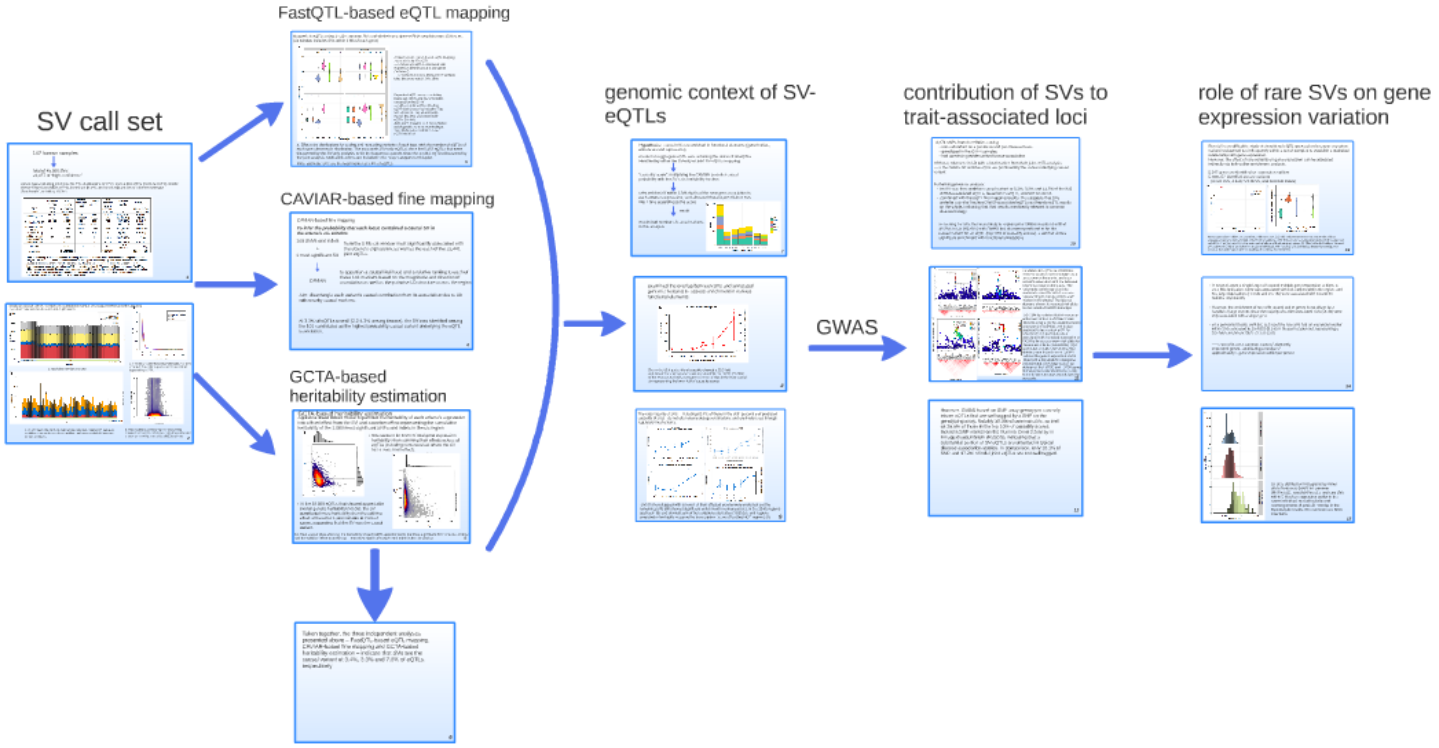
The impact of structural variation on human gene expression

--- a comprehensive human eQTL mapping study from deep WGS data
that directly measures the contribution of SVs, SNVs, and indels.



147 human samples

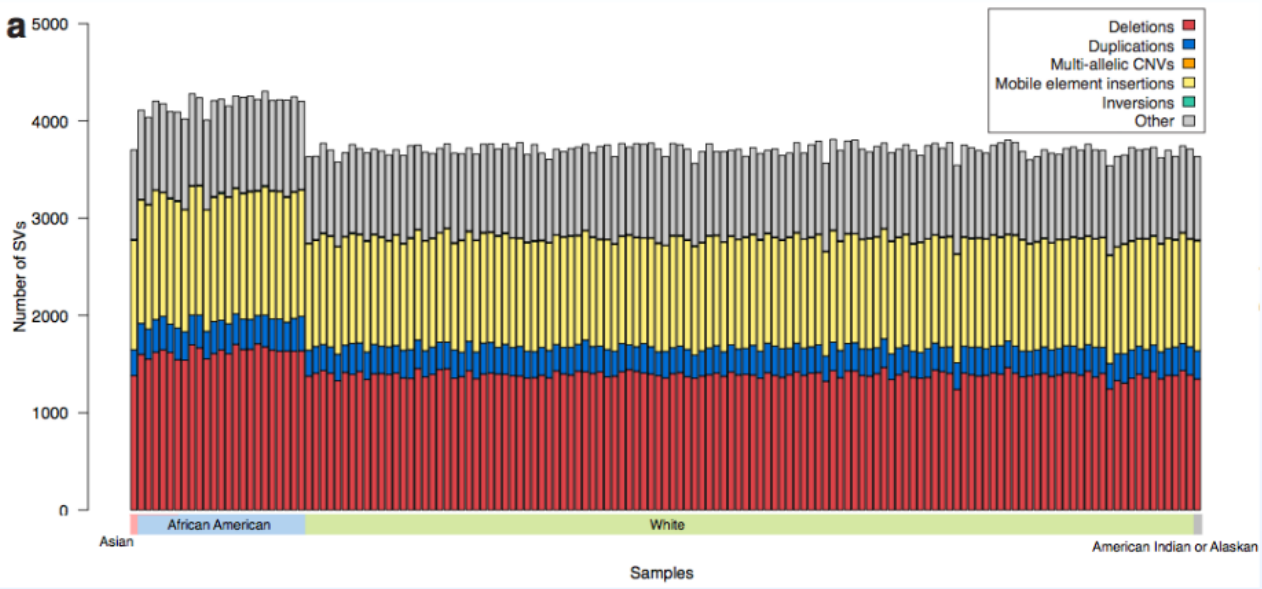
total of 45,968 SVs;
24,157 of “high confidence”

variant types including deletions (50.4%), duplications (14.7%), multi-allelic CNVs (mCNVs; 6.8%), mobile element insertions (MEIs; 8.5%), inversions (0.2%), and novel adjacencies of indeterminate type (“breakends”, or BNDs; 19.3%)

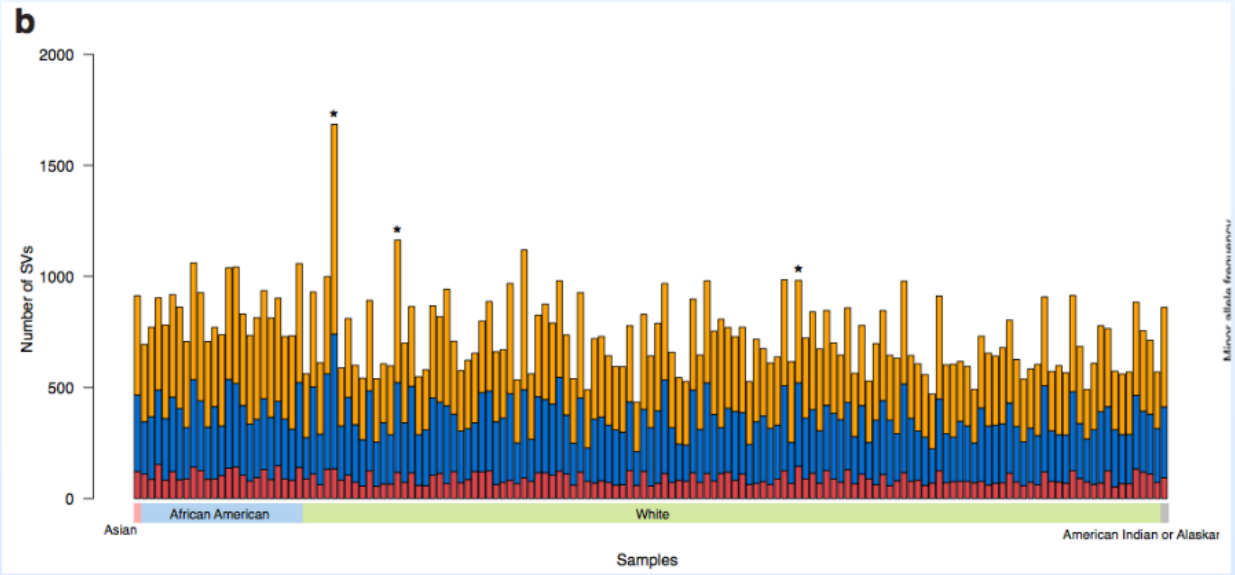
	Detection method	# of variants	Median resolution (bp)	Median size (bp)	# of common variants	SV-only eQTLs	Joint eQTLs
SNV	GATK	21,764,904	-	1	6,982,921	-	17,024 (0.2%)
Indel	GATK	3,030,964	-	3	834,255	-	2,245 (0.3%)
Deletion (DEL)	BP, RD	11,692	34	966	3,373	546 (16.2%)	27 (0.8%)
	RD	493	kilobase*	3,699	275	63 (22.9%)	16 (5.8%)
Duplication (DUP)	BP, RD	2,514	97	577	798	103 (12.9%)	3 (0.4%)
	RD	1,054	kilobase*	4,993	833	148 (17.8%)	70 (8.4%)
Multi-allelic CNV (mCNV)	RD	1,635	kilobase*	3,676	992	241 (24.2%)	100 (10.1%)
Inversion (INV)	BP	58	10	573	19	2 (10.5%)	0 (0.0%)
Mobile element insertion (MEI)	BP	2,053	1	21	1,610	269 (16.7%)	11 (1%)
Other SV (BND)	BP	4,658	33	-	2,220	298 (13.4%)	6 (0.7%)
All SVs	-	24,157	178	-	10,120	1,670 (16.5%)	233 (2.3%)
All variants	-	24,820,025	-	-	7,827,296	-	19,269 (0.2%)

Table 1. Summary of variant types and discovery methods. SNVs and indels were detected using the Genome Analysis Toolkit (GATK) and SVs were detected by breakpoint evidence (BP), read-depth evidence (RD), or both. A subset of common variants were tested for *cis*-eQTLs, and were required to be on the autosomes or X chromosome and present in a minimum of 10 of the individuals with RNA-seq in at least one tissue. The SV-only eQTL mapping excluded SNVs and indels for greater sensitivity, while the joint eQTL mapping included all variant types. *Resolution refers to the positional certainty at each breakpoint, with read-depth variants having approximate breakpoint precision on the kilobase scale.

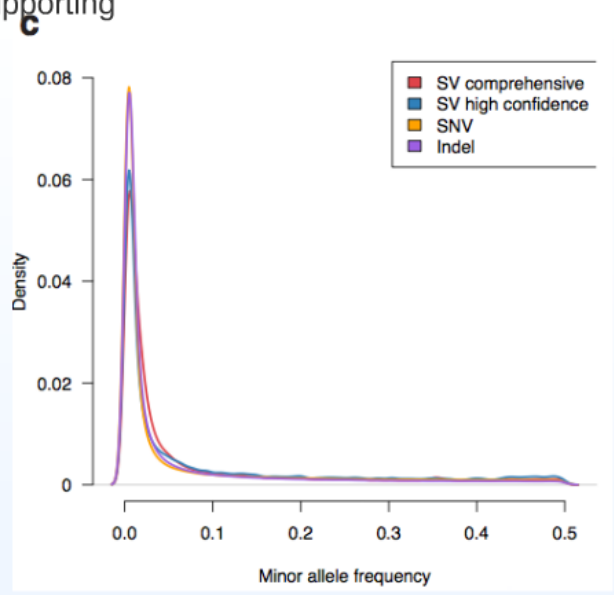
Structural variation call set. Number of SVs detected in sample by breakpoint evidence (with supporting read-depth evidence for deletions and duplications)



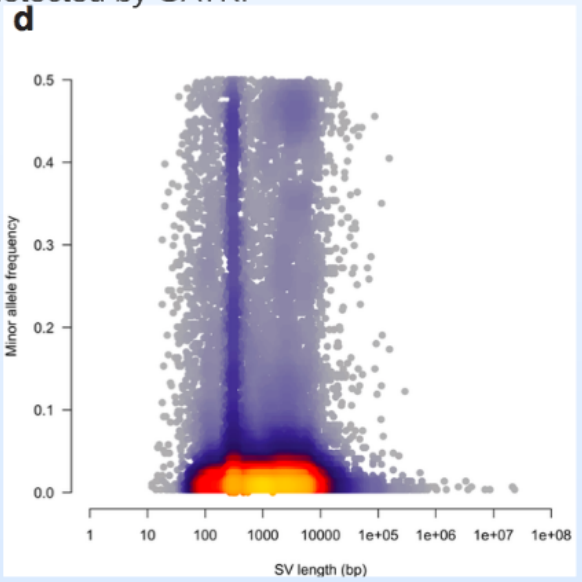
a. read-depth evidence alone



b. in 147 deep (30- 50X) human whole genomes. Starred (*) samples exhibited abnormal read-depth profiles, and were excluded from rare variant analyses.

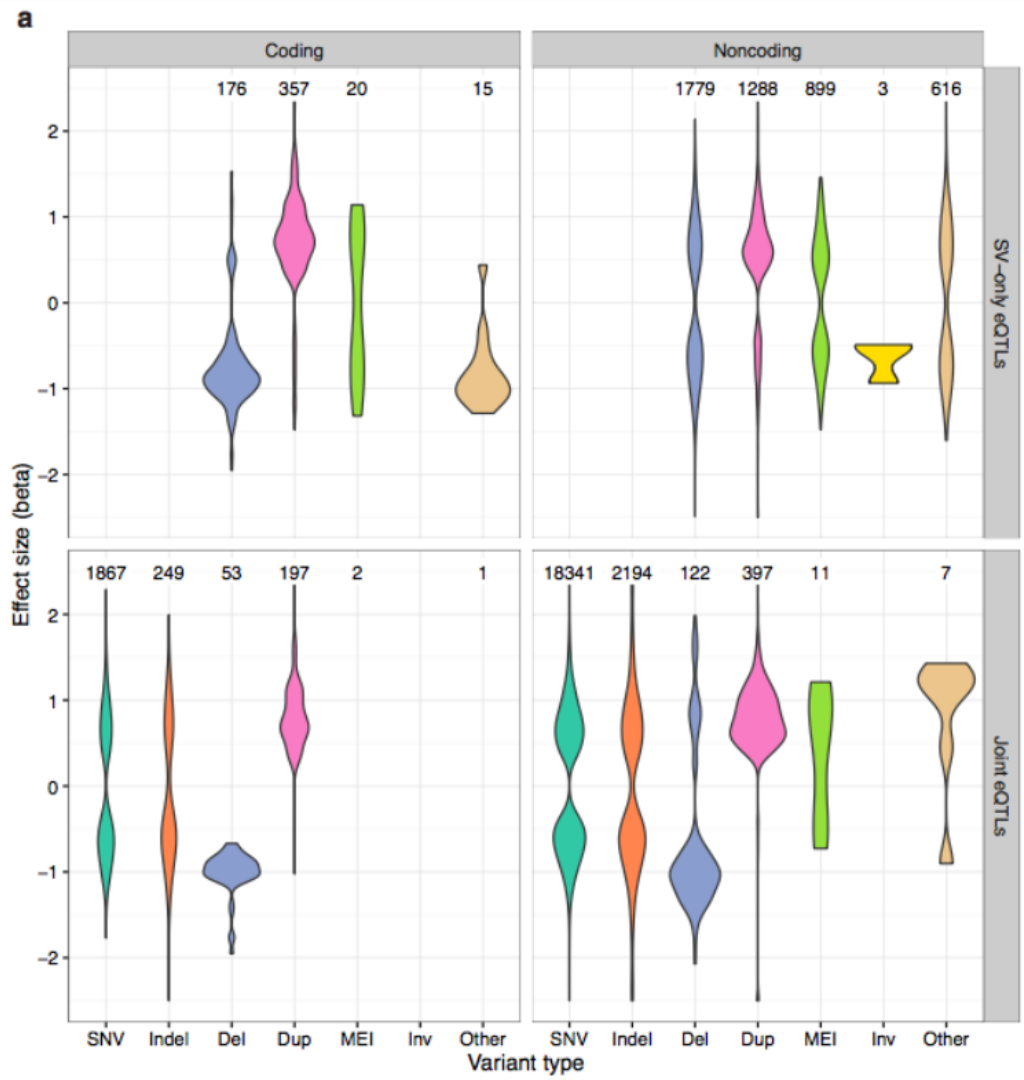


c. The minor allele frequency distribution of SVs mirrored that of high quality SNVs and indels detected by GATK.



d. Heat scatter plot showing the relationship between SV length and minor allele frequency, with a peak at ~300 bp due to Alu SINE insertions

Mapped cis eQTLs using 10,120 common SVs and whole transcriptome RNA-seq data from 13 tissues. (cis window includes SVs within 1 Mb of each gene)



Applied a permutation-based eQTL mapping approach using FastQTL
----> 5,153 SV-eQTLs associated with expression differences at 2,102 genes ("eGenes")
----> 1,670 distinct SVs (Benjamini-Hochberg false discovery rate (FDR): 10%)

Expanded eQTL analysis including 6,982,921 SNVs and 822,241 indels detected by the GATK
---->23,441 joint eQTLs affecting 9,547 distinct eGenes including 790 SV- eQTLs (3.4%), 20,208 SNV-eQTLs (86.2%), and 2,443 indel-eQTLs (10.4%)
Joint eQTL mapping with the complete set of genetic variants, nominating a most likely causal variant for each eQTL identified

a. Effect size distributions for coding and noncoding variants of each type, with the number of eQTLs of each type above each distribution. The top panels (SV-only eQTLs) show the 5,153 eQTLs that were discovered by the SV-only analysis, while the bottom two panels show the 23,441 eQTLs discovered by the joint analysis. Multi-allelic CNVs are included in the "Dup" category for this plot.

Initial estimate: SVs are the lead marker at 3.4% of eQTLs

CAVIAR-based fine mapping

To infer the probability that each locus contained a causal SV in the eGene's cis window

100 SNVs and indels

+

1 most sginificant SV

from the 1 Mb cis window most significantly associated with the eGene's expression, as well as the each of the 23,441 joint eQTLs.

CAVIAR

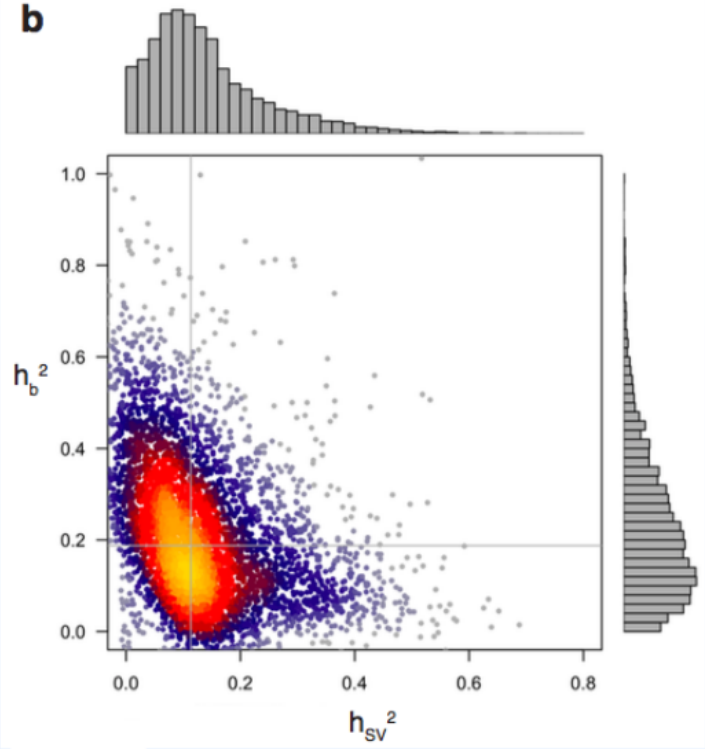
to apportion a causal likelihood and a relative ranking to each of these 101 markers based on the magnitude and direction of association as well as the pairwise LD structure across the region

Aim: disentangle each variant's causal contribution from its association due to LD with nearby causal markers.

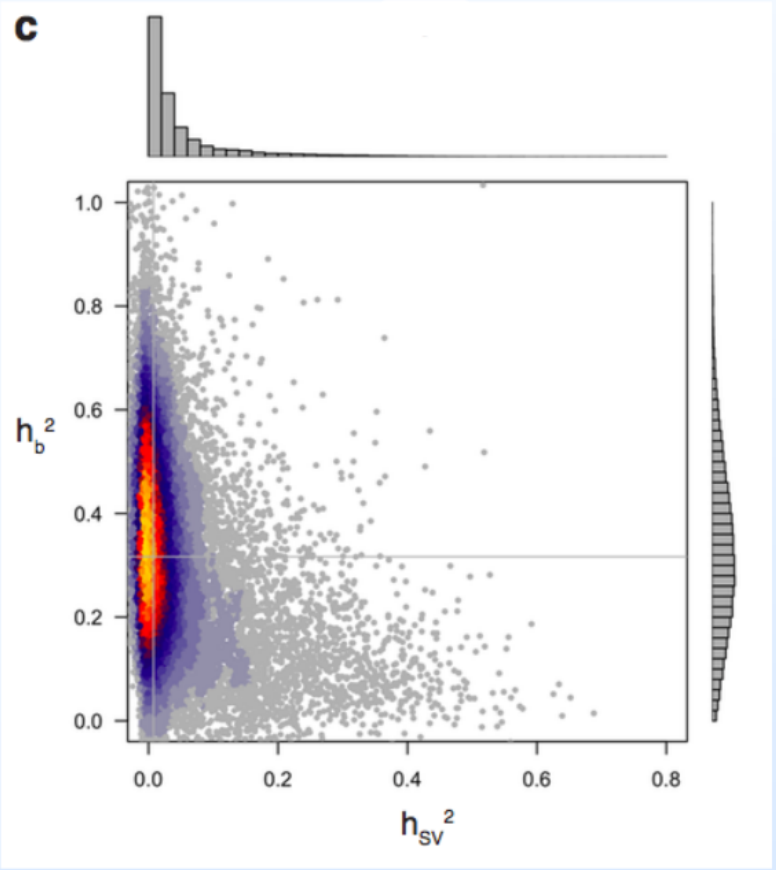
At 3.3% of eQTLs overall (2.2-4.3% among tissues), the SV was identified among the 101 candidates as the highest probability causal variant underlying the eQTL association.

GCTA-based heritability estimation

Applied a linear mixed model to partition the heritability of each eGene's expression into a fixed effect from the SV, and a random effect representing the cumulative heritability of the 1,000 most significant SNPs and indels in the cis region



- SVs account for 8.5% of total gene expression heritability when summing their effects across all eQTLs (including numerous loci where the SV has a very small effect).



- At the 22,289 eQTLs that showed appreciable overall genetic heritability (>0.05), the SV contributed more heritability than the additive effect of the other 1,000 variants in 7.0% of cases, suggesting that the SV was the causal variant.

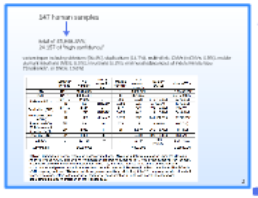
b,c Heat scatter plots showing the heritability of each eQTL apportioned to the most significant SV in the cis window and the additive effect from the top 1,000 most significant SNVs and indels in the cis window

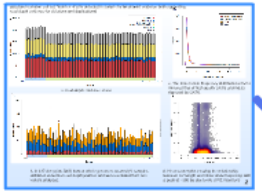
Taken together, the three independent analyses presented above – FastQTL-based eQTL mapping, CAVIAR-based fine mapping and GCTA-based heritability estimation – indicate that SVs are the causal variant at 3.4%, 3.3% and 7.0% of eQTLs, respectively

The impact of structural variation on human gene expression

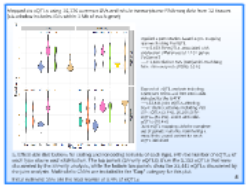
--- a comprehensive human eQTL mapping study from deep WGS data
that directly measures the contribution of SVs, SNVs, and indels.

SV call set

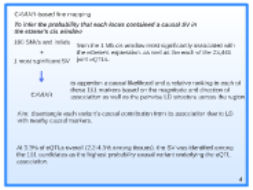




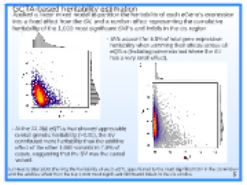
FastQTL-based eQTL mapping

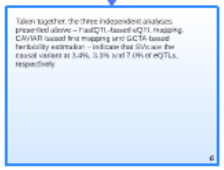


CAVIAR-based fine mapping

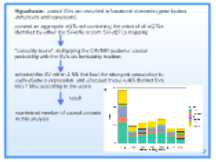


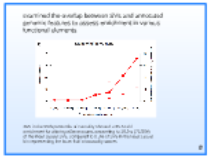
GCTA-based
heritability estimation

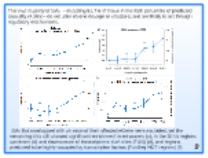




genomic context of SV-eQTLs

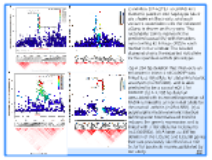




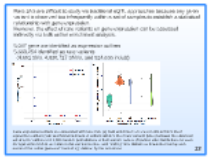


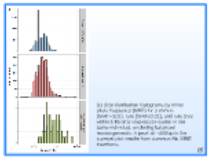
GWAS

contribution of SVs to trait-associated loci



role of rare SVs on gene expression variation





Hypothesis: causal SVs are enriched in functional elements (gene bodies, enhancers and repressors)

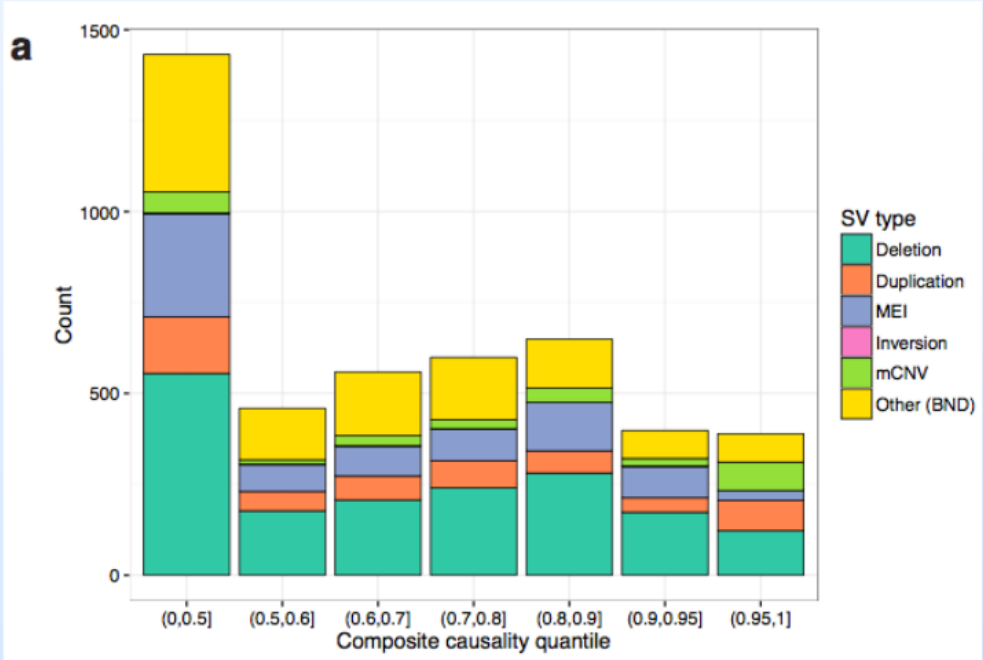
created an aggregate eQTL set containing the union of all eQTLs identified by either the SV-only or joint SV-eQTLs mapping

"casuality score": multiplying the CAVIAR posterior causal probability with the SV's cis heritability fraction

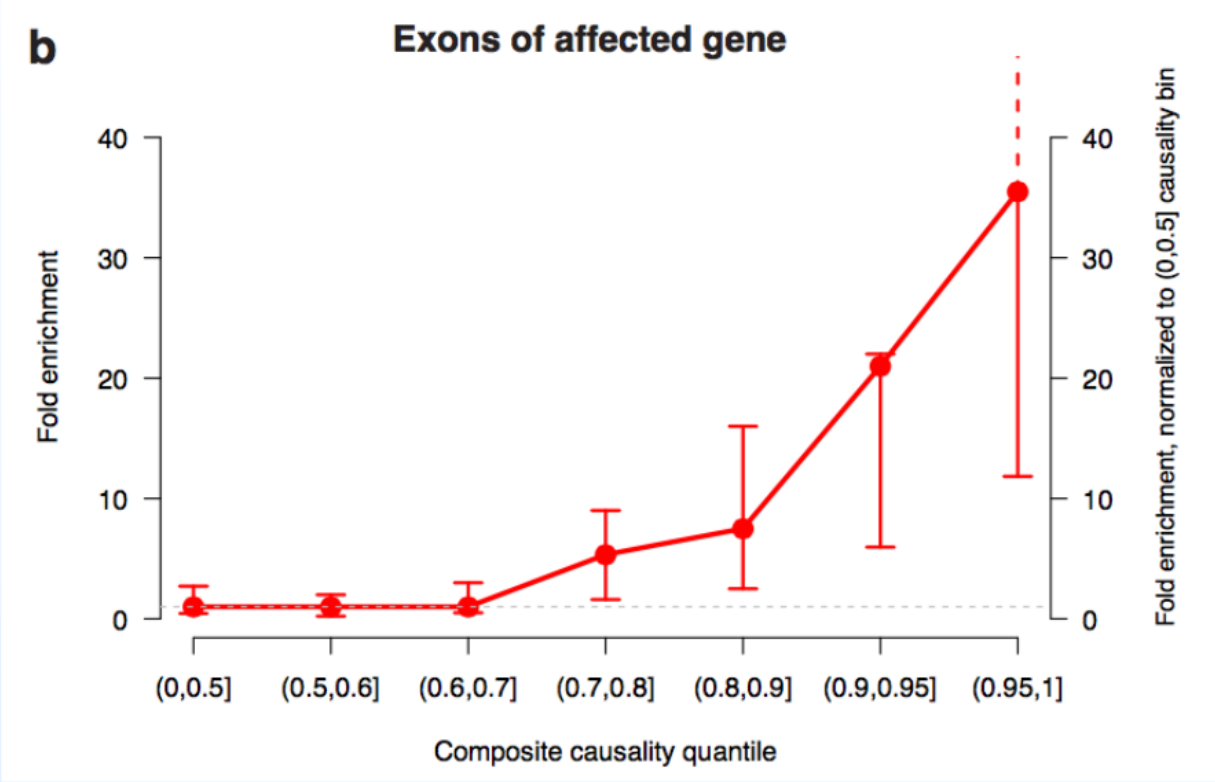
selected the SV within 1 Mb that had the strongest association to each eGene's expression, and allocated these 4,485 distinct SVs into 7 bins according to the score

result

maximized number of causal variants in this analysis

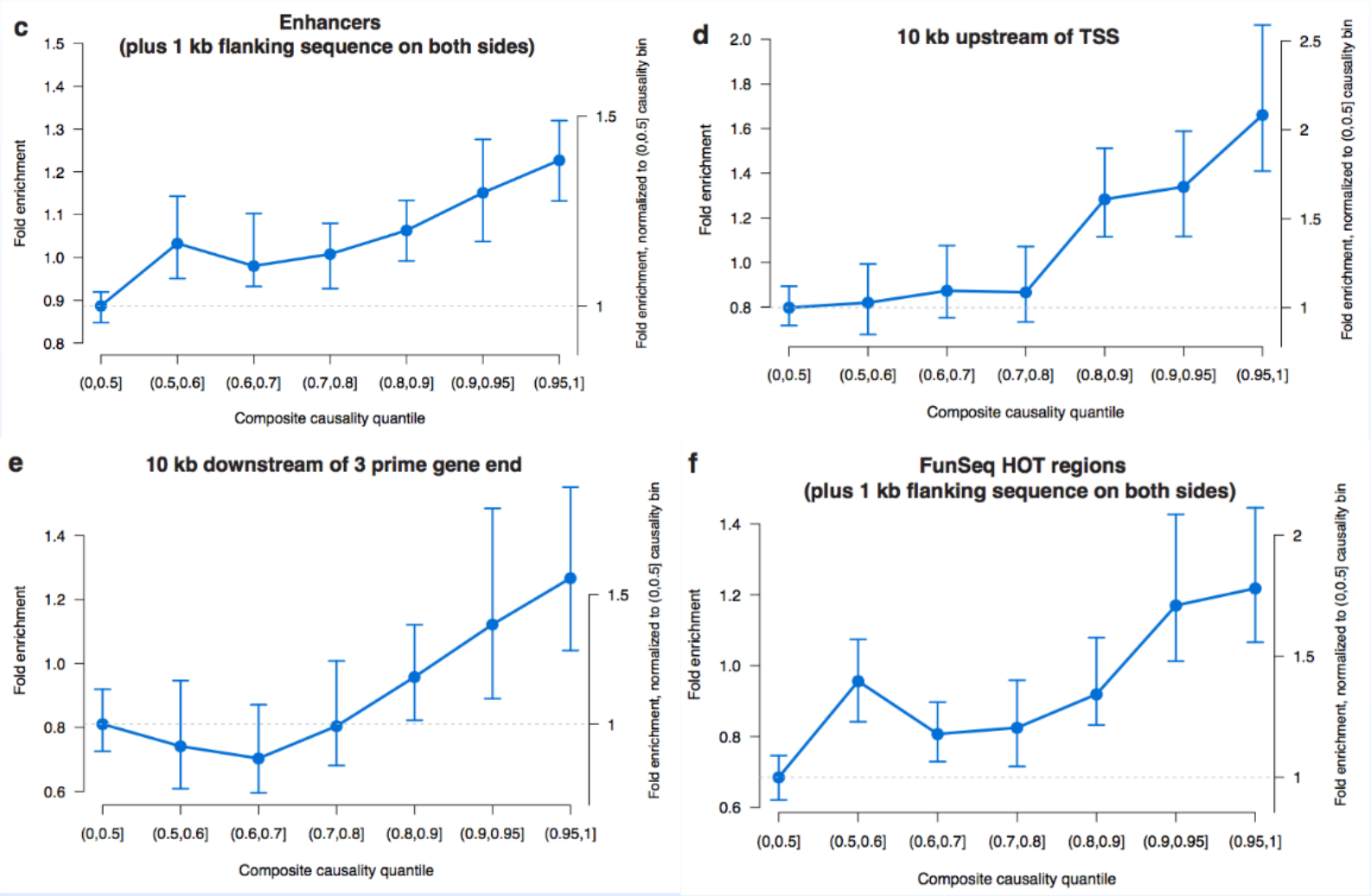


examined the overlap between SVs and annotated genomic features to assess enrichment in various functional elements.



SVs in the 95th percentile of causality showed a 35.5-fold enrichment for altering eGene exons, amounting to 18.3% (71/389) of the most causal SVs, compared to 0.3% of SVs in the least causal bin representing the lower half of causality scores

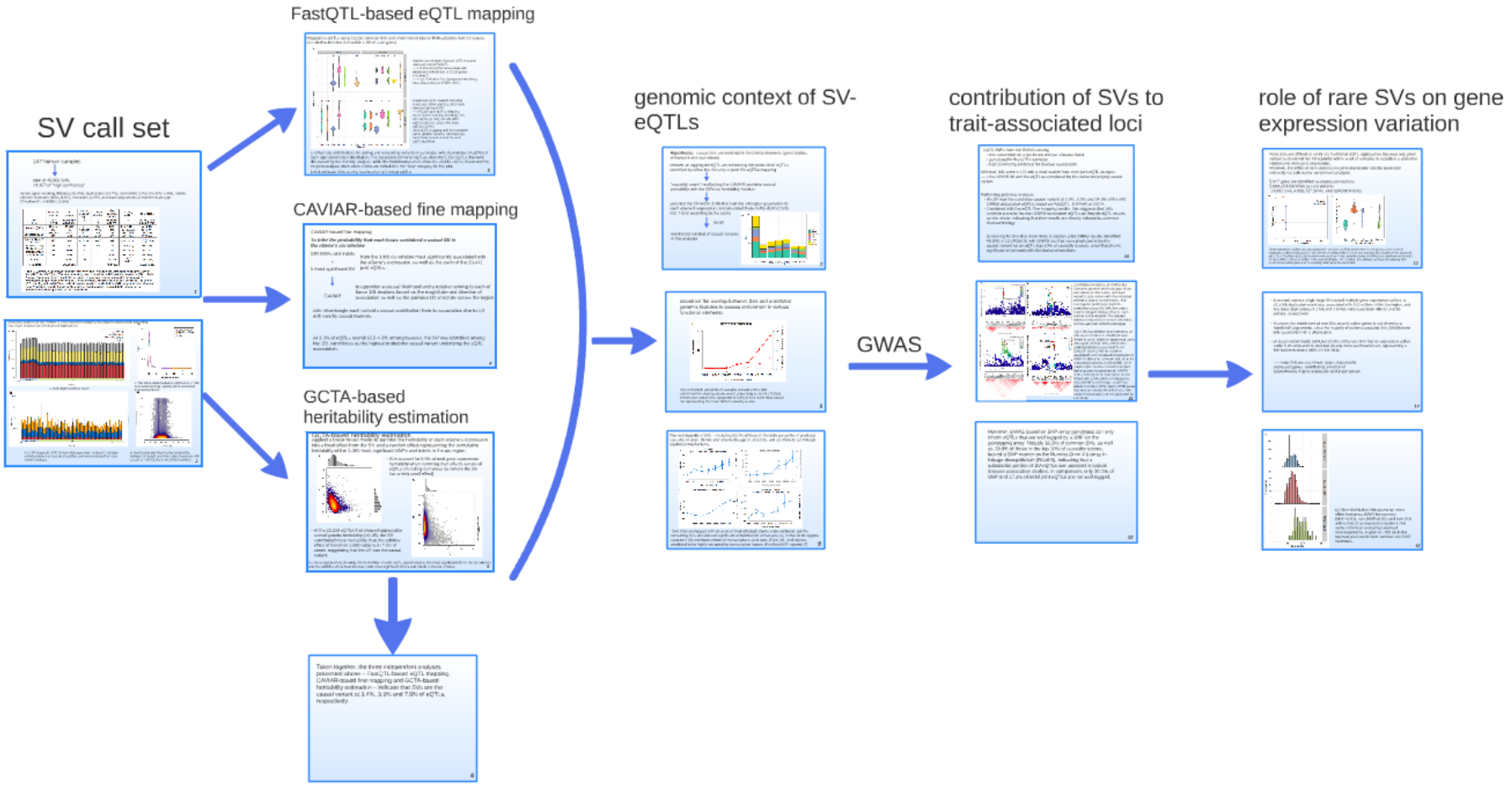
The vast majority of SVs – including 81.7% of those in the 95th percentile of predicted causality (4,363)– do not alter eGene dosage or structure, and are likely to act through regulatory mechanisms.



SVs that overlapped with an exon of their affected eGene were excluded, yet the remaining SVs still showed significant enrichment in enhancers (c), in the 10 kb regions upstream (d) and downstream of transcriptions start sites (TSS) (e), and regions predicted to be highly occupied by transcription factors (FunSeq HOT regions) (f).

The impact of structural variation on human gene expression

--- a comprehensive human eQTL mapping study from deep WGS data
that directly measures the contribution of SVs, SNVs, and indels.



4,874 SNPs from the GWAS catalog

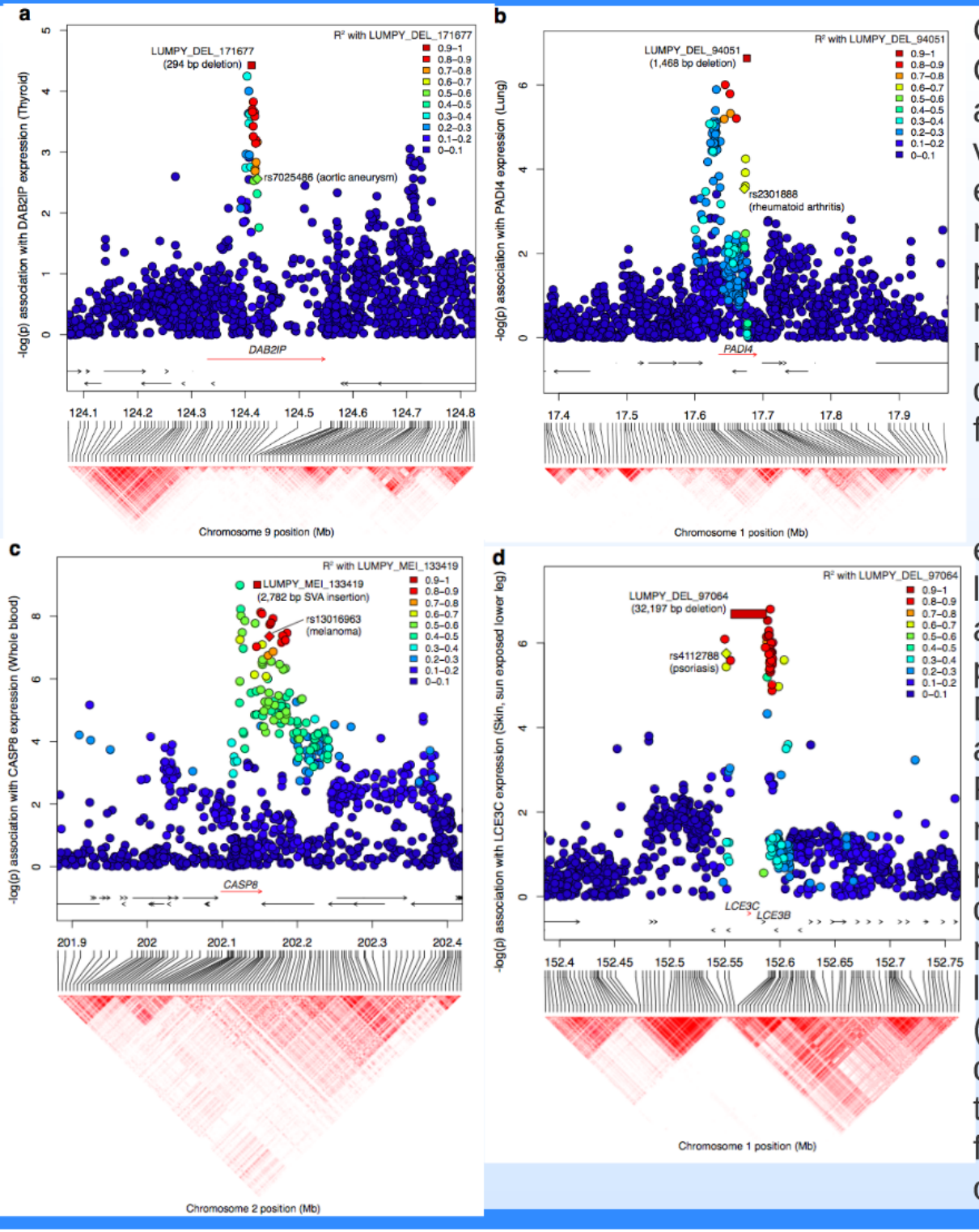
- non-redundant on a per-locus and per-disease basis
- genotyped in the GTEx samples
- had convincing evidence for disease association

Of these, 842 were in LD with a lead marker from their joint eQTL analysis ----> the GWAS hit and the eQTL are produced by the same underlying causal variant.

Performing previous analysis:

- An SV was the candidate causal variant at 3.2%, 3.3% and 14.3% of the 842 GWAS-associated eQTLs, based on FastQTL, CAVIAR or GCTA.
- Combined with the eQTL fine mapping results, this suggests that SVs underlie a similar fraction GWAS-associated eQTLs as they do eQTL results on the whole, indicating that their results are directly relevant to common disease biology

Screening for SVs that were likely to explain prior GWAS results identified 49 SVs in LD (R2≥0.5) with GWAS loci that were predicted to be the causal variant for an eQTL (top 10% of causality scores), a set that shows significant enrichment with functional annotations



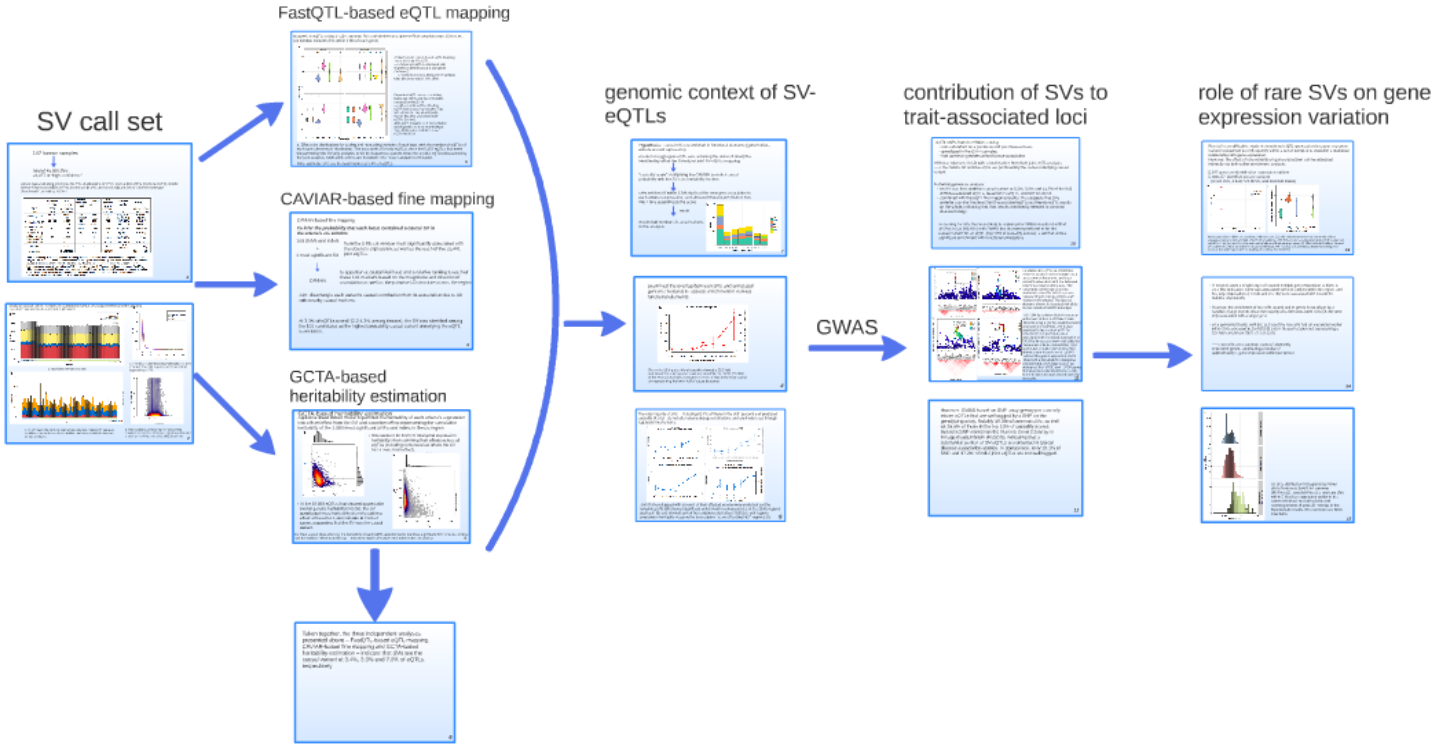
Candidate SV-eQTLs at GWAS loci. Genomic position and haplotype block are shown on the x-axis, and each variant's association with the indicated eGene is shown on the y-axis. The rectangular points represent the predicted causal SV, with the colors representing its linkage (R2) to each marker in the window. The labeled diamond shows the reported risk allele for the specified GWAS phenotype

(a) A 294 bp deletion that intersects an enhancer in intron 1 of DAB2IP was linked to a risk allele for abdominal aortic aneurysm (rs7025486), and is also predicted to be a causal eQTL for DAB2IP. (b) A 1,468 bp deletion associated with increased expression of PADI4 is linked to a known risk allele for rheumatoid arthritis (rs2301888). (c) A polymorphic mobile element insertion defining exon boundaries of CASP8 reduces the gene's expression and is linked with a risk allele for melanoma (rs13016963). (d) A large 32,197 bp deletion of the LCE3C and LCE3B genes that was previously identified as a risk factor for psoriasis was recapitulated by our study.

However, GWAS based on SNP array genotypes can only inform eQTLs that are well tagged by a SNP on the genotyping array. Notably 30.3% of common SVs, as well as 29.9% of those in the top 10% of causality scores, lacked a SNP marker on the Illumina Omni 2.5 array in linkage disequilibrium ($R2 \geq 0.5$), indicating that a substantial portion of SV-eQTLs are untested in typical disease association studies. In comparison, only 10.1% of SNP and 17.2% of indel joint eQTLs are not well tagged.

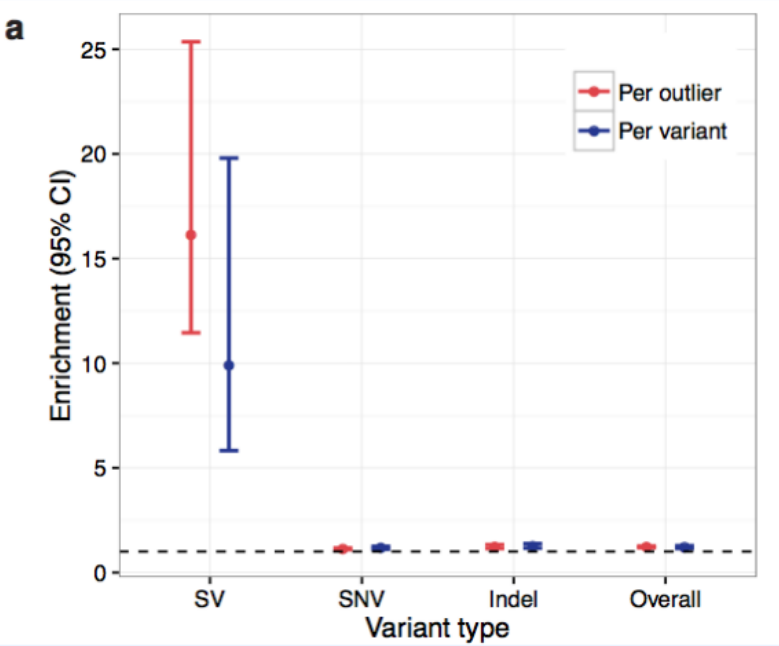
The impact of structural variation on human gene expression

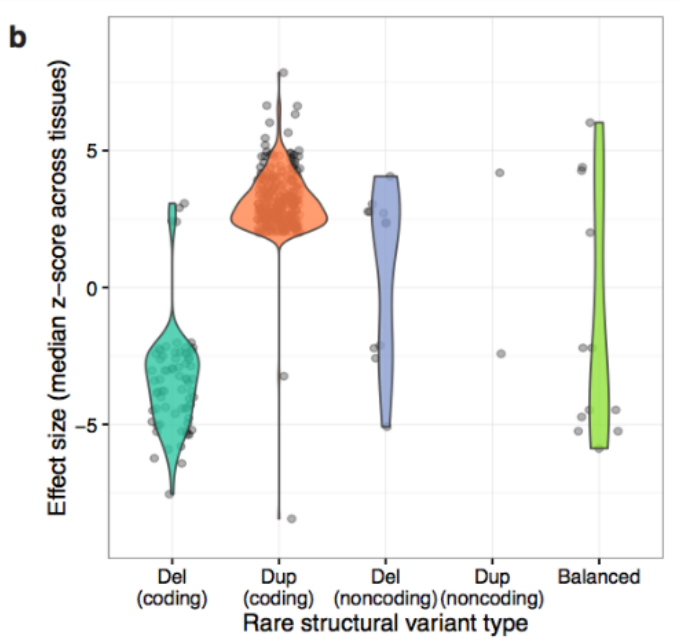
--- a comprehensive human eQTL mapping study from deep WGS data
that directly measures the contribution of SVs, SNVs, and indels.



Rare SVs are difficult to study via traditional eQTL approaches because any given variant is observed too infrequently within a set of samples to establish a statistical relationship with gene expression.
However, the effect of rare variants on gene expression can be assessed indirectly via bulk outlier enrichment analysis.

5,047 gene are identified as expression outliers
5,660,254 identified as rare variants
(4,691 SVs, 4,830,727 SNVs, and 824,836 indels)





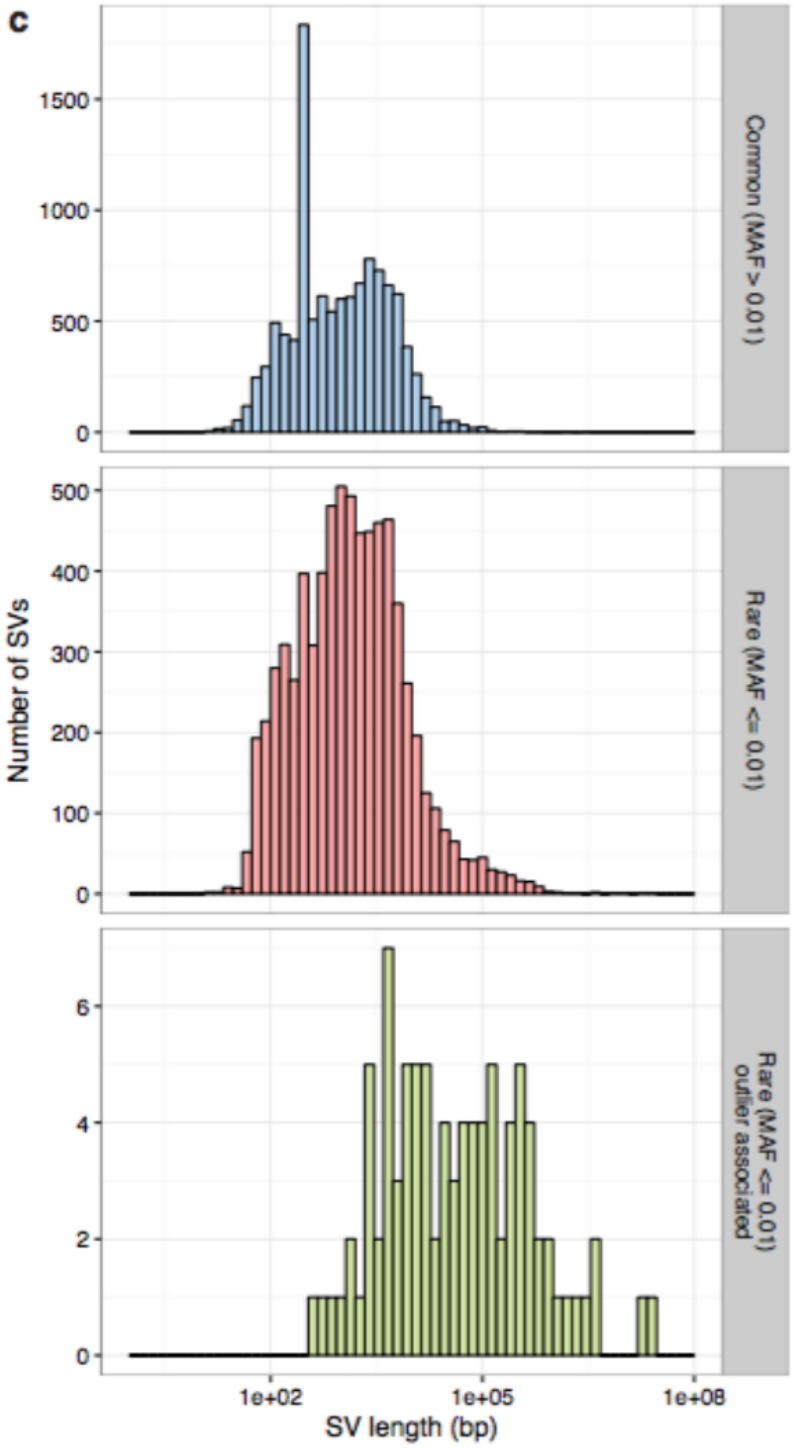
Gene expression outliers are associated with rare SVs. (a) Fold enrichment of rare variants within 5 kb of expression outliers (red) and fold enrichment of outliers within 5 kb of rare variants (blue) between the observed set of 5,047 outliers and 1,000 random permutations of their sample names. (b) Effect size distributions for each SV type within 5 kb of an outlier in the same individual, with “coding” SVs defined as those that overlap with exons of the outlier gene and “noncoding” defined by the remainder.

- In several cases a single large SV caused multiple gene expression outliers: a 21.3 Mb duplication event was associated with 161 outliers within the region, and two large duplications (4.1 Mb and 2.5 Mb) were associated with 11 and 30 outliers, respectively

- However, the enrichment of rare SVs around outlier genes is not driven by a handful of large events, since the majority of outlier-associated SVs (56/99) were only associated with a single gene

- on a per-variant basis, 99/4,691 (2.1%) of the rare SVs had an expression outlier within 5 kb compared to 10/4,691 (0.2%) in the permutation set, representing a 9.9- fold enrichment (95% CI: 5.8-19.8)

----->rare SVs are a common cause of aberrantly expressed genes, contributing a median of approximately 1 gene expression outlier per person



(c) Size distribution histograms by minor allele frequency (MAF) for common (MAF>0.01), rare (MAF≤0.01), and rare SVs within 5 kb of an expression outlier in the same individual, excluding balanced rearrangements. A peak at ~300 bp in the topmost plot results from common Alu SINE insertions.