RESEARCH STRATEGY

SIGNIFICANCE

Structural variations (SVs), such as deletions, duplications, insertions, inversions, copy number variations and translocations, are genetic variations and structurally diverse ranging from simple events to complex rearrangements. SVs affect far more bases than single-nucleotide polymorphisms (SNPs) combined. SVs can markedly affect phenotype in many ways, including modification of open reading frames, production of alternatively spliced mRNAs, alterations of transcription factor (TF) binding sites, and structural gains or losses within the regulatory regions. Consortium efforts such as the 1000 Genomes Project (1000GP) estimate that a typical genome contains 2.1–2.5 thousand SVs, affecting ~20 million bases, or ~5–6 times that of SNPs[1,2]. In the 1000GP, we also found that a typical genome contains ~150 LoF variants and discovered significant depletion of SVs (including deletions, duplications, inversions and multiallelic copy number variations) in coding sequences, untranslated regions, and introns of genes as compared to a random background model, implying strong purifying selection. Furthermore, some studies have shown that the complexity of SVs' breakpoints are much higher than estimated which suggests that SVs are widespread in human genomes and are appreciably more difficult to discover than previously thought[3].

SVs are common, larger in size, and more structurally diverse than SNPs, and they are likely to profoundly shape the regulation of many human phenotypes and disease states. SVs have long been associated with complex diseases. Their effect could derive, for example, from gene dosage, or by disruption coding regions[4]. The cause of a complex disease could derive either from the SV alone or in combination with other genetic or environmental factors[5]. SVs have been described as associated not only with sporadic Mendelian traits and disease susceptibility, but also with complex diseases such as mental disorders (autism[6–8], schizophrenia[9–12] and mental retardation[13–15]), asthma[16–20] and cardiovascular disorder[21–24], the last of which is the focus of this proposal.

Cardiovascular disease (CVD) is a public health concern affecting over 80,000,000 people and accounting nearly 801,000 deaths in the United States that is about 1 in 3 deaths. Globally, CVD is also the leading cause of death, accounting 17.9 million or 32.1% of all global deaths in 2015[25–27]. CVD is a class of complex pathologies of the heart and blood vessels and the most prevalent manifestations include coronary heart disease (e.g. heart attack), cerebrovascular disease (e.g. stroke), heart failure, cardiac arrhythmia, and heart valve problems. Most cardiovascular disease affects older adults. In the United States, 11% of people between 20 and 40 have CVD, 37% between 40 and 60, 71% of people between 60 and 80, and 85% of people over 80 have CVD. However, genetic factors may develop cardiovascular diseases in people who are less than 55 years-old and people having parents affected in CVDs increase their risk by 3 fold. Thus, this proposal mainly focuses on genetic structural variations in CVD.

Pedigree linkage studies[28] and genome-wide association studies (GWASs)[29–33] have shown that these diseases are influenced by inherited genetic variations and hundreds of loci associated with cardiovascular pathologies are identified. However, due to the complexity of cardiovascular disease, our knowledge of genetic contributions to CVD is still poor. The sensitivity for detecting the primary genetic defect is still approximately 50%. This elicits the importance of the use of next-generation sequencing and the essence of deciphering structural variations (SVs) to conquer the considerable proportion of the missing heritability of CVD[34–38].

Investigating SVs, could therefore hold the key to a deeper, more mechanistic understanding of the genetic basis of CVD. At present, most studies do not capture the spectrum of SVs present in genomes, so this complexity is not adequately accounted for in disease association studies. To the best of our knowledge, only a hotspot of short insertion-deletion polymorphisms in *NCX1*[21] and few copy number variations[23,24] are reported in relation to CVD. Furthermore, the functional impact of SVs, especially in non-coding regions[39], has not been investigated systematically. Surmounting these issues depends on stable computational methodologies for 1) mining whole genome sequencing datasets for SV discovery at high resolution and large scale, 2) functionally interpreting their origins and phenotypic effects, and 3) establishing associations between specific SVs and disease. A pipeline characterized with these three features has been developed by the team.

Here, we propose to apply our pipeline to understand the genetic basis of cardiovascular disease through computationally driven discovery, functional validation, and characterization of CVD-associated SVs within the CVD related cohorts being sequenced as part of the TOPMED program. Our SV detection pipeline embedded several SV-calling algorithms is able for a high-resolution SV discovery and for a comprehensive profile of all types of SVs. The pipeline will be applied on the four studies, San Antonio Family Studies (SAFS CVD), Framingham Heart Study (FHS), Jackson Heart Study (JHS), and Cardiovascular Health Study (CHS) in the

TOPMED program, which in total will be ~15,000 sequenced genomes (**Aim 1**). To examine the functional impact of the identified SVs, we will apply our method to an integration of RNA-seq data for functional annotation of variants and characterization of associated biological processes (**Aim 2**). Finally, we will use SVs from **Aim 1** and their impact scores from **Aim 2** to discern genotype-phenotype associations for disease-based SV association studies (**Aim 3**). Our deliverables will be the largest library of validated SVs discovered in a combined cohort of ~15,000 cardiovascular disease patients and related individuals, together with an unprecedented platform of cloud-based pipelines for comprehensive, high-resolution, and large-scale SV analysis.

Scientists participating in the proposed project are leaders in SV discovery and analysis. The three PIs, Charles Lee, Ph.D., Mark Gerstein, Ph.D. and Li Ding, Ph.D., have a history of productive scientific collaboration and bring complementary experience in SV detection (Lee), functional interpretation (Gerstein) and large-scale data analysis (all), particularly association analysis (Ding). Each also brings significant experience in leading (1000GP SV group, Lee; modENCODE AWG, Gerstein; ENCODE networks group, Gerstein; PsychENCODE AWG, Gerstein; exRNA AWG, Gerstein) and participating in (1000GP, Lee/Gerstein/Ding; ENCODE, Gerstein; ICGC, Gerstein/Ding; TCGA, Ding; CPTAC, Ding; KBase, Gerstein; GSP (Genome Sequencing Program), Gerstein) large-scale sequencing consortia. Under Dr. Lee's leadership, the 1000GP SV project identified SV events in 2,504 healthy genomes and helped define the methodologies for identifying and characterizing SVs from "lower depth" (mean depth = 7.4X) whole genome sequencing (WGS) datasets. Dr. Travis Hinson, co-Investigator, brings a wealth of knowledge about cardiovascular disease. He will serve as an integral member of the investigative team providing the essential clinical perspective and disease context to the characterization of SVs discovered in the TOPMED datasets and the association analyses of SVs to cardiovascular diseases.

INNOVATION

The originality of this proposal lies in the integration of cutting-edge computational methodologies—pioneered by the group—into a comprehensive, cloud-ready platform for novel SV discovery, characterization, and association with cardiovascular disease biology across the large assembled cohort of CVD patients and related individuals. Our proposed detection and genotyping strategy will meet the need for power and resolution for investigating association between SVs (that span a large size spectrum) and phenotypes, surpassing previous standard approaches employed in current SV association studies. The key innovations of our approach lie in its characteristics of: **1)** *Scalability*: Our cutting-edge SV detection and integration tools will provide the capability to perform high-resolution discovery and classification of SVs, and identify well-powered genotype-phenotype associations in a disease context. **2)** *Integration*: Our approach will integrate identified SVs with RNA-seq data and other functional data from coding and non-coding regions of the genome to provide scores for functional impact. **3)** *Extended functionality*: CVD has multiple and different manifestations so tools for mechanistic interpretation of SVs across different manifestations will allow us to make better inferences about each CVD manifestation associated SVs. **4)** *Sensitivity*: Association tests that integrate weighting methods for various biological considerations, such as allele frequency and impact score, will enable a generalized linear model to capture subtle association signals often missed by conventional approaches. **This systematic survey of SVs will yield the largest database of validated SVs associated with cardiovascular disease, together with an unparalleled system for high-dimensional, high-resolution studies of SV architecture and function.**

RESEARCH STRATEGY:

Specific Aim 1. Identifying complex structural variations on large-scale CVD-related genomes.

Rationale. To drive the discovery phase of thousands genomes in the TOPMED program, we will apply our SV detection pipeline, Structural Variation Engine (**SVE**), consisting of eight employed state-of-the-art SV-calling algorithms and **fusorSV** (manuscript in preparation, **Figure 1**). The eight SV-calling algorithms are BreakDancer[40], BreakSeq[41], cnMOPS[42], CNVnator[43], Delly[44], GenomeStrip[45], Hydra-Multi[46], and Lumpy[47], and each of them has its advantages and weaknesses for certain types of SV detection. To properly keep advantage and mitigate weaknesses of each SV-calling algorithm, we developed **fusorSV** to merge results from the eight SV-calling algorithm. **fusorSV** is an open source framework that takes a data mining approach by incorporating knowledge of the strengths of various existing SV callers (discovered using a truth set), and uses this knowledge to perform discovery on a novel cohort of genomes. The pipeline has multiple entrance points and for this project we will start at given BAM files. According to the reports from the TOPMED Informatics Research Center (IRC, http://nhlbi.sph.umich.edu/report/), we anticipate to receive quality controlled GRCh38 sequence alignment files for each sample. The pipeline will be applied to the entire set of CVD-related individuals being sequenced

by SAFS CVD, FHS, JHS, and CHS in the TOPMED program. The approximate sample size is ~15,000 which brings a great challenges of the pipeline's robustness and stability. Raw SV calls, described in VCFs, will be generated by each employed SV-calling algorithm and **fusorSV** will consolidate VCFs based on the pre-calculated model. The pre-calculated model is trained on high-coverage samples from the 1000GP. Afterwards, by using breakpoint assembly methods, we will perform *in silico* validation (**Figure 2)** of the SV events and use the assembled contigs to investigate the inherent complexity prevalent at breakpoints, as well as mechanisms of SV formation. **Ultimately, these studies will deliver the most comprehensive library of complex SVs discovered in people affected by cardiovascular disease and will enable us to make novel biological inferences at the population level.**

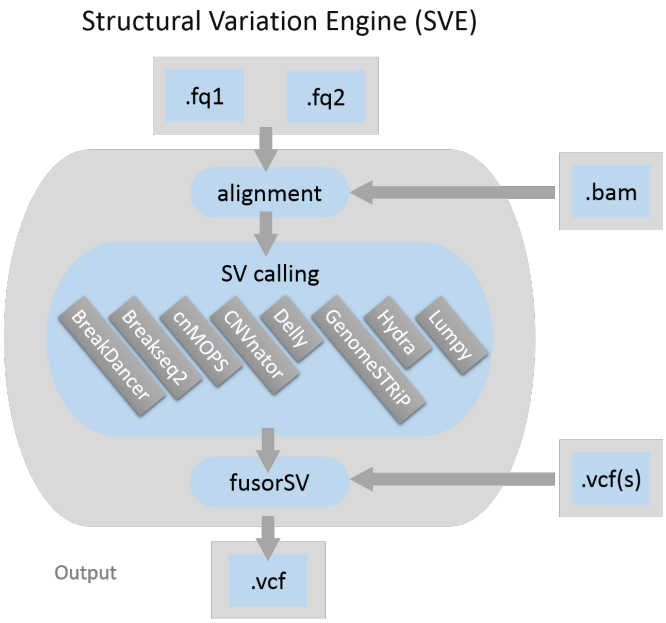

Figure 1. Structural Variation Engine. The overall work includes 1) Alignment, 2) SV calling and 3) VCF Consolidation. There are multiple entrance points of the pipeline to make the flexibility for users to process data.

Preliminary data. *A toolbox of methods for structural variation discovery.* As part of the 1000GP SV project, we have provided the research community with an unprecedented set of germline SVs from 2,504 normal human genomes that have been sequenced at low depth and have developed a large collection of complementary tools and methods, including: **1) Read-depth–based tools**. We developed CNVnator for copy number variant (CNV) discovery and genotyping from individual and trio-sequencing datasets. It utilizes a mean-shift approach, GC correction, and bandwidth partitioning to identify a wide range of CNV events. CNVnator can detect CNVs and provide genotype information on a population level, and also detects atypical CNVs including *de novo* and multi-allelic events. **2) Paired-end–based tools**. Meerkat[48], Hydra-Multi, PEMer[49] and BreakDancer cluster abnormally mapped paired-end reads to identify loci with a signature for an SV event. Meerkat remaps soft clipped and unmapped reads to generate clusters to identify breakpoints. Pindel-C[50,51] utilizes a pattern-growth approach to detect large deletions and insertions, including complex events, from WGS data. These methods have each already been successfully applied to hundreds of cancer genomes[48,52]. **3) Split-read-alignment–based tools**. We have also developed SRM[53],SRIC[54], and Tangram[55,56] for the high-resolution identification of SV events from WGS datasets. These tools specifically aim to provide single-nucleotide resolution of breakpoints—an invaluable feature that enables functional interpretation of the biology of these SV events. Tangram is a tool utilizing both paired-end and split-read approaches for mobile element insertion detection.

Breakpoint assembly tools for in silico validation. Pinpointing SV breakpoints with single-nucleotide resolution is essential to produce accurate individual genotypes in clinical samples. In our detection pipeline, we have already developed algorithms for identifying breakpoints at nucleotide resolution, thereby allowing us to validate SV breakpoints *"in silico"*. Primary short-read mappers, such as BWA[57], BOWTIE[58], and MOSAIK[59], do not usually map reads crossing SV breakpoints, and thus assembling those reads for SV breakpoints becomes a solution for SV *in silico* validation (**Figure 2**). As previously studies, we used assembly-based methods like SGA[60] or TIGRA-SV[61] for generating sequence contigs at SV breakpoints that improves breakpoint resolutions from 58.5% to 64.8%[52]. We also developed AGE[62], which performs sequence alignment at regions flanking SVs while considering large deletion and insertion blocks, which cannot be handled by conventional sequence alignment algorithms.

Ensemble approach to SV discovery. **SVE** (Figure 1, manuscript in preparation) consisting of eight employed state-of-the-art SV-calling algorithms and **fusorSV** is a stable pipeline and designed for large-scale complex SV analysis on the cloud or on traditional high-performance compute clusters. *fusorSV* takes a data mining approach by incorporating knowledge of the strengths of various existing SV callers, and uses this knowledge to perform discovery on a novel cohort of genomes. The pipeline has been tested on a dataset from 1000GP with 27 deep-coverage samples. Using the annotated SVs from the 1000GP Phase 3, we built a model using 18 samples and applied the model to the other 9 samples for SV discovery *ab initio*. This step was repeated 1000 times with random selection for the 18 learning samples and the 9 test samples. **Figure 3** shows the performance of **fusorSV** as compared to some other popular SV-calling algorithms that were integrated in **SVE**.

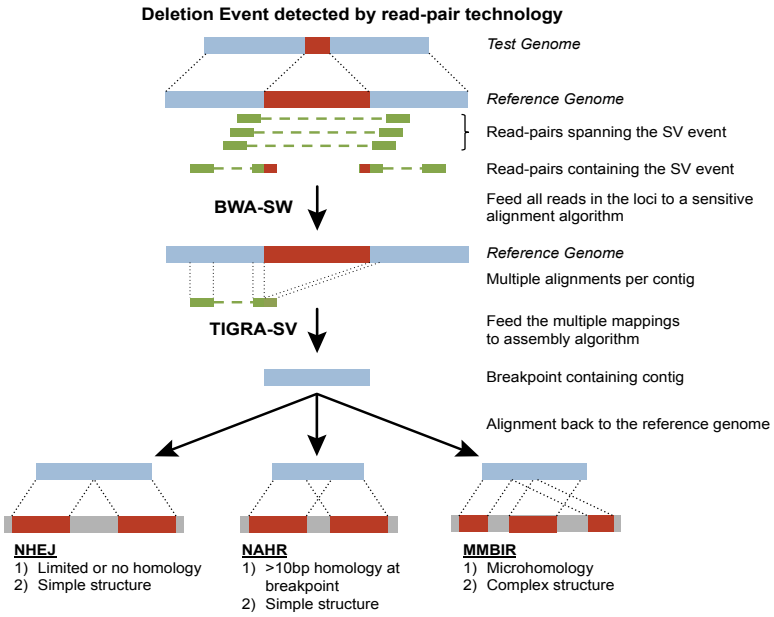

Figure 2. Breakpoint assembly for in silico validation. The top half of the figure shows a deletion SV event predicted by the read pairs spanning the event. All read pairs in the breakpoint locus are used for targeted *de novo* assembly and the resulting contig is aligned back to the genome.

As it can be seen, **fusorSV** outperforms all the SV callers by optimizing both precision and recall on the 1000GP Phase 3 callset. Precision and recall are defined as $prec = \frac{true\_positive}{true\_positive + false\_positive}$ and $recall = \frac{true\_positive}{true\_positive + false\_negative}$, respectively. *True_positive* is all retrieved calls by each SV-calling algorithm that overlap with calls reported by the 1000GP while *false_positive* is a set of calls that are reported by the algorithm but not by the 1000GP and *false_negative* then means calls reported by the 1000GP but not by the algorithm. Even with a strict metric such as the Jaccard Similarity score[63], **fusorSV** outperforms all other SV callers for SV discovery in the test set. Furthermore, **fusorSV** identified 562 (~10%) novel SV calls from the cohort of 27 genomes that were not reported by the 1000GP. We performed *in vitro* validation on a subset of SVs from this cohort and achieved a positive validation rate of 74.3% (78 positively validated SVs out of 105 tested SVs).

Research Plan. We plan to deploy and apply **SVE** on the cloud to identify and classify SVs across WGS datasets from the identified projects of the TOPMED program. We will deliver 1) integrated and comprehensive identification of a broad spectrum of SV types and 2) breakpoint resolution identification based on TIGRA-SV or similar assembly-based SV-calling algorithms.

Sample selection. Data storage and computing resource are required for SV discovery on the entirety of CVD-related genomes in the TOPMED program. We have identified four CVD related studies in the program, San Antonio Family Studies (SAFS CVD), Framingham Heart Study (FHS), Jackson Heart Study (JHS), and Cardiovascular Health Study (CHS). Combined, these cohorts plan to generate sequence from ~15,000 CVD patients and related individuals. We appreciate the enormity of the proposed analysis, and to ensure efficient use of resources, the entire dataset would be analyzed in multiple phases as described below.

Pipeline for population-level structural variant discovery. During phase 3 of the 1000GP SV project, we used an ensemble of eight algorithms for SV discovery. A callset of an individual was generated by each SV-calling algorithm and then merged into a single release of the sample by **fusorSV**. The proposed pipeline (**Figure 1**) for SV discovery will extend this work with the following salient features: **1)** Standard steps for quality control, duplicate removal, and alignment for all selected samples if necessary (a quality-controlled GRCh38 sequence alignment file for each sample is actually expected from the TOPMED program); **2)** A separate result and an ensemble of SV-calling algorithms including BreakDancer[40], BreakSeq[41], cnMOPS[42], CNVnator[43], Delly[44], GenomeStrip[45], Hydra-Multi[46], and Lumpy[47] for CVD genomes. This ensures that a particular algorithm does not bias the discovered SV set and increases our power to detect true SV events by asking for evidence by multiple methods; **3)** *In silico* validation for discovered set of SV sites using a library of known common variants; **4)** Complex SV identification using assembly-based tools for assessing breakpoints at nucleotide resolution.

The SV calling will be performed in two phases: ***Phase 1—Calibration*** *(Tasks 1 in* ***Figure 4****):* We will launch

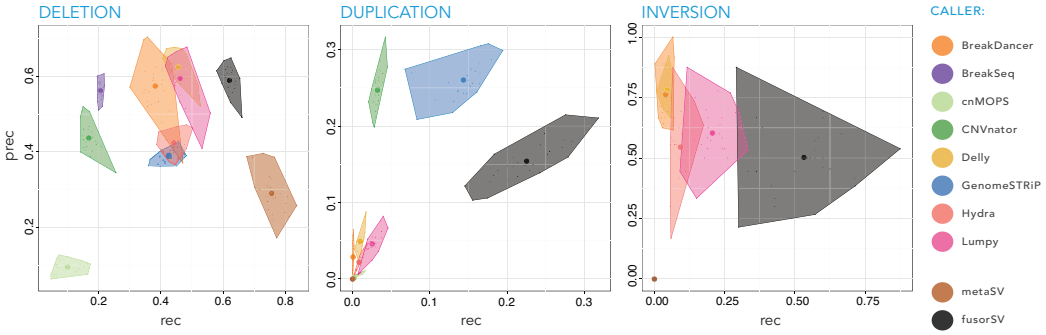

Figure 3. fusorSV cross fold validation using 1000GP samples. The 3 panels plot precision (y-axis) versus recall (x-axis) for Deletions, Duplications and Inversions.

a test within a selected cohort with about 100 CVD-related samples from the TOPMED program. The goal of the *Calibration* phase is to deploy the calibrated pipeline on the Google Cloud Platform and to test it for efficiency and eventual scale up in the next discovery phase. Based on the data access and the computational strategies descripted in the TOPMED program, we will explore parallelization where the tools already support this capability. According to the reports from the TOPMED Informatics Research Center (IRC, http://nhlbi.sph.umich.edu/report/), we anticipate to receive quality controlled GRCh38 sequence alignment files for each sample. The computational intensive steps in the **SVE** discovery pipeline that would be primary candidates for optimization are **1)** SV classification by the eight SV-calling algorithms, **2)** SV events consolidation, and most importantly **3)** clustering of aberrant reads for SV breakpoint assembling. ***Phase 2—Discovery*** *(Task 2):* The optimized system from Task 1 will be applied on the entire set of ~15,000 individuals sequenced as part of the four selected CVD related cohorts. We have done extensive preliminary analysis of the **SVE** detection pipeline on our own computing infrastructure at the Jackson Laboratory (JAX) to get an estimated amount of computing resources needed for the entire proposed computation. Preliminary results suggest that we need ~12 hours of CPU time per average 30X coverage whole genome sequence sample. Using a standard Google instance (n1-standard-16) with enough cores (16) and enough RAM (60GB) at an estimated cost of ~$28,080. The temporary storage required for BAM files is about 100GB per sample. Considering the stability of our pipeline, we aim to keep sample BAM files for 2 weeks estimated to be $5,120 for the project. We also estimate the cost of VCF storage (1GB per sample) to be $2,460. The total estimated cost of $35,660 for computing and storage is allocated across Years 1 and 2.

Aim	Task	Year 1				Year 2			
Aim 1	**Identify structural variations on large-scale CVD-related genomes**								
	Task 1: Deploy and optimize *SVE* detection pipeline on cloud platform	■							
	Task 2: Perform SV discovery on the entire cohort on cloud platform		■	■					
Aim 2	**Analyze the functional impact of structural variations**								
	Task 1: Deploy and calibrate *SVIM* on cloud platform using detected SVs, integrated RNA-Seq			■	■				
	Task 2: Process and annotate all discovered SVs using the *SVIM* pipeline					■	■		
Aim 3	**Association of structural variants with burden in CVD cohorts**								
	Task 1: Deploy and optimize *SV2Pheno* on cloud platform					■	■		
	Task 2: Perform associate studies with the discovered SVs and build models of CVD association							■	■

Figure 4. Project timeline

Calibration of method using known sites. Hundreds of sites across the human genome are polymorphic in a large fraction of the population[64,65]. Phase 3 of the 1000GP SV project[2] showed that a significant fraction of SVs (35%) occurs at a high frequency in the population (variant allele frequency ≥ 0.2%). For those common SVs, we will create a catalog of structural variation polymorphic sites across the genome and use them as validation sites for our SV-calling methods.

Validation of SV sites using in silico assembly–based methods. We demonstrated above that SVs can be validated *in silico* using targeted *de novo* assembly–based methods (TIGRA-SV or SGA). The same methodology was integrated into the **SVE** detection pipeline and will be used to process every discovered SV site for validation.

Complex SV identification. Complex SVs are a class of rearrangements of simple SVs, such as deletions, duplications, insertions, inversions, and copy number variations. Due to the limitation of the SV-calling algorithms, some types of SVs may be caught by certain SV-calling algorithms that never generate other types of SVs. We will use the two methods for complex SV identification. The first method will identifies SV clusters present in the same genomic region that have similar allele frequencies and copy number ratios. This will help to select SV that are part of the same complex SV event. The second method involves inspecting the mapping patterns of various parts of the assembled contig at the SV site. This would allow us to identify mislabeled SVs and SVs with more complexity than annotated by the individual SV-calling method.

Data access strategies: Total storage of the discovery cohort is expected to require ~1.5 PB. To manage the data corpus and computing requirements, we propose to use the Google Cloud Platform which will be available to all members of our team. JAX is currently expanding capabilities in cloud-based data analysis to address issues, including access to increased compute power, co-localization of novel and reference datasets and

reproducibility of analysis pipelines. JAX staff have adapted multiple pipelines for the cloud platform and evaluated the suitability of the cloud-based archival storage for genomics datasets. Dr. Ding's group has developed **GenomeVIP**, a secure, HIPAA-compliant, web-driven variant discovery and annotation platform through which multiple independent analysis tools can be applied to a given dataset. As it can call upon both local high-performance computing (HPC) and cloud resources, **GenomeVIP** is a tool that we may initially use to assist with variant discovery and to download results to local disks for subsequent analyses.

JAX is partnering and collaborating with commercial genomics cloud service providers (CSPs) on several important projects and has recently recruited cloud computing experts as part of the Research IT department. These activities are independent of this proposal and would aid us in providing the experience necessary for successful completion of various aspects of this project.

Expected results. This aim will yield a comprehensive catalog of validated SVs from CVD-related genomes in the TOPMED program that lay the foundation for subsequent functional interpretation and association studies (**Aims 2 and 3**). It will also help answer questions about SV formation and population-level associations of SVs across the various cardiovascular disease studies in the program. By making the **SVE** detection pipeline available as a community resource and demonstrating the correctness and comprehensiveness of the SV results, we expect this work to propel future genome-level SV analyses for the entirety of the TOPMED program and other large consortia.

Pitfalls and alternative approaches. A major challenge for this aim is the diversity of data that are being collected and of the variable availability of orthogonal data (genomic, transcriptomic, proteomic, etc.) across the various selected cohorts. In response, we will leverage the extensive experience of the team to handle complex datasets (see Preliminary data section) and design **SVE** to robustly handle diverse and complex datasets of the types that might be generated by the TOPMED Program. Another challenge of which we are mindful is the enormity of the proposed computation. The assembled team has extensive experience both in dealing with very large datasets and in developing a multi-phase strategy for the proposed computation that will make efficient use of resources. We are aware that **fusorSV**'s sensitivity and specificity values presented are moderate for proper genome wide associations, but we hypothesize that these are due to a small sample size, twenty seven genomes. Our preliminary results from a cohort of 100 simulated samples suggest that the discovery false discovery rate improves several folds, given enough number of datasets.

Specific Aim 2. Scoring the functional impact of structural variations.

Rationale. There is still little known about the functional impact of SVs at a genome-wide level. SVs are disproportionately observed in the non-coding part of the genome; hence, a comprehensive assessment of the functional impact of SVs will likely require the integration of large-scale data resources such as ENCODE, 1000GP and GTEx. To functionally prioritize SVs in preparation for disease association studies, we propose to use **SV Impact (SVIM)**, an analysis tool that integrates myriad datasets- including existing annotations, allelic activity from RNA-seq, and eQTLs from RNA-seq.

Preliminary data. *Tools for assessing functional impact of genomic variation in genes and pseudogenes.* We developed Variant Annotation Tool (VAT) to annotate the impact of protein sequence mutations[66]. VAT provides transcript-specific annotations of point mutations and insertions/deletions (indels) according to synonymous, missense, nonsense, or splice-site-disrupting changes. We observed that genes tolerant of loss-of-function (LoF) mutations are under the weakest selection. In 1000GP Phase 3, we found that a typical genome contains ~150 LoF variants and discovered significant depletion of SVs (including deletions, duplications, inversions and multiallelic CNVs) in coding sequences, untranslated regions, and introns of genes as compared to a random background model, implying strong purifying selection.

Tools for evaluating functional impact of variation in non-coding (nc) RNAs and regulatory regions. We have developed tools to specifically analyze ncRNAs. Our incRNA pipeline combines sequence, structural, and expression features to classify newly discovered, transcriptionally active regions into RNA biotypes, such as miRNA, snRNA, tRNA and rRNA[67]. Our ncVar pipeline further analyzes genetic variants across biotypes and subregions of ncRNAs, e.g., showing that miRNAs with more predicted targets show higher sensitivity to mutation in the human population[68].

To better understand nc regulatory regions, we developed tools to analyze ChIP-Seq data to identify genomic elements and interpret their regulatory potential. PeakSeq identifies regions bound by TFs and chemically modified histones[69]; it has been widely used in consortium projects such as ENCODE[70]. The second generation of PeakSeq is a newly developed tool that uses multiscale decomposition to help identify enriched regions in

cases where strict peaks are not apparent and robustly calls both broad and punctate peaks[71]. Peak calls and ChIP-Seq signal data can also be used to model gene expression and annotate target genes. We have developed methods that use both supervised and unsupervised machine-learning techniques to identify these regulatory regions (such as enhancers) and predict gene expression from ChIP-Seq data[72–75]. To investigate the evolutionary importance of these regions, we have analyzed patterns of single nucleotide variation within functional nc regions, along with their coding targets[68,75,76]. We used metrics such as diversity and fraction of rare variants to characterize selection pressure on various classes and subclasses of functional annotations[68]. We have also defined variants that are disruptive to a TF-binding motif in a regulatory region[70].

Tools for helping annotate functional impact based on network. We found that functionally significant and highly conserved genes tend to be more central in various biological networks[77] and are positioned at the top of regulatory networks[76]. Further studies showed relationships between selection and protein network topology (e.g., quantifying selection in hubs relative to proteins on the network periphery[77,78]). Incorporating multiple network and evolutionary properties, we developed NetSNP[77] to quantify the indispensability of genes. This method shows strong potential for interpreting the impact of variants involved in Mendelian diseases and in complex disorders probed by GWAS. We constructed regulatory networks for data from the ENCODE and modENCODE projects, identifying functional modules and network hierarchy[76]. To quantify the degree of hierarchy for a given hierarchical network, we defined a metric called hierarchical score maximization (HSM[79]).

FunSeq: Tools for integrated functional prioritization. We recently developed a prioritization pipeline called FunSeq[80,81] that identifies annotations under strong selective pressure as determined using genomes from many individuals from diverse populations. FunSeq links each nc mutations to target genes and prioritizes based on scaled network connectivity. FunSeq identifies deleterious variants in many nc functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites and detects their disruptiveness in TF-binding sites (both LoF and gain-of-function events). Due to the complexity of cardiovascular disease and multiple manifestations of CVD, we may classify SVs by manifestations and recalibrate the functional networks.

Mutational mechanisms of structural variants. The sequence content of SVs, especially around breakpoints, carries important information about origin and functional impact. Using datasets from the 1000GP, we studied the distinct features of SVs originating from different mechanisms[80,82]. We performed SV mechanism annotations for the 1000GP Phase 3 deletions using BreakSeq[41], categorizing 29,774 deletions by their creation mechanisms. Among these, non-homology-based rearrangement proved to be the most prevalent mechanism (~73% of all categorized deletions)[2]. These results inform us on the molecular mechanisms underlying SV formation and also indicate differences in functional impacts of different SV types.

Tools for uniform processing of RNA-seq data. We have considerable expertise in analyzing RNA-Seq data, including experience in developing and configuring pipelines for the processing of RNA-seq data, especially for long RNA-seq data for ENCODE, long and short RNA-seq data for the PsychENCODE[83] and Brainspan project, and a custom pipeline developed for the analysis of small exRNA-seq data for the Extracellular RNA Communication Consortium (ERCC). We have already developed an efficient in-house data processing workflow for RNA-seq data that includes data organization, format conversion, and quality assessment. RSeqTools[84] is a modular tool developed for the processing of RNA-seq data and generating either transcript, gene, or exon level quantifications. We also developed IQSeq[85] which calculates the relative and absolute abundance of contributing transcript isoforms to a gene from RNA-seq data using a fast algorithm based on the Fisher information matrix. Another tool we developed called FusionSeq[86] detects fusion transcript in RNA-seq data, which can be important biomarker for diseases such as cancer and neurological diseases.

Tools for allele activity and eQTL detection. We have also developed tools specifically for linking gene expression variation to genotype, including our Allele-Seq pipeline, which quantifies allele-specific gene expression by mapping reads onto a diploid personal genome built from called genetic variants, including SNPs, short indels, and structural variants[87]. We recently applied this pipeline on a population scale to RNA-Seq data from the 1000 Genomes Project and used this analysis to create AlleleDB, a database of genomic regions with high allelic activity[88]. Our expertise in eQTLs is demonstrated in our novel study on successfully utilizing expression-variant correlations to construct predicted genotypes. These predicted genotypes were then matched with known genotypes from a given dataset in order to demonstrate how the information security of the given dataset may be compromised[89].

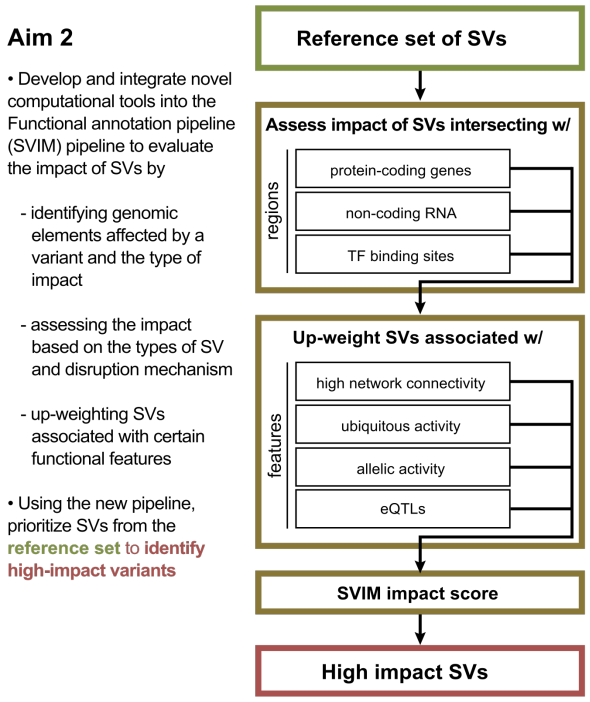

Figure 5. Overview of the functional prioritization and annotation pipeline

Research plan. To enable identification of SVs with high functional impact, we will use an extension of FunSeq/FunSeq2 within called **SVIM** (**S**tructural **V**ariation **IM**pact) (**Figure 6)**. We will evaluate the impact score for each SV, taking into account the functional annotation of the affected genomic region and the fraction of functional elements (i.e., genes, ncRNAs, nc regulatory elements). We will also upweight SVs based on ubiquitous activity, allelic activity and eQTLs. The impact score will also depend upon SV type (i.e., deletion, duplication, inversion or translocation).

For a given SV belonging to a particular SV type, we will use break point resolution coordinates to estimate the fraction of bases overlapping functional elements. Based on this fraction, we will categorize SVs into three classes (touch, cut, and engulf). Each overlapping class will have a different weight ($F_{svtype,\ class}$). We will divide genomic elements into three categories (coding region, nc region, TF binding site) and assign relative scores to them (S_{coding}, $S_{non\text{-}coding}$, S_{TFBS}), which will vary for different SV types. Relative scores F and S will be defined for class and functional elements analogous to the FunSeq2 tool[81].

$IS_{orig} = \sum_i \left(F_{j,k} \times S_{j,i} \times \delta_i\right) \times \prod_l g_l$; $IS_{norm} = \frac{IS_{orig} - \overline{IS_{random}}}{\sigma_{random}}$, where i is a functional element $\in \{protein\ coding, noncoding\ RNA, nonconding\ regulatory, allelic\ activity, eQTL\}$; k is a overlapping classification $\in \{cut(0.1 \leq f < 0.8), touch(f < 0.1), engulf(f \geq 0.8)\}$, and f is the fraction of functional element overlapping the SV; j is the type of SV ; $\delta \in \{0,1\}$; and l is a feature $\in \{connectivity, ubiquitous\ activity, allelic\ activity, eQTLs\}$;

SVs will be assigned an impact score by taking the sum over the product between weights of overlapping classes and scores of overlapping functional elements. The score (IS_{orig}) will also be upweighted based on activity of the affected region. The upweight factor is comprised of the product of four factors: i.e., allelic activity, eQTLs, network connectivity and ubiquitous activity. Significance level of an Impact score (IS_{orig}) will be estimated by running 1,000 Monte Carlo simulations generated by randomly shuffling the location of SVs.

Evaluating effect of structural variants on protein-coding genes. We will analyze loss of function (LoF) variants with mis-mapping, functional, evolutionary and network features of protein coding genes overlapping with SVs. We will first identify LoFs due to whole gene deletion, as well as putative LoF-causing mutations as those that induce premature stop codons, frameshifted open reading frames, or that we predict to produce truncated proteins due to deletion of RNA splice sites or either predicted or verified changes in splicing pattern from RNA-Seq data (see above). We will quantify the confidence of these LoFs using features such as whether they are in highly duplicated regions and the number of paralogs. For functional features, we will incorporate protein structures. For evolutionary properties, we will quantify the conservation of LoF variants, as well as truncated sequences. For network features, we will quantify the distance between genes with LoF variants and known disease-causing genes.

Prioritizing non-coding transcripts from structural variant data. To prioritize the effects of SVs in ncRNAs, we will focus on overlaps with regulatory elements and other functional regions. To perform this analysis, we will define categories of RNA regions that display human population-level conservation, and combine these features to generate RNA element scores. Note that we may further classify SVs by manifestations of CVD. We will mine RNA interactions between proteins (e.g., CLIP-Seq) and miRNAs (e.g., TargetScan) to create a compendium of biochemical interactions with RNA[90–94]. We will further investigate RNA secondary structure, looking for structured regions that are highly sensitive to mutation. For these regions, we will assess deleteriousness of mutations by differences in predicted free energy or structure ensembles[78] relative to wild type. We have found annotations of all of the above types—biochemical interactions, regulatory motifs, and structured regions—that are enriched for rare variants in the human population and will use these sensitive RNA regions to score and

prioritize potential deleterious SVs in ncRNA. Large SVs will ultimately be scored based on the highest scoring subregion disrupted (or created) by the SV.

Prioritizing non-coding regulatory elements from structural variant data. Unlike protein-coding genes and ncRNAs, TF binding motifs are relatively small in size. Thus, we are going to analyze duplications that occur close to these motifs and analyze where these duplications lead to the breakage of existing or creation of new motifs. In the prioritization scheme, we will also penalize changes in distance between motifs and newly created motifs if they occur close to an existing TF motif. We will use TF binding nc elements by leveraging better enhancer definitions provided by the Epigenome Roadmap[95–97] and ENCODE and also include new datasets.

Further variant prioritization based on networks, tissue specificity, eQTLs and allelic activity. After performing annotation-based assessment of identified SVs, the following functional features will be used for prioritization.
1) *Network connectivity.* We will update and use well established gene networks based on regulatory, phosphorylation signaling, metabolic, and protein-protein interaction data. We will integrate novel datasets from ENCODE and Epigenome RoadMap, update regulatory networks, and integrate new datasets from conservation and protein-protein interaction. We will then examine the network topological properties of the genomic elements affected by identified SVs. Variants disrupting regulatory elements with high connectivity—network hubs and bottlenecks—will be upweighted based on their scaled centrality scores.
2) *Ubiquitous activity.* We will evaluate the impact of SVs in an epigenetic context to identify tissue-specific phenotypic effects that are strongly influenced by SVs. We will prioritize SVs impacting genes, ncRNAs, and TF binding sites active in multiple tissues.
3) *Allelic activity.* We will use our existing AlleleSeq pipeline to annotate the transcripts produced at SV regions[87]. We will use this tool to create personal diploid genomes for each TopMed individual, and then will adapt our pipeline to perform RNA-Seq quantification specifically at SV regions. We will prioritize SVs that lead to strongly allelic expression. We will also prioritize SVs that overlap our database of strongly allelic regions throughout the genome, based on AlleleDB, our resource of such regions identified through allele-specific RNA-Seq analysis from over 300 individuals generated by the GEUVADIS consortium[88].
4) *eQTL association*. We will link SVs to the genes that they affect by performing genome-wide searches for eQTLs. Relative to SNVs, large SVs may be more manageable candidates in the search for distal eQTLs. We will use a framework similar to published earlier[89] in the search for SV-induced eQTLs. SV-induced eQTLs will be identified by performing genome-wide searches for CVD patterns in which the presence or absence of the SVs (from Aim 1) strongly correlate with the expression levels of a battery of genes throughout the genome. Specifically, we will use Matrix eQTL for eQTL identification[98]. We will perform multiple testing correction and will filter the list of putative eQTLs in order to achieve a false discovery rate of less than 5%. The SV-gene expression correlations reported by Matrix eQTL will be used as the strength-of-association measures between expression levels and genotypes. Of particular interest will be those genes previously implicated in CVD-associated pathways and network modules. SV-induced eQTLs with strong expression correlations that are associated with central network elements and known CVD-associated genes will be upweighted.

Expected results. We expect to estimate the impact scores of the SVs produced in Aim 1 using SVIM, will yield a prioritized set of SVs in Aim 2 that we can forward to Aim 3 (genotype and association) for further classification of their association to disease or a specific phenotype. We plan to make the prioritization results broadly available; therefore, the impact score produced by SVIM will be incorporated into a standard Variant Call Format (VCF). SVIM will be cloud-ready and will be available to the TOPMED consortium through a Docker image and a Common Workflow Language (CWL) file. Docker and CWL are standards for distributing computational pipelines, which will make SVIM amenable for compute cluster, local machine, and cloud execution.

Pitfalls and alternative approaches. We anticipate the main challenges being (i) possibly an overwhelming number of SV discovered in Aim 1 and (ii) the lack of standard format and increasing number and updates of annotation datasets. In order to overcome (i), we plan to gradually process the results into specific types of SVs. SVIM will also be based on the data context to optimally prioritize from WGS datasets. The overall modularization offers a flexible framework for users to incorporate the ever-increasing amounts of genomic data to both rebuild the underlying data context and prioritize case-specific variants. Regarding (ii), we will carefully engineer SVIM to be computationally efficient and to able to support the large-scale computing proposed for this aim. To build the data context, we will standardize large-scale publicly available data resources, such as SVs from the 1000 GP[2], conservation data from Bejerano *et al.*[99] and Cooper *et al.*[100], functional genomics data from ENCODE[70] and Roadmap Epigenomics Mapping Consortium[101].

Specific Aim 3. Association of structural variants with burden in CVD cohorts.

Rationale. Many high-impact SVs are expected to be relatively rare. To discover these important SVs, we have already developed a new association pipeline suitable for finding them and establishing their phenotype associations. We anticipate that building a reference database of structural variants in healthy individuals (Aim 1) will be essential for this goal.

Preliminary Results. *Power analysis for sample selection and association.* An important aspect will be performing full SV analysis for the entire discovery cohort of 15,000 individuals. The size of this discovery cohort sets the lower bound for minor allele frequency in genome wide associations we will examine. There is no general theory of discovery power currently used in SV algorithms, so we extended an existing statistical model of coverage[102] to estimate the discovery sample size. Bernoulli probabilities for two standard SV discovery methods, split reads and discordant read pairs, can be derived using probability theory considering read length, average and variance of insert length, SV length, etc. and subsequent incorporation of a detection rule, e.g. "≥3 split or discordant reads". Detection in each sample is binomial in the number of observations and discovery within sample set is likewise binomial in the detection and Minor Allele Frequency (MAF) probabilities.

Anticipated parameters for the WGS data to be generated for this project are 30X coverage per genome, average insert size of 400bp-600bp (20% coefficient of variation), 150bp reads, event detection based on ≥3 split reads or ≥5 discordant read pairs, and observation in at least 3 samples to constitute "discovery". The model predicts that split-read detection will predominate for simple SVs, as well as for complex events in which one sequence is replaced by another. Because split-reads depend only upon local alignment, power is essentially independent of the size of events (unlike for discordant read pairs), meaning it is primarily a function of sample size and MAF. **Figure 6a** shows power at MAF ≥ 0.1% is essentially 100% for 10K samples. It drops rapidly for lower MAFs, whose events are unlikely to be discovered in this study. Mosaicism is a potentially confounding factor, for example in blood samples where an event is not present in all cells. **Figure 6b** shows that power is not significantly impacted even for the 10K samples until mosaicism is quite significant.

The second aspect of "power" is variant-disease association. The issues are well-known[103], enabling the following "baseline" estimates of association power. General consensus recommends "collapsing" variants for low MAF in order to aggregate effects for increasing power. Analysis of the widely-used Li & Leal method for 10 collapsed variants at 4:1 risk ratio (**Figure 6c**) shows that groupings of 1% MAF variants having high (~50%) penetrance will require 15Ksamples for 50% power when Bonferroni-corrected. Power drops rapidly for lower MAF, penetrance, risk ratio, and sample size. Based on the analysis presented (**Figure 6d**), it is likely we will discover more variants than those for which solid associations can be established.

Association pipeline implementation and experience in discovering significant associations. We have developed a prototype pipeline incorporating extensive sample and variant level quality control (e.g, coverage, variant frequency and distribution), population stratification, pedigree segregation, etc. for population/family-based association analysis. It supports popular aggregation tests, including burden tests such as the Combined Multivariate Collapsing (CMC)[103], Exclusive Frequency Test (EFT)[104], Total Frequency Test (TFT)[104], and Cohort Allele Sum Test (CAST)[105], and variant component tests such as the Sequence Kernel Association Test (SKAT)[106]. We have already used it to discover associations by tailoring it to hypothesized genetic architectures of individual diseases. For example, assuming tumor suppressors are enriched for rare deleterious truncations, we grouped events by gene and used TFT to associate 13 genes with germline susceptibility in a >4,000 case cancer cohort[107].

Research Plan. SVs are characterized by size, type, penetrance, and multiple alleles. A critical step for association analysis of SVs is meaningful classification/annotation. By building on infrastructure and tools mentioned above, we will extend **SV2Pheno** to infer SV-phenotype associations (**Fig. 7**). It will use the impact scores for each SV (**Aim 2**) for integrated analysis of SNVs, indels, and SVs.

Extend SV2Pheno pipeline including improved burden tests considering impact score and annotation classification of various complex structure variants for CVD cohorts. We envision substantial extension of this pipeline in two major ways to address the ambitious goals of this proposal: 1) We plan to hybridize the pipeline with more recent methods that better account for non-contributing variants[108]. Likewise, annotation and functional prediction can help identify irrelevant variants, which can subsequently be removed from analysis. The pipeline will also process the information from the ENCODE & Epigenetics Roadmap analysis mentioned in **Aim 2**. 2) Variants are known to be associated with various diseases[109–111], but almost certainly contribute non-uniformly; assigning appropriate weights will be necessary to wring-out maximum power. Aggregation tests can be expressed in general by the linear regression equation $Y = \alpha + \beta \cdot \Sigma w_i g_i + \varepsilon$, where (left-to-right) is observed

trait, intercept, collective effect coefficient, weight of variant *i*, tally of variant *i* (0, 1, or 2), and normally distributed error residual. Assignment of weights will be based on a novel combination of four considerations: the Madsen-Browning equation[112] to account for allele frequency, consideration of "direction" (negative association) using e.g. aspects of the Pan-Shen approach[113], incorporation of our impact score (**Aim 2**) to account for biological strength, and RNA-seq data. The last aspect will weight expression impact, but must be implemented carefully because of variations in sample quality. Here, we will apply the method of Liu *et al.*[114], which essentially adds an extra adjustment to modulate contribution of higher-variability samples. In principle, this more sophisticated approach should capture signals that have been too subtle for earlier tests[115].

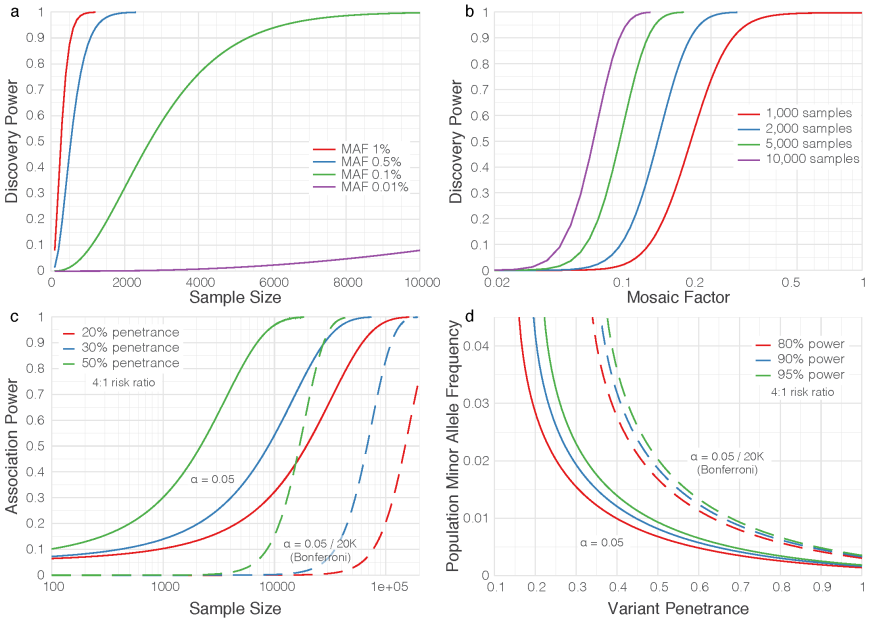

Figure 6. Power analysis for sample selection and association. a) Power vs sample size for selected MAFs from 0.01% to 1%. Events are assumed heterozygous and completely represented in the sample (no mosaicism). Curves are universal in that simple insertions and deletions, as well as complex indels, collapse and power is independent of indel size, since the "split reads" discovery mode dominates. b) Power vs "mosaic factor" (unity meaning event present in all cells; 0.5 meaning event present in half the cells, etc.) for selected samples sizes from 1K to 10K. All data plotted at 1% MAF. Split-read discovery again dominates and curves are universal. c) Association power for 10 collapsed variants (even numbers of cases and controls), each of 1% MAF and penetrance from 1% to 50%, at both single gene (α = 5%) and Bonferroni-corrected for 20K genes, as well as a 4:1 risk ratio for the Li and Leal (2008) collapsing strategy. d) Curves of constant power for 10K cases/10K controls, with other parameters the same as in c).

Since we anticipate that a high fraction of SVs will reside in non-coding regions, we will aggregate variants using a hierarchical approach based on three levels:

Level 1. *Prototypical Event level association analysis.* As the precise genomic region for a given SV may vary across samples, we will represent each set of similar SV events as a single prototypical SV event. The criterion constituting such events is given by the "80% reciprocal overlap" rule[61]. For large insertions and inter-chromosomal translations, we will require the breakpoints to be within 1kb of one another. We will then assess the significance of the associations using impact scores generated in **Aim 2**.

Level 2. *Functional Unit (Gene CDS/promoter/enhancer) level association analysis.* We annotate the prototypical SV events from Level 1 to identify any specific transcriptional regions (e.g., exons/CDS and cis-regulatory elements such as insulators, enhancers, and promoters) and gene(s). SVs in a given gene will be grouped as a single, effective functional unit based on annotation from **Aim 2** (**Figure 7**). We will then perform an association analysis on these functional units. In cases where multiple SV events may be affiliated with a given functional unit, we need a weighting scheme to combine the impact scores of the contributing SVs. This approach may reveal novel connections between non-coding functional regions and phenotypes.

Level 3. *Combined Functional Unit level analysis.* We will annotate the functional units in the previous step to identify known affiliated higher-order units (e.g., protein complexes and gene pathways) by recruiting various resources, including databases relating to gene-phenotype relationships (e.g., OMIM), gene pathways (e.g., KEGG, Reactome), gene ontology (e.g., GO database). The SVs affecting a given higher-order unit will be grouped as a single super-unit. We will again perform association analysis, considering the SV impact scores (**Aim 2**). This approach has the potential to discover novel combinations of SV-containing functional units.

We will apply this tiered approach and association analysis (**Figure 7)** to analyze all samples passing our extensive coverage and variant calling QC from various cohorts to identify promising candidate SVs associated with the cardiovascular disease phenotype.

Integrate various types of variants for association analysis. The most powerful analysis will come by combining information from SNVs, indels, and SVs for association analysis. Traditionally, weights in burden tests account for variants with different MAFs, but favoring those having lower MAFs[106,112]. Bioinformatic information, such as PolyPhen scores for SNVs, and SV impact scores from **Aim 2** will inform these weights. To the best of our knowledge, no previous approaches have aggregated variants of different types. Here, we propose two methods

for such integration: 1) We hypothesize that SVs would have stronger functional impacts than missense SNVs, on average, and we will extend our weighing scheme based on the size and genetic architecture of various variant types using the framework of previous weighting schemes. SNV, indel, and SV will be jointly calculated in a single burden analysis; 2) We hypothesize that alterations from functional regions, regardless of size, contribute to phenotype. Therefore, alternatively, we plan use SNV/indel and SV for independent burden analyses and combine the P-values from these independent tests.

Association between SNVs/indels and SVs. Under the null hypothesis that variation occurs randomly, it should be possible to correlate the numbers of SNVs/indels versus the number of SVs, the slope being indicative of differences in rates of occurrence, and also to check such correlation against established rates. We will perform association analysis for individual outlier cases in which SV census is significantly lower or higher than expected. It is possible that such outliers might harbor common germline alterations leading to genomic instability by affecting DNA repair pathways.

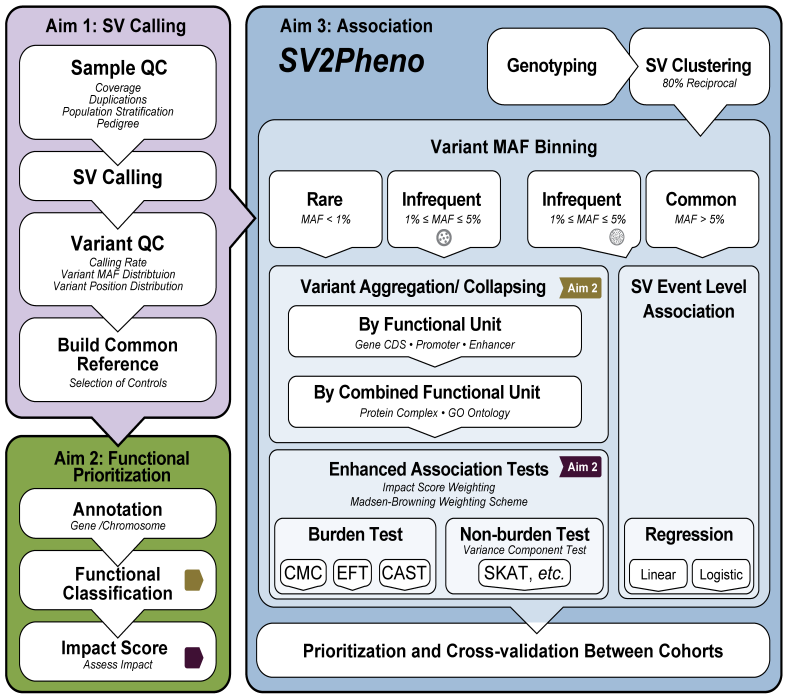

Figure 7. SV2Pheno Association Analysis Pipeline. The overall work flow includes QC, population stratification from Aim 1, functional classification and impact score generation from Aim 2 and single event test and burden analysis from Aim 3.

***Expected results*.** This aim will culminate in the **SV2Pheno** association pipeline and its tools for systematically discovering SVs associated with the cardiovascular disease phenotype. We expect to have increased statistical power to discover rare, novel SVs associated with phenotypes previously missed due to smaller sample size. We further anticipate revealing genetic changes associated with increased frequency of SVs genome-wide. The initial version of **SV2Pheno** will be distributed for broader community use, including on the cloud.

Pitfalls and alternative approaches. Our preliminary analysis indicates that we are well powered to detect SVs with MAFs around 0.5% to 1% using >10,000 cases. Although it is very likely that we will discover more SVs than we can establish associations for (discussed above), there are still some issues of selection of appropriate samples from the selected cohorts. There are several strategies for selecting datasets for discovery: 1) from one homogenous cohort; 2) from one CCDG center across multiple cohorts; 3) from multiple cohorts generated by multiple TOPMED centers. Regardless of choice, we will maintain high standards regarding coverage, read length, insert size, mapping rate, % mismatch etc. (rejecting samples when they don't meet our standards) to ensure accurate, representative detection of SVs across populations. To reduce the number of hypotheses to be tested, we can alternatively focus on SVs from regions indicated to have association with cardiovascular disease from previous studies using SNPs and Indels.

REFRENCES.

1. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, 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 ES, Altshuler DM, Gabriel SB, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Flicek P, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Bentley DR, Grocock R, Humphray S, James T, Kingsbury Z, Lehrach H, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo M-L, Mardis ER, Wilson RK, Fulton L, Fulton R, Sherry ST, 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 GA, Durbin RM, Balasubramaniam S, Burton J, Danecek P, Keane TM, Kolb-Kokocinski A, McCarthy S, Stalker J, Quail M, Schmidt JP, Davies CJ, Gollub J, Webster T, Wong B, Zhan Y, Auton A, Campbell CL, Kong Y, Marcketta A, Gibbs RA, Yu F, Antunes L, Bainbridge M, Muzny D, Sabo A, Huang Z, Wang J, Coin LJM, 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 GT, Garrison EP, Kural D, Lee W-P, Fung Leong W, Stromberg M, Ward AN, Wu J, Zhang M, Daly MJ, DePristo MA, Handsaker RE, Altshuler DM, Banks E, Bhatia G, del Angel G, Gabriel SB, Genovese G, Gupta N, Li H, Kashin S, Lander ES, McCarroll SA, Nemesh JC, Poplin RE, Yoon SC, Lihm J, Makarov V, Clark AG, Gottipati S, Keinan A, Rodriguez-Flores JL, Korbel JO, Rausch T, Fritz MH, Stütz AM, Flicek P, Beal K, Clarke L, Datta A, Herrero J, McLaren WM, Ritchie GRS, Smith RE, Zerbino D, Zheng-Bradley X, Sabeti PC, Shlyakhter I, Schaffner SF, Vitti J, Cooper DN, Ball E V., Stenson PD, Bentley DR, Barnes B, Bauer M, Keira Cheetham R, Cox A, Eberle M, Humphray S, Kahn S, Murray L, Peden J, Shaw R, Kenny EE, Batzer MA, Konkel MK, Walker JA, MacArthur DG, Lek M, Sudbrak R, Amstislavskiy VS, Herwig R, Mardis ER, Ding L, Koboldt DC, Larson D, Ye K, Gravel S, Swaroop A, Chew E, Lappalainen T, Erlich Y, Gymrek M, Frederick Willems T, Simpson JT, Shriver MD, Rosenfeld JA, Bustamante CD, Montgomery SB, De La Vega FM, Byrnes JK, Carroll AW, DeGorter MK, Lacroute P, Maples BK, Martin AR, Moreno-Estrada A, Shringarpure SS, Zakharia F, Halperin E, Baran Y, Lee C, Cerveira E, Hwang J, Malhotra A, Plewczynski D, Radew K, Romanovitch M, Zhang C, Hyland FCL, Craig DW, Christoforides A, Homer N, Izatt T, Kurdoglu AA, Sinari SA, Squire K, Sherry ST, Xiao C, Sebat J, Antaki D, Gujral M, Noor A, Ye K, Burchard EG, Hernandez RD, Gignoux CR, Haussler D, Katzman SJ, James Kent W, Howie B, Ruiz-Linares A, Dermitzakis ET, Devine SE, Abecasis GR, Min Kang H, Kidd JM, 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 GA, Marchini JL, Myers S, Churchhouse C, Delaneau O, Gupta-Hinch A, Kretzschmar W, Iqbal Z, Mathieson I, Menelaou A, Rimmer A, Xifara DK, Oleksyk TK, Fu Y, Liu X, Xiong M, Jorde L, Witherspoon D, Xing J, Eichler EE, Browning BL, Browning SR, Hormozdiari F, Sudmant PH, Khurana E, Durbin RM, Hurles ME, Tyler-Smith C, Albers CA, Ayub Q, Balasubramaniam S, Chen Y, Colonna V, Danecek P, Jostins L, Keane TM, McCarthy S, Walter K, Xue Y, Gerstein MB, Abyzov A, Balasubramanian S, Chen J, Clarke D, Fu Y, Harmanci AO, 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 GT, Garrison EP, Kural D, Lee W-P, Ward AN, Wu J, Zhang M, McCarroll SA, Handsaker RE, Altshuler DM, Banks E, del Angel G, Genovese G, Hartl C, Li H, Kashin S, Nemesh JC, Shakir K, Yoon SC, Lihm J, Makarov V, Degenhardt J, Korbel JO, Fritz MH, Meiers S, Raeder B, Rausch T, Stütz AM, Flicek P, Paolo Casale F, Clarke L, Smith RE, Stegle O, Zheng-Bradley X, Bentley DR, Barnes B, Keira Cheetham R, Eberle M, Humphray S, Kahn S, Murray L, Shaw R, Lameijer E-W, Batzer MA, Konkel MK, Walker JA, 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 DW, Homer N, Church D, Xiao C, Sebat J, Antaki D, Bafna V, Michaelson J, Ye K, Devine SE, Gardner EJ, Abecasis GR, Kidd JM, Mills RE, Dayama G, Emery S, Jun G, Shi X, Quitadamo A, Lunter G, McVean GA, Chen K, Fan X, Chong Z, Chen T, Witherspoon D, Xing J, Eichler EE, Chaisson MJ, Hormozdiari F, Huddleston J, Malig M, Nelson BJ, Sudmant PH, Parrish NF, Khurana E, Hurles ME, Blackburne B, Lindsay SJ, Ning Z, Walter K, Zhang Y, Gerstein MB, Abyzov A, Chen J, Clarke D, Lam H, Jasmine Mu X, Sisu C, Zhang J, Zhang Y, Gibbs RA, Yu F, Bainbridge M, Challis D, Evani US, Kovar C, Lu J, Muzny D, Nagaswamy U,

Reid JG, Sabo A, Yu J, Guo X, Li W, Li Y, Wu R, Marth GT, Garrison EP, Fung Leong W, Ward AN, del Angel G, DePristo MA, Gabriel SB, Gupta N, Hartl C, Poplin RE, Clark AG, Rodriguez-Flores JL, Flicek P, Clarke L, Smith RE, Zheng-Bradley X, MacArthur DG, Mardis ER, Fulton R, Koboldt DC, Gravel S, Bustamante CD, Craig DW, Christoforides A, Homer N, Izatt T, Sherry ST, Xiao C, Dermitzakis ET, Abecasis GR, Min Kang H, McVean GA, Gerstein MB, 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 SB, DeGorter MK, Khurana E, Tyler-Smith C, Chen Y, Colonna V, Xue Y, Gerstein MB, Balasubramanian S, Fu Y, Kim D, Auton A, Marcketta A, Desalle R, Narechania A, Wilson Sayres MA, Garrison EP, Handsaker RE, Kashin S, McCarroll SA, Rodriguez-Flores JL, Flicek P, Clarke L, Zheng-Bradley X, Erlich Y, Gymrek M, Frederick Willems T, Bustamante CD, Mendez FL, David Poznik G, Underhill PA, Lee C, Cerveira E, Malhotra A, Romanovitch M, Zhang C, Abecasis GR, Coin L, Shao H, Mittelman D, Tyler-Smith C, Ayub Q, Banerjee R, Cerezo M, Chen Y, Fitzgerald TW, Louzada S, Massaia A, McCarthy S, Ritchie GR, Xue Y, Yang F, Gibbs RA, Kovar C, Kalra D, Hale W, Muzny D, Reid JG, Wang J, Dan X, Guo X, Li G, Li Y, Ye C, Zheng X, Altshuler DM, Flicek P, Clarke L, Zheng-Bradley X, Bentley DR, Cox A, Humphray S, Kahn S, Sudbrak R, Albrecht MW, Lienhard M, Larson D, Craig DW, Izatt T, Kurdoglu AA, Sherry ST, Xiao C, Haussler D, Abecasis GR, McVean GA, Durbin RM, Balasubramaniam S, Keane TM, McCarthy S, Stalker J, Chakravarti A, Knoppers BM, Abecasis GR, Barnes KC, Beiswanger C, Burchard EG, Bustamante CD, Cai H, Cao H, Durbin RM, Gerry NP, Gharani N, Gibbs RA, Gignoux CR, Gravel S, Henn B, Jones D, Jorde L, Kaye JS, Keinan A, Kent A, Kerasidou A, Li Y, Mathias R, McVean GA, Moreno-Estrada A, Ossorio PN, Parker M, Resch AM, Rotimi CN, Royal CD, Sandoval K, Su Y, Sudbrak R, Tian Z, Tishkoff S, Toji LH, 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 JC, Oleksyk TK, Barnes KC, Mathias RA, Hennis A, Watson H, McKenzie C, Qadri F, LaRocque R, Sabeti PC, Zhu J, Deng X, Sabeti PC, 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, Dunstan SJ, Thuy Hang N, Fonnie R, Garry R, Kanneh L, Moses L, Sabeti PC, Schieffelin J, Grant DS, Gallo C, Poletti G, Saleheen D, Rasheed A, Brooks LD, Felsenfeld AL, McEwen JE, Vaydylevich Y, Green ED, Duncanson A, Dunn M, Schloss JA, Wang J, Yang H, Auton A, Brooks LD, Durbin RM, Garrison EP, Min Kang H, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature*. 2015 Sep 30;**526**(7571):68–74. doi:10.1038/nature15393 PMID: 26432245

2. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, St?tz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, 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 JM, Kong Y, Lameijer E-W, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO, Korbel JO. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015 Sep 30;**526**(7571):75–81. doi:10.1038/nature15394 PMID: 26432246
3. Quinlan AR, Hall IM. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet*. 2012 Jan;**28**(1):43–53. doi:10.1016/j.tig.2011.10.002 PMID: 22094265
4. Sharp AJ, Cheng Z, Eichler EE. Structural Variation of the Human Genome. *Annu Rev Genomics Hum Genet*. 2006 Sep;**7**(1):407–442. doi:10.1146/annurev.genom.7.080505.115618 PMID: 16780417
5. Stankiewicz P, Lupski JR. Structural Variation in the Human Genome and its Role in Disease. *Annu Rev Med*. 2010 Feb;**61**(1):437–455. doi:10.1146/annurev-med-100708-204735 PMID: 20059347
6. Kakinuma H, Sato H. Copy-number variations associated with autism spectrum disorder. *Pharmacogenomics*. 2008 Aug;**9**(8):1143–1154. doi:10.2217/14622416.9.8.1143 PMID: 18681787
7. AlSagob M, Colak D, Kaya N. Genetics of autism spectrum disorder: an update on copy number variations leading to autism in the next generation sequencing era. *Discov Med*. 2015 May;**19**(106):367–79. PMID: 26105700
8. Chung BH-Y, Tao VQ, Tso WW-Y. Copy number variation and autism: new insights and clinical implications. *J Formos Med Assoc*. 2014 Jul;**113**(7):400–8. doi:10.1016/j.jfma.2013.01.005 PMID: 24961180
9. Kirov G. The role of copy number variation in schizophrenia. *Expert Rev Neurother*. 2010 Jan 9;**10**(1):25–

32. doi:10.1586/ern.09.133 PMID: 20021318
10. Tam GWC, Redon R, Carter NP, Grant SGN. The Role of DNA Copy Number Variation in Schizophrenia. *Biol Psychiatry*. 2009 Dec 1;**66**(11):1005–1012. doi:10.1016/j.biopsych.2009.07.027 PMID: 19748074
11. Bassett AS, Scherer SW, Brzustowicz LM. Copy number variations in schizophrenia: critical review and new perspectives on concepts of genetics and disease. *Am J Psychiatry*. 2010 Aug;**167**(8):899–914. doi:10.1176/appi.ajp.2009.09071016 PMID: 20439386
12. Gulsuner S, McClellan JM. Copy Number Variation in Schizophrenia. *Neuropsychopharmacology*. 2015 Jan;**40**(1):252–254. doi:10.1038/npp.2014.216 PMID: 25482180
13. Hochstenbach R, Buizer-Voskamp JE, Vorstman JAS, Ophoff RA. Genome arrays for the detection of copy number variations in idiopathic mental retardation, idiopathic generalized epilepsy and neuropsychiatric disorders: lessons for diagnostic workflow and research. *Cytogenet Genome Res*. 2011;**135**(3–4):174–202. doi:10.1159/000332928 PMID: 22056632
14. Vissers LELM, de Vries BBA, Veltman JA. Genomic microarrays in mental retardation: from copy number variation to gene, from research to diagnosis. *J Med Genet*. 2010 May 1;**47**(5):289–297. doi:10.1136/jmg.2009.072942 PMID: 19951919
15. Bauters M, Weuts A, Vandewalle J, Nevelsteen J, Marynen P, Van Esch H, Froyen G. Detection and validation of copy number variation in X-linked mental retardation. *Cytogenet Genome Res*. 2008 Mar 11;**123**(1–4):44–53. doi:10.1159/000184691 PMID: 19287138
16. Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, Himes BE, Levin AM, Mathias RA, Hancock DB, Baurley JW, Eng C, Stern DA, Celed?n JC, Rafaels N, Capurso D, Conti D V, Roth LA, Soto-Quiros M, Togias A, Li X, Myers RA, Romieu I, Berg DJ Van Den, Hu D, Hansel NN, Hernandez RD, Israel E, Salam MT, Galanter J, Avila PC, Avila L, Rodriquez-Santana JR, Chapela R, Rodriguez-Cintron W, Diette GB, Adkinson NF, Abel RA, Ross KD, Shi M, Faruque MU, Dunston GM, Watson HR, Mantese VJ, Ezurum SC, Liang L, Ruczinski I, Ford JG, Huntsman S, Chung KF, Vora H, Li X, Calhoun WJ, Castro M, Sienra-Monge JJ, del Rio-Navarro B, Deichmann KA, Heinzmann A, Wenzel SE, Busse WW, Gern JE, Lemanske RF, Beaty TH, Bleecker ER, Raby BA, Meyers DA, London SJ, Gilliland FD, Burchard EG, Martinez FD, Weiss ST, Williams LK, Barnes KC, Ober C, Nicolae DL. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet*. 2011 Jul 31;**43**(9):887–892. doi:10.1038/ng.888
17. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, von Mutius E, Farrall M, Lathrop M, Cookson WOCM, GABRIEL Consortium. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med*. 2010 Sep 23;**363**(13):1211–21. doi:10.1056/NEJMoa0906312 PMID: 20860503
18. Meyers DA. Genetics of asthma and allergy: What have we learned? *J Allergy Clin Immunol*. 2010 Sep;**126**(3):439–446. doi:10.1016/j.jaci.2010.07.012 PMID: 20816180
19. Mersha TB. Mapping asthma-associated variants in admixed populations. *Front Genet*. 2015 Sep 29;**6**:292. doi:10.3389/fgene.2015.00292 PMID: 26483834
20. Rogers AJ, Chu J-H, Darvishi K, Ionita-Laza I, Lehmann H, Mills R, Lee C, Raby BA. Copy number variation prevalence in known asthma genes and their impact on asthma susceptibility. *Clin Exp Allergy*. 2013 Apr;**43**(4):455–62. doi:10.1111/cea.12060 PMID: 23517041
21. Kepp K, Org E, Sõber S, Kelgo P, Viigimaa M, Veldre G, Tõnisson N, Juhanson P, Putku M, Kindmark A, Kozich V, Laan M. Hypervariable intronic region in NCX1 is enriched in short insertion-deletion polymorphisms and showed association with cardiovascular traits. *BMC Med Genet*. 2010 Jan 28;**11**:15. doi:10.1186/1471-2350-11-15 PMID: 20109173
22. Costain G, Lionel AC, Ogura L, Marshall CR, Scherer SW, Silversides CK, Bassett AS. Genome-wide rare copy number variations contribute to genetic risk for transposition of the great arteries. *Int J Cardiol*. 2016 Feb 1;**204**:115–21. doi:10.1016/j.ijcard.2015.11.127 PMID: 26655555
23. Johnson AD, Hwang S-J, Voorman A, Morrison A, Peloso GM, Hsu Y-H, Thanassoulis G, Newton-Cheh C, Rogers IS, Hoffmann U, Freedman JE, Fox CS, Psaty BM, Boerwinkle E, Cupples LA, O'Donnell CJ. Resequencing and clinical associations of the 9p21.3 region: a comprehensive investigation in the Framingham heart study. *Circulation*. 2013 Feb 19;**127**(7):799–810. doi:10.1161/CIRCULATIONAHA.112.111559 PMID: 23315372
24. Shia W-C, Ku T-H, Tsao Y-M, Hsia C-H, Chang Y-M, Huang C-H, Chung Y-C, Hsu S-L, Liang K-W, Hsu F-R. Genetic copy number variants in myocardial infarction patients with hyperlipidemia. *BMC Genomics*. 2011 Nov 30;**12 Suppl 3**(Suppl 3):S23. doi:10.1186/1471-2164-12-S3-S23 PMID: 22369086
25. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age?sex specific all-cause and cause-specific mortality for 240 causes of death, 1990?2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2015 Jan 10;**385**(9963):117–171. doi:10.1016/S0140-

6736(14)61682-2 PMID: 25530442

26. Wang H, Naghavi M, Allen C, Barber RM, Bhutta ZA, Carter A, Casey DC, Charlson FJ, Chen AZ, Coates MM, Coggeshall M, Dandona L, Dicker DJ, Erskine HE, Ferrari AJ, Fitzmaurice C, Foreman K, Forouzanfar MH, Fraser MS, Fullman N, Gething PW, Goldberg EM, Graetz N, Haagsma JA, Hay SI, Huynh C, Johnson CO, Kassebaum NJ, Kinfu Y, Kulikoff XR, Kutz M, Kyu HH, Larson HJ, Leung J, Liang X, Lim SS, Lind M, Lozano R, Marquez N, Mensah GA, Mikesell J, Mokdad AH, Mooney MD, Nguyen G, Nsoesie E, Pigott DM, Pinho C, Roth GA, Salomon JA, Sandar L, Silpakit N, Sligar A, Sorensen RJD, Stanaway J, Steiner C, Teeple S, Thomas BA, Troeger C, VanderZanden A, Vollset SE, Wanga V, Whiteford HA, Wolock T, Zoeckler L, Abate KH, Abbafati C, Abbas KM, Abd-Allah F, Abera SF, Abreu DMX, Abu-Raddad LJ, Abyu GY, Achoki T, Adelekan AL, Ademi Z, Adou AK, Adsuar JC, Afanvi KA, Afshin A, Agardh EE, Agarwal A, Agrawal A, Kiadaliri AA, Ajala ON, Akanda AS, Akinyemi RO, Akinyemiju TF, Akseer N, Lami FH Al, Alabed S, Al-Aly Z, Alam K, Alam NKM, Alasfoor D, Aldhahri SF, Aldridge RW, Alegretti MA, Aleman A V, Alemu ZA, Alexander LT, Alhabib S, Ali R, Alkerwi A, Alla F, Allebeck P, Al-Raddadi R, Alsharif U, Altirkawi KA, Martin EA, Alvis-Guzman N, Amare AT, Amegah AK, Ameh EA, Amini H, Ammar W, Amrock SM, Andersen HH, Anderson BO, Anderson GM, Antonio CAT, Aregay AF, Ärnlöv J, Arsenijevic VSA, Artaman A, Asayesh H, Asghar RJ, Atique S, Avokpaho EFGA, Awasthi A, Azzopardi P, Bacha U, Badawi A, Bahit MC, Balakrishnan K, Banerjee A, Barac A, Barker-Collo SL, Bärnighausen T, Barregard L, Barrero LH, Basu A, Basu S, Bayou YT, Bazargan-Hejazi S, Beardsley J, Bedi N, Beghi E, Belay HA, Bell B, Bell ML, Bello AK, Bennett DA, Bensenor IM, Berhane A, Bernabé E, Betsu BD, Beyene AS, Bhala N, Bhalla A, Biadgilign S, Bikbov B, Abdulhak AA Bin, Biroscak BJ, Biryukov S, Bjertness E, Blore JD, Blosser CD, Bohensky MA, Borschmann R, Bose D, Bourne RRA, Brainin M, Brayne CEG, Brazinova A, Breitborde NJK, Brenner H, Brewer JD, Brown A, Brown J, Brugha TS, Buckle GC, Butt ZA, Calabria B, Campos-Nonato IR, Campuzano JC, Carapetis JR, Cárdenas R, Carpenter DO, Carrero JJ, Castañeda-Orjuela CA, Rivas JC, Catalá-López F, Cavalleri F, Cercy K, Cerda J, Chen W, Chew A, Chiang PP-C, Chibalabala M, Chibueze CE, Chimed-Ochir O, Chisumpa VH, Choi J-YJ, Chowdhury R, Christensen H, Christopher DJ, Ciobanu LG, Cirillo M, Cohen AJ, Colistro V, Colomar M, Colquhoun SM, Cooper C, Cooper LT, Cortinovis M, Cowie BC, Crump JA, Damsere-Derry J, Danawi H, Dandona R, Daoud F, Darby SC, Dargan PI, das Neves J, Davey G, Davis AC, Davitoiu D V, de Castro EF, de Jager P, Leo D De, Degenhardt L, Dellavalle RP, Deribe K, Deribew A, Dharmaratne SD, Dhillon PK, Diaz-Torné C, Ding EL, dos Santos KPB, Dossou E, Driscoll TR, Duan L, Dubey M, Duncan BB, Ellenbogen RG, Ellingsen CL, Elyazar I, Endries AY, Ermakov SP, Eshrati B, Esteghamati A, Estep K, Faghmous IDA, Fahimi S, Faraon EJA, Farid TA, Farinha CS e S, Faro A, Farvid MS, Farzadfar F, Feigin VL, Fereshtehnejad S-M, Fernandes JG, Fernandes JC, Fischer F, Fitchett JRA, Flaxman A, Foigt N, Fowkes FGR, Franca EB, Franklin RC, Friedman J, Frostad J, Fürst T, Futran ND, Gall SL, Gambashidze K, Gamkrelidze A, Ganguly P, Gankpé FG, Gebre T, Gebrehiwot TT, Gebremedhin AT, Gebru AA, Geleijnse JM, Gessner BD, Ghoshal AG, Gibney KB, Gillum RF, Gilmour S, Giref AZ, Giroud M, Gishu MD, Giussani G, Glaser E, Godwin WW, Gomez-Dantes H, Gona P, Goodridge A, Gopalani SV, Gosselin RA, Gotay CC, Goto A, Gouda HN, Greaves F, Gugnani HC, Gupta R, Gupta R, Gupta V, Gutiérrez RA, Hafezi-Nejad N, Haile D, Hailu AD, Hailu GB, Halasa YA, Hamadeh RR, Hamidi S, Hancock J, Handal AJ, Hankey GJ, Hao Y, Harb HL, Harikrishnan S, Haro JM, Havmoeller R, Heckbert SR, Heredia-Pi IB, Heydarpour P, Hilderink HBM, Hoek HW, Hogg RS, Horino M, Horita N, Hosgood HD, Hotez PJ, Hoy DG, Hsairi M, Htet AS, Htike MMT, Hu G, Huang C, Huang H, Huiart L, Husseini A, Huybrechts I, Huynh G, Iburg KM, Innos K, Inoue M, Iyer VJ, Jacobs TA, Jacobsen KH, Jahanmehr N, Jakovljevic MB, James P, Javanbakht M, Jayaraman SP, Jayatilleke AU, Jeemon P, Jensen PN, Jha V, Jiang G, Jiang Y, Jibat T, Jimenez-Corona A, Jonas JB, Joshi TK, Kabir Z, Kamal R, Kan H, Kant S, Karch A, Karema CK, Karimkhani C, Karletsos D, Karthikeyan G, Kasaeian A, Katibeh M, Kaul A, Kawakami N, Kayibanda JF, Keiyoro PN, Kemmer L, Kemp AH, Kengne AP, Keren A, Kereselidze M, Kesavachandran CN, Khader YS, Khalil IA, Khan AR, Khan EA, Khang Y-H, Khera S, Khoja TAM, Kieling C, Kim D, Kim YJ, Kissela BM, Kissoon N, Knibbs LD, Knudsen AK, Kokubo Y, Kolte D, Kopec JA, Kosen S, Koul PA, Koyanagi A, Krog NH, Defo BK, Bicer BK, Kudom AA, Kuipers EJ, Kulkarni VS, Kumar GA, Kwan GF, Lal A, Lal DK, Lalloo R, Lallukka T, Lam H, Lam JO, Langan SM, Lansingh VC, Larsson A, Laryea DO, Latif AA, Lawrynowicz AEB, Leigh J, Levi M, Li Y, Lindsay MP, Lipshultz SE, Liu PY, Liu S, Liu Y, Lo L-T, Logroscino G, Lotufo PA, Lucas RM, Lunevicius R, Lyons RA, Ma S, Machado VMP, Mackay MT, MacLachlan JH, Razek HMA El, Magdy M, Razek A El, Majdan M, Majeed A, Malekzadeh R, Manamo WAA, Mandisarisa J, Mangalam S, Mapoma CC, Marcenes W, Margolis DJ, Martin GR, Martinez-Raga J, Marzan MB, Masiye F, Mason-Jones AJ, Massano J, Matzopoulos R, Mayosi BM, McGarvey ST, McGrath JJ, McKee M, McMahon BJ, Meaney PA, Mehari A, Mehndiratta MM, Mejia-

Rodriguez F, Mekonnen AB, Melaku YA, Memiah P, Memish ZA, Mendoza W, Meretoja A, Meretoja TJ, Mhimbira FA, Micha R, Millear A, Miller TR, Mirarefin M, Misganaw A, Mock CN, Mohammad KA, Mohammadi A, Mohammed S, Mohan V, Mola GLD, Monasta L, Hernandez JCM, Montero P, Montico M, Montine TJ, Moradi-Lakeh M, Morawska L, Morgan K, Mori R, Mozaffarian D, Mueller UO, Murthy GVS, Murthy S, Musa KI, Nachega JB, Nagel G, Naidoo KS, Naik N, Naldi L, Nangia V, Nash D, Nejjari C, Neupane S, Newton CR, Newton JN, Ng M, Ngalesoni FN, de Dieu Ngirabega J, Nguyen Q Le, Nisar MI, Pete PMN, Nomura M, Norheim OF, Norman PE, Norrving B, Nyakarahuka L, Ogbo FA, Ohkubo T, Ojelabi FA, Olivares PR, Olusanya BO, Olusanya JO, Opio JN, Oren E, Ortiz A, Osman M, Ota E, Ozdemir R, PA M, Pain A, Pandian JD, Pant PR, Papachristou C, Park E-K, Park J-H, Parry CD, Parsaeian M, Caicedo AJP, Patten SB, Patton GC, Paul VK, Pearce N, Pedro JM, Stokic LP, Pereira DM, Perico N, Pesudovs K, Petzold M, Phillips MR, Piel FB, Pillay JD, Plass D, Platts-Mills JA, Polinder S, Pope CA, Popova S, Poulton RG, Pourmalek F, Prabhakaran D, Qorbani M, Quame-Amaglo J, Quistberg DA, Rafay A, Rahimi K, Rahimi-Movaghar V, Rahman M, Rahman MHU, Rahman SU, Rai RK, Rajavi Z, Rajsic S, Raju M, Rakovac I, Rana SM, Ranabhat CL, Rangaswamy T, Rao P, Rao SR, Refaat AH, Rehm J, Reitsma MB, Remuzzi G, Resnikoff S, Ribeiro AL, Ricci S, Blancas MJR, Roberts B, Roca A, Rojas-Rueda D, Ronfani L, Roshandel G, Rothenbacher D, Roy A, Roy NK, Ruhago GM, Sagar R, Saha S, Sahathevan R, Saleh MM, Sanabria JR, Sanchez-Niño MD, Sanchez-Riera L, Santos IS, Sarmiento-Suarez R, Sartorius B, Satpathy M, Savic M, Sawhney M, Schaub MP, Schmidt MI, Schneider IJC, Schöttker B, Schutte AE, Schwebel DC, Seedat S, Sepanlou SG, Servan-Mori EE, Shackelford KA, Shaddick G, Shaheen A, Shahraz S, Shaikh MA, Shakh-Nazarova M, Sharma R, She J, Sheikhbahaei S, Shen J, Shen Z, Shepard DS, Sheth KN, Shetty BP, Shi P, Shibuya K, Shin M-J, Shiri R, Shiue I, Shrime MG, Sigfusdottir ID, Silberberg DH, Silva DAS, Silveira DGA, Silverberg JI, Simard EP, Singh A, Singh GM, Singh JA, Singh OP, Singh PK, Singh V, Soneji S, Søreide K, Soriano JB, Sposato LA, Sreeramareddy CT, Stathopoulou V, Stein DJ, Stein MB, Stranges S, Stroumpoulis K, Sunguya BF, Sur P, Swaminathan S, Sykes BL, Szoeke CEI, Tabarés-Seisdedos R, Tabb KM, Takahashi K, Takala JS, Talongwa RT, Tandon N, Tavakkoli M, Taye B, Taylor HR, Ao BJ Te, Tedla BA, Tefera WM, Have M Ten, Terkawi AS, Tesfay FH, Tessema GA, Thomson AJ, Thorne-Lyman AL, Thrift AG, Thurston GD, Tillmann T, Tirschwell DL, Tonelli M, Topor-Madry R, Topouzis F, Towbin JA, Traebert J, Tran BX, Truelsen T, Trujillo U, Tura AK, Tuzcu EM, Uchendu US, Ukwaja KN, Undurraga EA, Uthman OA, Dingenen R Van, van Donkelaar A, Vasankari T, Vasconcelos AMN, Venketasubramanian N, Vidavalur R, Vijayakumar L, Villalpando S, Violante FS, Vlassov VV, Wagner JA, Wagner GR, Wallin MT, Wang L, Watkins DA, Weichenthal S, Weiderpass E, Weintraub RG, Werdecker A, Westerman R, White RA, Wijeratne T, Wilkinson JD, Williams HC, Wiysonge CS, Woldeyohannes SM, Wolfe CDA, Won S, Wong JQ, Woolf AD, Xavier D, Xiao Q, Xu G, Yakob B, Yalew AZ, Yan LL, Yano Y, Yaseri M, Ye P, Yebyo HG, Yip P, Yirsaw BD, Yonemoto N, Yonga G, Younis MZ, Yu S, Zaidi Z, Zaki MES, Zannad F, Zavala DE, Zeeb H, Zeleke BM, Zhang H, Zodpey S, Zonies D, Zuhlke LJ, Vos T, Lopez AD, Murray CJL. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016 Oct 8;**388**(10053):1459–1544. doi:10.1016/S0140-6736(16)31012-1 PMID: 27733281

27. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, de Ferranti SD, Floyd J, Fornage M, Gillespie C, Isasi CR, Jim?nez MC, Jordan LC, Judd SE, Lackland D, Lichtman JH, Lisabeth L, Liu S, Longenecker CT, Mackey RH, Matsushita K, Mozaffarian D, Mussolino ME, Nasir K, Neumar RW, Palaniappan L, Pandey DK, Thiagarajan RR, Reeves MJ, Ritchey M, Rodriguez CJ, Roth GA, Rosamond WD, Sasson C, Towfighi A, Tsao CW, Turner MB, Virani SS, Voeks JH, Willey JZ, Wilkins JT, Wu JH, Alger HM, Wong SS, Muntner P, American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics?2017 Update: A Report From the American Heart Association. *Circulation*. 2017 Mar 7;**135**(10):e146–e603. doi:10.1161/CIR.0000000000000485 PMID: 28122885
28. McBride KL, Garg V. Impact of Mendelian inheritance in cardiovascular disease. *Ann N Y Acad Sci*. 2010 Dec;**1214**:122–37. doi:10.1111/j.1749-6632.2010.05791.x PMID: 20958326
29. Kessler T, Vilne B, Schunkert H. The impact of genome-wide association studies on the pathophysiology and therapy of cardiovascular disease. *EMBO Mol Med*. 2016;**8**(7):688–701. doi:10.15252/emmm.201506174 PMID: 27189168
30. Allen NB, Lloyd-Jones D, Hwang S-J, Rasmussen-Torvik L, Fornage M, Morrison AC, Baldridge AS, Boerwinkle E, Levy D, Cupples LA, Fox CS, Thanassoulis G, Dufresne L, Daviglus M, Johnson AD, Reis J, Rotter J, Palmas W, Allison M, Pankow JS, O'Donnell CJ. Genetic loci associated with ideal cardiovascular health: A meta-analysis of genome-wide association studies. *Am Heart J*. 2016

May;**175**:112–20. doi:10.1016/j.ahj.2015.12.022 PMID: 27179730

31. Warren HR, Evangelou E, Cabrera CP, Gao H, Ren M, Mifsud B, Ntalla I, Surendran P, Liu C, Cook JP, Kraja AT, Drenos F, Loh M, Verweij N, Marten J, Karaman I, Lepe MPS, O'Reilly PF, Knight J, Snieder H, Kato N, He J, Tai ES, Said MA, Porteous D, Alver M, Poulter N, Farrall M, Gansevoort RT, Padmanabhan S, Mägi R, Stanton A, Connell J, Bakker SJL, Metspalu A, Shields DC, Thom S, Brown M, Sever P, Esko T, Hayward C, van der Harst P, Saleheen D, Chowdhury R, Chambers JC, Chasman DI, Chakravarti A, Newton-Cheh C, Lindgren CM, Levy D, Kooner JS, Keavney B, Tomaszewski M, Samani NJ, Howson JMM, Tobin MD, Munroe PB, Ehret GB, Wain L V, International Consortium of Blood Pressure (ICBP) 1000G Analyses L V, BIOS Consortium A, Lifelines Cohort Study R, Understanding Society Scientific group R, CHD Exome+ Consortium PJ, ExomeBP Consortium AM, T2D-GENES Consortium P, GoT2DGenes Consortium CP, Cohorts for Heart and Ageing Research in Genome Epidemiology (CHARGE) BP Exome Consortium HR, International Genomics of Blood Pressure (iGEN-BP) Consortium LM, UK Biobank CardioMetabolic Consortium BP working group GC, Hottenga J-J, Strawbridge RJ, Esko T, Arking DE, Hwang S-J, Guo X, Kutalik Z, Trompet S, Shrine N, Teumer A, Ried JS, Bis JC, Smith A V, Amin N, Nolte IM, Lyytikäinen L-P, Mahajan A, Wareham NJ, Hofer E, Joshi PK, Kristiansson K, Traglia M, Havulinna AS, Goel A, Nalls MA, Sõber S, Vuckovic D, Luan J, M FDG, Ayers KL, Marrugat J, Ruggiero D, Lopez LM, Niiranen T, Enroth S, Jackson AU, Nelson CP, Huffman JE, Zhang W, Marten J, Gandin I, Harris SE, Zemonik T, Lu Y, Evangelou E, Shah N, de Borst MH, Mangino M, Prins BP, Campbell A, Li-Gao R, Chauhan G, Oldmeadow C, Abecasis G, Abedi M, Barbieri CM, Barnes MR, Batini C, Blake T, Boehnke M, Bottinger EP, Braund PS, Brown M, Brumat M, Campbell H, Chambers JC, Cocca M, Collins F, Connell J, Cordell HJ, Damman JJ, Davies G, de Geus EJ, de Mutsert R, Deelen J, Demirkale Y, Doney ASF, Dörr M, Farrall M, Ferreira T, Frånberg M, Gao H, Giedraitis V, Gieger C, Giulianini F, Gow AJ, Hamsten A, Harris TB, Hofman A, Holliday EG, Jarvelin M-R, Johansson Å, Johnson AD, Jousilahti P, Jula A, Kähönen M, Kathiresan S, Khaw K-T, Kolcic I, Koskinen S, Langenberg C, Larson M, Launer LJ, Lehne B, Liewald DCM, Lin L, Lind L, Mach F, Mamasoula C, Menni C, Mifsud B, Milaneschi Y, Morgan A, Morris AD, Morrison AC, Munson PJ, Nandakumar P, Nguyen QT, Nutile T, Oldehinkel AJ, Oostra BA, Org E, Padmanabhan S, Palotie A, Paré G, Pattie A, Penninx BWJH, Poulter N, Pramstaller PP, Raitakari OT, Ren M, Rice K, Ridker PM, Riese H, Ripatti S, Robino A, Rotter JI, Rudan I, Saba Y, Pierre A Saint, Sala CF, Sarin A-P, Schmidt R, Scott R, Seelen MA, Shields DC, Siscovick D, Sorice R, Stanton A, Stott DJ, Sundström J, Swertz M, Taylor KD, Thom S, Tzoulaki I, Tzourio C, Uitterlinden AG, Völker U, Vollenweider P, Wild S, Willemsen G, Wright AF, Yao J, Thériault S, Conen D, John A, Sever P, Debette S, Mook-Kanamori DO, Zeggini E, Spector TD, van der Harst P, Palmer CNA, Vergnaud A-C, Loos RJF, Polasek O, Starr JM, Girotto G, Hayward C, Kooner JS, Lindgren CM, Vitart V, Samani NJ, Tuomilehto J, Gyllensten U, Knekt P, Deary IJ, Ciullo M, Elosua R, Keavney BD, Hicks AA, Scott RA, Gasparini P, Laan M, Liu Y, Watkins H, Hartman CA, Salomaa V, Toniolo D, Perola M, Wilson JF, Schmidt H, Zhao JH, Lehtimäki T, van Duijn CM, Gudnason V, Psaty BM, Peters A, Rettig R, James A, Jukema JW, Strachan DP, Palmas W, Metspalu A, Ingelsson E, Boomsma DI, Franco OH, Bochud M, Newton-Cheh C, Munroe PB, Elliott P, Chasman DI, Chakravarti A, Knight J, Morris AP, Levy D, Tobin MD, Snieder H, Caulfield MJ, Ehret GB, Barnes MR, Tzoulaki I, Caulfield MJ, Elliott P. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat Genet*. 2017 Mar 30;**49**(3):403–415. doi:10.1038/ng.3768 PMID: 28135244
32. Auer PL, Stitziel NO. Genetic association studies in cardiovascular diseases: Do we have enough power? *Trends Cardiovasc Med*. 2017; doi:10.1016/j.tcm.2017.03.005
33. Larson MG, Atwood LD, Benjamin EJ, Cupples LA, D'Agostino RB, Fox CS, Govindaraju DR, Guo C-Y, Heard-Costa NL, Hwang S-J, Murabito JM, Newton-Cheh C, O'Donnell CJ, Seshadri S, Vasan RS, Wang TJ, Wolf PA, Levy D. Framingham Heart Study 100K project: genome-wide associations for cardiovascular disease outcomes. *BMC Med Genet*. 2007 Sep 19;**8**(Suppl 1):S5. doi:10.1186/1471-2350-8-S1-S5 PMID: 17903304
34. Roberts R, Marian AJ, Dandona S, Stewart AFR. Genomics in cardiovascular disease. *J Am Coll Cardiol*. 2013 May 21;**61**(20):2029–37. doi:10.1016/j.jacc.2012.12.054 PMID: 23524054
35. Musunuru K. Personalized genomes and cardiovascular disease. *Cold Spring Harb Perspect Med*. 2014 Sep 25;**5**(1):a014068. doi:10.1101/cshperspect.a014068 PMID: 25256177
36. Kathiresan S, Srivastava D. Genetics of human cardiovascular disease. *Cell*. 2012 Mar 16;**148**(6):1242–57. doi:10.1016/j.cell.2012.03.001 PMID: 22424232
37. Cecconi M, Parodi MI, Formisano F, Spirito P, Autore C, Musumeci MB, Favale S, Forleo C, Rapezzi C, Biagini E, Davì S, Canepa E, Pennese L, Castagnetta M, Degiorgio D, Coviello DA, Group CW. Targeted next-generation sequencing helps to decipher the genetic and phenotypic heterogeneity of hypertrophic

cardiomyopathy. *Int J Mol Med*. 2016 Oct;**38**(4):1111–24. doi:10.3892/ijmm.2016.2732 PMID: 27600940
38. Faita F, Vecoli C, Foffa I, Andreassi MG. Next generation sequencing in cardiovascular diseases. *World J Cardiol*. 2012 Oct 26;**4**(10):288–95. doi:10.4330/wjc.v4.i10.288 PMID: 23110245
39. Rühle F, Stoll M. Long non-coding RNA Databases in Cardiovascular Research. *Genomics Proteomics Bioinformatics*. 2016 Aug;**14**(4):191–9. doi:10.1016/j.gpb.2016.03.001 PMID: 27049585
40. Fan X, Abbott TE, Larson D, Chen K. BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. *Curr Protoc Bioinforma*. Hoboken, NJ, USA; 2014 Mar 21;**45**:15.6.1-11. doi:10.1002/0471250953.bi1506s45 PMID: 25152801
41. Lam HYK, Mu XJ, Stütz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol*. 2010 Jan 27;**28**(1):47–55. doi:10.1038/nbt.1600 PMID: 20037582
42. 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*. 2012 May;**40**(9):e69. doi:10.1093/nar/gks003 PMID: 22302147
43. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 2011 Jun 1;**21**(6):974–984. doi:10.1101/gr.114876.110 PMID: 21324876
44. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012 Sep 15;**28**(18):i333–i339. doi:10.1093/bioinformatics/bts378 PMID: 22962449
45. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA. Large multiallelic copy number variations in humans. *Nat Genet*. 2015 Jan 26;**47**(3):296–303. doi:10.1038/ng.3200 PMID: 25621458
46. Lindberg MR, Hall IM, Quinlan AR. Population-based structural variation discovery with Hydra-Multi. *Bioinformatics*. 2015 Apr 15;**31**(8):1286–9. doi:10.1093/bioinformatics/btu771 PMID: 25527832
47. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014 Jun 26;**15**(6):R84. doi:10.1186/gb-2014-15-6-r84 PMID: 24970577
48. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh C-H, Zhang C, Ren X, Protopopov A, Chin L, Kucherlapati R, Lee C, Park PJ. Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes. *Cell*. 2013 May 9;**153**(4):919–929. doi:10.1016/j.cell.2013.04.010 PMID: 23663786
49. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol*. 2009 Feb 23;**10**(2):R23. doi:10.1186/gb-2009-10-2-r23 PMID: 19236709
50. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009 Nov 1;**25**(21):2865–71. doi:10.1093/bioinformatics/btp394 PMID: 19561018
51. Ye K, Wang J, Jayasinghe R, Lameijer E-W, McMichael JF, Ning J, McLellan MD, Xie M, Cao S, Yellapantula V, Huang K, Scott A, Foltz S, Niu B, Johnson KJ, Moed M, Slagboom PE, Chen F, Wendl MC, Ding L. Systematic discovery of complex insertions and deletions in human cancers. *Nat Med*. 2015 Dec 14;**22**(1):97–104. doi:10.1038/nm.4002 PMID: 26657142
52. Malhotra A, Lindberg M, Faust GG, Leibowitz ML, Clark RA, Layer RM, Quinlan AR, Hall IM. Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res*. 2013 May 1;**23**(5):762–76. doi:10.1101/gr.143677.112 PMID: 23410887
53. Malhotra A, Wang Y, Waters J, Chen K, Meric-Bernstam F, Hall IM, Navin NE. Ploidy-Seq: inferring mutational chronology by sequencing polyploid tumor subpopulations. *Genome Med*. 2015;**7**(1):6. doi:10.1186/s13073-015-0127-5 PMID: 25729435
54. Zhang ZD, Du J, Lam H, Abyzov A, Urban AE, Snyder M, Gerstein M. Identification of genomic indels and structural variations using split reads. *BMC Genomics*. 2011 Dec 25;**12**(1):375. doi:10.1186/1471-2164-12-375 PMID: 21787423
55. Wu J, Lee W-P, Ward A, Walker JA, Konkel MK, Batzer MA, Marth GT. Tangram: a comprehensive toolbox for mobile element insertion detection. *BMC Genomics*. 2014 Sep 16;**15**(1):795. doi:10.1186/1471-2164-15-795 PMID: 25228379
56. Lee W-P, Marth G, Wu J. Toolbox for Mobile-Element Insertion Detection on Cancer Genomes. *Cancer Inform*. 2015 Feb;**14**(Suppl 1):37. doi:10.4137/CIN.S24657 PMID: 25931804

57. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009 Jul 15;**25**(14):1754–60. doi:10.1093/bioinformatics/btp324 PMID: 19451168
58. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;**9**(4):357–360. doi:10.1038/nmeth.1923 PMID: 22388286
59. Lee W-P, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. Hsiao CK, editor. *PLoS One*. 2014 Jan;**9**(3):e90581. doi:10.1371/journal.pone.0090581 PMID: 24599324
60. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res*. 2012 Mar;**22**(3):549–56. doi:10.1101/gr.126953.111 PMID: 22156294
61. Chen K, Chen L, Fan X, Wallis J, Ding L, Weinstock G. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res*. 2014 Feb;**24**(2):310–7. doi:10.1101/gr.162883.113 PMID: 24307552
62. Abyzov A, Gerstein M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics*. 2011 Mar 1;**27**(5):595–603. doi:10.1093/bioinformatics/btq713 PMID: 21233167
63. LEVANDOWSKY M, WINTER D. Distance between Sets. *Nature*. 1971 Nov 5;**234**(5323):34–35. doi:10.1038/234034a0
64. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL. A Comprehensive Analysis of Common Copy-Number Variations in the Human Genome. *Am J Hum Genet*. 2007 Jan;**80**(1):91–104. doi:10.1086/510560 PMID: 17160897
65. Bailey JA, Kidd JM, Eichler EE. Human copy number polymorphic genes. *Cytogenet Genome Res*. 2009 Mar 11;**123**(1–4):234–243. doi:10.1159/000184713 PMID: 19287160
66. Habegger L, Balasubramanian S, Chen DZ, Khurana E, Sboner A, Harmanci A, Rozowsky J, Clarke D, Snyder M, Gerstein M. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics*. 2012 Sep 1;**28**(17):2267–9. doi:10.1093/bioinformatics/bts368 PMID: 22743228
67. Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C, Kato M, Miller DM, Slack F, Snyder M, Waterston RH, Reinke V, Gerstein MB. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res*. 2011 Feb 1;**21**(2):276–85. doi:10.1101/gr.110189.110 PMID: 21177971
68. Mu XJ, Lu ZJ, Kong Y, Lam HYK, Gerstein MB. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res*. 2011 Sep 1;**39**(16):7058–76. doi:10.1093/nar/gkr342 PMID: 21596777
69. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol*. 2009 Jan 4;**27**(1):66–75. doi:10.1038/nbt.1518 PMID: 19122651
70. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee B-K, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kellis M, Khatun J, Kheradpour P, Kundaje A, Lassmann T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SCJ, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LAL, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elnitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Stamatoyannopoulos JA, Tenenbaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ,

Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandhu KS, Schaeffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee B-K, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Ki Kim S, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge E, Trout D, Varley KE, Gasper C, Bansal A, Pepke S, Jain P, Amrhein H, Bowling KM, Anaya M, Cross MK, King B, Muratet MA, Antoshechkin I, Newberry KM, McCue K, Nesmith AS, Fisher-Aylor KI, Pusey B, DeSalvo G, Parker SL, Balasubramanian S, Davis NS, Meadows SK, Eggleston T, Gunter C, Newberry JS, Levy SE, Absher DM, Mortazavi A, Wong WH, Wold B, Blow MJ, Visel A, Pennachio LA, Elnitski L, Margulies EH, Parker SCJ, Petrykowska HM, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Chrast J, Davidson C, Derrien T, Despacio-Reyes G, Diekhans M, Ezkurdia I, Frankish A, Gilbert J, Gonzalez JM, Griffiths E, Harte R, Hendrix DA, Howald C, Hunt T, Jungreis I, Kay M, Khurana E, Kokocinski F, Leng J, Lin MF, Loveland J, Lu Z, Manthravadi D, Mariotti M, Mudge J, Mukherjee G, Notredame C, Pei B, Rodriguez JM, Saunders G, Sboner A, Searle S, Sisu C, Snow C, Steward C, Tanzer A, Tapanari E, Tress ML, van Baren MJ, Walters N, Washietl S, Wilming L, Zadissa A, Zhang Z, Brent M, Haussler D, Kellis M, Valencia A, Gerstein M, Reymond A, Guigó R, Harrow J, Hubbard TJ, Landt SG, Frietze S, Abyzov A, Addleman N, Alexander RP, Auerbach RK, Balasubramanian S, Bettinger K, Bhardwaj N, Boyle AP, Cao AR, Cayting P, Charos A, Cheng Y, Cheng C, Eastman C, Euskirchen G, Fleming JD, Grubert F, Habegger L, Hariharan M, Harmanci A, Iyengar S, Jin VX, Karczewski KJ, Kasowski M, Lacroute P, Lam H, Lamarre-Vincent N, Leng J, Lian J, Lindahl-Allen M, Min R, Miotto B, Monahan H, Moqtaderi Z, Mu XJ, O'Geen H, Ouyang Z, Patacsil D, Pei B, Raha D, Ramirez L, Reed B, Rozowsky J, Sboner A, Shi M, Sisu C, Slifer T, Witt H, Wu L, Xu X, Yan K-K, Yang X, Yip KY, Zhang Z, Struhl K, Weissman SM, Gerstein M, Farnham PJ, Snyder M, Tenenbaum SA, Penalva LO, Doyle F, Karmakar S, Landt SG, Bhanvadia RR, Choudhury A, Domanus M, Ma L, Moran J, Patacsil D, Slifer T, Victorsen A, Yang X, Snyder M, White KP, Auer T, Centanin L, Eichenlaub M, Gruhl F, Heermann S, Hoeckendorf B, Inoue D, Kellner T, Kirchmaier S, Mueller C, Reinhardt R, Schertel L, Schneider S, Sinn R, Wittbrodt B, Wittbrodt J, Weng Z, Whitfield TW, Wang J, Collins PJ, Aldred SF, Trinklein ND, Partridge EC, Myers RM, Dekker J, Jain G, Lajoie BR, Sanyal A, Balasundaram G, Bates DL, Byron R, Canfield TK, Diegel MJ, Dunn D, Ebersol AK, Frum T, Garg K, Gist E, Hansen RS, Boatman L, Haugen E, Humbert R, Jain G, Johnson AK, Johnson EM, Kutyavin T V., Lajoie BR, Lee K, Lotakis D, Maurano MT, Neph SJ, Neri F V., Nguyen ED, Qu H, Reynolds AP, Roach V, Rynes E, Sabo P, Sanchez ME, Sandstrom RS, Sanyal A, Shafer AO, Stergachis AB, Thomas S, Thurman RE, Vernot B, Vierstra J, Vong S, Wang H, Weaver MA, Yan Y, Zhang M, Akey JM, Bender M, Dorschner MO, Groudine M, MacCoss MJ, Navas P, Stamatoyannopoulos G, Kaul R, Dekker J, Stamatoyannopoulos JA, Dunham I, Beal K, Brazma A, Flicek P, Herrero J, Johnson N, Keefe D, Lukk M, Luscombe NM, Sobral D, Vaquerizas JM, Wilder SP, Batzoglou S, Sidow A, Hussami N, Kyriazopoulou-Panagiotopoulou S, Libbrecht MW, Schaub MA, Kundaje A, Hardison RC, Miller W, Giardine B, Harris RS, Wu W, Bickel PJ, Banfai B, Boley NP, Brown JB, Huang H, Li Q, Li JJ, Noble WS, Bilmes JA, Buske OJ, Hoffman MM, Sahu AD, Kharchenko P V., Park PJ, Baker D, Taylor J, Weng Z, Iyer S, Dong X, Greven M, Lin X, Wang J, Xi HS, Zhuang J, Gerstein M, Alexander RP, Balasubramanian S, Cheng C, Harmanci A, Lochovsky L, Min R, Mu XJ, Rozowsky J, Yan K-K, Yip KY, Birney E. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012 Sep 5;**489**(7414):57–74. doi:10.1038/nature11247 PMID: 22955616

71. Harmanci A, Rozowsky J, Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol*. 2014 Oct 8;**15**(10):474. doi:10.1186/s13059-014-0474-3 PMID: 25292436

72. Cheng C, Yan K-K, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol*. 2011;**12**(2):R15. doi:10.1186/gb-2011-12-2-r15 PMID: 21324173

73. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan K-K, Dong X, Djebali S, Ruan Y, Davis CA, Carninci P, Lassman T, Gingeras TR, Guigo R, Birney E, Weng Z, Snyder M, Gerstein M. Understanding transcriptional regulation by integrative analysis of transcription factor binding data.

Genome Res. 2012 Sep 1;**22**(9):1658–1667. doi:10.1101/gr.136838.111 PMID: 22955978

74. Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME, Boley NP, Booth BW, Cherbas L, Cherbas P, Di C, Dobin A, Drenkow J, Ewing B, Fang G, Fastuca M, Feingold EA, Frankish A, Gao G, Good PJ, Guig? R, Hammonds A, Harrow J, Hoskins RA, Howald C, Hu L, Huang H, Hubbard TJP, Huynh C, Jha S, Kasper D, Kato M, Kaufman TC, Kitchen RR, Ladewig E, Lagarde J, Lai E, Leng J, Lu Z, MacCoss M, May G, McWhirter R, Merrihew G, Miller DM, Mortazavi A, Murad R, Oliver B, Olson S, Park PJ, Pazin MJ, Perrimon N, Pervouchine D, Reinke V, Reymond A, Robinson G, Samsonova A, Saunders GI, Schlesinger F, Sethi A, Slack FJ, Spencer WC, Stoiber MH, Strasbourger P, Tanzer A, Thompson OA, Wan KH, Wang G, Wang H, Watkins KL, Wen J, Wen K, Xue C, Yang L, Yip K, Zaleski C, Zhang Y, Zheng H, Brenner SE, Graveley BR, Celniker SE, Gingeras TR, Waterston R. Comparative analysis of the transcriptome across distant species. *Nature*. 2014 Aug 27;**512**(7515):445–448. doi:10.1038/nature13424 PMID: 25164755
75. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol*. 2012 Sep 26;**13**(9):R48. doi:10.1186/gb-2012-13-9-r48 PMID: 22950945
76. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012 Sep 6;**489**(7414):91–100. doi:10.1038/nature11245 PMID: 22955619
77. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of Genomic Variants Using a Unified Biological Network Approach. Rzhetsky A, editor. *PLoS Comput Biol*. 2013 Mar 7;**9**(3):e1002886. doi:10.1371/journal.pcbi.1002886 PMID: 23505346
78. Kim PM, Korbel JO, Gerstein MB. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A*. 2007 Dec 18;**104**(51):20274–9. doi:10.1073/pnas.0710183104 PMID: 18077332
79. Cheng C, Andrews E, Yan K-K, Ung M, Wang D, Gerstein M. An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. *Genome Biol*. 2015 Mar 31;**16**(1):63. doi:10.1186/s13059-015-0624-2 PMID: 25880651
80. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gümüs ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GRS, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, 1000 Genomes Project Consortium ET, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C, Gerstein M. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*. 2013 Oct 4;**342**(6154):1235587. doi:10.1126/science.1235587 PMID: 24092746
81. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol*. 2014 Oct 2;**15**(10):480. doi:10.1186/s13059-014-0480-5 PMID: 25273974
82. Abyzov A, Li S, Kim DR, Mohiyuddin M, St?tz AM, Parrish NF, Mu XJ, Clark W, Chen K, Hurles M, Korbel JO, Lam HYK, Lee C, Gerstein MB. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun*. 2015 Jun 1;**6**:7256. doi:10.1038/ncomms8256 PMID: 26028266
83. PsychENCODE Consortium S, Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S, Geschwind DH, Mill J, Nairn AC, Abyzov A, Pochareddy S, Prabhakar S, Weissman S, Sullivan PF, State MW, Weng Z, Peters MA, White KP, Gerstein MB, Amiri A, Armoskus C, Ashley-Koch AE, Bae T, Beckel-Mitchener A, Berman BP, Coetzee GA, Coppola G, Francoeur N, Fromer M, Gao R, Grennan K, Herstein J, Kavanagh DH, Ivanov NA, Jiang Y, Kitchen RR, Kozlenkov A, Kundakovic M, Li M, Li Z, Liu S, Mangravite LM, Mattei E, Markenscoff-Papadimitriou E, Navarro FCP, North N, Omberg L, Panchision D, Parikshak N, Poschmann J, Price AJ, Purcaro M, Reddy TE, Roussos P, Schreiner S, Scuderi S, Sebra R, Shibata M, Shieh AW, Skarica M, Sun W, Swarup V, Thomas A, Tsuji J, van Bakel H, Wang D, Wang Y, Wang K, Werling DM, Willsey AJ, Witt H, Won H, Wong CCY,

Wray GA, Wu EY, Xu X, Yao L, Senthil G, Lehner T, Sklar P, Sestan N. The PsychENCODE project. *Nat Neurosci*. 2015 Dec 25;**18**(12):1707–12. doi:10.1038/nn.4156 PMID: 26605881

84. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*. 2011 Jan 15;**27**(2):281–3. doi:10.1093/bioinformatics/btq643 PMID: 21134889
85. Du J, Leng J, Habegger L, Sboner A, McDermott D, Gerstein M. IQSeq: integrated isoform quantification analysis based on next-generation sequencing. Rzhetsky A, editor. *PLoS One*. 2012 Jan 6;**7**(1):e29175. doi:10.1371/journal.pone.0029175 PMID: 22238592
86. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, Demichelis F, Rubin MA, Gerstein MB. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol*. 2010;**11**(10):R104. doi:10.1186/gb-2010-11-10-r104 PMID: 20964841
87. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M, Gerstein M. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol*. 2011 Aug 2;**7**(1):522. doi:10.1038/msb.2011.54 PMID: 21811232
88. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat Commun*. 2016 Apr 18;**7**:11101. doi:10.1038/ncomms11101 PMID: 27089393
89. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat Methods*. 2016 Mar 1;**13**(3):251–6. doi:10.1038/nmeth.3746 PMID: 26828419
90. Blin K, Dieterich C, Wurmus R, Rajewsky N, Landthaler M, Akalin A. DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res*. 2015 Jan 28;**43**(D1):D160–D167. doi:10.1093/nar/gku1180 PMID: 25416797
91. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*. 2014 Jan;**42**(Database issue):D92-7. doi:10.1093/nar/gkt1248 PMID: 24297251
92. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp A-C, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*. 2010 Apr 2;**141**(1):129–141. doi:10.1016/j.cell.2010.03.009 PMID: 20371350
93. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*. 2013 Apr 25;**153**(3):654–65. doi:10.1016/j.cell.2013.03.043 PMID: 23622248
94. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol*. 2011 Sep 11;**18**(10):1139–46. doi:10.1038/nsmb.2115 PMID: 21909094
95. Roadmap Epigenomics Consortium A, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M, Echipare L, Edsall L, Flowers D, Genbacev-Krtolica O, Gifford C, Gillespie S, Giste E, Glass IA, Gnirke A, Gormley M, Gu H, Gu J, Hafler DA, Hangauer MJ, Hariharan M, Hatan M, Haugen E, He Y, Heimfeld S, Herlofsen S, Hou Z, Humbert R, Issner R, Jackson AR, Jia H, Jiang P, Johnson AK, Kadlecek T, Kamoh B, Kapidzic M, Kent J, Kim A, Kleinewietfeld M, Klugman S, Krishnan J, Kuan S, Kutyavin T, Lee A-Y, Lee K, Li J, Li N, Li Y, Ligon KL, Lin S, Lin Y, Liu J, Liu Y, Luckey CJ, Ma YP, Maire C, Marson A, Mattick JS, Mayo M, McMaster M, Metsky H, Mikkelsen T, Miller D, Miri M, Mukame E, Nagarajan RP, Neri F, Nery J, Nguyen T, O'Geen H, Paithankar S, Papayannopoulou T, Pelizzola M, Plettner P, Propson NE, Raghuraman S, Raney BJ, Raubitschek A, Reynolds AP, Richards H, Riehle K, Rinaudo P, Robinson JF, Rockweiler NB, Rosen E, Rynes E, Schein J, Sears R, Sejnowski T, Shafer A, Shen L, Shoemaker R,

Sigaroudinia M, Slukvin I, Stehling-Sun S, Stewart R, Subramanian SL, Suknuntha K, Swanson S, Tian S, Tilden H, Tsai L, Urich M, Vaughn I, Vierstra J, Vong S, Wagner U, Wang H, Wang T, Wang Y, Weiss A, Whitton H, Wildberg A, Witt H, Won K-J, Xie M, Xing X, Xu I, Xuan Z, Ye Z, Yen C, Yu P, Zhang X, Zhang X, Zhao J, Zhou Y, Zhu J, Zhu Y, Ziegler S, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y-C, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K-H, Feizi S, Karlic R, Kim A-R, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L-H, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015 Feb 19;**518**(7539):317–30. doi:10.1038/nature14248 PMID: 25693563

96. Ziller MJ, Edri R, Yaffe Y, Donaghey J, Pop R, Mallard W, Issner R, Gifford CA, Goren A, Xing J, Gu H, Cacchiarelli D, Tsankov AM, Epstein C, Rinn JL, Mikkelsen TS, Kohlbacher O, Gnirke A, Bernstein BE, Elkabetz Y, Meissner A. Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature*. 2015 Feb 19;**518**(7539):355–9. doi:10.1038/nature13990 PMID: 25533951
97. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen C-A, Lin S, Lin Y, Qiu Y, Xie W, Yue F, Hariharan M, Ray P, Kuan S, Edsall L, Yang H, Chi NC, Zhang MQ, Ecker JR, Ren B. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*. 2015 Feb 19;**518**(7539):350–4. doi:10.1038/nature14217 PMID: 25693566
98. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012 May 15;**28**(10):1353–8. doi:10.1093/bioinformatics/bts163 PMID: 22492648
99. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. Ultraconserved Elements in the Human Genome. *Science (80-)*. 2004 May 28;**304**(5675):1321–1325. doi:10.1126/science.1098119 PMID: 15131266
100. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005 Jul 17;**15**(7):901–13. doi:10.1101/gr.3577405 PMID: 15965027
101. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*. 2010 Oct;**28**(10):1045–8. doi:10.1038/nbt1010-1045 PMID: 20944595
102. Wendl MC, Wilson RK. Statistical aspects of discerning indel-type structural variation via DNA sequence alignment. *BMC Genomics*. 2009 Aug 5;**10**(1):359. doi:10.1186/1471-2164-10-359 PMID: 19656394
103. Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am J Hum Genet*. 2008 Sep;**83**(3):311–321. doi:10.1016/j.ajhg.2008.06.024 PMID: 18691683
104. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*. 2004 Aug 6;**305**(5685):869–72. doi:10.1126/science.1099870 PMID: 15297675
105. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res*. 2007 Feb 3;**615**(1–2):28–56. doi:10.1016/j.mrfmmm.2006.09.003 PMID: 17101154
106. Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am J Hum Genet*. 2011 Jul 15;**89**(1):82–93. doi:10.1016/j.ajhg.2011.05.029 PMID: 21737059
107. Lu C, Xie M, Wendl MC, Wang J, McLellan MD, Leiserson MDM, Huang K-L, Wyczalkowski MA, Jayasinghe R, Banerjee T, Ning J, Tripathi P, Zhang Q, Niu B, Ye K, Schmidt HK, Fulton RS, McMichael JF, Batra P, Kandoth C, Bharadwaj M, Koboldt DC, Miller CA, Kanchi KL, Eldred JM, Larson DE, Welch JS, You M, Ozenberger BA, Govindan R, Walter MJ, Ellis MJ, Mardis ER, Graubert TA, Dipersio JF, Ley

TJ, Wilson RK, Goodfellow PJ, Raphael BJ, Chen F, Johnson KJ, Parvin JD, Ding L. Patterns and functional implications of rare germline variants across 12 cancer types. *Nat Commun*. 2015 Dec 22;**6**:10086. doi:10.1038/ncomms10086 PMID: 26689913

108. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team DC, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012 Aug 10;**91**(2):224–37. doi:10.1016/j.ajhg.2012.06.007 PMID: 22863193
109. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee Y-H, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimaki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King M-C, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M. Strong Association of De Novo Copy Number Mutations with Autism. *Science (80-)*. 2007 Apr 20;**316**(5823):445–449. doi:10.1126/science.1138659 PMID: 17363630
110. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King M-C, Sebat J. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*. 2008 Apr 25;**320**(5875):539–43. doi:10.1126/science.1155174 PMID: 18369103
111. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE. A copy number variation morbidity map of developmental delay. *Nat Genet*. 2011 Aug 14;**43**(9):838–846. doi:10.1038/ng.909 PMID: 21841781
112. Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. Schork NJ, editor. *PLoS Genet*. 2009 Feb 13;**5**(2):e1000384. doi:10.1371/journal.pgen.1000384 PMID: 19214210
113. Pan W, Shen X. Adaptive tests for association analysis of rare variants. *Genet Epidemiol*. 2011 Jul;**35**(5):381–388. doi:10.1002/gepi.20586 PMID: 21520272
114. Liu R, Holik AZ, Su S, Jansz N, Chen K, Leong HS, Blewitt ME, Asselin-Labat M-L, Smyth GK, Ritchie ME. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res*. 2015 Sep 3;**43**(15):e97. doi:10.1093/nar/gkv412 PMID: 25925576
115. Lee S, Abecasis G, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am J Hum Genet*. 2014 Jul 3;**95**(1):5–23. doi:10.1016/j.ajhg.2014.06.009 PMID: 24995866