nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-16482-4 OPEN

Properties of structural variants and short tandem repeats associated with gene expression and complex traits

David Jakubosky[1,2], Matteo D'Antonio[3], Marc Jan Bonder[4,5], Craig Smail[6,7], Margaret K. R. Donovan[2,8], William W. Young Greenwald[8], Hiroko Matsui[3], i2QTL Consortium*, Agnieszka D'Antonio-Chronowska[3], Oliver Stegle[4,5,9], Erin N. Smith[10], Stephen B. Montgomery[7,11], Christopher DeBoever[3] & Kelly A. Frazer[3,10]✉

Structural variants (SVs) and short tandem repeats (STRs) comprise a broad group of diverse DNA variants which vastly differ in their sizes and distributions across the genome. Here, we identify genomic features of SV classes and STRs that are associated with gene expression and complex traits, including their locations relative to eGenes, likelihood of being associated with multiple eGenes, associated eGene types (e.g., coding, noncoding, level of evolutionary constraint), effect sizes, linkage disequilibrium with tagging single nucleotide variants used in GWAS, and likelihood of being associated with GWAS traits. We identify a set of high-impact SVs/STRs associated with the expression of three or more eGenes via chromatin loops and show that they are highly enriched for being associated with GWAS traits. Our study provides insights into the genomic properties of structural variant classes and short tandem repeats that are associated with gene expression and human traits.

[1] Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA 92093-0419, USA. [2] Department of Biomedical Informatics, University of California San Diego, La Jolla, CA 92093-0419, USA. [3] Institute of Genomic Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA. [4] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK. [5] Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [6] Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA. [7] Department of Pathology, Stanford University, Stanford, California 94305, USA. [8] Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA, USA. [9] Division of Computational Genomics and Systems Genetics, German Cancer Research Center, Heidelberg, Germany. [10] Department of Pediatrics, University of California San Diego, La Jolla, CA 92093, USA. [11] Department of Genetics, Stanford University, Stanford, California 94305, USA. *A list of authors and their affiliations appears at the end of the paper. ✉email: kafrazer@ucsd.edu

Structural variants (SVs) and short tandem repeats (STRs) are important categories of genetic variation that account for the majority of base pair differences between individual genomes and are enriched for associations with gene expression[1–3]. SVs and STRs are comprised of several diverse classes of variants (e.g., deletions, insertions, multi-allelic copy number variants (mCNVs), and mobile element insertions (MEIs)), and multiple algorithmic approaches and deep whole genome sequencing are required to accurately identify and genotype variants in these different classes[4]. Due to the complexity of calling SVs and STRs, previous genetic association studies have generally not identified a comprehensive set of these variants but rather have focused on one or a few of the class types, and therefore the genomic properties of SVs and STRs associated with gene expression and/or complex traits are not well characterized.

SV classes and STRs vary in genomic properties including size, distribution across the genome, and impact on nucleotide sequences, but previous studies have not investigated whether these differences influence the likelihood of being an expression quantitative trait locus (eQTL), eQTL effect sizes, or the properties of eQTL genes (eGenes) such as gene type or level of evolutionary constraint[3,5–7]. Further, it is unknown if the variant classes may affect gene expression through different mechanisms such as altering gene copy number or three-dimensional spatial features of the genome. A comprehensive SV and STR data set generated using high-depth whole genome sequencing (WGS) from a population sample with corresponding RNA-sequencing data could be used to assess whether genomic features of SV classes and STRs are associated with properties of eGenes and eQTLs.

SVs and STRs have also been associated with complex traits, though they have been studied considerably less often in genome-wide association studies (GWAS) than single nucleotide variants (SNVs), and the overall contribution of SVs and STRs to complex traits is not well understood[8–15]. One difficulty with studying differences between SV classes and STRs in GWAS is that it is unknown whether the SV classes are differentially tagged by SNVs on genotyping arrays. A collection of hundreds of subjects genotyped for a full range of SVs, STRs, SNVs, and insertion/deletions (indels) could be used to assess the functional impact of SVs and STRs on complex traits using existing SNV-based GWAS and identify dark regions of the genome not captured by array GWAS.

In this study, as part of the i2QTL Consortium, we use RNA-sequencing data from induced pluripotent stem cells (iPSCs) from the iPSCORE and HipSci collections[7,16,17] along with a comprehensive call set of SVs and STRs from deep WGS data[4] to identify variants associated with iPSC gene expression and characterize the genomic properties of these SV and STR eQTLs. We observe that SVs are more likely to act as eQTLs than SNVs when in distal regions (> 100 kb from eGenes) and that duplications and mCNVs are more likely to have distal eQTLs and multiple eGenes compared to other SVs classes and STRs. eGenes for mCNV eQTLs are also less likely to be protein coding and more likely to have strong effect sizes relative to other SV classes and STRs. We examine the LD of SVs and STRs with GWAS variants and find that mCNVs and duplications are poorly tagged by GWAS SNVs compared to other variant classes. 11.4% of common SVs and STRs are in strong LD with a SNV associated with at least one of 701 unique GWAS traits; and deletion, rMEI, ALU, and STR lead eQTL variants are enriched for GWAS associations establishing that these variant classes have underappreciated roles in common traits. Finally, we find a highly impactful set of SVs and STRs located near high complexity loop anchors that localize near multiple genes in three dimensional space and are enriched for being associated with the expression of multiple genes and GWAS traits. This work establishes that different classes of SVs and STRs vary in their functional properties and provides a valuable, comprehensive eQTL data set for iPSCs.

Results

eQTL mapping. We performed a *cis*-eQTL analysis using RNA sequencing data from iPSCs derived from 398 donors in the iPSCORE and HipSci projects along with a comprehensive map of genetic variation (37,296 SVs, 588,189 STRs, and ~48 M SNVs and indels (Supplementary Data 1)) generated using deep WGS from these same donors[4]. These variants include several classes of SVs including biallelic duplications and deletions; multi-allelic copy number variants (mCNVs); mobile element insertions (MEIs) including LINE1, ALU, and SVA; reference mobile element insertions (rMEI); inversions; and unspecified break-ends (BNDs). We identified 16,018 robustly expressed autosomal genes and tested for *cis* associations between the genotypes of all common (MAF ≥ 0.05) SVs (9,313), STRs (33,608), indels (~1.52 M), and SNVs (~5.83 M) within 1 megabase of a gene body using a linear mixed model approach (Fig. 1a and Supplementary Data 2, "Methods" section). We detected associations between 11,197 eGenes (FDR < 5%, Methods) and 10,904 unique lead variants (lead eVariants), including 145 SVs (1.3%), 140 STRs (1.3%), 2648 indels (24.3%), and 7971 SNVs (73.1%, Fig. 1b and Table 1). We compared our eQTLs to those discovered by GTEx and 1000 Genomes and found that the number of eGenes we identified is consistent with the expected power from using 398 samples (Supplementary Figs. 1–3)[1,3]. While SVs and STRs accounted for only 0.1 and 0.38% of tested variants respectively in our analysis, they were highly enriched to be lead eVariants (SVs: OR = 17.9, p = 3.3e-91; STRs: OR = 4.14, p = 1.5e-24; Fisher's exact test (FET)) and collectively formed lead associations with 3.25% of eGenes (1.73% SVs and 1.52% STRs), indicating that these variant classes have a disproportionate effect on gene expression compared to SNVs and indels.

To conduct comparative analyses of the functional properties of the different SV classes and STRs, we performed an SV/STR-only eQTL analysis using the 42,921 common SVs and STRs and excluding SNVs and small indels (Fig. 1a and Supplementary Data 3). We identified 6,966 eGenes (FDR < 5%) associated with 5,343 unique lead eVariants (Table 1 and Fig. 1c). SVs were enriched among lead variants compared to STRs (OR = 1.15, p = 1.7e-5, FET) though the majority of lead eVariants were STRs (4087 eSTRs vs 1231 eSVs). Of the 11,197 eGenes identified in the joint eQTL analysis, 6,507 were also identified in the SV/STR-only eQTL analysis (Fig. 1d). Among these 6,507 shared eGenes, 94.6% (6,155) were mapped to a lead SNV or indel variant in the joint analysis, while the remaining 5.4% (352) were mapped to the same lead SV or STR identified in the SV/STR eQTL analysis. To evaluate how many of the 6,155 shared eGenes were likely driven by the same causal variant, we computed the linkage disequilibrium (LD) between SNV/indel lead variants in the joint eQTL analysis and eSVs and eSTRs from the SV/STR-only eQTL analysis. We found that lead SNVs or indels from the joint analysis were in strong LD (R^2 > 0.8) with the lead eSV or eSTR from the SV/STR-only analysis for 14.2% (872/6,155) of shared eGenes. While the true causal variant at these loci is unknown, these data suggest that a substantial number of eQTLs that can be identified using SNVs may be explained by SVs or STRs.

Variant size influences eQTL associations. Given that SVs and STRs have size ranges that span orders of magnitude[4], we sought to examine the relationship between variant length and the likelihood of being an eVariant across the different variant classes. We tested whether STRs or deletions, duplications, and

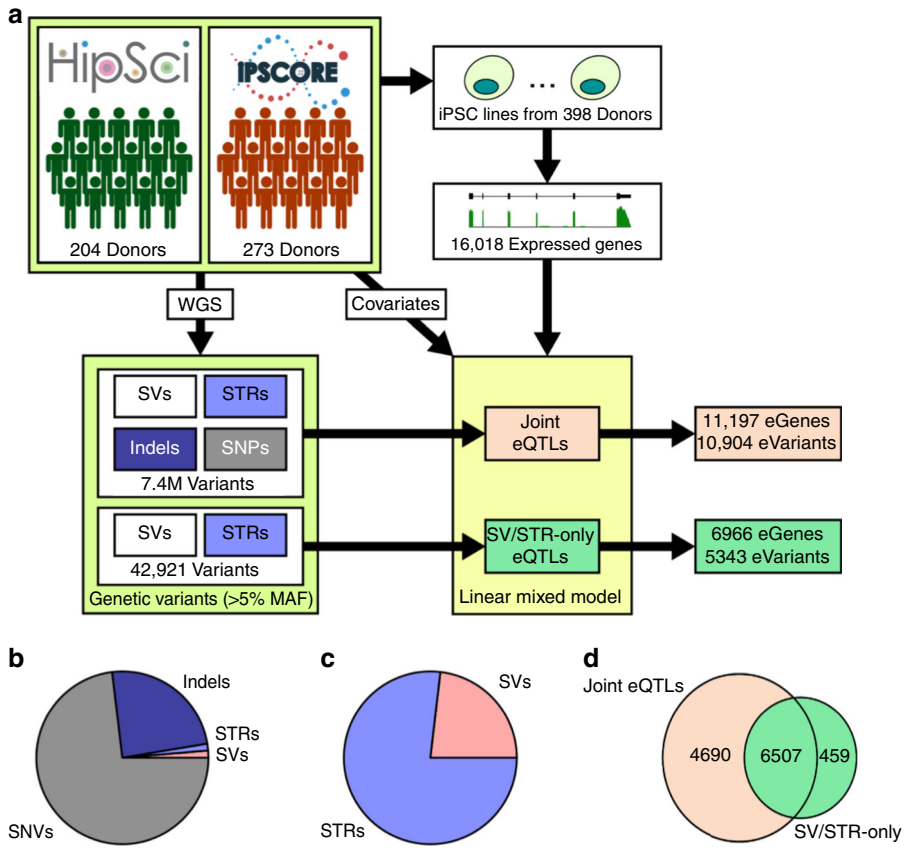


Fig. 1 eQTL mapping. a Overview of eQTL study design. We performed two eQTL analyses: a joint analysis that used all variants and identified 11,197 eGenes and an SV/STR-only analysis that only used SVs and STRs and identified 6,996 eGenes. **b**,**c** Pie charts showing the number of lead variants across the different variant classes for (**b**) joint and (**c**) SV/STR-only eQTL analyses. **d** Venn diagram showing the intersection between the eGenes detected in the joint and the SV/STR-only analysis.

Table 1 Summary of i2QTL variants and eQTL results.

Variant Class	No. Variants	No. Common Variants (tested)	Lead SV/STR-Only QTLs	Lead Joint QTLs
SNV	41,826,418	5,834,257		8,148
INDEL	7,040,457	1,520,762		2,685
Deletion (DEL)	16,238	3,073	661	51
Duplication (DUP)	2,693	391	55	9
Multiallelic CNV (mCNV)	1,703	947	294	111
Other SV (BND)	4,612	1,146	89	8
Inversion INV	210	84	11	0
Reference Mobile Element Insertions (rMEI)	2,343	1,448	243	3
ALU	7,880	1,932	294	9
LINE1	1,175	196	31	1
SVA	442	96	28	2
Short Tandem Repeats (STR)	588,189	33,608	5,260	170
Total SV	37,296	9,313	1,706	194
Total SV/STR	625,485	42,921	6,966	364
Total	49,492,360	7,397,940	6,966	11,197

Numbers in each category refer to the number of non-redundant variants that were within 1 Mb of a gene and used in the eQTL analyses. Variants used for eQTL mapping had ≥ 5% minor allele frequency for SNVs and indels and ≥5% non-mode allele frequency for SVs and STRs. Lead SV/STR-Only QTLs column shows the number of lead variants in the eQTL analysis using only SVs and STRs while Lead Joint QTLs column shows the number of lead variants in the eQTL analysis using SNVs, indels, SVs, and STRS.

mCNVs longer than a particular length threshold were more likely to be eVariants compared to variants shorter than the length threshold. We found that longer deletions, duplications and STRs were more likely to be eVariants and lead eVariants than shorter variants (Fig. 2a and Supplementary Fig. 4). The trend was especially strong for deletions where 36% of variants longer than 50 kb were lead eVariants (OR = 3.11, $p = 0.0095$, FET). Although a higher proportion of mCNVs were eVariants compared to other classes (Fig. 2a), mCNV length was not strongly associated with eQTL status; only mCNVs longer than 10 kb were significantly more likely to be lead eVariants (OR = 1.59, $p = 0.01$, FET).

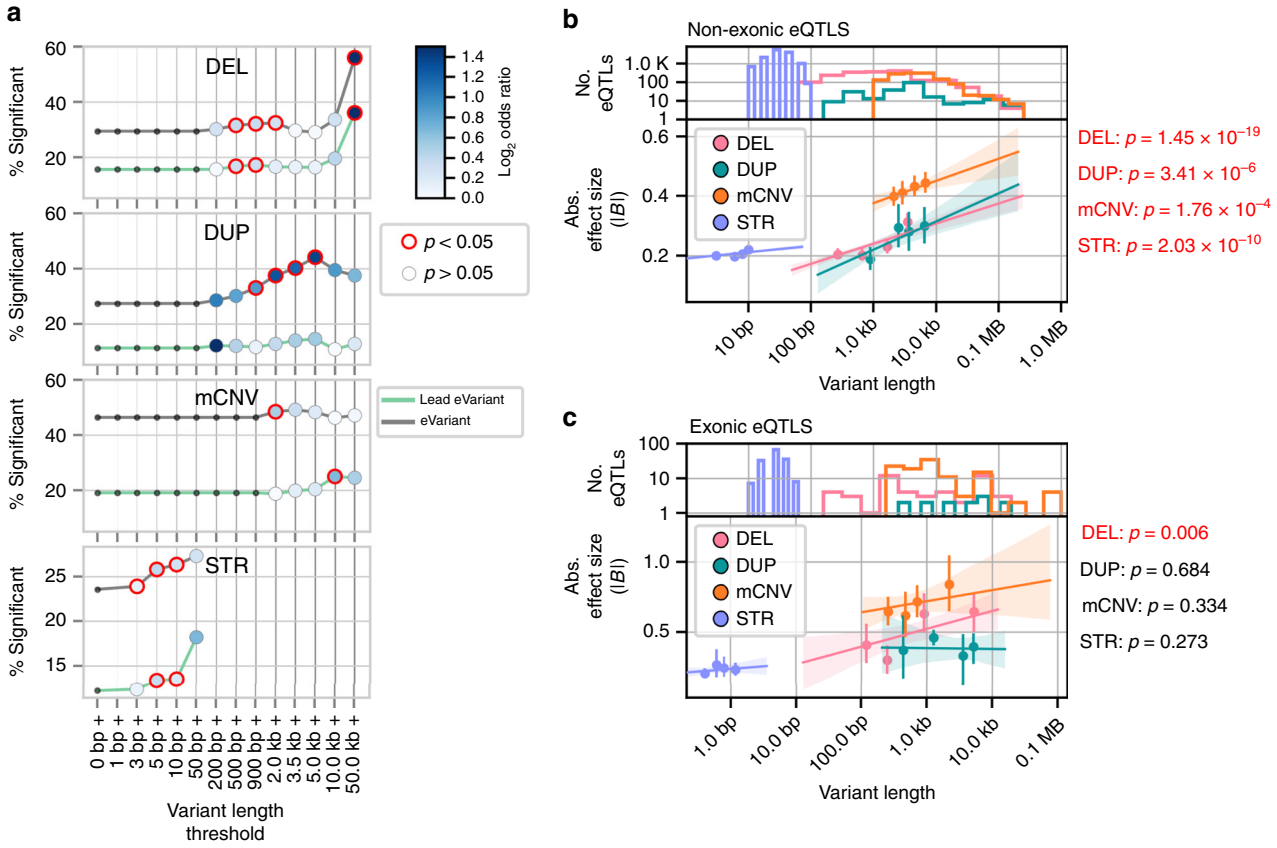
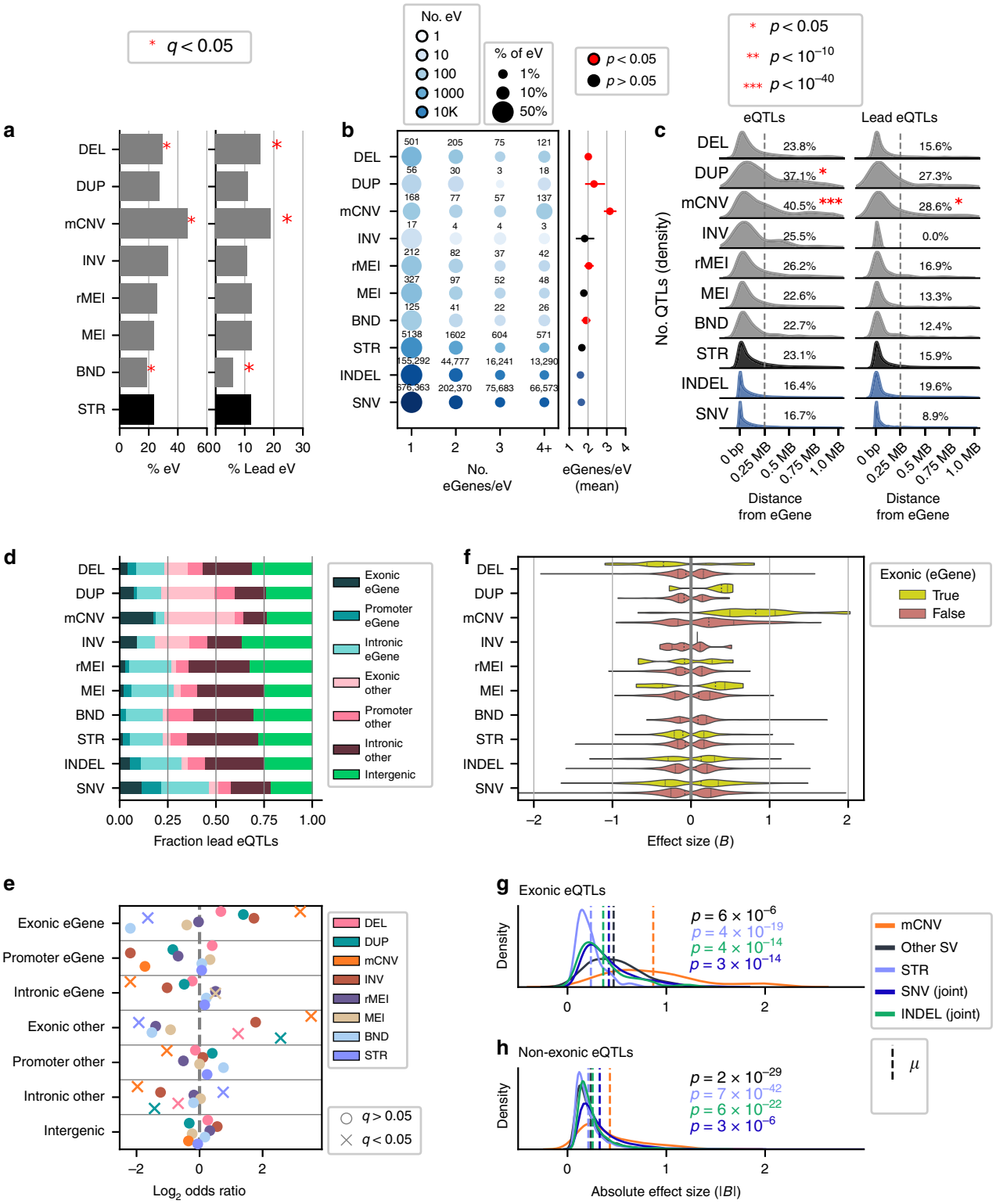


Fig. 2 Variant length influences the likelihood and effect size of eQTLs. a Percentage of variants with length greater than the thresholds on the *x* axis that were eVariants (grey lines) or lead eVariants (green lines). Points are colored according to the enrichment (log 2 odds ratio) of variants above each threshold among eVariants or lead eVariants relative to variants smaller than the threshold; points circled in red were significant (FET two-sided, $p < 0.05$). A complete list of *p* values and odds ratios is provided in Supplementary Fig. 4. (**b**,**c**) Association of variant length with eQTL effect size for (**b**) non-exonic eQTLs or (**c**) exonic eQTLs mapped to biallelic deletions ($n = 1{,}769$ non-exonic, $n = 48$ exonic), duplications ($n = 235$ non-exonic, $n = 13$ exonic), multiallelic CNVs ($n = 1{,}278$ non-exonic, $n = 111$ exonic) and STRs ($n = 13{,}085$ non-exonic, $n = 148$ exonic). Number of eQTLs for each variant class at defined length is shown in top panels. Points in bottom panels represent the centers of bins with equal numbers of observations and error bars indicate 95% confidence intervals around the mean (1000 bootstraps). Lines represent linear regressions, with 95% confidence intervals shaded, as calculated on unbinned data. *p* values at the right of each plot indicate the significance of the association between length and absolute effect size (linear regression, t-test) in a model that includes non-mode allele frequency and distance to TSS as covariates. *p* values presented are not adjusted for multiple testing.

We next sought to examine whether eVariant length for SVs and STRs was predictive of absolute eQTL effect size and if lead eQTLs that overlap (exonic) or do not overlap (non-exonic) exons of the eGene displayed similar effects (Fig. 2b, c). We found that lead eVariant length was significantly associated with the absolute effect size for non-exonic deletion, duplication, mCNV, and STR eQTLs independent of variant distance to the transcription start site and allele frequency (Fig. 2b). However, among exonic eQTLs, only those mapping to deletions had a significant correlation between length and effect size with longer deletions having larger effect sizes (Fig. 2c). These data show that longer variants are more likely to be eVariants for both SVs (excluding mCNVs) and STRs and that among eVariants that do not overlap exons, longer variants tend to have stronger effects on expression.

mCNVs and deletions are enriched for associations with multiple eGenes. We next investigated whether SVs from particular classes were more likely to be eVariants or associated with multiple eGenes compared to STRs which comprised 78% of all tested variants and 70% of eQTLs (Fig. 1c). We found that both mCNVs and deletions were more likely to be eVariants for at least one gene relative to STRs (mCNVs: OR = 2.81, $q = 1.57\text{e-}50$, FET; deletions: OR = 1.35, $q = 6.47\text{e-}12$, FET) and were also more likely to be lead eVariants compared to STRs (mCNVs: OR = 1.70, $q = 9.74\text{e-}9$, FET; deletions: OR = 1.32, $q = 4.56\text{e-}7$, FET) (Fig. 3a and Supplementary Fig. 5). Conversely, BNDs were less likely to be eVariants (OR = 0.75, $q = 2.18\text{e-}4$, FET) or lead eVariants (OR = 0.46, $q = 3.63\text{e-}11$, FET) compared to STRs (Fig. 3a). We next examined how often eVariants from each variant class were associated with multiple eGenes and found that, while many of the SV classes were more likely to be associated with multiple eGenes compared to STRs (Fig. 3b), mCNV eVariants and deletion eVariants were associated with two or more genes 61.7% (271/439) and 44.4% (401/902) of the time respectively. Moreover, 31.2% (137/439) of mCNV eVariants and 13.4% (121/902) of deletion eVariants were associated with at least 4 genes compared to only 7.2% (571/7,915) of STR eVariants (Fig. 3b). These results show that mCNV and deletion eVariants are more frequently associated with the expression of multiple genes compared to STRs and other SVs.

Genomic localization of SV and STR eQTLs. We examined how eVariants for each variant class were distributed with respect to genes and promoters by evaluating the distance of eVariants to their eGenes (5′-UTR or TSS) and their overlap with genic



elements. We found that, for all SV classes and STRs, most eQTLs were located near eGenes (<250 kb, Fig. 3c); however, a significantly larger proportion of eQTLs that were mCNVs or duplications were located far from their eGenes (>250 kb) compared to STRs (mCNVs: 40.5%, OR = 2.26, $q = 1.76e\text{-}41$, duplications: 37.1%, OR = 1.96, $q = 3.46e\text{-}5$; FET) suggesting increased distal regulatory activity for these variant types. SNV and indel eQTLs were generally closer to eGenes than all clases of SV and STR eQTLs (Fig. 3c). We next annotated each variant-gene pair tested for whether the variant overlapped an exon, promoter, or intron for the paired gene; overlapped an exon, promoter, or intron for a different gene; or was intergenic (Fig. 3d and Supplementary Fig. 6). Overall, we observed that 23.1% of lead eQTL variants directly overlapped the eGene with 205 overlapping exons (2.9%), 224 overlapping promoters (8.5%), and 1,180 overlapping only introns (17%) in the associated eGene.

Fig. 3 Properties of SV and STR eQTLs. a Percentage of tested variants from each class that are eVariants (eV), left) or lead eVariants (right) in the SV/STR-only eQTL. Asterisks indicate significant enrichment or depletion of variants among eVariants relative to STRs (FET two-sided, BH alpha < 0.05). **b** Left panel is a balloon plot where color indicates number of eVariants and size indicates fraction of eVariants in each bin. Right panel shows average number of eGenes per eVariant with 95% confidence intervals. Red points indicate significantly higher numbers of eGenes per eVariant (Mann–Whitney *U* one-sided, Bonferonni $p < 0.05$) compared to STRs. **c** Distribution of the distance of eQTL (left) and lead eQTL (right) variants to the boundary of their eGenes. Percentages indicate the proportion of eQTLs that were at least 250 kb distal to eGene, red asterisks indicate that eQTLs tended to be localized farther from eGenes as compared to STRs (Mann–Whitney *U* Test one-sided, Bonferonni $p < 0.05$). Distributions for lead SNV and indel eQTLs were not examined for significant differences relative to STRs. **d** Fraction of lead eQTLs that were intergenic or overlapped exons, promoters, or introns of their associated eGene or other genes. **e** Enrichment of lead eQTLs for each class that overlap each genic element compared to variants from all other classes (FET two-sided, BH alpha < 0.05). **f** Distribution of effect sizes for lead eQTLs that overlapped or did not overlap an exon of their eGene. (G,H) Distribution of absolute effect sizes for SV classes, STRs, and SNV/indels for exonic (**g**) and non-exonic (**h**) eQTLs. Vertical dashed lines indicate means. *P* values are derived from comparing the effect size distributions for mCNVs to the distributions for STRs, SVs, SNVs, and indels (Mann–Whitney *U* test one-sided, Bonferonni $p < 0.05$). All *p* values and odds ratios for (**a**) and (**e**) are in Supplementary Figs. S5 and S6. eQTL statistics for SNV and indels are from the joint analysis for all panels.

Interestingly, mCNVs were the only eQTL variant class whose lead variants were enriched for overlapping exonic regions of eGenes compared to all other variant classes (17.7%, OR = 9.16, $q = 4.7e\text{-}26$, Fig. 3d, e). mCNV lead variants were more likely to overlap gene exons even though a substantial number of mCNV eQTLs were also located far from their eGene (Fig. 3c) suggesting that a subset of mCNV eQTLs may be distal regulatory variants and a subset may affect expression by directly altering eGene copy number. Lead mCNVs, duplications, and deletions were also enriched for overlapping exonic regions of other genes besides their associated eGenes compared to other variant classes (mCNVs: OR = 11.64, $q = 1.81e\text{-}57$; duplications: OR = 5.95, $q = 3.29e\text{-}6$; deletions: OR = 2.34, $q = 1.42e\text{-}8$; FET); conversely, STRs were depleted in other gene exons (3.82%, OR = 0.26, $q = 7.7e\text{-}37$, FET, Fig. 3d, e) and eGene exons (1.98%, OR = 0.32, $q = 8.7e\text{-}14$). Overall, lead mCNV eVariants were more likely than other eVariants to overlap eGene exons while mCNVs and duplications had more distal eQTLs than other variant classes.

We next compared the direction and absolute effect sizes of lead eQTLs that overlapped or did not overlap exons of the eGene (exonic and non-exonic eQTLs) from each variant class to determine whether variants that alter gene copy number differ from regulatory region variants. Exonic and non-exonic lead eQTLs mapped to mCNVs and exonic lead eQTLs mapped to duplications had primarily positive associations with gene expression while exonic lead eQTLs mapped to deletions had mostly negative effects (Fig. 3f). Lead eQTLs mapped to all other variant classes, including SNVs and indels, had bimodal effect size distributions. Comparing the absolute effect sizes of lead eQTLs mapped to each variant class, we found that mCNV lead eQTLs also had significantly larger effect sizes in both exonic (Fig. 3g) and non-exonic (Fig. 3h) contexts compared to lead variants from other SV classes, STRs, SNVs, and indels (Fig. 3g, h). These data show that mCNV eQTLs are unique in that they tend to exert strong positive effects on gene expression, especially mCNV eQTLs that overlap exons which are almost always positively correlated with gene expression.

eVariant type is associated with eGene type and constraint. We next investigated whether the eGene type, such as protein coding or pseudogene, was associated with the variant class of lead variants. We annotated all eGenes with Gencode gene types and calculated whether a given variant class was more or less likely to be a lead variant for eGenes of a particular gene type (Fig. 4a, b and Supplementary Fig. 7a). Notably, a lower proportion of mCNV eGenes were protein coding (OR = 0.23, $q = 1.52e\text{-}28$, FET) and a higher proportion were pseudogenes (OR = 8.57, $q = 4.72e\text{-}34$, FET) or lincRNAs (OR = 2.5, $q = 2.29e\text{-}4$, FET) compared to other variant classes. Duplication and deletion eGenes followed the same trends but did not reach significance. However, STR eQTLs had the opposite pattern and were enriched for protein coding genes (OR = 1.83, $q = 8.38e\text{-}16$, FET) and depleted for pseudogenes (OR = 0.36, $q = 1.19e\text{-}16$, FET) and lincRNAs (OR = 0.62, $q = 8.68e\text{-}4$, FET). We looked at the effect sizes of associations among different gene types and found that lead eQTLs for protein coding eGenes tended to have lower effect sizes compared to lead eQTLs for genes that are not protein coding (Fig. 4c) which is consistent with non-protein coding genes being more tolerant of disruption[18]. Furthermore, the observation of higher effect sizes among mCNV eQTLs and their increased likelihood to overlap exons of their eGenes may be partly explained by their association with fewer protein coding genes, while the opposite properties were observed among lead eQTLs attributed to STRs, which were less frequently exonic.

Given the differences in eGene types between different variant classes, we hypothesized that eGenes might be under different levels of evolutionary constraint compared to non-eGenes. To test this, we obtained pLI scores (probability that a gene is intolerant to loss of one allele), pRec scores (probability that a gene is intolerant to loss of both alleles), and pNull scores (probability that a gene is tolerant of loss of both copies of gene) from ExAC for 13,012 of the 16,018 genes that were tested for eQTLs[18,19]. We examined the distributions of these constraint scores for eGenes with lead eVariants from each variant class and observed that eGenes were skewed towards low (<0.9) pLI and pNull scores but more evenly distributed between low and high pRec scores (Fig. 4d and Supplementary Fig. 7b). We found that across variant classes eGenes were significantly depleted for having high pLI scores (> 0.9) and generally enriched to have high pRec and pNull scores compared to non-eGenes (Fig. 4e). This result demonstrates that genes that are intolerant to mutation are less frequently eGenes while genes tolerant of heterozygous or null alleles are more likely to be eGenes, consistent with SNV eQTLs[20]. Examining this trend among variant classes, mCNVs had the lowest proportion of high pLI eGenes suggesting that mCNV protein coding eGenes are less constrained. Interestingly, eGenes mapped to deletions were most likely to be high pNull suggesting that, due to their severe negative effects on expression, deletion eVariants are under greater selection to affect dispensible genes. Given that some eGenes were classified as under high levels of constraint (pLI > 0.9), we sought to understand whether these genes are also sensitive to high levels of expression modulation. We compared the absolute effect size of lead QTLs to the pLI score of the eGene and found a strong and significant negative correlation between effect size and pLI (Fig. 4f) consistent with a previous report that there is less variation in the expression of

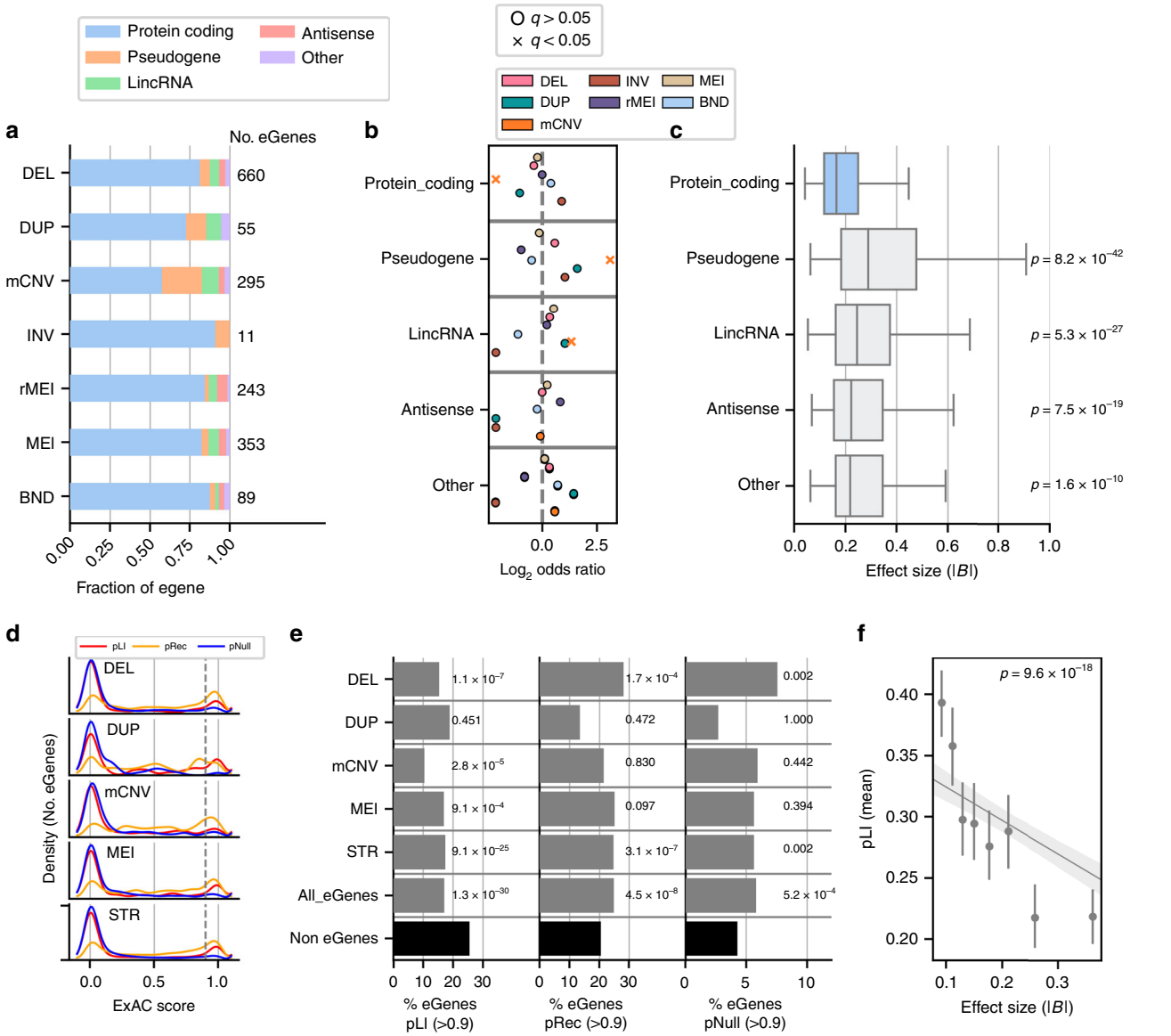
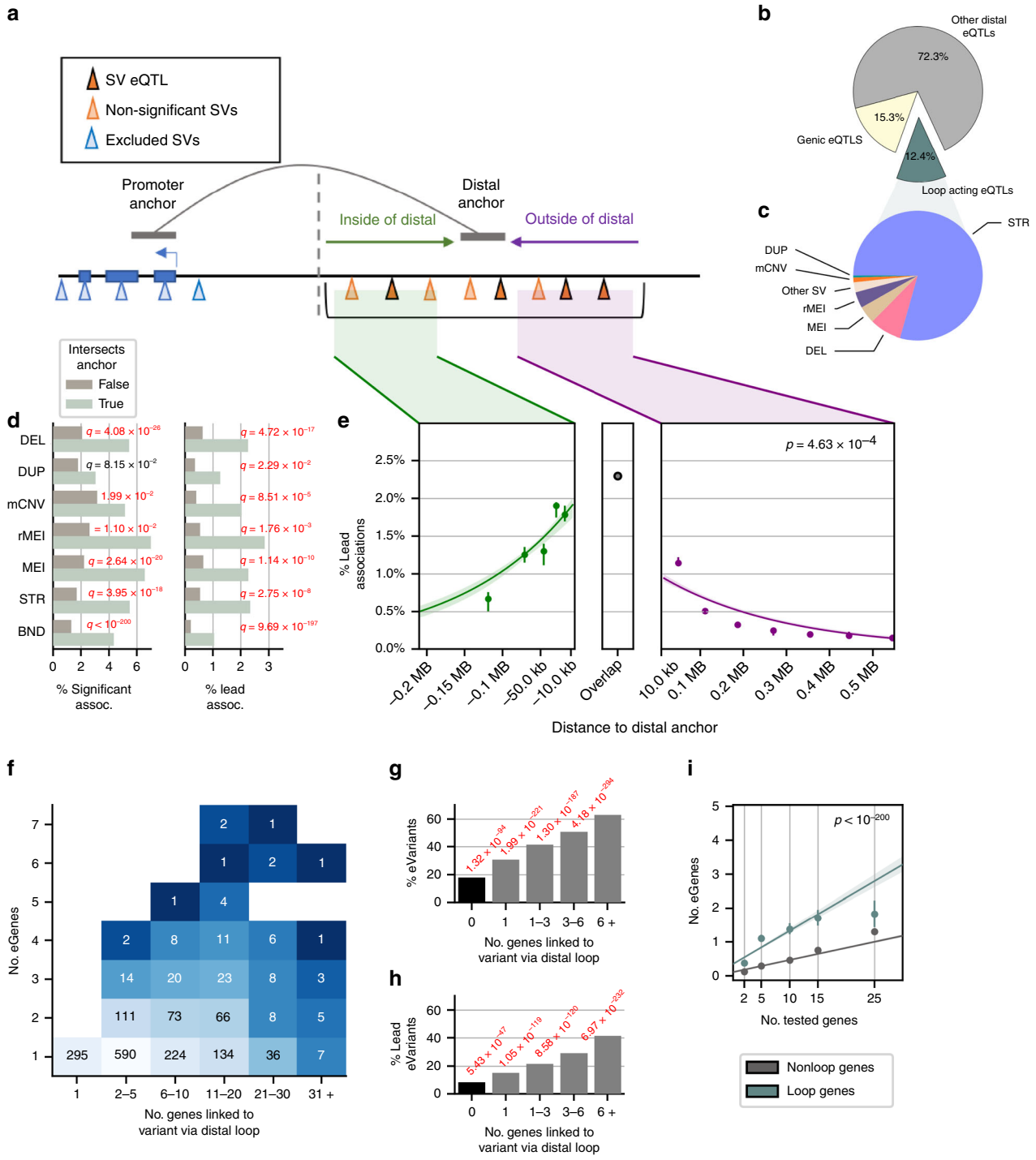


Fig. 4 Properties of eGenes associated with different variant classes. a Fraction of eGenes of each Gencode gene subtype mapped to lead variants of each class for the SV/STR-only eQTL. **b** Enrichment of the proportion of eGenes of each subtype mapped to a variant class compared with the proportion of other eGenes falling into that subtype. Significant associations (FET, BH FDR < 0.05) are indicated with x symbols. **c** Absolute effect size of associations for genes of each subtype among lead eQTLs. *P* values indicate significance of Mann–Whitney *U* test (one-sided) for difference in the effect size distributions of each category compared to protein coding genes (Bonferroni). For boxplots the minimum box edge indicates the first quartile while the maximum box edge indicates the third quartile and the center line indicates the median value. Whiskers of the box plot are drawn at the maximum point (upper whisker) or minimum point (lower whisker) that is within 1.5 times the interquartile range (quartile three—quartile one). **d** Distribution of ExAC scores for intolerance to loss-of-function variants in a single allele (pLI, red), intolerance to loss-of-function variants in both alleles (pRec, orange), and tolerance to loss-of-function variants in both alleles (pNull, blue) for 5,675 eGenes. **e** The percentage of eGenes (grey bars) mapped to lead variants that had high (>0.9) pLI (left), pRec (center), or pNull scores (right). Black bars show percentages for 7,337 non-eGenes from the SV/STR-only analysis. *P* values indicate the significance of the difference between the proportion of high score eGenes and high score non-eGenes for each group individually (FET, BH FDR < 0.05, within each probability score). **f** Absolute effect size versus pLI score for all eGenes with a pLI score ($n = 5{,}675$). Points are equally-sized bins and error bars show 95% confidence intervals ($n = 1000$ bootstraps) around the mean pLI. Line is a model predicting pLI by eQTL effect size after regressing out variant class and mean $\log_{10}$(TPM) expression level of the gene among expressed samples. *p* value is for the eQTL effect size term (*t* test).

highly constrained genes[20]. Taken together, these results suggest that while eGenes tend to be less constrained than other genes, the eGenes with mCNV or deletion lead eVariants are particularly tolerant of loss-of-function variation.

Multi-eGene eQTLs colocalize with distal chromatin loop anchors. Since chromatin looping has been shown to play a key role in the regulation of genes by positioning regulatory regions near gene promoters[21–24], we sought to determine whether distal eVariants are located near the promoters of their eGenes in three-dimensional space via chromatin looping. We obtained chromatin loop calls from iPSC promoter capture Hi-C data[25] that define promoter loops between gene promoters (promoter anchors) and distal sequences (distal anchors, Fig. 5a). We observed that 13,575 of the 16,018 genes tested for eQTLs had at least one promoter loop and that 29.2% of SVs and 30% of STRs



tested for eQTLs overlapped a distal anchor. Interestingly, mCNVs were significantly less likely to overlap a distal anchor than other variant classes with only 13.5% of mCNVs overlapping (OR = 0.37, $p < 1e\text{-}25$, FET) likely due to the difficulty of identifying loop anchors near segmental duplications, which frequently overlap mCNVs (Supplementary Fig. 8). Among the 13,575 genes that had at least one loop and were tested for SV/STR eQTLs, we identified 177,571 (31.8%) variant-gene pairs for which the variant was: (1) closer to the distal anchor than to the promoter anchor; (2) at least 50 kb away from the gene body; (3) did not overlap the exon of the tested gene; and (4) was a maximum of 200 kb from a distal anchor, which we defined as distal variant-gene pairs (Fig. 5a). Among these distal variant-gene pairs, we observed 1,598 eGenes (22% of all eGenes) with at least one eVariant located in the distal anchor of a loop to their promoter, 963 (12.4% of all eGenes) of which were mapped to a lead eVariant in the distal anchor; 82% (788/963) of these lead variants were STRs (Fig. 5b, c). Within each variant class, distal variant-gene pairs that overlapped the distal anchor of a loop to the promoter of the tested gene were highly enriched to be eQTLs or lead eQTLs (OR = 3.5–5.4; $q = 0.022\text{-}9.7e\text{-}197$; FET, Fig. 5d). The fraction of eQTLs from the joint analysis with SNVs/indels that were loop-acting was nearly the same as the SV/STR only analysis with 12.5% of eGenes having at least one loop-acting eVariant and 10.4% whose lead variant was loop-acting (Supplementary Figs. 5b and 9a,b). The majority of loop-acting eQTLs in the joint

Fig. 5 Localization of eQTLs near chromatin loops. a Diagram showing localization of SVs and STRs at loop anchors. eVariants closer to the distal anchor (right of grey dotted line) than the promoter anchor were considered loop-acting eQTLs. **b** Proportion of eQTLs that were genic (yellow), overlapping or close to distal anchors (green), or distal acting by some other mechanism (grey). **c** Distal loop-acting eQTLs ($n = 2,327$ eQTLs for 1,598 eGenes) per SV class. **d** Percentage of eVariant-eGene pairs where the eVariant (left) or lead eVariant (right) overlaps or does not overlap the distal anchor. p values derived by comparing proportions for each class (FET two sided, Benjamini–Hochberg). **e** Fraction of tested distal variant-gene pairs (**a**) that were lead eQTLs versus their distance to the distal anchor. Points represent the means of equally-sized bins; errors bars 95% confidence intervals. Curves are logistic regressions using distance to the loop anchor to predict whether the variant-gene pair was a lead association. Regressions were computed separately for variant-gene pairs inside the loop (left, $n = 47,831$) or outside the loop (right, $n = 294,796$). Center panel shows fraction of variant-gene pairs that overlapped distal anchors and were lead eQTLs ($n = 41,794$). **f** Number of eVariants connected to gene promoters through chromatin loops (x-axis) and number of these connected genes that are eGenes (y-axis). **g**,**h** Percentage of tested variants that were eVariants (**g**) or lead eVariants (**h**) stratified by number of genes the variant was linked to through a distal anchor. p values for each bar were derived by comparing the proportion of tested variants that were eVariants and linked to genes to the proportion of variants that were eVariants and not linked to genes (first bar). **i** Number of eGenes versus number of tested genes per eVariant stratified by whether the genes are linked by loops to the eVariant (blue) or not linked by loops (grey). Lines indicate relationship between number of eGenes per eVariant and number of genes tested for genes that were or were not linked by loops. p value is for loop/nonloop term (t-test).

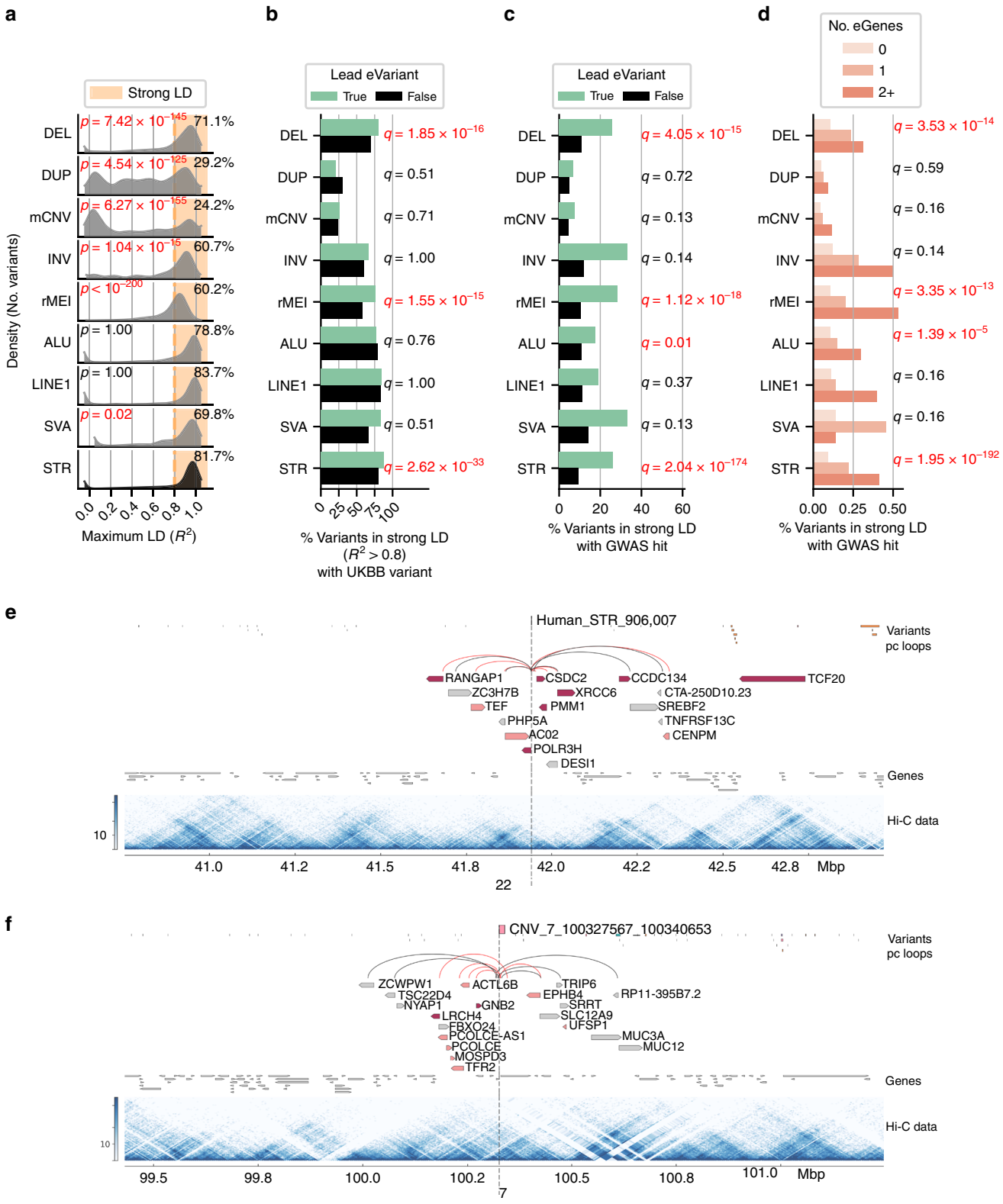
analysis are SNVs or indels reflecting the large number of SNV/indel eQTLs (Supplementary Fig. 9c,d). The fraction of all eQTLs versus lead eQTLs that were loop-acting was similar for SNVs and indels but differed for SV classes in the joint analysis such as MEI, rMEI, and duplication lead eQTLs which generally did not intersect loop anchors (Supplementary Fig. 9e,f). Overall, these results indicate that many eQTLs include variants that overlap distal chromatin loop anchors, that variants that overlap distal anchors are more likely to be eQTLs and lead variants, and that there are differences in the fraction of eQTLs and lead eQTLs that are loop acting between variant classes.

We next hypothesized that if variants near loops impact gene expression, the location of variants relative to the distal anchor should be related to the chance of that variant being an eQTL. We tested if the distance of a variant to the distal anchor or the variant's position inside or outside of the loop was predictive of whether the variant is an eVariant using a logistic model (Fig. 5a). For this model, we subsetted the variant-gene pairs to those whose variants were at least 100 kb away from the nearest TSS or 3′-UTR and a maximum of 200 kb from the nearest distal anchor to ensure we were examining interactions around the distal anchor. We observed that variants closer to the distal loop anchor were significantly more likely to be lead eQTLs ($p = 0.0005$) and that distal variant-gene pairs with a variant inside the loop were more likely to be lead eQTLs than those with variants outside the loop (Fig. 5e, OR = 1.5, $p = 2.1e\text{-}8$, FET). This suggests that variants near distal loop anchors are more likely to affect expression of the looped gene and that variants that do not directly overlap the loop anchor can still affect gene expression, potentially through changes in regulatory elements or loop structure.

Given that variants overlapping distal anchors are more likely to be eQTLs, we hypothesized that variants that are looped to multiple gene promoters may be associated with the expression of many of their looped targets. To examine this, we tested whether the number of looped genes to an eVariant was associated with the number of eGenes for that eVariant. We observed that variants overlapping distal anchors that were connected to multiple genes via chromatin loops tended to be multi-gene eVariants (Fig. 5f). We also found that the likelihood of a variant being an eQTL or lead eQTL increased significantly as the number of genes that the variant was looped to increased (Fig. 5g, h). For example, 41% of variants linked to 6 or more genes by a distal loop anchor were lead eVariants as compared to only 8.5% of distal variants that were not linked to an eGene by loop anchor (OR = 7.62, $q = 6.97e\text{-}232$, FET). One possible explanation of these results is that variants looping to multiple genes are located in gene-dense regions and are therefore tested for more eGenes. To address this, we compared, for each variant, the number of genes that were tested and the number that were identified as eGenes, stratified by whether the genes were connected by loops or not connected by loops, and found variants tended to have more eGenes among looped genes than genes not connected by loops (Fig. 5i, $p < 1e\text{-}200$, t-test, Methods). This trend was consistent across SV classes (Supplementary Fig. 10). These results suggest that variants located in high complexity loop anchors are more likely to be multi-gene eQTLs than variants simply located near many genes.

LD tagging and GWAS associations differ between variant classes. SVs and STRs are typically not assessed in GWAS, so the contribution of classes of non-SNV variation to complex traits and diseases is currently unclear. To examine the extent by which the different SV classes and STRs have been assayed by proxy in GWAS, we calculated LD between i2QTL variants and SNVs present in the UK Biobank (UKBB, ± 50 kb of each SV and STR). We observed strong LD ($R^2 > 0.8$) with UKBB SNVs for a large proportion of STRs (81.7%), ALU and LINE1 elements (79% and 83.7%), and deletions (71.1%), but a markedly lower proportion of duplications (29.2%) and mCNVs (24.2%) were in strong LD with a nearby variant (Fig. 6a). We stratified our analysis of duplications and mCNVs by whether they overlapped a segmental duplication (SD) and found that those that overlapped SDs were less likely to be in strong LD with UKBB variants (18.9% duplications and 16.8% of mCNVs $R^2 > 0.8$) than those that did not overlap SDs (33.3% duplications and 65.7% mCNVs $R^2 > 0.8$, Supplementary Fig. 11), indicating that poor tagging for these classes may in part be due to the presence of repetetive sequences. We also found that only 59% of multi-allelic STRs with four or more alleles were well-tagged by UKBB SNVs. These results suggest that the duplications and mCNVs are generally not assayed by proxy in GWAS, especially when located in segmental duplications.

Next, we investigated the extent to which SVs and STRs associated with gene expression were tagged by nearby UKBB SNVs ($R^2 > 0.8$) or linked to diseases and traits via GWAS. We observed that deletions, rMEI, and STR lead eVariants were more likely to be in strong LD with UKBB variants compared to non-lead eVariants of the same class (Fig. 6b and Supplementary Fig. 12a). While >65% of lead eVariants for most SV classes were in strong LD with any nearby UKBB variant, only 26% of mCNV and 21% of duplication lead eVariants were strongly tagged, further supporting that most mCNVs and duplications are not assayed by proxy in GWAS. We then examined how often variants in strong LD with UKBB variants were significantly associated with at least one GWAS trait ($p < 5e\text{-}8$) and found that



11.4% of common STRs and SVs were by proxy associated with at least one of 701 UKBB traits (Fig. 6c and Supplementary Fig. 12b). Lead eVariants were more likely to be in strong LD with significant GWAS variants across all classes; however, enrichment was only significant in STRs, deletions, rMEIs, and ALU elements likely because other classes had too few variants to reach significance (Fig. 6c). As a whole, SVs and STRs were respectively linked to 425 and 625 of 701 distinct GWAS traits, with 412 traits linked to variants of both types. Traits linked to eSVs and eSTRs included diseases such as type 1 diabetes, multiple sclerosis, arthritis, cancers, and heart disease, as well as quantitative traits such as height, body mass index and white blood cell count (Supplementary Data 4).

We hypothesized that multi-eGene eVariants may have greater impact on common traits and examined the LD of these eQTLs with GWAS variants. Interestingly, we found that multi-eGene

Fig. 6 Associations between SVs, STRs and GWAS. a Distribution of maximum LD score per i2QTL variant with UKBB variants within 50 kb for each variant type. *p* values calculated for the LD distribution of each SV class relative to STRs (Mann–-Whitney *U*, Bonferroni). **b** Fraction of variants of each class that are tagged by a UKBB variant ($R^2 > 0.8$) for lead eVariants (green) versus all other variants in that class (black). *Q* values indicate enrichment of lead eVariants to be in LD with a UKBB variant versus all other variants tested in the eQTL in the class (FET two-sided, Benjamini–Hochberg). **c** Fraction of variants of each class that are tagged by a UKBB variant that is associated with at least one trait in the UKBB ($p < 5e-8$). *q* values indicate enrichment of lead eVariants to be in strong LD with a UKBB variant that is associated with at least one trait versus all other variants tested in the eQTL in the class (FET two-sided, Benjamini–Hochberg). **d** Percentage of variants in LD ($R^2 > 0.8$) with a variant significantly linked to at least one GWAS trait when significantly associated with 0, 1, or 2 eGenes or more. To compute annotated *q* values, we utilized all variants tested in the SV/STR-only eQTL, and for each variant class we performed logistic regression to determine whether the number of eGenes for a variant was associated with whether the variant was in strong LD with a significant GWAS variant (z-test, Benjamini–Hochberg). **e** Example multi-gene eSTR on chromosome 22 with nine unique eGenes (pink/red) including four genes that the STR loops to. Genes for which the variant is a lead variant are colored red. iPSC Hi-C data is visualized as a heatmap of interaction frequencies. The variant is located between two chromatin subdomains that span ~100 kb on the left side of the variant and ~25 kb on the right side of the variant[22]. **f** Example of an mCNV on chromosome 7 that is a multi-gene eQTL associated with seven unique eGenes by looping. Exact *p* values and odds ratios for (**b**) and (**c**) are in Supplemental Fig. 12.

eVariants were highly enriched to be in strong LD (>0.8 R^2) with GWAS variants (Fig. 6d). ~40% of eVariants associated with two or more eGenes were in strong LD with a GWAS variant while only 20% of eVariants associated with one eGene were in strong LD with a GWAS variant. We also observed that ~70% of eVariants associated with three or more eGenes and localized near the eGenes' promoters via chromatin loops were in strong LD with GWAS traits. For example, a 20 bp eSTR was associated with nine eGenes (seven lead) connected via distal loops (Fig. 6e) and was in strong LD with a UKBB variant linked to 19 distinct traits including asthma and body fat percentage; two of the genes associated with this variant (*TCF20* and *POLR3H*) have also been previously linked to autism[26,27]. We observed that this variant appears to overlap a chromatin subdomain boundary visible in Hi-C data from iPSCORE[22] which is notable given that disease-causing STRs, such as the causal variant for Fragile X syndrome, have been reported to localize to subdomain boundaries[28]. Additionally, we found a 13 kb deletion on chromosome 7 linked to five eGenes via looping that was also linked to 14 traits (Fig. 6f). These data suggest that multi-gene associations mediated by chromatin looping are frequently linked to common traits, reflecting the impact of modulating the expression of several genes.

Discussion

We identified SVs and STRs associated with gene expression using a comprehensive SV/STR variant call set (Companion Paper) and RNA-sequencing from 398 iPSC samples. We discovered several genomic properties that were associated with gene expression. For deletions, duplications, and STRs, we found increased length was associated with a higher likelihood of being an eVariant and increased effect size for noncoding eVariants. We investigated the properties of eGenes associated with SVs and STRs and showed that they were less constrained than non-eGenes and that highly constrained protein coding eGenes tended to have smaller effect sizes. Distal SV and STR eVariants were enriched for being located near the promoters of their eGenes in three-dimensional space via chromatin looping. We have previously shown that loop detection may be affected by the presence of SVs[22], and therefore we have likely underestimated the proportion of distal SV eVariants that mediate their effects on gene expression via chromatin loops. We also show that SV and STR eVariants near high complexity loop anchors with multiple promoter-distal regulatory element interactions are more likely to affect the expression of several genes. These results demonstrate that chromatin looping may be an important mechanism by which SVs and STRs regulate gene expression. Our study presents one of the largest sets of SVs and STRs associated with gene expression and reveals important general genomic properties of both SV and STR eVariants and their corresponding eGenes.

While previous studies have cataloged SV eQTLs, the extent to which different SV classes and STRs differentially impact gene expression has not been thoroughly investigated[3,6,7,22]. We identified substantial differences between the different SV classes in their genomic locations relative to eGenes; their likelihood of being associated with multiple eGenes; the types of associated eGene (i.e., coding, noncoding, evolutionary constraint); their effect sizes; the extent of linkage disequilibrium with tagging single nucleotide polymorphisms (SNPs) used in GWAS; and their likelihood of being associated with GWAS traits. Interestingly, mCNV eQTLs differed in several respects compared to eQTLs for other variant classes. mCNVs eQTLs were more likely to be associated with the expression of multiple genes, had larger effect sizes, tended to affect noncoding genes, and were more likely to overlap the corresponding eGene or be located far from the eGene. mCNV eQTLS that overlapped exons were also highly enriched for positive associations between copy number and expression relative to other variant classes. Unlike other SV classes, the length of mCNVs was not strongly associated with the probability of being an eQTL. The differences in likelihood of being an eQTL, location, effect size, and types of eGenes for mCNVs are likely related; for instance, less constrained genes tend to have larger eQTL effect sizes, mCNVs tend to be eQTLs for less constrained genes, and mCNV eQTLs tend to have larger effect sizes. Our results indicate that a previous finding that mCNVs were enriched among predicted causal eQTL variants might be driven by the fact that mCNVs often overlap genes and likely cause eQTLs[1]. We also observed that deletion eQTLs were more likely to be associated with the expression of multiple genes but tended to have smaller effects on gene expression, not overlap genes, and affect less constrained genes. These observations are consistent with gene deletions and subsequent loss of expression having strong deleterious effects. Future studies may focus on whether the differences in eQTLs between variant classes are driven by selective pressures, genomic property differences between the SV classes, or some combination thereof.

The extent to which SVs and STRs contribute to variation in complex traits is not fully known because prior GWAS have generally not assessed SVs and STRs. We used our comprehensive SV/STR call set to estimate how well these variants are tagged by GWAS SNPs and whether they are associated with 701 traits from the UK Biobank. We found that only 26% of mCNV and 21% of duplication lead eVariants were tagged ($R^2 > 0.8$) by a SNP in the UK Biobank, likely due in part to these variants being located in or near segmental duplications, indicating that these variants are generally missed in GWAS studies based on genotyping arrays. Multiallelic STRs are also not tagged well by SNPs and are likely

not well-studied by current GWAS. We observed that 11.4% of common SVs and STRs are in strong LD with at least one significant GWAS SNP in the UK Biobank and that lead eSVs were more likely to be associated with traits compared to non-lead eSVs. We also identified a set of high-impact SVs and STRs associated with the expression of multiple genes and localized near the promoters of these genes via chromatin loops which are also highly enriched for GWAS associations. These high-impact variants that are associated with several seemingly unrelated GWAS traits may underly some of the observed pleiotropy in contempary genetic studies[29] and indicate that future fine-mapping efforts will greatly benefit from including SVs and STRs.

Our study demonstrates that SVs and STRs play an important role in the regulation of gene expression and that eQTLs for different classes of SVs and STRs differ in their effect sizes, genomic locations, and the types of eGenes they impact. We have also demonstrated that high-impact SVs and STRs, i.e., those associated with the expression of multiple genes via chromatin looping, are associated with a wide range of human traits. We anticipate that these properties of SV and STR eQTLs will be useful for identifying causal variants underlying eQTLs and delineating mechanisms by which structural variation can impact gene expression. The collection of eQTLs identified here, along with the catalog of high-quality SVs and STRs described in a companion paper[4], provide a powerful resource for future studies examining how these variants regulate gene expression and contribute to variation in complex traits.

Methods

Variant calls. Single nucleotide variant (SNV), insertion/deletion (indel), structural variant (SV) and short tandem repeat (STR) variant calls for iPSCORE and HipSci samples were discovered and rigorously analyzed in a companion paper (dbGaP: phs001325[4]). The iPSCORE collection was approved by the Institutional Review Board of the University of California at San Diego (Project #110776ZF). We have complied with all relevant ethical regulations for work with human participants and obtained informed consent. We used five variants callers to identify SVs and STRs. We used the SpeedSeq (SS) SV pipeline[30] that combines LUMPY[31] read-pair evidence with read depth support from CNVnator[32]. We also used the Genome STRiP CNVDiscovery pipeline (GS) and Genome STRiP LCNVDiscovery pipeline (GS LCNV)[33] that detect SVs based on read depth evidence. We used MELT[34] for mobile element insertion discovery and HipSTR[6] to identify and genotype short tandem repeats. LUMPY, GS, and GS LCNV each identified biallelic deletions (DEL), biallelic duplications (DUP), and multi-allelic copy number variants (mCNVs). mCNVs are defined as variants that have at least 3 predicted alleles. LUMPY identified inversions (INV) and generic breakends (BND) that can include deletions and duplications that lack read-depth evidence, balanced rearrangements (INVs), MEIs, or other uncategorized breakpoints. As part of the SpeedSeq pipeline we also identified reference mobile elements (rMEIs). For non-reference mobile element insertions we used MELT to identify Alu element insertions (ALU), LINE-1 element insertions (LINE1), and SINE-R/VNTR/Alu element insertions (SVA). HipSTR identifies short tandem repeats (STRs) where at least one individual differed in STR length compared to the reference. We considered copy number variants (CNVs) to include deletions, duplications, and mCNVs. We considered MEI to encompass non-reference mobile element insertion ascertained by MELT, including ALU, LINE1, and SVA elements.

RNA-Seq quality control and processing. As part of the i2QTL Consortium, we have collected a set of RNA sequencing (RNA-seq) samples from 1,367 human induced pluripotent stem cell (iPSC) lines derived from 948 unique donors from five studies: iPSCORE[7,17], HipSci[16,35], Banovich et al.[36], GENESiPS[37], and PhLiPS[38]. Sample processing and quality control was performed across all samples as described below, but the eQTL analysis presented here uses a subset of the total data set corresponding to 388 unique donors from iPSCORE and HipSci that have variant calls from deep whole genome sequencing[4]. The RNA-seq data were obtained from: (1) 210 iPSCORE RNA-seq samples from dbGaP (phs000924); (2) 288 HipSci cell lines (from 188 individuals) from the ENA project ERP007111 and several EGA projects (Supplementary Table 1); (3) Banovich et al.[36] (SRA: SRP093633, http://eqtl.uchicago.edu/yri_ipsc/); (4) GENESiPS (SRA—SRP072417, dbGaP: phs001139.v1.p1); (5) the PhLiPS projects (dbGaP: phs001341.v1.p1.). Data was available from these sources as either FASTQ, BAM or CRAM files. To ensure uniformity in processing, CRAM and BAM files were converted to FASTQ files. The reads in the FASTQ files were then trimmed to remove adapters and low quality bases (using Trim Galore!, http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), followed by read alignment using STAR (version: 020201)[39] with the two-pass alignment mode and default parameters as proposed by ENCODE (c.f. STAR manual). All alignments were relative to the GRCh37 reference genome, using Ensembl 75[40] for any of the necessary genome annotations. Gene-level RNA expression was quantified from the STAR alignments using featureCounts (v1.6.0)[41], which was applied to the primary alignments using the -B and -C options in stranded mode when applicable. In case multiple RNA-seq runs per iPSC-line were generated these were summed to one set of gene-counts per iPSC line.

After feature quantification high-quality RNA-seq samples were identified by applying filters on both Picard (https://broadinstitute.github.io/picard/) and VerifyBamID (http://csg.sph.umich.edu/kang/verifyBamID/) quality measures as well as gene expression levels. We defined high-quality samples as those with > 15 million reads, > 30% coding bases, > 65% coding mRNA bases, a duplication rate lower than 75%, Median 5′ bias below 0.4, a 3′ bias below 4, a 5′–3′ bias between 0.2 and 2, a median coefficient of variation of coverage of the 1000 most expressed genes below 0.8, and a free-mix value below 0.05.

Subsequently, gene expression values were normalized across lines that passed quality control. For this we derived edgeR[42,43] corrected transcript per million gene-level quantifications per iPSC line from the feature count information. After this normalization we removed samples that had low expression correlation (<0.6) with the average iPSC expression profile across our study, as measured per chromosome. For the purpose of the eQTL analyses presented here, we used gene expression estimates for 288 HipSci cell lines (188 individuals) and 210 iPSCORE cell lines (210 individuals) that had corresponding deep whole genome sequencing data (WGS) that allowed for comprehensive characterization of SNVs, indels, SVs, and STRs[4]. This joint data set of variant calls and iPSC gene expression data for 398 individuals is referred to as the i2QTL data set in this manuscript. Additional detail on these methods is described in a paper describing eQTLs on the full data set[44].

eQTL analysis. To find eQTLs we tested for associations between variants within a cis-region spanning 1MB up- and downstream of the gene body and 16,018 robustly expressed autosomal genes (expressed in >20% of samples at an average TPM > 0.5 among samples that expressed the gene) in 398 the HipSci and iPSCORE donors (Supplementary Data 1). We performed association tests using a linear mixed model (LMM), accounting for population structure and sample repeat structure as random effects (using a kinship matrix estimated using PLINK[45]). All models were fit using LIMIX[46] (https://limix.readthedocs.io/).

Before QTL testing the gene expression-levels were log transformed and standardized. Significance was tested using a likelihood ratio test. To adjust for global differences in expression across samples, we included the first 50 PEER factors (calculated across all 1,367 lines using log transformed expression values) as covariates in the model. In order to adjust for multiple testing, we used an approximate permutation scheme, analogous to the approach proposed in Ongen et al.[47]. Briefly, for each gene, we ran LIMIX on 1,000 permutations of the genotypes while keeping covariates, kinship, and expression values fixed. We then adjusted for multiple testing using this empirical null distribution. To control for multiple testing across genes, we used Storey's q values[48]. Genes with significant eQTLs were reported at an FDR < 5%.

eQTL input variants and post-processing. Because there are differences in types of SVs (e.g., copy number variants, mobile element insertions) and the output of SV variant callers, genotypes were preprocessed before use in the eQTL analysis. Since some STRs are highly multi-allelic, we used the difference in the number of base pairs with respect to the reference (expansion or contraction), as computed from the sum of the GB format tag in the HipSTR VCF file, as genotypes for eQTL analysis[49]. Genome STRiP CNVDiscovery and LCNVDiscovery[33] variants were encoded with integer diploid copy numbers (CN). SpeedSeq[30,31] variants were encoded using their allele balance (AB) fractions at each genotype, which ranges from 0 to 1 based on the amount of evidence for the variant at a site, for greater sensitivity and consistency with Genome STRiP variants, which use a continuous copy number. Finally, for MEIs identified by MELT[34], we used traditional genotypes (0/0, 0/1, 1/1) outputted by the software, as these SVs are expected to be largely biallelic and there is no continuous genotype outputs available. Before performing the eQTL analysis, genotypes for all SV callers (excluding MELT) were rank normalized and converted to a 0-2 scale. For MELT variants, reference, heterozygous, and homozygous alternate genotypes were converted to 0, 1, and 2 respectively. Missing genotypes from all variant callers were filled with the mean dosage among non-missing samples prior to the eQTL. GATK[50] SNV and indel genotypes were processed in the same way as MELT variants, converting them to 0, 1 and 2.

For the eQTL analysis, we utilized 5,834,257 SNVs, 1,520,762 indels that were present at a minor allele frequency of at least 5% and 9,313 SVs and 33,608 STRs that were present at a non-mode allele frequency of at least 5% among the 398 i2QTL donors with RNA-seq, passed QC, and were within 1MB of at least one of 16,018 expressed genes (see eQTL analysis). Notably, non-mode allele frequency was used for SVs and STRs in order to account for multi-allelic variants. For STRs, the non-mode allele frequency is computed from the difference in length of a genotype from the reference, as detailed in Gymrek et al.[49]. The structural variant call set includes variants generated from the same caller or different callers that

pass QC but may be redundant (overlapping or highly correlated)[4]. In a companion paper[4], we identified these redundant clusters and selected high-quality variants to create a non-redundant set of variants; here we chose to include all variants that passed quality control filters in the eQTL analysis including those that were marked as part of a redundancy cluster in order to maximize the chances of SV associations. Additionally, STRs were required to have a 99% call rate in both iPSCORE and HipSci samples in order to be included in the eQTL to prevent batch effects from affecting eQTLs[4]. To compute the number of unique variants in downstream analyses, variants were annotated with the redundancy clusters they belonged to[4] and variants in the same cluster were considered as a single variant. We thus tested 9,313 non-redundant SVs that were in *cis* windows of expressed genes. Because variants could be associated with multiple eGenes, we considered eQTLs to be an SV/eGene pair. We performed two independent eQTL analyses: (1) using STRs, SVs, small indels and SNVs (joint eQTL analysis) and (2) using only STRs and SVs (SV/STR-only eQTL analysis; Fig. 1a).

Association of variant length with likelihood and strength of eQTLs. To test whether longer variants were more likely to be eQTLs we restricted our analysis to tested variants from each variant class that was highly polymorphic in length (spanning orders of magnitude within variant class): duplications, deletions, mCNVs, and STRs. For variants of each of these classes, we computed the fraction of variants that were eVariants or lead eVariants that were longer than a given length threshold and calculated an enrichment p value by comparing the proportion of variants that were eVariants or lead eVariants among variants longer than the threshold to the proportion among variants smaller than the threshold (Fisher's Exact Test, Benjamini–Hochberg (BH)). To compare the association of length of variants with effect size, we utilized all significant eQTLs from each of the aforementioned variant classes and fit a logistic regression model comparing the absolute effect size of these associations with the length of the associated variant using the distance to the nearest TSS of the eGene and the non-mode allele frequency of the variant as covariates. p values from the regression were estimated with the Wald Test and then corrected using the Bonferroni correction.

Properties of SV-QTLs among different variant classes. To determine which SV classes were more likely to be associated with eGenes, we compared the proportion of variants that were eVariants for a given variant class to the proportion that were eVariants for STRs (Fisher's Exact Test, FDR < 5%, BH). To study the localization of eSVs with respect to eGenes, we used Gencode v19[51] annotations to measure distance to the nearest transcription start site for tested genes, as well as categorize variants based on their overlap of introns, exons, and promoters of tested genes, or elements of other genes. A variant was considered to overlap a particular genomic if feature if it overlapped by at least one base pair and each variant was categorized hierarchically into one of the following 7 categories, in order of precedence: 1) exonic eGene, 2) promoter eGene 3) intronic eGene 4) exonic other 5) promoter other, 6) intronic other 7) intergenic (overlapping none of the features).

eGenes properties and constraint. Gene types were annotated using Gencode v19 data for all expressed genes. We then performed enrichment analyses comparing the proportion of eGenes of a specific type mapped to each class to the proportion among all other variant classes (Fisher's Exact Test, BH). To compare the effect sizes of associations with each gene type, we compared the distribution of effect sizes for lead associations for protein coding eGenes to those of pseudogenes, lincRNA, antisense and all other genes (Mann–Whitney U test, BH). To investigate the constraint of eGenes, we obtained ExAC (v0.3.1)[20] pLI, pNull, and pRec estimates for 13,012 expressed genes and restricted our analyses to lead associations with these eGenes. We then compared the proportion of eGenes with high (> 0.9) ExAC scores mapped to either deletions, duplications, mCNV, MEI (LINE1, SVA, and Alu), STR, or all 13,012 eGenes to the proportion of genes with a high score among the 7,337 non-eGenes that were tested in our data set using Fisher's Exact Test and adjusting for multiple testing with the Bonferonni method. Finally, we fit a logistic model predicting the pLI of an eGene using the eGene's absolute effect size and including logTPM of the eGene as a covariate to test for an association between eGene effect size and pLI.

eQTL localization near distal anchors of chromatin loops. To examine the localization of SVs and eSVs with respect to chromatin loops in iPSCs, we downloaded previously published iPSC promoter capture Hi-C loop calls[25]. For this analysis, we obtained loops intersecting the promoter of 13,575 out of the 16,018 expressed genes included in the eQTL analysis, of which 5,803 were eGenes. To identify variants that might affect chromatin looping, we first intersected loop calls with all annotated Gencode v19 promoters. Then, for each variant, we computed the distance from each loop anchor and retained only the variants closer to the distal anchor (i.e. the anchor that does not overlap the promoter). We subsetted this set of variant-gene pairs to those where the variant was: (1) closer to the distal anchor than the promoter anchor (2) at least 50 kb from the promoter (3) at most 200 kb from the distal anchor which comprised 177,571 variant-gene pairs (31.8% of all tested variant-gene pairs). For all these variants, we determined whether they were included in the Hi-C loop (i.e. between the promoter anchor and the distal anchor) or outside the loop. To test whether variants that hit distal loop anchors are enriched to be eQTLs, we categorized variants within 10 kb of a loop anchor as intersecting that anchor. To calculate enrichment, we tested the proportion of eVariants that intersected a distal loop anchor to at least one expressed gene versus the proportion of eVariants that did not intersect distal loop anchors (Fisher's exact test). Next, we took this subset of variant-gene pairs and restricted it further to cases where the variant was at also least 100 kb from the gene body, to test whether the distance of a variant from the distal loop anchor was associated with its likelihood of being associated with gene expression. We fit a logistic model to this set of 161,793 variant gene pairs to see whether distance to the distal anchor is predictive of the likelihood of being associated with gene expression using distance to the gene body and non-mode allele frequency as covariates. We also compared the proportion of variants that were eVariants that were within 100 kb from the outside of the distal loop anchor to variants that were within 100 kb from inside of the distal loop anchor (Fisher's exact test). To test whether overlapping loop anchors associated with multiple promoters was more predictive of associations with multiple eGenes than variant localization in a gene dense window, we modeled the number of eGenes versus the number of genes tested for each eVariant, stratifying by the number of genes tested that were either connected by loops to the promoter or not connected by loops to the promoter (statsmodels.api.logit, statsmodels v0.9.0, https://pypi.org/project/statsmodels/). To visualize these regressions, seaborn (regplot)(https://pypi.org/project/seaborn/) was used in order to divide the X axis (number of genes tested that were connected or not connected by loops) into bins with points drawn at the center of the bin showing the mean and error bars indicating 95% confidence intervals. The same bins were used for both groups in order to enable direct comparison between groups, however, each bin does not contain equal numbers of observations.

SV/STR LD tagging and GWAS associations. We downloaded summary statistics for 4,357 human traits from the UK BioBank (UKBB) GWAS Round 2 (http://www.nealelab.is/uk-biobank, 01 August 2018). For each of the 42,921 non-redundant SVs and STRs, we used bcftools[52] to extract all SNPs 50 kb upstream and downstream. For each SV or STR, we calculated LD as the correlation (R^2) with the genotypes of each surrounding SNV or indel genotyped in i2QTL WGS. We selected the variant with strongest LD overall, as well as the variant with the strongest LD that was included in the UKBB data set (if the two were different). For each UKBB variant linked to an SV or STR, we obtained p values for the variant in all GWAS studies and considered it to be significantly associated with a trait if $p < 5 \times 10^{-8}$. For each variant type, we selected all lead SVs and STRs from the SV/STR-only eQTL analysis and tested if the lead eVariants were: (1) more likely to be in strong LD with UKBB variants in general, and (2) more likely to be in strong LD with UKBB variants significantly associated with a GWAS trait, as compared to non-lead eVariants, using the Fisher's exact test. To test the association of multi-eGene eQTLs with the likelihood of being in strong LD with a variant significantly associated with a GWAS trait, we divided tested variants by class and modeled the likelihood of a variant being linked to a trait versus the number associated eGenes (statsmodels.api.logit, statsmodels v0.9.0, https://pypi.org/project/statsmodels/). p values were calculated using the Wald test and then corrected for multiple testing using Benjamini–Hochberg.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Variant calls for iPSCORE samples and full eQTL summary statistics are available at dbGaP (phs001325 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001325.v2.p1]). Variant calls for HipSci samples are available from the Zenodo (https://doi.org/10.5281/zenodo.3835306). iPSCORE RNA-seq data are available at dbGaP (phs000924 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000924.v4.p1]); HipSci RNA-seq data are available in ENA project ERP007111 and several EGA projects (Supplementary Table 1). RNA-seq data from Banovich et al.[36] is available at SRA (SRP093633) and http://eqtl.uchicago.edu/yri_ipsc/. GENESiPS RNA-seq data is available at the SRA (SRP072417) and dbGaP (phs001139.v1.p1). RNA-seq data from the PhLiPS projects is available at dbGaP (phs001341.v1.p1).

Received: 26 July 2019; Accepted: 5 May 2020;
Published online: 10 June 2020

References

1. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
2. Schlattl, A., Anders, S., Waszak, S. M., Huber, W. & Korbel, J. O. Relating CNVs to transcriptome data at fine resolution: Assessment of the effect of

variant size, type, and overlap with functional regions. *Genome Res.* **21**, 2004–2013 (2011).
3. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
4. Jakubosky, D. et al. Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. *Nat Commun*. https://doi.org/10.1038/s41467-020-16481-5 (2020).
5. Li, X. et al. The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
6. Willems, T. et al. Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).
7. DeBoever, C. et al. Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells. *Cell Stem Cell* **20**, 533–546 (2017).
8. Den Dunnen, W. F. A. Trinucleotide repeat disorders. *Handb. Clin. Neurol.* **145**, 383–391 (2017).
9. Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).
10. Nelson, D. L., Orr, H. T. & Warren, S. T. The unstable repeats–three evolving faces of neurological disease. *Neuron* **77**, 825–843 (2013).
11. Beck, M. et al. Craniofacial abnormalities and developmental delay in two families with overlapping 22q12.1 microdeletions involving the MN1 gene. *Am. J. Med Genet A* **167A**, 1047–1053 (2015).
12. Brandler, W. M. et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
13. King, D. A. et al. Mosaic structural variation in children with developmental disorders. *Hum. Mol. Genet* **24**, 2733–2745 (2015).
14. Lupski, J. R. Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environ. Mol. Mutagen* **56**, 419–436 (2015).
15. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).
16. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
17. Panopoulos, A. D. et al. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Rep.* **8**, 1086–1100 (2017).
18. Ruderfer, D. M. *et al*. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat. Genet.* **48**, https://doi.org/10.1038/ng.3638 (2016).
19. Karczewski, K. J. et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017).
20. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
21. Duggal, G., Wang, H. & Kingsford, C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Res.* **42**, 87–96 (2014).
22. Greenwald, W. W. et al. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nat. Commun.* **10**, 1054 (2019).
23. Rao, S. S. P. et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping (vol 159, pg 1665, 2014). *Cell* **162**, 687–688 (2015).
24. Schoenfelder, S. et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
25. Montefiori, L. E. *et al*. A promoter interaction map for cardiovascular disease genetics. *Elife* **7**, https://doi.org/10.7554/eLife.35788 (2018).
26. Babbs, C. et al. De novo and rare inherited mutations implicate the transcriptional coregulator TCF20/SPBP in autism spectrum disorder. *J. Med. Genet.* **51**, 737–747 (2014).
27. Kong, S. W. et al. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS One* **7**, e49475 (2012).
28. Sun, J. H. et al. Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175**, 224–238 e215 (2018).
29. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
30. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
31. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
32. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
33. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
34. Gardner, E. J. *et al*. The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* https://doi.org/10.1101/gr.218032.116 (2017).
35. Streeter, I. et al. The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res.* **45**, D691–D697 (2017).
36. Banovich, N. E. et al. Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* **28**, 122–131 (2018).
37. Carcamo-Orive, I. et al. Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals Genetic and Non-genetic Determinants of Heterogeneity. *Cell Stem Cell* **20**, 518–532 e519 (2017).
38. Pashos, E. E. et al. Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. *Cell Stem Cell* **20**, 558–570 e510 (2017).
39. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
40. Flicek, P. et al. Ensembl 2014. *Nucleic Acids Res.* **42**, D749–D755 (2014).
41. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
42. Nikolayeva, O. & Robinson, M. D. edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. *Methods Mol. Biol.* **1150**, 45–79 (2014).
43. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
44. Bonder, M. J. *et al*. Systematic assessment of regulatory effects of human disease variants in pluripotent cells. Preprint at https://doi.org/10.1101/784967 (2019).
45. Slifer, S. H. PLINK: Key Functions for Data Analysis. *Curr. Protoc. Hum. Genet* **97**, e59 (2018).
46. Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. Preprint at https://doi.org/10.1101/003905 (2014).
47. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
48. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
49. Gymrek, M. M. *et al*. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **27**, 617–630 (2016).
50. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
51. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760–1774 (2012).
52. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Acknowledgements

This work was supported in part by supported by the National Science Foundation, a CIRM grant GC1R-06673 and NIH grants HG008118, HL107442, DK105541 and DK112155. D.J. and M.K.R.D. were supported by the National Library Of Medicine of the National Institutes of Health under Award Number T15LM011271. W.W.YG. was supported by the National Heart, Lung, And Blood Institute of the National Institutes of Health under Award Number F31HL142151. S.B.M. was supported by NIH grant U01HG009431.

Author contributions

Conceptualization, D.J., E.N.S., M.D., M.J.B., S.B.M, C.D and K.A.F.; Methodology, D.J., M.D., M.J.B., C.S., M.K.R.D., W.W.Y.G., S.B.M., O.S., S.B.M., C.D.; Formal Analysis, D.J., M.D., M.J.B.; Data Curation, D.J., M.J.B., C.S., A.D.C., H.M.; Writing, D.J., M.D., C.D. and K.A.F.; Visualization, D.J.; Supervision, O.S., S.B.M and K.A.F.; Funding Acquisition, O.S., S.B.M. and K.A.F.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16482-4.

Correspondence and requests for materials should be addressed to K.A.F.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

i2QTL Consortium

Marc J. Bonder[4,5], Na Cai[4,12], Ivan Carcamo-Orive[13], Matteo D'Antonio[3], Kelly A. Frazer[3,10], William W. Young Greenwald[8], David Jakubosky[1,2], Joshua W. Knowles[13], Hiroko Matsui[3], Davis J. McCarthy[4,14], Bogdan A. Mirauta[4], Stephen B. Montgomery[7,11], Thomas Quertermous[13], Daniel D. Seaton[4], Craig Smail[6,7], Erin N. Smith[10] & Oliver Stegle[4,5,9]

[12]Wellcome Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [13]Division of Cardiovascular Medicine and Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA 94305, USA. [14]St Vincent's Institute of Medical Research, Fitzroy, Victoria 3065, Australia.