2017

Integrating Human Population Genetics And Genomics To Elucidate The Etiology Of Brain Disorders

Arvis Sulovari
University of Vermont

INTEGRATING HUMAN POPULATION GENETICS AND GENOMICS TO ELUCIDATE THE ETIOLOGY OF BRAIN DISORDERS

A Dissertation Presented

by

Arvis Sulovari

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
Specializing in Cellular, Molecular, and Biomedical Sciences

October, 2017

Defense Date: June 14, 2017
Dissertation Examination Committee:

Dawei Li, Ph.D., Advisor
James J. Hudziak, M.D., Chairperson
Russell P. Tracy, Ph.D.
Jeffrey P. Bond, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of Graduate College

ABSTRACT

Brain disorders present a significant burden on affected individuals, their families and society at large. Existing diagnostic tests suffer from a lack of genetic biomarkers, particularly for substance use disorders, such as alcohol dependence (AD). Numerous studies have demonstrated that AD has a genetic heritability of 40-60%. The existing genetics literature of AD has primarily focused on linkage analyses in small family cohorts and more recently on genome-wide association analyses (GWAS) in large case-control cohorts, fueled by rapid advances in next generation sequencing (NGS). Numerous AD-associated genomic variations are present at a common frequency in the general population, making these variants of public health significance. However, known AD-associated variants explain only a fraction of the expected heritability. In this dissertation, we demonstrate that systems biology applications that integrate evolutionary genomics, rare variants and structural variation can dissect the genetic architecture of AD and elucidate its heritability.

We identified several complex human diseases, including AD and other brain disorders, as potential targets of natural selection forces in diverse world populations. Further evidence of natural selection forces affecting AD was revealed when we identified an association between eye color, a trait under strong selection, and AD. These findings provide strong support for conducting GWAS on brain disorder phenotypes. However, with the ever-increasing abundance of rare genomic variants and large cohorts of multi-ethnic samples, population stratification becomes a serious confounding factor for GWAS. To address this problem, we designed a novel approach to identify ancestry informative single nucleotide polymorphisms (SNPs) for population stratification adjustment in association analyses. Furthermore, to leverage untyped variants from genotyping arrays – particularly rare variants – for GWAS and meta-analysis through rapid imputation, we designed a tool that converts genotype definitions across various array platforms.

To further elucidate the genetic heritability of brain disorders, we designed approaches aimed at identifying Copy Number Variations (CNVs) and viral insertions into the human genome. We conducted the first CNV-based whole genome meta-analysis for AD. We also designed an integrated approach to estimate the sensitivity of NGS-based methods of viral insertion detection. For the first time in the literature, we identified herpesvirus in NGS data from an Alzheimer's disease brain sample.

The work in this dissertation represents a three-faceted advance in our understanding of brain disease etiology: 1) evolutionary genomic insights, 2) novel resources and tools to leverage rare variants, and 3) the discovery of disease-associated structural genomic aberrations. Our findings have broad implications on the genetics of complex human disease and hold promise for delivering clinically useful knowledge and resources.

CITATIONS

Material from this dissertation has been published in the following form:

Sulovari, A., Zhu, Z., Li, D.., (2017). Genome-wide Meta-analysis of Copy Number Variations with Alcohol Dependence. The Pharmacogenomics Journal, (00):1-8

Sulovari, A., Chen, Y., Hudziak, J.J., Li, D.., (2017). Atlas of Human Diseases Influenced by Genetic Variants with Extreme Allele Frequency Differences. Human Genetics, 136(1):39-54.

Sulovari, A., Kranzler, H. R., Farrer, L. A., Gelernter, J., Li, D.., (2015). Eye color: A potential Indicator of Alcohol Dependence Risk in European Americans. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 168B(5):347-353.

Sulovari, A., Kranzler, H. R., Farrer, L. A., Gelernter, J., Li, D.., (2015). Further Analyses Support the Association Between Light Eye Color and Alcohol Dependence. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 168(8):757-760.

Sulovari, A., and Li, D.., (2014). GACT: a Genome Build and Allele Definition Conversion Tool for SNP Imputation and Meta-analysis in Genetic Association Studies. BMC Genomics, 15:610

DEDICATION

This work is dedicated to my parents and my sister, who have always supported and loved me unconditionally, despite us being several thousand miles apart; to my life partner, Zoë, for making everything brighter and making me a better person. From the bottom of my heart: Thank you!

ACKOWLEDGEMENTS

I have many people I would like to acknowledge. First and foremost, I would like to thank my advisor Dr. Dawei Li, who took me into his lab as his first graduate student. Thank you for accepting me into your lab and for pushing me to improve at both a personal and professional level. I would also like to thank my lab mates, past and present: Xun Chen, Guangchen Liu, Michael Mariani, Jason Kost, David Miserak, and Acadia Moeyersoms. Our friendship will have a lasting positive impact on me for many more years to come. I am very grateful to my thesis committee members for ensuring that I became the best scientist that I could during my time at UVM, and for keeping my training and education at the forefront; Drs. James Hudziak, Russell Tracy and Jeffrey Bond: Thank you! I want to acknowledge the CMB community, particularly Drs. Nicholas Heintz and Matthew Poynter and the administrators over the years: Erin Montgomery, Kirstin van Luling, Jessica Deaette, Carrie Perkins, and Haley Bradstreet. I am grateful to the MMG department, especially Dr. Susan Wallace and to the administrators for their continued help and support: Barbara Drapelick, Helen Brunelle, France Roy and Anne MacLeod. I want to thank the UVM office of international education, especially Emma Swift and Evan Mills. I want to thank my first year mentors, especially Dr. Stephen Everse. Last but not least, the people who were there when I needed them the most: Drs. Alan Rubin and Claire Verschraegen. To all of you, and everyone else in the UVM community that I have had the pleasure to know and/or learn from: Thank you for making the past five years the best experience I could have wished for!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Brain disorders represent a major burden around the world, affecting at least 35% of the global population[1]. These disorders are often categorized into two primary types: psychiatric and neurological (e.g., neurodegenerative, neurobehavioral, neurocognitive and neurodevelopmental)[1]. In this dissertation, we focus primarily on alcohol dependence (AD) as a model for psychiatric disorders, and expand our search into Alzheimer's disease as a model for neurological disorders. Genomics and population genetics methods were developed and applied to samples ascertained for AD (Chapters 2-4) and Alzheimer's disease (Appendix A) to identify new disease-associated variants. In this introduction we review the genetic literature on AD and the most recent scientific paradigms of brain disease genetics, as they relate to the scope and purpose of our work.

Prevalence

Psychiatric disorders present an extreme burden to the health and overall well-being of affected individuals, their families, and indeed, our society as a whole. Specifically, alcohol use disorders represent one of the most costly diseases, with over $249 billion spent by USA alone, and around 3.3 million deaths across the globe (2010 statistics, NIAAA). According to the Diagnostic and Statistical Manual of Mental Disorders (DSM IV), AD is characterized as a "syndrome of persistent problems involving physiological tolerance, psychological cravings and behaviors centered around

alcohol use or the consequences of alcohol use" with an onset of mid-twenties[2]. According to the same source, the general population prevalence in the USA ranges from 13%, for alcohol abuse, to 5% for AD. Among adults, variations in prevalence exist across ethnicities, with the highest rates found among Native Americans and Native Alaskans (12.1%), European Americans (8.9%), Hispanics (7.9%), African Americans (6.9%) and Asian Americans and pacific islanders (4.5%).

Neurobiology

The negative effect of alcohol on human health became acknowledged as early as 1923 where physicians had noticed that pneumonia diagnoses was 32% higher in moderate alcohol users compared to abstainers (Capps and Coleman, 1923). Later, several psychoanalytic studies held the view that all men were born addicted, however, "alcoholics are notoriously slow to get over it" (Goodwin, 1968). However, the role of alcohol use on mental health was not clarified until early 1980, when individuals with problematic drinking behavior were observed to have been "depressed and unable to cope"[3]. Advances in physiology and functional neuroanatomy from the clinic and animal studies revealed a crucial pathway in the brain involved in etiology of AD and other addictive behaviors: the brain reward circuit.

One of the most well-annotate brain reward circuits comprises of dopaminergic neurons in the ventral tegmental area (VTA) projecting to the nucleus accumbens (NAc). The dopaminergic neurons of VTA-NAc innervate the prefrontal cortex, central and

basolateral amygdala, the hippocampus as well as other areas, making this circuit very important for the recognition and "consumption" of stimuli[4]. This area of the brain, well-conserved across mammalian brains, tells the organism if a stimulus is rewarding or aversive. The primary types of neurons in this area of the brain are GABAergic; however, there's input from other areas of the brain with glutamatergic neurons from hippocampus, basolateral and extended amygdala, pre-frontal cortex, and Orexinergic neurons from lateral hypothalamus. Amygdala is important in establishing whether an experience (e.g. food, stress, drug or abuse) is rewarding or aversive, while hippocampus plays a crucial role in declarative memory, i.e., association of places and experiences, therefore, playing a key role in drug use or abuse relapse. The least understood areas of the brain interacting with VTA-NAc are the pre-frontal cortex areas, such as medial, anterior cingulate cortex and orbitofrontal cortex, all of which may play a crucial role in the decision-making process, e.g., seeking reward.

Figure 1 demonstrates the interactions of ethanol in the VTA-NAc circuit of the brain under three different scenarios: control, acute and chronic ethanol exposure. As shown, ethanol molecules interact with both glutamate and dopamine axonal projections. Specifically, under the acute ethanol exposure scenario, ethanol molecules effect the brain on multiple fonts: inside the pre-synaptic terminal, GABAergic neuronal activity is decreased, leading to an increase in dopaminergic activity, ultimately increasing dopamine release in the synaptic cleft; meanwhile, in the synaptic cleft, ethanol acts as a blocker of N-methyl D-aspartate glutamate receptor (NMDA-R), further preventing Ca^{2+} influx into post-synaptic neuron, inhibiting synaptic plasticity induction in the post-

synaptic neuron. Lastly, in the post-synaptic neuron, prodynorphin (PDYN) is upregulated, downstream from CREB (Cyclic AMP-Responsive Element-Binding Protein 1, a transcription factor) upregulation, leading to decreased ethanol intake.

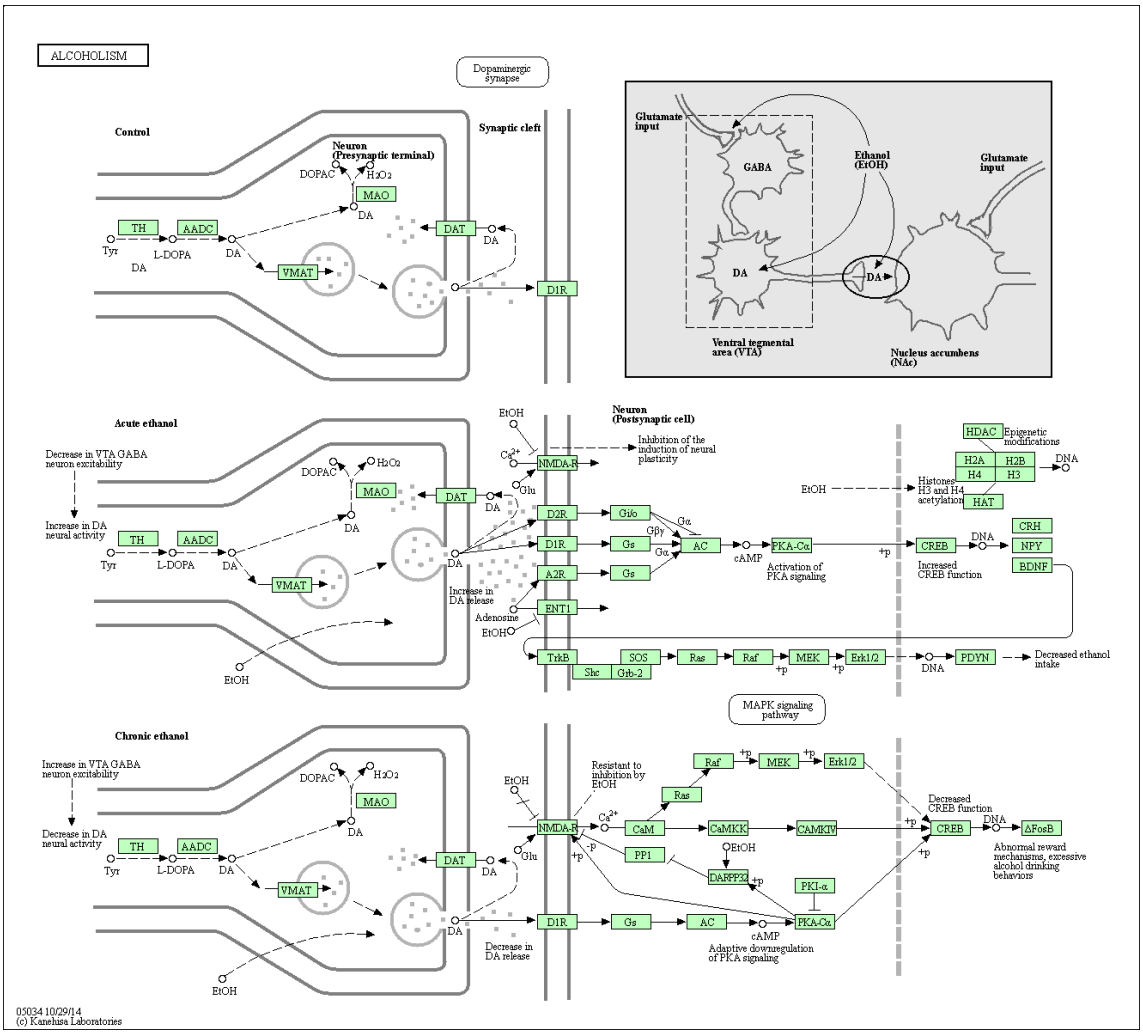


Figure 1: A molecular model for alcohol dependence (diagram from Kyoto Encyclopedia of Genes and Genomes; accessed on 05/01/2017)

However, under a chronic exposure scenario (**Figure 1**), GABAergic neurons in the VTA become increasingly excited, leading to a decrease in dopamine. Thus, less dopamine is released in synaptic cleft. On the surface of the post-synaptic neuron, NMDA-R has become resistant to ethanol inhibition, and Ca^{2+} influx into the post-synaptic neuron activates MAPK signaling, ultimately decreasing CREB function, and ultimately abnormal reward mechanisms and excessive drinking behaviors.

Genetic heritability

According to offspring data of adopted children registries from alcoholic parents, AD has an estimated heritability $h^2 \approx 39\%$ [5] and according to twin studies $h^2 \approx 64\%$ [6]. Additionally, the authors identified heritability across a period of 20 years, and found that their heritability estimate was consistent across different years. Another notable study[7] found heritability estimates consistent with Heath et al., and twin resemblance (i.e., phenotype concordance) was attributed to genetic factors (54%) and environmental factors (14%). It is important to note that AD is comorbid with other substance abuse and psychiatric disorders (such as antisocial personality disorder). It has been shown that this comorbidity has a heritability of 80%[8].

Genome-wide linkage studies

The first genome-wide linkage analyses were published by investigators in the intramural program of National Institute on Alcohol Abuse and Alcoholism (NIAAA)[9] and the Collaborative Study on Genetics of Alcoholism (COGA) group[10]. The NIAAA group ascertained southwestern American Native American tribe samples for A, while COGA recruited samples in six different sites across the United States. Both groups reported AD risk loci with LOD scores of 2 or higher, residing in vicinity of the alcohol dehydrogenase (*ADH*) gene on chromosome 4q. The only genome-wide linkage analysis of AD in African Americans was conducted in 2009 and reported genome-wide statistically significant loci in chromosome 10.

Candidate gene studies

The most abundant findings have come from candidate gene association studies. These genes were initially chosen based on their role in alcohol metabolism. For instance, the product of *ALDH2*, acetaldehyde dehydrogenase 2, is one of the primary acetaldehyde dehydrogenases involved in clearing the metabolic intermediary acetaldehyde. Acetaldehyde is known to produce a "flushing reaction", characterized by nausea and overall physiological discomfort. Thus, a variant known to decrease *ALDH2* function (commonly present in East Asian populations) is a protective variants for AD[11]. A highly replicated finding is that of an *ADH4* [12] [12] . One of the identified variants in this gene,

A75C, was shown to decrease promoter activity by more than twofold. Variants in *ADH1B* were also shown to be robustly associate with AD diagnosis, in the first meta-analysis for this gene[13].

Other candidate genes were selected due to their function in neurotransmission. In the 1990s several genes were reported, including dopamine receptor 2 (*DR2*)[14] and gamma amino-butyric acid (*GABA*) gene cluster, including *GABAβ2*, *GABAα6* and *GABAγ2*[15].However, some of these genes were not replicated; the *DR2* locus created much controversy after its publication[16]. Fine mapping of GABA gene cluster identified haplotypes and single alleles associated in the *GABRA2* gene to AD. Non-association studies were also reported for *GABRA2* [17] [18]. We performed the first meta-analysis of the GABA gene cluster with AD[19], where *GABRA2* gene provided the best evidence of association, i.e., association p-value $P = 9 \times 10^{-6}$ and odds ratio (OR) 95% confidence interval (CI) = 1.27 (1.15, 1.4) for SNP rs567926.

In addition to genes involved in alcohol metabolism and neurotransmission, several other genes have been published from candidate gene studies, including *CHRM2* (encoding muscarinic acetylcholine receptor M2) [20] and *OPRM1* (encoding the μ opiod receptor). The *OPRM1* gene contains a polymorphism resulting in amino acid substitution Asn40Asp, previously associated with AD; however, a meta-analysis found no overall association to AD. Interestingly, the same allele has been shown to lead to differential response to drug treatment, which will be discussed in further detail below.

Genome-wide association studies (GWAS)

After the completion of the human genome reference in 2000[21], and the rapid development of sequencing technologies, the genome-wide association studies (i.e., GWAS) era began[22]. The first GWAS of a substance dependence phenotype was carried out in nicotine dependence cohorts[23]. After Nicotine dependence, AD is the most studied substance dependence phenotype. The first GWAS of AD was conducted in 2009, in samples of German ancestry[24], and it reported nominal associations between AD and two previously associated genes, *CDH13* and *ADH1C*. The same study found a genome-wide significant locus rs7590720, located in the intergenic region of chromosome cytoband 2q35. Later studies identified autism-related gene *AUTS2*[25], intergenic variants in the previously published *ADH* gene cluster[26], intergenic region between *NKAIN1* and *SERINC2*[27]. The first GWAS in an African American study was conducted in 2014[28] and novel loci crossing genome-wide significance threshold were reported in *METAP* and rs1437396 in the intergenic space between *MTIF2* and *CCDC88A*.

The most recent catalogue of GWAS reports a total of 50 genes, each harboring association signals to AD that have been independently replicated at least once. The risk alleles found in these 50 genes are mostly common, with an average allele frequency (± standard deviation) of 0.29 ± 0.2.

Gene by environment interaction

Unlike other complex traits and diseases, the environment is a necessary component to AD onset. One cannot develop AD without exposure to alcohol. For other brain-related diseases, such as Alzheimer, no environmental factors are required to observe onset. The first gene-by-environment (G x E) study was conducted between an allele in the promoter region of the serotonin transporter gene (5-HTT), also known as of the s allele, and family relations using a Swedish cohort of adolescents between 16 and 19 years of age. The study reported an increase in alcohol intoxication frequency of 12-14 fold higher between carriers of the *s* allele with bad family relations and those with good family relations. A study conducted two years later reported a similar finding where college students carrying the *s* allele and experiencing stressful life events where at higher risk of abusing alcohol than s allele carriers who were not experiencing stressful life events [29].

Gene by drug interaction

The *OPRM1* gene has been assessed for mediating effects of opiod antagonist naltrexone. The Asp40 status on this gene was used to recruit patients and conduct a double blind study in a placebo-controlled trial where individuals were treated with placebo or naltrexone prior to intravenous alcohol challenge session[30]. Individuals heterozygous or homozygous for the Asp40 allele reported lower levels of alcohol

cravings; naltrexone weakened the positive effect of alcohol response, particularly in those carrying the Asp40 allele.

After several decades of ambiguity and controversy surrounding AD biology, work in the late 1980s and early 1990s suggested that drugs that activate neuronal production of 5HT (5 hydroxytryptamine receptors, also known as serotonin receptors) and block its re-uptake may reduce alcohol intake, particularly when combining psychopharmacological approaches with psychosocial therapies[31].

Missing heritability

For most complex human traits and disorders, relatives are more alike than unrelated individuals. This correlation between phenotypes of relatives underlies the fundamental premise of genetics of complex human diseases. However, this correlation is not fixed for a given phenotype, and the variation in its value is determined by many genetic and non-genetic components, thus it becomes necessary to specify which type of genetic component one is measuring. For instance, the total variance of a phenotype (V_P) is determined by variance in genetic components (V_G) and environmental components (V_E):

$$V_P = V_G + V_E$$

The genetic components are further divided into additive (V_A), dominant (V_D), and interaction (V_I) components, while the environmental components are divided into

common (V_C) and non-common (V_E) (i.e., everything else, that is, the remainder of environmental factors):

$$V_P = V_A + V_D + V_I + V_C + V_E$$

The additive genetic component (V_A) is usually the major contributor of resemblance between relatives. The proportion of phenotypic variance explained by this specific genetic component is known as narrow-sense heritability (h^2):

$$h^2 = V_A / V_P$$

From here onwards, h^2 will be referred to simply as genetic heritability. The additive genetic factors are presumed to have been passed down from parents to children. Importantly, additive genetic effect is difficult to ascertain; it's only when $V_G = V_A$ that dominant and epistatic effects of genes on the phenotype can be ignored, allowing for an empirical ascertainment of the V_A[32].

A commonly used method of estimating heritability is to use registries of twins (monozygotic, MZ, and dizygotic, DZ) data ascertained for the phenotype of interest. In this case, we assume that the phenotype is dichotomous, e.g., cases and controls, with a population prevalence of 8%. For MZ pairs we expect concordant cases $n_{11} = 62$, discordant pairs, $n_{10} = 791$, and concordant controls $n_{00} = 4{,}147$; while for DZ pairs, we observe $n_{11} = 45$, $n_{10} = 740$, and $n_{00} = 4{,}215$[32]. Assuming $V_D = 0$, we apply the following formula for the measurement of intra-class correlation coefficient (t):

$$t = \frac{n_{11}n_{00} - \left(\frac{n_{10}}{2}\right)^2}{n_{11}n_{00} + \left(\frac{n_{10}}{2}\right)^2}$$

Next, we obtain t_{MZ} = 0.24 and t_{DZ} = 0.16. Finally, genetic heritability is $h^2 = 2(t_{MZ} - t_{DZ}) = 0.16 = 16\%$.

Strategies for elucidating missing heritability

Using GWAS results, we could estimate the "explained heritability" by each associated variant using approaches discussed in detail elsewhere[33]. Studies of AD have demonstrated that only a fraction of AD heritability[34] is accounted for by the current association findings. The rest of the genetic heritability, i.e., the "missing heritability", is yet to be elucidated and novel disease-associated loci are being reported at an increasing rate. Next, we discuss the two major approaches used to discover new disease-associated genes, representing the two most significant aspects of this dissertation.

Rare variants

Several studies have demonstrated that common variants (primarily SNPs) capture <10% of genetic heritability of complex human diseases[35]. Alternative approaches have been designed to address this limitation of common variants, including association tests for rare variants and structural variations[36]. The current GWAS findings

of AD consist of 50 genes, with variants that have allele frequencies of 29% on average (see above), thus leveraging rare variants is a worthwhile effort to elucidate the missing heritability of this disease.

The advent and increased feasibility of NGS has enabled the discovery of genomic variants that are individually rare but commonly frequent. Since these variants are too rare to observe segregation in affected families, the traditional family study designs used during the linkage analysis era (see above) are not appropriate for association testing. Thus, a range of statistical methods have been developed, from methods that focus on comparison of cases-exclusive variants to controls-exclusive variants (RVE [37]) to combined multivariate and collapsing methods (CMC [38]) to weighted sum statistics (WSS [39]). The latter two methods have a power advantage over the RVE method. Intuitively, these methods rely on the concept of burden-testing, i.e., collapsing rare variants that fall within a pre-defined region, e.g., gene region to increase the "effective sample size" of that region. Nuances of this idea exist in the literature, where the collapsing is done for different minor allele frequency (MAF) categories. However, one common limitation of all these tests is that they assume the magnitude and direction of all rare variants under study is similar. To address this limitation, sequencing-based rare variation association testing with the sequence kernel association test or SKAT was developed [40].

Structural variants

Although the methods discussed above are not variant-specific, most studies of rare and common variants are focused on SNPs. After the discovery of CNVs in the general population in 2004[41], a plethora of studies started focusing on identification and association of CNVs with human diseases. Soon after the CNVs were found as ubiquitous sources of genomic variation across human populations, within two years, over 3,000 population-wide CNVs were identified[42]. Soon afterwards, both common (e.g., MAF > 5%) and rare (e.g., MAF ≤ 5%) CNVs were being associated with complex human diseases, with some of the most successful associations being observed in autism[43] and other neurodevelopmental disorders[44].

Importantly, common CNVs were observed to be in strong linkage disequilibrium (LD) with GWAS SNPs, while rare CNVs were located in regions with a paucity of GWAS SNPs. This observation had considerable impact on the efforts to elucidate the missing genetic heritability, since rare CNVS were more likely to reveal novel disease-associated loci than common CNVs. Thus, the next frontier of human genetics research was deemed to be the study and disease-association of rare CNVs[45]. In the recent years, several studies have demonstrated association of rare or *de-novo* CNVs to psychiatric[46] and neurodevelopmental disorders[47] through enrichment-based approaches. In addition to rare frequency CNVs being relatively independent of GWAS SNPs, they tend to be longer, and thus more likely to overlap with a gene region and have a pathogenic effect.

Scope and purpose

A common problem with the existing methodology and study paradigm of "rare variants - common disease" is population structure and inability to distinguish between neutral and truly disease-associated variants[36]. To identify genes under putative positive selection, we designed a novel method that identifies genes, pathways and complex diseases enriched with highly-differentiated alleles within populations of the same continent (**Chapter 2.1**). Next we demonstrate that eye color, a well-known pigmentation trait under positive selection, is associated with AD in European American population, further supporting the potential influence of natural selection forces on AD risk loci (**Chapters 2.2 and 2.3**).

One continuing challenge in population genetics is population structure, particularly for rare variants, or candidate gene sequencing studies. To address this issue, in **Chapter 3.1** we present a set of 325 panels of ancestry informative markers (AIMs) that may be used to adjust for population structure in a hierarchical fashion, representing a departure from traditional application of AIMs panels. We also report a novel method used for standardizing allele information, crucial for genotype imputation and meta-analyses (**Chapter 3.2**).

We expand on the existing post-GWAS studies by conducting the first CNV-based GWAS meta-analysis of AD in five cohorts of European and African ancestry (**Chapter 4.1**). Lastly, we present a framework for viral integrations from paired-end sequencing

data (**Chapter 4.2)** and a case study where we detected Human Herpesvirus 6 in the brain of an Alzheimer's disease patient (**Appendix A**).

CHAPTER 2: EVOLUTIONARY INSIGHTS

Chapter 2.1: Atlas of Human Diseases Influenced by Genetic Variants with Extreme Allele Frequency Differences

Arvis Sulovari[1], Yolanda H Chen[2], James J Hudziak[3] and Dawei Li[1,4,5*]

[1]*Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA*

[2]*Deptment of Plant and Soil Sciences, University of Vermont, Burlington, Vermont 05405, USA*

[3]*Vermont Center for Children, Youth, and Families, Department of Psychiatry, University of Vermont, Burlington, Vermont 05405, USA*

[4]*Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA*

[5]*Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, Vermont 05405, USA*

*To whom correspondence should be addressed:

Dawei Li, Ph.D., Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA. E-mail: dawei.li@uvm.edu

Number of words in the abstract: 181

Number of words in the text (excluding acknowledgments, financial disclosures, legends, and references): 5,405

Number of tables: 3

Number of figures: 3

Number of supplementary materials: 15 supplementary Tables and 9 supplementary Figures and Legends.

Abstract

Background: Genetic variants with extreme allele frequency differences (EAFD) may underlie some human health disparities across populations.

Method: To identify EAFD loci, we systematically analyzed and characterized 81 million genomic variants from 2,504 unrelated individuals of 26 world populations (phase III of the 1000 Genomes Project).

Results: Our analyses revealed a total of 434 genes, 15 pathways, and 18 diseases and traits influenced by EAFD variants from five continental populations. They included known EAFD genes, such as *LCT* (lactose tolerance), *SLC24A5* (skin pigmentation), and *EDAR* (hair morphology). We found many novel EAFD genes, including *TBC1D2B* (autophagy mediator), *TRIM40* (gastrointestinal inflammatory regulator), *KRT71*, *KRT75*, *KRT83* and *KRTAP10-1* (hair and epithelial keratin synthesis), *PIK3R3* (insulin receptor interaction), *DARS* (neurological disorders), and *NACA2* (skin inflammatory response). Our results also showed four complex diseases significantly enriched with EAFD loci, including asthma (adjusted enrichment $P = 4 \times 10^{-8}$), type I diabetes ($P = 6 \times 10^{-9}$), alcohol consumption ($P = 0.0002$), and attention-deficit/hyperactivity disorder ($P = 0.003$).

Conclusion: This study provides a comprehensive atlas of genes, pathways, and human diseases significantly influenced by EAFD variants.

Keywords: Missing heritability, Genomic variation, Fixation index (F_{ST}), Extreme allele frequency differences (EAFD), Population structure

Introduction

Evolutionary events such as migration, natural selection, and genetic drift have cumulatively changed the allele frequency spectrum of genomic variants in human populations. Sometimes, extreme allele frequency differences (EAFD) will exist even between pairs of closely related populations, e.g., populations from the same continent[48]. Due to their shared recent migration history, EAFD loci between related populations are unlikely driven by migration, but rather by genetic drift or selection. An allele under genetic drift with negative impact on fitness does not remain for long in a population due to purifying selection. Thus, disease-associated or pathogenic EAFD variants may be driven by balancing selection[49] (i.e., the heterozygote has selective advantage, such as the resistance to malarial infection[50-52]) or recent drift (i.e., random deleterious events that have either not been purified from the population yet[53], such as Tay-Sachs[54] disease, or have little effect on fitness[55]). An atlas of complex diseases influenced by EAFD would be of interest to many, as it could enable an improved understanding of the origins of various diseases as well as promote the discovery of novel etiological factors through evolutionary-driven hypotheses. To our knowledge, no such resource exists yet.

In this study, we identified a comprehensive list of diseases and traits influenced by EAFD in a global reference of human populations. In order to measure the relationship between EAFD and disease susceptibility, we designed an unbiased approach to systematically identify variants with EAFD between populations of a continent at the whole-genome level, and then determine associated genes, biological pathways, and complex phenotypes. An unbiased population differentiation estimator (fixation index or F_{ST}), specifically designed for sequencing data with abundance of rare variants, was used to identify EAFD loci. We analyzed over 81 million single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) from 2,504 unrelated individuals in the 1000 Genomes Project, Phase 3. These samples represent 26 world populations or five major post-migratory populations: African (AFR), European (EUR), East Asian (EAS), South Asian (SAS), and admixed American (AMR, **Supplementary Table 1**). We compared characteristics of whole-genome variants in different populations, and then identified genetic markers with EAFD. Then, we conducted systematic and rigorous enrichment analyses to identify genes, pathways, diseases and traits influenced by EAFD in the five major human populations (**Figure 1**).

Results

Whole-genome scans for EAFD

All of the biallelic variants (**Supplementary Table 2**) were used to estimate the fixation index (F_{ST}) for each of the 325 population pairs (see Methods; **Supplementary Tables 4**

and 5). To identify EAFD, we selected high-F_{ST} variants, i.e., those with F_{ST} values greater than the threshold (θ_H, which was the F_{ST} value at the 99.9^{th} percentile; see Methods). The variation of θ_H across chromosomes was minimal (range 5.8×10^{-6} - 0.0056). Our results showed that the θ_H values of our selected high-F_{ST} variants were well-correlated with population geographic distances (e.g., $\rho = 0.93$ and $R^2 = 0.87$ in Africans after removal of admixture samples; **Supplementary Table 6** shows both genetic and geographic distances for all the 325 population pairs; **Supplementary Figure 1** and **Supplementary Note 1**), and thus θ_H was considered a proxy for genetic distance between populations. As expected, we found that the θ_H values were significantly lower within (mean ± SD = 0.11 ± 0.05) than between continental groups (mean ± SD = 0.51 ± 0.16; Welch's two-sample t-test $P = 1 \times 10^{-54}$). To elucidate the structure of all 26 populations in detail, we analyzed each of the 325 population pairs by using θ_H as a proxy for their genetic distance. **Figure 2** shows the θ_H-based genetic distances among the 26 populations (**Supplementary Note 2**). Populations from the same continental group clustered closer to each other than those from different continental groups, with the exception of the admixed Americans. This structure was confirmed by principal component analysis (PCA, **Supplementary Figure 2**).

To screen for EAFD loci we identified variants with extreme allelic differentiation using θ_H as a threshold (see **Methods**). EAFD was applied only to populations that share continental origin, and each continental group was analyzed separately. As a result, we identified a total of 774,187 candidate EAFD loci across the five continental groups for

further analyses. The total number of EAFD variants was comparable to the 762,000 variants identified by 1000 Genomes Project, where an alternative method was used to identify alleles with large allele frequency differences [48]. Among our EAFD variants, 32,295 were shared by at least two continental groups (30,644, 1,489, 135, and 27 were shared by two, three, four, and five continental groups, respectively). **Supplementary Table 7** indicates that 19-33% of the EAFD variants were recurrent across population pairs. **Supplementary Table 8** and **Supplementary Figure 3** present the numbers of variants shared by different sets of populations.

Allele frequencies

Among our identified EAFD variants (high-F_{ST} variants *within* each continental group), we observed relatively more rare than common variants. Africans revealed the highest proportion of rare variants (product-moment correlation $\rho = -0.75$ and adjusted $R^2 = 0.56$) while Americans contained the lowest; the other three populations revealed a linear decline in abundance of EAFD variants with increasing allele frequencies ($\rho = -0.86$ and $R^2 = 0.73$ in EUR; $\rho = -0.80$ and $R^2 = 0.64$ in EAS; $\rho = -0.86$ and $R^2 = 0.75$ in SAS; **Figure 3A**). The allele frequency distribution of EAFD variants was vastly different from that of whole-genome variants (mean = 0.29 (AFR), 0.33 (EUR), 0.36 (EAS), 0.33 (SAS), and 0.38 (AMR) for the former, compared to 0.04 for whole-genome variants; $\rho = -0.05$ and $R^2 = 0.002$). The EAFD variants contained fewer rare variants (derived allele frequency, DAF < 5%) than whole-genome variants; i.e., only 0.6 % in Africans, 0.3% in

Europeans, 0.3% in East Asians, and 0% in South Asians and Americans, compared to 84.6% on the whole-genome level. This difference persisted for variants with DAF < 10%, i.e., 28% (AFR), 20% (EUR), 17% (EAS), 21% (SAS), and 6% (AMR), compared to whole-genome of 91.6%. Thus, both rare and common variants are targeted by EAFD; however, common variants are enriched. This observation supports the conclusion that our EAFD variant selection was not inflated by rare variants.

Among all EAFD variants, we observed a total of 506 indels, including 332 (0.09% of EAFD loci) from Africans, 62 (0.051%) from Europeans, 56 (0.053%) from East Asians, 235 from South Asians (0.17%), and 56 (0.063%) from Americans, compared to 3.7% indels from whole-genome. Among them, only one indel was genic (frameshift insertion rs111905334 in the *ITIH5* gene). This result reflects the potential preference of EAFD forces for SNPs against indels. Consistently, the EAFD indels showed markedly higher DAF (mean = 44%) than whole-genome (mean = 12%).

Linkage disequilibrium and physical proximity

We compared the linkage disequilibrium (LD) between EAFD and random SNPs (see **Methods** for the generation of matched random SNPs). We found that EAFD SNPs had longer LD ranges than random SNPs (OR = 1.72 (1.39 – 2.12) and $P = 2 \times 10^{-7}$; **Supplementary Figure 4**). This was consistent with the well-known property of extended haplotype homozygosity around positive selection variants[56].

The distances between consecutive EAFD variants followed a lognormal distribution (mean = 7,829 bp and SD = 59,795). These EAFD variants were more densely clustered than expected (e.g., 12% of the EAFD variants in Africans were in 100bp clusters; by comparison, only 7.2% of random variants were in such clusters; OR = 1.7 (1.67 – 1.72) and χ^2 test $P < 2 \times 10^{-16}$; **Supplementary Figure 5**).

Biological functions and disease susceptibility implications

Functional annotation

EAFD loci were enriched in genic regions: 28% of these loci were genic, compared to only 2.9% from whole-genome variants (χ^2 test odds ratio (95% confidence interval) or OR = 9.62 (9.56 - 9.7) and $P < 2 \times 10^{-16}$). Overall, the pathogenicity of the EAFD variants, as measured by their GERP scores, was lower than that of whole-genome variants (Wilcoxon rank sum test $P < 2 \times 10^{-16}$). As expected, on the whole-genome level, the coding regions had larger GERP scores than non-coding regions ($P < 2 \times 10^{-16}$, **Figure 3B**); however, for the EAFD variants, the two categories exhibited similar scores in all populations except East Asian ($P = 0.02$, **Figure 3B**). When the coding EAFD loci were decomposed into more specific functional categories, we observed a depletion of the most damaging variants (**Figure 3C** and **3D**); for instance, depletion for missense (OR = 0.6 (0.56 - 0.65) and $P < 2 \times 10^{-16}$), "stop gain" (OR = 0.3 (0.13 - 0.55) and $P < 2 \times 10^{-16}$), and the 3-prime untranslated region or 3'UTR (OR = 0.63 (0.59 - 0.67) and $P < 2$

$\times 10^{-16}$), as well as a decrease of the missense/synonymous variant ratio relative to all genic variants (e.g., 1.0, 1.3, 1.2, 1.1, and 1.1 for AFR, EAS, EUR, SAS, AMR, respectively, in comparison to a 1.9 ratio for genic variants).

The EAFD loci were enriched in the 5'UTR region (OR = 2.8 (2.45 - 3.1) and $P = 2 \times 10^{-5}$) and with synonymous variants (OR = 1.2 (1.10 - 1.26) and $P = 1.8 \times 10^{-5}$); **Figure 3D**). This reflects a possible preference of EAFD for regions that impact gene expression (e.g., 5'UTR), comparable to previous findings on local adaption drivers[57,58]. These findings were consistent among populations ($P > 0.05$ for all continental group pairs; **Supplementary Note 3**). The proportion of missense variants increased with decreasing genetic distance ($P = 0.002$, **Supplementary Table 9)**. Overall, we found that EAFD loci were approximately 10 times more likely to be genic, and were less damaging on average than whole-genome variants; however, some of these loci might compromise gene expression through regulatory regions such as 5'UTR.

Enrichment analyses of EAFD genes

We identified 434 EAFD genes: 138, 88, 91, 103 and 65 from Africans, Europeans, East Asians, South Asians, and Americans (**Supplementary Table 10**). These genes represent a highly distilled list from the total of 29,061 genes and pseudogenes containing at least one EAFD; of these, 21,614, 9,045, 7,875, 10,039 and 7,316 genes were found in Africans, Europeans, East Asians, South Asians and Americans, respectively. Each of the EAFD genes was significantly enriched with EAFD variants (**Supplementary Figure 6**

and Supplementary Table 11) and contained at least one nonsynonymous EAFD variant (average of 1.9 SNPs per gene)). The six EAFD genes containing the most damaging variants (i.e., prematurely halted protein synthesis and had CADD score ≥ 20) were involved in autoimmune (*HLA-DRB1*, *HLA-DRB5* and *LILRA3*), viral (*FUT2*) and parasite (*CD36*) infection response and olfaction (*OR52J3*) (**Table 1**). The well-known malarial resistance locus rs3211938[59] was identified in Africans, with the largest allele frequency difference in YRI and ESN. Of the other five damaging EAFD loci, the most striking allele frequency difference was observed in two East Asian populations with SNP rs138741442, which had allele frequencies of 0 and 12% in CDX and CHB, respectively.

Enrichment analyses of EAFD pathways

We found strong evidence of EAFD influence on pathways across all continental populations, such as asthma (e.g., R = 70 and adjusted $P = 4 \times 10^{-8}$), type I diabetes (e.g., R = 59 and $P = 6 \times 10^{-9}$) and autoimmune thyroid disease (e.g., R = 49 and $P = 2 \times 10^{-8}$, **Table 2**). In addition, we observed several population-specific pathways, including fat digestion and absorption in Africans (R = 14 and $P = 0.02$), endocytosis in Europeans (R = 7.5 and $P = 0.008$), osteoclast differentiation in East Asians (R = 7.6 and $P = 0.03$), type II diabetes in South Asians (R = 18 and $P = 0.008$), and primary immunodeficiency in Americans (R = 38 and $P = 0.004$; **Supplementary Table 12**). **Table 2** shows the identified pathways influenced by EAFD forces that replicated in all five continental groups.

Enrichment analyses of EAFD diseases and traits

To evaluate the influence of EAFD on disease susceptibility, we matched identified EAFD SNPs to diseases and trait associations maintained in the GWAS Catalogue[60]. After manual curation of the most recent GWAS Catalogue (**Supplementary Table 13**), we obtained 7,523 unique SNPs, each of which was associated with at least one of the 726 diseases and traits, and was replicated at least once. Among them, we found that from a total of 1,003 SNPs, 13% of these associations, were EAFD loci (i.e., 397, 204, 215, 181, and 135 SNPs from AFR, EUR, EAS, SAS, and AMR, respectively). To further identify diseases or traits significantly influenced by EAFD, we carried out enrichment analyses, identifying a total of 18 GWAS diseases and traits were significantly enriched with EAFD variants (after adjustment for multiple testing, **Table 3**).

These include 1) pigmentation traits, such as hair color (four among the six known hair color SNPs were EAFD SNPs or denoted as 4/6, R = 25 and adjusted $P = 0.0008$); 2) brain disorders, including attention deficit hyperactivity disorder (5/25, R = 11 and $P = 0.003$), alcohol consumption (4/5, R = 28 and $P = 0.0002$), drinking behavior (2/3, R = 23 and $P = 0.04$), frontotemporal dementia (both of the two associated SNPs were EAFD SNPs, R = 35 and $P = 0.018$), and white matter hyperintensity burden (both of the two associated SNPs were EAFD SNPs, R = 37 and $P = 0.02$); 3) metabolic traits, including trans fatty acid levels (34/131, R = 9.1 and $P = 1 \times 10^{-21}$; replicated in another continental

group: 23/131, R = 8.9 and $P = 5 \times 10^{-15}$), triglycerides (10/91, R = 3.8 and $P = 0.01$), gamma glutamyl transpeptidase (3/7, R = 15 and $P = 0.018$), glycemic traits (3/6, R = 17 and $P = 0.016$), and comprehensive strength and appendicular lean mass (both of the two associated SNPs were EAFD SNPs and this was consistent using data from two different populations, R = 55.8 and $P = 0.0078$); 4) infectious diseases, including chronic hepatitis B infection (3/9, R = 12 and $P = 0.03$), and response to hepatitis C treatment (2/3, R = 23 and $P = 0.036$); 5) nasopharyngeal carcinoma (3/11, R = 10 and $P = 0.04$), and 6) others, such as corneal curvature (3/8, R = 21 and $P = 0.0078$; **Table 3**). **Supplementary Table 14** shows the complete results of enrichment analyses of GWAS diseases and traits targeted by EAFD. Sharing patterns of EAFD genes, pathways, diseases and traits across the five continental groups are shown in **Supplementary Figure 7**.

All of the four alcohol consumption-associated SNPs identified as EAFD SNPs in East Asians (i.e., rs10849915 in intron of *CCDC63*, rs12229654 nearby *MYL2*, rs2074356 in intron of *HECTD4*, and rs2072134 in the 5' UTR of *OAS2*) originated from GWASs of also an East Asian population[61]. This concordance supports a strong relationship between EAFD and alcohol consumption or exposure, potentially due to local adaptation; e.g., environmental presence of fermented fruits[62]. All of the alcohol consumption-EAFD loci were in the same direction with respect to trait-increasing alleles. Furthermore, the derived allele showed protective effect, and this was true for all of the four EAFD SNPs (effect sizes[61] = -0.55, -1.06, -0.61, and -0.79 for the four derived alleles rs10849915-C, rs2074356-A, rs2072134-A, and rs12229654-G, respectively). Some of the EAFD

biological pathways converged with enriched GWAS diseases. For example, type I diabetes and glycemic traits[63] were identified via our pathway and disease/trait analyses, respectively. The EAFD SNPs used in both analyses were from East Asians; and the same was true for frontotemporal dementia (**Supplementary Table 12**) and neurotrophin signaling[64] (**Supplementary Table 14**).

To estimate the extent of the relationship between EAFD and disease susceptibility, we calculated the proportion of the cumulative effect sizes (i.e., odds ratios) explained by EAFD variants among the total effect sizes of all identified disease variants using data from the GWAS Catalogue (population matched). We found that the EAFD SNPs accounted for 26% and 70% of the total effects sizes of all known associated SNPs for alcohol consumption[61] and pigmentation traits[65], respectively. The results indicate local EAFD is likely to have considerable influences on traits and disease genetic heritability. In all, EAFD forces have influenced genes, biological pathways, and further affected traits and the susceptibility to a wide range of diseases - including infections, brain diseases, metabolic functions, and potentially cancer - across five major human populations.

Discussion

In this study, we characterized over 81 million whole-genome biallelic SNPs and indels from 26 human populations. With this data we were able to identify 434 candidate genes,

15 pathways, and 18 GWAS diseases and traits influenced by EAFD. We identified many known positive selection genes such as *LCT*[66] (lactose tolerance), *SLC24A5*[67] (skin pigmentation) and *EDAR*[68] (hair morphology). More importantly, we also detected novel EAFD genes (**Supplementary Table 10)**, such as *OR52J3* (smell perception), *TBC1D2B* (autophagy mediator [69]), *TRIM40* (gastrointestinal inflammatory regulator [70]), *KRT71*, *KRT75, KRT83* and *KRTAP10-1* (hair and epithelial keratin synthesis), *PIK3R3* (insulin receptor interaction [71]), *DARS*[72] (neurological disorders) and *NACA2* (skin inflammatory response[73]). Each of our novel genes was enriched with EAFD variants, where at least one was nonsynonymous, providing a tractable list for experimental validations. The individual gene functions converged with pathway and disease enrichment analyses.

We have identified a total of 15 pathways, such as olfactory transduction, asthma, type I diabetes, rheumatoid arthritis, viral myocarditis, allograft rejection, immune system disorders (**Table 2**), as well as 18 diseases and traits (**Table 3**). Most of the disease and traits are known for differential prevalence or disease risks across populations, such as nasopharyngeal carcinoma[74], alcohol consumption[75], and body strength and appendicular lean mass in East Asians, and skin pigmentation traits in Europeans (**Supplementary Note 4**). More importantly, we found evidence suggesting that EAFD may play an important role in population differences with regard to illnesses such as ADHD, dementia, brain white matter hyperintensity, trans fatty acid levels, and response to hepatitis C treatment (**Table 3**).

The EAFD genes presented in this study were enriched with targets of recent localized adaptation. We compared our approach with an independent approach: the F_{ST}-based population branch statistic (PBS), reported in the latest 1000 Genomes Project paper[76] (**Supplementary Figure 8**). Large PBS scores indicate possible positive selection or local adaptation[77]. The mean PBS score for our EAFD genes (427 of the 434 genes were matched) was 7.2, significantly higher than the mean score from whole-genome(1000 Genomes Project primary paper[76], mean = 4, Wilcoxon rank-sum $P < 2 \times 10^{-16}$) and mean scores of 373 adaptation genes from two well-known studies (mean = 5.1 and $P < 2 \times 10^{-16}$)[57,78]. These results suggest that our approach may be used to complement existing methods for identifying genes under positive selection. Ultimately, to identify putative selection loci among our EAFD loci, independent tests that rely on haplotype structure and frequency[79], or a mixture of independent tests[78] followed by functional analyses, are required.

We hypothesized that the drivers of pathogenic EAFD variants between two related populations were likely balanced selection and recent genetic drift. Firstly, we found that many EAFD variants are associated with both beneficial traits and diseases. For instance, rs1393350 (in intron of *TYR*), an EAFD SNP that we identified in South Asians, is associated with both eye color and melanoma [80,81] (**Supplementary Table 15**). Similarly, rs174547 (in intron of *FADS1*), an EAFD SNP from East Asians and Americans, is associated with both height and trans fatty acid levels[82,83], while rs1042602 (missense in *TYR*), another EAFD locus, is associated with both skin pigmentation and nicotine

dependence[84-87]. **Supplementary Table 15** indicates function of six EAFD SNPs potentially maintained in the population by balanced selection. Secondly, we identified numerous EAFD complex disorders that are non-lethal (e.g., alcohol consumption) or have a late onset (e.g., frontotemporal dementia), which supports the action of genetic drift on genetic loci of little effect on fitness[55]. The EAFD loci identified in this study are likely driven by a combination of balancing selection and genetic drift. Finally, we observed enrichment of EAFD in genic regions (28% versus 2.9% from whole-genome). This enrichment is unlikely biased by higher sequencing coverage in genic versus non-genic regions, since proportion of genic variants represents the actual proportion of gene regions in the genome (~2%).

In this study, we recognized and overcame several challenges and biases. For example, first, we adopted a recently evaluated, non-traditional derivation of the fixation index, F_{ST}[88], such that our approach adequately incorporated the effects of rare variants. Although it was previously shown that this F_{ST} estimator is appropriate for use in sequencing studies with abundance of rare variants[88], it is worth noting that the sample size determines the minimum frequency of alleles that may be analyzed (e.g., we cannot observe alleles with DAF<0.01 if sample size = 100). The F_{ST} estimator used here is robust to sample size, even for rare variants. For example, if allele frequency of a SNP is 0.01 and 0.04 in two populations of size 500 individuals each, the F_{ST} index will be 0.019; if sample sizes from both populations increased two-fold (to 1,000), the F_{ST} index remains 0.019. Second, instead of using a fixed threshold to define EAFD, such as F_{ST} =

0.65 as reported previously[89] we used a dynamic, data-driven approach, which determined reasonable thresholds based on each population, i.e., θ_H (**Supplementary Tables 5** and **6**). The 99.9th percentile of F_{ST} values represents a reasonably high threshold for selection of extremely differentiated alleles; we showed that this threshold resulted in a balanced number of genes, since lower or higher thresholds would have produced markedly higher or lower numbers of genes, respectively (**Supplementary Figure 9**). Third, due to rare variant inclusion and unbiased ascertainment, whole-genome sequencing data has higher power for demographic inference than SNP array data, used by many previous studies[90]. Rare variants have been shown to potentially inflate fixation index, but only marginally so, when ascertaining in the population in the pair (e.g., from 0.103 to 0.108 in CEU-CHB pair when ascertaining in CEU[88]). Fourth, we used four independent, but complementary, analyses to measure biological effects of EAFD, including the VEP[91], CADD[92], KEGG[93], and a curated GWAS Catalogue[60]. We confirmed our variant functional annotation[94] results using ANNOVAR[95].

In summary, we have demonstrated that a large number of genes, diseases, and traits are influenced by functional EAFD loci. We have provided a catalogue of highly distilled EAFD genes with functionally important variants for experimental validation. Future studies may demonstrate that a considerable portion of the genetic missing heritability in some complex human diseases is attributed to EAFD.

Methods

Whole-genome single nucleotide polymorphisms and small insertions and deletions

The SNP and indel data were derived from the most recent Phase 3 release of the 1000 Genomes Project (accessed as of August 20 2014). The program Tabix[96] was used to extract genotypes from the variant call format (VCF version 4.1) files, created using the human genome reference (build 37). The resulting 2,504 unrelated individuals represent a total of 26 world populations (**Supplementary Table 1**). Only autosomal biallelic SNPs and indels were used. All other structural variants (e.g., copy number variants) and multi-allelic variants, which occupied only 0.5% of the total variants, were excluded from the analyses.

Whole-genome fixation index

Fixation index (F_{ST}) is a measurement of genetic differentiation between two populations at a specific genetic locus. The conventional F_{ST} estimation methods by Weir and Cockerham[97] and Weir and Hill[98,99] have been widely used. In this study, we adopted a modified Hudson F_{ST}-estimation method[88] because this method does not overestimate F_{ST} and has adequate power for analysis of both common and rare variants, due to its insensitivity to sample size differences between populations. The latter is important since sample sizes between some populations are not well-matched (**Supplementary Table 1**). This new F_{ST} estimator is defined as:

$$F_{ST} = \frac{(\bar{p}_1 - \bar{p}_2)^2 - \frac{\bar{p}_1(1-\bar{p}_1)}{(n_1-1)} - \frac{\bar{p}_2(1-\bar{p}_2)}{(n_2-1)}}{\bar{p}_1(1-\bar{p}_2) + \bar{p}_2(1-\bar{p}_1)} \qquad (Equation\ 1)$$

, where $\bar{p}_1$ and $\bar{p}_2$ refer to derived allele frequencies (DAF) in samples from populations 1 and 2, and n_1 and n_2 refer to sample sizes of populations 1 and 2, respectively. The fixation index utilizes the DAF to measure allele frequency, instead of minor allele frequency (MAF), since it measures divergence of non-reference (i.e., derived) alleles. Intuitively, this estimator represents an average of the population specific F_{ST} estimators proposed by Weir and Hill[98], and has been shown to be independent of sample composition and not overestimate F_{ST}[88]. Bhatia *et al.* evaluated this estimator at depth and observed that when rare variants were used to calculate F_{ST} between CEU and CHB, F_{ST} was marginally inflated compared to when using common SNPs. The authors attribute this behavior to population bottlenecks being strong in both CEU and CHB, rather than recent population expansion. However, allele frequency dependence was removed when SNPs were ascertained in YRI.

We selected "high-F_{ST}" SNPs and indels based on population pair specific F_{ST} distributions, i.e., the threshold (defined here as θ_H) for high-F_{ST} SNPs was the F_{ST} value at the 99.9th percentile, consistent with previous studies[89]. Therefore, biallelic SNPs and indels with $F_{ST} > \theta_H$ between populations of same continental groups are referred as EAFD SNPs and indels, respectively, or variants (jointly). For computational reasons, the genome-wide θ_H between two populations was estimated as the weighted average $\overline{\theta}_H$ across autosomes:

$$\overline{\theta}_H = \frac{\sum_{i=1}^{22}(n_i \times \theta_{Hi})}{\sum_{i=1}^{22} n_i} \qquad (Equation\ 2)$$

, where n_i is the number of analyzed biallelic SNPs and indels in chromosome *i* (**Supplementary Tables 2 and 3**). The population pair specific θ_H analyzed in this study was the whole-genome θ_H (i.e., $\overline{\theta}_H$). For simplicity, we used the symbol θ_H to represent $\overline{\theta}_H$ throughout the study.

The F_{ST} was evaluated within and between the 26 populations (**Figure 1**). Thus, a total of 325 (i.e., $^{26}C_2 = 325$) pairwise population comparisons were analyzed, including 268 between continental groups pairs and 57 within continental group pairs (**Supplementary Table 1**), resulting in over 4.6 billion calculations of allelic differentiation (i.e., F_{ST}).

Population genetic distances and visualization

The (whole-genome) θ_H value was used to estimate genetic distances between any two populations. The programs, dendroscope 3[100] and circos[101], were adopted to draw population dendrogram and circos plots, respectively. To better visualize differences among population pairs, the θ_H values were exponentially transformed, i.e., $\text{width} = e^{20(1-\theta_H)}$, such that thicker connections correspond to more related populations, while thinner connections correspond to more distant populations. Furthermore, allele sharing was also adopted to evaluate genetic distances. PLINK/SEQ 0.10 (http://atgu.atgu.mgh.harvard.edu/plinkseq) was used to estimate pair-wise allele sharing

for a total of 3,133,756 (i.e., (2504^2 – 2504) / 2) unique sample pairs. The heatmap of resulting allele sharing counts was constructed using the heatmap.2 function in the R statistical programming language (www.r-project.org).

Geographic distances

The geographic distances between populations were determined using the geosphere function in R. The latitude and longitude for the 26 populations were determined using the Google Earth (https://earth.google.com). The center of country, region or city was used to represent the point of origin for each population.

Functional annotations of variants

All SNPs and indels were annotated with potential biological consequence terms. The functional annotation information was extracted from Variant Effector Predictor tool (VEP)[91]. The VEP database contains a total of 34 unique annotation categories, also known as sequence ontologies or SO terms. For comparison, all variants were also annotated using the latest version of ANNOVAR[95] (July 14 2014) and CADD[92]. The evolutionary conservation scores defined by the GERP[102] method were used to evaluate functional impacts of the variants. A positive GERP score represents conservation across mammals, and therefore, the greater the GERP score of the variant, the greater the level of evolutionary conservation at the particular genomic site[102]. Similarly, a high CADD score represents potential pathogenicity.

Linkage disequilibrium

We measured LD between consecutive variants for both the identified EAFD variants and random matched variants, using the African population. For the random matched variants, we randomly sampled variants from the whole genome while controlling for 1) the total number of markers (i.e., 10,000 EAFD and 10,000 random variants) and 2) their derived allele frequency (DAF) distribution. The DAF distribution of the randomly sampled variants had to match that of the EAFD variants (see Results). High LD was defined as $r^2 > 0.8$. We took each variant X and identified the length of genome until we encountered the first other variant Y such that r^2 (X,Y) ≤ 0.8 (defined as "LD range length"). This process led to exclusion of 40%-45% of variants in both EAFD and random matched variants since rare variants have $r^2 < 0.8$. The distributions of LD range lengths were compared between the EAFD and random matched variants. The LD calculations were conducted using SNAP[103].

Enrichment analyses of genes and biological pathways influenced by EAFD

EAFD variants refer to those with $F_{ST} \geq \theta_H$ *within* a continental group. Genic EAFD variants were identified for each of the five continental groups. Gene-level density of EAFD variants, i.e., $\frac{High-F_{ST}\ variants}{Total\ variants}$, was used to establish the genes that were likely influenced by EAFD forces. Specifically, genes under EAFD had to meet three criteria;

they had to: 1) have more than 1% of their variants designated as EAFD variants, 2) be significantly enriched with EAFD variants, and 3) contain at least one nonsynonymous variant. The significance of enrichment was evaluated using the hypergeometric probability model. To ensure minimal type I error (false positives), Bonferroni correction for multiple testing was set at 1.7×10^{-6} (0.05/29,061, the total number of genes). We only retained genes above genome-wide significant level for further analyses.

To identify EAFD enrichment in biological or disease pathways maintained in the latest version of Kyoto Encyclopedia of Genes and Genomes (KEGG)[93], the toolkit WebGestalt[104] was used. The statistical significance of enrichment was evaluated under the hypergeometric probability of the overlap between our gene sets and all the gene sets in the KEGG database (accessed on June 10, 2016). To correct for multiple testing, enrichment *P* values were adjusted using the false discovery rate (FDR) method[105].

Enrichment analyses of diseases and traits influenced by EAFD

We curated the genome-wide association findings of the latest GWAS Catalogue[60] (accessed on November 13, 2015) using multiple criteria (see Result section for a full list of quality controls). We mapped the EAFD variants that we identified to the curated GWAS Catalogue, and carried out enrichment analyses. The hypergeometric probability was used to calculate the statistical significance of enrichment. For instance, for each disease or trait *d*, there are l_d variants in the curated GWAS catalogue (set *G*) and k_d of

them are under EAFD (set *F*). The sizes of sets *F* and *G* are *m* and *n*, respectively. The null hypothesis H_0 states that the ratio of expected size of set *F* for disease *d* (i.e., $E(|F(k_d)|)$) to observed size (i.e., $O(|F(k_d)|)$) is *at most* equal to one: $\frac{E(|F(k_d)|)}{O(|F(k_d)|)} \leq 1$. The probability *P* of observing enrichment for disease *d* in set *F* (i.e., $\frac{E(k_d)}{O(k_d)} > 1$) is estimated using the hypergeometric probability distribution function:

$$P\left(\frac{E(|F(k_d)|)}{O(|F(k_d)|)} > 1 \mid H_0\right) = 1 - P\left(\frac{E(|F(k_d)|)}{O(|F(k_d)|)} \leq 1 \mid H_0\right)$$

$$= 1 - \sum_{i=0}^{k_d-1} \frac{\binom{l_d}{i} \cdot \binom{n-l_d}{m-i}}{\binom{n}{m}}$$

$$= \sum_{i=k_d}^{m} \frac{\binom{l_d}{i} \cdot \binom{n-l_d}{m-i}}{\binom{n}{m}} \qquad \text{(Equation 3)}$$

, where the expected number of variants for disease *d* is: $E(|F(k_d)|) = \frac{l_d}{n} \times m$.

Intuitively, $\frac{E(|F(k_d)|)}{O(|F(k_d)|)}$ represents the ratio between expected and observed disease-associated variants, taking values <1 in case of depletion and >1 in case of enrichment. The resulting *P* value was adjusted using the FDR method[105].

Acknowledgements

This work was supported by the Start-up Fund of The University of Vermont. The data used in this study were from the 1000 Genomes Project (Phase 3), the KEGG database, the GWAS catalogue (November 13th, 2015), and annotation databases of VEP and

ANNOVAR. We would like to thank Drs. Gonçalo Abecasis and Adam Auton for providing us with the gene-wide PBS statistics information. We would like to thank Drs. Hongyu Zhao and Joel Gelernter for their careful reviews of the manuscript. We are grateful to Dr. Xun Chen and Guangchen Liu for their critical comments and feedback throughout the process of preparing this manuscript. Finally, we thank the anonymous reviewers for their constructive comments. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors declare no potential conflict of interest.

References

Arnold M, Soerjomataram I, Ferlay J, Forman D. 2015. Global incidence of oesophageal cancer by histological subtype in 2012. Gut 64(3):381-7.

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. Nature 526(7571):68-74.

Baik I, Cho NH, Kim SH, Han BG, Shin C. 2011. Genome-wide association studies identify genetic loci related to alcohol consumption in Korean men. Am J Clin Nutr 93(4):809-16.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. Nat Genet 40(3):340-5.

Barrett JH, Iles MM, Harland M, Taylor JC, Aitken JF, Andresen PA, Akslen LA, Armstrong BK, Avril MF, Azizi E and others. 2011. Genome-wide association study identifies three new melanoma susceptibility loci. Nat Genet 43(11):1108-13.

Baughman RP, Teirstein AS, Judson MA, Rossman MD, Yeager H, Jr., Bresnitz EA, DePalo L, Hunninghake G, Iannuzzi MC, Johns CJ and others. 2001. Clinical characteristics of patients in a case control study of sarcoidosis. Am J Respir Crit Care Med 164(10 Pt 1):1885-9.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57(1):289-300.
Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74(6):1111-20.
Bertoni JM, Arlette JP, Fernandez HH, Fitzer-Attas C, Frei K, Hassan MN, Isaacson SH, Lew MF, Molho E, Ondo WG and others. 2010. Increased melanoma risk in Parkinson disease: a prospective clinicopathological study. Arch Neurol 67(3):347-52.
Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting FST: the impact of rare variants. Genome Res 23(9):1514-21.
Boyle P. 2012. Triple-negative breast cancer: epidemiological considerations and recommendations. Ann Oncol 23 Suppl 6:vi7-12.
Briscoe VJ, Tate DB, Davis SN. 2007. Type 1 diabetes: exercise and hypoglycemia. Appl Physiol Nutr Metab 32(3):576-82.
Carrigan MA, Uryasev O, Frye CB, Eckman BL, Myers CR, Hurley TD, Benner SA. 2015. Hominids adapted to metabolize ethanol long before human-directed fermentation. Proc Natl Acad Sci U S A 112(2):458-63.
Castren E, Tanila H. 2006. Neurotrophins and dementia--keeping in touch. Neuron 51(1):1-3.
Chang ET, Adami HO. 2006. The enigmatic epidemiology of nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev 15(10):1765-77.
Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15(7):901-13.
Currier RL, Payne DC, Staat MA, Selvarangan R, Shirley SH, Halasa N, Boom JA, Englund JA, Szilagyi PG, Harrison CJ and others. 2015. Innate Susceptibility to Norovirus Infections Influenced by FUT2 Genotype in a United States Pediatric Population. Clin Infect Dis 60(11):1631-8.
Dudley JT, Kim Y, Liu L, Markov GJ, Gerold K, Chen R, Butte AJ, Kumar S. 2012. Human genomic disease variants: a neutral evolutionary explanation. Genome Res 22(8):1383-94.
Ezzati M, Riboli E. 2013. Behavioral and dietary risk factors for noncommunicable diseases. N Engl J Med 369(10):954-64.
Fraser HB. 2013. Gene expression drives local adaptation in humans. Genome Res 23(7):1089-96.
Garcia-Barcelo MM, Yeung MY, Miao XP, Tang CS, Cheng G, So MT, Ngan ES, Lui VC, Chen Y, Liu XL and others. 2010. Genome-wide association study identifies a susceptibility locus for biliary atresia on 10q24.2. Hum Mol Genet 19(14):2917-25.
Genetic Analysis of Psoriasis C, the Wellcome Trust Case Control C, Strange A, Capon F, Spencer CC, Knight J, Weale ME, Allen MH, Barton A, Band G and others.

2010. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet 42(11):985-90.
Genome of the Netherlands C. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46(8):818-25.
Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA and others. 2015. A global reference for human genetic variation. Nature 526(7571):68-74.
Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, Bowden DW, Langefeld CD, Oleksyk TK, Uscinski Knob AL and others. 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science 329(5993):841-5.
Greenwood BM, Bradley AK, Wall RA. 1985. Meningococcal disease and season in sub-Saharan Africa. Lancet 2(8459):829-30.
Gregersen JW, Kranc KR, Ke X, Svendsen P, Madsen LS, Thomsen AR, Cardon LR, Bell JI, Fugger L. 2006. Functional epistasis on a common MHC haplotype associated with multiple sclerosis. Nature 443(7111):574-7.
Gronberg H. 2003. Prostate cancer epidemiology. Lancet 361(9360):859-64.
Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH and others. 2013a. Identifying recent adaptations in large-scale genomic data. Cell 152(4):703-13.
Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH and others. 2013b. Identifying recent adaptations in large-scale genomic data. Cell 152(4):703-13.
Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O and others. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327(5967):883-6.
Grossmann A, Benlasfer N, Birth P, Hegele A, Wachsmuth F, Apelt L, Stelzl U. 2015. Phospho-tyrosine dependent protein-protein interaction network. Mol Syst Biol 11(3):794.
Hamblin MT, Thompson EE, Di Rienzo A. 2002. Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70(2):369-83.
Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, Hankinson SE, Hu FB, Duffy DL, Zhao ZZ and others. 2008. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet 4(5):e1000074.
He M, Xu M, Zhang B, Liang J, Chen P, Lee JY, Johnson TA, Li H, Yang X, Dai J and others. 2015. Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. Hum Mol Genet 24(6):1791-800.
Heilmann S, Kiefer AK, Fricker N, Drichel D, Hillmer AM, Herold C, Tung JY, Eriksson N, Redler S, Betz RC and others. 2013. Androgenetic alopecia: identification of four genetic risk loci and evidence for the contribution of WNT signaling to its etiology. J Invest Dermatol 133(6):1489-96.

Hoglinger GU, Melhem NM, Dickson DW, Sleiman PM, Wang LS, Klei L, Rademakers R, de Silva R, Litvan I, Riley DE and others. 2011. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. Nat Genet 43(7):699-705.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet 10(9):639-50.

Hradetzky S, Balaji H, Roesner LM, Heratizadeh A, Mittermann I, Valenta R, Werfel T. 2013. The human skin-associated autoantigen alpha-NAC activates monocytes and dendritic cells via TLR-2 and primes an IL-12-dependent Th1 response. J Invest Dermatol 133(9):2289-92.

Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61(6):1061-7.

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24(24):2938-9.

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H and others. 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152(4):691-702.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42(Database issue):D199-205.

Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. Natural selection and infectious disease in human populations. Nat Rev Genet 15(6):379-93.

Key FM, Teixeira JC, de Filippo C, Andres AM. 2014. Advantageous diversity maintained by balancing selection in humans. Curr Opin Genet Dev 29:45-51.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310-5.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. Genome Res 19(9):1639-45.

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE and others. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310(5755):1782-6.

Lao O, Andres AM, Mateu E, Bertranpetit J, Calafell F. 2003. Spatial patterns of cystic fibrosis mutation spectra in European populations. Eur J Hum Genet 11(5):385-94.

Lee JW, Brancati FL, Yeh HC. 2011. Trends in the prevalence of type 2 diabetes in Asians versus whites: results from the United States National Health Interview Survey, 1997-2008. Diabetes Care 34(2):353-7.

Li H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27(5):718-9.

Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, Berrettini W, Knouff CW, Yuan X, Waeber G and others. 2010. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 42(5):436-40.
Love-Gregory L, Sherva R, Sun L, Wasson J, Schappe T, Doria A, Rao DC, Hunt SC, Klein S, Neuman RJ and others. 2008. Variants in the CD36 gene associate with the metabolic syndrome and high-density lipoprotein cholesterol. Hum Mol Genet 17(11):1695-704.
Madhava V, Burgess C, Drucker E. 2002. Epidemiology of chronic hepatitis C virus infection in sub-Saharan Africa. Lancet Infect Dis 2(5):293-302.
Malnic B, Godfrey PA, Buck LB. 2004. The human olfactory receptor gene family. Proc Natl Acad Sci U S A 101(8):2584-9.
McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, Donnelly P, Consortium W. 2014. Choice of transcripts and software has a large effect on variant annotation. Genome Medicine 6.
McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J and others. 2008. Lung cancer susceptibility locus at 5p15.33. Nat Genet 40(12):1404-6.
McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26(16):2069-70.
Mozaffarian D, Kabagambe EK, Johnson CO, Lemaitre RN, Manichaikul A, Sun Q, Foy M, Wang L, Wiener H, Irvin MR and others. 2015. Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. Am J Clin Nutr 101(2):398-406.
Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, Gudjonsson JE, Li Y, Tejasvi T, Feng BJ and others. 2009. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet 41(2):199-204.
Nikolaou V, Stratigos AJ. 2014. Emerging trends in the epidemiology of melanoma. Br J Dermatol 170(1):11-9.
Noguchi K, Okumura F, Takahashi N, Kataoka A, Kamiyama T, Todo S, Hatakeyama S. 2011. TRIM40 promotes neddylation of IKKgamma and is downregulated in gastrointestinal cancers. Carcinogenesis 32(7):995-1004.
Ober C, Yao TC. 2011. The genetics of asthma and allergic disease: a 21st century perspective. Immunol Rev 242(1):10-30.
Ordonez D, Sanchez AJ, Martinez-Rodriguez JE, Cisneros E, Ramil E, Romo N, Moraru M, Munteis E, Lopez-Botet M, Roquer J and others. 2009. Multiple sclerosis associates with LILRA3 deletion in Spanish patients. Genes Immun 10(6):579-85.
Platt OS, Brambilla DJ, Rosse WF, Milner PF, Castro O, Steinberg MH, Klug PP. 1994. Mortality in sickle cell disease. Life expectancy and risk factors for early death. N Engl J Med 330(23):1639-44.
Popovic D, Dikic I. 2014. TBC1D5 and the AP2 complex regulate ATG9 trafficking and initiation of autophagy. EMBO Rep 15(4):392-401.

Qiu C, Kivipelto M, von Strauss E. 2009. Epidemiology of Alzheimer's disease: occurrence, determinants, and strategies toward intervention. Dialogues Clin Neurosci 11(2):111-28.
Ralston SH. 2013. Clinical practice. Paget's disease of bone. N Engl J Med 368(7):644-50.
Risch N, Tang H, Katzenstein H, Ekstein J. 2003. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. Am J Hum Genet 72(4):812-22.
Robert-Gangneux F, Darde ML. 2012. Epidemiology of and diagnostic strategies for toxoplasmosis. Clin Microbiol Rev 25(2):264-96.
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ and others. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832-7.
Scally SW, Petersen J, Law SC, Dudek NL, Nel HJ, Loh KL, Wijeyewickrema LC, Eckle SB, van Heemst J, Pike RN and others. 2013. A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis. J Exp Med 210(12):2569-82.
Soler Artigas M, Loth DW, Wain LV, Gharib SA, Obeidat M, Tang W, Zhai G, Zhao JH, Smith AV, Huffman JE and others. 2011. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet 43(11):1082-90.
Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G and others. 2007. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet 39(12):1443-52.
Taft RJ, Vanderver A, Leventer RJ, Damiani SA, Simons C, Grimmond SM, Miller D, Schmidt J, Lockhart PJ, Pope K and others. 2013. Mutations in DARS cause hypomyelination with brain stem and spinal cord involvement and leg spasticity. Am J Hum Genet 92(5):774-80.
Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A and others. 2008. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 452(7187):638-42.
Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J and others. 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293(5529):455-62.
Tobacco, Genetics C. 2010. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet 42(5):441-7.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol 4(3):e72.
Wang J, Duncan D, Shi Z, Zhang B. 2013. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res 41(Web Server issue):W77-83.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164.
Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population-Structure. Evolution 38(6):1358-1370.
Weir BS, Hill WG. 2002. Estimating F-statistics. Annu Rev Genet 36:721-50.
Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L and others. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42(Database issue):D1001-6.
Wong GK, Yang Z, Passey DA, Kibukawa M, Paddock M, Liu CR, Bolund L, Yu J. 2003. A population threshold for functional polymorphisms. Genome Res 13(8):1873-9.
Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z and others. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46(11):1173-86.
Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS and others. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329(5987):75-8.
Zhang M, Song F, Liang L, Nan H, Zhang J, Liu H, Wang LE, Wei Q, Lee JE, Amos CI and others. 2013. Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. Hum Mol Genet 22(14):2948-59.

Tables

Table 3 Annotation of EAFD genes containing most pathogenic EAFD variants. (see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y)

Table 4 Replicated pathways influenced by EAFD. (see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y)

Table 5 Diseases and traits influenced by EAFD. (see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y)

Figure Legends

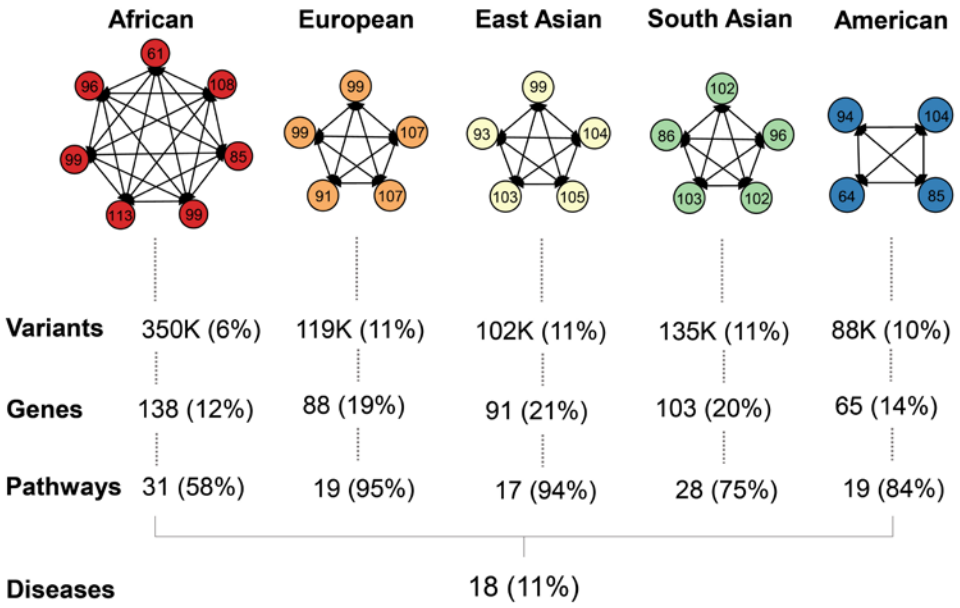


Figure 1 Identification of EAFD targets. From top to bottom: for each of the five continental groups, a population is represented by a circle, and the number inside the circle is the sample size. All possible population pairs, represented by solid black arrows, were considered in each continental group (21, 10, 10, 10, and 6 from the left to right), and for each pair, around 81 million loci (i.e., biallelic SNPs and indels) were used for fixation index (F_{ST}) estimation. For comparison, the average F_{ST} thresholds (averaged values of combined θ_H in populations of the same continental group) were 0.1, 0.11, 0.07, 0.09, and 0.2 for African, East Asian, South Asian, European, and American, respectively. An EAFD variant candidate is selected if its F_{ST} value is larger than θ_H. The overall range of the resulting F_{ST} values was consistent with previous estimates from HapMap samples[88,89]. Numbers in the following four rows are total numbers of identified variants, genes, pathways, and traits and diseases. The numbers in brackets correspond to the percentages of the population counts shared by at least one other population.

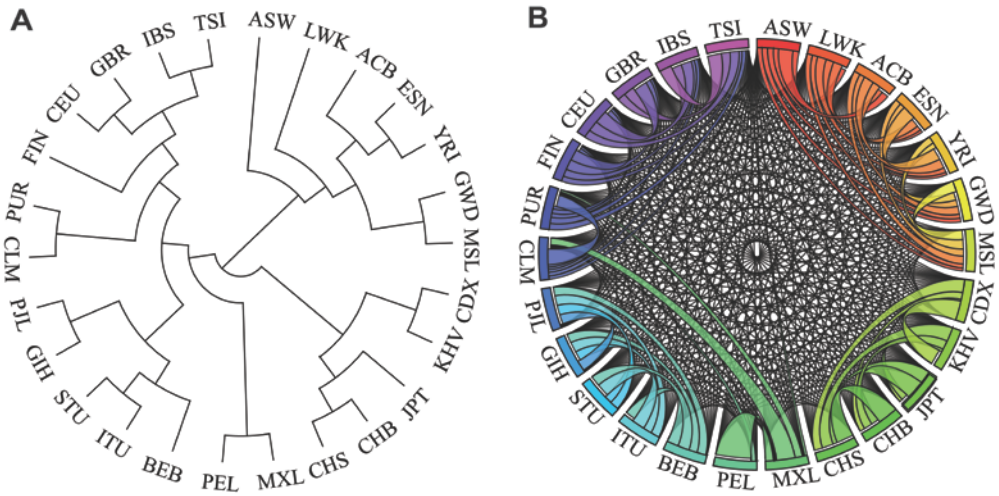


Figure 2 EAFD and population structure. A) Dendrogram of all populations defined by the θ_H values between population pairs. **B)** Circos plot of population pairwise θ_H for all of the 325 population pairs. Each colored segment corresponds to a different population. The width of each ribbon (i.e., connection) corresponds to an exponential transformation of the whole-genome θ_H value. The thicker the ribbon, the more similar the two populations.

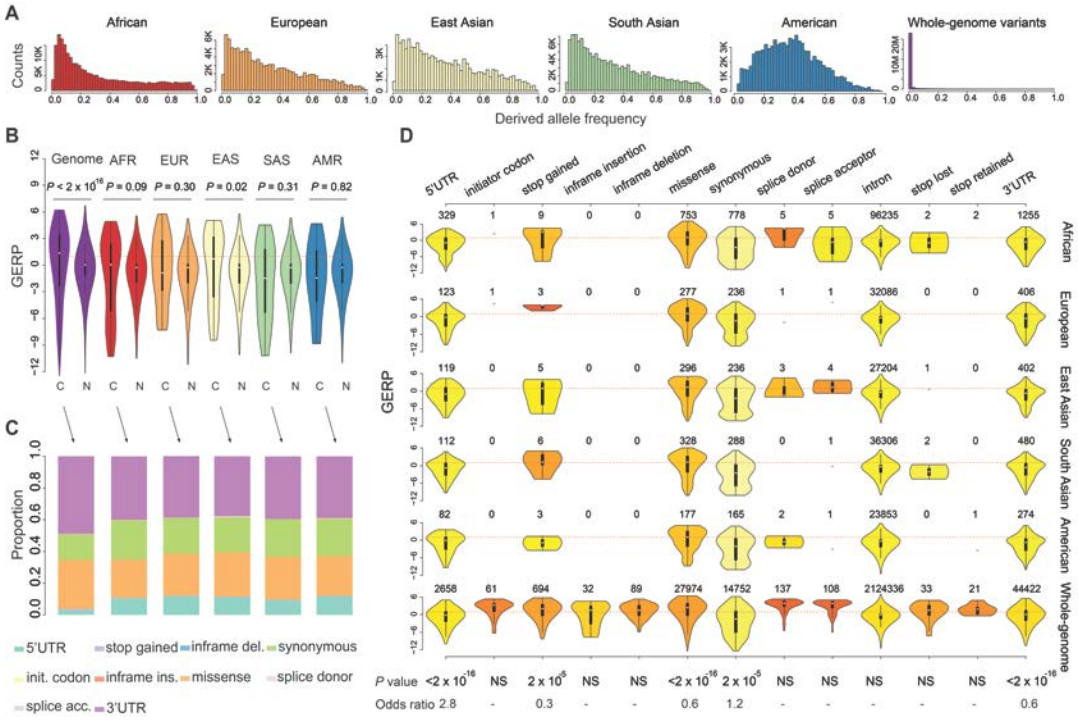


Figure 3 EAFD and functional annotation. A) Distributions of derived allele frequencies of EAFD variants identified within each continental group (K, thousand and M, million). By comparison, the high-F_{ST} variants identified *between* continental groups revealed more common variants. **B)** Comparison of GERP scores between coding (C) and non-coding (N) variants. From left to right: the whole-genome variants, the EAFD variants in African, European, East Asian, South Asian, and American. The *P* values were estimated using the rank sum test. The dotted red horizontal line represents GERP score of 1. **C)** Decomposition of functional categories of the coding variants under EAFD in each continental group, corresponding to **B. D)** Violin plots with y-axis representing GERP scores and x-axis indicating variant functional categories. The five rows of violin plots correspond to five different continental groups. The x-axis labels, variant functional categories, were ordered based on their relative genomic positions. The colors of the violin plots correspond to the degrees of mean GERP scores (yellow = low and red = high). The numbers on top of the violins correspond to the total number of variants represented by each violin. Around 93% of all variants had a GERP score. The boxplot in the center of the violin shows the quantiles, with the grey dot in the center being the median value. The two rows at the bottom show the *P* value and odds ratio (OR) from tests between EAFD and whole-genome variant counts (χ^2 test) using data from African samples. An OR < 1 represents depletion, while OR > 1 represents enrichment. NS, not significant.

Supplementary Results, Tables, and Figures

Supplementary Results

Supplementary Note 1

Associations between genetic and geographic distances

We found evidence of strong association between the populations' genetic (θ_H) and geographic distances. Particularly, when adjusted for recent migratory events, the geographic distance well reflected the genetic distances between populations. The association was moderate ($\rho = 0.67$ and $R^2 = 0.45$) when all African populations were combined; however, it became stronger ($\rho = 0.93$ and $R^2 = 0.87$) when the two admixed populations, ACB and ASW, were excluded. The admixed Americans showed no significant association, likely due to both recent migration and admixture patterns. A stronger association unraveled among Europeans after removal of the CEU samples ($\rho = 0.77$ and $R^2 = 0.6$). These results showed that θ_H was a reasonable proxy for genetic distances between populations.

Supplementary Note 2

Whole-genome scan for EAFD

The number of total high-F_{ST} variants decreased as the populations in pairs became distant, i.e., increased θ_H ($\rho = -0.52$ and $P = 0.05$, **Supplementary Table 9**). Population pairs CDX-ESN and ESN-YRI had the largest and smallest θ_H values, respectively. **Supplementary Table 9** shows three main population pairs while **Supplementary Table 10** shows θ_H for each of the 325 population pairs. All of the 268 between continental group pairs had $\theta_H > 0.1$ while the 57 within continental group pairs showed $\theta_H < 0.1$.

The variants with highest allele frequency difference between populations of the same continental group may, thus, be considered candidates of EAFD.

Supplementary Note 3

Functional annotation of EAFD variants

Genic annotation groups were affected equally by EAFD across the five continental groups (**Figure 3D**). However, when within- and between-population pairs were considered jointly (a total of 15 population pairs), the functional abundance differed across pairs. We carried out association analyses between the abundance of 14 common, variant functional annotation categories and θ_H values for each continental group pair (**Supplementary Table 9**). We found that the abundance significantly decreased ($P \leq 0.05$) as θ_H increased for several categories, particularly missense ($\rho = -0.73$), missense NMD ($\rho = -0.55$), intergenic ($\rho = -0.54$), upstream ($\rho = -0.51$), splice donor ($\rho = -0.51$), and 5'UTR ($\rho = -0.44$). After adjusting for multiple testing, the association remained significant for missense variants ($\rho = -0.73$ and adjusted $P = 0.03$). The ratio of missense/synonymous variants significantly associated with θ_H ($\rho = -0.77$ and adjusted $P = 0.01$). To our knowledge, this is the first report of a dose-dependent decrease in missense SNP abundance in response to increase in population genetic distance.

We estimated the number of times that the same high-F_{ST} variant was identified in different population pairs (i.e., recurrence). High recurrence is observed when a high-F_{ST} variant is observed multiple times in different population pairs, signifying that this variant is more likely a candidate of EAFD. In pairs of populations from different

continental groups, the recurrence (adjusted by number of samples in the population pair) was significantly higher than in pairs from the same continental group (KS-test, P = 0.0025; **Supplementary Table 7**). Thus, high-F_{ST} variants within continental groups are better candidates of EAFD that those between continental groups. This observation strongly supports the approach used in this study, where EAFD referred to within (but not between) continental group pairs.

Supplementary Note 4

Population-specific traits and diseases of high-prevalence or pathogen exposures

The findings from our enrichment analyses of the diseases and traits under EAFD were consistent with epidemiological reports, and reflected the published reports on disease or trait prevalence or exposure, for instance, in Africans: asthma[106], prostate cancer[107], breast cancer[108], chronic hepatitis infections (B and C)[109], African trypanosomiasis (WHO, March 2014), Malaria[110], Nephropathy[52], pathogenic E. *coli* infection[110], meningococcal disease[111], sickle cell anemia (haemolysis)[112], toxoplasmosis[113], sarcoidosis[114], AIDS[110]; in Europeans: Alzheimer's disease[115], Parkinson's disease[116], eye, hair and skin-color traits[81], male-pattern baldness[117], melanoma[118], Paget's disease[119], and cystic fibrosis severity[120]; in East Asians: biliary atresia[121], hepatitis B and C infections (WHO, March 2014), esophageal cancer (and related nasopharyngeal carcinoma)[122], type II diabetes (and related trait: retinol metabolism)[123]; in South Asians: type 2 diabetes and related trait, i.e., insulin signaling pathways[123] (most of the epidemiological literature grouped East and South Asians).

Supplementary Tables

Supplementary Table 1 Sample sizes of all 26 populations analyzed in this study

Continental groups	Populations	Sample sizes	Total
AFR	ACB, ASW, ESN, GWD, LWK, MSL, YRI	96, 61, 99, 113, 99, 85, 108	661
AMR	CLM, MXL, PEL, PUR	94, 64, 85, 104	347
EAS	CDX, CHB, CHS, JPT, KHV	93, 103, 105, 104, 99	504
EUR	CEU, FIN, GBR, IBS, TSI	99, 99, 91, 107, 107	503
SAS	BEB, GIH, ITU, PJL, STU	86, 103, 102, 96, 102	489

The three-letter codes represent the following populations: EAS, East Asian; SAS, South Asian; AMR, admixed populations from the Americas; EUR, European populations; AFR, African populations. The order of sample sizes corresponds to the populations order. Population codes correspond to African Carribbeans in Barbados (ACB); Americans of African Ancestry in Southwest of USA (ASW); Esan in Nigeria (ESN); Gambian in Western Divisions in the Gambia (GWD); Luhya in Webuye, Kenya (LWK); Mende in Sierra Leone (MSL); Yoruba in Ibadan, Nigeria (YRI); Columbians from Medellin, Colombia (CLM); Mexican Ancestry from Los Angeles USA (MXL); Peruvians from Lima, Peru (PEL); Puerto Ricans from Puerto Rico (PUR); Chinese Dai in Xishuangbanna, China (CDX); Han Chinese in Beijing, China (CHB); Southern Han Chinese (CHS); Japanese in Tokyo, Japan (JPT); Kinh in Ho Chi Minh City, Vietnam (KHV); Utah residents with Northern and Western European Ancestry (CEU); Finnish in Finland (FIN); British in England and Scotland (GBR); Iberian Population in Spain (IBS); Toscani in Italy (TSI); Bengali from Bangladesh (BEB); Gujarati Indian from Houston, Texas (GIH); Indian Telugu from the UK (ITU); Punjabi from Lahore, Pakistan (PJL); Sri Lankan Tamil from the UK (STU).

Supplementary Table 2 Summary of the total variants in the 1000 Genomes Project Phase 3 subjects

Variant Types	Counts	Percentages
SNPs	78,136,341	96.1%
Indels	3,135,424	3.9%
Biallelic SNPs and indels	80,800,311	99.4%

Multiallelic SNPs	259,370	0.3%
Multiallelic sites	416,023	0.5%
Others	58,671	0.1%
All	81,271,745	100%

Supplementary Table 3 Numbers of biallelic SNPs and indels by chromosomes

Chromosomes	**SNPs**	**indels**	**SNPs + indels**
1	6,196,151	236,961	6,433,722
2	6,786,300	256,128	7,043,032
3	5,584,397	214,796	5,799,690
4	5,480,936	217,939	5,699,315
5	5,037,955	197,094	5,235,493
6	4,800,101	194,243	4,994,802
7	4,517,734	171,699	4,689,864
8	4,417,368	152,173	4,569,905
9	3,414,848	124,884	3,540,028
10	3,823,786	145,438	3,969,564
11	3,877,543	144,615	4,022,530
12	3,698,099	147,887	3,762,572
13	2,727,881	113,548	2,841,649
14	2,539,149	100,450	2,639,834
15	2,320,474	90,444	2,411,151
16	2,596,072	84,920	2,681,201
17	2,227,080	88,730	2,316,023
18	2,171,378	82,671	2,254,259
19	1,751,878	69,034	1,821,116
20	1,739,315	63,315	1,802,809
21	1,054,447	43,974	1,098,537

22	1,055,454	41,022	1,096,558
Total	77,818,346	2,981,965	80,800,311

The average length of indels were three nucleotides (minimum length = 1 and maximum length = 60).

Supplementary Table 4 F_{ST} values across each chromosome in three representative population pairs

Chrs	**CEU-YRI**	**CEU-CHB**	**YRI-CHB**
	μ , se (θ_H)	**μ , se (θ_H)**	**μ , se (θ_H)**
1	0.056, 6.4×10^{-5} (0.67)	0.044, 6.9×10^{-5} (0.62)	0.061, 7.1×10^{-5} (0.73)
2	0.056, 6.1×10^{-5} (0.66)	0.045, 6.8×10^{-5} (0.64)	0.062, 7.0×10^{-5} (0.75)
3	0.057, 6.7×10^{-5} (0.66)	0.044, 7.1×10^{-5} (0.56)	0.062, 7.6×10^{-5} (0.76)
4	0.058, 6.8×10^{-5} (0.68)	0.043, 6.8×10^{-5} (0.56)	0.063, 7.5×10^{-5} (0.75)
5	0.055, 6.8×10^{-5} (0.66)	0.043, 7.2×10^{-5} (0.58)	0.06, 7.6×10^{-5} (0.74)
6	0.056, 6.8×10^{-5} (0.65)	0.043, 7.1×10^{-5} (0.58)	0.061, 7.8×10^{-5} (0.73)
7	0.056, 7.2×10^{-5} (0.67)	0.044, 7.6×10^{-5} (0.55)	0.061, 8.1×10^{-5} (0.74)
8	0.058, 7.8×10^{-5} (0.67)	0.042, 7.6×10^{-5} (0.54)	0.063, 8.6×10^{-5} (0.74)
9	0.056, 8.4×10^{-5} (0.66)	0.045, 9.2×10^{-5} (0.59)	0.060, 9.2×10^{-5} (0.71)
10	0.056, 7.9×10^{-5} (0.66)	0.046, 8.7×10^{-5} (0.60)	0.061, 8.6×10^{-5} (0.74)

11	0.056, 7.8×10^{-5} (0.65)	0.042, 8.0×10^{-5} (0.56)	0.060, 8.7×10^{-5} (0.74)
12	0.057, 8.4×10^{-5} (0.70)	0.047, 9.2×10^{-5} (0.62)	0.061, 9.1×10^{-5} (0.74)
13	0.055, 9.1×10^{-5} (0.65)	0.044, 0.0001 (0.60)	0.062, 0.00011 (0.71)
14	0.057, 0.0001 (0.68)	0.045, 0.0001 (0.56)	0.06, 0.00011 (0.72)
15	0.059, 0.00011 (0.73)	0.046, 0.00011 (0.63)	0.061, 0.00012 (0.72)
16	0.056, 9.9×10^{-5} (0.66)	0.043, 0.0001 (0.60)	0.061, 0.00011(0.74)
17	0.058, 0.00011 (0.74)	0.043, 0.00011 (0.58)	0.063, 0.00012 (0.81)
18	0.057, 0.00011 (0.66)	0.041, 0.0001 (0.50)	0.061, 0.00011 (0.70)
19	0.057, 0.00011 (0.66)	0.043, 0.00012 (0.56)	0.061, 0.00013 (0.71)
20	0.057, 0.00012 (0.68)	0.043, 0.00013 (0.62)	0.063, 0.00014 (0.76)
21	0.058, 0.00015 (0.65)	0.043, 0.00015 (0.58)	0.063, 0.00017 (0.71)
22	0.058, 0.00015 (0.64)	0.045, 0.00016 (0.57)	0.065, 0.00018 (0.77)

μ, the mean F_{ST} values across the entire chromosome, which are consistent with F_{ST} values reported previously[88,89].
se, the standard error, which increases as the length of chromosomes decreases.
$\theta_{H,}$ the threshold for identifying EAFD SNPs, which is the 99.9^{th} percentile of all F_{ST} values for a given population pair.

Supplementary Table 5 Estimates of θ_H values for each population pair on both chromosome- and genome-wide levels
(see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y).

Supplementary Table 6 Pair-wide physical distances and corresponding θ_H values (see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y)

Supplementary Table 7 Recurrence of genome-wide EAFD variants

Population pairs	**Total variants/sample**	**Recurrent variants/sample**	**Recurrence rates (%)**
Within-population			
AFR-AFR	403	133	33%
EUR-EUR	153	32	21%
SAS-SAS	174	33	19%
EAS-EAS	144	40	28%
AMR-AMR	168	40	24%
Between-population			
AFR-EUR	760	687	90%
AFR-SAS	806	727	90%
AFR-EAS	764	697	91%
AFR-AMR	692	585	85%
EUR-SAS	438	330	75%
EUR-EAS	425	367	86%
EUR-AMR	426	267	63%
SAS-EAS	447	371	83%
SAS-AMR	475	313	66%
EAS-AMR	453	343	76%

Recurrence rate refers to the percentage of recurrent variants (i.e. observed more than once in the corresponding population pairs) relative to the total EAFD variants. The numbers of variants were normalized by total sample size in the corresponding population pairs. As expected, the levels of recurrent variants were higher between than within continental groups, most likely due to common (between) *versus* rare frequency (within) variants by EAFD. Specifically, within African populations we observed the largest counts of genomic variants targeted by EAFD, as well as the highest level of recurrent EAFD (33%). Between continental groups, African - South Asian pair contained the largest count of EAFD variants and the African-East Asian pair contained the highest levels of recurrent variants.

Supplementary Table 8 The number of unique or shared EAFD variants across different continental groups

Population set	**No. EAFD variants without LD correction (no. unique variants)**	**No. EAFD variants with LD correction (no. unique variants)**
AFR	356,846 (333,814)	219,988 (209,874)
EUR	121,970 (106,577)	70,739 (64,020)
EAS	104,420 (90,856)	58,482 (52,835)
SAS	138,129 (121,609)	81,481 (73,881)
AMR	88,960 (78,537)	54,690 (49,870)
AFR ∩ EUR	6,623 (5,449)	2,828 (2,367)
AFR ∩ EAS	5,312 (4,278)	2,017 (1,537)
AFR ∩ SAS	7,546 (6,399)	3,593 (3,083)
AFR ∩ AMR	3,551 (2,814)	1,676 (1,319)
EUR ∩ EAS	3,167 (2,327)	1,400 (1,054)

EUR ∩ SAS	3,276 (2,503)	1,489 (1,189)
EUR ∩ AMR	2,327 (1,742)	1,002 (775)
EAS ∩ SAS	3,119 (2,412)	1,306 (980)
EAS ∩ AMR	1,966 (1,413)	924 (660)
SAS ∩ AMR	2,579 (2,104)	1,212 (916)
AFR ∩ EUR ∩ EAS	424 (308)	197 (136)
AFR ∩ EUR ∩ SAS	486 (398)	173 (122)
AFR ∩ EUR ∩ AMR	264 (161)	91 (37)
AFR ∩ EAS ∩ SAS	399 (269)	177 (91)
AFR ∩ EAS ∩ AMR	211 (67)	106 (56)
AFR ∩ SAS ∩ AMR	262 (203)	160 (81)
EUR ∩ EAS ∩ SAS	191 (118)	70 (23)
EUR ∩ EAS ∩ AMR	225 (138)	79 (29)
EUR ∩ SAS ∩ AMR	96 (37)	57 (17)
EAS ∩ SAS ∩ AMR	117 (16)	79 (4)
AFR ∩ EUR ∩ EAS ∩ SAS	51 (30)	29 (13)
AFR ∩ EUR ∩ EAS ∩ AMR	65 (44)	32 (16)
AFR ∩ EUR ∩ SAS ∩ AMR	37 (16)	22 (6)
EUR ∩ EAS ∩ SAS ∩ AMR	22 (1)	18 (2)
AFR ∩ EAS ∩ SAS ∩ AMR	79 (58)	57 (41)
AFR ∩ EUR ∩ EAS ∩ SAS ∩ AMR	21 (21)	16 (16)

Linkage disequilibrium (LD) correction was done by keeping only one EAFD variant within a window of 1,000bp.

Supplementary Table 9 Percentages of EAFD SNPs in different variant functional categories (see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y)

Supplementary Table 10 List of the 805 nonsynonymous EAFD variants found within 434 EAFD genes (see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y)

Supplementary Table 11 Results of gene enrichment analyses (see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y)

Supplementary Table 12 Results of KEGG pathway enrichment analysis (see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y)

Supplementary Table 13 GWAS catalogue quality control procedure

Filtering steps	**Before QC (number of SNPs)**	**After QC (number of SNPs)**
Pre-filtering	-	22,895
Non-missing P-values	22,895	22,521
GWAS significance $P \leq 5 \times 10^{-5}$	22,895	22,500
Replicated association only	22,500	10,168
rs SNP IDs only	10,168	10,120
Non-missing SNP IDs	10,120	10,118
Unique Disease-SNP pair	10,118	8,690

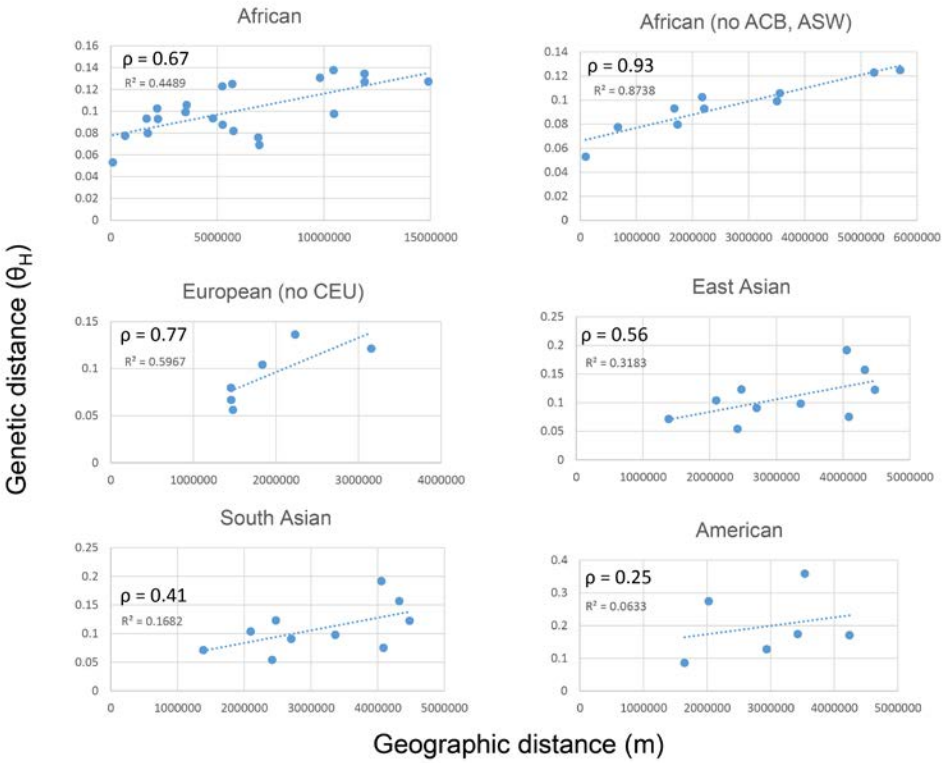
After quality control, the curated GWAS catalogue data contained information on 726 traits and 7,523 SNPs from 1,313 unique publications.

Supplementary Table 14 Results of enrichment analyses using EAFD SNPs matched to curated GWAS catalogue disease-associated SNPs (see https://link.springer.com/article/10.1007%2Fs00439-016-1734-y)

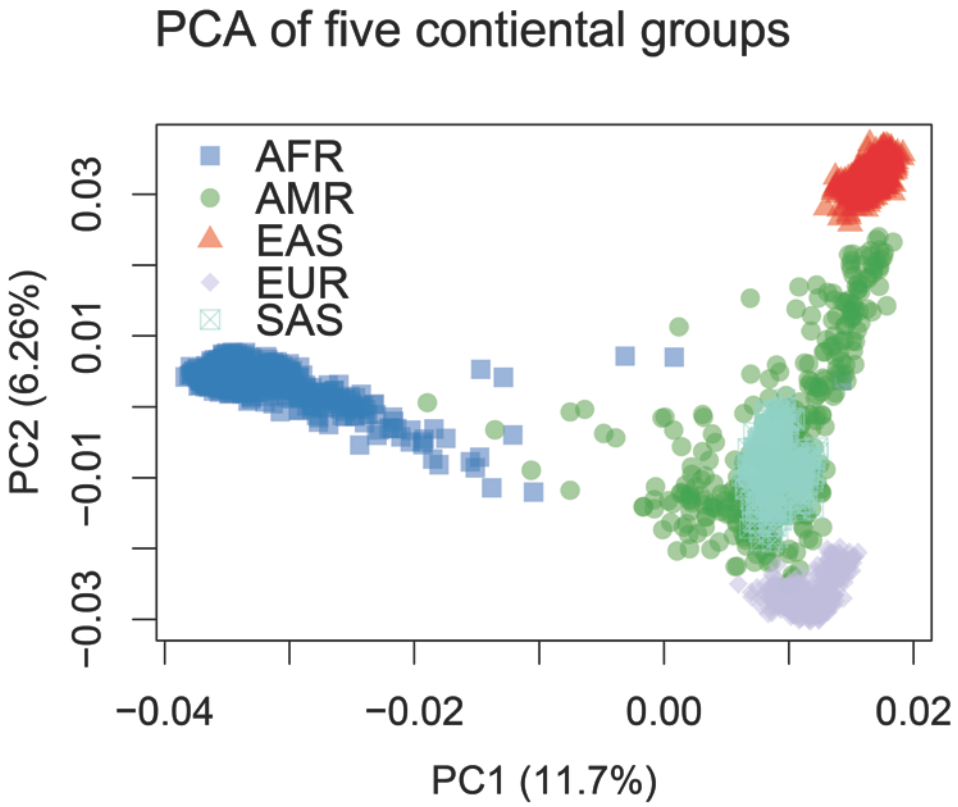
Supplementary Table 15 EAFD SNP with known associations with both a "beneficial" trait and a "harmful" disease

GWAS traits	GWAS diseases	EAFD SNPs	Population(s)
Eye color	Vitiligo, Melanoma	rs1393350[80,81]	SAS
Hair color	Non-melanoma skin cancer, Progressive supranuclear palsy	rs12203592[65,124,125]	EUR
Skin pigmentation	Lung cancer, Smoking behavior, Nicotine dependence	rs1042602[84-87]	EUR
Height	Psoriasis	rs2066808[83,126,127]	AFR
Height	Pulmonary function	rs2284746[83,128,129]	AFR
Height	Trans fatty acid levels	rs174547[82,83]	EAS, AMR

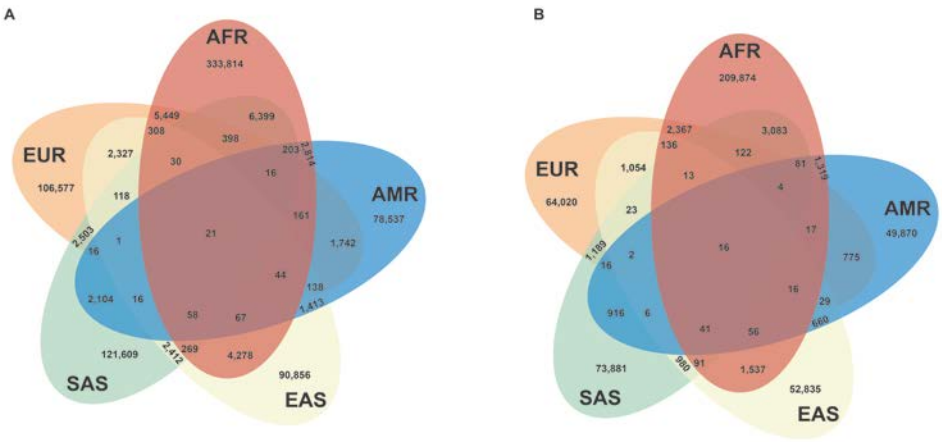
Supplementary Figures



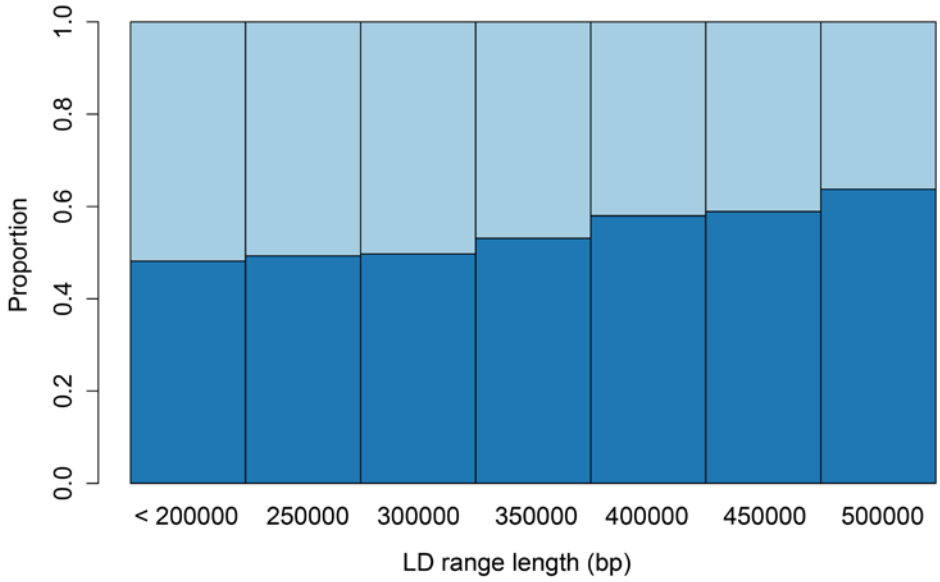
Supplementary Figure 1 Geography meets genetics: geographic distance versus θ_H. In each continental group the geographic distance between each population pair was estimated using the geosphere package in R. The correlation coefficients (ρ) and R^2 values are shown for each population pair. The African continental group underwent the same analysis twice, i.e., with and without the admixed sample of ACB and ASW. In the latter case, the correlation was stronger and more significant. For the European continental group, we excluded CEU, since these Europeans migrated to the USA, and their distance to continental European countries does not reflect their genetic distance.



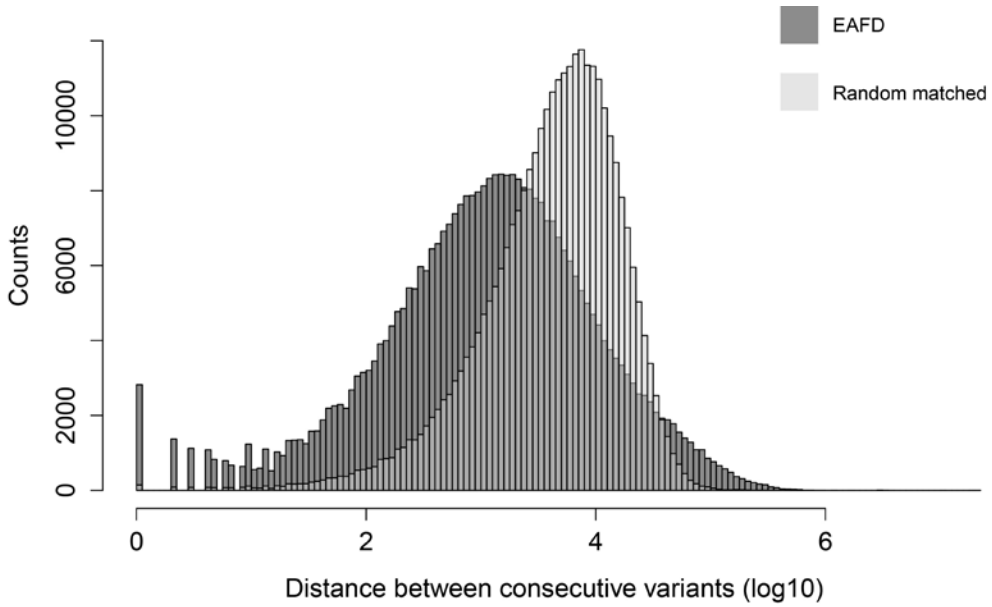
Supplementary Figure 2 PCA of the five continental groups. All 2,504 samples were included in the PCA. The most differentiated variants, those with extreme F_{ST} within and between continental groups, were used in the PCA. Four populations (AFR, EUR, EAS and SAS) separated very well from each-other, while American samples clustered closest to SAS, followed by EUR, AFR, and EAS.



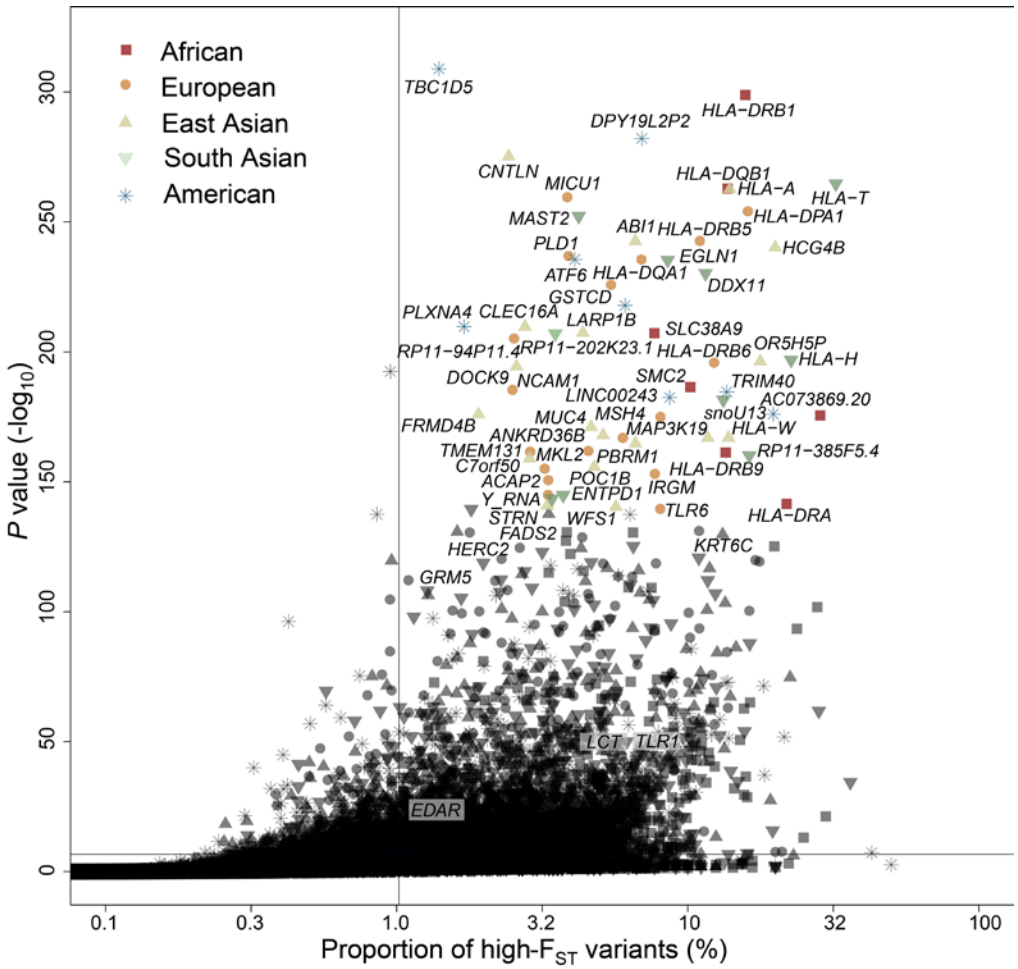
Supplementary Figure 3 Venn diagrams of variants shared among the five continental groups. Variant sharing patterns are shown before (A) and after LD correction (B), which was done by keeping one variant for any given window of 1,000bp.



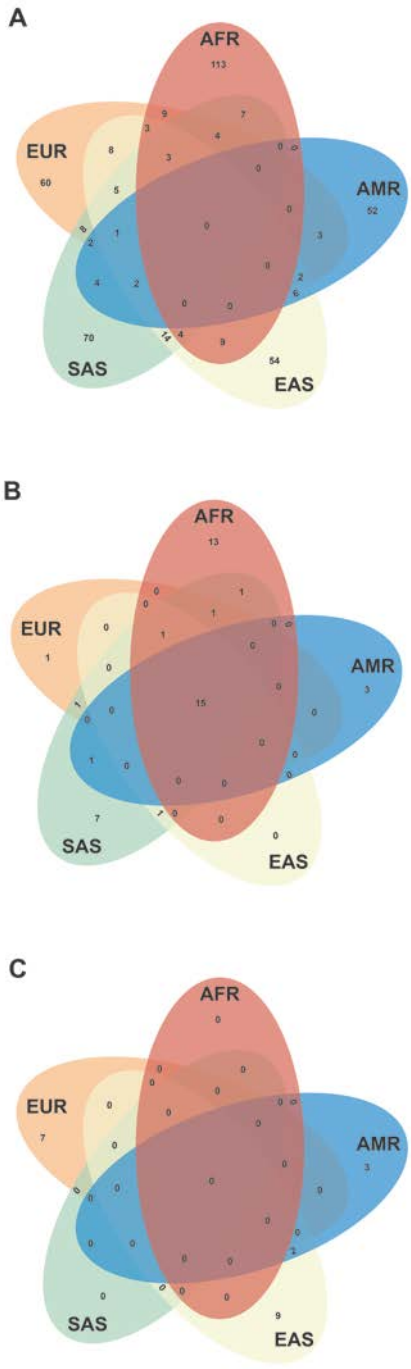
Supplementary Figure 4 Stacked bars of linkage disequilibrium (LD) range length distributions. Light and dark blue bars correspond to random matched and EAFD SNPs, respectively. The x-axis represents the lengths (up to 500,000bp) of the longest LD range for a SNP, while the y-axis represents proportion of total SNPs from each group. High LD was defined by $r^2 > 0.8$. The EAFD SNPs from the YRI population of 1000 Genomes Project (phase I) were used to measure the LD scores. The numbers of total SNPs under each bar, from left to right, are 3,766, 2,484, 1,585, 1,043, 740, 511, and 411, respectively.



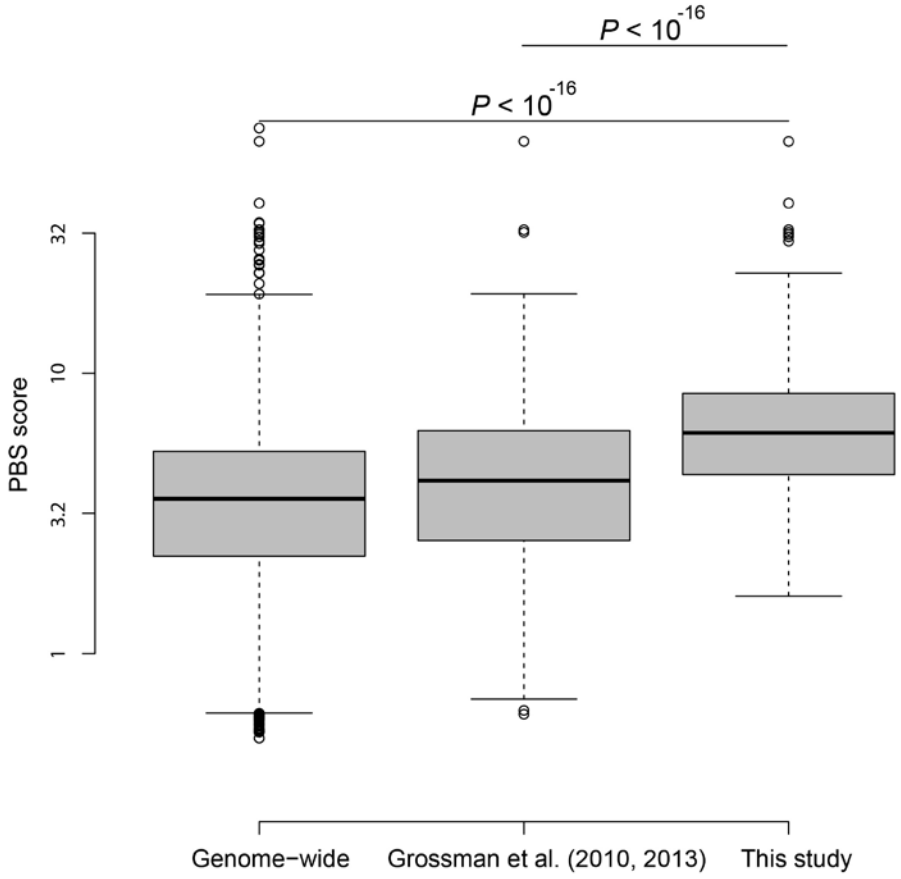
Supplementary Figure 5 Clustering of EAFD variants. Random matched variants (mean = 7,854 and SD = 57,318) were those randomly sampled from the whole-genome variants, and matched to the EAFD variants by 1) total variants (10,000) and 2) derived allele frequency distribution (the distribution of Africans in **Figure 3A** was used as reference). More than 12% of the EAFD variants, compared to only 7% of the matched random variants, were located within 100bp regions.



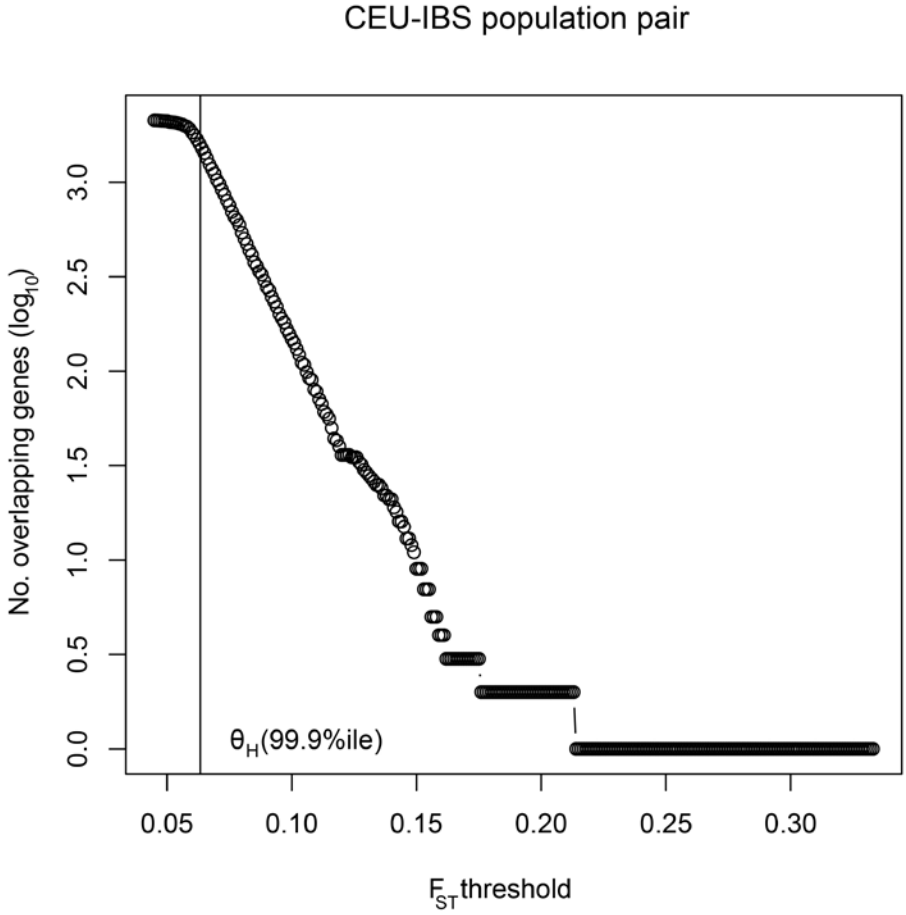
Supplementary Figure 6 Genes enriched with EAFD variants. The x-axis represents the proportion (logarithmic scale) of EAFD (or high-F_{ST}) variants among the total variants for each gene, and y-axis is the minus logarithmic (base 10) value of the raw enrichment *P* value. The hypergeometric distribution-based enrichment analysis was used (see **Methods**) and the significance threshold was 1.7×10^{-6} (horizontal line). The vertical line corresponds to the second threshold (i.e., proportion > 1%). Genes on the upper right quadrant underwent further analyses, and if a nonsynonymous EAFD variant was identified in them, it was selected as EAFD gene and retained for the further analyses. Selected genes were labelled and colored to indicate continental groups where they were identified.



Supplementary Figure 7 Venn diagrams of genes (A), pathways (B) and diseases and traits (C) shared among the five continental groups.



Supplementary Figure 8 EAFD genes are enriched with positive selection targets. Large PBS scores are indicative of positive selection or local adaptation. The mean PBS score for our EAFD genes was significantly higher (7.2) than the whole-genome scores (mean = 4, Wilcoxon rank-sum $P < 2 \times 10^{-16}$) and those from 373 adaptation genes from a combined list of two well-known studies on positive selection[78,130] (mean = 5.1 and $P < 2 \times 10^{-16}$).



Supplementary Figure 9 Robustness of F_{ST} threshold. The F_{ST} threshold was varied 289 times, from 0.045 to 0.333, by increments of 0.001. Each time, the overlap with genes from the lowest overlap, was measured (e.g., at threshold 0.045, the gene overlap is 100%). We carried out these measurements for a representative population pair, CEU – IBS. The 99.9th F_{ST} percentile (i.e., θ_H) for this population pair was 0.0634.

Chapter 2.2: Eye Color: A Potential Indicator of Alcohol Dependence Risk in European Americans

Arvis Sulovari[1,2], Henry R. Kranzler[3], Lindsay A. Farrer[4], Joel Gelernter[5,6,7] and Dawei Li[1,8,9*]

[1]*Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont*

[2]*Cell, Molecular and Biomedical Sciences Graduate Program, University of Vermont, Burlington, Vermont*

[3]*Department of Psychiatry, University of Pennsylvania School of Medicine and VISN 4 MIRECC, Philadelphia VAMC, Philadelphia, Pennsylvania*

[4]*Departments of Medicine (Biomedical Genetics), Neurology, Ophthalmology, Genetics & Genomics, Biostatistics, and Epidemiology, Boston University Schools of Medicine and Public Health, Boston, Massachusetts*

[5]*Department of Psychiatry, School of Medicine, Yale University, New Haven, Connecticut*

[6]*Department of Genetics, School of Medicine, Yale University, New Haven, Connecticut*

[7]*VA Connecticut Healthcare Center, West Haven, Connecticut and Department of Neurobiology, Yale University School of Medicine, New Haven, Connecticut*

[8]*Department of Computer Science, University of Vermont, Burlington, Vermont*

[9]*Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, Vermont*

*To whom correspondence should be addressed:

Dawei Li, Ph.D., Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, US. E-mail: dawei.li@uvm.edu

Number of words in the abstract: 240

Number of words in the text (excluding acknowledgments and financial disclosures sections, legends, and references): 2,996

Number of tables: 1

Number of figures: 3

Number of supplementary material: 3 supplementary Tables and 6 supplementary Figures and Legends.

Abstract

Light-eyed individuals have been found to consume more alcohol than dark-eyed individuals in archival samples of European-ancestry males and females. No published population-based studies have directly tested for an association between alcohol dependence (AD) and eye color. We hypothesize that light-eyed individuals have a higher prevalence of AD than dark-eyed individuals. A mixture model was used for selection of homogeneous sample and control for population stratification. After quality control, we conducted an association study using logistic regression, adjusting for confounders (age, sex, and genetic ancestry) in a sample of 1,263 European-Americans. We found evidence of association between AD and blue eye color (P = 0.0005 and odds ratio = 1.83 (1.31 - 2.57)), supporting light eye color as a risk factor relative to brown eye color. Network-based analyses revealed a statistically significant (P = 0.02) number of genetic interactions between eye color genes and AD-associated genes. We found evidence of linkage disequilibrium between AD-associated GABA receptor gene cluster, *GABRB3*/*GABRG3*, and eye color genes, *OCA2*/*HERC2* as well as between AD-associated *GRM5* and pigmentation- associated *TYR*. Our population-phenotype, network, and linkage disequilibrium analyses support a possible association between blue eye color and AD. Although we controlled for stratification we cannot exclude underlying occult stratification as a contributor to this observation. While replication is needed, our findings suggest that eye pigmentation information may be useful in future research of alcohol addiction. Further characterization of this association may unravel new AD etiological factors.

Key words: Alcohol Dependence; Drinking; Eye Pigmentation; Association; Ethanol; Melanogenesis

Introduction

Eye and hair color diversity is higher among Europeans than among any other populations, and these traits follow distinct geographic distributions. The blue eye color phenotype is more common in northern Europe than in the rest of Europe or, indeed, the rest of the world. A clear gradient of eye colors subsists across Europe, from dark-eyed populations in the south to light-eyed populations in the north. This gradient may be indicative of strong selection pressures that have acted on multiple genetic loci over a short evolutionary period[131]. Sexual selection, and adaptation to diet or climate partially explain the pigmentation diversity in Europe; e.g., the latter may have led to the observed correlation between ultraviolet radiation and skin pigmentation[132]. Recent research has indicated that positive selection on pigmentation variants in humans vary from 2% to 10% per generation, representing the strongest selection signals in humans[132]. A positive selection of this magnitude implicates multiple selection forces acting on pigmentation-related traits, such as eye color. Some selection pressures that affect eye color may be personality related. For instance, blue-eyed European individuals have been shown to be less agreeable than brown-eyed[133].

The main physiological determinant of eye color is the presence and distribution of melanin pigments within melanocytes of the uveal tract[134]. A molecular driver of melanin biogenesis pathway is the G-protein coupled receptor melanocortin 1 receptor (MC1R), which was found on the surface of melanocytes[134]. The *MC1R* gene is a key determinant of photosensitivity and harbors many variant alleles in European populations[135,136]. Penetrance of *MC1R* is mediated by oculocutaneous albinism type II (*OCA2*) [134]. Around 74% of the eye color variation is explained by a quantitative trait locus on intron 1 of *OCA2*[137]. Moreover, epistatic interactions between *OCA2* and *MC1R* have been reported to influence within-population skin pigmentation differences [138]. The melanogenesis cascade involves adenylyl cyclase 8 (encoded by *ADCY8*), which is to respond to MC1R and other factors in the cytosol of the melanocyte and convert ATP (adenosine triphosphate) to cAMP (cyclic adenosine monophosphate)[139]. Adenylyl cyclase 8 belongs to the family of adenylyl cyclase enzymes, which have been shown to play a role in substance addiction [140,141]. Interestingly, *ADCY8* has been reported to be associated with major depressive disorder and alcohol dependence (AD)[139], implying a possible connection between melanogenesis and etiological mechanisms of AD.

Northern Europeans may have evolved the blue eye trait as an adaptation to their darker environment (compared to southern Europeans) because blue eyes confer greater sensitivity to natural light[142]. However, heightened sensitivity to light might also confer a higher propensity for seasonal affective disorder (SAD), which is often comorbid with AD[143], via abnormal melatonin changes in response to varying light intensities[144]. Light-

eyed individuals have been found to consume more alcohol than dark-eyed individuals in archival samples of European-ancestry males and females[145]. Despite the indirect or sporadic evidence supporting the connection between eye color and alcohol drinking, no published population-based studies have directly tested for biological interactions, appropriately correcting for population stratification. In this study, we tested the hypothesis that light-eyed individuals have a higher prevalence of AD than dark-eyed individuals in European Americans (EAs).

Methods

Subjects

The samples analyzed in this study were recruited in multiple centers for alcohol and drug dependence studies, as described recently[146]. Subjects were ascertained using Diagnostic and Statistical Manual of Mental Disorders-fourth edition (DSM-IV) criteria[147] for substance use (e.g., alcohol, opioid, and cocaine dependence) or major psychiatric disorders. After a complete description of the study, written informed consent was obtained from each subject, as approved by the institutional review board at each site. All participants were interviewed using the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA). Control subjects had no diagnosed substance use or major psychotic disorders. Eye color was determined at interview, and by self-report. A total of 5,222 samples of European ethnicity underwent multiple quality control or filtering procedures to obtain homogenous groups with respect to population group,

exposure to alcohol (for controls), and the availability of phenotype and genotype information. The samples were filtered based on the exclusion criteria listed in Supplementary Table 1. Control subjects who had never been exposed to alcohol were excluded from the analysis.

Population Stratification

To explicitly model sample ancestry differences, we carried out principal component analysis (PCA) using the genotype data[28,146] from Illumina HumanOmni1 single nucleotide polymorphism (SNP) genotyping arrays. We adopted a mixture model approach to correct for structure and maximize genetic homogeneity. First, the noise was initialized by a Poisson method, which determined whether data points were noise or part of a cluster based on a Poisson-based process[148]. Second, the expectation-maximization-fitted Gaussian mixture model clustering method[149] was used to determine the boundary between the cluster and the noise. The first three components from PCA were used to evaluate the number of samples categorized as noise. The number of PCA dimensions that were selected as covariates in logistic regression analyses was determined on the basis of their contribution to genetic variation across samples. PCA and regression tests were applied independently to each population.

Association Analyses

Logistic regression was employed adjusting for confounding factors (age, sex, and genetic ancestry). The independent variable, eye color, was treated as categorical measure under three models. Model 1: each of the five categories (brown, blue, green, grey, and brown in the center) was analyzed separately; model 2: the three less-frequent light color categories (green, grey, and brown in the center) were combined as one group; and model 3: all of the four light color categories (blue, green, grey, and brown in the center) were combined as one group. In each of the three models, the eye color categories were regressed simultaneously. The first three principal components, which explained the vast majority of genetic ancestry variation, were used to correct for potential ancestry-based population stratification in the EA samples. The logistic regression model associates odds of AD and eye color, correcting for all of the aforementioned covariates:

$$\ln\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{m}\beta_i x_i + \beta_{m+1}age + \beta_{m+2}sex_j + \beta_{m+3}pc_1 + \beta_{m+4}pc_2 + \beta_{m+5}pc_3 \qquad (1)$$

, where p is probability of AD, $x = \{brown, blue, green, grey, brown - center\}$, $m = \{2,3,5\}$ corresponds to the three eye-color models (described above) and age, sex_j, pc_1, pc_2 and pc_3 are the covariates with pc representing principal components and $j = \{1,2\}$ denoting that sex is a categorical variable. The predictive capacity of eye color (i.e., x) towards odds of disease $\frac{p}{1-p}$ can be measured by: $odds_x = e^{\beta_0 + R}$, where R $= \sum_{i=1}^{m}\beta_i x_i + \beta_{m+1}age + \beta_{m+2}sex_j + \beta_{m+3}pc_1 + \beta_{m+4}pc_2 + \beta_{m+5}pc_3$. The brown

color is considered as the reference color when calculating odds ratio (OR): $OR_{x=brown} = e^{\beta_0} = 1$, for brown eye color and: $OR_{x \neq brown} = e^{R}$ for non-brown eye color. The glm package in R (v. 3.1.1) was used.

Network Analyses

A total of 26 AD-associated genes and 21 pigmentation genes were selected as the AD and eye color candidate genes, respectively, based on our previous meta-analyses[19,150-155] of genetic association studies and the literature[156] (Supplementary Table 2). The GeneMania database[157] was used to evaluate the number of genetic interactions between the 26 AD genes and 21 pigmentation genes (Supplementary Table 3 and Supplementary Figure 5). Here, two genes are considered to interact under "genetic interactions" if the effects of perturbing one gene are modified by perturbations to a second gene. To assess statistical significance, we randomly sampled 21 genes across the whole genome to replace the actual 21 pigmentation-related genes, and then measured their connectivity to AD genes. This procedure was repeated 1,000 times to generate a random distribution of genetic interaction connections. The significance levels were measured using Z scores. The statistical analysis was carried out and the histograms generated using R (version 3.1.1). The networks were simulated using Cytoscape[158].

Linkage Disequilibrium and Haplotype Analyses

HaploView[159] was used to calculate and visualize the linkage disequilibrium (LD) blocks in the selected chromosomal regions using genotype data from the HapMap samples of Utah residents of western and northern European ancestry (CEU) and Toscans in Italy (TSI). The D', r^2, and LOD metrics were used to calculate LD blocks. Besides these parameters, the method described by Gabriel et al.[160] was also applied for LD-block identification when intergenic distance was short (i.e., around 100 kilo base-pairs (bps)). Supplementary Figure 3 outlines the three different approaches used to test our hypothesis at the population, network, and genetic levels.

Results

A total of 1,263 unrelated AD cases and controls of EAs were analyzed in this study after quality control. The filtering procedure is shown in details in Supplementary Table 1. Supplementary Figure 1 shows a scatter plot of the first three principal components of the EA samples, indicating that our samples are moderately homogenous. This implied that there was a modest risk of observing false positive findings due to population stratification. Figure 1 shows the results of the model-based clustering method in combination with a Poisson-based process (see Methods). In our samples, the number of outliers was within < 5% of the total samples size (i.e., 4.2%), further indicating that our samples are relatively homogenous. The first three principal components were used to correct for potential population stratification in all of our association tests between eye color and AD.

We found evidence of significant phenotypic association between eye color and AD (P = 0.003; OR = 1.54 (1.15-2.04)) when compared the combined light eye colors (blue, green, grey, and brown-center) to brown eye color. Evidence of stronger association was observed between blue eye and AD when the blue eye color was analyzed separately ($P = 4.7 \times 10^{-4}$; OR = 1.83 (1.31-2.57) under model 1; and $P = 4.9 \times 10^{-4}$; OR = 1.82 (1.30-2.56) under model 2; Table 1). This result indicates that blue eye color is the most likely risk factor for AD among various light eye colors in EAs. Additionally, for the African Americans (AAs) included in our cohort, only 0.18% individuals (2,279) had blue eyes, indicating insufficient statistical power for the association tests (data available upon request).

To examine biological relevance, we carried out gene-gene interaction network analyses and LD measurements between known eye color genes and AD genes. We found evidence of a significant enrichment of genetic interactions between eye color genes and AD-associated genes (P = 0.02; Figure 2). Among these genes, the *MC1R* and gamma-aminobutyric acid A receptor α1 (*GABRA1*) genes showed the strongest genetic interaction (Supplementary Table 2). Genetic interactions may reflect complex biological interactions that include, but are not limited to, protein-protein interactions and possibly complex epistatic interactions[161].

Furthermore, we measured the LD between the chromosome 15q12 GABA receptor gene cluster, which has previously been reported to be involved in AD etiology[19] (Supplementary Table 2), and two eye color genes, *OCA2* and the ECT and RLD domain containing E3 ubiquitin protein ligase 2 gene (*HERC2*), which are also located on chromosome 15q12 at a distance of 221,887 bps (Supplementary Figure 2). We identified five strong LD blocks ($r^2 > 0.8$ and $D' > 0.8$), spanning a distance of around 200 kilo bps within the intergenic region between the GABA gene cluster (i.e., gamma-aminobutyric acid A receptor γ3, *GABRG3*) and eye color genes (i.e., *OCA2*). We used the similar approach to analyze all pairs of the AD and eye color genes residing on same chromosome (Supplementary Table 2) and found that the glutamate receptor, metabotropic 5 gene (*GRM5*; associated with AD) and tyrosinase gene (*TYR*; associated with pigmentation color of skin, hair, and eyes) were 111,507 bps apart on 11q14.3. This intergenic distance is spanned by five strong LD blocks, two of which overlap with the 5' UTR regions of *GRM5* and *TYR* (Supplementary Figure 6).

Additionally, SNPs from known AD-associated genes, including *ADCY8*, were tested for association with eye color, and vice versa, eye color genes were tested for association with AD. These tests revealed no evidence of statistically significant associations after correcting for multiple testing ($P > 0.05$, data not shown), suggesting that more investigation is needed regarding the underlying genes responsible for the potential AD-eye color association.

In all, the results from the three different types of analyses, i.e., population-phenotype, network, and LD, support that blue eye color may be associated with AD. The presence of genetic interactions between eye color genes and AD genes (Figure 3) implied a complex, potentially epistatic, genetic model. Figure 3 summarizes the results from these three approaches.

Discussion

In this study, we found a significant phenotypic association of AD with light eye colors, particularly blue eye color (Table 1), significant enrichment of genetic interactions between selected eye color genes and AD genes (Figure 2), and strong LD between pigmentation genes and AD-associated genes on chromosomes 15q12 and 11q14 (Supplementary Figures 2 and 6). The strengths of this study include 1) extensive control for potential population stratification of all samples using genome-wide SNP information, 2) leverage of genomic data to assess the extent of biologically relevant interactions between eye color genes and AD candidate genes, and 3) multilevel (i.e. population-phenotype, network and genetics) approaches to test our hypothesis.

Population stratification is a well-established source of false positive findings in association studies. To address this issue and assess the genetic homogeneity of our samples, we carefully selected only individuals who self-identified as EA and excluded admixed outliers such as individuals who were Hispanic based on self-report and our

principal component analysis. These quality control procedures are likely to lead to moderately homogenous samples (Figure 1 and Supplementary Figure 1). It should be noted that the PCA-based correction may not adequately correct for the south-north eye color cline in Europe or for the potential variation of this trait within countries of origin.

A few other lines of research support the observed AD-eye color association. Firstly, there is evidence of association between light eye color and SAD[162] (Supplementary Figure 4). SAD is often comorbid with AD[143]. While the relationship between eye color and SAD could plausibly be explained by varying light sensitivity, there is no readily available explanation for the association between eye color and AD. One possible physiological mechanism connecting eye color and AD is as follows: blue-eyed individuals have greater light sensitivity than brown-eyed individuals; and heightened sensitivity to varying light intensities has been associated with abnormal changes in endogenous melatonin production[162]. The latter has also been associated with SAD, which is often comorbid with AD (Supplementary Figure 4). Thus, we hypothesize that AD and eye color may have partially shared etiological factors. Terman et al. showed that light-eyed individuals were less likely to develop SAD than brown-eyed individuals during the winter[163]. However, this conclusion did not exclude the possibility that light-eyed individuals are at a higher risk for SAD than their dark-eyed counterparts when exposed to varying light intensities, which is known to alter endogenous levels of serotonin and melatonin in light-supersensitive individuals[162]. Furthermore, our results complement a recent paper where sunshine was shown to influence behavior [164]. This

study suggested that sunshine might facilitate suicidal behavior during the ten day period prior to suicide. Since AD is a known risk factor for suicidal behavior[165-167], our results imply that individuals with light eye color might be at higher susceptibility of sunshine-triggered behavior alteration (e.g., mood, aggression and impulsiveness) than dark-eye individuals. In sum, the inconsistent findings[144,162] in the literature reflect an incomplete understanding of the connection between eye color and psychiatric disorders.

Secondly, we observed strong LD blocks between eye color genes and GABA genes on chromosome 15q12. Interestingly, the 15q12 cytoband lies within the Prader-Willi syndrome (PWS) region. PWS presents with two relevant clinical features: hypopigmentation of the eyes and behavioral and psychiatric disturbances[168], which demonstrates that mutations in the 15q12 region can lead to both phenotypes. Similarly, we also observed strong LD between the *GRM5* (AD-associated) and *TYR* (pigmentation-associated) genes in cytoband 11q14.3. Interestingly, microdeletions in this region have been associated with leukodystrophy, a group of central nervous system disorders affecting the brain's white matter[169]. Additionally, variation in this region, specifically in *TYR*, has been associated with melanin production[170]. Overall, these observations support that two independent gene regions in the human genome may be concurrently associated with pigmentation variation and brain function.

Thirdly, animal experiments have also shown that hypopigmentation may correlate with behavioral changes (e.g., in the *Astyanax* cavefish model[171]). Despite lack of direct

evidence, these reports support the association between blue eye color and AD in EAs (Figure 3).

To conclude, our findings complement the existing research on the connection between eye color and mental illnesses and behavioral problems. Our study is the first to report an association between blue eye color and AD in EAs using well-diagnosed subjects and a moderate sample size. Our findings indicate that the selection pressures acting on the genetics of pigmentation might not only have implications for personality features, as previously reported[133], but also for AD susceptibility. Thus, integration of population-phenotype and gene and network analyses is helpful for the identification of risk factors in AD, and a broad range of mental illnesses, in general. Although we carefully controlled for stratification, we cannot exclude underlying occult stratification as a contributor to this observation. While replication is needed, our findings suggest that eye pigmentation information may be useful in the future research of AD and related alcohol consumption behaviors. Further characterization of this association may unravel novel etiological factors in alcohol addiction.

Acknowledgement

This work was supported by the Start-up Fund of the University of Vermont. We would like thank Dr. Richard M Sherva for help in genetic data preparation. We also thank the anonymous reviewers for their helpful suggestions and comments.

Conflict of Interest

Henry Kranzler has been a consultant or advisory board member for the following pharmaceutical companies: Alkermes, Lilly, Lundbeck, Otsuka, Pfizer, and Roche. He is also a member of the American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative, which is supported by Alkermes, Ethypharm, Lilly, Lundbeck, AbbVie, and Pfizer. All the other authors declare no potential conflict of interest.

References

Akey JM, Wang H, Xiong M, Wu H, Liu W, Shriver MD, Jin L. 2001. Interaction between the melanocortin-1 receptor and P genes contributes to inter-individual variation in skin pigmentation phenotypes in a Tibetan population. Hum Genet 108(6):516-20.

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

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263-5.

Bassett JF, Dabbs JM. 2001. Eye color predicts alcohol use in two archival samples. Personality and Individual Differences 31(4):535-539.

Beleza S, Johnson NA, Candille SI, Absher DM, Coram MA, Lopes J, Campos J, Araujo, II, Anderson TM, Vilhjalmsson BJ and others. 2013. Genetic architecture of skin and eye color in an African-European admixed population. PLoS Genet 9(3):e1003372.

Cao J, Hudziak JJ, Li D. 2013a. Multi-cultural association of the serotonin transporter gene (SLC6A4) with substance use disorder. Neuropsychopharmacology 38(9):1737-47.

Cao J, LaRocque E, Li D. 2013b. Associations of the 5-hydroxytryptamine (serotonin) receptor 1B gene (HTR1B) with alcohol, cocaine, and heroin abuse. Am J Med Genet B Neuropsychiatr Genet 162B(2):169-76.

Cao J, Liu X, Han S, Zhang CK, Liu Z, Li D. 2014. Association of the HTR2A gene with alcohol and heroin abuse. Hum Genet 133(3):357-65.

Cassidy SB, Schwartz S, Miller JL, Driscoll DJ. 2011. Prader-Willi syndrome. Genet Med.

DiRocco DP, Scheiner ZS, Sindreu CB, Chan GC, Storm DR. 2009. A role for calmodulin-stimulated adenylyl cyclases in cocaine sensitization. J Neurosci 29(8):2393-403.

Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Barta C, Lu RB, Zhukova OV, Kim JJ, Siniscalco M, New M and others. 2012. A global view of the OCA2-HERC2 region and pigmentation. Hum Genet 131(5):683-96.

Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA. 2007. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. Am J Hum Genet 80(2):241-52.

Elipot Y, Hinaux H, Callebert J, Launay JM, Blin M, Retaux S. 2014. A mutation in the enzyme monoamine oxidase explains part of the Astyanax cavefish behavioural syndrome. Nat Commun 5:3647.

Fraley C, Raftery AE. 2003. Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. Journal of Classification 20(2):263-286.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M and others. 2002. The structure of haplotype blocks in the human genome. Science 296(5576):2225-9.

Gardiner E, Jackson CJ. 2010. Eye color Predicts Disagreeableness in North Europeans: Support in Favor of Frost (2006). Current Psychology 29(1):1-9.

Gelernter J, Kranzler HR, Sherva R, Almasy L, Koesterer R, Smith AH, Anton R, Preuss UW, Ridinger M, Rujescu D and others. 2014. Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. Mol Psychiatry 19(1):41-9.

Goizet C, Coupry I, Rooryck C, Taine L, Dormoy V, Lacombe D, Arveiler B. 2004. Molecular characterization of an 11q14.3 microdeletion associated with leukodystrophy. Eur J Hum Genet 12(3):245-50.

Hennig C, Hausdorf B. 2010. prabclus: Functions for clustering of presence–absence, abundance and multilocus genetic data. R package version 2:2-2.

Higuchi S, Motohashi Y, Ishibashi K, Maeda T. 2007. Influence of eye colors of Caucasians and Asians on suppression of melatonin secretion by light. Am J Physiol Regul Integr Comp Physiol 292(6):R2352-6.

Inskip HM, Harris EC, Barraclough B. 1998. Lifetime risk of suicide for affective disorder, alcoholism and schizophrenia. Br J Psychiatry 172:35-7.

Kim KS, Lee KW, Lee KW, Im JY, Yoo JY, Kim SW, Lee JK, Nestler EJ, Han PL. 2006. Adenylyl cyclase type 5 (AC5) is an essential mediator of morphine action. Proc Natl Acad Sci U S A 103(10):3908-13.

Li B, Chen W, Zhan X, Busonero F, Sanna S, Sidore C, Cucca F, Kang HM, Abecasis GR. 2012a. A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet 8(10):e1002944.

Li D, Sulovari A, Cheng C, Zhao H, Kranzler HR, Gelernter J. 2014. Association of gamma-aminobutyric acid A receptor alpha2 gene (GABRA2) with alcohol use disorder. Neuropsychopharmacology 39(4):907-18.

Li D, Zhao H, Gelernter J. 2011. Strong Association of the Alcohol Dehydrogenase 1B Gene (ADH1B) with Alcohol Dependence and Alcohol-Induced Medical Diseases. Biol Psychiatry.

Li D, Zhao H, Gelernter J. 2012b. Strong protective effect of the aldehyde dehydrogenase gene (ALDH2) 504lys (*2) allele against alcoholism and alcohol-induced medical diseases in Asians. Hum Genet 131(5):725-37.

Li D, Zhao H, Kranzler HR, Li MD, Jensen KP, Zayats T, Farrer LA, Gelernter J. 2015. Genome-Wide Association Study of Copy Number Variations (CNVs) with Opioid Dependence. Neuropsychopharmacology 40(4):1016-26.

Lin A, Wang RT, Ahn S, Park CC, Smith DJ. 2010. A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. Genome Res 20(8):1122-32.

Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, Morris Q, Bader GD. 2010. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics 26(22):2927-8.

Olfson E, Bierut LJ. 2012. Convergence of genome-wide association and candidate gene studies for alcoholism. Alcohol Clin Exp Res 36(12):2086-94.

Pacchierotti C, Iapichino S, Bossini L, Pieraccini F, Castrogiovanni P. 2001. Melatonin in psychiatric disorders: a review on the melatonin involvement in psychiatry. Front Neuroendocrinol 22(1):18-32.

Procopio DO, Saba LM, Walter H, Lesch O, Skala K, Schlaff G, Vanderlinden L, Clapp P, Hoffman PL, Tabakoff B. 2013. Genetic markers of comorbid depression and alcoholism in women. Alcohol Clin Exp Res 37(6):896-904.

Rees JL. 2004. The genetics of sun sensitivity in humans. Am J Hum Genet 75(5):739-51.

Roecklein KA, Rohan KJ, Duncan WC, Rollag MD, Rosenthal NE, Lipsky RH, Provencio I. 2009. A missense variant (P10L) of the melanopsin (OPN4) gene in seasonal affective disorder. J Affect Disord 114(1-3):279-85.

Sher L. 2004. Alcoholism and seasonal affective disorder. Compr Psychiatry 45(1):51-6.

Sher L. 2006. Alcohol consumption and suicide. QJM 99(1):57-61.

Sturm RA. 2002. Skin colour and skin cancer - MC1R, the genetic link. Melanoma Res 12(5):405-16.

Sturm RA. 2009. Molecular genetics of human pigmentation diversity. Hum Mol Genet 18(R1):R9-17.

Sturm RA, Duffy DL. 2012. Human pigmentation genes under environmental selection. Genome Biol 13(9):248.

Sturm RA, Duffy DL, Zhao ZZ, Leite FP, Stark MS, Hayward NK, Martin NG, Montgomery GW. 2008. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. Am J Hum Genet 82(2):424-31.

Terman JS, Terman M. 1999. Photopic and scotopic light detection in patients with seasonal affective disorder and control subjects. Biol Psychiatry 46(12):1642-8.

Vyssoki B, Kapusta ND, Praschak-Rieder N, Dorffner G, Willeit M. 2014. Direct effect of sunshine on suicide. JAMA Psychiatry 71(11):1231-7.

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterlander M, Hollfelder N, Potekhina ID, Schier W, Thomas MG and others. 2014. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A 111(13):4832-7.

Wojnar M, Ilgen MA, Czyz E, Strobbe S, Klimkiewicz A, Jakubczyk A, Glass J, Brower KJ. 2009. Impulsive and non-impulsive suicide attempts in patients treated for alcohol dependence. J Affect Disord 115(1-2):131-9.

Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, Morris Q. 2013. GeneMANIA prediction server 2013 update. Nucleic Acids Res 41(Web Server issue):W115-22.

Tables

Table 1 Association results between eye colors and alcohol dependence in European-Americans

		Model 1		Model 2		Model 3	
Eye colors	**Cases/ Controls**	**OR**	***P***	**OR**	***P***	**OR**	***P***
Brown	368/130	-	-	-	-	-	-
Blue	377/70	1.83 (1.31-2.57)	**4.7×10^{-4}**	1.82 (1.30-2.56)	**4.9×10^{-4}**	1.54 (1.15-2.04)	**0.003**
Green	223/64	1.28 (0.90-1.83)	0.17	1.26 (0.90-1.78)	0.19		
Grey	5/5	0.34 (0.09-1.30)	0.11				
Brown-center	19/2	3.76 (1.04-24.14)	0.08				

Brown eye color is the reference color in the three models. Logistic regression analysis includes age, sex and the first three principal components as covariates. The OR column contains the OR values and 95% confidence intervals in brackets. The dotted vertical lines indicate the groupings of eye colors under Model 2 and 3. The *P* values in bold represents $P < 0.05$. In all three tests, blue eye color only (models 1 and 2) or all light eye colors together (model 3) were significantly associated with AD outcome. The three non-blue light eye colors represent a relatively small portion of the EA samples, which may explain their lack of statistical significance.

Figure Legends

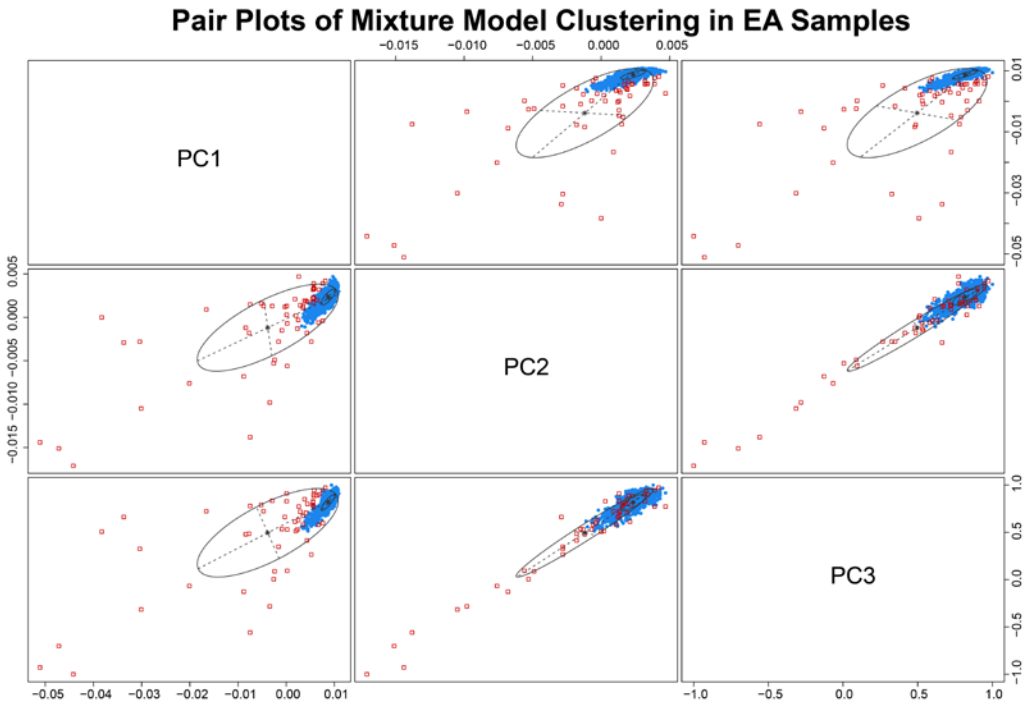


Figure 1 Pair plots of cluster analysis results in the EA population. The first three principal components from PCA were used as inputs for an EM-fitted Gaussian mixture model clustering method with Poisson-based noise initialization. Each point represents one of the 1,263 EA sample and the labels for all axes are either diagonally or on the sides of the plot. The blue points (1,211) represent the core cluster while the red points (53) represent potential outliers. The size of the two ellipses in each plot represents the covariance of the two mixture components (i.e., blue and red clusters).

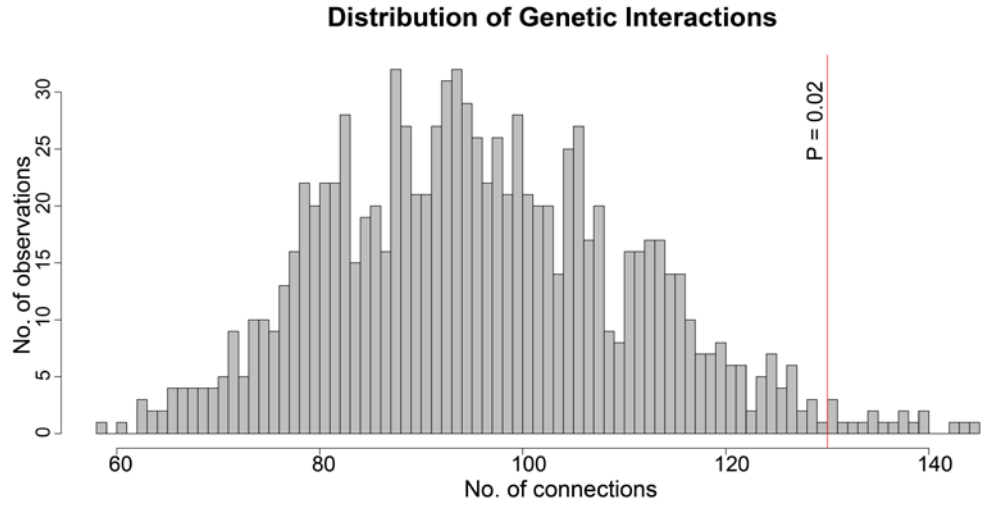


Figure 2 Distributions of genetic interactions between AD and eye color genes. Histogram represents distribution of connections (i.e., edges) between AD gene-set (vertical red line) and random gene-sets of 1,000 simulated networks (dark grey columns). The number of random genes was kept the same as the number of eye color genes in all simulations. There was evidence of enrichment of genetic interactions among the AD gene-set (P = 0.02).

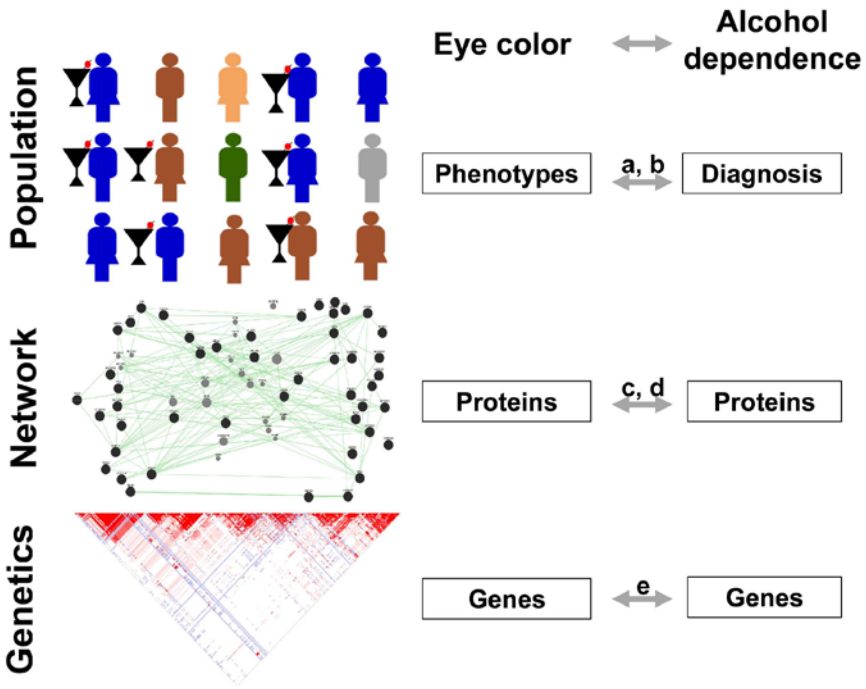


Figure 3 Summary of the association between eye color and alcohol dependence. Our study complements knowledge regarding associations between eye color and behavior problems. Our major observation is that blue eye color is a potential risk factor for alcohol dependence. Labels a-e correspond to the following evidence: a) association between blue eye phenotype and AD after adjustment for sex, age and ethnicity in our samples (Table 1); b) finding that light-eyed individuals consumed more alcohol than dark-eyed individuals in two archival samples from 1974 (10,860 Caucasian male prison inmates and 1,862 Caucasian females from a national survey)[145]; c) evidence of genetic interactions between addiction proteins and eye color proteins (Figure 2); d) literature evidence connects melanosome and dopamine synthesis using Astyanax cavefish model[171]; and e) evidence of LD between *GABRG3* and *OCA2* (Supplementary Figure 2) and LD between *GRM5* and *TYR* (Supplementary Figure 6).

Supplements

Supplementary Tables

Supplementary Table 1 The cumulative filtering procedure for the EA samples.

Steps	Samples sizes
All	5,222
Unrelated	4,726
Exposed to alcohol	4,643
Non-missing phenotype	3,862
Non-missing genotype	1,263

The first column displays the remaining samples after each step of the quality control process.

Supplementary Table 2: Summary of the AD and eye color genes paired from genetic interaction network analyses.

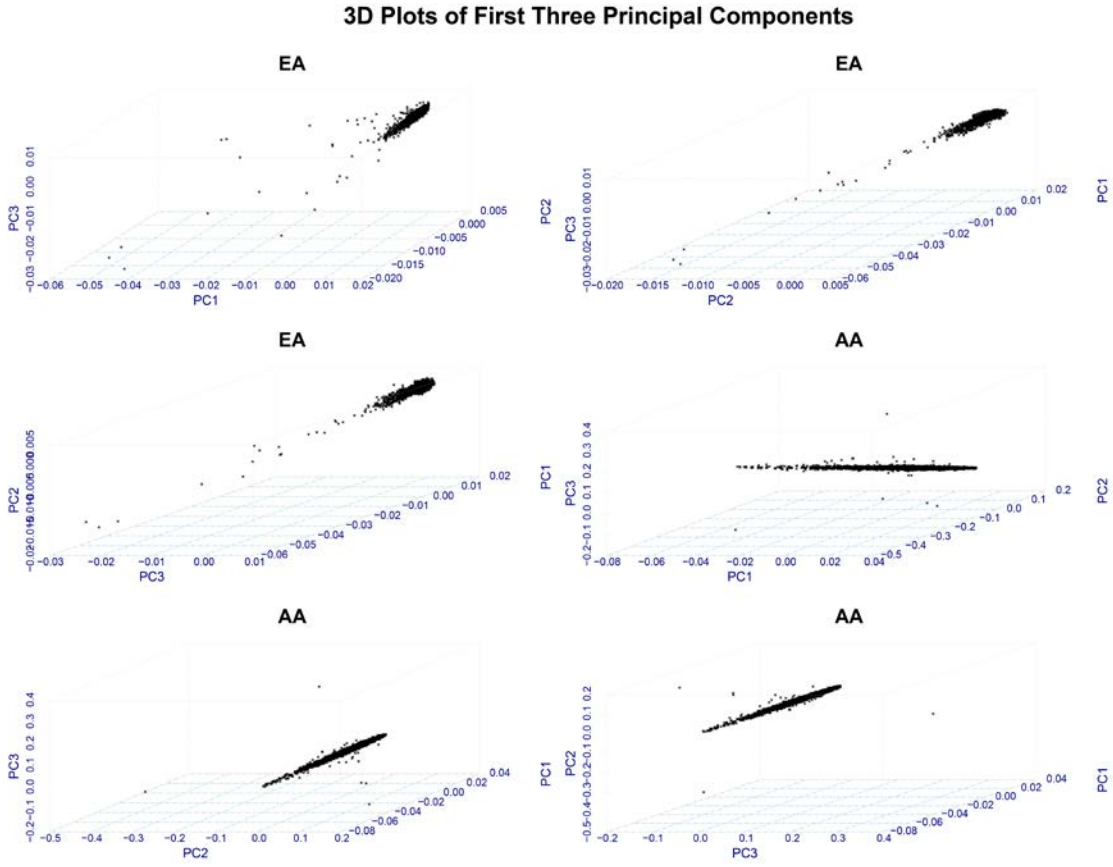
AD genes	Eye color genes	Numbers of Interactions
GABRA1	*MC1R*	1
ADH1B	*HERC2,ADCY8,TYR*	3
ALDH2	*SLC24A5,TTC3,FBXL17,TYRP1*	4
MREG	*OCA2,VASH2,FBXL17*	3
GABRG2	*EFR3A*	1
NXPH2	*TYR,SLC24A5*	2
METAP1	*TYRP1,ADCY8*	2
FAM44B	*NPLOC4*	1
TPK1	*OCA2,HERC2,TYRP1,EFR3A*	4
NXPH2	*TYR,SLA24A5*	2
PDLIM5	*KITLG,TTC3,FBXL17,SLC45A2,ADCY8,EFR3A*	6

GABRA6	*OCA2,SLC24A5*	2
GABRA3	*SLC24A4*	1
COL8A1	*TYRP1,EFR3A*	2
NOMO2	*EFR3A*	1
GRM5	*OCA2,FBXL17*	2
E2F8	*EFR3A,TYRP1,FBXL17*	3
PDLIM5	*EFR3A,ADCY8,SLC45A2,FBXL17,TTC3,KITLG*	5
GABRA2	*OCA2,SLC24A4,FBXL17,TTC3*	4
MREG	*OCA2,VASH2,FBXL17*	3

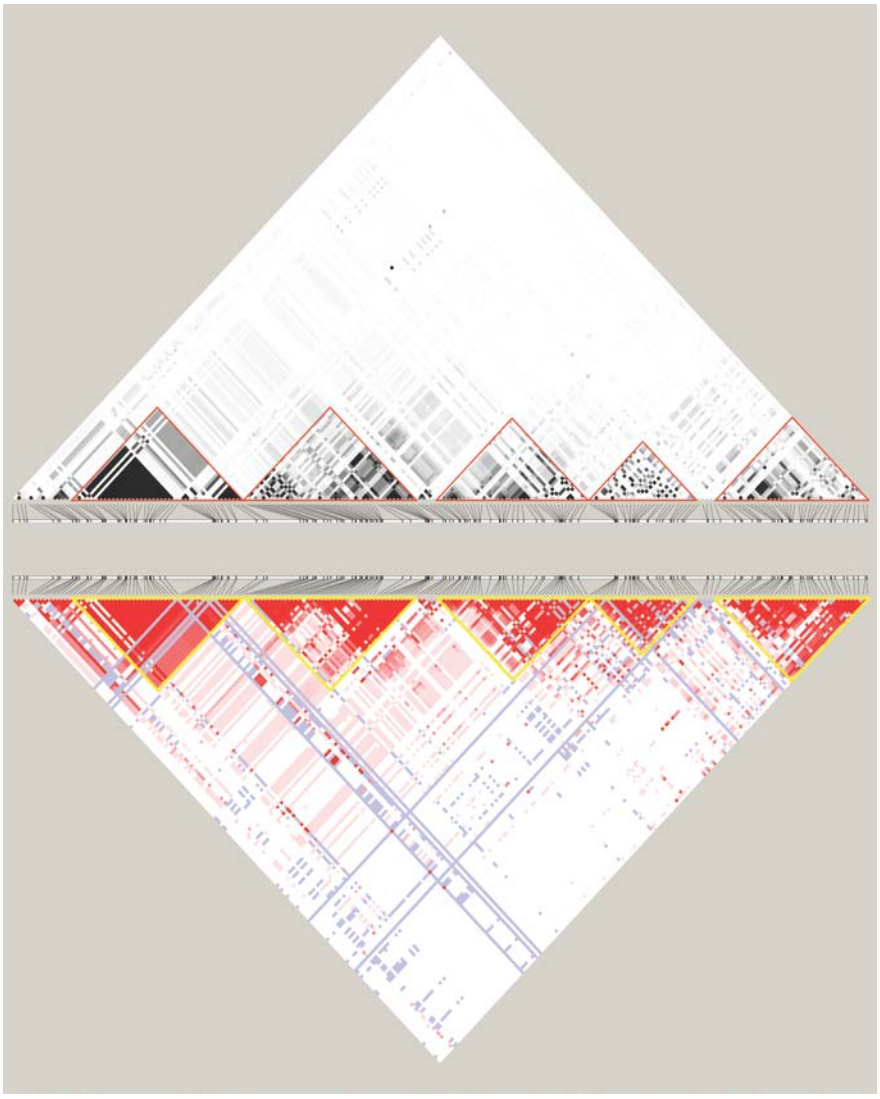
The set of 26 AD genes (*ADH1A,ADH1B,ALDH1A1,ALDH1B1,ALDH2,CC2D2B,COL8A1,E2F8, FAM44B,GABRA1,GABRA2,GABRA3,GABRA4,GABRA5,GABRA6,GABRG2,GRM5,METAP1,MREG, NOMO2,NXPH2, PDLIM5,PKNOX2,SH3BP5,TPK1,ZNF285A*) and that of 21 eye color genes (*ADCY8,ASIP,EFR3A,FBXL17,HERC2,HGS,IRF4,KITLG,MC1R,NPLOC4,OCA2,POLS, SLC24A4,SLC24A5,SLC45A2,TPCN2,TTC3,TYR,TYRP1,VASH2,PAX6*) were found to have a significant number of genetic interactions using GeneMANIA (P = 0.02, Figure 2; Supplementary Table 3). Genetic interactions are inferred from a database of radiation hybrid networks[161]. For each AD gene in the first column, the interacting eye color genes are shown in the second column with the total number of their connections in the third column. The table contains only gene pairs with non-zero interactions between the two gene sets. Interacting gene pairs are ordered by strength of genetic interaction, such as *GABRA1-MCR1* holds the highest weight.

Supplementary Table 3: The results of genetic interaction network analyses. (see http://onlinelibrary.wiley.com/doi/10.1002/ajmg.b.32316/abstract)

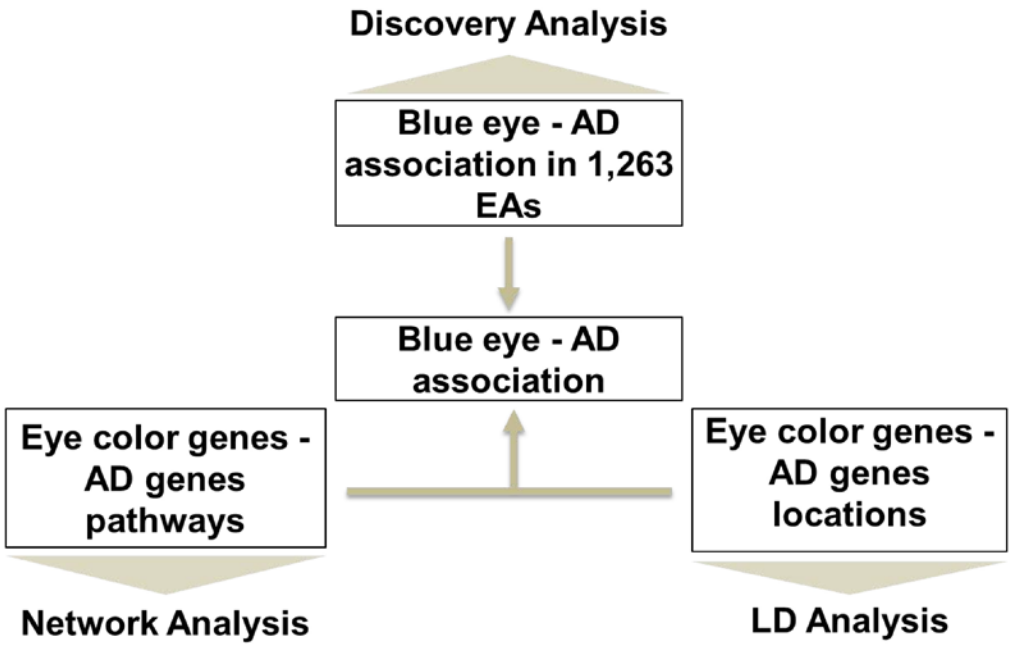
Supplementary Figures and Legends



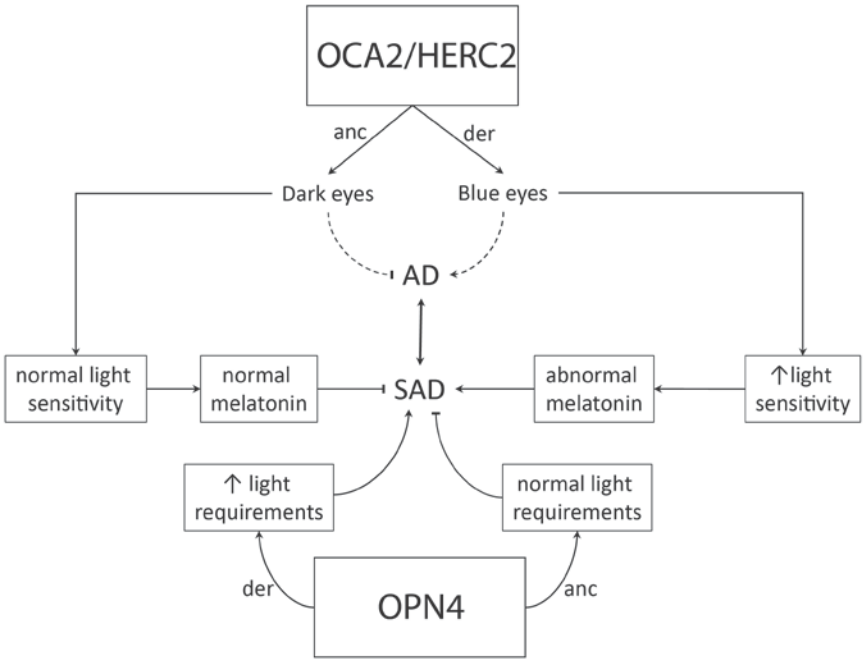
Supplementary Figure 1 Scatter plot of first three principal components for the EA and AA populations.



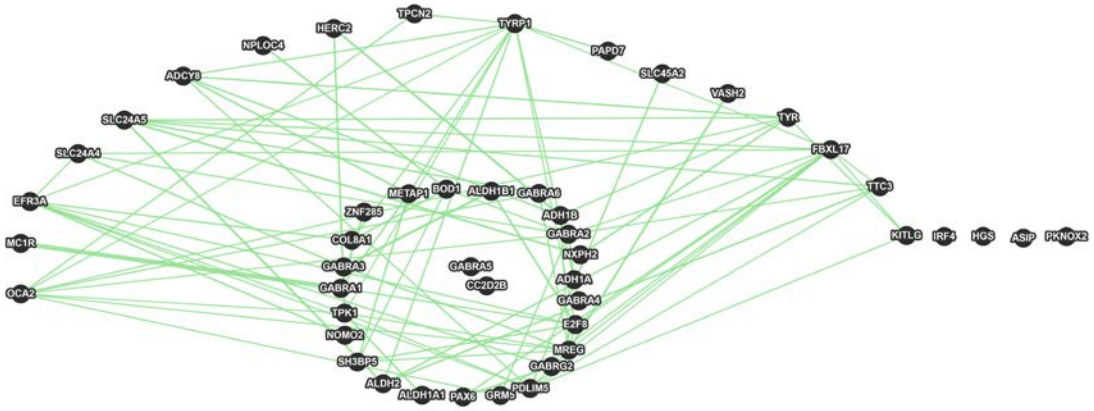
Supplementary Figure 2 Linkage disequilibrium blocks of the region encompassing *GABRG3* and *OCA2*. Yellow triangles mark the five linkage disequilibrium blocks. Bottom panel: the color of each pixel inside LD-blocks represents D'/LOD values (white (D' < 1 and LOD < 2), blue (D' = 1 and LOD < 2); shades of pink/red (D' < 1 and LOD ≥ 2), and bright red (D' = 1 and LOD ≥ 2)). Top panel: color of each pixel inside the LD-blocks represents r^2 values varying from 0 (white) to 1 (black). Of these five blocks, three lie exclusively intergenicly between *GABRG3* and *OCA2*, and two lie in the 3' and 5' UTR regions of *GABRG3* and *OCA2*.



Supplementary Figure 3 Approaches to testing the association between AD and eye color. The first type was at the population-phenotype level, connecting the eye color trait to AD in EAs. Then, the HapMap genetic data were utilized to measure the linkage disequilibrium between eye color gene regions and the GABA receptor genes regions on 15q12 and between the *GRM5-TYR* gene regions on 11q14 (genetic level). Finally, network analysis that leveraged genomic databases (GeneMANIA) provided insight into the type of biological interactions between selected AD candidate genes and eye color genes (network level).



Supplementary Figure 4 Proposed possible connections between eye color, light sensitivity, SAD and AD. Dotted lines represent our findings and solid lines represent literature evidence. Pointed arrows indicate positive association and flathead arrows indicate negative association. Single point mutations in the *OCA2/HERC2* (anc = ancestral, der = derived allele) region are determinants of blue-brown eye color trait in humans[172]. Blue eyed individuals are more sensitive to light when compared to brown eyed, which has been shown to infer significant melatonin production differences[144]. Melatonin production is one of several physiological factors that has been associated with supersensitivity to light variation in SAD subjects, via circadian-rhythm alterations[162]. SAD has been described to be comorbid with AD[143]. Another line of evidence supports the connection between light sensitivity and SAD, through melanopsin gene (*OPN4*) mutations[173]. The term "normal" refers to either ancestral allele or brown eye individuals' light sensitivity and melatonin levels, i.e., the base-lines.



Supplementary Figure 5 The gene-gene interaction network of selected AD-associated and eye color genes. The circle in the middle of the network corresponds to the AD gene set while the genes in the outer part are the eye color genes. The green lines depict the genetic interactions in gene pairs. Thickness of the green line corresponds to the strength of the interaction. Not all genes interact with each other and not all genes from one set have an interaction with genes in the other set. Supplementary Tables 2 and 3 show a list of all interacting gene pairs.

Supplementary Figure 6 Linkage disequilibrium blocks of the region encompassing *GRM5* and *TYR*. A total of five strong LD blocks span the intergenic distance between *GRM5* and *TYR*. Each pixel's color corresponds to D'/LOD values (white (D' < 1 and LOD < 2), blue (D' = 1 and LOD < 2); shades of pink/red (D' < 1 and LOD ≥ 2), and bright red (D' = 1 and LOD ≥ 2)). The number inside each pixel corresponds to the r^2 value ranging from 0 (i.e., 0.0) to 100 (i.e., 1.0).

Chapter 2.3: Further analyses support the association between light eye color and alcohol dependence

Arvis Sulovari[1,2], Henry R. Kranzler[3], Lindsay A. Farrer[4], Joel Gelernter[5,6,7] and Dawei Li[1,8,9*]

[1]*Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont*

[2]*Cell, Molecular and Biomedical Sciences Graduate Program, University of Vermont, Burlington, Vermont*

[3]*Department of Psychiatry, University of Pennsylvania School of Medicine and VISN 4 MIRECC, Philadelphia VAMC, Philadelphia, Pennsylvania*

[4]*Departments of Medicine (Biomedical Genetics), Neurology, Ophthalmology, Genetics & Genomics, Biostatistics, and Epidemiology, Boston University Schools of Medicine and Public Health, Boston, Massachusetts*

[5]*Department of Psychiatry, School of Medicine, Yale University, New Haven, Connecticut*

[6]*Department of Genetics, School of Medicine, Yale University, New Haven, Connecticut*

[7]*VA Connecticut Healthcare Center, West Haven, Connecticut and Department of Neurobiology, Yale University School of Medicine, New Haven, Connecticut*

[8]*Department of Computer Science, University of Vermont, Burlington, Vermont*

[9]*Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, Vermont*

*To whom correspondence should be addressed:
Dawei Li, Ph.D., Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, US. E-mail: dawei.li@uvm.edu

Key words: Alcohol Dependence; Association; Eye color; Meta-analysis; Linkage disequilibrium

Dear Editors,

We recently reported an association of eye color and alcohol dependence (AD)[174], on which Manzardo[175] commented. We agree with the author[175] that the identification of a benign trait like eye color as a risk factor for a complex disorder like AD warrants careful scrutiny of the study parameters and conclusions. To address the possible issues identified by Manzardo[176], we conducted additional analyses. The results continue to support the hypothesized association.

We assessed more fully the population structure of our research subjects. Ancestry information on great-grandparents, i.e., eight per subject, was used to evaluate the composition of the ancestral pool of our 1,263 European American (EA) subjects. This pool consisted of 8,075 ancestors, representing 24 European countries. Three European regions, Northern, Central, and Southern Europe, accounted for 41%, 30%, and 29% of the ancestral pool, respectively (see supplementary information for the definition of these regions). We assigned each sample to one of the three regions based on having more than one-third of their ancestry from that region. Samples with equal ancestry proportions in two groups, e.g., 40% northern and 40% southern, were removed from analysis. This process led to selection of 913 EA samples where ancestry could be defined. No evidence of significant heterogeneity was found among the three regions (Cochran's Q test P = 0.65). Meta-analysis of the datasets across the three regions showed evidence of significant association between light eye color and AD with OR (95% CI) = 1.44 (1.04 -

2.01) and $P = 0.029$ (Figure 1A). The results remained similar when the Northern and Central groups were merged. However, when the Central and Southern groups were merged, we observed evidence of stronger association with OR (95% CI) = 1.58 (1.14 - 2.18) and $P = 0.0059$ (Table 1, Supplementary Table 1, and Figure 1A). Figure 2 shows the results of principal component analysis where pink, blue, and green represent Northern, Central, and Southern Europe, respectively.

We assessed the diversity of genetic influences on AD by using the list of 334 genes reported by Manzardo[176]. First, we searched for pair-wise linkage disequilibrium (LD) between the 334 AD-related genes and the 21 eye color genes. We found two additional instances of LD: the F-box and leucine rich repeat protein 17 (*FBXL17*) and ephrin A-5 (*EFNA5*) genes on cytoband 5q21.3 (80kb apart, 15 LD blocks in the intergenic region, Figure 1B); and the nuclear protein localization protein 4 homolog (*NPLOC4*) and actin gamma 1 (*ACTG1*) genes on cytoband 17q25.3 (30kb apart, one strong LD block the intergenic region, Figure 1C). These findings complement our previous report of strong LD between eye color and AD-associated genes in 15q12 and 11q14.3. Second, we matched the 334 genes to the GeneMANIA[157] database (331 genes were mappable). Compared to our previously reported P value of 0.02, we observed evidence of more significant genetic interactions between the 21 eye color genes and 331 AD-related genes ($P = 0.0038$ and Figure 1D).

Additionally, we added two potential confounding parameters, household income and education level, in our logistic regression analysis to the previous covariates (which were age, sex, and the first three principal components). The association between blue eye color and alcohol dependence remained significant with OR = 1.86 (1.31 - 2.46) and $P = 5.2 \times 10^{-4}$ (Table 2).

Despite the lack of a readily available clinical explanation for the association, the additional analyses presented here provide more evidence supporting the hypothesis that light eye color may be a risk factor for alcohol dependence. Although we included several known potential confounders, we cannot exclude the possibility that our findings were affected by other population stratification factors or other unidentified confounders. Further investigation may clarify the contributions of genetic, behavioral, and cultural components to the reported association.

Acknowledgement
This work was supported by the Start-up Fund of the University of Vermont.

References

Manzardo A. 2015. Interpretation of Eye Color Associations with Alcohol Dependence Risk in European Americans. Am J Med Genet B Neuropsychiatr Genet.

Manzardo AM, McGuire A, Butler MG. 2015. Clinically relevant genetic biomarkers from the brain in alcoholism with representation on high resolution chromosome ideograms. Gene 560(2):184-94.

Sulovari A, Kranzler HR, Farrer LA, Gelernter J, Li D. 2015. Eye color: A potential indicator of alcohol dependence risk in European Americans. Am J Med Genet B Neuropsychiatr Genet 168(5):347-53.

Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, Morris Q. 2013. GeneMANIA prediction server 2013 update. Nucleic Acids Res 41(Web Server issue):W115-22.

Tables

Table 1 Meta-analyses of the selected samples with ancestry information

	OR (95% CI)	P(Z)	P(Q)
Northern + Central + Southern	1.44 (1.04-2.01)	**0.029**	0.66
(Northern+Central) + Southern	1.43 (1.03,1.98)	**0.033**	0.78
Northern + (Central+Southern)	1.58 (1.14-2.18)	**0.0059**	0.21

Table 2 Association results between eye color and alcohol dependence before (model 1) and after (model 1*) controlling for household income and education level

		Model 1		**Model 1* (Income and education)**	
Eye colors	**Cases/Controls**	**OR**	**P**	**OR**	**P**
Brown	368/130	-	-	-	-
Blue	377/70	1.83 (1.31-2.57)	**4.7×10^{-4}**	1.86 (1.31-2.46)	**5.2×10^{-4}**
Green	223/64	1.28 (0.90-1.83)	0.17	1.35 (0.94-1.95)	0.11
Grey	5/5	0.34 (0.09-1.30)	0.11	0.43 (0.01-1.78)	0.24
Brown-center	19/2	3.76 (1.04-24.14)	0.08	3.76 (0.99-24.8)	0.09

Figure Legends

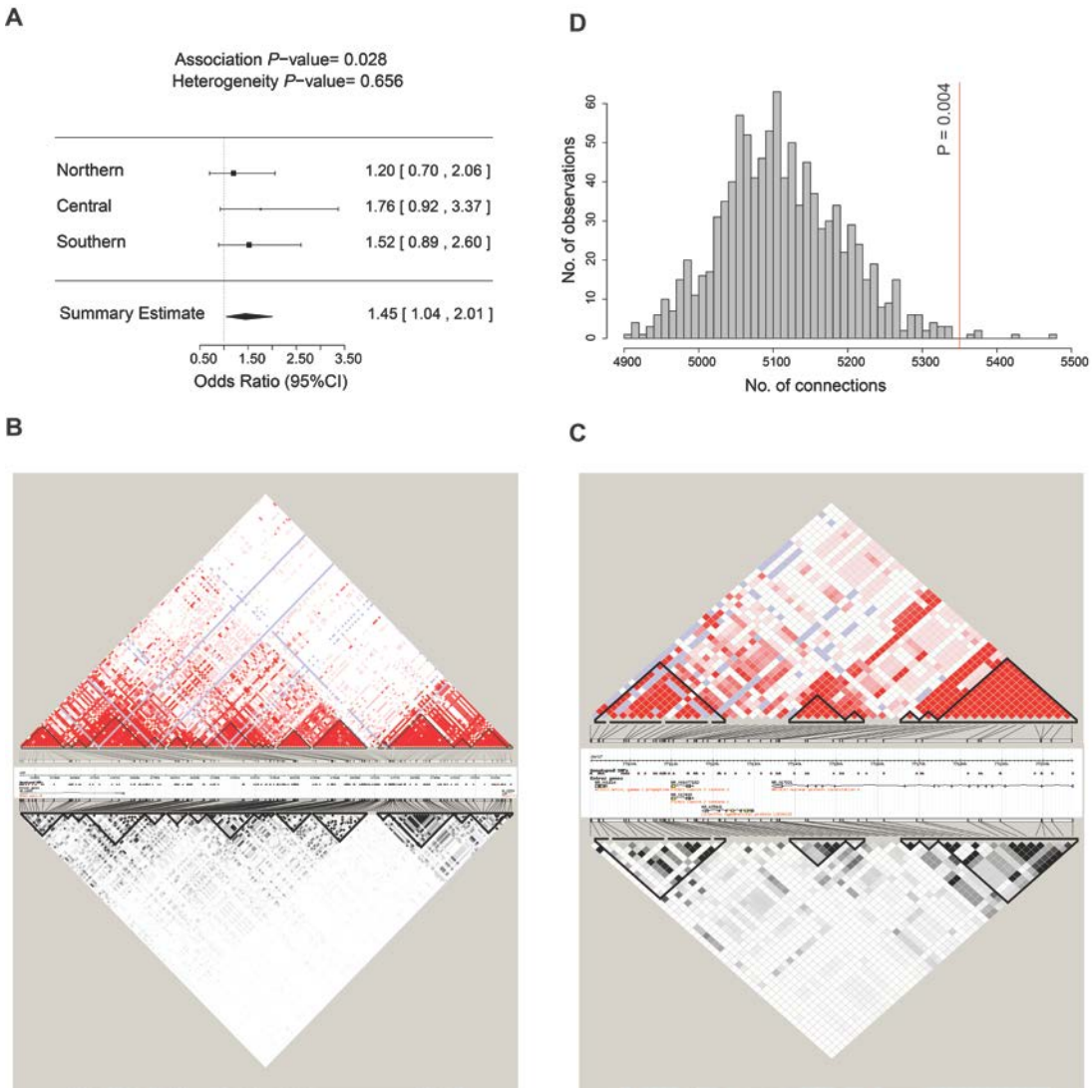


Figure 1 Panel of results from three analyses. A) Forrest plot of the meta-analysis of three major European regions: northern, central and southern. B-C) Linkage disequilibrium (LD) blocks in the regions encompassing *FBXL17* and *EFNA5* (B) and *NPLOC4* and *ACTG1* (C). Genotype data from the CEU and TSI populations of the HapMap project were used to estimate LD. The LD values are represented using D-prime (black and white) or R-square estimates (red and white). D) Distribution of genetic interactions between the 331 AD-related genes[176] and each of 1000 simulated gene sets of the 21 eye color genes. The vertical red line represents the number of genetic interactions between the 331 AD-related genes and 21 eye color genes, which are significantly higher than the 1,000 simulated genetic interactions (P = 0.0038).

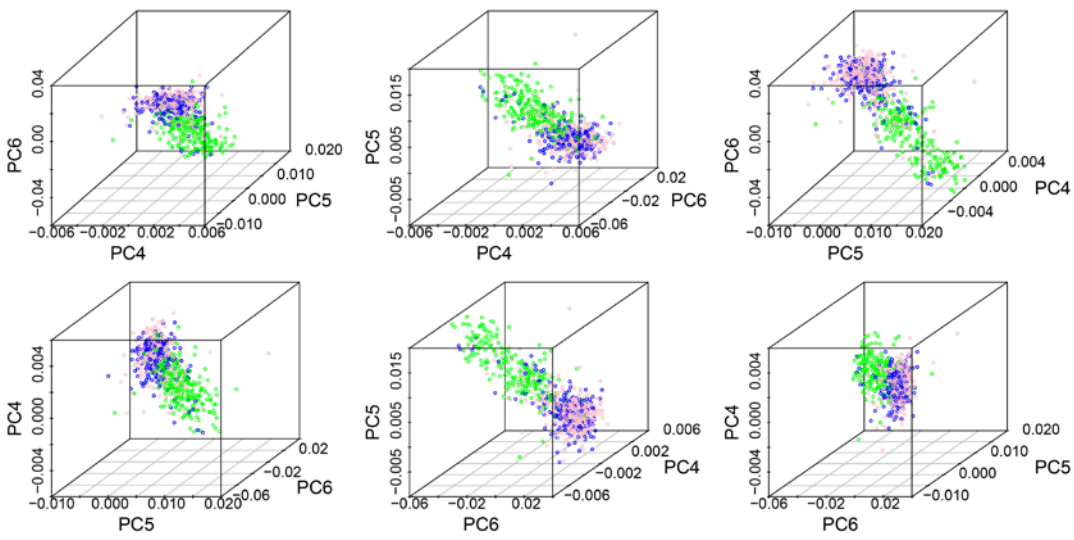


Figure 2 Results of principal component analysis. Pink, blue, and green represent Northern, Central, and Southern Europe, respectively. The 4th, 5th, and 6th principal components were used.

Supplementary Information

The three major European regions were defined according to the following nationality groupings:

Northern Europe: Danish, English, Finnish, Irish, Norwegian, Russian, Scottish, Swedish, and Welsh. Central Europe: Austrian, Belgian, Czechoslovakian, Dutch, French, German, Hungarian, Polish, Swiss. Southern Europe: Eastern Europeans (e.g., Albanian, Bulgarian), Greek, Italian, Portuguese, Spanish and Yugoslavian.

Supplementary Table

Supplementary Table 1 Results of individual association analyses of the selected samples with ancestry information

	Light (Case)	Light (Control)	Dark (Case)	Dark (Control)	OR (95% CI)	P
Northern	230	53	87	24	1.2 (0.7-2.06)	0.52
Central	136	23	74	22	1.76 (0.92-3.37)	0.089
Southern	80	29	100	55	1.52 (0.89-2.6)	0.13
Northern+Central	366	76	161	46	1.38 (0.91,2.07)	0.1275
Central+Southern	216	52	174	77	1.84 (1.23-2.75)	**0.0032**

CHAPTER 3: TOOLS AND RESOURCES FOR GENOME-WIDE ASSOCIATION STUDIES

Chapter 3.1: Multilevel ancestry informative markers (AIMs) for ancestry inferences and fine structures of world populations

-- A comprehensive AIMs panel set

Arvis Sulovari[1] and Dawei Li[1,2,3*]

[1]*Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA*

[2]*Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA*

[3]*Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, Vermont 05405, USA*

*To whom correspondence should be addressed:

Dawei Li, Ph.D., Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA. E-mail: dawei.li@uvm.edu

Number of words in the abstract: 191

Number of words in the text (excluding tables, figure, acknowledgments, financial disclosures, legends, and references): 3,343

Number of figures: 3

Number of supplementary material: 5 supplementary Tables and 6 supplementary Figures and Legends.

Abstract

Population stratification is a well-known source of false positive findings of disease genes in genetic association studies, particularly when research cohorts are genetically heterogeneous. With an increasing sample recruitment of multi-ethnic or international populations, it is necessary to identify powerful ancestry informative markers (AIMs) that can better capture between- and within-continental ancestry compositions. We analyzed 2,504 samples from the 1000 Genomes Project, representing five continental groups and 26 populations, and for each of the 325 possible population pairs we employed exhaustive whole-genome screen for new AIMs using the informativeness (I_N), fixation index (F_{ST}), and allele frequency difference (ΔDAF) methods. We constructed 325 AIMs panels, one for each population pair, with sizes from 136 to 735 markers per panel. 76 AIMs were highly recurrent in more than 120 population pairs. The panels have been demonstrated to separate population pairs of the same continental origin. The fine population structures inferred by our AIMs panels were also replicated by other methods, including principal component analysis, admixture analysis, and allele sharing. Our robust, multilevel AIMs panels can be used hierarchically to elucidate fine population structures in various studies using multi-ethnic or international samples.

Keywords: Ancestry informative marker (AIM), Genomic variation, Population structure**,** Ancestry prediction, Single nucleotide polymorphism (SNP), Genetic association study

Introduction

Genetic association studies have identified a large number of loci associated with complex human diseases, including those by us[152,177-179]. It is well-known that population genetic structure between cases and controls can confound associations leading to false positive or negative findings[180-182]. The increasing use of multi-ethnic or admixed populations in recent years has presented an unprecedented challenge due to the complex genetic heterogeneity . Ancestry informative markers (AIMs) are a set of genetic polymorphisms that exhibit substantial allele frequency differences between populations from different geographical regions of the world. AIMs have been widely used in genetic association studies to estimate the geographical origins of research subjects, such as continent-of-origin, and to evaluate the overall admixture proportions efficiently and inexpensively.

To correct for confounding factors by population stratification or estimate admixture, principal component analysis (PCA) of unlinked genotypes is commonly carried out in genetic association studies[174,183-185]. PCA captures latent variables that maximize variation between samples in high-dimensionality genotype data, serving as proxy for population structure and easily visualizing it based on allele frequency differences. This type of analysis can be conducted in studies with genome-wide genotypes; however, it is often not possible in studies with a smaller number of variants, such as candidate gene-based association studies[186] and targeted gene resequencing where only a small number of variants are genotyped or sequenced. Consequently, a panel of AIMs is required to

conduct PCA or similar analyses. In addition to inferring ancestry and controlling for population structure, AIMs panels have been proven to have a wide range of other applications, particularly in the identification of disease-associated genes. For example, population-specific AIMs have been successfully applied to associate sample admixture proportions to disease phenotypes, such as uterine blood flow in Andean samples[187] and breast cancer in Mexican women[188].

There are at least 21 recently-developed and widely-used AIM panels[189]; however, most of them were designed to identify only continent-of-origin[190-192] or for a specific population, e.g., Han Chinese[193] or Europeans[186]. For instance, a study of European and East Asian samples will use AIMs panels designed by different studies to capture ancestry differences between and within populations from the two continental groups. Multiple panels may exist for each scenario (e.g., European panels); however, they were likely designed using different approaches and genetic data sources, leading to a poor consensus across them. Indeed, a recent study identified an unexpectedly small overlap of 4% among ≥3 panels[189], and the overlapping markers could only predict continent-of-origin, but not the specific population-of-origin. This issue may be addressed by developing a multilevel set of AIMs panels for both between- and within-continent ancestry ascertainments using the same source of multi-ethnic genetic data. To our knowledge, no such panels have been published. Thus, a set of comprehensive AIMs panels that can ascertain sample ancestry or admixture proportion at global, continental, population, and particularly sub-population levels, is highly desirable.

Recently, we identified a large number of single nucleotide polymorphisms (SNPs) with large differences in allele frequencies between two or more continental populations from 78 million SNPs[194], most of which captured well the population structures. In this study, we systematically developed and validated a robust set of 325 AIMs panels (i.e., one per each possible population pair) for a total of 26 human populations[195]. All panels were built and calibrated using three different statistical methods, and their ancestry prediction value was evaluated on human samples from diverse populations.

Materials and Methods

Research subjects

The 1000 Genomes project included 2,504 unrelated individuals, representing 26 world populations from five continental groups (**Supp. Table S1**). The sample sizes were reasonably balanced with an average of 96 samples (standard deviation of 12) for each population.

Whole-genome single nucleotide polymorphisms (SNPs)

The SNP data was extracted from the most recent 1000 Genomes Project (Phase 3, last accessed on August 20 2015). The program Tabix[96] was used to extract genotypes from

the variant call format (VCF version 4.1) files, created using the human genome reference GRCh37. Only autosomal biallelic SNPs were used (**Supp. Table S2)**.

Panels of ancestry informative markers (AIMs)

To identify informative and new ancestry markers, we employed three independent methods, including informativeness (I_N), fixation index (F_{ST}), and allele frequency difference (ΔDAF). Each of the 26 populations was paired to every other population, resulting in a total of 325 population pairs. For each pair, we calculated I_N, F_{ST}, and ΔDAF using over 78 million SNPs, resulting in 76 billion calculations conducted in parallel on a high performance computing cluster, using in-house algorithms. First, genome-wide SNPs with F_{ST} value above the 99.9th percentile were identified, separately for each population pair. F_{ST} was calculated using an estimator, specifically derived for variants from sequencing studies, known to harbor large abundance of rare variants[88]. The F_{ST} estimator is defined as:

$$F_{ST} = \frac{(\bar{p}_1 - \bar{p}_2)^2 - \frac{\bar{p}_1(1-\bar{p}_1)}{(n_1-1)} - \frac{\bar{p}_2(1-\bar{p}_2)}{(n_2-1)}}{\bar{p}_1(1-\bar{p}_2) + \bar{p}_2(1-\bar{p}_1)}$$

, where $\bar{p}_1$ and $\bar{p}_2$ refer to allele frequencies in samples from populations 1 and 2, and n_1 and n_2 refer to sample sizes of populations 1 and 2, respectively. This method does not overestimate F_{ST} and has adequate power for analysis of both common and rare variants, due to its insensitivity to sample size differences between populations. The latter is important since sample sizes in real studies are often not perfectly matched between

populations (**Supp. Table S1**). Second, we calculated the I_N score for each SNP using the formula[196]:

$$I_N = \sum_{j=1}^{N}\left(-p_j \log p_j + \sum_{i=1}^{K}\frac{p_{ij}}{K}\log p_{ij}\right)$$

, where $p_j$is the average frequency of allele *j* over two populations (i.e., K=2), and p_{ij} is the frequency of allele *j* in population *i*; log represents natural logarithm with $0 \log 0 = 0$ [196]. Then, the SNPs were ranked based on their I_N score. Third, only the highest I_N-scoring SNP in LD-blocks as defined by $r^2 > 0.8$ were kept for further analyses. Cumulative informativeness was estimated using the top *n* markers for each population pair, such that the sum of top *n* I_N scores was varied from 5 to 50 in increments of 5. We pruned the AIMs by excluding markers with linkage disequilibrium $r^2 > 0.8$[197]. Lastly, ΔDAF (i.e., difference of derived allele frequencies) was calculated (i.e., $\Delta DAF_m = |p_i - p_j|_m$, where p_i and p_j are frequencies of SNP *m* in populations *i* and *j*) for each SNP. To ensure that the final AIMs had the largest difference in allele frequency, every SNP was required to satisfy $\Delta DAF_m \geq 0.05$. The AIMs with consensus results from all three methods were used for further analyses. If an AIM appeared in ≥ 120 population pairs, it was designated as a highly-recurrent AIM.

Evaluation of the AIMs panels

To measure the accuracy of our AIMs panels, three approaches were adopted for each population pair. First, we used the genotypes from each AIM panel, conducted PCA on

the respective samples, and assessed how well the AIMs panels clustered and assigned the samples in the two populations. Second, we carried out PCA on a new, validation set of 31 samples using the same AIMs panels. These were the relatives of the 2,504 unrelated samples used for the discovery of AIMs and they represent all five continental groups and 14 different populations. The k-means clustering was used to assign population-of-origin to each validation sample. Third, we repeated these analyses using randomly selected SNPs as a "negative control". The random sampling was conducted so that the probability of choosing a position on a given chromosome was proportional to its length, and the number of SNPs in the "negative control" panels was the same as that in the actual AIMs panel. Only SNPs with allele frequency ≥1% in both populations in the pair were used in the "negative control" set. The performance between our AIMs panels and the negative control panels was further compared by measuring the total variance in genetic ancestry explained by the first two principal components (i.e., PC1 and PC2). We focused on the population pairs within the same continental group (57 pairs in total), as those from different continental groups are known to be much easier to distinguish.

Principal component analysis (PCA)

The high performance computing toolset, SNPrelate[198], was used to carry out PCA using the VCF files of our identified AIMs. From the output of SNPrelate, the resulting eigenvectors and variance estimates of each principal component were utilized. PCA was conducted for each population pair, individual continental group, and all samples

combined. Since only unrelated samples underwent PCA, from the 31 validation samples, one (NA20336) was removed; in the 2,504 training samples, the relative, parent, child or siblings of the 31 validation samples were removed.

Population structures

In addition to the PCA plots, we employed STRUCTURE[199] on our identified AIMs panels to further elucidate the population structure. The program ADMIXTURE[200] was used to estimate the number of ancestries and genetic structures using the genotypic data from our AIMs panels in VCF. All statistical analyses and plots were conducted in the R statistical programming language (www.r-project.org).

Population genetic distances and visualization

To confirm the population structure indicated by PCA, allele sharing was measured as a proxy for genetic distances between all population pairs. PLINK/SEQ 0.10 (http://atgu.atgu.mgh.harvard.edu/plinkseq) was used to estimate pair-wise allele sharing for a total of 3,133,756 (i.e., $(2504^2 - 2504) / 2$) unique sample pairs. The heatmap of resulting allele sharing counts was constructed using the heatmap.2 function in the R statistical programming language.

Results

Identification of AIMs

After exhaustively screening all ten cumulative informativeness thresholds (see Methods) for each population pair, a threshold of 30 was selected as it produced a small number of markers and high population clustering accuracy. We identified a total of 325 AIMs panels with number of AIMs ranging from 136 (in PEL-JPT panel) to 735 (in CEU-GBR panel) in each panel. On average, each within- (i.e. populations of one continental group) and between-population (i.e. populations of different continental groups) pair had 415 (± standard deviation = 118) and 328 (± 73) AIMs, respectively, indicating that more AIMs are required to elucidate within- than between-population structures. Cumulatively, 2,919, 2,761, 1,910, 2,353 and 1,022 SNPs were identified specifically in Africans, Europeans, East Asians, South Asians, and Americans, respectively (**Supp. Table S3**). Most of the AIMs from population pairs within the same continental group were common SNPs with average allele frequencies of 42%, 33%, 38%, 32% and 37% in Africans, Europeans, East Asians, South Asians and Americans, respectively (**Supp. Figure S1**). Among these panels, 76 AIMs were recurrent in more than 120 population pairs, where the top two, rs7187359 and rs802566, occurred in 137 and 136 pairs (> 95% were between-population pairs), respectively.

Evaluation of AIMs panels

Each of the 57 within-population pairs underwent PCA using the respective AIMs panels. Overall, 44 of the population pairs (77%) separated perfectly from each other, while the rest have an almost perfect separation, in plots of the first two principal components (i.e., PC1 and PC2) with a very small number of exceptions (**Figure 1**). By comparison, the "negative control" AIMs produced PCA plots that did not distinguish between populations of the same continental group (**Supp. Figure S2**). To quantify differences in performance between AIMs panels, we calculated the cumulative genetic ancestry variance explained by the first two principal components. On average, the first two principal components of our AIMs panels explained nearly 24% more of the genetic ancestry variation (27.2 ± 11.5%) than the random set of SNPs (3.3 ± 1%). In addition, we assessed the accuracy of our AIMs panels by predicting ancestry of 30 different samples within each of the 14 respective populations. We found that 100% of the samples were correctly clustered (**Supp. Figure S3**).

Population structures from PCA

The first two principal components derived from all our identified AIMs were able to separate the samples very well by continental groups (**Supp. Figure S4**). As expected, the admixed American populations grouped closest to South Asians based on the principal component distances, followed by Africans, Europeans and East Asians. Next, each continental population was analyzed separately by combining AIMs panels from population pairs of the same continent, to identify fine population structures. In Africans,

all populations with exception of the admixed samples from ACB and ASW clustered in distinct regions. In Europeans, all populations grouped in distinct regions. Southern Europeans (TSI and IBS) were distinguishable from northern Europeans (GBR and CEU). In East Asians, we observed a relatively clear separation of the Japanese (JPT) samples from the Chinese populations. In South Asians, Gujarati separated distinctly from the other populations. In Americans, Puerto Ricans and Peruvians separated clearly from each other, while Mexicans in Los Angeles and Columbians displayed more heterogeneity. **Supp. Table S4** shows the contribution of first 10 principal components to genetic variance within each population.

Population structures from ADMIXTURE

The fine population structures describe above were well replicated by the admixture analyses[200]. **Figure 2** shows the estimated proportions of each ancestral group for a given genome. Under the assumption of two ancestral populations (i.e., K = 2) among the analyzed samples, Africans were separated from the rest of the populations. At K = 3, East Asians were separated from other non-Africans. At K = 4, South Asians became distinguishable from the rest. At K = 5, within- population structures appeared, e.g., Gujarati Indian from Texas became distinguishable from the rest of South Asians. At K = 7, non-admixed Africans separated from admixed Africans (ACB and ASW). At K = 8, Japanese became distinct from the rest of East Asians. The admixture plots using our AIMs panels produced consistent population structures to those from PCA.

American populations had average proportions of 53%, 35%, 8%, and 4% of European, East Asians, African, and South Asian ancestries, respectively (**Supp. Table S5** and **Supp. Figure S5**). The highest European, African, and South Asians ancestry proportions were all found in Puerto Ricans (65%, 16% and 6.4%, respectively), whereas the highest East Asian ancestry proportion was observed in Peruvians (62%). Furthermore, Peruvians contained the lowest European and African ancestries (33% and 3.5% respectively).

Population structures from allele sharing

Rare variants and distant ancestry

The abundance of rare alleles in the 1000 Genomes variants (84.6% of SNPs had DAF < 1%) allows us to asses rare allele sharing patterns between samples as a measure of population structure. The numbers of rare alleles shared by two individuals from the same continental group were significantly higher than those from different groups (**Figure 3** upper triangle; t-test $P < 2 \times 10^{-12}$), reflecting the more recent shared ancestry within a continental group.

Doubletons and recent ancestry

Doubletons are genetic variants shared by any two of the 2,504 individuals. Doubleton sharing, i.e., the proportion of doubleton variants shared by two individuals among total doubletons observed in both, elucidates recent ancestry[90,201]. High levels of doubleton sharing reflect identity-by-descent, i.e., genetic homogeneity, due to shared, recent,

population history. As expected, we observed higher doubleton sharing within the same, rather than between different, continental groups (**Figure 3** lower triangle). For instance, the average doubleton sharing between two African individuals was 3.3%, while it was only 0.0002% between an African and a non-African. All five continental groups revealed similar patterns of doubleton sharing ($P > 0.05$); however, inclusion of admixed samples significantly decreased doubleton sharing. For instance, the doubleton sharing in the total African samples, including the admixed ASW and ACB samples, were significantly lower when compared to doubleton sharing within Europeans ($P = 8 \times 10^{-6}$), East Asians ($P = 9 \times 10^{-6}$) or South Asians ($P = 1 \times 10^{-6}$). Therefore, combining admixed samples with samples from their ancestral population will exacerbate effects of population stratification.

Allele sharing patterns **(Figure 3)** clearly portray the recent admixture of European and African ancestries in modern Americans. The ancestral lineages (upper triangle) of Colombians and Puerto Ricans contained higher African components than those of Mexicans from Los Angeles and Peruvians (except one Peruvian individual). This observation was consistent with our admixture analysis results (**Supp. Table S5**), where Colombians and Puerto Ricans showed average African ancestry proportions of 9% and 16%, compared to 4.5% and 3.5% in Mexicans and Peruvians. The data also suggested that Japanese had rapid and recent population growth (the second strongest doubleton sharing among all 26 populations). Our results also support a recent admixture of East

Asians and South Asians (lower triangle). The fine population structures revealed by allele sharing analyses were consistent with those inferred by the AIMs panels.

Discussion

In the present study, we analyzed over 78 million biallelic SNPs and created a total of 325 panels of AIMs, corresponding to all possible pairs of 26 world populations from five continental groups. Each of these panels can be applied flexibly to discriminate between any specific population pair in genetic association studies, depending on sample ancestry composition. We have demonstrated the robustness of our panels based on the near-perfect separation of samples from closely related populations (e.g., CHS and CHB), and perfect prediction accuracy of validation samples.

On average, each panel had 343 ± 89 AIMs (range from 136 to 735), and 76 AIMs were highly recurrent among these panels. Our panels distinguished particularly well population pairs within continental groups, as demonstrated by the reasonably homogenous sample clusters by PCA (**Figure 1**). The resulting fine population structures and admixture proportions were consistent with the expected geographic and cultural differences in these samples. For instance, the Japanese (JPT) and Han Chinese (CHB) samples were separated more easily than the southwestern Chinese (CDX) and Vietnamese (KHV) samples. On the other hand, around 96% of the AIMs identified in this study are non-coding, common SNPs. However, we also found some

nonsynonymous AIMs located in genes that have been reported for positive selection, such as rs1871534 (L347V) in *SLC39A4*[202], rs16891982 (L374F) in *SLC45A2*[132], rs60910145 (I366M) in *APOL1*[52], and rs3827760 (V370A) in *EDAR* [68]. Additionally, two of the AIMs were recently published by our group[194] as highly pathogenic variants with extreme allele frequency differences in populations of the same continental group: rs200071340 (Gln39Ter) in Europeans and rs3211938 (Tyr325Ter) in Africans.

To build these AIMs panels, we adopted three statistical scoring systems, i.e., I_N, F_{ST}, and ΔDAF, which yielded highly correlated results in our study (**Supp. Figure S6**). The AIMs panels developed here were highly informative for ancestry, as measured by I_N. For example, among the top 12 AIMs of a recently published Han Chinese panel[193], four overlapped with our panel of the equivalent population pairs, i.e., CHB-CHS; however, our panel contained a larger number of high I_N markers, i.e., 192 AIMs with $I_N \geq 0.028$ in our panel compared to only two in the published panel. A detailed comparison between our panel and the one published by Qin et al. revealed that our panel has more informative markers, as measured by both F_{ST} and I_N statistics (**Supp. Figure S7**).

A recent study evaluated 21 published AIMs panels and found 1%, i.e., 14 AIMs, overlap among four or more panels or 3%, i.e., 46 AIMs, overlap among three or more panels[189]. By comparison, our panels contained all of the 14 AIMs or 42 of the 46 AIMs (markers in strong LD were also considered a match), indicating high consistency. It should be noted that in this study, only 2,504 whole-genomes were analyzed for our AIMs

development, and it is anticipated that future research with larger sample size may identify more new markers.

Regarding applications of our AIMs panels, we recommend using these panels hierarchically. For instance, a study that analyzes samples of African and East Asian ancestry may first use one or more of our AIMs panels that were designed for separating Africans from East Asian populations, then use the panels that separate specific African populations from one another, and those that separate specific East Asian populations from one another. This strategy prevents inclusion of AIMs designed for populations that are not represented in the underlying study. To the best of our knowledge, this study provides the first set of AIMs panels that can ascertain sample ancestry or admixture proportion with high accuracy at multiple resolutions, i.e., global, continental, population, and sub-population levels.

To conclude, in this study we have identified and validated a new set of multilevel AIMs panels. They have various potential applications, including ancestry inference at sub-population resolution, and gene-disease fine mapping studies in admixed or multi-ethnic cohorts.

Data archiving

The AIMs markers are available at: http://www.uvm.edu/genomics for download.

Acknowledgements

This work was supported by the Start-up Fund of The University of Vermont. The raw data analyzed in this study were from the 1000 Genomes Project (Phase 3). We would like to thank Zoe Furlong for her helpful feedback during preparation of the manuscript. We are grateful to Drs. Xun Chen and Guangchen Liu for their critical comments and feedback throughout the process of preparing this manuscript.

Conflict of Interest

The authors declare no potential conflict of interest.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655-64. doi: 10.1101/gr.094052.109

Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu YM, Wang J, Chang YQ, Feng Q, Fang XD, Guo XS, Jian M, Jiang H, Jin X, Lan TM, Li GQ, Li JX, Li YR, Liu SM, Liu X, Lu Y, Ma XD, Tang MF, Wang B, Wang GB, Wu HL, Wu RH, Xu X, Yin Y, Zhang DD, Zhang WW, Zhao J, Zhao MR, Zheng XL, Lander ES, Altshuler DM, Gabriel SB, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Flicek P, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Bentley DR, Grocock R, Humphray S, James T, Kingsbury Z, Lehrach H, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, et al. (2015) A global reference for human genetic variation. Nature 526: 68-+. doi: 10.1038/nature15393

Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting FST: the impact of rare variants. Genome Res 23: 1514-21. doi: 10.1101/gr.154831.113

Cao J, Hudziak JJ, Li D (2013a) Multi-cultural association of the serotonin transporter gene (SLC6A4) with substance use disorder. Neuropsychopharmacology 38: 1737-47. doi: 10.1038/npp.2013.73

Cao J, Liu X, Han S, Zhang CK, Liu Z, Li D (2013b) Association of the HTR2A gene with alcohol and heroin abuse. Hum Genet 133: 357-65. doi: 10.1007/s00439-013-1388-y

Cooper RS, Tayo B, Zhu X (2008) Genome-wide association studies: implications for multiethnic samples. Hum Mol Genet 17: R151-5. doi: 10.1093/hmg/ddn263

Engelken J, Carnero-Montoro E, Pybus M, Andrews GK, Lalueza-Fox C, Comas D, Sekler I, de la Rasilla M, Rosas A, Stoneking M, Valverde MA, Vicente R, Bosch E (2014) Extreme population differences in the human zinc transporter ZIP4 (SLC39A4) are explained by positive selection in Sub-Saharan Africa. PLoS Genet 10: e1004128. doi: 10.1371/journal.pgen.1004128

Fejerman L, Romieu I, John EM, Lazcano-Ponce E, Huntsman S, Beckman KB, Perez-Stable EJ, Gonzalez Burchard E, Ziv E, Torres-Mejia G (2010) European ancestry is positively associated with breast cancer risk in Mexican women. Cancer Epidemiol Biomarkers Prev 19: 1074-82. doi: 10.1158/1055-9965.EPI-09-1193

Genome of the Netherlands C (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46: 818-25. doi: 10.1038/ng.3021

Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65. doi: 10.1038/nature11632

Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, Bowden DW, Langefeld CD, Oleksyk TK, Uscinski Knob AL, Bernhardy AJ, Hicks PJ, Nelson GW, Vanhollebeke B, Winkler CA, Kopp JB, Pays E, Pollak MR (2010) Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science 329: 841-5. doi: 10.1126/science.1193032

Huckins LM, Boraska V, Franklin CS, Floyd JA, Southam L, Gcan, Wtccc, Sullivan PF, Bulik CM, Collier DA, Tyler-Smith C, Zeggini E, Tachmazidou I, Gcan, Wtccc (2014a) Using ancestry-informative markers to identify fine structure across 15 populations of European origin. Eur J Hum Genet 22: 1190-200. doi: 10.1038/ejhg.2014.1

Huckins LM, Boraska V, Franklin CS, Floyd JA, Southam L, Sullivan PF, Bulik CM, Collier DA, Tyler-Smith C, Zeggini E, Tachmazidou I (2014b) Using ancestry-informative markers to identify fine structure across 15 populations of European origin. Eur J Hum Genet 22: 1190-200. doi: 10.1038/ejhg.2014.1

Julian CG, Wilson MJ, Lopez M, Yamashiro H, Tellez W, Rodriguez A, Bigham AW, Shriver MD, Rodriguez C, Vargas E, Moore LG (2009) Augmented uterine artery blood flow and oxygen delivery protect Andeans from altitude-associated reductions in fetal growth. Am J Physiol Regul Integr Comp Physiol 296: R1564-75. doi: 10.1152/ajpregu.90945.2008

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, Powell A, Itan Y, Fuller D, Lohmueller J, Mao J, Schachar A, Paymer M, Hostetter E, Byrne E, Burnett M, McMahon AP, Thomas MG, Lieberman DE, Jin L, Tabin CJ, Morgan BA, Sabeti PC (2013) Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152: 691-702. doi: 10.1016/j.cell.2013.01.016

Kersbergen P, van Duijn K, Kloosterman AD, den Dunnen JT, Kayser M, de Knijff P (2009) Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. BMC Genet 10: 69. doi: 10.1186/1471-2156-10-69

Li D, He L (2008) Meta-study on association between the monoamine oxidase A gene (MAOA) and schizophrenia. Am J Med Genet B Neuropsychiatr Genet 147B: 174-8. doi: 10.1002/ajmg.b.30570

Li H (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27: 718-9. doi: 10.1093/bioinformatics/btq671

Li YR, Keating BJ (2014) Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med 6: 91. doi: 10.1186/s13073-014-0091-5

Liu J, Shi Y, Tang W, Guo T, Li D, Yang Y, Zhao X, Wang H, Li X, Feng G, Gu N, Zhu S, Liu H, Guo Y, Shi J, Sang H, Yan L, He L (2005) Positive association of the human GABA-A-receptor beta 2 subunit gene haplotype with schizophrenia in the

Chinese Han population. Biochem Biophys Res Commun 334: 817-23. doi: S0006-291X(05)01404-X [pii]

10.1016/j.bbrc.2005.06.167

Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. Science 201: 786-92.

Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF (2009) An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genet 10: 39. doi: 10.1186/1471-2156-10-39

Nievergelt CM, Maihofer AX, Shekhtman T, Libiger O, Wang X, Kidd KK, Kidd JR (2013) Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. Investig Genet 4: 13. doi: 10.1186/2041-2223-4-13

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-9. doi: 10.1038/ng1847

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-59.

Qin P, Li Z, Jin W, Lu D, Lou H, Shen J, Jin L, Shi Y, Xu S (2014) A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. Eur J Hum Genet 22: 248-53. doi: 10.1038/ejhg.2013.111

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73: 1402-22. doi: 10.1086/380416

Soundararajan U, Yun L, Shi M, Kidd KK (2016) Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. Forensic Sci Int Genet 23: 25-32. doi: 10.1016/j.fsigen.2016.01.013

Sulovari A, Chen YH, Hudziak JJ, Li D (2017) Atlas of human diseases influenced by genetic variants with extreme allele frequency differences. Hum Genet 136: 39-54. doi: 10.1007/s00439-016-1734-y

Sulovari A, Kranzler HR, Farrer LA, Gelernter J, Li D (2015a) Eye color: A potential indicator of alcohol dependence risk in European Americans. Am J Med Genet B Neuropsychiatr Genet 168B: 347-53. doi: 10.1002/ajmg.b.32316

Sulovari A, Kranzler HR, Farrer LA, Gelernter J, Li D (2015b) Further analyses support the association between light eye color and alcohol dependence. Am J Med Genet B Neuropsychiatr Genet 168: 757-60. doi: 10.1002/ajmg.b.32357

Traylor M, Lewis CM (2016) Genetic discovery in multi-ethnic populations. Eur J Hum Genet. doi: 10.1038/ejhg.2016.38

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterlander M, Hollfelder N, Potekhina ID, Schier W, Thomas MG, Burger J (2014) Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A 111: 4832-7. doi: 10.1073/pnas.1316513111

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28: 3326-8. doi: 10.1093/bioinformatics/bts606

Figure Legends

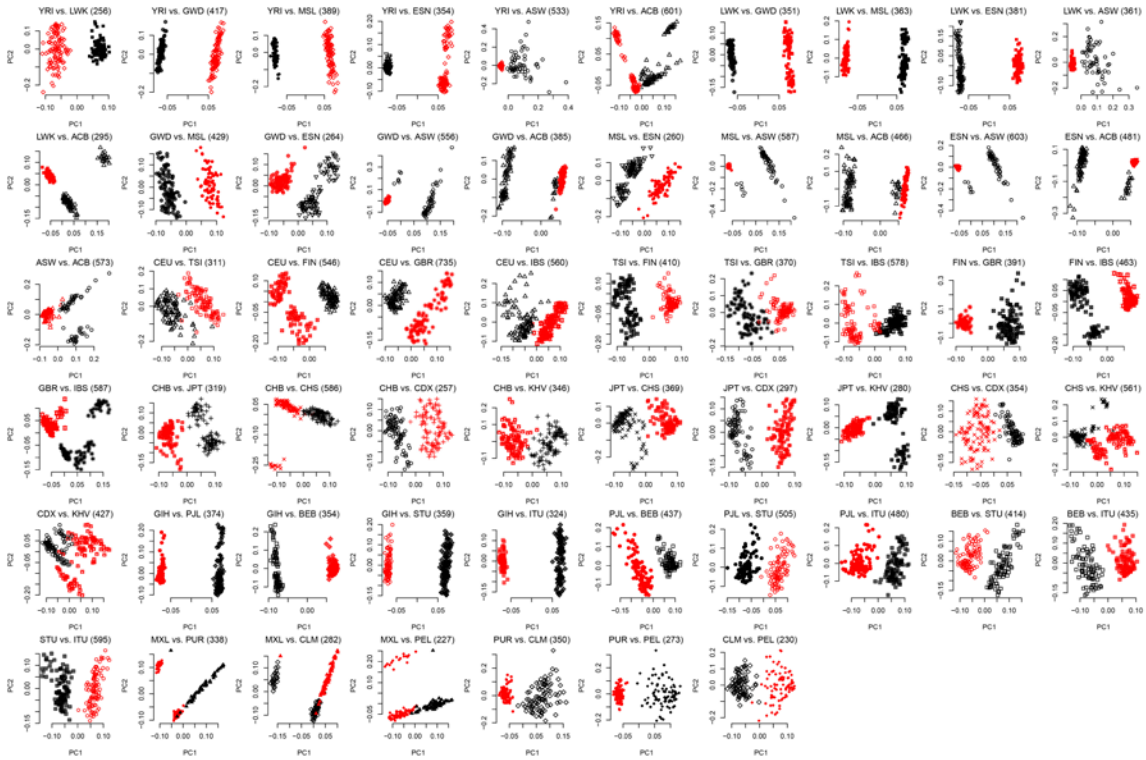


Figure 1 PCA plots of all 57 population pairs within the same continental group inferred using our AIMs panels. Every possible pair of populations was separated from each other at almost perfect levels. Some of the exceptions included CDX (Chinese Dai in Xishuangbanna, China) – KHV (Kinh in Ho Chi Minh City, Vietnam), and to a smaller degree CHS (Southern Han Chinese) – CHB (Han Chinese in Bejing, China) and MXL (Mexican ancestry from Los Angeles, USA) – CLM (Colombians from Mendellin, Colombia) pairs. In all three cases, the geographical proximity, admixture status or shared recent ancestry of these populations may account for the slight difficulty in distinguishing them. The order of population pairs is consistent with that on the official 1000 Genomes Project website (http://www.1000genomes.org/category/population).

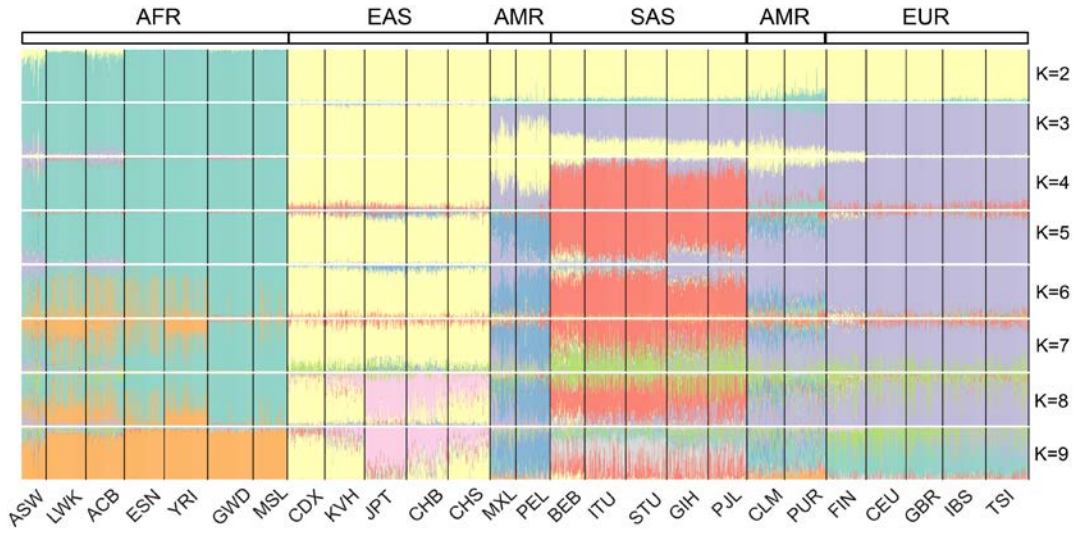


Figure 2 A population structure based on our AIMs panels. All of the 10,243 AIMs were used. Each color corresponds to an estimated ancestral group (referred to as K). The order of plots from top to bottom corresponds to K values of 2 to 9. The program ADMIXTURE was used to measure ancestral proportions in each sample. The order of populations was determined by genetic distance between them, based on pairwise F_{ST} measurements.

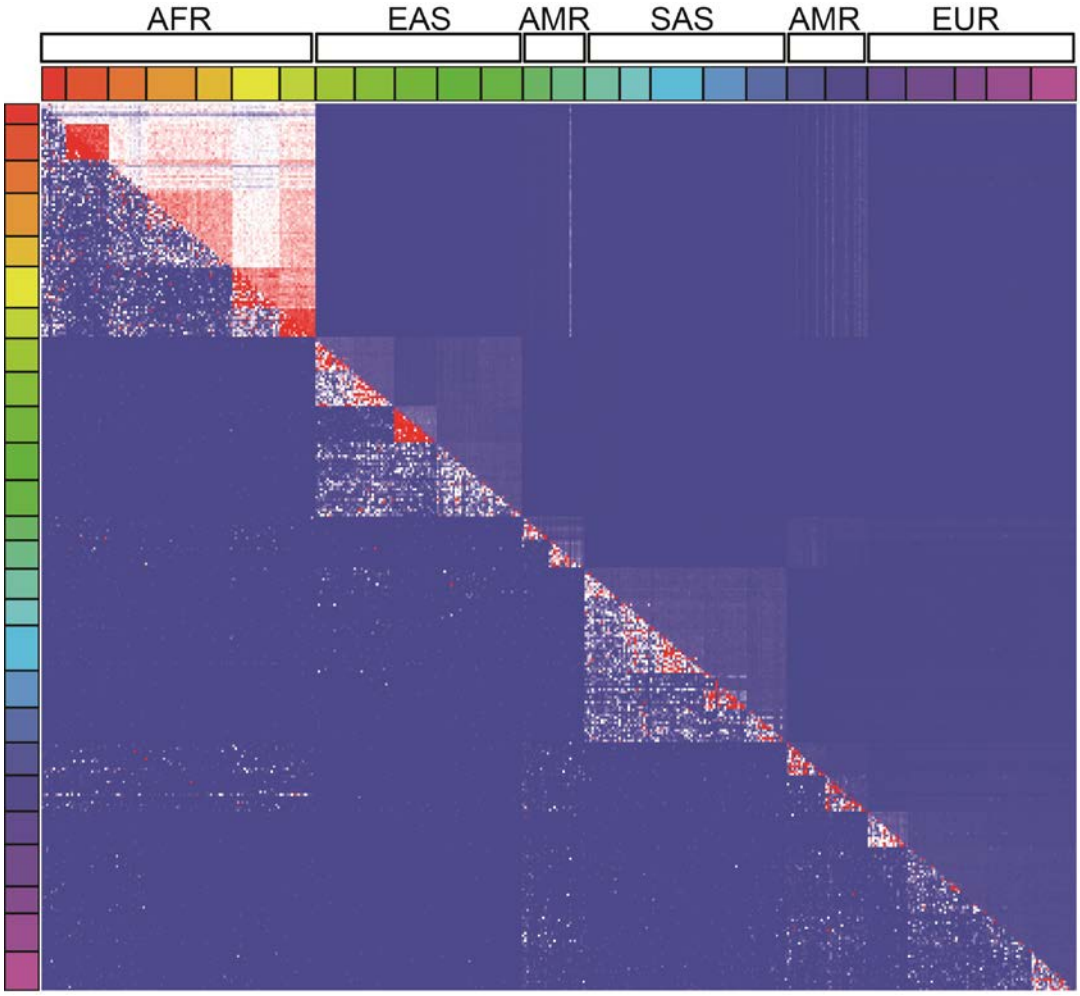


Figure 3 Allele sharing between individual pairs. Two allele sharing analyses between all possible unique pairs of unrelated individuals (3,133,756 sample pairs in total). The lower triangle of the heatmap corresponds to the recent ancestry measured by the doubleton sharing pattern defined in Plink/Seq (http://atgu.mgh.harvard.edu/plinkseq) (e.g., sample A has 1,000 doubletons and sample B has 2,000 doubletons; of these, 500 are shared by both; thus their doubleton sharing = (2 × 500)/(2000+1000) = 0.34 or 34%). The upper diagonal corresponds to the more ancient ancestry as measured by the sharing of variants with DAF < 1% between each sample pair. The blue (low), white (average), and red (high) color scheme is used in both halves of the heatmap.

Supplementary Tables and Figures

Supplementary Tables

Supp. Table S1 Summary of the samples analyzed in this study

Continental groups	Populations of each continental group	Sample sizes	Total
AFR	ACB, ASW, ESN, GWD, LWK, MSL, YRI	96, 61, 99, 113, 99, 85, 108	661
AMR	CLM, MXL, PEL, PUR	94, 64, 85, 104	347
EAS	CDX, CHB, CHS, JPT, KHV	93, 103, 105, 104, 99	504
EUR	CEU, FIN, GBR, IBS, TSI	99, 99, 91, 107, 107	503
SAS	BEB, GIH, ITU, PJL, STU	86, 103, 102, 96, 102	489

The three-letter codes represent the following continental groups: EAS, East Asian; SAS, South Asian; AMR, admixed populations from the Americas; EUR, European populations; AFR, African populations. The order of sample sizes corresponds to the populations order. Codes of populations within each continental group correspond to African Carribbeans in Barbados (ACB); Americans of African Ancestry in Southwest of USA (ASW); Esan in Nigeria (ESN); Gambian in Western Divisions in the Gambia (GWD); Luhya in Webuye, Kenya (LWK); Mende in Sierra Leone (MSL); Yoruba in Ibadan, Nigeria (YRI); Columbians from Medellin, Colombia (CLM); Mexican Ancestry from Los Angeles USA (MXL); Peruvians from Lima, Peru (PEL); Puerto Ricans from Puerto Rico (PUR); Chinese Dai in Xishuangbanna, China (CDX); Han Chinese in Beijing, China (CHB); Southern Han Chinese (CHS); Japanese in Tokyo, Japan (JPT); Kinh in Ho Chi Minh City, Vietnam (KHV); Utah residents with Northern and Western European Ancestry (CEU); Finnish in Finland (FIN); British in England and Scotland (GBR); Iberian Population in Spain (IBS); Toscani in Italy (TSI); Bengali from Bangladesh (BEB); Gujarati Indian from Houston, Texas (GIH); Indian Telugu from the UK (ITU); Punjabi from Lahore, Pakistan (PJL); Sri Lankan Tamil from the UK (STU).

Supp. Table S2 Summary of the total variants in the 1000 Genomes Project Phase 3 subjects

Variant Types	Counts	Percent of total variants
SNPs	78,136,341	96.1%
Indels	3,135,424	3.9%
Biallelic SNPs and indels	80,800,311	99.4%

Supp. Table S3 AIMs for population pairs among the primary CEU, CHB, JPT, and YRI populations (see https://www.uvm.edu/genomics/publications.html). The data contains positional information (build 37), I_N, F_{ST} and ΔDAF scores for each AIM. Additional AIMs are available upon request.

Supp. Table S4 Genetic variance explained by our AIMs panels

Principal component	Variance* explained (%)					
	All	AFR	EUR	EAS	SAS	AMR
1	11.7	8.79	4.14	5.82	8.76	18.72
2	6.26	4.04	2.37	4.14	2.16	8.77
3	2.79	2.08	1.95	2.85	1.83	1.96
4	1.93	1.79	1.81	2.18	1.36	1.70
5	0.81	1.29	1.66	1.81	1.21	1.06
6	0.68	0.99	1.09	1.61	1.11	0.84
7	0.66	0.84	0.95	1.43	0.97	0.80
8	0.59	0.80	0.79	1.41	0.74	0.77
9	0.51	0.64	0.72	1.26	0.70	0.70
10	0.44	0.62	0.68	1.10	0.62	0.66

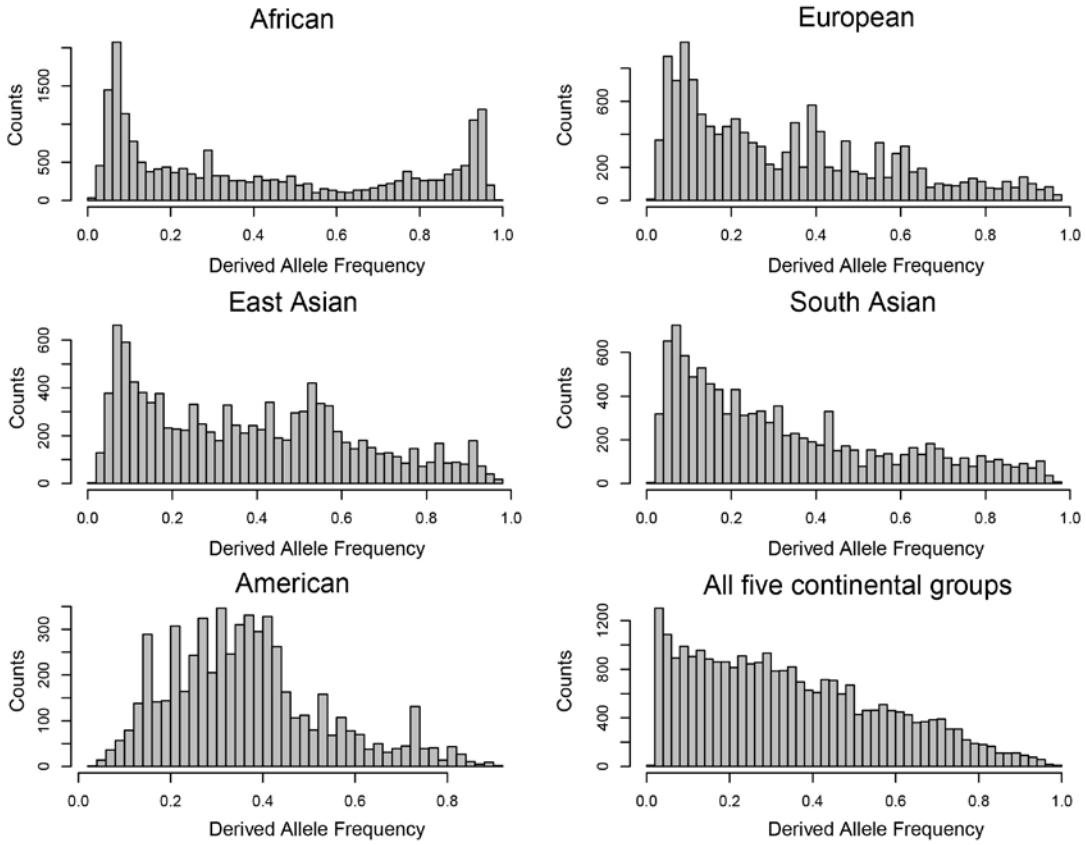
*Variation in genetic ancestry among the 2,504 samples. All, 10,243 SNPs; AFR, 2,919; EUR, 2,761; EAS, 1,910; SAS, 2,353; and AMR, 1,022 SNPs.

Supp. Table S5 The four major ancestral proportions of four American populations.

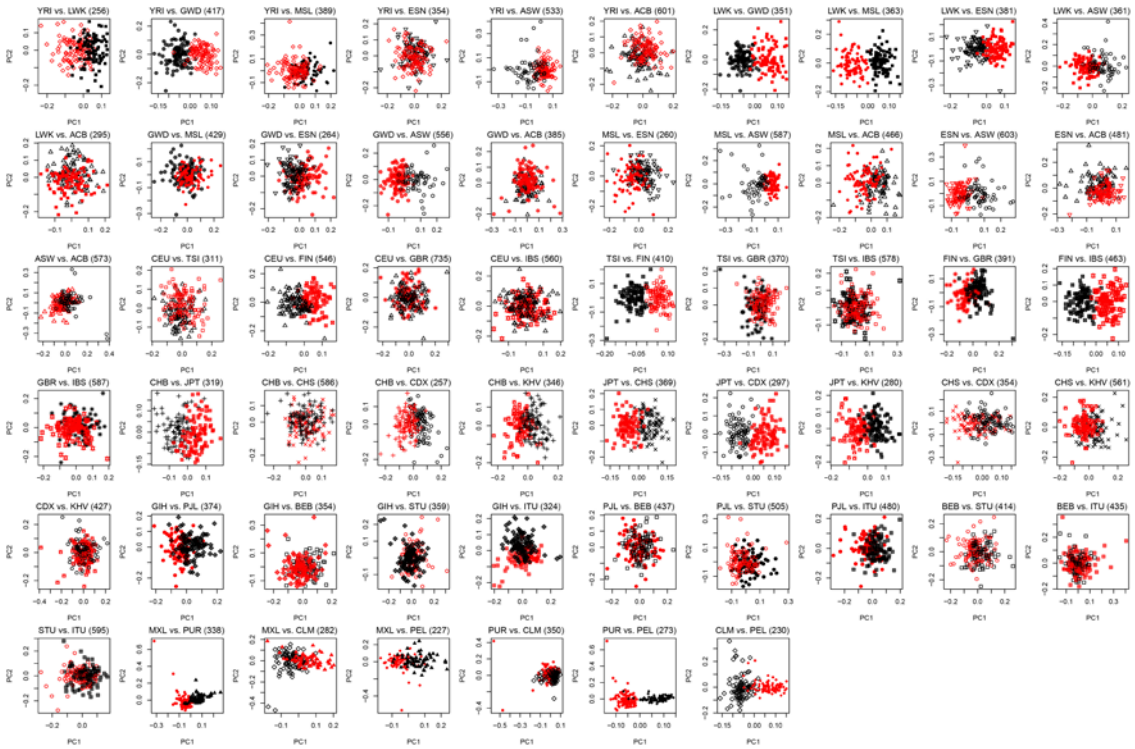
Ancestral group	**MXL**	**PEL**	**CLM**	**PUR**
European	52 (±14)%	33 (±9)%	62 (±12)%	65 (±11)%
East Asian	42 (±15)%	62 (±11)%	23 (±9)%	13 (±5)%
African	4.5 (±3)%	3.5 (±6)%	9 (±8)%	16 (±9)%
South Asian	1.7 (±4.4)%	2.1 (±4)%	6 (±9)%	6.4 (±7.5)%

The average (±standard deviation) ancestry proportions were estimated using results from ADMIXTURE, at K=4.

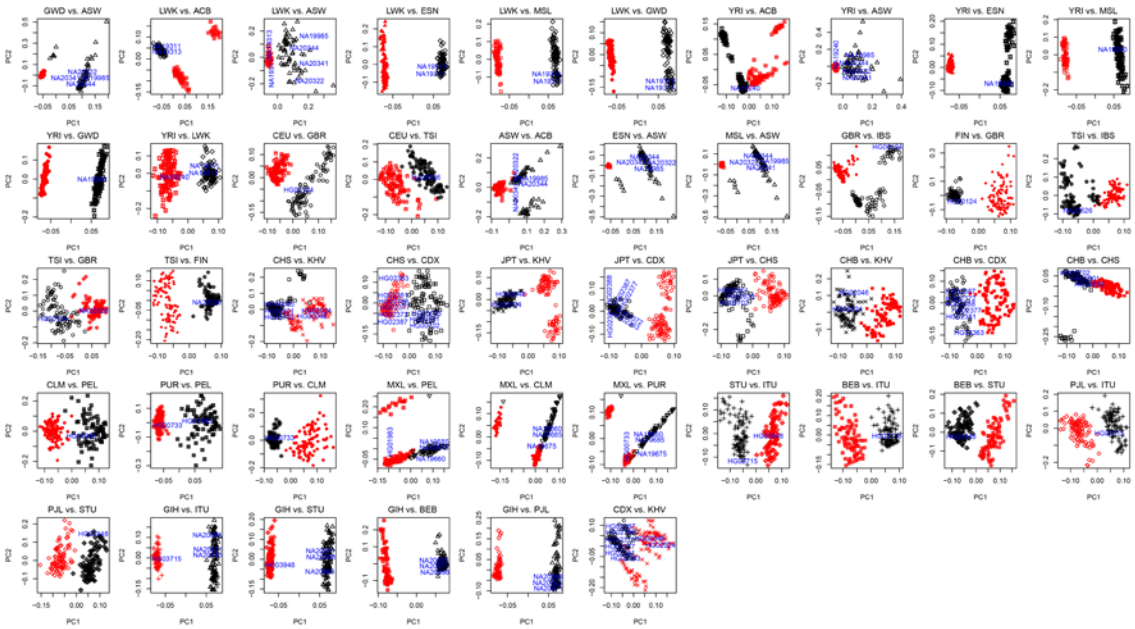
Supplementary Figures



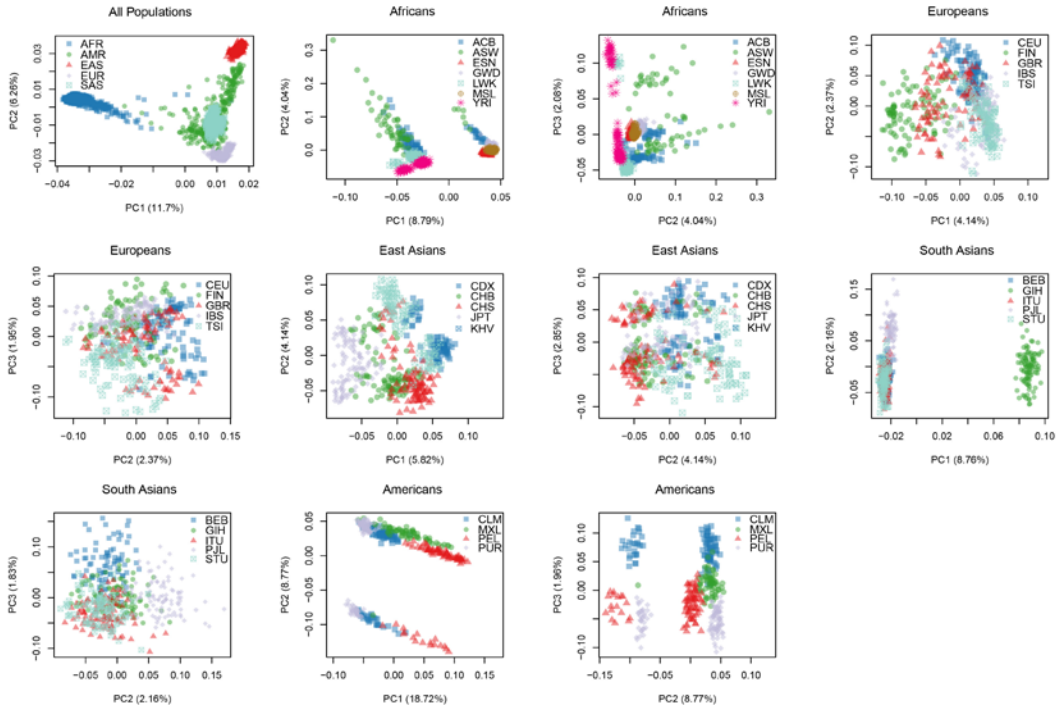
Supp. Figure S1 Allele frequency histograms for AIMs of each continental group. The values on the y-axis correspond to the number of markers with a specific derived allele frequency, denoted on the x-axis.



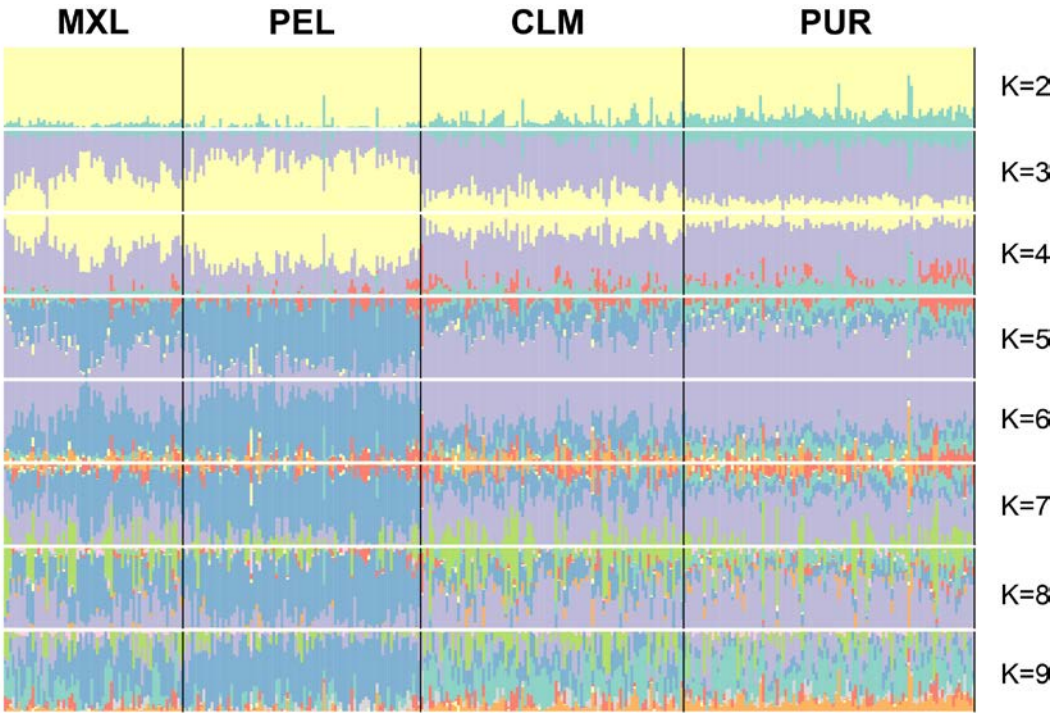
Supp. Figure S2 PCA plots of all 57 population pairs within the same continental group using random SNPs. The same numbers of randomly-selected SNPs as in the AIMs panel were used to conduct PCA on each within-population pair. The "negative control" panels failed to reveal expected population structures. The only populations pairs that seemed to separate well using the negative control set of AIMs were LWK (Luhya in Webuye, Kenya) – MSL (Mende in Sierra Leone), FIN (Finnish in Finland) – IBS (Iberian Population in Spain) and TSI (Toscani in Italy) – FIN (Finnish in Finland). The reason that these populations (i.e., 5%) separated from each other might be due to sufficiently large differences in genetic background or large number of markers applied.



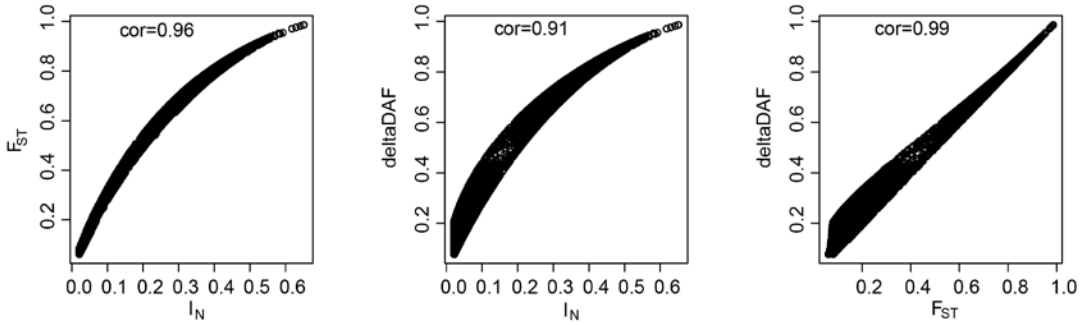
Supp. Figure S3 PCA plots of population pairs with new, validation samples. All of the 30 new samples were successfully clustered with the appropriate population. Each of these samples is a relative, parent, child or sibling of at least one of 2,504 unrelated 1000 Genomes Project samples. The sample IDs (shown in blue font) were clustered in the correct population in all possible within-population pairs.



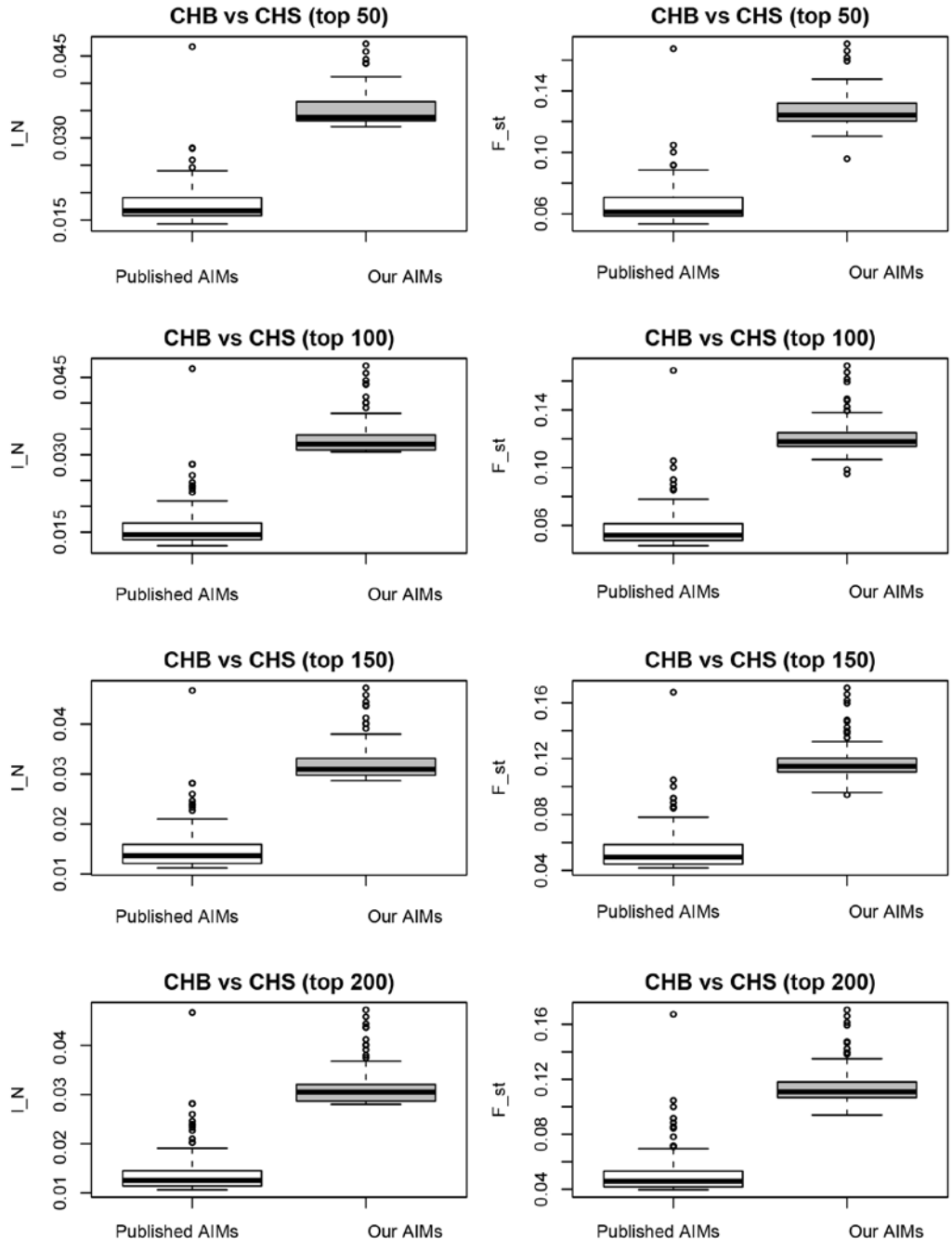
Supp. Figure S4 PCA plots of population pairs with new, testing samples. PCA was used to estimate first two principal components (PC) and their respective contribution to explained variance. For each continental group, PC1 vs PC2 and PC2 vs PC3 were plotted to show finer structures within populations. Our AIMs panels elucidated the global and fine within-population structures.



Supp. Figure S5 Structure of the American populations based on our AIMs panels. The analysis revealed differential levels of admixture in the four populations.



Supp. Figure S6 Correlation of the three AIM identification methods. The three statistical scoring systems used to identify AIMs were highly correlated to each other, particularly F_{ST} and ΔDAF.



Supp. Figure S7 AIMs panels quality differences between our findings and published panel of southern and northern Han Chinese samples. The I_N and F_{ST} values of the top 50, 100, 150 and 200 SNPs (ranked by I_N or F_{ST}, accordingly) were compared between our panel and that by Qin et al.

Chapter 3.2: GACT: A Genome build and Allele definition Conversion Tool for SNP imputation and meta-analysis in genetic association studies

Arvis Sulovari[1,2] and Dawei Li[1,3,4,*]

[1]*Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT 05405, USA*

[2]*Cell, Molecular and Biomedical Sciences Graduate Program, University of Vermont, Burlington, VT 05405, USA*

[3]*Department of Computer Science, University of Vermont, Burlington, VT 05405, USA*

[4]*Neuroscience, Behavior and Health Initiative, University of Vermont, Burlington, VT 05405, USA*

*To whom correspondence should be addressed:

Dawei Li, Ph.D., Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, US. E-mail: dawei.li@uvm.edu

Number of words in the abstract: 286

Number of words in the text (excluding abstract, acknowledgments and financial disclosures sections, legends, and references): 3,707

Number of Tables: 3

Number of Figures: 6

Number of supplementary materials: 1 supplementary Table; 6 supplementary Figures and Legends

Abstract

Background: Genome-wide association studies (GWAS) have successfully identified genes associated with complex human diseases. Although much of the heritability remains unexplained, combining single nucleotide polymorphism (SNP) genotypes from multiple studies for meta-analysis will increase the statistical power to identify new disease-associated variants. Meta-analysis requires same allele definition (nomenclature) and genome build among individual studies. Similarly, imputation, commonly-used prior to meta-analysis, requires the same consistency. However, the genotypes from various GWAS are generated using different genotyping platforms, arrays or SNP-calling approaches, resulting in use of different genome builds and allele definitions. Incorrect assumptions of identical allele definition among combined GWAS lead to a large portion of discarded genotypes or incorrect association findings. There is no published tool that predicts and converts among all major allele definitions.

Results: In this study, we have developed a tool, GACT, which stands for **G**enome build and **A**llele definition **C**onversion **T**ool, that predicts and inter-converts between any of the common SNP allele definitions and between the major genome builds. In addition, we assessed several factors that may affect imputation quality, and our results indicated that inclusion of singletons in the reference had detrimental effects while ambiguous SNPs had no measurable effect. Unexpectedly, exclusion of genotypes with missing rate > 0.001 (40% of study SNPs) showed no significant decrease of imputation quality (even

significantly higher when compared to the imputation with singletons in the reference), especially for rare SNPs.

Conclusion: GACT is a new, powerful, and user-friendly tool with both command-line and interactive online versions that can accurately predict, and convert between any of the common allele definitions and between genome builds for genome-wide meta-analysis and imputation of genotypes from SNP-arrays or deep-sequencing, particularly for data from the dbGaP and other public databases.

GACT software: www.uvm.edu/genomics/software/gact

Keywords: Allele definition (nomenclature); Genome build; Genome-wide association study (GWAS); Imputation; Meta-analysis

Background

Genome-wide association studies (GWASs) and next-generation deep sequencing studies have successfully identified genes associated with human diseases and traits, yet they suggest that the identified variants cumulatively explain a small percentage of the estimated inherited risk to develop these diseases. Combining samples from multiple GWASs or deep sequencing datasets of the same phenotype for large-scale meta-analyses will increase the statistical power to identify new or rare associated variants[203], particularly for complex traits where the disease variants may have moderate effect sizes, which may account for some of the missing heritability[204]. However, the raw single nucleotide polymorphism (SNP) genotype datasets might have been generated using different genotyping or sequencing platforms, array types[205] or SNP calling procedures, resulting in the use of different genome builds or allele definitions (nomenclatures). Thus, combining multiple GWASs or deep sequencing studies (e.g. the 1000 Genomes Project[206]) requires conversions of inconsistent allele definitions and genome builds between the datasets, as demonstrated in a large number of NHGRI (www.genome.gov) GWAS meta-analyses[203]. Likewise, imputation, one of the commonly-used approaches to predict the genotypes for un-assayed loci, requires the same consistency between the study and reference datasets, for example, imputation has been applied to almost half of the GWASs[203] in the NHGRI GWAS Catalog.

Four common nomenclatures exist for reporting biallelic SNPs, including: probe/target or A/B, Plus (+)/Minus (-), TOP/BOT, and Forward/Reverse[207]. The genotype data from

different studies are often not consistent or matched for genome builds or allele definitions, and thus, genotype and build conversions are required if an investigator combines multiple GWASs or imputes a reference dataset (e.g., the 1000 Genome data) into a study GWAS. For example, different genome builds, primarily build 36 (b36) and b37, and various allele definitions were adopted in the 15,541 NHGRI GWAS Catalog datasets. The solutions that disregard mismatched SNPs, i.e., direct allele-flipping or removal of mismatches[208], will lead to undesirable consequences. For example, allele-flip (i.e., from A1 to A2 and vice versa) ignores the allele frequencies of study population and may make the downstream analyses of the flipped SNPs irrelevant to the sample population; and genotype removal may significantly lower the SNP density of relevant regions. Thus, the build of the human genome that was used to call the study SNPs (or true-genotypes) and the allele definition have to be determined and converted where necessary prior to imputation and meta-analysis.

To our knowledge, there is no available tool that simultaneously predicts and converts human genome builds and allele definitions. The existing tools either convert between selected allele definitions alone (such as GenGen (www.openbioinformatics.org/gengen) where the Plus (+)/Minus (-) definition is not included) or between genome builds alone (such as the UCSC Genome Browser LiftOver (genome.ucsc.edu/cgi-bin/hgLiftOver)). In this study, we have developed a new and powerful genotype conversion tool, GACT, which stands for **G**enome build and **A**llele definition **C**onversion **T**ool, to aid in imputation, meta-analysis or both (Figure 1). GACT (Figure 2) directly inter-converts

among any of the four allele definitions and between the b36 and b37 genome builds. Since investigators who use datasets from existing GWAS repositories, such as the dbGaP, may not immediately know what allele definitions were used to call the SNPs, we built an artificial neural network (ANN) within GACT to predict the allele definitions. For next-generation sequencing (NGS) projects, since the sequence reads are aligned and mapped to the human reference genome, which is often in the Plus (+)/Minus (-) definition, the SNP genotypes will be of the same one definition. GACT can convert and match the SNP data from genotyping arrays to NGS data (SNP calls) for data merge and meta-analyses. Our example conversions from A/B definition b36 to Plus/Minus definition b37 consistently yielded high matches with the phased 1000 Genomes genotypes (Table 1), demonstrating the accuracy of our tool for converting the genome builds and allele definitions. GACT can be used as a powerful command line application as well as a user-friendly interactive web tool.

Imputation is often desirable before combining multiple genotype datasets from different recourses for meta-analysis. Our imputation analysis revealed higher quality for imputed SNPs when GACT was used, compared to when mismatched SNPs were excluded (Table S1). While GACT aims to convert between allele definitions and maximize the number of correctly matched alleles to a reference, there are many other factors that can affect imputation quality. Hence, we measured the effects of selected variant types (such as singletons (i.e. SNPs with only one copy of the minor allele among all samples), monomorphic SNPs, and ambiguous SNPs) and GWAS quality control procedures (such

as genotype missing rate) on imputation quality. We found that the exclusion of singletons and monomorphic SNPs from the reference improved imputation quality of rare SNPs with minor allele frequency (MAF) < 0.005 (the mean quality score increased from 0.52 to 0.57, which was the highest increase across all MAF ranges) but had no effect on SNPs with MAF > 0.005 (the mean score remained 0.91). The ambiguous SNPs had no measurable effect on imputation, while imputation quality decreased as the genotype missing thresholds became more conservative. Surprisingly, for imputed common SNPs (MAF > 0.1), the decrease in imputation quality started to emerge under very stringent genotype missing thresholds (0.004-0.001, instead of the commonly-used 0.05); by comparison, the imputation of relatively rare SNPs (MAF < 0.1) was even more robust, the decrease was not significant until the missing threshold reached a more stringent threshold of 0.0005 (corresponding to removal of 61.4% of the genotypes). Moreover, the physical locations of the SNPs that were excluded under these missing thresholds were distributed uniformly across the chromosomes. Our analyses provide novel insight into imputation insensitivity to genotype missingness, particularly for rare SNPs.

Implementation

Subjects and genotype data

A cohort of 3,096 subjects of Ashkenazi Jewish ethnicity were genotyped using the Illumina Human Omni 1 Quad arrays. The GWAS genotype data were obtained through the NIH dbGaP [phs000448].

GACT pipeline

GACT was designed for matching allele definitions between the study GWAS and reference data before imputation or merging multiple genome-wide genotype datasets before meta-analysis, where the genotypes were generated from SNP-arrays or deep-sequencing platforms (Figure 1). Figure 2 shows the study design and GACT pipeline, which can be directly connected to other commonly-used methods, including genotype phasing of GWAS (or deep sequencing) data, imputation, data merging, and meta-analysis (Figure 1). The proper execution in command line of GACT requires PLINK[209], GenGen, and the genotyping array annotation files in the same directory, which can be downloaded from our website. The command line follows this syntax (example): *./gact b36 b37 ab plus o1qd map_file_name*. The arguments represent the current genome build (b36), desired genome build (b37), current allele definition (ab), desired allele definition (plus), annotation file of SNP genotyping array (o1qd = Human Omni 1 Quad Duo), and input map file name, respectively. The input file should be in the same format as the PLINK binary map file, containing chromosome location and reference alleles of each SNP. The web version accesses the same command line options on the server-end after user uploads the input file, a PLINK format map file, and chooses the preferred options

on the web interface. Moreover, the web tool allows the user to view in real time a log of every step in the conversion process. The command line has no pre-defined limit on the input file size while the web tool has a limit of 40 megabytes (MB), which is sufficient for most SNP arrays (e.g, the entire map file of the Illumina Human Omni 1 Quad array is < 30 MB).

To build the allele definition prediction model, the 1000 Genomes data (2,046,145 SNPs on chromosome 1), dbSNP data (51,864 SNPs on chromosome 1), and our GWAS data (964,554 SNPs on chromosome 1) were used to extract the allele properties of the Plus (+)/Minus (-), Forward/Reverse, and TOP/BOT definitions, respectively (our findings were consistent across all chromosomes). The three genotypes (CT, TC, and GA, Figure 3) that showed the largest amount of differential enrichment among the allele definitions were used as the inputs for a feed-forward, back propagation, ANN with 3 input neurons, 2 hidden layers, and 1 output neuron. This ANN was trained using 10 random samples of various sizes (from 1,000 to 2,000,000 SNPs) from each of the three genotype sources. The ai4r ruby gem (ai4r.org) was used to implement the ANN. Similarly, the coordinates of selected common SNPs in both b36 and b37 datasets were used as the references to predict genome builds. We assessed the quality of implementing our tool to the GWAS data by counting the number of allele matches between the study data and 1000 Genomes Project data using SHAPEIT[210]. GACT was written using a set of Python, Ruby, Hypertext Preprocessor (PHP), and bash scripts. More details and frequently asked questions are available on our website.

Imputation quality assessment

The GWAS genotype data of the 3,096 Ashkenazi Jewish samples was in b36 genome build and A/B allele definition. GACT was used to convert the allele definition and genome build to the b37 and PLUS allele to keep them consistent with the 1000 Genomes panel. The genotype match rates between the study and reference datasets and imputation quality scores were used as primary measurements to assess conversion quality of GACT. After converting the genome builds and allele definitions in the map files using GACT, we recoded all the genotypes of the GWAS data using PLINK. The genotype phasing and imputation were carried out using SHAPEIT and Impute2[211], respectively. The latest phased 1000 Genomes genotypes of the European population (Phase 1 integrated release version 3) were used as the imputation reference. Imputation quality was assessed using the Impute2 information scores of the reference SNPs. The scores (equivalent to the r-squared metric reported by MaCH[212] and BEAGLE[213]) vary between 0 and 1, where values closer to 1 represent imputation with high certainty. The mean and standard deviation of these scores were used as measures of overall imputation quality of SNPs at specific MAF ranges. To compare the imputation quality between different MAFs, we used the Welch two sample t-test. All the statistical analyses and graphs were generated using the latest version of R (version 3.0.2), and the imputations were conducted using the multi-core cluster at the Vermont Advanced Computing Center.

Results

GACT prediction of genome build and allele definition

We measured the frequencies of all 16 possible genotype patterns under three allele definitions, including Plus (+)/Minus (-), Forward/Reverse, and TOP/BOT (the A/B or probe/target definition is differently coded). The distributions (Figure 3) were clearly distinguishable, and thus used to predict all the four designations. We observed the enrichment of two patterns A/G and G/A, two patterns A/G and C/T, and four patterns A/G, G/A, C/T and T/C for TOP/BOT, Forward/Reverse, and Plus/Minus, respectively. The prediction model matches relative ratios of the input genotypes to the expected ratios in each definition by measuring the proportions of CT, TC and GA alleles present. These three values acted as the input neurons into a multilayer perceptron that classified the input map file into one of the four SNP definitions (Figure S1). Thus, for users who have no knowledge about the allele definitions and (or) genome build, GACT will first notify the user of the predicted definition and build of the input SNPs prior to actual conversion. The prediction module is particularly useful when the datasets are obtained from public genotype repositories, such as the dbGaP.

GACT conversion of genome build and allele definition

GACT has been demonstrated to identify and clean all the convertible allele mismatches. Table 1 shows the amounts of genotypes that should be discarded if we incorrectly

assumed versus correctly converted the allele definitions between our GWAS data and the 1000 Genome data (Plus/Minus) during imputation. For instance, if we incorrectly converted our GWAS genotypes to the "Forward/Reverse" or "TOP/BOT" definition, and imputed with the 1000 Genome data, we had to discard 21.7% and 51.5% of the genotypes, respectively, due to mismatch. By comparison, if we correctly converted our genotypes to "Plus/Minus" by using GACT, only 7% needed to be discarded across all the chromosomes (Table 1). Moreover, since 3,344 SNPs existed in our data but not in the reference, when only the SNPs that existed in both datasets were used in the calculation, the discarded genotypes only accounted for 3.3%, which was significantly lower than commonly-observed mismatch rates in the literature. The reasons for the 3.3% mismatches are described in the discussion.

As expected, the imputation quality decreased when the mismatch rate increased (Table S1), which was primarily due to the decrease of SNP density in the study data. Figure 4 clearly shows evidence of a significant increase in the SNP density ($P = 3.2 \times 10^{-144}$ based on 2-sided paired t-test) of the study data across the entire chromosome. Likewise, the imputation quality (information scores) consistently increased by 1% across all MAFs after we converted the genome build and allele definition of our GWAS data from the Forward/Reverse definition (to the Plus/Minus definition) using GACT (Table S1). However, it should be noted that the improvement would be much higher if we converted the TOP/BOT definition (to the Plus/Minus definition) since without conversions (Table

1) the mismatch rate between the TOP/BOT and Plus/Minus definitions was larger than that between the Forward/Reverse and Plus/Minus definitions.

Imputation quality

We measured the effects of multiple SNP types and GWAS quality control procedures on imputation quality (i.e., using the information scores). The results (Table 2) showed that the imputation quality increased from 0.52 to 0.57 for the variants with $0.001 < MAF < 0.005$ when both the monomorphic variants and singletons were removed from the reference panel, however, no significant change was observed for more common variants with $MAF > 0.005$. When both of the ambiguous and singleton SNPs were removed from the study data (prior to phasing and imputation), the imputation quality showed no significant changes, which was consistent with previous studies[214].

Our results further showed that there was no noticeable effect on the imputation quality when the SNPs with genotype missing rate > 0.01 (667 SNPs) or 0.03 (939 SNPs) were excluded, regardless of the decrease of SNP density, when compared to the commonly-used genotype missing rate threshold of 0.05. This might be partially due to the fact that the assayed SNPs were of high quality, indicated by low genotype missing rates. For instance, the mean genotype missing rate was < 0.005 across all the SNPs with $0.001 < MAF < 0.5$ on chromosome 1 (Figures S2 and S3). We repeated the imputation procedures under new missing rate thresholds and measured their effects on imputation

quality (Figure 5). The new thresholds included 0.004, 0.002, 0.001, and 0.0005, corresponding to the removals of 10,279 (13.8%), 17,785 (23.8%), 29,307 (39.3%), and 45,856 (61.4%) SNPs, respectively. Table 2 and Figure 5 show the comparisons of imputation quality measurements at the four missing thresholds across six different MAF ranges. As the missing threshold became more conservative (i.e. < 0.05), we observed a decrease in imputation quality where the higher MAFs exhibited more sensitivity to less stringent thresholds. For instance, the decrease emerged for the most common SNP group ($0.1 < MAF < 0.5$) at the missing threshold of 0.004, for the SNP group with $0.05 < MAF < 0.5$ at the threshold of 0.002, and for the group containing rare SNPs ($0.001 < MAF < 0.5$) at the threshold of 0.0005. Surprisingly, we found that imputation of the rarest SNPs into genotyped genome regions tolerated very low SNP density (up to 39.3% lower when the missing threshold was 0.001) as long as the genotypes were of high quality (i.e. low missing rate). Moreover, exclusion of the SNPs with missing rate > 0.001 did not worsen imputation compared to the scenario where singletons were included in the reference (missing threshold = 0.05), particularly for SNPs with $0.001 < MAF < 0.005$ (Figure S4). Importantly, the locations of excluded SNPs (under the most conservative threshold) were distributed uniformly across the chromosome (Figure 6), indicating that the changes in imputation quality are very likely due to global, rather than local, changes in the SNP density of the genotype scaffold.

Discussion

Both genome builds and allele definitions should be well-matched before combing or imputing one genotype data with another. In this study, we have developed a new, powerful, and user-friendly tool that can predict, and convert the genome builds and allele definitions simultaneously between multiple GWAS or deep sequencing genotype datasets for meta-analyses, imputations or both. Our GWAS data demonstrated the accuracy of predictions and performance of conversions. Our further imputations showed that the inclusion of singletons in the reference panel significantly decreased imputation quality. However, the exclusion of SNPs with missing rate > 0.001 led to comparably high imputation quality with the commonly-used threshold of 0.05 for rare SNPs (Table 2 and Figures 5 and S5), which implied that approximately 600,000 well-typed SNPs were likely to be sufficient for high quality genome-wide imputation of rare SNPs in our GWAS data.

GACT pipeline

GACT achieved as low as 3.3% discarded genotypes (Table 1), which was significantly lower than commonly-observed mismatch rates. It should be noted that we always observe genotype mismatches in real datasets, particularly when one dataset is from microarray-based study and the other is from deep-sequencing-based study, like the case in Table 1. This is likely to be attributed to various factors, such as different experimental protocols, genotyping error rates, and disease statuses of research subjects. Interestingly, the genotype mismatch rates between different platforms are not significantly higher than

those within same platforms. For instance, a recent study[215] showed 0.6-1.6% genotype mismatch rate within two deep-sequencing studies(Li et al's data and the 1000 Genomes); by comparison, the 3.3% mismatch rate between two different platforms/samples is reasonably low. All these results demonstrated that it is required to correctly convert allele definitions prior to imputation or meta-analysis.

Table 3 shows the comparisons GACT with some of the existing tools that also include genome build and (or) allele definition conversion functions, including GWAMA[216], GenGen, METAL[217], and PLINK. The strengths of our tool include that it 1) can be easily connected to other commonly-used GWAS approaches (Figure 1); 2) can convert between any of the four commonly-used SNP allele definitions; 3) provides both the powerful command-line software and user-friendly web interface, where the latter can be easily used by biologists (no informatics training required except access to the internet); 4) can accurately predict allele definitions (and genome builds), which is particularly useful for investigators who use GWAS data from the dbGaP or other publicly available database; and 5) is computationally efficient, e.g., a typical conversion can be completed in a few seconds. In addition, the microarray-specific SNP definition information is used in GACT to flip the alleles and strands. Because it can convert data prior to association testing, meta-analysis and imputation, GACT complements existing tools and ensures allele definition and genome build consistency before using any of these tools. The limitation of our tool is that currently, the supported microarrays (primarily Illumina platforms) and genome-builds of the web version of GACT are not exhaustive (the

command-line version has no such limitation; users can convert between any platforms and arrays using the command-line version of GACT). However, we will actively include conversions of other existing allele definitions, e.g., numerical alleles. We will provide continued scientific and technical support, and expand the list of arrays, genome builds, and new modules as new technologies and platforms become available.

Imputation after GACT Conversion

Imputation before combining GWAS datasets is desirable because of 1) increased power for identifying disease-associated variants, e.g. by more than 10% as suggested previously[218]; 2) higher SNP coverage for fine-mapping disease genes; 3) additional rare SNPs and applicability to other variants such as copy number variations or classical leukocyte antigen alleles[208]; and 4) cost- and time-efficiency compared with the molecular genotyping or sequencing experiments. Various studies have been carried out to evaluate or identify the factors that might affect imputation quality[214,219], including ambiguous, monomorphic, and singleton SNPs. Phasing of singletons is known to be challenging, and imputation becomes faster with no burden in the downstream association tests when singletons are removed from the reference. We found that, additionally, the removal of either ambiguous or monomorphic SNPs alone from the study data prior to phasing and imputation had no detectable effect on imputation. However, the exclusion of monomorphic and singleton SNPs from the reference increased imputation quality, which is in accordance with previous studies[214,219]. We

further found that SNPs with very low MAF (0.001-0.005) showed the most significant increase of the imputation quality compared with the other MAF ranges (Table 2). This finding is important, particularly, for the rare variants, which are of increasing interest in the genetic studies of complex diseases and traits.

Balancing between genotype quality and genome coverage is important for imputation. The genotype missing thresholds of 0.05 to 0.02[219] are generally recommended for quality controls in GWAS. However, no published studies have explicitly evaluated the effects of more conservative missing thresholds (than the commonly-used values) on imputation quality. Our assessments might provide a new perspective on the selection of genotype missing thresholds in imputation. Based on our GWAS data, an approximate number of 600 thousand well-typed SNPs are likely to be sufficient for high quality genome-wide imputation of rare SNPs (high quality assayed SNPs may compensate for low true-genotype density). However, further analyses are warranted to replicate the findings in additional arrays. It should be noted that only the data on chromosome 1 were used for most of the analyses based on our observation of similar genotype missing patterns or comparable results across all the chromosomes (Figures S5 and S6).

Conclusion

Ignorance of inconsistent allele definitions and genome builds or incorrect conversions lead to incorrect genetic association “findings”. In this study, we developed a

comprehensive tool, GACT, with both powerful command-line and user-friendly web interface versions to predict, and convert both genome builds and allele definitions between multiple GWAS (or deep sequencing) genotype data, which is required for all imputations and genome-wide meta-analyses. GACT will facilitate and ease a broad use of the GWAS data from the dbGaP and other publicly available genotype repositories for large-scale secondary analyses and multi-laboratory collaborations in the genetic association studies of human diseases.

Availability and requirements

Project name: GACT: **G**enome build and **A**llele definition **C**onversion **T**ool

Project homepage: http://www.uvm.edu/genomics/software/gact

Operating system(s): Linux, UNIX (for command version) and Windows (for interactive web version)

Programming language: Python, Ruby, Hypertext Preprocessor (PHP), and Bash scripts

License: GPL-3

Availability: GACT (both command-line and web versions), including source code, documentation, and examples, is freely available for non-commercial use with no restrictions at http://www.uvm.edu/genomics/software/gact and http://asulovar.w3.uvm.edu/gact.

Competing interests

The authors declare that they have no competing interests

Author contributions

DL conceived, organized and supervised the project. AS wrote the source code and conducted the analyses. AS and DL drafted the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

This work was supported by the start-up fund from the University of Vermont, USA. The GWAS data in Ashkenazi Jewish that was described in this study were obtained from the database of Genotypes and Phenotypes (dbGaP) through accession number phs000448. Funding support for the GWAS was provided through the NIH RC2MH089964. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Robert Howe, a research student in our laboratory, for providing technical assistance in completing the web application of GACT. The authors acknowledge the Vermont Advanced Computing Core which is supported by NASA (NNX 06AC88G), at the University of Vermont for providing high performance computing resources that have contributed to the research results reported within this paper. We also thank reviewers for their helpful suggestions and comments.

References

1. Panagiotou OA, Willer CJ, Hirschhorn JN, Ioannidis JP: **The power of meta-analysis in genome-wide association studies**. *Annu Rev Genomics Hum Genet* 2013, **14**:441-465.
2. Panagiotou OA, Evangelou E, Ioannidis JP: **Genome-wide significant associations for variants with minor allele frequency of 5% or less--an overview: A HuGE review**. *Am J Epidemiol* 2010, **172**(8):869-889.
3. Nicolazzi EL, Picciolini M, Strozzi F, Schnabel RD, Lawley C, Pirani A, Brew F, Stella A: **SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock**. *BMC genomics* 2014, **15**:123.
4. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.
5. Nelson SC, Doheny KF, Laurie CC, Mirel DB: **Is 'forward' the same as 'plus'?...and other adventures in SNP allele nomenclature**. *Trends in genetics : TIG* 2012, **28**(8):361-363.
6. Marchini J, Howie B: **Genotype imputation for genome-wide association studies**. *Nature reviews Genetics* 2010, **11**(7):499-511.
7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *American journal of human genetics* 2007, **81**(3):559-575.
8. Delaneau O, Zagury JF, Marchini J: **Improved whole-chromosome phasing for disease and population genetic studies**. *Nature methods* 2013, **10**(1):5-6.
9. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies**. *PLoS genetics* 2009, **5**(6):e1000529.

10. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: **MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes**. *Genetic epidemiology* 2010, **34**(8):816-834.
11. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering**. *American journal of human genetics* 2007, **81**(5):1084-1097.
12. Liu EY, Buyske S, Aragaki AK, Peters U, Boerwinkle E, Carlson C, Carty C, Crawford DC, Haessler J, Hindorff LA *et al*: **Genotype imputation of Metabochip SNPs using a study-specific reference panel of ~4,000 haplotypes in African Americans from the Women's Health Initiative**. *Genetic epidemiology* 2012, **36**(2):107-117.
13. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR: **Low-coverage sequencing: implications for design of complex trait association studies**. *Genome research* 2011, **21**(6):940-951.
14. Magi R, Morris AP: **GWAMA: software for genome-wide association meta-analysis**. *BMC bioinformatics* 2010, **11**:288.
15. Willer CJ, Li Y, Abecasis GR: **METAL: fast and efficient meta-analysis of genomewide association scans**. *Bioinformatics* 2010, **26**(17):2190-2191.
16. Spencer CC, Su Z, Donnelly P, Marchini J: **Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip**. *PLoS genetics* 2009, **5**(5):e1000477.
17. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G *et al*: **Quality control procedures for genome-wide association studies**. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]* 2011, **Chapter 1**:Unit1 19.

Tables

Table 1 Genotype mismatches between the GWAS and 1000 Genomes datasets.

Study GWAS	1000 Genomes	Types	Incorrect conversions		Correct conversion
			Fwd-Plus	Top-Plus	Plus-Plus
T/C	C/T	FLIP	0	0	0
T/C	A/G	CSF	5,048	9,875	301
T/C	G/A	FLIP & CSF	8,556	27,648	1,840
T/A	*/*	AMBIG	432	432	432
/	-/-	NAR	3,344	3,344	3,344
Matches (%)			62,793 (78.3) (81.7)†	38,875 (48.5)	74,256 (92.6) (96.7)†

FLIP: switch both alleles with one another (from A1 to A2 and vice versa);
CSF: complimentary strand flip;
AMBIG: ambiguous SNPs in study GWAS;
NAR: not available in the reference;
/: any genotype;
-/-: missing genotype;
Fwd: Forward/Reverse;
Top: TOP/BOT;
Plus: Plus (+)/Minus (-);
†, percentages of matched genotypes after excluding the NAR genotype counts.
Both the "GWAS" (the 3,096 Ashkenazi Jewish samples) and "1000 Genome" columns show the example alleles in the A1/A2 order. The "Type" column indicates the changes required to match the study SNP to the reference. The last three columns refer to numbers of genotype mismatches on chromosome 1 (80,173 SNPs in total). The "Fwd-Plus" and "Top-Plus" columns show the numbers of genotype mismatches between the "Fwd" and "Top" definitions of our GWAS data (we first generated two versions of the same GWAS data: "Fwd" and "Top") and the "Plus" definition of the 1000 Genome data, respectively, while the "Plus" column refers to the numbers after we converted the GWAS data to "Plus" using GACT. The last row shows the numbers (percentages) of correct genotype matches (e.g., "T/C" and "T/C") between the GWAS and 1000 Genome data, where the (%) and (%)† represent the percentages measured by including and excluding the SNPs (NAR) unique to our GWAS data, respectively. Similar ratios were observed in other chromosomes.

Table 2 Quality scores of the imputed (I) and study (S) SNPs for each MAF category (see https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-610).

Table 3 Comparisons of tools for genome build and allele definition conversions.

Complementary Functionality	**GenGen**	**GWAMA**	**METAL**	**PLINK**	**GACT**
Allele definition prediction	No	No	No	No	Yes
Uninformed strand/allele flip[1]	No	Yes	Yes	Yes	No
Informed allele conversion[2]	Yes[3]	No	No	No	Yes
Automatic allele conversion	Yes[3]	No	No	No[4]	Yes
Genome build prediction	No	No	No	No	Yes
Genome build conversion	No	No	No	Yes[4]	Yes
Command line	Yes	Yes	Yes	Yes	Yes
Interactive web interface	No	No	No	No	Yes

[1]"Uninformed" refers to flipping without SNP allele annotation knowledge.
[2]"Informed" refers to use of the original SNP definition and microarray-specific annotation information.
[3]GenGen converts between Top, Forward, A/B and 1/2 allele definitions; by comparison, GACT converts between Top, Forward, A/B and Plus definitions while the Plus definition is used by the 1000 Genomes Project and most next-generation sequencing studies.
[4]PLINK can strand- or allele-flip but it cannot directly convert from one allele definition to another, unless the user manually extracts information from the microarray annotation file; by comparison, GACT automatically converts between genome builds and allele definitions.

Legends

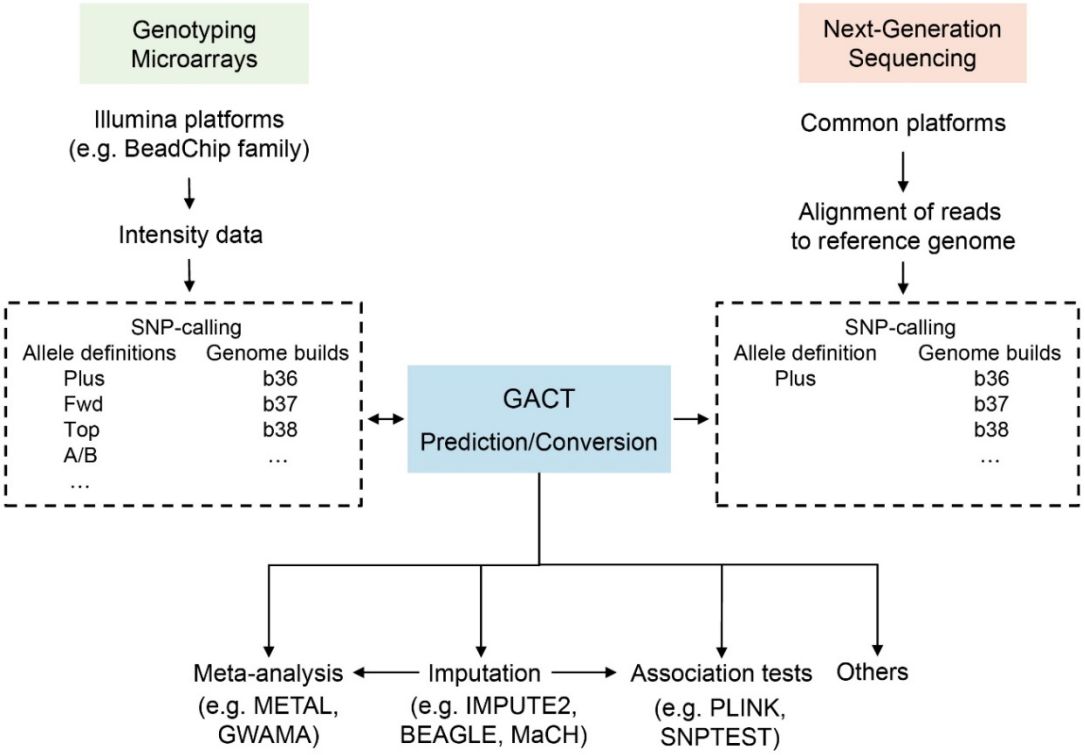


Figure 1 Study design and GACT functionality. The left side of the figure indicates that microarray data can be used to call SNPs in any of the four listed SNP definitions. Often, when genotypes are obtained from public repositories (e.g. dbGaP), allele definitions may not be immediately known to investigators. GACT will predict allele definition and genome build, and convert to any new definitions or builds. Since the SNP definition in the NGS data is determined during alignment to the human reference genome (Plus is a commonly-used definition), the SNP alleles from genotyping microarrays can be converted and matched to those from NGS. After GACT's conversion, imputation, meta-analysis and (or) other analyses may be carried out using the commonly-used tools such as GWAMA, METAL, PLINK, and IMPUTE2.

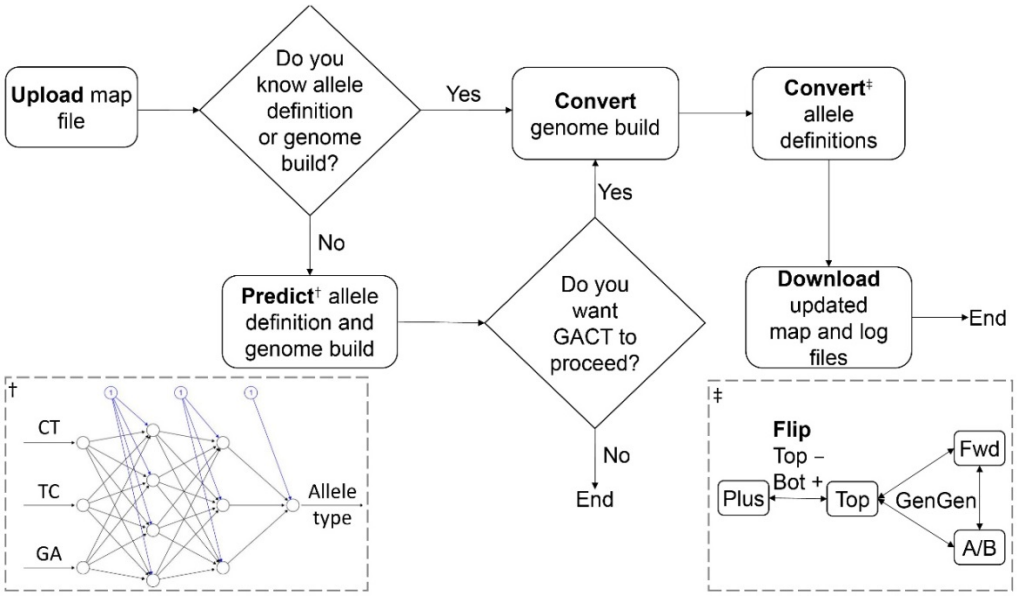


Figure 2 GACT pipeline. The flow diagram shows the major procedures in the GACT design. The bottom left panel shows the prediction model of allele definitions based on the distribution of each definition (Figure 2). The bottom right panel shows the allele conversion pathway among the four allele definitions. The input file to be uploaded is a PLINK format map file. This pipeline is implemented in both command-line and web interface.

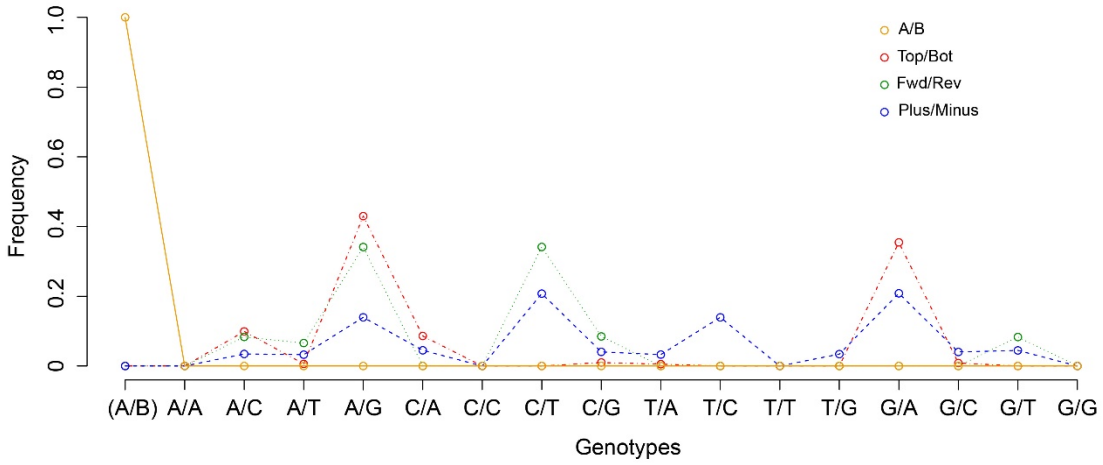


Figure 3 Frequencies and distributions of all possible genotypes of biallelic SNPs. The data were generated for the Plus/Minus, Forward/ Reverse, A/B, and TOP/BOT definitions based on the 1000 Genomes, dbSNP, and our GWAS datasets for the last two, respectively. The prediction model of allele definitions was trained using these distributions.

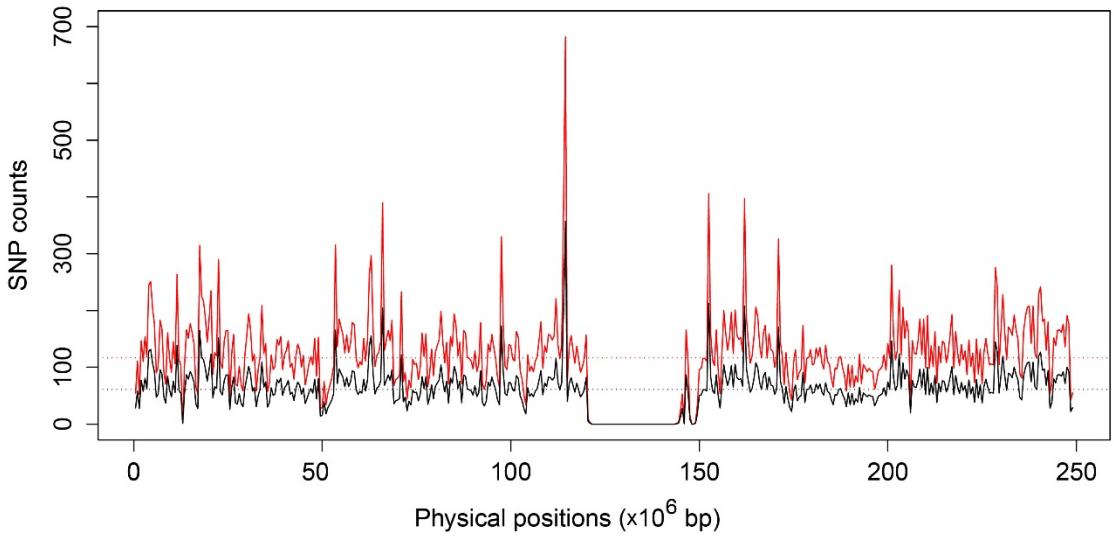


Figure 4 Comparison of SNP density plots before ("Top" allele definition; black line) and after ("Plus" allele definition; red line) GACT conversion. The SNP density

was measured per 500,000 bp window. It is clear that the SNP count (or density) increase after GACT converts all the mismatched loci, e.g., from 61.05 (median) to 117 SNPs per window. Moreover, it is evident that the increase is not biased with regard to physical location, which indicates that the allele definition mismatches are uniformly distributed across the chromosome. The dotted horizontal lines represent the median of values of each line matched by color. The median, instead of mean, was used since the former was less vulnerable to outliers (e.g. zero counts in the centromere region). The "Forward/Reverse" allele definition showed a similar distribution of mismatches with the 1000 Genomes, however, only the "TOP" definition is shown due to its higher level of mismatches (51.5% mismatches in "TOP" versus 21.7% mismatch in "Forward"). Other chromosomes showed similar patterns, and thus only the results of chromosome 1 are shown.

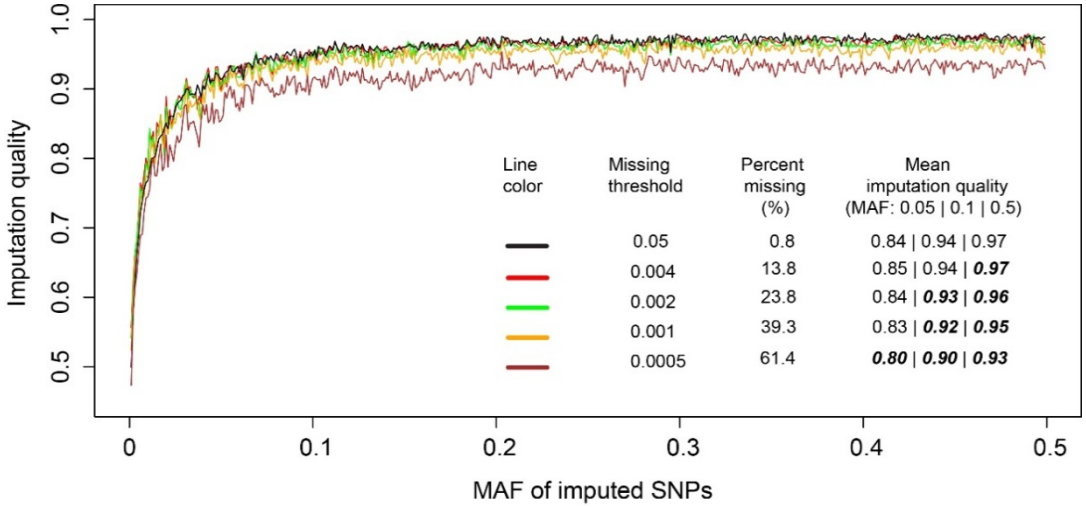


Figure 5 Comparison of imputation quality of imputed SNPs. The quality score columns list three SNP minor allele frequency (MAF) categories: very rare (0.001 < MAF < 0.05), rare (0.05< MAF <0.1), and common (0.1 < MAF < 0.5). The results under the missing thresholds of 0.03 and 0.01 showed the similar patterns to those under the threshold of 0.05, and thus are not shown. Bold indicates $P < 0.05$ in the Welch two sample t-test between the missing rate of 0.05 (black line) and the other thresholds.

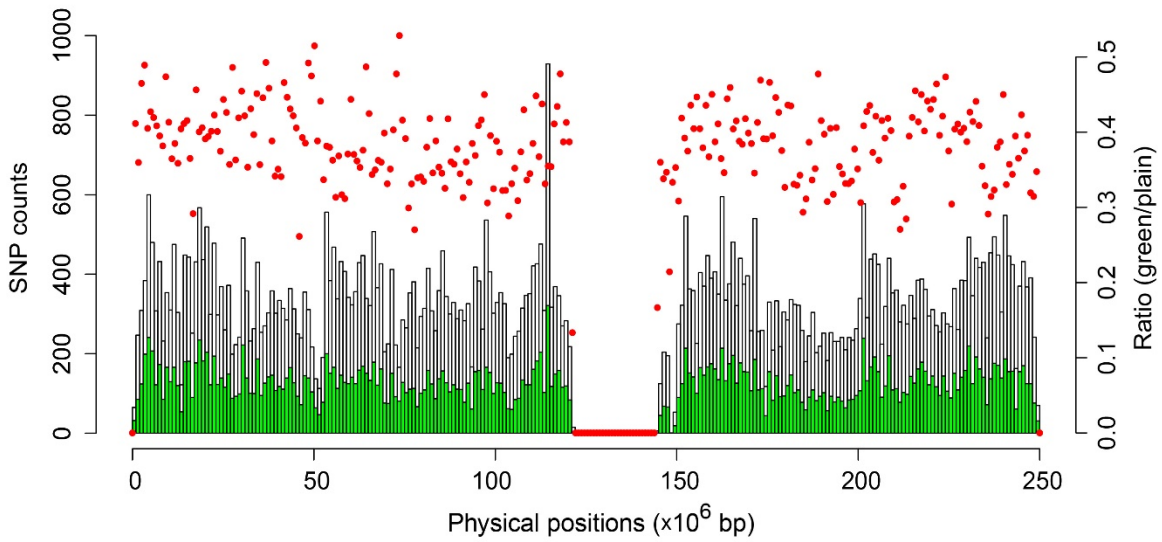


Figure 6 Distribution of SNP missing genotypes. The green histograms represent the numbers of remaining SNPs after removing the SNPs with missing rate > 0.05% while the plain histograms represent the total numbers of SNPs (on chromosome 1). The red circles represent the fractions of SNPs that passed the threshold. It is clear that the range of the fractions is narrow (i.e. 0.3-0.5).

Supplements

Supplementary Table

Table S1 Comparison of imputation quality before and after genotype conversion using GACT

MAFs	Before	After
0.001-0.005	0.56 (.30)	0.57 (.30)
0.005-0.01	0.72 (.22)	0.73 (.22)
0.01-0.05	0.84 (.18)	0.85 (.17)
0.05-0.1	0.93 (.12)	0.94 (.12)
0.1-0.3	0.96 (.09)	0.97 (.09)
0.3-0.5	0.97 (.08)	0.98 (.07)

Imputation is the process of using a reference haplotype panel at a dense set of SNPs (i.e., the 1000 Genomes Project) to impute into a sample of individuals genotyped for a subset of these SNPs (i.e., the GWAS data). The numbers in this table represent the mean imputation quality scores after the basic quality control of removing SNPs with missing genotype rate > 0.05. The standard deviations are shown in brackets. Imputing into less dense SNP regions (i.e. before GACT conversion) revealed lower imputation scores than denser SNP regions (i.e. after GACT conversion). This table shows the increase (improvement) of imputation quality based on our GWAS data ("Forward/Reverse") and the 1000 Genomes data ("Plus/Minus"). However, it should be noted that the improvement would be much *higher* if data with the "TOP/BOT" definition were used since the mismatch rate between the "TOP/BOT" and "Plus/Minus" definitions was larger (Table 1). Other chromosome showed similar patterns, and thus only the results of chromosome 1 are shown.

Supplementary Figures

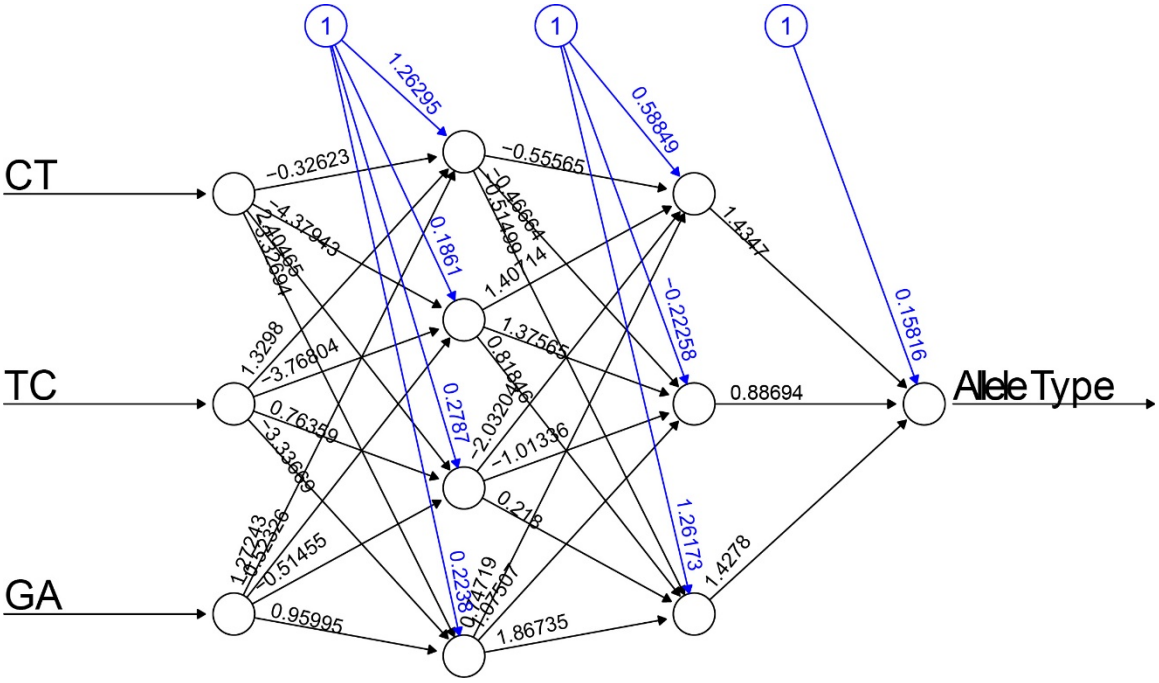


Figure S1 The feed-forward backpropagation neural network. The 3 input neurons correspond to the proportion of CT, TC and GA. The number in black next to each edge represents the weight of that edge. The numbers in blue represent the activation threshold for each hidden node, as defined by the activation function of the neural network, after training. There were three such networks in GACT, where each was trained to make an independent prediction on the likelihood that the input map file was using one of the three allele definitions: Plus (using the 1000 Genomes), Forward (using dbSNP) and Top (using our GWAS data). The artificial neural network that generated the largest likelihood determined the final allele definition. The A/B definition, which can be distinguished directly, was not included in the network.

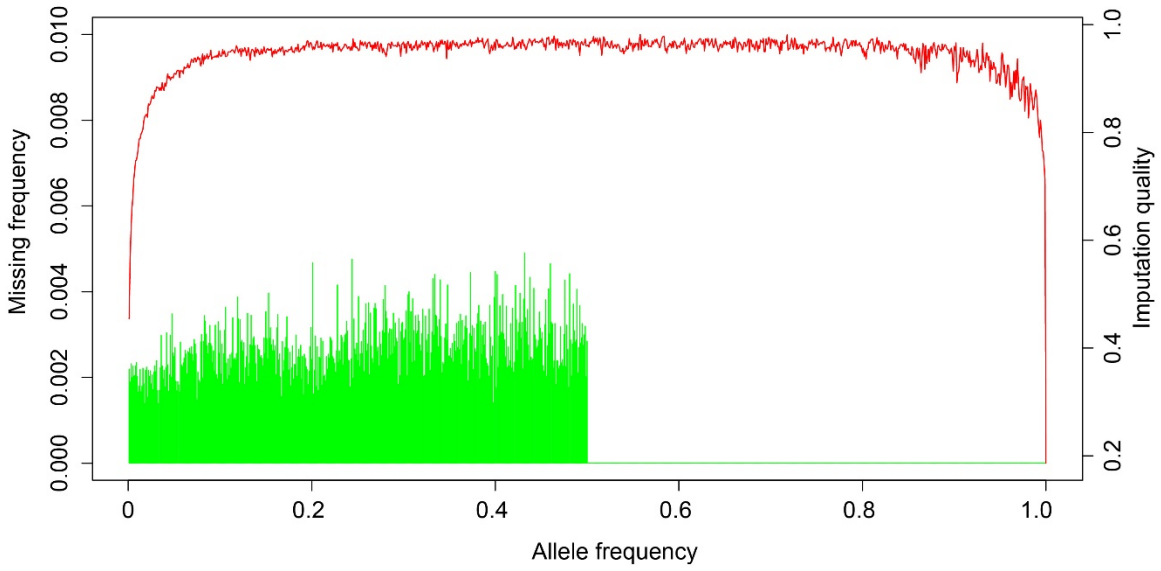
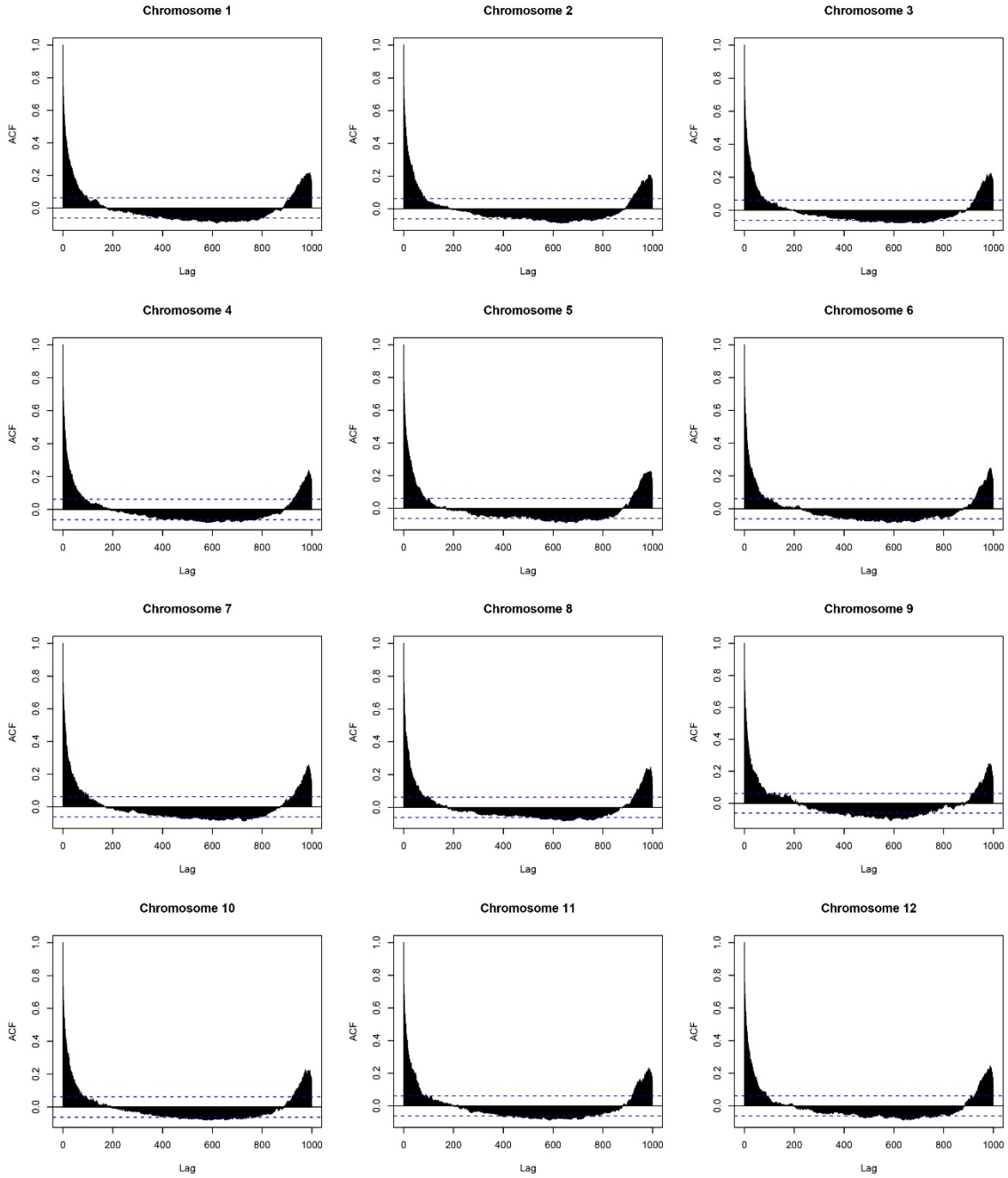


Figure S2 Imputation quality and genotype missing rate across allele frequencies. The missing frequency measurement is the average of missing genotype rates for all the SNPs at a given MAF. The numbers of the SNPs that were excluded were 45,856, 29,307, 17,785, 10,279, 4,667, and 939 (out of 74,638) when the genotype missing rate thresholds were set at 0.0005, 0.001, 0.002, 0.004, 0.01, and 0.03, respectively. The red curve shows the information (quality) scores of the imputed genotypes across the full allele frequency range (0-1). The green histogram shows the genotype missing rate distribution across the full range of MAFs (0-0.5) under the missing genotype threshold of 0.05. The MAF scale (0-0.5) was adopted, instead of a full scale (0-1), based on our autocorrelation analyses of the imputation quality curves which showed that the head-10% and tail-10% were significantly correlated (Figure S2). Other chromosome showed the similar patterns, and thus only the results of chromosome 1 are shown.


Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
ACF
Lag
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000

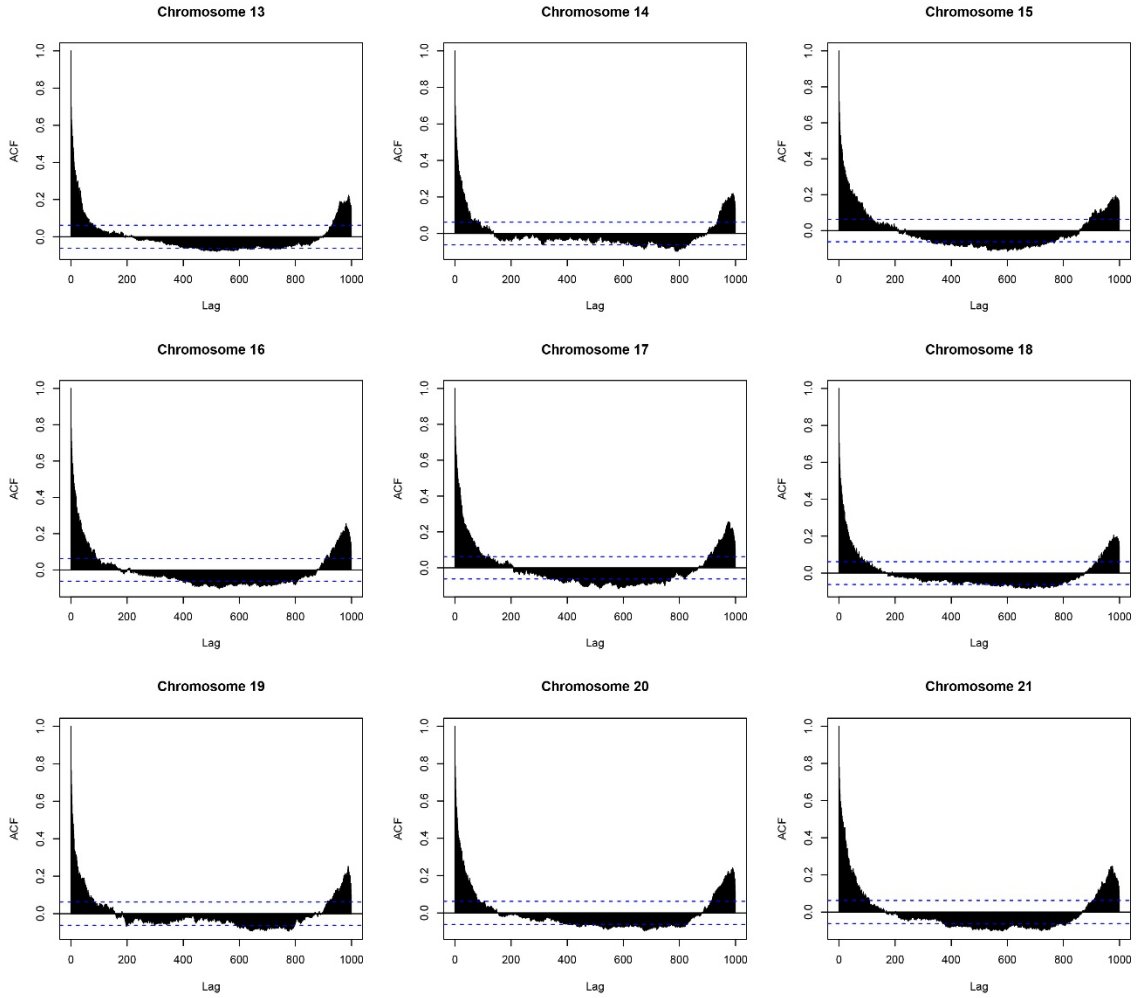


Figure S3 Autocorrelation plots of mean imputation scores. This figure corresponds to the full range of allele frequencies that is shown in Figure S1 (red line). The Lag axis represents the shift of the data points, one number at a time at a rate of 0.001, while the ACF axis represents an adjusted correlation factor between the "shifted" data and the original data. The histograms outside of the dotted blue lines represent the regions with higher correlation than expected by chance alone (at confidence level > 95%). Moreover, this autocorrelation plot indicated that the regions of allele frequency < 0.1 and > 0.9 were significantly correlated at the confidence level of > 0.95. Based on this result we combined both the upper and lower halves to generate MAFs (0-0.5), instead of the full range of allele frequencies (0-1).

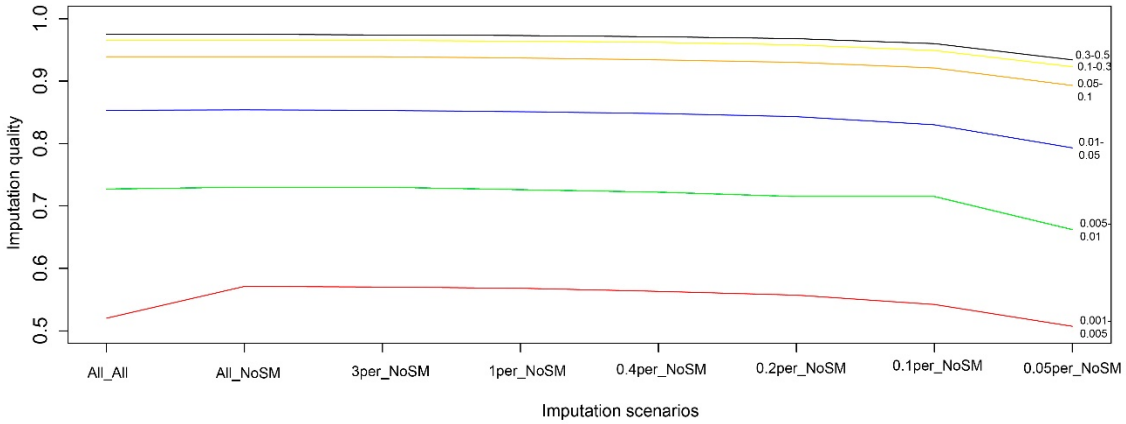
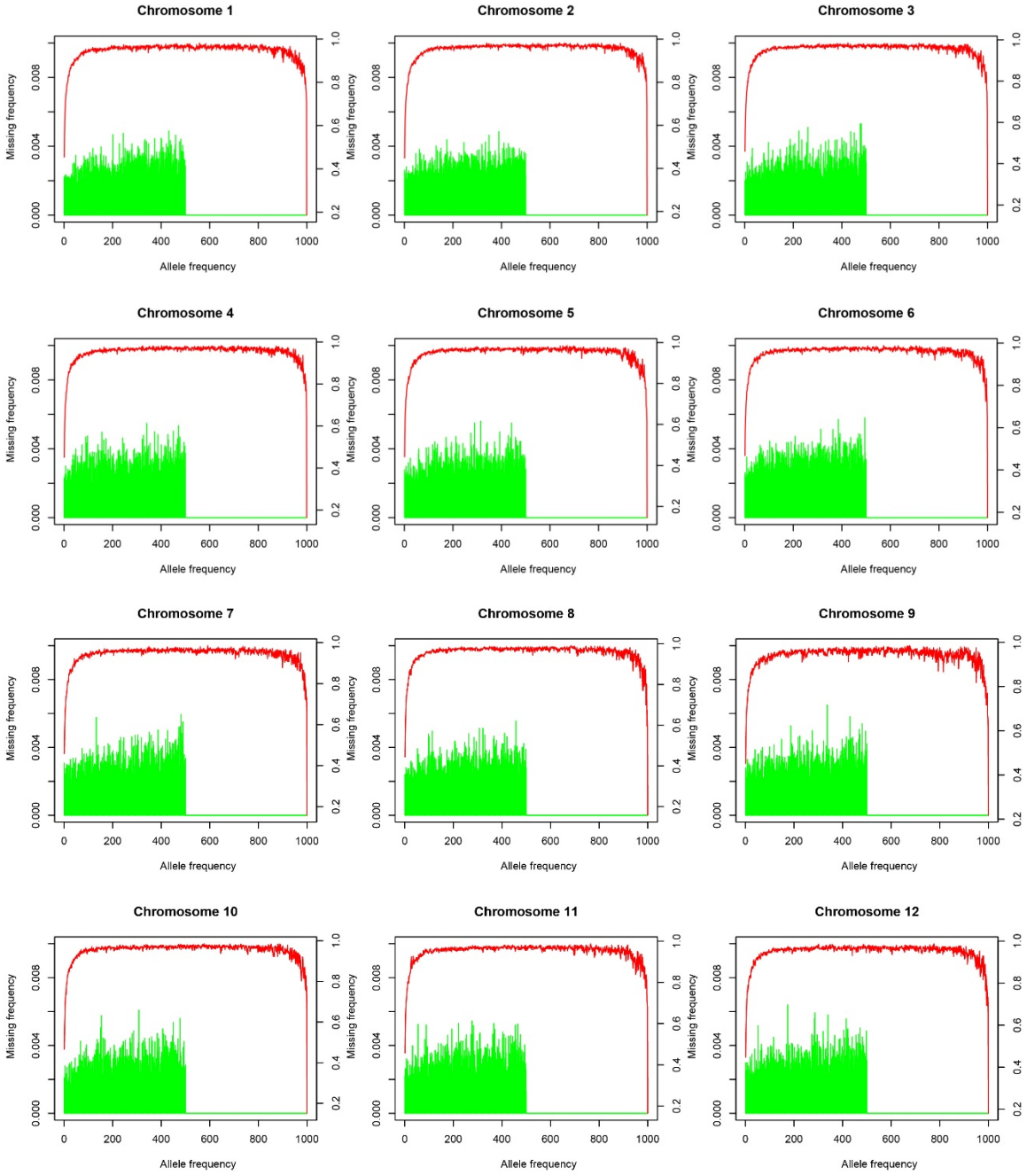


Figure S4 Changes of imputation quality across different genotype missing thresholds. When singleton and monomorphic sites were excluded from the reference, the highest imputation quality was achieved compared to other scenarios. When the entire reference was used, the imputation quality was particularly low for very rare SNPs (0.001 < MAF < 0.005; red line). The less rare and common SNPs (MAF > 0.005, i.e., green, blue, orange, yellow, and black lines) were not influenced as much by the removal of singletons and monomorphs in reference panel. Moreover, for very rare SNPs the exclusion of as many as 39.3% of the SNPs (i.e., "0.1per_NoSM" in the figure) led to a smaller decrease of imputation quality than inclusion of singletons and monomorphic SNPs in reference panel. NoSin: no reference singletons; NoAm: no reference ambiguous SNPs; NoSM: no reference singletons or monomorphs; *per: after removing study SNPs with genotype missing rate higher than *%.


Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
Missing frequency
Allele frequency
0.000
0.004
0.008
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000

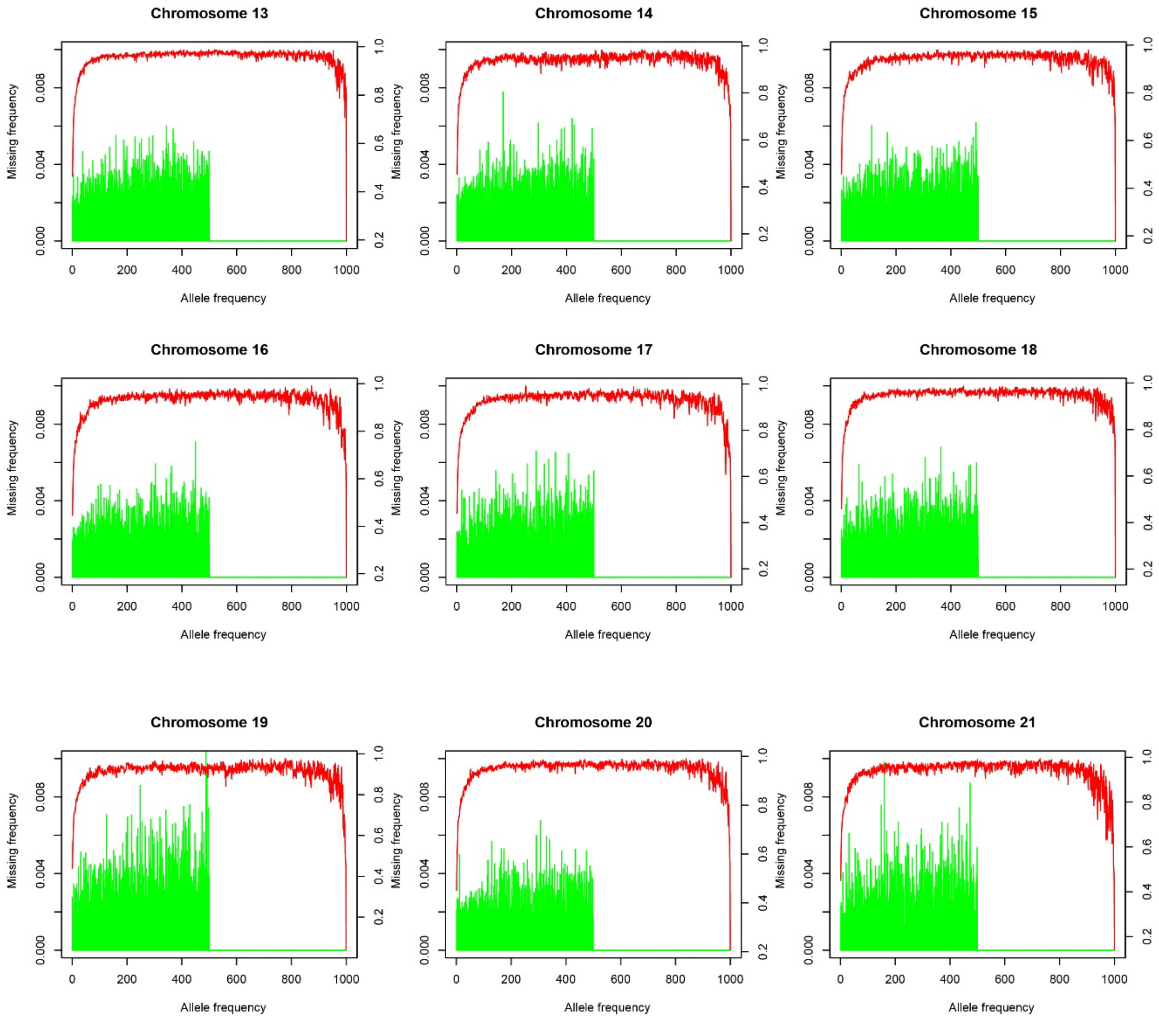


Figure S5 Imputation quality versus missing threshold across 21 autosomes. The green histograms represent genotype missing levels for SNPs that are measured using MAFs from 0.001 to 0.5 while the red curves represent imputation qualities for SNPs that are measured using the full allele frequency from 0.001 to 1.

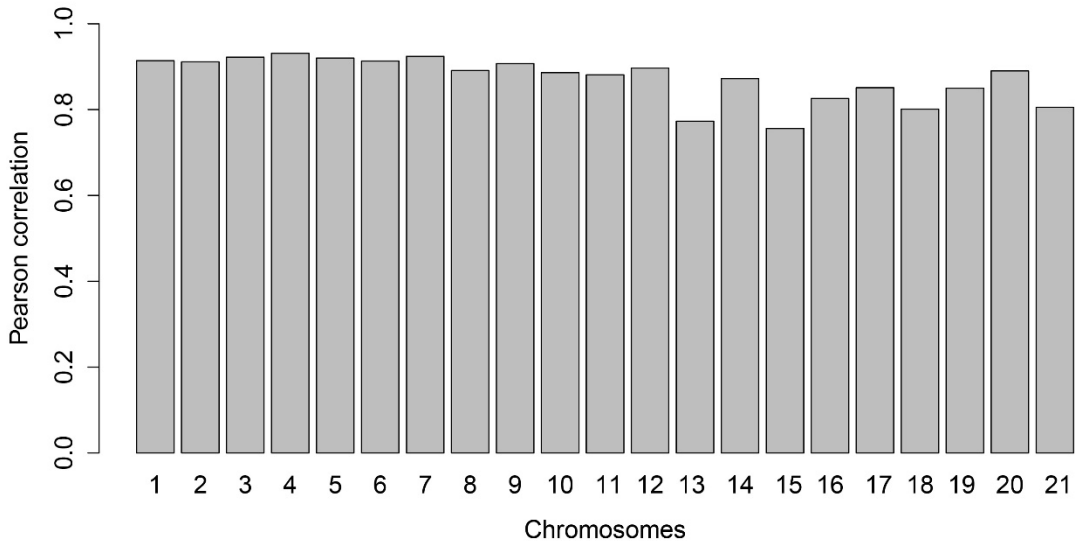


Figure S6 Pearson correlations of mean imputation quality scores between the MAF windows of 0-0.1 and 0.9-1.0. The plots show that the head 10% of the imputation curves is correlated with its tail 10% for all chromosomes, suggesting it is necessary to convert the allele frequencies of imputed SNPs from the range of 0.001-1 to range of 0.001-0.5.

CHAPTER 4: STUCTURAL GENOMIC ABERRATIONS IN BRAIN DISEASE

Chapter 4.1: Genome-wide meta-analysis of copy number variations (CNVs) with alcohol dependence

- The first genome-wide meta-analysis of CNVs with addiction

Arvis Sulovari[1], Zhen Liu[2], Zezhang Zhu[2*], and Dawei Li[1,3,4*]

[1]*Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont*

[2]*Spine Surgery, Drum Tower Hospital, Nanjing University Medical School, Nanjing, China*

[3]*Department of Computer Science, University of Vermont, Burlington, Vermont*

[4]*Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, Vermont*

*To whom correspondence should be addressed:

Dawei Li, Ph.D., Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, US. Tel: 802-656-9838; E-mail: dawei.li@uvm.edu or Zezhang Zhu, MD, Spine Surgery, Drum Tower Hospital, Nanjing University Medical School, Nanjing, 210008, China. Tel: 0086-25-68182022; E-mail: zhuzezhang@126.com

Number of words in the abstract: 206

Number of words in the text (excluding acknowledgments and financial disclosures, legends, and references): 3,600

Number of tables: 5

Number of figures: 4

Number of supplementary materials: 7 supplementary Tables and 4 supplementary Figures and Legends.

Abstract

Genetic association studies and meta-analyses of alcohol dependence (AD) have reported AD-associated single nucleotide polymorphisms (SNPs). These SNPs collectively account for a small portion of estimated heritability in AD. Recent genome-wide copy number variation (CNV) studies have identified CNVs associated with AD and substance dependence, suggesting that a portion of the missing heritability is explained by CNV. We applied PennCNV and QuantiSNP CNV calling algorithms to identify consensus CNVs in five AD cohorts of European and African origins. After rigorous quality control, genome-wide meta-analyses of CNVs were carried out in 3,243 well-diagnosed AD cases and 2,802 controls. We identified nine CNV regions, including a deletion in chromosome 5q21.3 with a suggestive association with AD (OR = 2.15 (1.41 - 3.29) and $P = 3.8 \times 10^{-4}$) and eight nominally significant CNV regions. All regions were replicated with consistent effect sizes across studies and populations. Pathway and gene-drug interaction enrichment analyses based on the resulting genes indicated mitogen-activated protein kinase signaling pathway (MAPK) and two drugs, recombinant insulin and hyaluronidase drugs, all relevant to AD biology or treatment. To our knowledge, this is the first genome-wide meta-analysis of CNVs with addiction. Further investigation of the AD-associated CNV regions will provide better understanding of the AD genetic mechanism.

Keywords: Copy number variation (CNV); Genome-wide meta-analysis; Alcohol dependence; Missing heritability; Structural variation

Introduction

Substance use disorders cost the United States over $200 billion a year (National Institute on Drug Abuse). Alcohol dependence (AD) is one of the most common substance use disorders. Twin studies have reported a genetic heritability of 50-60%[34]. Many genetic association studies and meta-analyses of AD, by our group and others[19,26,28,150-153,220-223], have reported AD-associated single nucleotide polymorphisms (SNPs). Each of the reported SNPs is likely to account for less than 1% of the AD heritability[224], and collectively, they explain a small portion of the estimated heritability in AD, leading to the phenomenon of missing heritability. Copy number variation (CNV) is the gain or loss of a segment of DNA sequence and it may influence thousands of genes or an estimated 12% of the human genome sequence[225]. CNV-based genome-wide association studies (GWASs) have identified CNVs associated with AD[226,227] and/or other substance dependence[146,228], suggesting that CNV also contributes to the missing heritability. Multiple large AD genetics projects have been established for sharing among the research community in the past year (**Table 1**), including the Study of Addiction: Genetics and Environment (SAGE), Collaborative Study on the Genetics of Alcoholism – Center for Inherited Disease Research (CIDR), and Genome-wide Association Study of Alcohol Use and Alcohol Use Disorder in Australian Twin-Families (OZALC). Individual case-control studies based on these cohorts have identified CNVs associated with AD, such as CNVs in 16q12.2[226] and 5q13.2[227]. However, it is unclear whether the associations can be replicated in other research cohorts or populations. A systematic meta-analysis is needed

to clarify the CNV associations. To our knowledge, no meta-analysis of CNV-based GWAS of AD has been published.

In this study, we carried out the first genome-wide meta-analysis between AD and CNVs. We analyzed a total of 6,045 well-diagnosed samples of European and African origins, including 3,243 cases and 2,802 controls. We applied our in-house pipeline of multiple CNV calling algorithms[229-231], which have been demonstrated to increase CNV calling accuracies compared to any single algorithm alone by our study[146] and others[231]. We identified nine CNVs associated with AD, and all of them showed consistent effect direction and magnitude across populations.

Materials and Methods

Research Subjects

The subjects were collected through three established studies, including SAGE, CIDR, and OZALC (the substance dependence cohort that we recently published[146] was not included here because no probes were found in the microarray for the top regions reported in this meta-analysis). All samples were ascertained for alcohol dependence (AD) diagnosis using the Diagnostic and Statistical Manual of Mental Disorders fourth edition (DSM-IV) or third (revised) edition (DSM-IIIR) (American Psychiatry Association, 1994). Controls were individuals who were exposed to alcohol but did not

meet the AD criteria defined by the DSM. The self-reported ancestry information was confirmed using principal component analysis (PCA). In samples where principal components were not readily available from the original studies, we conducted PCA based on autosomal genotypes using the GCTA tool '--pca 20' function[232]. PCA plots from this analysis identified two main ancestries, European and African, which were retained for further analysis. In total, we obtained 10,195 samples, including 3,953, 1,740, and 4,502 samples from SAGE, CIDR, and OZALC, respectively.

Genotyping

DNA extraction and genotyping experiments were carried out by each respective study, while the raw signal intensity information of each sample was obtained via the database of genotypes and phenotypes (dbGaP). As described in **Table 1**, DNA was extracted from saliva, buccal swabs, whole blood or immortalized cell lines, and the genotyping was carried out by the Illumina beadchip arrays (Illumina, San Diego, California).

CNV Calling

The raw intensity files were first processed using GenomeStudio software (Illumina, San Diego, California) where multiple algorithms were employed, including internal quality controls. The B allele frequency (BAF) and log R ratio (LRR) information, which were required for our CNV calling, were generated by the final report module. Our in-house CNV calling pipeline combined PennCNV[229] and QuantiSNP (version 2.0)[230], based on

the published software CNVision[231]. The two CNV callers combine different parameters, including LRR, BAF, and distance between neighboring probes, into a hidden Markov Model (PennCNV) or Bayes hidden Markov Model (QuantiSNP). **Figure 1** shows the workflow of our CNV detection and association analyses.

Statistical Analyses

Individual Cohort-Level Regression Analysis of Common CNVs

Each CNV was mapped to the supporting probe loci of the genotyping array. For each locus, logistic regression was adopted to identify associations between AD and CNVs with frequency > 1%, i.e., the AD diagnosis (dependent variable) was regressed against the copy number status (independent variable) at each probe. To control for potential confounders, multiple covariates were applied, including age, sex, DNA source, genotyping batch (the genotyping batch groups were labelled as “geno.batch”, “Sample_group”, and “Sample.group” in the SAGE, CIDR and OZALC studies, respectively; and the results with genotyping batch adjustment were similar to those without adjustment in most of these tests), and first five principal components. The CNVs that exhibited both copy gain and loss were encoded with three categories, i.e., copy loss, normal copy and copy gain, and the copy number of two was used as the reference. From the association analysis, we obtained an effect size, i.e., odds ratio (OR) with 95% confidence interval (CI), and *P* value for each probe locus. Each of the five populations,

i.e., two African and three European populations, was analyzed separately. Male and female samples were also analyzed separately for CNVs on the X chromosome.

Individual Study-Level Collapsing-based Analyses of Rare CNVs

To identify AD-associated genes with rare CNVs, we projected all CNVs to the 51,509 coding and non-coding gene region reference (UCSC Genome Browser, HG18/NCBI36, last accessed on April 28, 2016), and conducted permutation testing for each gene region using the PLINK[209] label-swapping permutation function '--mperm'. The analysis was performed separately in four CNV frequency windows, i.e., 0-0.25%, 0-1%, 0-2% and 0-5%. Our published tool, GACT[233], was used to test the consistency of genome builds among datasets from the three cohorts.

Random Effects Meta-analyses

For each probe locus, a two-by-two table was populated with counts of cases and controls with or without CNVs. The random effects model, implemented in the DerSimonian-Laird estimator[234], was used in the meta-analyses. For each probe we obtained an OR with 95% CI, *P* value from meta-analysis, and *P* value from heterogeneity test (Q test). The package *metaphor* in the statistical programming language R (version 3.3.0) was used for all the meta-analyses[235]. Only the probes shared among the cohorts were included in the meta-analyses. The meta-analyses were conducted separately for deletions and duplications. The genome-wide significance threshold was $\alpha = 1.8 \times 10^{-5}$, based on

the total number of CNV regions defined by the probes shared across the three studies; the suggestive threshold was $\alpha = 5 \times 10^{-4}$, based on the distribution of *P* values from the meta-analyses.

Analyses of Gene Pathways and Gene-Drug Interactions

Pathway enrichment analysis was performed using all genes near or overlapping with CNV regions that showed meta-analysis *P* values ≤ 0.1 in Europeans or Africans. Since the effect of deletion is abolishment of gene activity, compared to the ambiguous effect of duplication, deletion CNVs were analyzed for enrichment both separately and in combination with duplication CNVs. WebGestalt[104] (last accessed April 18, 2017) was used to test whether these genes were enriched in certain biological or disease pathways maintained in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[93]. The statistical significance was evaluated under the hypergeometric probability of the overlap between our meta-analysis gene sets and KEGG pathway-specific gene sets (last accessed on April 18, 2017), as described in our recent study[194]. Enrichment was calculated using all the genes available in the Entrez Gene database[236] as the background pool of genes, from which our query genes were presumed to have been sampled. Webgestalt[104] was also applied to identify whether any of the meta-analysis gene sets were associated with known drugs based on its curated gene-drug interaction database (the drug terms and associated genes were obtained from PharmGKB[237] and MEDLINE, respectively). The database consists of 758 drug terms with at least five associated genes for each drug[104].

To correct for multiple testing, the resulting *P* values were adjusted using the false discovery (FDR) method[105].

The human genome is nonrandom, and genes from the same pathway tend to cluster together[238]. To replicate the results from pathway enrichment analyses of CNV-derived genes, we conducted permutation tests. Specifically, we generated 20 "null" datasets, where the phenotypes of all samples were permuted independently to generate random distributions. The phenotypes, instead of CNVs, were permuted since we had a fair sample size to produce independent shuffled phenotypes and to preserve the complex relationship between CNVs[239]. For each "null" dataset, we repeated the same meta-analyses, identified "significant" genes, and carried out the exact same pathway enrichment analyses using these genes. For each significant pathway from the real data, we generated a permutation rank, which was defined as the rank of the observed *P* value among all 21 *P* values (20 from the "null" datasets and one from the real dataset, in ascending order).

Results

Sample-Level Quality Controls

Among the 10,195 samples, a total of 504 samples were excluded due to large standard deviation of LRR or BAF, as described in **Table 2**; 1,866 samples were removed due to

family relationship based on identity by descent estimation; 330 samples were excluded due to missing AD diagnosis; and 1,252 samples were excluded because they were genotyped in both SAGE and CIDR. After these quality control measures were applied, a total of 6,243 samples remained.

CNV-Level Quality Controls

Each CNV had to 1) be identified by both PennCNV and QuantiSNP, and 2) contain at least two probes (93% of our identified CNV had at least 5 probes). If the overlap between the CNV regions from the two callers was ≥50%, the two CNVs were considered to be the same CNV, as previously described[146,231]. If a CNV region was designated as deletion by one caller but duplication by the other, it was excluded. Lastly, we removed all the CNVs that did not overlap with those identified by the 1000 Genomes[240] or ExAC[241] projects, resulting in the removal of around 4% of all CNVs (i.e., 7.8%, 7.8%, and 1.9% in SAGE, CIDR, and OZALC, respectively). **Table 3** shows the number of CNVs before and after each CNV-level quality control.

After all the sample- and CNV- level quality controls, we obtained a total of 6,045 samples, including 1,229 Africans and 4,816 Europeans; 3,243 AD patients and 2,802 controls or 3,880 males and 2,165 females (**Table 4**). These quality controls were effective at removing outliers, as indicated in **Figure 2** (combined cohorts) and **Supplementary Figure 1** (individual cohorts). Overall, after applying all quality control

measures, we obtained a total of 321,189 high-quality CNVs. On average, each genome contained 40 CNVs ± 22.5 standard deviation (48, 55, and 19 in SAGE, CIDR, and OZALC, respectively). The majority (i.e., 85%, 85%, and 75% in SAGE, CIDR and OZALC, respectively) of CNVs were between 1 kilo basepairs (kb) and 100kb, and the average CNV lengths were 50 ± 24kb (48kb, 45kb, and 71kb in SAGE, CIDR and OZALC, respectively; **Supplementary Figure 2**). As expected, the vast majority (93% - 97%) of the CNVs were rare (frequency < 1%) with similar patterns in Europeans and Africans (**Supplementary Figure 3**).

Reproducibility of CNV genotyping

We identified a total of 1,252 samples that were genotyped in both CIDR and SAGE datasets. On average, 7.1% of the total CNVs derived from the 1,252 samples were discordant between CIDR and SAGE (**Supplementary Figure 4**). We randomly selected three samples and measured the percentages of CNV boundary concordance. Concordant CNV regions included those with identical start and end positions and those where the shorter CNV was entirely within the boundaries of the longer CNV. We found an average of 90.4% concordance of CNV boundaries based on all 327 CNVs derived from these samples (**Supplementary Table 1**).

Burden Analyses

The CNV burden, i.e., the number of CNVs per sample, varied by study due to the density of microarray probes. The average burden was 48, 55, and 19 CNVs per sample for the SAGE, CIDR, and OZALC datasets, respectively. We found that in the same dataset, CNV burden was slightly higher in AD cases than controls (**Table 4**); across same-ethnicity cohorts, on average, 51.8 versus 49.3 in African cases and controls, respectively (t test $P = 0.02$).

Individual Study-Level Association Analyses

For the common CNVs (frequency > 1%), we found evidence of nominally significant associations with AD at five CNV regions (**Supplementary Table 2**). They included (1) a deletion on chromosome 5q21.3 in Europeans (OR = 3.05 (1.5-6.2) and $P = 0.0019$ in the SAGE cohort); (2) a 14q33.32 deletion in Europeans from CIDR (OR = 3.52 (1.25-9.9) and $P = 0.017$); (3) a 8p23.2 deletion in Africans from SAGE (OR = 1.8 (1.07-3) and $P = 0.03$); (4) a 4p11 duplication in Europeans from CIDR (OR = 2.65 (1.14-7) and $P = 0.03$; and (5) a 6p21.32 deletion in Europeans from CIDR (OR = 2.66 (1.05-7.66) and $P = 0.05$). For rare CNVs (frequency ≤1%), we found evidence of association with AD in the tyrosine phosphatase receptor type D gene (*PPTRD*) in Europeans from SAGE (**Supplementary Table 3**, FDR adjusted $P = 0.02$). *PPTRD* is involved in neuronal signaling and has been implicated in alcohol response[242].

Meta-analyses

Overall, the meta-analyses identified one CNV region, the 5q21.3 deletion (the same CNV described above), with suggestive association with AD (suggestive threshold $\alpha = 5 \times 10^{-4}$, see Methods). The OR was 2.15 (1.41-3.29) and *P* value was 3.8×10^{-4} (OR = 4.13 (0.72-23.6) and *P* = 0.11 in Africans and OR = 2.07 (1.34-3.2) and *P* = 0.001 in Europeans; **Table 5** and **Supplementary Table 4**). This deletion had a frequency of 2.4% and 1.1% in cases and controls, respectively. It is 77.5kb in length, and located upstream of a Ras-oncogene family pseudogene (*RAB9P1*).

We also identified eight additional CNV regions with nominally significant associations with AD (**Figure 4**). They included (1) a 4.3kb deletion in 8p23.2 with frequency of 6.6% and 5.4% in cases and controls, respectively, and OR = 1.38 (1.11-1.73) and *P* = 0.004. This deletion overlaps with *CSMD1*, a gene that has been associated with bipolar disorder[243], autism spectrum disorder[244], and cannabis dependence[245]; (2) a rare 221kb deletion in 14q32.33 with frequency of 1.8% and 0.6% in cases and controls, respectively, and OR = 2.4 (1.25-4.6) and *P* = 0.008. This region overlaps with an immunoglobulin heavy chain pseudogene, and has been associated with several psychiatric disorders, including intellectual disability[246] and Dubowitz syndrome[247]; (3) a 72.7kb duplication in 22q11.21 with frequency of 0.8% and 0.3% in cases and controls, respectively, and OR = 2.88 (1.24-6.69) and *P* = 0.014. This CNV overlaps with the gamma-glutamyltransferase gene (*GGT2*), which has been associated with alcohol consumption and addiction[248]; (4) a 26.7kb deletion in 9p21.1 with frequency of 0.4% and 0.2% in cases and controls, respectively, and OR = 2.8 (1.2-6.4) and *P* = 0.017. This

deletion overlaps with *LINGO2*, a gene that has been associated with essential tremor in Parkinson's disease[249]; (5) a 65.8kb deletion in 9p13.1 with frequency of 0.1% and 0.5% in cases and controls, respectively, and OR = 0.3 (0.1-0.91) and *P* = 0.03. This deletion intersects with *CNTNAP3*, which has been associated with autism spectrum disorder[250]; (6) a 5.3kb deletion in 6p21.32 with frequency of 3.1% and 2.1% in cases and controls, respectively, and OR = 1.44 (1.03-2.01) and *P* = 0.03. The cytogenic region has been associated with alcoholism[251]; (7) a 44kb duplication in 16p11.2 with frequency of 1.8% and 1.5% in cases and controls, respectively, and OR = 1.88 (1.18-3.0) and *P* = 0.035. The cytogenic region has been associated with neuropsychiatric disorders[252]; and (8) a 28.7kb duplication in 12p13.2 with frequency of 4.1% and 3.5% in cases and controls, respectively, and OR = 1.31 (1.0-1.72) and *P* = 0.05. This CNV overlaps with the basic salivary proline-rich protein gene cluster (*PRB1*, *PRB2* and *PRB3)*, which has been reported as important biomarkers in salivary-secretion related phenotypes[253]. Four of the deletions (i.e., 5q21.3, 14q32.33, 9p21.1, and 6p21.32) also showed nominal significance in the individual study-level association analyses (**Supplementary Table 2**).

In silico validation of CNVs

To visualize the CNVs, we plotted the raw LRR and BAF values of each probe to manually curate each CNV "call" reported in **Table 5**; these CNVs contained a range of 8 to 131 probes. **Figure 3** shows the results of the 5q21.3 deletion (28 probes) in three randomly-selected samples, and two samples had one-copy deletion and one had two-

copies. Furthermore, all of the nine CNV regions in **Table 5** were also observed in the 1000 Genomes Project (phase III) samples[240], and the CNV boundary coordinates as well as their population-level frequencies were almost 100% consistent. The converging results support an accurate *in silico* calling of our reported CNVs.

Gene Pathways and Gene-Drug Interactions

Our gene pathway analyses showed that the genes from meta-analyses were enriched in the MAPK signaling pathway (R = 6.6 and $P = 0.05$). Further permutation tests confirmed that this *P* value ranked at the top, compared to those from the 20 permutations (**Supplementary Table 5**). MAPK plays a pivotal role in signal transduction of alcohol across tissues[254], and has been reported as a potential mediator of AD and opioid dependence[255]. The gene-drug interaction analyses based on the same gene set showed two associated drugs, recombinant insulin (enriched with CNVs, R = 13 and $P = 0.02$) and hyaluronidase (enriched with deletion CNVs, R = 138 and $P = 9\times10^{-5}$). Similarly, further permutation tests revealed that these *P* values ranked first, compared to those from permutations (**Supplementary Table 6**). Insulin secretion has been shown to increase in response to alcohol[256] and associated with alcohol craving in AD patients [257]; additionally, one of the drug-interacting genes harboring CNVs is *PTPRN2*, which was previously associated with response to amphetamines[258], schizophrenia, and bipolar disorder[259]. Hyaluronidase is an enzyme and often used as an adjuvant to help increase absorption and dispersion of injected drugs and fluids[260]; one of the drug-interacting

genes harboring deletion CNVs is *WWOX*, which has been associated with smoking behavior[261]. Additionally, hyaluronidase cleaves hyaluronan, which interacts with the extracellular matrix (ECM); recent work has demonstrated the importance of the interaction between the brain ECM and alcohol in AD[262].

Discussion

We report the first genome-wide meta-analysis between CNVs and AD. We systematically identified CNV regions based on three established substance use disorder cohorts. The CNVs were called using our in-house pipeline based on PennCNV[229] and QuantiSNP(v2.0)[230]. Previous genome-wide CNV studies from our group and others have demonstrated that the consensus CNV regions independently genotyped by these two callers were highly replicated by qPCR experiments[146,263]. Our quality control procedures (**Tables 2 and 3**) effectively removed outlier samples and false positive CNVs, leading to the expected distribution of CNV burden across analyzed samples (**Figure 2** and **Supplementary Figure 1**). Meta-analyses of the curated high-quality CNVs showed nine nominally significant regions with AD (**Figure 4**), six deletions and three duplications; although the individual studies might be underpowered, they collectively revealed consistent effect sizes, in both direction and magnitude (**Table 5** and **Supplementary Table 4**). The nine CNVs ranged from 4.3kb to 221.7kb in size and had ORs from 1.31 to 2.88; and eight of them had frequency ≤ 5% (no CNV imputation conducted in this study due to low frequencies of these CNVs). The most significant AD association was found

with the 5q21.3deletion (OR = 2.15 and $P = 3.8 \times 10^{-4}$). This cytogenetic band has been associated with alcohol cravings in a Native American population[264]. This meta-analysis, for the first time, identified a specific CNV in this region associated with AD.

A careful review of the literature revealed that the majority of these CNV regions or intersecting genes identified in this meta-analysis have been associated with AD (e.g., 5q21.3 and 6p21.32) or psychiatric disorders (e.g., 8p32.2, 14q32.33, 9p21.1, 9p13.1, and 16p11.2), although not all of them were GWAS-replicated regions. The *GGT2* gene, overlapping with the 22q11.21 duplication, has been associated with alcohol consumption and addiction[248]; *CNTNAP3*, overlapping with the 9p13.1 deletion, has been associated with autism spectrum disorder[250]; and *CSMD1*, overlapping with the 8p23.2 deletion, has been associated with bipolar disorder[243], autism spectrum disorder[244], and schizophrenia[265]. The *PPTRD* gene identified in our collapsing-based analyses has also been implicated in alcohol response[242]. Our findings support the roles of rare CNVs in addiction, as described in our recent CNV study of opioid dependence[146]. Interestingly, the gene-drug interaction analyses based on the meta-analysis genes revealed one drug (recombinant insulin) relevant to AD biology and another (hyaluronidase) known to interact with a gene associated with smoking behavior.

Limitations of our study include lack of genome-wide significance and molecular validation. First, the lack of genome-wide statistical significance may indicate that our study was underpowered for the specific CNVs analyzed. Indeed, our analyses showed

that the statistical power for detection of the CNVs with frequencies from 0.5% to 3.5% and odds ratios from 1.3 to 2.8 at $\alpha = 1.8 \times 10^{-5}$ was under 80% (**Supplementary Table 7)**, indicating that larger sample sizes are required in future studies. For instance, to achieve 80% power for detecting the CNVs with frequency = 0.5% and OR=2, a cohort of 11,838 samples is required. In addition to increasing sample sizes, collapsing rare CNVs may also increase power, particularly for rare CNVs[266]. Second, since the five cohorts analyzed in this meta-analysis were recruited by different institutions and investigators, a timely collection of sufficient DNA from all of these cohorts for molecular validation is complicated for most individual investigator. Future collaboration through related research consortium is needed. In all, replication of the findings in larger samples is warranted and further investigation of the reported structural variations may lead to identification of novel AD genes.

Acknowledgements

This work was supported by the Start-up Fund of The University of Vermont. The raw signal intensity data described in this study were obtained from the database of Genotypes and Phenotypes through accession numbers phs000092 (SAGE), phs000125 (CIDR), and phs000181 (OZALC). The authors acknowledge the Vermont Advanced Computing Core for providing high performance computing resources at the University of Vermont. The authors thank Gina Castellano and Addison Marcus for their very

careful review of some of the reported CNVs. The authors also thank Zoe Furlong for her careful review of the manuscript.

Conflict of Interest

The authors declare no potential conflict of interest.

References

1. Gelernter J, Kranzler HR. Genetics of alcohol dependence. *Human genetics* 2009; **126**(1)**:** 91-99.

2. Gelernter J, Kranzler HR, Sherva R, Almasy L, Koesterer R, Smith AH, *et al*. Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. *Molecular psychiatry* 2014; **19**(1)**:** 41-49.

3. Frank J, Cichon S, Treutlein J, Ridinger M, Mattheisen M, Hoffmann P, *et al*. Genome-wide significant association between alcohol dependence and a variant in the ADH gene cluster. *Addict Biol* 2012; **17**(1)**:** 171-180.

4. Li D, Zhao H, Gelernter J. Further clarification of the contribution of the ADH1C gene to vulnerability of alcoholism and selected liver diseases. *Human genetics* 2012; **131**(8)**:** 1361-1374.

5. Li D, Zhao H, Gelernter J. Strong association of the alcohol dehydrogenase 1B gene (ADH1B) with alcohol dependence and alcohol-induced medical diseases. *Biological psychiatry* 2011; **70**(6)**:** 504-512.

6. Li D, Zhao H, Gelernter J. Strong protective effect of the aldehyde dehydrogenase gene (ALDH2) 504lys (*2) allele against alcoholism and alcohol-induced medical diseases in Asians. *Human genetics* 2012; **131**(5)**:** 725-737.

7. Cao J, Hudziak JJ, Li D. Multi-cultural association of the serotonin transporter gene (SLC6A4) with substance use disorder. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 2013; **38**(9)**:** 1737-1747.

8. Cao J, LaRocque E, Li D. Associations of the 5-hydroxytryptamine (serotonin) receptor 1B gene (HTR1B) with alcohol, cocaine, and heroin abuse. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 2013; **162B**(2)**:** 169-176.

9. Cao J, Liu X, Han S, Zhang CK, Liu Z, Li D. Association of the HTR2A gene with alcohol and heroin abuse. *Human genetics* 2014; **133**(3)**:** 357-365.

10. Li D, Sulovari A, Cheng C, Zhao H, Kranzler HR, Gelernter J. Association of gamma-aminobutyric acid A receptor alpha2 gene (GABRA2) with alcohol use disorder. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* 2014; **39**(4)**:** 907-918.

11. Sulovari A, Kranzler HR, Farrer LA, Gelernter J, Li DW. Further Analyses Support the Association Between Light Eye Color and Alcohol Dependence. *Am J Med Genet B* 2015; **168**(8)**:** 757-760.

12. Sulovari A, Kranzler HR, Farrer LA, Gelernter J, Li DW. Eye color: A potential indicator of alcohol dependence risk in European Americans. *Am J Med Genet B* 2015; **168**(5)**:** 347-353.

13. Palmer RH, McGeary JE, Francazio S, Raphael BJ, Lander AD, Heath AC, *et al*. The genetics of alcohol dependence: advancing towards systems-based approaches. *Drug and alcohol dependence* 2012; **125**(3)**:** 179-191.

14. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* 2007; **39**(7 Suppl)**:** S16-21.

15. Ulloa AE, Chen JY, Vergara VM, Calhoun V, Liu JY. Association Between Copy Number Variation Losses and Alcohol Dependence Across African American and European American Ethnic Groups. *Alcoholism-Clinical and Experimental Research* 2014; **38**(5)**:** 1266-1274.

16. Lin P, Hartz SM, Wang JC, Agrawal A, Zhang TX, McKenna N, *et al*. Copy Number Variations in 6q14.1 and 5q13.2 are Associated with Alcohol Dependence. *Alcoholism-Clinical and Experimental Research* 2012; **36**(9)**:** 1512-1518.

17. Li D, Zhao H, Kranzler HR, Li MD, Jensen KP, Zayats T, *et al*. Genome-wide association study of copy number variations (CNVs) with opioid dependence. *Neuropsychopharmacology* 2015; **40**(4)**:** 1016-1026.

18. Cabana-Dominguez J, Roncero C, Grau-Lopez L, Rodriguez-Cintas L, Barral C, Abad AC, *et al*. A Highly Polymorphic Copy Number Variant in the NSF Gene is Associated with Cocaine Dependence. *Scientific reports* 2016; **6:** 31033.

19. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, *et al*. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007; **17**(11)**:** 1665-1674.

20. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, *et al*. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic acids research* 2007; **35**(6)**:** 2013-2025.

21. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, *et al*. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 2011; **70**(5)**:** 863-885.

22. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* 2011; **88**(1)**:** 76-82.

23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 2007; **81**(3)**:** 559-575.

24. Sulovari A, Li D. GACT: a Genome build and Allele definition Conversion Tool for SNP imputation and meta-analysis in genetic association studies. *BMC genomics* 2014; **15:** 610.

25. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in medicine* 1997; **16**(7)**:** 753-768.

26. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *J Stat Softw* 2010; **36**(3)**:** 1-48.

27. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic acids research* 2013; **41**(Web Server issue)**:** W77-83.

28. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research* 2014; **42**(Database issue)**:** D199-205.

29. Sulovari A, Chen YH, Hudziak JJ, Li D. Atlas of human diseases influenced by genetic variants with extreme allele frequency differences. *Human genetics* 2017; **136**(1)**:** 39-54.

30. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* 2011; **39**(Database issue)**:** D52-57.

31. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, *et al*. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic acids research* 2002; **30**(1)**:** 163-165.

32. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* 1995; **57**(1)**:** 289-300.

33. Thevenin A, Ein-Dor L, Ozery-Flato M, Shamir R. Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic acids research* 2014; **42**(15)**:** 9854-9861.

34. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nature reviews Genetics* 2014; **15**(5)**:** 335-346.

35. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, *et al*. An integrated map of structural variation in 2,504 human genomes. *Nature* 2015; **526**(7571)**:** 75-81.

36. Ruderfer DM, Hamamsy T, Lek M, Karczewski KJ, Kavanagh D, Samocha KE, *et al*. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nature genetics* 2016; **48**(10)**:** 1107-1111.

37. Joslyn G, Ravindranathan A, Brush G, Schuckit M, White RL. Human variation in alcohol response is influenced by variation in neuronal signaling genes. *Alcoholism, clinical and experimental research* 2010; **34**(5)**:** 800-812.

38. Xu W, Cohen-Woods S, Chen Q, Noor A, Knight J, Hosang G, *et al*. Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1. *BMC medical genetics* 2014; **15:** 2.

39. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, *et al*. Excess of rare, inherited truncating mutations in autism. *Nature genetics* 2015; **47**(6)**:** 582-588.

40. Sherva R, Wang Q, Kranzler H, Zhao H, Koesterer R, Herman A, *et al*. Genome-wide Association Study of Cannabis Dependence Severity, Novel Risk Variants, and Shared Genetic Risks. *JAMA psychiatry* 2016; **73**(5)**:** 472-480.

41. Maurin ML, Brisset S, Le Lorc'h M, Poncet V, Trioche P, Aboura A, *et al*. Terminal 14q32.33 deletion: genotype-phenotype correlation. *American journal of medical genetics Part A* 2006; **140**(21)**:** 2324-2329.

42. Stewart DR, Pemov A, Johnston JJ, Sapp JC, Yeager M, He J, *et al*. Dubowitz syndrome is a complex comprised of multiple, genetically distinct and phenotypically overlapping disorders. *PloS one* 2014; **9**(6)**:** e98686.

43. Franzini M, Fornaciari I, Vico T, Moncini M, Cellesi V, Meini M, *et al*. High-sensitivity gamma-glutamyltransferase fraction pattern in alcohol addicts and abstainers. *Drug and alcohol dependence* 2013; **127**(1-3)**:** 239-242.

44. Wu YW, Prakash KM, Rong TY, Li HH, Xiao Q, Tan LC, *et al*. Lingo2 variants associated with essential tremor and Parkinson's disease. *Human genetics* 2011; **129**(6)**:** 611-615.

45. Vaags AK, Lionel AC, Sato D, Goodenberger M, Stein QP, Curran S, *et al*. Rare deletions at the neurexin 3 locus in autism spectrum disorder. *American journal of human genetics* 2012; **90**(1)**:** 133-141.

46. Demirhan O, Tastemir D. Cytogenetic effects of ethanol on chronic alcohol users. *Alcohol and alcoholism* 2008; **43**(2)**:** 127-136.

47. Zufferey F, Sherr EH, Beckmann ND, Hanson E, Maillard AM, Hippolyte L, *et al*. A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. *Journal of medical genetics* 2012; **49**(10)**:** 660-668.

48. Azen EA, Latreille P, Niece RL. PRBI gene variants coding for length and null polymorphisms among human salivary Ps, PmF, PmS, and Pe proline-rich proteins (PRPs). *American journal of human genetics* 1993; **53**(1)**:** 264-278.

49. Aroor AR, Shukla SD. MAP kinase signaling in diverse effects of ethanol. *Life sciences* 2004; **74**(19)**:** 2339-2364.

50. Zamora-Martinez ER, Edwards S. Neuronal extracellular signal-regulated kinase (ERK) activity as marker and mediator of alcohol and opioid dependence. *Frontiers in integrative neuroscience* 2014; **8:** 24.

51. Huang Z, Sjoholm A. Ethanol acutely stimulates islet blood flow, amplifies insulin secretion, and induces hypoglycemia via nitric oxide and vagally mediated mechanisms. *Endocrinology* 2008; **149**(1)**:** 232-236.

52. Leggio L, Ray LA, Kenna GA, Swift RM. Blood glucose level, alcohol heavy drinking, and alcohol craving during treatment for alcohol dependence: results from the Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence (COMBINE) Study. *Alcoholism, clinical and experimental research* 2009; **33**(9)**:** 1539-1544.

53. Hart AB, Engelhardt BE, Wardle MC, Sokoloff G, Stephens M, de Wit H, *et al*. Genome-wide association study of d-amphetamine response in healthy volunteers

identifies putative associations, including cadherin 13 (CDH13). *PloS one* 2012; **7**(8)**:** e42646.

54. Curtis D, Vine AE, McQuillin A, Bass NJ, Pereira A, Kandaswamy R, *et al*. Case-case genome-wide association analysis shows markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes. *Psychiatric genetics* 2011; **21**(1)**:** 1-4.

55. Dunn AL, Heavner JE, Racz G, Day M. Hyaluronidase: a review of approved formulations, indications and off-label use in chronic pain management. *Expert opinion on biological therapy* 2010; **10**(1)**:** 127-131.

56. Park SL, Carmella SG, Chen M, Patel Y, Stram DO, Haiman CA, *et al*. Mercapturic Acids Derived from the Toxicants Acrolein and Crotonaldehyde in the Urine of Cigarette Smokers from Five Ethnic Groups with Differing Risks for Lung Cancer. *PloS one* 2015; **10**(6)**:** e0124841.

57. Lasek AW. Effects of Ethanol on Brain Extracellular Matrix: Implications for Alcohol Use Disorder. *Alcoholism, clinical and experimental research* 2016; **40**(10)**:** 2030-2042.

58. Kim SY, Kim JH, Chung YJ. Effect of Combining Multiple CNV Defining Algorithms on the Reliability of CNV Calls from SNP Genotyping Data. *Genomics & informatics* 2012; **10**(3)**:** 194-199.

59. Ehlers CL, Wilhelmsen KC. Genomic scan for alcohol craving in Mission Indians. *Psychiatric genetics* 2005; **15**(1)**:** 71-75.

60. Schizophrenia Psychiatric Genome-Wide Association Study C. Genome-wide association study identifies five new schizophrenia loci. *Nature genetics* 2011; **43**(10)**:** 969-976.

61. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. *American journal of human genetics* 2013; **93**(1)**:** 42-53.

Tables

Table 1 Description of the samples analyzed in the meta-analyses prior to quality controls (see http://www.nature.com/tpj/journal/vaop/ncurrent/full/tpj201735a.html)

Table 2 Summary of sample-level quality controls (see

http://www.nature.com/tpj/journal/vaop/ncurrent/full/tpj201735a.html)

Table 3 Summary of CNV-level quality controls (see

http://www.nature.com/tpj/journal/vaop/ncurrent/full/tpj201735a.html)

Table 4 Demographic information of all samples after sample- and CNV-level quality control procedures (see
http://www.nature.com/tpj/journal/vaop/ncurrent/full/tpj201735a.html)

Table 5 Results of meta-analyses between CNV and AD (see

http://www.nature.com/tpj/journal/vaop/ncurrent/full/tpj201735a.html)

Figure Legends

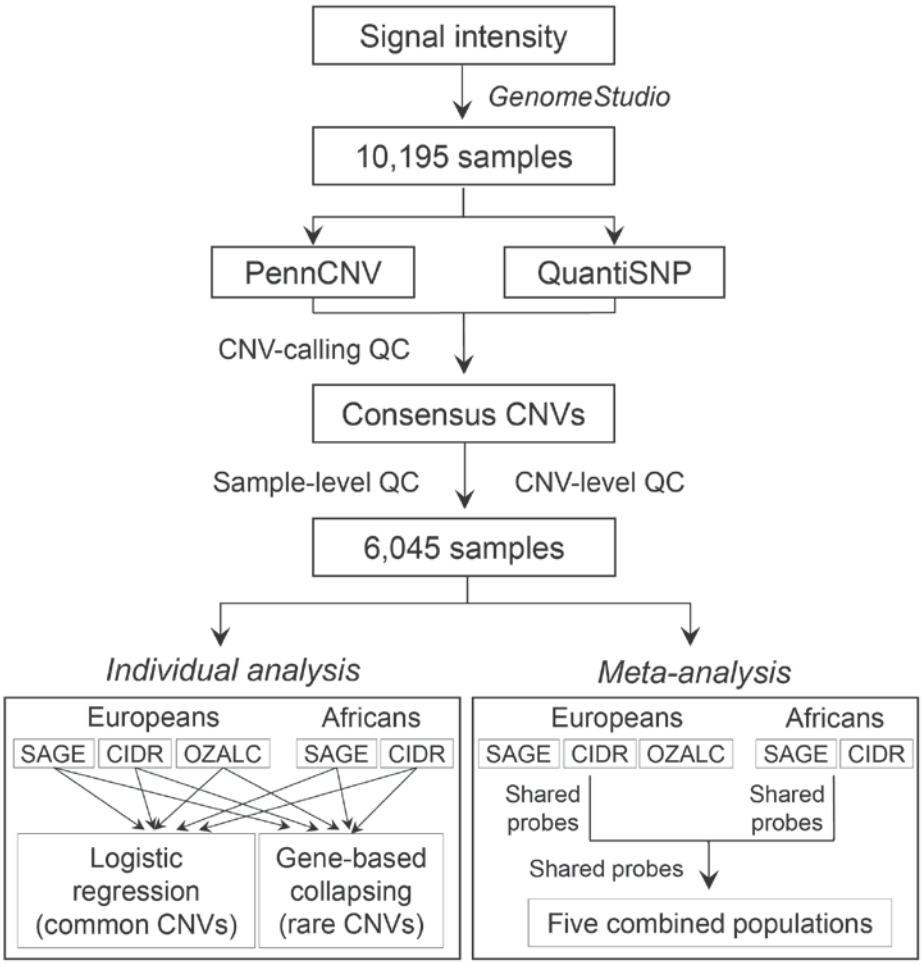


Figure 1 Workflow for CNV calling and association analyses. The raw signal intensity data from Illumina GenomeStudio were used to call CNVs in a total of 10,195 samples using both PennCNV and QuantiSNP. After three rounds of quality controls, i.e., CNV calling, sample-level, and CNV-level, we obtained a total of 6,045 samples, in five populations. We only kept the consensus CNVs called by both algorithms for further analyses. For each individual population (study), logistic regression and gene collapsing methods were applied to analyze the common and rare CNVs, respectively. Meta-analyses of the CNV regions were performed based on the probes shared by the two genotyping arrays.

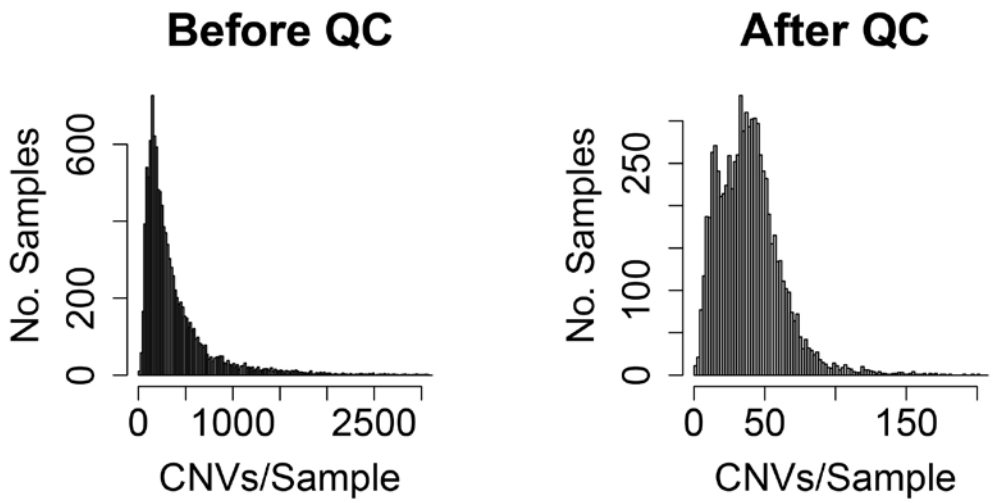


Figure 2 Number of CNVs per sample before and after sample- and CNV-based quality controls. CNVs were pooled across the three cohorts, and the distribution of CNVs per sample was plotted before and after quality controls. Our quality control filters were effective at removing outlier samples, as indicated by the Gaussian shape of the plot on the right (after quality controls), i.e., lack of extreme outliers compared to plot on the left (before quality controls). Plot on the right indicates two genotyping arrays used, one by OZALC and the other by CIDR and SAGE. QC: quality controls.

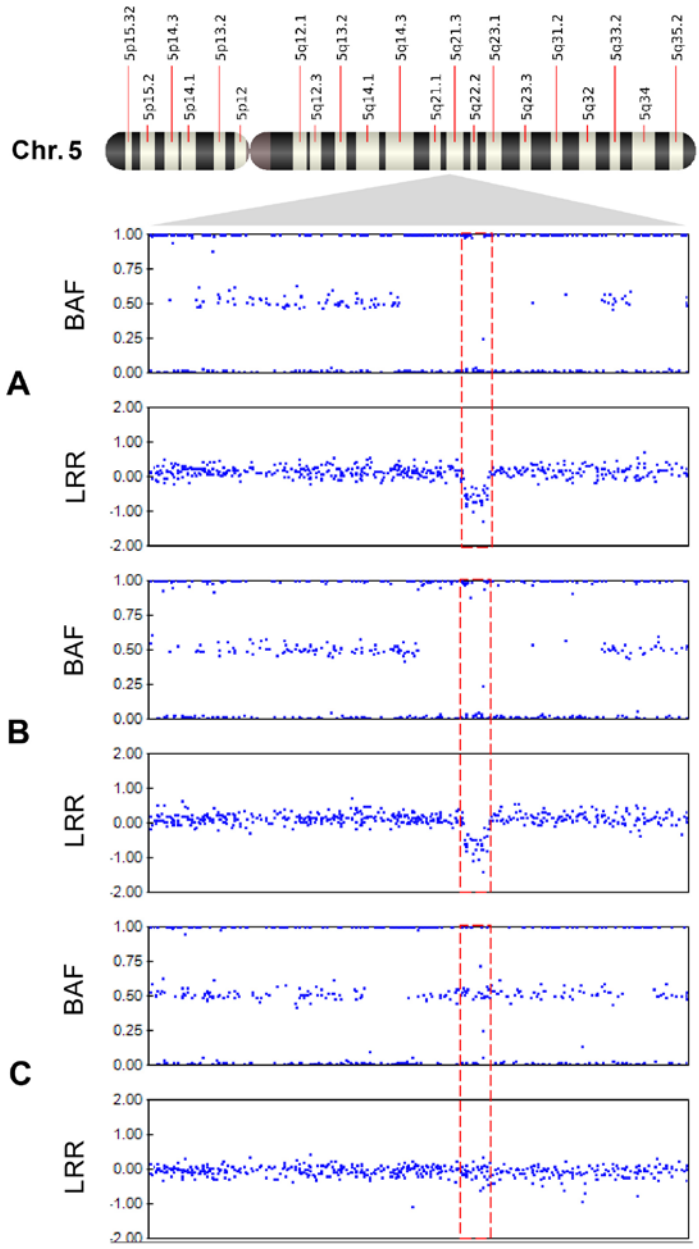


Figure 3 Plots of log R ratio (LRR) and B allele frequency (BAF) of the 5q21.3 deletion. The plots show the LRR and BAF of this deletion in three samples: A, a European case (sample ID: 40721162398); B, a European control (sample ID: 4072116332); and C, a European case (sample ID: 1954615060). Each blue dot represents a microarray probe, and the highlighted window indicates the 5q21.3 deletion region. Plots A and B show single deletion events (copy number of one), while plot C represents normal copy number of two (i.e., negative control). A total of 28 probes were detected in this CNV region by both PennCNV and QuantiSNP.

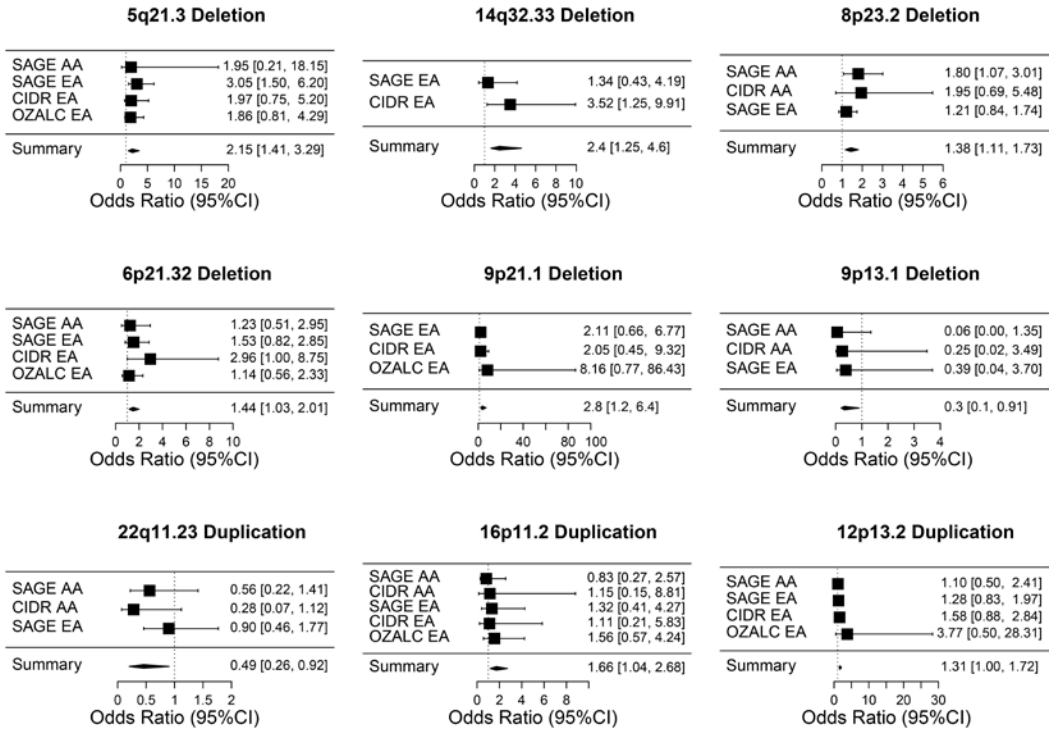


Figure 4 Forrest plot of the individual studies and meta-analysis results. The detailed information of the nine CNVs is shown in **Supplementary Tables 2 and 3**. The odds ratios of each individual study were calculated using logistic regression with correction for appropriate covariates; while the odds ratios of meta-analysis, labelled as "Summary", was calculated using the random effects model (see Methods).

Supplementary Tables and Figures

Supplementary Tables

Supplementary Table 1 Concordance of the CNV boundaries between the CIDR and SAGE datasets

Sample ID	#CNV regions in SAGE	#CNV regions in CIDR	Concordance
4059931034	77	74	98% (100% of CIDR)
4059931127	41	41	90.2% (90.2% of CIDR)
4059931355	47	47	83% (83% of CIDR)

The concordance is calculated as: Number of overlapped regions × 2 / total number of CNVs from both SAGE and CIDR. The numbers in brackets in the last column represent the concordance based on the calculation of number of overlapped regions / total number of CNVs from CIDR.

Supplementary Table 2 Results of logistic regression analyses for nominally significant CNVs identified by individual studies or meta-analyses (http://www.nature.com/tpj/journal/vaop/ncurrent/suppinfo/tpj201735s1.html)

Supplementary Table 3 *P* values of gene-base collapsing analysis of rare CNVs

	CNV Frequency Bin			
Gene	**0-0.25%**	**0-1%**	**0-2%**	**0-5%**
PTPRD (European, SAGE)	0.1	**0.021**	**0.023**	**0.02**

Four frequency bins of rare CNVs were collapsed to known gene regions. Numbers in the table represent the FDR adjusted *P* values based on 10,000 label-swapping permutation tests. *P* values ≤ 0.05 are in bold.

Supplementary Table 4 Results of meta-analyses between CNV and AD (full version; http://www.nature.com/tpj/journal/vaop/ncurrent/suppinfo/tpj201735s1.html)

Supplementary Table 5 Results of pathway enrichment analyses using KEGG

Biological pathway	Contributing genes	Enrichment ratio	*P* value (FDR-adjusted)	*Permutation Rank
MAPK signaling	*HSPA1A*, *DUSP22*	6.6	0.05	1/21

Enrichment ratio is the ratio between the observed and expected numbers of genes for a given pathway.
*, The permutation rank was calculated based on ranking of the observed enrichment *P* value against 20 null enrichment *P* values.
Note: No pathway enrichment was observed when only the genes overlapping with deletion CNVs (meta-analysis $P \leq 0.1$) were analyzed.

Supplementary Table 6 Results of enrichment analyses of gene-drug interactions

Drug pathway	Contributing genes	Enrichment ratio	*P* value (FDR-adjusted)	*Permutation Rank
†Hyaluronidase	*WWOX*, *CTDSPL*	138.4	9×10^{-5}	1/21
Insulin recombinant	*PTPRN2*, *RLN1*	12.8	0.02	1/21

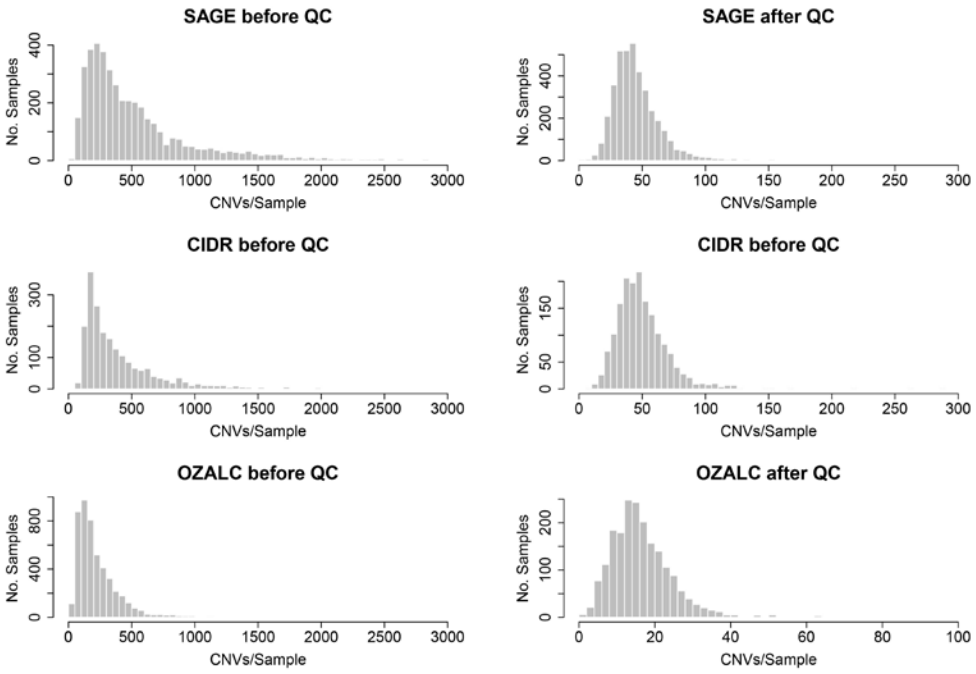
Enrichment ratio is the ratio between the observed and expected number of genes for a given pathway.
*, The permutation rank was calculated based on ranking of the observed enrichment *P* value against 20 null enrichment *P* values.
†, The enrichment analysis was carried out using deletion CNVs only (meta-analysis $P \leq 0.1$).

Supplementary Table 7 Results from statistical power analysis

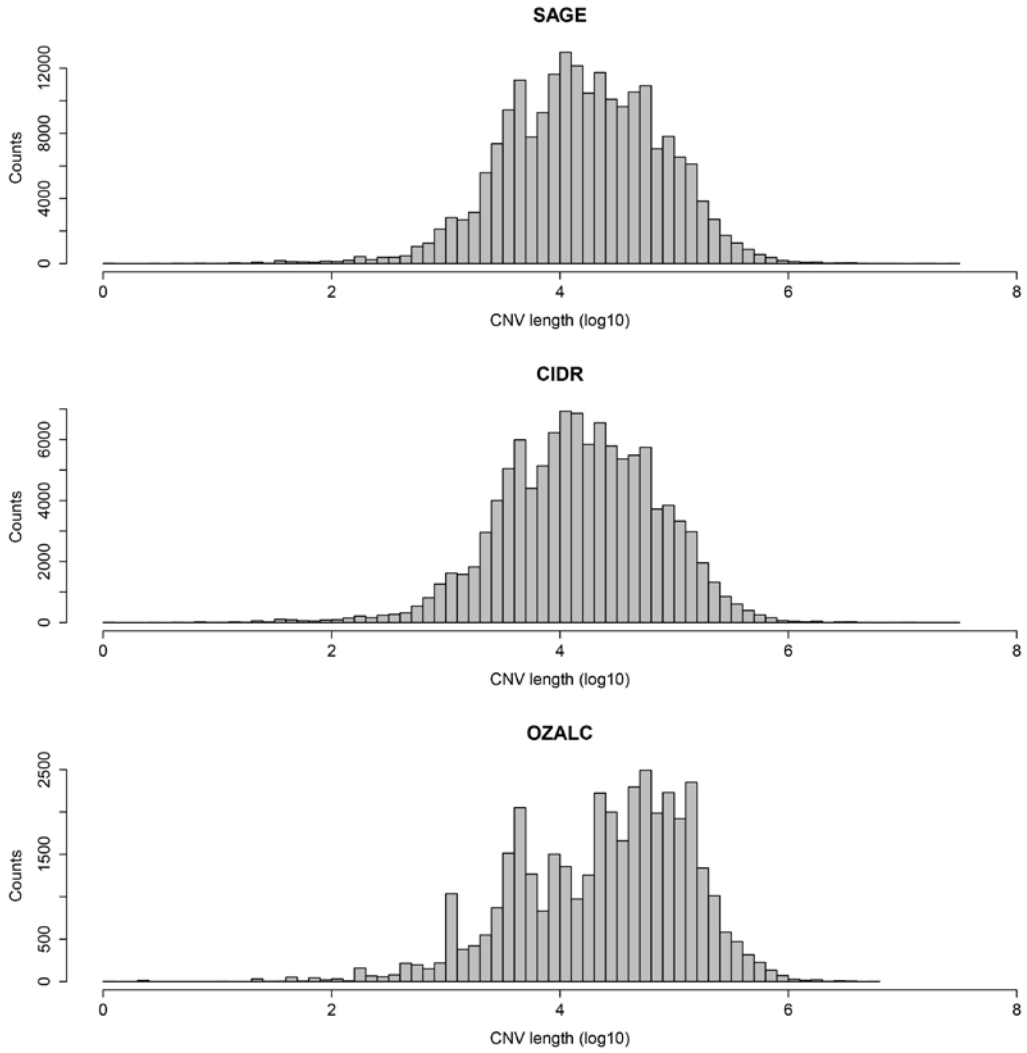
CNV frequency (%)	Odds Ratio	Power (%)
0.005	1.4	0.07
0.015	1.4	0.9
0.035	1.3	1.8
0.025	1.4	3.5
0.005	2.1	7.6
0.005	2.4	21.9
0.005	2.8	51.2
0.015	2.1	72.8

The power analysis was carried out using our in-house scripts, designed to interact with the online tool PGC (pngu.mgh.harvard.edu/~purcell/gpc/). The prevalence of AD was set to 6.2%, as reported by the National Survey on Drug Use and Health (NIAAA, 2015), while the linkage disequilibrium (D prime) parameter was set to 0.8; the odds ratio and allele frequency varied according to the range of our reported CNVs in **Supplementary Table 4**.

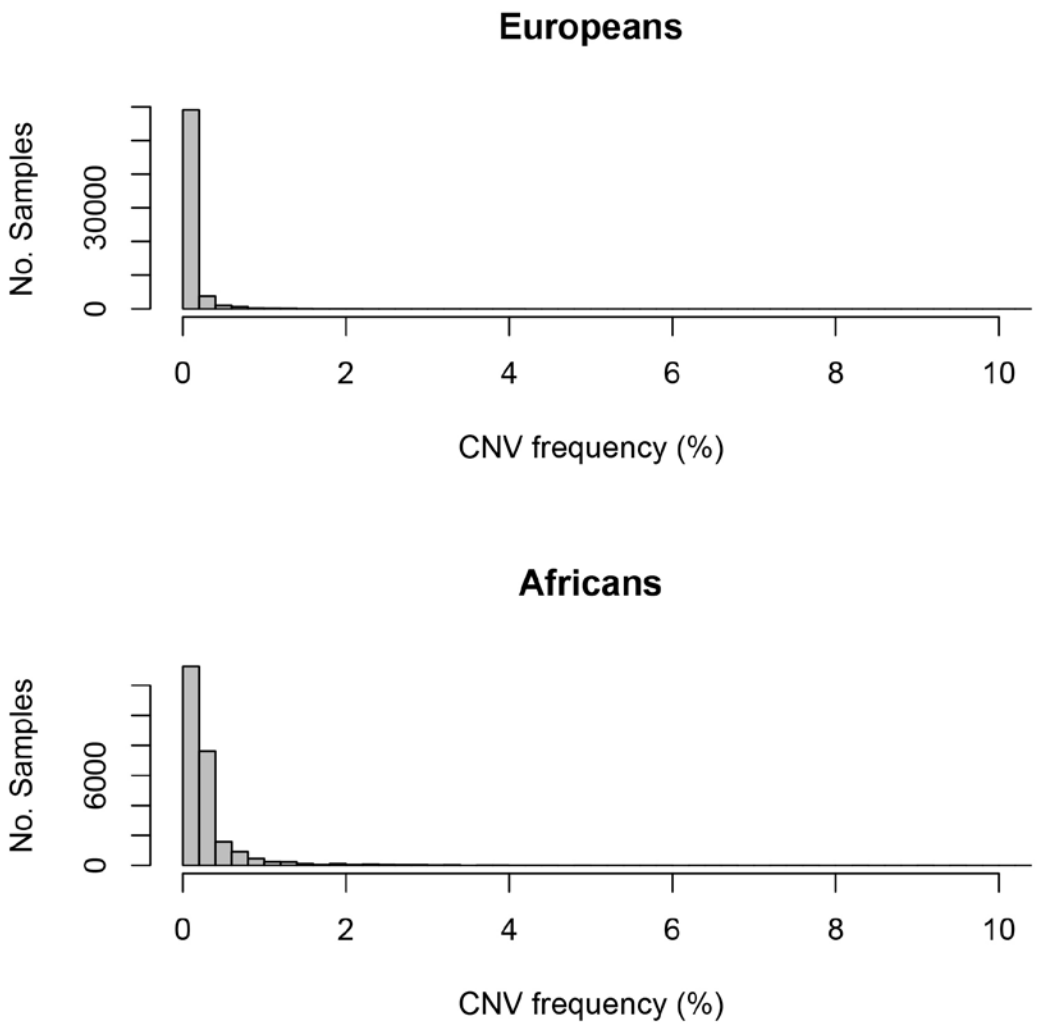
Supplementary Figure Legends



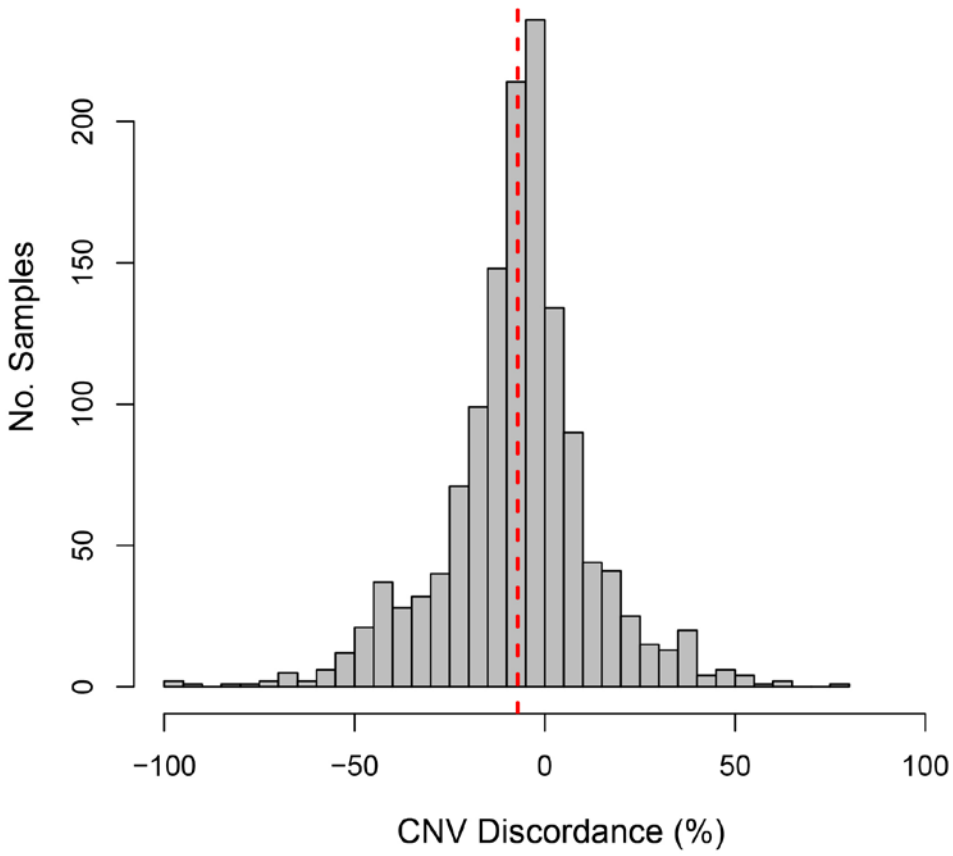
Supplementary Figure 1 Individual study-level number of CNVs per sample before and after sample- and CNV-based quality controls. All samples analyzed in this study were included with the exception of six samples (five unique IDs). Their IDs and respective number of CNVs were 4068221273 (1,844), 4072116227 (2,264), and 4068221885 (3,366) in CIDR; and 4186068211 (820), 4192409004 (1,196), and 4072116227 (2,674) in SAGE. Our quality control filters were effective at removing outlier samples, as indicated by the Gaussian shape of the plots on the right (after quality controls), i.e., lack of extreme outliers compared to plots on the left (before quality controls). QC: quality controls.



Supplementary Figure 2 Distribution of lengths of CNVs discovered by our CNV calling pipeline. The x-axis and y-axis represent the $\log_{10}$ values of CNV lengths and their counts, respectively, for each of the three cohorts.



Supplementary Figure 3 Distribution of frequencies of CNVs discovered by our CNV calling pipeline. The African and European samples are displayed at the top and bottom plots, respectively. The x-axis represents the CNV frequencies (%) while the y-axis shows their sample count of each bin. The average CNV frequencies were 0.3 ± 1.1 (0.4 ± 1.4 and 0.2 ± 1 in Africans and Europeans, respectively).



Supplementary Figure 4 Distribution of the percentages of discordant CNVs in all the 1,252 samples shared by the CIDR and SAGE datasets. The vertical line represents the average of 7.1% (i.e., a CIDR sample had an average of 7.1% discordant/more CNVs than a SAGE samples). The distribution is Gaussian, suggesting that there is no directional bias regarding the CNV calling between the two datasets. Discordance was measured as the difference of CNVs in the same sample from SAGE and CIDR, divided by the maximum number of CNVs that the sample had between the two datasets.

Chapter 4.2: VIpower: power analysis for viral integration detection using next-generation sequencing

-- A novel tool for viral integration detection

Arvis Sulovari[1] and Dawei Li[1,2,3*]

[1]*Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA*

[2]*Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA*

[3]*Neuroscience, Behavior, and Health Initiative, University of Vermont, Burlington, Vermont 05405, USA*

*To whom correspondence should be addressed:

Dawei Li, Ph.D., Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405, USA. E-mail: dawei.li@uvm.edu

Number of words in the abstract: 217

Number of words in the text (excluding declarations, legends, and references): 1,872

Number of figures: 3

Number of supplementary materials: 2 supplementary Tables and 10 supplementary Figures and Legends.

Abstract

Viral integrations have been associated with many human diseases. Next-generation sequencing (NGS) allows for accurate detection of novel viral sequences integrated into the human genome. However, the experimental factors influencing the detection power of viral integration events using NGS data have not been well-characterized. We designed a novel viral integration detection framework, including *in silico* generation of paired-end human and viral sequencing reads, alignment to the human and viral reference genomes, and detection of viral integration events. A total of 15 selected key molecular and bioinformatics factors were incorporated into the power calculation. We found that the power for detecting viral integration events was significantly associated with six molecular and bioinformatics factors ($P < 2\times10^{-16}$), including the proportion of cells with viral integrations (Pearson's $\rho = 0.64$), sequencing depth ($\rho = 0.37$), viral integration length ($\rho = 0.37$), NGS insert size (0.23), minimum number of supporting reads required to determine a viral integration ($\rho = -0.19$), and read length ($\rho = -0.09$). We developed VIpower for accurate and fast estimation of viral integration detection power. To detect viral integration events in the human genomes, we have designed VIpower to guide NGS library preparation, sequencing experiments, and bioinformatics analyses. The tool can be used in the general population and disease cohort or germline and somatic scenarios. VIpower is available as user-friendly web interface and command-line application (www.uvm.edu/genomics/software/VIpower).

Importance

Viral etiologies have been speculated in various human diseases. Next-generation sequencing (NGS) allows for the detection of viral sequences integrated into the human genome. However, accurate identification of viral integrations remains challenging due to limited knowledge on how to better design NGS experiments and analyze the resulting data for viral integration identification. This study, for the first time, addresses these gaps in knowledge. Through a large amount of simulation and empirical data, we evaluated the key factors for experimental designs as well as bioinformatics analyses for viral integration detection. The results from this study, including the power calculation tool, allow investigators to design better NGS experiments for conducting viral integration screening in various disease samples. Additionally, in a separate study (manuscript in preparation), we have applied our approach to several disease cohorts and successfully identified (and validated) viral integrations in both germline and somatic scenarios.

Keywords: Next-generation sequencing (NGS), Viral etiology, Viral integration (VI), Power analysis

Introduction

Viral etiology has been speculated in various human diseases, such as cancers[267,268], amyotrophic lateral sclerosis[269], Alzheimer's disease[270], chronic fatigue syndrome[271], type I diabetes[272,273], Crohn's disease[274], and asthma[275]. Many infectious viruses are able to insert their genetic material into host chromosomes[276-279], and the resulting viral integrations may play roles in disease pathogenesis and development by disrupting or dysregulating gene functions. Use of next generation sequencing (NGS) allows for the discovery of viral integrations (i.e., virus-human-virus sequences) in both somatic and germline cells[277]. However, accurate identification of viral integrations in the human genome remains challenging due to limitations of the available bioinformatics methods[280-284] and insufficient empirical data to guide experimental designs of viral integration detection and related data analyses. To accurately capture novel viral sequences integrated in the human genome, systematic research is required to determine the key molecular and bioinformatics factors that affect the power to detect viral integrations.

In this study, we have carefully evaluated 15 selected key molecular and bioinformatics factors related to viral integration detection, and found six factors that was significantly associated with the viral integration detection power. We further developed the first tool for accurate and fast estimation of detection power of viral integrations for public use. The results and tool from this study allow biologists and physicians to design NGS

experiments for conducting virome-wide viral integration screening in various human disease and healthy samples.

Results

We identified a total of 15 key molecular and bioinformatics factors that were important to NGS-based viral integration detection. We first designed a viral integration detection framework, and then, developed an implementation pipeline. Based on this pipeline, we further developed a novel computational tool, VIpower, to estimate the viral integration detection power.

To identify the molecular and bioinformatics factors that significantly influence viral integration detection power, we ran VIpower to estimate detection power for various expected values of the 15 key factors (a total of 23,040 combinations). We found that six factors were significantly associated with detection power (**Figure 2**), including cellular proportion (Pearson's $\rho = 0.64$ and $P < 2\times10^{-16}$), sequencing depth ($\rho = 0.37$ and $P < 2\times10^{-16}$), length of integrated viral sequence ($\rho = 0.37$, $P = 1\times10^{-13}$), insert size ($\rho = 0.23$ and $P < 2\times10^{-16}$), minimum number of supporting reads required (threshold) to determine viral integration event ($\rho = -0.19$ and $P < 2\times10^{-16}$), and read length ($\rho = -0.09$ and $P < 2\times10^{-16}$ when the total data volume/sequencing depth was fixed; $\rho = 0.1$ and $P < 2\times10^{-16}$ when the total read number was fixed). The first molecular factor, cellular proportion, is particularly relevant when sequencing a heterogeneous population of cells, such as cancer

biopsies[285]. Additionally, we observed marginal association with minimum mappable length (ρ = -0.02 and P = 0.0003). **Figure 3** shows the pairwise correlations among all these seven molecular and bioinformatics factors, numbers of supporting (chimeric and split) reads, and the resulting detection power. As expected, the observed numbers of supporting reads were strongly associated with detection power. **Supplementary Figure 7** shows the distributions of supporting reads and threshold to determine viral integration events. Moreover, we compared the detection power of rare and common viral integrations, and found no evidence of significant difference (r^2 = 0.96; **Supplementary Figure 8**), implying the feasibility to study the roles of rare viral integration events in the etiologies of human diseases.

We compared the power estimates from our viral integration detection framework with those from Virus-Clip[286] for each of the six significant factors. We found our framework consistently showed higher power (**Supplementary Figure 9**). Our framework uses both split and chimeric reads to detect viral integrations while Virus-Clip uses split reads only. It should also be noted that our framework detects multiple viruses simultaneously (such as virome-wide) while Virus-Clip, like other similar tools, only detects one virus at a time.

VIpower is available as a user-friendly web interface for live runs of power analyses (www.uvm.edu/genomics/software/VIpower/live). Users can also query the precomputed power estimates (**Supplementary Table 2**). This tool is also available as a Linux

command line version where advanced users may calculate power for other NGS scenarios by modifying the reference files, such as the viral integration profile, distance to repeats, and distribution of GC content-specific PE read coordinates.

Discussion

VIpower is the first viral integration detection power calculator. It can be used to guide NGS experimental designs and data analyses. Using VIpower, we have identified six factors significantly associated with the detection power. Compared to use of only slit reads, use of both chimeric and split reads, as used by VIpower, increased the detection power. VIpower also allows for testing of complex interactions among the key molecular and bioinformatics factors. For instance, when the sequencing read length increased from 100 bp to 300 bp (the total sequence volume was fixed), the number of total supporting reads decreased by an average of 37%; however, the proportion of split reads increased 4.7 fold (**Supplementary Figure 10**). This design may be beneficial for more precise mapping of integration breakpoints. Because it stores and processes viral integration information by genomic features, instead of actual sequences, VIpower has a very short runtime. For example, each of our simulations (**Supplementary Table 2**) can be completed by one standard laptop in an average of nine seconds (range from 0.6 to 62 seconds). Similarly, the live web interface can conduct a power calculation within one minute. A limitation of this study was that the empirical viral integrations were derived from the clinical HBV integrations. However, VIpower allows replacement of the viral

integration references to any viruses or a combination of them. This makes it possible to conduct virome-wide viral integration screens of various human samples. We will update the VIpower viral integration references as soon as additional data becomes available.

To conclude, we developed a fast computational framework to detect virome-wide viral integrations in the human genome, and validated six key molecular and bioinformatics factors significantly associated with the detection power. The results in this study provide the fundamental guidance to the NGS-based experimental designs and data analyses of viral etiological studies of various human diseases.

Methods

The detection of viral integration events was implemented in four modules (**Figure 1**), including modules 1 and 2: the simulation of virtual human and viral sequences, respectively; module 3: the simulation of paired-end (PE) sequencing reads and *in silico* alignment of the reads to the human and viral reference genomes; and module 4: the detection of viral integration events and power calculation. The whole-genome empirical distributions of four features, including GC content, length of repeat region, characteristics of known viral integrations (e.g., location and distance to repeat region), and GC-specific Illumina PE read positions[287], were used for the simulation of viral integrations (**Supplementary Figure 1**).

Human sequence simulation

The human sequences were simulated according to the whole-genome distributions of empirical GC content (**Supplementary Figure 2**) and repeat regions (**Supplementary Figure 3**). The GC content was calculated employing 200 base pair tiling windows using the human reference genome (Genome Browser, GRCh37/hg19)[287]. The lengths and frequencies (17 repeats/10,000 bp) of repeat regions were extracted from RepeatMasker[288]. The whole-genome distributions of the two features were randomly sampled with replacement, and assigned to our simulated human sequences.

Viral integration simulation

The viral integration events were simulated based on the properties of known viral integrations. The lengths of viral integrations were created based on the widely-studied and validated Hepatitis B virus (HBV) integrations maintained in the dr.VIS database[289]. The locations of the viral integrations were assigned according to the distances between the known viral integration sites and repeat regions provided by RepeatMasker (**Supplementary Figure 4**).

In silico read alignment

Each PE read was assigned physical coordinates according to the empirical distribution of sequencing depth by GC content (**Supplementary Figure 5**), which was generated using

known Illumina PE read counts measured by 200 bp tiling windows across the human genome[287]. To remove low quality reads, several commonly-used quality control procedures were employed, including minimum mappable read length, read trimming, PCR duplicate removal, and non-uniquely mapped read removal (**Supplementary Table 1**). In a simulated example with commonly-used NGS parameters, the quality controls removed low quality reads, particularly those mapped to regions with very high sequencing depth (**Supplementary Figure 6**). All of the remaining PE reads were further aligned to the hybrid human and viral reference genome. For somatic viral integration events, we adjusted the number of reads in the integrated viral sequence region to match the corresponding cellular proportion.

Viral integration detection and power analysis

Each PE read was labelled either chimeric or split when one entire read or a portion of a single read mapped to the viral reference genome, respectively, while the remaining portion mapped to the human genome. Both split and chimeric reads were used as supporting evidence to determine viral integration events. The power to detect viral integrations is defined as:

$$\text{Detection power}\ (\%) = \frac{\text{Number of identified viral integrations}}{\text{Number of simulated viral integrations}} \times 100$$

Identification of factors associated with detection power

Pearson's correlation test was used to measure the association between detection power and each of the key molecular and bioinformatics factors (**Supplementary Table 2**). The statistical significance threshold was adjusted for the number of multiple tests using Bonferroni correction, resulting in $P < 0.0001$.

Evaluation of viral integration detection framework

We compared the power of our viral integration detection framework with an existing viral integration detection tool Virus-Clip[286]. First, we randomly selected 100 sequences of equal lengths from the HBV reference sequences and inserted into randomly-selected positions of human chromosome 22 (hg19). This process was repeated with viral integration lengths of 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 500, and 1,000 bp, and the resulting sequences were stored in FASTA format. Second, these FASTA files were used to generate PE sequencing reads (i.e., FASTQ format) of library designs with varying sequencing depths (1, 2, 4, 6, 8, 10, 20, and 40 fold), read lengths (75 and 100 bp), and insert sizes (600, 1,300, and 2,200 bp) using pIRS[287]. Third, we ran these FASTQ files to detect HBV integrations using Virus-Clip. As Virus-Clip was designed to use split reads only, we tested our framework by using split and chimeric reads as well as split reads only. Three replications, each corresponding to different HBV sequences and integration breakpoints, were carried out. The average detection powers were compared between the two approaches using in-house R scripts.

Web application

The source code was written primarily in R (version 3.3.0). The web interface was designed using HTML and PHP (version 5.3.3) scripts. MySQL was used to store pre-computed power estimates.

Availability of data and software

The web application can be accessed at www.uvm.edu/genomics/software/VIpower/live, or downloaded for command-line application at www.uvm.edu/genomics/software/VIpower/downloads. The database of results presented here can be accessed at www.uvm.edu/genomics/software/VIpower.

The datasets supporting the conclusions of this article are included within the article and its additional files.

Acknowledgements

This work was supported by the Start-up Fund of The University of Vermont. We would like to thank Dr. Xun Chen for his critical comments and feedback, and Michael Mariani for his help with the website design.

Authors' contributions: DL and AS conceived and organized the project. DL supervised the project. AS wrote the source code and conducted the analyses. AS and DL wrote the manuscript. Both authors read and approved the final manuscript.

Conflict of Interests: The authors declare no potential competing interests.

References

1. **Sung, W.K., Zheng, H., Li, S., Chen, R., Liu, X., Li, Y., Lee, N.P., Lee, W.H., Ariyaratne, P.N., Tennakoon, C. *et al.*** 2012. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nature genetics, **44**, 765-769.
2. **Khoury, J.D., Tannir, N.M., Williams, M.D., Chen, Y., Yao, H., Zhang, J., Thompson, E.J., Network, T., Meric-Bernstam, F., Medeiros, L.J. *et al.*** 2013. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. Journal of virology, **87**, 8916-8926.
3. **Douville, R., Liu, J., Rothstein, J. and Nath, A.** 2011. Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. Annals of neurology, **69**, 141-151.
4. **Carbone, I., Lazzarotto, T., Ianni, M., Porcellini, E., Forti, P., Masliah, E., Gabrielli, L. and Licastro, F.** 2014. Herpes virus in Alzheimer's disease: relation to progression of the disease. Neurobiol Aging, **35**, 122-129.

5. **Mikovits, J.A., Lombardi, V.C., Pfost, M.A., Hagen, K.S. and Ruscetti, F.W.** 2009. Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. Virulence, **1**, 386-390.

6. **Smyth, D.J., Cooper, J.D., Bailey, R., Field, S., Burren, O., Smink, L.J., Guja, C., Ionescu-Tirgoviste, C., Widmer, B., Dunger, D.B. *et al.*** 2006. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nature genetics, **38**, 617-619.

7. **Foxman, E.F. and Iwasaki, A.** 2011. Genome-virome interactions: examining the role of common viral infections in complex disease. Nat Rev Microbiol, **9**, 254-264.

8. **Karst, S.M., Wobus, C.E., Lay, M., Davidson, J. and Virgin, H.W.** 2003. STAT1-dependent innate immunity to a Norwalk-like virus. Science, **299**, 1575-1578.

9. **Gern, J.E.** 2009. Rhinovirus and the initiation of asthma. Curr Opin Allergy Cl, **9**, 73-78.

10. **Klenerman, P., Hengartner, H. and Zinkernagel, R.M.** 1997. A non-retroviral RNA virus persists in DNA form. Nature, **390**, 298-301.

11. **Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J.M. *et al.*** 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature, **463**, 84-87.

12. **Belyi, V.A., Levine, A.J. and Skalka, A.M.** 2010. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. PLoS Pathog, **6**, e1001030.

13. **Taylor, D.J. and Bruenn, J.** 2009. The evolution of novel fungal genes from non-retroviral RNA viruses. BMC Biol, **7**, 88.

14. **Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J.M. *et al.*** 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature, **463**, 84-U90.

15. **Wang, Q.G., Jia, P.L. and Zhao, Z.M.** 2015. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome medicine, **7**.

16. **Wang, Q.G., Jia, P.L. and Zhao, Z.M.** 2013. VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data. PloS one, **8**.

17. **Chen, Y.X., Yao, H., Thompson, E.J., Tannir, N.M., Weinstein, J.N. and Su, X.P.** 2013. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics, **29**, 266-267.

18. **Ho, D.W.H., Sze, K.M.F. and Ng, I.O.L.** 2015. Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. Oncotarget, **6**, 20959-20963.

19. **Katz, J.P. and Pipas, J.M.** 2014. SummonChimera infers integrated viral genomes with nucleotide precision from NGS data. Bmc Bioinformatics, **15**.

20. **Meyerson, M., Gabriel, S. and Getz, G.** 2010. Advances in understanding cancer genomes through second-generation sequencing. Nature reviews. Genetics, **11**, 685-696.

21. **Ho, D.W., Sze, K.M. and Ng, I.O.** 2015. Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. Oncotarget, **6**, 20959-20963.

22. **Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., Chen, Y., Mu, D., Zhang, H., Li, N. *et al.*** 2012. pIRS: Profile-based Illumina pair-end reads simulator. Bioinformatics, **28**, 1533-1535.

23. **Tarailo-Graovac, M. and Chen, N.** 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], **Chapter 4**, Unit 4 10.

24. **Yang, X., Li, M., Liu, Q., Zhang, Y., Qian, J., Wan, X., Wang, A., Zhang, H., Zhu, C., Lu, X. *et al.*** 2015. Dr.VIS v2.0: an updated database of human disease-related viral integration sites in the era of high-throughput deep sequencing. Nucleic acids research, **43**, D887-892.

Figure Legends

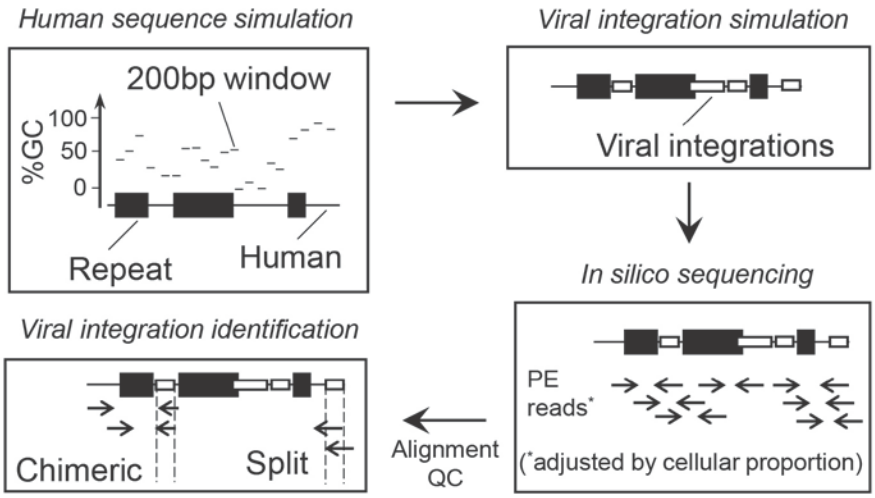


Figure 1 Overview of the VIpower flow diagram. The simulation and detection of viral integrations in the human genome are composed of four modules. The first two modules simulate features of human and viral sequences; while the last two align PE reads to the human and viral reference sequences and detect viral integration events.

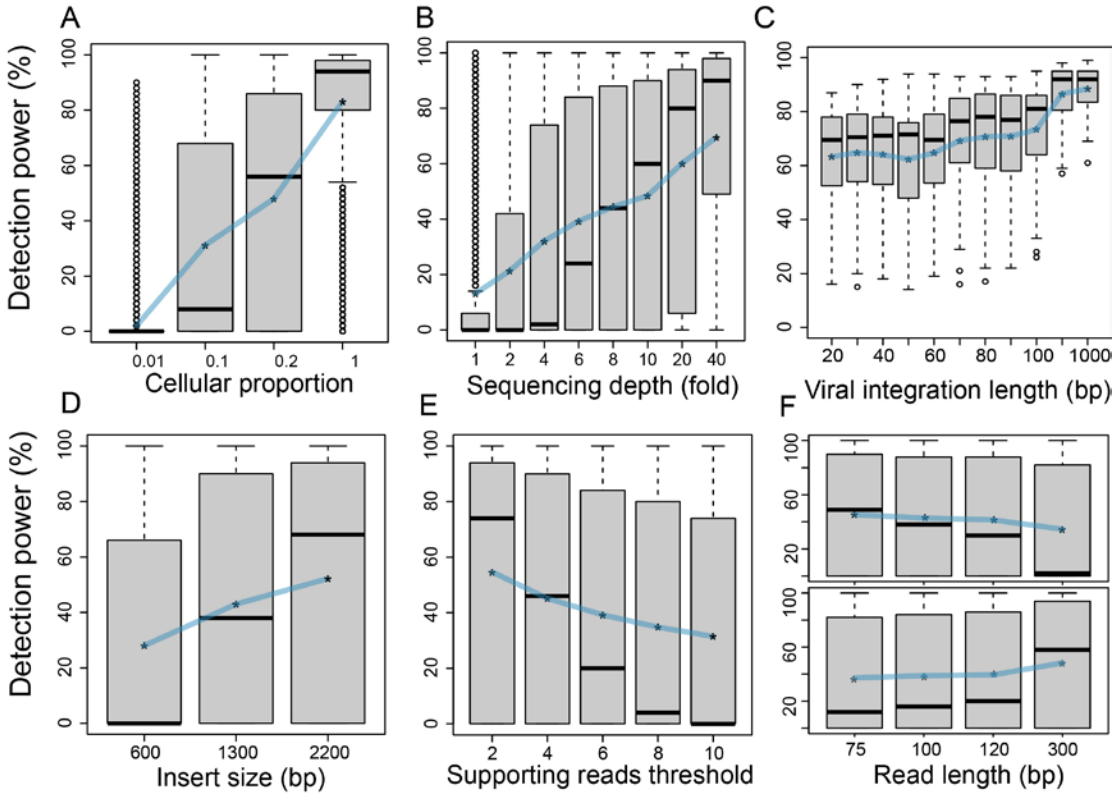


Figure 2 Six factors significantly associated with viral integration detection power. The six factors are ordered by significance level of correlation. The box plots indicate five quantiles, and the star symbol (*) represents the average value. The correlation coefficients ρ and *P* values for each factor were (A) cellular proportion ($\rho = 0.64$, $P < 2\times10^{-16}$), (B) sequencing depth ($\rho = 0.37$, $P < 2\times10^{-16}$), (C) viral integration length ($\rho = 0.37$, $P = 1\times10^{-13}$) (D) insert size ($\rho = 0.23$, $P < 2\times10^{-16}$), (E) supporting reads threshold ($\rho = -0.19$, $P < 2\times10^{-16}$), (F) read length (the top panel represents a scenario where the sequencing depth is fixed, $\rho = -0.09$, $P < 2\times10^{-16}$; the bottom panel shows represents a scenario where the read number is fixed, $\rho = 0.1$, $P < 2\times10^{-16}$) , respectively. In each box plot, all other involved variables were simulated in equal proportion of representation to ensure balanced comparisons among data points.

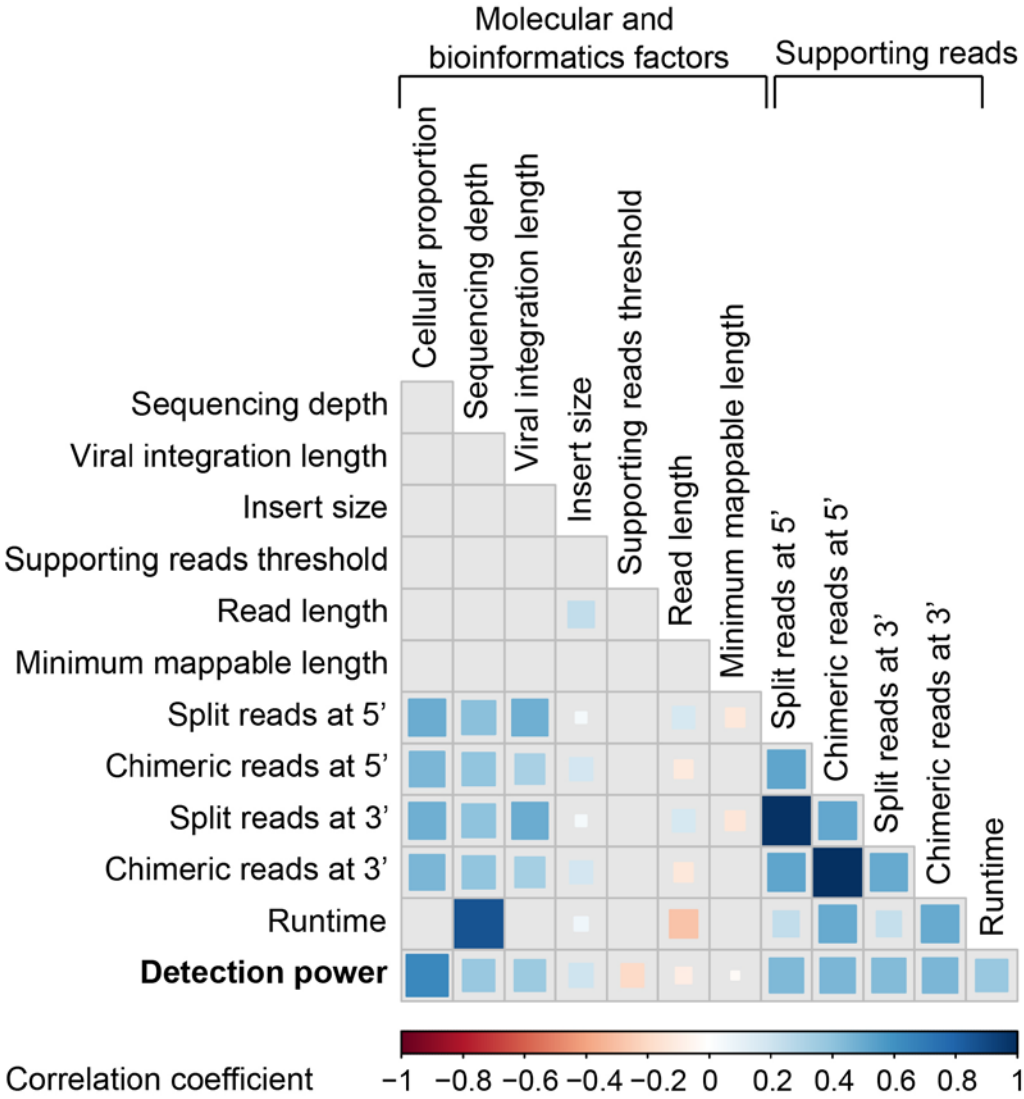


Figure 3 Pairwise correlations of detection power with key molecular and bioinformatics factors. The color of each square corresponds to correlation coefficient ρ (darker color corresponds to stronger correlation) while the size corresponds to the *P* value (smaller *P* value corresponds to bigger square size). The six significant factors ($P \leq 0.0001$), ordered by their correlation coefficient with detection power, are cellular proportion, sequencing depth, viral integration length, insert size, supporting reads threshold, and read length. All parameters represent their average values, except minimum mappable length, cellular proportion, runtime, and detection power.

Supplementary Tables and Figures

Supplementary Tables

Supplementary Table 1 List of quality control procedures implemented in VIpower

Quality control procedure	Default value	Note
Minimum mappable length*	20 bp	Required minimum read length mappable to either human or viral genome
Trim reads*	0.1	A proportion of the 3' end of a read to be trimmed
Remove PCR duplicates	Yes	Reads with identical coordinates are removed.
Remove non-uniquely aligned PE reads	Yes	Reads aligned in repeat regions (< 20 bp in non-repeat regions) are removed.

*, The parameter can be changed by users.

Supplementary Table 2 Key molecular and bioinformatics factors and reference files used by VIpower

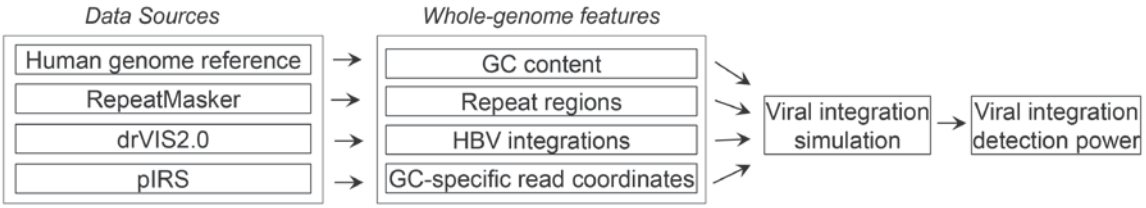
Key factor	Values	Description
Cellular proportion	0.01; 0.1; 0.2;1	Proportion of cells with viral integrations (e.g., germline, 1 and somatic, <1)
Human sequence length	1,000,000	Total sequence length, including integrated viral sequences (bp)
Number of viral integration events	50	Number of viral integration events
Length of integrated viral sequences (mean)	500	Average length of viral integrations (bp)

Length of integrated viral sequences (standard deviation)	5,000	Standard deviation of lengths of viral integrations (bp)
Length of integrated viral sequences (minimum)	10; 50; 200	Minimum length of viral integrations (bp)
Sequencing depth	1; 2; 4; 6; 8; 10; 20; 40	Sequencing depth (fold or X)
Read length	75; 100; 120; 300	Read length (bp)
Insert size (mean)	600; 1,300; 2,200	Average of insert size (bp)
Insert size (standard deviation)	read_insert_mean/20[287]	Standard deviation of insert size (bp)
Supporting reads required	2; 4; 6; 8; 10	Required minimum number of supporting (chimeric and split) reads
Minimum mappable length	20; 40	Minimum read length uniquely mapped to either human or viral reference
Reads in repeat regions (proportion)	0.05	Proportion of reads completely mapped inside repeat regions (whole-genome).
Reads to trim (proportion)	0.05	Proportion of number of reads that are trimmed
Nucleotides of a read to trim (proportion)	0.15	Proportion of number of nucleotides (of a read) that is trimmed
seed_value	[random]	Simulation seed (for reproducible results)
Repeat regions	[matrix]	Repeat sequence distribution (~5.2 million repeat regions from RepeatMasker)
GC content	[matrix]	GC content distribution specific to the human genome

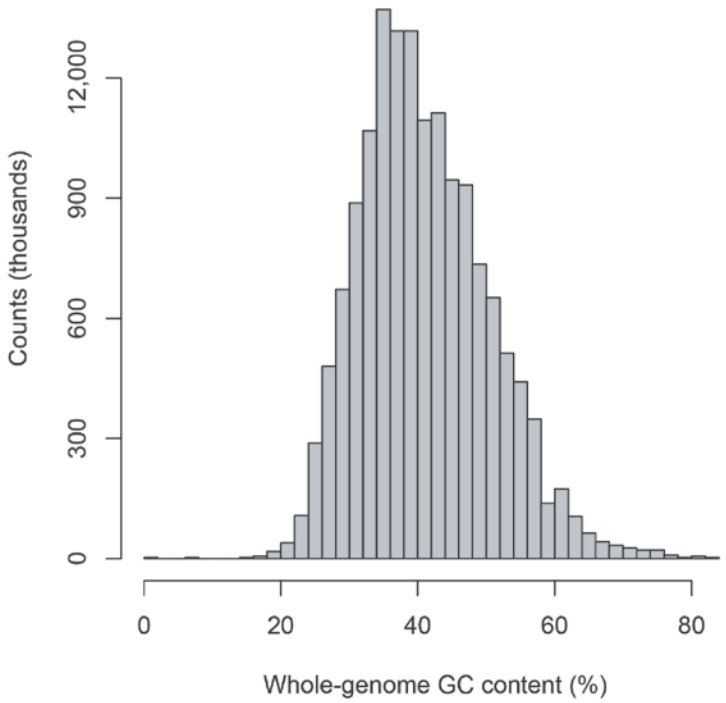
A total of 23,040 unique combinations of the listed values involving 15 molecular and bioinformatics factors were used to measure their correlations with detection power.

Additionally, two reference files, i.e., repeat sequence information and GC content distributions, can also be modified by users.

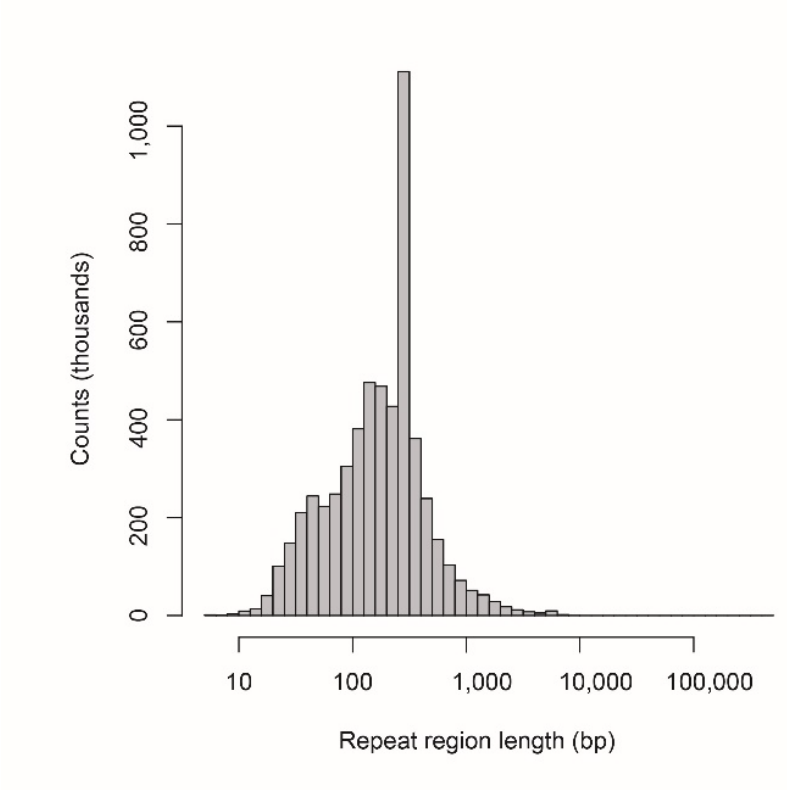
Supplementary Figures



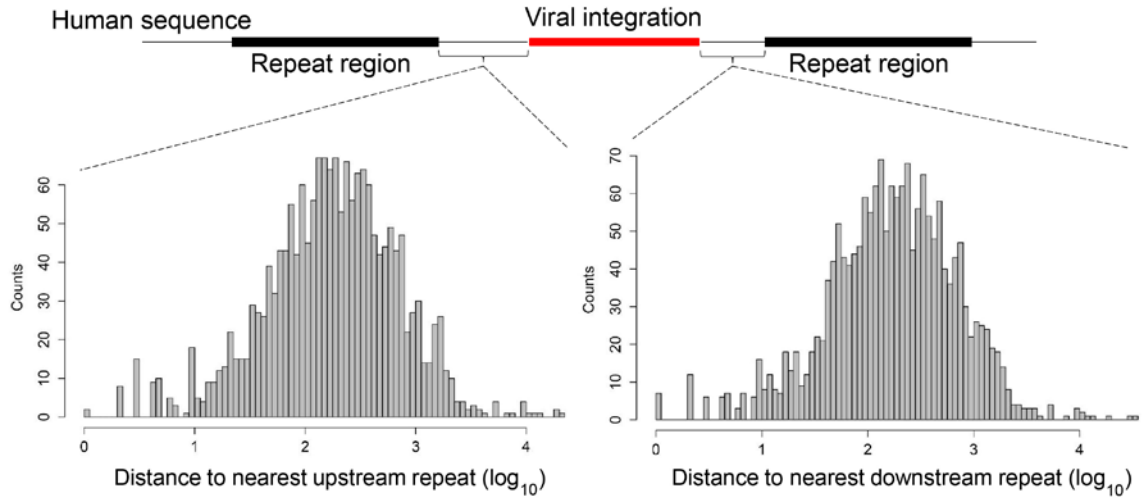
Supplementary Figure 1 Empirical features and data sources included in the simulation of viral integration events.



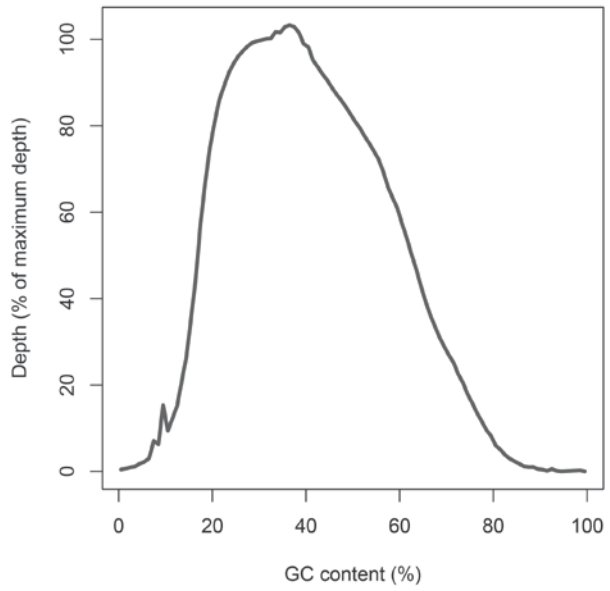
Supplementary Figure 2 Whole-genome distribution of GC content. The whole-genome GC content values were binned by 200 bp tiling windows, and then used to draw the distribution. Each of our simulated human sequences was assigned a GC content value according to this distribution.



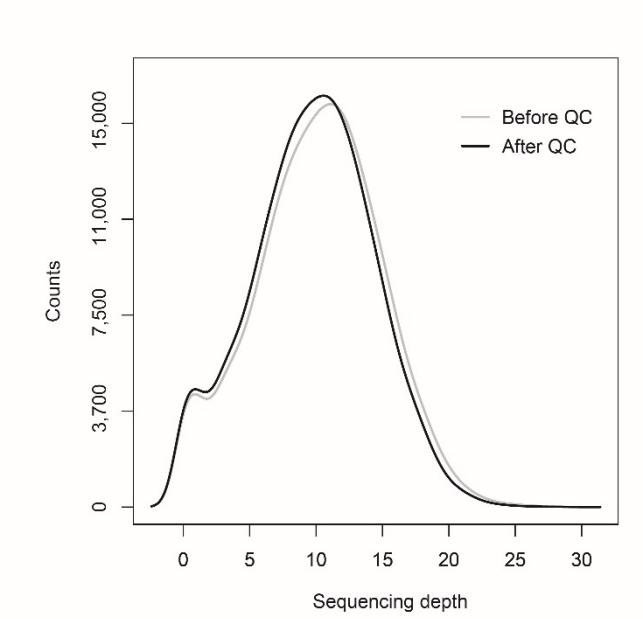
Supplementary Figure 3 Whole-genome distribution of lengths of repeat regions. The ReapeatMasker (hg19 version) was used to extract over 5.2 million repeat regions. This empirical distribution was randomly sampled to assign repeat region characteristics, i.e., location, to our simulated human sequences.



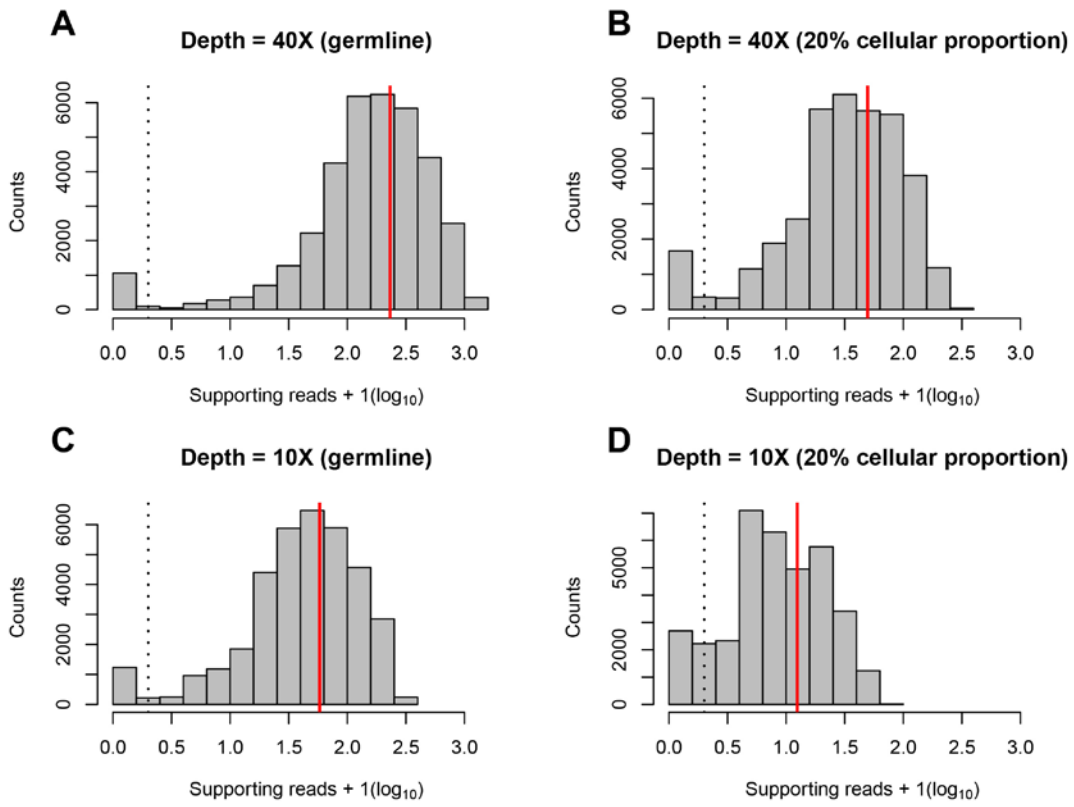
Supplementary Figure 4 Empirical distribution of repeat regions around known viral integration sites. The RepeatMasker database (hg19 version) was used to build a reference of human repeat regions. The distributions of these repeat regions were further used in our simulation of human sequences. Distances from upstream (left) and downstream (right) of integrated viral sequence to the nearest repeat region were measured separately.



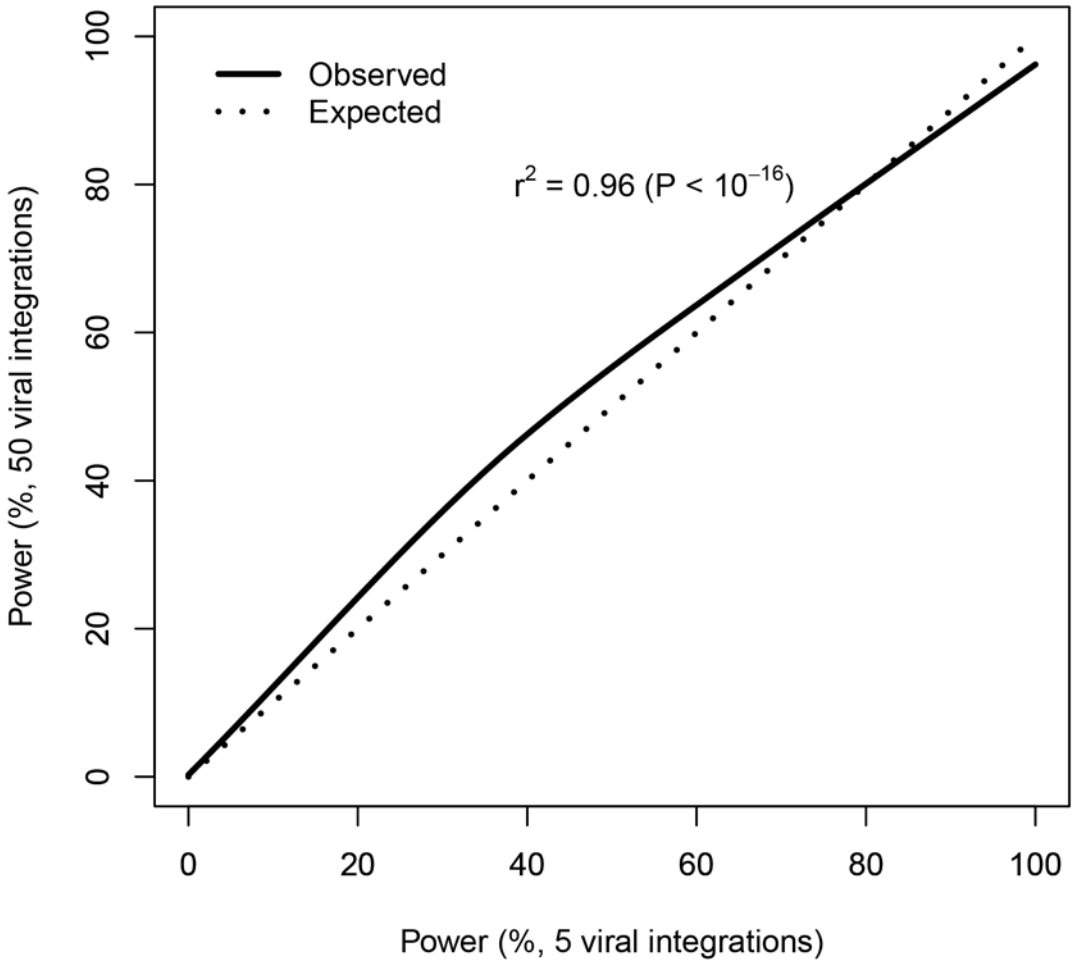
Supplementary Figure 5 Influence of GC content on sequencing depth. The whole-genome GC content (in 200 bp tiling windows) and sequencing depth were calculated based on a ~30X paired-end sequencing data[287]. This distribution was converted into a probability distribution function to determine sequencing depth for each 200 bp tiling window of the simulated human sequence.



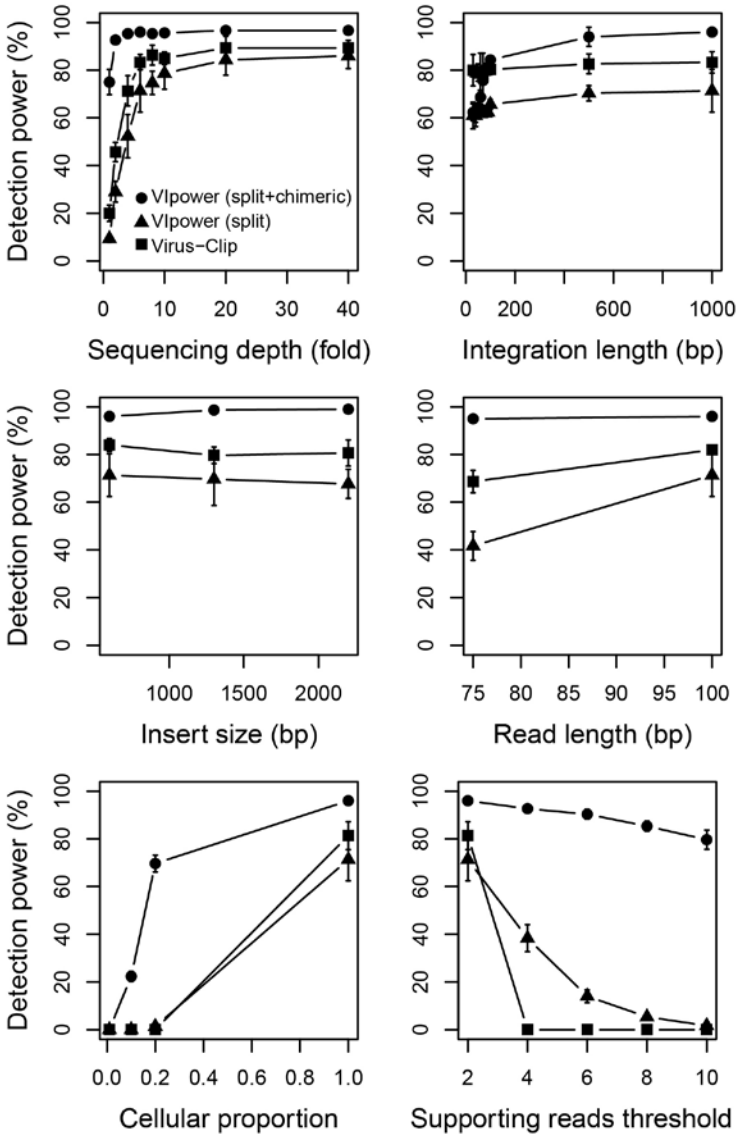
Supplementary Figure 6 Distributions of mapped read depth before and after quality controls. We simulated a commonly-used sequencing library design: read length = 100 bp, insert size = 500 ± 25 bp, and average sequencing depth = 10. The resulting 10,000 simulated PE reads were mapped to a 200,000 bp human region. The sequencing depth distributions were plotted before and after quality controls according to the quality control procedures described in **Supplementary Table 1**. For example, for the regions with depth >13 (75 percentile), the total sequence volume decreased by 17% after quality controls, demonstrating the effectiveness of the quality control procedures.



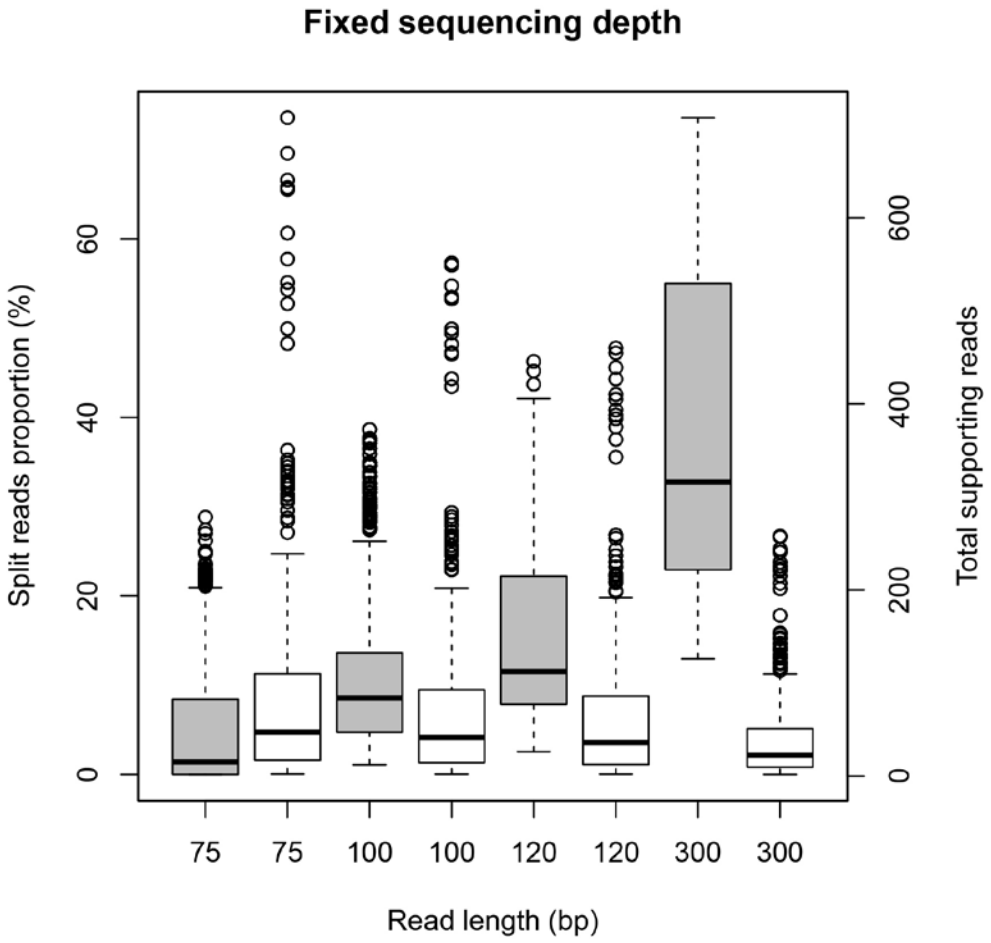
Supplementary Figure 7 Distribution of sequencing depth at viral integration breakpoints. The expected (red vertical line) and observed (grey bars) numbers of supporting PE reads of all simulated viral integration sites are shown for different sequencing depths and cellular proportions: (A) depth of 40X in germline viral integrations, (B) depth of 40X in somatic viral integrations (20% cellular proportion), (C) depth of 10X in germline integrations, and (D) depth of 10X in somatic integrations (20% cellular proportion). The dotted line represents a threshold of two supporting reads, which is one of the thresholds used in our detection; in this case, only viral integrations to the right of the dotted line are considered successfully detected. The expected number of supporting reads was calculated as: (insert size) × (number of reads) / sequence length. The distributions of numbers of supporting reads were consistent for different sequencing depths and cellular proportions.



Supplementary Figure 8 Comparison of detection power for common and rare viral integrations. Five and 50 viral integration events were simulated into a one million bp human sequence to represent (relatively) rare and common viral integrations, respectively. For each case, a total of 23,040 unique input combinations of the 15 factors (**Supplementary Table 2**) were used to estimate power. The power estimates in both cases were highly correlated (r^2 = 0.96; the expected line corresponds to the perfect correlation of r^2 = 1), indicating no significant difference in detection power between common and rare viral integrations.



Supplementary Figure 9 Evaluation of our viral integration detection framework. We compared the power of our viral integration detection framework with Virus-Clip. The overlap between predicted HBV (RefSeq access: NC_003977.2) integration positions and actual positions was used to calculate the detection power. Each plot corresponds to one of the six significant factors, while keeping the other factors fixed (sequencing depth = 6X, viral integration length = 1,000bp, insert size = 600bp, read length =100, cellular proportion = 1, and supporting reads threshold = 2). The three curves in each plot correspond to VIpower based on both split and chimeric reads (circle), VIpower based on split reads only (triangle), and Virus-Clip (square), respectively.



Supplementary Figure 10 Balance between integration breakpoint precision and detection power. The open and solid bars represent the total supporting reads, and proportion of split reads, respectively (all viral integrations were assumed to be germline). Based on the existing viral integrations from our 23,040 simulations (**Supplementary Table 2**), which combined PE reads from various sequencing library designs, under the assumption of fixed total sequence volume (sequencing depth), when read length increases, the total number of supporting reads decreases (**Figure 3**), however, both the proportion and actual number of split reads per viral integration increases.

CHAPTER 5: CONCLUSIONS

The work presented in this dissertation describes three integrated approaches for elucidation of brain disorders: evolutionary genomics (chapter 2), bioinformatics tools and resources (chapter 3) and, identification and disease association of structural genomic aberrations (chapter 4). Below we discuss each of these contributions individually.

In chapter 2 we demonstrated that a large number of genes, diseases, and traits are influenced by functional SNPs with extreme allele frequency differences (EAFD) between populations of the same continental origin. Some of the identified phenotypes included brain disorders, such as ADHD, frontotemporal dementia, white matter hyperintensity burden, alcohol consumption and drinking behavior. Future studies may demonstrate that indeed, a considerable portion of the genetic missing heritability in these complex brain disorders is attributed to EAFD.

Next, we found that light eye color was significantly associated with AD; an association which held true after controlling for population stratification and socio-economic factors. This finding supports the idea that selection forces may have indirectly acted on AD risk loci. Our findings complement the existing research on the connection between eye color and mental illnesses and behavioral problems. Our study is the first to report an association between blue eye color and AD in EAs using clinically-ascertained subjects and a moderate sample size. Our findings indicate that the selection pressures acting on the genetics of pigmentation might have implications for AD susceptibility. Thus, integration of population-phenotype and gene and network analyses is helpful for

the identification of risk factors in AD, and a broad range of mental illnesses, in general. While replication is needed, our findings suggest that eye pigmentation information may be useful in the future research of AD and related alcohol consumption behaviors. Further characterization of this association may unravel novel etiological factors in alcohol addiction.

Findings presented in chapters 2.1-2.3 support the idea that positive selection may increase disease risk, a hallmark of antagonistic pleiotropy. Importantly, this mechanism is central to the theory of aging proposed by G. C. Williams in 1957, who observed that while high *p53* gene activity (as a tumor suppressor) increased fitness early in life, it also led to increased aging-related disorders later in life (i.e., cellular senescence). Since the detrimental health effects occurred after reproductive age, negative selection would not be effective at removing the *p53* alleles from the population. A well-known pleiotropic functional variant is located in the *p53* gene, causing a Proline (Pro) to Arginine (Arg) amino acid change in residue 72. The Pro/Pro carriers were found to be at higher risk of developing cancer than the Arg/Arg carriers, by 2.54 fold[290]. However, the Pro/Pro carriers had a 41% increased longevity[291]. Thus, this *p5*3 variant protects from cancer at a cost of shorter life span. Similarly, in chapter 2.1 we present other examples of antagonistic pleiotropy, derived from the GWAS catalogue, such as adaptation traits (skin color, or eye color) and Melanoma, or height and psoriasis.

Since selection drives emergence of common allele frequencies with strong effect on phenotype, the statistical power for detecting these variants through GWAS is higher than it would be for neutral variants. Indeed, Sabeti and colleagues used the GWAS

catalogue to demonstrate that GWAS of variants under positive selection had a smaller association p-value than the rest of disease-associated variants in the catalog[110]. More recently, Scott Williams and colleagues demonstrated that populations with extreme disease resistance in the face of extensive pathogen exposure can increase the statistical power to detect associations with complex human diseases[292]. Similarly, Rasmus Nielsen and colleagues demonstrated that Greenlandic Inuit populations have had positive selection for genetic variants involved in omega-3 polyunsaturated fatty acids metabolism; thus, when this population was leveraged in a GWAS, novel fatty-acid metabolism risk loci were discovered[293]. Thus, whole-genome association studies hold a promise for discovery of disease-associated loci, particularly in populations where disease genes are expected to be under selection.

In the post-GWAS era, the genetic etiology of brain diseases and other complex human diseases will likely be surveyed under the lens of rare variants and by leveraging multi-ethnic cohorts. Thus, we developed a new resource for genetic association analysis of multi-ethnic cohorts (chapter 3.1) and a tool to improve accuracy of inferring unassayed alleles in microarray data (i.e., genotype imputation) (chapter 3.2).

Tools and resources for rare genomic variants

To enable rapid discovery of disease-associated variants, particularly when using a multi-ethnic or other complex population structures, we constructed a panel of AIMs to control for population structure (chapter 3.1). The constructed AIMs panels were highly

informative for ancestry, as measured by I_N. For example, among the top 12 AIMs of a recently published Han Chinese panel[193], four overlapped with our panel of the equivalent population pairs, i.e., CHB-CHS; however, our panel contained a larger number of high I_N markers, i.e., 192 AIMs with $I_N \geq 0.028$ in our panel compared to only two in the published panel. A detailed comparison between our panel and the one published by Qin et al. revealed that our panel has more informative markers, as measured by both F_{ST} and I_N statistics.

We recommend using these panels hierarchically. For instance, a study that analyzes samples of African and East Asian ancestry may first use one or more of our AIMs panels that were designed for separating Africans from East Asian populations, then use the panels that separate specific African populations from one another, and those that separate specific East Asian populations from one another. This strategy prevents inclusion of AIMs designed for populations that are not represented in the underlying study. To the best of our knowledge, this study provides the first set of AIMs panels that can ascertain sample ancestry or admixture proportion with high accuracy at multiple resolutions, i.e., global, continental, population, and sub-population levels. These panels would be particularly useful in two scenarios: target sequencing studies where whole-genome data is not available to extract AIMs, and GWAS of complex population structures (e.g., multiethnic samples).

In chapter 3.2 we introduce a tool that improves quality of genotype imputation, and accuracy of downstream association analyses or meta-analyses. Importantly, we found that approximately 600 thousand well-typed SNPs are likely to suffice for high quality genome-wide imputation of rare SNPs. Inconsistent allele definitions and genome builds or incorrect conversions lead to incorrect genetic association "findings". In this chapter, we developed a comprehensive tool, GACT, with both powerful command-line and user-friendly web interface versions to predict, and convert both genome builds and allele definitions between multiple GWAS (or deep sequencing) genotype data, which is required for all imputations and genome-wide meta-analyses. GACT will ease a broad use of the GWAS data from the dbGaP and other publicly available genotype repositories for large-scale secondary analyses and multi-laboratory collaborations in the genetic association studies of human diseases.

The chapters above focused primarily on single nucleotide polymorphisms (SNPs). Thus, in the last part of this dissertation, we focused on identification and disease-association of two types of structural variants: CNVs (chapter 4.1) and viral insertions (chapter 4.2).

Structural variation detection and disease-association

In chapter 4.1 we identified CNVs using an integrated approach to discover CNVs de-novo, followed by the meta-analyses of the curated high-quality CNVs. We identified nine nominally significant regions with AD, six deletions and three duplications;

although the individual studies might be underpowered, they collectively revealed consistent effect sizes, in both direction and magnitude. The nine CNVs ranged from 4.3kb to 221.7kb in size and had ORs from 1.31 to 2.88; and eight of them had frequency ≤ 5% (no CNV imputation conducted in this study due to low frequencies of these CNVs). The most significant AD association was found with the 5q21.3deletion (OR = 2.15 and $P = 3.8 \times 10^{-4}$). This cytogenetic band has been associated with alcohol cravings in a Native American population[264]; however, our meta-analysis, for the first time, identified a specific CNV in this region associated with AD.

In chapter 4.2, we present an in-silico method to simulate viral insertions (VIpower), according to empirical genomic information. Our primary findings include the discovery of six factors that are most important at discovery of VIs: cellular proportion, sequencing depth, length of integrated viral sequence, insert size, minimum number of required supporting reads (user-defined), and read length. We also developed a fast computational framework to detect virome-wide viral integrations in the human genome, and validated the six factors above using an independent NGS tool. The results in this study provide the fundamental guidance to the NGS-based experimental designs and data analyses of viral etiological studies of various human diseases.

In Appendix A, we applied an existing VI discovery approach to Alzheimer's disease brain samples. We identified HHV6B of a very specific strain (Z29) present at sufficiently high-abundance that the entire virus genome was sequenced at around 15 fold

depth. Due to the complex integration mechanism, whereby HHV6 genome relies on homologous recombination to integrate into sub-telomeric regions of the human genome, we were unable to provide definitely proof of integration. Chimeric reads with perfect repeats of the motif (TAACCC) were challenging to designate as uniquely human or virus since the motif pattern is shared by both human and HHV genome. To address this issue, we have proposed a statistical framework (borrowed from RNA-seq transcript quantification) to provide a probabilistic solution to the question of confidence of viral integration detection.

Future directions

Our work contributes to the "growing wave" of post-GWAS studies for brain disorders as well as complex human diseases. The post-GWAS era is a term coined around 2010, and it refers to the genetic and/or genomic analyses conducted to identify disease-causing variants, and not simply disease-associated variants. It is clear that the post-GWAS future of genetic research for human brain disease will rely on the well-integrated application of multi-disciplinary approaches such as human evolution, anthropology, epidemiology, psychiatry, molecular genetics and genomics. Immediate next steps that will need to be taken in the near future to bring the impact of our work one step closer to the clinic are:

"Deep phenotyping" Collection of hundreds of phenotype data (i.e., biological phenotypes as well as environmental information) for each individual has several advantages. First and foremost, disease-predicting models would not be limited to just genetic information if phenotypic and environmental information was available. Second, we would be able to test for shared genetic causality or genetic architectures between phenotypes using Mendelian randomization and LD-score regression, respectively. For instance, our finding of eye color association with alcohol dependence could be further elucidated using the LD-score regression approach. More recently, Beirut and colleagues reported a genetic correlation between smoking behavior and schizophrenia using "deeply phenotyped" samples[294]. Third, the abundance of phenotypes would allow us to identify potentially beneficial phenotypes caused by the pleiotropic, disease-associated variants we reported in chapter 2.1. Further evidence supporting antagonistic pleiotropy would help pinpoint disease mechanisms for the brain disorders and other complex diseases identified in our study. Fourth, endophenotype information can be very valuable in discovering disease loci, particularly in brain disorders. For instance we may have a higher statistical power to detect associations with activity in different brain regions associated with substance addiction, rather than associations between genomic loci and the addiction diagnosis itself. This may occur due to the complex nature of addiction etiology, composed of genetic, epigenetic, environmental and socio-economic factors; all of the non-genetic factors may cause incomplete penetrance of the risk loci. Fifth, detailed phenotyping information would allow for testing gene-environment interaction hypotheses. Sixth, artificial intelligence methods would be able to find information that

traditional statistical approaches may not be able to easily identify, such as disease-progression patterns, using unsupervised and semi-supervised learning.

"Pathogenic structural variant map" Rare and de-novo CNVs with large effect sizes in brain disorders are being discovered at an increasing pace. The most recent catalog of such disorders includes nearly 33,000 de-novo CNVs discovered from 23,098 trios[295]. Thus, current efforts to build a map of pathogenic CNVs are very promising at delivering disease-causing CNVs and genes with recurrent CNVs (i.e., hotspots). A similar map can be constructed for viral insertions. The primary advantage of having a reference for pathogenic CNVs or VIs, is that targeted (re)sequencing experiments can be carried out at ultra-high depth and lower cost than whole-genome sequencing, allowing for accurate genotyping, in the case of CNVs, or cellular proportion measurements in the case of VIs.

BIBLIOGRAPHY

Akey, J.M. et al. Interaction between the melanocortin-1 receptor and P genes contributes to inter-individual variation in skin pigmentation phenotypes in a Tibetan population. Hum Genet 108, 516-20 (2001).

Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19, 1655-64 (2009).

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. J Mol Biol 215, 403-10 (1990).

Altshuler, D.M. et al. A global reference for human genetic variation. Nature 526, 68-+ (2015).

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

Arnold, M., Soerjomataram, I., Ferlay, J. & Forman, D. Global incidence of oesophageal cancer by histological subtype in 2012. Gut 64, 381-7 (2015).

Aroor, A.R. & Shukla, S.D. MAP kinase signaling in diverse effects of ethanol. Life Sci 74, 2339-64 (2004).

Auton, A. et al. A global reference for human genetic variation. Nature 526, 68-74 (2015).

Azen, E.A., Latreille, P. & Niece, R.L. PRBI gene variants coding for length and null polymorphisms among human salivary Ps, PmF, PmS, and Pe proline-rich proteins (PRPs). Am J Hum Genet 53, 264-78 (1993).

Baik, I., Cho, N.H., Kim, S.H., Han, B.G. & Shin, C. Genome-wide association studies identify genetic loci related to alcohol consumption in Korean men. Am J Clin Nutr 93, 809-16 (2011).

Barreiro, L.B., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. Nat Genet 40, 340-5 (2008).

Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263-5 (2005).

Barrett, J.H. et al. Genome-wide association study identifies three new melanoma susceptibility loci. Nat Genet 43, 1108-13 (2011).

Bassett, J.F. & Dabbs, J.M. Eye color predicts alcohol use in two archival samples. Personality and Individual Differences 31, 535-539 (2001).

Baughman, R.P. et al. Clinical characteristics of patients in a case control study of sarcoidosis. Am J Respir Crit Care Med 164, 1885-9 (2001).

Beleza, S. et al. Genetic architecture of skin and eye color in an African-European admixed population. PLoS Genet 9, e1003372 (2013).

Belyi, V.A., Levine, A.J. & Skalka, A.M. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. PLoS Pathog 6, e1001030 (2010).

Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57, 289-300 (1995).

Bersaglieri, T. et al. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74, 1111-20 (2004).

Bertoni, J.M. et al. Increased melanoma risk in Parkinson disease: a prospective clinicopathological study. Arch Neurol 67, 347-52 (2010).

Bhatia, G., Patterson, N., Sankararaman, S. & Price, A.L. Estimating and interpreting FST: the impact of rare variants. Genome Res 23, 1514-21 (2013).

Bierut, L.J. et al. Novel genes identified in a high-density genome wide association study for nicotine dependence. Hum Mol Genet 16, 24-35 (2007).

Biggerstaff, B.J. & Tweedie, R.L. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Stat Med 16, 753-68 (1997).

Blum, K. et al. Allelic Association of Human Dopamine-D2 Receptor Gene in Alcoholism. Jama-Journal of the American Medical Association 263, 2055-2060 (1990).

Bolos, A.M. et al. Population and Pedigree Studies Reveal a Lack of Association between the Dopamine-D2 Receptor Gene and Alcoholism. Jama-Journal of the American Medical Association 264, 3156-3160 (1990).

Bonafe, M. et al. The different apoptotic potential of the p53 codon 72 alleles increases with age and modulates in vivo ischaemia-induced cell death. Cell Death Differ 11, 962-73 (2004).

Boyle, P. Triple-negative breast cancer: epidemiological considerations and recommendations. Ann Oncol 23 Suppl 6, vi7-12 (2012).

Briscoe, V.J., Tate, D.B. & Davis, S.N. Type 1 diabetes: exercise and hypoglycemia. Appl Physiol Nutr Metab 32, 576-82 (2007).

Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81, 1084-97 (2007).

Burmeister, M., McInnis, M.G. & Zollner, S. Psychiatric genetics: progress amid controversy. Nat Rev Genet 9, 527-40 (2008).

Cabana-Dominguez, J. et al. A Highly Polymorphic Copy Number Variant in the NSF Gene is Associated with Cocaine Dependence. Sci Rep 6, 31033 (2016).

Cao, J. et al. Association of the HTR2A gene with alcohol and heroin abuse. Hum Genet 133, 357-65 (2013).

Cao, J. et al. Association of the HTR2A gene with alcohol and heroin abuse. Hum Genet 133, 357-65 (2014).

Cao, J., Hudziak, J.J. & Li, D. Multi-cultural association of the serotonin transporter gene (SLC6A4) with substance use disorder. Neuropsychopharmacology 38, 1737-47 (2013).

Cao, J., LaRocque, E. & Li, D. Associations of the 5-hydroxytryptamine (serotonin) receptor 1B gene (HTR1B) with alcohol, cocaine, and heroin abuse. Am J Med Genet B Neuropsychiatr Genet 162B, 169-76 (2013).

Carbone, I. et al. Herpes virus in Alzheimer's disease: relation to progression of the disease. Neurobiology of Aging 35, 122-129 (2014).

Carrigan, M.A. et al. Hominids adapted to metabolize ethanol long before human-directed fermentation. Proc Natl Acad Sci U S A 112, 458-63 (2015).

Carter, N.P. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39, S16-21 (2007).

Cassidy, S.B., Schwartz, S., Miller, J.L. & Driscoll, D.J. Prader-Willi syndrome. Genet Med (2011).

Castren, E. & Tanila, H. Neurotrophins and dementia--keeping in touch. Neuron 51, 1-3 (2006).

Challoner, P.B. et al. Plaque-associated expression of human herpesvirus 6 in multiple sclerosis. Proc Natl Acad Sci U S A 92, 7440-4 (1995).

Chang, E.T. & Adami, H.O. The enigmatic epidemiology of nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev 15, 1765-77 (2006).

Chen, Y.X. et al. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics 29, 266-267 (2013).

Cloninger, C.R., Bohman, M. & Sigvardson, S. Inheritance of Alcohol-Abuse - Cross-Fostering Analysis of Adopted Men. Archives of General Psychiatry 38, 861-868 (1981).

Cohen, J.C. et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305, 869-72 (2004).

Colella, S. et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35, 2013-25 (2007).

Cooper, G.M. et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15, 901-13 (2005).

Cooper, R.S., Tayo, B. & Zhu, X. Genome-wide association studies: implications for multiethnic samples. Hum Mol Genet 17, R151-5 (2008).

Covault, J. et al. Interactive effects of the serotonin transporter 5-HTTLPR polymorphism and stressful life events on college student drinking and drug use. Biol Psychiatry 61, 609-16 (2007).

Curtis, D. et al. Case-case genome-wide association analysis shows markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes. Psychiatr Genet 21, 1-4 (2011).

Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10, 5-6 (2013).

Demirhan, O. & Tastemir, D. Cytogenetic effects of ethanol on chronic alcohol users. Alcohol Alcohol 43, 127-36 (2008).

DiLuca, M. & Olesen, J. The cost of brain diseases: a burden or a challenge? Neuron 82, 1205-8 (2014).

DiRocco, D.P., Scheiner, Z.S., Sindreu, C.B., Chan, G.C. & Storm, D.R. A role for calmodulin-stimulated adenylyl cyclases in cocaine sensitization. J Neurosci 29, 2393-403 (2009).

Donati, D. et al. Detection of human herpesvirus-6 in mesial temporal lobe epilepsy surgical brain resections. Neurology 61, 1405-11 (2003).

Donnelly, M.P. et al. A global view of the OCA2-HERC2 region and pigmentation. Hum Genet 131, 683-96 (2012).

Douville, R., Liu, J., Rothstein, J. & Nath, A. Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. Ann Neurol 69, 141-51 (2011).

Drgon, T., D'Addario, C. & Uhl, G.R. Linkage disequilibrium, haplotype and association studies of a chromosome 4 GABA receptor gene cluster: Candidate gene variants for addictions. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics 141B, 854-860 (2006).

Dudley, J.T. et al. Human genomic disease variants: a neutral evolutionary explanation. Genome Res 22, 1383-94 (2012).

Duffy, D.L. et al. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. Am J Hum Genet 80, 241-52 (2007).

Dunn, A.L., Heavner, J.E., Racz, G. & Day, M. Hyaluronidase: a review of approved formulations, indications and off-label use in chronic pain management. Expert Opin Biol Ther 10, 127-31 (2010).

Ehlers, C.L. & Wilhelmsen, K.C. Genomic scan for alcohol craving in Mission Indians. Psychiatr Genet 15, 71-5 (2005).

Eichler, E.E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11, 446-50 (2010).

Elipot, Y. et al. A mutation in the enzyme monoamine oxidase explains part of the Astyanax cavefish behavioural syndrome. Nat Commun 5, 3647 (2014).

Engelken, J. et al. Extreme population differences in the human zinc transporter ZIP4 (SLC39A4) are explained by positive selection in Sub-Saharan Africa. PLoS Genet 10, e1004128 (2014).

Ezzati, M. & Riboli, E. Behavioral and dietary risk factors for noncommunicable diseases. N Engl J Med 369, 954-64 (2013).

Fejerman, L. et al. European ancestry is positively associated with breast cancer risk in Mexican women. Cancer Epidemiol Biomarkers Prev 19, 1074-82 (2010).

Foxman, E.F. & Iwasaki, A. Genome-virome interactions: examining the role of common viral infections in complex disease. Nature Reviews Microbiology 9, 254-264 (2011).

Fraley, C. & Raftery, A.E. Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. Journal of Classification 20, 263-286 (2003).

Frank, J. et al. Genome-wide significant association between alcohol dependence and a variant in the ADH gene cluster. Addict Biol 17, 171-80 (2012).

Franzini, M. et al. High-sensitivity gamma-glutamyltransferase fraction pattern in alcohol addicts and abstainers. Drug Alcohol Depend 127, 239-42 (2013).

Fraser, H.B. Gene expression drives local adaptation in humans. Genome Res 23, 1089-96 (2013).

Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. Nat Rev Genet 10, 241-51 (2009).

Fumagalli, M. et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science 349, 1343-7 (2015).

Gabriel, S.B. et al. The structure of haplotype blocks in the human genome. Science 296, 2225-9 (2002).

Garcia-Barcelo, M.M. et al. Genome-wide association study identifies a susceptibility locus for biliary atresia on 10q24.2. Hum Mol Genet 19, 2917-25 (2010).

Gardiner, E. & Jackson, C.J. Eye color Predicts Disagreeableness in North Europeans: Support in Favor of Frost (2006). Current Psychology 29, 1-9 (2010).

Gelernter, J. & Kranzler, H.R. Genetics of alcohol dependence. Hum Genet 126, 91-9 (2009).

Gelernter, J. et al. Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. Mol Psychiatry 19, 41-9 (2014).

Genetic Analysis of Psoriasis, C. et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet 42, 985-90 (2010).

Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46, 818-25 (2014).

Genomes Project, C. et al. A global reference for human genetic variation. Nature 526, 68-74 (2015).

Genomes Project, C. et al. A map of human genome variation from population-scale sequencing. Nature 467, 1061-73 (2010).

Genomes Project, C. et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65 (2012).

Genovese, G. et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science 329, 841-5 (2010).

Gern, J.E. Rhinovirus and the initiation of asthma. Current Opinion in Allergy and Clinical Immunology 9, 73-78 (2009).

Goizet, C. et al. Molecular characterization of an 11q14.3 microdeletion associated with leukodystrophy. Eur J Hum Genet 12, 245-50 (2004).

Greenwood, B.M., Bradley, A.K. & Wall, R.A. Meningococcal disease and season in sub-Saharan Africa. Lancet 2, 829-30 (1985).

Gronberg, H. Prostate cancer epidemiology. Lancet 361, 859-64 (2003).

Grossman, S.R. et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327, 883-6 (2010).

Grossman, S.R. et al. Identifying recent adaptations in large-scale genomic data. Cell 152, 703-13 (2013).

Grossmann, A. et al. Phospho-tyrosine dependent protein-protein interaction network. Mol Syst Biol 11, 794 (2015).

Hamblin, M.T., Thompson, E.E. & Di Rienzo, A. Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70, 369-83 (2002).

Han, J. et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet 4, e1000074 (2008).

Hart, A.B. et al. Genome-wide association study of d-amphetamine response in healthy volunteers identifies putative associations, including cadherin 13 (CDH13). PLoS One 7, e42646 (2012).

Hartz, S.M. et al. Genetic correlation between smoking behaviors and schizophrenia. Schizophr Res (2017).

He, M. et al. Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. Hum Mol Genet 24, 1791-800 (2015).

Heath, A.C. et al. Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. Psychological Medicine 27, 1381-1396 (1997).

Heilmann, S. et al. Androgenetic alopecia: identification of four genetic risk loci and evidence for the contribution of WNT signaling to its etiology. J Invest Dermatol 133, 1489-96 (2013).

Hennig, C. & Hausdorf, B. prabclus: Functions for clustering of presence–absence, abundance and multilocus genetic data. R package version 2, 2-2 (2010).

Hewett, M. et al. PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res 30, 163-5 (2002).

Hicks, B.M., Krueger, R.F., Iacono, W.G., McGue, M. & Patrick, C.J. Family transmission and heritability of externalizing disorders - A twin-family study. Archives of General Psychiatry 61, 922-928 (2004).

Higuchi, S., Motohashi, Y., Ishibashi, K. & Maeda, T. Influence of eye colors of Caucasians and Asians on suppression of melatonin secretion by light. Am J Physiol Regul Integr Comp Physiol 292, R2352-6 (2007).

Ho, D.W., Sze, K.M. & Ng, I.O. Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. Oncotarget 6, 20959-63 (2015).

Ho, D.W.H., Sze, K.M.F. & Ng, I.O.L. Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. Oncotarget 6, 20959-20963 (2015).

Hoglinger, G.U. et al. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. Nat Genet 43, 699-705 (2011).

Holsinger, K.E. & Weir, B.S. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet 10, 639-50 (2009).

Horie, M. et al. Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463, 84-U90 (2010).

Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5, e1000529 (2009).

Hradetzky, S. et al. The human skin-associated autoantigen alpha-NAC activates monocytes and dendritic cells via TLR-2 and primes an IL-12-dependent Th1 response. J Invest Dermatol 133, 2289-92 (2013).

Hu, X. et al. pIRS: Profile-based Illumina pair-end reads simulator. Bioinformatics 28, 1533-5 (2012).

Huang, Z. & Sjoholm, A. Ethanol acutely stimulates islet blood flow, amplifies insulin secretion, and induces hypoglycemia via nitric oxide and vagally mediated mechanisms. Endocrinology 149, 232-6 (2008).

Huckins, L.M. et al. Using ancestry-informative markers to identify fine structure across 15 populations of European origin. Eur J Hum Genet 22, 1190-200 (2014).

Huson, D.H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61, 1061-7 (2012).

Iafrate, A.J. et al. Detection of large-scale variation in the human genome. Nat Genet 36, 949-51 (2004).

Inskip, H.M., Harris, E.C. & Barraclough, B. Lifetime risk of suicide for affective disorder, alcoholism and schizophrenia. Br J Psychiatry 172, 35-7 (1998).

Johnson, A.D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24, 2938-9 (2008).

Joslyn, G., Ravindranathan, A., Brush, G., Schuckit, M. & White, R.L. Human variation in alcohol response is influenced by variation in neuronal signaling genes. Alcohol Clin Exp Res 34, 800-12 (2010).

Julian, C.G. et al. Augmented uterine artery blood flow and oxygen delivery protect Andeans from altitude-associated reductions in fetal growth. Am J Physiol Regul Integr Comp Physiol 296, R1564-75 (2009).

Kamberov, Y.G. et al. Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell 152, 691-702 (2013).

Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42, D199-205 (2014).

Karlsson, E.K., Kwiatkowski, D.P. & Sabeti, P.C. Natural selection and infectious disease in human populations. Nat Rev Genet 15, 379-93 (2014).

Karst, S.M., Wobus, C.E., Lay, M., Davidson, J. & Virgin, H.W. STAT1-dependent innate immunity to a Norwalk-like virus. Science 299, 1575-1578 (2003).

Katz, J.P. & Pipas, J.M. SummonChimera infers integrated viral genomes with nucleotide precision from NGS data. Bmc Bioinformatics 15(2014).

Kendler, K.S., Prescott, C.A., Neale, M.C. & Pedersen, N.L. Temperance board registration for alcohol abuse in a national sample of Swedish male twins, born 1902 to 1949. Archives of General Psychiatry 54, 178-184 (1997).

Kersbergen, P. et al. Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. BMC Genet 10, 69 (2009).

Key, F.M., Teixeira, J.C., de Filippo, C. & Andres, A.M. Advantageous diversity maintained by balancing selection in humans. Curr Opin Genet Dev 29, 45-51 (2014).

Khoury, J.D. et al. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. J Virol 87, 8916-26 (2013).

Kim, K.S. et al. Adenylyl cyclase type 5 (AC5) is an essential mediator of morphine action. Proc Natl Acad Sci U S A 103, 3908-13 (2006).

Kim, S.Y., Kim, J.H. & Chung, Y.J. Effect of Combining Multiple CNV Defining Algorithms on the Reliability of CNV Calls from SNP Genotyping Data. Genomics Inform 10, 194-9 (2012).

Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310-5 (2014).

Klein, R.J. et al. Complement factor H polymorphism in age-related macular degeneration. Science 308, 385-389 (2005).

Klenerman, P., Hengartner, H. & Zinkernagel, R.M. A non-retroviral RNA virus persists in DNA form. Nature 390, 298-301 (1997).

Krumm, N. et al. Excess of rare, inherited truncating mutations in autism. Nat Genet 47, 582-8 (2015).

Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. Genome Res 19, 1639-45 (2009).

Lamason, R.L. et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310, 1782-6 (2005).

Lander, E.S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001).

Lao, O., Andres, A.M., Mateu, E., Bertranpetit, J. & Calafell, F. Spatial patterns of cystic fibrosis mutation spectra in European populations. Eur J Hum Genet 11, 385-94 (2003).

Lasek, A.W. Effects of Ethanol on Brain Extracellular Matrix: Implications for Alcohol Use Disorder. Alcohol Clin Exp Res 40, 2030-2042 (2016).

Lee, J.W., Brancati, F.L. & Yeh, H.C. Trends in the prevalence of type 2 diabetes in Asians versus whites: results from the United States National Health Interview Survey, 1997-2008. Diabetes Care 34, 353-7 (2011).

Lee, S., Teslovich, T.M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet 93, 42-53 (2013).

Leggio, L., Ray, L.A., Kenna, G.A. & Swift, R.M. Blood glucose level, alcohol heavy drinking, and alcohol craving during treatment for alcohol dependence: results from the Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence (COMBINE) Study. Alcohol Clin Exp Res 33, 1539-44 (2009).

Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83, 311-21 (2008).

Li, B. et al. A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet 8, e1002944 (2012).

Li, D. & He, L. Meta-study on association between the monoamine oxidase A gene (MAOA) and schizophrenia. Am J Med Genet B Neuropsychiatr Genet 147B, 174-8 (2008).

Li, D. et al. Association of gamma-aminobutyric acid A receptor alpha2 gene (GABRA2) with alcohol use disorder. Neuropsychopharmacology 39, 907-18 (2014).

Li, D. et al. Genome-Wide Association Study of Copy Number Variations (CNVs) with Opioid Dependence. Neuropsychopharmacology 40, 1016-26 (2015).

Li, D., Zhao, H. & Gelernter, J. Further clarification of the contribution of the ADH1C gene to vulnerability of alcoholism and selected liver diseases. Hum Genet 131, 1361-74 (2012).

Li, D., Zhao, H. & Gelernter, J. Strong Association of the Alcohol Dehydrogenase 1B Gene (ADH1B) with Alcohol Dependence and Alcohol-Induced Medical Diseases. Biol Psychiatry (2011).

Li, D., Zhao, H. & Gelernter, J. Strong association of the alcohol dehydrogenase 1B gene (ADH1B) with alcohol dependence and alcohol-induced medical diseases. Biol Psychiatry 70, 504-12 (2011).

Li, D., Zhao, H. & Gelernter, J. Strong protective effect of the aldehyde dehydrogenase gene (ALDH2) 504lys (*2) allele against alcoholism and alcohol-induced medical diseases in Asians. Hum Genet 131, 725-37 (2012).

Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-60 (2009).

Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27, 718-9 (2011).

Li, Y., Sidore, C., Kang, H.M., Boehnke, M. & Abecasis, G.R. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res 21, 940-51 (2011).

Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34, 816-34 (2010).

Li, Y.R. & Keating, B.J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med 6, 91 (2014).

Lin, A., Wang, R.T., Ahn, S., Park, C.C. & Smith, D.J. A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. Genome Res 20, 1122-32 (2010).

Lin, P. et al. Copy Number Variations in 6q14.1 and 5q13.2 are Associated with Alcohol Dependence. Alcoholism-Clinical and Experimental Research 36, 1512-1518 (2012).

Lin, W.R., Wozniak, M.A., Cooper, R.J., Wilcock, G.K. & Itzhaki, R.F. Herpesviruses in brain and Alzheimer's disease. J Pathol 197, 395-402 (2002).

Liu, E.Y. et al. Genotype imputation of Metabochip SNPs using a study-specific reference panel of ~4,000 haplotypes in African Americans from the Women's Health Initiative. Genet Epidemiol 36, 107-17 (2012).

Liu, J. et al. Positive association of the human GABA-A-receptor beta 2 subunit gene haplotype with schizophrenia in the Chinese Han population. Biochem Biophys Res Commun 334, 817-23 (2005).

Liu, J.Z. et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 42, 436-40 (2010).

Loh, E.W. et al. Association between variants at the GABA(A)beta 2, GABA(A)alpha 6 and GABA(A)gamma 2 gene cluster and alcohol dependence in a Scottish population. Molecular Psychiatry 4, 539-544 (1999).

Long, J.C. et al. Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an American Indian population. American Journal of Medical Genetics 81, 216-221 (1998).

Love-Gregory, L. et al. Variants in the CD36 gene associate with the metabolic syndrome and high-density lipoprotein cholesterol. Hum Mol Genet 17, 1695-704 (2008).

Luczak, S.E., Glatt, S.J. & Wall, T.L. Meta-analyses of ALDH2 and ADH1B with alcohol dependence in Asians. Psychological Bulletin 132, 607-621 (2006).

Luo, X.G. et al. ADH4 gene variation is associated with alcohol dependence and drug dependence in European Americans: Results from HWD tests and case-control association studies. Neuropsychopharmacology 31, 1085-1095 (2006).

Madhava, V., Burgess, C. & Drucker, E. Epidemiology of chronic hepatitis C virus infection in sub-Saharan Africa. Lancet Infect Dis 2, 293-302 (2002).

Madsen, B.E. & Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5, e1000384 (2009).

Magi, R. & Morris, A.P. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11, 288 (2010).

Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 39, D52-7 (2011).

Manzardo, A. Interpretation of Eye Color Associations with Alcohol Dependence Risk in European Americans. Am J Med Genet B Neuropsychiatr Genet (2015).

Manzardo, A.M., McGuire, A. & Butler, M.G. Clinically relevant genetic biomarkers from the brain in alcoholism with representation on high resolution chromosome ideograms. Gene 560, 184-94 (2015).

Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. Nat Rev Genet 11, 499-511 (2010).

Matthews, A.G., Hoffman, E.K., Zezza, N., Stiffler, S. & Hill, S.Y. The role of the GABRA2 polymorphism in multiplex alcohol dependence families with minimal comorbidity: Within-family association and linkage analyses. Journal of Studies on Alcohol and Drugs 68, 625-633 (2007).

Maurin, M.L. et al. Terminal 14q32.33 deletion: genotype-phenotype correlation. Am J Med Genet A 140, 2324-9 (2006).

McCarroll, S.A. & Altshuler, D.M. Copy-number variation and association studies of human disease. Nat Genet 39, S37-42 (2007).

McCarroll, S.A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40, 1166-74 (2008).

McCarthy, D.J. et al. Choice of transcripts and software has a large effect on variant annotation. Genome Medicine 6(2014).

McKay, J.D. et al. Lung cancer susceptibility locus at 5p15.33. Nat Genet 40, 1404-6 (2008).

McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26, 2069-70 (2010).

Mefford, H.C. et al. A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. Genome Res 19, 1579-85 (2009).

Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. Science 201, 786-92 (1978).

Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11, 685-96 (2010).

Mikovits, J.A., Lombardi, V.C., Pfost, M.A., Hagen, K.S. & Ruscetti, F.W. Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. Virulence 1, 386-90 (2009).

Montojo, J. et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics 26, 2927-8 (2010).

Morissette, G. & Flamand, L. Herpesviruses and chromosomal integration. J Virol 84, 12100-9 (2010).

Mozaffarian, D. et al. Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. Am J Clin Nutr 101, 398-406 (2015).

Nair, R.P. et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet 41, 199-204 (2009).

Nassir, R. et al. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genet 10, 39 (2009).

Nelson, S.C., Doheny, K.F., Laurie, C.C. & Mirel, D.B. Is 'forward' the same as 'plus'?...and other adventures in SNP allele nomenclature. Trends Genet 28, 361-3 (2012).

Nicolazzi, E.L. et al. SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics 15, 123 (2014).

Nievergelt, C.M. et al. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. Investig Genet 4, 13 (2013).

Nikolaou, V. & Stratigos, A.J. Emerging trends in the epidemiology of melanoma. Br J Dermatol 170, 11-9 (2014).

Noguchi, K. et al. TRIM40 promotes neddylation of IKKgamma and is downregulated in gastrointestinal cancers. Carcinogenesis 32, 995-1004 (2011).

Ober, C. & Yao, T.C. The genetics of asthma and allergic disease: a 21st century perspective. Immunol Rev 242, 10-30 (2011).

Olfson, E. & Bierut, L.J. Convergence of genome-wide association and candidate gene studies for alcoholism. Alcohol Clin Exp Res 36, 2086-94 (2012).

Pacchierotti, C., Iapichino, S., Bossini, L., Pieraccini, F. & Castrogiovanni, P. Melatonin in psychiatric disorders: a review on the melatonin involvement in psychiatry. Front Neuroendocrinol 22, 18-32 (2001).

Palmer, R.H. et al. The genetics of alcohol dependence: advancing towards systems-based approaches. Drug Alcohol Depend 125, 179-91 (2012).

Panagiotou, O.A., Evangelou, E. & Ioannidis, J.P. Genome-wide significant associations for variants with minor allele frequency of 5% or less--an overview: A HuGE review. Am J Epidemiol 172, 869-89 (2010).

Panagiotou, O.A., Willer, C.J., Hirschhorn, J.N. & Ioannidis, J.P. The power of meta-analysis in genome-wide association studies. Annu Rev Genomics Hum Genet 14, 441-65 (2013).

Park, S.L. et al. Mercapturic Acids Derived from the Toxicants Acrolein and Crotonaldehyde in the Urine of Cigarette Smokers from Five Ethnic Groups with Differing Risks for Lung Cancer. PLoS One 10, e0124841 (2015).

Platt, O.S. et al. Mortality in sickle cell disease. Life expectancy and risk factors for early death. N Engl J Med 330, 1639-44 (1994).

Popovic, D. & Dikic, I. TBC1D5 and the AP2 complex regulate ATG9 trafficking and initiation of autophagy. EMBO Rep 15, 392-401 (2014).

Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904-9 (2006).

Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945-59 (2000).

Procopio, D.O. et al. Genetic markers of comorbid depression and alcoholism in women. Alcohol Clin Exp Res 37, 896-904 (2013).

Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-75 (2007).

Qin, P. et al. A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. Eur J Hum Genet 22, 248-53 (2014).

Qiu, C., Kivipelto, M. & von Strauss, E. Epidemiology of Alzheimer's disease: occurrence, determinants, and strategies toward intervention. Dialogues Clin Neurosci 11, 111-28 (2009).

Ralston, S.H. Clinical practice. Paget's disease of bone. N Engl J Med 368, 644-50 (2013).

Ray, L.A. & Hutchison, K.E. Effects of naltrexone on alcohol sensitivity and genetic moderators of medication response: a double-blind placebo-controlled study. Arch Gen Psychiatry 64, 1069-77 (2007).

Rees, J.L. The genetics of sun sensitivity in humans. Am J Hum Genet 75, 739-51 (2004).

Reich, T. et al. Genome-wide search for genes affecting the risk for alcohol dependence. American Journal of Medical Genetics 81, 207-215 (1998).

Risch, N., Tang, H., Katzenstein, H. & Ekstein, J. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. Am J Hum Genet 72, 812-22 (2003).

Robert-Gangneux, F. & Darde, M.L. Epidemiology of and diagnostic strategies for toxoplasmosis. Clin Microbiol Rev 25, 264-96 (2012).

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 12, R22 (2011).

Roecklein, K.A. et al. A missense variant (P10L) of the melanopsin (OPN4) gene in seasonal affective disorder. J Affect Disord 114, 279-85 (2009).

Rosenberg, N.A., Li, L.M., Ward, R. & Pritchard, J.K. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73, 1402-22 (2003).

Ruderfer, D.M. et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. Nat Genet 48, 1107-11 (2016).

Russo, S.J. & Nestler, E.J. The brain reward circuitry in mood disorders. Nature Reviews Neuroscience 14, 609-625 (2013).

Sabeti, P.C. et al. Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832-7 (2002).

Sanders, S.J. et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70, 863-85 (2011).

Schizophrenia Psychiatric Genome-Wide Association Study, C. Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43, 969-76 (2011).

Schumann, G. et al. Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. Proc Natl Acad Sci U S A 108, 7119-24 (2011).

Sebat, J. et al. Strong association of de novo copy number mutations with autism. Science 316, 445-9 (2007).

Sellers, E.M., Higgins, G.A. & Sobell, M.B. 5-Ht and Alcohol-Abuse. Trends in Pharmacological Sciences 13, 69-75 (1992).

Sham, P.C. & Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet 15, 335-46 (2014).

Sharp, A.J. et al. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet 40, 322-8 (2008).

Sher, L. Alcohol consumption and suicide. QJM 99, 57-61 (2006).

Sher, L. Alcoholism and seasonal affective disorder. Compr Psychiatry 45, 51-6 (2004).

Sherva, R. et al. Genome-wide Association Study of Cannabis Dependence Severity, Novel Risk Variants, and Shared Genetic Risks. JAMA Psychiatry 73, 472-80 (2016).

Smyth, D.J. et al. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nature Genetics 38, 617-619 (2006).

So, H.C., Gui, A.H., Cherny, S.S. & Sham, P.C. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol 35, 310-7 (2011).

Sobota, R.S. et al. A Locus at 5q33.3 Confers Resistance to Tuberculosis in Highly Susceptible Individuals. Am J Hum Genet 98, 514-24 (2016).

Soler Artigas, M. et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet 43, 1082-90 (2011).

Soundararajan, U., Yun, L., Shi, M. & Kidd, K.K. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. Forensic Sci Int Genet 23, 25-32 (2016).

Spencer, C.C., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5, e1000477 (2009).

Stefansson, H. et al. Large recurrent microdeletions associated with schizophrenia. Nature 455, 232-6 (2008).

Stewart, D.R. et al. Dubowitz syndrome is a complex comprised of multiple, genetically distinct and phenotypically overlapping disorders. PLoS One 9, e98686 (2014).

Sturm, R.A. & Duffy, D.L. Human pigmentation genes under environmental selection. Genome Biol 13, 248 (2012).

Sturm, R.A. et al. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. Am J Hum Genet 82, 424-31 (2008).

Sturm, R.A. Molecular genetics of human pigmentation diversity. Hum Mol Genet 18, R9-17 (2009).

Sturm, R.A. Skin colour and skin cancer - MC1R, the genetic link. Melanoma Res 12, 405-16 (2002).

Sudmant, P.H. et al. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75-81 (2015).

Sulem, P. et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet 39, 1443-52 (2007).

Sulovari, A. & Li, D. GACT: a Genome build and Allele definition Conversion Tool for SNP imputation and meta-analysis in genetic association studies. BMC Genomics 15, 610 (2014).

Sulovari, A., Chen, Y.H., Hudziak, J.J. & Li, D. Atlas of human diseases influenced by genetic variants with extreme allele frequency differences. Hum Genet 136, 39-54 (2017).

Sulovari, A., Kranzler, H.R., Farrer, L.A., Gelernter, J. & Li, D. Eye color: A potential indicator of alcohol dependence risk in European Americans. Am J Med Genet B Neuropsychiatr Genet 168B, 347-53 (2015).

Sulovari, A., Kranzler, H.R., Farrer, L.A., Gelernter, J. & Li, D. Further analyses support the association between light eye color and alcohol dependence. Am J Med Genet B Neuropsychiatr Genet 168, 757-60 (2015).

Sung, W.K. et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet 44, 765-9 (2012).

Taft, R.J. et al. Mutations in DARS cause hypomyelination with brain stem and spinal cord involvement and leg spasticity. Am J Hum Genet 92, 774-80 (2013).

Tanzi, R.E. & Bertram, L. Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective. Cell 120, 545-55 (2005).

Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4, Unit 4 10 (2009).

Taylor, D.J. & Bruenn, J. The evolution of novel fungal genes from non-retroviral RNA viruses. BMC Biol 7, 88 (2009).

Tenesa, A. & Haley, C.S. The heritability of human disease: estimation, uses and abuses. Nat Rev Genet 14, 139-49 (2013).

Terman, J.S. & Terman, M. Photopic and scotopic light detection in patients with seasonal affective disorder and control subjects. Biol Psychiatry 46, 1642-8 (1999).

Thevenin, A., Ein-Dor, L., Ozery-Flato, M. & Shamir, R. Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. Nucleic Acids Res 42, 9854-61 (2014).

Thomasson, H.R. et al. Alcohol and Aldehyde Dehydrogenase Genotypes and Alcoholism in Chinese Men. American Journal of Human Genetics 48, 677-681 (1991).

Thorgeirsson, T.E. et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 452, 638-42 (2008).

Tishkoff, S.A. et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293, 455-62 (2001).

Tobacco & Genetics, C. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet 42, 441-7 (2010).

Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511-5 (2010).

Traylor, M. & Lewis, C.M. Genetic discovery in multi-ethnic populations. Eur J Hum Genet (2016).

Treutlein, J. et al. Genome-wide association study of alcohol dependence. Arch Gen Psychiatry 66, 773-84 (2009).

Turner, S. et al. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet Chapter 1, Unit1 19 (2011).

Turner, T.N. et al. denovo-db: a compendium of human de novo variants. Nucleic Acids Res 45, D804-D811 (2017).

Ulloa, A.E., Chen, J.Y., Vergara, V.M., Calhoun, V. & Liu, J.Y. Association Between Copy Number Variation Losses and Alcohol Dependence Across African American and European American Ethnic Groups. Alcoholism-Clinical and Experimental Research 38, 1266-1274 (2014).

Vaags, A.K. et al. Rare deletions at the neurexin 3 locus in autism spectrum disorder. Am J Hum Genet 90, 133-41 (2012).

Vaillant, G.E. Natural history of male psychological health: VIII. Antecedents of alcololism and "orality". Am J Psychiatry 137, 181-6 (1980).

van Heemst, D. et al. Variation in the human TP53 gene affects old age survival and cancer mortality. Exp Gerontol 40, 11-5 (2005).

Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software 36, 1-48 (2010).

Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. PLoS Biol 4, e72 (2006).

Vyssoki, B., Kapusta, N.D., Praschak-Rieder, N., Dorffner, G. & Willeit, M. Direct effect of sunshine on suicide. JAMA Psychiatry 71, 1231-7 (2014).

Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res 41, W77-83 (2013).

Wang, J.C. et al. Evidence of common and specific genetic effects: association of the muscarinic acetylcholine receptor M2 (CHRM2) gene with alcohol dependence and major depressive syndrome. Hum Mol Genet 13, 1903-11 (2004).

Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17, 1665-74 (2007).

Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164 (2010).

Wang, Q.G., Jia, P.L. & Zhao, Z.M. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Medicine 7(2015).

Wang, Q.G., Jia, P.L. & Zhao, Z.M. VirusFinder: Software for Efficient and Accurate Detection of Viruses and Their Integration Sites in Host Genomes through Next Generation Sequencing Data. Plos One 8(2013).

Weir, B.S. & Cockerham, C.C. Estimating F-Statistics for the Analysis of Population-Structure. Evolution 38, 1358-1370 (1984).

Weir, B.S. & Hill, W.G. Estimating F-statistics. Annu Rev Genet 36, 721-50 (2002).

Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42, D1001-6 (2014).

Wilde, S. et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A 111, 4832-7 (2014).

Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190-1 (2010).

Wojnar, M. et al. Impulsive and non-impulsive suicide attempts in patients treated for alcohol dependence. J Affect Disord 115, 131-9 (2009).

Wong, G.K. et al. A population threshold for functional polymorphisms. Genome Res 13, 1873-9 (2003).

Wood, A.R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46, 1173-86 (2014).

Wu, M.C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89, 82-93 (2011).

Wu, Y.W. et al. Lingo2 variants associated with essential tremor and Parkinson's disease. Hum Genet 129, 611-5 (2011).

Xu, W. et al. Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1. BMC Med Genet 15, 2 (2014).

Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88, 76-82 (2011).

Yang, X. et al. Dr.VIS v2.0: an updated database of human disease-related viral integration sites in the era of high-throughput deep sequencing. Nucleic Acids Res 43, D887-92 (2015).

Yi, X. et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329, 75-8 (2010).

Zamora-Martinez, E.R. & Edwards, S. Neuronal extracellular signal-regulated kinase (ERK) activity as marker and mediator of alcohol and opioid dependence. Front Integr Neurosci 8, 24 (2014).

Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821-9 (2008).

Zhang, M. et al. Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. Hum Mol Genet 22, 2948-59 (2013).

Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326-8 (2012).

Zuberi, K. et al. GeneMANIA prediction server 2013 update. Nucleic Acids Res 41, W115-22 (2013).

Zufferey, F. et al. A 600 kb deletion syndrome at 16p11.2 leads to energy imbalance and neuropsychiatric disorders. J Med Genet 49, 660-8 (2012).

Zuo, L. et al. NKAIN1-SERINC2 is a functional, replicable and genome-wide significant risk gene region specific for alcohol dependence in subjects of European descent. Drug Alcohol Depend 129, 254-64 (2013).

APPENDIX A: NGS-based Human-herpes 6 virus detection in Alzheimer brain

Abstract

The health burden of Alzheimer's disease (AD) is significant with prevalence of 5%-8% among individuals over 65 years old, or 15%-20% among individuals over 75 (WHO 2017). The biochemical pathways have been found to involve amyloid precursor genes which lead to an increase in Aβ aggregation and/or decrease in Aβ clearance, such as in the case of *APOE*- ε4 allele carriers[296]. However, around 40% of the genetic heritability of AD has not been accounted for. Recent studies have indicated a potential etiological role for viruses in AD[297]. However, no published studies have been able to identify fusion events between human and HHV6 DNA. In this study we identified one AD brain sample with HHV6 infection and potential integration, using whole-genome paired-end read NGS data. A statistical framework is proposed to estimate the probability of an integration event. This finding represents the first instance in the published literature of identifying a putative viral integration in AD brain.

Introduction

Viral etiologies in the context of brain disorders were initially described in multiple sclerosis[298], mesial temporal lobe epilepsy[299] and then Alzheimer disease[297]. The existing literature has focused on the use of polymerase chain reaction (PCR) methods to identify virus DNA for all these disorders. However, PCR presents several disadvantages:

(i) cannot conduct a hypothesis-free survey for viruses in brain tissue, (ii) it is not possible to sequence the entire genome of the virus that has been identified, and (iii) it is not possible to identify variations in the genome of the identified virus. All of these shortcoming can be addressed using NGS data. Here, we conducted alignment of 101bp paired-end reads from 20 brain samples to identify HHV-6B infection and/or insertion.

Methods

High-throughput alignment

The sequencing reads (saved in fastq format) for all 20 Alzheimer diagnosed brains were accessed and downloaded into a local server from dbGAP (accession code: phs000572.v7.p4). The NGS library was prepared with 500bp fragment sizes and sequenced at 33-fold depth. Next, the fastq-formatted reads were aligned against the HHV-6B reference genome (NCBI Nucleotide database accession code: NC_000898.1) using bwa[300]. Both single-end and completely aligned reads aligning to HHV-6B were considered. The coverage was calculated using: 2 × (read-length) × (read number) / genome-length, where read-length was 101bp and genome-length is the virus reference length of 162,114bp.

Local alignment

Since bwa chooses a random position between two or more equally likely alignments, we decided to use a conservative approach and align against the complete nucleotide NCBI database using BLAST[301]. For each read (i.e., each of the two read ends were considered independently), we kept only the most confident alignment result, as determined by the E-value. If and only if the read with the smallest E-value aligned to the virus (for different levels of virus specificity, see Table 1), it was considered to be a unique viral read.

Splice junctions

First, we assign each paired-end read to one of three classes: human-only, human-virus chimera, and virus only. A paired-end read (i.e., fragment) is a chimera if one end of the pair covers completely or partially a human-virus splice junction. The read alignment is examined together as reported in the SAM format, following *bwa* alignment. The following rules are applied to resolve multiple alignments for fragments, as previously described in the Cufflinks paper[302]. Only fragments with the highest rank are reported. Let *a* and *b* be two fragment alignments of the same fragment (i.e., read-pair), such that *a* is ranked lower than *b* if any of the following are true (in this order):

1) *a* is single end mapped, while b has both ends mapped,
2) *a* crosses more splice junctions than *b*

3) The reads from *a* map significantly apart according to the library's fragment length distribution (≥3 standard deviations), while the reads from *b* do not.
4) The reads from *a* are significantly closer together than expected (following Z-score normalization of the fragment size distribution), while reads from *b* are not.
5) The reads in alignment *a* map more than a read-length (e.g., 100bp) apart than the *b* alignment
6) *a* has more mismatches (reflected by a lower alignment score) than *b*.

Note that alignments of equal quality are all reported (e..g, *n* alignments), and the probability of each of them being correct is 1/*n*.

Likelihood of viral integrations

The statistical framework underlying the Cufflings[302] method for quantifying transcript abundance by RNA-seq was adopted for quantifying confidence of viral integration. Although the biology of RNA-seq is different from that of viral insertion, the statistical framework for estimating transcript abundance is similar. We assume that a region of the genome, for example a gene locus, is integrated by viral insertions at a certain cellular proportion ≤ 1. The integration results in formation of at least two isoforms: the human-only sequence and the human-virus chimeric sequence. More isoforms may form, if the integration site is a hotspot where multiple viruses can integrate. Since we do not know *a priori* the location of these integration sites, we slide a

window of a fixed size across the genome. Each window represents a distinct locus, labelled as *g* (from here onward we refer to the sliding window as locus *g*). The likelihood is a function of the relative isoform abundance (ρ) such that $\sum_{t \in T} \rho_t = 1$, where ρ_t is the relative abundance for individual isoform *t* relative to the entire genome, and *T* is the set of all isoforms across the human genome. The length of each isoform, *l(t)*, is fixed if locus *g* contains human-only reads, however, insertion of a viral sequence increases the value of *l(t)* by the same length as the integrated viral sequence. Since locus *g* is defined as a region that contains a set of overlapping isoforms; hence, $\rho_t = \beta_g \gamma_t$, where β_g is the relative abundance of locus *g* in which *t* is contained, and γ_t is the relative abundance of *t* within the *g* locus. The entire human genome is denoted by *G*.

The probability of selecting a fragment from single isoform *t*, conditioned on locus g, such that $t \in g$, is the locus-specific relative abundance γ_t, which is equal to:

$$\gamma_t = \frac{\tau_t \cdot \tilde{l}(t)}{\sum_{m \in g} \tau_m \tilde{l}(u)}$$

, where $\tau_t$represents the locus-specific proportion of isoform *t* (i.e., viral integration cellular proportion), such that $\tau_t = \frac{\rho_t}{\sum_{t \in g} \rho_t}$ and $\tilde{l}(t)$ represents the adjusted isoform length such that: $\tilde{l}(t) = \sum_{i=1}^{l(t)} F(i) \cdot (l(t) - i + 1)$. The adjusted isoform length is required since the probability of selecting a fragment of length *k* from isoform *t* at one of the positions is: $\frac{1}{l(t)-k}$.

The full likelihood model has been derived elsewhere for RNA-seq expression estimates[303]. The following likelihood function represents "the probability that a fragment selected at random originates from isoform *t*"

$$L(\rho|R) = \prod_{r\in R} Probability\ (read\ alignment = r)$$

$$= \left(\prod_{g\in G} {\beta_g}^{X_g}\right)\left(\prod_{g\in G}\left(\prod_{r\in R: r\in g}\sum_{t\in G}\gamma_t \cdot \frac{F(I_t(r))}{l(t) - I_t(r) + 1}\right)\right)$$

, where R is the complete set of aligned reads, *F* is the distribution of all fragment lengths (5' and 3' ends of a single fragment are sequenced by each read in a read-pair), such that *F(i)* represents the probability that a fragment has length *i* (although this is NGS-library specific, we assume F is normally distributed) and $\sum_{i=1}^{\infty} F(i) = 1$; X_g is the total number of fragments (i.e., read pairs) in a locus *g*, $I_t(f)$ is the implied length of a fragment *f*, assuming that it originated from the isoform *t*, and finally $l(t)$ is the length of the sliding window. Remember that *l(t)* is fixed if locus *g* contains human-only reads, however, insertion of a viral sequence increases the value of *l(t)* by the same length as the integrated viral sequence.

Results

In this study we identified 26,616 reads that aligned to the HHV-6B reference genome (**Figure 1** and **Table 1**). Of these reads, 16,825 (63.2%) reads were confirmed to align uniquely to the HHV-6B strain *Z29* (**Table 1**). In addition to the brain sample (sample ID SRR987641), we also identified a smaller coverage of the same viral genome in a blood sample of an Alzheimer patient (SRR1105833, no brain sample data was available for this sample). The average coverage of the HHV-6B genome from viral reads identified in the brain sample was around 16-fold, or 10-fold when considering the unique reads only (**Table 1**).

The unique reads were further used to construct contigs using the de-novo assembler Velvet[304]. A total of 154 contigs were generated, and each of them was found to uniquely align to a different positon on the reference genome. Thus, the entire genome of HHV6B (strain z29) was represented by the unique contigs we assembled. The assembled contigs were further used to identify SNPs and short indels in the virus' genome. A total of 106 variants were identified, including 102 SNPs and 4 short indels.

Lastly, we observed that 14 paired-end reads supported the existence of a circular episomal structure for the HHV6B genome. These were aligned with a high confidence to the reference genome (i.e., average alignment score of 96, out of 101).

Discussion

In this study we have identified 16,825 NGS reads that uniquely align to the z29 strain of the HHV-6B genome, leading to a uniquely-aligned coverage of around 10-fold. To our knowledge this is the first report of identifying a complete HHV-6B viral genome in an Alzheimer brain. The strengths of our study include: (i) identification of a complete HHV-6B genome in an Alzheimer's disease brain tissue, (ii) identification of 106 virus-specific variants (102 SNPs and 4 short indels). A weakness of our study is the lack of definitive evidence that we have observed a viral insertion. However, we have proposed a statistical framework that would allow us to quantify the confidence of the viral insertion into the human genome (see Methods).

Of particular interest is the integration mechanism by which HHV6B may infect or integrate into human neuron DNA. It is known that subtelomeric regions with the repeat signature of $(TAACCC)_n$ are preferred targets of herpesvirus integration via homologous recombination[305]. Furthermore, given that an NGS experiment is able to detect presence of HHV-6B implies that the cellular proportion (i.e., proportion of viral DNA copies out of all human and virus DNA copies) is from 10/33=30% (for uniquely-aligned viral reads) to 16/33=48% (for all bwa virus-aligned reads).

An immediate next step for our study would be the estimate the parameters β_g and γ_t in the likelihood function, using either a variable order markov model that leverages empirical sequencing data, or an analytical expectation maximization approach.

Tables and Figures

Table 1: The majority of NGS viral reads align with the highest confidence to HHV6B strain Z29.

Reference sequence type	**% of BWA-aligned reads***	
	SRR987641	SRR1105833
All	100% (26,616)	100% (1,316)
Virus	99.6% (26,513)	95.6% (1,258)
Herpes virus	99% (26,329)	91.1% (1,199)
Human herpesvirus	98.4% (26,198)	90.4% (1,190)
Human herpesvirus 6	98.4% (26,190)	90.4% (1,189)
Human herpesvirus 6B	89.3% (23,781)	81.4% (1,071)
Human herpesvirus 6B strain Z29	72% (19,180)	64.5% (849)
Human herpesvirus 6B strain Z29 only†	63.2% (16,825)	55.5% (730)

Note: all NGS reads that were aligned to HHV6B virus (gi:9633069) by *bwa*, were aligned against the entire nucleotide database of NCBI using the blastn algorithm.
*, a total of 27,932 NGS reads were available, 26,616 for sample SRR987641 and 1,316 reads for SRR1105833

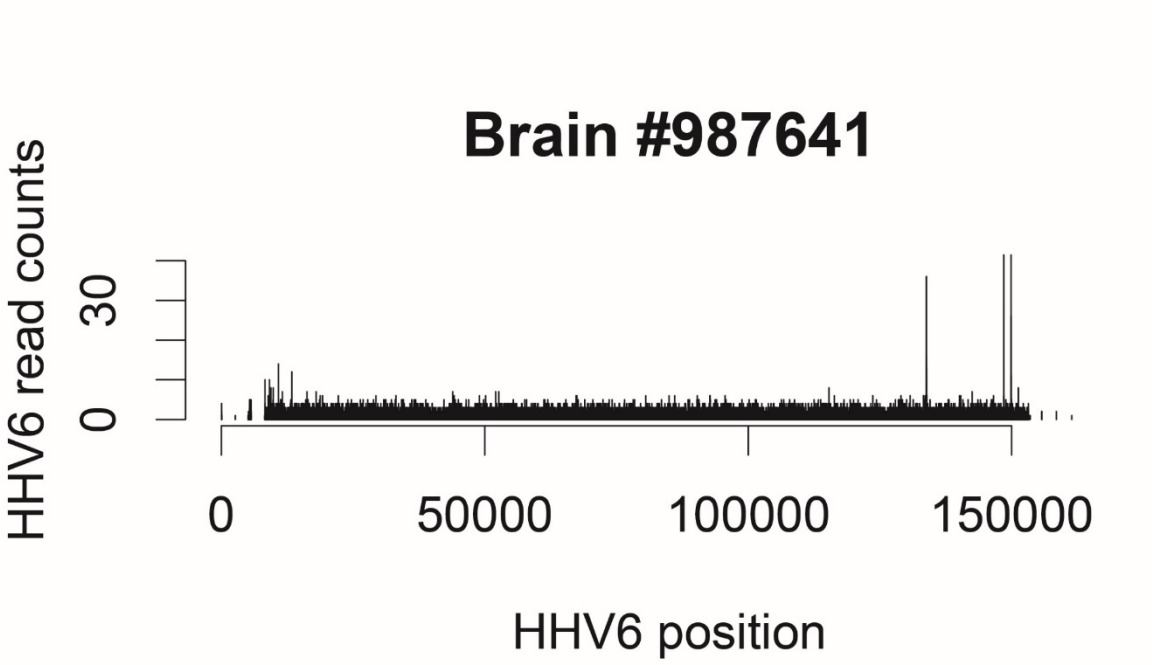


Figure 1: Brain sample SRR987641 contains sequencing reads that align to the entire genome of HHV6 reference genome. All physical positions where paired-end reads 'anchored' on the reference viral genome were collected and plotted into the histogram. It is clear from the figure that each position on the reference genome is 'anchored' around times by a paired end read.

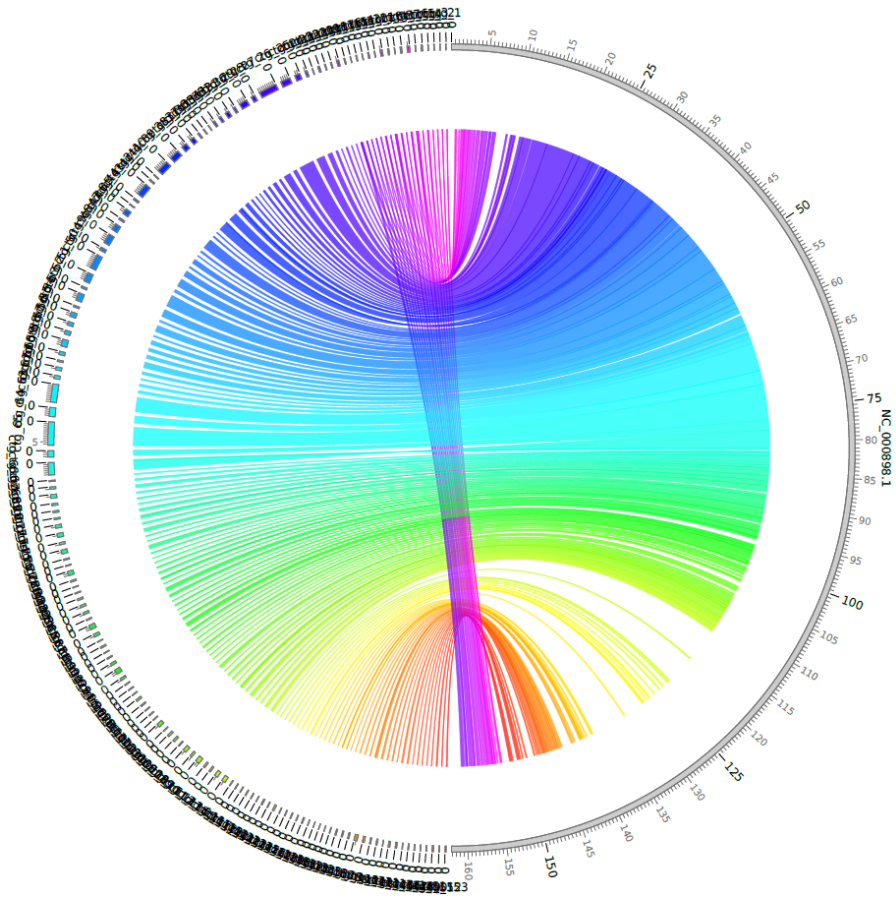


Figure 2: The mapping of contigs built using the 16,825 uniquely mapped reads (Table 1) to the HHV6 reference genome. The left side of the circle represents the 154 contigs and the right side represents the reference genome. The rainbow-colored ribbons indicate the position in the reference genome where each the contigs align.