HHS Public Access

Author manuscript

Pharmacogenomics. Author manuscript; available in PMC 2016 June 12.

Published in final edited form as:

Pharmacogenomics. 2015 August ; 16(13): 1487–1498. doi:10.2217/pgs.15.91.

Dissecting ancestry genomic background in substance dependence genome-wide association studies

Renato Polimanti[1,2], **Can Yang**[1,3], **Hongyu Zhao**[3,4], and **Joel Gelernter**[*,1,2,4,5]

[1]Department of Psychiatry, Yale University School of Medicine, VA CT 116A2, 950 Campbell Avenue, West Haven, CT 06516, USA

[2]VA CT Healthcare Center, West Haven, CT 06516, USA

[3]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520-8034, USA

[4]Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA

[5]Department of Neurobiology, Yale University School of Medicine, New Haven, CT 06510, USA

Abstract

Aims—To understand the role of ancestral genomic background in substance dependence (SD) genome-wide association studies (GWAS), we analyzed population diversity at genetic loci associated with SD traits and evaluated its effect on GWAS outcomes.

Materials & methods—We investigated 24 genes with variants associated with SD by GWAS; and 82 loci with putative subordinate roles with respect to SD-associated genes.

Results—We observed high ancestry-related frequency differences in common functional alleles in GWAS relevant genes and their interactive partners. Common functional alleles with high frequency differences demonstrated significant effects on the GWAS outcomes.

[*]Author for correspondence: Tel.: +1 203 932 5711/3590, Fax: +1 203 937 3897, joel.gelernter@yale.edu.

Financial & competing interests disclosure

This study was supported by National Institutes of Health grants RC2 DA028909, R01 DA12690, R01 DA12849, R01 DA18432, R01 AA11330, R01 AA017535 and the VA Connecticut MIRECC. This work was also supported in part by the facilities of the Yale University Faculty of Arts and Sciences High Performance Computing Center. The publicly available dbGaP datasets used for the analyses were obtained from at http://www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000092.v1.p1 through dbGaP accession number phs000092.v1.p. Funding support for the Study of Addiction: Genetics and Environment (SAGE) was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01 HG004422). SAGE is one of the genome-wide association studies funded as part of the Gene Environment Association Studies (GENEVA) under GEI. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Support for collection of data sets and samples was provided by the Collaborative Study on the Genetics of Alcoholism (COGA; U10 AA008401), the Collaborative Genetic Study of Nicotine Dependence (COGEND; P01 CA089392) and the Family Study of Cocaine Dependence (FSCD; R01 DA013423). Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438), the National Institute on Alcohol Abuse and Alcoholism, the National Institute on Drug Abuse and the NIH contract 'High throughput genotyping for studying the genetic contributions to human disease' (HHSN268200782096C). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Conclusion—Population differences in SD GWAS outcomes seem not to be influenced by general variation across the genome, but by ancestry-related local haplotype structures at SD-associated loci.

Keywords

African–Americans; ethnicity; European–Americans; substance dependence

Substance dependence (SD) is an important problem in the US population, and for many others worldwide, with a substantial impact on medical health, quality of life, security and economics [1]. Genetic epidemiology studies have shown that drug dependence has high hereditability, highlighting a genetic component regulating risk for these phenotypes [2,3]. Recent genome-wide association studies (GWAS) of SD traits have identified numerous significant risk alleles across the human genome [4–7]. Although replication studies and functional investigations have, in some cases, confirmed the role of these alleles in the predisposition to drug dependence [8–11], certain risk alleles have failed in replication efforts in independent study populations, likely due in part to the presence of heterogeneity and to other several confounding factors, including differences in ancestry in the samples being studied [12,13] (as well as, in some cases, the original results being false positives). Indeed, in studies of other complex phenotypes, including SD, there are three situations in which ancestry confounding effects are seen: genes significantly associated with the phenotype in one ancestry group, but not in other ancestry groups; genes associated with the phenotype in more than one ancestry group, but with different groups presenting specific associated alleles; and alleles associated with the phenotype in different ancestry groups with ancestry-related differences in association strength. These confounders are largely attributable to the basic genetic differences that are present among human populations and that distinguish them. Indeed, large allele frequency differences (ΔF) are, of course, present among human populations, generating human phenotype diversity [14]. Many studies have investigated human genetic variation to understand the role that population demographic history or the environment plays in shaping the evolution of the human genome through natural selection [15,16]. Furthermore, some investigations have deepened the understanding of the role ancestry-related genetic variation plays in identifying the significant differences in genetic predisposition to health-related phenotypes among human populations [17–20]. Regarding drug dependence specifically, genetic studies have highlighted significant differences in genetic predisposition among populations, especially between European and African ancestries, which are the most studied [4,5]. Furthermore, epidemiologic studies indicated ethnic disparities in the prevalence of substance dependence traits among ancestry groups [21–23]. However, to our knowledge, no studies have endeavored to explain the genetic mechanisms at the basis of ancestry-related differences in genetic predisposition to drug dependence. To do this, we hypothesized that common alleles with large interethnic ΔFs and/or interethnic variation in rare variants occurrences (i.e., instances where multiple rare alleles are observed in a gene in only one population) in genes associated to substance dependence traits may have effects on risk alleles, partially explaining the association differences observed among human populations. These differences may also be attributable to the ancestry-related variation of genes encoding

proteins that interact with the protein products of other SD-associated genes via protein–protein interactions.

To gain insight into the genetic mechanisms at the basis of ancestry differences in the significant outcomes of GWAS of traits related to alcohol (AD), nicotine (ND) and opioid dependencies (OD), we analyzed the relationship of ancestral genomic background to GWAS results. Specifically, we investigated common and rare variation in genes with alleles significantly associated with SD traits based on GWAS and the genes that interact with them in four ancestry groups: African, admixed American (defined in the Materials and methods), Asian and European. Then, to verify the effect of this genetic variation on GWAS findings, we analyzed our GWAS discovery samples (Yale–Penn) on AD [4] and SAGE (Study of Addiction: Genetics and Environment) samples available in dbGAP (accession number phs000092.v1.p) [24]. Specifically, we tested whether variants with high allele ΔFs show significant effects on the associations between GWAS-relevant alleles (i.e., genetic variants identified by GWAS) and AD, explaining the differences observed between African–Americans (AA) and European–Americans (EA) – by evaluating whether the inclusion of ancestry-differentiated variants as covariates significantly modifies the associations observed. To exclude the possibility that the nonreplication of GWAS between AAs and EAs in AD GWAS is due to power differences attributable to allele ΔFs for GWAS-relevant alleles, we considered GWAS-relevant alleles that do not have large ΔFs between African and European ancestry subjects, and that showed genome-wide significance in one ancestry ($p < 5*10^{-8}$) and no significance in the other population ($p > 0.05$).

Materials & methods

Identification of relevant genes via GWAS

To find alleles potentially associated with SD traits, we searched Medline in November 2013 using combinations of the following keywords: ‘drugs’, ‘alcohol’, ‘nicotine’, ‘opioid’, ‘addiction’, ‘dependence’, ‘genome-wide association studies’ and ‘GWAS’. We identified 16 articles, in which at least a gene showed near-significant association with a trait related to at least one of the considered drug dependencies ($p < 5*10^{-7}$). A total of 24 genes (hereafter indicated as GWAS-relevant genes) were identified, as shown in Supplementary Table 1. We observed that the GWAS results are more consistent within the ancestral groups than between them, as anticipated. This trend is clearer for SD traits that were investigated by independent GWAS.

Identification of interacting genes/proteins

We used several interaction/pathway tools to identify genes and proteins interacting with the identified GWAS-relevant genes: STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [25], MINT (Molecular INTeraction Database) [26], KEGG (Kyoto Encyclopedia of Genes and Genomes) [27] and Pharm-GKB (The Pharmacogenomics Knowledgebase) [28]. For STRING analysis, we excluded text-mining from the active prediction methods and considered outcomes with the highest confidence (STRING score >0.900). For the MINT analysis, we considered Homo sapiens, as the reference organism,

and interactions with a score greater than 0.50. For KEGG and PharmGKB, we considered interacting genes/proteins those involved in direct interactions: either involvement in the same catalytic reactions or immediately preceding/subsequent catalytic reactions with proteins encoded by GWAS-relevant genes. With these criteria, we identified 82 interacting proteins. Supplementary Figures 1, 2 & 3 portray the interacting networks based on study of AD, ND and OD, respectively.

Population information & genomic data for ancestry analysis

Phase 1 of the 1000 Genomes (1KG) Project was used to obtain genotypic data. Specifically, we downloaded Variant Call Format files relevant to each GWAS-relevant or interacting gene. The 1KG Phase 1 consisted of 1092 individuals from 14 human populations. According to the 1KG project definition, they can be classified into four continental ancestry groups, as follows. The African group included: African ancestry in Southwest USA (ASW), Luhya in Webuye, Kenya (LWK) and Yoruba in Ibadan, Nigeria (YRI); the American group included: Colombian in Medellin, Colombia (CLM), Mexican ancestry in Los Angeles, CA (MXL) and Puerto Rican in Puerto Rico (PUR); the Asian group included: Han Chinese in Beijing, China (CHB), Han Chinese South (CHS) and Japanese in Tokyo, Japan (JPT); the European group included: Utah residents with northern and western European ancestry (CEU), Finnish from Finland (FIN), British from England and Scotland (GBR), Iberian populations in Spain (IBS) and Toscani in Italy (TSI). Since 1KG American populations reflect an admixture of Amerindian, African and European ancestries [29], we refer to them in the text as ‘admixed Americans’.

Functional annotation analysis

We used three different tools for functional annotation analysis. To distinguish between variants with functional effects (i.e., variants that affect gene regulation and/or protein activity) and variants with no regulatory or activity effects, we used VARIANT (VARIant ANalysis Tool) [30]. Based on the information obtained from VARIANT, we distinguished: variants likely to be nonfunctional (i.e., variants mapping to putative nonfunctional regions), variants with low potential for functional effect (i.e., variants with only a generic annotation: ‘located in a regulatory region annotated by Ensembl’) and variants with high potential for functional effect (i.e., variants with a specific annotation: ‘CpG island’, ‘miRNA target site’, ‘splice site’, ‘splice donor variant’, ‘RNA polymerase promoter’, ‘transcription factor binding site’, ‘splice acceptor variant’, ‘non-synonymous variant’ or ‘stop codon’). In the annotation analysis performed by VARIANT based on *in silico* evidence, we considered only information related to transcripts annotated in the Consensus Coding Sequence (CCDS) database. We also used rSNPBase (database for curated regulatory SNPs) and RegulomeDB [31,32] to further investigate variants potentially associated with epistatic effects on SD risk alleles. Both rSNPBase and RegulomeDB perform functional annotation on the basis of *in silico* and experimental evidences.

Identification of ancestry-related differences in common & rare variants

To identify the ancestry-related differences in GWAS-relevant genes and their interacting partners, we performed distinct analyses for variants with minor allele frequency (MAF) ≥1% (common variants) and variants with a MAF <1% (rare variants; RVs).

To identify the allele ΔF for common variants in the ancestry groups, we used the method proposed by Hofer and colleagues [14]. We chose this metric based on allele ΔF rather than others commonly used in population genetics (e.g., Wright's fixation index) in order to make our analysis clear also to nonexperts in the field. For each allele *i*, we computed the average allele frequency p_{ij} within each ancestry group *j*, as well as the difference with the average frequency computed over all other populations as $\Delta F = |p_{ij} - p_{-ij}|$, where p_{-ij} is the average frequency of allele *i* in all populations not belonging to the ancestry group *j*. A permutation test (n = 10,000) was performed to determine whether the MAF between the populations within ancestry groups and the rest of the populations was significantly different than expected by the chance.

For RVs, we analyzed the occurrences of functional RVs among all of the ancestry groups and used an equation to estimate ancestry differences in the occurrence of functional RVs. Using VARIANT annotation analysis, we distinguished functional variants from nonfunctional variants. Based on this information, for each gene we estimated the occurrence of functional RVs as the ratio of functional RVs to all observed variants. Specifically, for each gene *i*, we computed the average ratio r_{ij} within each ancestry group *j*, as well as the difference with the average ratio computed over all other populations as $r\Delta F = |r_{ij} - r_{-ij}|$, where r_{-ij} is the average ratio of gene *i* in all populations not belonging to the ancestry group *j*. A permutation test (n = 10,000) was performed to evaluate the significance of differences between the populations within an ancestry group compared with the rest of the populations.

Analysis in GWAS data of the effect of common variants with ΔF>0.10

To verify the effect of common variants with high ΔF on genome-wide significant associations, we analyzed two independent datasets used in a recent published AD GWAS [4]. Specifically, the datasets used in the present analysis comprise our Yale–Penn samples and SAGE samples, the latter obtained via dbGAP application. The Yale–Penn dataset includes 3318 AAs and 2379 EAs. The SAGE dataset includes 1195 AAs and 2528 EAs. Information about the genotyping, quality control and imputation analysis was published previously [4]. Considering the outcomes obtained by our previous AD GWAS, ten variants were found to be genome-wide significant ($p < 5*10^{-8}$) for AD symptom count in an ancestry group with nonsignificant results for the other one (Supplementary Table 2). Due to the strong linkage disequilibrium (LD) among some significant associations in this AD GWAS, we selected three independent variants from the total of ten: *PDLIM5* rs10031423, *ADH1B* rs1693457 and *ADH1C* rs6846835. Specifically, *PDLIM5* rs10031423 showed $R^2 = 0$ with respect to *ADH1B* rs1693457 and *ADH1C* rs6846835 in AAs and EAs, whereas *ADH1B* rs1693457 and *ADH1C* rs6846835 showed $R^2 = 0$ in EAs and $R^2 = 0.15$ in AAs. Slight differences exist for these variants in the present association results compared with the published AD GWAS (Supplementary Table 3), because in the present study we used the

original SAGE-imputed data, and symptom counts instead of substance dependence diagnosis to adjust for multiple dependencies. To estimate the effect of common variants on the genome-wide significant associations with Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV) symptom counts for AD, among the common variants investigated in the first part of the study, we selected those with ΔF>0.10 in African or European ancestries that are present in Yale–Penn and SAGE datasets (12,969 variants for African ancestry, and 8721 variants for European ancestry). We chose this threshold in order to exclude those variants with minimal allele ΔF among human populations. Then, performing separate analyses for AAs and EAs and for the Yale–Penn and SAGE datasets, we estimated the association of rs10031423, rs1693457 and rs6846835 with AD symptom counts in accordance with two different models using the R package genome-wide association/interaction analysis and rare variant analysis with family data (GWAF) to fit a generalized estimating equations model to correct for correlations among related individuals [33]. The first model ('A') tested the association of the imputed minor allele dosage with the DSM-IV symptom counts for AD considered as phenotype and using DSM-IV cocaine dependence symptom counts, DSM-IV OD symptom counts, DSM-IV ND symptom counts, sex, age and the first three ancestry principal components, as covariates. The second model ('B') performed the same analysis with the addition of a further covariate, a variant with ΔF>0.10. Then, we meta-analyzed the results obtained in the Yale–Penn and SAGE datasets for each ancestry group, applying the following equations:

$$\beta_{META} = \left(\beta_{Yale-Penn} * W_{Yale-Penn} + \beta_{SAGE} * W_{SAGE}\right) / \left(W_{Yale-Penn} + W_{SAGE}\right)$$

where β_{META}, $\beta_{Yale–Penn}$ and β_{SAGE} are the β values in the meta-analysis, Yale–Penn and SAGE datasets, respectively. The meta-analyzed p-values were calculated using METAL software [34]. To estimate the effect of each tested variant with ΔF >0.10, we calculated the z-score according to the following equation:

$$Z = \left(\left[\beta_{META1} - \beta_{META0}\right] - average\left[\beta_{META1} - \beta_{META0}\right]\right) / SD\left(\beta_{META1} - \beta_{META0}\right)$$

where META1 defined the meta-analyzed β values obtained from model B and META0 for the meta-analyzed parameter of model A. In accordance with Bonferroni correction for multiple testing, Z scores >|4.388| and |4.473| were considered significant in EAs and AAs, respectively.

Results

Considering the results of our gene–gene/protein–protein interaction's investigation, we constructed three interaction networks, one each for AD, ND and OD. Supplementary Figure 1 shows the AD interaction network. We observed only *ADH1B* and *ADH1C* interacting between themselves and with other common interacting partners. The other AD GWAS-relevant genes were related only to specific interacting partners (i.e., *CTBP2*, *HTR1A*, *GSS*, *KCNB2*) or were not related to any interacting partners (i.e., *THSD7B*, *SERINC2*, *NALCN*, *PKNOX2*, *DSCAML1*, *METAP1*, *KIAA0040*, *C15orf53*, *PDLIM5*). In the ND interaction network, we observed that *CHRNA3* and *CHRNA5* have common interacting proteins.

Conversely, *CHRNB3* and *ARHGAP10* did not interact (Supplementary Figure 2). In the OD interaction network, *NCK2* and *KCNG2* showed many interacting partners, having common interactions with *PARVA* and *KCNC1*, respectively (Supplementary Figure 3). In the OD interaction analysis, we saw no interactions for the *APBB2* gene.

The ΔF analysis of common variants (n = 51,079) indicated that allelic differences are greater in subjects of African ancestry (99.9th percentile of African ΔF = 0.690) than in those of Asian (99.9th percentile of Asian ΔF = 0.627), European (99.9th percentile of European ΔF = 0.422) or admixed-American ancestries (99.9th percentile of admixed-American ΔF = 0.281) (Figure 1). Considering each SD diagnosis separately, we observed high ΔF values in variants potentially associated with a large functional effect in GWAS-relevant genes and their interacting partners in AD, ND and OD (Supplementary Tables 4, 5 & 6, respectively). Table 1 reports the common variants with top ΔF that are also potentially associated with regulatory functions observed in each GWAS-relevant gene.

In the analysis of RVs, we observed that the distribution of r values differed significantly between those of African ancestry and those of Asian ($p_{Bonferroni}$ <0.05), European ($p_{Bonferroni}$ <0.001) and admixed-American ancestries ($p_{Bonferroni}$ <0.001) (Figure 2).

Considering the rΔF values for each analyzed gene, we observed only two significant outcomes for African ancestry: *ZEB1* (Africa rΔF = 0.186; p = 0.005), and *HDAC1* (Africa rΔF = 0.316; p = 0.004). These are interactive partners; we did not observe any significant difference for genome-wide significant genes. Supplementary Table 6 shows the top 1% of rΔF values for each ancestry group. Although no ancestry-related differences were observed for the occurrence of RVs in GWAS-relevant genes, high r values in GWAS-relevant genes can be seen in the overall 1KG population ($r_{overall}$ = 0.529 ± 0.215; Table 2).

As mentioned above, we hypothesized that allele ΔFs in variants with functional impact may explain the ancestry-related differences observed in drug dependence GWAS. To verify this hypothesis, we analyzed two different datasets with AA and EA samples used in our recently published AD GWAS, as described above. In this study, ten GWAS-significant variants were observed associated with AD symptom counts in one ancestry group but not in the other (Supplementary Table 2 & 3) – i.e., in either AAs or EAs but not both. These genetic variants are located in *ADH1B*, *ADH1C* and *PDLIM5*. In *PDLIM5*, we observed large ΔF values for African ancestry, but the *PDLIM5* significant variant (i.e., rs10031423) showed low ΔF values in all ancestry groups (Supplementary Figure 4). In *ADH1B*, we observed large ΔF values for Asian ancestry (i.e., ΔF>0.6), but low ΔF values of *ADH1B* variants for European and African ancestries (Supplementary Figure 5). In *ADH1C*, extreme ΔF values were observed for Asian ancestry (i.e., ΔF>0.6), and most of the Asian ΔF peaks are genome-wide significant variants for AD symptom counts (Supplementary Figure 6). Both *ADH1B* and *ADH1C* ΔF top values are included in the top 0.1% of the distribution of Asian ΔFs. To check whether the high ΔF values of *ADH1B* and *ADH1C* in Asians and *PDLIM5* in Africans are due to human demographic history or to natural selection processes, we verified the integrated Haplotype Score of these loci using the Haplotter application [35]. Significant signatures of natural selection are confirmed in Asians for

ADH1B (p = 0.011) and *ADH1C* (p = 0.009) (Supplementary Figure 7), while a nonsignificant outcome was observed for *PDLIM5*.

As noted above, due to the high LD present among the ten genome-wide significant variants considered here, we chose three independent variants (i.e., *PDLIM5* rs10031423, *ADH1B* rs1693457 and *ADH1C* rs6846835).

In EAs *PDLIM5* rs10031423 showed a significant association with AD symptom counts. In EAs, we identified 59 variants with significant effect on this association. Except for *ADH1B* rs1229984 (i.e., a genome-wide significant variant for AD symptom counts in EAs), they are all located in the *PDLIM5* gene region, and 36 of them have an $r^2<0.2$ with rs10031423 both in the Yale–Penn and SAGE datasets (Supplementary Table 7). Among these non LD variants, there are *ADH1B* rs1229984 and *PDLIM5* rs2452594. This latter may have functional impact on the *PDLIM5* function (RegulomeDB score = 2b, i.e., TF binding + any motif + DNase Footprint + DNase peak) and this variant reduces the association of rs10031423 with AD symptom count (meta-analyzed p-value from 7.3510^{-6} to $5.75*10^{-4}$). In AAs, we identified 88 variants with significant z-scores (Supplementary Table 8). Except for *KCND2* rs12333476, they are all located on chromosome 4, as rs10031423. Among these variants, rs6853490 (z = −20.085) showed that a high African ΔF, may play a role in *PDLIM5* regulation (RegulomeDB score = 2b, i.e., TF binding + any motif + DNase Footprint + DNase peak), and increase the association of rs10031423 with AD symptom count in AAs (meta-analyzed p-value from 0.276 to 0.097). Regarding *ADH1B* rs1693457 (significant in AAs), through our analysis of AAs, we identified 27 variants with significant z-scores. They are all located in the surrounding regions of rs1693457 (Supplementary Table 9). Among these, 18 variants showed r^2 <0.2 with rs10031423 both in the Yale–Penn and SAGE datasets. In accordance with RegulomeDB, none of these seem to have functional impact on *PDLIM5* gene function. However, rs12639887 (z = 16.997) showed the highest African ΔF among these (Africa ΔF = 0.39). Both the VARIANT tool and rSNPbase identified it as regulatory SNP, and rs12639887 reduces the association significance of rs1693457 with AD symptom counts (meta-analyzed p-value from $3.40*10^{-9}$ to $1.10*10^{-5}$). In EAs, we identified 106 variants with significant z-scores. All of these are located in the regions surrounding rs10031423 (Supplementary Table 10). According to RegulomeDB, none of them showed notable evidence of functional regulatory effects, and the strongest increase of significance was observed for rs1229984 (z = 13.369; meta-analyzed p-value from 0.726 to 0.189). In AAs, the analysis of *ADH1C* rs6846835 (significant in AAs) revealed 15 variants with significant z-scores. All of these are located on chromosome 4, as rs6846835 (Supplementary Table 11). Among them, only ten variants present an r^2 with rs6846835<0.2. Considering the non LD variants, rs12639887 showed significant effects on both *ADH1B* rs1693457 and *ADH1C* rs6846835. In the analysis of *ADH1C* rs6846835, rs12639887 showed a strong effect on the association with AD symptom count (z = −6.015; meta-analyzed p-value from $5.65*10^{-9}$ to 0.237). Stratifying the AA and EA samples for rs12639887 genotype, we observed the same trend in both ancestry groups for the association between *ADH1C* rs6846835 and AD symptom count (Supplementary Table 12). Regarding the analysis of *ADH1C* rs6846835 in EAs, we observed 72 variants with significant z-scores. They are all located on chromosome 4 (Supplementary Table 13).

Among these variants, only rs4147541 showed evidence of regulatory function (RegulomeDB score = 3a, i.e., TF binding + any motif + DNase peak) and an effect that increased the association significance of *ADH1C* rs6846835 with AD symptom counts in EAs (z = 39.468; meta-analyzed p-value from 0.950 to 0.041).

Considering the *PDLIM5*, *ADH1B* and *ADH1C* analyses, we observed that the most significant findings are related to common functional alleles with high African or European ΔFs located in the regions surrounding genome-wide significant variants. To analyze the local haplotype structure of the *PDLIM5-ADH1B-ADH1C* region (chr4:95,379,741-100,274,157), we used 1KG Phase 1 data from the ASW populations to investigate AAs and from the CEU to investigate EAs, considering those ancestry variants with relevant effects on *PDLIM5* rs10031423, *ADH1B* rs1693457 and *ADH1C* rs6846835 associations with AD outcomes. In accordance with the method of Gabriel and colleagues [36], we observed 14 and 11 haplotype blocks in ASW and CEU, respectively (Supplementary Figures 8 & 9). In ASW, *PDLIM5* included ten haplotype blocks. The ancestry variants with effects on *PDLIM5* genome-wide significant association are located in the haplotype blocks closer to rs10031423. Regarding the *ADH1B* rs1693457 genome-wide significant association, the variants with relevant effects are mainly located within *METAP1* and *ADH1B* haplotypes. For the *ADH1C* rs6846835 genome-wide significant association, most variants with relevant effects are located in *METAP1* and *ADH1C* haplotypes, and they showed relevant effects also on *ADH1B* association. In CEU, the *PDLIM5* gene region includes eight haplotype blocks, and the genome-wide significant variant is located in last (i.e., 3′-most) haplotype block (i.e., Block 8). Most variants with significant effects on the genome-wide association are located in Block 1 and Block 2; single haplotypes are present for each of the remaining genes (i.e., *METAP1*, *ADH1B* and *ADH1C*). The variants with relevant effects on the *ADH1B* rs1693457 genome-wide significant association are mainly located in *METAP1* and *ADH1B* haplotypes, while the relevant findings related to *ADH1C* rs6846835 genome-wide association are located in an *ADH1C* haplotype and most of them showed significant effects also on the *ADH1B* rs1693457 genome-wide association.

Discussion

GWAS for drug dependence traits have identified several significant risk alleles, but replication studies in different ancestry groups often fail to reproduce these outcomes [12,13]. In the present study, we tested the effect of genetic background (variants located in the surrounding regions or in functional partners) on the significantly-associated variants. Our results provided significant evidence that local ancestry differences can partially explain the ancestry difference observed in GWAS. Specifically, we observed that when adjusting for these local ancestry-related variants, the differences between African–Americans and European–Americans tend to diminish.

Our investigation of gene–gene/protein–protein interactions highlighted multiple interactions of GWAS-relevant genes with other genome-wide relevant loci or other genes involved in the same molecular pathways. Genetic variations among these interacting partners may have an effect on the predisposition to drug dependence. However, we also found that certain genome-wide relevant genes were not integrated into the network. This

suggests that there is missing information in the experimental evidence pertaining to gene–gene/protein–protein interaction in the predisposition to drug dependence – a result that is hardly surprising. Other approaches have been used to integrate drug dependence GWAS data with the human protein interactome [37]. Comparing these approaches, we observed some common features and some differences, which are likely attributable to the fact that the present network analysis is based only on experimental evidence independent from genetic data. Therefore, by analyzing our interaction networks, we have confidently investigated human variation in genes that directly interact with the identified GWAS-relevant genes.

Our analysis of the ancestry-related allele ΔF of common variants indicated that African ancestry is the most divergent group, in agreement with the current knowledge about genetics of human demographic history [38,39]. Conversely, the admixed-American group showed lower allele ΔFs than the other ancestry groups. It is highly likely that this is due to the complex admixture of African, Amerindian and European ancestries present in these populations [29]. When considering the observed ancestry-related differences in common variants in both GWAS-relevant genes and their interactive partners, we detected a number of variants with noteworthy allele ΔFs which are also potentially involved in regulatory mechanisms, suggesting their role in the diverse outcomes observed in drug dependence GWAS-among human populations. Regarding the rare variants analysis, we observed that there are differences among ancestry groups for the occurrence of regulatory rare variants. These differences are significant when we considered the entire genome, but, when we focused on specific gene regions, they are not statistically significant. The result observed, considering genomic differences, is in accordance with recent evidence that has indicated differences in the stratification of functional RVs among human populations [40,41]. As is the case for common variants, the differences in RVs between African and non-African populations may be attributed to human demographic history. However, the specific-gene analysis suggests that, in drug dependence GWAS, confounding effect due to rare variation is not linked to the ancestry of the investigated populations. Nevertheless, we cannot exclude that RVs could have a population-specific effect on an SD-phenotype, since their effect could depend on a permissive epistatic genetic background present in one population group but not another. Moreover, the high rates of functional RVs in both GWAS-relevant genes and their interactive partners also indicate that rare variation may strongly confound outcomes of drug dependence GWAS, as postulated for other complex phenotypes related to drug response [42,43].

The ΔF analysis of genes with AD GWAS ancestry-specific significant alleles showed the strong diversity of *ADH1B* and *ADH1C* variation between Asian and non-Asian populations. Previous studies indicated that *ADH1B-ADH1C* genetic diversity between Asians and non-Asians is due in part to a selective pressure [44–46], suggesting a potential effect on ADH-associated traits, such as AD. A recent GWAS on AD traits in a Chinese study population indicated that *ADH1B* and *ADH1C* did not contain alleles significantly associated with AD traits, but *ALDH2*, another alcohol metabolism-related gene and an interactive partner of both *ADH1B* and *ADH1C*, showed AD risk alleles consistent with many prior observations [47]. This difference between non-Asian ancestries (i.e., AAs and EAs: genome-wide

significance of ADH genes for AD traits) and Asian one (non-genome-wide significance of ADH genes for AD traits) may ultimately be attributable to the effect of natural selection on genetic variation of ADH-gene cluster.

Our analysis on AD GWAS datasets indicated that multiple variants, located in the surrounding regions of genome-wide significant alleles, have effects modifying the associations between significant alleles and AD symptom counts. Most of them showed high allele ΔFs in African or European ancestries and are potentially involved in regulatory mechanisms. Among them, in AAs, rs12639887 showed significant effects both on *ADH1B* rs1693457 and *ADH1C* rs6846835 association with AD symptom counts. This variant is located in a *PDLIM5* intronic region, approximately 5 Mb downstream with respect to *ADH1B* and *ADH1C*. Both VARIANT and rSNPbase defined it as a regulatory SNP. In particular, VARIANT describes it as being located in an H3K36me3 region, whereas rSNPbase designates it as being involved in proximal, distal and RNA binding protein-mediated regulations. Unfortunately, to best of our knowledge, no information is currently available regarding the role of rs12639887 in *ADH1B* and *ADH1C* expression or methylation status in brain tissues. However, according to annotation analysis, we can hypothesize that rs12639887 plays a role in determining the differences observed between AAs and EAs in AD GWAS outcomes of *ADH1B* and *ADH1C*. In addition to this single case, our data generally revealed that ancestry-related variability in AD GWAS-relevant genes and their surrounding regions can explain some differences in the outcomes of drug dependence GWAS among human populations. Consequently, the differences among ancestry groups in terms of AD genetic predisposition seem not to be influenced by general variation across the genome – that is, these differences may reside in a collection of identifiable ancestry-specific risk alleles. Accordingly, the analysis of local haplotype structure of the *PDLIM5-ADH1B-ADH1C* region confirmed how population differences in haplotype structure can affect GWAS findings.

Conclusion

Our data furnish novel information not only about the relationship between AD and ancestry, but also about the interactions between GWAS-relevant alleles and cis-regulatory variants. Further analyses based on an evolutionary approach may provide other relevant knowledge about the predisposition of AD.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Henry R Kranzler for assisting with data collection and manuscript review. The authors are also grateful to the research groups of the 1000 Genomes Project and the SAGE project for their publicly available data.

References

Papers of special note have been highlighted as:

• of interest;

•• of considerable interest

1. Office of National Drug Control Policy. The Economic Costs of Drug Abuse in the United States, 1992–2002. Executive Office of the President; Washington, DC, USA: 2004. www.ncjrs.gov/ondcppubs/publications/pdf/economic_costs

2. Gelernter J, Kranzler HR. Genetics of alcohol dependence. Hum Genet. 2009; 126(1):91–99. [PubMed: 19533172]

3. Goldman D, Oroszi G, Ducci F. The genetics of addictions: uncovering the genes. Nat Rev Genet. 2005; 6(7):521–532. [PubMed: 15995696]

4••. Gelernter J, Kranzler HR, Sherva R, et al. Genome-wide association study of alcohol dependence: significant findings in African– and European–Americans including novel risk loci. Mol Psychiatry. 2014; 19(1):41–49. GWAS of alcohol dependence used to test the effect of ancestry genomic background in the present study. [PubMed: 24166409]

5. Gelernter J, Kranzler HR, Sherva R, et al. Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. Biol Psychiatry. 2014; 76(1):66–74. [PubMed: 24143882]

6. Gelernter J, Sherva R, Koesterer R, et al. Genome-wide association study of cocaine dependence and related traits: *FAM53B* identified as a risk gene. Mol Psychiatry. 2014; 19(6):717–723. [PubMed: 23958962]

7. Gelernter JKH, Kranzler HR, Sherva R, et al. Genome wide association study of nicotine dependence in American populations: identification of novel risk loci in both African– and European–Americans. Biol Psychiatry. 2014; 77(5):493–503. [PubMed: 25555482]

8. Johnson C, Drgon T, Walther D, Uhl GR. Genomic regions identified by overlapping clusters of nominally-positive SNPs from genome-wide studies of alcohol and illegal substance dependence. PLoS ONE. 2011; 6(7):e19210. [PubMed: 21818250]

9. Derringer J, Krueger RF, Dick DM, et al. The aggregate effect of dopamine genes on dependence symptoms among cocaine users: cross-validation of a candidate system scoring approach. Behav Genet. 2012; 42(4):626–635. [PubMed: 22358648]

10. Park BL, Kim JW, Cheong HS, et al. Extended genetic effects of ADH cluster genes on the risk of alcohol dependence: from GWAS to replication. Hum Genet. 2013; 132(6):657–668. [PubMed: 23456092]

11. Biernacka JM, Geske JR, Schneekloth TD, et al. Replication of genome wide association studies of alcohol dependence: support for association with variation in *ADH1C*. PLoS ONE. 2013; 8(3):e58798. [PubMed: 23516558]

12. Verweij KJ, Zietsch BP, Liu JZ, et al. No association of candidate genes with cannabis use in a large sample of Australian twin families. Addict Biol. 2012; 17(3):687–690. [PubMed: 21507154]

13. Enoch MA. Genetic influences on the development of alcoholism. Curr Psychiatry Rep. 2013; 15(11):412. [PubMed: 24091936]

14•. Hofer T, Ray N, Wegmann D, Excoffier L. Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet. 2009; 73(1):95–108. Method used in the present study to identify the most differentiated loci among ancestry groups. [PubMed: 19040659]

15. Elhaik E, Tatarinova T, Chebotarev D, et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nat Commun. 2014; 5:3513. [PubMed: 24781250]

16. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nat Rev Genet. 2014; 15(6):379–393. [PubMed: 24776769]

17. Polimanti R, Piacentini S, Manfellotto D, Fuciarelli M. Human genetic variation of *CYP450* superfamily: analysis of functional diversity in worldwide populations. Pharmacogenomics. 2012; 13(16):1951–1960. [PubMed: 23215887]

18. Chanock SJ. A twist on admixture mapping. Nat Genet. 2011; 43(3):178–179. [PubMed: 21350496]

19. Gelernter J. *SLC6A4* polymorphism, population genetics, and psychiatric traits. Hum Genet. 2014; 133(4):459–461. [PubMed: 24385047]
20. Bentley AR, Chen G, Shriner D, et al. Gene-based sequencing identifies lipid-influencing variants with ethnicity-specific effects in African Americans. PLoS Genet. 2014; 10(3):e1004190. [PubMed: 24603370]
21. Chartier KG, Scott DM, Wall TL, et al. Framing ethnic variations in alcohol outcomes from biological pathways to neighborhood context. Alcohol Clin Exp Res. 2014; 38(3):611–618. [PubMed: 24483624]
22. Mack KA. Centers for Disease C and Prevention. Drug-induced deaths – United States, 1999–2010. MMWR Surveill Summ. 2013; 62(Suppl 3):161–163. [PubMed: 24264508]
23. Rayens MK, Hahn EJ, Fernander A, Okoli CT. Racially classified social group differences in cigarette smoking, nicotine dependence, and readiness to quit. J Addict Nurs. 2013; 24(2):71–81. [PubMed: 24621484]
24••. Bierut LJ, Agrawal A, Bucholz KK, et al. A genome-wide association study of alcohol dependence. Proc Natl Acad Sci USA. 2010; 107(11):5082–5087. GWAS of alcohol dependence used to test the effect of ancestry genomic background in the present study. [PubMed: 20202923]
25. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013; 41:D808–D815. [PubMed: 23203871]
26. Licata L, Briganti L, Peluso D, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012; 40:D857–D861. [PubMed: 22096227]
27. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28(1):27–30. [PubMed: 10592173]
28. Thorn CF, Klein TE, Altman RB. Pharmacogenomics and bioinformatics: PharmGKB. Pharmacogenomics. 2010; 11(4):501–505. [PubMed: 20350130]
29. Gravel S, Zakharia F, Moreno-Estrada A, et al. Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet. 2013; 9(12):e1004023. [PubMed: 24385924]
30. Medina I, De Maria A, Bleda M, et al. VARIANT: command line, web service and web interface for fast and accurate functional characterization of variants found by nextgeneration sequencing. Nucleic Acids Res. 2012; 40:W54–W58. [PubMed: 22693211]
31. Guo L, Du Y, Chang S, Zhang K, Wang J. rSNPBase: a database for curated regulatory SNPs. Nucleic Acids Res. 2014; 42:D1033–D1039. [PubMed: 24285297]
32. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012; 22(9):1790–1797. [PubMed: 22955989]
33. Chen MH, Yang Q. GWAF: an R package for genome-wide association analyses with family data. Bioinformatics. 2010; 26(4):580–581. [PubMed: 20040588]
34. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26(17):2190–2191. [PubMed: 20616382]
35. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006; 4(3):e72. [PubMed: 16494531]
36. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science. 2002; 296(5576):2225–2229. [PubMed: 12029063]
37. Han S, Yang BZ, Kranzler HR, et al. Integrating GWASs and human protein interaction networks identifies a gene subnetwork underlying alcohol dependence. Am J Hum Genet. 2013; 93(6): 1027–1034. [PubMed: 24268660]
38. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008; 319(5866):1100–1104. [PubMed: 18292342]
39. Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet. 2008; 9:403–433. [PubMed: 18593304]
40. Moore CB, Wallace JR, Frase AT, Pendergrass SA, Ritchie MD. Using BioBin to explore rare variant population stratification. Pac Symp Biocomput. 2013:332–343. [PubMed: 23424138]

41. Moore CB, Wallace JR, Wolfe DJ, et al. Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. PLoS Genet. 2013; 9(12):e1003959. [PubMed: 24385916]

42. Polimanti R, Iorio A, Piacentini S, Manfellotto D, Fuciarelli M. Human pharmacogenomic variation of antihypertensive drugs: from population genetics to personalized medicine. Pharmacogenomics. 2014; 15(2):157–167. [PubMed: 24444406]

43. Drake KA, Torgerson DG, Gignoux CR, et al. A genome-wide association study of bronchodilator response in Latinos implicates rare variants. J Allergy Clin Immunol. 2014; 133(2):370–378. [PubMed: 23992748]

44. Suo C, Xu H, Khor CC, et al. Natural positive selection and north-south genetic diversity in East Asia. Eur J Hum Genet. 2012; 20(1):102–110. [PubMed: 21792231]

45. Evsyukov A, Ivanov D. Selection variability for Arg48His in alcohol dehydrogenase ADH1B among Asian populations. Hum Biol. 2013; 85(4):569–577. [PubMed: 25019189]

46. Li H, Gu S, Han Y, et al. Diversification of the *ADH1B* gene during expansion of modern humans. Ann Hum Genet. 2011; 75(4):497–507. [PubMed: 21592108]

47. Quillen EE, Chen XD, Almasy L, et al. *ALDH2* is associated to alcohol dependence and is the major genetic determinant of “daily maximum drinks” in a GWAS study of an isolated rural chinese sample. Am J Med Genet B Neuropsychiatr Genet. 2014; 165B(2):103–110. [PubMed: 24277619]

Executive summary

Background

- Genome-wide association studies (GWAS) of substance dependence traits have identified numerous significant risk alleles, but replication studies on different ancestry groups have often failed to reproduce these outcomes.

Aim

- To understand the role of ancestral genomic background in substance dependence (SD) GWAS, we analyzed population diversity at genetic loci associated with SD traits and evaluated its effect on the significant outcomes of GWAS in different ancestry groups.

Results

- We observed high ancestry-related frequency differences in common functional alleles in GWAS-relevant genes and their interactive partners.
- We also identified significant ancestry differences in the genome-wide occurrence of regulatory rare variants between African and non-African population, but gene-specific analysis confirmed few significant ancestry differences.
- The analysis of SD GWAS datasets indicated that common functional alleles with high African or European frequency differences have significant effects on the outcomes of genome-wide significant loci observed in African– and European–Americans.

Conclusion

- We observed that population differences in SD GWAS outcomes seem not to be influenced by general variation across the genome, but rather are due to ancestry-related local haplotype structures at SD-associated loci.

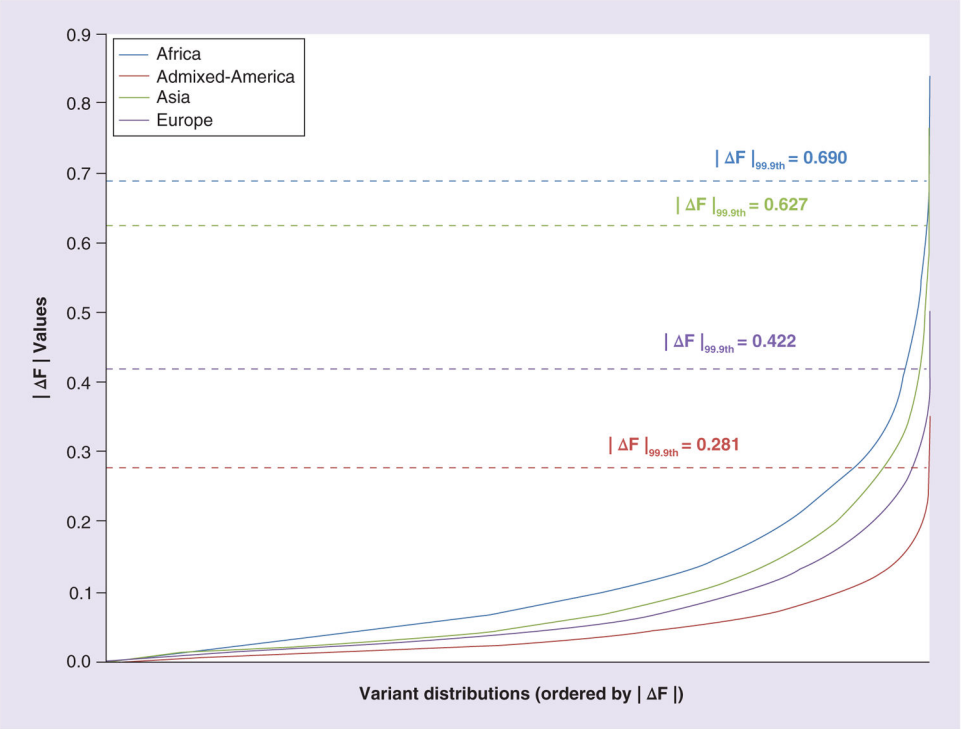

Figure 1. Frequency differences values of common variants among ancestry groups

Each ancestry-specific line is made up by symbols that represent single variants. For color figures, see online at: http://www.futuremedicine.com/doi/full/10.2217/PGS.15.91

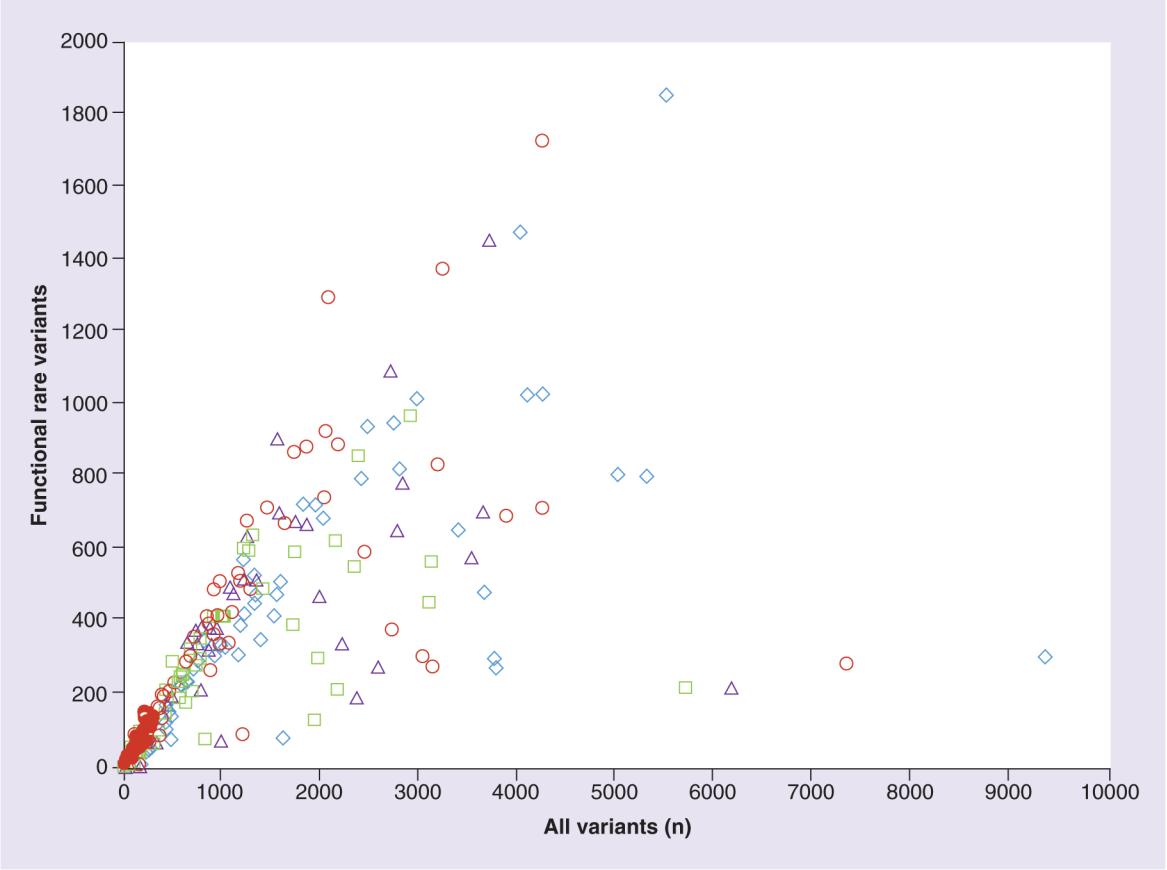

Figure 2. Occurrence of functional rare variants in drug-dependence genes and their interactive partners among ancestry groups (Africa: triangles; admixed America: red circles; Asia: green square; Europe: blue diamonds)

Each symbol represents a gene. Trend lines for each ancestry groups are also reported (Africa: $r^2 = 0.92$, $p < 0.001$; admixed America: $r^2 = 0.90$; $p < 0.001$; Asia $r^2 = 0.90$; $p < 0.001$; and Europe: $r^2 = 0.90$; $p < 0.001$).

Table 1

Top frequency difference value of common variants potentially associated with functional effects for each gene.

Gene	Addiction	Id	Annotation	Ancestry	IΔFI	p-value
ADH1B	Alcohol	rs1229984	non_synonymous_codon	Asia	0.695	0.001*
ADH1C	Alcohol	rs1693426	splice_region_variant	Asia	0.143	0.020
C15orF53	Alcohol	rs12595568	miRNA_target_site	Asia	0.293	0.002*
CTBP2	Alcohol	rs1041191	CpG_island	Asia	0.430	0.001*
DSCAML1	Alcohol	rs3741283	CpG_island	Asia	0.365	0.005*
GSS	Alcohol	rs17092185	RNA_polymerase_promoter	Africa	0.195	0.003*
HTR1A	Alcohol	rs6294	non_synonymous_codon	Africa	0.298	0.003*
KCNB2	Alcohol	rs34259157	non_synonymous_codon	Africa	0.068	0.001*
KIAA0040	Alcohol	rs2272785	RNA_polymerase_promoter	Africa	0.309	0.001*
METAP1	Alcohol	rs2167981	RNA_polymerase_promoter	Europe	0.279	0.004*
NALCN	Alcohol	rs66537139	splice_region_variant	Africa	0.179	0.187
PDLIM5	Alcohol	rs12294	miRNA_target_site	Africa	0.314	0.001*
SERINC2	Alcohol	rs7417775	RNA_polymerase_promoter	Africa	0.429	0.002*
THSD7B	Alcohol	rs5012365	miRNA_target_site	Africa	0.339	0.002*
ARHGAP10	Nicotine	rs11099669	RNA_polymerase_promoter	Africa	0.688	0.002*
CHRNA3	Nicotine	rs71541945	miRNA_target_site	Asia	0.285	0.002*
CHRNA5	Nicotine	rs16969968	non_synonymous_codon	Europe	0.262	0.001*
CHRNB3	Nicotine	rs11264221	RNA_polymerase_promoter	Africa	0.257	0.004*
APBB2	Opioid	rs4861358	non_synonymous_codon	Africa	0.483	0.003*
KCNC1	Opioid	rs16934718	CpG_island	Asia	0.125	0.002*
KCNG2	Opioid	rs35235239	CpG_island	Africa	0.466	0.001*
NCK2	Opioid	rs746437	RNA_polymerase_promoter	Africa	0.308	0.001*
PARVA	Opioid	rs11022331	RNA_polymerase_promoter	Africa	0.543	0.003*

Author Manuscript

Author Manuscript

Author Manuscript

* Significant p-values ($p < 0.01$).

Table 2

Occurrence of functional rare variants in significant genes. $r_{overall}$ is the ratio between functional rare variants and all variants observed in a gene considering all ancestry groups together.

Gene	Addiction	All variant (n)	Functional rare variant (n)	$r_{overall}$
ADH1B	Alcohol	202	125	0.619
ADH1C	Alcohol	228	22	0.096
C15orF53	Alcohol	84	60	0.714
CTBP2	Alcohol	2956	1998	0.676
DSCAML1	Alcohol	5910	3898	0.660
GSS	Alcohol	275	203	0.738
HTR1A	Alcohol	27	20	0.741
KCNB2	Alcohol	5628	743	0.132
KIAA0040	Alcohol	505	348	0.689
METAP1	Alcohol	865	577	0.667
NALCN	Alcohol	5428	883	0.163
PDLIM5	Alcohol	2850	1990	0.698
SERINC2	Alcohol	359	215	0.599
THSD7B	Alcohol	13590	868	0.064
ARHGAP10	Nicotine	4542	2696	0.594
CHRNA3	Nicotine	390	245	0.628
CHRNA5	Nicotine	392	142	0.362
CHRNB3	Nicotine	648	277	0.427
APBB2	Opioid	6085	2774	0.456
KCNC1	Opioid	583	412	0.707
KCNG2	Opioid	580	337	0.581
NCK2	Opioid	2133	1473	0.691
PARVA	Opioid	2260	1464	0.648