# Impact of Variant-level Batch Effects on Identification of Genetic Risk Factors in Large Sequencing Studies

**Daniel P Wickland**
Mayo Clinic Department of Health Sciences Research
**Yingxue Ren**
Mayo Clinic Department of Health Sciences Research
**Jason P Sinnwell**
Mayo Clinic Department of Health Sciences Research
**Joseph S Reddy**
Mayo Clinic Department of Health Sciences Research
**Cyril Pottier**
Mayo Clinic's Campus in Florida
**Vivekananda Sarangi**
Mayo Clinic Department of Health Sciences Research
**Minerva M Carrasquillo**
Mayo Clinic's Campus in Florida
**Owen A Ross**
Mayo Clinic's Campus in Florida
**Steven G Younkin**
Mayo Clinic's Campus in Florida
**Nilüfer Ertekin-Taner**
Mayo Clinic's Campus in Florida
**Rosa Rademakers**
Mayo Clinic's Campus in Florida
**Matthew E Hudson**
University of Illinois at Urbana-Champaign
**Liudmila Sergeevna Mainzer**
University of Illinois at Urbana-Champaign
**Joanna M Biernacka**
Mayo Clinic Department of Health Sciences Research
**Yan Asmann** ( asmann.yan@mayo.edu )
Mayo Clinic https://orcid.org/0000-0002-8896-2647

**Research article**

**Keywords:** Batch effects, Exome capture, Alternative allele fraction

**Posted Date:** August 13th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-44710/v1

MAYO CLINIC

Mayo Clinic Jacksonville
4500 San Pablo Road
Jacksonville, Florida 32224

July 16, 2020

Dear Editors,

I am pleased to submit an original research article entitled, "***Impact of variant-level batch effects on identification of genetic risk factors in large sequencing studies***", for consideration for publication in *BMC Genetics*.

Large sequencing based genetic studies typically include samples collected by different protocols, sequenced at different sequencing centers, and/or processed by different exome capture kits. These batch effects are conventionally controlled by including batch covariates in the association models. In this manuscript, we demonstrated that this approach is not effective, and observed that there were substantial batch differences remaining at variant level in variant qualities and characteristics. Current genome/exome-wide association analyses of sequencing variants treat all variants equally as long as the variants pass a pre-defined QC threshold. This approach might be problematic. Using the Alzheimer's Disease Sequencing Project (ADSP) exome dataset of 9,904 cases and controls, we studied and profiled the variant-level differences between sample batches of three sequencing centers and two exome capture kits that were not removed by current approaches.

More importantly, we demonstrated that automatically filtering out variants with batch differences may lead to false negatives if the batch discordances largely came from quality differences and if the variants from one batch had better qualities. We found that one of the two exome capture kits used by ADSP was more effective at capturing the alternative alleles, which resulted in exome-wide significant associations for a set of variants in the entire cohort even though the signals were contributed only by samples processed by this kit. This set of variants should not be dismissed simply due to the lack of signals from samples processed by a technically deficient exome capture kit.

We believe that this manuscript is appropriate for publication by *BMC Genetics* because our findings suggest that the approaches to deal with batch effects for genetic association analyses of sequencing variants should be different from those utilized in the SNP-chip based GWAS. Additional investigations in variant-level batch differences should be conducted, and the source of batch differences should be examined before dismissing all variants with batch discordance. This manuscript is likely of great interest to the readers of *BMC Genetics* because of the large number of on-going sequencing-based genetic studies that unavoidably need to deal with similar challenges.

This manuscript has not been published and is not under consideration for publication elsewhere. We have no conflicts of interest to disclose. All authors have approved the manuscript.

Thank you for your consideration!

Sincerely,

Yan Asmann, Ph.D.

Associate Professor of BioMedical Informatics, Consultant
Mayo College of Medicine
Division of Biomedical Statistics and Informatics
Department of Health Sciences Research
Mayo Clinic

# Impact of variant-level batch effects on identification of genetic risk factors in large sequencing studies

Short title: Variant-level batch effects in sequencing studies

Daniel P. Wickland[1,2], Yingxue Ren[1], Jason P. Sinnwell[3], Joseph S. Reddy[1], Cyril Pottier[4], Vivekananda Sarangi[3], Minerva M. Carrasquillo[4], Owen A. Ross[4,5], Steven G. Younkin[4], Nilüfer Ertekin-Taner[4,6], Rosa Rademakers[4], Matthew E. Hudson[2,7], Liudmila Sergeevna Mainzer[2,7], Joanna M. Biernacka[3], Yan W. Asmann[1]*

[1] Department of Health Sciences Research, Mayo Clinic, Jacksonville, FL 32224, USA

[2] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[3] Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA

[4] Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA

[5] Department of Clinical Genomics, Mayo Clinic, Jacksonville, FL 32224, USA

[6] Department of Neurology, Mayo Clinic, Jacksonville, FL 32224, USA

[7] Carl R. Woese Institute for Genomic Biology, Carver Biotechnology Center and Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

* Corresponding author

E-mail: Yan.Asmann@mayo.edu

# ABSTRACT

## Background

Genetic studies have shifted to sequencing-based rare variants discovery after decades of success in identifying common disease variants by Genome-Wide Association Studies using Single Nucleotide Polymorphism chips. Sequencing-based studies require large sample sizes for statistical power but often inadvertently introduce batch effects because samples are typically collected, processed, and sequenced at multiple centers. Conventionally, batch effects are first detected and visualized using Principal Components Analysis and then controlled by including batch covariates in the disease association models. For sequencing-based genetic studies, because all variants included in the association analyses have passed quality control measures, this conventional approach treats every variant as equal and ignores the substantial differences still remaining in variant qualities and characteristics such as genotype quality scores, alternative allele fractions (fraction of reads supporting alternative allele at a variant position) and sequencing depths.

## Results

In the Alzheimer's Disease Sequencing Project (ADSP) exome dataset of 9,904 cases and controls, we discovered hidden variant-level differences between sample batches of three sequencing centers and two exome capture kits. Although sequencing centers were included as a covariate in our association models, we observed differences at the variant level in genotype quality and alternative allele fraction between samples processed by different exome capture kits that significantly impacted both the confidence of variant detection and the identification of disease-associated variants. Furthermore, we found that the association signals of a subset of top disease risk variants came exclusively from samples processed by one exome capture kit that was more effective at capturing the alternative alleles compared to the other kit.

## Conclusions

Our findings highlight the importance of additional variant-level quality control for large sequencing-based genetic studies. More importantly, we demonstrate that automatically filtering out variants with batch

differences may lead to false negatives if the batch discordances came largely from quality differences and if the variants from one batch had better quality scores.


## KEYWORDS

Batch effects, Exome capture, Alternative allele fraction


## BACKGROUND

Genetic studies have shifted from Single Nucleotide Polymorphism (SNP) chip-based genome-wide association study (GWAS) of common variants to rare variants discovery by exome and whole-genome sequencing. The large samples required for statistical power in sequencing-based searches for rare disease-associated variants often inadvertently introduce batch effects and systematic biases. Batch effects refer to sources of variation arising not from the targeted biological differences between sample classes but from differences between experimental or technological groups of samples [1]. If not adequately addressed in the analysis, batch effects reduce statistical power and lead to both false-positive and false-negative associations. Practices that may introduce batch effects include dividing samples among multiple sequencing centers, collecting samples under different protocols, and extracting exomes using different target capture kits. For example, the Alzheimer's Disease Sequencing Project (ADSP) sequenced exomes of more than 10,000 cases and controls to identify genetic factors associated with Alzheimer's disease (AD) [2]. Sequencing of ADSP samples took place at three centers: Broad Institute (Broad), the McDonnell Genome Institute at Washington University (WashU), and the Human Genome Sequencing Center at Baylor College of Medicine (Baylor). Broad prepared sequencing libraries using the Illumina Rapid Capture Exome kit, while WashU and Baylor used the Roche NimbleGen VCRome v2.1 kit.

The standard approach to manage batch effects is first to identify and visualize batch effects using Principal Components Analysis (PCA) of the variant genotypes, and then to adjust association models using batch

77 covariates [3–6]. However, this approach, which was developed during the SNP chip GWAS era, may not be
78 sufficient for large genetic studies by sequencing [7]. The qualities and characteristics of variants identified by
79 sequencing vary substantially even when all variants included in the association analysis have passed quality
80 control thresholds such as those defined by the Variant Quality Score Recalibration (VQSR) model from the
81 Genome Analysis Toolkit (GATK) [8]; some quality-related characteristics differ between samples at the single-
82 variant level. For example, exome kit capture efficiency may differ between the reference and alternative
83 alleles, and the template amplification and sequencing chemistry differ between variants located in GC-rich
84 and AT-rich regions. This variation will lead to differences in depth of coverage (DP), genotype quality (GQ)
85 and alternative allele fraction (AAF; fraction/percentage of reads supporting the alternative allele), etc.
86 Because the variants that pass VQSR are all treated equally for disease association analysis, the impact of
87 differences in individual variant-level quality and characteristics is often masked.
88
89 We studied the ADSP cohort of 9,904 exomes and found that conventional PCA of the genotypes did not
90 reveal the magnitude of the batch differences between samples sequenced at three centers and processed by
91 two exome target capture kits. Furthermore, genetic association analyses that adjusted for center batches did
92 not sufficiently remove hidden batch differences at the variant level. We found differences between capture kits
93 in both variant GQ and AAF that significantly impacted the identification of AD-associated risk variants. Our
94 findings highlight the importance of additional variant-level quality control to help researchers find truly
95 meaningful genetic variants that are masked by batch effects.
96
97 # RESULTS
98 ## Description of dataset
99 The Sequence Read Archive (SRA) files containing the raw sequencing data of 10,993 AD cases and controls
100 were downloaded from dbGaP (https://www.ncbi.nlm.nih.gov/gap/) and converted to FASTQ read files using
101 the SRA Toolkit (https://www.ncbi.nlm.nih.gov/books/NBK158899). Access to this public dataset was approved
102 by dbGaP and the Institutional Review Boards of Mayo Clinic and University of Illinois. The Alzheimer's cases

satisfied the National Institute on Aging and the Alzheimer's Association criteria [9] for definite, possible or probable Alzheimer's disease. These cases included patients with and without *APOE* [10] risk alleles. The controls were at least 60 years old, showed no sign of dementia based on cognitive testing, and scored low on risk assessment [2]. Of the 10,993 samples, 9,904 passed sample-level quality control (QC) [11] based on the following criteria: variant call rate $\geq$ 95% per sample for SNPs and 90% for INDELs; coverage $\geq$ 10x for at least 90% of variants; *APOE* genotype match between cohort meta-data and sequenced genotypic data; average transition/transversion ratio $\geq$ 2.75; FREEMIX sample contamination estimate < 0.02 [12]; sex check (PLINK F estimate $\geq$ 0.7 for males and $\leq$ 0.3 for females); lack of first-, second- or third-degree relatedness as defined by the KING-robust algorithm [13]; and removal of duplicate samples. Samples were sequenced at three centers: Broad Institute (Broad; 4,427 samples passing QC), Washington University (WashU; 3,260 samples passing QC), and Baylor College of Medicine (Baylor; 2,217 samples passing QC). Of the 1,584,609 variants detected in 9,904 ADSP exomes, 166,947 variants passed VQSR and the additional filtering steps detailed in the methods section.

## Population substructure explains sample clusters in PCA

The 9,904 ADSP subjects that passed sample-level QC were homogeneous based on 99.8% reporting European ancestry [11,14] and on comparison to 1000 Genomes reference samples (**Figure S1**). However, PCA of the ADSP genotypes using a pruned set of high-quality common variants (see methods section) identified multiple sample clusters (**Fig. 1A**) despite the homogeneous European ancestry of the study subjects. To determine whether the observed sample clusters represented population substructure or batch effects, we performed a *de novo* estimate using Admixture [15] that identified 9 sub-populations. As shown in **Fig. 1B**, these 9 sub-populations overlapped with and accounted for the sample clusters identified in PCA. The sample batches of different sequencing centers were not clearly visible by PCA, nor were gender or case/control status (**Fig. 1C-E**). Therefore, the sample clusters visualized by PCA can be explained largely by population substructure alone, at first glance. To control for this substructure, eigenvectors from the first four principal components were included as covariates in the association analysis.

## Association analysis

Of the 1,584,609 variants detected in 9,904 exomes, 166,947 variants passed filtering criteria for association analysis detailed in the methods section (**Figure S2**). Association analysis was conducted using four models, including those described in a previous ADSP report [14], with different combinations of the following covariates: sequencing center, the first four principal components from PCA underlying population substructure, sex, age and *APOE* genotype (**Fig. 2**). Combined, these four association models identified 52 SNPs associated with AD with exome-wide significance, including 8 SNPs reported by the ADSP consortium [14] and 7 SNPs (3 reported by ADSP) in the known AD genes *APOE* and *TOMM40* (**Figure S3**). In this paper, we focus on 1) the 29 "novel" SNPs that remained significant after adjustment for *APOE* and sex and that have not previously been linked to AD, as well as on 2) the 7 SNPs in *APOE* and *TOMM40* as positive controls (**Table 1**).

**Table 1**. Seven SNPs in *APOE* and *TOMM40* (indicated by * of the SNP IDs) and 29 novel SNPs reaching exome-wide significance ($p < 3.0 \times 10^{-7}$, Bonferroni-corrected cutoff of $p < 0.05$ / # tests): Minor allele frequency (MAF) in cases and controls, MAF in controls processed by Illumina or NimbleGen exome capture kit, and MAF in Non-Finish European (NFE) cohort of the ExAC database (http://exac.broadinstitute.org/).

| Chr | Position | Ref | Alt | Association p-value | Gene | SNP ID | Cases MAF | Controls MAF | Controls MAF, Illumina | Controls MAF, NimbleGen | ExAC AAF, NFE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 45411941 | T | C | 2.44E-185 | APOE | rs429358* | 0.2293 | 0.0701 | 0.074 | 0.067 | 0.1504 |
| 19 | 45396144 | C | T | 4.67E-103 | TOMM40 | rs11556505* | 0.2023 | 0.0879 | 0.085 | 0.090 | 0.0875 |
| 19 | 45395714 | T | C | 2.39E-75 | TOMM40 | rs157581* | 0.2766 | 0.1629 | 0.165 | 0.162 | 0.2326 |
| 19 | 45412079 | C | T | 1.79E-48 | APOE | rs7412* | 0.042 | 0.0988 | 0.116 | 0.087 | 0.0813 |
| 15 | 75913319 | T | G | 2.72E-35 | SNUPN | rs1004285543 | 0.0825 | 0.0255 | 0.063 | 0.001 | NA |
| 17 | 25973604 | A | C | 1.43E-34 | LGALS9 | rs761436847 | 0.0823 | 0.0256 | 0.056 | 0.006 | 0.0906 |
| 6 | 36979483 | T | G | 1.65E-29 | FGD2 | rs769719224 | 0.0998 | 0.0404 | 0.072 | 0.019 | 0.0294 |
| 14 | 99976645 | A | C | 7.49E-28 | CCNK | rs745936510 | 0.0899 | 0.0348 | 0.068 | 0.013 | 0.0713 |
| 13 | 114188430 | C | T | 2.22E-27 | TMCO3 | rs77834374 | 0.1068 | 0.0459 | 0.080 | 0.024 | 0.1336 |
| 14 | 99976639 | G | C | 3.67E-26 | CCNK | rs778243462 | 0.0932 | 0.0372 | 0.071 | 0.015 | 0.0808 |
| 17 | 25973598 | A | C | 1.10E-21 | LGALS9 | rs760143837 | 0.0472 | 0.014 | 0.033 | 0.002 | 0.0243 |
| 14 | 77706020 | A | C | 6.28E-21 | TMEM63C | rs774212969 | 0.0577 | 0.019 | 0.036 | 0.008 | 0.009 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | 45397229 | G | A | 9.62E-21 | TOMM40 | rs1160983* | 0.0173 | 0.0416 | 0.045 | 0.039 | 0.0718 |
| 11 | 117280516 | A | C | 2.48E-20 | CEP164 | rs756182128 | 0.0748 | 0.0296 | 0.050 | 0.016 | 0.081 |
| 3 | 42739737 | T | G | 9.73E-20 | HHATL | rs763168412 | 0.0539 | 0.0182 | 0.034 | 0.008 | 0.0919 |
| 3 | 48451952 | A | C | 2.69E-19 | PLXNB1 | rs770786389 | 0.0562 | 0.02 | 0.037 | 0.009 | 0.0255 |
| 19 | 45397307 | C | T | 1.48E-18 | TOMM40 | rs112849259* | 0.0308 | 0.0107 | 0.005 | 0.014 | 0.0011 |
| 12 | 56622883 | A | C | 1.17E-17 | NABP2 | rs757798976 | 0.0714 | 0.0301 | 0.054 | 0.014 | 0.0476 |
| 2 | 85662149 | A | C | 4.27E-16 | SH2D6 | rs748669078 | 0.068 | 0.0309 | 0.044 | 0.022 | 0.0026 |
| 19 | 10946797 | G | C | 2.51E-15 | TMED1 | rs767166604 | 0.0421 | 0.0128 | 0.029 | 0.002 | 0.0007 |
| 14 | 105932775 | G | C | 3.14E-15 | MTA1 | rs782227993 | 0.0627 | 0.0259 | 0.047 | 0.012 | 0.0208 |
| 6 | 29429950 | A | C | 3.62E-15 | OR2H1 | rs746691570 | 0.0402 | 0.0132 | 0.022 | 0.007 | 0.0207 |
| 11 | 117280522 | A | C | 3.35E-14 | CEP164 | rs758240656 | 0.0529 | 0.0198 | 0.037 | 0.009 | 0.0768 |
| 2 | 85662154 | A | C | 3.97E-14 | SH2D6 | rs760146451 | 0.0617 | 0.0288 | 0.040 | 0.021 | 0.0018 |
| 13 | 88330245 | A | C | 1.13E-13 | SLITRK5 | rs773717935 | 0.0277 | 0.0065 | 0.014 | 0.002 | 3.1E-05 |
| 19 | 45409167 | C | G | 9.66E-13 | APOE | rs440446* | 0.3332 | 0.3817 | 0.361 | 0.395 | 0.4346 |
| 19 | 10946802 | T | C | 3.07E-12 | TMED1 | rs776909029 | 0.0366 | 0.0117 | 0.028 | 0.001 | 0.0009 |
| 9 | 34564740 | A | C | 3.57E-12 | CNTFR | rs774039930 | 0.0516 | 0.0222 | 0.039 | 0.011 | 0.0008 |
| 3 | 108474687 | T | G | 5.77E-12 | RETNLB | rs199707443 | 0.0328 | 0.0107 | 0.025 | 0.001 | 0.0493 |
| 19 | 43025485 | T | G | 7.69E-12 | CEACAM1 | rs763190977 | 0.0921 | 0.0523 | 0.107 | 0.016 | 0.0026 |
| 3 | 31659462 | A | T | 9.21E-12 | STT3B | rs74346226 | 0.0891 | 0.0514 | 0.076 | 0.035 | 0.131 |
| 12 | 109719316 | T | G | 2.11E-10 | FOXN4 | rs760573591 | 0.0309 | 0.0115 | 0.025 | 0.003 | 1.5E-05 |
| 8 | 145112936 | T | C | 5.02E-10 | OPLAH | rs781948612 | 0.0331 | 0.0114 | 0.026 | 0.002 | 0.0364 |
| 19 | 42799299 | T | C | 1.39E-09 | CIC | rs745695673 | 0.019 | 0.0043 | 0.011 | 0.000 | 0 |
| 13 | 111164389 | A | C | 1.61E-08 | COL4A2 | rs199702442 | 0.0517 | 0.0274 | 0.041 | 0.018 | 0.0285 |
| 22 | 30951295 | T | G | 1.80E-08 | GAL3ST1 | rs762634521 | 0.0204 | 0.0056 | 0.013 | 0.001 | 0.028 |

## Batch effects between exome capture kits among AD-associated variants

Additional analyses of the 36 SNPs listed in **Table 1** revealed that the significance of association came more from the individuals sequenced at Broad using the Illumina exome capture kit than from those sequenced at WashU and Baylor using the NimbleGen exome capture kit. Association analyses of center-specific cohorts demonstrated that this center bias occurred in previously known AD SNPs in *APOE* and *TOMM40* (**Fig. 3A)** as well as in the 29 novel SNPs (**Fig. 3B**). The significance of AD association of the 29 novel SNPs came exclusively from the Broad cohort processed using the Illumina exome kit (**Fig. 3B**), which may explain why the ADSP consortium did not report these SNPs.

We next studied these 29 novel SNPs using PCA. The SNP genotypes showed no differences between sub-populations or genders (**Fig. 4A-B**). However, we observed a clear and significant difference between samples captured by the Illumina kit (Broad) *vs.* those captured by the Roche NimbleGen kit (WashU and Baylor) (**Fig.**

**4C**), consistent with the observation that these SNPs are only associated with AD in Broad samples as shown in **Fig. 3B**. Since all our association models included sequencing centers as a covariate, the batch differences visualized in **Fig. 4C** clearly were due to other factors, probably at the individual-variant level.

## Identification of variant-level differences in genotype quality and alternative allele fraction between two exome capture kits

As the PCA plot of the 29 novel SNPs showed significant batch differences between samples processed by two exome capture kits, we next examined variant-level factors that could explain why exclusively the Illumina kit-captured exomes yielded this set of highly significant novel SNPs. Although all analyzed SNPs passed VQSR quality thresholds, we speculated that there remained significant differences in qualities and characteristics at the individual-variant level between exome kits that might explain the greater significance of association in samples captured by one exome kit *vs.* the other. First, we compared several quality measures of the variants called in cohorts captured by the two kits. For the 166,947 SNPs used in the association analyses, we studied the distribution of the three most common variant quality parameters: GQ, DP and AAF (**Fig. 5**). The overall distributions of mean GQ, DP, and AAF were very similar between capture kits. For the 29 SNPs of interest, the GQ and DP were also similar between capture kits. Interestingly, we observed a bi-modal distribution of AAF values from both capture kits, with the first mode closer to zero indicating approximately equal percentage of reads supporting reference and alternative alleles at the variant positions, and a second mode significantly deviated from zero and centered around log2 value of -6. This second mode represented variants with significantly lower percentage of reads supporting the alternative alleles than the reference alleles. More importantly, the AAF of the 29 SNPs captured by the Illumina exome kit are mostly located close to the first mode and have a more balanced ratio of reference to alternative allele-supporting reads, while the AAF of the 29 novel SNPs captured by the NimbleGen kit are located approximately around the second mode. This observation implies that the NimbleGen kit lost most of the alternative alleles during exome capture, leading to the lack of variant calls in the NimbleGen cohort. Indeed, when we examined the population minor allele frequencies (MAF) of these 29 SNPs in the non-Finish European (NFE) population of the ExAC database (http://exac.broadinstitute.org/) (**Table 1**; **Fig. 6**), the MAFs in the ADSP control cohort

captured by the Illumina exome kit are generally higher compared to those in the NimbleGen-captured control cohort (**Fig. 6**), and the MAFs from the Illumina-processed controls are closer to the MAFs observed in ExAC NFE. This observation supports our hypothesis of more effective capture of the alternative alleles by the Illumina kit.

To further investigate variant quality differences between capture kits beyond the top 29 SNPs, we calculated distributions of log2 ratios of mean GQ, DP, and AAF between capture kits for all 166,947 variants (**Figure S4**). We divided these variants into two groups: (1) 10% of the variants with the largest differences in GQ, DP, or AAF between two capture kits (5% at each of the two tails); and (2) the remaining 90% of variants. The PCA plots showed that the 10% of variants with the biggest quality differences clearly contributed to the batch differences between the two exome capture kits (**Fig. 7A)**. Intriguingly, dividing these SNPs into those with better quality in Illumina-captured samples (right tail of the 5% variants; **Fig. 7B**) and with better quality in NimbleGen-captured samples (left tail of the 5% variants; **Fig. 7C**) demonstrated that the batch differences came mostly from the latter. We observed no obvious batch contribution from the 90% of variants with similar GQ, DP, or AAF (**Fig. 7D**).

## DISCUSSION

We analyzed 9,904 exomes from ADSP to study the impact of batch effects on variant calling and association analysis. We focused on the known batches of three different sequencing centers and two exome capture kits. At first, no batch differences between sequencing centers or exome capture kits were visible from the PCA plots, and visually we were able to attribute all sample clusters in PCA to population substructure (**Fig. 1B**). Sequencing center was included as a covariate in all association models in our analyses; however, exome capture kit was not included in the association models because it confounds with sequencing center and therefore is not an independent variable. Our association analyses identified 29 novel SNPs in addition to variants in previously known AD genes including *APOE* and *TOMM40* as well as variants reported by the ADSP consortium [14]. These 29 SNPs exemplified the impact of batch effects that were later attributed to the

differences between the two exome capture kits. First, the significance of association with AD came exclusively from the samples processed by the Illumina exome kit at Broad, while those processed by the NimbleGen kit at WashU or Baylor lacked significance (**Fig. 3B**). Second, PCA of these 29 SNPs clearly showed the separation of samples according to capture kit (**Fig. 4C**). The qualities of the variants between the two captured kits were similar overall at first glance but significantly different for a substantial number of variants (**Fig. 5,** and **Fig. 7**). The fact that the majority of the variants had similar quality measures suggested that batch effects impacted individual variants rather than all variants.

Further analyses of the 29 SNPs of interest (**Fig. 5**) pinpointed the difference of AAF (Alternative Allele Fraction; percent of reads supporting alternative allele) between the Illumina- and NimbleGen- processed cohorts. It is clear that the NimbleGen kit did not effectively capture the alternative alleles for these 29 SNPs, which implies that the variant calling of Illumina-captured was more reliable, and that the AD association of these 29 SNPs unique to the Illumina exome kit is likely real. We speculate that in previous publications the ADSP consortium applied an undisclosed filtering step to exclude variants with discordant significance between cohorts processed at different sequencing centers and/or by exome capture kits, which would explain why these 29 SNPs have not been reported (e.g. [14]). Additional investigation of the genes connected to these 29 SNPs supports this speculation. Multiple genes underlying these 29 SNPs have previously been linked to neural function or pathology, including *Transmembrane P24 Trafficking Protein 1* (*TMED1*; rs767166604, association $p = 7.22 \times 10^{-15}$ for Illumina kit cohort; rs776909029, $p = 3.25 \times 10^{-11}$ ), *Plexin B1* (*PLXNB1*; rs770786389, $p = 7.49 \times 10^{-23}$), *Capicua Transcriptional Repressor* (*CIC*; rs745695673, $p = 3.70 \times 10^{-9}$), *Centrosomal Protein 164* (*CEP164*; rs756182128, $p = 5.66 \times 10^{-25}$; rs758240656, $p = 1.94 \times 10^{-16}$), and *Cyclin K* (*CCNK;* rs745936510, $p = 3.98 \times 10^{-30}$; rs778243462, $p = 1.50 \times 10^{-29}$). TMED1 is reported to interact with Amyloid Precursor Protein (*APP*) [16], whose cleavage into amyloid beta generates one of the key components of AD-associated pathological protein aggregation [17]. PLXNB1 influences amyloid beta load [18], and CIC is a transcriptional repressor whose inactivation promotes gliomagenesis, the formation of glial tumors in the brain [19]. CEP164 binds to TTBK2, a kinase that phosphorylates tau [20] and contributes to

neurodegeneration in frontotemporal dementia [21]. Finally, copy-number mutations in the transcriptional regulator *CCNK* have been associated with neurodevelopmental abnormalities [22].

## CONCLUSIONS

In summary, we discovered batch effects at the individual-variant level resulting from differences in GQ and AAF between different exome capture kits. In particular, we found that the signal of some variants most significantly associated with AD came exclusively from samples processed by one exome capture kit that was more effective at capturing the alternative alleles compared to the other kit. In general, a subset of variants with the biggest disparities in GQ and AAF values between the two exome capture kits contributed to the remaining batch effects. Our findings highlight the importance of additional variant-level quality control for large sequencing-based genetic studies. More importantly, we demonstrated that automatically filtering out variants with batch differences may lead to false negatives if the batch discordances largely came from quality differences and if the variants from one batch had better quality scores.

## METHODS

### Variant calling

The paired-end sequence reads were aligned to the human reference genome build 37 using Novoalign (http://www.novocraft.com) (default parameters), which was selected on the basis of its greater accuracy in read placement relative to other methods [23,24] and its lack of prior application to this dataset for association testing (e.g. [14,25]). The alignment files were then sorted by read position using Novosort (http://www.novocraft.com), realigned around small insertions and deletions (INDELs) using Picard (https://broadinstitute.github.io/picard/), and subjected to base recalibration using the Genome Analysis Toolkit (GATK) version 3.4 [26]. Variant calling followed GATK's best practices guidelines for germline variants (https://gatk.broadinstitute.org/hc/en-us): per-sample variant calling on the realigned, recalibrated BAM files was performed using HaplotypeCaller, and multi-sample joint genotyping of all 9,904 samples was performed using GenotypeGVCFs. Variant calling was conducted only on the exome regions common between the two

exome capture kits (Illumina Rapid Capture Exome kit and NimbleGen VCRome v2.1 kit). Variants were annotated by SnpEff [27] and ANNOVAR [28].

## Ancestry estimation

We implemented a series of filtering steps to select a pruned set of high-quality variants from the 1,584,609 variants detected across the cohort. Variants were excluded if they (1) failed VQSR; (2) lay within the highly variable HLA, LCT, 8p and 17q regions; (3) had MAF below 5% or deviation from Hardy-Weinberg Equilibrium below $p < 1.0$ E-4; or (4) had linkage disequilibrium $r^2$ value above 0.2 within 0.1 megabase sliding windows. In total, 16,187 variants passed all filters, and 12,100 variants overlapped with variants detected in the 1000 Genomes samples of known ethnicities. The genotypes of the ADSP samples were combined with those of the 1000 Genomes samples at these 12,100 loci. PCs of the genotypes were then computed and plotted using the software plinkQC (https://meyer-lab-cshl.github.io/plinkQC/). The first four PCs of the ADSP genotypes had eigenvalues above 1 and were retained as covariates in the association analyses.

## Sub-population estimation

The software Admixture [15] was used to estimate sub-population number from the ADSP data *de novo*, using the 16,187 variants described above. Sub-population *k* values between 1 and 20 were tested using default settings. A value of 9 produced the lowest cross-validation standard error, indicating that 9 sub-populations best fit the data.

## Variant-level quality control for association analyses

Several steps were undertaken to minimize the number of false-positive variant calls prior to running the association models. The Variant Quality Score Recalibration (VQSR) step implemented in GATK uses machine learning algorithms to compute new, well-calibrated quality scores for each variant based on the annotations of a high-quality subset of the analyzed data. In accordance with GATK Best Practices for whole-exome data, the variables included in the VQSR model consisted of QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR and

InbreedingCoeff for SNPs; and QD, MQRankSum, ReadPOsRankSum, FS, SOR and InbreedingCoeff for INDELs [26]. A sensitivity threshold of 99.5 was used for SNPs and 99.0 for INDELs. Detected variants were excluded from association analysis if they failed VQSR, deviated significantly ($p < 1.0 \times 10^{-6}$) from Hardy-Weinberg equilibrium (HWE) in the control samples, or had an alternative allele call supported by fewer than 10 reads across the cohort (**Fig S2**).

### Association tests, statistical model and variant filtering

After quality control, association testing using disease status (case or control) as the phenotype was performed on the variants under an additive logistic regression model implemented in Plink 1.9 [29]. Covariates included sequencing center, sex, *APOE* genotype, and the first four principal components (PCs) underlying population substructure. PCs were calculated using Plink. The association tests were conducted on all 9,904 samples together for the full-cohort analysis, as well as on sets of samples stratified by sequencing center for the center-based association analyses. Variants were considered exome-wide statistically significant at the Bonferroni-corrected threshold of $p < 0.05 / \#\text{ tests}$, as in [14]. Only bi-allelic SNPs with a recalibrated variant quality score (VQSLOD) > 0 were retained for further analysis.

## LIST OF ABBREVIATIONS

**SNP**: Single Nucleotide Polymorphism

**GWAS**: Genome-Wide Association Study

**ADSP**: Alzheimer's Disease Sequencing Project

**PCA**: Principal Components Analysis

**VQSR**: Variant Quality Score Recalibration

**GATK**: Genome Analysis Toolkit

**DP**: Sequencing Depth of Coverage

**GQ**: Genotype Quality

**AAF**: Alternative Allele Fraction

**QC**: Quality Control

**MAF**: Minor Allele Frequency

# DECLARATIONS

## Ethics approval and consent to participate

All aspects of the study were approved the Institutional Review Boards from both Mayo Clinic and University of Illinois at Urbana-Champaign. This study was also approved by dbGAP. Written informed consent obtained from all participants and surrogates were conducted by ADSP.

## Consent for publication

Not applicable

## Availability of data and materials

The exome sequencing data and participants metadata from ADSP were obtained from dbGAP with a project approval.

## Competing interests

None


## Funding

This research was supported in part by the Mayo Clinic Center for Individualized Medicine and Illinois Alliance Fellowships for Technology-Based Healthcare Research program. This research is also part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-

0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications.

## Authors' contributions

**Daniel P. Wickland:** Writing – Original Draft Preparation, Formal Analysis, Investigation, Visualization, Methodology, Software, Data Curation, Writing – Review & Editing

**Yingxue Ren:** Data Curation, Investigation

**Jason P. Sinnwell:** Data Curation, Methodology

**Joseph S. Reddy:** Data Curation

**Cyril Pottier:** Data Curation

**Vivekananda Sarangi:** Formal Analysis

**Minerva M. Carrasquillo:** Investigation

**Owen A. Ross:** Investigation

**Steven G. Younkin:** Investigation

**Nilüfer Ertekin-Taner:** Investigation

**Rosa Rademakers:** Investigation

**Matthew E. Hudson: -** Funding Acquisition, Supervision, Resources, Writing – reviewing and editing

**Liudmila Sergeevna Mainzer:** Conceptualization, Supervision, Resources, Methodology

**Joanna M. Biernacka:** Supervision, Methodology

**Yan W. Asmann:** Conceptualization, Funding Acquisition, Supervision, Resources, Writing – reviewing and editing

## Acknowledgements

We thank the Mayo Clinic Bioinformatics Core for their help to download and process all exome sequencing data.

## REFERENCES


1. Goh WW Bin, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol. 2017;35(6):498–507.
2. Beecham GW, Bis JC, Martin ER, Choi SH, DeStefano AL, Van Duijn CM, et al. Clinical/Scientific Notes: The Alzheimer's disease sequencing project: Study design and sample selection. Neurol Genet. 2017;3(5):doi:10.1212/NXG.0000000000000194.
3. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):2074–93.
4. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9.
5. Vansteelandt S, Goetgeluk S, Lutz S, Waldman I, Lyon H, Schadt EE, et al. On the adjustment for covariates in genetic association analysis: A novel, simple principle to infer direct causal effects. Genet Epidemiol. 2009;33:394–405.
6. Zhao H, Mitra N, Kanetsky PA, Nathanson KL, Rebbeck TR. A practical approach to adjusting for population stratification in genome-wide association studies: Principal components and propensity scores (PCAPS). Stat Appl Genet Mol Biol. 2018;17(6):doi:10.1515/sagmb-2017-0054.
7. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics. 2016;17(1):29–39.
8. McKenna A, Hanna M, Banks E, Sivacheko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
9. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-

Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement. 2011;7(3):263–9.

10. Corder E, Saunders A, Strittmatter W, Schmechel D, Gaskell P, Small G, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science (80- ). 1993 Aug;261:921–3.

11. Ren Y, Reddy J, Pottier C, Sarangi V, Tian S, Sinnwell J, et al. Identification of missing variants by combining multiple analytic pipelines. BMC Bioinformatics. 2018;19:doi:10.1186/s12859-018-2151-0.

12. Jun G, Flickinger M, Hetrick K, Romm J, Doheny K, Abecasis G, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet. 2012 Nov;91:839–48.

13. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–73.

14. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Mol Psychiatry. 2018;

15. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.

16. Del Prete D, Suski JM, Oulès B, Debayle D, Gay AS, Lacas-Gervais S, et al. Localization and processing of the amyloid-β protein precursor in mitochondria-associated membranes. J Alzheimer's Dis. 2017;55:1549–70.

17. Penke B, Bogár F, Fülöp L. β-amyloid and the pathomechanisms of Alzheimer's disease: A comprehensive view. Molecules. 2017;22:doi:10.3390/molecules22101692.

18. Mostafavi S, Gaiteri C, Sullivan SE, White CC, Tasaki S, Xu J, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. Nat Neurosci. 2018;21:811–9.

19. Yang R, Chen LH, Hansen LJ, Carpenter AB, Moure CJ, Liu H, et al. Cic loss promotes gliomagenesis

via aberrant neural stem cell proliferation and differentiation. Cancer Res. 2017;77(22):6097–108.

20. Liao J, Yang T, Weng R, Kuo C, Chang C. TTBK2: A tau protein kinase beyond tau phosphorylation. Biomed Res Int. 2015;doi:10.1155/2015/575170.

21. Taylor LM, McMillan PJ, Liachko NF, Strovas T, Ghetti B, Bird T, et al. Pathological phosphorylation of tau and TDP-43 by TTBK1 and TTBK2 drives neurodegeneration. Mol Neurodegener. 2018;13(7):doi:10.1186/s13024-018-0237-9.

22. Fan Y, Yin W, Hu B, Kline AD, Zhang VW, Liang D, et al. De novo mutations of CCNK cause a syndromic neurodevelopmental disorder with distinctive facial dysmorphism. Am J Hum Genet. 2018;103(3):448–55.

23. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;1303.3997(00):http://arxiv.org/abs/1303.3997.

24. Thankaswamy-Kosalai S, Sen P, Nookaew I. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics. 2017;109:186–91.

25. Patel T, Brookes KJ, Turton J, Chaudhury S, Guetta-Baranes T, Guerreiro R, et al. Whole-exome sequencing of the BDR cohort: evidence to support the role of the PILRA gene in Alzheimer's disease. Neuropathol Appl Neurobiol. 2018;44:506–21.

26. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. 2013. doi:10.1002/0471250953.bi1110s43.

27. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.

28. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):doi:10.1093/nar/gkq603.

29. Chang C, Chow C, Tellier L, Vattikuti S, Purcell S, Lee J. Second-generation PLINK: rising to the

challenge of larger and richer datasets. Gigascience. 2015;4:doi:10.1186/s13742-015-0047-8.

# FIGURES

**Fig.1. Principal Component (PC) eigenvector plots using genotypes of a pruned set of 16,187 high-quality common variants for 9,904 ADSP individuals.** Each data point represents a single individual. Clustering of samples for a particular variable signifies genotypic similarity between individuals for the trait represented by that color. (A) PCs of the genotypes. (B) PCs color coded based on sub-population. (C) PCs color coded based on center. (D) PCs color coded based on Gender. (E) PCs color coded based on AD phenotype. As expected, clustering is apparent only by sub-population.

**Fig. 2. Covariates included in each model for association tests.**

**Fig. 3. Sequencing center specific association p-values of SNPs that reached exome-wide significance (denoted by the dashed horizontal lines) in the full-dataset analysis.** (A) Seven SNPs in *TOMM40* and *APOE.* (B) Twenty-nine novel SNPs

**Fig. 4. PC eigenvector plots of genotypes at 29 exome-wide significant SNPs.** Each data point represents a single individual. Clustering of samples for a particular variable indicates genotypic similarity between individuals for the trait represented by that color. (A) PCs color coded based on sub-population. (B) PCs color coded based on gender. (C) PCs color coded based on capture kit. The NimbleGen-captured samples cluster tightly together, indicating their genotypic similarity that is distinct from the Illumina-captured samples.

**Fig. 5. Density plots of variant quality parameters between two exome capture kits.** Mean values were computed across all samples for each variant. The solid lines show the distributions of all 166,947 variants used in the association analyses, and the scattered dots represent the 29 novel SNPs.

**Fig. 6. Minor Allele Frequency (MAF) of 29 exome-wide significant SNPs in AD control exomes processed by two capture kits and in the ExAC Non-Finnish European (NFE) population.**

**Fig. 7. PC eigenvector plots of genotypes at variants lying in different sections of quality-metric ratio distributions.** Each data point represents a single individual, color coded according to capture kit. (A) PCs of variants in either 5% tail. (B) PCs of variants in right 5% tail. (C) PCs of variants in left 5% tail. (D) Variants in middle 90% of distributions. Variants in the tails, in particular the left 5% tail (better quality in NimbleGen kit), show clear separation by capture kit in both cases and controls.

Fig 1.

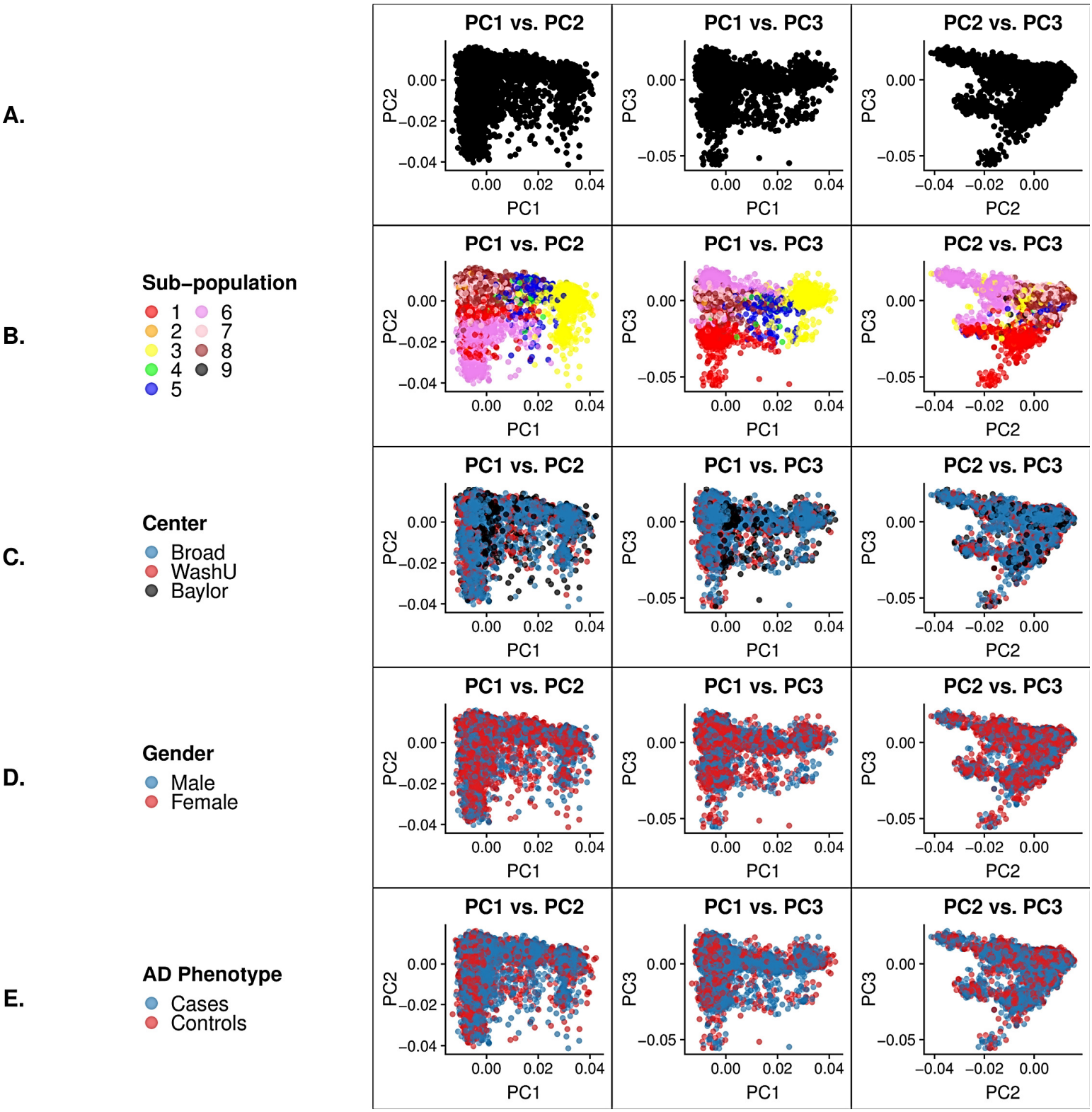

A.
B.
C.
D.
E.
PC1 vs. PC2
PC1 vs. PC3
PC2 vs. PC3
PC1
PC2
PC3
0.00
−0.02
−0.04
−0.05
0.02
0.04
−0.04
−0.02
Sub−population
1
2
3
4
5
6
7
8
9
Center
Broad
WashU
Baylor
Gender
Male
Female
AD Phenotype
Cases
Controls

Fig 2.

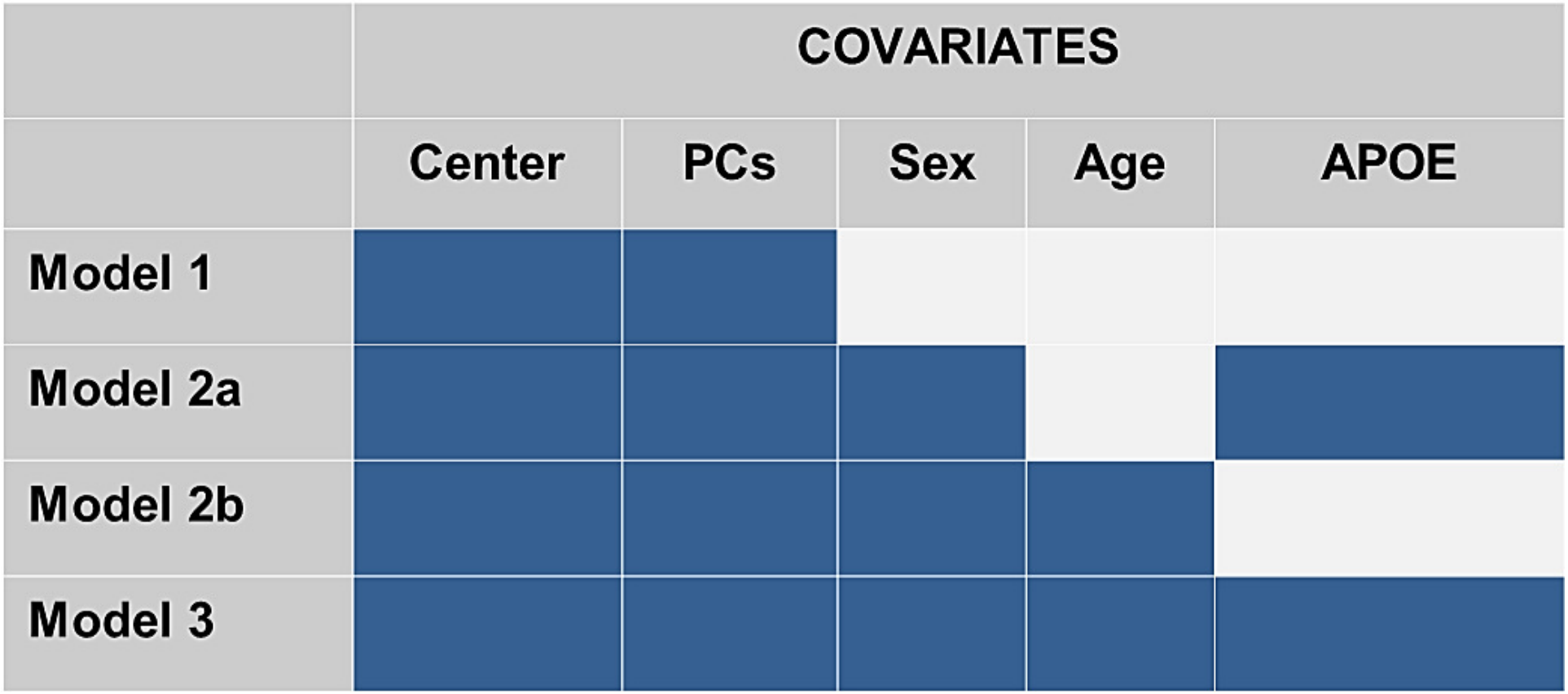

COVARIATES
Center
PCs
Sex
Age
APOE
Model 1
Model 2a
Model 2b
Model 3

Fig 3


A.
APOE / 19:45411941
TOMM40 / 19:45396144
TOMM40 / 19:45395714
APOE / 19:45412079
TOMM40 / 19:45397229
TOMM40 / 19:45397307
APOE / 19:45409167
−log10 p−value
75
50
25
0
B.
SNUPN / 15:75913319
LGALS9 / 17:25973604
FGD2 / 6:36979483
CCNK / 14:99976645
TMCO3 / 13:114188430
CCNK / 14:99976639
LGALS9 / 17:25973598
TMEM63C / 14:77706020
CEP164 / 11:117280516
HHATL / 3:42739737
PLXNB1 / 3:48451952
NABP2 / 12:56622883
SH2D6 / 2:85662149
TMED1 / 19:10946797
MTA1 / 14:105932775
OR2H1 / 6:29429950
CEP164 / 11:117280522
SH2D6 / 2:85662154
SLITRK5 / 13:88330245
TMED1 / 19:10946802
CNTFR / 9:34564740
RETNLB / 3:108474687
CEACAM1 / 19:43025485
STT3B / 3:31659462
FOXN4 / 12:109719316
OPLAH / 8:145112936
CIC / 19:42799299
COL4A2 / 13:111164389
GAL3ST1 / 22:30951295
30
20
10
0
Broad
WashU
Baylor

Fig 4.

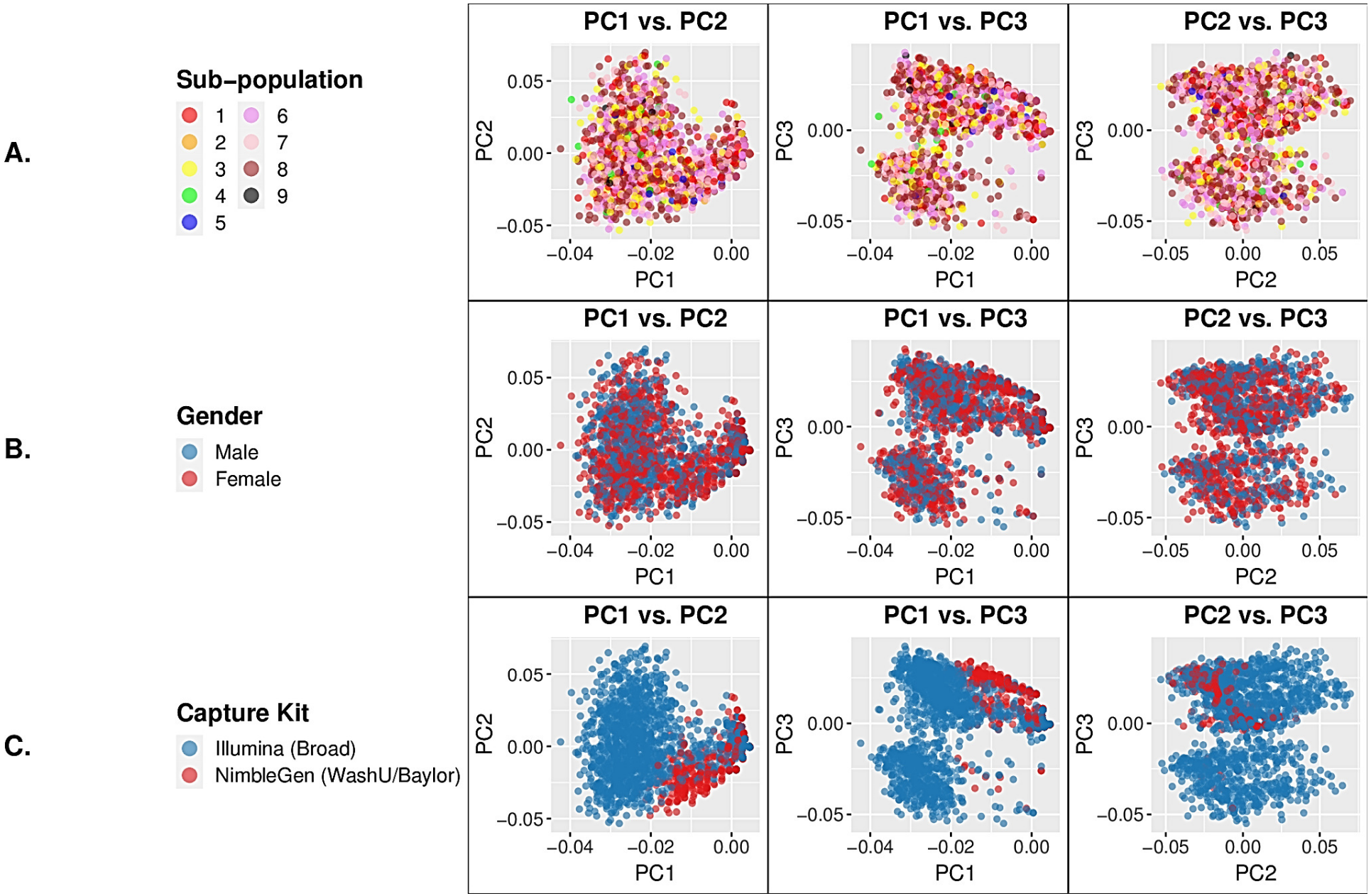

A.
Sub−population
1
2
3
4
5
6
7
8
9
B.
Gender
Male
Female
C.
Capture Kit
Illumina (Broad)
NimbleGen (WashU/Baylor)
PC1 vs. PC2
PC1 vs. PC3
PC2 vs. PC3
PC1
PC2
PC3
0.05
0.00
−0.05
−0.04
−0.02
0.00
−0.05
0.00
0.05

Fig 5.

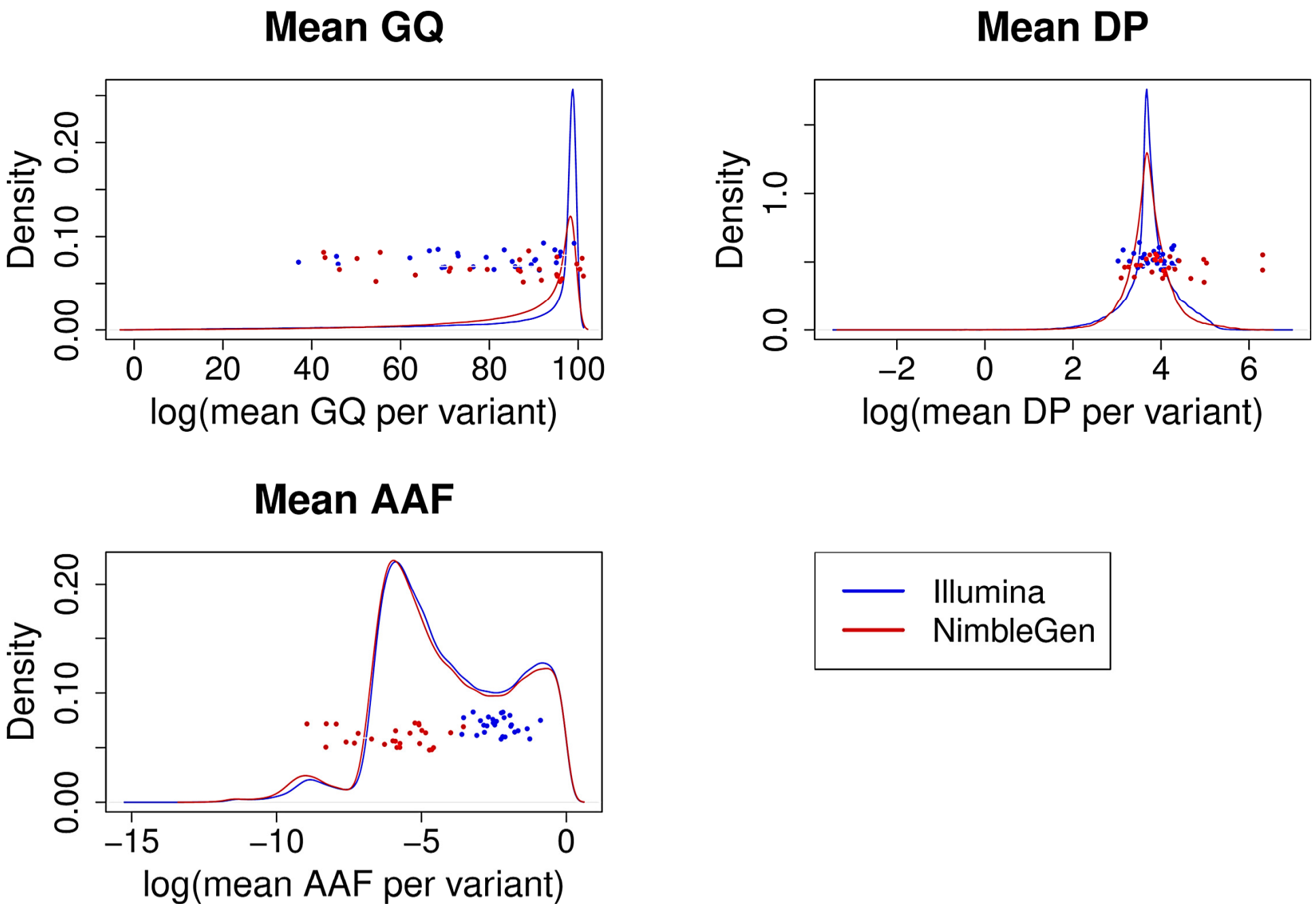
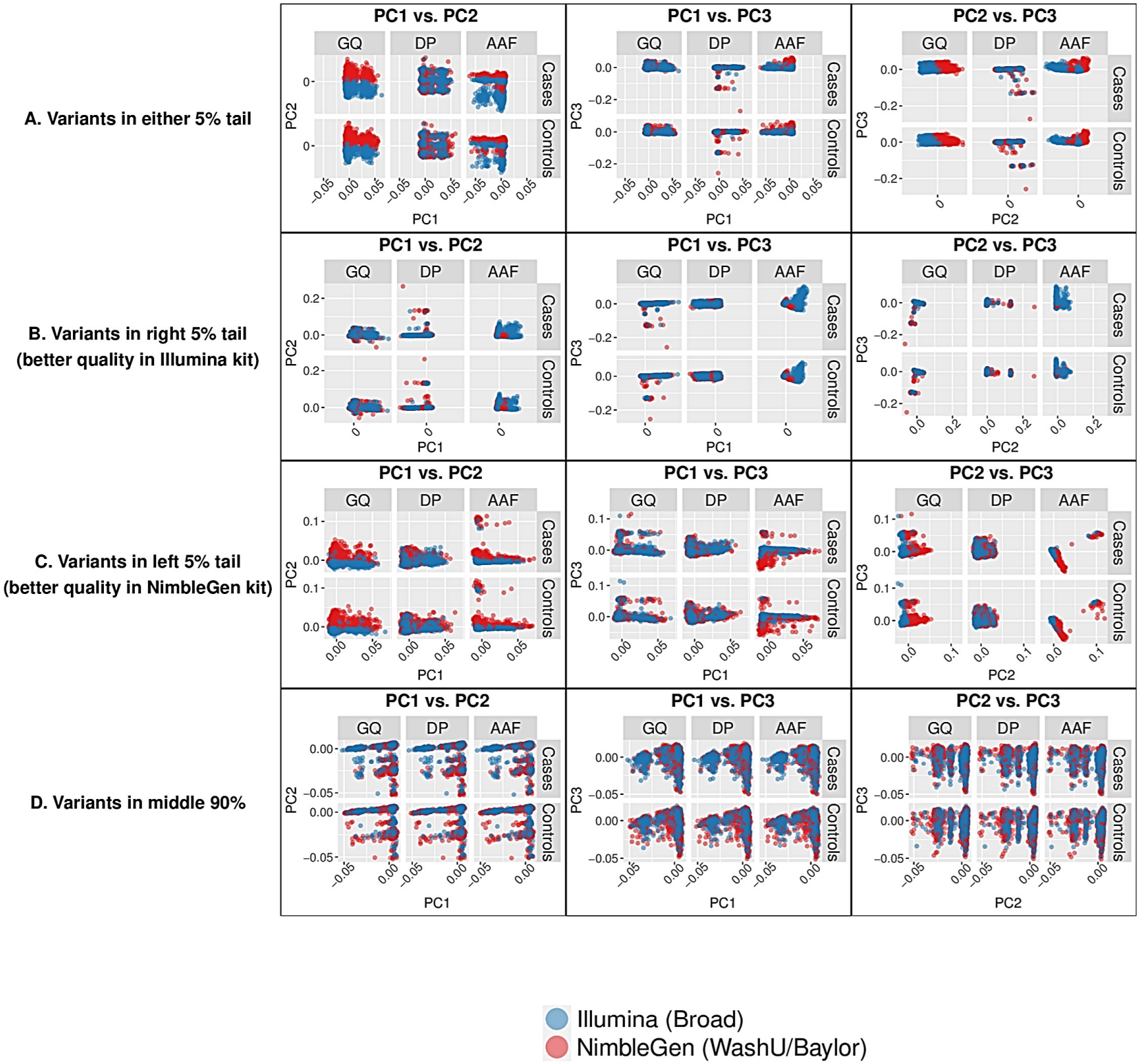

Mean GQ
Density
0.00
0.10
0.20
0
20
40
60
80
100
log(mean GQ per variant)
Mean DP
Density
0.0
1.0
−2
0
2
4
6
log(mean DP per variant)
Mean AAF
Density
0.00
0.10
0.20
−15
−10
−5
0
log(mean AAF per variant)
Illumina
NimbleGen

Fig 6.


SNUPN / 15:75913319
LGALS9 / 17:25973604
FGD2 / 6:36979483
CCNK / 14:99976645
TMCO3 / 13:114188430
CCNK / 14:99976639
LGALS9 / 17:25973598
TMEM63C / 14:77706020
CEP164 / 11:117280516
HHATL / 3:42739737
PLXNB1 / 3:48451952
NABP2 / 12:56622883
SH2D6 / 2:85662149
TMED1 / 19:10946797
MTA1 / 14:105932775
OR2H1 / 6:29429950
CEP164 / 11:117280522
SH2D6 / 2:85662154
SLITRK5 / 13:88330245
TMED1 / 19:10946802
CNTFR / 9:34564740
RETNLB / 3:108474687
CEACAM1 / 19:43025485
STT3B / 3:31659462
FOXN4 / 12:109719316
OPLAH / 8:145112936
CIC / 19:42799299
COL4A2 / 13:111164389
GAL3ST1 / 22:30951295
MAF
0.20
0.15
0.10
0.05
0.00
Illumina controls
NimbleGen controls
ExAC NFE

Fig 7.


A. Variants in either 5% tail
B. Variants in right 5% tail
(better quality in Illumina kit)
C. Variants in left 5% tail
(better quality in NimbleGen kit)
D. Variants in middle 90%
PC1 vs. PC2
PC1 vs. PC3
PC2 vs. PC3
GQ
DP
AAF
Cases
Controls
PC1
PC2
PC3
Illumina (Broad)
NimbleGen (WashU/Baylor)

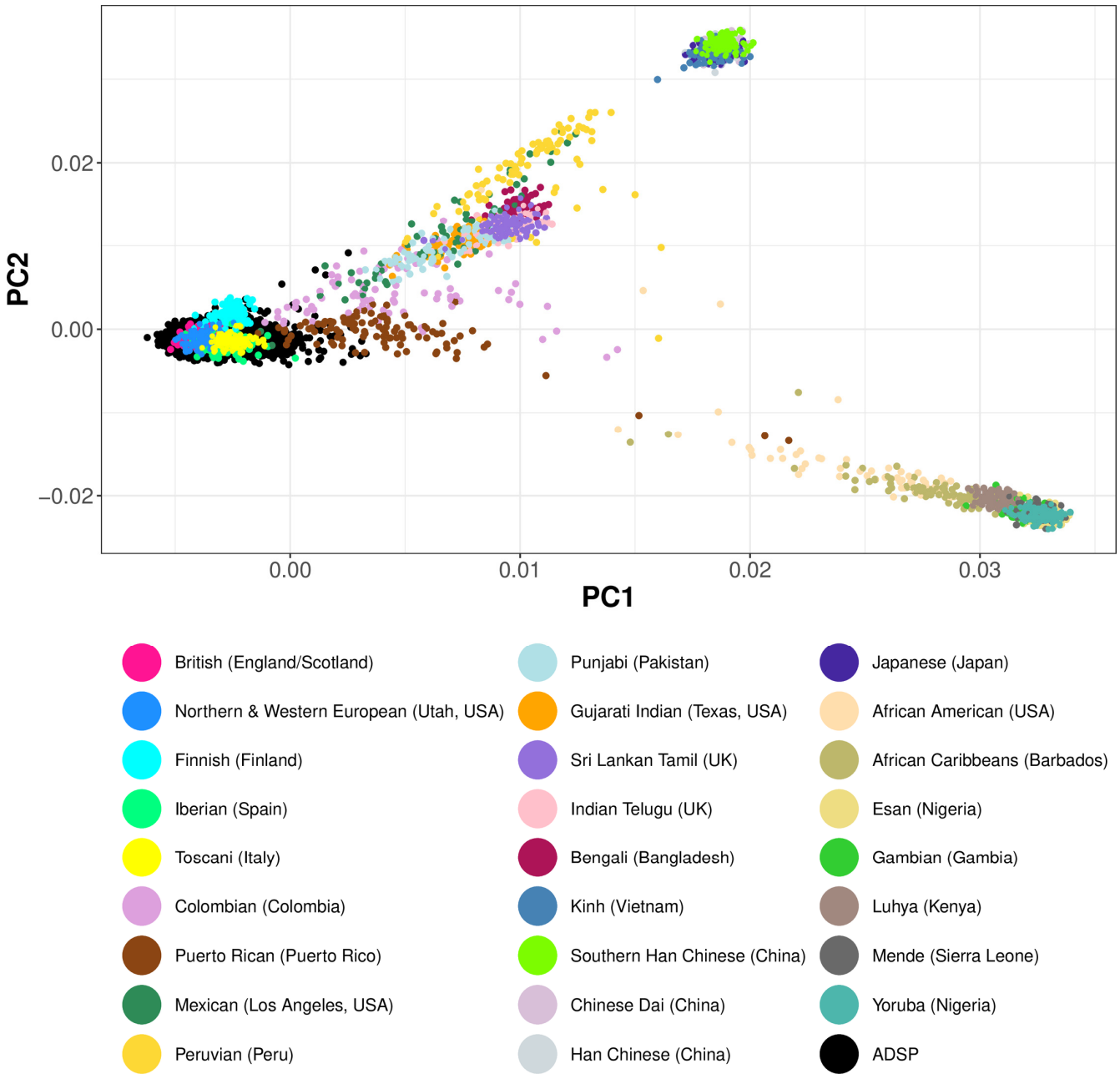



**Fig S1. Principal Component (PC) eigenvector plot of combined 1000 Genomes and ADSP genotypes.** Each data point represents a single individual. 1000 Genomes reference individuals are color-coded by ancestry. ADSP samples are shown in black. The position of ADSP samples relative to the 1000 Genomes reference samples indicates their genotypic similarity, which reflects ancestry. Most ADSP samples cluster near European reference samples (e.g. Finland and Spain).

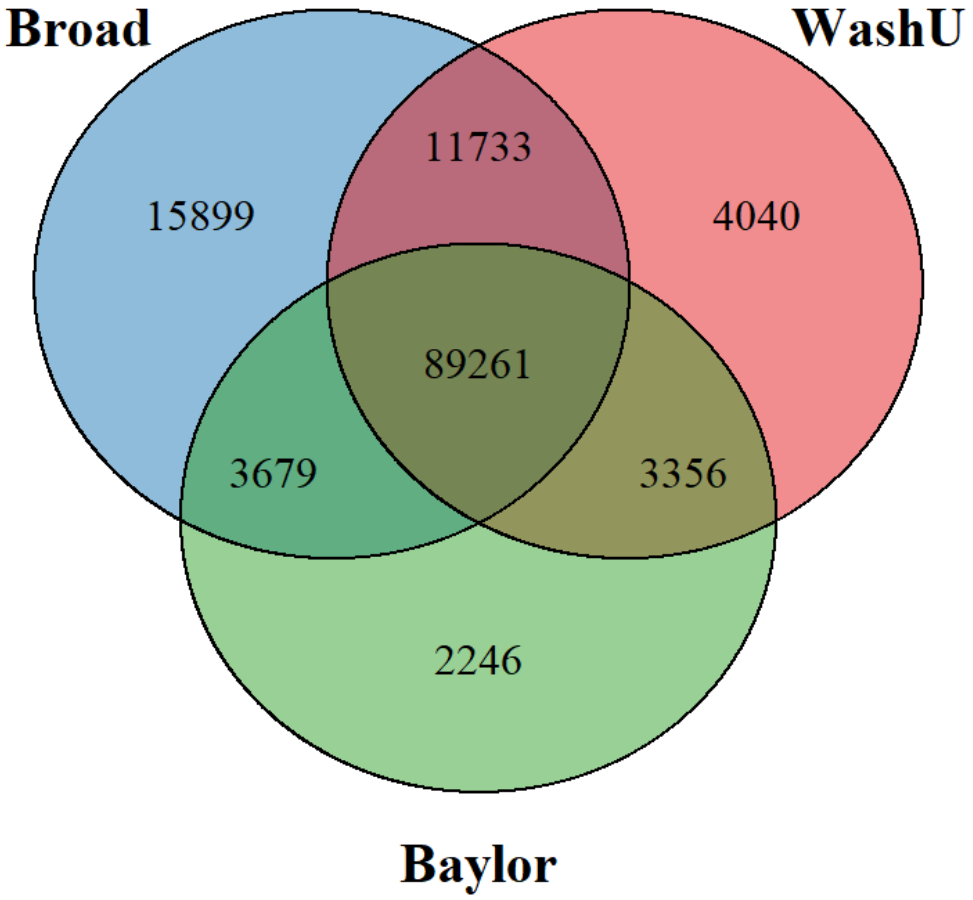



**Fig S2. Number of QCed variants used for association analysis, and their overlap among samples from three sequencing centers.** These variants totaled 120,572 from the Broad samples; 108,390 from the WashU samples; and 98,542 from the Baylor samples. Approximately 70% of variants were shared among samples from all three sequencing centers. The larger number of variants detected in Broad samples was likely due to the larger number of individuals sequenced by Broad compared to the other two centers.

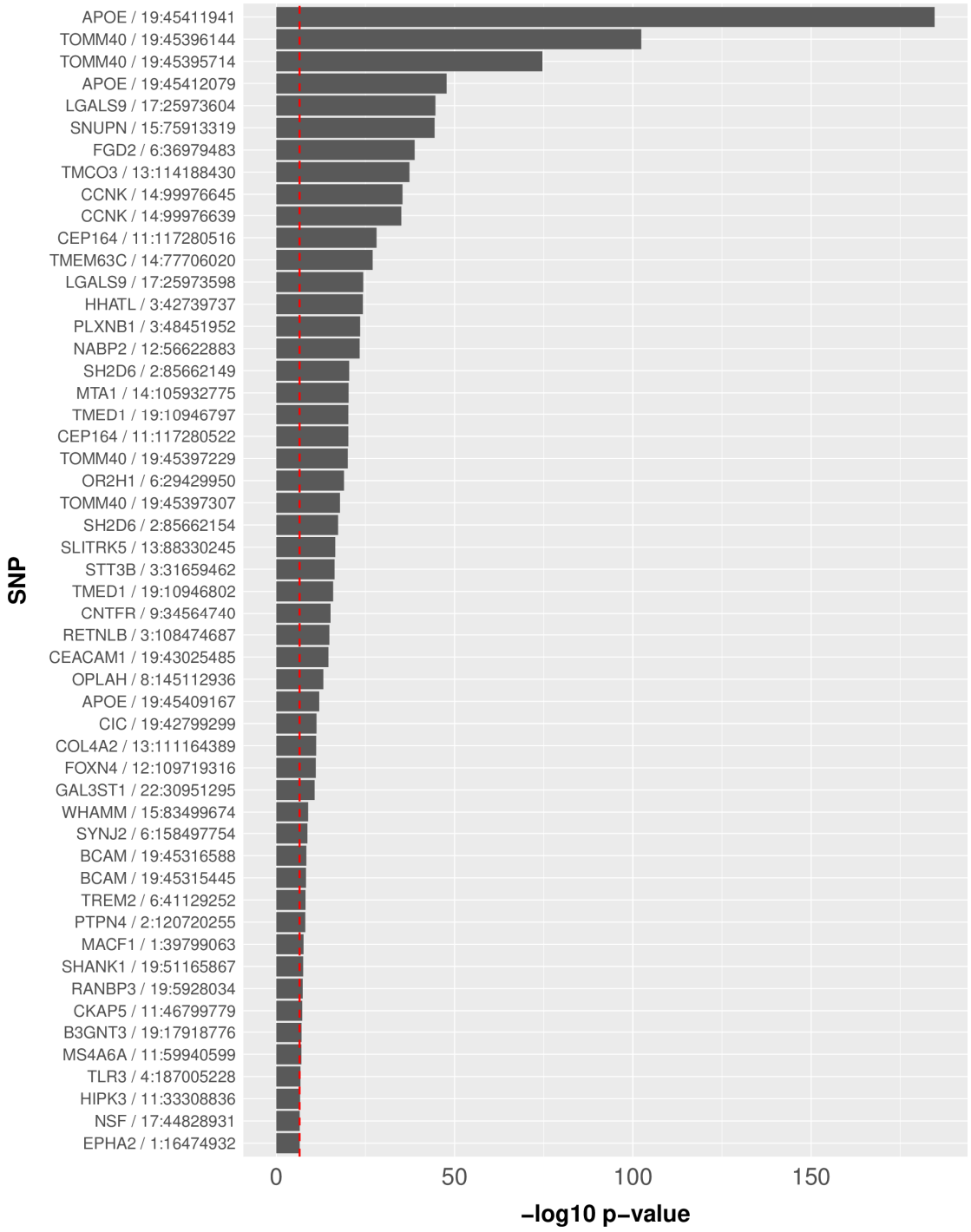



**Fig S3. Log-transformed p-values of all 52 SNPs reaching exome-wide significance under any association model (largest p-value shown for each SNP).** The red vertical line denotes exome-wide statistical significance ($p < 3.0 \times 10^{-7}$).

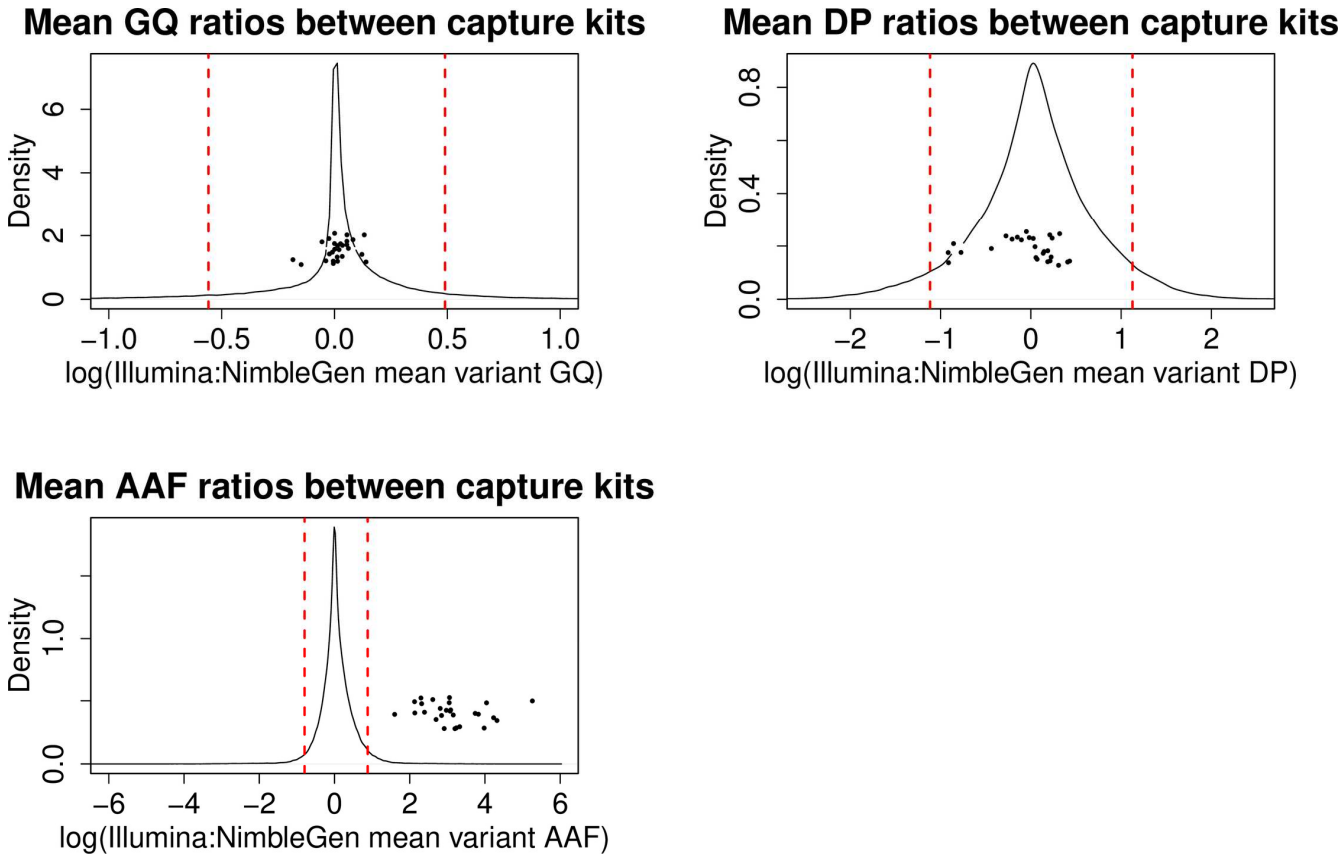



**Fig S4. Distributions of log-adjusted ratios of mean quality metrics for QCed variants shared between capture kits.** Mean values were computed across all samples for each variant. Solid vertical lines demarcate the boundaries separating the 5% tails from the middle 90% of the distribution. The scattered dots represent the positions of the top 29 SNPs within the distributions. Almost all top SNPs lie far to the right of the mean AAF ratio distribution, indicating that these variants are highly discrepant between capture kits.

# Figures

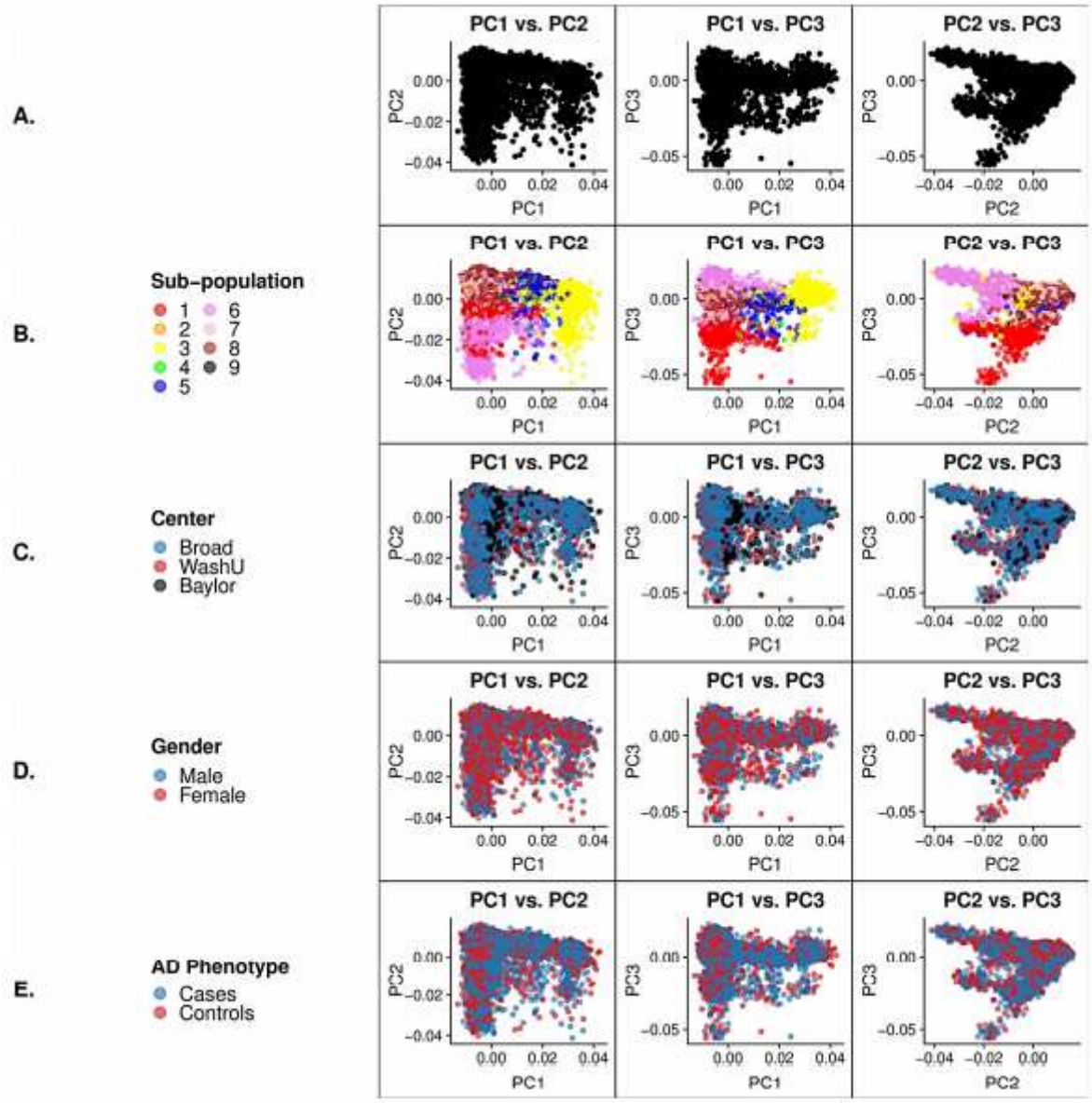



**Figure 1**

Principal Component (PC) eigenvector plots using genotypes of a pruned set of 16,187 high quality common variants for 9,904 ADSP individuals. Each data point represents a single individual. Clustering of samples for a particular variable signifies genotypic similarity between individuals for the trait represented by that color. (A) PCs of the genotypes. (B) PCs color coded based on sub-population. (C) PCs color coded based on center. (D) PCs color coded based on Gender. (E) PCs color coded based on AD phenotype. As expected, clustering is apparent only by sub-population.

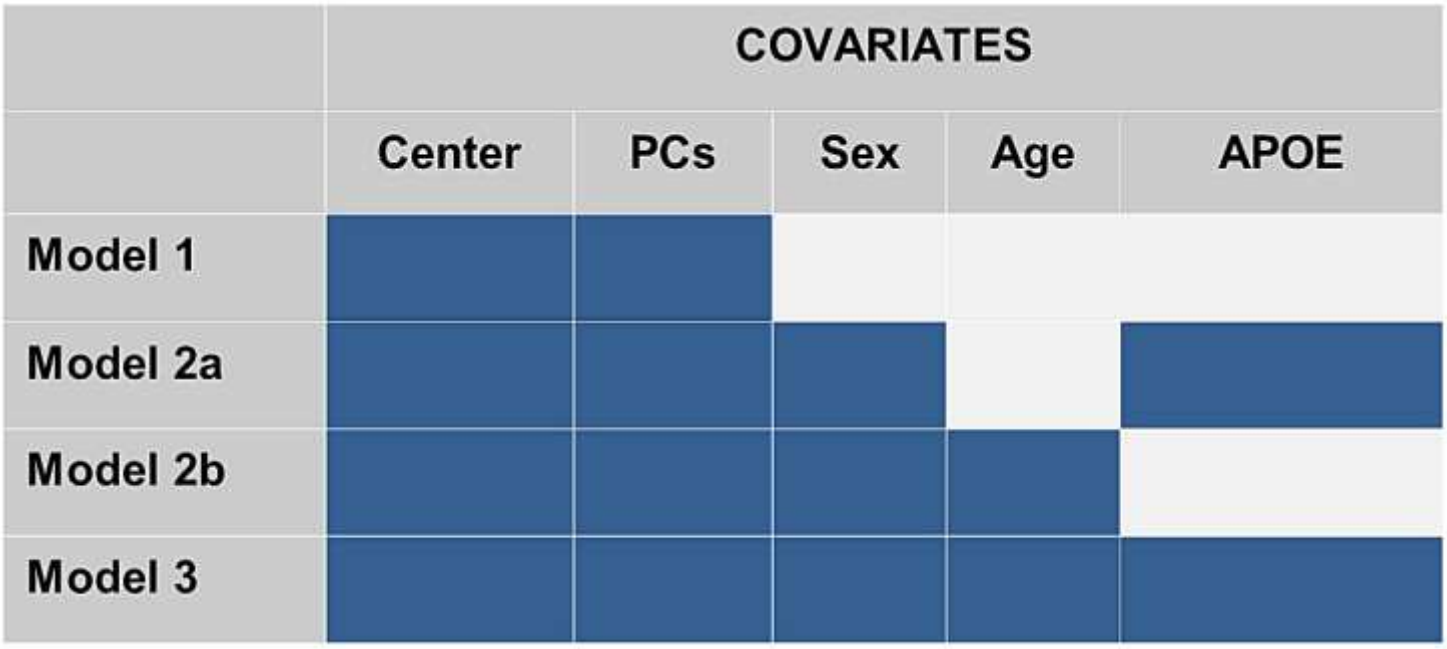
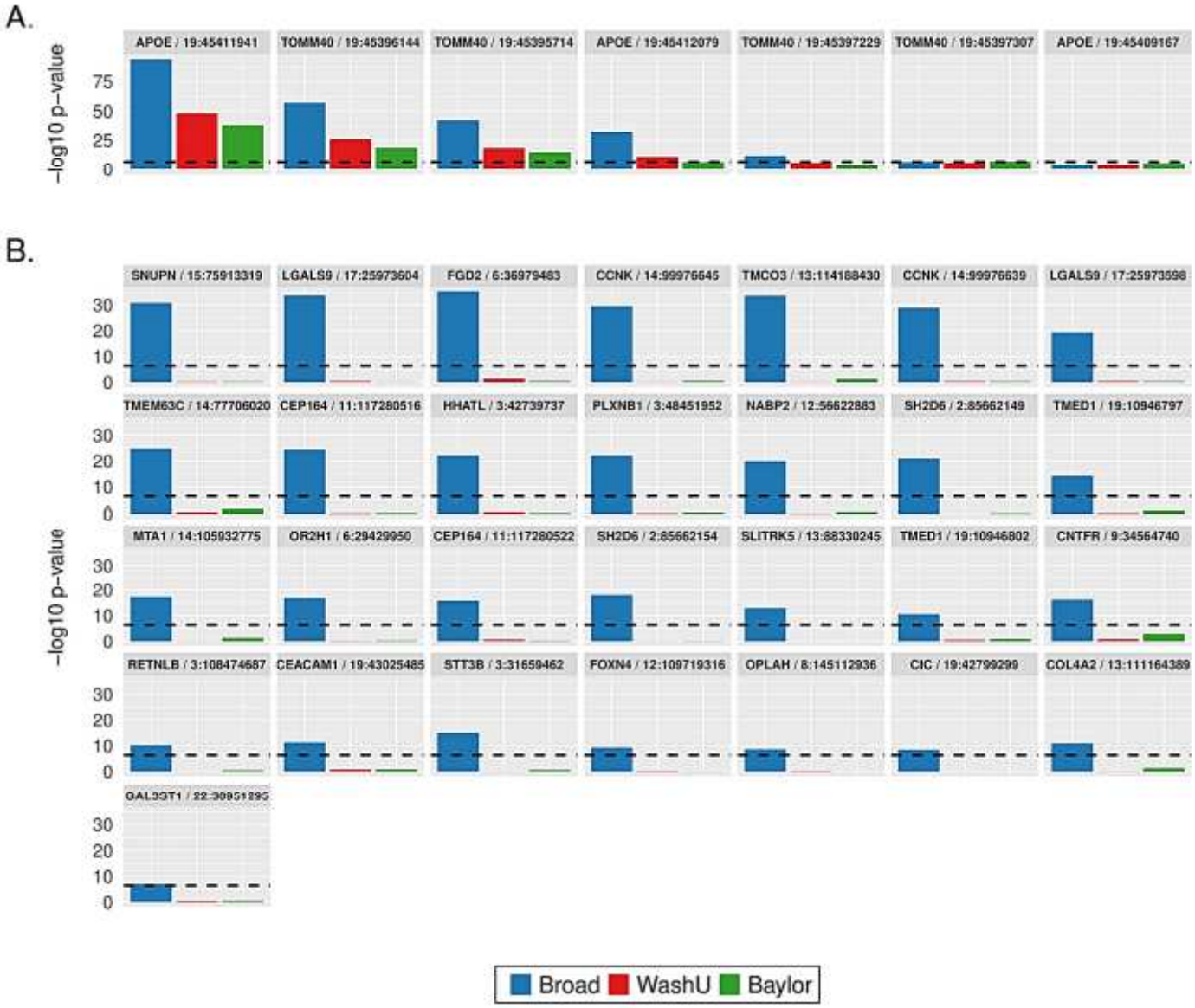


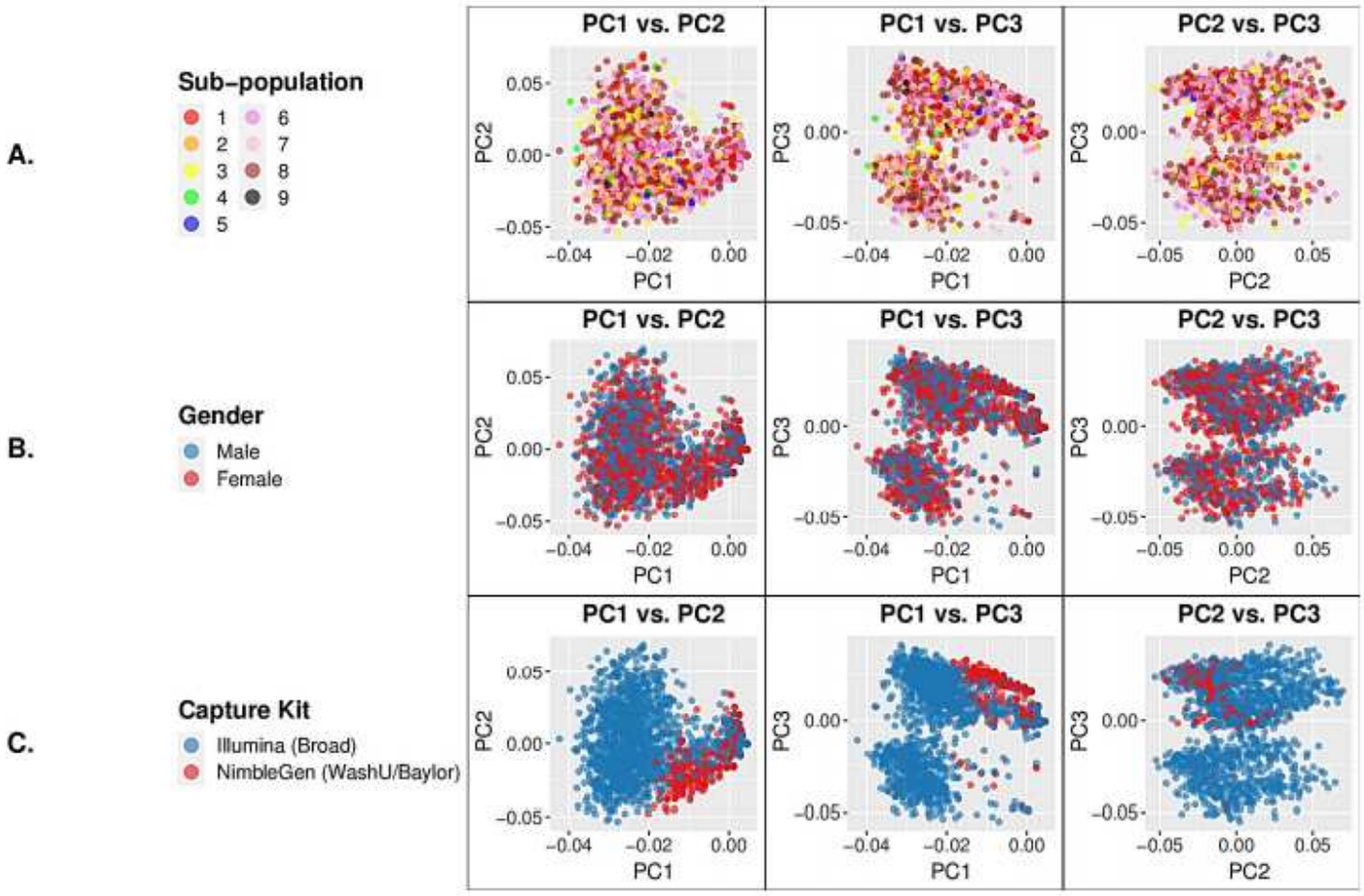

**Figure 2**

Covariates included in each model for association tests.




## Figure 3

Sequencing center specific association p-values of SNPs that reached exome-wide significance (denoted by the dashed horizontal lines) in the full-dataset analysis. (A) Seven SNPs in TOMM40 and APOE. (B) Twenty-nine novel SNPs




## Figure 4

PC eigenvector plots of genotypes at 29 exome-wide significant SNPs. Each data point represents a single individual. Clustering of samples for a particular variable indicates genotypic similarity between individuals for the trait represented by that color. (A) PCs color coded based on sub-population. (B) PCs color coded based on gender. (C) PCs color coded based on capture kit. The NimbleGen-captured samples cluster tightly together, indicating their genotypic similarity that is distinct from the Illumina-captured samples.

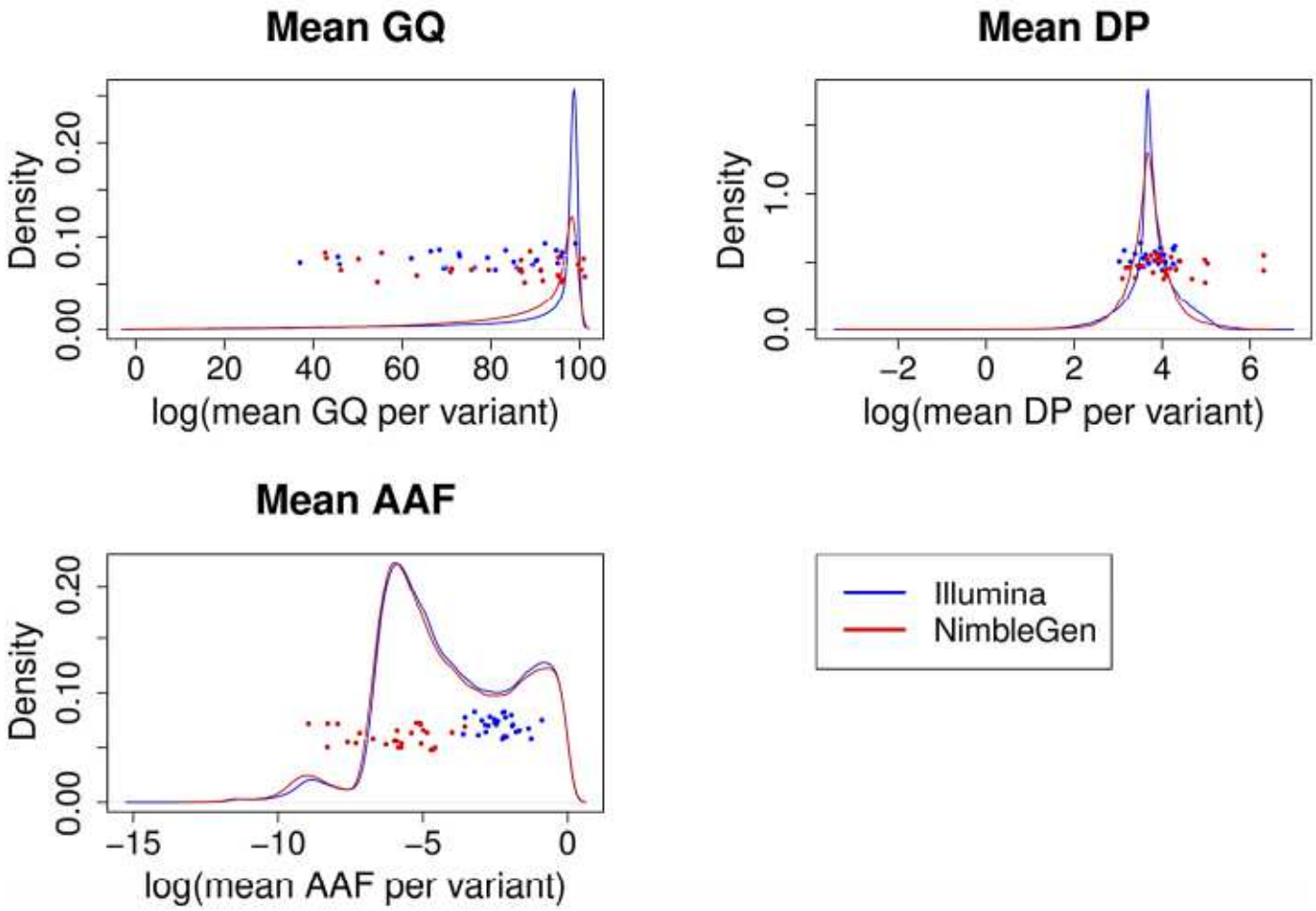



**Figure 5**

Density plots of variant quality parameters between two exome capture kits. Mean values were computed across all samples for each variant. The solid lines show the distributions of all 166,947 variants used in the association analyses, and the scattered dots represent the 29 novel SNPs.


## Figure 6

Minor Allele Frequency (MAF) of 29 exome-wide significant SNPs in AD control exomes processed by two capture kits and in the ExAC Non-Finnish European (NFE) population.

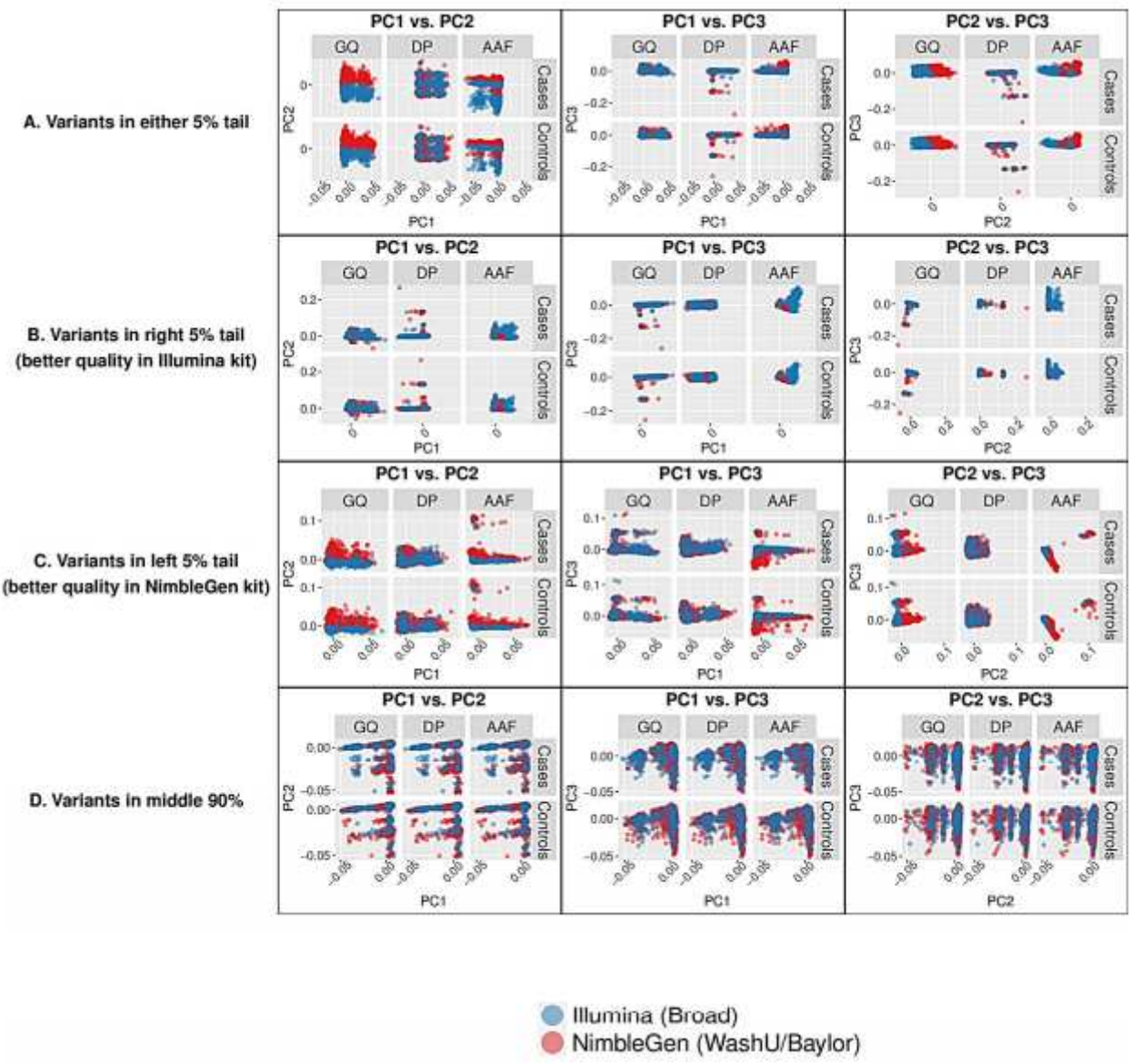



**Figure 7**

PC eigenvector plots of genotypes at variants lying in different sections of quality-metric ratio distributions. Each data point represents a single individual, color coded according to capture kit. (A) PCs of variants in either 5% tail. (B) PCs of variants in right 5% tail. (C) PCs of variants in left 5% tail. (D) Variants in middle 90% of distributions. Variants in the tails, in particular the left 5% tail (better quality in NimbleGen kit), show clear separation by capture kit in both cases and controls.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- figS4.png
- figS1.png

- figS2.png
- figS3.png