UNIVERSITY OF CALIFORNIA SAN DIEGO

Identification and Functional Impact of Structural Variants and Short Tandem Repeats in the Human Genome

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Biomedical Sciences

by

David Jakubosky

Committee in charge:

Professor Kelly Frazer, Chair
Professor Kyle Gaulton
Professor Christopher Glass
Professor Bing Ren
Professor Jonathan Sebat

2019

The Dissertation of David Jakubosky is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

To my father Andrew Jakubosky for his tireless support. To my greatest heroes, my mother Ann LeBlanc Jakubosky and grandfather Robert LeBlanc, may they rest in peace.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Kelly Frazer, Dr. Erin Smith, Dr. Christopher DeBoever and Dr. Matteo D'Antonio as mentors throughout my doctoral process. I would not be the scientist that I am today without their guidance.

Chapter 1, in full, has been submitted for publication of the material as it may appear in Nature Communications, 2019, David Jakubosky, Erin N. Smith, Matteo D'Antonio, Marc Jan Bonder, William W. Young Greenwald, Agnieszka D'Antonio-Chronowska, Hiroko Matsui, Oliver Stegle, Stephen B. Montgomery, Christopher DeBoever, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, has been submitted for publication of the material as it may appear in Nature Communications, 2019, David Jakubosky, Matteo D'Antonio, Marc Jan Bonder, Craig Smail, Margaret K.R. Donovan, William W. Young Greenwald, Agnieszka D'Antonio-Chronowska, Hiroko Matsui, Oliver Stegle, Erin N. Smith, Stephen B. Montgomery, Christopher DeBoever, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

VITA

2007-2012 Bachelor of Science, Behavioral Neuroscience, Northeastern University

2014-2019 Doctor of Philosophy, Biomedical Sciences, University of California San Diego

PUBLICATIONS

First Author

Genomic properties of structural variants and short tandem repeats that impact gene expression and complex traits in humans. *In Review: Nature Communications.* 2019

Discovery and Quality Analysis of a Comprehensive Set of Structural Variants and Short Tandem Repeats. *In Review: Nature Communications.* 2019

Other Authorship

Genetic regulation of gene expression in iPSC derived cardiomyocytes. *In Review: Nature Communications. 2019*

Full resolution HLA typing of 273 individuals from deep whole-genome sequencing data enables genetic studies of human 6p21.3. *In Review: eLife. 2019*

Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature Communications.* 2019

Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. *Cell Reports*. 2018

iPSCORE: A resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types. *Stem Cell Reports.* 2017

Decreased STARD10 expression is associated with defective insulin secretion in humans and mice. *American Journal of Human Genetics.* 2017

Whole genome sequencing revealed mutations in two independent genes as the underlying cause of retinal degeneration in an ashkenazi jewish pedigree. *Genes (Basel)*. 2017

Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell*. 2017

Establishing the involvement of the novel gene AGBL5 in retinitis pigmentosa by whole genome sequencing. *Physiological Genomics*. 2016

ABSTRACT OF THE DISSERTATION

Identification and Functional Impact of Structural Variants and Short Tandem Repeats in the Human Genome

by

David Jakubosky

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2019

Professor Kelly Frazer, Chair

Over the last decade, a substantial amount of work in genetics has been done with the goal of understanding how genetic variation affects human traits and diseases, primarily via genome-wide association studies (GWAS). Until recently, these studies have focused on associations with single nucleotide variants (SNVs) largely because they have traditionally been easier to genotype. However, the

genome contains diverse classes of non-SNV variation such as short tandem repeats (STRs) and structural variants (SVs) that have been shown in some cases to affect human traits. The increasing availability of deep whole genome sequencing (WGS) data has now enabled algorithms to robustly detect high resolution structural variants and STRs, unlocking the possibility of a deeper understanding of the effects of these variants. Here I present two studies that focus on characterizing the extent and functional impact of SVs and STRs in the human genome. First, I present a study in which I built a comprehensive high quality map of SVs and STRs using over 700 deeply sequenced whole genomes. I also describe a novel method of filtering variants using reproducibility of genotypes within genetically duplicate sample pairs, and use this information to make insights into the quality of diverse classes of variants called using different methods. I then utilize this high quality map of genetic variation to assess the impact of different classes of variation on gene expression, and show that the functional properties of unique classes of genetic variation are associated with their likelihood to affect genes and linkage to complex traits in humans.

CHAPTER 1 DISCOVERY AND QUALITY ANALYSIS OF A COMPREHENSIVE SET OF STRUCTURAL VARIANTS AND SHORT TANDEM REPEATS

Abstract

Structural variants (SVs) and short tandem repeats (STRs) are important sources of genetic diversity but are not routinely analyzed in genetic studies because they are difficult to accurately identify and genotype. Because SVs and STRs range in size and type, it is necessary to apply multiple algorithms that incorporate different types of evidence from sequencing data and employ complex filtering strategies to discover a comprehensive set of high-quality and reproducible variants. Here we assembled a set of 719 deep whole genome sequencing (WGS) samples (mean 42x) from 477 distinct individuals which we used to discover and genotype a wide spectrum of SV and STR variants using five algorithms. We used 177 unique pairs of genetic replicates to identify factors that affect variant call reproducibility and developed a systematic filtering strategy to create of one of the most complete and well characterized maps of SVs and STRs to date.

Introduction

Structural variants (SVs) and short tandem repeats (STRs) respresent a significant fraction of polymorphic bases in the human genome and have been shown to cause monogenic diseases and contribute to complex disease risk [1-14]. STRs are polymorphic 1-6 base pair (bp) sequence repeats whose total size can range from ~10bp to more than 1kb while SVs capture diverse sequence variation greater

than 50bp in size such as insertions, duplications, deletions, and mobile element insertions (MEIs). The full contribution of STRs and SVs to disease risk, quantitative molecular traits, and other human phenotypes is currently not understood because previous studies have typically genotyped SVs and STRs using arrays or low coverage sequencing which are limited in their ability to accurately identify and genotype these variants in many samples across different variant classes and sizes [15-18]. The increasing adoption of high coverage whole genome sequencing data (WGS), however, has recently enabled the development of improved methods to identify STRs and different classes of SVs [19-21].

While high-depth WGS data has made it possible to profile a wider spectrum of genetic variation, the variability in the size and characteristics of SV classes necessitates the use of several algorithmic approaches that differ in the types of evidence used to capture all classes of SVs. For instance, some algorithms specialize in identifying small SVs (50-5,000 bp) by using split or discordant read (abnormal insert size) information to determine the location of SV breakpoints with high resolution [22-25]. Other algorithms detect large SVs (>5 kb) by comparing the amount of reads that align to the reference genome to identify regions that differ in copy number between samples [26-29], but with lower resolution breakpoint precision [20,30-32]. Finally, algorithms have also been designed to contend with more complex multi-allelic signatures, including regions with multiple copy number or repeat alleleles that are more challenging to genotype than biallelic variants [27,29]. Genotyping SVs and STRs across many samples thus requires using several highly parameterized algorithms to discover each class of SVs, processing schemes to combine results from different algorithms, and detailed filtering to remove false positives or inconsistely genotyped variants. Such pipelines for SV/STR identification must also be sensitive to study-specific parameters such as library prepration methods, sequencing depth, cell/tissue type, and read length [19-21,30-32]. Thus, due to the diversity of SV/STR calling algorithms and the need for complex downstream

processing, it remains difficult to create a comprehensive SV and STR call set with consistent quality that covers the spectrum of variant sizes and subclasses.

In addition to difficulties associated with complex pipelines for calling SVs and STRs, the need to perform *de novo* discovery and subsequent genotyping of variants across hundreds or thousands of samples leads to inconsistencies between variant calls across studies. A comprehensive catalog of SVs and STRs in the human genome would make it possible for different studies to genotype this same set of variants. While several efforts are underway to establish such catalogs of SVs [18-20,32-36] and STRs [37,38], most are limited in their number and diversity of samples or do not capture all types of variants due to the sequencing depth or algorithms employed. There is also a need to understand the extent to which differences in sample collection and preparation may impact SV and STR calling by measuring the reproducibility of variants called on genetic duplicate samples that share the same genome but were collected and prepped separately. A comprehensive reference catalog of high quality SVs and STRs discovered in a large set of subjects with deep WGS data could therefore be useful for calling and genotyping the full spectrum of variants across future studies involving hundreds to thousands of subjects.

In this study, as part of the i2QTL consortium, we profiled 719 whole genomes from iPSCORE [39-41] and HipSci [42,43] with five variant calling algorithms to capture a wide spectrum of SVs including biallelic deletions and duplications, multi-allelic copy number variants (mCNVs; regions that have more than two copy number alleles segregating in the population), MEIs, reference MEIs (rMEI), inversions, unspecified breakends (BND), and STRs. We identified algorithm-specific quality metrics and SV genomic properties associated with the reproducibility of variant calling using 177 pairs of genetic replicates embedded in our collection (25 monozygotic twin pairs and 152 fibroblast-iPSC pairs) and devised filtering and processing approaches to obtain a highly accurate, non-redundant call set across variant classes and algorithmic approaches. We compared our set of SVs with those identified in GTEx [19]

and 1KGP [18] and found that we capture the vast majority of common SVs likely discoverable in Europeans with short read sequencing and add novel, high-quality variants at lower allele frequencies. Finally, we characterized the extent to which different classes of SVs and STRs are tagged by SNPs and indels. This study establishes methods for filtering SVs and STRs to obtain reproducible variant calls and provides a high-quality reference catalog of SVs and STRs that will benefit studies that investigate how these variants contribute to human disease.

Results

We generated the i2QTL variant calls dataset by calling SNVs, indels, SVs, and STRs using 719 human WGS samples from 477 unique donors. (Figure 1.1A). The samples were obtained by combining data from two large induced pluripoitent stem cell (iPSC) resources: 1) iPSCORE (273 individuals, mean WGS coverage 50X, range 36-126X) [39-41] and 2) HipSci (446 samples from 204 individuals, mean WGS coverage 37X, range 35-78X) [43,44] (Figure 1.1B, C). The 477 individuals include members of all five 1KGP superpopulations [45]: 415 European, 34 East Asian, 15 Admixed American, 7 South Asian, and 6 African. While all 204 HipSci donors were unrelated, there were 183 donors in iPSCORE that are part of 56 unique families (2-14 individuals/family), including 25 monozygotic (MZ) twin pairs (Figure 1.1D). For 152 HipSci individuals, we also obtained matched fibroblast and iPSC WGS data (Figure 1.1D). Between these 152 matched samples and 25 MZ twin pairs, we had WGS data for 177 genetic replicates, which we used to determine quality filtering thresholds and to calculate reproducibility of calls across all variant classes.

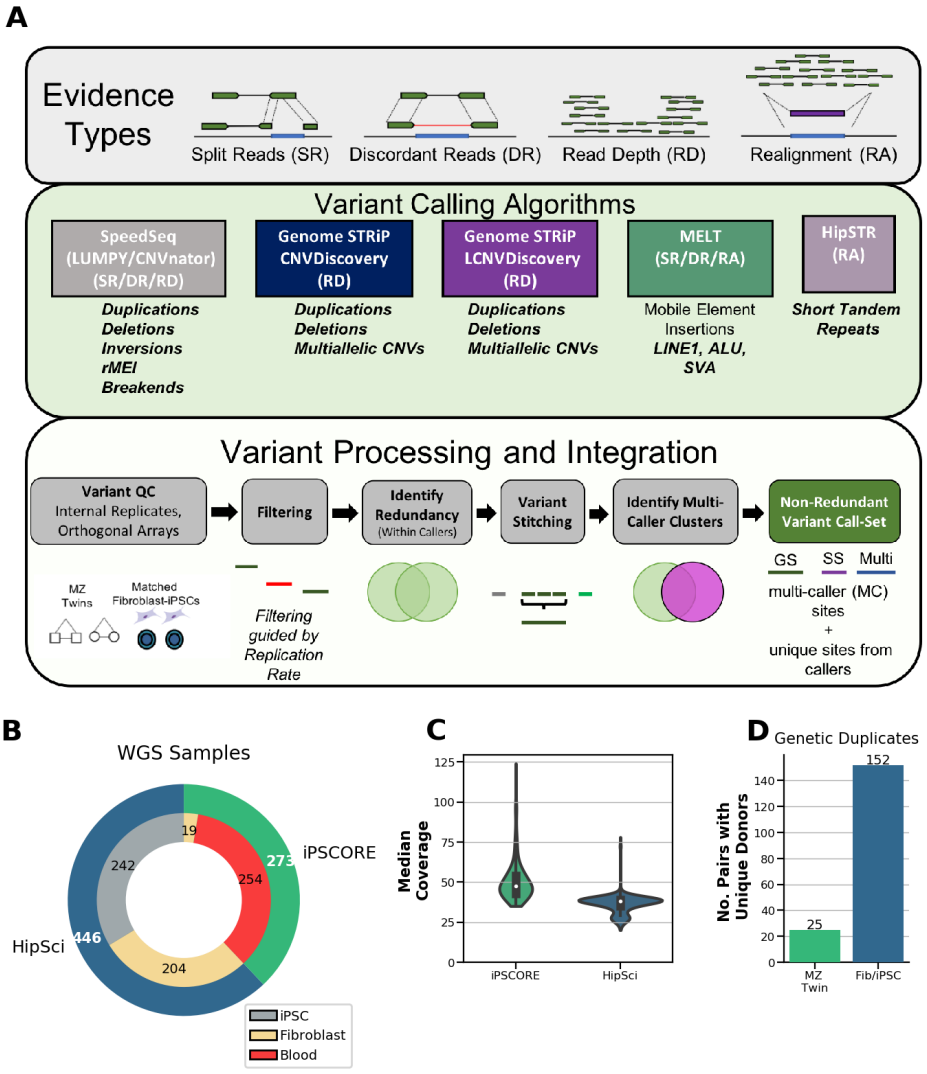


Figure 1.1 Variant calling, processing and i2QTL WGS samples
(A) Illustration of the evidence types from short read sequencing data utilized in variant calling (top). Description of the variant callers utilized, the types of variants they identify, and the evidence they use (middle). Flowchart showing the processing, quality control (see Methods), and integration of SVs from different variant callers (bottom). (B) Pie chart showing the number of whole genome sequencing samples from the iPSCORE or HipSci studies used for variant calling and the cell type from which DNA was obtained. (C) Distribution of the median coverage of whole genomes from iPSCORE (green) and HipSci (blue). (D) Number of genetic replicate samples included in the collection, including 25 monozygotic twin pairs (iPSCORE) and fibroblast-iPSC pairs from 152 unique donors (HipSci). These data enable robust variant calling for all classes of genetic variation along with reproducibility analysis.

Comprehensive structural variant call set

To identify SVs across a wide range of sizes (50 bp to > 1Mb) and classes, we called variants using four algorithms (Figure 1.1A): SpeedSeq (LUMPY/with CNVnator support)[24,26,46] , Genome STRiP CNVDiscovery, Genome STRiP LCNVDiscovery[29], and MELT [47]. Together, these algorithms incorporate information from two evidence types: 1) read-pair signal (LUMPY and MELT), which includes detection of split reads (two portions of the same read map to different genomic locations) and discordant read pairs (aligned to the genome with abnormal insert size or orientation); and 2) read-depth (Genome STRiP CNVDiscovery, Genome STRiP LCNVDiscovery, CNVnator). Generally, read-pair signal enables discovery of shorter variants (50bp) and balanced events, while read-depth signal is limited to discovery of longer (>1kb) copy number variants (CNVs) which include biallelic deletions, biallelic duplications and multi-allelic copy number variants (mCNVs). When variant calling algorithms utilize information from a group of samples to predict genotypes, study-specific differences in the WGS data (cell type assayed, library preparation technique) can cause erroneous variant calls. To account for this, we performed variant calling and genotyping separately in HipSci and iPSCORE samples for Genome STRiP and combined variant calls afterward to avoid batch effects during variant calling (Methods). Using read-pair signals we detected 223,371 SVs consisting of CNVs, inversions, MEIs, and novel adjacencies of indeterminate type referred to as "breakends" (BND). Among these SVs, biallelic deletions and biallelic duplications were also supported by supplementary read-depth evidence (CNVnator). Using read-depth signals alone (Genome STRiP), we detected 28,417 biallelic deletions, biallelic duplications, and mCNVs, bringing the initial call set to a combined 251,788 SVs, before additional processing.

Reproducibility of SV calling is associated with quality metrics

Because there is considerable diversity in subtypes of SVs and disparities between detection algorithms, measuring structural variant quality is challenging. Here we used 177 genetic replicates (25 MZ twin pairs and 152 matched fibroblast and iPSC pairs) to measure reproducibility of SV calls for each variant class and SV calling approach under a range of quality metric filter thresholds. Because of complications in variant calling on sex chromosomes due to dosage differences in males and females, we analyzed reproducibility among 198,651 autosomal SVs. Notably, we were able to assess the reproducibility of most variants in the SV call set since 44% of autosomal SVs (88,496) segregated in at least one monozygotic twin pair, 65.4% (129,937) segregated in at least one fibroblast-iPSC pair (Figure 1.2A), and 71.8% (142,678) segregated in any of the 177 genetic replicates. For each variant that segregated in at least one genetic replicate pair, we assessed reproducibility by calculating how often a non-reference genotype in one replicate pair sample was called concordantly in the other replicate sample, which we define as "replication rate" (RR, Methods). Replication rates were calculated for each SV separately among MZ twin pairs and fibroblast-iPSC pairs. The 25 MZ twin pairs were used to select filters because they have matched cell types and fewer somatic differences[39] while the 152 matched fibroblasts-iPSC pairs were used to confirm the performance of these thresholds in the HipSci collection.

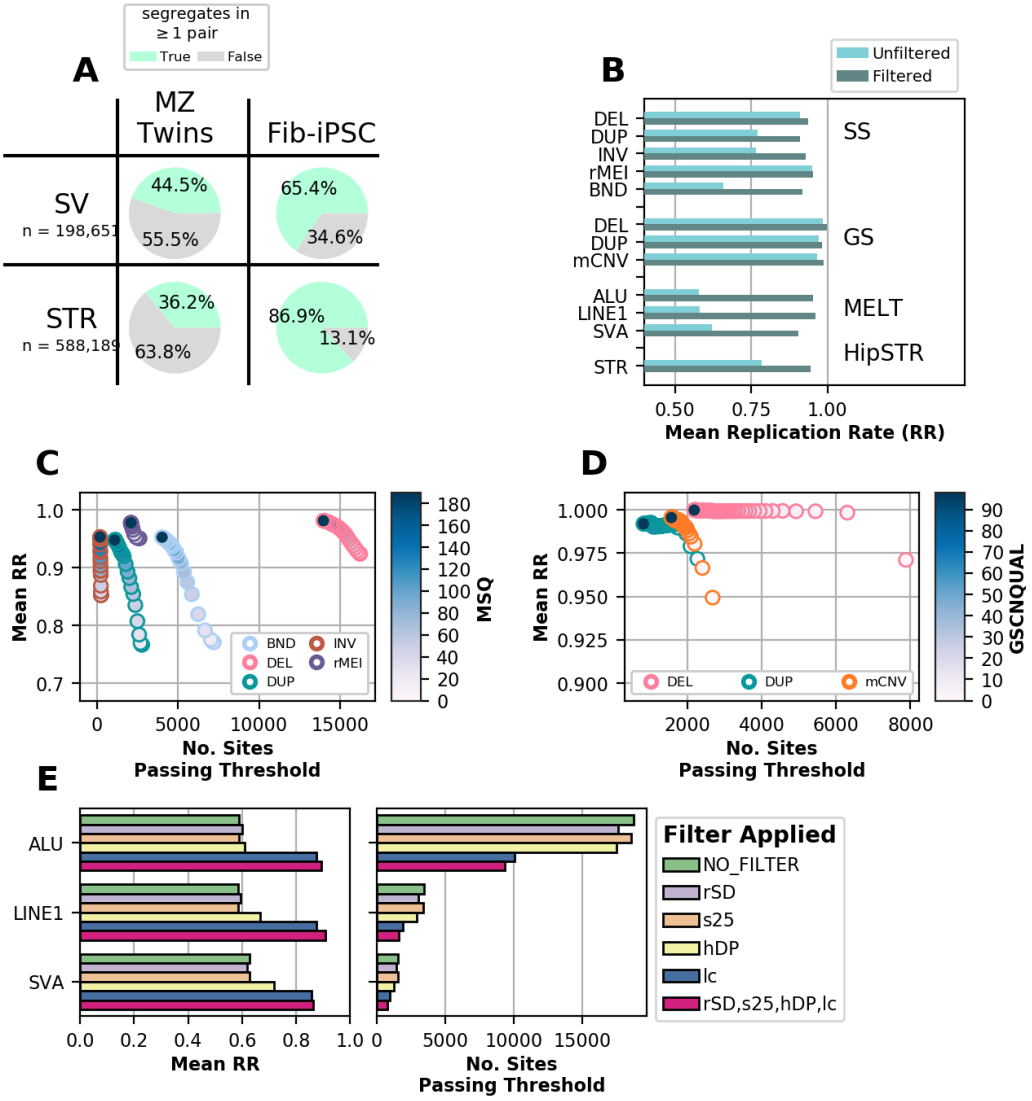


Figure 1.2 Replication rate is associated with reported quality metrics
(A) Proportion of SVs and STRs that were non-reference (green) in at least one of the iPSCORE MZ twin pairs or HipSci fibroblast-iPSC pairs prior to filtering. (B) Replication rate of variants before and after filtering and deduplicating within caller. (C) Replication rate in MZ twins versus the number of total SpeedSeq (LUMPY) sites remaining that pass criteria when filtering variants to different thresholds for MSQ score (indicated by color). (D) Replication rate versus the number of total Genome STRiP sites remaining that pass criteria when filtering variants to different thresholds for GSCNQUAL score (indicated by color). (E) Replication rate in MZ twins for MELT sites that pass criteria when filtering variants under suggested hard site filters (left). Pink represents the result of filtering using all 4 exclusion criteria (rSD, s25, hDP, lc; see Methods). The number of total sites remaining that passed criteria is shown at right.

Prior to filtering on quality metrics, we observed that within the 25 MZ twin pairs CNVs (deletions, duplications and mCNVs) detected with Genome STRiP showed high reproducibility (RR > 0.96) as did the SpeedSeq deletions (RR > 0.9) and rMEIs (RR > 0.95), whereas SpeedSeq duplications and inversions (both RR < 0.77), BND (RR=0.65), and MELT MEIs (RR = 0.59) had lower reproducibility (Figure 1.2B). Interestingly, we found that for all variant callers, increasingly strict quality metric filters yielded variant sets with higher average replication rates, supporting the premise that reproducibility is a predictor of variant quality (Figure 1.2 C-E). For example, we found strong relationships between Median Sample Quality (MSQ) score from SpeedSeq, the GSCNQUAL score from Genome STRiP, and qualitative filters from MELT and the average RR of filtered variants (Figure 1.2 C-E). Notably, filtering MELT variants called in low complexity regions ("lc" tag in FILTER) improved reproducibility from 59% to 87.5% in MZ twins and applying all four MELT filters improved RR to ~95% (Figure 1.2E, Methods). Using RR we selected strict quality metric thresholds for each caller and variant class to achieve high specificity without removing a significant number of variants. We observed that within each algorithm, different variant classes required different levels of filtering stringency to attain the same reproducibility (Figure 1.2C,D). For instance, insertions and duplications were less reliably genotyped than deletions regardless of detection method [19,21,32], and SpeedSeq duplications required an MSQ score of 100 to attain >0.9 RR while deletions had an RR of 0.92 with no MSQ filtering (Figure 1.2C) in MZ twin pairs.

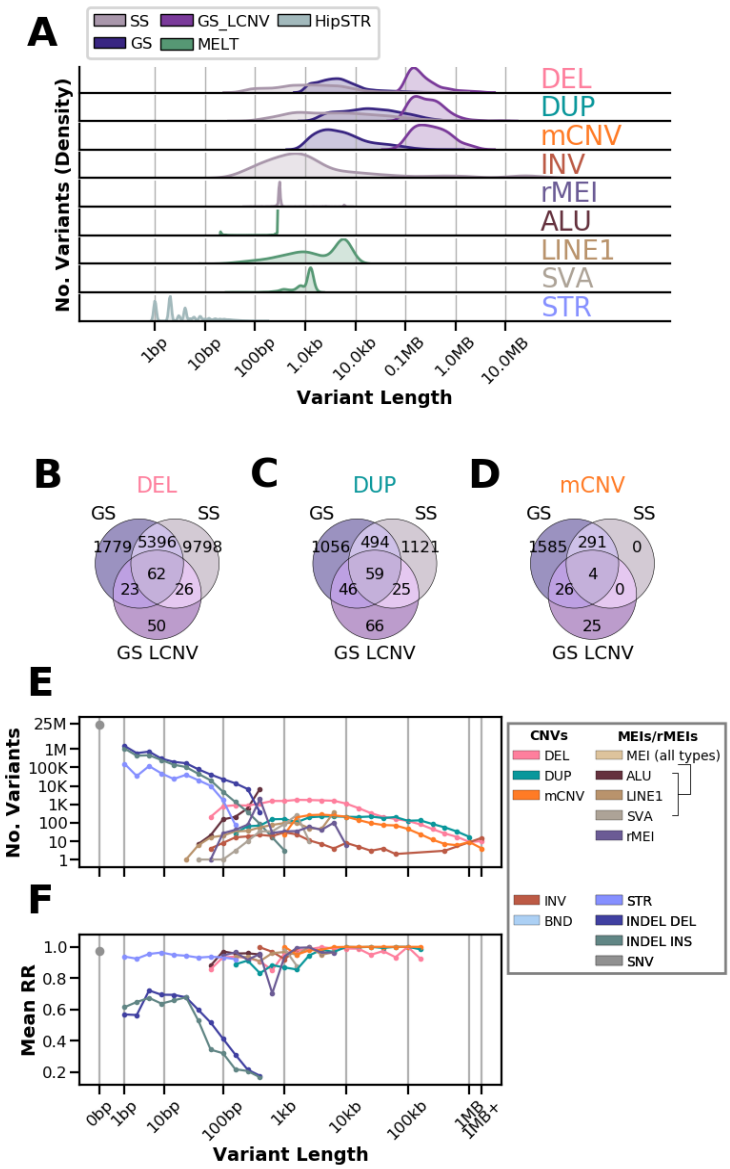


Figure 1.3 Length distributions and intersection between variants identified with each algorithm
(A) Density plot showing the size spectrum of each variant caller before identifying multi-caller clusters. (B-D) Number of overlapping variants after identifying multi-caller clusters for deletions (B) duplications (C) and mCNVs (D). (E) Number of variants in the non-redundant call set separated by variant class and grouped in log linear bins by variant length. Points are drawn at the upper limit of each bin (eg. a bin from 50-100bp is drawn at 100bp). For STRs length represents the maximum number of bases different from the reference at each site (largest insertion or deletion observed). (F) The average replication rate of variants segregating in the 25 monozygotic twin pairs is represented for each length bin that contains at least 10 variants. GATK SNVs and indels previously discovered in iPSCORE samples (DeBoever et al., 2017) were used for (E) and (F).

After filtering, we obtained 50,980 autosomal variants (20.2% of initial call set) with generally high RR (>0.9) for all callers, although variants called by SpeedSeq and MELT tended to have lower RR than those called by Genome STRiP (Figure 1.2B) suggesting that variants called using read pair signal are less reproducibly genotyped between genetic replicates than than those called by read depth signals. We tested for batch effects by comparing allele frequencies between iPSCORE and HipSci samples and found that they largely agreed across algorithms. We compared the CNV genotypes to those called from SNP arrays for 216 iPSCORE samples and found that the FDR for CNVs ranged from 3-7.8% depending on the SV type and algorithm, consistent with previous reports [18,19]. We also found that biallelic SVs generally obeyed Hardy-Weinberg across algorithms after filtering. Together, these results suggest that our stringent filtering approach can be used to obtain comparable, high quality variants across SV classes and algorithms.

Creating a high quality, non-redundant SV call set

SV calling algorithms overlap in the types and sizes of variants they identify (Figure 1.3A) which can lead to the same genetic variant being called with slightly different breakpoints by different algorithms in the same subject or by the same algorithm in different subjects. To obtain a non-redundant map of structural variation, we devised a graph-based approach to consolidate overlapping sites that are redundant with each other (Methods). We first clustered overlapping variants that were detected using the same algorithm and showed high genotype correlation and designated each cluster as a single distinct SV with a breakpoint defined by the highest quality variant (Figure 1.1A). We next stitched together neighboring variants from Genome STRiP whose genotypes were correlated because they likely represent a single variant that Genome STRiP called as multiple adjacent variants 19. Finally we clustered overlapping variants identified by different algorithms with high genotype correlation and designated

each “multi-caller” cluster as a single distinct SV (Figure 1.3B-D, Methods). We inspected variants identified by multiple algorithms and found that overlap between Genome STRiP and SpeedSeq was highest among deletions (55%), while duplications and mCNVs were only co-discovered 17% and 15% of the time respectively reflecting both the different size spectrums captured by the two methods (SpeedSeq captures smaller variants) and that evidence types (read-pair/read depth) do not always co-occur. SVs identified by more than one algorithm (i.e. with support from both read pair and read depth signals) had higher replication rates than SVs detected with a single algorithm, supporting the premise that the highest quality sites also tend to be the most reproducible. Overall, we collapsed 50,980 variants to 37,296 non-redundant SVs which were used for downstream analyses (Table 1.1). We examined the numbers and proportions of non-reference calls for each of the 719 i2QTL samples (from 477 individuals) across variant calling algorithms and variant classes. We observed high consistency in the number of variants per sample except for individuals with African ancestry who had more SVs per sample, consistent with other variant types [18,48]. Taken together, these results show that the set of i2QTL SVs is of high quality and demonstrates the utility of using genetic replicate samples for SV filtering and processing.

Table 1.1 Summary of i2QTL variants called from samples in the HipSci and iPSCORE collections.
Common variants are defined as those with ≥ 5% non-mode allele frequency (NMAF) for SVs and STRs and ≥ 5% MAF for SNVs and indels.

	Variant Class	No. Variants	No. Common Variants
	SNV	41,826,418	7,013,178
	INDEL	7,040,457	1,862,365
Copy Number Variants (CNV)	Deletion (DEL)	16,238	3,490
	Duplication (DUP)	2,693	416
	Multiallelic CNV (mCNV)	1,703	949
	Other SV (BND)	4,612	1,377
	Inversion (INV)	210	92
	Reference Mobile Element Insertions (rMEI)	2,343	1,689
Mobile Element Insertions (MEI)	ALU	7,880	2,385
	LINE1	1,175	262
	SVA	442	115
	Short Tandem Repeats (STR)	588,189	381,053
	Total SV	*37,296*	*10,775*
	Total SV/STR	*625,485*	*391,828*
	Total	*49,492,360*	*9,267,371*

Variant length, allele frequency and reproducibility

Since SVs can vary widely in size and we are using short read data to call SVs, we assessed whether replication rate was related to SV length. While we could detect many more short SVs (< 1kb) than long SVs, we observed that long SVs had higher RR (Figure 3E, F). Generally, SVs greater than 1kb were highly reproducible (> 95% RR) while shorter duplications and insertions tended to have the lowest RR, reflecting the relative lack of consistency in genotyping small read-pair based SVs. This dependence on length was observed across variant calling approaches and independent of allele frequency. We also found that rare variants were slightly less reproducible than common variants across SV classes. These

results highlight that it remains challenging to identify SVs in intermediate size ranges (~200 bp to 1 kb) using short read sequencing, because the interval is: 1) too small to distinguish from "noise" in read depth signal; 2) within the bounds of variability in insert size, making discordant read-signal undetectable; and 3) too long to be directly sequenced with a single read. While challenges in the discovery of SVs in the ~200 bp to 1 kb range still exist, the i2QTL call set consists of high-quality SVs across a wide size range of SVs (~50 bp to >1 Mb).

Comparison between SVs in i2QTL, GTEx and 1KGP

We next investigated what proportion of the SVs in the i2QTL call set are novel compared to previous SV call sets by comparing the 37,296 non-redundant i2QTL SVs with the existing 1KGP [18] and GTEx [19] SV call sets. GTEx used 148 deeply sequenced genomes and the 1KGP project used 2,504 shallowly sequenced genomes (7.4x) to call the same SV classes present in i2QTL (excluding BNDs in 1KGP and nonreference MEIs in GTEx) and are therefore strong benchmark datasets. The i2QTL SV call set captured the vast majority of common deletions, duplications, mCNVs, inversions, rMEIs and MEIs present with non-mode allele frequency (NMAF) greater than 0.05 in either study, including 77% of variants present in 1KGP Europeans and 79% of variants present in GTEx (Figure 1.4A,B). Out of all SV classes, we captured the smallest proportion of common GTEx duplications (49%) and BNDs (17%), likely due to differences in filtering stringency, WGS data quality, and breakpoint merging approaches. In total, 83% of common i2QTL SVs (NMAF > 0.05) were co-discovered by one or both of these studies (Figure 1.4C). Common deletions had the highest co-discovery rate (87%) while mCNVs had the lowest (~66%), consistent with the idea that mCNV discovery benefits from high read-depth and large numbers of samples [29]. Rare variants (NMAF < 0.05) were more likely to be unique to either set, with ~40% of sites from GTEx and 1KGP represented in the i2QTL call set (Figure 1.4C). In total, 43% of i2QTL SVs

were not found in either GTEx or 1KGP. These novel variants were predominantly rare, tended to have shorter lengths, and, excluding those identified by Genome STRiP, had on average 12% lower replication rates than co-discovered variants (Figure 1.4C). This is expected given that small SVs are the most difficult to genotype and rare variants are more likely to be false positives or negatives. Overall the i2QTL call set contains a significant number of novel, high-quality variants at lower allele frequencies missing from 1KGP and GTEx and captures most common SVs present in 1KGP and GTEx indicating that the call set contains most common SVs in Europeans.

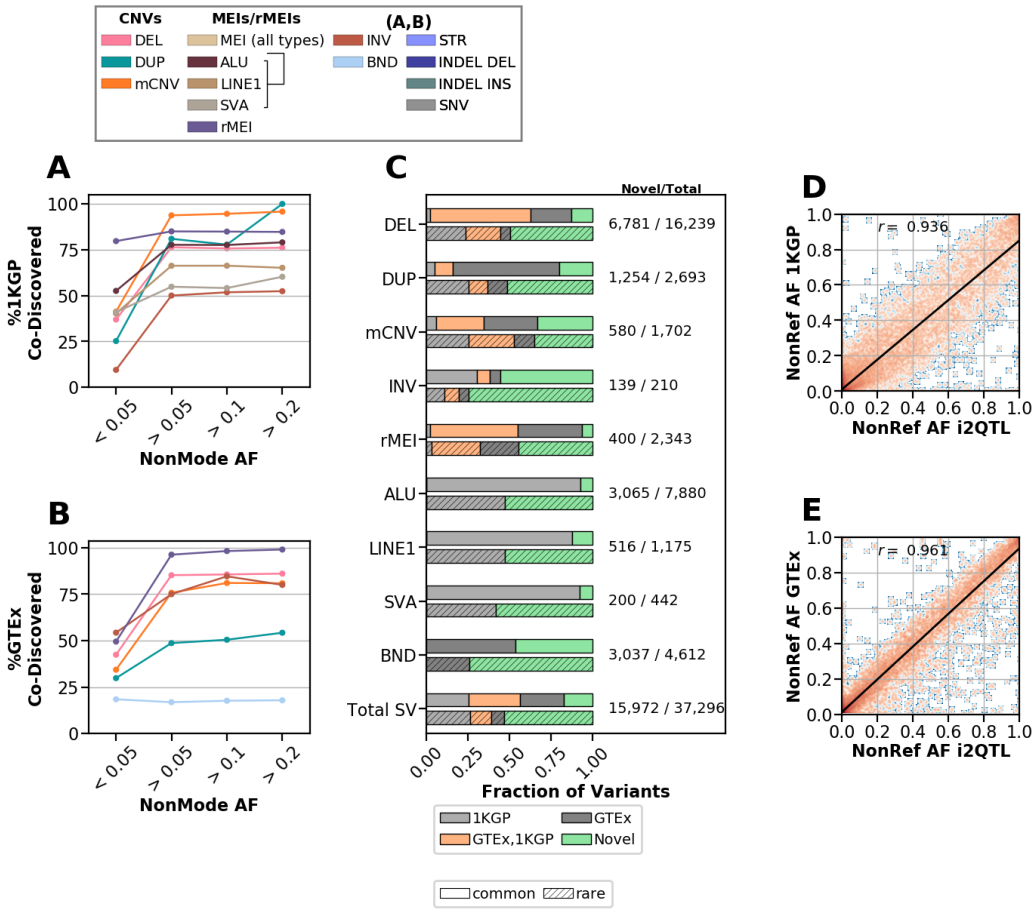


Figure 1.4 Comparison to other SV calling studies
(A,B) The fraction of variants from either (A) 1KGP (European population) or (B) GTEx that were also captured in our study in different non-mode allele frequency (NMAF) bins. (C) Fraction of i2QTL SVs that were co-discovered in 1KGP, GTEx, both 1KGP and GTEx, or were unique to i2QTL (novel), divided by whether variants were common (> 0.05 NMAF) or rare (< 0.05 NMAF) in unrelated i2QTL samples indicated by absence or presence of hatching respectively. (D,E) Non-reference allele frequency of variants co-discovered in i2QTL and (D) 1KGP (Europeans) or (E) GTEx in their respective discovery samples. Here, the non-reference allele frequency among unrelated i2QTL donors is used, and the density is plotted with orange indicating more observations, and blue fewer.

To assess how similar genotyping sensitivity was between studies, and confirm that overlapping sites were likely to have the same breakpoint, we compared the non-reference allele frequencies of sites that we classified as co-discovered. We found that overall the non-reference allele frequencies of i2QTL

variants were highly correlated (r >0.9) with their matched GTEx and 1KGP variants (Figure 1.4D,E). This was true across variant classes in both studies, with the exception of duplications in 1KGP, which were less correlated (r=0.74), likely as a result of limited genotyping sensitivity in 1KGP due to the use of low coverage WGS data. These results support that the i2QTL SV call set is accurate and contains most common SVs discoverable using short read sequencing data as well as novel, rare SVs, making it a valuable resource for examining functional differences between the SV classes [49].

STR Genotyping

We genotyped STR variants at over 1.6 million reference sites using HipSTR [38] which employs a hidden Markov model to realign reads around each STR locus (Figure 1.1A). HipSTR models PCR stutter artifacts to genotype STRs and because of such artifacts, greater genotyping sensitivity and accuracy of predicted *de-novo* STR alleles can be achieved with PCR-free WGS data. In light of this, HipSci samples, which were generated with a PCR-free library preparation, were genotyped separately and then these alleles were used as a reference to genotype iPSCORE samples, which were prepared using a PCR-based library prep, and the results for both sample sets were combined into one call set with consistent alleles. To retain only high-quality STR calls, we applied the genotype specific filters suggested by HipSTR [38] and required all sites to have an 80% call rate in iPSCORE or HipSci samples. This resulted in 588,189 autosomal variants with high reproducibility across the range of gentotyped expansion/deletion sizes (1 bp to ~150 bp) (overall 94.5%, > 90% in all size bins); these variants were substantially more reproducible than indels in this same size range called by GATK in the i2QTL call set, which overall showed low replication rates (62%) (Figure 1.3E-F). Because HipSTR STRs and GATK indels overlap in size and location, it is likely that some variants are present in both datasets. To compare the genotyping quality of these possibly redundant variants, we intersected GATK indels with 1.6 million HipSTR STR

reference loci. Interestingly, we found that indels (2-100 bp) called by GATK that overlapped an STR locus that was genotyped non-reference in at least one sample by HipSTR had higher RR (77.3%) than those that overlapped STR loci not genotyped as polymorphic by HipSTR (56%), or those that did not overlap an STR region (64.7%). These findings suggest that it is useful to filter large GATK indels (>30bp) because they have low RR (42%), and that STR genotypes are more reproducible than GATK indels.

Linkage disequilibrium tagging for SVs and STRs

Given the large amount of GWAS and QTL studies performed using genotyping arrays, we next asked to what extent different classes of SVs and STRs are tagged by SNPs and indels. We calculated the maximum linkage disequilibrium (LD) between common SVs and STRs (NMAF > 0.05) within 1Mb of an expressed gene in iPSCs and SNPs/indels (Methods) [49] within 50kb of the SVs/STRs. We found that 97.7% of STRs are tagged by a SNP or indel with $R^2 > 0.8$ while SVs classes ranged from 44.2% to 86.7% of variants tagged with $R^2 > 0.8$ (Figure 1.5). Duplications and mCNVs were the most poorly tagged classes likely because they are often located near segmental duplications where SNPs and indels are poorly genotyped [18,29,33]. These results indicate that most common STRs and some classes of SVs are assayed well by proxy using SNP and indel genotypes, but to increase the coverage of SVs, particularly mCNVs and duplications, studies need to include the genotyping of these variant classes in their samples.

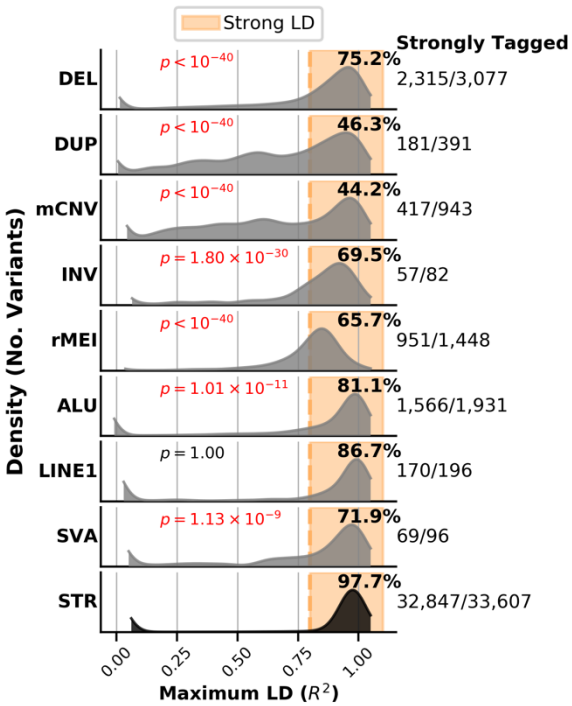


Figure 1.5 Linkage disequilibrium of structural variants and short tandem repeats with nearby SNVs and indels
Distribution of maximum linkage disequilibrium (R^2) between common SVs and STRs (non-mode allele frequency > 0.05) and SNVs or indels within 50kb, considering only SVs/STRs that are within 1MB of an expressed gene in iPSCs.

Discussion

In this study, we discovered and genotyped SVs and STRs in 719 high-coverage WGS samples from 477 unique donors. We detected a wide spectrum of variants across different sizes as most STRs are in the 10bp to 1kb range whereas SVs may span more than 100 kb. We leveraged genetic replicates, such as twin pairs and fibroblast-iPSC matched samples, to test variant calling accuracy and determine filtering approaches to retain only high-quality SVs and STRs. Our filtered call set has very high replication rate (on average >90% for all SV callers), indicating high genotype quality for detected SVs and STRs. The call set captures most of the common variants identified in 1KGP [18] and GTEx SV variant calling efforts and also contributes novel short (~100-1,000 bp) and rare (NMAF < 5%) variants. The high quality, non-redundant i2QTL SV set described here will serve as a useful reference for other studies and is

particularly valuable for genetic association analyses that aim to identify SVs that influence disease risk or quantitative molecular traits like gene expression.

We used five algorithms designed for calling variants across many samples to detect different classes of SVs and STRs and compared the RR in genetic replicates (MZ twin pairs and fibroblast-iPSC pairs) to identify factors that impact RR. We found that we needed to call variants separately in the iPSCORE and HipSci WGS collections and implement specific filtering strategies to account for dataset-specific features such as library preparation techniques to achieve high RR. Given the variability in library preparation methods, future improvements to SV calling algorithms may explicitly adjust for specific library features such as PCR-free sequencing. We also observed differences in RR between different classes and sizes of SVs and different algorithms. We found that SVs in the 100-1,000bp range remain harder to identify and genotype likely due to the limitation of using short reads. We also observed that accuracy was highest for large (>10 kb) duplications, deletions, and mCNVs suggesting that FDR estimates from orthogonal data sets such as arrays may overestimate accuracy for SV call sets since they generally assess the largest and easiest-to-genotype variants. Future studies that combine deep short read WGS with long read sequencing data may be able to improve the detection and genotyping of SVs in the 100-1,000bp range by directly sequencing them or assembling the short and long reads.

We used genetic replicates to identify algorithm- and SV-specific thresholds and applied these thresholds to filter the initial set of SV calls and create a high quality catalog of SVs and STRs that complements previous SVs identified using low depth WGS or fewer samples [18,19]. We also developed approaches for collapsing redundant SVs and harmonizing SVs called by different algorithms across hundreds of samples. Comparing our SV catalog to previous sets of SVs from the 1KGP and GTEx projects shows that the i2QTL SV call set captures most common (NMAF > 0.05%) SVs in Europeans. However, consistent with others types of genetic variants, we found that African ancestry samples had

more SVs than Europeans. Future sequencing studies are needed to fully catalog SVs in other ancestries and identify rare, population-specific SVs. Such multi-ancestry SV catalogs will be indispensable for population sequencing studies such as All of Us [50] that aim to integrate genetic and health data for patients from diverse and admixed ancestries.

The filtering scheme and catalog of SVs and STRs presented here can be used in future genetic association and sequencing studies that aim to study the impact of SVs/STRs. One method for utilizing this catalog for calling SVs and STRs is to impute variants via tagging SNPs and indels; a benefit of this approach is that imputation is possible using both array- and sequenced-based genotyping. A second option when sequencing data is available is to skip the de novo SV and STR discovery step and instead genotype the reproducible variants reported here. This will restrict genotyping to high-quality sites and may lessen the burden of filtering variant calls. A third option is to perform de novo discovery, genotyping, processing, and filtering using the approaches and thresholds that we have identified. While it may be possible that some filtering thresholds need to be adjusted for specific studies, the thresholds provided here likely provide a good starting point for genotyping and filtering de novo discovered SVs and STRs in other datasets.

Overall, this study provides a roadmap for discovering and genotyping SVs from WGS data and establishes a high-quality catalog of SVs and STRs that can be used in future genotyping efforts. A companion paper [49] examines how the i2QTL SVs and STRs characterized here influence gene expression and contribute to disease risk. These studies demonstrate that SVs and STrs can be reliably identified and genotyped for hundreds of samples and used to study the impact of this class of genetic variation on human health.

Acknowledgments

This work was supported in part by the National Science Foundation, CIRM grant GC1R-06673-B, and NIH grants HG008118, HL107442, DK105541 and DK112155. D.J. was 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. M.J.B. was supported by a fellowship from the EMBL Interdisciplinary Postdoc (EI3POD) program under Marie Skłodowska-Curie Actions COFUND (grant number 664726). S.B.M. was supported by NIH grant U01HG009431.

Author Contributions

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

Conflicts of Interest

The authors have no conflicts of interest.

Methods

Abbreviations

1KGP: 1000 Genomes Project

Indel: Small insertion/deletion variant

SV: Structural Variant

SNV: Single nucleotide variant

SNP: Single nucleotide-polymorphism

WGS: Whole-genome sequencing

FDR: False discovery rate

MAF: Minor Allele Frequency

NMAF: Non-Mode Allele Frequency

MSQ: Median Sample Quality

Variant Callers

- SS: SpeedSeq SV pipeline (LUMPY read-pair evidence with read depth support from CNVnator)
- GS: Genome STRiP CNVDiscovery pipeline (read depth evidence),
- GS LCNV: Genome STRiP LCNVDiscovery pipeline (read depth evidence)
- MELT: MELT mobile element insertion discovery
- HipSTR: HipSTR short tandem repeat genotyper

Types of Genetic Variants Detected

- DEL: Biallelic deletion ascertained by LUMPY, GS, GS LCNV
- DUP: Biallelic duplication ascertained by LUMPY, GS, GS LCNV

- mCNV: multiallelic copy number variant ascertained by LUMPY, GS, GS LCNV. This is defined as a variant that has at least 3 predicted alleles.
- INV: inversion ascertained by LUMPY
- rMEI: reference mobile element insertion
- BND: generic "breakend" ascertained by LUMPY. May include deletions and duplications that lack read-depth evidence, balanced rearrangements (INV), MEI or other uncategorized break points.
- ALU: Non-reference Alu element insertion identified by MELT
- LINE1: Non-reference LINE-1 element insertion identified by MELT
- SVA: Non-reference SVA (SINE-R/VNTR/Alu) element insertion identified by MELT
- STR: short tandem repeat variant, detected by HipSTR. Included variants have at least one individual with a change in length from the reference.
- CNV: copy number variant (deletion or duplication structural variant). Encompasses DEL, DUP, mCNV
- MEI: Non-reference mobile element insertion ascertained by MELT, including ALU, LINE1, and SVA elements

Subject enrollment

273 subjects were recruited as part of the iPSCORE study, of which 215 subjects have been previously described [39-41]. Data for additional 204 subjects was obtained from the HipSci Collection [42,43]. The iPSCORE collection was approved by the Institutional Review Board of the University of California at San Diego (Project #110776ZF). Each of the subjects provided consent, filled out a questionnaire, had blood drawn, and had a 1 mm skin biopsy taken from which fibroblasts were obtained. Five individuals

provided consent only for cardiovascular studies, therefore they were removed from downstream analyses. Family relatedness, sex, age, and ethnicity were recorded in the questionnaire. Detailed pedigree information for iPSCORE available in Panopoulos et al. [39-41] (dbGAP: phs001035). In total, we utilized a total of 477 HipSci and iPSCORE subjects, 276 were females and 201 were males, and collectively subjects ranged in age from 5 and 89 years of age. Notably, iPSCORE individuals were included in 56 families composed of two or more subjects (range: 2 to 14 subjects) and 86 single individuals. Overall, 167 iPSCORE individuals were unrelated. All iPSCORE individuals were grouped into one of five superpopulations (European, African, Admixed American, East Asian, and South Asian) on the basis of genotype data [39-41] and HipSci samples were similarly categorized [42]. For HipSci, some subjects had multiple iPSC clones with WGS. For these subjects, we chose the pair of fibroblast and iPSC WGS samples that had the highest reproducibility for Genome STRiP calls.

WGS processing

IPSCORE: WGS sequencing for iPSCORE individuals has previously been described in detail DeBoever et al. [40] and is available on dbGaP (dbGAP: phs001035). DNA isolated from either blood (254 samples) or fibroblasts (19 samples) (Figure 1.1A), was PCR-amplified and sequenced on Illumina HiSeqX (150 base paired-end). We obtained an average of 180.9 billion total raw bases per sample (range 117.81 to 523.49 billion bases). The quality of raw fastq files was assessed using FASTQC [51]. Reads were then aligned to the human b37 genome assembly with decoy sequences included and a Sendai virus contig with the BWA-mem algorithm under default parameters [52].

HipSci: We downloaded cram files associated with 446 genomes (mean depth 36.3X) generated with a PCR free protocol from 204 healthy donors (ENA Study Accession: ERA828) [42]. Genomes were previously aligned to hs37d5 genome, a reference identical to the one used for iPSCORE alignments with

the exception of the inclusion of a Sendai virus contig. Cram files were converted to the bam file format and merged when necessary using samtools[53].

Bam files from both iPSCORE and HipSci were sorted with sambamba [54] and duplicates were marked with biobambam2 (https://gitlab.com/german.tischler/biobambam2).

Code availability

Code used for analyses and variant processing can be found on GitHub (https://github.com/frazer-lab/i2QTL-SV-STR-analysis).

Replication rate and filtering strategy for SVs and STRs

To minimize the number of poorly genotyped structural variants (SVs) and maximize quality across multiple variant calling approaches, we used the replication rate (RR) metric, calculated as the proportion of non-reference genotypes that were also called non-reference in a paired genetic replicate, as a measure of the reproducibility (and quality) of a variant. The rationale behind this approach is that variants that have high genotyping accuracy should be genotyped consistently in different samples with the same genome and that variants with low genotyping accuracy will differ between samples with the same genome. Under this logic, variants should be consistently genotyped in samples with the same genomes (e.g. technical duplicates, monozygotic twins) and discrepancies would result from false negative or false positive genotypes.

To determine RR for all variant classes, we used genetic duplicate samples in the form of monozygotic twin pairs (n=25) and fibroblast iPSC pairs (n = 152). We used RR to assess the reproducibility of variants under different filtering conditions; the filters were specific to the unique quality metrics measured by each calling algorithm. Using the relationships between filters and RR that

we identified, we selected filtering criteria for each variant class in each caller to maximize the quality (specificity) and the number of variants (sensitivity) called. Because there may be a greater number of somatic variations between fibroblasts and iPSC clones [39] due to reprogramming, replication rates in monozygotic twins were used to select thresholds, and iPSC-fibroblast pairs were used for additional confirmation. For this analysis, one member of each pair of genetic duplicates was chosen arbitrarily as the "comparison sample", and the concordance of non-reference sites in this sample was assessed with respect to the other sample. Replication rate was calculated on all autosomal SVs on a site-by-site basis as the number of pairs with matching non-reference genotypes divided by the total number of pairs with at least one non-reference genotype. Average RRs reported for particular SV classes were calculated as the average RR over all SVs in that class.

Batch effects, Hardy Weinberg equilibrium, and sample consistency analysis

The i2QTL Consortium includes WGS data from iPSCORE and HipSci [42,55], which are different in aspects which may affect variant calling: 1) mean coverage is higher for iPSCORE (50.4X, compared with 36.6X); 2) while most iPSCORE donors had WGS from blood and only 14 from skin fibroblasts, all HipSci donors had WGS from skin fibroblasts; and 3) HipSci samples were sequenced using a PCR-free protocol (Figure 1). To limit the batch effects associated with these differences, in cases where a variant caller used information from the entire set of samples to build a global model (Genome STRiP [29] and HipSTR [38]), we genotyped or performed discovery separately in iPSCORE samples and HipSci samples, which were additionally divided into two groups for fibroblast and iPSC samples.

We compared allele distributions for autosomal variants ascertained for unrelated members of each collection (167 unrelated iPSCORE samples and 204 HipSci samples) after variant calling and filtering to ensure that differences between WGS from each collection did not create widespread

systematic artifacts in variant calling. Allele distributions were compared between the studies using a chi-squared test with a Bonferroni correction. For instance, for an insertion, the number of samples with zero, one, or two copies of the insertion in iPSCORE were compared to the number of samples with zero, one, or two copies of the insertion in HipSci using the chi-squared test. Variants with Bonferroni-corrected $p<0.05$ were tagged in the VCF file. For this analysis, missing genotypes were also included as a unique allele when present.

We also calculated Hardy Weinberg Equilibrium to identify variants that could be affected by batch effects in variant calling or that were poor quality. We used all unrelated blood/fibroblast samples and considered autosomal biallelic duplications and deletions from Genome STRiP [29] as well as all variant classes ascertained by SpeedSeq [46] and MELT [47]. We tested HWE using a chi-squared test to compare the counts of the observed genotypes to those expected given HWE. SVs with Bonferroni corrected HWE $p<0.05$ were flagged as potentially not obeying HWE.

Consistency in the number of non-reference calls per sample is associated with variant calls from high-quality WGS sequencing data, samples of similar ancestry, and algorithm performance. We counted the number of calls per sample for all algorithms to assess whether there were differences in the number of SVs identified in samples from each study, ancestry, or cell type from which the WGS was derived.

SpeedSeq variant calling and quality control

We used the split and discordant read pair-based structural variant caller LUMPY (v0.2.13)[56] under its implementation in SpeedSeq (v0.1.2)[46] to call duplications, deletions, inversions and other novel adjacencies referred to as "breakends" (BNDs). We ran LUMPY on each of 719 samples (446 from the HipSci collection and 273 from the iPSCORE collection) using the "speedseq sv" command with the -P option to retain probability curves in the output VCFs, -d to CNVnator (v0.3.3)[26] to calculate absolute

copy number information on each sample, and -x to exclude a published list of genomic regions (ceph18.b37.lumpy.exclude.2014-01-15.bed) known to be potentially misassembled regions [56,57]. Calls from individual samples were then genotyped using SVTyper (v0.1.4), before being combined into a single VCF file. Individual VCF files were sorted, and merged using svtools (v0.3.2) with the "sort" and "merge" command (slop 20bp) to remove overlapping breakpoints, resulting in a single VCF file with the most probable sites. Each sample was then genotyped at these merged sites using SVTyper and annotated with an absolute copy number using the "svtools copynumber" command. Variants were merged into a single VCF file, pruned, and reclassified under suggested parameters[19]. Individual VCFs were merged using "svtools vcfpaste" and further processed to remove additional identical variants using "svtools prune". This set of breakpoints was then reclassified by using "svtools classify" to identify high confidence copy number variants by regressing the estimated copy number and "allele balance" information (non-reference/reference reads at an SV site) as well as to identify mobile element insertions in the reference genome (rMEI, which appear as deletions in our call set).

Because metrics such as replication rate may select variants that are reproducible artifacts, to remove as many known low-quality sites as possible, we first applied filtering guidelines suggested in a previous study [19] as follows : 1) deletions that were less than 418bp were required to have split read support; 2) all non-BND variants were required to be at least 50bp in length; 3) BND calls required 25% percent support from either split or paired end reads; and 4) QUAL > 100 inversions were required to have at least 10% of evidence from split or paired end reads. Finally, to ensure a baseline level of genotyping consistency at each site, variants were filtered if they had a missing rate of > 10%.

After running the Speedseq/SVtools pipelines and filtering variants as described above, the variant call set still contained overlapping variants suspected to be identical. To produce a single set of non-overlapping unique calls, we performed additional pruning steps. To identify and prune putatively

identical calls that remained in our call set we implemented a graph-based approach: 1) we constructed a graph where SVs with reciprocal overlap of at least 50% are nodes connected by an edge; 2) we created a correlation matrix for each set of connected components using the allele balance (non-reference/reference reads at an SV site) at each site across individuals; 3) we refined the graph, retaining only the edges between SVs with $r > 0.25$ at a given site, which are likely to represent a single breakpoint; 4) we iterated through connected components, and chose variants with the highest median sample quality (MSQ) score, pruning other variants in the subgraph; and 5) in cases where one call was fully contained within another call and there was a correlation of at least 0.5 in allele balance between them, indicating that both calls were genotyped as non-reference in the same individual(s), only the site with the highest MSQ score was retained.

During SpeedSeq quality analysis we investigated supporting reads (SU) and median sample quality (MSQ) as possible filtering criteria. MSQ was strongly associated with RR in iPSCORE twins and HipSci fibroblast iPSC pairs while the number of supporting reads was not (data not shown). For variant filtering, we determined variant class-specific MSQ thresholds, with the goal of ensuring at least 90% replication rate across all variant classes and retaining the maximal number of variants. Classes of variation that were highly reproducible before quality score filtering (>90% replication rate) were filtered at a 20 MSQ score, a threshold used in previous efforts [19]. With this approach, we performed additional filtering as follows: 1) deletions and rMEIs must have MSQ > 20; 2) duplications, inversions and BND calls must have MSQ > 100, 90, and 90 respectively. Deletions and rMEIs were genotyped most reproducibly, prior to filtering, while duplications were less reliably genotyped, reflecting the sensitivity of split read versus discordant read signal. After filtering, RR was on average 97% in twin pairs and slightly worse (92%) in fibroblasts-iPSC pairs.

We tested variants on autosomes that passed the filters described above and 196 variants with missing rate > 10% but that otherwise passed filters for differences in allele distribution or deviations from HWE as described above. We found that only 544 of 25,537 sites tested had different allele distributions (2.1%). We also observed that 1,256 variants (4.9%) deviated from HWE, suggesting that batch effects do not affect SpeedSeq variant calls. We also observed that allele frequencies were highly correlated between variants detected in iPSCORE and HipSci.

After variant calling, we found that the number of SVs identified was consistent across samples, regardless of the study or cell type. In agreement with previous SV discovery studies, we observed on average 10.2% more SpeedSeq variants per sample for those of African ancestry (4,260/sample) [18,19] as compared to samples that were not predicted to be of African ancestry (3,863/sample).

Genome STRiP CNVDiscovery variant calling and quality control

Genome STRiP (svtoolkit 2.00.1611) CNVDiscovery [29], a population level read-depth based caller, was used to identify and genotype biallelic duplications and deletions as well as multiallelic CNVs (mCNVs) with suggested discovery parameters for deeply sequenced genomes (window size: 1000bp, window overlap: 500bp, minimum refined length: 500bp, boundary precision: 100bp, reference gap length: 1000). Because Genome STRiP is sensitive to differences in cell replication rate between samples derived from different cell types as well as in sequencing depth, we ran CNVDiscovery separately for iPSCORE fibroblast and blood samples and HipSci fibroblast samples. At the midstage of Genome STRiP discovery, 10 iPSCORE samples and 6 HipSci samples were removed from their respective discovery runs due to excessive variation in the number of calls per sample (exceeding the median call rate across all samples plus 3 median absolute deviations). To produce a call set where all sites were genotyped in all samples, sites discovered in either iPSCORE or HipSci samples were next genotyped

using SVGenotyper in the opposite set (genotyping separately within these respective sets) and the combined list of discovered sites was genotyped in the remaining HipSci iPSC samples, which were excluded from discovery. Using this strategy, the Genome STRiP dataset was not biased by the presence of somatic CNVs in iPSCs, and differences due to WGS library preparation specific to each study were minimized. Additionally, output VCF files from genotyping each subset of samples were annotated to match those from variant discovery using the SVAnnotator ("-A CopyNumberClass, \ -A CNQuality \ -A VariantsPerSample \ -A NonVariant \ -A Redundancy"), to ensure that quality metric information was available for each variant within each subset of samples for downstream processing.

A commonly suggested filtering parameter for SV detection is the per site quality score GSCNQUAL, described as being comparable for filtering of both duplication and deletion events [58]. We thus tested the RR of Genome STRiP variants ascertained in iPSCORE samples as well as the replication of variants ascertained in the HipSci fibroblast samples. Here we found that GSCNQUAL was highly correlated with RR in both twin pairs and iPSCs, but duplications and mCNVs had higher RR among twin pairs than iPSC-fibroblast pairs. Furthermore, deletions in both iPSCORE and HipSci sites were more reproducible under less stringent filtering than duplications and mCNVs. We selected 2, 12, and 14 as the minimum GSCNQUAL score required for deletions, mCNVs and duplications, respectively. We then filtered variants that were monoallelic in the data set as well as sites that had more than 10% of non-iPSC genotypes marked as low quality (LQ format field). These standard filters were applied before proceeding to combine the discovery sets of iPSCORE and HipSci and other data processing.

To collapse redundant variants that were obtained through separate SV discovery for iPSCORE and HipSci samples, we first filtered the HipSci discovery set and the iPSCORE discovery set to those passing filters described above, and then intersected the call sets using bedtools[59,60]. Overlapping sites were required to meet the following criteria in order to be considered redundant: 1) at least 50%

reciprocal overlap; 2) Pearson correlation coefficient in the copy numbers of non-iPSC samples > 0.95; and 3) differences in less than 5% of non-mode genotypes in non-iPSC samples. To process these overlaps, we considered cases where two sites exactly overlapped (same coordinates), choosing the site with the largest sum of GSCNQUAL scores from iPSCORE and HipSci (Non-iPSC) samples sets as the high quality "primary" site and marking the other as "redundant". Pairs of sites with exact overlaps were then removed from the analysis, and the remaining intersections were processed using a graph-based method similar to the one developed for Speedseq. Briefly, overlapping sites (nodes) were connected by edges weighted according to the average percentage overlap (the average of the percentage overlap of site B with A and the percentage overlap of site A with B) and which variant had the largest sum of GSCNQUAL scores from iPSCORE and HipSci (Non-iPSC) samples. Then, we iterated through connected components of the graph; chose the pair of sites that had the highest average overlap; and marked the variant with the largest sum of GSCNQUAL scores as the "primary site" and the other variants in the cluster as redundant. For the X chromosome, the computation of correlation and differences among non-mode samples was done separately for males and females, requiring that sites pass criteria in males, females or both males and females, depending on whether each subgroup had variability. This was done to control for bias in correlation coefficients due to the difference in reference copy number for males and females on the X chromosome. Overall, this process resulted in 7,987 sites being reduced to 3,856 non-redundant primary sites.

Genome STRiP occasionally reports a single CNV as several adjacent CNVs [19]. To address this issue, we analyzed sites that passed filtering, and were non-redundant, computing the correlation and distance between every pair of adjacent sites. We observed high genotype correlation between sites that overlapped or were close to each other (within ~40kb). Pairs of sites were considered for stitching into a single CNV if they had high overall correlation ($r > 0.9$) between copy number genotypes and at least

80% concordance between copy number genotypes of non-mode samples for each variant (union). Because variants that are very far from one another are less likely to be fragmented variant calls, we also selected a maximum distance between a pair of variants to consider for stitching. To do so, we examined the number and percentage of adjacent variant pairs that passed genotype correlation requirements at different distance thresholds, and selected 30 kb, which maximized the number and percentage of pairs passing these requirements. We then identified correlated adjacent CNVs to be stitched using a graph-based method: 1) a genotype correlation matrix was created for all the CNVs on each chromosome using estimated copy numbers across samples; 2) a graph was drawn with CNVs as nodes, connecting a pair of CNVs with an edge if they resided on the same chromosome and had correlation from their copy number estimates >0.9; 3) for each connected component in the graph with more than a single CNV, CNVs were sorted by position and each adjacent pair was examined for potential stitching; and 4) CNVs were merged if they passed the correlation/concordance criteria described above and were within 30kb of one another. This approach ensured that only highly correlated adjacent CNVs were merged. In cases where a set of CNVs was chosen to be stitched, a new breakpoint spanning the start point of the first CNV to the end point of the last CNV (sorted by start point) was defined, referred to hereafter as the "stitch" breakpoint, while the other CNVs in the cluster were considered "constituent" sites. Note that in cases when a stitch cluster was made up of a single CNV containing one or more smaller CNVs, the large CNV was identified as a stitch breakpoint. Overall, this process lead to 3,558 sites being combined into 1,252 putative "stitch" breakpoints, 355 of which were large breakpoints in the call set that contained smaller breakpoints, and 897 were new breakpoints. The set of 897 new stitch breakpoints (not already genotyped in our set), were then genotyped across all samples using Genome STRiP SVGenotyper (CNVDiscovery), separately for iPSCORE samples, HipSci fibroblast samples, and HipSci iPSC samples. Finally, we compared the genotypes of the stitched breakpoint with the genotypes of the constituent sites, and those

that did not have high correlation (average $r < 0.9$ across all constituents) were "unstitched", and if the stitch breakpoint was one of the 897 new breakpoints genotyped, it was marked for filtering. If the new stitched breakpoint had over 10% low quality flagged genotypes (LQ) or was non polymorphic, the stitch cluster was also unstitched, and the breakpoint marked for filtering.

The vast majority of new stitch breakpoints were closely correlated with the constituents (862/897, 96%), suggesting that our stitching strategy indeed identified single CNVs that were broken into fragments. An additional 7/862 correlated sites failed low quality genotype filtering criteria, yielding 855/897 (95%) new stitch breakpoints which passed all criteria. Overall, the process yielded 1207 unique sites (855 newly stitched sites and 353 sites that had been previously genotyped) comprised of 2-30 distinct CNVs each. For analysis of the non-redundant set, we filtered these constituent sites and retained the stitch breakpoints. After the filtering, deduplication, and stitching process, remaining non-redundant variants had high replication fractions in each individual twin pair and fibroblast iSPC pair and high average replication rates on a per site basis (Figure 1.3A).

After filtering, variant collapsing and stitching, we tested for differences in allele distribution and deviations from HWE as described above. Non-mode allele frequency was highly correlated between unrelated samples from iPSCORE and HipSci though a small number of variants (276/10,302 autosomal CNVs) were identified as having possible differences in allele distribution or deviation from HWE.

After variant calling and collapsing, we observed approximately the same number of calls per sample among iPSCORE and HipSci fibroblast samples, and no notable outliers among them. As with other variant callers, we saw larger numbers of calls per sample among samples from the African predicted superpopulation (~28% more calls per sample). Additionally, we found a small number of low quality genotypes per sample on the samples from which we performed discovery. HipSci iPSCs have higher rates of low quality genotypes because they were excluded from filtering that of sites based on

their percentage of genotypes that were tagged as low quality (FORMAT = LQ) because they were genotyped separately and excluded from the CNVDiscovery pipeline. These results suggest that the discovery and genotyping approach was successful in preventing systematic batch effect variants.

Genome STRiP LCNVDiscovery variant calling and quality control

To identify CNVs longer than 100 kb, which we refer to as long CNVs (LCNVs) we used the LCNVDiscovery module of the Genome STRiP toolkit (svtoolkit 2.00.1611). This pipeline uses information from depth of coverage in fixed-size bins across the genome, and while sample normalization is performed across samples, individual samples are called separately. Prior to LCNVDiscovery, we generated depth profiles for all genomes using GenerateDepthProfiles with suggested parameters (maximumReferenceGapLength = 1000, profileBinSize = 10000). Then, similar to our approach in Genome STRiP CNVDiscovery, iPSCORE samples, Hipsci fibroblasts and HipSci iPSCs were processed separately when running the LCNVDiscovery module (maxDepth=50). We collected the calls from each sample and filtered them with the suggested parameters (NBINS ≥10 and a SCORE ≥1000). Sites that were entirely contained within the centromere or overlapped the entire centromere were removed and variant sites were required to have an absolute copy number greater than 2.75 or less than 1.25 for duplications and deletions, respectively.

Genome STRiP LCNVDiscovery identifies sites per individual sample, so it is necessary to identify redundant sites that are called in different samples. To find redundant CNVs representing a single breakpoint, sites with a reciprocal overlap of at least 80% were grouped into clusters and a single breakpoint spanning the minimum start position to the maximum end position of CNVs in the group was used to represent the merged site. Individual CNVs that were within these clusters were marked as merged constituents, and excluded from non-redundant set, while those that didn't overlap with CNVs

from another individual were considered unique variants that were present in only a single sample. Absolute copy number estimates were rounded in order to produce integer copy number estimates similar to Genome STRiP CNVDiscovery. We identified 73 redundant sites comprised of 2 to 19 CNVs detected in individuals. On average, twin replication rates of the filtered variants was > 75% but very few large common variants were identified. After filtering and collapsing variants, we obtained 432 unique LCNV sites, with 200 duplications, 166 deletions, and 66 mCNV (size range: 100 kb to 5 Mb). On average each individual had 4 large duplications and 3 large deletions.

MELT variant calling and quality control

Mobile element insertions (MEIs) were called using the Mobile Element Locator Tool (MELT) [47]. We used the MELT (v2.0.2) SPLIT workflow to discover, genotype, merge and annotate MEI calls for ALU, SVA and LINE1 elements. We also included discovered 1KGP MEI sites [18] as priors in "MELT GroupAnalysis".

While MELT does not output quantitative quality scores, it does flag variants that meet one or more of several criteria. These criteria include: 1) sites that overlap low complexity regions (lc), 2) have more than 25% missing genotypes (s25), 3) have a ratio of evidence for the left and right breakpoint (LP/RP) that is > 2 standard deviations from the ratio among all other sites (rSD), or 4) have a larger than expected number of discordant read pairs that are also split reads (hDP). We tested whether the flags, or combinations or flags, were associated with RR and found that filtering on all suggested criteria improved RR considerably for detected MEIs, raising it from below 0.6 to ~0.9 for ALU, LINE, and SVA elements. Among these quality metrics, filtering on low complexity resulted in the best improvement compared with the other individual filters; however, filtering on all quality tags was necessary to improve RR to 0.9. Additionally, MELT outputs a quality tranche score from 1-5 (defined as "ASSESS") that describes the

types of evidence used to determine the location of the insertion site. For example, the highest quality insertion sites are given a score of 5, and has a target site duplication sequence flanking the MEI supported by split reads. Filtering with higher quality tranche score thresholds also improved replication rate, either before or after filtering using all flags (data not shown). We chose to filter variants that that were flagged for any criteria, and also required a quality tranche score of 5, for maximum stringency and best RR improvement. After filtering, individual twin and fibroblast-iPSC pairs had high replication percentages (RR>0.9).

We tested all MELT variants for differences in allele distribution and deviation from HWE as described above and found that only 527/9,566 autosomal MEIs had differences in allele distribution (49/527) or showed deviation from HWE (492/527). Additionally, non-reference allele frequency in iPSCORE and HipSci was highly correlated ($r > 0.9$), suggesting batch effects did not influence MEI calls.

MELT variants were highly consistent in calls per sample in both studies, (mean 1,107 and 1,097 calls/sample in iPSCORE and HipSci fibroblast samples respectively) and in all cell types, while having very few missing genotypes (median 1/sample). We observed an increased number of ALU, LINE1, and SVA elements per sample in samples from individuals of African ancestry (1,144 ALU/118 LINE1/53 SVA per sample versus 952 ALU/ 105 SVA/ 45 SVA sample for Non-African samples from iPSCORE).

HipSTR variant calling and quality control

Short tandem repeat (STR) variants were genotyped using the HipSTR algorithm (v0.5.61) [38], on a set of 1,527,077 GRCh37 autosomal STR regions that were provided by the tool (https://github.com/HipSTR-Tool/HipSTR-references/raw/master/human/GRCh37.hipstr_reference.bed.gz). Because only HipSci WGS data was

PCR-free, special considerations were required to run HipSTR, as it uses PCR stuttering models to genotype repeats and assumes all WGS samples were generated using the same pipeline. For STR genotyping, PCR-free data produces more accurate genotypes, thus we first ran HipSTR at STR sites in all 446 HipSci samples under standard settings. Next, we genotyped the iPSCORE samples using the HipSci genotypes as references (--ref option). Finally, we genotyped iPSCORE samples separately without using the HipSci genotypes as reference alleles. We used only the diploid genotype option, as we lacked phased SNVs for all samples.

To filter HipSTR variants, we first used the supplied "filter_vcf.py" script with recommended thresholds for individual genotypes (min-call-qual = 0.9, max-call-flank-indel = 0.15, max-call-stutter = 0.15, --min-call-allele-bias= -2, min-call-strand-bias= -2). This procedure converts genotypes that do not pass these thresholds to "missing". We examined the number of variant calls per sample and the number of missing genotypes when variants were genotyped in iPSCORE, iPSCORE using HipSci reference alleles, and in HipSci samples. Among iPSCORE samples, we observed a median of 122,249 calls per sample in African ancestry individuals and 111,613 calls per samples in non-African ancestry individuals. While four samples from non-African ancestry individuals had a surprisingly large number of STRs, all but one individual self-reported as having partial African ancestry. iPSCORE genotypes at HipSci reference alleles had similar numbers of calls per sample (median 110,023/sample) compared to the genotypes from iPSCORE alone. African ancestry samples, however, had a smaller number of calls using the HipSci reference alleles likely because HipSci did not include African ancestry samples, so the African samples in iPSCORE were only genotyped for STRs discovered in Europeans. HipSci samples had about twice as many calls per sample (222,321/sample for HipSci fibroblast samples) compared to iPSCORE and fewer missing calls per sample, demonstrating that using PCR-free WGS provides better accuracy for STR genotyping. To obtain a high-quality set of STRs, we required >80% call rate for

variants from each subset. We excluded one iPSCORE sample from this missingness calculation that had more than 70,000 missing calls. This filter resulted in high replication rates (> 92%) in each twin pair for both genotyping methods in iPSCORE, and even higher replication rates (>95%) in fibroblast-iPSC pairs for HipSci genotyping likely due to more accurate STR genotyping in the PCR-free WGS. Overall, the replication rate before all filtering and after processing improved from ~78% to ~94.4% in iPSCORE twins (Figure 1.3B).

HipSTR genotypes were combined between iPSCORE and HipSci by creating a single combined VCF file using the HipSci genotypes and iPSCORE genotypes at HipSci alleles. We additionally added iPSCORE genotypes for STRs that were unique to iPSCORE to the VCF file.

Unifying SpeedSeq and Genome STRiP call sets

Since different variant callers may detect the same variants using different methods, we developed a strategy to integrate variants from Genome STRiP and SpeedSeq call sets that were likely to represent the same site. To approach this problem, we used a graph-based method similar to those used to identify duplicates within SpeedSeq and Genome STRiP prior to this step. To generate clusters of overlapping SVs, we first intersected our filtered Genome STRiP calls (redundant sites removed, GSCNQUAL filtered, stitching sites included, stitched constituents excluded) with filtered SpeedSeq variants (redundant sites removed, standard filters, MSQ filtered) and retained all SV pairs with >50% reciprocal overlap or where one site completely contains another. SV pairs were required to have the same SV types, with exception being that mCNVs were allowed to match with both duplications and deletions and deletions were allowed to match with rMEI (as they appear as deletions). We built a graph where edges were represented by connected SV pairs that pass these overlap thresholds and SV type compatibility parameters. We iterated through connected components, testing every combination of

elements in each connected component, and generating a new graph, connecting pairs of variants if they passed correlation thresholds between copy number genotypes (Genome STRiP) variants or allele balance ratios (SpeedSeq) at the sites. If the connected component contained a duplication and deletion from SpeedSeq and an mCNV from Genome STRiP, SV pairs were allowed to connect if their genotype evidence had a correlation (R^2) > 0, while other components required an R^2 > 0.5. We then iterated through connected components of this new graph and selected the highest degree variant (connected to the most other variants) from each caller with the highest quality score (GSCNQUAL for Genome STRiP, MSQ for SpeedSeq) from which we chose one variant as the "primary" variant and all other variants as "secondary." All variants in each cluster were marked with a cluster ID. In cases where a Genome STRiP deletion overlapped a SpeedSeq rMEI, the SpeedSeq variant was chosen as the primary site, and the Genome STRiP variant was assigned as secondary. In all other scenarios, the Genome STRiP variant was chosen as the primary variant and the SpeedSeq variant was the secondary due to the comparably higher replication rates for Genome STRiP variants and the granularity of having integer copy numbers.

This method assures that highly correlated variants with significant overlap are clustered together, and that generally, the larger, higher quality variants are chosen as representative primary sites. Sites that were assigned as primary sites from the intersection clusters, as well as unique variants from either variant call set that were not included in intersection clusters, were then selected to produce a non-redundant set of sites necessary for global analyses of SVs (Figure 1.4,5).

Estimating FDR in arrays using intensity rank sum test

To estimate the false discovery rate (FDR) of the merged CNV call set we used 216 MEGA_Consortium_v2 arrays available for iPSCORE samples to perform an intensity rank sum (IRS)

test to assess whether the SV genotypes after filtering agree with genotypes from array data. SNP arrays were analyzed using the Illumina GenomeStudio software (v2011.1) and were required to have an overall call rate of <97%. The log(R ratio) was obtained from the final report. We used the Genome STRiP Intensity Rank Sum Annotator to compare genotypes for a subset of the SV calls that were present in the 216 samples for which we had array data using the log R ratio as input. Before testing, the intensity matrix was first adjusted for covariates by regressing out the effects of batch and plate on a probe-wise basis using the statsmodels (v0.9.0) linear regression module. To assess our filtering strategy we tested 2,563 /15,437 SpeedSeq duplications and deletions, and 4,233/18,171 Genome STRiP CNVs that were present in at least one of the 216 individuals (before any filtering) and contained at least 3 probes and computed IRS FDR as in 1KGP [18]. Restricting our analysis to 2,376 filtered and deduplicated SpeedSeq variants with array probes, we observed that deletions and duplications had an FDR of 5.35% and 3% respectively. Similarly, among 1,848 filtered and deduplicated Genome STRiP variants containing array probes, we observed that deletions, duplications and mCNVs had an FDR of 5.4%, 7.8%, and 7% respectively. These FDR estimates were similar to those in 1KGP and GTEx, although the probe density of arrays limited the number of sites we could test.

Comparison of i2QTL SVs to 1000 Genomes Project and GTEx SV Call-Sets

To investigate the quality and completeness of our SV calls, we compared them to GTEx v6p SV calls [19] which used 147 deeply sequenced whole genomes (median 49.9X depth), and the robustly characterized 1000 Genomes Project Phase-3 call-set [18] derived from 2,504 shallowly sequenced samples (7.4X depth). While the GTEx call-set contains relatively few samples, the whole genome sequencing data and variant calling approach were similar to the approach used in i2QTL (Genome STRiP and SpeedSeq), and were thus used as a benchmark. Before analysis, we obtained VCF files with genotypes

from 1KGP phase 3 (link?) and GTEx V6p (dbGaP accession number phs000424.v7.p1). Phased genotypes from 1KGP SVs were converted to unphased genotypes using the alternative allele information to enable comparison with the unphased SVs from i2QTL and GTEx. This enabled us to compute non-mode allele frequency for 1KGP and GTEx SVs to match the frequency measures used in this study. Because of the significant diversity of the 1KGP cohort (26 populations, 70% European) as compared to i2QTL (6 subpopulations, 80% European), we filtered the 1KGP data to 1,755 European samples, and used variants present in at least one of these samples. For co-discovery analyses, we used non-redundant sites from i2QTL as well as variants that passed filters and were part of redundancy clusters to maximize the potential overlap between sets. To identify putative co-discovered sites between i2QTL and either GTEx or 1KGP, CNVs (DUP, DEL, mCNV), rMEI and inversions from each call-set were intersected using "bedtools intersect" and co-discovered sites were selected using the following approach: 1) excluding inversions, all variants were required to have at least 25% reciprocal overlap, or if one variant was fully contained within the other, it was required to span at least 20% of the larger variant; inversions were required to have 80% reciprocal overlap; 2) variant classes were required to match with the exception of mCNVs, which were allowed to match with either duplications or deletions; for BND sites, we considered breakpoints within 50bp of each other to be matching; and 3) because we included 1KGP MEIs as priors in our MELT pipeline, MEIs co-discovered with 1KGP were known, and did not require overlap analysis. For overlap reported with i2QTL, we computed the fraction of sites co-discovered by one or both call-sets, considering non-redundant clusters a single site.

SNV and Indel Calling

For SNV and indel genotype calling we followed the GATK [61] best practices (version 3.8 accessed June 2018). Unless otherwise mentioned settings for the tools are taken from the best practices

or left default. As described above, the HipSci WGS data was aligned to the GRCh37 [62] build of the human reference genome using bwa [52]. After alignment Picard was used to mark duplicates. GATK was used for indel realignment and base-recalibration, and genotypes were called using the GATK haplotype caller in GVCF mode. iPSCORE GVCFs were obtained from dbGAP (phs001325) and were used to perform joint genotyping across all iPSCORE and HipSci samples. We used GATK variant recalibration (TS filter level 99.0) to filter low quality genotype calls for the called SNVs and indels separately.

LD Tagging

For each of the 42,921 total non-redundant SVs and STRs that were within 1MB of an expressed gene in iPSCs [49], we used bcftools [53] to extract all SNPs 50 kb upstream and downstream. For each SV or STR, we calculated LD as the correlation (Pearson R^2) with the genotypes of each surrounding SNV or indel genotyped in i2QTL WGS and selected the variant with the strongest LD.

Chapter 1, in full, is a reprint of the material as it appears in bioRxiv, 2019, David Jakubosky, Erin N. Smith, Matteo D'Antonio, Marc Jan Bonder, William W. Young Greenwald, Agnieszka D'Antonio-Chronowska, Hiroko Matsui, Oliver Stegle, Stephen B. Montgomery, Christopher DeBoever, Kelly A. Frazer. The dissertation author was the primary investigator and author of this paper.

CHAPTER 2 GENOMIC PROPERTIES OF STRUCTURAL VARIANTS AND SHORT TANDEM REPEATS THAT IMPACT GENE EXPRESSION AND COMPLEX TRAITS IN HUMANS

Abstract

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 show that different SV classes and STRs differentially impact gene expression and complex traits. Functional differences between SV classes and STRs include their genomic 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 also identified a set of high-impact SVs/STRs associated with the expression of three or more eGenes via chromatin loops and showed 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 impact gene expression and human traits.

Introduction

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 [17-19]. 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 [63]. 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.

One important problem that has not been addressed by previous SV and STR studies is whether the diverse variant classes differ in their functional impact on gene expression [18,38,40,64]. Therefore, while SV classes and STRs vary in genomic properties including size, distribution across the genome, and impact on nucleotide sequences, it is unknown whether these differences result in the variant classes having differential impact on gene expression such as their likelihood of being associated with a gene (eGene), their effect size, and the types of associated eGene (e.g., coding, noncoding, level of evolutionary constrant). Further, it is unknown if the variant classes exert their effects on gene expression through different mechanisms such as directly altering eGene copy number or altering three-dimensional spatial features of the genome. A comprehensive SV and STR dataset generated using high-depth whole genome sequencing (WGS) from a population sample with corresponding RNA-sequencing data could be used to compare how the different genomic properties of SV classess and STRs correspond to differential functional impacts on gene expression.

SVs and STRs have also been associated with complex traits, though they have been studied considerably less often in GWAS than single nucleotide variants (SNVs), and the overall contribution of SVs and STRs to complex traits is not well understood [2,6,9,12,14,65-67]. One difficulty with studying differences between SV classes and STRs in GWAS is that it is unknown whether the variant 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 indels could be used to determine whether any particular classes of SVs and STRs are not currently captured by genotyping arrays and indicate “dark” regions of the genome

not assessed by array-based GWAS. A comprehensive set of variants with corresponding linkage disequilibrium (LD) estimates would also be valuable for assessing the functional impact of SVs and STRs on complex traits using existing SNV-based GWAS.

In this study, as part of the i2QTL Consortium, we used RNA-seq data from induced pluripotent stem cells (iPSCs) from the iPSCORE and HipSci collections [40,42,55] along with a comprehensive call set of SVs and STRs from deep WGS data [63] to identify variants associated with iPSC gene expression and characterize the genomic properties of these SV and STR eQTLs. We observed that SVs were more likely to act as eQTLs than SNVs when in distal regions (> 100kb from eGenes) and that duplications and mCNVs were more likely to have distal eQTLs and multiple eGenes compared to other SVs classes and STRs. eGenes for mCNV eQTLs were also less likely to be protein coding and more likely to have strong effect sizes relative to other SV classes and STRs. We examined LD of SVs and STRs with GWAS variants and found that mCNVs and duplications are poorly tagged by GWAS SNVs compared to other variant classes. 11.4% of common SVs and STRs were 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 were enriched for GWAS associations establishing that these variant classes have underappreciated roles in common traits. Finally, we found 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 multi-gene eVariants and associated with GWAS traits. This work establishes that different classes of SVs and STRs vary in their functional properties and provides a valuable, comprehensive eQTL dataset 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 ~48M SNVs and indels) generated using deep WGS from these same donors [63]. 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 $\geq$ 0.05) SVs (9,313), STRs (33,608), indels (~1.86M), and SNVs (~7M) within 1 megabase of a gene body using a linear mixed model approach (Figure 2.1A, Methods). 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%), 2,648 indels (24.3%), and 7,971 SNVs (73.1%, Figure 1B, 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 [18,19]. 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.

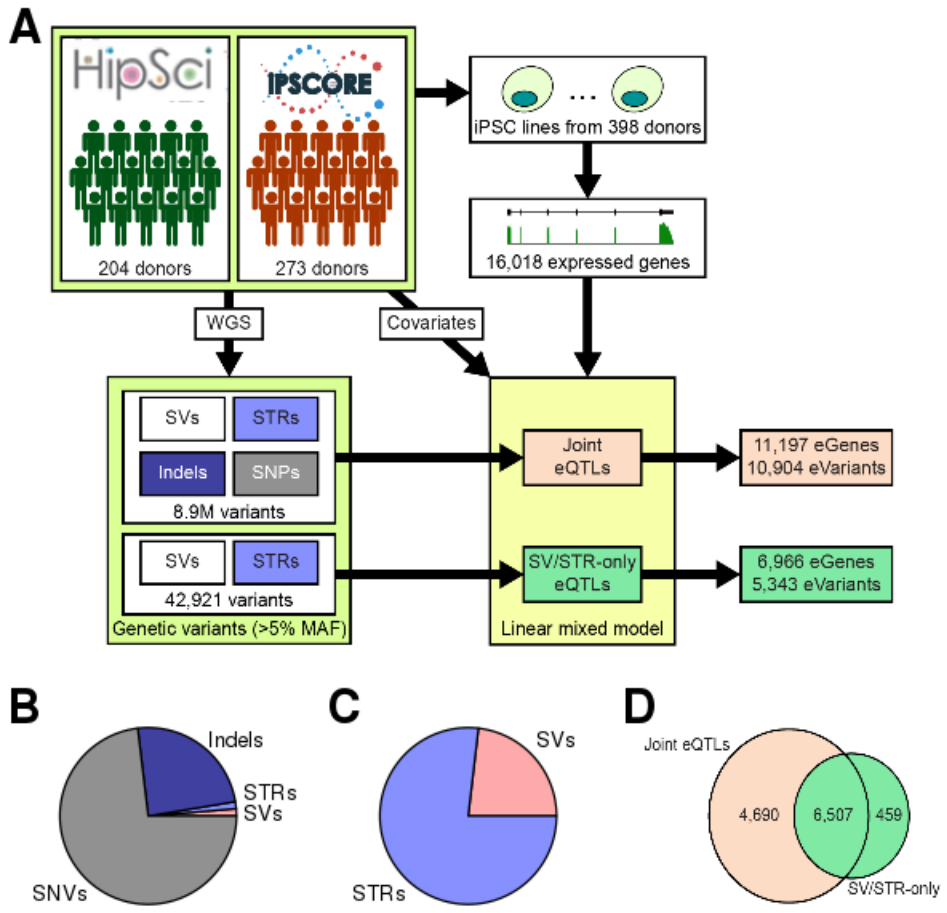


Figure 2.1 eQTL mapping
.(A) Overview of eQTL study design. We used SNPs, indels, SVs and STRs called using deep WGS from 477 donors[63] in conjunction with iPSCs reprogrammed from 398 of these donor[40,42,55] and detected 16,018 expressed genes. We performed two eQTL analyses: a joint analysis that 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.

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 (Figure 2.1A,B). We identified 6,966 eGenes (FDR<5%) associated with 5,343 unique lead eVariants (Table 2.1). SVs were more likely to be lead variants compared to STRs (OR=1.15, p=1.7e-5, FET) though the majority of lead eVariants were STRs (4,087 eSTRs versus 1,231

eSVs)(Figure 2.1C). Of the 11,197 eGenes identified in the joint eQTL analysis, 6,507 were also identified in the SV/STR-only eQTL analysis (Figure 2.1D). Among these 6,507 shared eGenes, 94.6% were mapped to a lead SNV or indel variant in the joint analysis, while the remaining 5.4% 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 in both analyses, 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 (R2> 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.

Table 2.1 Summary of i2QTL variants and eQTL results
Numbers in each category refer to the number of non-redundant variants that were within 1Mb 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.

	Variant Class	No. Variants	No. Common Variants	Lead SV/STR-Only QTLs	Lead Joint QTLs
	SNV	41,826,418	7,013,178		8,148
	INDEL	7,040,457	1,862,365		2,685
Copy Number Variants (CNV)	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
Mobile Element Insertions (MEI)	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*	*8,918,464*	*6,966*	*11,197*

Variant size influences eQTL associations

Given that SVs and STRs have size ranges that span orders of magnitude [63], 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 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 (Figure 2.2A). The trend was especially strong for deletions where 38% of variants longer than 50kb were lead eVariants (OR=3.11, p=0.0065, FET). Although a higher proportion of mCNVs were eVariants compared to other classes (Figure 2.2A), mCNV length was not

strongly associated with eQTL status; only mCNVs longer than 10kb were significantly more likely to be lead eVariants (OR=1.59, p=0.01, FET).

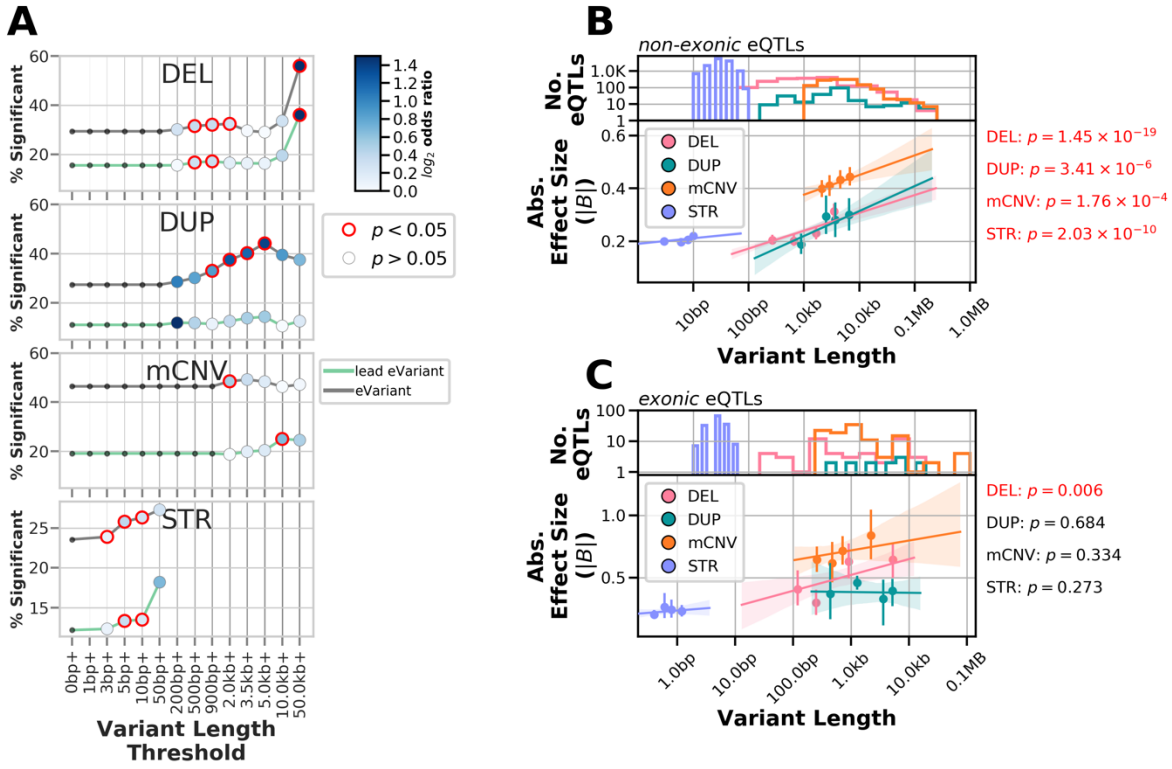


Figure 2.2 Variant length influences the likelihood and effect size of eQTLs
(A) The percentage of tested variants that were eVariants (grey lines) or lead eVariants (green lines) greater than or equal to threshold given on the x-axis for each variant class. Points are colored according to their log2 odds ratio for enrichment when comparing the fraction significant at or above the threshold to the fraction significant for variants smaller than the threshold; points circled in red were significant (FET, p<0.05). (B, C) Association of variant length with effect size for (B) non-exonic eQTLs or (C) exonic eQTLs mapped to biallelic deletions, duplications, multi-allelic CNVs and STRs. Number of eQTLs for each variant class at defined length is shown (top panels). Points represent the centers of bins with equal numbers of observations and error bars indicate 95% confidence intervals around the mean (1000 bootstraps) (bottom panels). 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) when also including non-mode allele frequency and distance to TSS as covariates.

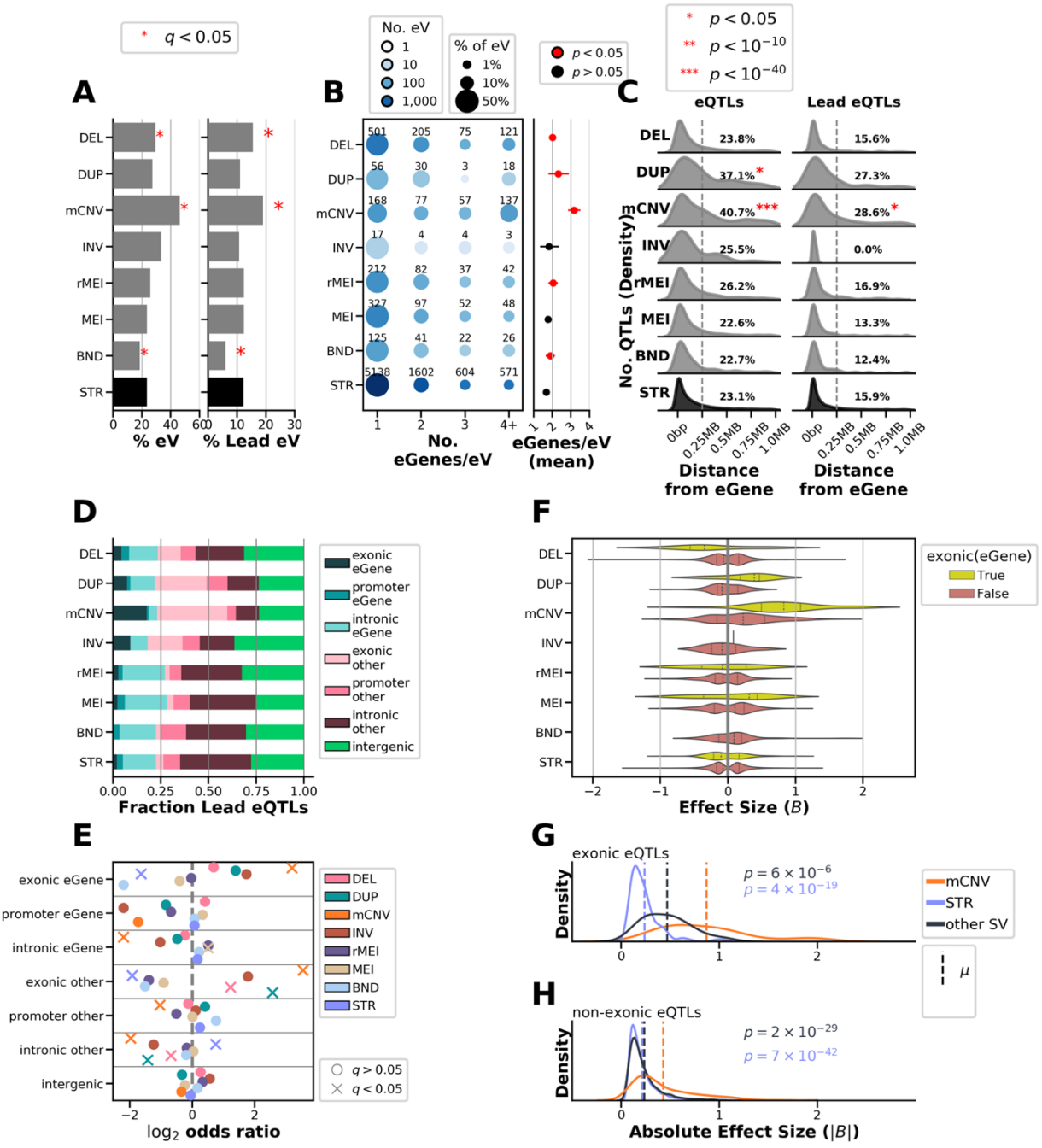
We next sought to examine whether eVariant length for SVs and STRs was predictive of absolute eQTL effect size and if lead eQTLs that overlap (exonic) or do not overlap (non-exonic) exons of the eGene displayed similar effects (Figure 2.2B-C). We found that lead eVariant length was significantly

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 (Figure 2.2B). 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 (Figure 2.2C). 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 (Figure 1C). We found that both mCNVs and deletions were more likely to be eVariants for at least one gene relative to STRs (mCNVs: OR=2.74, q=7.9e-50, FET; deletions: OR=1.31, q = 3.9e-10, FET) and were also more likely to be lead eVariants compared to STRs (mCNVs: OR=1.68, q=2.27e-8, FET; deletions: OR=1.32, q=3.7e-7, FET) (Figure 2.3A). Conversely, BNDs were less likely to be eVariants (OR=0.7, q= 4.2e-6, FET) or lead eVariants (OR=0.44, q=1.95e-12, FET) compared to STRs (Figure 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 affect multiple eGenes compared to STRs (Figure 2.3B), these effects were most prominent among mCNV eVariants and deletion eVariants which affected 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 7.6% (604/7,915) of STR eVariants (Figure 2.3B). These results show that mCNV and deletion eVariants are more frequently associated with the expression of multiple genes compared to STRs and other SVs.

Figure 2.3 Properties of SV and STR eQTLs
(A) Percentage of tested variants from each class that are eVariants (left) or lead eVariants (right) in the SV/STR eQTL. Asterisks indicate significant enrichment or depletion, comparing the likelihood of variants from a specific class to be eVariants to that of STRs (FET, BH alpha < 0.05). (B) Balloon plot with color showing the number of eVariants and size indicating the fraction of eVariants within the variant class represented in the bin (left), and the average number of eGenes per eVariant (right) with 95% confidence intervals. Red points indicate significantly higher numbers of eGenes/eVariant (Mann Whitney U Test, Bonferonni corrected p-values < 0.05) compared to STRs. (C) Distribution of the distance of eQTL (left) and lead eQTL (right) variants to the boundary of their eGenes (5' UTR or TSS). Proportion of eQTLs that were at least 250kb distal to eGene, red asterisks indicate that these percentages were significantly different from the other SV classes (Mann Whitney U Test, Bonferonni corrected p-values < 0.05). (D) For each variant class, the fraction of lead eQTLs that overlapped exons, promoters, or introns of their associated eGene, other genes, or that were intergenic is shown. (E) Enrichment (log2 odds ratio) comparing the proportions of lead eQTLs for each class that overlap each genic element with all other classes. (F) Distribution of effect sizes for lead eQTLs which overlapped or did not overlap an exon of their eGene. (G,H) Distribution of absolute effect sizes for mCNVs (orange), SVs excluding mCNVs (blue) and STRs (light blue) for exonic (G) and non-exonic (H) eQTLs. Annotated p-values show significantly higher effect sizes comparing the effect size distributions of mCNVs to STRs or other SVs, and are colored to match these groups (Mann Whitney U Test, Bonferonni corrected p-values < 0.05).


A
* q < 0.05
DEL
DUP
mCNV
INV
rMEI
MEI
BND
STR
0 20 40 60
% eV
0 10 20 30
% Lead eV
B
No. eV
1
10
100
1,000
% of eV
1%
10%
50%
p < 0.05
p > 0.05
501 205 75 121
56 30 3 18
168 77 57 137
17 4 4 3
212 82 37 42
327 97 52 48
125 41 22 26
5138 1602 604 571
1 2 3 4+
No. eGenes/eV
1 2 3 4
eGenes/eV (mean)
C
* p < 0.05
** p < 10^-10
*** p < 10^-40
eQTLs
Lead eQTLs
No. QTLs (Density)
23.8%
15.6%
37.1%*
27.3%
40.7%***
28.6%*
25.5%
0.0%
26.2%
16.9%
22.6%
13.3%
22.7%
12.4%
23.1%
15.9%
0bp
0.25MB
0.5MB
0.75MB
1.0MB
Distance from eGene
D
0.00 0.25 0.50 0.75 1.00
Fraction Lead eQTLs
exonic eGene
promoter eGene
intronic eGene
exonic other
promoter other
intronic other
intergenic
E
exonic eGene
promoter eGene
intronic eGene
exonic other
promoter other
intronic other
intergenic
−2 0 2
log2 odds ratio
DEL
DUP
mCNV
INV
rMEI
MEI
BND
STR
q > 0.05
q < 0.05
F
−2 −1 0 1 2
Effect Size (B)
exonic(eGene)
True
False
G
exonic eQTLs
Density
p = 6 × 10^-6
p = 4 × 10^-19
0 1 2
mCNV
STR
other SV
μ
H
non-exonic eQTLs
Density
p = 2 × 10^-29
p = 7 × 10^-42
0 1 2
Absolute Effect Size (|B|)

Genomic localization of SV and STR eQTLs

We next examined how eVariants for each variant class were distributed with respect to genes and promoters by evaluating the distance of eVariants to their eGenes and their overlap with genic elements. We found that, for all SV classes and STRs, most eQTLs were located near eGenes (<250kb, Figure 2.3C); however, a significantly larger proportion of eQTLs that were mCNVs or duplications were located far from their eGenes (>250kb) compared to STRs (mCNVs: 40.7%, OR=2.23, q<1e-40 ; duplications: 37.1%, OR=1.93, q=4.03e-5; FET) suggesting increased distal regulatory activity for these variant types. 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 (Figure 2.3D). 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. 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=9.15, q=4.6e-26, Figure 2.3D,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 (Figure 2.3C) 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, p<1e-40; duplications: OR=5.94,3.29e-6; deletions: OR=2.33, p=1.41e-8; FET); conversely, STRs were depleted in other gene exons (0.25%, OR=0.26, q=1.7e-16, FET, Figure 2.3D,E) and eGene exons (1.98%, OR=0.32, q=8.7e-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.

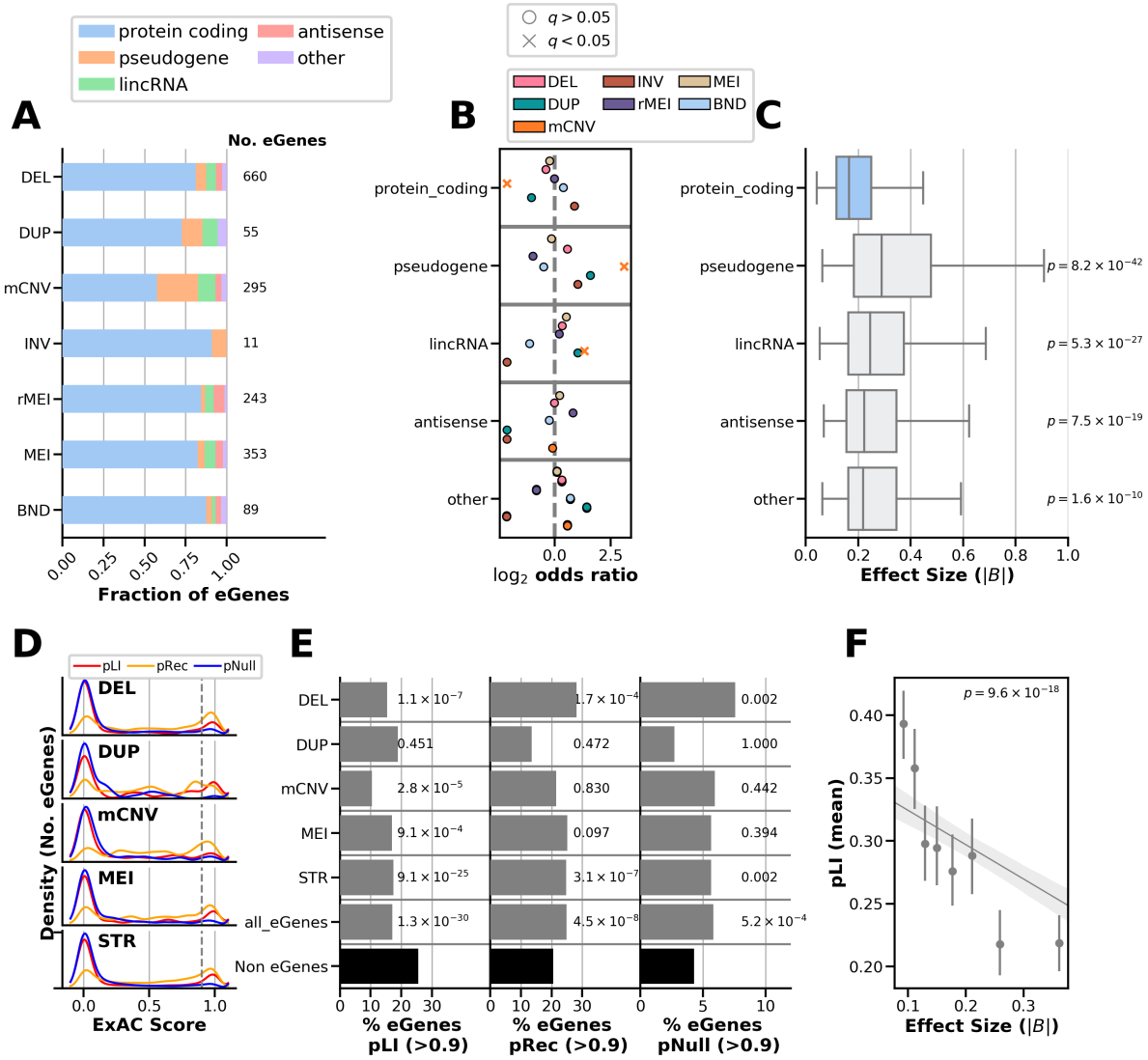
We next compared the direction and absolute effect sizes of lead eQTLs that overlapped or did not overlap exons of the eGene (exonic and non-exonic eQTLs) from each variant class to determine whether variants that alter gene copy number differ from variants that affect regulatory regions. Exonic and 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 (Figure 2.3F). Lead eQTLs mapped to all other variant classes 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 (Figure 2.3G) and non-exonic (Figure 2.3H) contexts compared to lead variants from other SV classes or to STRs (Figure 2.3G,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 (Figure 2.4A,B). Notably, a lower proportion of mCNV eGenes were protein coding (OR=0.22, q=3.33e-28, FET) and a higher proportion were pseudogenes (OR=8.56, q=1.04e-33, FET) or lincRNAs (OR=2.5, q=5e-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=1.84e-15, FET) and depleted for pseudogenes (OR=0.36, q=2.62e-16, FET) and lincRNAs (OR=0.62, q=1.9e-3, 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 (Figure 2.4C) which is consistent with non-protein coding genes being more tolerant of disruption [68]. 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.

Figure 2.4 Properties of eGenes associated with different variant classes
(A) Fraction of eGenes of each Gencode subtype mapped to lead variants of each class for the SV/STR only eQTL. (B) Enrichment $\log_2$ odds ratios for 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 in the SV/STR only eQTL. *p*-values indicate significance of Mann Whitney U test for difference in the effect size distributions of each category as compared to protein coding genes (Bonferroni corrected). (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 of each class that had high (> 0.9) pLI (left), pRec (center), or pNull scores (right). The percentage of non-eGenes (7,337 genes that were tested in the SV/STR eQTL but not significantly associated with an SV or STR) with high scores (black bars) is also included. Annotated 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 after regressing out the effects of variant class and mean $\log_{10}$(TPM) expression level of the gene among expressed samples. Points represent binned effect size bins with equal number of observations per bin and error bars showing 95% confidence intervals (n=1000 bootstraps) around the mean pLI for the bin. Line represents a linear regression predicting pLI by eQTL effect size with variant class and the $\log_{10}$(TPM) expression level of the gene among expressed samples as covariates, Here the *p*-value represents the significance of the eQTL effect size term (t-test).

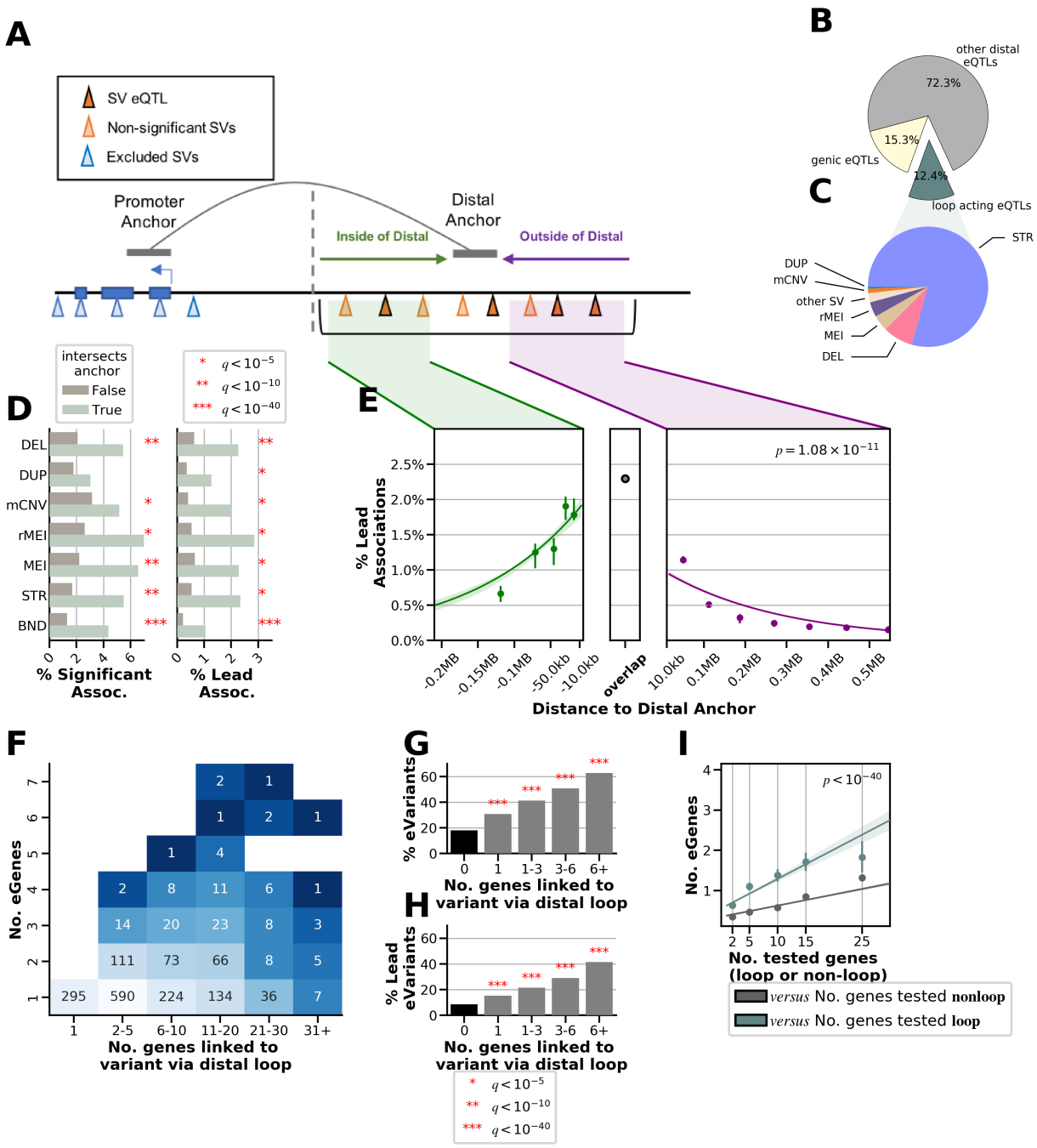

A
protein coding
antisense
pseudogene
other
lincRNA
No. eGenes
DEL
DUP
mCNV
INV
rMEI
MEI
BND
660
55
295
11
243
353
89
0.00
0.25
0.50
0.75
1.00
Fraction of eGenes
B
q > 0.05
q < 0.05
DEL
INV
MEI
DUP
rMEI
BND
mCNV
protein_coding
pseudogene
lincRNA
antisense
other
0.0
2.5
log2 odds ratio
C
protein_coding
pseudogene
lincRNA
antisense
other
p = 8.2 × 10−42
p = 5.3 × 10−27
p = 7.5 × 10−19
p = 1.6 × 10−10
0.0
0.2
0.4
0.6
0.8
1.0
Effect Size (|B|)
D
pLI
pRec
pNull
DEL
DUP
mCNV
MEI
STR
Density (No. eGenes)
0.0
0.5
1.0
ExAC Score
E
DEL
DUP
mCNV
MEI
STR
all_eGenes
Non eGenes
1.1 × 10−7
0.451
2.8 × 10−5
9.1 × 10−4
9.1 × 10−25
1.3 × 10−30
1.7 × 10−4
0.472
0.830
0.097
3.1 × 10−7
4.5 × 10−8
0.002
1.000
0.442
0.394
0.002
5.2 × 10−4
0
10
20
30
0
5
10
% eGenes
pLI (>0.9)
% eGenes
pRec (>0.9)
% eGenes
pNull (>0.9)
F
p = 9.6 × 10−18
pLI (mean)
0.40
0.35
0.30
0.25
0.20
0.1
0.2
0.3
Effect Size (|B|)

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 [69,70]. 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 pLI scores (< 0.9) and pNull scores but more evenly distributed between low and high pRec scores (Figure 2.4D). 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 (Figure 2.4E). 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 [71]. 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 (Figure 2.4F) consistent with a previous report that there is less variation in the expression of highly constrained genes [71]. 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 [72-75], 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 [76] that define promoter loops between gene promoters (promoter anchors) and distal sequences (distal anchors, Figure 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-25, FET) likely due to the difficulty of identifying loop anchors near segmental duplications, which frequently overlap mCNVs. 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 50kb away from the gene body 3) did not overlap the exon of the tested gene and 4) was a maximum of 200kb from a distal anchor, which we defined as "distal variant-gene pairs" (Figure 2.5A). 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 (Figure 2.5B,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; p=0.022, <1e-40; FET, Figure 2.5D). These results indicate that many eQTLs include variants that overlap distal chromatin loop anchors and that variants that overlap distal anchors are more likely to be eQTLs and lead variants.

Figure 2.5 Localization of eQTLs near chromatin loops
(A) Cartoon showing localization of SVs and STRs at loop anchors. We selected all eVariants overlapping or close to distal anchors and associated with the expression of eGenes at the promoter anchor. Only eVariants closer to the distal anchor (right of grey dotted line) than to the promoter anchor were considered. (B) Proportion of eQTLs in the SV/STR eQTL analysis that were genic (overlapping an intron, exon, or promoter of the eGene; 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 to 1,598 eGenes) mapped to SV classes. (D) Percentage of eVariant-eGene pairs where the eVariant (left) or lead eVariant (right) overlaps or does not overlap the distal anchor. Significance is calculated with FET comparing these proportions. (E) Fraction of tested distal variant-gene pairs (A) that were lead eQTLs versus their distance to the distal anchor. Points represent the centers of equally sized bins and show the mean and 95% confidence interval. Regression lines were calculated using logistic regression testing association between distance to the loop anchor (with anchors padded by 10kb) and whether the variant-gene pair was a lead association. These were computed separately for variant-gene pairs inside the loop or outside the loop. (F) Heatmap showing the number of eVariants connected to gene promoters through chromatin loops (X axis) and the number of these connected genes that are associated eGenes (Y axis). (G,H) Barplot showing the percentage of tested variants that were eVariants (G) or lead eVariants (H) as a function of the number of genes the variant was linked to via overlapping a distal anchor. P-values were calculated by FET (Benjamini-Hochberg FDR < 5%) comparing the proportions of tested variants that were eVariants for variants linked to some number of genes by chromatin loops (grey) to the proportion of variants that were eVariants among variants that were not linked to any genes by chromatin loops. (I) Number of eGenes versus the 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). We fit a linear model comparing the number of eGenes/eVariant versus the number of genes tested using whether those genes were linked by loops or not as a covariate. The *p* value indicates the significance of the covariate of whether the genes were or were not linked by loops.


A
SV eQTL
Non-significant SVs
Excluded SVs
Promoter Anchor
Distal Anchor
Inside of Distal
Outside of Distal
B
other distal eQTLs
72.3%
15.3%
genic eQTLs
12.4%
loop acting eQTLs
C
STR
DUP
mCNV
other SV
rMEI
MEI
DEL
D
intersects anchor
False
True
* $q<10^{-5}$
** $q<10^{-10}$
*** $q<10^{-40}$
DEL
DUP
mCNV
rMEI
MEI
STR
BND
% Significant Assoc.
% Lead Assoc.
E
$p = 1.08 \times 10^{-11}$
% Lead Associations
2.5%
2.0%
1.5%
1.0%
0.5%
0.0%
-0.2MB
-0.15MB
-0.1MB
-50.0kb
-10.0kb
overlap
10.0kb
0.1MB
0.2MB
0.3MB
0.4MB
0.5MB
Distance to Distal Anchor
F
No. eGenes
2 1
1 2 1
1 4
2 8 11 6 1
14 20 23 8 3
111 73 66 8 5
295 590 224 134 36 7
1
2-5
6-10
11-20
21-30
31+
No. genes linked to variant via distal loop
G
% eVariants
***
0
1
1-3
3-6
6+
No. genes linked to variant via distal loop
H
% Lead eVariants
No. genes linked to variant via distal loop
* $q<10^{-5}$
** $q<10^{-10}$
*** $q<10^{-40}$
I
$p < 10^{-40}$
No. eGenes
2 5 10 15 25
No. tested genes (loop or non-loop)
versus No. genes tested nonloop
versus No. genes tested loop

We next hypothesized that if variants near loops are regulating 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 (Figure 2.5A). For this model, we subsetted the variant-gene pairs to those whose variants were at least 100kb away from the nearest TSS and a maximum of 200kb 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 ($p=0.003$) 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 (Figure 2.5E, OR=1.5, $p=2.1e-8$, 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.

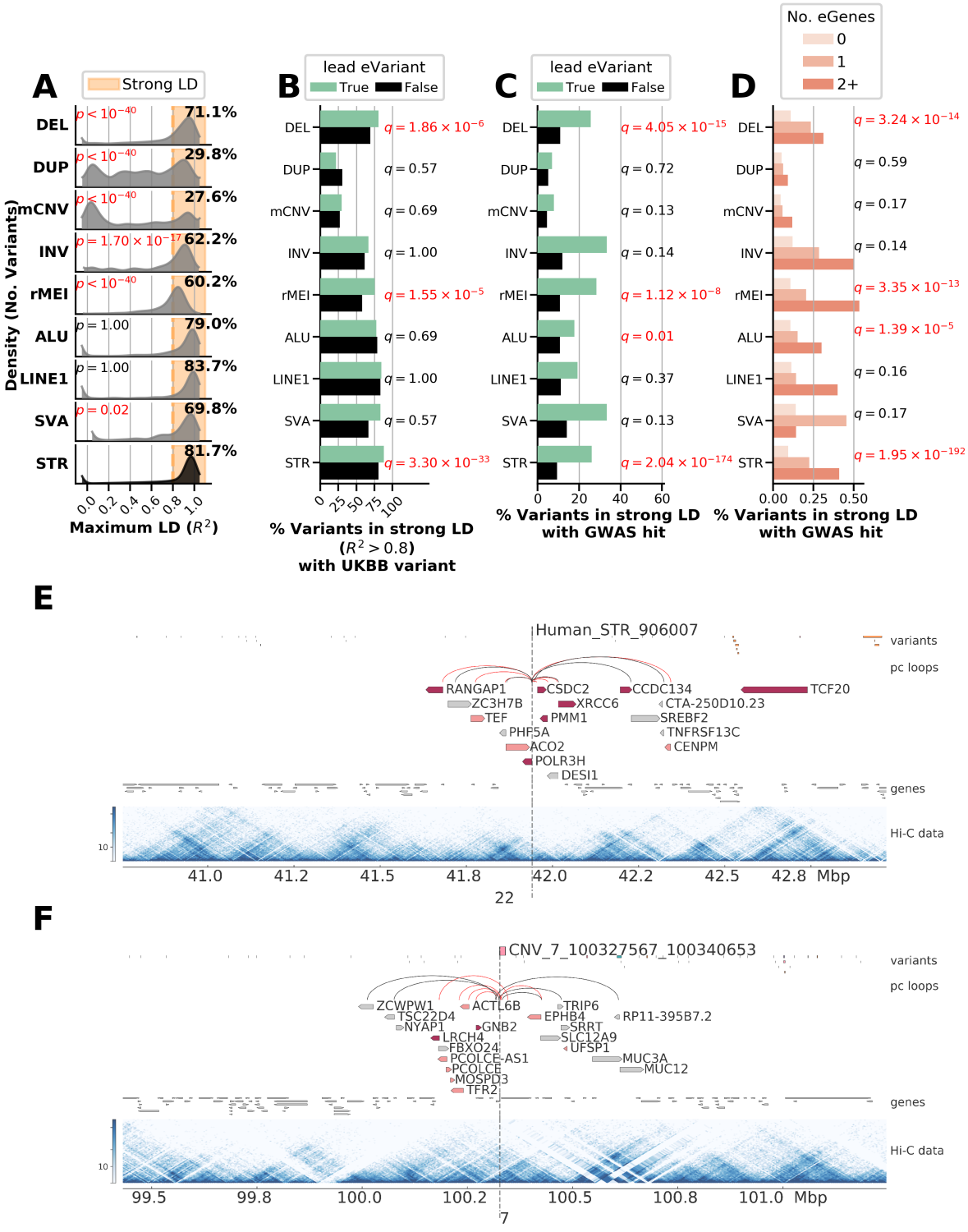
Given that variants overlapping distal anchors are more likely to be eQTLs, we hypothesized that variants that are looped to multiple gene promoters would affect the expression of many of their looped targets. To examine this, we tested whether the number of looped genes to an eVariant was associated 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 (Figure 2.5F). 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 (Figure 2.5G,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, $p<1e-40$, 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 (Figure 2.5I, p<1e-5, t-test, Methods). This trend was consistent across SV classes. 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, ± 50kb 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 (71.2%), but a markedly lower proportion of duplications (29.8%) and mCNVs (27.4%) were in strong LD with a nearby variant (Figure 2.6A). 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), 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, suggesting that the impact of these SV classes on traits and diseases needs to be further investigated.

Figure 2.6 Associations between SVs, STRs and GWAS
(A) Distribution of maximum LD score per i2QTL variant with UKBB variants nearby (within 50kb) for each variant type. Annotated p-values indicate results of Mann Whitney U test for the LD distribution of each SV class to be skewed lower than STRs after Bonferroni correction for multiple tests. (B) Fraction of variants of each class that are strongly tagged by a UKBB variant ($R^2 > 0.8$) for lead eVariants (green) versus all other variants in that class (black). Annotated q values indicate enrichment of lead eVariants to be in strong LD with a UKBB variant versus all other variants tested in the eQTL in the class (FET, Benjamini Hochberg FDR). (C) Fraction of variants of each class that are strongly tagged by a UKBB variant ($R^2 > 0.8$) that is significantly associated with at least one trait in the UKBB (p<5e-8). Annotated q values indicate enrichment of lead eVariants to be in strong LD with a UKBB variant that is significantly associated with at least one trait versus all other variants tested in the eQTL in the class (FET, Benjamini Hochberg FDR). (D) Percentage of variants in strong 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 significant GWAS variant. The p-values from these regressions were then corrected using Benjamini Hochberg resulting in annotated q-values, marked in red when $q < 0.05$. (E) Example of an STR on chromosome 22 that is a multi-gene eQTL associated with a total of nine unique eGenes including four genes that the STR loops to. Genes for which the variant is a lead variant are colored dark red and genes for which the variant is significantly associated are colored pink. iPSC Hi-C data is visualized as a heatmap of interaction frequencies. The variant is located between two chromatin subdomains that span ~100kb on the left side of the variant and ~25kb on the right side of the variant [73]. (F) Example of an mCNV on chromosome 7 that is a multi-gene eQTL associated with seven unique eGenes by looping.


A
Strong LD
DEL
DUP
mCNV
INV
rMEI
ALU
LINE1
SVA
STR
71.1%
29.8%
27.6%
62.2%
60.2%
79.0%
83.7%
69.8%
81.7%
Density (No. Variants)
Maximum LD (R^2)
B
lead eVariant
True
False
$q = 1.86 \times 10^{-6}$
$q = 0.57$
$q = 0.69$
$q = 1.00$
$q = 1.55 \times 10^{-5}$
$q = 0.69$
$q = 1.00$
$q = 0.57$
$q = 3.30 \times 10^{-33}$
% Variants in strong LD ($R^2 > 0.8$) with UKBB variant
C
lead eVariant
True
False
$q = 4.05 \times 10^{-15}$
$q = 0.72$
$q = 0.13$
$q = 0.14$
$q = 1.12 \times 10^{-8}$
$q = 0.01$
$q = 0.37$
$q = 0.13$
$q = 2.04 \times 10^{-174}$
% Variants in strong LD with GWAS hit
D
No. eGenes
0
1
2+
$q = 3.24 \times 10^{-14}$
$q = 0.59$
$q = 0.17$
$q = 0.14$
$q = 3.35 \times 10^{-13}$
$q = 1.39 \times 10^{-5}$
$q = 0.16$
$q = 0.17$
$q = 1.95 \times 10^{-192}$
% Variants in strong LD with GWAS hit
E
Human_STR_906007
variants
pc loops
RANGAP1
CSDC2
CCDC134
TCF20
ZC3H7B
XRCC6
CTA-250D10.23
TEF
PMM1
SREBF2
PHF5A
TNFRSF13C
ACO2
CENPM
POLR3H
DESI1
genes
Hi-C data
41.0
41.2
41.5
41.8
42.0
42.2
42.5
42.8 Mbp
22
F
CNV_7_100327567_100340653
variants
pc loops
ZCWPW1
ACTL6B
TRIP6
TSC22D4
EPHB4
RP11-395B7.2
NYAP1
GNB2
SRRT
LRCH4
SLC12A9
FBXO24
UFSP1
PCOLCE-AS1
MUC3A
PCOLCE
MUC12
MOSPD3
TFR2
genes
Hi-C data
99.5
99.8
100.0
100.2
100.5
100.8
101.0 Mbp
7

Next, we investigated the extent to which SVs and STRs linked to gene expression were tagged by nearby UKBB SNVs (R2 > 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 (Figure 2.6B). While >65% of lead eVariants for most SV classes were in strong LD with any nearby UKBB variant, only 26% of mCNV and 24% 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 ($p<5e-8$) and found that 11.4% of common STRs and SVs were by proxy associated with at least one of 701 UKBB traits (Figure 2.6C). 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 (Figure 2.6C). 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 were diverse including 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.

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 eVariants were highly enriched to be in strong LD (> 0.8 R2) with GWAS variants (Figure 2.6D). In fact, ~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 20bp eSTR was

associated with nine eGenes (seven lead) connected via distal loops (Figure 2.6E) 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 [16,77]. We observed that this variant appears to overlap a chromatin subdomain boundary visible in Hi-C data from iPSCORE [73] 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 [78]. Additionally, we found a 13kb deletion on chromosome 7 linked to five eGenes via looping that was also linked to 14 traits (Figure 2.6F). 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 across many if not all the SV classes and STRs. For deletions, duplications, and STRs, we found that increased length tended to be associated with a higher likelihood of being an eVariant and was also associated with increased effect size for non-coding 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 [73], 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 though more work is needed to establish the underlying details of this potential mechanism. 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.

One outstanding question is whether different SV classes and STRs differentially impact gene expression. 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 SNPs used in GWAS; and their likelihood of being associated with GWAS traits. Interestingly, mCNVs 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 which likely causes differences in gene expression [19]. 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 24% 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 [79] 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. The collection of eQTLs identified here, along with the catalog of high-quality SVs and STRs described in a companion paper [63], provide a powerful resource for future studies examining how these variants regulate gene expression and contribute to variation in complex traits.

Acknowledgments

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.C.D., H.M.; Writing – Original Draft, 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 that they have no conflicts of interest.

Methods

Abbreviations

1KGP: 1000 Genomes Project

eQTL: Expression quantitative trait locus. Defined by a (eGene - eVariant pair)

eGene: Gene implicated in an eQTL

eVariant: Variant implicated in an eQTL

eSV: Structural eVariant

eSTR: Short tandem repeat eVariant

eSNV: Single nucleotide eVariant

eIndel: Small insertion/deletion eVariant

FET: Fisher's exact test

GWAS: Genome-wide association studies

Indel: Small insertion/deletion variant

LD: Linkage disequilibrium

SV: Structural Variant

SNV: Single nucleotide variant

SNP: Single nucleotide-polymorphism

WGS: Whole-genome sequencing

Types of Genetic Variants Detected:

- DEL: Biallelic deletion ascertained by LUMPY, GS, GS LCNV
- DUP: Biallelic duplication ascertained by LUMPY, GS, GS LCNV

- mCNV: multiallelic copy number variant ascertained by LUMPY, GS, GS LCNV. This is defined as a variant that has at least 3 predicted alleles.
- INV: inversion ascertained by LUMPY
- rMEI: reference mobile element insertion
- BND: generic "breakend" ascertained by LUMPY. May include deletions and duplications that lack read-depth evidence, balanced rearrangements (INV), MEI or other uncategorized break points.
- ALU: Non-reference Alu element insertion identified by MELT
- LINE1: Non-reference LINE-1 element insertion identified by MELT
- SVA: Non-reference SVA (SINE-R/VNTR/Alu) element insertion identified by MELT
- STR: short tandem repeat variant, detected by HipSTR. Included variants have at least one individual with a change in length from the reference.
- CNV: copy number variant (deletion or duplication structural variant). Encompasses DEL, DUP, mCNV
- MEI: Non-reference mobile element insertion ascertained by MELT, including ALU, LINE1, and SVA elements

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, [63]).

RNA-Seq quality control and processing

As part of the i2QTL Consortium, we have collected a set of RNA sequencing (RNA-seq) samples from 2,954 human induced pluripotent stem cell (iPSC) lines derived from 1,600 unique donors from five studies: iPSCORE [40,41], HipSci[42,43], Banovich et al.[80], GENESiPS[81], and PhLiPS[82]. 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 dataset corresponding to 388 unique donors from iPSCORE and HipSci that have variant calls from deep whole genome sequencing [63]. 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; (3) Banovich et al. (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) [83] 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 [84] for any of the necessary genome annotations. Gene-level RNA expression was quantified from the STAR alignments using featureCounts (v1.6.0) [85], 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’ to 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 [86,87] 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. This resulted in 1,378 iPSC lines derived from 1,001 donors. 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 [63]. 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.

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. 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 [88]). All models were fit using LIMIX [89] (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 1,378 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 [90]. 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 [91]. 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 [37]. Genome STRiP CNVDiscovery and LCNVDiscovery [29] variants were encoded with integer diploid copy numbers (CN). SpeedSeq [24,46] 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 [47], we used traditional genotypes (0/0, 0/1, 1/1) outputted by the software, as these SVs are expected to be largely bi-allelic 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 [61] 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 7,013,178 SNVs, 1,862,365 indels that were present at a minor allele frequency of at least 5% and 13,804 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. 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.[37]. 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) [63]. In a companion paper [63], 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 [63]. To compute the number of unique variants in downstream analyses, variants were annotated with the redundancy clusters they belonged to [63] 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; Figure 2.1A).

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). 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%, Benjamini-Hochberg). To study the localization of eSVs with respect to eGenes, we used Gencode v19 [92] 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, Benjamin-Hochberg correction). 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, Benjamini-Hochberg correction). To investigate the constraint of eGenes, we obtained ExAC (v0.3.1) [71] 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 9,052 non-eGenes that were tested in our dataset 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 variant class 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 iPSC promoter capture Hi-C loop call data from a previous study [93]. 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 50kb from the promoter 3) at most 200kb 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 10kb 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 used this same subset of variant-gene pairs to test whether the distance of a variant from the distal loop anchor was associated with its likelihood of being associated with gene expression. To do so, we fit a logistic model 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 100kb from the outside of the distal loop anchor to variants that were within 100kb 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 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, August 1, 2018). For each of the 42,921 non-redundant SVs and STRs, we used bcftools[53] 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 calculate using the Wald test and then corrected for multiple testing using Benjamini-Hochberg FDR.

Chapter 2, in full, is a reprint of the material as it appears in bioRxiv, 2019, David Jakubosky, Matteo D'Antonio, Marc Jan Bonder, Craig Smail, Margaret K.R. Donovan, William W. Young Greenwald, Agnieszka D'Antonio-Chronowska, Hiroko Matsui, Oliver Stegle, Erin N. Smith, Stephen B. Montgomery, Christopher DeBoever, Kelly A. Frazer. The dissertation author was one of the primary investigators and authors of this paper.

REFERENCES

1 Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**, 224-238, doi:10.1038/nrg.2015.25 (2016).

2 Brandler, W. M., Antaki, D., Gujral, M., Kleiber, M. L., Whitney, J., Maile, M. S., Hong, O., Chapman, T. R., Tan, S., Tandon, P., Pang, T., Tang, S. C., Vaux, K. K., Yang, Y., Harrington, E., Juul, S., Turner, D. J., Thiruvahindrapuram, B., Kaur, G., Wang, Z., Kingsmore, S. F., Gleeson, J. G., Bisson, D., Kakaradov, B., Telenti, A., Venter, J. C., Corominas, R., Toma, C., Cormand, B., Rueda, I., Guijarro, S., Messer, K. S., Nievergelt, C. M., Arranz, M. J., Courchesne, E., Pierce, K., Muotri, A. R., Iakoucheva, L. M., Hervas, A., Scherer, S. W., Corsello, C. & Sebat, J. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327-331, doi:10.1126/science.aan2261 (2018).

3 Malhotra, D., McCarthy, S., Michaelson, Jacob J., Vacic, V., Burdick, Katherine E., Yoon, S., Cichon, S., Corvin, A., Gary, S., Gershon, Elliot S., Gill, M., Karayiorgou, M., Kelsoe, John R., Krastoshevsky, O., Krause, V., Leibenluft, E., Levy, Deborah L., Makarov, V., Bhandari, A., Malhotra, Anil K., McMahon, Francis J., Nöthen, Markus M., Potash, James B., Rietschel, M., Schulze, Thomas G. & Sebat, J. High Frequencies of De Novo CNVs in Bipolar Disorder and Schizophrenia. *Neuron* **72**, 951-963, doi:10.1016/j.neuron.2011.11.007 (2011).

4 Malhotra, D. & Sebat, J. CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. *Cell* **148**, 1223-1241, doi:10.1016/j.cell.2012.02.039 (2012).

5 Michaelson, Jacob J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A., Wu, W., Corominas, R., Peoples, Á., Koren, A., Gore, A., Kang, S., Lin, Guan N., Estabillo, J., Gadomski, T., Singh, B., Zhang, K., Akshoomoff, N., Corsello, C., McCarroll, S., Iakoucheva, Lilia M., Li, Y., Wang, J. & Sebat, J. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell* **151**, 1431-1442, doi:10.1016/j.cell.2012.11.019 (2012).

6 Beck, M., Peterson, J. F., McConnell, J., McGuire, M., Asato, M., Losee, J. E., Surti, U., Madan-Khetarpal, S., Rajkovic, A. & Yatsenko, S. A. 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, doi:10.1002/ajmg.a.36839 (2015).

7 Spielmann, M. & Klopocki, E. CNVs of noncoding cis-regulatory elements in human disease. *Current Opinion in Genetics & Development* **23**, 249-256, doi:10.1016/j.gde.2013.02.013 (2013).

8 Pearson, C. E. Slipping while sleeping? Trinucleotide repeat expansions in germ cells. *Trends Mol Med* **9**, 490-495 (2003).

9 Mirkin, S. M. Expandable DNA repeats and human disease. *Nature* **447**, 932-940, doi:10.1038/nature05977 (2007).

10 La Spada, A. R. & Taylor, J. P. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat Rev Genet* **11**, 247-258, doi:10.1038/nrg2748 (2010).

11 McMurray, C. T. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet* **11**, 786-799, doi:10.1038/nrg2828 (2010).

12 Nelson, D. L., Orr, H. T. & Warren, S. T. The unstable repeats--three evolving faces of neurological disease. *Neuron* **77**, 825-843, doi:10.1016/j.neuron.2013.02.022 (2013).

13 Spielmann, M. & Mundlos, S. Structural variations, the regulatory landscape of the genome and their alteration in human disease. *BioEssays : news and reviews in molecular, cellular and developmental biology* **35**, 533-543, doi:10.1002/bies.201200178 (2013).

14 Den Dunnen, W. F. A. Trinucleotide repeat disorders. *Handb Clin Neurol* **145**, 383-391, doi:10.1016/B978-0-12-802395-2.00027-4 (2017).

15 Gamazon, E. R., Nicolae, D. L. & Cox, N. J. A Study of CNVs As Trait-Associated Polymorphisms and As Expression Quantitative Trait Loci. *Plos Genet* **7**, doi:ARTN e1001292 10.1371/journal.pgen.1001292 (2011).

16 Kong, S. W., Collins, C. D., Shimizu-Motohashi, Y., Holm, I. A., Campbell, M. G., Lee, I. H., Brewster, S. J., Hanson, E., Harris, H. K., Lowe, K. R., Saada, A., Mora, A., Madison, K., Hundley, R., Egan, J., McCarthy, J., Eran, A., Galdzicki, M., Rappaport, L., Kunkel, L. M. & Kohane, I. S. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS One* **7**, e49475, doi:10.1371/journal.pone.0049475 (2012).

17 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 Research* **21**, 2004-2013, doi:10.1101/gr.122614.111 (2011).

18 Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. a., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. a., Untergasser, A., Walker, J. a., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. a., McCarroll, S. a., Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E. & Korbel, J. O. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81, doi:10.1038/nature15394 (2015).

19 Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., Montgomery, S. B., Battle, A., Conrad, D. F. & Hall, I. M. The impact of structural variation on human gene expression. *Nature Genetics* **49**, 692-699, doi:10.1038/ng.3834 (2017).

20 Hehir-Kwa, J. Y., Marschall, T., Kloosterman, W. P., Francioli, L. C., Baaijens, J. A., Dijkstra, L. J., Abdellaoui, A., Koval, V., Thung, D. T., Wardenaar, R., Renkens, I., Coe, B. P., Deelen, P., de Ligt, J., Lameijer, E. W., van Dijk, F., Hormozdiari, F., Genome of the Netherlands, C., Uitterlinden, A. G., van Duijn, C. M., Eichler, E. E., de Bakker, P. I., Swertz, M. A., Wijmenga, C., van Ommen, G. B., Slagboom, P. E., Boomsma, D. I., Schonhuth, A., Ye, K. & Guryev, V. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**, 12989, doi:10.1038/ncomms12989 (2016).

21 Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M. & Kamatani, Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**, 117, doi:10.1186/s13059-019-1720-5 (2019).

22 Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V. & Korbel, J. O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339, doi:10.1093/bioinformatics/bts378 (2012).

23 Fan, X., Abbott, T. E., Larson, D. & Chen, K. BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. *Curr Protoc Bioinformatics* **45**, 15 16 11-11, doi:10.1002/0471250953.bi1506s45 (2014).

24 Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: A probabilistic framework for structural variant discovery. *Genome biology* **15**, R84, doi:10.1186/gb-2014-15-6-r84 (2014).

25 Kronenberg, Z. N., Osborne, E. J., Cone, K. R., Kennedy, B. J., Domyan, E. T., Shapiro, M. D., Elde, N. C. & Yandell, M. Wham: Identifying Structural Variants of Biological Consequence. *PLoS Comput Biol* **11**, e1004572, doi:10.1371/journal.pcbi.1004572 (2015).

26 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 Research* **21**, 974-984, doi:10.1101/gr.114876.110 (2011).

27 Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer, U. & Hochreiter, S. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res* **40**, doi:ARTN e69 10.1093/nar/gks003 (2012).

28 Zhu, M., Need, A. C., Han, Y., Ge, D., Maia, J. M., Zhu, Q., Heinzen, E. L., Cirulli, E. T., Pelak, K., He, M., Ruzzo, E. K., Gumbs, C., Singh, A., Feng, S., Shianna, K. V. & Goldstein, D. B. Using ERDS to infer copy-number variants in high-coverage genomes. *American Journal of Human Genetics* **91**, 408-421, doi:10.1016/j.ajhg.2012.07.004 (2012).

29 Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M. & Mccarroll, S. A. Large multiallelic copy number variations in humans. *Nature Genetics* **47**, 296-303, doi:10.1038/ng.3200 (2015).

30 Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G. & de Ridder, D. Making the difference: integrating structural variation detection tools. *Brief Bioinform* **16**, 852-864, doi:10.1093/bib/bbu047 (2015).

31 Becker, T., Lee, W. P., Leone, J., Zhu, Q., Zhang, C., Liu, S., Sargent, J., Shanker, K., Mil-Homens, A., Cerveira, E., Ryan, M., Cha, J., Navarro, F. C. P., Galeev, T., Gerstein, M., Mills, R. E., Shin, D. G., Lee, C. & Malhotra, A. FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol* **19**, 38, doi:10.1186/s13059-018-1404-6 (2018).

32 Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., Hastie, A. R., Antaki, D., Anantharaman, T., Audano, P. A., Brand, H., Cantsilieris, S., Cao, H., Cerveira, E., Chen, C., Chen, X., Chin, C. S., Chong, Z., Chuang, N. T., Lambert, C. C., Church, D. M., Clarke, L., Farrell, A., Flores, J., Galeev, T., Gorkin, D. U., Gujral, M., Guryev, V., Heaton, W. H., Korlach, J., Kumar, S., Kwon, J. Y., Lam, E. T., Lee, J. E., Lee, J., Lee, W. P., Lee, S. P., Li, S., Marks, P., Viaud-Martinez, K., Meiers, S., Munson, K. M., Navarro, F. C. P., Nelson, B. J., Nodzak, C., Noor, A., Kyriazopoulou-Panagiotopoulou, S., Pang, A. W. C., Qiu, Y., Rosanio, G., Ryan, M., Stutz, A., Spierings, D. C. J., Ward, A., Welch, A. E., Xiao, M., Xu, W., Zhang, C., Zhu, Q., Zheng-Bradley, X., Lowy, E., Yakneen, S., McCarroll, S., Jun, G., Ding, L., Koh, C. L., Ren, B., Flicek, P., Chen, K., Gerstein, M. B., Kwok, P. Y., Lansdorp, P. M., Marth, G. T., Sebat, J., Shi, X., Bashir, A., Ye, K., Devine, S. E., Talkowski, M. E., Mills, R. E., Marschall, T., Korbel, J. O., Eichler, E. E. & Lee, C. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**, 1784, doi:10.1038/s41467-018-08148-z (2019).

33 Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J. & Eichler, E. E. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-U163, doi:10.1038/nature13907 (2015).

34 Collins, R. L., Brand, H., Redin, C. E., Hanscom, C., Antolik, C., Stone, M. R., Glessner, J. T., Mason, T., Pregno, G., Dorrani, N., Mandrile, G., Giachino, D., Perrin, D., Walsh, C., Cipicchio, M., Costello, M., Stortchevoi, A., An, J. Y., Currall, B. B., Seabra, C. M., Ragavendran, A., Margolin, L., Martinez-Agosto, J. A., Lucente, D., Levy, B., Sanders, S. J., Wapner, R. J., Quintero-Rivera, F., Kloosterman, W. & Talkowski, M. E. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol* **18**, 36, doi:10.1186/s13059-017-1158-6 (2017).

35 Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., Warren, W. C., Magrini, V., McGrath,

S. D., Li, Y. I., Wilson, R. K. & Eichler, E. E. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663-675 e619, doi:10.1016/j.cell.2018.12.019 (2019).

36 Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A. K. Y., McCaffrey, J., Young, E., Lam, E. T., Hastie, A. R., Wong, K. H. Y., Chung, C. Y. L., Ma, W., Sibert, J., Rajagopalan, R., Jin, N., Chow, E. Y. C., Chu, C., Poon, A., Lin, C., Naguib, A., Wang, W. P., Cao, H., Chan, T. F., Yip, K. Y., Xiao, M. & Kwok, P. Y. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat Commun* **10**, 1025, doi:10.1038/s41467-019-08992-7 (2019).

37 Gymrek, M. M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M. J., Price, A. L., Pritchard, J., Sharp, A. & Erlich, Y. Abundant contribution of short tandem repeats to gene expression variation in humans. **27**, 617-630, doi:10.1016/j.ccell.2015.04.006.SRSF2 (2016).

38 Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M. & Erlich, Y. Genome-wide profiling of heritable and de novo STR variations. *Nature Methods* **14**, 590-592, doi:10.1038/nmeth.4267 (2017).

39 D'Antonio, M., Benaglio, P., Jakubosky, D., Greenwald, W. W., Matsui, H., Donovan, M. K. R., Li, H., Smith, E. N., D'Antonio-Chronowska, A. & Frazer, K. A. Insights into the Mutational Burden of Human Induced Pluripotent Stem Cells from an Integrative Multi-Omics Approach. *Cell Reports* **24**, 883-894, doi:10.1016/j.celrep.2018.06.091 (2018).

40 DeBoever, C., Li, H. H., Jakubosky, D., Antonio-chronowska, A. D., Farley, E. K. E. K., Frazer, K. A. K. A., DeBoever, C., Li, H. H., Jakubosky, D., Benaglio, P., Reyna, J., Olson, K. M. K. M., Huang, H., Biggs, W., Sandoval, E., D'Antonio, M., Jepsen, K., Matsui, H., Arias, A., Ren, B., Nariai, N., Smith, E. N., D'Antonio-Chronowska, A., Farley, E. K. E. K. & Frazer, K. A. K. A. Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell* **20**, 533-546, doi:10.1016/j.stem.2017.03.009 (2017).

41 Panopoulos, A. D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S. I., Schuldt, B. M., DeBoever, C., Arias, A. D., Garcia, M., Nelson, B. C., Harismendy, O., Jakubosky, D. A., Donovan, M. K. R., Greenwald, W. W., Farnam, K. J., Cook, M., Borja, V., Miller, C. A., Grinstein, J. D., Drees, F., Okubo, J., Diffenderfer, K. E., Hishida, Y., Modesto, V., Dargitz, C. T., Feiring, R., Zhao, C., Aguirre, A., McGarry, T. J., Matsui, H., Li, H., Reyna, J., Rao, F., O'Connor, D. T., Yeo, G. W., Evans, S. M., Chi, N. C., Jepsen, K., Nariai, N., Müller, F. J., Goldstein, L. S. B., Izpisua Belmonte, J. C., Adler, E., Loring, J. F., Berggren, W. T., D'Antonio-Chronowska, A., Smith, E. N. & Frazer, K. A. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Reports* **8**, 1086-1100, doi:10.1016/j.stemcr.2017.03.012 (2017).

42 Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F. P., Culley, O. J., Danecek, P., Faulconbridge, A., Harrison, P. W., Kathuria, A.,

McCarthy, D., McCarthy, S. A., Meleckyte, R., Memari, Y., Moens, N., Soares, F., Mann, A., Streeter, I., Agu, C. A., Alderton, A., Nelson, R., Harper, S., Patel, M., White, A., Patel, S. R., Clarke, L., Halai, R., Kirton, C. M., Kolb-Kokocinski, A., Beales, P., Birney, E., Danovi, D., Lamond, A. I., Ouwehand, W. H., Vallier, L., Watt, F. M., Durbin, R., Stegle, O. & Gaffney, D. J. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370-375, doi:10.1038/nature22403 (2017).

43 Streeter, I., Harrison, P. W., Faulconbridge, A., The HipSci, C., Flicek, P., Parkinson, H. & Clarke, L. The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res* **45**, D691-D697, doi:10.1093/nar/gkw928 (2017).

44 Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F. P., Culley, O. J., Danecek, P., Faulconbridge, A., Harrison, P. W., Kathuria, A., McCarthy, D., McCarthy, S. A., Meleckyte, R., Memari, Y., Moens, N., Soares, F., Mann, A., Streeter, I., Agu, C. A., Alderton, A., Nelson, R., Harper, S., Patel, M., White, A., Patel, S. R., Clarke, L., Halai, R., Kirton, C. M., Kolb-Kokocinski, A., Beales, P., Birney, E., Danovi, D., Lamond, A. I., Ouwehand, W. H., Vallier, L., Watt, F. M., Durbin, R., Stegle, O. & Gaffney, D. J. Corrigendum: Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 686, doi:10.1038/nature23012 (2017).

45 Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. a., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. a., Nickerson, D. a., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Gibbs, R. a., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E. S., Altshuler, D. M., Gabriel, S. B., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D. R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. a., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E. R., Wilson, R. K., Fulton, L., Fulton, R., Sherry, S. T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G. a., Durbin, R. M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J. P., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C. L., Kong, Y., Marcketta, A., Gibbs, R. a., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L. J. M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly,

M. J., DePristo, M. a., Handsaker, R. E., Altshuler, D. M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S. B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E. S., McCarroll, S. a., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Clark, A. G., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Korbel, J. O., Rausch, T., Fritz, M. H., Stütz, A. M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W. M., Ritchie, G. R. S., Smith, R. E., Zerbino, D., Zheng-Bradley, X., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Bentley, D. R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. a., Konkel, M. K., Walker, J. a., MacArthur, D. G., Lek, M., Sudbrak, R., Amstislavskiy, V. S., Herwig, R., Mardis, E. R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., Swaroop, A., Chew, E., Lappalainen, T., Erlich, Y., Gymrek, M., Frederick Willems, T., Simpson, J. T., Shriver, M. D., Rosenfeld, J. a., Bustamante, C. D., Montgomery, S. B., De La Vega, F. M., Byrnes, J. K., Carroll, A. W., DeGorter, M. K., Lacroute, P., Maples, B. K., Martin, A. R., Moreno-Estrada, A., Shringarpure, S. S., Zakharia, F., Halperin, E., Baran, Y., Lee, C., Cerveira, E., Hwang, J., Malhotra, A., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Hyland, F. C. L., Craig, D. W., Christoforides, A., Homer, N., Izatt, T., Kurdoglu, A. a., Sinari, S. a., Squire, K., Sherry, S. T., Xiao, C., Sebat, J., Antaki, D., Gujral, M., Noor, A., Ye, K., Burchard, E. G., Hernandez, R. D., Gignoux, C. R., Haussler, D., Katzman, S. J., James Kent, W., Howie, B., Ruiz-Linares, A., Dermitzakis, E. T., Devine, S. E., Abecasis, G. R., Min Kang, H., Kidd, J. M., Blackwell, T., Caron, S., Chen, W., Emery, S., Fritsche, L., Fuchsberger, C., Jun, G., Li, B., Lyons, R., Scheller, C., Sidore, C., Song, S., Sliwerska, E., Taliun, D., Tan, A., Welch, R., Kate Wing, M., Zhan, X., Awadalla, P., Hodgkinson, A., Li, Y., Shi, X., Quitadamo, A., Lunter, G., McVean, G. a., Marchini, J. L., Myers, S., Churchhouse, C., Delaneau, O., Gupta-Hinch, A., Kretzschmar, W., Iqbal, Z., Mathieson, I., Menelaou, A., Rimmer, A., Xifara, D. K., Oleksyk, T. K., Fu, Y., Liu, X., Xiong, M., Jorde, L., Witherspoon, D., Xing, J., Eichler, E. E., Browning, B. L., Browning, S. R., Hormozdiari, F., Sudmant, P. H., Khurana, E., Durbin, R. M., Hurles, M. E., Tyler-Smith, C., Albers, C. a., Ayub, Q., Balasubramaniam, S., Chen, Y., Colonna, V., Danecek, P., Jostins, L., Keane, T. M., McCarthy, S., Walter, K., Xue, Y., Gerstein, M. B., Abyzov, A., Balasubramanian, S., Chen, J., Clarke, D., Fu, Y., Harmanci, A. O., Jin, M., Lee, D., Liu, J., Jasmine Mu, X., Zhang, J., Zhang, Y., Li, Y., Luo, R., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G. T., Garrison, E. P., Kural, D., Lee, W.-P., Ward, A. N., Wu, J., Zhang, M., McCarroll, S. a., Handsaker, R. E., Altshuler, D. M., Banks, E., del Angel, G., Genovese, G., Hartl, C., Li, H., Kashin, S., Nemesh, J. C., Shakir, K., Yoon, S. C., Lihm, J., Makarov, V., Degenhardt, J., Korbel, J. O., Fritz, M. H., Meiers, S., Raeder, B., Rausch, T., Stütz, A. M., Flicek, P., Paolo Casale, F., Clarke, L., Smith, R. E., Stegle, O., Zheng-Bradley, X., Bentley, D. R., Barnes, B., Keira Cheetham, R., Eberle, M., Humphray, S., Kahn, S., Murray, L., Shaw, R., Lameijer, E.-W., Batzer, M. a., Konkel, M. K., Walker, J. a., Ding, L., Hall, I., Ye, K., Lacroute, P., Lee, C., Cerveira, E., Malhotra, A., Hwang, J., Plewczynski, D., Radew, K., Romanovitch, M., Zhang, C., Craig, D. W., Homer, N., Church, D., Xiao, C., Sebat, J., Antaki, D., Bafna, V., Michaelson, J., Ye, K., Devine, S. E., Gardner, E. J., Abecasis, G. R., Kidd, J. M., Mills, R. E., Dayama, G., Emery, S., Jun, G., Shi, X., Quitadamo, A., Lunter, G., McVean, G. a., Chen, K., Fan, X., Chong, Z., Chen, T., Witherspoon, D., Xing, J., Eichler, E. E., Chaisson, M. J., Hormozdiari, F., Huddleston, J., Malig, M., Nelson, B. J., Sudmant, P. H., Parrish, N. F., Khurana, E., Hurles, M. E., Blackburne, B., Lindsay, S. J., Ning, Z., Walter, K., Zhang, Y., Gerstein, M. B., Abyzov, A., Chen, J., Clarke, D., Lam, H., Jasmine Mu, X., Sisu, C., Zhang, J., Zhang, Y., Gibbs, R. a., Yu,

F., Bainbridge, M., Challis, D., Evani, U. S., Kovar, C., Lu, J., Muzny, D., Nagaswamy, U., Reid, J. G., Sabo, A., Yu, J., Guo, X., Li, W., Li, Y., Wu, R., Marth, G. T., Garrison, E. P., Fung Leong, W., Ward, A. N., del Angel, G., DePristo, M. a., Gabriel, S. B., Gupta, N., Hartl, C., Poplin, R. E., Clark, A. G., Rodriguez-Flores, J. L., Flicek, P., Clarke, L., Smith, R. E., Zheng-Bradley, X., MacArthur, D. G., Mardis, E. R., Fulton, R., Koboldt, D. C., Gravel, S., Bustamante, C. D., Craig, D. W., Christoforides, A., Homer, N., Izatt, T., Sherry, S. T., Xiao, C., Dermitzakis, E. T., Abecasis, G. R., Min Kang, H., McVean, G. a., Gerstein, M. B., Balasubramanian, S., Habegger, L., Yu, H., Flicek, P., Clarke, L., Cunningham, F., Dunham, I., Zerbino, D., Zheng-Bradley, X., Lage, K., Berg Jespersen, J., Horn, H., Montgomery, S. B., DeGorter, M. K., Khurana, E., Tyler-Smith, C., Chen, Y., Colonna, V., Xue, Y., Gerstein, M. B., Balasubramanian, S., Fu, Y., Kim, D., Auton, A., Marcketta, A., Desalle, R., Narechania, A., Wilson Sayres, M. a., Garrison, E. P., Handsaker, R. E., Kashin, S., McCarroll, S. a., Rodriguez-Flores, J. L., Flicek, P., Clarke, L., Zheng-Bradley, X., Erlich, Y., Gymrek, M., Frederick Willems, T., Bustamante, C. D., Mendez, F. L., David Poznik, G., Underhill, P. a., Lee, C., Cerveira, E., Malhotra, A., Romanovitch, M., Zhang, C., Abecasis, G. R., Coin, L., Shao, H., Mittelman, D., Tyler-Smith, C., Ayub, Q., Banerjee, R., Cerezo, M., Chen, Y., Fitzgerald, T. W., Louzada, S., Massaia, A., McCarthy, S., Ritchie, G. R., Xue, Y., Yang, F., Gibbs, R. a., Kovar, C., Kalra, D., Hale, W., Muzny, D., Reid, J. G., Wang, J., Dan, X., Guo, X., Li, G., Li, Y., Ye, C., Zheng, X., Altshuler, D. M., Flicek, P., Clarke, L., Zheng-Bradley, X., Bentley, D. R., Cox, A., Humphray, S., Kahn, S., Sudbrak, R., Albrecht, M. W., Lienhard, M., Larson, D., Craig, D. W., Izatt, T., Kurdoglu, A. a., Sherry, S. T., Xiao, C., Haussler, D., Abecasis, G. R., McVean, G. a., Durbin, R. M., Balasubramaniam, S., Keane, T. M., McCarthy, S., Stalker, J., Chakravarti, A., Knoppers, B. M., Abecasis, G. R., Barnes, K. C., Beiswanger, C., Burchard, E. G., Bustamante, C. D., Cai, H., Cao, H., Durbin, R. M., Gerry, N. P., Gharani, N., Gibbs, R. a., Gignoux, C. R., Gravel, S., Henn, B., Jones, D., Jorde, L., Kaye, J. S., Keinan, A., Kent, A., Kerasidou, A., Li, Y., Mathias, R., McVean, G. a., Moreno-Estrada, A., Ossorio, P. N., Parker, M., Resch, A. M., Rotimi, C. N., Royal, C. D., Sandoval, K., Su, Y., Sudbrak, R., Tian, Z., Tishkoff, S., Toji, L. H., Tyler-Smith, C., Via, M., Wang, Y., Yang, H., Yang, L., Zhu, J., Bodmer, W., Bedoya, G., Ruiz-Linares, A., Cai, Z., Gao, Y., Chu, J., Peltonen, L., Garcia-Montero, A., Orfao, A., Dutil, J., Martinez-Cruzado, J. C., Oleksyk, T. K., Barnes, K. C., Mathias, R. a., Hennis, A., Watson, H., McKenzie, C., Qadri, F., LaRocque, R., Sabeti, P. C., Zhu, J., Deng, X., Sabeti, P. C., Asogun, D., Folarin, O., Happi, C., Omoniwa, O., Stremlau, M., Tariyal, R., Jallow, M., Sisay Joof, F., Corrah, T., Rockett, K., Kwiatkowski, D., Kooner, J., Tịnh Hiê`n, T. n., Dunstan, S. J., Thuy Hang, N., Fonnie, R., Garry, R., Kanneh, L., Moses, L., Sabeti, P. C., Schieffelin, J., Grant, D. S., Gallo, C., Poletti, G., Saleheen, D., Rasheed, A., Brooks, L. D., Felsenfeld, A. L., McEwen, J. E., Vaydylevich, Y., Green, E. D., Duncanson, A., Dunn, M., Schloss, J. a., Wang, J., Yang, H., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Min Kang, H., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. a. & Abecasis, G. R. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

46 Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R. & Hall, I. M. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods* **12**, 966-968, doi:10.1038/nmeth.3505 (2015).

47 Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., Mills, R. E., 1000 Genomes Project Consortium, G. P. & Devine, S. E. The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome research*, gr.218032.218116, doi:10.1101/gr.218032.116 (2017).

48 Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. & Cavalli-Sforza, L. L. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* **102**, 15942-15947, doi:10.1073/pnas.0507611102 (2005).

49 Jakubosky, D., D'Antonio, M., Bonder, M. J., Smail, C., Donovan, M. K. R., Young Greenwald, W. W., D'Antonio-Chronowska, A., Matsui, H., Stegle, O., Smith, E. N., Montgomery, S. B., DeBoever, C. & Frazer, K. A. Structural variant classes and short tandem repeats differentially impact gene expression and complex traits. *bioRxiv*, 714477, doi:10.1101/714477 (2019).

50 Sankar, P. L. & Parker, L. S. The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. *Genet Med* **19**, 743-750, doi:10.1038/gim.2016.183 (2017).

51 Brown, J., Pirrung, M. & McCue, L. A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, doi:10.1093/bioinformatics/btx373 (2017).

52 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

53 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

54 Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034, doi:10.1093/bioinformatics/btv098 (2015).

55 Panopoulos, A. D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S. I., Schuldt, B. M., DeBoever, C., Arias, A. D., Garcia, M., Nelson, B. C., Harismendy, O., Jakubosky, D. A., Donovan, M. K. R., Greenwald, W. W., Farnam, K., Cook, M., Borja, V., Miller, C. A., Grinstein, J. D., Drees, F., Okubo, J., Diffenderfer, K. E., Hishida, Y., Modesto, V., Dargitz, C. T., Feiring, R., Zhao, C., Aguirre, A., McGarry, T. J., Matsui, H., Li, H., Reyna, J., Rao, F., O'Connor, D. T., Yeo, G. W., Evans, S. M., Chi, N. C., Jepsen, K., Nariai, N., Muller, F. J., Goldstein, L. S. B., Izpisua Belmonte, J. C., Adler, E., Loring, J. F., Berggren, W. T., D'Antonio-Chronowska, A., Smith, E. N. & Frazer, K. A. iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Reports* **8**, 1086-1100, doi:10.1016/j.stemcr.2017.03.012 (2017).

56 Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84, doi:10.1186/gb-2014-15-6-r84 (2014).

57 Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics (Oxford, England)*, 1-9, doi:10.1093/bioinformatics/btu356 (2014).

58 Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, a. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. a., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M. & Sandelin, A. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461, doi:10.1038/nature12787 (2014).

59 Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-34, doi:10.1002/0471250953.bi1112s47 (2014).

60 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

61 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

62 Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H. C., Agarwala, R., McLaren, W. M., Ritchie, G. R., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E. E., Weinstock, G., Mardis, E. R., Wilson, R. K., Howe, K., Flicek, P. & Hubbard, T. Modernizing reference genome assemblies. *PLoS Biol* **9**, e1001091, doi:10.1371/journal.pbio.1001091 (2011).

63 Jakubosky, D., Smith, E. N., D'Antonio, M., Bonder, M. J., Young Greenwald, W. W., Matsui, H., D'Antonio-Chronowska, A., Stegle, O., Montgomery, S. B., DeBoever, C. & Frazer, K. A. Discovery and Quality Analysis of a Comprehensive Set of Structural Variants and Short Tandem Repeats using Deep Whole Genome Sequencing Data. *bioRxiv*, 713198, doi:10.1101/713198 (2019).

64 Li, X., Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C., Hess, G. T., Zappala, Z., Strober, B. J., Scott, A. J., Li, A., Ganna, A., Bassik, M. C., Merker, J. D., Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Karczewski, K. J., Lek, M., Li, X., MacArthur, D. G., Nedzel, J. L., Nguyen, D. T.,

Noble, M. S., Segrè, A. V., Trowbridge, C. A., Tukiainen, T., Abell, N. S., Balliu, B., Barshir, R., Basha, O., Bogu, G. K., Brown, A., Brown, C. D., Castel, S. E., Chen, L. S., Conrad, D. F., Cox, N. J., Delaneau, O., Dermitzakis, E. T., Engelhardt, B. E., Eskin, E., Ferreira, P. G., Frésard, L., Gamazon, E. R., Garrido-Martín, D., Gewirtz, A. D. H., Gliner, G., Gloudemans, M. J., Guigo, R., Hall, I. M., Han, B., He, Y., Hormozdiari, F., Howald, C., Im, H. K., Jo, B., Kang, E. Y., Kim-Hellmuth, S., Lappalainen, T., Li, G., Liu, B., Mangul, S., McCarthy, M. I., McDowell, I. C., Mohammadi, P., Monlong, J., Muñoz-Aguirre, M., Ndungu, A. W., Nicolae, D. L., Nobel, A. B., Oliva, M., Ongen, H., Palowitch, J. J., Panousis, N., Papasaikas, P., Park, Y., Parsana, P., Payne, A. J., Peterson, C. B., Quan, J., Reverter, F., Sabatti, C., Saha, A., Sammeth, M., Shabalin, A. A., Sodaei, R., Stephens, M., Stranger, B. E., Sul, J. H., Urbut, S., Van De Bunt, M., Wang, G., Wen, X., Wright, F. A., Xi, H. S., Yeger-Lotem, E., Zaugg, J. B., Zhou, Y. H., Akey, J. M., Bates, D., Chan, J., Claussnitzer, M., Demanelis, K., Diegel, M., Doherty, J. A., Feinberg, A. P., Fernando, M. S., Halow, J., Hansen, K. D., Haugen, E., Hickey, P. F., Hou, L., Jasmine, F., Jian, R., Jiang, L., Johnson, A., Kaul, R., Kellis, M., Kibriya, M. G., Lee, K., Billy Li, J., Li, Q., Lin, J., Lin, S., Linder, S., Linke, C., Liu, Y., Maurano, M. T., Molinie, B., Nelson, J., Neri, F. J., Park, Y., Pierce, B. L., Rinaldi, N. J., Rizzardi, L. F., Sandstrom, R., Skol, A., Smith, K. S., Snyder, M. P., Stamatoyannopoulos, J., Tang, H., Wang, L., Wang, M., Van Wittenberghe, N., Wu, F., Zhang, R., Nierras, C. R., Branton, P. A., Carithers, L. J., Guan, P., Moore, H. M., Rao, A., Vaught, J. B., Gould, S. E., Lockart, N. C., Martin, C., Struewing, J. P., Volpi, S., Addington, A. M., Koester, S. E., Little, A. R., Brigham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W. F., Lonsdale, J. T., McDonald, A., Mestichelli, B., Myer, K., Roe, B., Salvatore, M., Shad, S., Thomas, J. A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B. A., Gillard, B. M., Karasik, E., Kumar, R., Miklos, M., Moser, M. T., Jewell, S. D., Montroy, R. G., Rohrer, D. C., Valley, D. R., Davis, D. A., Mash, D. C., Undale, A. H., Smith, A. M., Tabor, D. E., Roche, N. V., McLean, J. A., Vatanian, N., Robinson, K. L., Sobin, L., Barcus, M. E., Valentino, K. M., Qi, L., Hunter, S., Hariharan, P., Singh, S., Sung Um, K., Matose, T., Tomaszewski, M. M., Barker, L. K., Mosavel, M., Siminoff, L. A., Traino, H. M., Flicek, P., Juettemann, T., Ruffier, M., Sheppard, D., Taylor, K., Trevanion, S. J., Zerbino, D. R., Craft, B., Goldman, M., Haeussler, M., Kent, W. J., Lee, C. M., Paten, B., Rosenbloom, K. R., Vivian, J., Zhu, J., Battle, A. & Montgomery, S. B. The impact of rare variation on gene expression across tissues. *Nature* **550**, 239-243, doi:10.1038/nature24267 (2017).

65 King, D. A., Jones, W. D., Crow, Y. J., Dominiczak, A. F., Foster, N. A., Gaunt, T. R., Harris, J., Hellens, S. W., Homfray, T., Innes, J., Jones, E. A., Joss, S., Kulkarni, A., Mansour, S., Morris, A. D., Parker, M. J., Porteous, D. J., Shihab, H. A., Smith, B. H., Tatton-Brown, K., Tolmie, J. L., Trzaskowski, M., Vasudevan, P. C., Wakeling, E., Wright, M., Plomin, R., Timpson, N. J., Hurles, M. E. & Deciphering Developmental Disorders, S. Mosaic structural variation in children with developmental disorders. *Hum Mol Genet* **24**, 2733-2745, doi:10.1093/hmg/ddv033 (2015).

66 Lupski, J. R. Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environ Mol Mutagen* **56**, 419-436, doi:10.1002/em.21943 (2015).

67 Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223-1241, doi:10.1016/j.cell.2012.02.039 (2012).

68 Ruderfer, D. M., Hamamsy, T., Lek, M., Karczewski, K. J., Kavanagh, D., Samocha, K. E., Consortium, E. A., Daly, M. J., MacArthur, D. G., Fromer, M. & Purcell, S. M. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nature genetics* **48**, doi:10.1038/ng.3638 (2016).

69 Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B., Birnbaum, D., The Exome Aggregation, C., Daly, M. J. & MacArthur, D. G. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* **45**, D840-D845, doi:10.1093/nar/gkw971 (2017).

70 Ruderfer, D. M., Hamamsy, T., Lek, M., Karczewski, K. J., Kavanagh, D., Samocha, K. E., Exome Aggregation, C., Daly, M. J., MacArthur, D. G., Fromer, M. & Purcell, S. M. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat Genet* **48**, 1107-1111, doi:10.1038/ng.3638 (2016).

71 Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H. H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G. & Exome Aggregation, C. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

72 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, doi:10.1093/nar/gkt857 (2014).

73 Greenwald, W. W., Li, H., Benaglio, P., Jakubosky, D., Matsui, H., Schmitt, A., Selvaraj, S., D'Antonio, M., D'Antonio-Chronowska, A., Smith, E. N. & Frazer, K. A. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nat Commun* **10**, 1054, doi:10.1038/s41467-019-08940-5 (2019).

74 Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping (vol 159, pg 1665, 2014). *Cell* **162**, 687-688, doi:10.1016/j.cell.2015.07.024 (2015).

75 Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B. M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W., Dimitrova, E., Dimond, A., Edelman, L. B., Elderkin, S., Tabbada, K., Darbo, E., Andrews, S., Herman, B., Higgs, A., LeProust, E., Osborne, C. S., Mitchell, J. A., Luscombe, N. M. & Fraser, P. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research* **25**, 582-597, doi:10.1101/gr.185272.114 (2015).

76 Montefiori, L. E., Sobreira, D. R., Sakabe, N. J., Aneas, I., Joslin, A. C., Hansen, G. T., Bozek, G., Moskowitz, I. P., McNally, E. M. & Nobrega, M. A. A promoter interaction map for cardiovascular disease genetics. *Elife* **7**, doi:10.7554/eLife.35788 (2018).

77 Babbs, C., Lloyd, D., Pagnamenta, A. T., Twigg, S. R., Green, J., McGowan, S. J., Mirza, G., Naples, R., Sharma, V. P., Volpi, E. V., Buckle, V. J., Wall, S. A., Knight, S. J., International Molecular Genetic Study of Autism, C., Parr, J. R. & Wilkie, A. O. De novo and rare inherited mutations implicate the transcriptional coregulator TCF20/SPBP in autism spectrum disorder. *J Med Genet* **51**, 737-747, doi:10.1136/jmedgenet-2014-102582 (2014).

78 Sun, J. H., Zhou, L., Emerson, D. J., Phyo, S. A., Titus, K. R., Gong, W., Gilgenast, T. G., Beagan, J. A., Davidson, B. L., Tassone, F. & Phillips-Cremins, J. E. Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175**, 224-238 e215, doi:10.1016/j.cell.2018.08.005 (2018).

79 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, doi:10.1038/nrg3461 (2013).

80 Banovich, N. E., Li, Y. I., Raj, A., Ward, M. C., Greenside, P., Calderon, D., Tung, P. Y., Burnett, J. E., Myrthil, M., Thomas, S. M., Burrows, C. K., Romero, I. G., Pavlovic, B. J., Kundaje, A., Pritchard, J. K. & Gilad, Y. Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res* **28**, 122-131, doi:10.1101/gr.224436.117 (2018).

81 Carcamo-Orive, I., Hoffman, G. E., Cundiff, P., Beckmann, N. D., D'Souza, S. L., Knowles, J. W., Patel, A., Papatsenko, D., Abbasi, F., Reaven, G. M., Whalen, S., Lee, P., Shahbazi, M., Henrion, M. Y. R., Zhu, K., Wang, S., Roussos, P., Schadt, E. E., Pandey, G., Chang, R., Quertermous, T. & Lemischka, I. 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, doi:10.1016/j.stem.2016.11.005 (2017).

82 Pashos, E. E., Park, Y., Wang, X., Raghavan, A., Yang, W., Abbey, D., Peters, D. T., Arbelaez, J., Hernandez, M., Kuperwasser, N., Li, W., Lian, Z., Liu, Y., Lv, W., Lytle-Gabbin, S. L., Marchadier, D. H., Rogov, P., Shi, J., Slovik, K. J., Stylianou, I. M., Wang, L., Yan, R., Zhang, X., Kathiresan, S., Duncan, S. A., Mikkelsen, T. S., Morrisey, E. E., Rader, D. J., Brown, C. D. & Musunuru, K. 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, doi:10.1016/j.stem.2017.03.017 (2017).

83 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

84 Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kahari, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadissa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T. J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R. & Searle, S. M. Ensembl 2014. *Nucleic Acids Res* **42**, D749-755, doi:10.1093/nar/gkt1196 (2014).

85 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

86 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, doi:10.1007/978-1-4939-0512-6_3 (2014).

87 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, doi:10.1093/bioinformatics/btp616 (2010).

88 Slifer, S. H. PLINK: Key Functions for Data Analysis. *Curr Protoc Hum Genet* **97**, e59, doi:10.1002/cphg.59 (2018).

89 Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. *bioRxiv*, 003905, doi:10.1101/003905 (2014).

90 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, doi:10.1093/bioinformatics/btv722 (2016).

91 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445, doi:10.1073/pnas.1530509100 (2003).

92 Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R. & Hubbard, T. J. GENCODE: the reference human

genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).

93 Montefiori, L. E., Sobreira, D. R., Sakabe, N. J., Aneas, I., Joslin, A. C., Hansen, G. T., Bozek, G., Moskowitz, I. P., McNally, E. M. & Nóbrega, M. A. A promoter interaction map for cardiovascular disease genetics. *eLife* **7**, 1-35, doi:10.7554/eLife.35788 (2018).