HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2017 October 03.

Published in final edited form as:

Nat Genet. 2017 May ; 49(5): 692–699. doi:10.1038/ng.3834.

The impact of structural variation on human gene expression

Colby Chiang[1], **Alexandra J. Scott**[1], **Joe R. Davis**[2,3], **Emily K. Tsang**[2,4], **Xin Li**[2], **Yungil Kim**[5], **Tarik Hadzic**[6], **Farhan N. Damani**[5], **Liron Ganel**[1], **GTEx Consortium**[†], **Stephen B. Montgomery**[2,3,7], **Alexis Battle**[5], **Donald F. Conrad**[8,9,*], and **Ira M. Hall**[1,8,10,*]

[1]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA

[2]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA

[3]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

[4]Biomedical Informatics Program, Stanford University School of Medicine, Stanford, CA, USA

[5]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

[6]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA

[7]Department of Computer Science, Stanford University, Stanford, CA, USA

[8]Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO, USA

[9]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

[10]Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA

Abstract

Structural variants (SVs) are an important source of human genetic diversity but their contribution to traits, disease, and gene regulation remains unclear. We mapped *cis* expression quantitative trait loci (eQTLs) in 13 tissues via joint analysis of SVs, single nucleotide (SNV), and short insertion/deletion (indel) variants from deep whole genome sequencing (WGS). We estimate that SVs are causal at 3.5–6.8% of eQTLs – a substantially higher fraction than prior estimates – and that

[*]Correspondence should be addressed to I.M.H. (ihall@wustl.edu); D.F.C. (don.conrad@wustl.edu).
[†]a full list of authors appears in the Supplementary Note

URLs.

LUMPY, https://github.com/arq5x/lumpy-sv; SVTyper, https://github.com/hall-lab/svtyper; FastQTL, http://fastqtl.sourceforge.net; FunSeq, http://archive.gersteinlab.org/funseq2.1.0_data; TAD domains, http://compbio.med.harvard.edu/modencode/webpage/hic/hESC_domains_hg19.bed.

Data availability

Sequencing data, and SNV/indel genotypes are available under dbGaP accession phs000424.v6.p1.

Author contributions

C.C., A.B., S.B.M., D.F.C., and I.M.H. designed the experiments. C.C. and A.J.S. performed SV discovery and genotyping. C.C. performed common eQTL mapping, causality analyses, LD tagging, and candidate GWAS analyses. J.R.D., E.K.T., X.L., Y.K., F.N.D. identified gene expression outliers. C.C. and A.J.S. analyzed rare SVs. L.G. and I.M.H. designed SVScore annotation. D.F.C. and T.H. performed microarray-based CNV detection. C.C., D.F.C., and I.M.H. wrote the manuscript.

Competing financial interests

D.F.C. is a paid consultant of PierianDx. The authors declare no other competing financial interests.

expression-altering SVs have larger effect sizes than SNVs and indels. We identified 789 putative causal SVs predicted to directly alter gene expression: most (88.3%) are noncoding variants enriched at enhancers and other regulatory elements, and 52 are linked to genome-wide association study loci. We observe a notable abundance of rare, high impact SVs associated with aberrant expression of nearby genes. These results suggest that comprehensive WGS-based SV analyses will increase the power of common and rare variant association studies.

Introduction

Over the past decade, genome-wide association studies (GWAS) have linked thousands of common genetic variants to human traits and diseases. Fine-mapping causal variants at GWAS loci has proven difficult because the vast majority (~88%) reside in noncoding genomic regions, and in most cases the causal variant(s) and relevant gene(s) or functional element(s) are not known[1]. This has confounded the identification of therapeutic targets for precision medicine. To bridge the gap between molecular and clinical phenotypes, genome-wide eQTL scans have sought to identify genetic determinants of gene expression variation as markers of functional effect and a bridge connecting germline genetic variation to somatic cell biology[2–4]. These studies have successfully identified tens of thousands of eQTLs in a variety of human tissues.

A notable limitation of most extant eQTL studies is that, due to their reliance on SNV genotyping arrays, it has been difficult to identify the causal variants underlying eQTL associations and to judge the relative contribution of different variant classes to genetically regulated expression. Of particular interest is structural variation, a broad class of variation that includes copy number variants (CNVs), balanced rearrangements and mobile element insertions (MEIs). Structural variation is recognized to be an important source of genetic diversity – 5,000 to 10,000 SVs are detectable in a typical human genome using short-read DNA sequencing technologies – but little is known about the mechanisms through which SVs affect gene expression and phenotypic variation. Although SVs are less abundant than SNVs, which represent ~4 million variant sites per genome[5], SVs account for a greater number of nucleotide sequence differences due to their size, and may therefore exhibit outsized phenotypic effects[6,7]. Indeed, SVs have been identified as causal contributors to a number of rare and common diseases, and are generally presumed to act through their effects on gene expression[8].

Despite many noteworthy examples linking SVs to gene expression and phenotypic variation in humans, more general and quantitative questions regarding the contribution of SVs relative to other variant classes remain a matter of debate. Several studies have used low-resolution microarray technologies to study the relationship between CNVs and gene expression, but their conclusions were limited to large CNVs that are now known to comprise a small fraction of SV[9–12]. A recent study from the 1000 Genomes Project represents the most comprehensive analysis to date, using RNA-seq expression profiles from lymphoblastoid cell lines (LCLs) of 446 individuals[2] and SVs identified from low-coverage (median 7.4X) WGS data[13]. This analysis identified 9,591 eQTLs, of which 54 had an SV as the lead marker (denoted SV-eQTLs), implying that SVs are the causal variant at 0.56% of

eQTLs. However, the study's shallow sequencing depth limited SV detection power and genotyping accuracy, which are known to suffer in low-coverage data[7]. Furthermore, although gene expression is differentially regulated across tissues, prior SV-eQTL studies have focused solely on LCLs, and it is not known whether these observations extend to other cell types.

Here, we utilized multi-tissue RNA-seq expression data from the Genotype-Tissue Expression (GTEx) project to perform the first comprehensive human eQTL mapping study from deep WGS (median 49.9X) data that directly measures the contribution of SVs, SNVs, and indels.

Results

Structural variation call set

We analyzed 147 human samples using the SpeedSeq[14] pipeline for alignment (via BWA-MEM[15]), data processing and per-sample SV breakpoint detection via LUMPY[16], followed by cohort-level breakpoint merging, refinement, classification, and genotyping (Online Methods). We used complementary read-depth analysis with Genome STRiP to detect additional CNVs[17]. Together, these methods yielded a total of 23,602 "high confidence" SVs that met strict quality filters and are the basis for all subsequent analyses (Table 1).

Structural variation is known to be a difficult class of genome variation to detect and genotype accurately, and variant call sets may vary considerably in quality depending on sequencing technologies, depth of coverage, data quality and bioinformatics approaches. Several features of our call set suggests that it is extremely high quality: we detected consistent numbers and proportions of SVs per sample (Fig. 1b, Supplementary Table 1), African samples had an average of 29% more heterozygous LUMPY deletions compared to other samples (in accordance with previous observations[13]), and the site frequency spectrum for SVs mirrored that of SNVs and indels (Fig. 1c). Moreover, although we cannot directly measure genotyping accuracy, a detailed comparison to the 1000 Genomes Project SV call set shows that we detect a larger number of SVs per genome, that SVs have a similar size distribution (Fig. 1a), and that our call set has a similar (if not higher) CNV validation rate based on array-based intensity rank sum (IRS) statistics (Supplementary Note, Supplementary Figs. 1–4). This comprehensive variant call set is a powerful resource for functional analyses due to its high resolution (median breakpoint confidence interval: 34 bp) and diverse variant types including deletions (50.7%), duplications (15.0%), multi-allelic CNVs (mCNVs; 6.5%), reference mobile element insertions (rMEIs; 8.7%), inversions (0.2%), and novel adjacencies of indeterminate type (hereafter denoted as "breakends", or BNDs; 18.9%)[18].

Common eQTL mapping

We mapped *cis* eQTLs using 8,980 common SVs with minor allele frequency (MAF) ≥ 0.05 and whole transcriptome RNA-seq data from 13 tissues including 34,053 expressed genes, 18,126 of which were protein-coding (Online Methods, Supplementary Fig. 5). We defined an eQTL as an eVariant/eGene pair detected in a given tissue, and the *cis* window to include

SVs within 1 Mb of each gene transcription start site (TSS). We applied a permutation-based eQTL mapping approach using FastQTL, revealing 5,128 SV-eQTLs associated with expression differences at 2,064 distinct eGenes and 1,634 distinct eSVs (Benjamini-Hochberg false discovery rate (FDR): 10%)[19] (Supplementary Table 2).

SVs altered exons at 11.0% of eQTLs, providing a testable framework for their causal effects. Loss of function variants such as deletions or exon-disrupting MEIs are expected to decrease gene expression, exon duplications should increase gene expression, and neutral markers that tag a nearby causal variant through linkage disequilibrium (LD) should show bidirectional effects. Indeed, 507/552 (91.8%) of exon-altering eQTLs showed patterns of expression consistent with the SV class (Fig. 2a). This finding establishes strong evidence of a causal role for SVs at a subset of eGenes. In contrast, the remaining 4,566 non-exonic eQTLs (89.0%) generally exhibited bidirectional expression effects (Fig. 2a). This may reflect a complex regulatory landscape of both enhancing and repressing DNA elements, or loci at which the SV is merely in LD with the true causal variant.

To assess the relative contribution of SV, we expanded our eQTL analysis to include 6,394,161 biallelic SNVs and 801,431 indels detected by the Genome Analysis Toolkit (GATK)[20] with MAF ≥ 0.05 (Supplementary Note, Supplementary Fig. 6, Supplementary Tables 1,3). We performed joint eQTL mapping with the complete set of genetic variants, nominating a most likely causal variant for each eQTL identified. This produced 23,554 joint eQTLs across 13 tissues affecting 9,634 distinct eGenes including 828 SV-eQTLs (3.5%), 20,148 SNV-eQTLs (85.5%), and 2,578 indel-eQTLs (10.9%). The observation that SVs are the lead marker at 3.5% of eQTLs provides an initial estimate of their contribution to gene expression variation, ranging from 2.4% in transformed fibroblasts to 4.5% in skin (Supplementary Table 4). Per-tissue estimates were influenced by the number of available samples for each tissue type, and controlling for the number of available samples recapitulates relative rates of eQTLs per tissues reported in previous studies[3] (Supplementary Fig. 7). In whole blood, we observed a nearly 4-fold larger contribution of SVs to protein-coding eQTLs (2.2%) than a similar estimate from the 1000 Genomes Project, where merely 0.56% of eQTLs identified in LCLs had an SV as the lead marker[13] (Supplementary Note, Supplementary Figs. 8–12).

Fine-mapping causal variants

We next applied fine-mapping approaches to infer the probability that each locus contained a causal SV in the eGene's *cis* window. At each of the 23,554 joint eQTLs, we identified the 100 SNVs and indels in the 1 Mb *cis* window that were most significantly associated with the eGene's expression by their FastQTL nominal p-value, as well as the single most significant SV. We then used the CAVIAR software package to apportion a causal likelihood and a relative ranking to each of these 101 markers based on the magnitude and direction of association as well as the pairwise LD structure across the region[21]. This approach aims to disentangle each variant's causal contribution from its association due to LD with nearby causal markers. At 3.5% of eQTLs overall (2.4–4.4% among tissues), the SV was identified among the 101 candidates as the highest probability causal variant underlying the eQTL association.

As an orthogonal estimate of contribution of SV, we applied a linear mixed model to partition the heritability of each eGene's expression into a fixed effect from the SV and a random effect representing the cumulative heritability of the 1,000 most significant SNVs and indels in the *cis* region (Fig 2b,c, Supplementary Fig. 13). This method mirrors that of several prior studies that have examined relative contributions of distinct variant classes on a quantitative trait[22–24]. Heritability partitioning revealed that SVs account for 8.4% of total gene expression heritability when summing their effects across all eQTLs, although we note that this includes numerous loci where the SV has a very small effect. More importantly, at the 22,448 eQTLs that showed appreciable overall genetic heritability (>0.05), the SV contributed more heritability than the additive effect of the other 1,000 variants in 6.8% of cases, suggesting that the SV was the causal variant.

Taken together, the analyses presented above indicate that SVs are the causal variant at 3.5–6.8% of eQTLs, depending on the causal variant inference method. These are likely to be underestimates because the genotyping error rate for SVs is typically higher than for SNVs and indels, giving the latter a relative advantage to "win" causal variant prediction tests in regions of strong LD. For example, simulation experiments show that a 5% increase in SV genotyping error leads to a 19.6% decrease in the SV-eQTL mapping rate (Supplementary Figure 9). Although the absolute contribution of SVs to heritable expression variation is small compared to SNVs and indels, on a per-variant basis, an SV is 28 to 54 times more likely to modulate expression than an SNV or an indel. Moreover, SVs showed a 1.3-fold larger median effect size on gene expression than SNVs and indels (p-value: $< 1 \times 10^{-15}$, Mann-Whitney U test), and deletions showed a 1.4-fold larger median effect size, with direction of effect predominantly correlating with SV type (Fig. 2a). This result is unlikely to stem from differences in statistical power given the observed allele frequency distributions of each variant class, and the fact that SVs have consistently greater effect sizes across matched allele frequency bins (Supplementary Fig. 14). Together, these results demonstrate that SVs play an important and outsized role in defining the landscape of genetically regulated gene expression.

Functional context of eQTLs

We next sought to examine the genomic context of SV-eQTLs for clues into their molecular mechanisms. We hypothesized that causal SVs would be enriched in functional elements such as gene bodies, enhancers and repressors. To maximize the number of causal variants in this analysis, we first created an aggregate eQTL set containing the union of all eQTLs identified by either the SV-only or joint eQTL mapping (24,884 eQTLs affecting 10,165 distinct eGenes). We then derived a composite "causality score" that incorporates the aforementioned CAVIAR and GCTA estimates of SV causality at each eQTL by multiplying the CAVIAR posterior causal probability with the SV's *cis* heritability fraction (h^2_{SV}/h^2_{cis}) (Supplementary Fig. 15). At each eGene we selected the SV within 1 Mb that had the strongest association to the eGene's expression, and allocated these 4,398 distinct SVs into 6 bins according to their composite score quantile, with the least causal bin comprising the bottom half of composite scores. Different SV classes were represented in roughly consistent proportions across the lower causality bins, but the most causal bin had higher concentrations of multi-allelic CNVs and duplications (Fig. 3a). SVs in the most causal bin

were also enriched in segmental duplications and noncoding gene classes (Supplementary Fig. 16), which is consistent with the known concentration of SVs in architecturally complex genomic regions[7,25].

We examined the overlap between SVs and annotated genomic features to assess enrichment in various functional elements. SVs in the 90th percentile of causality scores – hereafter referred to as "predicted causal eSVs" – showed a 23-fold enrichment for altering eGene exons, amounting to 11.7% (92/789) of the predicted causal eSVs, compared to 0.4% of SVs in the least causal bin representing the lower half of causality scores (Fig. 3b). Recapitulating the trend from SV-alone eQTL mapping (Fig. 2a), the expression effect direction was highly correlated with SV type (94/106 showing the expected direction), strongly suggesting that this set of exon-altering SVs are the causal variant at their respective eQTLs. Importantly, this analysis also demonstrates that our causality score effectively distinguishes neutral from causal SVs: little to no enrichment of exon-altering SVs is observed in bins beneath the 80th percentile of scores, and enrichment rises precipitously from the 80th to the 90th percentile.

However, the majority of SVs – including 88.1% of predicted causal eSVs – do not alter eGene dosage or structure, and thus are likely to act through regulatory mechanisms. We analyzed these 4,272 noncoding SVs for enrichment in other functional elements of the genome with potential regulatory consequences. We found that several functional elements were stratified by causality score and significantly enriched in the most causal bins, including the regions within 1 kb of enhancers, the regions 10 kb upstream or downstream of gene transcripts, and regions predicted by FunSeq to be highly occupied by transcription factors[26–28] (Fig. 3c–f, Supplementary Fig. 17). In all cases, regulatory element enrichment was most pronounced in the top causality score bin – providing further evidence of the effectiveness of our scoring method – yet more moderate enrichments were also observed in lower bins.

GWAS associated SV-eQTLs

To investigate the contribution of SVs to trait-associated loci, we identified 4,874 SNVs from the GWAS catalog that were non-redundant on a per-locus and per-disease basis, were genotyped in the GTEx samples, and that had convincing evidence for disease association ($p < 5 \times 10^{-8}$)[29]. Of these, 851 were in LD ($r^2 \geq 0.5$) with a lead marker from our joint SV/SNV/indel eQTL analysis, suggesting that the GWAS hit and the eQTL are produced by the same underlying causal variant. An SV was the candidate causal variant at 3.2–14.2% of the 851 GWAS-associated eQTLs, depending on whether causality is judged based on eQTL p-value ranks or heritability partitioning via GCTA (as in the prior causal SV analysis). Combined with the eQTL fine mapping results presented above, this suggests that SVs underlie a significant fraction GWAS-associated eQTLs, indicating that our results are directly relevant to common disease biology.

We next screened for eSVs that were likely to explain prior GWAS results. We identified 52 predicted causal eSVs in LD ($r^2 \geq 0.5$) with GWAS loci, a set that shows significant enrichment with functional annotations (Supplementary Table 5, Fig. 3). Ultimately, experimental validation will be required to definitively establish the causal relationship

between any given variant and GWAS result. However, there are a number of promising candidates among these 52 loci. In one case, a 294 bp deletion is associated with decreased expression of the DAB2IP gene in thyroid tissue – apparently by disrupting an intronic enhancer – and is linked to a risk allele for abdominal aortic aneurysm[30] (r^2 = 0.57; Fig 4a). In another case, a 1,468 bp deletion in intron 10 of the PADI4 gene is linked (r^2 = 0.70) to a risk allele for rheumatoid arthritis[31] (Fig. 4b). Multiple studies have reported significant association between haplotypes of the PADI4 gene and rheumatoid arthritis, and PADI4 mRNA is expressed in pathological synovial tissues[32,33], yet none have implicated this deletion, which flanks an annotated enhancer and is predicted to be the causal variant for increased PADI4 expression in lung. Finally, we recapitulate several SVs previously recognized as clinically associated markers, including an SVA retroelement insertion to a GWAS risk allele for melanoma and esophageal cancer (r^2 = 0.85)[34–36], a ~32 kb deletion conferring risk for psoriasis[37], and a ~37 kb deletion linked to circulating liver enzyme levels (gamma-glutamyl transferase)[38] (Supplementary Fig. 18).

The extent to which SVs are tagged by other genetic markers via LD is an important consideration in the design of trait mapping studies. Notably only 58.2% of common, autosomal SVs (as well as only 51.4% of predicted causal eSVs) were in strong LD ($r^2 \geq$ 0.8) with a SNV or indel ascertained by WGS in our study, compared to 79.4% of common SNVs and 77.6% of eSNVs (by joint eQTL mapping) (Supplementary Note, Supplementary Fig. 19). This is markedly lower than a previous estimate that 79% of CNVs detected by microarray were well-tagged by nearby markers[39]. Moreover, although modern genotyping arrays are designed to detect large CNVs directly via probe intensity analysis, we found that only 3.8% of common CNVs and 4.9% of eCNVs found in our study were detectable by 5 or more contiguous probes on the Omni 2.5 platform (Supplementary Note, Supplementary Fig. 20). Indeed, when we omitted SV genotypes from joint eQTL mapping, 41.2% (341/828) of eQTLs originally ascribed to SVs did not meet genome-wide significance through SNV or indel markers (Supplementary Table 6).

Impact of rare SVs

We next sought to assess the role of rare SVs in shaping gene expression variation. In contrast to common variant eQTLs, which are caused by ancient mutations that have been subjected to natural selection, most rare variants arose recently and are more likely to have larger effect sizes and deleterious consequences[40]. Rare variants are difficult to study via traditional eQTL approaches because any given variant is observed too infrequently within a set of samples to establish a statistical relationship with gene expression[41]. However, the effect of rare variants on gene expression can be assessed indirectly via bulk outlier enrichment analyses[42]. We thus identified 5,047 gene expression outliers (median: 30 per person; range: 10–298) in which an individual exhibited aberrant transcript dosage compared to the data set as a whole (Online Methods). Next, we identified 5,660,254 rare variants (4,671 SVs, 4,830,727 SNVs, and 824,836 indels) that were positively genotyped in at most two individuals. To reduce the effects of population stratification, we limited this analysis to the 117 individuals of European ancestry with RNA-seq data in at least 5 tissues.

Rare variants were significantly enriched by 1.2-fold (95% CI: 1.2–1.3) within the gene body and the 5 kb flanking sequence of expression outliers (Fig. 5a, Supplementary Table 7). This enrichment is most pronounced for SVs (16.1-fold, 95% CI: 11.5–25.4), in which 355/5,047 (7.0%) of gene expression outliers harbored a rare SV compared to the null expectation of 22/5,047 (0.4%) in 1,000 random permutations of the sample expression values. Notably, expression-altering SVs were significantly larger than rare SVs on the whole (p-value: $< 1 \times 10^{-15}$, Mann-Whitney U test), and duplications were disproportionately represented (Fig. 5b,c). In several cases a single large SV caused multiple gene expression outliers: a 21.3 Mb duplication event was associated with 161 outliers within the region, and two large duplications (4.1 Mb and 2.5 Mb) were associated with 11 and 30 outliers, respectively. However, the enrichment of rare SVs around outlier genes was not driven by a handful of large events, since the majority of outlier-associated SVs (56/99) were only associated with a single gene (Supplementary Fig. 21), nor was it a consequence of subpopulation structure (Supplementary Note, Supplementary Figs. 22–25). Moreover, on a per-variant basis, 99/4,671 (2.1%) of the rare SVs had an expression outlier within 5 kb compared to 10/4,671 (0.2%) in the permutation set, representing a 9.9-fold enrichment (95% CI: 5.8–19.8) (Fig. 5b). These findings demonstrate that rare SVs are a common cause of aberrantly expressed genes, contributing a median of approximately 1 gene expression outlier per person. We expect this to be a large underestimate given the strict definition of expression outliers used in this study – rare variants are likely to contribute to more modest changes in expression as well.

Our data show that rare SVs alter gene expression through diverse mechanisms. Of the 99 rare SVs predicted to causally alter gene expression (permutation-based FDR: 0.2%), 79 (79.8%) are CNVs that change dosage of the aberrantly expressed gene (Supplementary Table 8). Most gene expression changes occur in the expected direction relative to the dosage alteration (Fig. 5c), but we observed 4 deletions and 2 duplications with expression effects in the opposite direction; all involve partial gene alterations, which suggests complex regulatory effects rather than simple dosage compensation. The next most common class (11, 11.1%) are noncoding CNVs that appear to act through regulatory effects and – as in the case of the SV-eQTLs (Fig 2a) – show bidirectional effects on transcription. Remarkably, we identified a number of atypical SVs with strong yet unpredicted effects on gene expression. These include a 3.6 Mb inversion associated with altered expression of 3 genes found at or near the breakpoints (one with increased and two with decreased expression), a 391 bp intronic inversion that appears to cause increased expression, a complex 3-breakpoint balanced rearrangement associated with decreased gene expression, and 9 complex CNVs involving a combination of multiple copy number variable segments and/or adjacent balanced rearrangements, including one highly complex 6-breakpoint event that resembles chromothripsis (Supplementary Table 9). These results are consistent with prior studies describing the prevalence of complex SVs in “normal” human genomes, and reveal for the first time the diversity of gene expression effects caused by rare, complex SVs[13,43].

We compared the relative contribution of rare SVs, SNVs, and indels to expression outliers. Although the overall enrichment of SNVs and indels at gene expression outliers is mild due to the high background prevalence of rare variants in these classes, enrichment increases dramatically when analyses are restricted to high impact mutations (as judged by

CADD[44,45]; Supplementary Fig. 26). Overall, we observed a net excess of 441 outliers within 5 kb of a rare variant in the same individual compared to the expected number from permutation tests, or 8.7% (441/5,047) of total outliers. Moreover, by partitioning excess outliers among SVs, SNVs and indels, we estimate that 70.0% of gene expression outliers with a genetic basis are likely explained by structural variation, whereas merely 16.0% and 13.9% are due to SNVs and indels, respectively (Online Methods, Supplementary Fig. 27). We note that this approximation assumes similar proportions of causal variants for each variant type, so may under-estimate the contribution of SNVs and indels. It also only captures the effects of rare variants within 5 kb of the outlier gene and depends on our definition of expression outliers. While the strength of the SV effect is due in part to 8 very large CNVs (> 1 Mb), GTEx individuals should be representative of the general population in terms of the prevalence of large CNVs, and the relative contribution of SVs remains noteworthy even when individuals with megabase-scale CNVs are excluded from the analysis (SV: 40.7%, SNV: 33.5%, indel: 25.9%).

Discussion

Structural variation is an important source of genetic diversity, but assessing its functional consequences has been hindered by technical challenges in detecting and genotyping SVs in large cohorts. Here, we mapped *cis*-eQTLs from 147 individuals in the GTEx project, which for the first time leverages deep WGS data and multi-tissue RNA-seq to elucidate the functional role of SVs in a broader genomic context. We estimate that 3.5–6.8% of cis-eQTLs are driven by a causal SV, a several-fold greater contribution than previously recognized, and we present novel findings demonstrating an outsized role for rare SVs on gene expression outliers.

SV detection and genotyping is known to be a challenging endeavor, and results can vary widely due to different methodological approaches and sequencing technologies. However, our study improves upon previous SV-eQTL mapping efforts in two ways. First, it harnesses SV genotypes derived directly from deep WGS reads rather than microarrays or haplotype-based genotype refinement of low coverage sequencing. Second, we capture the expression profiles of 12 human tissues and transformed fibroblasts rather than a single derived cell line. Using these methods, we observed a nearly 4-fold greater contribution of SVs to protein-coding eQTLs than a similar estimate from the 1000 Genomes Project in LCLs[13], a discrepancy that is unlikely to stem from trivial methodological differences given the similarity of eQTL mapping methods used in the two studies. Our analyses of the methodological consequences of genotyping error and haplotype-based refinement suggest that the key difference between these results is the greater sensitivity and accuracy of SV genotypes afforded by deep WGS data, underscoring the power and novelty of our study.

Bridging the gap between disease-associated loci and mechanism is a driving motivation for eQTL studies, since noncoding variants encompass approximately 88% of GWAS loci[1], but their gene targets and regulatory effects are often difficult to predict. We applied fine-mapping approaches to identify 789 putative causal SV-eQTLs. We confirmed previous reports that coding SV-eQTLs generally exhibit an effect direction consistent with SV type[10], and we observed that noncoding SVs with strong causality predictions were

significantly enriched for overlap with known regulatory elements. Given the paucity of causal variants discovered in the human genome to date, the 789 putatively causal SV-eQTLs identified here – of which 52 are linked ($r^2 \geq 0.5$) with GWAS findings – will be a valuable resource for future functional studies.

Finally, we analyzed the functional impact of rare variants, the majority of which are relatively new alleles with limited exposure to purifying selection. Our study assessed bulk enrichment of rare variants in proximity to gene expression outliers, demonstrating for the first time that rare SVs are a common cause of aberrantly expressed genes in the human population, and that rare SVs contribute a large fraction of gene expression outliers relative to SNVs and indels. This result implies that thorough ascertainment of SV will significantly increase the power of rare variant association studies and the efficacy of WGS-based disease diagnosis.

An important extension of this work lies in guiding the design of future trait-mapping studies. Our results show that SVs comprise a significant and outsized fraction of expression-altering genetic variants, a substantial portion of which are untested in typical association studies. As human genetics moves deeper into the era of whole genome sequencing, it has become possible to include all forms of genetic variation in cohort studies and clinical practices. Comprehensive analysis of structural variation will be a critical aspect of these efforts.

Online methods

SV call set generation

We acquired 148 deep whole genome BAM files from the GTEx V6 data release (dbGaP accession phs000424.v6.p1). Post-mortem donors were consented by their next-of-kin, as described previously[3]. We excluded one sample (GTEX-WHWD-0002) due to an abnormal insert size distribution, which confounds SV detection. We realigned the remaining 147 whole genomes to GRCh build 37 plus a contig for Epstein-Barr virus using SpeedSeq v0.0.3 (BWA-MEM v0.7.10-r789) according to published practices[14,15]. We ran LUMPY v0.2.9 on each sample with the default parameters in the LUMPY Express script, using the published list of excluded genomic regions from SpeedSeq as well as the -P option to output probability curves for each breakpoint[16]. We merged the 147 VCF files using the l_sort.py and l_merge.py scripts included in LUMPY with the “-product” option and 20 bp of slop, simultaneously combining variants with overlapping breakpoint intervals while refining their spatial precision based on the probability curves to create a cohort-level VCF. We pruned remaining variants with nearly overlapping breakpoint intervals (within 50 bp) by selecting the single variant with the highest allele frequency among the overlapping set. Next, we genotyped each sample with SVTyper v0.0.3, which performs breakpoint sequencing of paired-end and split-read discordants[14]. We define the term “allele balance” as the ratio of non-reference to total reads at each breakpoint. Allele balance serves a proxy for genotype that is tolerant to inefficiencies in aligning the alternate allele for SVs, and is used for most analyses in this paper. We then used CNVnator v0.3 to annotate the copy number of each spanning variant (putative deletions, duplications, and inversions).

We applied several filters to the LUMPY call set to flag low quality SVs. Since 68 samples were sequenced on the Illumina HiSeq 2000 platform and 79 on the Illumina HiSeq X Ten platform, we flagged variants whose linear correlation (r^2) between genotype and sequencing platform exceeded 0.1. We further flagged deletions lacking split-read support that were smaller than 418 bp, which was measured to be the empirical minimum deletion size at which all insert size libraries were able to discriminate between concordant and discordant reads with 95% certainty. We determined that three samples (GTEX-NPJ8–0004, GTEX-T2IS-0002, GTEX-OIZI-1026) had abnormal read-depth profiles and we therefore flagged SVs private to any of those samples, effectively excluding them from rare variant analyses. Finally, we flagged variants with a mean sample quality (MSQ, a measure of genotype quality among positively genotyped samples that is independent of allele frequency) of less than 20 as low quality.

Next, we reclassified variant types, requiring that deletions and duplications exhibit correlation between read-depth and the allele balance at the breakpoint. For SVs positively genotyped in at least 10 samples, we fit a linear regression and required a slope of at least 1.0 in the appropriate direction (positive for duplications, negative for deletions) and $r^2 \geq 0.2$. For the remaining low frequency SVs we required that > 50% of positively genotyped samples must be read-depth of > 2 MAD (median absolute deviation) (in the correct direction for deletion/duplication) and > 0.5 absolute copies from the median of reference genotyped samples. For low frequency SVs on the sex chromosomes, we limited the above criteria to the gender with more non-reference individuals to avoid gender confounders. We identified mobile elements insertions in the reference genome (rMEIs) as SVs with breakpoint orientations indicative of deletions that had > 0.9 reciprocal overlap with an annotated SINE, LINE, or SVA element with sequence divergence of less than 200 milliDiv, based on RepeatMasker annotations. Due to limitations of our pipeline, we were only able to detect mobile elements inserted into the reference genome based on their absence in other genotyped samples.

We ran Genome STRiP 2.00.1602 according to the best practices workflow for deeply sequenced genomes, using a window size of 1,000 bp, window overlap of 500 bp, reference gap length of 1,000 bp, boundary precision of 100 bp, and minimum refined length of 500 bp. We flagged CNVs for platform bias and the three samples with abnormal coverage profiles as described above. For rare SVs detected by Genome STRiP (private or doubletons in our call set) we merged fragmented variants of matching types with identical genotypes within 10 Mb of each other whose combined footprint encompassed at least 10% of their span.

We then unified the LUMPY and Genome STRiP call sets while collapsing redundancies. Because LUMPY variants are substantially more precise and have well-defined confidence intervals, we retained LUMPY calls when an SV was detected by both algorithms with a reciprocal overlap of > 0.5 and a matching variant type (mCNVs were allowed to merge with either LUMPY duplications or LUMPY deletions). To ensure that SVs would be merged even when the Genome STRiP call was fragmented, which occurs fairly often with GTEx WGS data, we also merged calls where > 0.9 of a Genome STRiP CNV was contained in a LUMPY SV of the same type (or mCNV) and their correlation between LUMPY allele

balance and copy number had $r^2 > 0.25$. This last step ensures that the merged variants have a high degree of co-occurrence among samples, and are not simply independent variants that inhabit the same genomic interval.

We measured MAF for LUMPY SVs as the ratio of minor alleles to total alleles in the population. For Genome STRiP variants we defined MAF as the fraction of samples that deviate from the mode copy number value in the population.

We defined a high confidence call set from the variants that had not been flagged by the aforementioned filters, at least 50 bp in size, and located on the autosomes or the X chromosome. In general, high confidence SVs had to be supported by multiple independent evidence types. Since LUMPY deletions and duplications were identified by paired-end and/or split-read evidence and had also met requisite read-depth support from reclassification they were automatically considered high confidence. Similarly, rMEIs were considered high confidence based on support from reference genome repeat annotations. LUMPY inversions and BNDs (unclassified breakends) were required to have a minimum variant quality score of 100. Inversions were further required to show evidence from both sides of the event, and at least 10% of supporting reads derived from each of split-read and paired-end evidence types. LUMPY BND variants were required to have at least 25% of supporting reads derived from each of split-read and paired-end evidence types. Genome STRiP variants that were merged as described above and those with GSCNQUAL score ≥ 10 were considered high confidence. This set of 23,602 variants served as the basis for all analyses in this paper.

We estimated the FDR of this SV call set with Genome STRiP's Intensity Rank Sum (IRS) annotator for *in silico* CNV validation using Illumina Omni 5M SNV genotyping array. Array data was available for 131 of 147 samples, and we used the log R ratio ($\log_2(R_{observed}/R_{expected})$) of intensity values from Illumina GenomeStudio as IRS input. We tested 7,575 of 17,040 CNVs (deletions or duplications, excluding reference MEIs) that spanned at least one probe.

Array-based CNV calling

For array-based CNV calling using for quality control, DNA samples from each GTEx donor were run on Illumina DNA arrays (N=186 samples on the Illumina 5M platform, N=275 samples on the 2.5M platform). GenomeStudio software was used to generate B Allele Frequency and Log R Ratio data for each array experiment, and these normalized probeset summaries were used as the primary data for calling CNVs using plumbCNV, an R package based on the popular and widely used PennCNV algorithm[46,47]. We constructed custom *.pfb files for each array platform using the full set of GTEx data for each platform. Prior to CNV calling, we perform sample QC on each array experiment, and removed samples with abnormalities in either the mean or variance of the Log R Ratio across the entire genome. Principal components analysis was then used to correct for batch effects in the Log R Ratio data. PennCNV was then used to call CNVs with default parameters. Raw CNV calls were cleaned by a) merging adjacent CNVs separated by a gap < 20% of the size of the smaller CNV, and b) removing CNVs with >50% overlap with immunoglobulin loci, telomeric and centromeric regions. Post-calling sample QC was performed to identify and

remove individual with an excess of CNV calls based on the poisson expectation derived from the total set of samples.

Common eQTL mapping

We mapped *cis*-eQTLs to scan for significant associations between common variant genotypes and gene expression in all tissues for which there were ≥ 70 individuals with both WGS data and RNA-seq data. These include the following 12 tissues: whole blood, skeletal muscle, lung, tibial artery, aortic artery, adipose (subcutaneous), thyroid, esophagus mucosa, esophagus muscularis, skin (sun-exposed), tibial nerve, muscle (skeletal), as well as transformed fibroblasts. For convenience, we refer to the transformed fibroblasts as a tissue type throughout this study. Biospecimen collection was performed as previously described[3]. RNA-seq data from each tissue was aligned with Tophat v1.4 using GENCODE v19 gene annotations by taking the union of exons for gene level quantification, and RPKM values were calculated with RNA-SeQC[48–50]. Reads were required to align exclusively within exons or span them (without aligning to intronic regions), align in proper pairs, contain a maximum of six non-reference bases, and map uniquely to the gene. We note that because these gene-level expression values are normalized to the reference transcript length, partial exonic copy number variants that alter the transcript length are expected to modulate RPKM values even if the absolute number of transcripts remains stable. Samples were quantile normalized within each tissue followed by inverse quantile normalization of each gene to control outliers.

We selected common genetic markers with MAF ≥ 0.05 for eQTL mapping. We performed two independent *cis*-eQTL mapping runs. The first, an “SV-only” eQTL analysis, used only common SV markers as genotype input for improved sensitivity under a reduced multiple-testing burden. The second, a “joint” eQTL analysis, included the 8,980 common SVs as well as 6,394,161 SNVs and 801,431 indels detected by the Genome Analysis Toolkit HaplotypeCaller v3.1–144-g00f68a3[20], allowing a fair comparison of the relative contribution of different variant types.

We mapped *cis*-eQTLs with FastQTL v2.184 using a *cis* window of 1 Mb on either side of the TSS of autosomal and X chromosome genes with a permutation analysis to identify the most significant marker for each gene[51]. We customized the FastQTL software to include an SV for genotype-expression associations when the span of a deletion, duplication, mCNV, or rMEI fell within the *cis* window for a particular gene TSS, or when the breakpoints of an inversion or uncharacterized breakend (BND) fell within the *cis* window. For each tissue, we applied a set of covariates including sex, three genotyping principal components, genotyping platform (HiSeq 2000 or HiSeq X Ten), and a variable number of PEER (probabilistic estimation of expression residuals) factors determined by number of samples per tissue, N ($N < 150$: 15 PEERs, $150 \leq N < 250$: 30 PEERs, $N \geq 250$: 35 PEERs)[52]. Note that PEER factor sample sizes include RNA-seq data from individuals lacking WGS, providing more samples for PEER correction than the 147 individuals in the remainder of this study. We performed gene level multiple-testing correction for each of the SV-only and joint analyses using Benjamini-Hochberg at a 10% FDR.

Fine mapping of causal variants at eQTLs

We used CAVIAR to untangle linkage disequilibrium to predict a causal variant for each eQTL[21]. CAVIAR assesses summary statistics in conjunction with LD across an associated locus to rank the causal probability of each variant in a region. Thus, we ran FastQTL once again on the 24,884eQTLs that had previously met FDR thresholds in either the SV-only or joint *cis*-eQTL mapping analyses to generate the nominal t-statistic for every common variant in the *cis* window. For each eQTL, we selected the most significant SV as well as the 100 most significant SNVs or indels (based on nominal p-value) in the *cis* window and estimated their pairwise LD using linear regression. For SVs, we used allele balance rather than discrete genotype for computing LD. We ran CAVIAR at each of these eQTLs using the t-statistics and signed *r* values of LD among the 101 variants with a causal set size of 1.

As an alternate estimate of the causal role of structural variation, at each eQTL discovered by either the SV-only or joint analyses, we applied a linear mixed model (LMM) to partition the heritability of each eGene's expression into a fixed effect from the SV, and a random effect representing the cumulative heritability of the 1,000 most significant SNVs and indels in the *cis* region. This method mirrors that of several other studies that have examined relative contributions of distinct variant classes on a quantitative trait in n individuals[22–24]. We first corrected for the same covariates as in the *cis*-eQTL mapping analyses above by linear regression residualization, and then applied a linear model of the form

$$\vec{y}_g = \beta_{j,g}\vec{x}_j + \vec{u}_g + \vec{\varepsilon}_{j,g}$$

where $\vec{y}_g$ is a vector of the normalized expression values, $\beta_{j,g}$ is the effect of allele dosage of SV *j* on gene *g*, x_j is a n-length vector of genotypes at SV *j*, $\vec{u}_g$ is a n-length vector of random effects drawn from the genetic relatedness matrix (GRM) with $\vec{u}_g \sim MVN\left(0, \sigma^2_{u_g} K_g\right)$, and $\vec{\varepsilon}$ is a random error term drawn from $N(0, I\sigma^2_{\varepsilon,j,g})$ representing unexplained variance. We defined the n × n dimensional GRM (K_g) with entries $k_{i,j} = \frac{\sum_q (z_{iq} - \overline{z}_q)(z_{jq} - \overline{z}_q)}{\sqrt{\sum_q (z_{iq} - \overline{z}_q)^2 \sum_q (z_{jq} - \overline{z}_q)^2}}$ for the 1,000 SNV and indel variants (z_q) that are most significantly associated with the expression of eGene *g*.

Solving this equation with GCTA produces an estimate of variance where $var(\vec{x}_j) \approx h^2_{SV}$.

$$var\left(\vec{y}_g\right) = \beta^2_{j,g} var(\vec{x}_j) + \sigma^2_{u_g} K_g + \sigma^2_{\varepsilon_{j,g}} I_n$$

$$\sigma^2_{\mathrm{P}} = h^2_{SV} + h^2_b + \sigma^2_\varepsilon$$

$$h^2_{cis}=h^2_{SV}+h^2_b$$

Heritability estimates for a small number of eQTLs could not be calculated due to non-positive definite matrices likely arising from small sample sizes (788/23,554 of joint eQTLs and 892/24,884 of eQTLs detected by either SV-only or joint mapping). These loci were excluded from the heritability analysis and composite causality scores described below. To estimate the fraction of *cis* heritability attributable to SVs across all eQTLs in our data set, we counted the number of eQTLs where $h^2_{SV} > h^2_b$ as a fraction of joint eQTLs at which the overall heritability (h^2_{cis}) was at least 0.05.

We combined the CAVIAR and heritability estimates of causality into a single composite score for each eQTL by taking the product of the CAVIAR causal probability and the fraction of heritability attributed to the SV ($h^2_{SV} \,/\, h^2_{SV}+h^2_b$). To bound the heritability fraction between 0 and 1, we set the minimum h^2_{SV} and h^2_b to 10^{-6} and the maximum to 1 before taking the quotient. For SVs that were associated with multiple eQTLs, including those that were independently ascertained in multiple tissues, we selected the eQTL (tissue, gene pair) in which the SV had the highest causality score, resulting in a set of 4,485 distinct SVs.

Though we calculated causality scores (CAVIAR, heritability, composite) for all eQTLs detected by the SV-only or joint analyses, we restricted estimates comparing the relative contributions and effect sizes of SVs, SNVs, and indels to only those 23,554 detected by the joint analysis to eliminate confounding differences in statistical power between the eQTL mapping runs.

Feature enrichment

We performed intersections between SVs across the range of composite causality score quantiles and various annotated genomic features (Fig. 3, Supplementary Fig. 17). We first allocated SVs into 6 bins by the quantile (bottom 50% and 5 deciles of the top 50%) of their composite causality score and then counted the number that intersected with each feature, allowing 1 kb of flanking distance except for the following: exon-altering plot, no flanking distance; proximity to TSS, 10 kb of directional flanking distance; GENCODE genes, no flanking distance; GENCODE exons, no flanking distance; and topologically associated domain boundaries, 5 kb of flanking distance. SVs involved in multiple eQTLs were considered to touch an eGene if they overlapped the exons of genes at any of those eQTLs. SVs from each causality bin were shuffled with BEDTools into non-gapped regions of the genome within 1 Mb of a gene transcription start site[53]. We calculated the fold enrichment of observed feature intersections compared to the median of 100 random shuffled sets of the elements of each bin to control for each bin's composition of SV types and size distributions. The 95% confidence intervals were derived from the empirical distributions of feature intersections from the shuffled set for each bin.

Enhancer positions were defined as those in the Dragon ENhancers database (DENdb) with a minimum support of 2[28] (Fig. 3b). Positions 10 kb upstream and downstream of the TSS were defined from GENCODE v19 gene positions (Fig. 3c,d). FunSeq 2.1.0 regions and topologically associated domain boundaries from human embryonic stem cell lines were downloaded from the authors' websites (see URLs)[27,54]. All other regions were defined by the ENCODE project and downloaded from the UCSC genome browser[26,55].

eQTL linkage to GWAS hits

We defined a set of phenotype-associated SNVs from the GWAS catalog v1.0.1 (downloaded 2016–02–04)[29]. We selected for results with a p-value better than 5×10^{-8} and tested in Europeans. When multiple markers within a 100 kb window met this criteria in a single study and a single phenotype, we selected the most significant marker in the window to reduce redundancy, resulting in a set of 4,951 SNVs, of which 4,874 were genotyped in our cohort of 147 samples. We calculated LD between these GWAS hits and variants in our cohort using a linear regression to approximate r^2, using allele balance rather than discrete genotypes for SVs detected with LUMPY.

Rare variant association with expression outliers

We began by defining gene expression outliers in each of 544 individuals with RNA-seq data across the 44 tissues available from the GTEx project. Since quantile normalization of expression values (as applied in *cis*-eQTL mapping) can reduce the signal from true expression outliers, we derived PEER-corrected expression values without quantile normalization to define expression outliers. For each tissue, we filtered for genes on the autosomes or the X chromosome in which at least 10 individuals had an RPKM (reads per kilobase of transcript per million mapped reads) > 0.1 and raw read counts > 6. Next we took the $\log_2$(RPKM + 2) transformation of the data, followed by Z-transformation across each gene. We then removed PEER factors by linear regression residualization (using the same number of factors per tissue as described above, see *Common eQTL mapping*) followed by Z-transformation.

We then subsampled the 544 individuals above to select the 117 who were of European ancestry (since this was the largest subpopulation in our cohort) and had available WGS sequence data. Among these 117 individuals, we identified (sample, gene) pairs at which an individual's absolute median Z-score of a gene's expression was at least 2, and there were at least 5 tissues with available expression data for the gene. This amounted to 5,047 gene expression outliers (median: 30 per person, range: 10–298). Next, we identified rare variants that were present in at most two individuals in our cohort of 147 individuals and positively genotyped in at least one of the 117 European ancestry individuals, amounting to 5,660,256 rare variants (4,671 SVs, 4,830,727 SNVs, and 824,836 indels).

We counted the number of rare SVs, SNVs, and indels that co-occurred in the outlier individual that resided within the outlier transcript or 5 kb of flanking sequence. To define the frequency that this occurs by chance, we performed 1,000 random permutations of the outlier individual names in our set to determine the number of rare variants of each type that co-occur with an outlier in a random individual. Notably, this strategy retained the relative

number of outliers per individual in each permutation so that individuals with many outliers were still over-represented in the permuted sets.

We performed two reciprocal measures of enrichments. The “outlier-centric” approach tested for a significant difference in the number of outliers that had a rare variant within 5 kb (Fig. 5a, red points). However, for SVs in particular, a single large variant may be in proximity to many outliers, and we controlled for this phenomenon with the “variant-centric” approach to test for a significant difference in the number of rare variants that had an outlier gene within 5 kb (Fig. 5a, blue points).

To judge the enrichment across thresholds of variant functional impact, we computed a predicted impact score with CADD v1.2 for all variants in our data set[44]. For SVs, we used the highest-scoring base across the affected interval and the confidence intervals around the SV breakpoints. We then computed to the percentile of these impact scores for SVs, SNVs, and indels separately across the full set of allele frequencies. We show the fold-enrichment for the outlier-centric (Supplementary Fig. 26a–c) and variant-centric (Supplementary Fig. 26d–f) approaches across the range of impact score percentiles for each variant class.

To estimate the relative contribution of each variant type to expression outliers, we first defined the fraction of outliers with a likely genetic basis. Across 1,000 shuffled permutations of the data, we observed a median of 1,976 outliers (95% CI: 1,917–2,036) with a rare variant of any type in the outlier individual. We identified 2,417 outliers with a rare variant, representing a net excess of 441 over the expected value (95% CI: 381–500). Thus, of the 5,047 total outliers in our data set, an estimated 8.7% (95% CI: 7.6–9.9%) of outliers have a genetic basis.

We then apportioned these 8.7% of genetically determined outliers amongst SVs, SNVs, and indels according to the net excess of observed outliers within 5 kb of each variant types (Supplementary Fig. 27). For outliers that were within 5 kb of multiple variant types (overlaps on the Venn diagram), we allocated the net excess percentage based on the relative strength of their overall fold-enrichment. To achieve this, we first estimated the fraction of expression outliers with a genetic basis within each variant type ($T=\{SV,SNV,indel\}$) as

$G_T = \frac{s - \hat{s}}{s}$, where s is the number of observed outliers with 5 kb of a rare variant of type T and $\hat{s}$ is the median from the permuted sets (G_{SV}=0.94; G_{SNV}=0.12; G_{indel}=0.19 in our data set). Then, for each overlapping region of the Venn diagram, we multiplied the net excess by $\frac{G_T}{\sum_{T \in A} G_T}$ for each of the variant types in each Venn diagram area A.

To identify complex variants, we clustered rare SVs with breakpoint evidence located no more than 100 kb away from each other and present in the same individual(s). Rare SVs with only read-depth support were not included in this clustering because of their imprecise boundaries. We joined separate clusters if they contained two sides of the same uncharacterized BND. Clusters containing SVs that were previously found to be associated with outlier gene expression were reclassified as a complex deletion, complex duplication, or balanced complex rearrangement by manual curation. During this manual curation, rare

SVs with only read-depth support and present in the appropriate sample(s) were added to the rare variant cluster if they overlapped other variants included in the cluster. Upon manual inspection, one outlier-associated SV (LUMPY_BND_195398) with inverted breakpoint orientation was visually determined to have amplified read-depth over the interval and thus reclassified as a complex duplication.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank R.E. Handsaker for advice on Genome STRiP, H.J. Abel for helpful statistical discussions, and R.M. Layer for software contributions. This work is supported by the NIH (MH101810) (D.F.C.), the NIH/NHGRI (1UM1HG008853) (I.M.H.), a Burroughs Wellcome Fund Career Award (I.M.H.), the Mr. and Mrs. Spencer T. Olin Fellowship for Women in Graduate Study (A.J.S.), the Lucille P. Markey Biomedical Research Stanford Graduate Fellowship (J.R.D.), the Stanford Genome Training Program (SGTP; NIH/NHGRI T32HG000044) (J.R.D.), a Hewlett-Packard Stanford Graduate Fellowship (E.K.T.), and a doctoral scholarship from the Natural Science and Engineering Council of Canada (E.K.T.). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI/SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplements to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St. Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.v6.p1 on 05/11/2015.

References

1. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: Illuminating the Dark Road from Association to Function. The American Journal of Human Genetics. 2013; 93:779–797. [PubMed: 24210251]
2. Lappalainen T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–511. [PubMed: 24037378]
3. The GTEx Consortium. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015; 348:648–660. [PubMed: 25954001]
4. Battle A, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research. 2014; 24:14–24. [PubMed: 24092820]
5. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
6. Conrad DF, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010; 464:704–712. [PubMed: 19812545]
7. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011; 12:363–375. [PubMed: 21358748]
8. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013; 14:125–138. [PubMed: 23329113]
9. Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007; 315:848–853. [PubMed: 17289997]

10. Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO. Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. Genome Research. 2011; 21:2004–2013. [PubMed: 21862627]
11. Bryois J, et al. Cis and Trans Effects of Human Genomic Variants on Gene Expression. PLoS Genet. 2014; 10:e1004461. [PubMed: 25010687]
12. Gamazon ER, Nicolae DL, Cox NJ. A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. PLoS Genet. 2011; 7:e1001292. [PubMed: 21304891]
13. Sudmant PH, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015; 526:75–81. [PubMed: 26432246]
14. Chiang C, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Meth. 2015; 12:966–968.
15. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXivorg 13033997. 2013
16. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for structural variant discovery. Genome Biol. 2014; 15:R84. [PubMed: 24970577]
17. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nature Genetics. 2011; 43:269–276. [PubMed: 21317889]
18. Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27:2156–2158. [PubMed: 21653522]
19. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics. 2016; 32:1479–1485. [PubMed: 26708335]
20. McKenna A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; 20:1297–1303. [PubMed: 20644199]
21. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014; 198:497–508. [PubMed: 25104515]
22. Gymrek M, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nature Genetics. 2016; 48:22–29. [PubMed: 26642241]
23. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88:76–82. [PubMed: 21167468]
24. Gusev A, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014; 95:535–552. [PubMed: 25439723]
25. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–291. [PubMed: 27535533]
26. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
27. Fu Y, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. 2014; 15:153.
28. Ashoor H, Kleftogiannis D, Radovanovic A, Bajic VB. DENdb: database of integrated human enhancers. Database. 2015; 2015
29. Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014; 42:D1001–6. [PubMed: 24316577]
30. Gretarsdottir S, et al. Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. Nature Genetics. 2010; 42:692–697. [PubMed: 20622881]
31. Okada Y, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2013; 506:376–381. [PubMed: 24390342]
32. Suzuki A, et al. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nature Genetics. 2003; 34:395–402. [PubMed: 12833157]
33. Yang XK, et al. Associations Between PADI4 Gene Polymorphisms and Rheumatoid Arthritis: An Updated Meta-analysis. Archives of Medical Research. 2015; 46:317–325. [PubMed: 26043831]

34. Wu C, et al. Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. Nature Genetics. 2014; 46:1001–1006. [PubMed: 25129146]
35. Barrett JH, et al. Genome-wide association study identifies three new melanoma susceptibility loci. Nature Genetics. 2011; 43:1108–1113. [PubMed: 21983787]
36. Stacey SN, et al. Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer. Human Molecular Genetics. 2016; 25:1008–1018. [PubMed: 26740556]
37. de Cid R, et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nature Genetics. 2009; 41:211–215. [PubMed: 19169253]
38. Chambers JC, et al. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nature Genetics. 2011; 43:1131–1138. [PubMed: 22001757]
39. Wellcome Trust Case Control Consortium. et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010; 464:713–720. [PubMed: 20360734]
40. Li X, et al. The impact of rare variation on gene expression across tissues. 2016; doi: 10.1101/074443
41. Li X, Montgomery SB. Detection and Impact of Rare Regulatory Variants in Human Disease. Front Genet. 2013; 4
42. Li X, et al. Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. Am J Hum Genet. 2014; 95:245–256. [PubMed: 25192044]
43. Quinlan AR, Hall IM. Characterizing complex structural variation in germline and somatic genomes. Trends in Genetics. 2012; 28:43–53. [PubMed: 22094265]
44. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. 2014; 46:310–315. [PubMed: 24487276]
45. Ganel L, Abel HJ, FinMetSeq Consortium. Hall IM. SVScore: an impact prediction tool for structural variation. Bioinformatics. 2016; doi: 10.1093/bioinformatics/btw789
46. Cooper NJ, et al. Detection and correction of artefacts in estimation of rare copy number variants and analysis of rare deletions in type 1 diabetes. Human Molecular Genetics. 2015; 24:1774–1790. [PubMed: 25424174]
47. Wang K, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research. 2007; 17:1665–1674. [PubMed: 17921354]
48. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
49. Harrow J, et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006; 7:S4–9. [PubMed: 16925838]
50. Deluca DS, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012; 28:1530–1532. [PubMed: 22539670]
51. Ongen H, Buil A, Brown A, Dermitzakis E, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. bioRxiv. 2015; doi: 10.1101/022301
52. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012; 7:500–507. [PubMed: 22343431]
53. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]
54. Ho J, et al. Comparative analysis of metazoan chromatin organization. Nature. 2014
55. Kent WJ, et al. The Human Genome Browser at UCSC. Genome Research. 2002; 12:996–1006. [PubMed: 12045153]

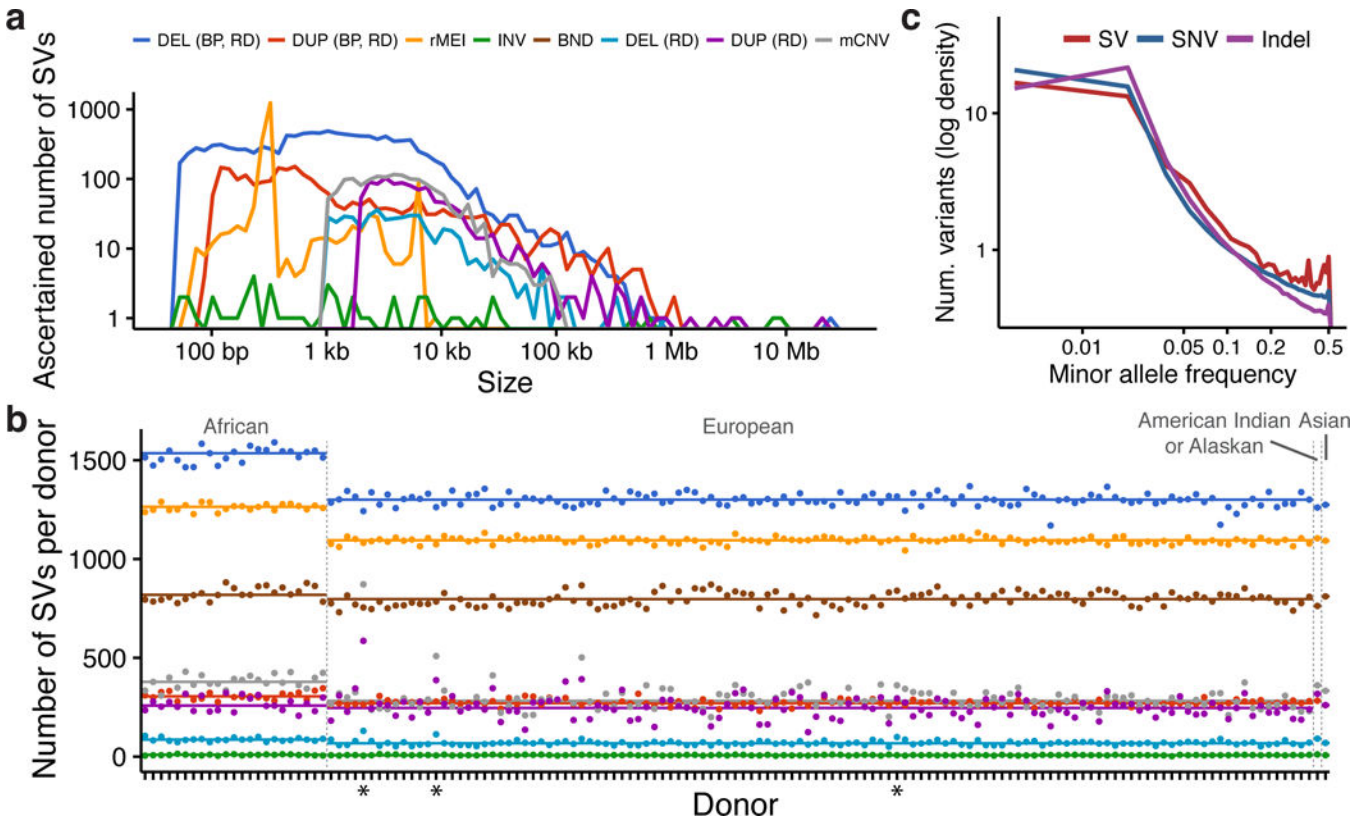

Figure 1.

Structural variation call set. (**a**) Size distribution of ascertained SVs by variant type and (**b**) number of SVs detected in each sample. Starred (*) samples exhibited abnormal read-depth profiles, and were excluded from rare variant analyses. (**c**) The site frequency spectrum of SVs compared to SNVs and indels detected by GATK.

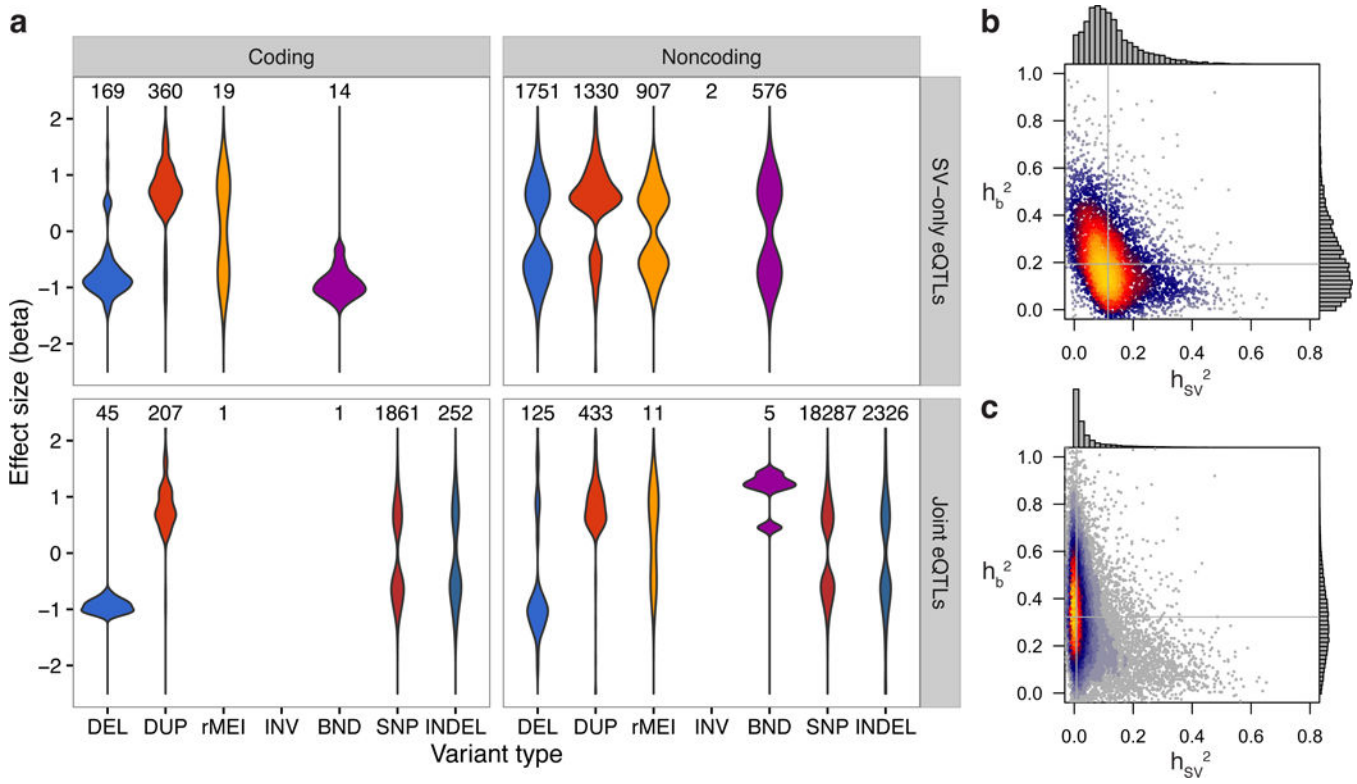

Figure 2.

eQTL effect size distributions and heritability partitioning with linear mixed models. (**a**) Effect size distributions for coding and noncoding variants of each type, with the number of eQTLs of each type above each distribution. The top panels (SV-only eQTLs) show the 5,128 eQTLs that were discovered by the SV-only analysis, while the bottom two panels show the 23,554 eQTLs discovered by the joint analysis. The “DUP” category includes duplications and mCNVs, and the alternate allele for rMEIs is the insertion. (**b,c**) Heat scatter plots 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 (y-axis) for (**b**) SV-only and (**c**) joint eQTL mapping analyses. Gray lines denote the median of values for each axis.

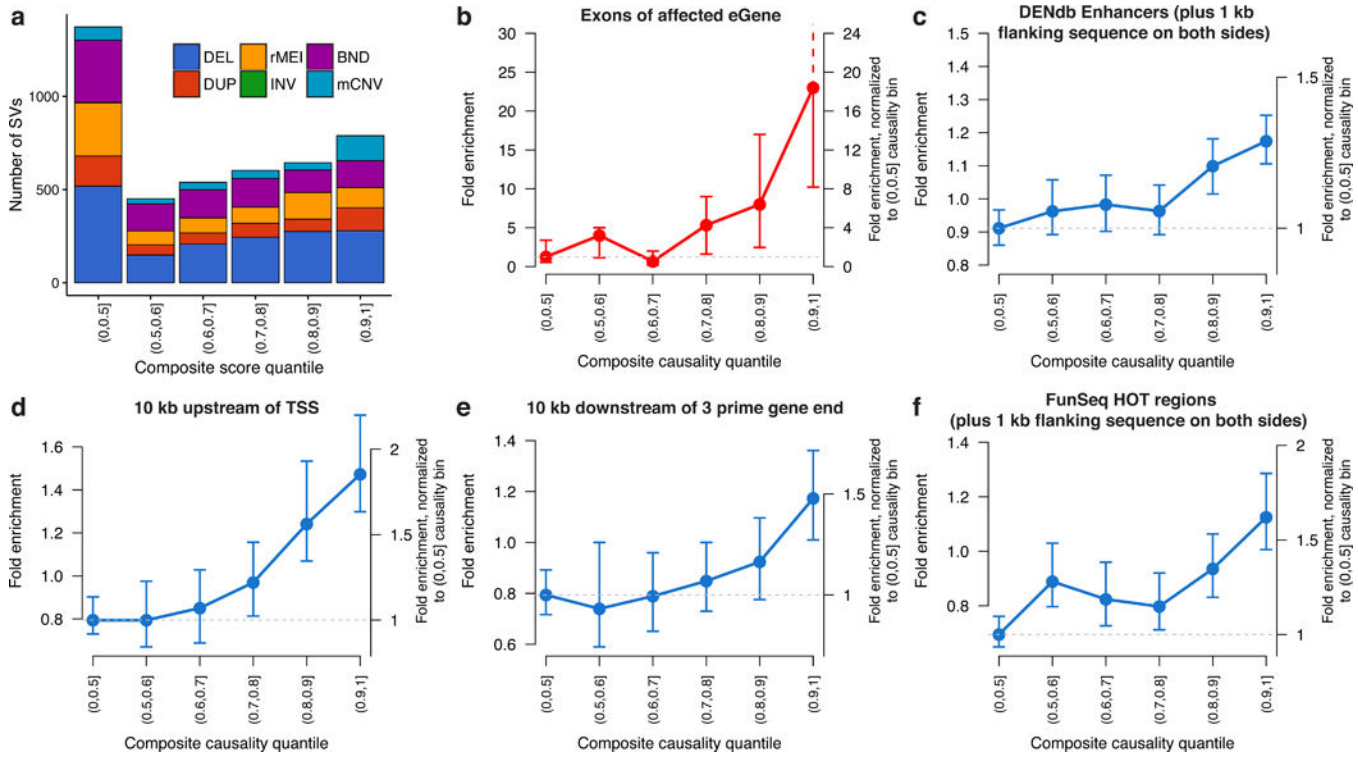

Figure 3.

Feature enrichment of SV-eQTLs. Fold enrichment and 95% confidence intervals (based on 100 random shuffled sets of the positions of SVs in each bin) for the overlap between the most significant SV and various annotated genomic features at the union of eQTLs discovered by SV-only or joint eQTL mapping. (**a**) Composition of each causality score bin by SV type. (**b**) Enrichment for an SV in each bin of causality to touch exons of the affected eGene. For the remaining plots in blue (**c**-**f**), SVs that overlapped with an exon of their affected eGene were excluded, yet the remaining SVs still showed significant enrichment in (**c**) enhancers from the Dragon Enhancers Database (DENdb), (**d**) in the 10 kb regions upstream and (**e**) downstream of transcriptions start sites (TSS), and (**f**) regions predicted to be highly occupied by transcription factors (FunSeq HOT regions).

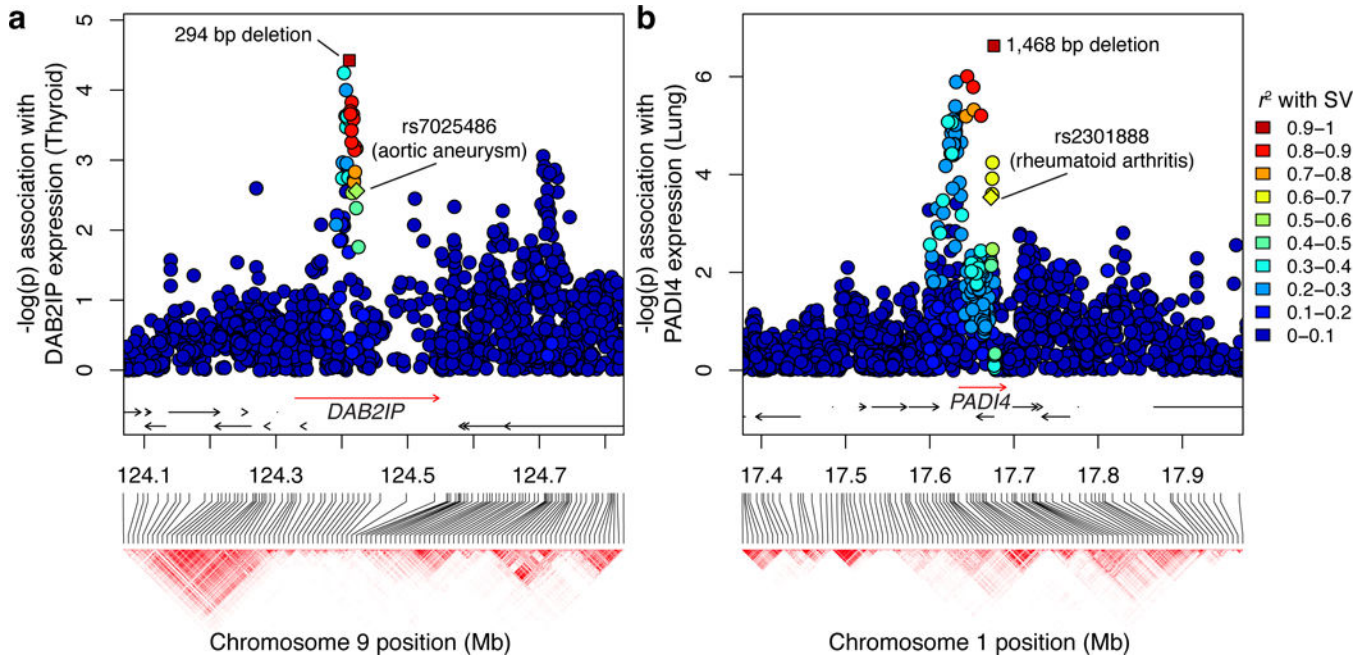

Figure 4.

Candidate SV-eQTLs at GWAS loci. Genomic position and haplotype blocks are shown on the x-axis, and each variant's association with the indicated eGene is shown on the y-axis. The rectangular points represent the predicted causal SV, with the colors representing its linkage (r^2) to each marker in the window. The labeled diamonds show the reported risk allele for the specified GWAS phenotype. (**a**) A 294 bp deletion that intersects an enhancer in intron 1 of DAB2IP was linked to a risk allele for abdominal aortic aneurysm (rs7025486), and is also predicted to be a causal eQTL for DAB2IP. (**b**) A 1,468 bp deletion associated with increased expression of PADI4 is linked to a known risk allele for rheumatoid arthritis (rs2301888).

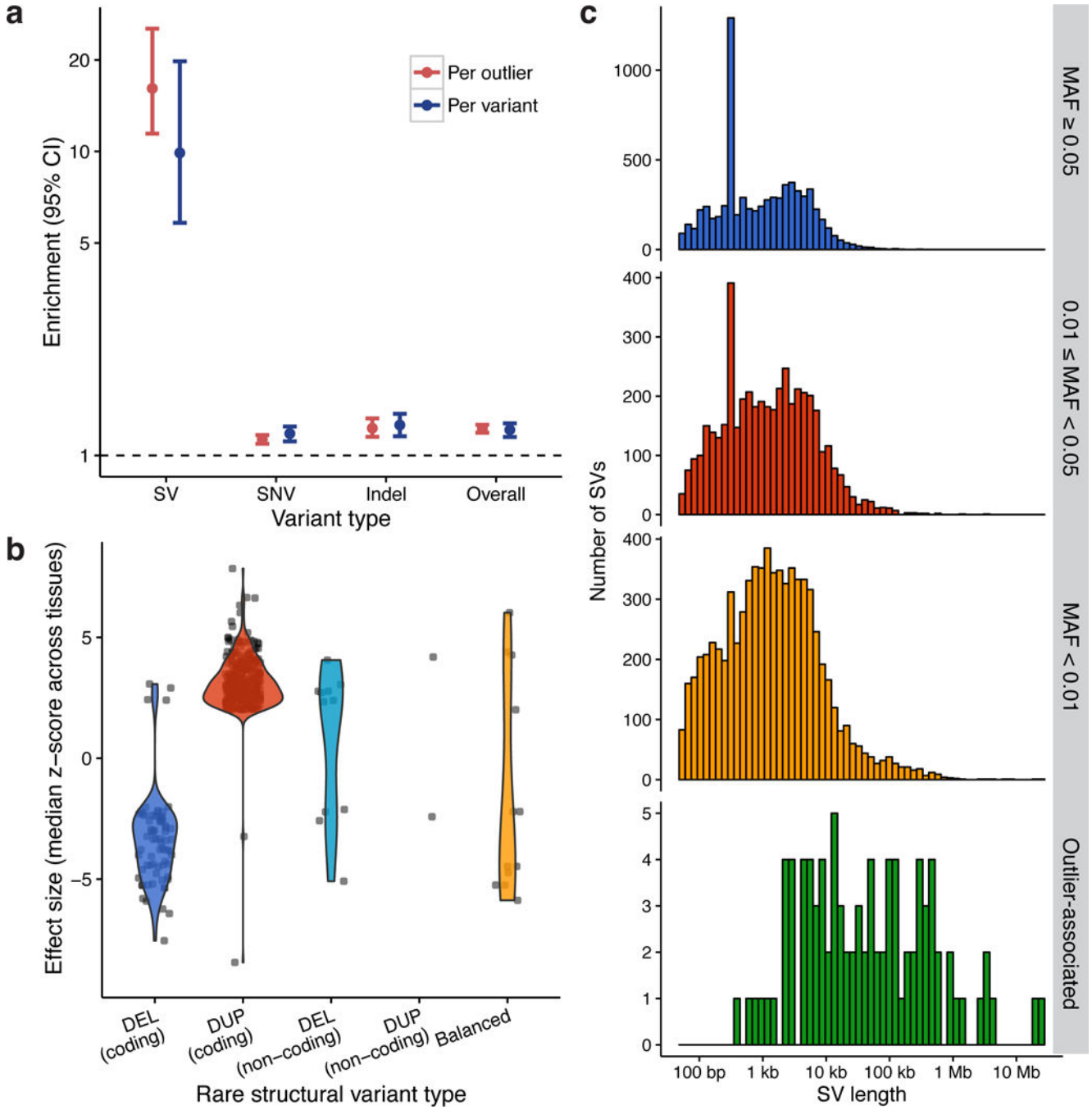

Figure 5.

Gene expression outliers are associated with rare SVs. (**a**) Fold enrichment of rare variants within 5 kb of expression outliers (red) and fold enrichment of outliers within 5 kb of rare variants (blue) between the observed set of 5,047 outliers and 1,000 random permutations of their sample names (y-axis is log-scaled). (**b**) Effect size distributions for each SV type within 5 kb of an outlier in the same individual, with “coding” SVs defined as those that overlap with exons of the outlier gene and “noncoding” defined by the remainder. (**c**) Size distribution histograms by minor allele frequency (MAF) classes and rare SVs within 5 kb of

an expression outlier in the same individual, excluding balanced rearrangements. A peak at ~300 bp in the top two plots results from Alu SINE insertions in the reference genome.

Table 1

Summary of variant types and discovery methods. SNVs and indels were detected using the Genome Analysis Toolkit (GATK) and SVs were detected by breakpoint evidence (BP) and supported by read-depth evidence (BP, RD), or only detected by read-depth evidence (RD). Common variants (MAF ≥ 0.05) were tested for *cis* eQTLs. The SV-only eQTL mapping excluded SNVs and indels for greater sensitivity, while the joint eQTL mapping included all variant types.

	Detection method	# of variants	Median resolution (bp)	Median size (bp)	# of common variants	eVariants (SV-only)	eVariants (joint)
SNV	GATK	21,764,904	–	1	6,394,161	–	16,959
Indel	GATK	3,030,964	–	3	801,431	–	2,130
Deletion (DEL)	BP, RD	11,492	35	993	2,939	510	25
	RD	473	kilobase*	3,819	284	68	17
Duplication (DUP)	BP, RD	2,506	96	576	676	97	3
	RD	1,035	kilobase*	4,999	684	148	76
Multi-allelic CNV (mCNV)	RD	1,534	kilobase*	3,847	1060	264	118
Inversion (INV)	BP	51	15	1045	14	0	0
Reference mobile element insertion (rMEI)	BP	2,051	1	307	1,535	265	10
Other SV (BND)	BP	4,460	34	–	1,788	281	4
All SVs	–	23,602	39	–	8,980	1,634	253
All variant types	–	24,819,470	–	–	7,204,572	–	19,342

*Resolution refers to the positional certainty at each breakpoint, with read-depth variants having approximate breakpoint precision on the kilobase scale.