**Gerstein lab experience in constructing and mapping genomics datasets to diploid personal genomes**

***Structural variant detection*:** A major component of personal genome construction is accurately identify and annotate structural variants (SVs) in individual genomes compared to the reference genome25–28. The Gerstein lab has significant experience with developing tools for detection of SVs in a variety of contexts. In particular, we have extensively studied the characteristics of SVs originating from different mechanisms, indicating specific creation processes and potentially divergent functional impacts29,30. In fact, we have performed SV mechanism annotations for the 1000 Genomes Phase 3 deletions using BreakSeq, and categorized 29,774 deletions into NAHR, NHR, TEI and VNTR-based mechanisms28,31. Among these, NHR is the most prevalent mechanism32 (~73% of all categorized deletions). Additionally, we have also quantified the constraint induced by SVs on various genomic elements, finding that SVs affecting genic regions (coding, UTRs, and intronic) and regulatory elements including enhancers and TF binding sites are highly depleted32. In contrast, we found SVs to be enriched among pseudogenes consistent with their formation mechanism, which involves either duplication or retrotransposition. We performed similar analysis as part of the Human Genome Structural Variant Consortium (HGSVC), identifying the mechanisms of formation for annotated SVs identified 34 assembled genomes constructed from deep long-read sequencing data9. Finally, we recently built an agnostic machine learning tool called SVFX to assign pathogenicity scores to somatic and germline SVs in various diseases33. Using SVFX, we identified key roles for gene regulatory elements and conserved sequences towards SV pathogenicity33.

***Personal genome construction*:** The alignment of assay reads is one of the main steps in processing functional genomic datasets. Conventionally, reads are aligned to the human reference genome. However, a systematic reference bias is introduced when reads are mapped to the haploid human reference sequence, as reads that harbor an alternate allele are less likely to be aligned. In addition, reads can be improperly mapped to the reference genome in regions or samples with more genetic variation, especially when indels and larger structural variants are involved. This reduced mappability impairs estimation of read abundance and therefore compromises variant calling and any downstream analyses. For constructing linear personal genomes, we have developed a computational tool called vcf2diploid10. The tool integrates an individual’s genomic variation data (SNVs, indels, and SVs) into the reference genome. Phase information of heterozygous variants is also incorporated, producing maternal and paternal haplotypes in the resulting genome. Chain files generated by the program can be used to account for coordinate offsets between the individual’s parental haplotypes and the original reference genomic sequence. The versatility to convert between reference and personal genome coordinates allows for mapping of genomic annotated regions (e.g. gene or peak coordinates for RNA-seq and ChIP-seq, respectively) between genomes using available tools, such as the UCSC liftOver tool34. We have previously constructed the personal diploid genome, splice-junction libraries, and personalized gene annotations for NA1287810. We have made this assembly available as a resource at alleleseq.gersteinlab.org and regularly update the reference assembly as new versions of the human reference genome, genomic annotations, and NA12878 genomic variation data are released. Furthermore, the availability of a computational tool enables the construction of personal genomes in a high-throughput fashion, as demonstrated in a recent publication where we built 382 personal genomes using the variant call sets from the 1000 Genomes Project35.

In addition to vcf2diploid, we recently led an effort within the ENCODE and GTEx consortia to develop the EN-TEx personal genome-based resource, consisting of linked sets of >1600 functional genomics assays (RNA-Seq, DNAse- and ATAC-Seq, ChIP-Seq, Hi-C, proteomics, and methylation assays) across 30 tissues in four individuals with European ancestry24. As part of this effort, we constructed personal genomes using phased SVs and SNPs from high-quality short- and long-read sequencing datasets. Structural variants were identified from a combination of PacBio CLR and Oxford Nanopore PromethION sequencing to an average read length of 15kbp and 60kbp, respectively. SVs were identified using an intersection of Sniffles and pbsv, and were refined using Iris package, including refinement of novel insertions36,37. We then used a novel pipeline to phase large variants by matching reads supporting the variants to phased haplotypes determined from long-read Hi-C sequencing data in each sample. The phased variants were then integrated within the CrossStitch software package to generate fully phased personal genomes12. The coordinates in these genomes can be mapped back to human reference coordinates to facilitate comparisons with other reference-based resources, such as gene annotations. Overall, we identified ~2.4M SNPs and >15,000 SVs per individual using this pipeline; among the SVs identified were ~6,000 private SVs and ~7,000 SVs located within transposable elements per person24.

***Mapping functional genomics datasets to personal genomes:*** In the framework of EN-TEx, we carried out an extensive evaluation of how personal genomes improve mapping and quantification of different types of genomic data compared to the reference genome24. Specifically, we mapped 1,635 datasets (15 assays in 30 tissues) to phased diploid and reference genomes, giving rise to three signal tracks for each assay (two haplotypes and the reference). Mapping of RNA-Seq reads was performed using STAR-2.6.0c49, and other datasets (i.e., DNAse-Seq, Hi-C, and ChIP-Seq) were mapped using bwa50. We found that mapping sequences to the derived haplotypes rather than the reference genome resulted in an overall improvement in mapping accuracy for each assay24. By applying conventional mapping criteria, we observed an increase in the number of mapped reads by about 0.5–1% in the personal genome-mapped data; with more stringent filtering criteria for high-quality and uniquely mapping sequences, we observed a much larger improvement of 2–4% more mapped reads across the four individuals24. The use of personal genome-mapped data allowed us to more accurately assess variant effects towards gene regulation, especially for SVs. In fact, we identified ~300 SV-QTLs that often co-occurred with disrupted chromatin peaks24, and further observed a depletion of open chromatin regions near transposable elements, potentially due to silencing of transposable loci. These findings highlight the utility of personal genome approaches to better identify functional genomics elements in individuals.

***Determining allele specificity of regulatory elements:*** We have spearheaded allele specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project10,24,35,51–53. For example, in the 1000 Genomes project, we annotated variants associated with allele-specific expression (ASE) and binding (ASB) by matching functional datasets35 (955 RNA-Seq and 165 ChIP-Seq in total) with genotype data54. Overall, we detected more than 6K and 63K SNVs associated with ASB and ASE, respectively, which we made available as an online resource, AlleleDB35. We further constructed a high-resolution map of allelic imbalances in DNA methylation, histone marks, and transcription in 71 epigenomes from 36 distinct cell and tissue types from 13 donors53. To measure ASE, ASB, and AS chromatin activity (ASCA) in diploid genomes, we developed a pipeline dubbed AlleleSeq2 to incorporate personal variation, including large SVs, to account for reference bias in a straightforward way10,35,55. We included additional filters to mitigate ambiguous mapping biases35,56. To account for the over-dispersed nature of functional genomics read count data, the significance of allelic imbalance is assessed using the beta-binomial test35. In our EN-TEx datasets, we found that 1.1-7.3% of assay reads across all samples that were preferentially aligned to a haplotype overlapped with hetSNVs, with the allelic imbalance measured by the fraction of unique reads mapping to each haplotype24. Using AlleleSeq2, we generated a catalog of >1 million allele-specific loci from EN-Tex data24. These loci exhibit coordinated activity along haplotypes and are less conserved than corresponding, non-allele-specific ones. Combining these data with existing annotations also revealed strong associations between allele-specific and GWAS loci24.