Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes

Alexandra J. Scott[1,2], Colby Chiang[1,2], Ira M. Hall[1,2,3,*]

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

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

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

* Corresponding author

Corresponding email: ira.hall@yale.edu (I.M.H.)

Keywords: Structural variation, eQTLs, gene expression outliers

ABSTRACT

Structural variants (SVs) are an important source of human genome diversity but their functional effects are not well understood. We mapped 61,668 SVs in 613 individuals with deep genome sequencing data from the GTEx project and measured their effects on gene expression. We estimate that common SVs are causal at 2.66% of eQTLs, which is a 10.5-fold enrichment relative to their abundance in the genome and consistent with prior work using smaller sample sizes. Duplications and deletions were the most impactful variant types, whereas the contribution of mobile element insertions was surprisingly small (0.12% of eQTLs, 1.9-fold enriched). Multi-tissue analysis of expression effects revealed that gene-altering SVs show significantly more constitutive effects than other variant types, with 62.09% of coding SV-eQTLs active in all tissues with known eQTL activity compared to 23.08% of coding SNV- and indel-eQTLs, while noncoding SVs, SNVs and indels show broadly similar patterns. We also identified 539 rare SVs associated with nearby gene expression outliers. Of these, 62.34% are noncoding SVs that show strong effects on gene expression yet modest enrichment at known regulatory elements, demonstrating that rare noncoding SVs are a major source of gene expression differences but remain difficult to predict from current annotations. Remarkably, both common and rare noncoding SVs often show strong regional effects on the expression of multiple genes: SV-eQTLs affect an average of 1.82 nearby genes compared to 1.09 genes affected by SNV- and indel-eQTLs, and 21.34% of rare expression-altering SVs show strong effects on 2-9 different genes. We also observe significant effects on rare gene expression changes extending 1 Mb from the SV. This provides a mechanism by which individual noncoding SVs may have strong and/or pleiotropic effects on phenotypic variation and disease.

INTRODUCTION

Structural variants (SVs) are a diverse class of genetic variation that include copy number variants (CNVs), mobile element insertions (MEIs) and balanced rearrangements at least 50 base pairs (bp) in length. While SVs are relatively rare compared to single-nucleotide variants (SNVs) and small insertion or deletion (indel) variants, their size and diversity mean that SVs can disrupt protein-coding genes and genomic regulatory elements through diverse mechanisms. Furthermore, SVs often have more severe consequences compared to smaller variants and previous studies have found that SVs have an outsized impact on human gene expression compared to their relative abundance in the genome (Chiang et al. 2017; Stranger et al. 2007; Sudmant et al. 2015). SVs have also been implicated in the biology of human diseases such as autism spectrum disorder (Brandler et al. 2018; Sebat et al. 2007; Turner et al. 2017; Weiss et al. 2008) and schizophrenia (International Schizophrenia Consortium 2008; Marshall et al. 2017; McCarthy et al. 2009; Walsh et al. 2008). However, SVs are difficult to detect from short-read DNA sequencing data and are often excluded from complex trait association studies.

Advances in high-throughput sequencing technologies that have allowed for widespread use of whole genome sequencing (WGS), combined with advances in scaling SV detection algorithms, mean that comprehensive studies of all forms of genetic variation are now possible for large human cohorts. Recent studies of SV in large, deeply-sequenced human cohorts have found that SVs account for 4.0-11.2% of rare high-impact coding alleles (Abel et al. 2020) and are responsible for 25-29% of rare protein-truncating events per genome (Collins et al. 2020). However, few studies to date have examined the functional effects of SV on gene expression and these studies are limited to relatively small cohort sizes or only a few tissue types with available gene expression data (Chiang et al. 2017; Han et al. 2020; Jakubosky et al. 2020; Sudmant et al. 2015).

Here, we use deep WGS data and multi-tissue RNA-seq expression data from 613 individuals in the Genotype-Tissue Expression (GTEx) project to comprehensively map SVs and

to evaluate their impact on both common and rare gene expression changes in up to 48 tissue types (**Supplemental Table S1**). This study expands on our prior analysis of SV in 147 human samples from the GTEx cohort with RNA-seq expression data from 13 different tissues (Chiang et al. 2017) and is the most comprehensive study of SV-eQTLs to date. The expanded cohort size provides greater power to evaluate the impact and mechanisms of SV-associated gene expression changes, particularly for rare SVs.

RESULTS

Variant calling

We mapped SVs in 613 individuals from the GTEx v7 release using LUMPY (Chiang et al. 2015; Layer et al. 2014), svtools (Larson et al. 2019), GenomeSTRiP (Handsaker et al. 2011, 2015) and the Mobile Element Locator Tool (MELT) (Gardner et al. 2017) (see **Methods**). Variant calls were filtered and merged using the same approach as in our previous GTEx study (Chiang et al. 2017; Li et al. 2017), resulting in a total of 61,668 "high confidence" SVs that are the basis for all subsequent analyses (**Table 1**). Single nucleotide (SNV) and small insertion deletion (indel) variants were mapped using GATK (McKenna et al. 2010) as part of the official v7 release from the GTEx Consortium.

	Detection method	No. variants	Median size (bp)	# of common variants	eVariants
SNV	GATK	37,087,030	1	9,609,545	178,000
Indel	GATK	3,081,270	3	818,401	16,460
Deletion (DEL)	BP	20,954	1,311	4,385	210
	RD	10,252	2,151	8,166	66
Duplication (DUP)	BP	3,388	2,632	1,090	64
	RD	1,598	6,891	896	233
Multi-allelic CNV (mCNV)	RD	4,365	3,602	3,238	460
Inversion	BP	295	1,054	96	2
Reference mobile element insertion (MEI-del)	BP	2,681	306	2,026	88
Non-reference mobile element insertion (MEI-ins)	BP	13,066	280	4,496	91
Other (BND)	BP	5,069	-	2,010	57
All SVs	-	61,668	-	26,409	1,271
All Variants	-	40,229,968	-	10,454,355	195,731

Table 1. Summary of variant types and eQTL mapping. SVs were detected based on breakpoint evidence (BP) or read-depth evidence (RD). SNVs and indels were called using the Genome Analysis Toolkit (GATK). Common variants (MAF ≥ 0.01) were used to map cis-eQTLs.

Effects of common SVs

We performed *cis*-eQTL mapping of common variants (MAF ≥ 0.01) using a permutation-based mapping approach with FastQTL (Ongen et al. 2016), limiting comparisons to variants within 1 Mb of the transcription start site (TSS) of each gene. We performed eQTL analyses in each of the 48 tissues for which expression data was available for at least 70 individuals (**Supplemental Table S1**) and defined an eQTL as an eVariant/eGene pair detected in a given tissue. We performed a "joint" eQTL mapping analysis in which SVs, SNVs and indels were simultaneously queried for eQTL status, allowing for direct comparisons between their properties and identification of a likely causal variant. An SV was the lead marker in 2.66% (7,960/299,187) of eQTLs (**Supplemental Table S2**), although this is likely an underestimate of SV causality due to inferior genotyping accuracy for SVs, which biases eQTL fine-mapping analyses against SVs. While this estimate of the contribution of SVs is relatively small, it represents an 10.5-fold enrichment over the abundance of SVs in the genome. This result is consistent with our prior analysis of the initial 147 individuals from the GTEx cohort (Chiang et al. 2017). In the same 13 tissues evaluated in this previous study, the increased sample size used here allowed us to identify 617 genes with SV-eQTLs that were not identified in the smaller study, though 57 genes from the initial study are no longer SV-eQTLs. Interestingly, 71.82% (5,717/7,960) of all SV-eQTLs identified in this study are noncoding (**Supplemental Fig. S1**), meaning the SV does not intersect with any exons of its associated eGene. This figure is even more striking when eQTLs are collapsed across tissues, where 1,907/2,318 (82.27%) of unique eGene/eSV pairs are noncoding. This also suggests that coding SV-eQTLs are more constitutive as more of them are identified in multiple tissues.

A novel aspect of this study is that we used MELT to sensitively map mobile element insertion (MEI) variants, including non-reference insertions that were not detected in our prior GTEx studies. It has been proposed that MEIs may have broad effects on gene expression due

to their ability to disrupt genes, promote epigenetic gene silencing, and serve as alternate promoters (Payer and Burns 2019; Chuong et al. 2017); however, there has been scant data in humans to address this. We found that only 0.12% (353/299,187) of eQTLs had an MEI as the lead marker. Although this is a 1.9-fold enrichment of predicted causal MEIs relative to their abundance (0.06% of common variants), MEIs were far less likely than other SV types to be the lead marker (e.g., mCNVs are enriched 45-fold, duplications 38-fold and deletions 3.3-fold). Thus, despite compelling molecular evidence for the functional potential of MEIs, our results suggest that they are only slightly enriched as causal eQTL variants relative to SNVs and indels and are depleted relative to other SVs, on average.

We found that not only do SVs have larger effect sizes compared to SNPs and indels, as noted in previous studies (**Supplemental Fig. S1**) (Jakubosky et al. 2020; Chiang et al. 2017), they are also more likely to alter the expression of multiple nearby genes. Each eSV affects an average of 1.82 unique eGenes while SNVs and indels affect an average of 1.09 unique eGenes. Although this effect is partially explained by large SVs that alter the copy number of multiple adjacent genes, there is also a significant difference for genes affected by noncoding eVariants: on average, eSVs affect 1.50 unique eGenes for which they do not intersect any exons of the eGene, compared to an average of 1.04 unique eGenes for SNVs and indels ($p=1.02x10^{-55}$, one-sided Mann-Whitney U test) (**Fig. 1B-D**). These noncoding effects are most pronounced for duplications ($p=6.10x10^{-53}$) and mCNVs ($p=4.75x10^{-56}$), which are the only two categories of noncoding SVs that affect significantly more eGenes than point variants. This result indicates that causal SVs are generally more impactful than causal point variants, both in terms of their per-gene effect sizes as well as their potential to affect multiple genes. These results also suggest that SVs are more likely to disrupt key regulatory elements and/or alter higher-order genome architecture, allowing individual variants to affect multiple genes.

To investigate the functional mechanisms of expression-altering SVs, we defined a set of putative causal SVs using a score generated by taking the product of the causal probability

calculated using CAVIAR (Hormozdiari et al. 2014) and the fraction of heritability attributed to the SV calculated using GCTA (Yang et al. 2011) (**Supplemental Table S3**), as described previously (Chiang et al. 2017). At each eGene we selected the SV within the *cis*-region that had the strongest association with the eGene's expression and allocated these 10,911 unique SVs into six bins on the basis of causality score quantiles, with the least-causal bin containing the 50% of SVs with the lowest scores. Next, we measured the enrichment of SVs in each causality bin at a diverse set of genomic annotations and in the core 15 chromatin segmentation states from the Roadmap Epigenomics Project using a permutation test based on shuffled genomic positions (**Supplemental Fig. S2-S3; see Methods)**. SVs in the most causal quantiles were strongly enriched in the exons of their associated eGenes, which is expected and confirms that our causality score is informative. We also observed an enrichment of causal SVs in the 10 kb regions upstream of the TSS and downstream of the 3' UTR of the associated eGene. Additionally, there is a small enrichment of the causal SVs in segmental duplications, which is likely driven by large mCNVs at multi-copy genes. However, predicted causal SVs were not enriched in any other genomic features tested, which suggests that while eSVs are generally found relatively close to their eGenes, they may be altering expression through diverse mechanisms and our study is underpowered to identify enrichments in specific regulatory element classes. Alternatively, existing annotations may be insufficiently informative to detect functional enrichments for the variants and tissues analyzed here.

The number and diversity of tissues with available expression data allows us to evaluate the tissue specificity of eQTLs. We hypothesized that SVs might have more ubiquitous effects on gene expression than point variants due to constitutively-acting dosage changes or due to complete deletion or duplication of regulatory elements rather than more subtle effects, for example, on transcription factor binding. To allow for facile comparisons between variant types, we limited this analysis to variant-gene pairs with a significant association in our eQTL analysis for which expression data was available across all 48 tissues. We used METASOFT (Han and

Eskin 2011) to evaluate eQTL activity across all tissues and limited this analysis to eQTLs for which active (m>0.9) or inactive (m<0.1) status could be determined in at least 43 tissues. We found that coding SV-eQTLs are more constitutive than other eQTL classes, showing activity across a larger proportion of tissues compared to SNV- and indel-eQTLs (**Fig. 1E**). Whereas 92.16% of coding SV-eQTLs are constitutively active – defined here as active in >75% of tissues with known status – only 74.12% of coding SNV- and indel-QTLs are constitutive. However, the result at noncoding eQTLs is less clear: 74.86% of noncoding SV-eQTLs are constitutively active as defined above and 74.12% of noncoding SNV- and indel-eQTLs are constitutive, which suggests that there are not significant differences between these variant categories. However, when we examine noncoding eQTLs that are active in 100% of tissues with known activity, 44.44% of noncoding SV-eQTLs are active in all known tissues compared to 26.23% of noncoding SNV- and indel-eQTLs (**Supplemental Fig. S4**). Overall this analysis shows that coding SVs typically impact expression across many tissues, whereas smaller and noncoding variants tend to affect gene expression on a more tissue-specific basis. In contrast to coding SV-eQTLs, noncoding SV-eQTLs show similar patterns of tissue specificity to noncoding SNV- and indel-eQTLs, indicating that these variant types are likely to function through similar mechanisms. However, it is important to note that noncoding SV-eQTL activity could not be determined by METASOFT in many tissues (**Supplemental Fig. S5**), so it is possible that the true tissue specificity of noncoding SVs may differ from noncoding SNVs and indels. This appears to be the result of relatively large effect-size standard errors for SV-eQTLs that result from genotyping inaccuracies. While METASOFT can determine cross-tissue eQTL activity when effect sizes are large despite large standard errors, as seen in coding SV-eQTLs, when effect sizes are small but effect size errors are large, the algorithm often cannot confidently judge activity (**Supplemental Fig. S6**).

Effects of rare SVs

Rare SVs are enriched near genes with highly aberrant expression (Chiang et al. 2017) and are more likely to have large effect sizes compared to other variant types (Li et al. 2017). To assess the effects of rare SVs on gene expression, we identified genes in which individuals displayed highly aberrant gene expression levels compared to the dataset as a whole. We limited this analysis to the 513 individuals of European descent to reduce the effects of population stratification and limited our analyses to the 47 tissues in which data were available for at least 70 European individuals (**Supplemental Table S1**). We defined 26,289 autosomal multi-tissue gene expression outliers (median $|Z| \geq 2$ across all tissues in an individual) and 173,061 autosomal "tissue-restricted" outliers with highly aberrant expression ($|Z| \geq 4$) in two or more tissues in the same individual. Next, we identified 13,768 "singleton" SVs no larger than 1 Mb in size that were positively genotyped in one individual. These rare SVs are strongly enriched within the gene body and flanking sequence of multi-tissue gene expression outliers compared to the null expectation in 1,000 random permutations of the outlier sample names, with enrichment decreasing as flanking distance increases (**Supplemental Fig. S7**). The enrichment of rare SVs in close proximity (14.1-fold enriched within 5 kb; 95% confidence interval (CI), 8.7-25.1; $p<0.001$) to multi-tissue gene expression outliers is consistent with our prior work (Chiang et al. 2017), but the increased power in this study allows us to observe enrichment at greater distances as well. At flanking distances as large as 50 kb we observe a 6.4-fold enrichment (95% CI 4.9-8.8; $p<0.001$) of rare SVs around multi-tissue outliers, suggesting that rare SVs contribute to rare expression differences even from relatively large genomic distances. Importantly, because gene expression values can only decrease to 0, a conservative Z-score limit such as the one used for tissue-restricted outliers favors gene expression outliers with increased expression, thus limiting our ability to detect SVs associated with decreased expression (**Supplemental Fig. S8**). However, these conservative outlier definitions, combined with the above enrichment results, provide confidence in the set of outlier-associated SVs.

A total of 539 unique outlier-associated SVs are located in the gene body and 50 kb flanking region of gene expression outliers (**Fig. 2A; Supplemental Table S4**). Notably, 62.34% (336/539) of these are noncoding SVs that do not affect the coding sequence of one or more expression outliers. This contradicts the general assumption that rare SVs typically act through gene dosage effects. In total, 16.92% (31,978/188,988) of expression outliers are associated with a rare SV, although outliers can also arise via non-genetic mechanisms. To evaluate the relative potential of different SV types or sizes to cause expression outliers, we calculated the odds ratio (OR) of being outlier-associated for the SV category of interest compared to all other SVs. Duplications (OR 4.07) and mCNVs (OR 1.87) are most likely to be associated with an expression outlier, MEIs are least likely (OR 0.25) (**Fig. 2B**) and larger SVs are more likely to be outlier-associated regardless of type (**Fig. 2C**). However, many outlier-associated SVs are smaller in size (**Fig. 2D**). For example, 13.33% (50/375) of SVs associated with tissue-restricted outliers are smaller than 1 kb and nearly half (49.33%; 185/375) are smaller than 10 kb. Multi-tissue outlier-associated SVs tend to be slightly larger, with only 4.98% (12/241) smaller than 1 kb and 35.27% (85/241) smaller than 10 kb. Overall these results provide further evidence that rare SVs often affect gene expression through more complex mechanisms than large, dosage-altering events.

We next sought to determine if rare outlier-associated SVs are enriched in annotated genomic features. Although there was little signal in our enrichment analysis of common SVs, as described above, rare variants typically have larger effect sizes and are more likely to be deleterious. For this analysis, we defined a set of “control” SVs that are located within or near genes but do not exhibit expression effects. We identified 1,405 singleton SVs (1,327 noncoding) located within 50 kb of autosomal genes that showed consistent expression levels ($|Z| < 1$) across all tissues in an individual. Although this is not an ideal set of control SVs considering that some may in fact alter gene expression in tissues or developmental timepoints for which expression was not measured, it is nonetheless a relatively conservative set of likely-

nonfunctional SVs that can be used for comparison to outlier-associated SVs. We examined the overlap of both outlier- and control-associated noncoding SVs with annotated genomic features and with segmentation states from the Roadmap Epigenomics Project core 15-state model (**Fig. 3A**). We observed significant enrichment of outlier-associated SVs in 5 of the 34 evaluated features and chromatin states (Fisher's Exact test; Bonferonni $p < 0.05$). Most of these significant associations are in Roadmap Epigenomics Project segmentation states in close proximity to transcribed genes, including transcription at the 5' and 3' end of genes showing both promoter and enhancer signatures, active transcription start sites and regions flanking active transcription start sites. We also observed significant enrichment in the Roadmap Epigenomics Project segmentation state associated with zinc finger protein genes and in enhancer annotations from Genehancer. It is important to note, however, that the number of overlaps observed in this analysis is small and increased power might change these results. Thus, while rare SVs appear to have dramatic effects on gene expression, most existing functional annotations are not very informative. Consistent with this, the distribution of SV impact scores (Ganel et al. 2017) is not significantly different between expression-altering SVs and control SVs (**Supplemental Fig. S9**).

Interestingly, we found that 115 (21.34%) outlier-associated SVs are associated with more than one expression outlier and that 8 (1.48%) are associated with 5-9 expression outliers, suggesting that many rare SVs may have regional effects. In order to evaluate these broader regional effects of rare expression-altering SVs, we relaxed the definition for aberrant expression to generate a set of "secondary" expression outliers in which the tissue-restricted ("primary") outlier absolute Z-score cutoff was reduced to 3 in at least two tissues. We found significantly more primary and secondary outliers within 1 Mb of the 469 tissue-restricted outlier-associated SVs compared to the 1,496 control-associated SVs and to a null distribution in which we randomly shuffled the sample names of outlier-associated SVs 1,000 times and calculated the median number of associated outlier genes (**Fig. 3B,C**). This increase is especially

pronounced for secondary outliers whose coding regions do not overlap with the associated SV. We observe that noncoding outlier-associated SVs are associated with an average of 1.44 primary outliers (|Z|≥4) compared to an average of 0.02 associated primary outliers surrounding the shuffled null SVs (p-value=2.78×10^{-106}; one-sided Mann-Whitney U test). These differences remain for secondary outliers, with an average of 3.34 secondary outliers found in the expanded region surrounding noncoding outlier-associated SVs compared to an average of 0.54 secondary outliers for the shuffled null (p-value=4.94×10^{-76}; one-sided Mann-Whitney U test). These results suggest that rare SVs have far-reaching effects on gene expression and that these effects are primarily driven by noncoding regulatory mechanisms rather than changes to gene copy number.

DISCUSSION

We have comprehensively mapped SVs from WGS data in 613 individuals from the GTEx dataset and analyzed the impact of both common and rare SV on human gene expression. Our findings confirm results from previous analyses that SVs make an outsized contribution to common gene expression changes compared to their abundance in the genome and play an important role in rare gene expression differences (Chiang et al. 2017). A novel aspect of this study is the inclusion of a comprehensive set of MEI insertions, including those present in the GTEx samples but not the reference genome. We observed that MEIs do not play an especially important role in determining gene expression differences. In contrast, we found that mCNVs play an extremely impactful role, being 45-fold enriched among eQTL lead markers compared to their abundance in the genome and more likely to be associated with gene expression outliers (OR=1.88). mCNVs were found to give rise to most human variation in gene dosage (Handsaker et al. 2015), but our findings indicate that noncoding functional mCNVs are also abundant in the human genome.

One of the major motivators for studies such as this one is to understand the role of genetic variation in affecting gene transcription. Unfortunately, expression-altering SVs were not well correlated with any specific functional annotations other than proximity to genes, and thus existing annotations are unlikely to be informative for modeling functional variant effects. This may simply be due to a lack of power given that SVs are such a diverse class of variants that can affect large genomic segments and have the potential to affect gene expression through diverse mechanisms, and our sample size is limited to 11,026 common SVs and 539 rare SVs predicted to be functional. Alternatively, the annotations currently available may be inadequate.

Nonetheless, it is clear that SVs have broad regional impacts on human gene expression, with individual variants frequently affecting multiple genes. Interestingly, these effects are not driven by large CNVs that alter the dosage of multiple coding sequences, as one might naively expect, but are most commonly observed for noncoding variants: common noncoding eSVs affect an average of 1.50 unique genes and rare noncoding SVs are associated with an average of 1.44 primary expression outlier genes. This observation suggests a mechanism by which rare noncoding SVs may be especially deleterious, and may help explain why prior work has estimated that a surprisingly large number of rare noncoding deletions – an average of 19.1 per individual – appear to be under strong purifying selection (Abel et al. 2020). Furthermore, the burden of *de novo* CNVs has been associated with autism spectrum disorder, including for noncoding variants (Turner et al. 2017; Turner and Eichler 2019). Our results provide a mechanism through which individual noncoding SVs can have strong and potentially pleiotropic effects, and thus a higher potential to contribute to disease.

While this study represents the most comprehensive analysis of the impact of SVs on human gene expression to date, our callset is missing some of the most repetitive classes of SV, such as short tandem repeats. As long read sequencing and variant calling methods improve, we will be able to gain additional insights into repetitive variants in the most complex regions of the genome. Despite the limitations of short-read sequencing data, this study

demonstrates the importance of comprehensive variant detection when evaluating genomic variants that contribute to gene expression and disease. SVs have a disproportionately large effect on common and rare gene expression changes and often affect multiple genes. Our findings reinforce the importance of comprehensive variant detection in the design of future trait mapping studies.

METHODS

Call set generation

We obtained 613 whole-genome sequencing BAM files from the GTEx v7 release (dbGap accession phs000424.v7.p2, accessed 1 June 2016). Structural variant calls were generated using both the SpeedSeq v0.1.1 pipeline (Chiang et al. 2015), which performs sample-level breakpoint detection via LUMPY v.0.2.13 (Layer et al. 2014) followed by population-scale merging and genotyping of SV calls via svtools v0.3.1 (Larson et al. 2019) and the GenomeSTRiP v2.00.1636 read-depth analysis pipeline (Handsaker et al. 2011), as described in our preliminary GTEx study (Chiang et al. 2017). GenomeSTRiP false discovery rate (FDR) was evaluated based on available Illumina Human Omni 5M gene expression array data (n=161) using the GenomeSTRiP IntensityRankSumAnnotator. We limited GSCNQUAL to ≥ 1 for GenomeSTRiP deletions and to ≥8 for multiallelic copy number variants, corresponding to an FDR of 10%. The GSCNQUAL cutoff for GenomeSTRiP duplications was set at ≥17, the point at which the FDR plateaued at 15.1% and did not fluctuate more than ±1% for over 50 steps of increasing GSCNQUAL score. Redundant Lumpy and GenomeSTRiP calls were merged as previously described (Chiang et al. 2017). Additionally, we ran the Mobile Element Locator Tool (MELT) v2.1.4 using MELT-SPLIT to identify ALU, SVA and LINE1 insertions into the test genomes (Gardner et al. 2017). We retained MELT calls categorized as “PASS” in the VCF info field that had an ASSESS score ≥3 and SR count ≥3. Genome Analysis Toolkit (GATK) HaplotypeCaller v3.4 (McKenna et al. 2010) SNV and indel calls were obtained from the GTEx

consortium (dbGap accession phs000424.v7.p2, accessed 1 June 2016). We use allele balance instead of genotype for analyses described in this paper because it is tolerant to alignment inefficiencies for the alternate SV allele. For MEIs identified by MELT, we converted generated genotypes (0/0, 0/1, 1/1) to integer values (0, 1, 2) that were used as a proxy for allele balance to allow for comparable analyses on these variants.

Common eQTL mapping

We mapped *cis*-eQTLs in each of the 48 tissues for which both WGS data and RNA-seq data was available in ≥70 individuals. Available tissues and those used in each analysis are listed in **Supplemental Table S1**. We refer to EBV-transformed lymphocytes and transformed fibroblasts as tissue types throughout this study for convenience. Biospecimen collection, RNA-seq data alignment, RPKM calculations and data normalization were previously described (Lappalainen et al. 2013; Chiang et al. 2017).

We selected common genetic markers, defined as having MAF ≥ 0.01, for eQTL mapping. We performed a joint *cis*-eQTL analysis that included 26,409 common SVs, as well as 9,609,545 common SNVs and 818,401 common indels detected using GATK, to allow for a fair comparison of the contribution of different variant types. We used FastQTL v2.184 (Ongen et al. 2016) to perform *cis*-eQTL mapping, customized to accomidate the unique architecture of SVs (Chiang et al. 2017), using a *cis* window of 1 Mb on either side of the TSSs of autosomal and X-chromosome genes with a permutation analysis to identify the most significant marker for each gene. For each tissue we applied the same covariates described in Chiang et al. 2017. We corrected for multiple-testing at the gene-level using the Benjamini-Hochberg method with a 10% FDR.

Feature enrichment

To evaluate whether SVs that cause common gene expression changes are enriched in particular genomic features, we calculated a previously described causality score (Chiang et al. 2017) generated by taking the product of the SV heritability fraction obtained from GCTA (Yang et al. 2011) and the causal probability generated by CAVIAR (Hormozdiari et al. 2016) for the strongest-associated SV within the *cis* region of each eQTL. No associated SVs were identified in 199 eQTLs due to the subset of samples with available data in the relavent tissue and thus were not included in enrichment analyses. GCTA heritability estimates could not be calculated for a small number of eQTLs (6,146/299,187) due to nonpositive definite matrices, likely resulting from small sample sizes, and these loci were excluded from feature enrichment analyses. For SVs that were associated with multiple eQTLs or the same eQTL in multiple tissues, we selected the eQTL (tissue/gene pair) for which the SV had the highest causality score. SVs were allocated into bins based on causality score quantiles, with the first bin consisting of SVs in the bottom 50% of causality scores and the other five consisting of deciles of the top 50% of scores.

Next, we counted the number of SVs in each bin that intersected with various genomic annotations. We allowed 1 kb of flanking distance surrounding all annotations with the following exceptions: GENCODE exons, no flanking distance; proximity to TSS and 3' gene end, 10 kb of directional flanking distance; topologically associated domain boundaries, 5 kb of flaking distance; Roadmap Epigenomics segmentation states, no flanking distance. SVs associated with multiple eGenes were considered to touch an eGene if they overlapped with the exons of any associated gene. SVs that touched an exon of an associated eGene were excluded from all feature enrichment analyses except for the enrichment of affected eGenes. To generate a shuffled null for comparison, SVs within each causality bin were shuffled with BEDTools v2.23.0 (Quinlan and Hall 2010) into non-gapped regions of the genome within 1 Mb of the TSS of a gene. We did not allow shuffled SVs to intersect any exons of their new eGene. We calculated the fold enrichment of the number of SVs that intersect with each genomic feature compared to

the median number of intersections observed for 100 randomly shuffled sets within each causality bin. These shuffled sets were also used to empirically derive the 95% confidence intervals.

Regions 10 kb upstream of TSS and downstream of 3' gene end were defined based on GENCODE v19 gene positions. DNAse hypersensitive regions and enhancer regions with a minimum support of 2 were obtained from the Dragon ENhancers database (DENdb) (Ashoor et al. 2015). We downloaded FunSeq 2.1.0 (Fu et al. 2014) regions and topologically associated domain boundaries from human embryonic stem cells from author websites (http://archive.gersteinlab.org/funseq2.1.0_data/ and http://compbio.med.harvard.edu/modencode/webpage/hic/hESC_domains_hg19.bed). GeneHancer (Fishilevich et al. 2017) enhancer regions for b38 were downloaded from the UCSC genome browser (Kent et al. 2002) and lifted over to b37 using CrossMap v0.2.6 (Zhao et al. 2014). Regions defined by the ENCODE (Encode Project Consortium 2012) project were downloaded from the UCSC genome browser. To evaluate the intersection with the chromatin segmentation state annotations from the Roadmap Epigenomics Project (Kundaje et al. 2015), we downloaded the core 15-state model annotations for all 127 available epigenomes (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final). We used BEDTools multiIntersectBed (Quinlan and Hall 2010) to identify genomic intervals where each of the 15 annotations is found in at least 10 of the 127 available epigenomes and used these collapsed regions as the annotation intervals for SV intersections.

eQTL tissue specificity

We selected significant gene-variant pairs identified in eQTL mapping with available expression data available across all 48 tissues in which eQTL analyses were performed. These pairs were only required to have a significant eQTL in one tissue. We used METASOFT v2.0.0 (Han and

Eskin 2011) to perform a meta-analysis of the selected eQTL effect sizes and their standard errors across all 48 tissues. METASOFT employs a mixed effects model (RE2) to generate a posterior probability that an effect exists in each tissue (*m*-value) (Han and Eskin 2012). To allow computational feasibility with the relatively large number of tissues sampled, the Markov Chain Monte Carlo (MCMC) method was used to approximate these values. The *m*-values generated indicate whether a tested eQTL is active ($m>0.9$), inactive ($m<0.1$), or has ambiguous activity ($0.1\leq m\leq 0.9$). Only eQTLs with at least 43 tissues having known (active or inactive) activity were included in analyses. eQTLs with active status in at least 75% of tissues with known activity were defined as "constitutively active."

Identification of expression outliers

We limited outlier analyses to the 513 European individuals, the largest subpopulation in the cohort, who had available WGS data. We performed Z transformation of PEER-corrected expression values without quantile normalization across the 47 tissues for which RNA-seq data was available from the GTEx consortium for at least 70 European individuals (**Supplemental Table S1**). We defined two sets of gene expression outliers (gene/sample pairs) among these individuals: "multi-tissue" expression outliers in which an individual's absolute median Z-score of a gene's expression across all available tissues was ≥2, as previously described in (Chiang et al. 2017), and "tissue-restricted" outliers in which an individual's absolute Z score for a gene's expression was ≥4 in at least two different tissues. The two tissue requirement was necessary to eliminate false positive expression outliers resulting from individual tissues with systematically aberrant gene expression profiles for an individual. Additionally, we defined a set of control gene/sample pairs in which an individual's absolute Z score of a gene's expression was less than 1 across all tissues for which RNA-seq data was available. For all definitions we limited to gene/sample pairs with data available in at least 5 tissues. We removed one individual (GTEX-14753) from this analysis due an excessive number of expression outliers.

Rare variant association with expression outliers

We identified 15,016 structural variants that were positively genotyped in no more than one individual in the European cohort. Because large rare structural variants tend to affect gene expression through dosage changes, we removed 12 variants larger than 1 Mb in size from this analysis. We calculated the enrichment of singleton SVs overlapping with multi-tissue outlier transcripts and the flanking 5 kb sequence by randomly shuffling the outlier individual names 1,000 times to determine the median number of times a rare variant randomly co-occurred with an outlier, as described in (Chiang et al. 2017). We also performed the reciprocal analysis counting the number of outliers that co-occurred within 5 kb of a rare SV. We repeated these calculations for increased outlier-flanking regions of 10 kb, 25 kb, 50 kb and 100 kb. We calculated the odds ratio of being outlier-associated by dividing the ratio of outlier-associated SVs to non-outlier associated SVs in a category of interest (SV type or size) by the ratio of outlier-associated SVs to non-outlier associated SVs for all SVs not included in the category.

Feature enrichment for outlier-associated SVs

We performed intersections between the 369 noncoding outlier-associated SVs and the same genomic features and chromatin segmentation states evaluated for eSVs. The above intersections were repeated for the 1,416 noncoding control-associated SVs. We calculated the fold enrichment of outlier-associated SVs in each feature compared to control-associated SVs and determined significant enrichments using a Fisher's Exact Test with Bonferroni correction for multiple testing.

Regional effect of rare SVs

To evaluate the broader regional effects of rare, gene expression-altering SVs, we counted the number of tissue-restricted outlier genes, referred to as "primary" outliers, located in the spanning region and 1 Mb of flanking sequence both upstream and downstream of the 469 SVs

previously identified as being associated with an expression outlier. We repeated this analysis with a relaxed definition of tissue-restricted expression outliers, referred to as "secondary" outliers, in which the absolute Z score cutoff was reduced from $|Z| \geq 4$ to $|Z| \geq 3$. We compared the number of primary and secondary outliers found in the expanded region surrounding outlier-associated SVs to the expanded region surrounding the 1,224 control-associated SVs. Finally, because the controls defined above do not represent a null expectation, we performed 1,000 random permutations of the outlier-associated SV sample names and calculated the median number of associated primary and secondary outliers for each SV in order to determine how frequently rare expression-altering SVs co-occurred with primary and secondary outliers in random individuals.

DATA ACCESS

The SV genotype data generated in this study have been submitted to AnVIL (https://app.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V7_hg19) under dbGaP accession number phs000424.

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGMENTS

The authors thank E.J. Gardner for advice on MELT and R.E. Handsaker for advice on Genome STRiP. This work was supported by a Mr. and Mrs. Spencer T. Olin Fellowship for Women in Graduate Study (A.J.S.) and by the NIH/NHGRI UM1 HG008853 (I.M.H).

REFERENCES

Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, et al. 2020. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**: 83–89.

Ashoor H, Kleftogiannis D, Radovanovic A, Bajic VB. 2015. DENdb: database of integrated human enhancers. *Database* **2015**. http://dx.doi.org/10.1093/database/bav085.

Brandler WM, Antaki D, Gujral M, Kleiber ML, Whitney J, Maile MS, Hong O, Chapman TR, Tan S, Tandon P, et al. 2018. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**: 327–331.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* **12**: 966–968.

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, GTEx Consortium, et al. 2017. The impact of structural variation on human gene expression. *Nature Genetics* **49**: 692–699.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**: 71–86.

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581**: 444–451.

Encode Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al. 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**. http://dx.doi.org/10.1093/database/bax028.

Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M. 2014. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**: 480.

Ganel L, Abel HJ, FinMetSeq, Consortium, Hall IM. 2017. SVScore: an impact prediction tool for structural variation. *Bioinformatics* **33**: 1083–1085.

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, Genomes Project, Consortium, Devine SE. 2017. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* **27**: 1916–1929.

Han B, Eskin E. 2012. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet* **8**: e1002555.

Han B, Eskin E. 2011. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* **88**: 586–598.

Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**: 269–276.

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. 2015. Large multiallelic copy number variations in humans. *Nat Genet* **47**: 296–303.

Han L, CommonMind Consortium, Zhao X, Benton ML, Perumal T, Collins RL, Hoffman GE, Johnson JS, Sloofman L, Wang HZ, et al. 2020. Functional annotation of rare structural variation in the human brain. *Nature Communications* **11**. http://dx.doi.org/10.1038/s41467-020-16736-1.

Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. 2014. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* **198**: 497–508.

Hormozdiari F, van de Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. 2016. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* **99**: 1245–1260.

International Schizophrenia Consortium. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**: 237–241.

Jakubosky D, D'Antonio M, Bonder MJ, Smail C, Donovan MKR, Young Greenwald WW, Matsui H, i2QTL Consortium, D'Antonio-Chronowska A, Stegle O, et al. 2020. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat Commun* **11**: 2927.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–511.

Larson DE, Abel HJ, Chiang C, Badve A, Das I, Eldred JM, Layer RM, Hall IM. 2019. svtools: population-scale analysis of structural variation. *Bioinformatics* **35**: 4782–4787.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, Hess GT, Zappala Z, Strober BJ, Scott AJ, et al. 2017. The impact of rare variation on gene expression across tissues. *Nature* **550**: 239–243.

Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, Antaki D, Shetty A, Holmans PA, Pinto D, et al. 2017. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* **49**: 27–35.

McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, et al. 2009. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* **41**: 1223–1227.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. 2016. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**: 1479–1485.

Payer LM, Burns KH. 2019. Transposable elements in human genetic disease. *Nat Rev Genet* **20**: 760–772.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. 2007. Strong association of de novo copy number mutations with autism. *Science* **316**: 445–449.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848–853.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, Kronenberg ZN, Hormozdiari F, Raja A, Pennacchio LA, et al. 2017. Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**: 710–722.e12.

Turner TN, Eichler EE. 2019. The Role of De Novo Noncoding Regulatory Mutations in Neurodevelopmental Disorders. *Trends Neurosci* **42**: 115–127.

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al. 2008. Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science* **320**: 539–543. http://dx.doi.org/10.1126/science.1155174.

Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, et al. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**: 667–675.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**: 76–82.

Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**: 1006–1007.

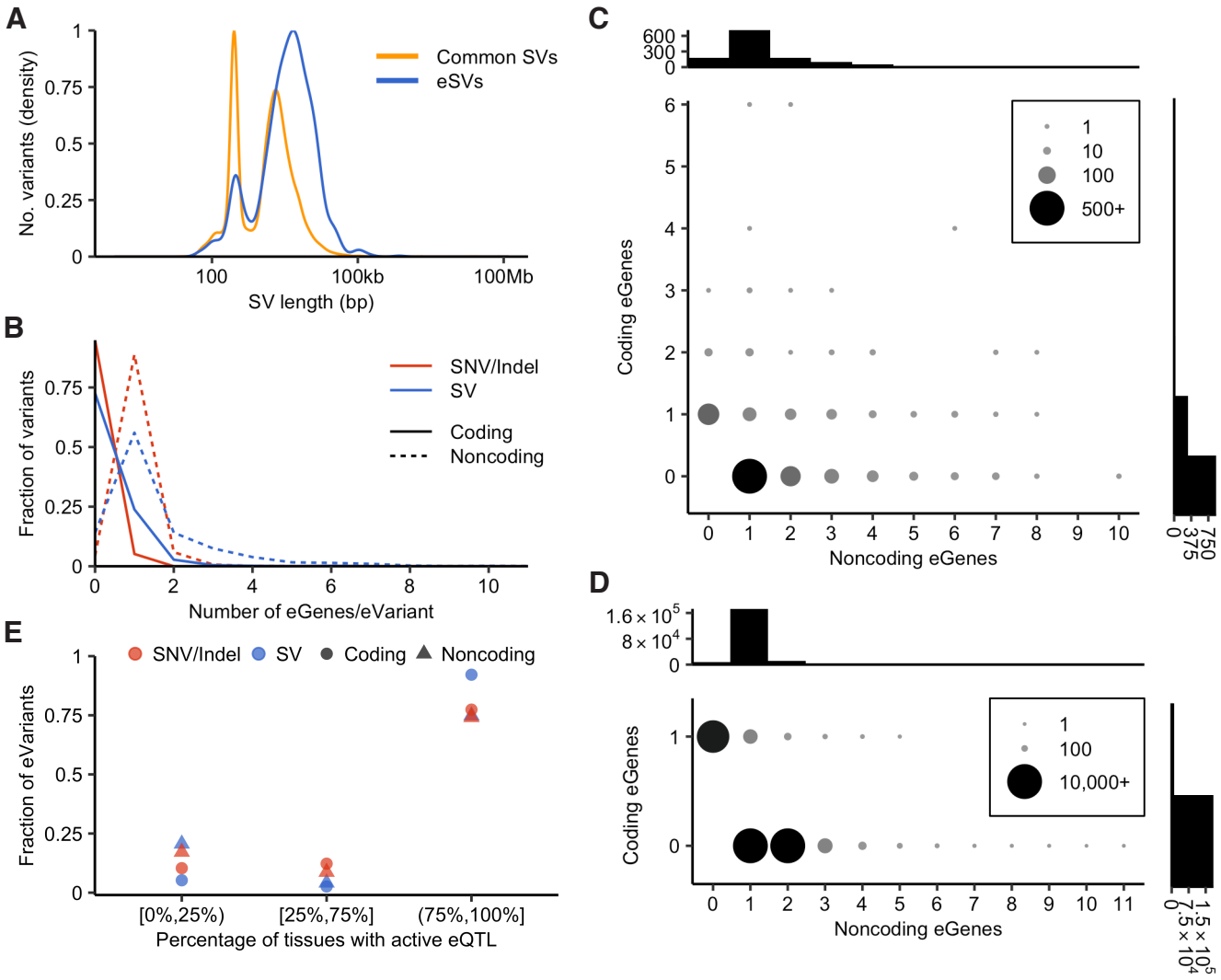

Figure 1. Features of SV-eQTLs. (**A**) Size distribution of eSVs compared to all common SVs. (**B**) Distribution of the number of eGenes per eVariant for SVs compared to SNVs and indels. "Coding" eGenes refer to eGenes whose exons are intersected by the associated eVariant and "noncoding" eGenes are not intersected by the associated eVariant. Counts are shown for every eVariant, thus eVariants with zero coding or zero noncoding eGenes are included in the distributions. (**C,D**) The number of eVariants, as shown by dot size and color, with the indicated combination of coding and noncoding eGenes, as defined above. Shown for SVs (**C**) and SNV/indels (**D**), with histograms showing the total number of eVariants with the indicated number of associated coding or noncoding eGenes above the y- and x-axes, respectively. (**E**) Distribution of tissue specificity of eQTLs across tissues as evaluated by METASOFT, separated into the lowest quartile, middle two quartiles and top quartile, for eQTLs in which the activity status is known in at least 43 of 48 evaluated tissues. The points indicate the fraction of SV-eQTLs or SNV- and indel-eQTLs that are active (m>0.9) in the proportion of tissues indicated on the x-axis.

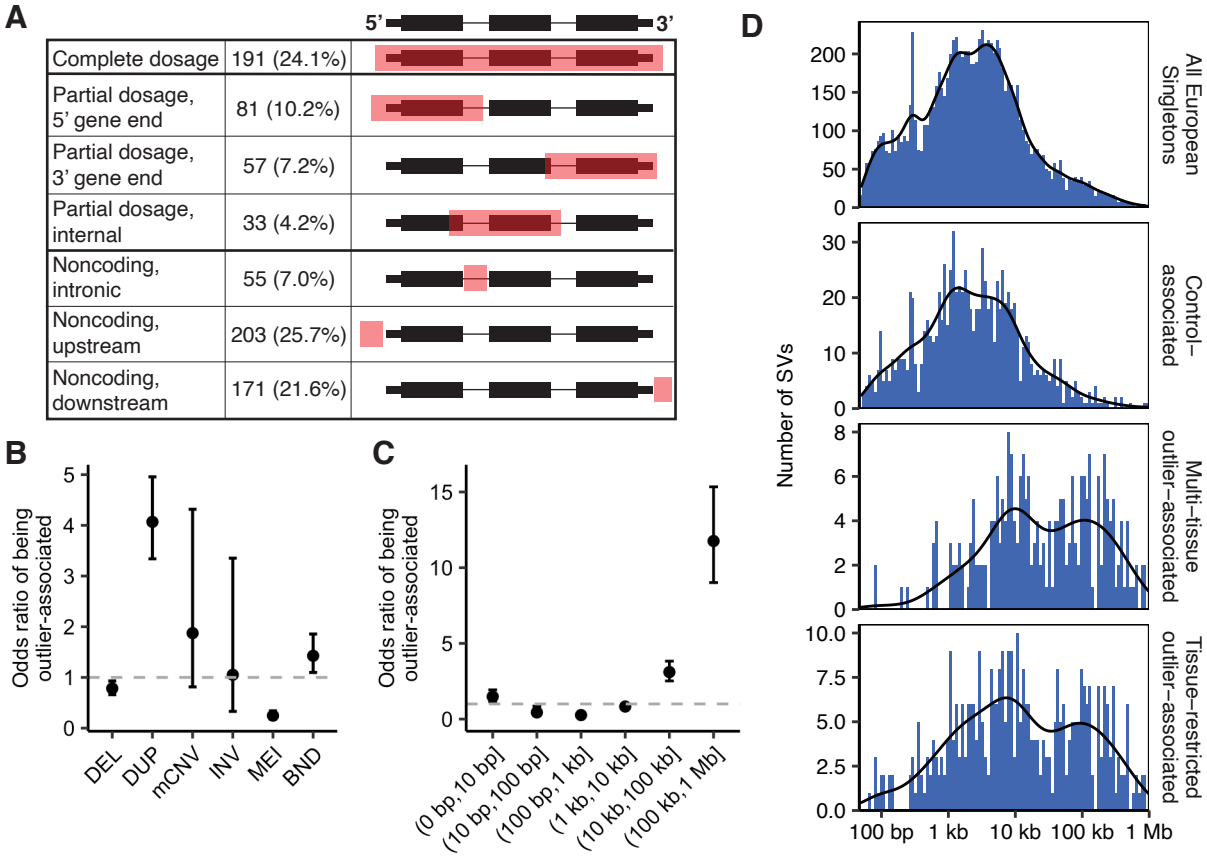

Figure 2. Features of outlier-associated SVs. (**A**) Location of outlier-associated SVs relative to their associated outlier gene and the number of SV/outlier gene associations identified in each category. Percentages indicate the fraction of outlier/SV pairs found at each relative location compared to the total number of SV/outlier gene associations. Note that this definition allows one SV to be associated with multiple outlier genes and thus the SV is counted in multiple categories. Gene diagrams provide examples of possible SV location, shown in red, relative to the outlier gene. (**B,C**) Odds ratio (OR) of being outlier-associated by SV type (**B**) and SV size (**C**) for the SV category of interest compared to all other SVs. Note that BNDs were excluded from the size OR calculations due to their ambiguous nature and thus size. (**D**) Distribution of SV sizes for singleton SVs smaller than 1 Mb identified in European individuals that were used in outlier analyses. Panels depict size distributions for all European-cohort singletons, control-associated singletons, multi-tissue outlier-associated singletons and tissue-restricted outlier-associated singletons.

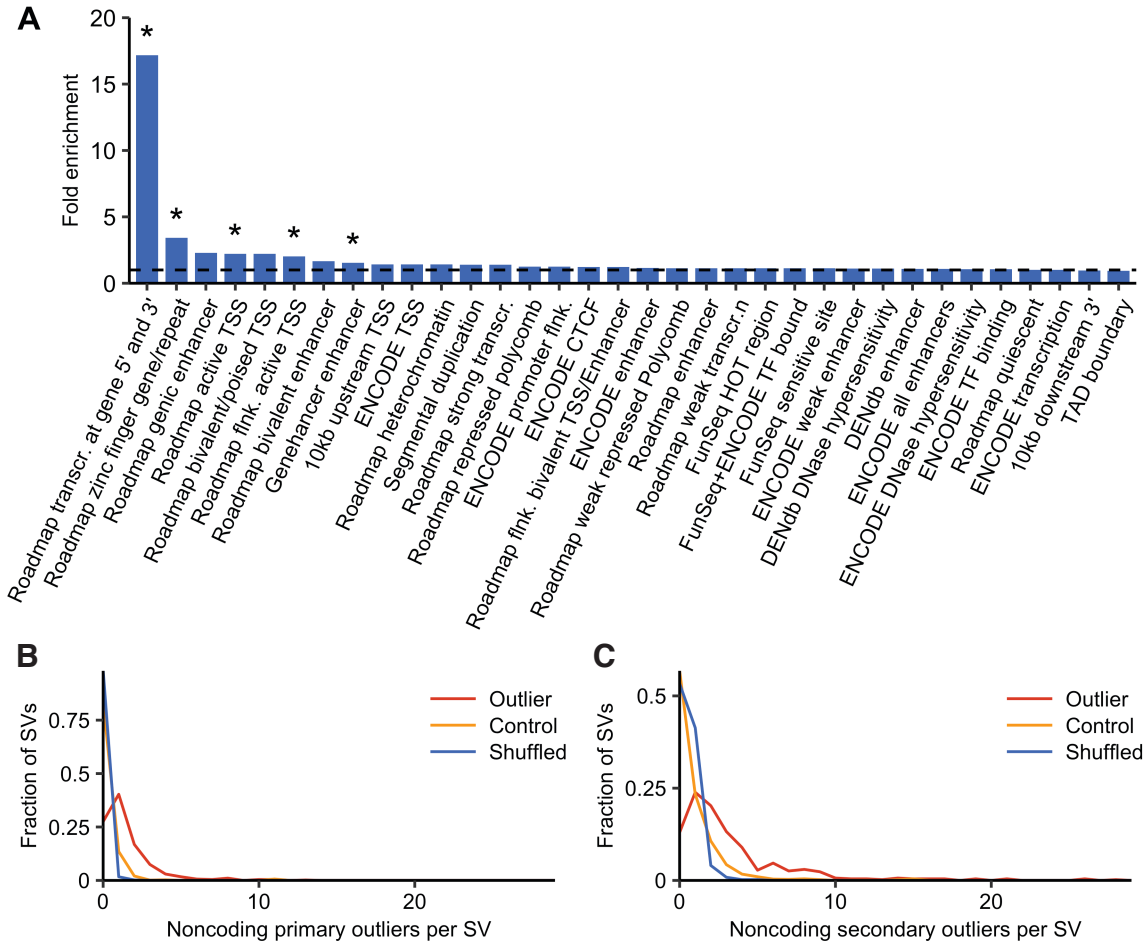

Figure 3. Mechanistic insights into outlier-associated SVs. (**A**) Enrichment of outlier-associated SVs in functional genomic annotations compared to control-associated SVs. Asterisks indicate statistical significance based on a Fisher's exact test with Bonferroni correction for multiple testing. (**B,C**) The distribution of the number of noncoding primary (**B**) and secondary (**C**) outliers found within 1 Mb of the region surrounding tissue-restricted outlier-associated SVs, control-associated SVs and a shuffled null.