Published in final edited form as:

Ann N Y Acad Sci. 2008 October ; 1141: 318–381. doi:10.1196/annals.1441.018.

Molecular genetics of addiction and related heritable phenotypes: genome wide association approaches identify “connectivity constellation” and drug target genes with pleiotropic effects

George R Uhl^{1,*}, Tomas Drgon¹, Catherine Johnson¹, Chuan-Yun Li^{1,2}, Carlo Contoreggi³, Judith Hess¹, Daniel Naiman⁴, and Qing-Rong Liu¹

¹Molecular Neurobiology Branch, NIH-IRP (NIDA), Suite 3510, 333 Cassell Drive Baltimore, Maryland 21224

²Office of the Clinical Director, NIH-IRP (NIDA), Baltimore MD 21224

³Center for Bioinformatics, College of Life Sciences, Peking University. Beijing, China 100871

⁴Department of Mathematics, Johns Hopkins University

Abstract

Genome wide association (GWA) can elucidate molecular genetic bases for human individual differences in “complex” phenotypes that include vulnerability to addiction. Here, we review: a) evidence that supports polygenic models with (*at least*) modest heterogeneity for the genetic architectures of addiction and several related phenotypes; b) technical and ethical aspects of importance for understanding genome wide association data: genotyping in individual samples vs DNA pools, analytic approaches, power estimation and ethical issues in genotyping individuals with illegal behaviors; c) the samples and the data that shape our current understanding of the molecular genetics of individual differences in vulnerability to substance dependence and related phenotypes; d) overlaps between GWA datasets for dependence on different substances; e) overlaps between GWA data for addictions vs other heritable, brain-based phenotypes that include: i) bipolar disorder, ii) cognitive ability, iii) frontal lobe brain volume, iv) ability to successfully quit smoking, v) neuroticism and vi) Alzheimer’s disease. These convergent results identify potential targets for drugs that might modify addictions and play roles in these other phenotypes. They add to evidence that individual differences in the quality and quantity of brain connections make pleiotropic contributions to individual differences in vulnerability to addictions and to related brain disorders and phenotypes. A “connectivity constellation” of brain phenotypes and disorders appears to receive substantial pathogenic contributions from individual differences in a constellation of genes whose variants provide individual differences in the specification of brain connectivities during development and in adulthood. Heritable brain differences that underlie addiction vulnerability thus lie squarely in the midst of the repertoire of heritable brain differences that underlie vulnerability to other common brain disorders and phenotypes.

Keywords

pleiotropic; cell adhesion; monte carlo

*To whom correspondence should be addressed at: Molecular Neurobiology, Box 5180, Baltimore, MD 21224 phone: (410) 550-2843 x 146, fax: (410) 550-1535, guhl@intra.nida.nih.gov.

I. Introduction

Genome wide association is now increasingly the method of choice for identifying allelic variants that contribute to complex genetic disorders, especially those with **polygenic** genetic bases (*eg* derived from effects at many gene loci, each with modest effects, as well as from environmental determinants; see also Glossary)[1–9]. Substance dependence was one of the first complex phenotypes for which replicated association-based genome scanning data was reported [3, 10–14]. There is now a torrent of information from genome wide association studies of a number of other complex, brain-based phenotypes that both display substantial heritability and are unlikely (based on linkage study results) to manifest many common gene variants that produce large effects [1, 2, 4, 5]. A number of these other heritable, brain-based phenotypes co-occur with addictions and are thus good candidates to display genetic overlaps with addiction.

No single approach to designing genome wide association studies or to analyzing genome wide association data is now universally accepted. There is now no universal standard for considering genome-wide association results “significant” in ways that allow us to identify polygenic allelic variants in reasonably-sized single experiments. Here, we describe specific sets of working hypotheses about the genetic architecture of addiction (*eg* vulnerability to develop dependence on an addictive substance). This set of hypotheses is also useful for considering the molecular genetic bases for other common, complex phenotypes that, like addictions, display both substantial evidence for heritability and little evidence for large influences from any single gene (*eg* single gene, Mendelian influences or oligogenic effects that come from a few genetic loci, each with moderate effects on the phenotype). We then detail experimental design and analytic approaches that arise from working hypotheses about underlying genetic architecture and likely sources of false positive results.

A number of samples provide the bases for these analyses. In analyzing data from these samples, we focus on clusters of genomic markers whose allele frequencies distinguish control individuals from those with substance dependence or addiction-related phenotypes. We describe identification of chromosomal regions that contain clusters of such nominally-positive results in replicate samples for addiction vulnerability. We then describe evidence for generalization that arises from identification of overlapping chromosomal locations of clustered positive results for different phenotypes. These data thus support **pleiotropic** influences (*eg* contributions of the same allelic variants to multiple phenotypes) of common allelic variants on several of the brain based phenotypes. The data thus document overlapping heritable influences on several interesting brain phenotypes.

We focus here on clinical phenotypes that co-occur with addiction and a structural brain phenotype, individual differences in frontal cortical volume. Twin studies document sizable heritable components for individual differences in the volumes of brain regions. High heritabilities are especially evident for individual differences in frontal and temporal cerebral cortical regions [15]. Volumes of these brain regions have been reported to be reduced in substance dependent individuals [16–19]. Increasing evidence from fMRI and PET studies identifies functional differences in these brain regions in studies of individuals with substance dependence and related phenotypes [20, 21]. We thus focus on this “frontal cortical volume” phenotype.

A number of the genes identified here encode classical “druggable” targets for pharmacological modulation, including enzymes, receptors and transporters. Other genes encode cell adhesion related molecules. We discuss genes in each of these classes below.

A. What is genome wide association and why might it be useful for studies of the molecular bases of heritable influences on vulnerability to addiction and related phenotypes? One way to view genome wide association is in relationship to linkage based genome scanning, since most of the efforts to “positionally clone” gene variants for complex human disorders (*eg* those that are likely to be caused by multiple genetic and environmental factors) have used **linkage**-based methods. Linkage asks how addiction phenotypes and genetic markers (typically genotyped approximately every $1/400^{\text{th}}$ – $1/1,000^{\text{th}}$ of the genome) move together through pedigrees of closely-related individuals. Chromosomal regions that contain marker alleles that move through pedigrees together with the trait are said to be “linked” to the trait. Many loci with nominally-significant linkage to addiction phenotypes have been identified [22–31] (further references in [32]). Several of the loci identified in independent linkage studies do overlap. However, the large numbers of reported linkage-based studies of addictions yield large numbers of nominally-positive results that cover virtually all chromosomes. These widespread results may not converge more than expected by chance, as we have documented in recent analyses of the reported data for linkage to smoking [32].

Such inconsistent linkage data is consistent with the idea that most of the genetic architecture for human addiction vulnerability is polygenic in most populations. There is now an increasing consensus that **genome-wide association** (GWA; also termed “whole genome association” or “association genome scanning”) is more likely than linkage approaches to yield positive results in polygenic complex disorders, such as addictions [1, 3, 10–12, 33–37]. Association asks how addiction phenotypes and genetic markers (genotyped approximately every $1/500,000^{\text{th}}$ to every $1/1,000,000^{\text{th}}$ of the genome in current datasets) are found together in nominally-unrelated individuals (although we are all distantly related to each other, of course). We and others have developed these methods, relying on the increasing densities of single nucleotide polymorphism (SNP) markers that can be assessed using “SNP chip” microarrays of increasing sophistication [10–12, 33, 34, 36, 37]. Genome wide association gains power as densities of genomic markers increase. Association identifies much smaller chromosomal regions than linkage-based approaches. Association thus allows us to identify variants in specific genes rather than in large chromosomal regions. Genome wide association fosters pooling strategies that preserve confidentiality and reduce costs, as we discuss below [10–12, 33, 38–44]. Genome wide association provides ample genomic controls. Proper genomic controls can minimize the chances that disease *vs* control differences are confounded by occult stratification, such as the stratification that might arise from unintended occult ethnic mismatches between disease and control samples.

B. A likely underlying genetic architecture for addiction (and other common heritable brain-based phenotypes for which linkage data fails to provide evidence for “genes of major effect”) We approach analyses of the molecular genetic bases of addiction and related disorders from perspectives that are based on sets of working hypotheses concerning the underlying genetic architectures of these disorders and phenotypes. In general, the best experimental design and analytic approaches are likely to arise from the best possible working hypotheses concerning: 1) the genetic architectures of the disorders or phenotypes being evaluated 2) the population genetics of the samples being tested 3) the anticipated association signals. It is also desirable to consider and provide controls for alternative hypotheses that might provide alternative explanations for systematic differences between disease and control samples. Alternative hypotheses include: 1) unintended stratification based on *eg* racial/ethnic differences between “disease” and “control” samples, as noted above, 2) uneven distribution of noise in some assays, so that SNPs with the largest variance might be identified rather than the SNPs whose allelic frequencies truly differ between disease and control individuals, 3) stochastic, chance differences between disease and control samples (at least some of these are highly likely in any single study that uses so

many repeated comparisons) and 4) sampling issues so that genetics related to the ways in which the samples are ascertained and obtained (*eg* features such as differential willingness to consent) is identified rather than true disease/control differences. Many of these concerns become more prominent as hundreds of thousands or millions of repeated comparisons are made using single sample sets. Many of these concerns may be more acute as attempts to rapidly assemble samples with larger and larger “n” drive investigators to include subsamples that may well contribute occult heterogeneities to overall disease and/or control samples (*see below*).

C. Family, adoption and twin data that each support substantial polygenic heritability for addictions Current models for the genetic architecture for substance dependence in the population are based on information from: 1) family, adoption and twin data that each support substantial heritability for addictions, 2) twin data (in which concordance in genetically-identical monozygotic and genetically-half-identical dizygotic twins are compared) that document that most of this heritable influence is not substance-specific, 3) linkage based (and genome wide association) studies that fail to provide evidence for genes of major effect (*eg* for any single gene whose variants produce substantial differences in addiction vulnerability) for substance dependence.

Support for the idea that vulnerability to addictions is a complex trait with strong genetic influences that are largely shared by abusers of different legal and illegal addictive substances [45–48] comes from classical genetic studies. Family studies document that first degree relatives (*eg* sibs) of addicts display greater risk for developing substance dependence than more distant relatives [45, 49]. Adoption studies find greater similarities between levels of substance abuse between adoptees *vs* biological relatives than adoptees *vs* members of the adoptive families [45]. In twin studies, differences in concordance between genetically identical and fraternal twins also support heritability for vulnerability to addictions [47, 50–56]. Twin data allows quantitation of the amount, about half, of addiction vulnerability that is heritable. Twin data also supports the idea that the environmental influences on addiction vulnerability that are not shared among members of twin pairs are much larger than those that are shared by members of twin pairs (*eg* $e^2 \gg c^2$ in virtually every such study). Most environmental influences on human addiction vulnerability are thus likely to come from outside of the immediate family environment (Figure 1).

D. Twin data document that most of this heritable influence is not substance-specific, but provides “higher order” pharmacogenomics We are fortunate to have data from studies of identical *vs* fraternal twin pairs that evaluate the degree to which one twin’s dependence on a substance enhances the chance that his or her co-twin will become dependent on a substance of a different class. Results of these analyses document that most of the genetic influences on addiction vulnerability are common to dependence on multiple different substances, though others do appear to be substance-specific [46, 53, 54].

Elsewhere [32] we have suggested levels of analysis for pharmacogenomics and pharmacogenetics: 1) “**primary**” **pharmacogenomics** that describes the genetics of individual differences in the adsorption, distribution, metabolism and/or excretion of a drug; 2) “**secondary**” **pharmacogenomics** that describes individual differences in drug targets, such as the G-protein coupled receptors, transporters and ligand gated ion channels that are the primary targets of opiates, psychostimulants and barbiturates, respectively and 3) “**higher order**” **pharmacogenomics** that provide individual differences in post-receptor drug responses. Such post-receptor drug responses are more likely to be common to actions of abused substances that come from several different chemical classes and act at distinct primary receptor or transporter sites in the brain. Based on the twin data that are currently

available, we thus postulate that much, if not most, of the human genetics of addiction vulnerability represents “higher order” pharmacogenomics.

E. Failure to document evidence for substance dependence genes of major effect in most populations There are few careful studies of the ways in which most human addiction vulnerabilities move through families (*eg segregation analyses*). No such study indicates a “major” gene effect on addiction vulnerability in most current populations. There is an exception: the “flushing syndrome” variants at the aldehyde (ALDH) and alcohol (ADH) dehydrogenase loci in Asian individuals do provide genes of major effect in this population. Individuals with these gene variants are at lower risk for becoming dependent on alcohol than individuals with other genotypes [57] in Chinese [58, 59], Korean [60], Japanese [61–66] and other populations [67, 68]. Homozygous ALDH2*2 individuals are strongly protected from alcohol dependence [61, 62]. This locus thus provides a good example of “primary” pharmacogenomics, though in a restricted population.

Quantity-frequency data for smoking also provide evidence for a replicable “secondary” pharmacogenomic effect of moderate magnitude. Markers in the chromosome 15 gene cluster that encodes the $\alpha 3$, $\alpha 5$ and $\beta 4$ nicotinic acetylcholine receptors display different allelic frequencies in heavy vs light smokers in each of several studies [3, 35, 69]. This chromosome 15 locus is likely to provide a good example of “secondary” pharmacogenomics, since it has not been associated as reproducibly with dependence on other substances.

Linkage-based analyses for addiction vulnerabilities would be expected to reproducibly identify many of the genes whose variants exerted major influences on human addiction vulnerability. However, existing linkage data for human dependence on alcohol, nicotine and a number of other substances fails to provide any highly-reproducible results that would support any major gene locus ([22–31] and references in [32]). These results add to the conclusion that no locus individually appears to contribute a large fraction of the vulnerability to dependence on any addictive substance. There is a caveat: these data come from subjects with largely European ethnic/racial backgrounds [22, 23, 70–84]. Nevertheless, as with many complex human disorders in which initial hopes for a tractable (*eg oligogenic*) underlying genetic architecture supported use of linkage approaches, the linkage peaks that are identified in each individual study may be more likely to arise on other bases when the underlying architecture is in fact polygenic. Apparent linkage signals identified in single studies might result from polygenic influences from several genes that each happen to lie near each other on human chromosomes or to be found on stochastic bases when there is no true major effect from any single gene variant, for example [85].

F. Current models for the genetic architecture of human dependence on legal and illegal addictive substances in the population thus postulate that each is influenced roughly 50% by polygenic genetic influences, that is by variants in individual genes that each contribute modest amounts to this overall genetic vulnerability. These genetic architectural models posit that many of these genetic vulnerabilities increase risk for addiction to several pharmacologic classes of abused substances, but that some of these genetic influences are specific to drugs of one class [32].

Analyses of twin data for vulnerability to develop dependence on a substance fit with large additive genetic components (a^2), large components for nonshared environmental influences (e^2) and small components for c^2 terms that represents familial or other environmental influences that are shared between members of the twin pair [47, 50–56]. What about the possibility that there could be large interactions between these genetic and environmental terms ($G \times E$ interactions), invalidating additive models for genetic and environmental

contributions? $G \times E$ correlations of three types have been described [86, 87]. In one terminology, “passive” $G \times E$ correlation occurs when parents transmit both genes and environmental influences that are relevant for a trait [88, 89], “active” $G \times E$ correlation occurs where subjects of a certain genotype actively select environments that are correlated with that genotype and “reactive” $G \times E$ correlation occurs when an individual’s genotype provides different reactions from the environment. Small values for c^2 influences of common environments shared by members of sibpairs appear to provide evidence against “passive” $G \times E$ correlations. On these bases, “active” and “reactive” $G \times E$ correlations remain of theoretical interest. However, one influential train of thought [88, 90] suggests that $G \times E$ correlations are best regarded as parts of the genetic variance because “... the non-random aspects of the environment are... consequence(s) of the genotype(es)...”.

Large interactions between genetic and environmental components would be likely to lead to 1) differences in estimates of heritability from samples obtained in different environments, and 2) differences in molecular genetic findings in individuals from different environments. Data from studies of twins who were sampled from a number of different environments is nevertheless largely convergent. Such convergence supports relatively modest upper limits on ($G \times E$) interactions between genetic and environmental influences on addiction vulnerability. Modest $G \times E$ influences are also consistent with molecular genetic results that identify substantial overlaps between molecular genetics of vulnerability to dependence on illegal substances in samples from substantially different environments, such as the United States and Asia (*see below*).

Gene - gene interactions ($G \times G$) of some magnitude appear likely, *a priori*, to make at least some contributions to addiction vulnerability. However, if there were large amounts of **epistasis**, $G \times G$ interactions in which specific alleles at one gene locus are required for expression of the effects of allelic variants at a second gene locus, segregation analysis data might provide uneven patterns of familiarity. With large amounts of epistasis, second degree relatives (*eg* cousins) of addicts would be less likely to display specific combinations of $G \times G$ alleles than first degree relatives (*eg* sibs). Substance dependence rates would thus drop more precipitously between first and second degree relatives of addicts than they would if most risk alleles exerted largely independent effects on addiction vulnerability.

There is only a modest amount of family data that allows us to compare concordance in first vs second degree relatives. However, the existing evidence does not support less concordance in second degree relatives than we would anticipate based on the observed concordance in first degree relatives and the assumption that most risk alleles produce largely independent effects [91].

G. The genetic architecture for substance dependence in individuals What about the genetic architecture for substance dependence in individuals? Both “*between locus*” **heterogeneity** and “*within locus*” **heterogeneity** are likely. If we follow the implications of polygenic genetic models for addiction vulnerability, we can infer that each dependent individual might even display a nearly-distinct set of risk-elevating or risk-reducing allelic variants. As an illustrative example, we might postulate that a) an individual must display at least 50 risk alleles to robustly elevate his/her likelihood of acquiring a substance dependence disorder and b) there are 200 genes that contain common allelic variants that can augment addiction risk. Under such circumstances, it is easy to see that the exact genetic recipe for addiction vulnerability found in one addicted individual might be replicated in only a relatively few other addicted individuals. Such an underlying genetic architecture would be consistent with the failure of linkage-based methods to provide reproducible results in addictions, since linkage relies on identifying consistent patterns in the ways that

specific DNA markers and phenotypes move through many families that display high densities of the disorder.

As noted above, the best documented genetic heterogeneity for addictions comes from the chromosome 4 major gene effects found in poorly-alcohol-metabolizing (“flushing”) Asian individuals [61–63, 67]. The best documented substance-specific influence comes from the chromosome 15 nicotinic acetylcholinergic receptor gene cluster. There are likely to be other examples of between-locus genetic heterogeneity and of genes whose variants exert substance-specific effects on use and/or dependence that have yet to be elucidated.

We also postulate that *within locus heterogeneity* is likely, though not yet clearly documented in addiction, to our knowledge. Many common Mendelian disorders and rarer Mendelian phenocopies of common disorders display substantial heterogeneity within their pathogenic loci. A number of variants in the same CFTR gene produce cystic fibrosis disorders [92]. α synuclein missense variants and copy number variants can each provide phenocopies of idiopathic Parkinson’s disease [93]. Evidence for within-locus heterogeneity in complex disorders is just beginning to be accrued; such evidence now includes data from neurexin gene family variants in autism [94–97].

H. “Epigenetics” and individual differences in vulnerability to addiction and related phenotypes

“Epigenetics” is now used with both classical and recently-revised definitions. Classical definitions of “**epigenetic**” emphasize influences of variations that are not encoded in primary DNA sequence but nevertheless inherited “... a change in the state of expression of a gene that does not involve a mutation, but that is nevertheless inherited in the absence of the signal (or event) that initiated the change” [98]. However, more recent definitions of “epigenetic” emphasize gene regulatory mechanisms that do not alter primary DNA sequence while paying less attention to documenting heritability [98].

In the context of this review, heritable epigenetic influences are most relevant. One example of a classical, heritable epigenetic influence is imprinting. Imprinting conveys information from parent to child through mechanisms that include DNA methylation or histone acetylation. These mechanisms retain the primary DNA sequence but can dramatically alter function of specific genes. DNA methylation at CpG sequences in the promoter regions of genes can profoundly alter gene transcription. Since methylation during the course of maternal oocyte (or paternal sperm) development is key to this process, familial patterns of gender-specific transmission can provide evidence for this subset of heritable epigenetic influence. The modest quality of current family datasets for addiction renders them a relatively weak basis for any strong inferences concerning parent-of-origin effects. Nevertheless, there is no segregation data of which we are aware that supports strong parent-of-origin effects on substance dependence. Thus, while there are obvious and large roles for nonheritable “epigenetic” influences in the biology of addiction, there is no current compelling evidence that there are any strong effects of overall heritable “epigenetic” influences, as classically defined. We nevertheless need to be alert for such influences as we unravel effects of variants in specific genes.

I. The nature (and likely evolutionary sources) of the allelic variants likely to contribute to individual differences in vulnerability to addiction and related phenotypes

The analytic strategies described below are based on postulates that *common disease/common allele models* hold for many of the variants that alter vulnerabilities to addiction and related phenotypes. Rare variants may also explain significant fractions of the genetic risk for common diseases. However, increasing evidence supports roles in addiction

vulnerability for allelic variants that are currently common and are thus likely to be “old” in an evolutionary sense. Data indicating that such variants can be identified in diseased individuals from European, African and Asian genetic backgrounds also point, in general, to variants of substantial age. How could **genetic selection** act on such common functional allelic variants over the large number of generations that are implied by this “substantial age”? It is conceivable that some currently-common allelic variants could exert polygenic influences on addiction vulnerability without exerting positive or negative selective effects during lengthy evolutionary histories. However, it also seems likely that many allelic variants that influence addiction vulnerability can provide **balancing selection**; favorable in some individuals or organs or circumstances and unfavorable in other individuals or organs or circumstances. Balancing selection might thus maintain relatively high frequencies of multiple functional allelic variants in the population over long periods of time.

The biology of some genes might allow for common, functional allelic variants that could escape selective pressures or exert balancing selection over many generations. However, other genes might not be able to harbor such allelic variations without engendering selective pressures that would reduce the frequency of one of the allelic variants in the population over time. Common allelic variants that are able to influence addiction vulnerability are thus likely to be restricted to a subset of the genes whose products are involved in addictive processes. An important consequence of this logic follows: if a gene fails to display variants that influence vulnerability to addiction, the gene’s products are not at all excluded from involvement in addiction. On the other hand, convincing data that implicates a gene’s common variants in addiction should prompt us to consider mechanisms whereby such variants might provide balancing (*eg* both positive and negative) selective influences in the differing environments through which the ancestors of current human populations have passed.

How does this discussion of common disease/common allele hypotheses relate to the postulates of genetic heterogeneity noted above? None of the above discussion about common alleles and common variants precludes (or even reduces the likelihood of) contributions of rarer (or even “private”) allelic variants, including those that have arisen more recently in evolutionary time. Recently-arising variations would be much more likely to persist for a number of generations even in the face of even moderately-negative influences on survival or fertility. Indeed, based on experience with other genetic disorders, it may be worthwhile to actively search for effects of rarer “phenocopy” variants in genes that are initially identified based on common (and evolutionarily older) allelic variants [99]. A rarer copy number variant might contribute to addiction vulnerability by altering levels of expression of a gene that also contains more common allelic variants that alter expression *via* SNPs in other gene elements, for example [94–97]. Such considerations support searches within identified loci for molecular genetic heterogeneity relevant to addiction.

In the analyses presented in this review, we focus on addiction-associated allelic variants that lie in genes. Evolutionarily-old common haplotypes (*eg* groups of nearby variants that travel together through generations) that lie within genes are among the most likely to be tagged by SNP markers that are represented on current microarrays. Haplotypes that involve genes are thus among the most likely variants to exist in currently-reported datasets. It seems reasonable to postulate that many of these allelic variants that lie within genes provide regulatory variants that alter expression or regulation. Other variants are likely to alter mRNA half-lives or mRNA splicing. Variants that alter mRNA splicing could occur at the locus of the affected gene (*cis*) or at genes at different loci that alter generic mRNA splicing processes (*trans*). Reproducible association of A2BP1 gene variants with addiction vulnerability, for example [12], provide a good candidate for *trans* effects on mRNA splicing, since this gene’s product regulates splicing and thus are likely to modify the

functions of a number of other genes expressed in brain. It also seems likely that a minority of the addiction-associated variants will involve missense effects on expressed proteins.

It also seems likely that many addiction-associated variants will lie outside of genes, at least as we currently understand them. Loci reproducibly associated with diabetes/body mass, for example, lack conventional hallmarks of “genes”, such as expressed sequences [7]. While the analyses in this review focus on the identification of variants within genes, we should also remain alert for roles for “intergenic” variations in chromosomal regions that lie between currently-understood genes.

II. Samples for genome studies of human addiction vulnerabilities and related phenotypes (Figure 2)

A. Sample 1) *European-American polysubstance abusers and controls.* European-Americans volunteered for research at the NIH IRP (NIDA) in Baltimore, Maryland based on word of mouth referrals and newspaper advertisements. Volunteers self-reported their ethnicities, provided drug use histories and provided DSM (diagnostic and statistical manual) diagnoses of substance use disorders [10, 100, 101]. “Abusers” displayed heavy lifetime use of illegal substances [33] and dependence on at least one illegal substance. “Controls” displayed neither abuse nor dependence on any addictive substance and reported no significant lifetime histories of use of any addictive substance. Individuals with intermediate levels of lifetime substance use without dependence were thus not included in analyses of substance dependence, although a number of them were included among samples studied for cognitive abilities (*see Samples 15 and 16, below*). “Control” individuals thus combined those with no lifetime experience with any addictive substance with those who had modest to moderate exposures to legal addictive substances.

B. Sample 2) *African-American polysubstance abusers and controls:* African-Americans who volunteered for research at the NIH IRP (NIDA) were also characterized and separated into “abuser”, “control” and samples with intermediate lifetime uses of substance use, as noted above.

Efficacies of recruiting subjects for association studies (and comparison with recruitment for linkage based studies): Little currently-available data documents the features of drug abuse research volunteers who might consent to participate in molecular genetic studies for linkage or association. We thus describe the details of the recruitment of subjects for these studies at the NIH (IRP) facility in Baltimore, MD during a 30 month period (J H, GRU *et al*, *unpublished observations*). During this period, 13,969 individuals were screened by telephone and 2633 were interviewed in person for all (genetic and nongenetic) studies at this research facility. These individuals were 68% African American, 29% European American, 1% Hispanic and 72% male, with average age 35.

613 unrelated proband individuals from the group of interviewed subjects were offered participation in this genetic study, based on the availability of screening resources. No individual who was offered participation during this time period refused to participate. The individuals who accepted participation averaged age 34 and were 72% male, 72% African American, 26% European-American and 2% of other ethnicities, based on self-report. Subjects accepting research participation in this study thus appear representative of this overall research volunteer population in this area. They also share some characteristics of the drug abusing population in Baltimore, based on population trends and 1981 data from the Baltimore Epidemiological Catchment Area (ECA) site that identified 57% of the males and 41% of the females who displayed substance abuse and/or dependence as European-American (J. Anthony, *personal communication*, 1998).

Each volunteering subject was offered three choices concerning family member contacts. 112 (18%) probands provided Type I consents which allowed investigators to contact their family members. Seventy-three percent of these individuals were male, 66% were African-American, 30% European-American and 3% of other ethnicities. 312 (51%) probands gave Type II consents, stating that they would contact their family members. Seventy-two percent were male, 74% were African-American, 23% European-American and 3% of other ethnicities. 189 (31%) refused family member contacts. Seventy-two percent were male, 74% were African-American, 23% European-American and 3% of other ethnicities.

For 33% of the pedigrees for which the proband had provided Type I consents, at least one member could be reached by telephone or mail. At least one member of 12.5% of the pedigrees for whom the proband had provided Type II consent called an investigator and kept an appointment for study participation.

Of the 79 pedigrees from which family members made and kept appointments for study participation over this 30 month time period, 75% had African American probands and 69% had male probands. The sizes of the potentially-accessible sibships in the pedigrees of these individuals, as described by the probands, were: 1 (22%), 2 (23%), 3 (13%), 4 (11%), 5 (12%), 6 (9%), 7 (7%), 8 (1%), and 11 (2%). The numbers of accessible parents from these pedigrees was 0 (10%), 1 (28%) and 2 (62%). Average families for which more than the proband were accessible thus had 1.5 accessible parents and ca. 3.5 accessible sibs.

Over this 30 month time period, DNA and clinical information were collected from 2/3 of the members of the average pedigree from which any member came for an appointment. More complete sampling was obtained from smaller than from larger pedigrees. Of the 79 pedigrees for which DNA and clinical information were obtained, fifty-four had 2 members, 15 had 3 members, 7 had 4 members and 3 had 5 members. By the end of this period, DNA and clinical information was successfully collected from 2.5 members of the average pedigree. 14 of these pedigrees contained same-gender and 15 opposite-gender sibs within 5 years of the age of the proband who were discordant for drug abuse phenotype. 18 sibs were within 5 years of age of the proband and concordant for substance abuse phenotype. Thirteen sibs were more than 5 years different in age from the probands and discordant for phenotype, while 4 were concordant.

Reliability of pedigree structure reports was assessed by comparing family structure information provided by the proband and another first degree relative informant from these 79 families. Parents agreed (completely) with 100% of the proband's pedigree assignments, while sibs agreed (completely) with 70% of all possible pedigrees. Disagreements were largely due to differential reporting of half vs full siblings.

Reliability of information about drug histories was assessed by comparing drug use survey (DUS) and family history/research diagnostic criteria (FH/RDC) estimates of drug use by the proband with FH/RDC estimations from first-degree relatives. 73% of parental evaluations agreed (completely) with the probands' evaluations of all pedigree members' status with respect to all abused substances. 81% of sib evaluations were concordant (completely) with the proband's evaluations of his/her own drug use. Differences were most prominent in both cases due to family member underestimation and/or under-reporting of offspring/sib drug use.

As estimate of the validity of drug use survey (DUS) quantity/frequency estimates was obtained by comparing DUS ratings for substances with criteria identified as strongly heritable in work of Tsuang and colleagues [102] "never used, used fewer than 5 times, and used more than 5 times". All individuals who denied using alcohol, cocaine, heroin,

cannabis or nicotine when questioned as part of the addiction severity index screening received “0” scores on the DUS for the appropriate substance. All but two of the individuals who scored 2+ or 3+ on the DUS reported use more than 5 times with a separate instrument on a separate occasion. One individual who reported use of a substance five or fewer times during screening obtained a 2+ and one a 3+ score on the DUS for cannabis. Individuals who reported 1–5 lifetime uses during screening with the addiction severity index received 1+ DUS ratings on several occasions. The fraction of individuals who were rated as 1+ on DUS and who reported use of substances between one and 5 times on a different scale was 7%, 12%, 24%, 32% and 39% for alcohol, nicotine, cocaine, cannabis and heroin, respectively.

Although it is encouraging that no individual who was offered participation in this study refused to participate themselves, only one fifth of the probands agreed, during initial evaluations, to provide permission for research staff to directly contact other members of their families. Individuals who agree to contact their own family members do so at rates substantially lower than the contact rates achieved when family contacts are initiated by research staff members. Such differential participation may well provide occult differences between samples collected for association *vs* linkage studies.

It may also be important to consider that these subjects consented to studies in which genotyping was performed using pooling approaches that provide maximal protection from research risks (*see below*). Samples collected in studies which propose to conduct unlimited high density individual genotyping in any number of different laboratories might well experience different consent rates from some subgroups of participants, providing an additional confounding factor.

Genotyping: The primary data reviewed here is based on assessments of allele frequencies in multiple DNA pools that each contained equal amounts of DNA from 20 individuals of the same racial and phenotype group. Each DNA pool was assessed on four sets of four arrays, two from “100k” and two from “500k” Affymetrix microarray sets. We also document some of the features of assessments using “1M SNP” Affymetrix 6.0 arrays.

600k methods used for these samples revealed $r = 0.95$ correlations between pooled and individual genotyping in extensive validation studies [12] (*see below*). Much of this variation (1–2%) can be attributed to variations in pipetting and DNA quantification required for pool construction (TD and DW, unpublished observations, 2002–2006). The remainder of the variance is reduced for arrays used at the end of a study, in comparison to validating experiments that use data from some of the first arrays of this type that were studied within this laboratory. The $r = 0.95$ correlation is thus likely to reflect an upper limit of the variance noted in actual disease *vs* control comparisons. 1M SNP assays can reveal even higher correlations between individual genotyping and pooled genotyping datasets (Liu *et al*, *in preparation*).

Data for each SNP provides a score that provides a continuous measure of the percentage of its two alleles in the DNAs from the hetero- and homozygous individuals represented in each pool. For SNPs that displayed nominally-significant differences between addicted and control individuals in 600k studies, there were $r = 0.9$ correlation between the magnitude of the differences between pooled *vs* individual genotyping approaches. Variance from array-to-array (assessing the same DNA pool) and variance from pool-to-pool were both modest, *ca* 3%, supported good validity of this pooling data. Results from 1M SNP assays provide even more modest pool to pool and array to array variances (Liu *et al*, *in preparation*).

C. Sample 3) *European-American alcohol dependent and control* [33] Unrelated individuals sampled from pedigrees collected by the Collaborative Study on the Genetics of Alcoholism provide an interesting sample for several reasons. Dependence on alcohol and other substances has been carefully characterized in these individuals using validated instruments. Unrelated control individuals free from substance abuse or dependence diagnoses, largely individuals who marry into these pedigrees, are available. We thus identified 120 unrelated alcohol dependent individuals and 160 unrelated unaffected controls with self-reported European-American ethnicities.

Genotyping: Allele frequencies were assessed in 14 DNA pools, each containing equal amounts of DNA from 20 individuals of the same phenotype group, using four sets of four arrays, two from “100k” and two from “500k” Affymetrix SNP arrays, using approaches that were extensively validated as noted above.

D. Sample 4) *Taiwanese methamphetamine-dependent vs control:* Unrelated subjects recruited in Taipei included 140 methamphetamine-dependent individuals diagnosed independently by each of two psychiatrists using DSM IV criteria [103–108] and 240 matched Han Chinese controls who denied any history of use of illegal drugs and denied any histories of psychotic symptoms. Subjects were 30% female and averaged 32.5 $\pm$ 10 years of age. Dependent individuals reported methamphetamine use > 20 times/year, or described well-documented methamphetamine psychosis with lower levels of regular use. They denied histories of psychosis either prior to methamphetamine use or in relation to other psychedelic drugs. Most reported use of at least one other addictive substance. Controls denied illegal drug use or psychotic symptoms and were matched for gender and age.

E. Sample 5) *JGIDA Japanese methamphetamine-dependent vs control:* Japanese subjects were 21% female and averaged 40 years of age. One hundred methamphetamine dependent subjects were in- or outpatients of psychiatric hospitals in these regions that participate in the Japanese Genetics Initiative for Drug Abuse (JGIDA) [12, 104–116] and met ICD-10-DCR criteria F15.2 and F15.5 for methamphetamine dependence in independent diagnoses made by each of two trained psychiatrists based on interviews and review of records. Ninety one percent revealed histories of methamphetamine psychosis, 89% used methamphetamine intravenously, 62% also abused organic solvents and most abused at least one other substance. Subjects who displayed clinical diagnoses of schizophrenia, other psychotic disorders, or organic mental syndromes were excluded. Controls were one hundred age-, gender-, and geographically-matched staff recruited at the same institutions who denied use of any illegal substance, abuse or dependence on any legal substance, any psychotic psychiatric illness, or any family history of substance dependence or psychotic psychiatric illness during interviews with trained psychiatrists.

Genotyping: We assessed allele frequencies for methamphetamine-dependent and control subjects in DNA pools, each containing equal amounts of DNA from 20 individuals of the same phenotype group, on four sets of arrays, two arrays from “100k” and two from “500k” Affymetrix sets (Sample 5) or two arrays from “500k” Affymetrix sets (Sample 4).

F. Sample 6) *Australian and US dependent vs nondependent smokers of European ancestry* [3].

Dependent smokers of European ancestry were diagnosed using Fagerström Test for Nicotine Dependence (FTND score ≥ 4) criteria and compared to smokers did not display FTND dependence (scores = 0) [3]. About 1/4 of the individuals who displayed “dependence” by FTND standards did not display DSM IV nicotine dependence. About 1/4 of the individuals used in this study as “controls” *did* display DSM nicotine dependence.

Genotyping: Allele frequency data was assessed in 16 DNA pools, each containing equal amounts of DNA from 60 individuals using a single set of 49 arrays that assessed 2,427,354 SNPs. Methods used for these samples revealed $r = 0.85$ overall correlations between pooled and individual genotyping for all SNPs. However, there were much more modest, $r = 0.58$ correlations between dependent vs nondependent data derived from pooled vs individual genotyping for the SNPs that displayed nominally-significant differences. Individual genotyping followed up nominally-positive results for 39,213 SNPs in 1,050 dependent and 879 nondependent smokers. The convergence analyses presented here use data from the subset of these 39,213 SNPs that lie within genes.

G. Sample 7) WTCCC bipolar disease vs control: Genome wide association for bipolar disorder compared controls to 1868 UK individuals of European descent with bipolar disorders [1]. Bipolar mood disorders were diagnosed using Research Diagnostic Criteria. Uncharacterized control samples include: a) 1480 individuals from a 1958 Birth Cohort sample b) 1458 individuals from a United Kingdom Blood Service sample of consenting blood donors and c) individuals with disease phenotypes whose genetics was deemed unlikely to overlap with the genetics of bipolar disorder [1].

Genotyping for 436,604 autosomal SNPs analyzed was performed using Affymetrix 500k arrays with allele calls made by a CHIAMO algorithm with a 0.9 *a posteriori* probability threshold [1]. 436,604 of the 469,557 SNPs assessed could be assigned confident chromosomal localizations. A p value for each SNP was determined based on its χ^2 test for significance of allele frequency differences in bipolar vs control (actually control plus other disease) subjects [1].

H. Sample 8) NIMH bipolar vs control: Genome wide association was assessed in controls compared to 461 unrelated bipolar I probands of self-reported European-American ancestry who were selected from families which included at least one affected sibling pair who participated in the NIMH Genetics Initiative (<http://nimhgenetics.org>) [2]. Probands were assigned a 'confident' diagnosis of DSM-IV bipolar I disorder by each of two trained clinicians. 563 unrelated control individuals of European-American ancestry who failed to display evidence for DSM-IV criteria for major depression, any history of bipolar disorder or any history of psychosis were recruited by a marketing firm.

I. Sample 9) German bipolar vs control: 536,288 autosomal SNP genome-wide association was assessed in controls compared with 772 bipolar I patients diagnosed using DSM-IV criteria who were recruited from consecutive hospital admissions [2]. 876 population-based controls were randomly recruited; individuals with histories of affective disorder or schizophrenia were excluded.

For *genotyping*, NIMH samples (Sample 8) were divided into seven bipolar and nine control pools of 50–80 subjects/pool. German samples (Sample 9) were divided into 13 bipolar and 10 control pools of 42–60 subjects/ pool. SNP allelic distributions were assessed using duplicate Illumina HumanHap550 assays (Illumina Inc., La Jolla, CA, USA) [2]. Normalized allele frequencies were calculated from raw intensity data averaged across duplicate pools to obtain a relative allele frequency estimate for each SNP in each pool. SNP with allele frequencies that displayed > 2% variance between replicate pools were excluded. *t* tests compared pool-to-pool variation within phenotypes to phenotype-to-phenotype differences.

J. Sample 10) Unrelated members of NHLBI twin pairs for assessment of frontal brain volume: 242 unrelated individuals were selected randomly from members of twin pairs from a population-based registry of European-American male World War II veteran twin pairs

who received volumetric MRI studies as part of the National Heart, Lung, and Blood Institute (NHLBI) Twin Study. Studies were performed when subjects averaged 72.6 years of age and reported 13.6 years of education. Frontal lobar volumes, corrected for intracranial volumes, were obtained [117] in ways that produce inter-rater reliabilities > 0.90.

For *genotyping*, DNAs were carefully quantitated and combined into 12 pools that each represented about 20 subjects, based on estimates for frontal brain corrected for total cranial volume. We thus constructed 4 DNA pools from individuals with the lowest, 4 pools from individuals with intermediate, and 4 pools from individuals with the highest estimates of total frontal brain volumes and subjected these DNA pools to Affymetrix 500k genotyping as noted above. *t* tests compared differences between highest and lowest brain volume groups. SNPs that displayed nominally-significant *t* values and that also displayed rank order of allelic frequencies with either highest tercile > intermediate tercile > lowest tercile or the converse are included in these analyses.

K. Sample 11) *Framingham study participants for assessment of frontal brain volume:*

Subjects were 705 stroke- and dementia-free participants in the Framingham study, age 62 +/- 9 yrs, 50% male, who received volumetric brain MRI studies that were analyzed as noted above for the NHLBI subjects.

Genotyping provided data from 70,987 autosomal SNPs using Affymetrix 100K arrays. Allele frequencies for SNPs that displayed a) minor allele frequencies > 0.10, b) genotype success > 0.80, and c) Hardy-Weinberg equilibrium *p* > 0.001 were used. A generalized estimating equation provided corrections for familial relatedness and other covariates [8].

L. Sample 12) *European American smokers who successfully vs unsuccessfully quit smoking in trials in Philadelphia, Washington DC and Buffalo, New York* responded to advertising and physician referrals for help in quitting smoking [118, 119]. Subjects aged 18 – 65 enrolled in a randomized clinical trials for smoking cessation accompanied by standardized behavioral counseling that utilized a blinded, placebo controlled trial of bupropion 300 mg/d or matching placebo × 10 weeks or an open label trial of nicotine nasal spray versus nicotine patch × 8 weeks [118]. 126 individuals with biochemically-confirmed abstinence for at least the 7 days prior to both 8 week and 24 week assessments were contrasted with 140 unsuccessful quitters who were not abstinent at either time point.

M. Sample 13) *European American smokers who successfully vs unsuccessfully quit smoking in trials in North Carolina:* Participants received either active 21 mg/day or placebo nicotine skin patches for two weeks before the targeted quit date as well as mecamlamine 10 mg/day po prior to the target quit-smoking date [120]. After the quit-date, participants were randomly assigned to mecamlamine 10 mg/day vs matching placebo and 21 mg/24 h vs 42 mg/24 h nicotine skin patch doses. Fifty-five individuals reported continuous abstinence from smoking when assessed 6 weeks after the quit date with biochemical confirmation; 79 were not abstinent.

N. Sample 14) *European American smokers who successfully vs unsuccessfully quit smoking in trials in Rhode Island:* Participants engaged in a ten week, double-blind placebo controlled trial of placebo or bupropion (150 mg/day for the first 3 days, then 300 mg/day) with a target quit date one week following initiation of drug or placebo [119]. Sixty individuals with biochemically-confirmed abstinence for at least the 7 days prior to both the end of treatment and 24 week assessments were contrasted with 90 unsuccessful quitters who were not abstinent at either time point.

Genotyping for samples 12–14 used Affymetrix 500k arrays and multiple pools of DNAs (*n* = 16 to 20), as noted above.

O. Sample 15) African American individuals (ages 18 – 65) with different levels of general cognitive ability as assessed by the Shipley Institute of Living scale. Research volunteers with a variety of different levels of lifetime use of addictive substances volunteered for research protocols at the NIH (NIDA) facility in Baltimore, Maryland, as described above. Eighteen pools were constructed with DNA from 20 individuals each (33% female, average age 32.1). Mean cognitive function scores estimated from Shipley Institute of Living scales ranged from “IQ” equivalents of 75.9 to 109.2 for the individuals in these DNA pools.

P. Sample 16) African American individuals (ages 18 – 65) with different levels of general cognitive ability as assessed by the Shipley Institute of Living scale. Eleven pools were constructed with DNA from 16 individuals each (34% female, average age 31.8). Mean estimated “IQ” scores ranged from 79.1 to 112.3 for the individuals represented in these DNA pools.

Genotyping for samples 15–16 used Affymetrix 500k arrays and pools of DNAs ($n = 16$ to 20), as noted above. The nominal significance of the correlations between pool-to-pool differences in assessments of allele frequency and pool-to-pool differences in Shipley scores was assessed for each SNP.

Q. Sample 17) Alzheimer’s disease vs control: brain donors: Subjects were 1086 brain donors of mean age 82 who were 43% male and at least 65 years of age at death. Brains and clinical data met pathological criteria for Alzheimer’s disease or control status. DNAs were subjected to Affymetrix 500k genotyping. From the files TGEN_WGA_DATA_recode_ped.txt [4], genotype calls for 552 control and 859 AD individuals allowed us to calculate χ^2 values for the AD vs control differences for each SNP, select SNPs which displayed χ^2 values with $p < 0.05$ as nominally-positive and assess which of these nominally-positive SNPs fell into chromosomal clusters such that at least 3 nominally-positive SNPs that represent both array types lie 25 kb from each other.

R. Sample 18) Alzheimer’s disease vs control: memory clinic participants: Seven hundred fifty-three AD cases and 736 controls with European ancestry were recruited in Canadian memory clinics. Probable Alzheimer’s disease was diagnosed by clinical criteria, and controls selected who displayed no histories of memory impairment or any impairment on neuropsychological tests [5]. DNAs were subjected to Affymetrix 500k genotyping. From files in http://ctr.gsk.co.uk/Summary/genetic_observational/studylist.asp, p values for each SNP, derived from Fisher’s exact tests, were extracted and data was analyzed as described above.

S. Sample 19) Individuals with scores on tests of neuroticism 1038 individuals from southwestern England sites with European backgrounds and with high neuroticism (“N”) scores on the revised Eysenck Personality Questionnaire and 1016 individuals with low neuroticism scores (63% female) were studied along with a replication sample of 831 high vs 702 low neuroticism individuals (61% female) [121]. Genotyping of eight pools of DNA from mouth swabs compared: (1) men with high N scores ($n = 112$), (2) men with low N score ($n = 158$), (3) men with very high N score ($n = 245$), (4) men with very low N score ($n = 238$), (5) women with high N score ($n = 320$), (6) women with low N score ($n = 205$), (7) women with very N high score ($n = 340$) and (8) women with very N low score ($n = 436$). Very high or low N scores were defined as more than 1.5 s.d. from the mean score adjusted to age and sex (on average 2 s.d.), while high and low N scores were between 1 and 1.5 s.d. from the mean score (on average 1.3 s.d.). Data for relative allele score from the 452 574 SNPs from 100 and 500k Affymetrix arrays with minor allele frequencies above 5% were obtained from five replicate arrays used to assess each sample.

III. Selected Methodological Issues

A. Genotyping using pooled DNAs

Genotyping for genome wide association studies can be performed in either individual samples or in pools of DNAs from individuals with the same racial/ethnic background and the same phenotype. Pooling strategies have several advantages. Pooling fits well with association genetics, can efficiently perform allele typing, preserve confidentiality and reduce costs [10, 122–124].

Not all pooling strategies are alike. “Single-pool” strategies seek differences in allele frequencies in comparing data between a single pool of DNA from diseased individuals vs a single pool of DNA from control individuals. Such results generate hypotheses. However, such designs, and related designs with very few pools, provide little ability to differentiate between: 1) the variability between disease vs control samples with 2) the variability within disease or within control samples.

We focus here instead on “multiple pool” strategies. With careful attention to a large number of small details, these approaches can provide accurate allele typing. Multiple pool approaches provide estimates of a) differences between disease vs control samples, b) variability within disease samples and c) variability within control samples. Assessment of the differences between disease and control in the context of assessments of the variability within disease samples and within control samples allows us to use standard statistical approaches to assess the significance of the results.

Here, we provide assessment of several of the steps necessary to validate and characterize features of the power and sensitivity of multipool genome scanning strategies. Used with high densities of genomic markers, carefully-performed multiple pool studies can 1) provide increased study feasibility and 2) preserve virtually absolute genetic confidentiality with 3) only modest effects on the sensitivity and specificity of genome wide association.

1. DNA quality, quantity and contamination: Care in assessing and maintaining the quantity and quality of DNA in every sample is crucial for pooling studies. Rough DNA quantitation procedures that are routinely used in most genotyping laboratories are likely to introduce such substantial errors that many of the apparent disease/control differences will actually arise from occult over- or under-representation of genotypes of individuals with misquantitated DNAs. Uneven DNA quality can provide the same selective over- and under-representation of genotypes of selected individuals in each pool, leading to more false positives and less sensitivity for detection of true positive results. Contamination of pooled DNA samples with even a modest amount of DNA from laboratory personnel or other sources can also provide difficulties for pooling procedures.

2. Numbers of individuals per pool and numbers of pools: To obtain maximal benefits from multiple pool genome wide association: 1) The numbers of individuals in each pool should be sufficient that even sophisticated analyses of pooled data cannot reconstruct individual identities or genotypes. Treatments of this subject suggest that pools need to contain more than 4–5 individuals for maximal confidentiality protection [125]. 2) The numbers of individuals in each pool should allow significant cost/time savings compared to individual genotyping. 3) The numbers of pools should be sufficient to provide good estimates of pool-to-pool variability, which can then be used to compare to the differences between disease and control individuals using standard statistical tests.

Multiple genotyping assessments of each DNA pool can aid the precision of estimates of pool-to-pool variability as well. We have used three to four microarrays to assess DNAs

from each pool. These numbers are based on preliminary studies that seek to optimize estimates of “true” relative allele frequencies at acceptable cost. We construct each pool using DNAs from 20 individuals of the same self-reported ethnicity and same disease or “control” phenotype. Thus, we obtain results in quadruplicate at 1/5 – 1/7 the reagent costs of individual genotyping using a single array set per person.

3. Number of different replicate samples: Power calculations assess the likelihood that an experiment can detect a difference of a certain magnitude in a specific SNP. Experiments require reasonable levels of protection against false-positive results, α . They also require reasonable power, β . Genome wide association requires many repeated measures; considerations of α and β thus need to be applied to hundreds of thousands or millions of SNPs.

One approach to the dilemma raised by these large numbers of multiple comparisons has been to propose single studies with larger and larger “n”. However, accretion of very large samples is expensive. Attempts to assemble large samples from smaller subsamples collected at various sites also run greater and greater risks of incorporating increasing numbers of occult heterogeneities that could provide confounding influences on the results obtained [126].

The approaches that we outline here rely on initial use of achievable sample sizes that may be more likely to be more homogeneous. Initial samples can nominate sets of SNP markers, genomic regions and genes that can be studied in additional independent replicate samples. A requirement that genes display SNPs whose allelic frequencies distinguish disease from control individuals in multiple samples is one of the few assurances against false-positive results that is likely to pass ultimate statistical muster and also to yield feasible experimental designs. There is a downside to this approach: rates of false negative results are also unavoidably elevated by requirements for replication (*see below*).

4. Validation studies assess the fits between individual and pooled genotyping in a number of different ways. These include assessments of the concordance between results from sense and antisense probes for the same SNP and concordance between results for the same SNPs obtained using arrays of different types. The same DNAs can be pooled multiple times, and the same pool can be analyzed multiple times, in order to further assess concordance.

We focus here on a core validating test for pooling. This core test comes from analyses of the relationships between a) “observed allele ratios”, background-subtracted, normalized hybridization intensity ratio values obtained from different pools of DNAs and b) “expected allele ratios” the fraction of *eg.* “A” and “B” alleles obtained from individual genotypes. We and others have compared allelic determinations from individual DNAs *vs* results from pools with equal or differing amounts of DNA from small numbers or larger numbers of control individuals in use of HuSNP, Affymetrix 10k, 100k, 500k and 1million SNP products, Perlegen arrays and Illumina 300 and 500k arrays [11, 33, 39, 41, 42, 122, 123, 127].

Using 500k Affymetrix arrays, we have evaluated pooling using equal and varying amounts of DNA from CEPH individuals. Overall data for 150,000 SNPs from these comparisons produces correlations between pooled and individually-determined genotypes of $r^2 = 0.95$ (Figure 3). These overall results derive from studies of 5475 and 6230 informative *Nsp* I and *Sty* I SNPs in experiments in which equal amounts of DNA from homozygotes were mixed (correlation ca 0.9); 10,032 informative *Nsp* I and 10,249 informative *Sty* I SNPs in experiments in which these same DNAs were mixed in 1:1, 1:5, and 1:15 ratios (correlations 0.96 and 0.98, respectively) and 31,201 informative *Nsp* I and 39,827 *Sty* I SNPs for studies

of one homozygote and one heterozygote (correlations 0.89 and 0.92, respectively). When we compare these results with those reported using long range PCR products, using the reported procedure of eliminating the 9% of SNPs that yielded the more problematic correlations [127], the overall correlations for the remaining SNPs is 0.98. Correlations using 1M Affymetrix SNP arrays are at least as strong, with $r^2 > 0.98$.

B. Power assessments

Approaches to assessing power of genome wide association have used a variety of assumptions about the frequencies of disease-causing alleles, the heterogeneity and penetrance of disease-causing alleles, marker frequencies, and the nature and distribution of linkage disequilibrium across the genomic intervals surveyed [128, 129]. Many approaches to this problem use linkage disequilibrium distributions identified in HapMap samples, even though these HapMap individuals represent only very small subsets of several current human populations.

We have been impressed by the variability in the detailed distribution of linkage disequilibrium across different genomic loci in different samples [130, 131]. We have also been impressed by the possibilities that approximations of this variability could be modeled, on average, by simple functions. Under these circumstances, estimates of effects of sample size, locus-specific effect sizes for the underlying functional alleles, genetic heterogeneity, penetrance and marker density can produce reasonable models that can allow assessments of the effects of variation in these parameters on power.

We have focused on diallelic markers and disease/no disease phenotypes. We have developed a model to simulate the effects of varying these parameters that has resulted in a program “Gene Detective”. We can use this approach (*see Supplement for details*) to simulate the *ca* 620,000 diallelic markers reported to date for samples with $n = 400$ case and $n = 400$ controls with nominal 0.05 α levels (Figure 4). We can observe effects of sample size, heterogeneity/penetrance ratios, marker minor allele frequencies and disease frequencies. Such effects are each relatively modest over a reasonable range of values for genome-wide distributions of linkage disequilibrium. However, there is a striking relationship between power and effect size. Power to detect effects that would produce odds ratios of less than 1.2 –fold is modest, while power to detect effects as high as 1.7 fold is relatively good.

Increments in marker density from 630,000 to 1,000,000 and increases in sample sizes from $n = 400$ to $n = 2,000$ samples in case and control improve power (Figure 5). However, the steep relationships between power and effect sizes are also found in these simulations. Such results underscore the distinctions that we have made above concerning analytic approaches to “oligogenic” disorders in which variants at individual gene loci produce relatively large differences in risk vs “polygenic” disorders in which the effects at each locus are likely to be modest.

These power calculations apply only to the initial genome scans. As we note above, replicate studies aid in distinguishing false- from true positive associations, but also increase the cumulative number of false-negative results.

Limits on the precision of the power calculations that derive from the approach outlined here include limits of the precision of estimation of the parameters whose estimation is required. Several of these parameters can be estimated based on substantial empirical data. These include the size of the genome or genomic segment under consideration, the sizes of the samples of disease and nondisease control subjects studied and the value for α desired. The frequency $P[D]$ of the disease in the population under study is available from

epidemiological studies. It is thus important that the samples for association genome scanning display characteristics similar to those of the populations in which disease probabilities were determined, so that estimates of $P[D]$ are as accurate as possible.

To assess the statistical power of analysis, we have also used more standard power calculations. The program PS v2.1.31 [132] a) $\alpha = 0.05$, b) sample sizes equal to the numbers of pools from the current dataset, c) mean abuser/control differences of 0.05 and 0.1 and d) standard deviations from the SNPs that provided the largest differences between control and abuser population are used in several of the discussions below. We have also used data from the Genetic Power Calculator (<http://pngu.mgh.harvard.edu/~purcell/gpc/>) for some analyses.

C. Analyses

1. Achieving significant genome wide association in single samples vs seeking replication and generalization in multiple samples—*a) Single sample approaches:*

As noted above, genome wide association gains power to detect variants in more and more of the genome as more and more genetic markers, generally SNPs and/or copy number variants, are assayed. Since many hundreds of thousands of SNPs and/or copy number variants are assayed in current datasets, stringent approaches to correct for the large number of multiple comparisons are needed.

There is no clearcut consensus about any single method that will produce only true results from any single sample. One approach to concerns about the large numbers of comparisons that are key components of GWA focuses on achieving “*genome wide*” significance in *single samples*. Single samples that demonstrate genome wide significance in this way must contain single SNPs whose association displays a striking nominal p value, often in the neighborhood of $ca\ 10^{-8}$ [133, 134]. Such results may be the most likely to be published in prominent journals. However, in most studies with findings of this statistical magnitude, effects of variants at a single locus are sufficiently large that linkage studies also provided significant evidence at the same chromosomal locus [135, 136]. For “oligogenic” contributions to common, complex disorders, seeking association with “genome wide” significance in single samples thus provides a reasonable approach. When there is a large effect of a single gene, a number of corrections for multiple comparisons can be applied without creating many false negatives. Bonferroni corrections for multiple comparisons are advocated by some investigators, though they are generally agreed to provide a conservative correction [137, 138]. False discovery rate corrections can also be applied [139–141]. Permutation and Monte Carlo tests provide additional approaches [6, 142, 143].

As the expected effect of each locus falls from the large effects characteristic of oligogenic influences to the small effects that characterize polygenic influences, however, the sample sizes needed to generate p values in these ranges provide a daunting problem. Costs of individually genotyping such large samples become limiting in all but the best-supported enterprises [144]. The risks of introducing occult heterogeneities increase when subsamples are collected at a variety of distinct sites [126]. As more occult heterogeneities are included as disease and control samples need to be assembled from more and more diverse sources to achieve sufficient “ n ”, more and more of the results obtained may well represent “false positives” based on such occult sample heterogeneity for genetic background or for heritable traits that are not (nominally) being studied.

b) Replicate sample approaches: Here we use an alternative analytic approach that focuses on step-wise assessments. These step wise analyses address the problem of multiple testing by seeking nominally-significant results that can be replicated in several independent

samples. We can assess the significance of these replicated, nominally-significant results through use of Monte Carlo methods that correct for multiple comparisons.

It is important to emphasize the ways in which the step wise analyses presented here a) first identify evidence for genes that contain haplotypes that are found at different frequencies in single disease vs control sample comparisons and b) then identify evidence for genes that display haplotypes with such differing frequencies in multiple different samples that results are unlikely to be due to chance.

We thus a) first identify nominally-significant SNPs in each sample, 2) identify the clustering of such SNPs (within small chromosomal regions) in each sample, 3) seek replication, identifying small genomic areas in which clusters from multiple replicate samples from the same phenotype also identify clustered nominally-significant SNPs 4) seek generalization, identifying genes that contain clusters of nominally-positive SNPs from studies of related, genetically determined phenotypes.

The criterion used here identifies clustering based on chromosomal position. This approach allows direct comparison between datasets that assess different sets of SNPs in samples that may well differ in the details of their patterns of linkage disequilibrium. The Monte Carlo simulation methods used here do not make assumptions about the underlying distribution of the data assessed. Monte Carlo methods provide empirical p values based on repeated random samples from the actual datasets analyzed. Such approaches are especially useful when we seek to assess the significance of apparently-reproducible results from convergent data from multiple independent datasets which differ from each other in “n”, number and types of genomic markers, racial/ethnic background of the subjects and other key features. No alternative method of which we are aware provides as tractable a method for assessing the significance of results obtained in multiple samples without assumptions about underlying distributions of the data as do Monte Carlo approaches. We use 10,000 Monte Carlo trials in circumstances in which moderately-high significance is anticipated, and 100,000 trials in circumstances in which extremely-high significance is anticipated.

This approach seeks to identify genes with variants that are likely to play roles in addition and in related phenotypes. This approach allows for “locus heterogeneity”, and thus does not use the more stringent criterion that the same SNP is required to display nominal significance in each of the samples in which association data is said to support association at a specific gene locus. This approach allows for differences in phase of association, and thus does not use the more stringent criterion that the same allele of the SNP (or haplotype) must be associated with nominal significance in each of the samples in which association data is said to support association at a specific gene locus. The approach allows for different details of the patterns of linkage disequilibrium between marker and functional haplotype from sample to sample. The approach allows us to combine datasets in which different marker sets are used. With each of these limitations, it is clear that subsequent followup analyses are required. Analyses in the same and in additional independent samples are required to untangle any locus heterogeneity, to unequivocally identify which individual SNPs are associated, to identify pathological haplotypes and the phases with which they are associated with phenotypes in samples from different racial and ethnic backgrounds. While we describe examples of such follow-up studies for the NrCAM and NRXN3 genes below, it is important to note the limited numbers of genes for which such confirmatory follow-up data is available. In many circumstances, we believe that this sort of follow-up requires molecular biologic, behavioral and other evidence to buttress the data that comes from association genetics alone.

2. Stepwise approaches to analyses—a) *Determination of nominally-significant markers*

Nominal p values that come from “t” (for pooled data), “ χ^2 ” (for individual genotype frequencies) or “p” (for correlational approaches) statistics delineate the nominal significance of the differences between disease and control groups for each SNP. For pooled assessments, proper definition of the pool to pool variability is crucial for proper assignment of the appropriate nominal t value. However, the continuous results that come from pooled datasets do provide the additional statistical power characteristic of statistics based on continuous measures.

b) Identifying chromosomal clusters of nominally significant markers in single samples:

We focus on the SNPs whose chromosomal positions can be accurately determined. Since gender ratios differ substantially in many of these datasets, we omit data from sex chromosomes for most of these samples.

We focus on clusters of nominally-positive autosomal SNPs that lie within 100 or 25 kb of each other, depending on the density of markers available; the latter figure is closer to the average “haplotype block” length in the samples studied here [11, 12, 33, 34, 36, 37]. To use a valuable technical control that is possible with Affymetrix 500k reagents, we require that SNPs in each cluster come from both *Sty* I and *Nsp* I array types where possible [12]. In assessment of the data from each sample set, these criteria thus provide some assurance that haplotypes do occur at differing frequencies in disease vs controls. These criteria provide significant technical controls, based on requirements that multiple nearby SNPs must display positive results and that these positive results must come from two array types.

It is important to note that, if stochastic events produce a nominally “significant” association at a given SNP in a single sample, linkage disequilibrium with nearby SNPs might provide “cluster” of several SNPs with nominal significance in this single sample on stochastic grounds alone. Control for the possibility that these differences in haplotype frequencies are due to stochastic differences between samples thus awaits the next analytic step (c, *below*).

We test the nonrandomness of clustering of nominally-significant SNPs using Monte Carlo simulations. We can also use these approaches to identify the nonrandomness of clustering within genes. For each simulation trial, a random set of SNPs from the database that contains the results from these studies is subjected to the same analytic procedures that had been used for the actual data analysis. The number of trials for which the results from the randomly-selected set of SNPs match or exceeded the results actually observed from the SNPs identified in the current study is tabulated. Empirical p values are calculated by dividing the number of trials for which the observed results are matched or exceeded by the total number of Monte Carlo simulation trials performed. This method examines the properties of the actual SNPs contained in each dataset. It is therefore relatively robust despite the uneven distributions of SNP markers across the genome, differences in linkage disequilibrium across the genome in different samples and the different SNPs genotyped using different assays.

c) Identifying the clustered, nominally positive SNPs within the strongest positive support from several “replication” datasets: We next seek convergence between data from several “replicate” samples. We focus on samples that test the same underlying hypothesis (*eg* that common allelic variants contribute to genetic components of vulnerability to develop substance dependence). Some of these samples and their matched controls differ from each other on other bases (*eg* racial/ethnic background, primary substances of abuse). We thus use “replication” here in a restricted sense. Such a restricted use of “replication” allows us to reserve use of the term “generalization” to denote identification of genes whose pleiotropic influences are evident in studies of other heritable phenotypes that often co-occur with

addictions (*see below*). Obviously, there is also an aspect of “generalization” when comparing data from a) polysubstance dependent *vs* control samples collected from individuals of two racial/ethnic backgrounds with b) methamphetamine- dependent *vs* control samples collected from individuals with a third racial/ethnic background (*see below*).

Analyses focus on genes that are identified by clustered positive results from several samples. This approach, rather than focus on individual SNPs whose informativeness might differ in different samples, allows for some degree of genetic heterogeneity and for some sample-to-sample differences in the detailed patterns of linkage disequilibrium.

Clustering of positive results in the same gene in each of several independent samples is much less likely to represent purely stochastic effects than observations made in any single sample. Such clustering in multiple samples is more likely to reflect true differences related to the phenotype of interest, *eg* dependence on addictive substances. However, it is important to emphasize again that these criteria are aimed at identification of genes, rather than precise definition of exact disease-associated haplotypes. We thus allow the phase of association to differ between samples at this level of analysis. Detailed studies of the phase of association can provide a very valuable fine mapping tool to allow identification of the exact pathogenic haplotype [131].

d) Identifying the clustered, nominally positive SNPs within the strongest positive support from several “generalization” datasets: To seek possible *generalization* of results, we have sought chromosomal locations where the clustered positive data from several substance-dependence genome wide association samples lies near clustered, nominally-positive (and reproducibly positive) results from studies from other “related, heritable” phenotypes.

Baysian approaches to these analyses suggest that the stronger the evidence for coheritabilities of substance dependence and these “related” phenotypes, the higher the likelihood that molecular genetic studies will demonstrate true overlaps [145]. We focus on phenotypes that display good evidence for heritability from classical genetic studies, including evidence that complex genetics plays substantial etiologic roles. We focus first on heritable phenotypes that co-occur with addictions at frequencies much greater than those that we would expect if they were independent of each other. For example, even though substance dependence and bipolar disorder are each common, the product of their population frequencies does not nearly explain the *ca.* 2/3 of bipolar individuals who report abuse of or dependence on an addictive substance [146, 147].

Twin data that compares co-occurrence frequencies in monozygotic *vs* dizygotic twin pairs provides evidence for shared heritability for some of these phenotypes. For other phenotypes, the magnitude of genetic influences and the frequency of co-occurrence with substance dependence each indicate the likelihood of **pleiotropic** influences of some of the same allelic variants on both phenotypes. Finally, we also present here the idea that “**transitive**” genetic approaches may also identify evidence for generalization of effects of some pleiotropic influences. If substance dependence and intermediate heritable phenotypes share genetic overlap and co-occur, then a third heritable phenotype that is documented to co-occur with the intermediate phenotype might also share substantial heritability with addiction vulnerability.

Examples of heritable phenotypes for which twin data document shared genetic determinants include frontal lobe brain volume and cognitive abilities [15, 117]. Examples of heritable phenotypes for which co-occurrence makes shared genetics highly likely, *a priori*, include substance dependence and bipolar disorder [148, 149]. A possible “transitive” genetics comes from the shared genetics of substance dependence and cognitive

abilities/brain volume on the one hand, and the likely shared genetics of cognitive ability/brain volume and vulnerability to Alzheimer's disease, on the other hand. Data from cognitive function and frontal brain volume genetics thus provide potential intermediate phenotypes to link the genetics of addiction with that of Alzheimer's disease. Such links might or might not have been anticipated, based on equivocal evidence from current epidemiologic studies [150].

As we seek to document the extent of "generalization" of effects of alleles that were initially identified in studies of addiction, we test the null hypothesis that clustered positive results from the genome wide association data from addiction vulnerability do not converge with the chromosomal positions of clustered nominally-positive SNPs in comparisons of other phenotypes, *eg* bipolar *vs* control samples. Monte Carlo simulations that test this null hypothesis sample from data within the SNP datasets noted above. 100,000 trials allow estimates of the significance of the generalization of the effects of the alleles identified in studies of addiction vulnerability, as noted above.

e) Controls for the alternative possibilities that results could come from occult racial/ethnic stratification or assay noise. Several alternative hypotheses might explain observed results. To test some of these alternative hypotheses, we compare the clustered- positive SNPs from different samples with SNPs that display the largest allele frequency differences in appropriate control datasets. These comparison datasets include those that contrast allele frequencies in 1) European-American *vs* African American control individuals from NIDA samples [12]; 2) Japanese *vs* Han Chinese individuals from HapMap samples (JPT: Japanese from Tokyo and HCB: Han Chinese from Beijing), 3) control individuals sampled in different portions of the United Kingdom [1], and 4) SNPs that display the largest variances from array to array [12]. We can thus compare data from the true comparisons in our experiments to similarly-analyzed data from samples that test alternative hypotheses, providing substantial additional control evidence.

f) Results from alternative approaches: principal components (PCA) and hierarchical clustering. Several alternative approaches to analyses of genome wide association datasets can also provide interesting results that assess the structure of the pool-to-pool variance, based on data from 500 or 600k SNP sets. Principal components analyses (PCA) of the pool-to-pool differences in 500k data from sample sets from European-American, African-American and Asian samples divides the data along these racial/ethnic lines, as we might expect. However, PCA analyses also subdivide data from experiments studying two distinct United States samples of nominally-equivalent genetic background, Sample 1 NIDA European-American samples recruited in Baltimore, Maryland *vs* Sample 3 COGA European-American samples recruited in St Louis, Bronx New York, San Diego, California and other sites. Similarly, these PCA analyses separate the Asian samples recruited in Japan from those recruited in or near Taipei that are self-characterized as Han Chinese. Each of these results underscores the need for extremely careful matching of the racial/ethnic backgrounds of control and disease samples.

Hierarchical clustering is most conveniently limited to data from individual genes. Large 500k SNP datasets provide substantial limitations based on computer time. When we examine data from several genes using this approach in pools from a single racial/ethnic background, we can identify relatively clear patterns of separation of data from pools containing substance dependent individuals from pools containing control individuals. These hierarchical clustering approaches are independent of the principal analyses noted above. These results reassure us that modest association signals can be identified in many of these addiction-associated genes using a variety of different statistical approaches.

D. Ethical issues in high density genotyping of individuals who are selected due to self-reported illegal behaviors

Individuals who are individually genotyped in relationship to addiction and related phenotypes are subject to a number of potential risks. Some of these risks are shared with individuals who are subjected to high density genotyping in relationship to other disorders and phenotypes. Other risks are more likely to come to the fore in studies of illegal behaviors.

Concerns relating to insurability, employability, paternity determination and providing (or not providing) genotyped individuals with access to their genotypes and/or genetic counseling are shared by individuals with other complex disorders [151–153]. Pending legislation in the United States may mitigate several of these concerns, and they are reviewed elsewhere. We therefore will not consider these issues further here.

High density, individual genotyping of DNA from individuals who are addicted to illegal substances raises additional issues. Many of these individuals are likely to have experienced involvement in criminal activities that goes beyond use of illegal substances. Since the risks of high density individual genotyping in this population have not been as generally discussed elsewhere, we provide several lines of information that may inform thinking about these special ethical issues.

Increasingly-ubiquitous DNA testing related to criminal activities lies at the heart of these concerns. In the United States, each state has a DNA database that collects information from crime scenes and from offenders convicted of particular offenses. A combined DNA index system (CODIS) operates local, State, and national DNA profile databases from convicted offenders, unsolved crime scenes and missing persons. Numerous suspects have been identified through matches between DNA profiles from crime scenes and profiles from convicted offenders. A relevant website reports that the “success of CODIS is demonstrated by the thousands of matches that have linked serial cases to each other and cases that have been solved by matching crime scene evidence to known convicted offenders”. The European Union is just one of the other international entities with a similar system (<http://www.interpol.com/Public/Forensic/dna/dnafaq.asp>).

“Core” CODIS data comes from genotypes at 13 simple sequence length polymorphic (SSLP) loci. These loci lie near SNP markers that provide information about virtually all of these loci, providing a ready means of translating between SNP and SSLP genotypes. Other mitochondrial, sex chromosome and autosomal markers are also genotyped on substantial numbers of these DNA samples.

A recent, October 2007 analyses of the CODIS-linked DNA index system revealed individually-identifying genotype profiles for more than 5 million convicted offenders, as well as almost 200,000 DNA profiles from crime scenes (www.fbi.gov/hq/lab/codis/).

Almost 40% of males and 15% of females in cohorts from the areas of Baltimore from which Sample 1 and Sample 2 research volunteers come have experienced significant adult criminal justice system involvement (*eg* incarceration as adults) by the time they reach their late 20's (N. Ialongo, *personal communication*, 2008). It thus seems reasonable to conclude that several of the >3400 research volunteers from whom DNAs were collected to form Samples 1 and 2 might be at potential risk for matches with crime scene DNA profiles. Similar potential risks might also be incurred through genetic study participation by individuals who report dependence on illegal substances in other parts of the United States. While this problem is not unique to studies of the genetics of illegal behaviors, it appears to be much more likely in this than in most other areas of complex genetics.

Our laboratory, along with most other laboratories that work in this field, has established elaborate means of coding, providing physical protections and providing electronic protections for the electronic and paper records that might identify our research volunteers. Subjects are protected by confidentiality certificates obtained through the Department of Health and Human Services. Data from these studies is analyzed and reported in ways that do not identify individual subjects.

However, the strongest protection for individuals who volunteer for this work comes from development and use of DNA pooling approaches. Since these approaches never generate high densities of genotypes for any individual, it is impossible to abuse or misuse this pooled data for other unintended purposes. Pooling approaches provide these research volunteers with the strongest confidentiality protections that are currently available. Pooling may also merit increasing attention in other settings in which the downside risks of DNA-based personal identification are of significant concern.

IV. Results of tests of clustering within individual samples and tests of convergence between replicate samples for the same phenotypes

A. Polysubstance dependence vs controls (*Samples 1 and 2*)

NIDA substance dependence samples were analyzed [12] by selecting “nominally positive SNPs” that displayed p values < 0.05 for comparisons between substance dependent vs control samples in both European-American and African-American samples. We assessed the extent to which nominally-positive SNPs that were identified in both samples by SNPs represented on at least two different array types cluster together in small chromosomal regions. Clusters contain at least three SNPs that display $p < 0.05$ in both samples and lie within 100 kb of each other. In this dataset, 6,666 of the 639,401 tested SNPs displayed reproducible, nominally-significant abuser vs control allele frequency differences ($p < 0.05$) in both samples. The criterion that the same SNP display nominally-significant association in each of two samples is more stringent than criteria used in other comparisons (*see below*). This criterion was applied to reduce the number of false positive results, but does not allow as much within-locus heterogeneity. 1,158 of these 6,666 reproducibly-positive SNPs lie in 320 chromosomal clusters. 184 of these clusters identify 244 annotated genes.

Monte Carlo simulation trials that assess the probabilities that these results are due to chance find that none of 100,000 simulation trials identify as many SNPs that display nominally-positive differences between substance dependent vs control samples in both European and African American samples as observed here (thus $p < 0.00001$) [12]. 2,100 of 100,000 Monte Carlo simulation trials that each began by selecting 6,666 random SNPs provide chromosomal clustering as marked as that observed for the true reproducibly-positive SNPs ($p = 0.021$).

B. Methamphetamine dependent vs control (*Samples 4 and 5*)

Nominally-positive SNPs from each of these two samples cluster together with 25 kb separation between nominally positive SNPs more than anticipated by chance. 846 clusters contain 3,749 of the 15,569 nominally-positive SNPs from Sample 4. 1,787 clusters contain 8,388 of the 25,538 nominally-positive SNPs from Sample 5. Such clustering is found in no Monte Carlo simulation trial ($p < 0.0001$ for both Sample 4 and Sample 5).

When we evaluated the genes that were identified by clustered, nominally-positive results from both Samples 4 and 5, we obtained evidence for replication and results that could not be expected by chance alone. This criterion for genes to be identified by clustered, nominally positive SNPs from each of two samples is not as stringent as the criterion that

the same SNPs produce nominally positive results in each of two samples. It does allow for within-locus heterogeneity. The degree of convergent identification of genes by data from each of these two samples was never observed by chance in any of 100,000 Monte Carlo simulation trials ($p < 0.00001$).

C. Bipolar disorder vs control (Samples 7, 8 and 9)

WTCCC bipolar vs control 28,192 of the 426,604 SNPs analyzed in the WTCCC bipolar disorder collection displayed χ^2 values with $p < 0.05$ [1]. 12,560 of these SNPs fell into 1,775 clusters in which at least 4 SNPs that each displayed $p < 0.05$ (and were sampled on at least two array types) lay within 25 kb of each other. Monte Carlo simulation trials that assessed the probabilities that these results were due to chance found that none of 100,000 simulation trials identified as many clusters of SNPs that displayed nominally-positive differences between bipolar vs controls (control and other disease) samples as were actually identified (thus $p < 0.00001$) [1].

NIMH bipolar vs control 32,835 of the 536,288 SNPs analyzed in the NIMH bipolar collection [2] displayed t values with $p < 0.05$. 9,971 of these SNPs fell into 1,770 clusters in which at least 4 SNPs that each displayed t values corresponding to $p < 0.05$ lay within 25 kb of each other. Monte Carlo simulation trials that assessed the probabilities that these results were due to chance found that none of 100,000 simulation trials identified as many clustered SNPs that displayed nominally-positive differences between bipolar vs control samples as were actually identified from this work (thus $p < 0.00001$).

German bipolar vs control 27,057 of the 532,835 SNPs analyzed in the German samples displayed t values that corresponded to $p < 0.05$. 6,110 of these SNPs fell into 1,137 clusters in which at least 4 SNPs that each displayed t values corresponding to $p < 0.05$ lay within 25 kb of each other [2]. Monte Carlo simulation trials that assessed the probabilities that these results were due to chance found that none of 100,000 simulation trials identified as many clustered SNPs that displayed nominally-positive differences between bipolar vs control samples as observed in this work (thus $p < 0.00001$).

Convergent data from at least two of the three bipolar vs control comparisons: Simulation trials assessed the likelihood that the clusters of nominally-positive SNPs from at least two of these three bipolar samples identified the same genes. Monte Carlo simulation trials that assessed the probabilities that these results were due to chance found that none of 100,000 simulation trials identified as many clustered SNPs that displayed nominally-positive differences between bipolar vs control samples in the same genes in multiple samples as we actually observe (thus $p < 0.00001$).

D. Frontal brain volume (Samples 10 and 11)

10,266 SNPs provide nominally-positive results in 500k Affymetrix data from unrelated members of NHLBI twin pairs (Uhl *et al*, *submitted*). 583 of these 10,266 nominally-positive SNPs fall into 169 chromosomal clusters that each contain at least three nominally-positive SNPs that lie within 25 kb of each other and come from both the *Sty* I and *Nsp* I array types. Monte Carlo trials did not identify such a degree of clustering in any of 100,000 simulation trials (thus $p < 0.00001$).

5786 SNPs provide nominally-positive results in 100k Affymetrix data from Framingham samples (REF). 2729 of these nominally-positive SNPs fall into 613 chromosomal clusters with at least 3 SNPs that lie within 100 kb of each other. Monte Carlo trials do not identify this degree of clustering in any of 100,000 simulation trials (thus $p < 0.00001$).

Only 212 of 100,000 Monte Carlo trials provided random fits between NHLBI and Framingham datasets that were as strong as those identified from the true datasets. Thus, Monte Carlo $p = 0.00212$ for the convergence between Sample 10 (500k) and Sample 11(100k) datasets.

E. Success in smoking cessation (Samples 12, 13, 14)

In comparing data from successful vs unsuccessful quitters from samples 12, 13 and 14, we identified 5,411, 4,539 and 4,894 SNPs whose allele frequencies differ between these successful and unsuccessful quitters with nominal $p < 0.01$ [36]. These nominally-positive SNPs cluster together to extents much greater than expected by chance if their allele frequencies were independent of each other (Monte Carlo $p < 0.00001$). For Sample 12, 1434 of the 5411 nominally-positive SNPs lay in 308 clusters in which each positive SNP lay within 100 kb of at least two other positive SNPs with representation from SNPs on both array types. For Samples 13 and 14, 2258 of the 4539 nominally-positive SNPs and 2184 of the 4894 nominally-positive SNPs lay in 820 and 861 clusters along with at least one other nominally-positive SNP. We thus observed clustering in each of these independent samples. We would anticipate such clustering if many of these nominally-positive SNPs identified haplotypes that were present in different frequencies in our samples of successful vs unsuccessful quitters, but not if they represented chance independent observations.

Convergent data from at least two of the three quit success comparisons: Nominally-positive clustered SNPs from successful vs unsuccessful quitter comparisons from Samples 12 – 14 also cluster together on small chromosomal regions to extents much greater than chance [36]. The Monte Carlo p values for the replication for samples 12 vs 13, 12 vs 14 and 13 vs 14 are 0.00054, 0.0016 and 0.00063, respectively.

F. Cognitive ability (Samples 15 and 16)

Of the 489,922 autosomal SNPs assessed in Sample 15, 46,361 displayed correlations with cognitive ability assessments with nominal $p < 0.05$ (Uhl *et al*, *submitted*). 19,781 of these nominally-significant SNPs fell into 4,016 clusters in which at least three nominally-positive SNPs sampled on two array types lay within 25 kb of the neighboring nominally-positive SNP. Similarly, of the 489,922 autosomal SNPs that were assessed in Sample 16, 36,689 displayed correlations with cognitive ability with nominal $p < 0.05$. 11,376 of these nominally-significant SNPs fell into 2,556 clusters. None of 100,000 Monte Carlo simulation trials that sampled at random from the same databases found this degree of clustering of nominally-positive SNPs by chance for either sample ($p < 0.00001$).

348 genes contained at least three clustered nominally-positive SNPs in data from each of the two cognitive ability samples (Uhl *et al*, *submitted*). Monte Carlo simulations never identified convergence of this magnitude by chance ($p < 0.00001$).

G. Alzheimer's disease (Samples 17 and 18)

There was significant clustering of the SNPs that displayed nominally-positive data in the Alzheimer's disease genome wide association studies [4, 5]. For Alzheimer's disease sample 17, 5,428 of 16,224 nominally-positive SNPs identified 1,125 clusters (*reanalysis of* [4]). For the Alzheimer's disease sample 18, 7,469 of the 21,600 nominally-positive SNPs fall into 1,436 chromosomal clusters (*reanalysis of* [5]). We never observed the extents of clustering identified here in any of 100,000 Monte Carlo simulation trials for either sample ($p < 0.00001$).

Monte Carlo simulations also document the significant convergence between the results from these two independent studies of Alzheimer's disease; clustered nominally positive

SNPs from both samples identify 99 genes (Uhl et al, *submitted*). The observed overall convergence between these two datasets was never observed by chance in any of 10,000 simulation trials (thus Monte Carlo $p < 0.0001$).

V. “Generalization” of selected genome wide association results between related disorders and phenotypes

A. Convergence between human genome wide association studies of dependence on a variety of addictive substances (Samples 1 – 6)

There is a remarkable degree of overall convergence between the clustered, reproducibly positive SNPs in the NIDA samples and those identified by comparisons of alcohol dependent and methamphetamine dependent individuals [32] (Fig 6; Tables I and II).

The “methamphetamine dependence” genes described above display convergence with genes identified by: a) clustered- positive results from genome wide association studies of polysubstance abuse in NIDA-MNB European- and African- American samples [12] b) nominally-positive SNPs from genome wide association studies of alcohol dependence [3, 33] and c) nominally-positive SNPs in comparisons of nicotine-dependent vs nondependent smokers [3]. Data from Samples 5 and 6 converge with these other addiction vulnerability datasets with Monte Carlo p values of a) 0.0412, b) 0.0016 and c) 0.0003, respectively.

B. Convergence between human genome wide association studies of dependence on addictive substances and success in smoking cessation (Samples 5 and 6 vs 12–14)

A number of the reproducibly-positive genes identified by studies of methamphetamine dependence are also identified by clustered- positive results studies of Samples 12, 13 and 14 of European-American smokers who are successful vs unsuccessful in abstaining from smoking during clinical trials for smoking cessation. p values were 0.002 [34] for convergence with sample 13 and < 0.00001 for convergence with samples 12–14 taken together [36, 37] (Figure 5).

C. Bipolar disorder (Samples 7, 8 and 9) vs substance dependence (Samples 1, 2) genes

There was remarkable convergence between the genes identified by clustered, nominally-positive SNPs for addiction and those identified in similar fashion for bipolar disorder (Uhl *et al, submitted; reanalysis of* [1, 2]). Few of 100,000 Monte Carlo simulation trials identified chance overlaps as significant as those noted in actual data for comparisons between the genes identified by clustered positive SNPs for the NIDA substance dependence vs control comparisons and any of the three bipolar disorder vs control comparison groups. Monte Carlo p values were < 0.00001 for the overlaps between data from samples 1 and 2 vs data from samples 7 and 9. Simulation p value was $p = 0.00012$ for comparison between data from Samples 1 and 2 vs Sample 8.

D. Neuroticism (Sample 19) vs substance dependence (Samples 1, 2) genes

There was significant convergence between the genes identified by clustered, nominally-positive SNPs for addiction and those identified in generally similar fashion for association with neuroticism (Uhl *et al, in preparation, reanalysis of* [121]). Of the 31,110 SNPs with nominal $p < 0.05$ for comparisons of individuals with high vs low neuroticism scores, 4,623 SNPs lie in 903 clusters that each contain 4 such SNPs.

When we compared these clusters to those identified in Samples 1 and 2, only 30 of 10,000 Monte Carlo simulation trials identified chance overlap that was at least as robust as the true datasets ($p = 0.003$).

E. Frontal brain volume (Samples 10, 11) vs cognitive function (Samples 15, 16) genes

The frontal brain volume genes identified by both NHLBI and Framingham datasets overlap with those identified in genome wide association studies for cognitive function, whose genetics is linked to frontal lobe function (Uhl *et al*, *submitted*).

Nineteen genes identified by clustered nominally positive SNPs for frontal lobe volume in the NHLBI dataset are also identified by clustered nominally-positive SNPs from each of the two genome wide association assessments for cognitive function. Such overlap was identified by chance in only 212 of 100,000 Monte Carlo simulation trials that correct for repeated comparisons ($p < 0.00212$). One of these genes, CLSTN2, was also identified in a genome wide association study of individual differences in memory function [154].

F. Alzheimer's disease (Samples 17, 18) vs cognitive function (Samples 15, 16) genes

Each of these overlapping Alzheimer's disease datasets overlaps significantly with each of datasets for cognitive function. None of 100,000 Monte Carlo simulation trials identifies as many overlapping genes as those identified by these true datasets (Monte Carlo $p < 0.00001$) (Uhl *et al*, *submitted*). These comparisons thus strongly support generalization of the observations that we have made in studying general cognitive ability to vulnerability to AD. Pleiotropic effects of variants in a number of genes are thus likely to influence both cognitive ability and AD.

G. Alzheimer's disease (Samples 17, 18) vs frontal brain volume (Sample 10) genes

These overlapping AD datasets overlap significantly with the dataset for frontal brain volume (Uhl *et al*, *submitted*). None of 100,000 Monte Carlo simulation trials identifies as many overlapping genes as those identified by these true datasets (Monte Carlo $p < 0.00001$). Pleiotropic effects of variants in a number of genes are thus likely to influence both AD and frontal brain volume.

H. Alzheimer's disease (Samples 17, 18) vs substance dependence (Samples 1, 2) genes

Data from the overlapping Alzheimer's disease datasets also overlaps significantly data from the overlapping NIDA addiction vulnerability samples (Samples 1 and 2) [12]. None of 100,000 Monte Carlo simulation trials identifies as many overlapping genes as those identified by these true datasets (Monte Carlo $p < 0.00001$). These comparisons thus strongly support the "transitive" genetic overlaps between addiction, cognitive function/ brain volume and AD based on pleiotropic effects of variants in a number of genes.

I. Failure of control experiments to support alternative hypotheses for the observed genome wide association results

There is also no evidence that many of the clustered, reproducibly positive SNPs identified in these sample sets result from alternative effects of several different classes.

There were no significant overlaps between the SNPs identified in these studies and the SNPs that provided the largest racial/ethnic differences. Comparison groups included: 1) SNPs that provide the largest racial/ethnic differences in comparisons between European- and African-American control individuals from samples 1 and 2) 2) (for Asian samples) SNPs that provide the largest racial/ethnic differences between Japanese and Chinese HapMap samples 3) (for samples of European heritage) SNPs that display the largest allele frequency differences between United Kingdom samples of self-reported European heritage [155].

There were no significant overlaps between the SNPs identified in these studies and the SNPs that display the largest variances from array to array. Thus, there is no significant support for either of the alternative hypotheses: 1) that the positive results of any of these studies coming from occult racial/ethnic stratification or 2) that the positive results of these studies come from assay noise alone.

Finally, comparisons with data from a complex phenotype outside the brain provide a useful contrast with the observed convergence identified in the datasets above. Data for type I diabetes from the Wellcome trust case control consortium provide such a contrast [1]. Using Monte Carlo simulation approaches identical to those used to identify the highly significant overlaps noted above, we identify no significant overlap between genes that contain clustered, reproducibly nominally-positive SNPs for substance dependence (samples 1 and 2) and genes that are identified in similar fashion for type I diabetes (C J and GRU, *unpublished observations*, 2007).

VI. Limitations and cautions

It is important to consider several limitations for this convergent, replicated genome wide data. The sizes available for many of the samples from which data is reviewed here provide moderate power, at best, to detect gene variants. False negative results are likely since we require positive data from each of the several samples. The likelihood of false negatives is also increased since we require positive results from several SNPs from at least two array types that cluster within small chromosomal regions, making it easier to miss modest association signals within small genes that contain few SNPs or genes whose SNPs lie on only one array type. We have recently reassessed the relatively large numbers (6,734) of currently-annotated genes for which there is no possibility of detection, using 500k array data and the criteria that a gene must be identified by “at least three clustered (25kb) SNPs from at least two array types lying within the gene’s exons, introns or 10kb flanking sequences”. Genes that do not have the requisite numbers of SNPs with these characteristics are listed in the Supplement, Table SI. This list includes genes for which there is substantial positive association data from candidate gene studies, including the DRD4 dopamine receptor [156–158]. A number of these genes could be detected with the addition of 100 + 500k data and even more with more recently available arrays that assess *ca.* 1 million SNPs; these genes are also indicated in this table.

Differences in allele frequencies in different populations could explain why some genes are strongly associated with substance dependence in European but not in African or Asian samples. Apparent failure to “replicate” in the present analyses could thus derive from the differential informativeness of SNP markers in samples from different racial/ethnic backgrounds.

We focus only on data from autosomal regions. This focus allows us to combine data from male and female samples. However, such analyses will miss potentially important contributions from genes on sex chromosomes.

Many of the subjects of these studies volunteered for demanding clinical protocols, potentially rendering them not totally representative of all individuals in the population with the same phenotypes. Some of the subjects for this work have also come to clinical attention due to specific clinical features, including presentation to emergency rooms with methamphetamine psychosis or presentation to memory clinics with memory complaints. These selection criteria might not allow these individuals to completely represent all individuals who display the phenotype or disorder.

Ethical considerations arise in studying substance abuse, an illegal behavior often associated with other difficult individual and family pathologies (*see also above*). Such ethical considerations suggest that it is unlikely that any large, truly population-based sample for dependence on illegal substances will be readily obtained. Sampling research volunteers and examining the extent to which the volunteers' demographic characteristics match or deviate from those of the larger drug abuser communities provides one approach to this issue. We cannot quantitate the selection biases that might be imposed by factors not reflected in the demographic features (*eg* gender and self-reported racial/ethnic background) that are often used to characterize subgroups.

For studies of genetics of brain volume, it is worth noting that at least some of the genetic influences on frontal cortical volume are likely to provide "pleiotropic" influences on other brain regions. Some gene variants that influence frontal lobes as well as addiction, Alzheimer's disease and cognitive functions could even provide effects on these behavioral phenotypes through actions outside of the frontal cortex.

For the current analyses, we employ principal analytic approaches that represent only one of many current approaches to analyzing GWA data (see discussion above, also at www.nhlbi.nih.gov/resources/listserv/number35.htm (RFA-HL-07-010); [6, 85]. Despite this limitation, the replicated positive results obtained here and the failure of control experiments to support alternative hypotheses do provide significant confidence in roles for most of the genes reported.

The "generalization" to results from replicated samples to data from other samples that are likely, *a priori*, to display genetic overlap also adds to confidence in the results that are obtained. While it is not possible to supply precise estimates for this enhanced confidence, the generalization reported here does add to the overall confidence in the sets of genes listed. Identifying the same gene in multiple studies also enhances confidence that the gene contains relatively common allelic variants that alter its function sufficiently to produce association signals in several GWA datasets.

However, as noted above, it is important to note that, while the analytic approach that we use here seeks to identify genes with variants that are likely to play roles in addiction and in related phenotypes, there are limitations to interpretation of this data that come from our belief that flexible approaches to identification of such genes are likely to benefit analyses such as those that are presented here. If we allow for locus heterogeneity, for sample-to-sample differences in the details (including the phase) of linkage disequilibrium between marker alleles and pathogenic alleles, and for the use of different marker sets with different datasets, then we cannot demand that all samples display the more stringent criterion: significant association of the same SNP with the same phase in all samples. This approach thus seeks to use stepwise analyses to identify and confirm statistically-significant association at the level of individual gene loci, assuming that subsequent followup analyses will use further steps to untangle any locus heterogeneity, to unequivocally identify which individual SNPs are associated, to identify pathological haplotypes and the phases with which they are associated with phenotypes in samples from different racial and ethnic backgrounds. Indeed, our belief that, in many circumstances, molecular biologic, behavioral and other evidence will be required to have confidence in pathological haplotypes allows the analyses that we present here to represent open door invitations for just these sorts of ancillary studies to guide further molecular genetic fine mapping efforts.

For diagnoses of individuals selected due to dependence on a specific substance, *eg* "methamphetamine dependence", the specificity of effects of a gene's variants is likely to be limited by the fact that many of the subjects for these studies also report use of additional

addictive substances (*eg* inhalants for the methamphetamine dependence Samples 4 and 5). These clinical considerations, as well as the overlap between the “methamphetamine dependence” genes and the genes identified in other genome wide association work, support the idea that many, but not all, of the methamphetamine dependence (or nicotine dependence) loci are likely to contain allelic variants that provide a more general vulnerability to addictive substances. When we term some genes “methamphetamine dependence” genes, for example, to denote the fact that variants in these genes are likely to alter vulnerability to developing dependence on this substance, it is likely that many of these allelic variants also predispose individuals to dependence on other addictive substances.

Relatively few of the controls for the studies of substance dependence report any significant use of illegal addictive substances. The genes identified herein thus could influence vulnerabilities to a number of features important for eventually developing dependence including: a) initiation of use, b) persistence of this use, c) failure of attempts at quitting, d) transition from persistent use to dependence, or other steps, in ways that are not completely unraveled by simple comparisons of dependent *vs* control samples.

The largest limitation to the analyses presented here, however, is likely to come from the modest size of the effects reported for each of these phenotypes (with the exception of the APOE influences in Alzheimer’s disease, of course). In each polygenic disorder, replication of effects of polygenic loci is unlikely to be sufficiently robust that each study will provide nominally-significant observations. Final confidence in true effects requires both replication and generalization, including studies of more and more samples with the phenotypes that we study in the present review. However, we also need to display caution in considering the significance of apparent nonreplications. Recent analyses underscore the role that heterogeneity can play in engendering false-negative results from attempts at replication [126].

VII. Classes of genes that are identified in multiple GWA samples for multiple phenotypes: Focus on cell adhesion related genes

One approach to describing the convergence between the datasets, presented above, relies on the overall convergence between the results obtained in each study. A different approach focuses on convergence of data concerning specific genes and classes of genes, especially when most are expressed in the brain. Many of the genes that we identify in this analysis of convergent genome wide association findings are involved in “cell adhesion” processes whereby neurons recognize and respond to features of their environments that are important for establishing and maintaining proper connections (Table I). Others are involved in enzymatic activities, protein translation, trafficking and degradation; transcriptional regulation, receptor, ion channel and transport processes, disease processes and cell structures. A subset of this latter group is of especial interest since they represent classically “druggable” targets for potential small molecule therapeutics (Table II).

A. Cell adhesion related genes: The genes whose products are involved in cell adhesion processes provide a number of especially interesting results (Table I). Cell adhesion mechanisms are central for properly establishing and regulating neuronal connections during development. Cell adhesion mechanisms can play major roles in mnemonic and other neuroadaptive processes in adults [159, 160]. It is interesting to note that most of the cell adhesion related genes that we identify in these genome wide association studies are expressed in developing and adult brains. Altered expression of several of these genes can alter neurite extension [161–163], activate signaling pathways [164–169] and alter mnemonic processes [163]. Almost all of these cell adhesion-related genes are expressed in memory-associated brain regions that include hippocampus and cerebral cortex (<http://brain->

map.org) [170–173]. By contrast, substantial expression in mesolimbic/ mesocortical dopamine “reward system” neurons is not documented for many of them.

“Cell adhesion” related genes identified by these genome wide association studies encode members of several structural cell adhesion molecule subfamilies. Those that are anchored to cell membranes by glycosphosphoinositol (GPI) anchors, those that display apparent single-transmembrane topologies, those that display apparent seven transmembrane topologies and those that produce soluble products are each represented.

Here, we discuss several of the cell adhesion related genes that have been identified in multiple sets of replicated genome wide association studies (Table I). Monte Carlo trials that seek to establish the level of significance of identification of these genes in multiple datasets establish levels for this convergent identification that range from 0.006 to < 0.00001 . We also discuss follow up data for two cell adhesion molecule genes, NrCAM and NRXN3, which were initially identified by lower density genome wide association studies [10, 11]. Results from these two loci provide examples of the ways in which initial identification in genome wide association studies can lead to subsequent identification of specific haplotypes that alter levels of expression of the gene or levels of expression of a specific set of splice variants.

CAMs with the strongest levels of cumulative support: One of the cell adhesion molecules that achieves the most striking nominal p values in these analyses is an “atypical” member of the cadherin gene family, CDH13. Cadherin 13 is a glycosphosphoinositol (GPI)-anchored cell adhesion molecule. CDH13 is expressed in neurons in brain regions that are likely to play roles in addiction, including hippocampus, frontal cortex, and ventral midbrain [171]. CDH13 can inhibit neurite extension from select neuron populations [161, 171] and activate a number of signaling pathways [164–167]. It is thus a strong candidate for roles in brain mechanisms important for both developing and quitting addictions.

Other cell adhesion related genes that manifest cumulative p values $< 10^{-6}$ in these analyses include BAI3, CLSTN2, CNTNAP2, CSMD1, CTNNA2, DAB1, DSCAM, NRXN1, PTPRD and SGCZ. Data from NRXN1 associations in smoking have been recently reviewed. We discuss several of the other genes here.

DSCAM is a single-transmembrane domain cell adhesion molecule with immunoglobulin and fibronectin domains that is expressed strongly in brain [169, 174] and in hippocampus in ways that are required for appropriate neuronal connections to form in memory-associated circuits in model organisms [162, 163]. Different dendritic processes of the same neuron do not often cross each other; this self-avoidance mechanism depends on expression of a large array of tightly-regulated DSCAM isoforms [175, 176]. Simplifying this repertoire substantially disrupts appropriate formation of neuronal networks *in vivo* [177]. Indeed, flies with altered DSCAM expression display altered memories for both rewarded and punished behaviors [163].

CLSTN2 contains allelic variants that are identified in genome wide association studies of individual differences in memory and executive function as well as the cognitive ability/ Alzheimer’s disease vulnerability and frontal brain volume phenotypes reviewed here [178, 179]. CLSTN2 is expressed in frontal cortex and hippocampus [172]. CLSTN2 is well-positioned to provide calcium-dependent cell adhesion functions in the brain regions that include hippocampus and in the postsynaptic densities where it is highly expressed. The structure and expression of CLSTN2 make it a good candidate to function as a single transmembrane domain cell adhesion molecule in which variants could alter the ways in

which neuronal and synaptic connections develop, the ways in which they are maintained and reorganized in adult brains or both.

DAB1 interacts with and participates in signaling from several cell adhesion molecules. DAB1 has long been identified with signaling through the cell adhesion molecule reelin in ways that alter formation and maintenance of neuronal processes [180]. More recent evidence also supports roles for DAB1 in signaling through other cell adhesion/cell regulatory mechanisms, including those that utilize the amyloid precursor protein cell adhesion molecule [181]. DAB1 expression in many brain neurons includes those in hippocampus and mid to deep cerebral cortical layers [172] (<http://brain-map.org>). Mice with DAB1 disruption display substantial alterations in cerebral cortical development accompanied by gross motor and other behavioral phenotypes [182].

BAI3, a seven transmembrane domain cell adhesion molecule, as well as PTPRM, a single transmembrane receptor tyrosine kinase that mediates homophilic cell recognition and is supported at a more modest level of statistical confidence, are both expressed in vasculature [183, 184]. Identifying these genes fits with the idea that control and regulation of angiogenesis and vascular functions plays important roles in determining the richness of cerebral cortex and other areas of adult brains [184] in ways that have consequences for a variety of interesting brain based phenotypes. In addition, there is substantial neuronal expression of PTPRM in cortical and cerebellar cortical neurons [172, 183].

PTPRD is expressed in brain, and displays prominent hippocampal expression. Its extracellular ligands have not been elucidated, though it can bind to liprin [185]. PTPRD knockout mice display altered hippocampal long-term potentiation and spatial learning [186], which fit well with the human phenotypes related to cognitive function. Mice with deletions of both PTPRD and a related PRP sigma (but not with either knockout alone) die at birth due to failure to innervate appropriately [187]. SCGZ participates in protein complexes with cell adhesion-like [188]. High levels of SCGZ expression in the brain are confirmed by Allen brain atlas images [172]. Biochemical studies identify expression in schwann cells of peripheral nerves [188]. SCGZ can be found in complexes with $\alpha\delta$ or with $\epsilon\beta\delta$ sarcoglycans, demonstrating specificity of the context of its function in brain [189].

CSMD1 is substantially expressed in adult brain regions that include hippocampus [170]. High levels of CSMD1 expression in growth cones of neurons cultured from developing brain support substantial roles in development as well [170]. Less striking levels of evidence implicate variants in CSMD family members CSMD2 and CSMD3 in several of these brain related phenotypes [190].

CAMs with intermediate levels of cumulative support: Cell adhesion related genes with nominal cumulative significance levels between 0.00007 and 0.006 include ANKS1B, ASTN2, CNTN4, CNTN5, CNTN6, CTNNA3, CTNND2, LRP1B, LRRN6C NRG1, ITGB8, PTPRM, ROR1, TRIO, CSMD2, CNTN5 and SEMA3C. We discuss several of these molecules here.

ASTN family members play substantial roles in migration and organization of neuronal processes along glial fibers in ways that are key to proper development of laminated brain structures [191]. ASTNs is expressed at high levels in the adult hippocampus [172]. ASTNs knockouts slow the development of these structures [192]. ASTNs is upregulated after hippocampal lesions that induce synaptogenesis [193].

CNTN6 is a GPI-anchored cell adhesion molecule that can act as a Notch ligand, triggering nuclear translocation of the Notch intracellular domain. This interaction can promote formation of oligodendroglia from progenitor cells and increase expression of myelin-

associated glycoprotein in oligodendrocytes [194]. Neuronal expression in hippocampal, cerebral cortical, cerebellar and thalamic subdivisions in adult brain could contribute to the motor incoordination phenotypes noted in knockout mice [195, 196].

SEMA3C is one of the cell adhesion molecules that has been implicated in development of dopaminergic projections [197]. SEMA3C and its neuropilin 2 receptor are regulated by neuronal injury [198]. Significant levels of SEMA3C expression in hippocampus and connected regions, such as the medial septal nucleus, provide a good rationale for involvement in memory-related phenotypes [199].

LRRN6C displays interesting patterns of brain expression, with dense expression in cells of the dentate gyrus of the hippocampus and entorhinal cortical regions [172]. However, little additional data concerning this gene has appeared in the literature to date.

CTNNA2 is another protein whose strong association with cell adhesion mechanisms, especially cadherin mechanisms, renders it of substantial interest here. Mice with impaired CTNNA2 expression display impairments in prepulse inhibition and fear conditioning, altered dendritic spine morphogenesis in hippocampal neurons, unstable synaptic junctions and defective anterior commissure formation [200, 201].

Potential roles for cell adhesion related genes: The cell adhesion genes identified here provide an attractive way to bridge the gap between 1) the remarkable observed overlap between the molecular genetics of the clinical and cognitive phenotypes reviewed here and 2) the brain differences, especially those that might manifest in the quantity and/or quality of neuronal connections, that might underlie these shared heritable influences.

These cell adhesion related genes also provide an attractive bridge between the genetics of phenotypes that are not accompanied by gross brain pathology and those that are. Some of the cell adhesion genes that we identify in studies of cognitive ability might conceivably be directly involved in Alzheimer's disease pathological processes [202]. However, the majority of the cognitive ability genes identified here fit with emerging views that relatively subtle differences in brain connections that contribute to individual differences in cognitive abilities also provide individual differences in "cognitive reserves" [203]. If "cognitive reserves" mitigate the cognitive impact from a given density of Alzheimer's disease senile plaques and neurofibrillary tangles, for example, individuals with greater cognitive reserves might die without dementia despite neuropathological brain burdens that would otherwise produce significant dementia in individuals with lesser cognitive reserves [204].

Cell adhesion molecule genes with substantial follow up information: 1) **Neurexin 3** (NRXN3). Neurexins are cell adhesion molecules that help to specify and stabilize synapses and provide receptors for neuroligins, neuroligins, dystroglycans, and α -latrotoxins [168, 205, 206]. Neurexins function in the nervous system as cell adhesion molecules at excitatory and inhibitory synapses [168, 205–214]. The mammalian neurexin genes, *NRXN1*–*NRXN3*, each display multiple promoters from which longer α -neurexins and shorter β -neurexins are transcribed. Differential promoter usage and/or differential splicing events provide many neurexin isoforms [213, 215–218]. Neurexin splicing variants also provide opportunities to produce both membrane bound and soluble isoforms [215–217].

NRXN3 expression in specific cerebral cortical regions and layers, especially layers 2–3 and 5–6 is documented in rodent brain atlas images [172]. While *NRXN3* is also expressed in other brain regions of interest for addiction, its expression in glutamatergic projections that arise from prefrontal cortex to innervate nucleus accumbens and other striatal regions and/or in GABAergic neurons that project from cortical area to cortical area provide sites at which altered expression could readily influence circuits that are key for addictive behaviors.

Alterations in the size or strength of synapses in these circuits could readily provide alteration in behavioral features, including vulnerability to addictions.

From 10,000 SNP genome wide association data, we initially reported that SNP rs760288, located in the 3' region of *NRXN3*, displayed allele frequencies that distinguished individuals dependent on illegal substances from control individuals in both Sample 1 and Sample 2 [11]. These observations have received support by results from Sample 6 [3] and a linkage study of opioid dependence [219].

We genotyped nine SNPs in the 3' *NRXN3* regions that lie near two splicing sites, termed SS#4 and SS#5, in 144 European-American alcohol dependent and 188 control individuals from Sample 3 [130]. Minor allele frequencies at four SNPs near SS#5 were each higher in alcohol dependent than in control samples, supporting initial observations in other samples. A "T" allele at the rs8019381 (close to the original rs760288 SNP) produced a 2.46 odds ratio for belonging to the alcohol dependent group. Common 3 SNP haplotypes in this area were also strongly associated with alcohol dependence (nominal $p = 0.000598$, corrected to $p = 0.00580$ based on permutations).

The rs8019381 SNP was located 23 bp 3' from the 3' end of exon 23, near a branch point that could change the splicing of *NRXN3* primary transcripts [220, 221]. In mRNAs isolated from postmortem cerebral cortex, individuals with one or two copies of the minor rs8019381 T allele expressed less of the major transmembrane isoform ex22a24b and the minor transmembrane isoform ex22a24a mRNA levels than did CC homozygotes ($p = 0.0008$ and 0.021 for genotypes by two-tailed Mann Whitney test, respectively). By contrast, expression of the major ex22a23a soluble isoform did not correlate with rs8019381 genotype ($p = 0.171$ by two-tailed Mann Whitney test). As a consequence, ratios between transmembrane and soluble isoforms were significantly increased in individuals with one or two T alleles in comparison to CC homozygotes ($p = 0.018$, two-tailed Mann Whitney test).

Standard models for development and alteration of excitatory and inhibitory synapses posit that neurexins provide pre-synaptic cell adhesion molecules that bind heterophilically and with calcium-dependence to postsynaptic cell adhesion molecules to form trans-synaptic complexes that help to stabilize excitatory or inhibitory synapses [205]. Repertoires of diffusible entities enrich this picture. Neurexophilins provide soluble neurexin ligands that are expressed in neurons in a number of brain regions that also express *NRXN3*. Diffusible *NRXN3* isoforms are expressed in brain regions that express several neuroligins and dystroglycan [222]. Conceivably, soluble neurexins and neurexophilins might diffuse to sites at which membrane-bound neuroligins and neurexins were localized. Enhancing ratios between diffusible and membrane bound *NRXN3* isoforms in the CT/TT genotypes of rs8019381 might thus provide more diffused signaling in ways that could alter pruning, spreading and/or strength of synaptic contacts [223, 224].

2) *NrCAM* *NrCAM* encodes a single-transmembrane-domain cell adhesion molecule with six immunoglobulin domains, four to five fibronectin III repeats, a transmembrane domain and a C-terminal cytoplasmic domain with tyrosine kinase phosphoacceptor sites [225]. Multiple *NrCAM* isoforms can be identified as the products of differential RNA splicing events which produce *NrCAM* translation products that contain short inserted peptide sequences. *NrCAM* mRNA expression is high in hippocampus, mid- to deep layers of the cerebral cortex, cerebellum purkinje cell layers and striatal interneurons. Cells of the substantia nigra and ventral tegmental area (VTA) express hybridization densities that are almost as dense as those identified in hippocampal pyramidal neurons [131]. Many, but not all of the *NrCAM* immunopositive neurons in the VTA also express dopamine transporter (DAT) immunoreactivity, a marker for dopaminergic neurons [131].

We initially identified NrCAM based on convergence between data from genome wide association and studies of drug regulated gene expression [226]. Our initial 1.5 k SNP genome wide association study identified a reproducibly-nominally-positive SNP that was mapped roughly to the region that contained NrCAM as well as other genes. Follow up studies provided confirmatory observations: 11 of 37 tested mid-chromosome seven simple sequence length polymorphisms (SSLPs) display allele frequency differences between abusers and controls in NIDA European- and/or African-American samples that reach nominal significance. Three displayed nominally-significant differences in both populations; fewer than one would have been expected by chance. These nominally-significant results encompassed markers located between 98–103 Mb of chromosome 7.

Subtracted differential display PCR (SDD) identified NrCAM as a morphine-regulated gene. Sequences that encoded 3' untranslated regions of NrCAM were among subcloned SDD cDNAs that corresponded to mRNAs whose expression was altered in striata of rats sacrificed 4 hr after treatment with 20 mg/kg morphine. These biochemical results led us to focus on NrCAM. We thus identified 3' and 5' blocks of restricted haplotype diversity in NrCAM in European- and African-American samples, and sought association between addiction and allelic frequencies of markers in both of these haplotype blocks. The 3' haplotype was associated with addiction vulnerability in European American individuals from Samples 1 and 3 (nominal $p = 0.0006$ and $p = 0.003$, respectively) and from African-American individuals from Sample 2 (nominal $p = 0.0006$). However, the phase of the association was opposite in the African-American samples in comparison to the phase in the two European-American samples. Frequencies of the addiction-associated 3' *NrCAM* haplotype were thus higher in African American abusers than in controls but lower in two samples of European-American substance abusers than in the corresponding control samples. We interpreted these phase differences as indicating that the 3' NrCAM haplotype block was close to, but not identical with, the pathogenic haplotype.

NrCAM 5' flanking region haplotypes, however, displayed association with the same phase in each of these three samples ($p = 0.0002$, $p = 0.06$ and $p = 0.02$, respectively). Further, an additional independent sample that compared haplotype frequencies in 288 alcohol-dependent Japanese subjects vs 472 matched controls also displayed highly-significant association ($p < 0.01$) with the same phase. These cumulative observations strongly supported the idea that *NrCAM* variants provide polygenic contributions to human interindividual differences in addiction vulnerability. They contrast with the failure to identify such reproducibly-positive findings at the adjacent genes including LAMB1, LAMB4 and iPLA2.

We sought possible functional effects of these 5' *NrCAM* haplotypes by assessing patterns of allele specific expression of NrCAM haplotypes in mRNAs extracted from human postmortem brains. mRNA corresponding to the addiction-associated 5' NrCAM haplotype was expressed at an average of 26% of the levels of expression of mRNAs encoded by the alternative haplotypes.

When we compared *NrCAM* mRNA levels in cerebral cortex, midbrain and hippocampal samples from individuals who were heterozygotes for this 5' NrCAM haplotype to samples from individuals who lacked it, expression was about 40% lower in the brains of individuals who displayed the disease-associated haplotype. None of the adjacent genes' expression revealed such evidence for haplotype-specific differential expression.

The core 5' NrCAM haplotype that provided the largest effects on differential expression provided associations that were significant in Sample 1 ($p=0.0003$), Sample 2 ($p=0.02$) and Sample 3 ($p=0.05$).

We then asked if mice with altered levels of NrCAM expression display differences in conditioned place preference, which provides a relatively robust test for alterations in the reward and reward-memories induced by abused substances. Heterozygous and/or homozygous NrCAM knockout mice display striking reductions in their preferences for the places where they received either morphine or methamphetamine, in comparison to wildtype control mice. By contrast, mice of all genotypes displayed similar sensitivities to the acute locomotor stimulant properties of all three drugs.

These data thus provide convergent results from searches for drug- regulated genes and from association-based genome scans for drug abuse vulnerability alleles. The data support the idea that NrCAM and NRXN3 haplotypes contribute to human individual differences in addiction vulnerability in ways that are likely to depend on differences in levels of NrCAM expression and/or regulation and NRXN3 splicing.

VIII. A possible underlying neurobiology: connectivity constellations

Many brain-based phenotypes and disorders are accompanied by gross brain anatomic changes or differences when compared to brains from individuals without such phenotypes. These anatomical differences provide foci for many pathophysiological inquiries. However, in many other disorders and phenotypes, brains do not reveal reproducible gross pathologies. In these circumstances, investigations more often focus on molecular and biochemical pathways. Pharmacological mechanisms can even provide foci for exploration, when drugs that ameliorate disease symptoms are discovered serendipitously. Roles for mechanisms involved in “no gross pathology” phenotypes can be inferred in a number of disorders in which gross brain pathologies have been identified. Cognitive abilities at baseline, accompanied by no gross pathology, can play substantial roles in individual differences in vulnerability to dementing illnesses with gross pathological brain changes, such as Alzheimer’s disease, for example [227].

Most of the common brain phenotypes and disorders that lack gross neuropathological underpinnings have been shown to represent “complex” disorders (or phenotypes) from a genetic perspective [228]. Twin studies support heritability of at least half of total vulnerability to many such disorders or phenotypes. Major psychiatric disorders such as bipolar disorder, major traits such as general cognitive abilities and major brain phenotypes such as the volume of frontal and temporal cerebral lobes each display at least 50% heritabilities in a number of well-performed twin studies (Table III). Linkage studies, which have been good at identifying individual genes whose variants exert substantial effects on phenotypes or diseases, have established a lack of genes of major effect for most of these phenotypes or disorders. Genome-wide association approaches, however, are now revealing more and more of the gene loci that contain variants that contribute to such disorders, and thus more and more of the classes into which genes that contribute to these disorders or traits fall.

It is remarkable that virtually all of the brain-based disorders and traits for which genome wide association data that comes from a variety of laboratories can now be analyzed, as noted above (Table I), appear to receive substantial contributions from variants in cell adhesion related genes.

Could these repeated observations come simply from unexpectedly large representation of such genes in the genome? We have recently performed bioinformatic searches that attempt to enumerate the number of “cell adhesion molecules”. We constructed and performed these searches in ways that were independent of the genome wide association data described here. These searches used a combination of natural-language and keyword-based literature and database searches with DNA sequence motif searches using sequence elements common to

cell adhesion molecule families (CYL, QRL and GRU, *in preparation*). Briefly, we (CYL) integrated Gene Ontology annotations, domain structure information and keyword queries of NCBI Entrez Gene annotations in ways that were independent of the hand-annotated lists of cell adhesion related genes identified by genome wide association. First, we extracted 196 genes that encoded 281 proteins using the gene ontology term “cell adhesion (GO: 0007155)” [229]. Second, we extracted detailed domain features for six sub-families of cell adhesion molecules, including cadherins, IgCAMs, integrins, neuroligins, catenins and neuroligins [230–236]. Using these features and Perl scripts, we identified 218 human genes that encoded 532 proteins based on standardized related InterPro domain architectures and proteins mapped onto these architectures [237]. Fourth, we identified lists of cell adhesion genes for rat and mouse as noted above, and searched for human orthologs using homologue [238]. Finally, we identified 1487 results from searches of Entrez Gene using “adhesion AND Homo sapiens [organism]” [238]. The second through fourth approaches added 136 additional genes, yielding a total of 496 human cell adhesion molecule genes (CYL, QRL and GRU, *in preparation*).

With this set of genes in hand, we were now able to estimate the fraction of the genome that they represent. These 496 genes represent 1.6% of the 31,227 human genes now annotated by RefSeq. These 496 genes are often large, occupying a total of 73,695,264 bp of the genome. These sequences thus represent 2.4% of the 3,080,436,051 bp currently-elucidated human genome sequence. These genes represent 5.8% of the 1,271,259,295 bp currently annotated as gene sequences (*eg* exon, intron and 10kb 3' and 5' flanking sequence).

We can seek overlap between these “bioinformatic search cell adhesion molecule genes” and the 99 genes that we have 1) identified by clustered, nominally-positive SNPs in one or more of the genome wide association studies described above and 2) independently (GRU) annotated as “cell adhesion related”. Sixty-four of the 99 genes that we identified as “cell adhesion related” in genome wide association datasets are included on this independently-derived list of “bioinformatic search cell adhesion molecules”. These 64 overlapping genes represent 12.9% of all of the “bioinformatic search cell adhesion molecule genes” that we can identify in the genome. These 64 genes total 35,619,485 bp, representing 1.2% of the genome and 2.8% of currently annotated gene sequences.

Despite this relatively modest representation in the genome, then, cell adhesion related genes are substantially overrepresented in several of the genome wide association datasets reviewed here when compared to their genomic representation. Cell adhesion related genes represent 21–26% of the genes identified by the replicated searches for addiction vulnerability genes in Samples 1 and 2 and Samples 4 and 5. They represent 27% of the genes identified for vulnerability to bipolar disorder, 17% of the genes implicated in individual differences in regional cerebral volumes, 20% of genes implicated in cognitive abilities and 20% of the genes implicated in success in smoking cessation.

Further, when we compare the data from the sixty-four genes identified in both bioinformatic and hand searches with the data for the remaining 35 genes, strong cases can be made for inclusion of most of the remaining genes as “cell adhesion related”. CSMD1 and CSMD2, for example, represent two of the three currently-identified cub and sushi multiple domain genes, though they lack literature support for cell adhesion functions that is sufficiently strong to include them on the list of “bioinformatic search cell adhesion molecule genes”. Reviews and individual papers cite receptor protein tyrosine phosphatases as “cell adhesion molecules”, in ways that correspond to our hand annotation but not to their identification in the independent bioinformatic searches as cell adhesion molecules [239–243].

B. Quality and/or quantity of connections involved in vulnerabilities

Quantitative morphometric studies have documented the large fraction of volume of cerebral cortical grey matter that is composed of “neuropil”, neuronal processes and their supporting/ensheathing glial elements [244, 245]. By comparison, vasculature and neuronal cell bodies contribute much less to the volume of grey matter. The ways in which neuropil and the neuronal connections that it contains 1) develop 2) change with experience (or disease) and 3) change with aging all provide interesting avenues for functional impact of individual differences in neuronal connections.

Heritability estimates are available for several of the human phenotypes that are likely to be influenced by the quality and/or quantity of such neuronal connections. Regional brain volumes have been among the most studied using classical twin methods [15]. Some of the best data that is currently available comes from studies of the volumes of the cerebrum, frontal lobe, temporal lobe and hippocampus. These twin studies support heritable interindividual differences in these regions that range from 0.4 to 0.6 [15]. Since the majority of the volumes of the grey matter aspects of these regions is comprised of connection-rich neuropil, it seems highly likely that differences in the quantity of connections provides one substantial source of these heritable interindividual differences.

Twin methods also provide data that links these heritable determinants of brain volume to heritable determinants of cognitive functions that include memory and estimates of different aspects of cognitive function [246, 247]. In one of the largest available twin datasets, virtually all of the heritable individual differences in brain regional volumes could be attributed to genetic influences shared with influences on cognitive abilities (measured primarily by memory tasks) [247]. In another twin dataset, the overwhelming majority of genetic influences on frontal lobe brain volume were shared with those that influence cognitive ability measured primarily by tests of executive function [117].

These data strongly imply roles for individual differences in the quantity of neuronal connections in heritable determinants of cognitive function. No human data of which we are aware directly address qualitative differences in neuronal connections. However, it seems unlikely that genes that influence quantity would fail to influence quality of connections as well. Interrelations between neurons and glia are also likely to provide significant contributions to the anatomic coherence of white matter pathways. It seems likely that the interindividual differences in diffusion tensor imaging signals that come from white matter differences and that can readily be observed in humans [248] will also receive contributions from “connectivity constellation” genes, to the extent that these individual differences are heritable [249, 250].

Rodent models provide support for quantitative and qualitative effects of variations in connectivity constellation genes. Hints of redundancies in the ways in which these genes might work come from studies of effects of gene $\times$ gene interactions in brains of knockout mice. Mice with knockout of the addiction-associated gene NrCAM and another immunoglobulin + fibronectin domain cell adhesion molecule gene, L1, display substantial reductions in brain volume and substantial developmental brain-based phenotypes (NrCAM L1KO). No observable differences between the brains of wildtype mice and knockouts of either NrCAM or L1 alone can be identified, however. While gross brain differences in knockouts of only a few of the “connectivity constellation” genes can be identified, it appears likely that more subtle, more qualitative differences might be induced by variants at many more single loci from this list. Larger quantitative differences that can be grossly observed may require contributions from variants at several of these gene loci.

C. Overlapping effects of “connectivity constellation” gene variants, comorbidities and co-occurrence frequencies. One of the predictions that comes from postulates that variants in a limited number of genes contribute to genetic components of susceptibility to a variety of phenotypes is the prediction that these phenotypes would occur together more often than expected by chance. The magnitude of co-occurrence, of course, would be expected to differ based on differential penetrance of each locus on each phenotype as well as other factors.

There is relatively strong evidence for co-occurrence of vulnerability to a number of the phenotypes listed in Table III. Bipolar disorder and addiction occur together much more frequently than we would expect by chance alone; almost three quarters of individuals with bipolar disorder also manifest substance use disorders in a recent review and elsewhere [146, 147, 251]. This co-occurrence of these two heritable phenotypes fits with the substantial overlap that we documented above in comparisons between genome wide association results from addiction and bipolar disorder. Many of the genes that receive substantial support from both of these studies encode cell adhesion molecules and other “connectivity constellation” genes.

As noted above, twin study data supports substantial shared overall genetic influences on cognitive abilities and brain volumes [246, 247]. Comparison of genome wide association datasets for these two traits documents substantial overlaps that are much greater than chance and that again identify “connectivity constellation” genes. One of the strongest risk factors for Alzheimer’s disease that can be identified in the literature is baseline cognitive abilities, as reflected in individual differences in educational attainment, for example [252, 253]. We have identified strong overlap between genome-wide association results for cognitive ability and those for Alzheimer’s disease that includes a number of “connectivity constellation” genes.

Neuroticism is a heritable personality trait [254] that differs most in comparisons of addicted vs control individuals [255, 256]. We have also identified a strong overlap between genome-wide association datasets for neuroticism and addiction [121]; Johnson *et al*, *in preparation*). Again, many “connectivity constellation” genes are identified by both of these approaches.

As denser and denser datasets for genome wide association for disorders such as antisocial personality disorder and schizophrenia become available, it will be interesting to seek such overlaps with other “connectivity constellation” disorders.

D. Contributions of “connectivity constellation” genes to morbidity/mortality from brain disorders in the United States One way to assess the impact of complex polygenic genetics on brain disorders is to estimate US costs associated with the disorder, estimate heritabilities from twin study data, estimate Mendelian and oligogenic contributions from family study, segregation, linkage and genome wide association analyses and thus identify the overall impact of polygenic contributions. We have estimated that the polygenic influences on addiction vulnerability are responsible for \$212.20 billion in cost to the US in 2004 [257–259]. Other “connectivity constellation” disorders and the relevant estimates for costs of polygenic contributions include Alzheimer’s disease and dementias, with perhaps 74 billion in such costs, pain and migraine with 58.8 billion in such costs, anxiety disorders with 24 billion in such costs, schizophrenia with 39 billion in such costs, depressive illnesses with almost 19 billion in such costs, developmental disorders with perhaps 5 billion in such costs and stroke, Parkinson’s disease, multiple sclerosis and seizures with 2, 1.4, almost 3 and 0.6 billion, respectively, that can be attributed to complex genetics. If variants in the “connectivity constellation” of genes contribute 20% to the complex genetic impact of these

disorders, then the costs attributable to these cell adhesion related gene variants could amount to almost 90 billion in the US annually.

IX. Classes of genes that are identified in multiple GWA samples for multiple phenotypes: focus on “druggable” targets

The emphasis that we have placed on cell adhesion molecule genes here should not obscure the identification of a large number of genes that encode proteins that are more classically “druggable”. Enzymes, receptors and ligands (G-protein coupled receptors, ligand-gated ion channels and peptide ligands), ion channels and transporters, in particular, provide targets for many currently-effective drugs and are thus thought of as more “druggable” than cell adhesion molecules, for which there is less precedent for efficacious small molecule therapeutics. It is therefore instructive to identify the potential targets that come from genes in these families that are supported from results in replicated genome wide association studies of at least three of the phenotypes assessed in this review. It is worth noting that many of these drug targets in fact can be targeted by promising lead compounds that have already been identified (*see below*).

“Druggable” genes that are most strongly supported by the current analyses: Cumulative nominal p values, based on the strongest clusters on nominally positive SNPs in each of these individual genes in studies of a variety of different phenotypes, provide support for individual genes that is as striking as $10^{-3} - < 10^{-5}$ for several of these “druggable” genes.

Enzyme-related genes are most numerous group, with PRKG1, CAMK1D, CHN2, FHIT and SERPINA1 providing the most significant convergent observations. PRKG1, the cyclic-G dependent protein kinase 1, and FHIT, the fragile histone triad gene that represents a diadenosinase, each display clustered, nominally-positive SNPs in several of these studies in ways that make them especially unlikely to represent chance findings. PRKG1 is expressed in brain and in hippocampal, cerebellar and other neurons [260]. Nitric oxide dramatically modulates brain cGMP systems; PRKG1 thus provides a major target for the products of nitric oxide synthases (NOS). Mnemonic and addictive functions can each be altered by changes in cGMP-dependent protein kinase and/or NOS [261]. Identification of potent imidazopyridine PRKG inhibitors for use in lower species could well provide leads for development of similar compounds for use in humans [262].

FHIT was initially named the fragile histone triad gene based on its location at the chromosome 3p Fra3B locus that provides one of the most common neoplasia-associated human genomic fragile sites. FHIT is highly expressed in brain, with robust expression in and localized in regions that include hippocampus. The high levels of brain expression of FHIT protein and its characterization as the major human enzyme that can hydrolyze diadenosine-polyphosphates provides a strong link to purinergic signaling [263]. Diadenosine polyphosphates are released from synaptic vesicles along with more classical transmitters. They can activate P2Y(1), P2Y(2), and P2Y(4) receptors, homomeric P2X(1), P2X(2), P2X(3), P2X(4), and P2X(6) receptors, and even a P4 receptor-operated Ca^{++} channel [261]. Recent reports have also implicated FHIT in a nonenzymatic function, altering expression of the cell adhesion related protein beta catenin [264]. Robust inhibition of FHIT's diadenosine triphosphatase activity by suramin provides a starting place for developing more selective inhibitors [265, 266]. However, recent identification of FHIT as a modulator of beta catenin-induced transcriptional regulation may also place this gene in the cell adhesion signaling cascade as well [263].

CAMK1D is expressed in neurons, including those of the hippocampus, at relatively high levels. While little direct information about the functional effects of its brain expression is

available, other calcium/calmodulin dependent protein kinases play large roles in memory like processes. Useful reagents to study CAMK1D activities in peripheral systems might well come from studies of mimics of CAMK1D sequences [267–269].

LARGE encodes a likely enzyme; this gene also receives much cumulative support from these genome wide association datasets. LARGE encodes a protein with high homology to acetylglucosaminyltransferases. Brain provides many of the LARGE expressed sequence tags, while *in situ* hybridization documents high levels of expression in apparent neurons in a number of brain regions, including hippocampus and cortex [172]. In muscle, interrelationships between dystroglycans, LARGE and its family members have been so well documented that interactions with other dystroglycans in brain also appear likely [270]. Humans with LARGE mutations can express developmental delay as well as muscular dystrophies; a spontaneously occurring mouse model is also available [271].

The phosphodiesterase PDE4D, that also receives substantial support, is an enzyme expressed in neurons and other cells in brain in a number of regions that include hippocampus. PDE4D knockouts display altered results in “antidepressant” testing [272]. Variants at the PDE4D locus have been associated with a number of additional human phenotypes. PDE4D SNPs have been associated with neuroticism in the initial analyses of data from Sample 19 [9]. A number of studies, but not all, have linked and/or associated PDE4D haplotypes with altered risk of stroke [273]. A PDE4D SNP has been associated with sleepiness in Framingham study subjects in 100,000 SNP genome wide association [274]. Rolipam and other PDE4D inhibitors or stimulators have been extensively characterized, providing starting points for more synthesis of more selective drugs [275]. Memory-enhancing effects of PDE4 inhibitors documented in rodent behavioral assays, and provide relatively straightforward links with the genome wide association data that we review here [12].

RYR3 is the channel gene that is most heavily supported by this data. RYR3 is a site at which calcium activated calcium efflux from sequestered intracellular stores elevates free cytoplasmic calcium levels. Brain RYR3 expression is highest in hippocampus [276, 277]. Significant evidence links RYRs to associative learning [278]. RYR3 knockout mice display alterations in several sorts of learning tasks [279]. Ryanodine and a number of other small molecule appear to provide excellent starting points for synthesis of selective RYR3 inhibitors [280].

Other genes on this list of classically-druggable targets display substantial cumulative nominal p values include genes that encode a peptide/growth factor ligand, G protein coupled receptors, and transporters.

FGF14 isoforms are widely expressed in a number of brain regions during development, and expressed in more focal fashion in adult brain [281]. FGF14 lacks signal sequences and may thus play intracellular roles that include interactions with ion channels and intracellular signal transducing proteins [282]. The FGF14 gene is a candidate to play roles in control of midbrain dopamine neurons, based on mouse strain comparison studies [283, 284]. FGF14 expression is upregulated by methylphenidate [285]. FGF14 knockout mice display substantial alterations in learning and aspects of long term potentiation [286]. Humans with FGF14 mutations also display cognitive changes [121]. While we are aware of no small molecule FGF14 agonists or antagonists, such development might well benefit from the substantial progress in producing small molecule peptidomimetics in other fields.

Recent progress in identifying ligands for G protein coupled glutamate receptors (*see Carroll et al, this volume*) provides substantial hope that mGluR7 ligands will soon be available to

test roles for this receptor in substance dependence and other brain functions. The high significance attained by ABCC4 also supports strong consideration of drugs targeting this interesting ABC cassette transporter.

The genes discussed here provide examples of potential therapeutic targets; many of the other genes identified by these studies may also provide equally tractable drug target sites. Indeed, some of the genes that are identified in genome wide data from only one or two phenotypes might conceivably provide targets for drugs with more favorable specificities and therapeutic indices than genes identified by GWA data for many phenotypes.

X. Directions, summary and conclusions

It is an exciting time to be able to summarize and review the rapidly-emerging data on the complex genetics of human addiction vulnerability and of related phenotypes. Genome wide association results for dependence on several different classes of addictive substances converge with each other in striking fashion that is highly unlikely to be due to chance. Studies of dependence phenotypes in samples of individuals from several different racial and ethnic backgrounds support the idea that many of the allelic variants that predispose to these common disorders are so evolutionarily old that they are present in members of each major current human population. These data, combined with the varying results from linkage-based studies, fit a genetic architecture for addiction that is based on polygenic contributions from common allelic variants. Such a genetic architecture is quite consistent with data from family, adoption and twin classical genetic studies.

The identification of genes with markers whose allelic frequencies distinguish addicts of several different ethnicities from matched controls supports “common disease/common allele” genetic architecture [287] for at least much of addiction vulnerability. The convergent data derived from studies of individuals with addictions to substances in several different pharmacological classes supports the idea that “higher order pharmacogenomic/pharmacogenetic” variations enhance vulnerability to many addictions. These results do not exclude additional contributions to addiction vulnerability from genomic variants that influence vulnerability to specific substances or variants that are found only in specific populations. Nevertheless, the findings presented here provide promise for enhancing understanding of features that are common to human addictions in ways that could facilitate efforts to personalize prevention and treatment strategies for debilitating addictive disorders.

Identification of addiction-associated variants in genes that are likely to alter the quality of brain connections provides a first step toward defining a new neurobiology for the underpinnings of specific diseases and phenotypes. For many of these diseases and phenotypes, only little current research focuses on direct study of brain connections. The “connectivity constellation” concepts that we introduce here support studies that develop and use current and novel means for assessing the qualities and quantities of brain connections, especially in contexts in which they assess their functional properties. We have identified contributions of connectivity constellation genes to volumes of the same brain regions in which many of these genes are expressed. This convergence may provide new insights into data that documents individual differences in frontal lobe volume and/or in function, detected by volumetric, deoxyglucose PET and/or fMRI imaging, for virtually all of the “connectivity constellation” phenotypes or disorders noted here [8, 117].

The addiction vulnerability genes identified in this work contribute to the growing body of data that implicates cell adhesion and related memory-like and other cognitive processes in addiction. Studies that alter reconsolidation and other memory-related processes using knockout mice, protein synthesis inhibitors and/or pharmacologic treatments demonstrate

powerful influences on addictions [288, 289]. This empirical evidence enriches theoretical work that increasingly recognizes memory-like features for addiction [36] and work that implicates memory-associated brain regions in relapse to addiction. Such work also complements clinical observations which document that addicts' enhanced vulnerabilities to substance abuse relapse can persist for decades after their last prior use of addictive substances.

There is also substantial evidence for generalization of these results from addiction. This evidence comes from the significant overlaps between the molecular genetics of addiction and the molecular genetics of a number of related phenotypes and disorders. Overlap with bipolar disorder provides one of several likely psychiatric diagnoses for which shared genetic influences are likely *a priori*, based on the substantial heritabilities of both addiction and the high frequency of addiction/bipolar disorder comorbidity [146, 147]. This same logic suggests that abundant shared genetics may well also underpin the frequent comorbidities between addictions and antisocial personality/conduct disorders [251]. Less compelling evidence points to overlaps with other depressive, anxiety and schizophrenic disorders as well [251].

We have sought evidence for genetic influences that are shared between addiction and 1) frontal lobe brain volumes, 2) cognitive function and 3) Alzheimer's disease. Hypotheses about such shared genetic influences are based, in part, on initial observations that so many of the genes that we and others have identified in addiction genome wide association relate to cell connections. These molecularly-based hypotheses were reinforced by the evidence for substantial, complex genetic components to each of these phenotypes. These hypotheses were strengthened by evidence, though often from small samples, that appears to document 1) small frontal lobe volumes in samples of addicts [19, 290], 2) lower performance levels on tests of cognitive and executive function in samples of addicts [291–293], and 3) large roles of heritability vs little role for the drug exposure itself in determining the cognitive abilities of twin pair members who are discordant for cannabis use [294]. These hypotheses are further reinforced by 1) twin data that document strong shared genetic influences on frontal brain volumes and cognitive function measures [246, 247], 2) smaller head (and thus likely brain) sizes in individuals who go on to develop Alzheimer's disease decades later [295], and 3) lower levels of educational attainment (and thus cognitive function) in individuals who go on to develop Alzheimer's disease many years later [252, 253]. Little convincing prior evidence linked addiction liability directly with Alzheimer's disease, although there are reasonably strong links between addiction vulnerability and the related, often-dementing neurodegenerative illness, Parkinson's disease [296] (further reviewed in [297]). Identification of the genes documented here should lead to elucidation of features of addiction vulnerability that might arise preferentially from the "connectivity constellation" gene variants that also alter vulnerability to Alzheimer's disease.

The focus on genes in the current analyses should not obscure the fact that many of the SNPs and loci identified in each of these genome wide association studies lie between annotated genes. As the biological roles for "intergenic" regions become better understood generally, it should be increasing.

Disease-associated markers both within and between genes can all begin to allow us to assess individual differences in vulnerability to addiction based on profiles of genotypes. In settings in which prevention of addiction is sought, addiction vulnerability genomic profiles could help to target more (or different) prevention resources to individuals at the most (or at different) genetic risk. When a therapeutic opiate is being considered for chronic, noncancer pain, for example, the costs of engendering substance dependence are likely to be sufficient to justify genotyping even if the results provide only partial information about risk

assessment and minimization for prescribing physicians. When treatment for an established dependence on nicotine, opiates or alcohol is being contemplated, a number of different therapeutic options with different pharmacological mechanisms of action are now available [298]. Subsets of the SNPs that we have associated with success in quitting smoking appear to provide selective influence success in responding to bupropion, while others appear to provide selective influences on success in response to nicotine replacement. Replication and extension of these observations to treatments for alcohol, opiates and other addictive substances will make it more and more likely that SNP markers will increasingly aid “personalization” of antiaddiction therapies within the near future, in ways that are now impacting the design of clinical trials in this area.

The shared genetic influences on addiction with these other brain based phenotypes and our implication of roles for variants in “cell adhesion” genes in many of these phenotypes underscore the likelihood that the biology of these systems will demand better and better approaches to understanding more of the details of the estimated ca. 10^{14} synaptic connections in the cerebral cortex, and those in other areas [299]. Gross phenotypes, such as lobar brain volumes, are likely to vastly underrepresent the true complexity of the qualitative differences that are driven by these gene variants, and provide only weak reporters for quantitative differences. The present genetic analyses underscore the urgency of developing better approaches to elucidating individual differences in connectivities in human brain, both living and postmortem.

As we elaborate and extend these molecular genetic approaches, it seems likely that the number of genes with variants that contribute to individual differences in human vulnerabilities to addiction and related phenotypes will continue to grow, and that many of the genes listed here will receive additional support. By the time that this review is printed, more detailed data from arrays that contain markers for almost 1 million SNP and 1 million copy number variant probes will enhance the picture that we present here (Liu *et al*, *in preparation*). These studies will add power to detect differences in many of the genes and intragenic regions that are not well studied using the 500 – 600k datasets analyzed in this review, and to allow us to understand potential roles of copy number variation in these addiction and related phenotypes. More and more such studies will provide more and more ability to both better understand, better prevent and better treat these common and debilitating conditions.

This work, taken together, supports the idea that the heritable brain bases for individual differences in addiction vulnerability lie squarely in the midst of the repertoire of common complex determinants of individual differences that are manifest in many heritable complex brain disorders and phenotypes. Such conclusions place the biology of addictions squarely in the midst of important biologies of a number of brain phenotypes and disorders, hopefully in ways that will benefit them all.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank subjects for each of these samples. We thank collaborators who include the H Ujike and the JGIDA methamphetamine investigators, SK Li and the Taiwan methamphetamine investigators, Jed Rose, Caryn Lerman, Ray Niaura, Sean David, Gary Swan, Christina Lessov-Schlaggar. We are grateful to TGen, Glaxo Smith Kline, Framingham study investigators, the NicSNP group and the Wellcome Trust Case-Control Consortium for access to genotype data analyzed here. Rachel Gibson, S Seshadri and J Pollock were of especial help with obtaining access to the GSK, Framingham and NicSNP datasets cited here. We are grateful for dedicated help with clinical characterization of NIDA subjects from Dan Lipstein, Fely Carillo, Carlo Contoreggi, Fred Snyder and other Johns

Hopkins-Bayview support staff. We benefited from passionate discussions of statistical issues with Dr Daniel Naiman and from the Baltimore Epidemiology Catechment Area follow up study that was generously provided by Dr. J. Anthony. We thank NHLBI staff and the twin study technicians and investigators, PA Wolf, BL Miller, T Reed, L Epstein, L Hawk, P Shields, F Patterson, A Pinto, M Rukstalis, W Berrettini, R Brown, E Richardson, F M. Behm, P Kukovich, E C Westman and G Samsa for their rigor in overseeing data collection at their research sites. We acknowledge financial support from NIH-IRP (NIDA), DHHS and are also grateful for support for some of the studies discussed in detail here from the Taiwan and Japanese Ministries for Science and Technology, NIH grants P50CA/DA84718, RO1CA 63562, HL32318, DA08511, P50CA84719, 1K08 DA14276-05, HL51429, support from the Wellcome Trust (076113), the Pennsylvania Department of Health (which specifically disclaims responsibility for any analyses, interpretations, or conclusions), GlaxoSmithKline, Inc and unrestricted support for studies of adult smoking cessation from Phillip Morris USA. Some human brain tissues were obtained from the Brain and Tissue Bank for Developmental Disorders supported through NIH contract NO1-HD-1-3138.

References

1. WellcomeTrustConsortium. 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]
2. Baum AE, et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry*. 2007
3. Bierut LJ, et al. Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet*. 2007; 16(1):24–35. [PubMed: 17158188]
4. Coon KD, et al. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J Clin Psychiatry*. 2007; 68(4):613–618. [PubMed: 17474819]
5. Li H, et al. Candidate Single-Nucleotide Polymorphisms From a Genomewide Association Study of Alzheimer Disease. *Arch Neurol*. 2007
6. Dudbridge F, Gusnanto A, Koeleman BP. Detecting multiple associations in genome-wide studies. *Hum Genomics*. 2006; 2(5):310–317. [PubMed: 16595075]
7. Frayling TM. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet*. 2007; 8(9):657–662. [PubMed: 17703236]
8. Seshadri S, et al. Genetic correlates of brain aging on MRI and cognitive test measures: a genome-wide association and linkage analysis in the Framingham Study. *BMC Med Genet*. 2007; 8(Suppl 1):S15. [PubMed: 17903297]
9. Gottlieb DJ, O'Connor GT, Wilk JB. Genome-wide association of sleep and circadian phenotypes. *BMC Med Genet*. 2007; 8(Suppl 1):S9. [PubMed: 17903308]
10. Uhl GR, et al. Polysubstance abuse-vulnerability genes: genome scans for association, using 1,004 subjects and 1,494 single-nucleotide polymorphisms. *Am J Hum Genet*. 2001; 69(6):1290–1300. [PubMed: 11704927]
11. Liu QR, et al. Pooled association genome scanning: validation and use to identify addiction vulnerability loci in two samples. *Proc Natl Acad Sci U S A*. 2005; 102(33):11864–11869. [PubMed: 16091475]
12. Liu QR, et al. Addiction molecular genetics: 639,401 SNP whole genome association identifies many "cell adhesion" genes. *Am J Med Genet B Neuropsychiatr Genet*. 2006; 141(8):918–925. [PubMed: 17099884]
13. Hung RJ, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*. 2008; 452(7187):633–637. [PubMed: 18385738]
14. Thorgeirsson TE, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*. 2008; 452(7187):638–642. [PubMed: 18385739]
15. Peper JS, et al. Genetic influences on human brain structure: a review of brain imaging studies in twins. *Hum Brain Mapp*. 2007; 28(6):464–473. [PubMed: 17415783]
16. Sim ME, et al. Cerebellar gray matter volume correlates with duration of cocaine use in cocaine-dependent subjects. *Neuropsychopharmacology*. 2007; 32(10):2229–2237. [PubMed: 17299505]
17. Brody AL, et al. Differences between smokers and nonsmokers in regional gray matter volumes and densities. *Biol Psychiatry*. 2004; 55(1):77–84. [PubMed: 14706428]
18. Thompson PM, et al. Structural abnormalities in the brains of human subjects who use methamphetamine. *J Neurosci*. 2004; 24(26):6028–6036. [PubMed: 15229250]

19. Franklin TR, et al. Decreased gray matter concentration in the insular, orbitofrontal, cingulate, and temporal cortices of cocaine patients. *Biol Psychiatry*. 2002; 51(2):134–142. [PubMed: 11822992]
20. Baicy K, London ED. Corticolimbic dysregulation and chronic methamphetamine abuse. *Addiction*. 2007; 102(Suppl 1):5–15. [PubMed: 17493049]
21. Tomasi D, et al. Thalamo-cortical dysfunction in cocaine abusers: implications in attention and perception. *Psychiatry Res*. 2007; 155(3):189–201. [PubMed: 17582746]
22. Guerrini I, et al. Genetic linkage analysis supports the presence of two susceptibility loci for alcoholism and heavy drinking on chromosome 1p22.1-11.2 and 1q21.3–24.2. *BMC Genet*. 2005; 6(1):11. [PubMed: 15740611]
23. Zhong X, Zhang H. Linkage analysis and association analysis in the presence of linkage using age at onset of COGA alcoholism data. *BMC Genet*. 2005; 6(Suppl 1):S31. [PubMed: 16451641]
24. Clarimon J, et al. Linkage disequilibrium and association analysis of alpha-synuclein and alcohol and drug dependence in two American Indian populations. *Alcohol Clin Exp Res*. 2007; 31(4): 546–554. [PubMed: 17374033]
25. Gelernter J, et al. Genomewide linkage scan for nicotine dependence: identification of a chromosome 5 risk locus. *Biol Psychiatry*. 2007; 61(1):119–126. [PubMed: 17081504]
26. Hopfer CJ, et al. A genome-wide scan for loci influencing adolescent cannabis dependence symptoms: evidence for linkage on chromosomes 3 and 9. *Drug Alcohol Depend*. 2007; 89(1):34–41. [PubMed: 17169504]
27. Kuo PH, et al. Identification of susceptibility loci for alcohol-related traits in the Irish Affected Sib Pair Study of Alcohol Dependence. *Alcohol Clin Exp Res*. 2006; 30(11):1807–1816. [PubMed: 17067344]
28. Li MD, et al. Linkage and association studies in African- and Caucasian-American populations demonstrate that SHC3 is a novel susceptibility locus for nicotine dependence. *Mol Psychiatry*. 2007; 12(5):462–473. [PubMed: 17179996]
29. Loh EW, et al. Association analysis of GABA receptor subunit genes on 5q33 with heroin dependence in a Chinese male population. *Am J Med Genet B Neuropsychiatr Genet*. 2007; 144(4):439–443. [PubMed: 17440936]
30. Pomerleau OF, et al. Genome-wide linkage analysis for smoking-related regions, with replication in two ethnically diverse populations. *Nicotine Tob Res*. 2007; 9(9):955–958. [PubMed: 17763112]
31. Saccone SF, et al. Genetic linkage to chromosome 22q12 for a heavy-smoking quantitative trait in two independent samples. *Am J Hum Genet*. 2007; 80(5):856–866. [PubMed: 17436240]
32. Uhl GR, et al. "Higher order" addiction molecular genetics: convergent data from genome-wide association in humans and mice. *Biochem Pharmacol*. 2008; 75(1):98–111. [PubMed: 17764662]
33. Johnson C, et al. Pooled association genome scanning for alcohol dependence using 104,268 SNPs: validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism. *Am J Med Genet B Neuropsychiatr Genet*. 2006; 141(8):844–853. [PubMed: 16894614]
34. Uhl GR, et al. Molecular genetics of nicotine dependence and abstinence: whole genome association using 520,000 SNPs. *BMC Genet*. 2007; 8(1):10. [PubMed: 17407593]
35. Saccone SF, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet*. 2007; 16(1): 36–49. [PubMed: 17135278]
36. Uhl GR, et al. Molecular genetics of successful smoking cessation: convergent genome-wide association results. (in press). *Arch Gen Psychiatry*. 2008
37. Uhl GR, et al. Genome-wide association for methamphetamine dependence: convergent results from two samples (in press). *Arch Gen Psychiatry*. 2008
38. Meaburn E, et al. Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics*. 2005; 6(1):52. [PubMed: 15811185]
39. Meaburn E, et al. Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. *Nucleic Acids Res*. 2006; 34(4):e27. [PubMed: 16478714]

40. Butcher LM, et al. Association analysis of mild mental impairment using DNA pooling to screen 432 brain-expressed single-nucleotide polymorphisms. *Mol Psychiatry*. 2005; 10(4):384–392. [PubMed: 15452586]
41. Butcher LM, et al. SNPs, microarrays and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6000 children. *Hum Mol Genet*. 2005; 14(10):1315–1325. [PubMed: 15800012]
42. Butcher LM, et al. Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav Genet*. 2004; 34(5): 549–555. [PubMed: 15319578]
43. Sham P, et al. DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet*. 2002; 3(11): 862–871. [PubMed: 12415316]
44. Pearson JV, et al. Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. *Am J Hum Genet*. 2007; 80(1): 126–139. [PubMed: 17160900]
45. Uhl, GR., et al. Genetic influences in drug abuse. In: Gloom, FE.; Kupfer, DJ., editors. *Psychopharmacology: The Fourth Generation of Progress*. New York: Raven Press; 1995. p. 1793-2783.
46. Tsuang MT, et al. Co-occurrence of abuse of different drugs in men: the role of drug-specific and shared vulnerabilities. *Arch Gen Psychiatry*. 1998; 55(11):967–972. [PubMed: 9819064]
47. Karkowski LM, Prescott CA, Kendler KS. Multivariate assessment of factors influencing illicit substance use in twins from female-female pairs. *Am J Med Genet*. 2000; 96(5):665–670. [PubMed: 11054775]
48. True WR, et al. Interrelationship of genetic and environmental influences on conduct disorder and alcohol and marijuana dependence symptoms. *Am J Med Genet*. 1999; 88(4):391–397. [PubMed: 10402507]
49. Merikangas KR, et al. Familial transmission of substance use disorders. *Arch Gen Psychiatry*. 1998; 55(11):973–979. [PubMed: 9819065]
50. Woodward CE, et al. Tobacco, alcohol and drug use in 8–16 year old twins. *NIDA Res. Monograph*. 1996; 162:309.
51. Tsuang MT, et al. Genetic influences on DSM-III-R drug abuse and dependence: a study of 3,372 twin pairs. *Am J Med Genet*. 1996; 67(5):473–477. [PubMed: 8886164]
52. Kendler KS, Prescott CA. Cocaine use, abuse and dependence in a population-based sample of female twins. *Br J Psychiatry*. 1998; 173:345–350. [PubMed: 9926041]
53. Kendler KS, et al. Illicit psychoactive substance use, abuse and dependence in a population-based sample of Norwegian twins. *Psychol Med*. 2006; 36(7):955–962. [PubMed: 16650346]
54. Agrawal A, et al. A twin study of early cannabis use and subsequent use and abuse/dependence of other illicit drugs. *Psychol Med*. 2004; 34(7):1227–1237. [PubMed: 15697049]
55. Grove WM, et al. Heritability of substance abuse and antisocial behavior: a study of monozygotic twins reared apart. *Biol Psychiatry*. 1990; 27(12):1293–1304. [PubMed: 2364118]
56. Gynther LM, et al. A twin study of non-alcohol substance abuse. *Psychiatry Res*. 1995; 56(3):213–220. [PubMed: 7568543]
57. Chen CC, et al. Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism. *Am J Hum Genet*. 1999; 65(3):795–807. [PubMed: 10441588]
58. Thomasson HR, et al. Alcohol and aldehyde dehydrogenase genotypes and alcoholism in Chinese men. *Am J Hum Genet*. 1991; 48(4):677–681. [PubMed: 2014795]
59. Chen WJ, et al. Alcohol-metabolising genes and alcoholism among Taiwanese Han men: independent effect of ADH2, ADH3 and ALDH2. *Br J Psychiatry*. 1996; 168(6):762–767. [PubMed: 8773821]
60. Shen YC, et al. Polymorphism of ADH and ALDH genes among four ethnic groups in China and effects upon the risk for alcoholism. *Alcohol Clin Exp Res*. 1997; 21(7):1272–1277. [PubMed: 9347089]
61. Higuchi S. Polymorphisms of ethanol metabolizing enzyme genes and alcoholism. *Alcohol Alcohol Suppl*. 1994; 2:29–34. [PubMed: 8974313]

62. Higuchi S, et al. Aldehyde dehydrogenase genotypes in Japanese alcoholics. *Lancet*. 1994; 343(8899):741–742. [PubMed: 7907720]
63. Higuchi S, et al. Alcohol and aldehyde dehydrogenase polymorphisms and the risk for alcoholism. *Am J Psychiatry*. 1995; 152(8):1219–1221. [PubMed: 7625477]
64. Maezawa Y, et al. Alcohol-metabolizing enzyme polymorphisms and alcoholism in Japan. *Alcohol Clin Exp Res*. 1995; 19(4):951–954. [PubMed: 7485844]
65. Nakamura K, et al. Characteristics of Japanese alcoholics with the atypical aldehyde dehydrogenase 2*2. I. A comparison of the genotypes of ALDH2, ADH2, ADH3, and cytochrome P-4502E1 between alcoholics and nonalcoholics. *Alcohol Clin Exp Res*. 1996; 20(1):52–55. [PubMed: 8651462]
66. Tanaka F, et al. Polymorphism of alcohol-metabolizing genes affects drinking behavior and alcoholic liver disease in Japanese men. *Alcohol Clin Exp Res*. 1997; 21(4):596–601. [PubMed: 9194910]
67. Luczak SE, et al. Genetic risk for alcoholism relates to level of response to alcohol in Asian-American men and women. *J Stud Alcohol*. 2002; 63(1):74–82. [PubMed: 11925062]
68. Schuckit MA, Doby J. Alcohol-related flushing and the risk for alcoholism in sons of alcoholics. *J Clin Psychiatry*. 1982; 43(10):415–418. [PubMed: 7118836]
69. Berrettini W, et al. alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol Psychiatry*. 2008
70. Porjesz B, et al. Amplitude of visual P3 event-related potential as a phenotypic marker for a predisposition to alcoholism: preliminary results from the COGA Project. Collaborative Study on the Genetics of Alcoholism. *Alcohol Clin Exp Res*. 1998; 22(6):1317–1323. [PubMed: 9756048]
71. Dick DM, et al. Association of GABRG3 with alcohol dependence. *Alcohol Clin Exp Res*. 2004; 28(1):4–9. [PubMed: 14745296]
72. Gelernter J, et al. Results of a genomewide linkage scan: support for chromosomes 9 and 11 loci increasing risk for cigarette smoking. *Am J Med Genet B Neuropsychiatr Genet*. 2004; 128(1):94–101. [PubMed: 15211640]
73. Saccone NL, et al. A genome screen of maximum number of drinks as an alcoholism phenotype. *Am J Med Genet*. 2000; 96(5):632–637. [PubMed: 11054770]
74. Schuckit MA, et al. A genome-wide search for genes that relate to a low level of response to alcohol. *Alcohol Clin Exp Res*. 2001; 25(3):323–329. [PubMed: 11290841]
75. Bergen AW, et al. Genomic regions linked to alcohol consumption in the Framingham Heart Study. *BMC Genet*. 2003; 4(Suppl 1):S101. [PubMed: 14975169]
76. Bierut LJ, et al. A genomic scan for habitual smoking in families of alcoholics: common and specific genetic factors in substance dependence. *Am J Med Genet A*. 2004; 124(1):19–27. [PubMed: 14679582]
77. Dick DM, et al. Endophenotypes successfully lead to gene identification: results from the collaborative study on the genetics of alcoholism. *Behav Genet*. 2006; 36(1):112–126. [PubMed: 16341909]
78. Dick DM, et al. Suggestive linkage on chromosome 1 for a quantitative alcohol-related phenotype. *Alcohol Clin Exp Res*. 2002; 26(10):1453–1460. [PubMed: 12394277]
79. Dick DM, et al. Association between GABRA1 and drinking behaviors in the collaborative study on the genetics of alcoholism sample. *Alcohol Clin Exp Res*. 2006; 30(7):1101–1110. [PubMed: 16792556]
80. Edenberg HJ, Foroud T. The genetics of alcoholism: identifying specific genes through family studies. *Addict Biol*. 2006; 11(3–4):386–396. [PubMed: 16961766]
81. Pinnaduwa D, Briollais L. Comparison of genotype- and haplotype-based approaches for fine-mapping of alcohol dependence using COGA data. *BMC Genet*. 2005; 6(Suppl 1):S65. [PubMed: 16451678]
82. Porjesz B, et al. Linkage and linkage disequilibrium mapping of ERP and EEG phenotypes. *Biol Psychol*. 2002; 61(1–2):229–248. [PubMed: 12385677]
83. Reck BH, et al. Analysis of alcohol dependence phenotype in the COGA families using covariates to detect linkage. *BMC Genet*. 2005; 6(Suppl 1):S143. [PubMed: 16451603]

84. Yang HC, et al. A genome-wide scanning and fine mapping study of COGA data. *BMC Genet.* 2005; 6(Suppl 1):S30. [PubMed: 16451640]
85. Gudmundsson J, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet.* 2007; 39(5):631–637. [PubMed: 17401366]
86. Plomin R, DeFries JC, Loehlin JC. Genotype-environment interaction and correlation in the analysis of human behavior. *Psychol Bull.* 1977; 84(2):309–322. [PubMed: 557211]
87. Scarr S, McCartney K. How people make their own environments: a theory of genotype greater than environment effects. *Child Dev.* 1983; 54(2):424–435. [PubMed: 6683622]
88. Posthuma D, et al. Theory and practice in quantitative genetics. *Twin Res.* 2003; 6(5):361–376. [PubMed: 14624720]
89. Lytton H, Martin NG, Eaves L. Environmental and genetical causes of variation in ethological aspects of behavior in two-year-old boys. *Soc Biol.* 1977; 24(3):200–211. [PubMed: 563105]
90. Falconer, DS.; MacKay, TFC. Introduction to quantitative genetics. 4 ed.. London: Longan Group Ltd.; 1996.
91. Buster MA, Rodgers JL. Genetic and environmental influences on alcohol use: DF analysis of NLSY kinship data. *J Biosoc Sci.* 2000; 32(2):177–189. [PubMed: 10765609]
92. Stanke F, et al. Diversity of the basic defect of homozygous CFTR mutation genotypes in humans. *J Med Genet.* 2008; 45(1):47–54. [PubMed: 18178635]
93. Douglas MR, Lewthwaite AJ, Nicholl DJ. Genetics of Parkinson's disease and parkinsonism. *Expert Rev Neurother.* 2007; 7(6):657–666. [PubMed: 17563249]
94. Stephan DA. Unraveling autism. *Am J Hum Genet.* 2008; 82(1):7–9. [PubMed: 18179879]
95. Alarcon M, et al. Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet.* 2008; 82(1):150–159. [PubMed: 18179893]
96. Arking DE, et al. A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am J Hum Genet.* 2008; 82(1):160–164. [PubMed: 18179894]
97. Bakkaloglu B, et al. Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. *Am J Hum Genet.* 2008; 82(1):165–173. [PubMed: 18179895]
98. Ptashne M. On the use of the word 'epigenetic'. *Curr Biol.* 2007; 17(7):R233–R236. [PubMed: 17407749]
99. Tan EK. The role of common genetic risk variants in Parkinson disease. *Clin Genet.* 2007; 72(5):387–393. [PubMed: 17868389]
100. Smith SS, et al. Genetic vulnerability to drug abuse. The D2 dopamine receptor Taq I B1 restriction fragment length polymorphism appears more frequently in polysubstance abusers. *Arch Gen Psychiatry.* 1992; 49(9):723–727. [PubMed: 1355337]
101. Persico AM, et al. D2 dopamine receptor gene TaqI A1 and B1 restriction fragment length polymorphisms: enhanced frequencies in psychostimulant-preferring polysubstance abusers. *Biol Psychiatry.* 1996; 40(8):776–784. [PubMed: 8894071]
102. Goldberg J, et al. A strategy for assembling samples of adult twin pairs in the United States. *Stat Med.* 1993; 12(18):1693–1702. [PubMed: 8248662]
103. Bucholz KK, et al. A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. *J Stud Alcohol.* 1994; 55(2):149–158. [PubMed: 8189735]
104. Itoh K, et al. Association study between brain-derived neurotrophic factor gene polymorphisms and methamphetamine abusers in Japan. *Am J Med Genet B Neuropsychiatr Genet.* 2005; 132(1):70–73. [PubMed: 15459944]
105. Iwata N, et al. No association is found between the candidate genes of t-PA/plasminogen system and Japanese methamphetamine-related disorder: a collaborative study by the Japanese Genetics Initiative for Drug Abuse. *Ann N Y Acad Sci.* 2004; 1025:34–38. [PubMed: 15542697]
106. Ujiie H, et al. Genetic studies on substance dependence. *Seishin Shinkeigaku Zasshi.* 2004; 106(12):1598–1603. [PubMed: 15770963]

107. Morita Y, et al. A nonsynonymous polymorphism in the human fatty acid amide hydrolase gene did not associate with either methamphetamine dependence or schizophrenia. *Neurosci Lett*. 2005; 376(3):182–187. [PubMed: 15721218]
108. Nishiyama T, et al. Haplotype association between GABAA receptor gamma2 subunit gene (GABRG2) and methamphetamine use disorder. *Pharmacogenomics J*. 2005; 5(2):89–95. [PubMed: 15772696]
109. Hashimoto T, et al. A functional glutathione S-transferase P1 gene polymorphism is associated with methamphetamine-induced psychosis in Japanese population. *Am J Med Genet B Neuropsychiatr Genet*. 2005; 135(1):5–9. [PubMed: 15729709]
110. Morita Y, et al. The X-box binding protein 1 (XBP1) gene is not associated with methamphetamine dependence. *Neurosci Lett*. 2005; 383(1–2):194–198. [PubMed: 15936534]
111. Ikeda M, et al. Positive association of AKT1 haplotype to Japanese methamphetamine use disorder. *Int J Neuropsychopharmacol*. 2005:1–5.
112. Ide S, et al. Gene polymorphisms of the mu opioid receptor in methamphetamine abusers. *Ann N Y Acad Sci*. 2004; 1025:316–324. [PubMed: 15542732]
113. Harano M, et al. A polymorphism of DRD2 gene and brain atrophy in methamphetamine psychosis. *Ann N Y Acad Sci*. 2004; 1025:307–315. [PubMed: 15542731]
114. Inada T, et al. No association found between the type 1 sigma receptor gene polymorphisms and methamphetamine abuse in the Japanese population: a collaborative study by the Japanese Genetics Initiative for Drug Abuse. *Ann N Y Acad Sci*. 2004; 1025:27–33. [PubMed: 15542696]
115. Koizumi H, et al. Association between the glutathione S-transferase M1 gene deletion and female methamphetamine abusers. *Am J Med Genet B Neuropsychiatr Genet*. 2004; 126(1):43–45. [PubMed: 15048646]
116. Ujike H, et al. Nine- or fewer repeat alleles in VNTR polymorphism of the dopamine transporter gene is a strong risk factor for prolonged methamphetamine psychosis. *Pharmacogenomics J*. 2003; 3(4):242–247. [PubMed: 12931138]
117. Carmelli D, et al. Quantitative genetic modeling of regional brain volumes and cognitive performance in older male twins. *Biol Psychol*. 2002; 61(1–2):139–155. [PubMed: 12385673]
118. Lerman C, et al. Role of functional genetic variation in the dopamine D2 receptor (DRD2) in response to bupropion and nicotine replacement therapy for tobacco dependence: results of two randomized clinical trials. *Neuropsychopharmacology*. 2006; 31(1):231–242. [PubMed: 16123753]
119. David SP, et al. Pharmacogenetic clinical trial of sustained-release bupropion for smoking cessation. *Nicotine & Tobacco Research*. 2007 (in press).
120. Rose JE, Behm FM, Westman EC. Nicotine-mecamylamine treatment for smoking cessation: the role of pre-cessation therapy. *Exp Clin Psychopharmacol*. 1998; 6(3):331–343. [PubMed: 9725117]
121. Shifman S, et al. A whole genome association study of neuroticism using DNA pooling. *Mol Psychiatry*. 2008; 13(3):302–312. [PubMed: 17667963]
122. Barcellos LF, et al. Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Genet*. 1997; 61(3):734–747. [PubMed: 9326338]
123. Germer S, Holland MJ, Higuchi R. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res*. 2000; 10(2):258–266. [PubMed: 10673283]
124. Hacia JG, et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet*. 1999; 22(2):164–167. [PubMed: 10369258]
125. Yang Y, et al. Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc Natl Acad Sci U S A*. 2003; 100(12):7225–7230. [PubMed: 12777616]
126. Moonesinghe R, et al. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc Natl Acad Sci U S A*. 2008; 105(2):617–622. [PubMed: 18174335]
127. Hinds DA, et al. Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. *Hum Genomics*. 2004; 1(6):421–434. [PubMed: 15606997]

128. Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* 1998; 8(12): 1273–1288. [PubMed: 9872982]
129. Teng J, Risch N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.* 1999; 9(3):234–241. [PubMed: 10077529]
130. Hishimoto A, et al. Neurexin 3 polymorphisms are associated with alcohol dependence and altered expression of specific isoforms. *Hum Mol Genet.* 2007; 16(23):2880–2891. [PubMed: 17804423]
131. Ishiguro H, et al. NrCAM in addiction vulnerability: positional cloning, drug-regulation, haplotype-specific expression, and altered drug reward in knockout mice. *Neuropsychopharmacology.* 2006; 31(3):572–584. [PubMed: 16123759]
132. Dupont WD, Plummer WD Jr. Power and sample size calculations. A review and computer program. *Control Clin Trials.* 1990; 11(2):116–128. [PubMed: 2161310]
133. Kruglyak L, Lander ES. High-resolution genetic mapping of complex traits. *Am J Hum Genet.* 1995; 56(5):1212–1223. [PubMed: 7726179]
134. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet.* 1995; 11(3):241–247. [PubMed: 7581446]
135. Maller J, et al. Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat Genet.* 2006; 38(9):1055–1059. [PubMed: 16936732]
136. Li M, et al. CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat Genet.* 2006; 38(9):1049–1054. [PubMed: 16936733]
137. Lin DY. Evaluating statistical significance in two-stage genomewide association studies. *Am J Hum Genet.* 2006; 78(3):505–509. [PubMed: 16408254]
138. Gao X, Starmer J. Human population structure detection via multilocus genotype clustering. *BMC Genet.* 2007; 8:34. [PubMed: 17592628]
139. Sabatti C, Service S, Freimer N. False discovery rate in linkage and association genome screens for complex disorders. *Genetics.* 2003; 164(2):829–833. [PubMed: 12807801]
140. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A.* 2003; 100(16):9440–9445. [PubMed: 12883005]
141. Benjamini Y, et al. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res.* 2001; 125(1–2):279–284. [PubMed: 11682119]
142. Sawcer S, et al. Empirical genomewide significance levels established by whole genome simulations. *Genet Epidemiol.* 1997; 14(3):223–229. [PubMed: 9181352]
143. Zhao LP, et al. On the assessment of statistical significance in disease-gene discovery. *Am J Hum Genet.* 1999; 64(6):1739–1753. [PubMed: 10330362]
144. Manolio TA, et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet.* 2007; 39(9):1045–1051. [PubMed: 17728769]
145. Wakefield J. Reporting and interpretation in genome-wide association studies. *Int J Epidemiol.* 2008
146. Sonne SC, Brady KT. Substance abuse and bipolar comorbidity. *Psychiatr Clin North Am.* 1999; 22(3):609–627. ix. [PubMed: 10550858]
147. Shih RA, Belmonte PL, Zandi PP. A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. *Int Rev Psychiatry.* 2004; 16(4): 260–283. [PubMed: 16194760]
148. Antila M, et al. Heritability of cognitive functions in families with bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet.* 2007; 144(6):802–808. [PubMed: 17907246]
149. McGuffin P, et al. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch Gen Psychiatry.* 2003; 60(5):497–502. [PubMed: 12742871]

150. Almeida OP, et al. Smoking as a risk factor for Alzheimer's disease: contrasting evidence from a systematic review of case-control and cohort studies. *Addiction*. 2002; 97(1):15–28. [PubMed: 11895267]
151. Wolf SM, Kahn JP. Genetic testing and the future of disability insurance: ethics, law & policy. *J Law Med Ethics*. 2007; 35(2 Suppl):6–32. [PubMed: 17543053]
152. Kristman VL, Kreiger N. Information Disclosure in Population-Based Research Involving Genetics: A Framework for the Practice of Ethics in Epidemiology. *Ann Epidemiol*. 2007
153. Lowrance WW, Collins FS. Ethics. Identifiability in genomic research. *Science*. 2007; 317(5838): 600–602. [PubMed: 17673640]
154. Papassotiropoulos A, et al. Common Kibra alleles are associated with human memory performance. *Science*. 2006; 314(5798):475–478. [PubMed: 17053149]
155. 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]
156. George SR, et al. Polymorphisms of the D4 dopamine receptor alleles in chronic alcoholism. *Biochem Biophys Res Commun*. 1993; 196(1):107–114. [PubMed: 8216280]
157. Muramatsu T, et al. Association between alcoholism and the dopamine D4 receptor gene. *J Med Genet*. 1996; 33(2):113–115. [PubMed: 8929946]
158. Li T, et al. Association analysis of the DRD4 and COMT genes in methamphetamine abuse. *Am J Med Genet B Neuropsychiatr Genet*. 2004; 129(1):120–124. [PubMed: 15274053]
159. Welzl H, Stork O. Cell adhesion molecules: key players in memory consolidation? *News Physiol Sci*. 2003; 18:147–150. [PubMed: 12869614]
160. Benson DL, et al. Making memories stick: cell-adhesion molecules in synaptic plasticity. *Trends Cell Biol*. 2000; 10(11):473–482. [PubMed: 11050419]
161. Fredette BJ, Miller J, Ranscht B. Inhibition of motor axon growth by T-cadherin substrata. *Development*. 1996; 122(10):3163–3171. [PubMed: 8898229]
162. Chen BE, et al. The molecular diversity of Dscam is functionally required for neuronal wiring specificity in *Drosophila*. *Cell*. 2006; 125(3):607–620. [PubMed: 16678102]
163. Keene AC, et al. *Drosophila* dorsal paired medial neurons provide a general mechanism for memory consolidation. *Curr Biol*. 2006; 16(15):1524–1530. [PubMed: 16890528]
164. Kipmen-Korgun D, et al. T-cadherin mediates low-density lipoprotein-initiated cell proliferation via the Ca(2+)-tyrosine kinase-Erk1/2 pathway. *J Cardiovasc Pharmacol*. 2005; 45(5):418–430. [PubMed: 15821437]
165. Philippova M, et al. RhoA and Rac mediate endothelial cell polarization and detachment induced by T-cadherin. *Faseb J*. 2005; 19(6):588–590. [PubMed: 15703273]
166. Hug C, et al. T-cadherin is a receptor for hexameric and high-molecular-weight forms of Acrp30/adiponectin. *Proc Natl Acad Sci U S A*. 2004; 101(28):10308–10313. [PubMed: 15210937]
167. Ivanov DB, Philippova MP, Tkachuk VA. Structure and functions of classical cadherins. *Biochemistry (Mosc)*. 2001; 66(10):1174–1186. [PubMed: 11736639]
168. Dean C, et al. Neurexin mediates the assembly of presynaptic terminals. *Nat Neurosci*. 2003; 6(7):708–716. [PubMed: 12796785]
169. Yamakawa K, et al. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. *Hum Mol Genet*. 1998; 7(2):227–237. [PubMed: 9426258]
170. Kraus DM, et al. CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. *J Immunol*. 2006; 176(7):4419–4430. [PubMed: 16547280]
171. Takeuchi T, et al. Expression of T-cadherin (CDH13, H-Cadherin) in human brain and its characteristics as a negative growth regulator of epidermal growth factor in neuroblastoma cells. *J Neurochem*. 2000; 74(4):1489–1497. [PubMed: 10737605]
172. Lein ES, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. 2007; 445(7124):168–176. [PubMed: 17151600]

173. Kim YS, et al. GLIS3, a novel member of the GLIS subfamily of Kruppel-like zinc finger proteins with repressor and activation functions. *Nucleic Acids Res.* 2003; 31(19):5513–5525. [PubMed: 14500813]
174. Barlow GM, et al. Down syndrome cell adhesion molecule is conserved in mouse and highly expressed in the adult mouse brain. *Cytogenet Cell Genet.* 2001; 94(3–4):155–162. [PubMed: 11856873]
175. Wojtowicz WM, et al. A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. *Cell.* 2007; 130(6):1134–1145. [PubMed: 17889655]
176. Gao FB. Molecular and cellular mechanisms of dendritic morphogenesis. *Curr Opin Neurobiol.* 2007; 17(5):525–532. [PubMed: 17933513]
177. Hattori D, et al. Dscam diversity is essential for neuronal wiring and self-recognition. *Nature.* 2007; 449(7159):223–227. [PubMed: 17851526]
178. Koob GF. The neurobiology of addiction: a neuroadaptational view relevant for diagnosis. *Addiction.* 2006; 101(Suppl 1):23–30. [PubMed: 16930158]
179. Liu F, et al. A genomewide screen for late-onset Alzheimer disease in a genetically isolated dutch population. *Am J Hum Genet.* 2007; 81(1):17–31. [PubMed: 17564960]
180. MacLaurin SA, Krucker T, Fish KN. Hippocampal dendritic arbor growth in vitro: regulation by Reelin-Disabled-1 signaling. *Brain Res.* 2007; 1172:1–9. [PubMed: 17825270]
181. Young-Pearse TL, et al. A critical function for beta-amyloid precursor protein in neuronal migration revealed by in utero RNA interference. *J Neurosci.* 2007; 27(52):14459–14469. [PubMed: 18160654]
182. Sheldon M, et al. Scrambler and yotari disrupt the disabled gene and produce a reeler-like phenotype in mice. *Nature.* 1997; 389(6652):730–733. [PubMed: 9338784]
183. Koop EA, et al. Receptor protein tyrosine phosphatase mu expression as a marker for endothelial cell heterogeneity; analysis of RPTPmu gene expression using LacZ knock-in mice. *Int J Dev Biol.* 2003; 47(5):345–354. [PubMed: 12895029]
184. Kee HJ, et al. Expression of brain-specific angiogenesis inhibitor 3 (BAI3) in normal brain and implications for BAI3 in ischemia-induced brain angiogenesis and malignant glioma. *FEBS Lett.* 2004; 569(1–3):307–316. [PubMed: 15225653]
185. Paul S, Lombroso PJ. Receptor and nonreceptor protein tyrosine phosphatases in the nervous system. *Cell Mol Life Sci.* 2003; 60(11):2465–2482. [PubMed: 14625689]
186. Uetani N, et al. Impaired learning with enhanced hippocampal long-term potentiation in PTPdelta-deficient mice. *Embo J.* 2000; 19(12):2775–2785. [PubMed: 10856223]
187. Uetani N, et al. Mammalian motoneuron axon targeting requires receptor protein tyrosine phosphatases sigma and delta. *J Neurosci.* 2006; 26(22):5872–5880. [PubMed: 16738228]
188. Cai H, et al. The sarcoglycan complex in Schwann cells and its role in myelin stability. *Exp Neurol.* 2007; 205(1):257–269. [PubMed: 17397833]
189. Shiga K, et al. Zeta-sarcoglycan is a functional homologue of gamma-sarcoglycan in the formation of the sarcoglycan complex. *Exp Cell Res.* 2006; 312(11):2083–2092. [PubMed: 16635485]
190. Lau WL, Scholnick SB. Identification of two new members of the CSMD gene family small star, filled. *Genomics.* 2003; 82(3):412–415. [PubMed: 12906867]
191. Zheng C, Heintz N, Hatten ME. CNS gene encoding astrotactin, which supports neuronal migration along glial fibers. *Science.* 1996; 272(5260):417–419. [PubMed: 8602532]
192. Adams NC, et al. Mice that lack astrotactin have slowed neuronal migration. *Development.* 2002; 129(4):965–972. [PubMed: 11861479]
193. Price M, et al. Seven cDNAs enriched following hippocampal lesion: possible roles in neuronal responses to injury. *Brain Res Mol Brain Res.* 2003; 117(1):58–67. [PubMed: 14499481]
194. Cui XY, et al. NB-3/Notch1 pathway via Deltex1 promotes neural progenitor cell differentiation into oligodendrocytes. *J Biol Chem.* 2004; 279(24):25858–25865. [PubMed: 15082708]
195. Takeda Y, et al. Impaired motor coordination in mice lacking neural recognition molecule NB-3 of the contactin/F3 subgroup. *J Neurobiol.* 2003; 56(3):252–265. [PubMed: 12884264]

196. Lee S, et al. Expression and regulation of a gene encoding neural recognition molecule NB-3 of the contactin/F3 subgroup in mouse brain. *Gene*. 2000; 245(2):253–266. [PubMed: 10717476]
197. Hernandez-Montiel HL, et al. Semaphorins 3A, 3C, and 3F in mesencephalic dopaminergic axon pathfinding. *J Comp Neurol*. 2008; 506(3):387–397. [PubMed: 18041777]
198. Ara J, et al. Modulation of sciatic nerve expression of class 3 semaphorins by nerve injury. *Neurochem Res*. 2004; 29(6):1153–1159. [PubMed: 15176472]
199. Steup A, et al. Sema3C and netrin-1 differentially affect axon growth in the hippocampal formation. *Mol Cell Neurosci*. 2000; 15(2):141–155. [PubMed: 10673323]
200. Park C, et al. Deletion in *Catna2*, encoding alpha N-catenin, causes cerebellar and hippocampal lamination defects and impaired startle modulation. *Nat Genet*. 2002; 31(3):279–284. [PubMed: 12089526]
201. Uemura M, Takeichi M. Alpha N-catenin deficiency causes defects in axon migration and nuclear organization in restricted regions of the mouse brain. *Dev Dyn*. 2006; 235(9):2559–2566. [PubMed: 16691566]
202. Caltagarone J, Jing Z, Bowser R. Focal adhesions regulate Abeta signaling and cell death in Alzheimer's disease. *Biochim Biophys Acta*. 2007; 1772(4):438–445. [PubMed: 17215111]
203. Fein G, Di Sclafani V. Cerebral reserve capacity: implications for alcohol and drug abuse. *Alcohol*. 2004; 32(1):63–67. [PubMed: 15066705]
204. Borenstein AR, et al. Developmental and vascular risk factors for Alzheimer's disease. *Neurobiol Aging*. 2005; 26(3):325–334. [PubMed: 15639310]
205. Dean C, Dresbach T. Neuroligins and neuexins: linking cell adhesion, synapse formation and cognitive function. *Trends Neurosci*. 2006; 29(1):21–29. [PubMed: 16337696]
206. Ushkaryov YA, et al. Neuexins: synaptic cell surface proteins related to the alpha-latrotoxin receptor and laminin. *Science*. 1992; 257(5066):50–56. [PubMed: 1621094]
207. Ushkaryov YA, et al. Conserved domain structure of beta-neuexins. Unusual cleaved signal sequences in receptor-like neuronal cell-surface proteins. *J Biol Chem*. 1994; 269(16):11987–11992. [PubMed: 8163501]
208. Ushkaryov YA, Sudhof TC. Neuexin III alpha: extensive alternative splicing generates membrane-bound and soluble forms. *Proc Natl Acad Sci U S A*. 1993; 90(14):6410–6414. [PubMed: 8341647]
209. Boucard AA, et al. A splice code for trans-synaptic cell adhesion mediated by binding of neuroligin 1 to alpha- and beta-neuexins. *Neuron*. 2005; 48(2):229–236. [PubMed: 16242404]
210. Graf ER, et al. Structure function and splice site analysis of the synaptogenic activity of the neuexin-1 beta LNS domain. *J Neurosci*. 2006; 26(16):4256–4265. [PubMed: 16624946]
211. Graf ER, et al. Neuexins induce differentiation of GABA and glutamate postsynaptic specializations via neuroligins. *Cell*. 2004; 119(7):1013–1026. [PubMed: 15620359]
212. Kattenstroth G, et al. Postsynaptic N-methyl-D-aspartate receptor function requires alpha-neuexins. *Proc Natl Acad Sci U S A*. 2004; 101(8):2607–2612. [PubMed: 14983056]
213. Missler M, Fernandez-Chacon R, Sudhof TC. The making of neuexins. *J Neurochem*. 1998; 71(4):1339–1347. [PubMed: 9751164]
214. Missler M, et al. Alpha-neuexins couple Ca²⁺ channels to synaptic vesicle exocytosis. *Nature*. 2003; 423(6943):939–948. [PubMed: 12827191]
215. Rowen L, et al. Analysis of the human neuexin genes: alternative splicing and the generation of protein diversity. *Genomics*. 2002; 79(4):587–597. [PubMed: 11944992]
216. Tabuchi K, Sudhof TC. Structure and evolution of neuexin genes: insight into the mechanism of alternative splicing. *Genomics*. 2002; 79(6):849–859. [PubMed: 12036300]
217. Occhi G, et al. Identification and characterization of heart-specific splicing of human neuexin 3 mRNA (NRXN3). *Biochem Biophys Res Commun*. 2002; 298(1):151–155. [PubMed: 12379233]
218. Ullrich B, Ushkaryov YA, Sudhof TC. Cartography of neuexins: more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron*. 1995; 14(3):497–507. [PubMed: 7695896]
219. Lachman HM, et al. Genomewide Suggestive Linkage of Opioid Dependence to Chromosome 14q. *Hum Mol Genet*. 2007

220. Nissim-Rafinia M, Kerem B. Splicing regulation as a potential genetic modifier. *Trends Genet.* 2002; 18(3):123–127. [PubMed: 11858835]
221. Nissim-Rafinia M, Kerem B. The splicing machinery is a genetic modifier of disease severity. *Trends Genet.* 2005; 21(9):480–483. [PubMed: 16039004]
222. Gorecki DC, Derry JM, Barnard EA. Dystroglycan: brain localisation and chromosome mapping in the mouse. *Hum Mol Genet.* 1994; 3(9):1589–1597. [PubMed: 7833916]
223. Nikonenko A, et al. Tenascin-R-deficient mice show structural alterations of symmetric perisomatic synapses in the CA1 region of the hippocampus. *J Comp Neurol.* 2003; 456(4):338–349. [PubMed: 12532406]
224. Sorra KE, Harris KM. Overview on the structure, composition, function, development, and plasticity of hippocampal dendritic spines. *Hippocampus.* 2000; 10(5):501–511. [PubMed: 11075821]
225. Grumet M. Nr-CAM: a cell adhesion molecule with ligand and receptor functions. *Cell Tissue Res.* 1997; 290(2):423–428. [PubMed: 9321706]
226. Niculescu AB 3rd, Kelsoe JR. Convergent functional genomics: application to bipolar disorder. *Ann Med.* 2001; 33(4):263–271. [PubMed: 11405548]
227. Ngandu T, et al. Education and dementia: what lies behind the association? *Neurology.* 2007; 69(14):1442–1450. [PubMed: 17909157]
228. Uhl GR, Grow RW. The burden of complex genetics in brain disorders. *Arch Gen Psychiatry.* 2004; 61(3):223–229. [PubMed: 14993109]
229. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25(1):25–29. [PubMed: 10802651]
230. Shapiro L, Love J, Colman DR. Adhesion molecules in the nervous system: structural insights into function and diversity. *Annu Rev Neurosci.* 2007; 30:451–474. [PubMed: 17600523]
231. Yamada S, Nelson WJ. Synapses: sites of cell recognition, adhesion, and functional specification. *Annu Rev Biochem.* 2007; 76:267–294. [PubMed: 17506641]
232. Thiery JP. Cell adhesion in development: a complex signaling network. *Curr Opin Genet Dev.* 2003; 13(4):365–371. [PubMed: 12888009]
233. Rodriguez-Fernandez JL, Corbi AL. Adhesion molecules in human dendritic cells. *Curr Opin Investig Drugs.* 2005; 6(11):1103–1111.
234. Hirano S, Suzuki ST, Redies C. The cadherin superfamily in neural development: diversity, function and interaction with other molecules. *Front Biosci.* 2003; 8:d306–d355. [PubMed: 12456358]
235. Bamji SX. Cadherins: actin with the cytoskeleton to form synapses. *Neuron.* 2005; 47(2):175–178. [PubMed: 16039559]
236. Salinas PC, Price SR. Cadherins and catenins in synapse development. *Curr Opin Neurobiol.* 2005; 15(1):73–80. [PubMed: 15721747]
237. Mulder N, Apweiler R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol.* 2007; 396:59–70. [PubMed: 18025686]
238. Wheeler DL, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2008; 36:D13–D21. Database issue. [PubMed: 18045790]
239. Ohashi K, et al. Cloning of the cDNA for a novel receptor tyrosine kinase, Sky, predominantly expressed in brain. *Oncogene.* 1994; 9(3):699–705. [PubMed: 8108112]
240. Brady-Kalnay SM, Rimm DL, Tonks NK. Receptor protein tyrosine phosphatase PTPmu associates with cadherins and catenins in vivo. *J Cell Biol.* 1995; 130(4):977–986. [PubMed: 7642713]
241. Brady-Kalnay SM, Tonks NK. Protein tyrosine phosphatases as adhesion receptors. *Curr Opin Cell Biol.* 1995; 7(5):650–657. [PubMed: 8573339]
242. McAndrew PE, et al. Identification and characterization of RPTP rho, a novel RPTP mu/kappa-like receptor protein tyrosine phosphatase whose expression is restricted to the central nervous system. *Brain Res Mol Brain Res.* 1998; 56(1–2):9–21. [PubMed: 9602027]
243. Beltran PJ, Bixby JL. Receptor protein tyrosine phosphatases as mediators of cellular adhesion. *Front Biosci.* 2003; 8:d87–d99. [PubMed: 12456340]

244. DeFelipe J, et al. Estimation of the number of synapses in the cerebral cortex: methodological considerations. *Cereb Cortex*. 1999; 9(7):722–732. [PubMed: 10554995]
245. Schuz A, Palm G. Density of neurons and synapses in the cerebral cortex of the mouse. *J Comp Neurol*. 1989; 286(4):442–455. [PubMed: 2778101]
246. Posthuma D, et al. Genetic correlations between brain volumes and the WAIS-III dimensions of verbal comprehension, working memory, perceptual organization, and processing speed. *Twin Res*. 2003; 6(2):131–139. [PubMed: 12723999]
247. Posthuma D, et al. The association between brain volume and intelligence is of genetic origin. *Nat Neurosci*. 2002; 5(2):83–84. [PubMed: 11818967]
248. Mori S, Zhang J. Principles of diffusion tensor imaging and its applications to basic neuroscience research. *Neuron*. 2006; 51(5):527–539. [PubMed: 16950152]
249. Pfefferbaum A, Sullivan EV, Carmelli D. Genetic regulation of regional microstructure of the corpus callosum in late life. *Neuroreport*. 2001; 12(8):1677–1681. [PubMed: 11409738]
250. Sullivan EV, et al. Heritability of hippocampal size in elderly twin men: equivalent influence from genes and environment. *Hippocampus*. 2001; 11(6):754–762. [PubMed: 11811670]
251. Regier DA, et al. Comorbidity of mental disorders with alcohol and other drug abuse. Results from the Epidemiologic Catchment Area (ECA) Study. *Jama*. 1990; 264(19):2511–2518. [PubMed: 2232018]
252. Stern Y, et al. Influence of education and occupation on the incidence of Alzheimer's disease. *Jama*. 1994; 271(13):1004–1010. [PubMed: 8139057]
253. Stern Y, et al. Rate of memory decline in AD is related to education and occupation: cognitive reserve? *Neurology*. 1999; 53(9):1942–1947. [PubMed: 10599762]
254. Bergeman CS, et al. Genetic and environmental effects on openness to experience, agreeableness, and conscientiousness: an adoption/twin study. *J Pers*. 1993; 61(2):159–179. [PubMed: 8345444]
255. Terracciano A, Costa PT Jr. Smoking and the Five-Factor Model of personality. *Addiction*. 2004; 99(4):472–481. [PubMed: 15049747]
256. Brooner RK, et al. Antisocial personality disorder among drug abusers. Relations to other personality diagnoses and the five-factor model of personality. *J Nerv Ment Dis*. 1993; 181(5):313–319. [PubMed: 8501448]
257. Weitzdoerfer R, et al. Neuronal nitric oxide synthase knock-out mice show impaired cognitive performance. *Nitric Oxide*. 2004; 10(3):130–140. [PubMed: 15158692]
258. Itzhak Y, et al. The role of neuronal nitric oxide synthase in cocaine-induced conditioned place preference. *Neuroreport*. 1998; 9(11):2485–2488. [PubMed: 9721919]
259. Kleppisch T, et al. Hippocampal cGMP-dependent protein kinase I supports an age- and protein synthesis-dependent component of long-term potentiation but is not essential for spatial reference and contextual memory. *J Neurosci*. 2003; 23(14):6005–6012. [PubMed: 12853418]
260. Liang GB, et al. Synthesis and SAR studies of potent imidazopyridine anticoccidial agents. *Bioorg Med Chem Lett*. 2007; 17(13):3558–3561. [PubMed: 17475489]
261. Asensio AC, et al. Biochemical and immunochemical characterisation of human diadenosine triphosphatase provides evidence for its identification with the tumour suppressor Fhit protein. *Biochimie*. 2006; 88(5):461–471. [PubMed: 16359767]
262. Pintor J, et al. Diadenosine polyphosphate receptors. from rat and guinea-pig brain to human nervous system. *Pharmacol Ther*. 2000; 87(2–3):103–115. [PubMed: 11007994]
263. Weiske J, Albring KF, Huber O. The tumor suppressor Fhit acts as a repressor of beta-catenin transcriptional activity. *Proc Natl Acad Sci U S A*. 2007; 104(51):20344–20349. [PubMed: 18077326]
264. Grewal PK, et al. Characterization of the LARGE family of putative glycosyltransferases associated with dystroglycanopathies. *Glycobiology*. 2005; 15(10):912–923. [PubMed: 15958417]
265. Longman C, et al. Mutations in the human LARGE gene cause MDC1D, a novel form of congenital muscular dystrophy with severe mental retardation and abnormal glycosylation of alpha-dystroglycan. *Hum Mol Genet*. 2003; 12(21):2853–2861. [PubMed: 12966029]

266. Grewal PK, Hewitt JE. Mutation of Large, which encodes a putative glycosyltransferase, in an animal model of muscular dystrophy. *Biochim Biophys Acta*. 2002; 1573(3):216–224. [PubMed: 12417403]
267. Hyman SE. Addiction: a disease of learning and memory. *Am J Psychiatry*. 2005; 162(8):1414–1422. [PubMed: 16055762]
268. Kelley AE. Memory and addiction: shared neural circuitry and molecular mechanisms. *Neuron*. 2004; 44(1):161–179. [PubMed: 15450168]
269. Nestler EJ. Neurobiology. Total recall-the memory of addiction. *Science*. 2001; 292(5525):2266–2267. [PubMed: 11423644]
270. Mori F, et al. Developmental changes in expression of the three ryanodine receptor mRNAs in the mouse brain. *Neurosci Lett*. 2000; 285(1):57–60. [PubMed: 10788707]
271. Alkon DL, et al. Time domains of neuronal Ca²⁺ signaling and associative memory: steps through a calyculin, ryanodine receptor, K⁺ channel cascade. *Trends Neurosci*. 1998; 21(12):529–537. [PubMed: 9881851]
272. Ross OA, Worrall BB, Meschia JF. Advancing stroke therapeutics through genetic understanding. *Curr Drug Targets*. 2007; 8(7):850–859. [PubMed: 17630939]
273. Giovannoni MP, et al. Synthesis of pyrrolo[2,3-d]pyridazinones as potent, subtype selective PDE4 inhibitors. *J Enzyme Inhib Med Chem*. 2007; 22(3):309–318. [PubMed: 17674813]
274. Zhang HT, et al. Effects of the novel PDE4 inhibitors MEM1018 and MEM1091 on memory in the radial-arm maze and inhibitory avoidance tests in rats. *Psychopharmacology (Berl)*. 2005; 179(3):613–619. [PubMed: 15672274]
275. Verploegen S, et al. Characterization of the role of CaMKI-like kinase (CKLiK) in human granulocyte function. *Blood*. 2005; 106(3):1076–1083. [PubMed: 15840691]
276. Balschun D, et al. Deletion of the ryanodine receptor type 3 (RyR3) impairs forms of synaptic plasticity and spatial learning. *Embo J*. 1999; 18(19):5264–5273. [PubMed: 10508160]
277. Kouzu Y, et al. Mutant mice lacking ryanodine receptor type 3 exhibit deficits of contextual fear conditioning and activation of calcium/calmodulin-dependent protein kinase II in the hippocampus. *Brain Res Mol Brain Res*. 2000; 76(1):142–150. [PubMed: 10719224]
278. Ostrovskaya O, et al. Inhibition of ryanodine receptors by 4-(2-aminopropyl)-3,5-dichloro-N,N-dimethylaniline (FLA 365) in canine pulmonary arterial smooth muscle cells. *J Pharmacol Exp Ther*. 2007; 323(1):381–390. [PubMed: 17640951]
279. Wang Q, McEwen DG, Ornitz DM. Subcellular and developmental expression of alternatively spliced forms of fibroblast growth factor 14. *Mech Dev*. 2000; 90(2):283–287. [PubMed: 10640713]
280. Olsen SK, et al. Fibroblast growth factor (FGF) homologous factors share structural but not functional homology with FGFs. *J Biol Chem*. 2003; 278(36):34226–34236. [PubMed: 12815063]
281. Vadasz C, et al. Mesencephalic dopamine neuron number and tyrosine hydroxylase content: Genetic control and candidate genes. *Neuroscience*. 2007; 149(3):561–572. [PubMed: 17920205]
282. Adriani W, et al. Short-term effects of adolescent methylphenidate exposure on brain striatal gene expression and sexual/endocrine parameters in male rats. *Ann N Y Acad Sci*. 2006; 1074:52–73. [PubMed: 17105903]
283. Wozniak DF, et al. Impaired spatial learning and defective theta burst induced LTP in mice lacking fibroblast growth factor 14. *Neurobiol Dis*. 2007; 26(1):14–26. [PubMed: 17236779]
284. Xiao M, et al. Impaired hippocampal synaptic transmission and plasticity in mice lacking fibroblast growth factor 14. *Mol Cell Neurosci*. 2007; 34(3):366–377. [PubMed: 17208450]
285. Dalski A, et al. Mutation analysis in the fibroblast growth factor 14 gene: frameshift mutation and polymorphisms in patients with inherited ataxias. *Eur J Hum Genet*. 2005; 13(1):118–120. [PubMed: 15470364]
286. Zhang HT, et al. Antidepressant-like profile and reduced sensitivity to rolipram in mice deficient in the PDE4D phosphodiesterase enzyme. *Neuropsychopharmacology*. 2002; 27(4):587–595. [PubMed: 12377395]
287. Miller CA, Marshall JF. Molecular substrates for retrieval and reconsolidation of cocaine-associated contextual memory. *Neuron*. 2005; 47(6):873–884. [PubMed: 16157281]

288. Vorel SR, et al. Relapse to cocaine-seeking after hippocampal theta burst stimulation. *Science*. 2001; 292(5519):1175–1178. [PubMed: 11349151]
289. Volkow ND, Li TK. Drugs and alcohol: treating and preventing abuse, addiction and their medical consequences. *Pharmacol Ther*. 2005; 108(1):3–17. [PubMed: 16098597]
290. Liu X, et al. Smaller volume of prefrontal lobe in polysubstance abusers: a magnetic resonance imaging study. *Neuropsychopharmacology*. 1998; 18(4):243–252. [PubMed: 9509492]
291. Bolla KI, et al. Neural substrates of faulty decision-making in abstinent marijuana users. *Neuroimage*. 2005; 26(2):480–492. [PubMed: 15907305]
292. Bolla K, et al. Prefrontal cortical dysfunction in abstinent cocaine abusers. *J Neuropsychiatry Clin Neurosci*. 2004; 16(4):456–464. [PubMed: 15616172]
293. Grant S, Contoreggi C, London ED. Drug abusers show impaired performance in a laboratory test of decision making. *Neuropsychologia*. 2000; 38(8):1180–1187. [PubMed: 10838152]
294. Tsuang MT, et al. The Harvard Twin Study of Substance Abuse: what we have learned. *Harv Rev Psychiatry*. 2001; 9(6):267–279. [PubMed: 11600486]
295. Mortimer JA, Snowden DA, Markesbery WR. Head circumference, education and risk of dementia: findings from the Nun Study. *J Clin Exp Neuropsychol*. 2003; 25(5):671–679. [PubMed: 12815504]
296. Godwin-Austen RB, et al. Smoking and Parkinson's disease. *J Neurol Neurosurg Psychiatry*. 1982; 45(7):577–581. [PubMed: 7119826]
297. Allam MF, et al. Smoking and Parkinson's disease: systematic review of prospective studies. *Mov Disord*. 2004; 19(6):614–621. [PubMed: 15197698]
298. Gottschalk PC, Jacobsen LK, Kosten TR. Current concepts in pharmacotherapy of substance abuse. *Curr Psychiatry Rep*. 1999; 1(2):172–178. [PubMed: 11122921]
299. Tang Y, et al. Total regional and global number of synapses in the human brain neocortex. *Synapse*. 2001; 41(3):258–273. [PubMed: 11418939]
300. Reiman EM, et al. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron*. 2007; 54(5):713–720. [PubMed: 17553421]
301. Feil S, et al. Distribution of cGMP-dependent protein kinase type I and its isoforms in the mouse brain and retina. *Neuroscience*. 2005; 135(3):863–868. [PubMed: 16154279]
302. Demyanenko GP, et al. Abnormal neocortical development in mice lacking cGMP-dependent protein kinase I. *Brain Res Dev Brain Res*. 2005; 160(1):1–8.

Glossary

<i>A priori</i>	Existing in the mind prior to and independent of experiments.
Balancing selection	A natural process that results in the survival and reproductive success of individuals or groups best adjusted to their environment and that leads to the perpetuation of genetic qualities best suited to that particular environment.
Between locus heterogeneity	A single disorder, trait, or pattern of traits caused by mutations in genes at different chromosomal loci.
Common disease and common allele model	The illness results from the cumulative impact of multiple common small-effect, genetic variants, interacting with environmental exposures to exceed a biological threshold.
Complex genetic phenotype (polygenic and multifactorial traits)	Any <i>phenotype</i> that results from the effect of multiple <i>genes</i> at two or more loci, with possible environmental influences too.
Epigenetic	Changes in the regulation of the expression of gene activity without alteration of DNA sequence.

Epistasis	A mutation in one gene masks the expression of a different gene.
Genetic heterogeneity	A single disorder, trait, or pattern of traits caused by genetic factors in some cases and non-genetic factors in others.
Genetic selection	Differential and non-random reproduction of different genotypes, operating to alter the gene frequencies within a population.
Genome wide association study	Any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition.
Linkage	The tendency for genes or segments of DNA closely positioned along a chromosome to segregate together at meiosis and therefore be inherited together.
Linkage analysis	Testing DNA sequence polymorphisms that are near or within a gene of interest to track within a family the inheritance of a disease-causing in a given gene.
Linkage disequilibrium	In a population, co-occurrence of a specific DNA marker and a disease at a higher frequency than would be predicted by random chance.
Pharmacogenomics	The study of how an individual's genomic system affects the body's response to drugs.
Pharmacogenetics	The study focused on specific genes, such as drug-metabolizing enzymes.
Pleiotropy	Multiple, often seemingly unrelated, physical effects caused by a single altered gene or pair of altered genes.
Segregation analysis	The determination of the number of progeny that have inherited distinct and mutually exclusive phenotypes.
Susceptibility gene	A gene mutation that increases the likelihood that an individual will develop a certain disease or disorder. When such a mutation is inherited, development of symptoms is more likely but not certain.
Transitive	Passing over to or affecting something else.
Within locus heterogeneity	A single disorder, trait, or pattern of traits influenced by several different variants at a single chromosomal locus

Figure 1. Pie graph model for the genetic architecture of human vulnerability to dependence on addictive substances

Polygenic additive genetic influences and environmental influences that are largely those that are not shared between members of sibships are depicted. Potential roles for $g \times g$ and $g \times e$ interactions are not depicted here.

Figure 2. Venn diagram of overlapping genetic contributions to several of the phenotypes discussed here based on genome wide association datasets of about 500000 SNPs
Note that the area of overlap in the figure does not necessarily represent the area of the overlap in the datasets. See text for more details.

Figure 3. Validation of SNP genotyping in DNA pools (From [12])

The relationship ($r = 0.95$) between individual and pooled genotyping using 500k SNP Affymetrix arrays provides an opportunity to assess the sensitivity of pooled genotyping. Since these validation experiments were the first ones performed with new array sets, these data provide a lower limit. Current results from 1M SNP arrays (6.0) provide relationships ca 0.98 (Drgon *et al*, *in preparation*).

Figure 4. Power of genome wide association as assessed using Gene Detective

Simulation of power with 620,000 diallelic markers for samples with $n = 400$ case and $n = 400$ controls with nominal 0.05 α levels. Note the striking relationship between power and effect size. Power to detect effects that would produce odds ratios of less than 1.2 –fold is modest, while power to detect effects as high as 1.7 fold is relatively good.

Figure 5. Power of genome wide association as assessed using Gene Detective
Simulation of power with 1,000,000 diallelic markers for samples with $n = 2000$ case and $n = 2000$ controls with nominal 0.05 α levels. Note that the striking relationship between power and effect size is retained.

Figure 6. Distributions of data mapped onto cartoons of human chromosomes

Chromosomes 1– 8 (*top row*), 9 – 16 (*second row*), 17–21 (*third row*), 22 (*fourth row*).

Red triangles to the left of the (black) main axis for each chromosome mark locations where genes are identified by clustered positive SNPs from Samples 1 & 2, and support is provided from at least one other sample.

Green triangles to the left of the main axis mark locations where the gene is also identified by clustered positive SNPs from both Methamphetamine Samples 4 & 5.

Blue triangles to the left of the main axis mark locations where the gene is also identified by clustered positive SNPs from 2 of the 3 Nicotine Abstinence Samples 11–14.

Black squares to the right of the main axis where the gene is also identified by clustered positive SNPs from 2 of the 3 Bipolar Disease Samples 7–9.

Purple squares to the right of the main axis mark locations where the gene is also identified by clustered positive SNPs from both Cognitive Function Samples 15 & 16.

Grey squares to the right of the main axis mark locations where the gene is also identified by clustered positive SNPs from both Alzheimer's Disease Samples 17 & 18.

Yellow squares to the right of the main axis mark locations where the gene is also identified by clustered positive SNPs from the NHLBI Frontal Brain Volume Sample 10.

Chromosomal positions are based on National Center for Biotechnology Information MAPVIEWER Build 36.1 coordinates and supplemental data from NETAFFX.

“Cell adhesion related” genes identified in multiple genome wide association studies of addiction and related disorders

Table 1

Gene	Description	Chr	bp	Samples with clustered positive SNPs	p
BAI3	brain-specific angiogenesis inhibitor 3	6	69404158	1,2,3,7,9,10,19	<0.00001
CDH13	cadherin 13	16	81218079	1,2,3,4,5,6,7,8,9,11,12,13,14,15,16,19	<0.00001
CLSTN2	calsyntenin 2	3	141136897	3,4,5,7,8,10,12,13,14,15,16,17,18,19	<0.00001
CNTNAP2	contactin associated protein-like 2	7	145444386	3,4,5,6,7,8,9,11,14,15,16,18,19	<0.00001
CSMD1	CUB and Sushi multiple domains 1	8	2782789	1,2,3,4,5,6,7,8,9,10,11,13,14,15,16,17,18,19	<0.00001
CTNNA2	catenin α 2	2	79593634	3,4,5,6,7,8,9,13,14,15,16,19	<0.00001
DAB1	disabled homolog 1	1	57236167	3,4,5,6,7,8,9,12,14,15,16,18,19	<0.00001
DSCAM	Down syndrome cell adhesion molecule	21	40306213	1,2,3,4,5,7,8,12,13,15,16,17,19	<0.00001
NRXN1	neurexin 1	2	50000992	3,4,5,6,9,11,14,15,16,17,18,19	<0.00001
PTPRD	receptor protein tyrosine phosphatase D	9	8307268	1,2,3,4,5,6,8,11,13,14,15,16,17	<0.00001
SGCZ	sarcoglycan zeta	8	13991744	1,2,3,4,5,8,9,10,11,13,14,15,16,17,19	<0.00001
ASTN2	astrotactin 2	9	118227328	1,2,4,5,6,8,9,13,15,16,17,18,19	0.000070
CNTN4	contactin 4	3	2117247	1,2,3,6,8,9,11,15,16,17,18	0.000110
CNTN6	contactin 6	3	1109629	1,2,3,4,6,8,9,10,15,18	0.000120
LRP1B	low density lipoprotein-related protein 1B	2	140705466	1,2,3,5,6,8,9,10,11,13,15,16,18,19	0.000150
NRG1	neuregulin 1	8	32525295	3,4,5,6,7,11,15,16	0.000240
ITGB8	integrin β 8	7	20337250	1,2,8,9,15,16	0.000260
PTPRM	receptor protein tyrosine phosphatase M	18	7557817	1,2,3,4,5,7,11,13,15,16,18,19	0.000290
ROR1	receptor tyrosine kinase-like orphan rec 1	1	64012302	4,5,7,8,9,13,15,16,18	0.000290
TRIO	triple functional domain/PTPRF interact	5	14196829	1,2,9,11,13,14,15	0.000690
CSMD2	CUB and Sushi multiple domains 2	1	33752196	1,2,3,4,7,8,11,14,17,18,19	0.000830
CNTN5	contactin 5	11	98397081	1,2,3,5,6,7,8,9,11,14,15,19	0.000980
CTNNA3	catenin α 3	10	67349937	1,2,3,5,6,7,8,13,15,16,19	0.001090
LRRN6C	leucine rich repeat neuronal 6C	9	27938528	1,2,3,4,5,6,7,8,9,11,15,16,17,19	0.001340
CTNND2	catenin δ 2	5	11024952	1,2,3,5,6,9,15,16	0.003270
ANKS1B	ankyrin repeat sterile α domain 1B	12	97653202	1,2,5,6,9,15,16,19	0.003410
SEMA3C	semaphorin 3C	7	80209790	1,2,9,16	0.006310

Columns list gene symbol, gene description, chromosome, basepair of gene start, samples (from this review) in which clustered nominally-positive SNPs are found and overall p value for this gene in this entire dataset, based on 100,000 Monte Carlo simulation trials. Note that genes that are identified strongly in a single sample will not be included on this list.

“Druggable” genes identified in multiple genome wide association studies of addiction related disorders

Table II

Gene	Class	Description	Chr	Start	samples	p
RVR3	CHA	ryanodine receptor 3	15	31,390,469	1,2,3,4,5,6,7,8,9,13,15,16,17,19	<0.00001
CAMK1D	ENZ	calcium/calmodulin-dependent protein kinase ID	10	12,431,589	1,2,7,8,9,13,15,16,17,18,19	<0.00001
CHN2	ENZ	chimerin 2	7	29,200,646	1,2,3,7,8,10,13,14,15,16	<0.00001
FHIT	ENZ	fragile histidine triad gene	3	59,710,076	1,2,3,4,5,6,7,8,9,10,13,15,16,17,18,19	<0.00001
PRKG1	ENZ	cGMP-dependent protein kinase I	10	52,504,299	1,2,3,4,5,6,7,8,9,11,12,13,14,15,16,18,19	<0.00001
SERPINA1	ENZ	serpin peptidase inhibitor A 1	14	93,914,451	1,2,5,13,14,18	<0.00001
LARGE	ENZ	like-glycosyltransferase	22	31,999,063	3,4,5,7,9,13,14,15,16,17,18	0.00002
UST	ENZ	uronyl-2-sulfotransferase	6	149,110,157	1,2,6,15,16,17,18	0.00017
DAPK1	ENZ	death-associated protein kinase 1	9	89,302,576	1,2,3,5,13,14,15,16	0.00024
GALNTL4	ENZ	UDPNAc-α-D-galactosamine:polypeptide NAcgalactosaminyltransferase-like 4	11	11,248,999	1,2,3,4,5,9,15,16,19	0.00040
CAPN13	ENZ	calpain 13	2	30,799,141	1,2,3,5,6,8,15	0.00042
PDE4D	ENZ	cAMD-specific phosphodiesterase 4D	5	58,302,468	1,2,4,5,6,8,9,15,16,19	0.00055
AKAP13	ENZ	protein kinase A anchor protein 13	15	83,724,875	1,2,4,6,8,13,14,15,16	0.00062
ST6GALNAC3	ENZ	αNAcneuraminyl-2,3-β-galactosyl-1,3-NAcgalactosaminide α2,6-sialyltransferase 3	1	76,312,992	4,5,6,9,11,13,14,15,16,17,19	0.00063
SERPINA2	ENZ	serpin peptidase inhibitor A 2	14	93,900,404	1,2,13,14	0.00080
DGKB	ENZ	diacylglycerol kinase β	7	14,153,770	1,2,3,8,9,13,15,16,19	0.00138
FGF14	LIGAND	fibroblast growth factor 14	13	101,173,036	1,2,3,6,7,12,15,16,17,18	0.00001
GRM7	REC	metabotropic glutamate receptor 7	3	6,877,927	1,2,3,5,6,7,8,11,14,15,17	<0.00001
GPR39	REC	G protein-coupled receptor 39	2	132,891,086	1,2,8,14,15,18	0.00033
ABCC4	TRANSP	ATP-binding cassette C 4	13	94,470,090	1,2,5,10,15,16,17	<0.00001
SLC24A3	TRANSP	solute carrier family 24 member 3	20	19,141,290	3,4,5,7,8,11,14,15,16,18	0.00033
XKR4	TRANSP	XK family member 4	8	56,177,571	1,2,3,4,5,11,15,16,19	0.00091
SLC9A9	TRANSP	solute carrier family 9 member 9	3	144,466,754	1,2,3,4,6,8,9,13,15,18	0.00134

Columns list gene symbol, gene description, chromosome, basepair of gene start, samples (from this review) in which clustered nominally-positive SNPs are found and overall p value for this gene in this entire dataset, based on 100,000 Monte Carlo simulation trials. Note that genes that are identified strongly in a single sample will not be included on this list.

Table III

“Connectivity constellation” of phenotypes and disorders, each with sizable heritable influences

A. Phenotypes and disorders for which genome wide association data support involvement of “connectivity constellation” gene variants in human interindividual differences

Addiction (*vulnerability to substance dependence*) [3, 12, 33, 34, 37]
 Alzheimer’s disease (reanalyses of data in [4, 5, 300]
 Bipolar disorder (reanalyses of data in [1, 2])
 Brain volume (frontal lobe, Uhl *et al submitted*, reanalyses of data in [8, 117, 121, 154])
 Cognitive ability (Uhl *et al, submitted*)
 Memory [34, 36]
 Neuroticism (reanalysis of data in [228])
 Success at abstaining from smoking [301, 302]

B. Examples of other phenotypes and disorders that are strong candidates for roles for variants in “connectivity constellation” genes

Antisocial personality disorder
 Attention-deficit hyperactivity disorder
 Autism
 Depression
 HIV-associated dementia
 Functional recovery from stroke
 Functional recovery from traumatic brain injury
 Functional recovery from relapses of multiple sclerosis
 Parkinsonian dementia
 Schizophrenia (*especially negative symptoms*)