**Summary of experience in analyzing the functional impact of structural variants.**

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.

***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. VAT provides transcript-specific annotations of point mutations and indels according to synonymous, missense, nonsense or splice-site-disrupting changes[26](#_ENREF_26). 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 the coding sequences, untranslated regions and introns of genes 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 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[27](#_ENREF_27). 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[28](#_ENREF_28).

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[29](#_ENREF_29),[30](#_ENREF_30); it has been widely used in consortium projects such as ENCODE[29](#_ENREF_29),[31](#_ENREF_31). 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[30](#_ENREF_30). 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[32-35](#_ENREF_32). 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[28](#_ENREF_28) [35](#_ENREF_35),[36](#_ENREF_36). We used metrics such as diversity and fraction of rare variants to characterize selection pressure on various classes and subclasses of functional annotations[28](#_ENREF_28). We have also defined variants that are disruptive to a TF-binding motif in a regulatory region[31](#_ENREF_31).

*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[37](#_ENREF_37) and are positioned at the top of regulatory networks[36](#_ENREF_36). Further studies showed relationships between selection and protein network topology (e.g., quantifying selection in hubs relative to proteins on the network periphery[37](#_ENREF_37),[38](#_ENREF_38)). Incorporating multiple network and evolutionary properties, we developed NetSNP[37](#_ENREF_37) 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[36](#_ENREF_36). To quantify the degree of hierarchy for a given hierarchical network, we defined a metric called hierarchical score maximization (HSM[39](#_ENREF_39)).

*FunSeq: Tools for integrated functional prioritization.*We recently developed a prioritization pipeline called FunSeq[40](#_ENREF_40),[41](#_ENREF_41) 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).

*Mutational mechanisms of structural variants.*The sequence content of SVs, especially around breakpoints, carries important information about origin and functional impact. Using datasets from 1000GP, we studied the distinct features of SVs originating from different mechanisms[40](#_ENREF_40),[42](#_ENREF_42). We performed SV mechanism annotations for the 1000GP Phase 3 deletions using BreakSeq[17](#_ENREF_17), categorizing 29,774 deletions by their creation mechanisms. Among these, NHR proved to be the most prevalent mechanism (~73% of all categorized deletions)[23](#_ENREF_23). 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 setting up pipelines for the processing of RNA-seq data; specially for long RNA-seq data for ENCODE, long and short RNA-seq data for the PsychENCODE[43](#_ENREF_43) and Brainspan project as well as 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[44](#_ENREF_44) 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[45](#_ENREF_45) 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[46](#_ENREF_46) was to detect fusion transcript in RNA-seq data, which can be important biomarker for diseases such as various types of 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 [47](#_ENREF_47). 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[48](#_ENREF_48). 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[49](#_ENREF_49).