Supplementary Note

Abbreviations

1KGP: 1000 Genomes Project
eQTL: Expression quantitative trait locus. Defined by a (eGene, eVariant, tissue) triplet.
eGene: Gene implicated in an eQTL
eVariant: Variant implicated in an eQTL

- eSNV: Single nucleotide eVariant
- eIndel: Small insertion/deletion eVariant
- eSV: Structural eVariant

FDR: False discovery rate
GWAS: Genome-wide association studies
Indel: Small insertion/deletion variant
IRS: Intensity rank sum
LD: Linkage disequilibrium
MAF: minor allele frequency
PCA: Principal component analysis
SNP: Single nucleotide polymorphism
SNV: Single nucleotide variant
SV: Structural variant
WGS: Whole-genome sequencing

Types of structural variation

- DEL (BP, RD): deletion ascertained by LUMPY with breakpoint evidence, with supporting read-depth evidence from Genome STRiP or CNVnator
- DUP (BP, RD): duplication ascertained by LUMPY with breakpoint evidence, with supporting read-depth evidence from Genome STRiP or CNVnator
- DEL (RD): deletion ascertained by Genome STRiP, without breakpoint evidence from LUMPY
- DUP (RD): deletion ascertained by Genome STRiP, without breakpoint evidence from LUMPY
- mCNV: multi-allelic copy number variant ascertained by Genome STRiP, without breakpoint evidence from LUMPY
- rMEI: reference mobile element insertion
- INV: inversion ascertained by LUMPY
- BND: generic breakend ascertained by LUMPY. These included small deletions/duplications lacking read-depth evidence, balance rearrangements, mobile element insertions, and other uncategorized structural variation.
- CNV: copy number variant (deletion or duplication structural variant). Encompasses DEL, DUP, and mCNV.

Table of Contents

4. Author list – The GTEx Consortium
5. Supplementary Tables
6. Supplementary Figures

1. Comparison to 1000 Genomes Project

1.1 Comparison to 1000 Genomes Project variant call set

SNVs and indels

To estimate the accuracy of our variant call set, we compared it to the well-characterized 1000 Genomes Project (1KGP) Phase 3 call set derived from low-coverage (median 7.4X) WGS of 2,504 individuals from diverse ancestries[1,2]. Despite considerable differences in experimental design, such as cohort size, sequencing coverage depth, and population ancestry, we detected roughly similar numbers of SNVs and indels per person to 1KGP and similarly elevated variant counts in individuals of African ancestry compared to Europeans (**Supplementary Table 1**). Our call set is also comparable by typical quality control metrics such as transition/transversion ratio (GTEx: 2.13; 1KGP: 2.08) and the fraction of exonic indels that are out-of-frame (GTEx: 0.80; 1KGP: 0.82) (**Supplementary Table 3**). Furthermore, we recapitulate 94.8% (7,743,012/8,167,029) of 1KGP biallelic SNVs and 61.8% (648,474/1,049,038) of 1KGP biallelic indels with European variant allele frequency ≥ 0.01 (the predominant ancestry in the GTEx cohort).

SVs

SVs detected in this study demonstrated similar call set summary characteristics to 1KGP. Both studies showed consistent trends in the relationship between SV length and minor allele frequency, with larger variants tending to be rarer, as well as a dense band of Alu SINE insertions at approximately 300 bp (**Supplementary Fig. 1a,b**). The two studies also showed similar distributions in the number of SVs ascertained of a given size (**Supplementary Fig. 1c**). Tandem duplications detected by LUMPY using read-pair and/or split-read evidence in this study (DUP) were considerably smaller than those in 1KGP, reflecting a difference in detection algorithms and the difficulty in identifying small CNVs with read-depth evidence.

We observed a similar number of SVs per person to the 1000 Genomes Project, with each exception of tandem duplications (DUPs), for which we find significantly more variants (**Supplementary Table 1**). This is due to the fact that the vast majority (89%) of DUPs reported by 1KGP were larger than 10 kb, whereas we report many smaller DUPs as well (83% less than 10 kb). We also find a somewhat larger number of CNVs by read-depth analysis, presumably due to the greater resolution afforded by deep coverage data (median 49.9X (GTEx) vs. 7.4X (1KGP)).

We compared the overlap between our SV calls and those reported by 1KGP and found that 38.7% of our high confidence calls were previously reported, including 37.2% of the SVs used for eQTL mapping, and thus are presumably not false positives. Importantly, there are not any obvious differences in quality between the "known SVs" that were previously reported by 1KGP, or the "novel SVs" that are unique to our study. When we map eQTLs using solely SVs (in the absence of competing SNVs and indels), we find that known SVs and novel SVs map eQTLs at the same rate, showing that they are equally effective at tagging haplotypes with the exception of tandem duplications which are known to have very different size profiles (**Supplementary Fig. 2a**). Consistent with this, known SVs and novel SVs have similar patterns of linkage disequilibrium (LD) as judged by their maximal r^2 value to flanking SNVs (**Supplementary Fig. 3**). Known and novel SVs also comprise a similar fraction of putatively "causal" SVs predicted to underlie eQTLs (**Supplementary Fig. 2b**), have similar validation rates by IRS statistics (**Supplementary Fig. 2c**), and show a similar pattern of effect size direction when gene coding regions are duplicated or deleted (**Supplementary Fig. 4**).

Moreover, we estimated the false discovery rate (FDR) of high confidence GTEx CNVs to be 2.9%, using the Genome STRiP Intensity Rank Sum annotator and the log R ratio ($\log_2(R_{observed}/R_{expected})$) of intensity values from Illumina Omni 5M genotyping arrays. This FDR is similar to the 2-4% FDR estimated by the 1000 Genomes Project using the same algorithm and Affymetrix SNP6 or Illumina Omni 2.5M arrays.

1.2 Comparison to 1000 Genomes Project SV-eQTL mapping

Our analysis attributes a substantially higher portion of eQTLs to SVs than the 1000 Genomes Project[1]. In whole blood, joint eQTL mapping of SVs, SNVs, and indels revealed an SV to be the lead marker at 2.2% (41/1,899) of protein-coding eQTLs, compared to the 0.56% (54/9,591) of SV-eQTLs mapped by 1KGP in lymphoblastoid cell lines (LCLs). Here we investigate potential biological and technical sources for the discrepancy between these two findings.

First, we note that while whole blood and LCLs represent similar underlying cell types, they have biologically distinct expression profiles that are further subject to procedural differences in RNA isolation and bioinformatics algorithms[3]. Indeed, only 14,750 of the 18,969 transcripts from 1KGP (and 8,593 of the 9,591 eQTLs from 1KGP) were expressed at sufficient levels to be tested in our study. Nonetheless, for the purpose of this comparison we evaluate whole blood because (1) it is the most similar tissue comparator in our data set to LCLs and (2) it is the tissue for which we have the greatest number of

available samples (133 individuals) to compare with the 1KGP cohort of 446 individuals. We have also restricted analyses in this section to protein-coding genes from GENCODE v19 used in 1KGP, even though our broader analysis included non-coding RNA and pseudogenes that are enriched for SV-eQTLs.

Sample size can greatly affect the sensitivity of eQTL mapping studies. Previous work has demonstrated that eQTL discovery increases approximately linearly with sample size[4]. Indeed, serial downsampling of the number of individuals used in eQTL mapping for each tissue recapitulates this linear trend, as well as the tissue-specific differences in eQTL discovery rates when controlling for sample size (**Supplementary Fig. 7a,b**). By linear extrapolation, we estimate that with an equal number of samples to 1KGP, our methods would identify 10,148 protein-coding eQTLs in whole blood. This number closely approximates that 9,591 eQTLs actually discovered by 1KGP, and suggests a similar eQTL mapping efficiency between the two studies. A caveat is that eQTL mapping experiments with fewer samples are biased toward identifying loci with larger effect sizes, which may be a characteristic of SV-eQTLs (**Supplementary Fig. 7c**). Indeed the fraction of SV-eQTLs is slightly elevated in tissues with fewer available samples (**Supplementary Fig. 7d**). However, the relationship between sample number and effect size appears to plateau in the tissues with larger sample sizes (including whole blood), and is therefore unlikely to fully explain the SV-eQTL mapping difference between our study and 1KGP.

Variant detection sensitivity and genotyping accuracy can also impact eQTL mapping efficiency. As described above, the GTEx call set is ostensibly similar to that of 1KGP (**Supplementary Note 1.1**). However, due to limitations of variant detection using low coverage sequencing, the 1000 Genomes Project performed a series of genotype refinement procedures to infer genotypes, in part, from predicted population haplotypes. While the resulting 1KGP call set is extremely high quality by most standards, the refinement procedure introduces haplotype dependence to the genotypes of distinct variants. Since assigning a causal marker to an eQTL is, at its core, a fine-mapping problem, the haplotype-based genotype refinement may confound the results in two ways: (1) an artificially strong interdependence between marker genotypes reduces the power to distinguish causal variants from the overall haplotype; and (2) variants with non-discriminating *a priori* genotype likelihoods (as is often the case for SVs) or those that are poorly tagged by a haplotype are likely to be systematically penalized. Indeed, among a set of 3,063 SVs that were detected by both our study and 1KGP (50% reciprocal overlap, matching variant type, and MAF $\geq$ 0.05), the genotypes from the 1KGP cohort exhibit markedly higher LD to the best linked SNV within 100 kb (**Supplementary Fig. 8**). The effect was similar whether

the “best linked SNV” was defined by the GTEx call set (**Supplementary Fig. 8a**) or by the 1KGP call set (**Supplementary Fig. 8b**). SNVs detected by both studies show a similar trend that is unlikely caused by differences in genotyping quality due to the ease in genotyping these variants from deep WGS data (**Supplementary Fig. 8c,d**). In contrast, all genotype information in our study is derived solely from the primary read alignments, such that each individual SV, SNV or indel genotype was calculated independently from any other variant’s genotype using the raw sequencing data, with extremely deep coverage (median 49.9x), affording greater power to disentangle causal variants at eQTLs.

Next, we evaluated eQTL mapping of the 2,577 SVs that were detected in both GTEx and 1KGP, and that had similar MAFs in the two studies (within 10%). This subset included 28.7% of the 8,980 eQTL-eligible GTEx SVs and 19.6% of the 14,531 eQTL-eligible 1KGP SVs. In our study, these 2,577 SVs are the lead marker at 6 whole blood protein-coding eQTLs (compared with 10 eQTLs from 1KGP). Thus, our study maps 60% as many eQTLs for the same set of input SVs despite detecting 19.8% (1,899/9,591) as many eQTLs overall as a result of differences in sample size (**Supplementary Fig. 7**). This comparison indicates that on a per variant basis, we detect an approximately 3-fold as many SV-eQTLs as 1KGP. The similarity of eQTL mapping sensitivity overall (see above) suggests that this difference is specifically due to SV genotype information, not other factors such eQTL mapping methodology, RNA expression data quality, etc.

Most importantly, the ultimate effect of variant genotyping error is reduced power to map eQTLs, therefore any issues related to SV genotyping accuracy will result in false negative eQTLs, not false positive eQTLs. Thus, in the context of our WGS-based study, where most eQTLs can be detected by multiple linked variants, an increased SV genotyping error rate would decrease the number and fraction of SV-eQTLs relative to SNV-eQTLs or indel-eQTLs, and cause an underestimate of the impact of SV.

We performed a simulation experiment to investigate the effect of genotyping error on the ability to map SV-eQTLs. A mere 5% increase in the genotyping error rate in SVs is sufficient to reduce SV-eQTL mapping rate by 19.6% (**Supplementary Fig. 9**).

Finally, we compared the properties of SV-eQTLs discovered in this study to those in the 1000 Genomes Project. The small number of SV-eQTLs (54) identified by 1KGP limited the interpretation of these data, but the two studies showed similar trends in the size distribution and number of each SV class (**Supplementary Figs. 10-12**). The exception to this similarity lies in tandem duplication SVs, for

which methodologies in our study allowed detection of far more smaller events (**Supplementary Fig. 1**). Overall, we recapitulated 32 of 47 (68.1%) previously identified LCL SV-eQTLs at eGenes also expressed in available tissues from our study.

2. SV-eQTL detectability by alternative means

While deep WGS provides greater sensitivity and genotyping accuracy for SV detection, its utility must be balanced against its relative cost compared to other technologies. To aid in the experimental design of future studies, we have conducted a series of experiments to estimate the number of SV-eQTLs that could have been detected with high throughput genotyping arrays.

2.1 Tagging of structural variants by linked markers

The extent to which structural variants are tagged by other genetic markers via linkage disequilibrium (LD) is an important consideration in the design of trait-mapping studies. Our analyses indicate that SVs exhibit weaker linkage to the surrounding haplotype than other variant types. Of 8,577 autosomal SVs in our study with minor allele frequency (MAF) ≥ 0.05, only 58.2% had a well-tagged (r^2 ≥ 0.8) SNV or indel detected by WGS within 50 kb. This fraction is markedly lower than that of SNVs (79.4%) and slightly lower than indels (64.5%) (randomly downsampled to 10,000 variants of each type) (**Supplementary Fig. 19a**). Moreover, the weaker linkage of SVs is not likely to result solely from differences in genotyping quality because it is apparent in various subsets of SVs that had sufficiently high quality genotype information to map eQTLs and/or be judged as causal through various measures. For example, only 56.7% of the SV-eQTLs identified from the "SV-only" mapping exercise were well-tagged based on r^2 ≥ 0.8, as were 19.3% of the 243 autosomal eSVs identified by joint eQTL mapping, and 51.4% of the 766 autosomal SVs in the top 10% of composite causality scores (the set used for all functional analyses). In contrast, 77.6% of the SNVs judged as causal by joint eQTL analysis were well-tagged.

When tagging markers were limited to the 1,980,784 SNVs present on the widely used Illumina Omni 2.5M genotyping array and detected by WGS in our GTEx cohort, only 46.7% of SVs (including 41.3% of predicted causal eSVs) had a probe with r^2 ≥ 0.8 and only 69.6% (66.8% of predicted causal eSVs) had a probe with r^2 ≥ 0.5 (**Supplementary Fig. 19b**).

To investigate the consequences of omitting SVs from trait mapping studies, we assessed the fraction of SV-eQTLs that would have been discovered by linked SNVs or indels when SVs were excluded from the analysis. We ran FastQTL on the SNVs and indels alone and tracked the fate of the 828 SV-eQTLs

(across 13 tissues) originally discovered by joint eQTL mapping. Overall 41.2% (341/828) of eQTLs did not meet genome-wide significance through SNV and indel eQTL mapping, demonstrating that a substantial portion of eQTL effects caused by SVs are invisible through linkage disequilibrium with nearby SNVs and indels detected by WGS (**Supplementary Table 6**). We note the important caveat that the power to detect eQTLs through non-causal markers is heavily influenced by sample size in addition to LD, a trend which is apparent in our data. However, even in whole blood, the tissue with the greatest number of available samples, 20.8% of eGenes originally mapped to SV-eQTLs did not meet genome-wide significance through other markers.

2.2 Detection of SVs by genotyping array probe intensities

Genotyping microarrays are a high-throughput and cost-effective technology that can detect CNVs through the signal intensities of genotyping probes. However, due to their low-resolution (commonly 2-5 million probes per array), they are only sensitive to large CNVs that comprise the minority of genomic structural variation. In a typical array-based CNV detection workflow, aberrant signal intensity must be observed for at least 5 consecutive probes, and of the 17,040 CNVs identified in this study, only 12.9% and 24.2% spanned 5 probes for the Illumina Omni 2.5M and Omni 5M genotyping arrays respectively (**Supplementary Fig. 20**). Moreover, since common SVs are generally smaller than rare events, only 3.8% (Omni 2.5M) and 13.9% (Omni 5M) of the CNVs with MAF ≥ 0.05 spanned 5 probes. CNVs that were in the 90th percentile of causality scores were spanned by 5 probes at similar frequencies, (Omni 2.5M: 4.9%; Omni 5M: 16.0%).

Finally, we compared CNV calls detected by WGS to those identified in any sample by either of these two array platforms in our data set (Omni 2.5M: 270 samples; Omni 5M: 178 samples). This included an additional 301 samples for which microarray data was available. Only 11.0% (1,873/17,040) CNVs (3.7% (208/5,643) of CNVs with MAF ≥ 0.05) were detected in any sample with either array platform, when requiring 50% reciprocal overlap. CNVs with a causality score in the 90th percentile were only detected on arrays at a rate of 6.2% (33/536).

3. Examination of population substructure in rare variant analysis

We examined the subpopulation structure within the 117 Caucasians used for our rare variant analysis to exclude the possibility that it may lead to non-causal co-occurrence of rare variants and expression outliers. Principal components analysis of the 117 Caucasian individuals using SNVs did not reveal clear population clusters (**Supplementary Fig. 22**), which suggests that subpopulation architecture is not a major confounding factor. A single outlier individual (GTEX-WHPG) who clustered with admixed

Hispanic ethnicity did not account for an excess number of RNA expression outliers (30) or an excess number of "genetically explained" expression outliers that have a rare variant within 5 kb (11) (**Supplementary Fig. 23**), and exclusion of this individual did not significantly change our results or conclusions (**Supplementary Fig. 24**).

None of the principal components calculated above correlate with the number of RNA expression outliers identified per individual (**Supplementary Fig. 25a**), or with the number of genetically explained expression outliers (**Supplementary Fig. 25b**). Thus, whatever population structure may exist in the data set, there is no evidence that it affects the comparison of rare variants and gene expression outliers.

4. Author list – The GTEx Consortium

Laboratory, Data Analysis and Coordinating Center (LDACC) - Analysis Working Group (AWG)
Kristin G. Ardlie[1], Gad Getz[1,2], Ellen T. Gelfand[1], Ayellet V. Segrè[1], François Aguet[1], Timothy J. Sullivan[1], Xiao Li[1], Jared L. Nedzel[1], Casandra A. Trowbridge[1], Daniel G. MacArthur[1,3], Monkol Lek[1,3], Taru Tukiainen[3,4], Kane Hadley[4], Katherine H. Huang[4], Michael S. Noble[4], Duyen T. Nguyen[4], Beryl B. Cummings[3,4]

Funded Statistical Methods groups - Analysis Working Group (AWG)
Andrew B. Nobel[5], Fred A. Wright[6], Andrey A. Shabalin[7], John J. Palowitch[8], Yi-Hui Zhou[9], Emmanouil T. Dermitzakis[10,11,12], Mark I. McCarthy[13,14,15], Anthony J. Payne[13], Tuuli Lappalainen[16,17], Stephane Castel[16,17], Sarah Kim-Hellmuth[16,17], Pejman Mohammadi[16,17], Alexis Battle[18], Princy Parsana[18], Sara Mostafavi[19], Andrew Brown[10,11,12], Halit Ongen[10,11,12], Olivier Delaneau[10,11,12], Nikolaos Panousis[10,11,12], Cedric Howald[10,11,12], Martijn van de Bunt[13,14], Roderic Guigo[20,21,22], Jean Monlong[20,21,23], Ferran Reverter[20,24], Diego Garrido[20,21], Manuel Munoz[20,21], Gireesh Bogu[20,21], Reza Sodaei[20,21], Panagiotis Papasaikas[20,21], Anne W. Ndungu[13], Stephen B. Montgomery[25], Xin Li[25], Laure Fresard[25], Joe R. Davis[25], Emily K. Tsang[25,26], Zachary Zappala[25], Nathan S. Abell[25], Michael J. Gloudemans[25,26], Boxiang Liu[25,27], Farhan N. Damani[28], Ashis Saha[28], Yungil Kim[18], Benjamin J. Strober[29], Yuan He[29], Matthew Stephens[30,31], Jonathan K. Pritchard[30,32,33], Xiaoquan Wen[34], Sarah Urbut[30], Nancy J. Cox[35,36], Dan L. Nicolae[37], Eric R. Gamazon[35,36], Hae Kyung Im[38], Christopher D. Brown[39], Barbara E. Engelhardt[40], YoSon Park[39], Brian Jo[41], Ian C. McDowell[42], Ariel Gewirtz[41], Genna Gliner[43], Don Conrad[44,45], Ira Hall[46,47,48], Colby Chiang[46], Alexandra Scott[46], Chiara Sabatti[49], Eleazar Eskin[50], Christine Peterson[51], Farhad Hormozdiari[52], Eun Yong Kang[52], Serghei Mangul[52], Buhm Han[53], Jae Hoon Sul[54]

Enhancing GTEx (eGTEx) funded groups
Andrew P. Feinberg[55], Lindsay F. Rizzardi[56], Kasper D. Hansen[57], Peter Hickey[58], Joshua Akey[59], Manolis Kellis[4,60], Jin Billy Li[61], Michael Snyder[61], Hua Tang[61], Lihua Jiang[61], Shin Lin[61,62], Barbara E. Stranger[63], Marian Fernando[64], Meritxell Oliva[64], John Stamatoyannopoulos[65], Rajinder Kaul[65], Jessica Halow[65], Richard Sandstrom[65], Eric Haugen[65], Audra Johnson[65], Kristen Lee[65], Daniel Bates[65], Morgan Diegel[65], Brandon L. Pierce[66], Lin Chen[66], Muhammad G. Kibriya[66], Farzana Jasmine[66], Jennifer Doherty[67], Kathryn Demanelis[66], Stephen B. Montgomery[25], Emily K. Tsang[25], Kevin S. Smith[25], Qin Li[61], Rui Zhang[61]

NIH Common Fund
Concepcion R. Nierras[68]

NIH/NCI
Helen M. Moore[69], Abhi Rao[69], Ping Guan[69], Jimmie B. Vaught[69], Philip A. Branton[69], Latarsha J. Carithers[70]

NIH/NHGRI
Simona Volpi[71], Jeffery P. Struewing[71], Casey G. Martin[71], Lockhart C. Nicole[71]

NIH/NIMH
Susan E. Koester[72], Anjene M. Addington[72]

NIH/NIDA
A. Roger. Little[73]

Biospecimen Collection Source Site - NDRI
William F. Leinweber[74], Jeffrey A. Thomas[74], Gene Kopen[74], Alisa McDonald[74], Bernadette Mestichelli[74], Saboor Shad[74], John T. Lonsdale[74], Michael Salvatore[74], Richard Hasz[75], Gary Walters[76], Mark Johnson[76], Michael Washington[76], Lori E. Brigham[77], Christopher Johns[78], Joseph Wheeler[78], Brian Roe[79], Marcus Hunter[79], Kevin Myer[79]

Biospecimen Collection Source Site - RPCI
Barbara A. Foster[80], Michael T. Moser[80], Ellen Karasik[80], Bryan M. Gillard[80], Rachna Kumar[80], Jason Bridge[81], Mark Miklos[81]

Biospecimen Core Resource - VARI
Scott D. Jewell[82], Daniel C. Rohrer[82], Dana Valley[82], Robert G. Montroy[82]

Brain Bank Repository - U Miami
Deborah C. Mash[83], David A. Davis[84]

Leidos Biomedical - Project Management
Anita H. Undale[85], Anna M. Smith[86], David E. Tabor[86], Nancy V. Roche[86], Jeffrey A. McLean[86], Negin Vatanian[86], Karna L. Robinson[86], Leslie Sobin[86], Mary E. Barcus[87], Kimberly M. Valentino[86], Liqun Qi[86], Stephen Hunter[86], Pushpa Hariharan[86], Shilpi Singh[86], Ki Sung Um[86], Takunda Matose[86], Maria M. Tomadzewski[86]

ELSI Study
Laura A. Siminoff[88], Heather M. Traino[89], Maghboeba Mosavel[90], Laura K. Barker[91]

Genome Browser Data Integration, and Visualization - EBI
Daniel R. Zerbino[92], Thomas Juettmann[92], Kieron Taylor[92], Magali Ruffier[92], Dan Sheppard[92], Steven Trevanion[92], Paul Flicek[92]

Genome Browser Data Integration, and visualization - UCSC Genomics Institute, University of California Santa Cruz
W. James Kent[93], Kate R. Rosenbloom[93], Maximilian Haeussler[93], Christopher M. Lee[93], Benedict Paten[93], John Vivan[93], Jingchun Zhu[93], Mary Goldman[93], Brian Craft[93]

Other members of the Analysis Working Group (AWG)
Gen Li[94], Pedro G. Ferreira[95,96], Esti Yeger-Lotem[97,98], Matthew T. Maurano[99], Ruth Barshir[97], Omer Basha[97], Hualin S. Xi[100], Jie Quan[100], Michael Sammeth[101], Judith B. Zaugg[102]

1. The Broad Institute of Massachusetts Institute of Technology and Harvard University. Cambridge, Massachusetts 02142, USA.
2. Massachusetts General Hospital Cancer Center and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA, Boston, MA 02114, USA
3. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston MA 02114, USA.
4. The Broad Institute of Massachusetts Institute of Technology and Harvard University Cambridge, Massachusetts 02142, USA.
5. Department of Statistics and Operations Research and Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-3260
6. Bioinformatics Research Center and Departments of Statistics and Biological Sciences, North Carolina State University, Raleigh NC, 27695
7. Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, Richmond, VA 23298-0581
8. Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260
9. Bioinformatics Research Center and Department of Biological Sciences, North Carolina State University, Raleigh NC, 27695
10. Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland
11. Institute for Genetics and Genomics in Geneva (iG3), University of Geneva, 1211 Geneva, Switzerland
12. Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland
13. Wellcome Trust Centre for Human Genetics Research, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK OX3 7BN
14. Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK, OX3 7LE
15. Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, UK, OX3 7LJ
16. New York Genome Center, 101 Avenue of the Americas, New York, NY, 10013
17. Department of Systems Biology, Columbia University Medical Center, New York, NY 10032
18. Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
19. Department of Computer Science, Stanford University, Stanford, CA 94305, USA
20. Center for Genomic Regulation (CRG), Barcelona, Catalonia, Spain
21. Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain
22. Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), 08003 Barcelona, Spain
23. Human Genetics Dept., McGill University, Montréal Canada
24. Universitat de Barcelona, 08028 Barcelona, Catalonia, Spain
25. Departments of Genetics and Pathology, Stanford University, Stanford, CA, 94305
26. Biomedical Informatics Program, Stanford University, Stanford, CA, 94305
27. Department of Biology, Stanford University, Stanford, CA, 94305
28. Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218
29. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, 21218
30. University of Chicago, Department of Human Genetics, Chicago, IL 60637

31. University of Chicago, Department of Statistics 5734 S. University Avenue, Chicago, IL 60637
32. Dept of Genetics and Biology, Stanford University
33. Howard Hughes Medical Institute
34. Dept of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109
35. Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Canter, Nashville, TN 37232
36. Department of Clinical Epidemiology, Biostatistics and Bioinformatics and Department of Psychiatry, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands
37. University of Chicago, Section of Genetic Medicine, Department of Medicine, Department of Statistics and Department of Human Genetics, 900 East 57th Street KCBD 3220, Chicago, IL 60637
38. Section of Genetic Medicine, Department of Medicine, The University of Chicago, 900 East 57th Street KCBD 3220, Chicago, IL 60637
39. University of Pennsylvania, Perelman School of Medicine, Department of Genetics, Philadelphia, PA, 19104
40. Princeton University, Department of Computer Science, Center for Statistics and Machine Learning, 35 Olden Street, Princeton, NJ 08540
41. Lewis Sigler Institute, Princeton University, Princeton, NJ 08540
42. Computational Biology & Bioinformatics Graduate Program, Duke University, Durham, NC 27708
43. Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540
44. Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, 63108 USA
45. Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, Missouri, 63108, USA
46. McDonnell Genome Institute, Washington University School of Medicine, Saint Louis, MO, 63108
47. Department of Medicine, Washington University School of Medicine, Saint Louis, MO, 63108
48. Department of Genetics, Washington University School of Medicine, Saint Louis, MO, 63108
49. Departments of Biomedical Data Science and Statistics, HRP Redwood building, Stanford, CA 94305-5404
50. Department of Computer Science, Department of Human Genetics, University of California, Los Angeles, CA 90095
51. Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030
52. Department of Computer Science, University of California, Los Angeles, CA 90095
53. Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, Korea
54. Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA 90095, USA
55. Center for Epigenetics, Johns Hopkins University School of Medicine, Departments of Medicine, Biomedical Engineering, and Mental Health, Johns Hopkins University Schools of Medicine, Engineering, and Public Health, Baltimore, MD, 21205
56. Center for Epigenetics, Johns Hopkins University School of Medicine, Baltimore, MD, 21205

57. McKusick-Nathans Institute of Genetic Medicine, Center for Epigenetics, Johns Hopkins School of Medicine, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, 21205
58. Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205
59. Department of Genome Sciences, University of Washington, Seattle, WA 98195
60. CSAIL, MIT, Cambridge MA
61. Department of Genetics, Stanford University, Stanford, CA, 94305
62. Division of Cardiology, University of Washington, Seattle, WA 98195
63. University of Chicago, Section of Genetic Medicine, Department of Medicine, Institute for Genomics and Systems Biology, Center for Data Intensive Science, Chicago, IL 60637
64. University of Chicago, Section of Genetic Medicine, Department of Medicine, Institute for Genomics and Systems Biology, Chicago, IL 60637
65. Altius Institute for Biomedical Sciences, Seattle, WA 98121
66. University of Chicago, Department of Public Health Sciences, Chicago, IL 60637
67. Department of Epidemiology, The Geisel School of Medicine at Dartmouth, Lebanon, NH 03756
68. Office of Strategic Coordination, Division of Program Coordination, Planning and Strategic Initiatives, Rockville, MD 20852-9305
69. Biorepositories and Biospecimen Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD 20892
70. National Institute of Dental and Craniofacial Research, 6701 Democracy Blvd, Bethesda, MD 20892
71. Division of Genomic Medicine, National Human Genome Research Institute, Rockville, MD
72. DNBBS/NIMH/NIH, Bethesda, MD 20892
73. National Institute on Drug Abuse, NIH, HHS. Bethesda, Maryland USA 20892
74. National Disease Research Interchange, Philadelphia, PA 19103
75. Gift of Life Donor Program, Philadelphia, PA 19103
76. LifeNet Health, Virginia Beach, VA 23453
77. Washington Regional Transplant Community, Annandale, VA 22003
78. Center for Organ Recovery and Education, Pittsburgh, PA 15238
79. LifeGift, Houston, TX 77054
80. Roswell Park Cancer Institute Pharmacology & Therapeutics, Buffalo NY 14263
81. 110 Broadway, Buffalo, NY 14203
82. Van Andel Research Institute, Grand Rapids, MI, 49503
83. Univ. Miami Miller School of Medicine, Dept. Neurology, Miami, FL 33136
84. Univ. Miami Brain Endowment Bank, Miller School of Medicine, Miami, FL 33136
85. NIH/NIAID, 5601 Fishers Lane, Rockville, MD 20852
86. 6110 Executive Blvd, Suite 250, Rockville MD 20852
87. 8560 Progress Drive, Room C3021, Frederick MD 21701
88. Temple University, Philadelphia, PA 19122
89. Temple University, 1301 Cecil B. Moore Avenue, Ritter Annex 9th Floor, Philadelphia, PA 19122
90. Virginia Commonwealth University, Richmond, VA 23219
91. Temple University, Philadelphia, PA 19122

92. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB 10 1SD, United Kingdom
93. UCSC Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064
94. Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032
95. i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Rua Alfredo Allen, 208, 4200-135 Porto, Portugal
96. IPATIMUP - Institute of Molecular Pathology and Immunology, University of Porto, Rua Dr. Roberto Frias s/n, 4200-625 Porto, Portugal
97. Ben-Gurion University of the Negev, Beer-Sheva, 84105 Israel
98. National Institute for Biotechnology in the Negev, Beer-Sheva, 84105 Israel
99. Institute for Systems Genetics, New York University Langone Medical Center, New York, New York 10016, USA
100. Computational Sciences, Pfizer Inc, 610 Main st, Cambridge, MA02140
101. Institute of Biophysics Carlos Chagas Filho (IBCCF), Federal University of Rio de Janeiro (UFRJ), 21941902 Rio de Janeiro, Brazil
102. European Molecular Biology Laboratorium, Meyerhofstrasse 1, 69117 Heidelberg, Germany

References

1. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
2. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
3. Min, J. L. *et al.* Variability of gene expression profiles in human blood and lymphoblastoid cell lines. *BMC Genomics* **11,** 96 (2010).
4. The GTEx Consortium *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).

Cohort (ancestry)		1KGP (European)	GTEx (European)	1KGP (African)	GTEx (African)
Number of individuals		503	122	661	23
Median number of variants per individual	SNVs	3.53M	3.39M	4.31M	4.07M
	Indels	546k	368k	625k	441k
	Deletions	1223	1369	1,431	1,620
	Duplications	10	516	14	564
	mCNVs	166	282.5	179	378
	Inversions	29	7	33	9
	Reference MEIs	661	1,095	764	1,264

Supplementary Table 1. Median number of variants of each class per individual for this study (GTEx) and the 1000 Genomes Project (1KGP).

[supp_table.02.xlsx]

Supplementary Table 2. Excel file of all SV-only and joint eQTLs, along with causality scores

Study	Number of samples	SNVs				Indels		
		Number of variants	Ti/Tv	Percent singletons	Singleton Ti/Tv	Number of variants	Percent out of frame exonic	Percent singleton
GTEx	147	21,764,904	2.13	34.4%	2.12	3,030,964	80%	33.6%
1KGP Phase 3	2,504	81,443,083	2.08	43.9%	1.97	3,363,851	82%	2.2%

Supplementary Table 3. Number and characteristics of SNVs and indels discovered in GTEx and 1KGP studies.

Tissue	# samples	# expressed genes	# expressed genes (protein-coding)	# joint eQTLs	# joint SNV-eQTLs	# joint indel-eQTLs	# joint SV-eQTLs
Whole blood	133	23,931	15,335	2,596	2,205	314	77
Cells (transformed fibroblasts)	116	23,745	15,036	3,573	3,083	404	86
Muscle (skeletal)	116	23,906	15,487	1,813	1,550	208	55
Lung	105	28,631	16,940	2,035	1,749	205	81
Artery (tibial)	98	25,262	15,914	1,918	1,623	233	62
Adipose (subcutaneous)	97	27,133	16,539	1,684	1,424	189	71
Thyroid	89	28,472	16,795	2,032	1,746	217	69
Esophagus (mucosa)	88	25,914	16,256	1,782	1,522	185	75
Skin (sun exposed lower leg)	87	27,763	16,852	1,320	1,132	129	59
Nerve (tibial)	82	27,762	16,604	1,520	1,298	162	60
Esophagus (muscularis)	80	25,270	16,129	1,607	1,376	167	64
Artery (aorta)	72	25,253	15,926	1,048	903	101	44
Heart (left ventricle)	70	23,668	15,467	626	537	64	25
Overall	145	34,053	18,126	23,554	20,148	2,578	828
Overall distinct eGenes	-	-	-	9,634	8,825	1,999	224
Overall distinct eVariants	-	-	-	19,342	16,959	2,383	253

Supplementary Table 4. Number of samples, expressed (protein-coding) genes, and joint eQTLs from each tissue type.

[supp_table.05.xlsx]

Supplementary Table 5. Excel file of all SV-eQTL GWAS hits

	# samples	# joint SV-eQTLs	# Attributed to SNP or indel	% Attributed to SNP or indel	# Did not meet genome-wide significance	% Did not meet genome-wide significance
Whole blood	133	77	61	79.2%	16	20.8%
Cells (transformed fibroblasts)	116	86	65	75.6%	21	24.4%
Muscle (skeletal)	116	55	36	65.5%	19	34.5%
Lung	105	81	44	54.3%	37	45.7%
Artery (tibial)	98	62	35	56.5%	27	43.5%
Subcutaneous adipose	97	71	41	57.7%	30	42.3%
Thyroid	89	69	39	56.5%	30	43.5%
Esophagus (mucosa)	88	75	46	61.3%	29	38.7%
Skin (sun exposed lower leg)	87	59	33	55.9%	26	44.1%
Nerve (tibial)	82	60	27	45.0%	33	55.0%
Esophagus (muscularis)	80	64	30	46.9%	34	53.1%
Artery (aorta)	72	44	18	40.9%	26	59.1%
Heart (left ventricle)	70	25	12	48.0%	13	52.0%
Overall	145	828	487	58.8%	341	41.2%

Supplementary Table 6. Fate of SV-eQTLs when performing eQTL mapping in the absence of SVs.

	Rare variant type	Num. outliers with rare variant within 5 kb	Num. outliers	Shuffle median	Shuffle 2.5-%tile	Shuffle 97.5-%tile	Fold enrichment of outliers	Fold enrichment (95% CI)
Per outlier	SV	355	5,047	22	14	31	16.1	(11.5, 25.4)
	SNV	1,965	5,047	1,738	1,679	1,797	1.1	(1.1, 1.2)
	Indel	690	5,047	561	519	600	1.2	(1.2, 1.3)
	Any	2,417	5,047	1,974	1,912	2,035	1.2	(1.2, 1.3)
	Rare variant type	Num. rare variants with outlier within 5 kb	Num. rare variants	Shuffle median	Shuffle 2.5-%tile	Shuffle 97.5-%tile	Fold enrichment of rare variants	Fold enrichment (95% CI)
Per variant	SV	99	4,691	10	5	17	9.9	(5.8, 19.8)
	SNV	4,188	4,830,727	3,536	3,349	3,762	1.2	(1.1, 1.3)
	Indel	917	824,836	727	664	786	1.3	(1.2, 1.4)
	Any	5,204	5,660,254	4,275	4,071	4,528	1.2	(1.1, 1.3)

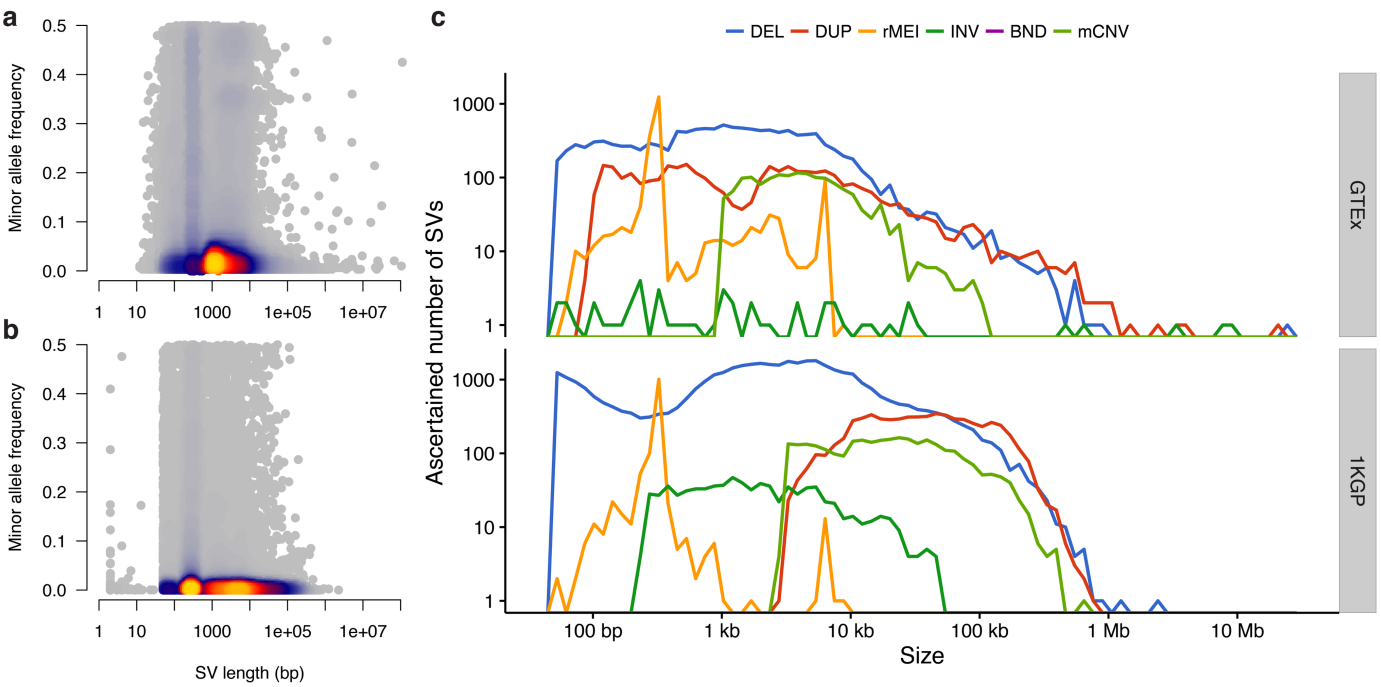
Supplementary Table 7. Fold enrichment of the co-occurrence of gene expression outliers and rare variants in same sample on a per-outlier (top) and per-variant (bottom) basis. Shuffled medians and percentiles represent the number of co-occurrences expected by chance based on 1,000 random permutations of the outlier sample names.

	Type	Variants	Outliers
Deletions	Simple	47	70
	Complex	3	4
Duplications	Simple	32	263
	Complex	6	13
Balanced	Inversions	2	4
	Complex	1	1

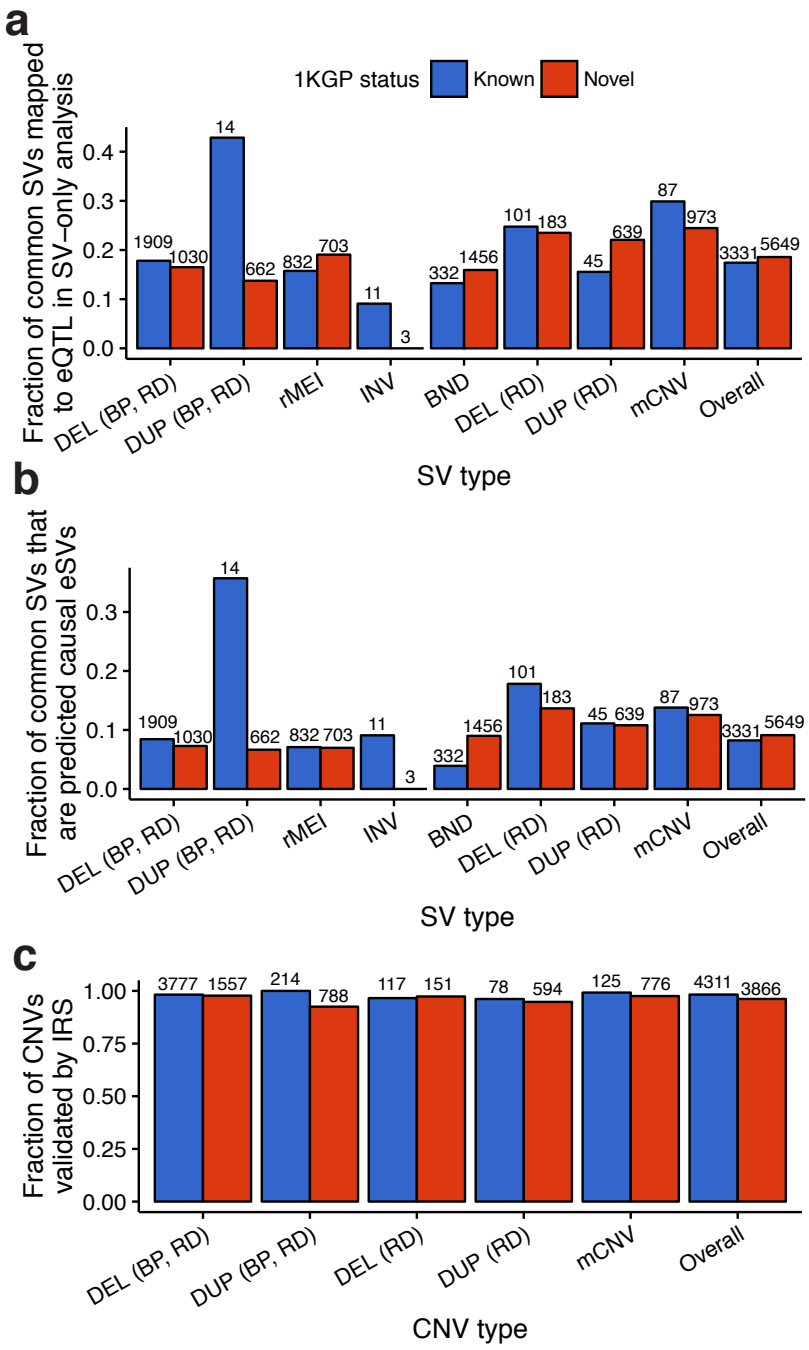
Supplementary Table 8. Distribution of simple and complex rearrangements associated with gene expression outliers. After clustering SVs into complex variants present in the same individual(s) and located no more than 100 kb away from each other, a total of 99 SVs associated with expression outliers were collapsed into 91 events.

Cluster ID	Locus	SV IDs	Sample	Class	Coding Region	Outlier Genes
1565	1:25551621-25761207	LUMPY_BND_184573, LUMPY_DUP_176134	GTEX-NPJ7	Complex dup	Yes	ENSG00000117614.5, ENSG00000117616.13, ENSG00000183726.6
1868	1:1388772-1429798	LUMPY_BND_93489, LUMPY_DUP_175996	GTEX-XGQ4	Complex dup	Yes	ENSG00000215915.5
1902	20:32168930-55372800	LUMPY_DEL_135568, LUMPY_DEL_136038	GTEX-P4QR	Complex del	Yes	ENSG00000124126.9
258	11:47153961-47186142	LUMPY_BND_186174, LUMPY_BND_186175, GS_DEL_CNV_11_47153934_47166318, GS_DEL_CNV_11_47173052_47186140	GTEX-Q2AG	Complex del	Yes	ENSG00000149179.9, ENSG00000149182.10
3276	6:127656006-127656010	LUMPY_BND_182569, LUMPY_BND_193281	GTEX-OXRL	Balanced	Yes	ENSG00000093144.14
339	11:77413211-77786061	LUMPY_DUP_177173, LUMPY_DUP_177174	GTEX-UPIC	Complex dup	Yes	ENSG00000087884.10, ENSG00000149262.12
4274	X:78417460-78425402	LUMPY_DEL_174258, LUMPY_DEL_174259	GTEX-X8HC	Complex del	No	ENSG00000147138.1
1126	16:26052128-26052227, 16:26457178-26551538, 16:26910809-27287111	LUMPY_BND_119970, LUMPY_BND_178606, LUMPY_BND_188219, LUMPY_BND_188221, LUMPY_BND_188222, LUMPY_DUP_178610	GTEX-QV31	Complex dup	Yes	ENSG00000155666.7, ENSG00000169189.12
1629	19:50401535-50401536, 19:52871602-52970915	LUMPY_BND_179537, LUMPY_DUP_179549	GTEX-X261	Complex dup	Yes	ENSG00000269834.1, ENSG00000221923.4, ENSG00000167555.9
4136	X:100747271-100747272	LUMPY_BND_195398	GTEX-OXRN	Complex dup	Yes	ENSG00000196440.7, ENSG00000198960.6

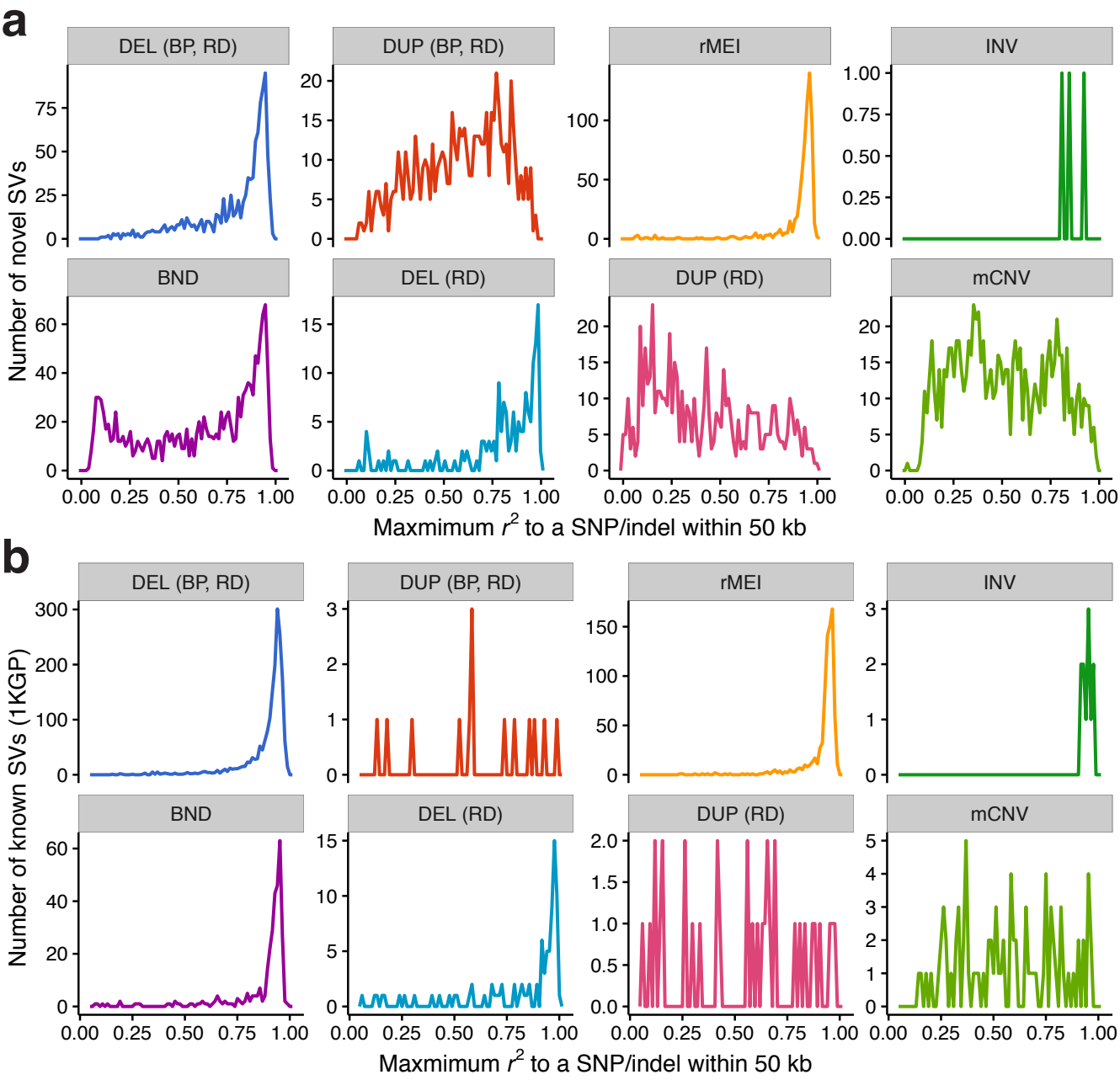
Supplementary Table 9. Complex SVs associated with expression outliers. Complex SVs were identified by clustering rare SVs located no more than 100 kb away from each other and present in the same individual(s).



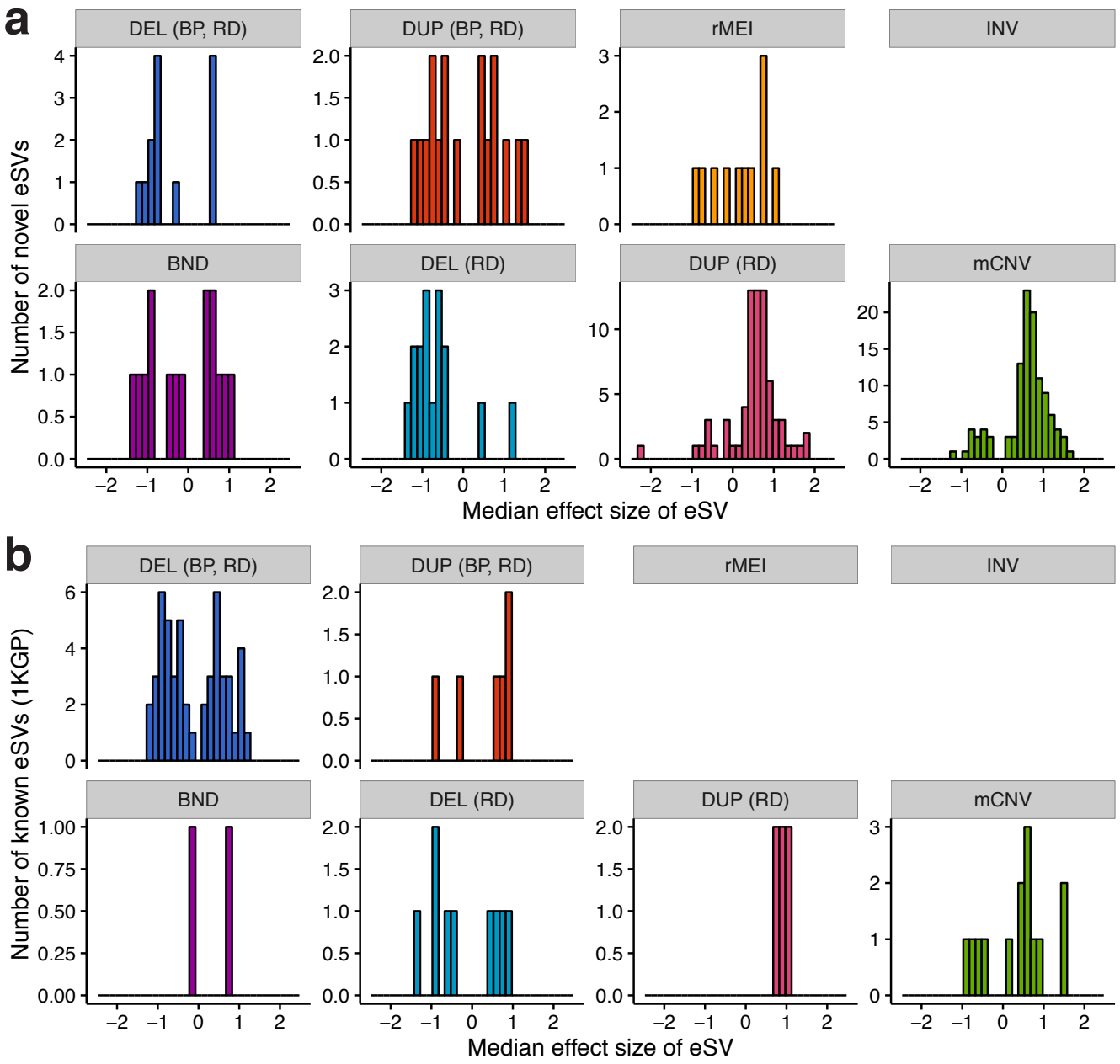
Supplementary Figure 1. Plots of the GTEx SV call set compared to the 1000 Genomes SV call set. (**a**) Heat scatter plots of SV size by minor allele frequency (MAF) showing the GTEx SV call set compared to (**b**) the 1000 Genomes Project SV call set from 2,504 individuals (Sudmant *et al.* 2015). (**c**) Size and number of ascertained variants for each SV type on a log-log axis scale for GTEx (top panel) and 1000 Genomes Project (1KGP) (bottom panel).



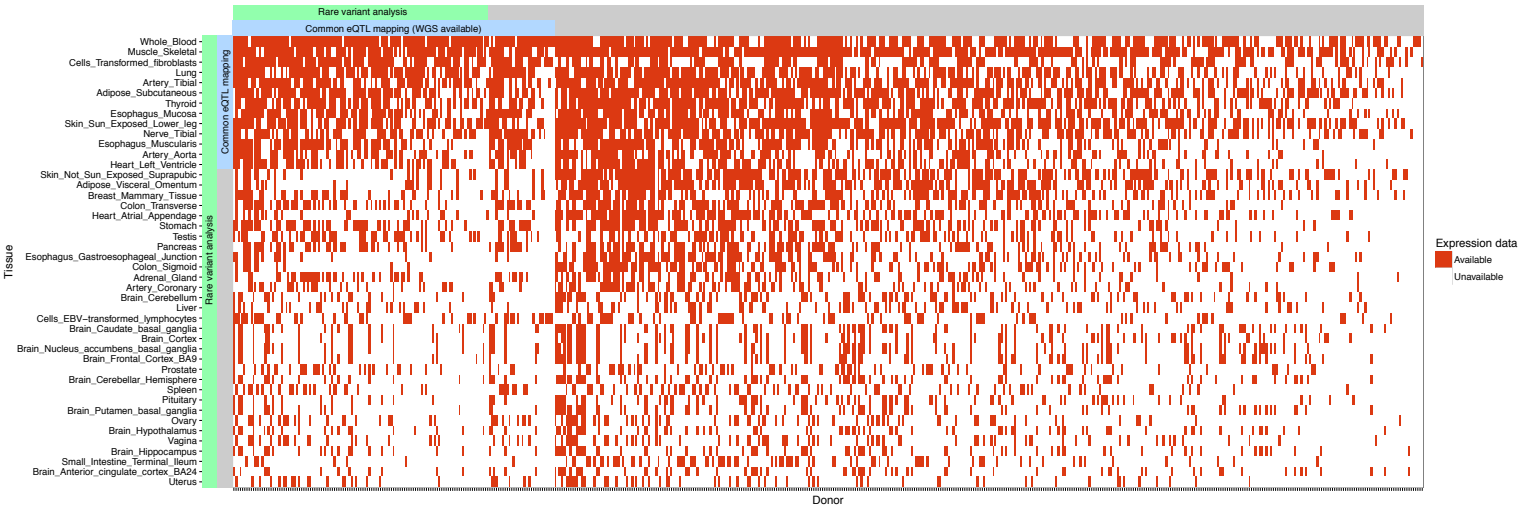
Supplementary Figure 2. Comparison between (**a**) eQTL mapping rates in when SVs are mapped to expression phenotypes in the absence of SNVs and indels ("SV-only eQTL mapping"), (**b**) the fraction of common SVs predicted to be causal eSVs from the composite causality score, and (**c**) validation rates by Intensity Rank Sum (IRS) annotator. Note that the difference in eQTL mapping rate and validation for the DUP class is most likely due to the size distribution difference apparent in Supplementary Figure 1. Text above each bar denotes its denominator.



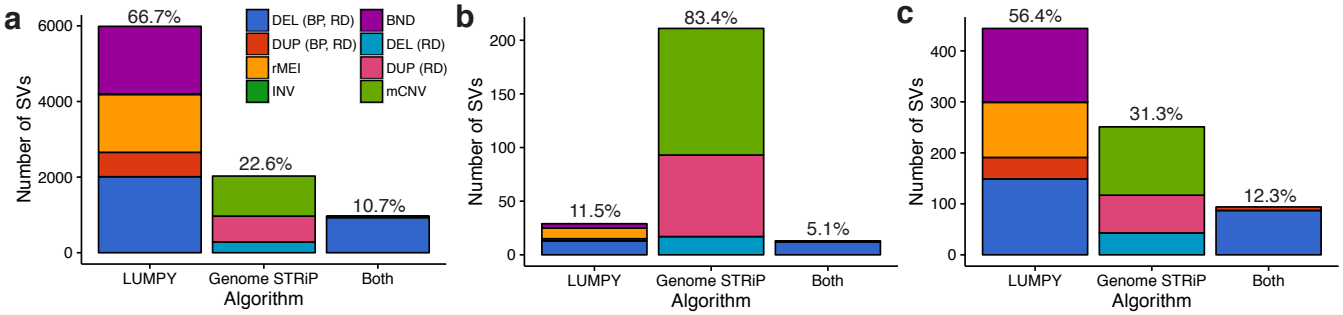
Supplementary Figure 3. Maximum LD between each SV (MAF ≥ 0.05) and a marker within 50 kb for (**a**) novel SVs and (**b**) SVs previously detected by 1KGP.



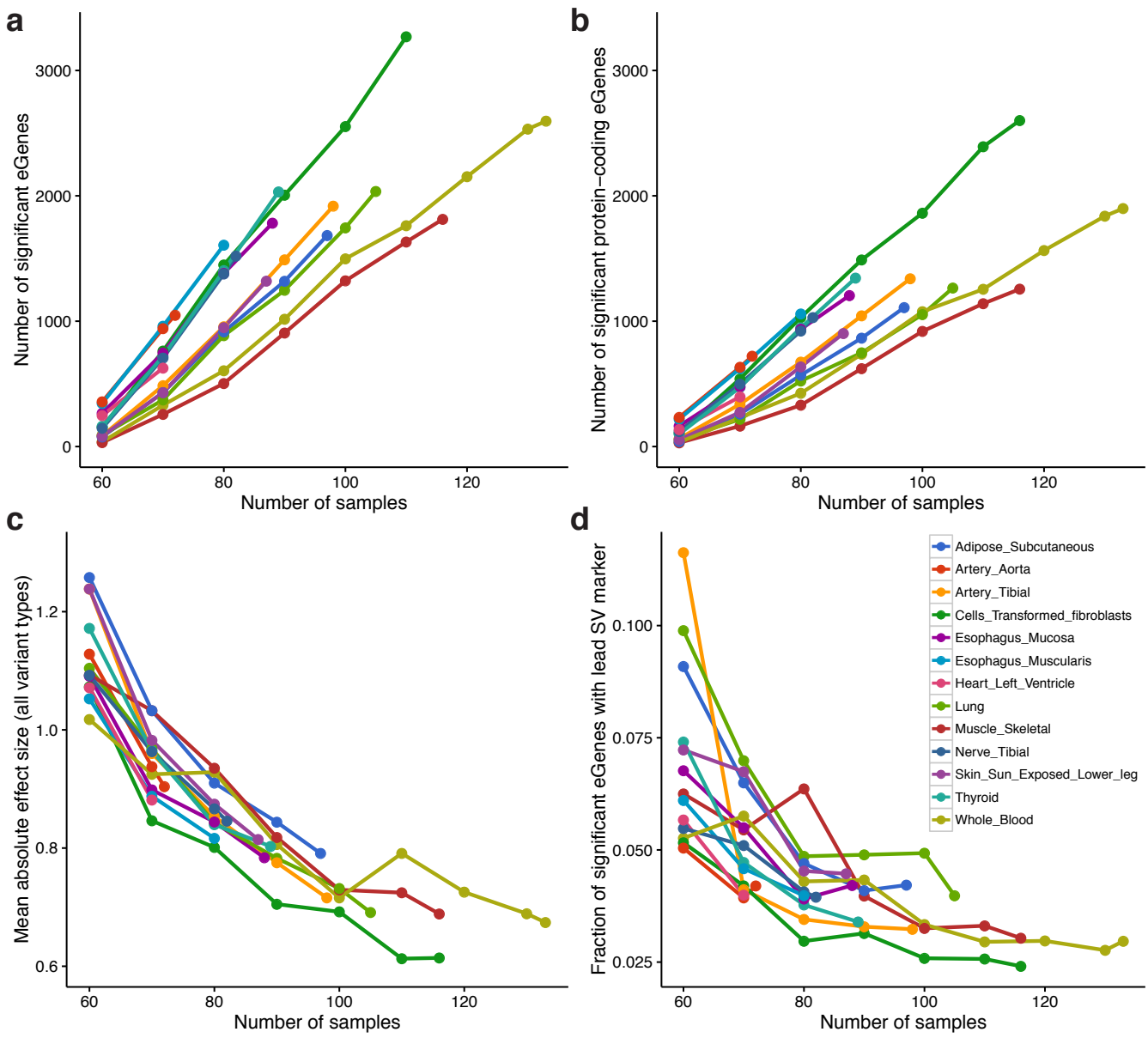
Supplementary Figure 4. Median effect size (across all tissues, eGenes) for each eSV from SV-only eQTL mapping that overlapped with at least one exon of any gene for (**a**) novel SVs and (**b**) SVs detected by 1KGP.



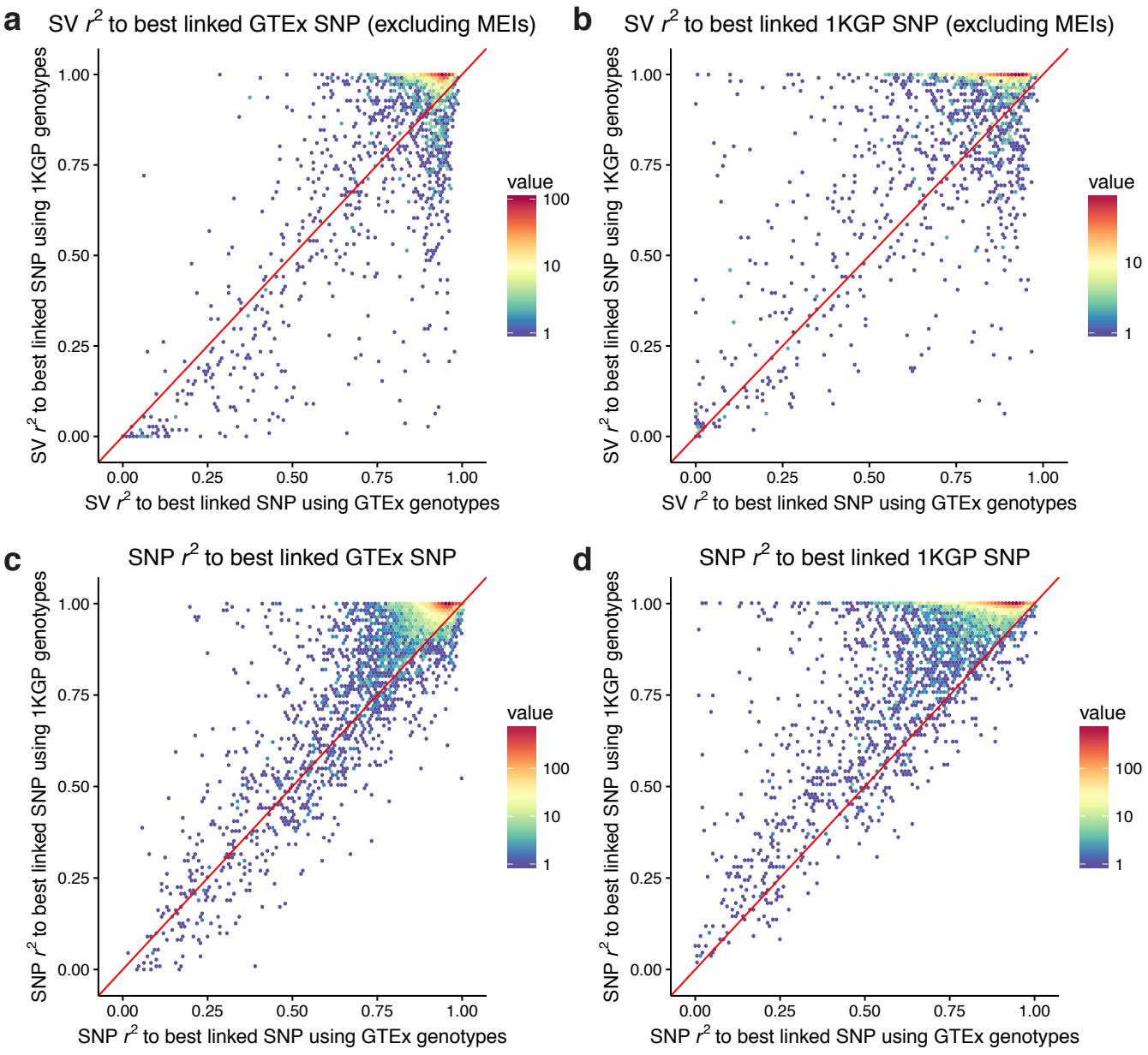
Supplementary Figure 5. Availability of RNA-seq expression data for 544 samples in the GTEx project, of which 147 samples had whole genome sequencing (WGS) data that passed quality control. Common eQTL mapping was performed on the 13 tissues with at least 70 individuals with both WGS and expression data (bounded by blue bars). Rare variant analysis was conducted on 117 individuals of European ancestry with at least 5 tissues with expression data per individual (green bars). Expression outliers for the rare variant analysis were defined using all 544 individuals and all 44 tissues.



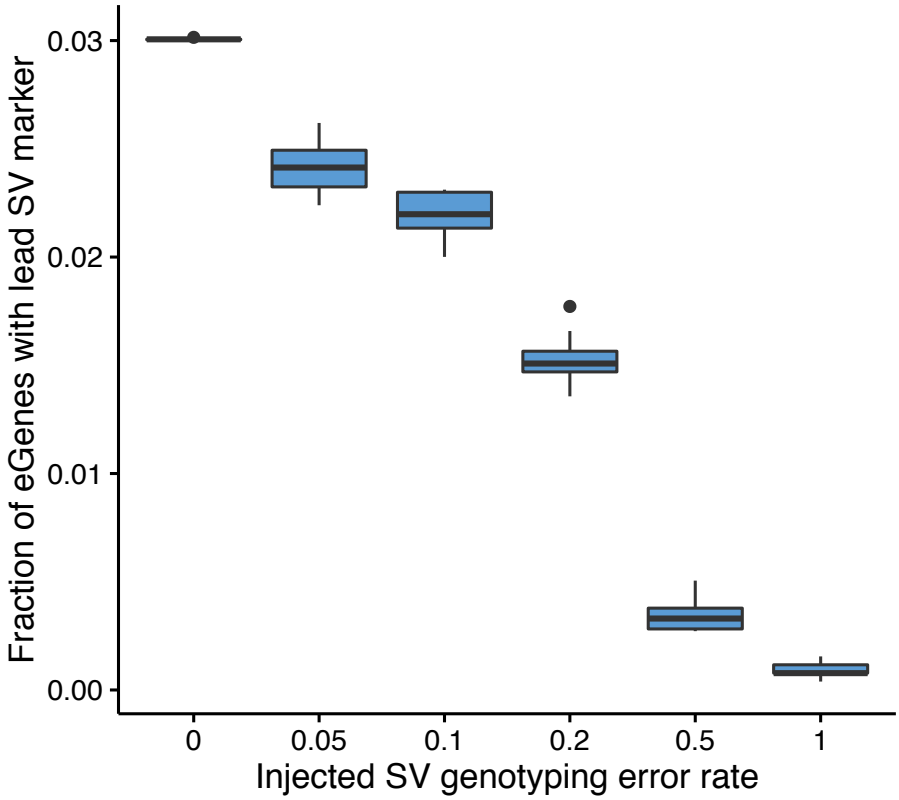
Supplementary Figure 6. Number and fraction of SVs ascertained by LUMPY, Genome STRiP, or both algorithms for (**a**) common SVs eligible for eQTL mapping, (**b**) joint eQTL mapping winners, and (**c**) predicted causal eSVs in the 90th percentile of causality scores



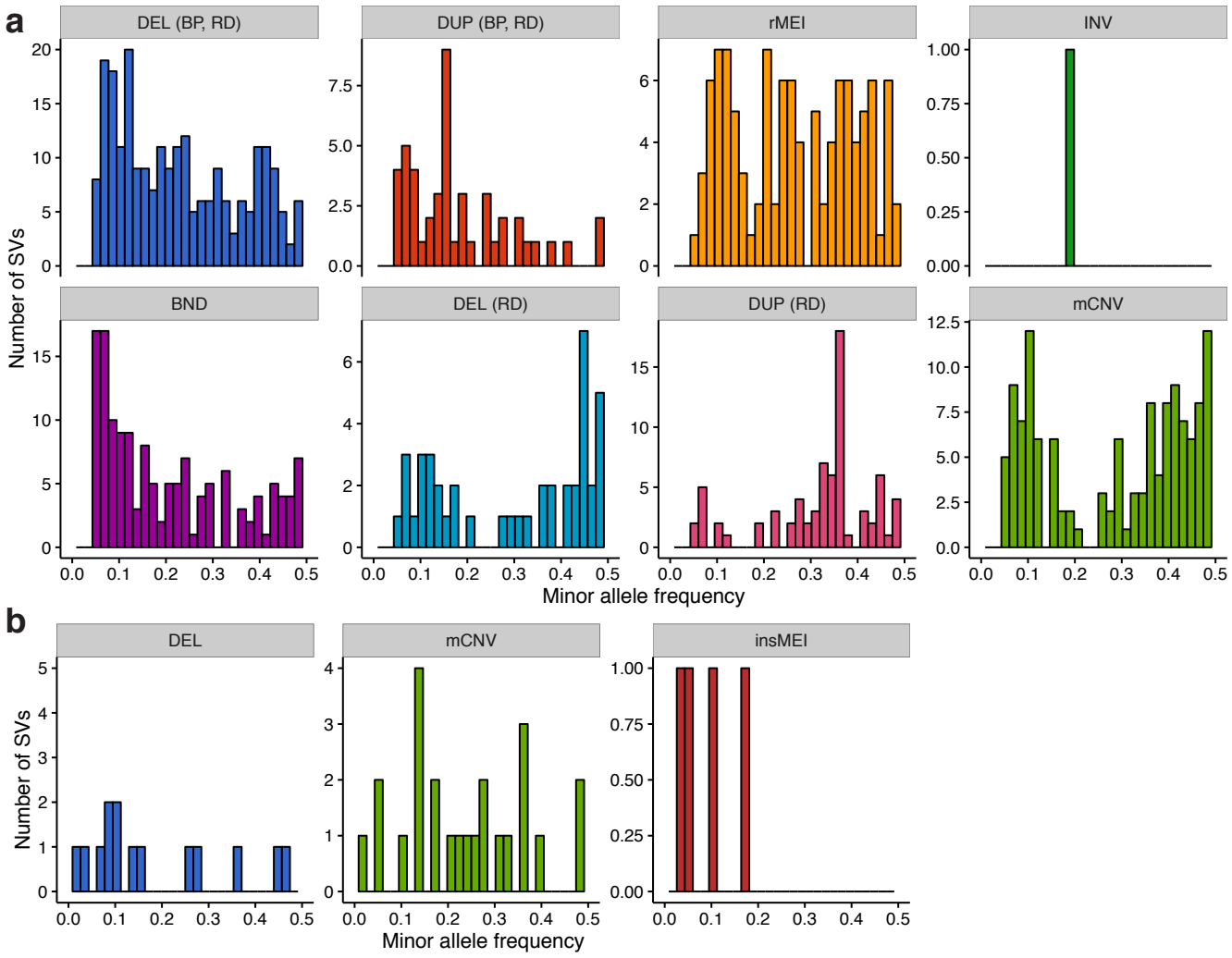
Supplementary Figure 7. Number of significant eGenes per tissue for (**a**) all genes and (**b**) all protein-coding genes. (**c**) Mean effect size for eQTLs and (**d**) fraction of eGenes with lead SV marker detected by serial downsampling within each tissue.



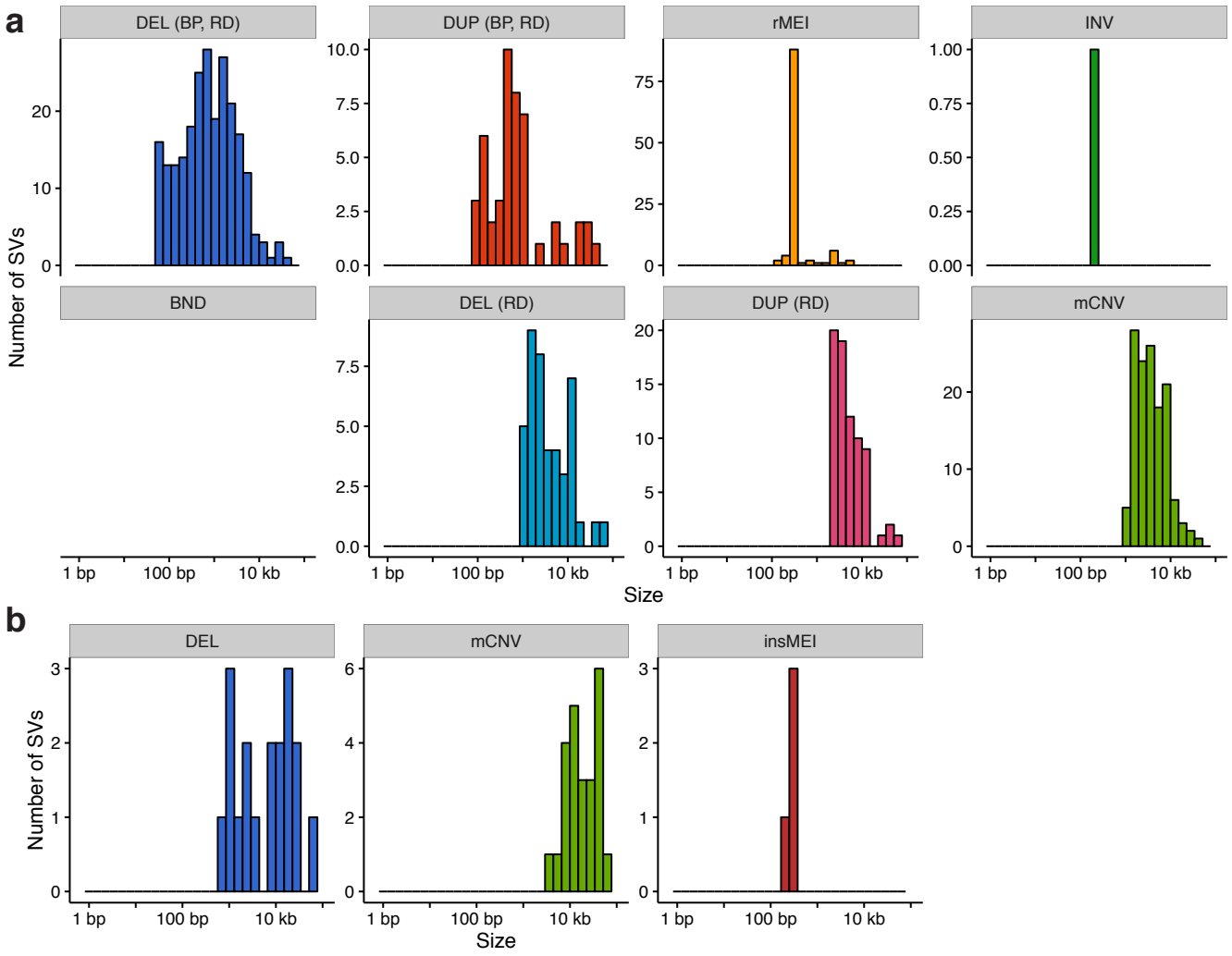
Supplementary Figure 8. Linkage disequilibrium patterns at SVs (**a**,**b**) and SNVs (**c**,**d**) discovered by both GTEx and 1KGP studies and with MAF ≥ 0.05. Shown is the maximal r^2 value to SNVs within 100 kb detected by both GTEx and 1KGP, using the most tightly linked SNV based on genotypes from GTEx (**a**,**c**) and 1KGP (**b**,**d**) studies.



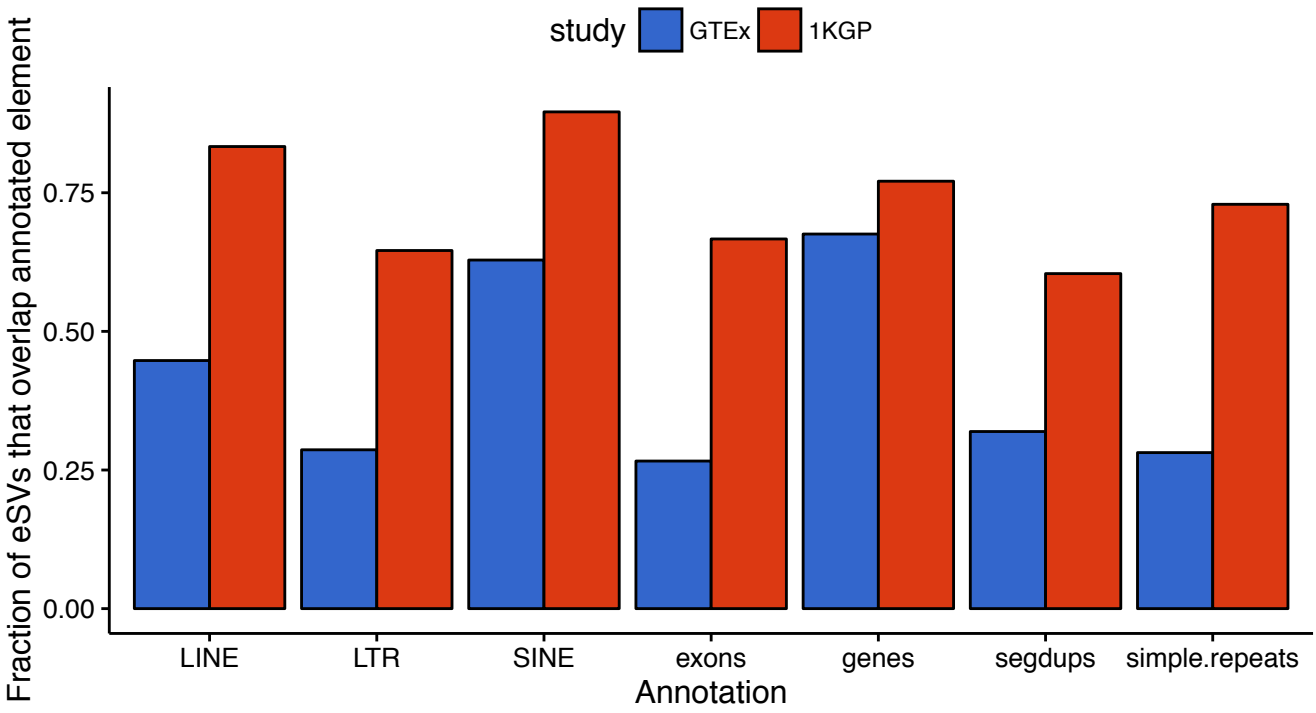
Supplementary Figure 9. The fraction of eQTLs in whole blood with an SV as the lead marker as a function of injected SV genotyping error, for which a fraction of the samples were assigned a random genotype value drawn from the allele frequency distribution at each site.



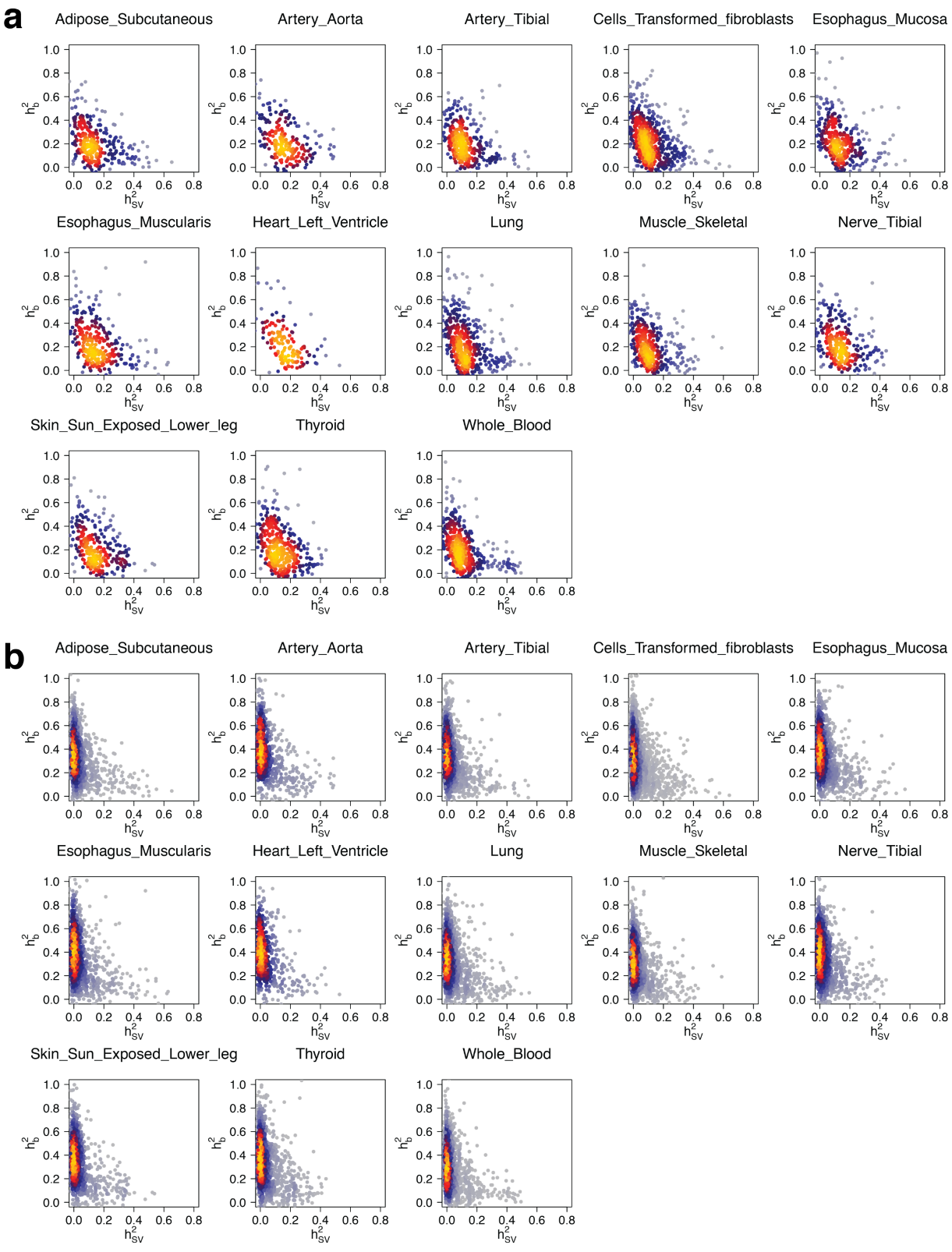
Supplementary Figure 10. Minor allele frequency of (**a**) eSVs in the 90th percentile of causality scores in our study, compared with (**b**) eSVs identified by the 1000 Genomes Project (Sudmant *et al.*, 2015).



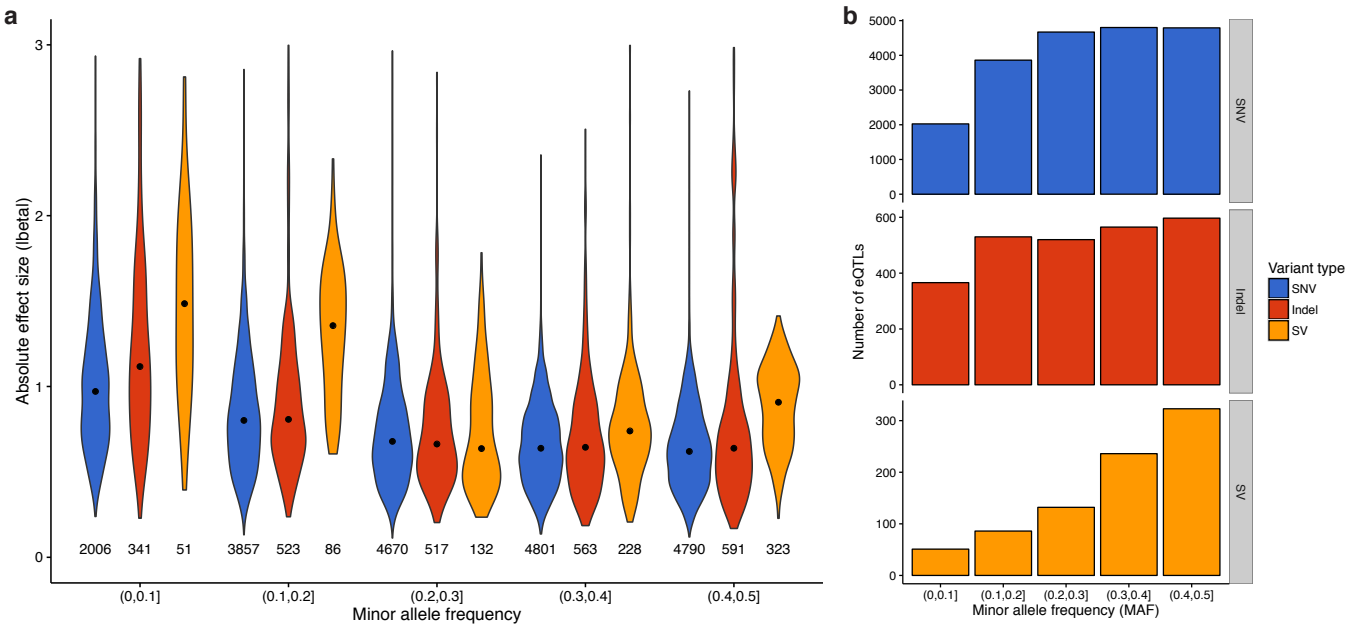
Supplementary Figure 11. Size distribution of (**a**) eSVs in the 90th percentile of causality scores in our study, compared with (**b**) eSVs identified by the 1000 Genomes Project (Sudmant *et al.*, 2015).



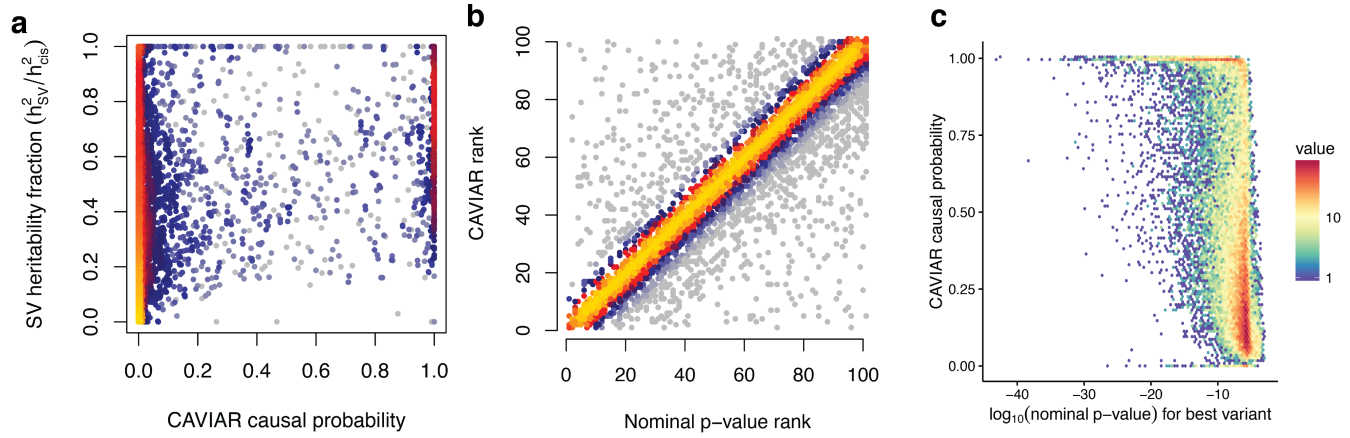
Supplementary Figure 12. Overlap with genomic elements for eSVs in the 90th percentile of causality scores in our study (blue), compared with eSVs identified by the 1000 Genomes Project (red) (Sudmant *et al.*, 2015). GTEx N=789; 1KGP N=48.



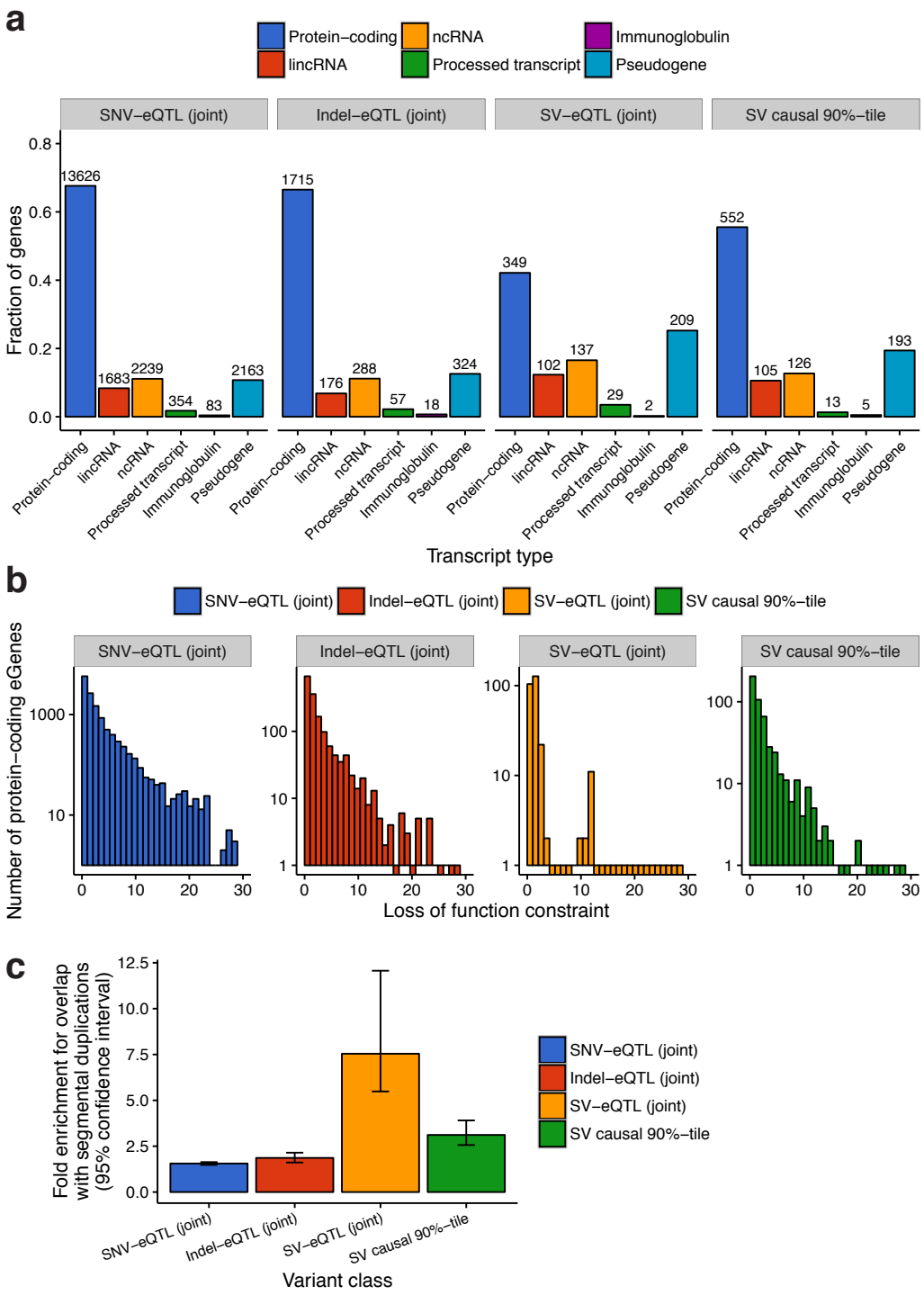
Supplementary Figure 13. Heat scatter plots (grouped by tissue) showing the heritability of each eQTL apportioned to the most significant SV in the *cis* window (x-axis) and the additive effect from the top 1,000 most significant SNVs and indels in the *cis* window for (**a**) SV-only and (**b**) joint eQTL mapping analyses.



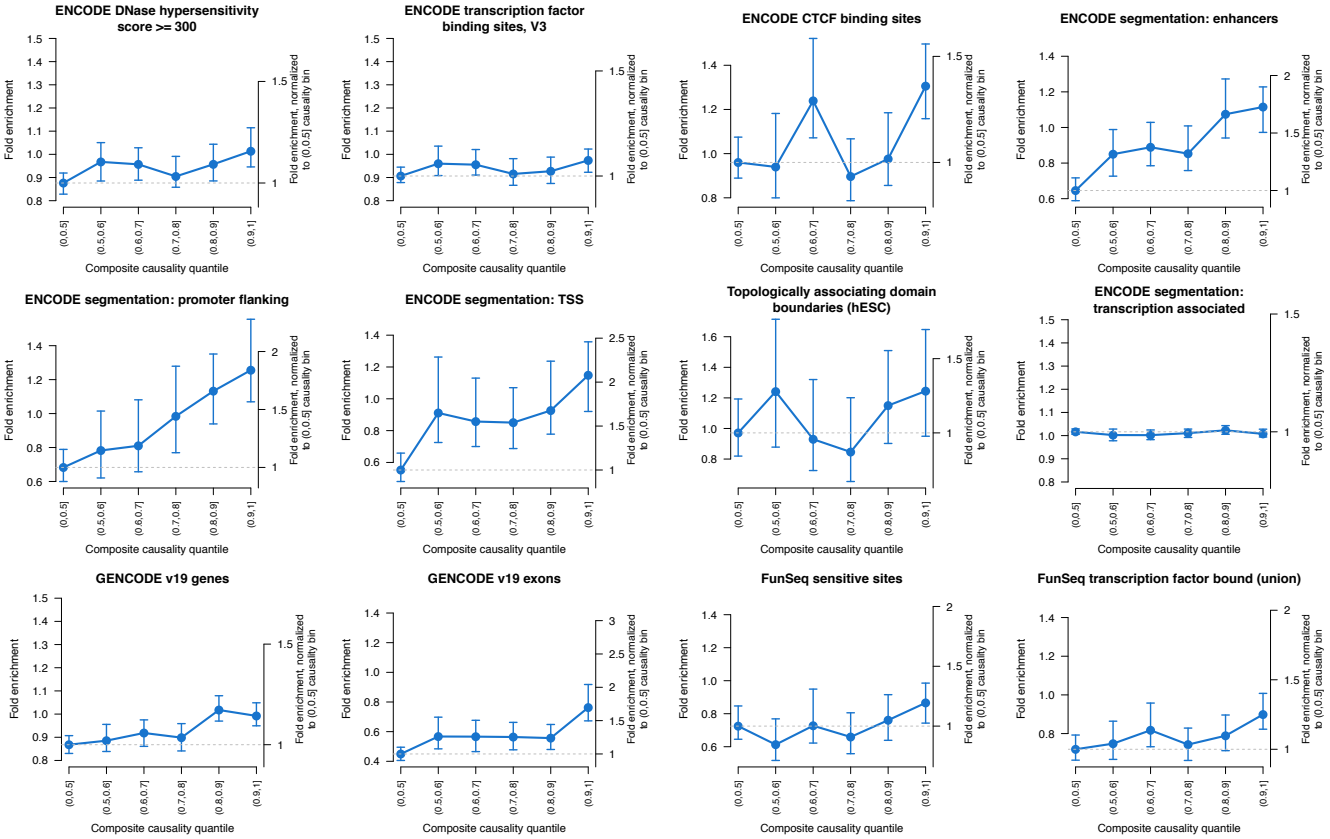
Supplementary Figure 14. Relationship between eQTL effect size and minor allele frequency (MAF). (**a**) Absolute effect size of joint eQTLs within each bin of minor allele frequency (MAF) for SVs, SNVs, and indels. Black dots represent the median of each distribution, and values beneath indicate the number of observations in each distribution. (**b**) Number of eQTLs in each bin of minor allele frequency, by variant type.



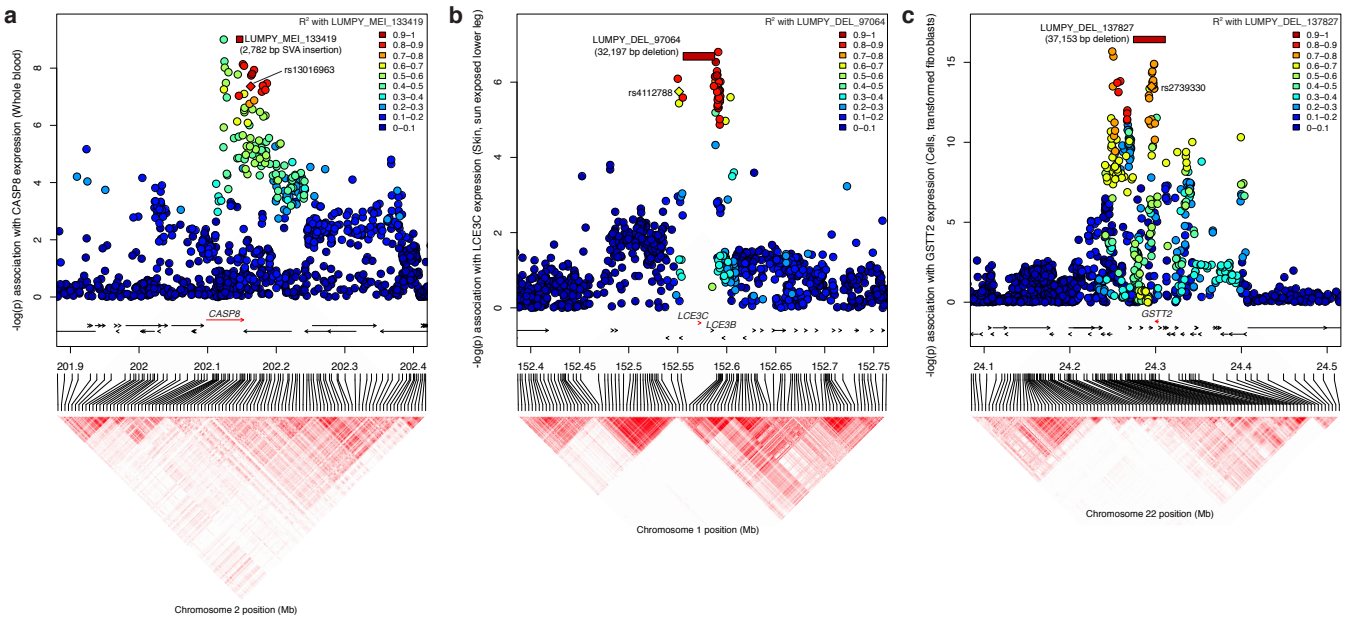
Supplementary Figure 15. (**a**) Comparison between CAVIAR causal probabilities and the SV heritability fraction (h_{sv}^2/h_{cis}^2) from the GCTA linear mixed model analysis. (**b**) Relationship between nominal p-value from FastQTL and CAVIAR causal probability and (**c**) the ranking among the 101 variants included for each eQTL by nominal p-value and CAVIAR causal probability.



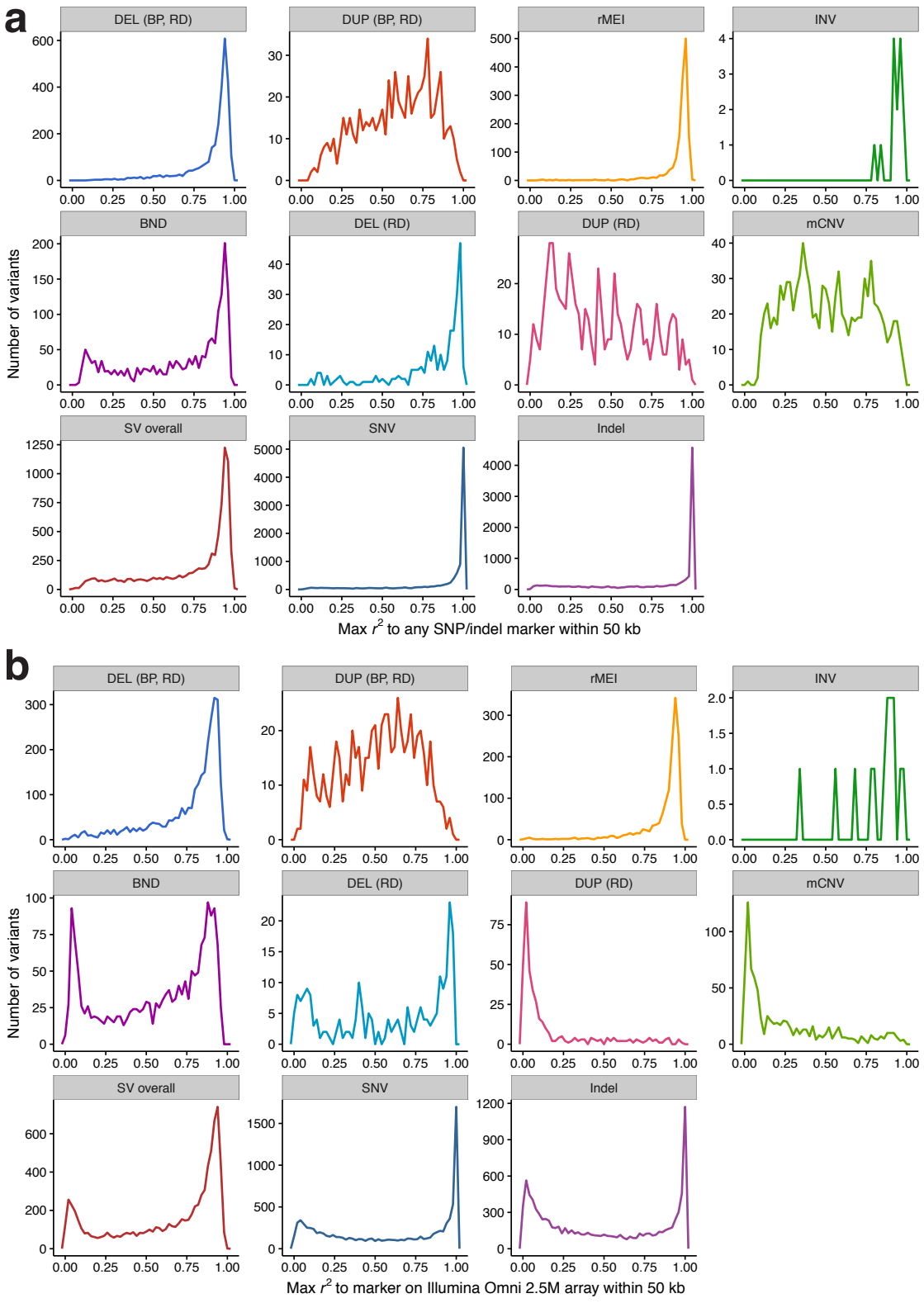
Supplementary Figure 16. (**a**) eGene classes for eQTLs with SNV, indel, or SV lead markers through joint eQTL mapping, as well as the eGene classes for predicted causal eSVs. (**b**) Loss of function gene constraint score (-$\log_{10}$(probability loss of function intolerance)) from ExAC for eQTLs with SNV, indel, or SV lead markers through joint eQTL mapping, as well as the eGene classes for predicted causal eSVs. (**c**) Fold enrichment for overlap with segmental duplications (error bars: 95% confidence interval) compared to 1,000 randomly shuffled permutations non-gapped genomic regions within 1 Mb of a gene transcript.



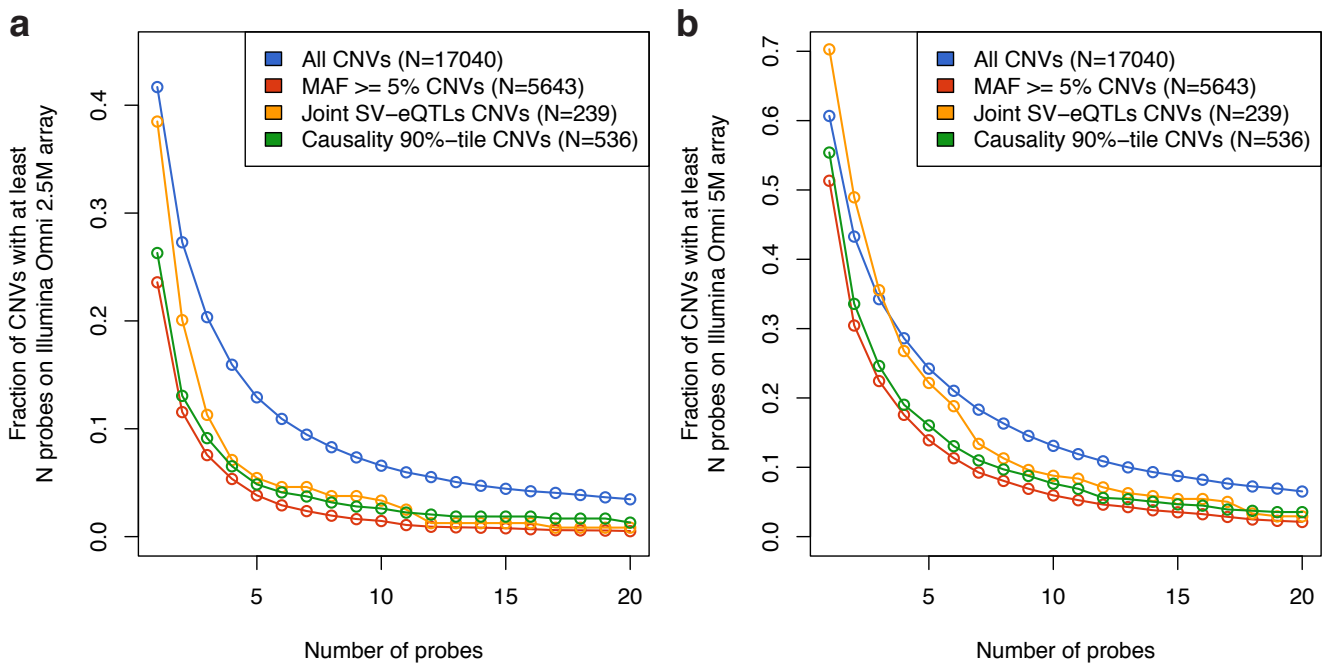
Supplementary Figure 17. Additional genomic features showing fold enrichment for SVs in each composite causality quantile bin compared to the median of 100 permutations with randomly shuffled genomic positions. SVs that overlap with exons of the eGene were excluded. Each annotated feature was allowed 1 kb of flanking sequence on either side for intersection, except GENCODE genes and GENCODE exons (no flanking sequence) and topologically associated domain boundaries (5 kb flanking sequence).



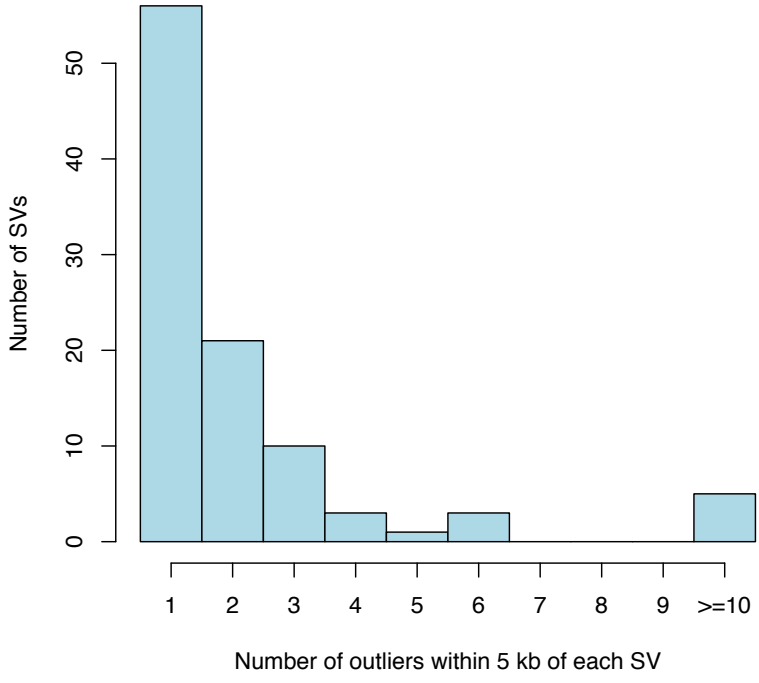
Supplementary Figure 18. (**a**) A polymorphic mobile element insertion defining exon boundaries of CASP8 reduces the gene's expression and is linked with a risk allele for melanoma (rs13016963). (**b**) A large 32,197 bp deletion of the LCE3C and LCE3B genes that was previously identified as a risk factor for psoriasis was recapitulated by our study. (**c**) A ~37 kb deletion of the GSTT2 (glutathione S-transferase theta-2) linked to a GWAS marker of circulating gamma-glutamyl transferase levels (rs2739440).



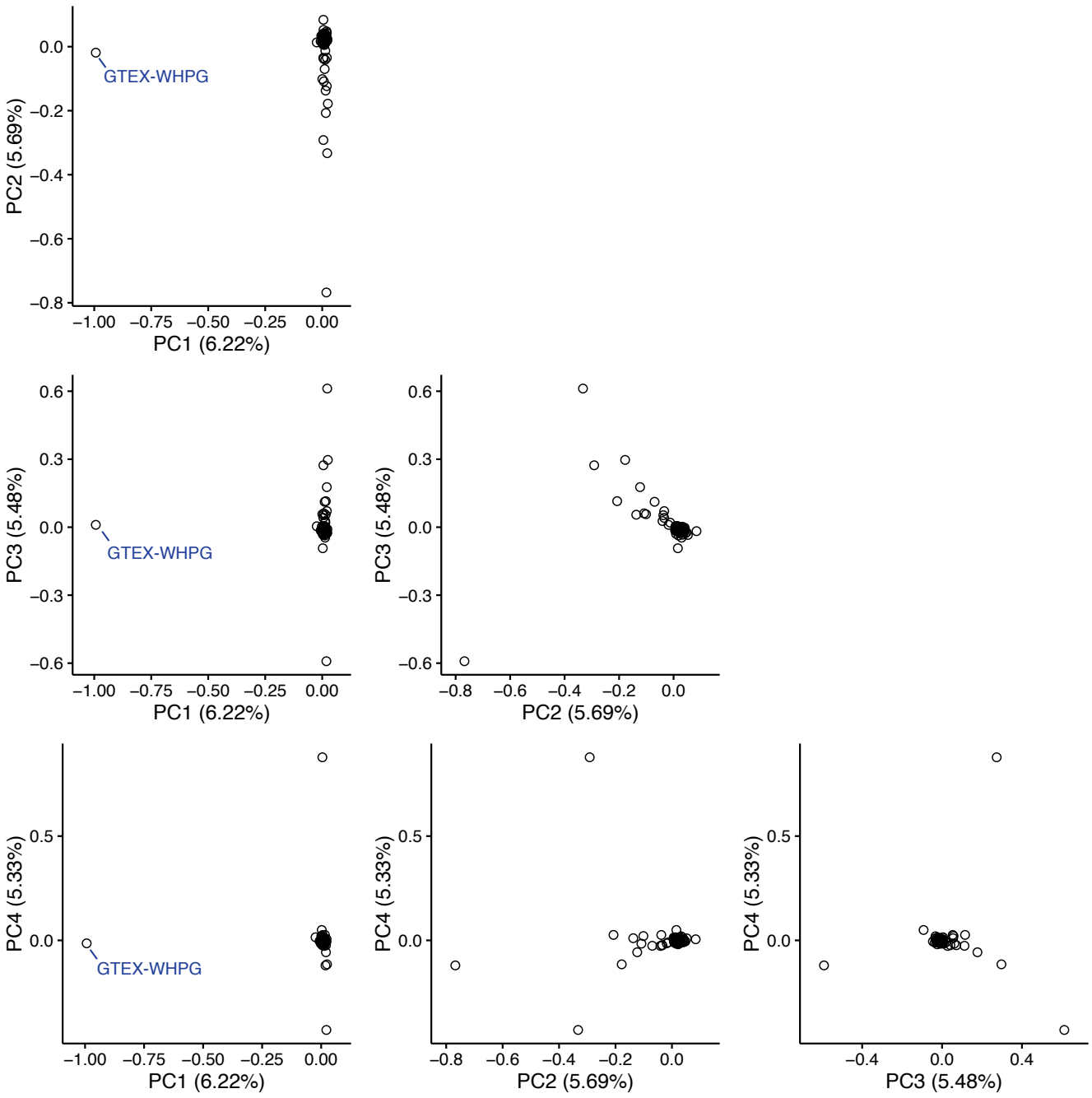
Supplementary Figure 19. Distribution of maximum LD (r^2) from variants of each type (MAF ≥ 0.05) to a marker within 50 kb among (**a**) all SNVs and indels detected by WGS and (**b**) only SNVs present on the Illumina Omni 2.5M genotyping array.



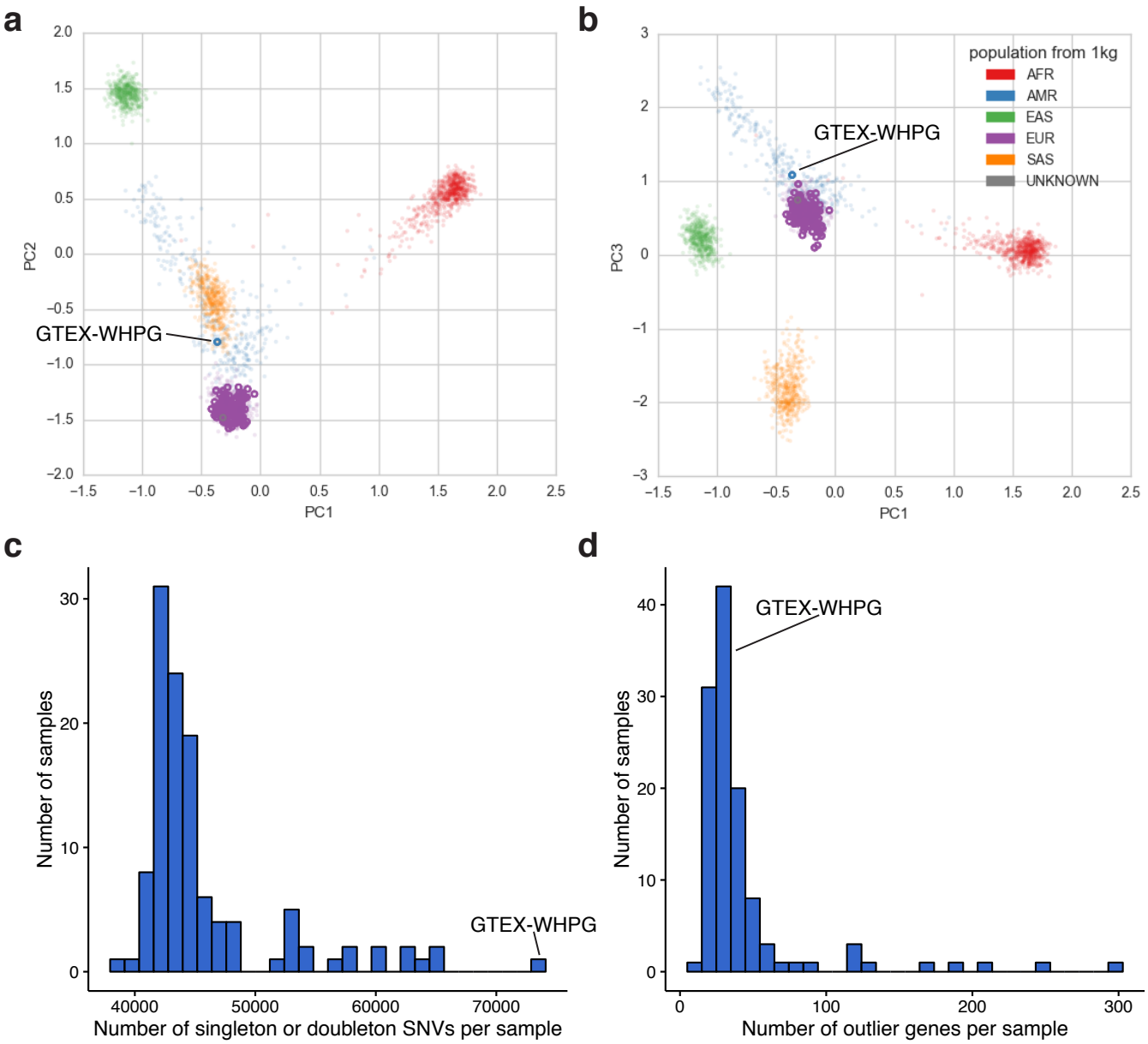
Supplementary Figure 20. Number of probes spanned by each CNV (DEL, DUP, or mCNV) on (**a**) the Illumina Omni 2.5M genotyping array and (**b**) the Illumina Omni 5M genotyping array.



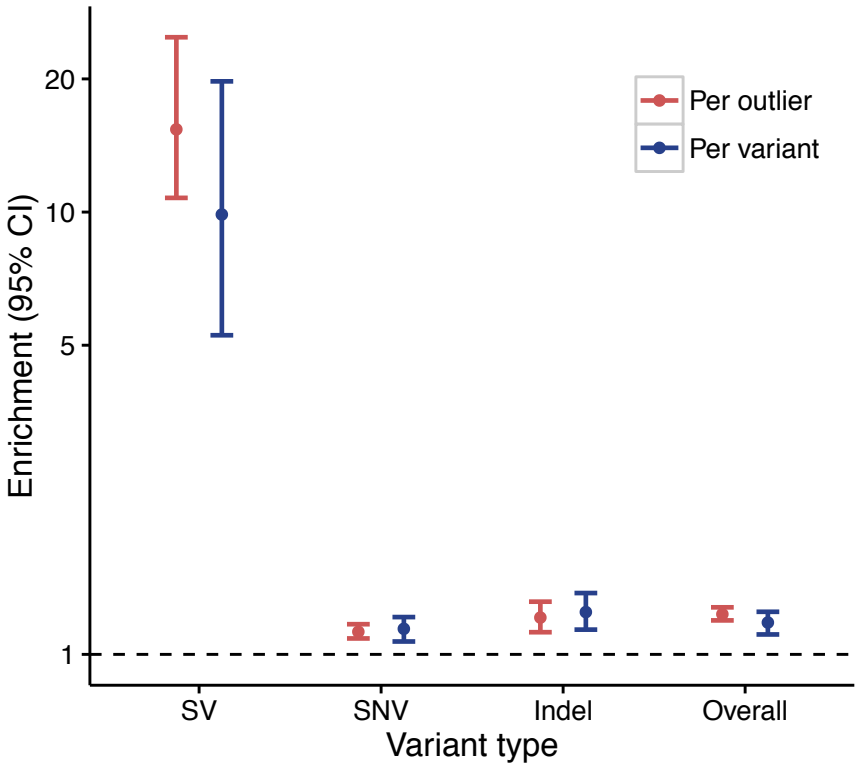
Supplementary Figure 21. Histogram showing the number of outlier genes per SV (among the SVs within 5 kb of an outlier gene in the same individual).



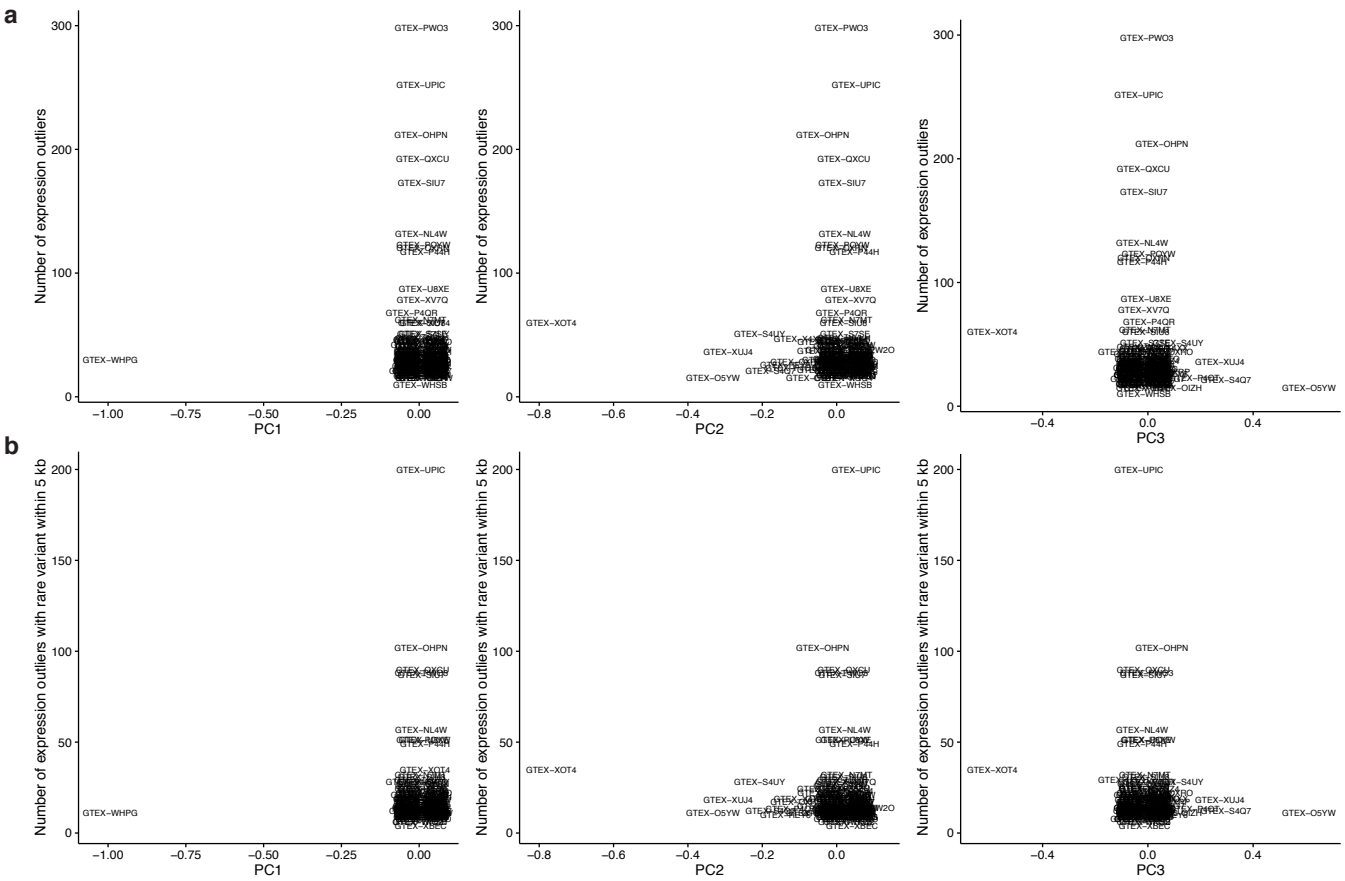
Supplementary Figure 22. Principal components based on SNV genotypes from the 117 rare variant samples.



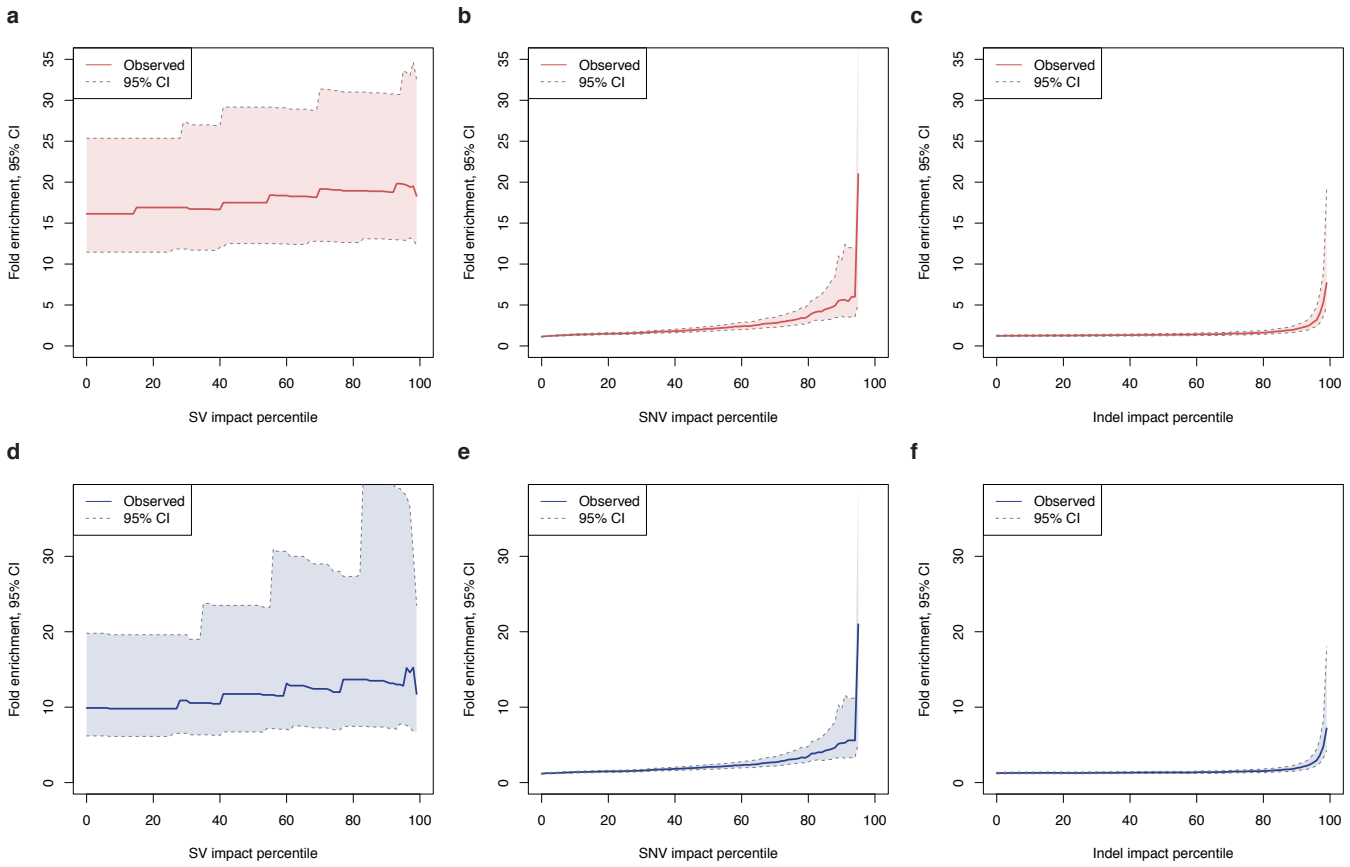
Supplementary Figure 23. Population clustering of the 117 samples used in the rare variant analysis (hollow circles) based on 1000 Genomes Project architecture showing (**a**) principal components 1,2 and (**b**) principal components 2,3. A single genetic outlier (GTEX-WHPG) clusters with admixed Americans, and has a greater burden of singleton or doubleton SNVs (**c**). However, this individual does not exhibit a greater burden of expression outliers (**d**).



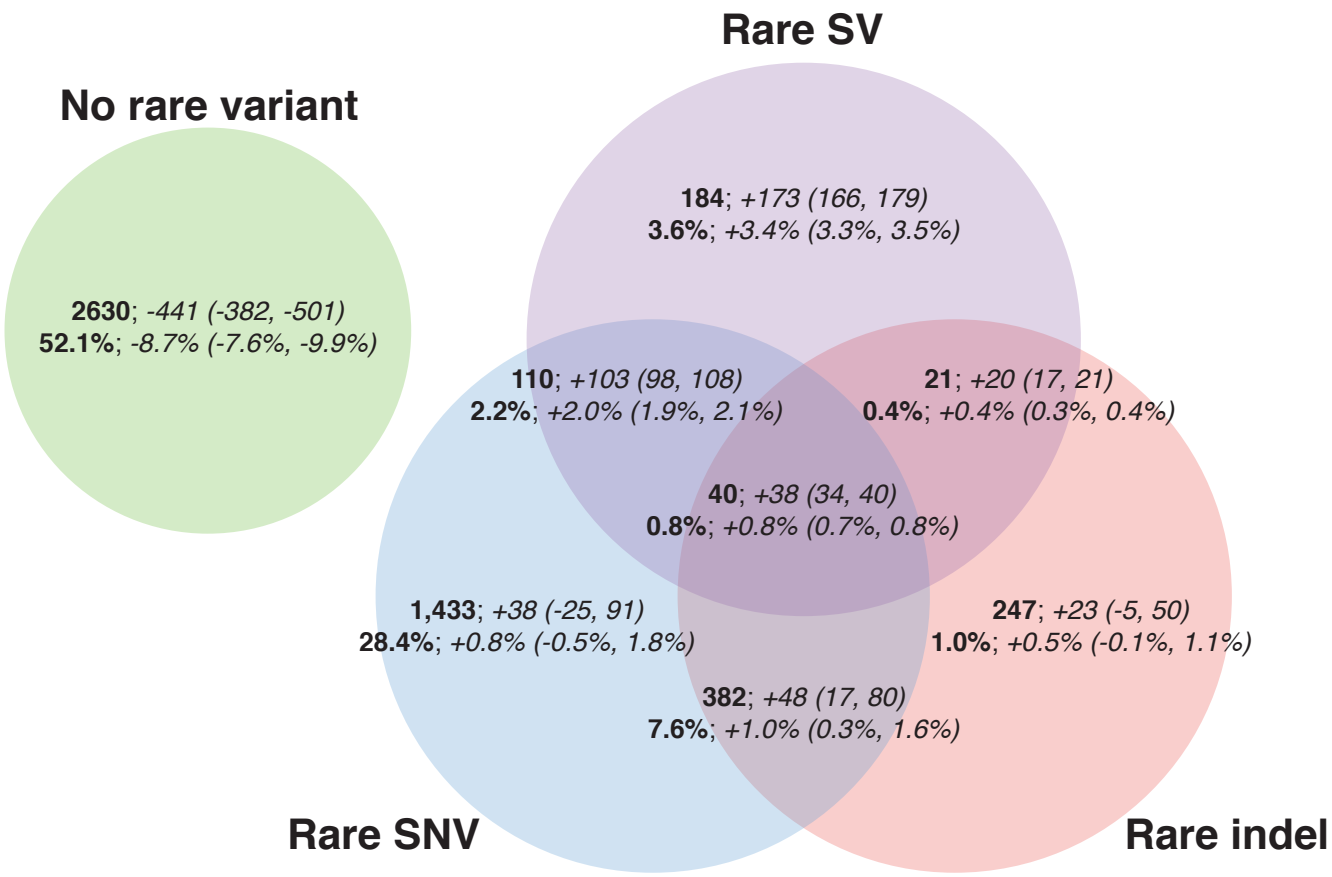
Supplementary Figure 24. Result of rare variant analysis excluding the 19 samples with more than 50,000 singleton or SNVs.



Supplementary Figure 25. Correlation between principal components for the 117 samples in the rare variant analysis and (**a**) the number of expression outliers or (**b**) the number of expression outliers with a rare variant within 5 kb in the same sample.



Supplementary Figure 26. Fold enrichment of rare variants within 5 kb of expression outliers for (**a**) SVs, (**b**) SNVs, and (**c**) indels gated on impact score percentile. Panels (**d-f**) show the fold enrichment of expression outliers within 5 kb of rare variants for (**d**) SVs, (**e**) SNVs, and (**f**) indels. For SVs, impact score percentile was based on the highest CADD scoring base in the affected interval and the confidence intervals around the SV breakpoints. For SNVs and indels the impact score percentile was derived from the CADD score of the variant.



Supplementary Figure 27. Number and percent of gene expression outliers that have a rare variant of each type within 5 kb of the gene. For each area of the Venn diagram, bold text shows the number (top) and percent (bottom) of the 5,047 expression outliers observed to be within 5 kb of a rare variant in the same individual. Italic text shows the number and percentage of outliers in excess of the median from 1,000 random permutations of the outlier dataset, with the 95% confidence intervals in parentheses.