7264, University of North Carolina at Chapel Hill, NC 27599-7264 [4]Department of Pathology, University of Otago, Christchurch P.O. Box 4345, Christchurch, New Zealand [5]Division of Epidemiology, Services and Prevention Research, National Institute on Drug Abuse, 6001 Executive Boulevard, Suite 5185, Bethesda, MD 20892-9589 [6]Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University School of Pharmacy, Richmond VA 23298 [7]Thurston Arthritis Research Center, University of North Carolina at Chapel Hill, CB# 7280, 3330 Thurston Building, Chapel Hill, NC 27599-7280

Abstract

The importance of including developmental and environmental measures in genetic studies of human pathology is widely acknowledged, but few empirical studies have been published. Barriers include the need for longitudinal studies that cover relevant developmental stages and for samples large enough to deal with the challenge of testing gene-environment-development interaction. A solution to some of these problems is to bring together existing data sets that have the necessary characteristics. As part of the NIDA-funded Gene-Environment-Development Initiative (GEDI) our goal is to identify exactly which genes, which environments, and which developmental transitions together predict the development of drug use and misuse. Four data sets were used whose common characteristics include (1) general population samples including males and females; (2) repeated measures across adolescence and young adulthood; (3) assessment of nicotine, alcohol and cannabis use and addiction; (4) measures of family and environmental risk; and (5) consent for genotyping DNA from blood or saliva. After quality controls, 2,962 individuals provided over 15,000 total observations. In the first gene-environment analyses, of alcohol misuse and stressful life events, some significant gene-environment and gene-development effects were identified. We conclude that in some circumstances, already-collected data sets can be combined for gene-environment and gene-development analyses. This greatly reduces the cost and time needed for this type of research. However, care must be taken to ensure careful matching across studies and variables.

Corresponding author: E. J.Costello, Ph.D., Professor, Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Box 3454 DUMC, Durham NC 27710; Mailing address: Suite 22, Brightleaf Square, 905 West Main Street, Durham NC 27701, Telephone 919 687 4686 estension 230, Fax 919 687 4737, jcostelKS)psych.duhs.duke.edu.

INTRODUCTION

"The focus of the next generation of gene-environment research will add development into the equation and focus upon gene-environment-development interactions" (Rose and Dick 2010) (pp.1854-1855).

It is clear from twin studies that the relative significance of genetic and environmental factors changes across some stages of development, notably childhood and adolescence (reviewed in Dick 2011) Indeed, some have suggested that one reason for failures to replicate gene-environment findings may be the different developmental status of the samples. For example, Cole and colleagues have argued that one explanation for the evolutionary maintenance of genetic features that are maladaptive late in life is that they must be selectively advantageous earlier in life "or else they would have been eradicated from the gene pool by adverse selection" (Cole, Arevalo et al. 2011) p. 1174. But there are many genetically driven changes with a strong developmental component, such as puberty or menopause, or the timing and synchronization of myelination in the central nervous system (van Ijzendoorn, Bakermans-Kranenburg et al. 2011).

Grappling with gene-environment-development in the same study is important not only empirically but also methodologically. "We have known for decades that failure to incorporate both genetic and environmental factors in a joint analysis will weaken the observed associations between a true risk factor and disease occurrence. Because the pools of susceptible and non-susceptible persons are mixed, the observed associations tend to be shifted toward the null..Theoretically, if we are able to measure gene-environment interactions, we should sharpen our measurements of effects in subsets of the population and even potentially increase our statistical power in measuring such effects." (Khoury and Wacholder 2009) (p.228). The same argument can be applied to the importance of incorporating developmental measures (e.g., Chiang, McMahon et al. 2011).

Longitudinal studies with repeated, prospective assessments using standardized measures of phenotype and envirotype offer several opportunities for improved data quality compared with the standard methods of genetic case-control studies, which tend to rely on lifetime retrospective assessments using multiple diagnosticians and uncertain control over the uniformity of diagnostic criteria. In a longitudinal study diagnoses are likely to be more consistent and there is likely to be less recall bias in participants' psychiatric histories, which means fewer false negatives. Both of these qualities increase statistical power in hypothesis testing (Anastasi 1950; Luan, Wong et al. 2001; Wong, Day et al. 2003; Wong, Day et al. 2004). The timing of environmental events relative to the onset of a disorder is also likely to be more accurate and less vulnerable to "seeking after meaning" (Spatola, Scaini et al. 2011…).

A fourth reason to explore the use of longitudinal studies for genetics in that each participant provides multiple measures of phenotype and envirotype. This increases the total number of observations, even after the corrections necessary to deal with the wave-to-wave correlations within individuals {Dunlap, 1996…}.

Despite theoretical discussions of the importance of GWAS-based gene-environment studies (sometimes called gene-environment-wide interactions studies: GEWIS (Khoury and Wacholder 2009)), and even some discussion of developmental GEWIS (Rose and Dick 2010; Lenroot and Giedd 2011), there are few published results of GEWIS analyses in the behavioral sciences. The few studies of developmental effects on gene-environment interplay have focused on individual genes such as BDNF(Casey, Glatt et al. 2009); none yet published uses a genome-wide approach to psychiatric disorders (Ackermann, Adams et al. 2001; Lenroot and Giedd 2011).

It is not hard to see why developmental GEWIS studies are so rare. The costs of initiating and maintaining large longitudinal studies are extremely high, and it is important that subjects be assessed using similar measures of both phenotype and environment over key developmental stages. Clinical studies rarely cover sufficiently long periods of time, so epidemiological samples are needed. Birth cohorts or other life-course epidemiologic studies, such as the nascent US National Children's Study, are potential sources.

Another solution to this problem is to bring together multiple data sets to conduct joint analyses or meta-analysis. This approach depends crucially on the ability to combine the data across studies. Even before genetic analyses can begin, it is necessary to develop and test methods for harmonizing data across studies (Cornelis, Agrawal et al. 2010; Bookman, McAllister et al. 2011; Fortier, Doiron et al. 2011).

The National Institute on Drug Abuse (NIDA) and the National Cancer Institute (NCI) recognized both the promise and the problems of developmental GEWIS when they wrote in the Request for Applications "Over many years, NIDA, other NIH Institutes, and other organizations have funded numerous high-quality longitudinal and developmental studies that contain a wealth of data from individuals who are at risk for, or are in the course of development, progression, and desistance of, substance abuse and related phenotypes….The GEDI seeks to build on this substantial public investment by soliciting applications that integrate environmental and developmental variables with genotypic information in order to permit comprehensive model-building and hypothesis testing for determining genetic, environmental, and developmental contributions to substance abuse and related phenotypes."{R01DA024413}.

NIDA and NCI hoped to take already-existing materials and see if they could be woven into something that, if created from scratch, would have taken twenty years and untold millions of dollars. If successful, GEDI would be a proof-of-concept that could lead perhaps to an expansion of the collaborative group of studies.

In summary, we report here on a proof-of-concept study to carry out gene-environment, gene-development, and gene-environment-development analyses (both parallel and meta-analytic) using longitudinal, population-based data sets with repeated measures over childhood, adolescence, and early adulthood, with DNA available or obtainable, with comparable measures of drug and alcohol use, abuse, and dependence, and also of key environmental exposures,

MATERIALS AND METHODS

Common GEDI study characteristics

The data sets that make up the consortium have the following characteristics in common: (1) general population samples; (2) multiple waves of data collection across childhood, adolescence, and young adulthood; (3) detailed assessments of drug use, abuse, and dependence (substance use disorders: SUD) and drug abuse symptoms; (4) assessments of comorbid psychiatric disorders, diagnosed using the Diagnostic and Statistical Manual (American Psychiatric Association 1994) and psychiatric symptoms scores; (5) measures of a range of environmental exposures including serious life events. Methods used to collect information on diagnoses, symptoms, and environmental factors are described first, followed by brief descriptions of each study. Table 1 presents a summary of similarities and differences. Further details can be found in study-specific publications cited below.

GEDI samples

1. Virginia Twin Study on Adolescent Behavioral Development (VTSABD) (Simonoff, Pickles etal.1997)—The VTSABD is a cohort-longitudinal study of twins born between 1974 and 1983 ascertained primarily through the state school system and participating private schools in Virginia. Of 1894 putative twin pairs, 1412 families (75%, 2775 children) participated and were included in the first wave of data collection. Three subsequent waves of data collection occurred at approximately lVi-year intervals, and a fifth wave when participants were in their mid-20s (see Table 1). The study was limited to subjects of European ancestry as insufficient numbers from other ancestry groups were ascertainable. Parents completed a similar assessment on both twins. After age 18, the twins alone were interviewed individually by telephone. Over 8,500 family interview sets (parents and twins 8-17, twins 18+) have been completed. Variable numbers of subjects completed each interview wave, and 2289 (82%) of the Wave 1 sample have completed the fifth wave.

2. Great Smoky Mountains Study(Costello, Angold et ai 1996; Costello, Farmer et al. 1997)—Three cohorts of boys and girls, aged 9,11, and 13 at intake in 1993, were selected from a rural population of some 20,000 children using a household equal probability design. A two-phase procedure was used for White and African-American youth, to increase power by oversampling children at risk for psychiatric and substance use disorders. Parents (usually mothers) of the first stage random population sample completed a questionnaire about their child's behavioral problems. Of 4,195 subjects selected, 95% (N=3,896) of parents completed the screen. All children scoring above a predetermined threshold (the top 25% of the total scores), plus a 10% random sample of the remaining 75%, were recruited for detailed interviews. Results can be back-weighted to population levels for analyses. Half of the sample are female, and 6% are African American, reflecting the population of the study area. The interviewed sample of white and Africian American subjects was 1,070 (80% of those recruited). American Indian youth were oversampled (100%) because they are an understudied group known to be at high risk for stressful events, substance disorders, and mood disorders. Of 431 age-eligible children, 350 (81%, 49% girls) participated. Thus, the size of total GSMS sample is 1,070+350 =1,420. Data collection is complete for ages 9-26, and age 30 interviews are in progress. By age 26 a total of 9858 interviews had been completed; the average number of interviews per subject was 7, and by age 26 97.3% completed two or more interviews.

3. The Caring for Children in the Community study (CCC) (Angold, Erkanli et ai 2002)—This representative study of psychiatric illness and service use in African American and White youth took place in four rural counties in the southeastern USA. The two-stage sampling design and methods are similar to those used in the GSMS. Of 4,500 youth randomly selected from the 17,117 9-17-year-olds in the public school's database, 3613 (80.0%) were successfully contacted and agreed to complete the behavioral screen. Of the 1302 selected to participate in the study, 920 (70.7%) interviews were completed. Because CCC was also the only study in GEDI to contain more than a very few African American participants, these were omitted from the multi-site analyses.

4. Child Health and Development Study (CHDS) (Fergusson and Horwood 2001)—The CHDS is a longitudinal study of a birth cohort from New Zealand. The cohort was based on an unselected sample of 1,265 consecutive births (635 male; 630 female) occurring in the Christchurch urban region in mid-1977. The cohort has been studied at birth, 4 months of age, 1 year of age, annual intervals to the age of 16 years, and again at ages 18, 21, 25 and 30 years. Sample retention rates were high throughout the study and at age 30 the study was still able to assess over 80% of the surviving cohort.

Informed consent in each study

Participants in all studies gave consent for their DNA to be genotyped. However, depositing biological samples and genetic data in controlled-access biorepositories (e.g., dbGaP (Mailman, Feolo et al. 2007) required a different level of consent. This was obtained for GSMS and VTSABD participants in Year 1 of GEDI. Further consents were not required from CCC as the study was closed. CHDS subjects gave consent for genotyping only.

Each IRB had slightly different requirements for consent forms, but in general study participants were given the opportunity to consent to (1) completing only the assessment instruments; (2) assessment plus DNA collection for internal use only; or (3) assessment and DNA collection, the anonymized data to be put into a repository.

Blood samples and genotyping

VTSABD—Nine ml. of blood were collected from VTSABD participants in the first year of the GEDI study, i.e., when subjects were aged 25 to 34. Blood and informed consent for genotyping and storage in dbGaP were obtained from 913 participants, of whom 281 were co-twins.

GSMS and CCC Blood from the GSMS and CCC samples were collected at each assessment: 10 finger-stick samples were collected on specially prepared paper, dried, and shipped to the study laboratory where they were stored at −23°C until they were assayed. A pilot study showed that even after 10 years of storage adequate DNA could be extracted from these samples. Most subjects (94%) provided at least one sample; the one collected as close to age 19 as possible was used for genotyping. Because there were so few African American participants any of the data sets except for CCC, the multi-site analyses excluded them, leaving 196 CCC and 784 GSMS participants with adequate genotype data and, in the case of GSMS, consent to deposit data in dbGaP. Since CCC was a closed study, no further consents were needed.

CHDS—Beginning in 2004 (at age 28) participants were asked for consent to provide saliva sample for DNA, and 918 (90% of the surviving cohort) consented. Consent for DNA collection was separate from consent for the rest of the study. In 2008-2009 participants were asked for consent for the GEDI multi-site GWAS, and 813 consented. Of these, 86% provided peripheral blood samples, 8% provided saliva, and 6% provided buccal swabs (the latter proved not to provide samples of sufficient quality for genotyping). After quality control checks, good quality data were obtained on 747 participants. The New Zealand government does not permit the data to be deposited in dbGaP.

Blood samples from the VTSABD, GSMS, and CCC samples were sent to the Rutgers University Cell and DNA Repository for DNA extraction, and to the Genotyping Shared Resource at the Mayo Clinic Cancer Center for genotyping. DNA for the CHDS sample was prepared in New Zealand and also sent to Mayo for genotyping. DNA samples were randomized to plates within studies. All samples were genotyped using Illumina Human660W-Quad vl DNA Analysis BeadChips. Quality control was carried out in the Department of Genetics at the University of North Carolina, Chapel Hill. In each of the fours samples, single nucleotide polymorphisms (SNPs) with missing rate > 0.01, minor allele frequency (MAF) <0.05, or extreme deviation ($p< l"^{6}$) from Hardy-Weinberg equilibrium (HWE) were removed from further analysis. Subjects with missing rate > 0.01 or unusual genome-wide homozygosity (|normalized homozygosity rate| > 5) were excluded. Sex was investigated using the no-call proportions of chrY SNPs and heterozygosity proportions for chrX SNPs. Mislabeled sex information was corrected after double-checking with the original data, and subjects with unexplainable results were deleted. In addition,

pairwise identical-by-descent (IBD) estimation was evaluated to identify unexpected duplicates and relative pairs. We imputed SNP dosages in all samples using MACH (Liu, McRae et al. 2010). The imputation reference was HapMap3 CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) for subjects of European ancestry. All subjects with pcl<0 were grouped as “white” and were imputed using HapMap3 CEU as reference; all subjects with pcl>0 were grouped together as “other” and imputed using HapMap3 CEU+YRI as reference. Prior to imputation, the studies had between 496k and 515k SNPs that passed quality control. After imputation, all studies had over 1,193k total SNP values for analysis. The number of SNPs used in particular studies may differed back on cut-offs used for minor allele frequency.

Unobserved population admixture due to ancestry is a well-known confound in GWAS. To protect against false-positives due to ancestry, we extracted five principal components from each sample to correspond with ancestral and cryptic population stratification. To improve the efficiency of the principal components analysis (PCA) for control of population stratification, a subset of independent SNPs were selected using the PUNK function—indep with proper parameters (window size = 50, the number of SNPs to shift the window at each step = 5, and the VIF threshold = 2). PCA was applied to the selected SNPs using the smartpca module of EigenSoft (Price, Butler et al. 2008). Between 77,155 and 79,517 SNPs were used for each of the samples that we analyzed in the ancestry PCA. All genotyping and QC was done blind to phenotype.

The VCU samples included in the analysis were all White, and HapMap3 CEU was used as the reference data to do the imputation for all the subjects from this data set. Most of the New Zealand sample were also White, although some were either Maori or mixed Maori and White. Since the number of Maori was small, we ignored the Maori’s Asian genetic background and used HapMap3 CEU samples as the reference data to run the imputation for all subjects. The imputation quality will not be perfect for Maori, but we used Rsq (the imputation quality score from MACH) to remove badly imputed SNPs.

The Duke samples contained a range of populations, including White, African American, American Indian, Hispanic and Asian, but most were either White or Black. For the imputation purpose, we split the samples into two groups. All the subjects with pcl>0 were grouped as “Black” and were imputed using HapMap3 CEU+YRI as reference. All the subjects with pcl<0 were grouped as “White” and were imputed using HapMap3 CEU as reference.

Environmental exposures

Each study collected extensive information on individual, family, and community risk for psychopathology. For the first analyses we selected a measure of exposure to potentially stressful life events (SLE), because several candidate-gene-based gene-environment studies have demonstrated significant gene-environment interplay using this measure (Karg, Burmeister et al. 2011). At each assessment participants in all studies were asked to indicate whether they had experienced any of several potentially stressful like events such as losing a friend, or moving. Each study provided count variables of the total number of stressful life events recently. The period assessed varied by study from the previous 12 months to the previous 3 months. Stressful life event terms were centered to the study mean to reduce multicollinearity with interaction terms. Although studies employed different primary periods, the parameter estimates for the association between stressful life events and substance-related outcomes were similar across studies. The study will only focus on effects that are robust to modest between-study differences in measurement or period assessed.

Data Analysis

Data analytic methods for each substance (cannabis, alcohol, nicotine) varied, and will be described in the empirical reports. However, there are some general principles that we discuss here.

The overall goal is to determine which, if any, of the measured or imputed SNPs contribute to the explained variance in substance involvement, after controlling for ancestry, sex, and age. We expect that a SNP may contribute to substance involvement directly, via a main effect of the SNP on substance involvement, but that a SNP may also have a heterogeneous effect across individuals due to differential exposure to environmental exposures, including time. The degree to which genetic information influences substance involvement across the lifespan may vary over time as a function of time-specific life circumstances.

One of the issues that strongly influence the success of multi-site analyses is that of identifying measures of phenotypes or environmental factors that are comparable across data sets. For example, in the analyses described below, a factor score measuring alcohol involvement was estimated using Mplus 6.0 (http://www.statmodel.com/),from measures of quantity of use, frequency of use, and symptoms count common to all the data sets.

One of the unique features of GEDI is the rich developmental data inherent in each of the samples that allows us to investigate if there are genetic variants that influence substance use in a key period of development. As mentioned above each of the samples covers a different age range (although they overlaps across some ages) and each empirical paper takes a different approach to data harmonization and handling the developmental piece. As our first step toward data harmonization for GEDI, in consultation with our colleagues at Duke's Social Science Research Institute we adopted the "transform and recode" procedure most commonly used in harmonization studies. (BATH, DEEG et al. 2010) First, a key member of each study team is tasked to achieve consensus regarding whether it is possible to find variables (and associated response categories) that have the same "face value". Next, a new harmonized variable is created for each "comparable" existing variable set by applying the transform and recode procedure to one or both of the original study measures, such that existing codes for categories can be merged and re-labelled in each study depending on the precise wording and ordering of the categories. (Fortier, Doiron et al. 2011)

For example, two primary, longitudinal measures of over-time alcohol consumption were generated to study the main effects of alcohol consumption on genetic variants: 1) a mixed model which explicitly models the developmental alcohol consumption trajectory spanning adolescence and early adulthood (ages 12-30), and 2) a simple mean of alcohol consumption (drinks per week) repeated measures collected across adolescence (ages 12-21) for each individual. We selected these two specifications because the trajectory outcome (1) was found to be the best fitting longitudinal model, taking advantage of all repeated measures, while the mean adolescent consumption outcome (2) provided a simpler summary of individuals' drinking behavior, and thus provided greater continuity to existing literature(Agrawal, Grant et al. 2009; Grant, Agrawal et al. 2009; Agrawal, Freedman et al. 2012) The harmonization methods used focused in this first instance on measures that are relatively constant in meaning across development, such as number of drinks per week. With the help of our colleagues from the Data Harmonization team at Duke, we will then tackle measures that may change either content or meaning across development. Other proposed analyses will rely upon using multiple items to estimate a singi **Power for analysis of gene-environment interplay.**

In principle, biostatistical methods for testing for genetic association and gene-environment interplay do not differ from those for testing any other association, interaction, or

correlation. The problem, of course, is the vast number of SNPs and environments,(van den Oord 2002) and the importance of controlling for false discoveries; i.e., concluding that a marker affects an outcome when in reality it does not. We use an approach to control false discoveries based on the false discovery rate (FDR)(Benjamini and Hochberg 1995). In comparison to controlling a family-wise error rate, e.g. Bonferroni correction, the FDR (1) provides a better balance between the competing goals of finding true effects versus controlling false discoveries, (2) results in comparable standards for declaring significance across studies because it does not directly depend on the number of tests, and (3) is relatively robust against correlated tests (Borden, Brown et al. 1987; Sabatti, Service et al. 2003; Tsai, Hsueh et al. 2003; Fernando, Nettleton et al. 2004; Korn, Troendle et al. 2004). The FDR is commonly used in many high-dimensional applications and has also successfully been applied in the context of GWAS{Lei, 2009 #506;Liu, 2009 #1515;Beecham, 2009 #1527}. We chose a FDR threshold of 0.1 for declaring genomewide significance,(van den Oord and Sullivan 2003) which means that on average 10% of the SNPs declared significant are expected to be false discoveries. Operationally{Black, 2004 #1284} the FDR is controlled using *q*-values that are FDRs calculated using the p-value of the markers as thresholds for declaring significance(Storey 2003). It is important to note that performing many GWAS analyses does not present a problem for the FDR because it controls the expected ratio of false to all discoveries. Thus, when many GWAS are performed the number of false positives will increase and so will the number of true positives. The expected ratio of false to all discoveries will therefore remain 0.1, our threshold for declaring genomewide significance.

RESULTS, CROSS-VALIDATION, AND REPLICATION

The results of the first set of papers focus on the main effect of the SNP and the interaction term between the SNP and SLE exposure. The first 3 papers, currently under review, focus on problem alcohol use, number of cigarettes per week, and any cannabis use in the past 3 months. In each case, the environmental factor used was a measure of severe life events, developed to be the same for all data sets.

GEDI takes two approaches to testing the validity of the results: cross-validation and replication. For the former we present the analyses separately for each data set and compare size and direction of effects across studies. This provides a more powerful test than the standard replication study, because it involves complete genome-wide comparisons rather than simply comparisons of a few sites selected from one data set. The disadvanatage is that individual data sets are necessarily smaller than the combined GEDI data set. We are therefore working to find other data sets with which we can carry out standard replication studies: comparing results on the “top hits” from GEDI. there are few other data sets with the characteristics of the studies included in GEDI (multiple measures across adolescence and early adulthood of both substance ue and abuse and relevant environmental risk factors), but we have identified three, with whom we are currently working (the Minnesota Twin and Family Study (Derringer, Krueger, McGue, & Iacono, 2008), Finn Twin (Pagan et al., 2006) and the Center for Education and Drug Abuse Research (CEDAR) sample (Tarter & Vanyukov, 1994)), with other collaborations under development.

2. Further analyses

The next stage in the program of data analysis is to broaden it to include gene-based analyses (Neale and Sham 2004), pathway analyses (Wang, Li et al. 2007), and polygenic risk score analyses (Purcell, Wray et al. 2009). Gene-based analyses test whether any genes harbor an excess of SNPs with small P-values. Such analyses must account for both gene length and linkage disequilibrium between SNPs (see VEGAS (Liu, McRae et al. 2010) for one example). Pathway analyses similarly test for an enrichment of SNPS with low P values

in genes involved in specific functional pathways (such as those in the Gene Ontology and Kyoto Encyclopedia of Genes and Genomes databases). Optimal approaches must account for varying gene size and SNP density, linkage disequilibrium within and between genes, and overlapping genes with similar annotations (see INRICH (Lee, O'Dushlaine et al. 2012) for one example). Finally, the polygenic risk score analyses test a polygenic basis for the phenotype by looking at the variance accounted for by a given set of top SNPs determined by a *P* value threshold (e.g., 0.005, 0.01, 0.10, or 0.25). In the first step, the sample is partitioned into a discovery and replication sets. Parameter estimates, derived in the discovery sample, are used as weights to calculate scores in the replication set. Subsequently, a regression is performed on the disease state in the replication set from the polygenic score and then *P* values and pseudo lvalues presented (see (International Schizophrenia Consortium 2009) for example). In each case we propose to analyze the individual data sets and also to perform a meta-analysis of the entire group. In these cases, we are applying analytic approaches already in use in other GWAS studies to GEDI studies as is, but by focusing on the model term related to the interaction between the environmental exposure and SNP status, for example, this standard approach allows us to address a novel outcome -genes that moderate the association between the exposure and the outcome.

Next steps: 3. Candidate gene selection for next generation sequencing

In addition to GWAS, we will employ targeted capture(Gnirke, Melnikov et al. 2009) and massively-parallel next generation sequencing(McKernan, Peckham et al. 2009), to exhaustively determine all genetic variation at selected genomic loci with evidence for involvement in SUD etiology. This approach uses a solution-based capture method(Gnirke, Melnikov et al. 2009), where genomic DNA from each subject is mixed with a "library" of synthetic oligonucleotides, designed to be complementary to the genomic regions of interest. Molecular tags on these oligonucleotides allow them to be pulled out of solution, bringing the bound, complementary genomic DNA with them. This "captured" DNA from each individual is labeled with a unique identifier and sequenced using next-generation, ultra high throughput technology(Smith, Heisler et al. 2010). This approach uses similar methods to exome sequencing(Ng, Turner et al. 2009), but here we will sequence the entire genomic region of each gene of interest, rather than just the exons, in order to capture all relevant variation.

We selected regions to include in our targeted capture library as follows:

1. GEDI GWAS findings for smoking and alcohol. For each SNP showing significant association at the genome-wide level (q-value < 0.1), we targeted the genomic region encompassing it (+/−25 kb) for sequencing, plus any genes that fall within this 50 kb window. For SNPs showing "potentially interesting" associations {*q*-values 0.1-0.2), we chose only those that fell within 25kb of a gene, based on the principle that potentially interesting associations are more likely to be real if they are in, or close to, a gene. These criteria led us to select 17 loci covering 3.5 Mb of genomic DNA sequence.
2. Genes identified through published GWAS alcohol(Schumann, Coin et al. 2011) and smoking GWAS meta-analyses(2010; Liu, Tozzi et al. 2010; Thorgeirsson, Gudbjartsson et al. 2010), plus gene nominations by expert colleagues (17 loci covering 2.5 Mb).
3. All human alcohol and aldehyde dehydrogenases, the key enzymes involved in alcohol metabolism(Edenberg 2007) (28 loci covering 1.3 Mb).

4. Reward system genes, including dopaminegic(Di Chiara and Imperato 1988), opioid(Le Merrer, Becker et al. 2009) and cannabinoid(Solinas, Yasar et al. 2007) receptors and related metabolic genes (16 loci covering 1.09 Mb).
5. All remaining human nicotinic acetyl choline receptors not already selected (12 loci covering 0.42 Mb).
6. Additional priority genes close to GWAS hits (4 loci covering 0.35 Mb).
7. Prioritized candidate genes. We compiled 3 lists of candidate genes, the first based on previous associations in the literature with SUDs, the second comprising all known human genes involved in absorption, distribution, metabolism and excretion (ADME) of drugs (www.pharmaadme.org), and the third included all human neuroactive ligand receptors from the KEGG database(Kanehisa and Goto 2000; Kanehisa, Goto et al. 2011). We ranked genes by the number of times they co-occurred in the literature with the search terms “smoking”, “alcohol” or “cannabis” and filled the remainder of our targeted capture library with the top-ranked genes from this list (31 loci covering 1.1 Mb).

After removing overlap and collapsing neighboring genes into single loci, our selection encompassed 121 unique loci, covering a total of 10.2 Mb. However, human genomic DNA includes repetitive elements, including RNA and DNA transposons, which constitute approximately 45% of the human genome(Lander, Linton et al. 2001). These contribute no useful sequence information, because they are difficult to align to unique positions. After elimination of these repetitive elements, our final library encompassed approximately 5.5 Mb. We are currently sequencing this library in 1000 individuals selected from the VTSABD and CHDS.

DISCUSSION

As noted earlier, there is an abundance of literature recommending that genomics move in the direction of genome-wide gene-environment-development analyses, but very few data - in fact, we have found no empirical “developmental GEWIS” (Khoury and Wacholder 2009) studies so far. The few developmental gene-environment studies published have used a candidate gene approach (e.g., Adkins, Daw, McClay, & Van den Oord, 2012; Casey et al., 2009; Cole et al., 2011), and carry many of the limitations that have long plagued these studies (Sullivan, Eaves et al. 2001).

Thus, we undertook GEDI partly as a proof of concept of the feasibility of a developmental GEWIS. Despite the lack of empirical data, most discussions in the literature take a gloomy view of the feasibility of development GEWIS, concentrating on the large sample sizes needed and the unreliability of measures of the envirotype (Thomas 2010). We acknowledge these problems unreservedly. On the other hand, there are other aspects of the situation that should be considered. First, estimates of sample sizes tend to be based on experience with early genetic studies that have collected cases and controls using what is often very unreliable data: lifetime psychiatric histories, or clinical diagnoses from hundreds or thousands of different clinicians. These methods result not only in false positives, but also in considerable numbers of false negatives (cases included non-cases because subjects have forgotten past episodes of illness). Both of these type of error inflate the sample size needed. In their paper on estimating the size of gene-environment interactions in the presence of measurement error (Wong, Day et al. 2004), and related papers(Luan, Wong et al. 2001; Wong, Day et al. 2003), Wong and colleagues pointed out that accuracy in measuring all the related elements — genotype, phenotype, and exposure - critically affect the sample size needed for a given power. For example, “the difference between unreliable (correlation with true score=0.4) and reliable (r=0.7) measurements corresponds to a 20-fold difference in

sample size" (Moffitt, Caspi et al. 2005) p.476. Furthermore, Wong et al. noted that "Improving the measurement can be achieved by taking repeated measurements"(Wong, Day et al. 2003) p.54; thus, the longitudinal studies preferred for developmental analyses will also increase their power to test hypotheses.

Advantages of the GEDI consortium are that it includes only data sets with longitudinal, repeated assessments of subjects, taken across the period of adolescence and young adulthood. We can expect subjects to be less vulnerable to either false remembering or false forgetting than those in studies using lifetime retrospective data. The use of repeated assessments also means that even a relatively small number of subjects yield a large number of person-observations (for example, the 1,420 GSMS subjects yielded 9,858 person-observations by age 26). Even after controlling for non-independence of observations, this approach substantially increases the effective sample size and therefore power. Third, the studies used reliable assessments of symptoms and diagnoses created using a single taxonomy (DSM-IV) and highly-structured diagnostic algorithms. Fourth, the studies used reasonably similar measures of key environmental risk factors.

The results of our first analyses (Copeland et al. Stressful life events and Alcohol use: A longitudinal GxE GWAS Meta-analysis, submitted) appear to provide tentative empirical evidence that the combination of prospective, longitudinal assessment and careful attention to data harmonization can, to some extent, compensate for a modest sample sizes. However, regardless of these benefits, there remains an acute need to build a broader, more inclusive consortium of qualifying longitudinal data sets. Our strongest recommendation from this experiment is for the creation of an international "developmental dbGaP" of such data sets to optimize power in future investigations of environmentally and developmentally contigent genetic effects on behavioral outcomes.

Acknowledgments

Work on this project was carried out with support from NIDA (R01DA024413). Dr. Adkins was supported by K01MH093731, and Dr. Copeland by K23MH080230. We are grateful to all the study participants who contributed to this work.

REFERENCES

Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010; 42(5):441–447. [PubMed: 20418890]

Ackermann KH, Adams N, et al. Elliptic flow in Au+Au collisions at square root(S)NN = 130 GeV. Phvs Rev Lett. 2001; 86(3):402–407.

Agrawal A, Freedman ND, et al. Measuring alcohol consumption for genomic meta-analyses of alcohol intake: opportunities and challenges. American Journal of Clinical Nutrition. 2012; 95(3): 539–547. [PubMed: 22301922]

Agrawal A, Grant JD, et al. Developing a Quantitative Measure of Alcohol Consumption for Genomic Studies on Prospective Cohorts. Journal of Studies on Alcohol and Drugs. 2009; 70(2):157–168. [PubMed: 19261227]

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders Fourth Edition (DSM-IV). American Psychiatric Press, Inc.; Washington, DC: 1994.

Anastasi A. The concept of validity in the interpretation of test scores. Journal of Psychology and Educational Measures. 1950; 10:67–78.

Angold A, Erkanii A, et al. Psychiatric disorder, impairment, and service use in rural African American and White youth. Archives of General Psychiatry. 2002; 59:893–901. [PubMed: 12365876]

BATH PA, DEEG D, et al. The harmonisation of longitudinal data: a case study using data from cohort studies in The Netherlands and the United Kingdom. Ageing & Society. 2010; 30(08):1419–1437.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995; 57:289–300. Series B.

Bookman EB, McAllister K, et al. Gene-environment interplay in common complex diseases: forging an integrative model-recommendations from an NIH workshop. Genet Epidemiol. 2011

Borden KA, Brown RT, et al. Achievement attributions and depressive symptoms in Attention Deficit Disordered and Normal Children. Journal of School Psychology. 1987; 25:399–404.

Casey BJ, Glatt CE, et al. Brain-derived neurotrophic factor as a model system for examining gene by environment interactions across development. Neuroscience. 2009; 164(1):108–120. [PubMed: 19358879]

Chiang MC, McMahon KL, et al. Genetics of white matter development: a DTI study of 705 twins and their siblings aged 12 to 29. Neuroimage. 2011; 54(3):2308–2317. [PubMed: 20950689]

Cole SW, Arevalo JM, et al. Antagonistic pleiotropy at the human IL6 promoter confers genetic resilience to the pro-inflammatory effects of adverse social conditions in adolescence. Dev Psychol. 2011; 47(4):1173–1180. [PubMed: 21639625]

Cornells MC, Agrawal A, et al. The Gene, Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet Epidemiol. 2010; 34(4):364–372. [PubMed: 20091798]

Costello E, Farmer E, et al. Psychiatric disorders among American Indian and White youth in Appalachia: The Great Smoky Mountains Study. American Journal of Public Health. 1997; 87:827–832. [PubMed: 9184514]

Costello EJ, Angold A, et al. The Great Smoky Mountains Study of Youth: Goals, designs, methods, and the prevalence of DSM-III-R disorders. Archives of General Psychiatry. 1996; 53:1129–1136. [PubMed: 8956679]

Di Chiara G, Imperato A. Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats. Proc Natl Acad Sci U S A. 1988; 85(14):5274–5278. [PubMed: 2899326]

Dick DM. Developmental changes in genetic influences on alcohol use and dependence. Chald Development Perspectives. 2011; 5(4):223–230.

Edenberg HJ. The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. Alcohol Res Health. 2007; 30(1):5–13. [PubMed: 17718394]

Fergusson D, Horwood L. The Christchurch health and development study: Review of findings on child and adolescent mental health. Australian and New Zealand Journal of Psychiatry. 2001; 35:287–296. [PubMed: 11437801]

Fernando R, Nettleton D, et al. Controlling the proportion of false positives in multiple dependent tests.. Genetics. 2004; 166:611–619. [PubMed: 15020448]

Fortier I, Doiron D, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. International Journal of Epidemiology. 2011; 40(5):1314–1328. [PubMed: 21804097]

Gnirke A, Melnikov A, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009; 27(2):182–189. [PubMed: 19182786]

Grant JD, Agrawal A, et al. Alcohol Consumption Indices of Genetic Risk for Alcohol Dependence. Biological Psychiatry. 2009; 66(8):795–800. [PubMed: 19576574]

International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460(7256):748–752. [PubMed: 19571811]

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28(1):27–30. [PubMed: 10592173]

Kanehisa M, Goto S, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2011

Karg K, Burmeister M, et al. The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation. Arch Gen Psychiatry. 2011; 68(5):444–454. [PubMed: 21199959]

Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies—challenges and opportunities. Am J Epidemiol. 2009; 169(2):227–230. discussion 234-225. [PubMed: 19022826]

Khoury MJ, Wacholder S. Invited Commentary: From Genome-Wide Association Studies to Gene-Environment-Wide Interaction Studies—Challenges and Opportunities. Am. J. Epidemiol. 2009; 169(2):227–230. [PubMed: 19022826]

Korn E, Troendle J, et al. Controlling the number of false discoveries: Application to high-dimensional genomic data. Journal of Statistical Planning and Inference. 2004; 124:379–398.

Lander ES, Linton LM, et al. Initial sequencing and analysis of the human genome. Nature. 2001; 409(6822):860–921. [PubMed: 11237011]

Le Merrer J, Becker JA, et al. Reward processing by the opioid system in the brain. Phvsiol Rev. 2009; 89(4):1379–1412.

Lee PH, O'Dushlaine C, et al. INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics. 2012; 28(13):1797–1799. [PubMed: 22513993]

Lenroot RK, Giedd JN. Annual Research Review: Developmental considerations of gene by environment interactions. J Child Psychol Psychiatry. 2011; 52(4):429–441. [PubMed: 21391998]

Liu JZ, McRae AF, et al. A Versatile Gene-Based Test for Genome-wide Association Studies. American Journal of Human Genetics. 2010; 87(1):139–145. [PubMed: 20598278]

Liu JZ, McRae AF, et al. A Versatile Gene-Based Test for Genome-wide Association Studies. The American Journal of Human Genetics. 2010; 87(1):139–145.

Liu JZ, Tozzi F, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. 2010; 42(5):436–440. [PubMed: 20418889]

Luan JA, Wong MY, et al. Sample size determination for studies of gene-environment interaction. International Journal of Epidemiology. 2001; 30(5):1035. [PubMed: 11689518]

Mailman MD, Feolo M, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007; 39(10):1181–1186. [PubMed: 17898773]

McKernan KJ, Peckham HE, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 2009; 19(9):1527–1541. [PubMed: 19546169]

Moffitt TE, Caspi A, et al. Strategy for investigating interactions between measured genes and measured environments. Archives of General Psychiatry. 2005; 62:473–481. [PubMed: 15867100]

Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. Am J Hum Genet. 2004; 75(3):353–362. [PubMed: 15272419]

Ng SB, Turner EH, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461(7261):272–276. [PubMed: 19684571]

Price AL, Butler J, et al. Discerning the Ancestry of European Americans in Genetic Association Studies. PLoS Genet. 2008; 4(1):e236. [PubMed: 18208327]

Purcell SM, Wray NR, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460(7256):748–752. [PubMed: 19571811]

Rose RJ, Dick DM. Commentary on Agrawal et al. (2010): Social environments modulate alcohol use. Addiction. 2010; 105(10):1854–1855. [PubMed: 20860079]

Sabatti C, Service S, et al. False discovery rate in linkage and association genome screens for complex disorders. Genetics. 2003; 164(2):829–833. [PubMed: 12807801]

Schumann G, Coin LJ, et al. Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. Proc Natl Acad Sci U S A. 2011; 108(17):7119–7124. [PubMed: 21471458]

Simonoff E, Pickles A, et al. The Virginia Twin Study of adolescent behavioral development: Influences of age, sex and impairment on rates of disorder. Archives of General Psychiatry. 1997; 54:801–808. [PubMed: 9294370]

Smith AM, Heisler LE, et al. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res. 2010; 38(13):el42.

Solinas M, Yasar S, et al. Endocannabinoid system involvement in brain reward processes related to drug abuse. Pharmacol Res. 2007; 56(5):393–405. [PubMed: 17936009]

Spatola CA, Scaini S, et al. Gene-environment interactions in panic disorder and CO(2) sensitivity: Effects of events occurring early in life. Am J Med Genet B Neuropsvchiatr Genet. 2011; 156B(1): 79–88.

Storey J. The positive false discovery rate: A Bayesian interpretation and the q-value. Annals of Statistics. 2003; 31:2013–2035.

Sullivan PF, Eaves LJ, et al. Genetic case-control association studies in neuropsychiatry. Archives of General Psychiatry. 2001; 58:1015–1024. [PubMed: 11695947]

Sung M, Erkanli A, et al. Effects of age at first substance use and psychiatric comorbidity on the development of substance use disorders. Drug and Alcohol Dependence. 2004; 75:287–299. [PubMed: 15283950]

Thomas D. Gene—environment-wide association studies: emerging approaches. Nat Rev Genet. 2010; 11(4):259–272. [PubMed: 20212493]

Thorgeirsson TE, Gudbjartsson DF, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet. 2010; 42(5):448–453. [PubMed: 20418888]

Tsai C, Hsueh H, et al. Estimation of false discovery rates in multiple testing: application to gene microarray data. Biometrics. 2003; 59:1071–1081. [PubMed: 14969487]

van den Oord E. Association studies in psychiatric genetics: what are we doing? Molecular Psychiatry. 2002; 7(8):827–828. [PubMed: 12232772]

van den Oord EJ, Sullivan P. False discoveries and models for gene discovery. Trends in Genetics. 2003; 19:537–542. [PubMed: 14550627]

van Ijzendoom MH, Bakermans-Kranenburg MJ, et al. Gene-by-environment experiments: a new approach to finding the missing heritability. Nat Rev Genet. 2011; 12(12):881. author reply 881.

Wang K, Li M, et al. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007; 81(6):1278–1283. [PubMed: 17966091]

Wong MY, Day NE, et al. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? International Journal of Epidemiology. 2003; 32(1):51. [PubMed: 12690008]

Wong MY, Day NE, et al. Estimation of magnitude in gene-environment interactions in the presence of measurement error. Stat Med. 2004; 23(6):987–998. [PubMed: 15027084]

Table 1

Characteristics of the studies

	GSMS	ccc	vcu	CHDS
Number of data waves	7-10	3	5	18
2 or more developmental periods? Age range from start to most recent wave	Yes 9-26	Yes 9-17	Yes 8-30	Yes 1-30
Number of participants recruited at baseline	1,420	920	913 Twin pairs	1,265
Number of participants successfullYes genotYesped	784	518	913 (632 unique families)	747
Representative population sample?	Yes	Yes	Yes (twins)	Yes

=Great Smoky Mountains Study

CCC= Caring for Children and the Community

VCUABD=Virginia Twin Study of Adolescent Behavioral Development

CHDS=Christchurch Health and Development Study