Preliminary data.

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 have studied the distinct features of SVs originating from different mechanisms[25,26]. For example, non-allelic homologous recombination (NAHR), is associated with active enhancers and an open chromatin environment. Our analysis also showed that microinsertions, flanking non-homologous breakpoints, originate from late-replicating genome loci with characteristic distances from breakpoints. These results inform us on the molecular mechanisms underlying SV formation and also indicate differences in functional impacts of different SV types.

We further performed SV mechanism annotations for the 1000GP Phase 3 deletions using BreakSeq[27], categorizing 29,774 deletions by their creation mechanisms. Among these, NHR proved to be the most prevalent mechanism (~73% of all categorized deletions)[17].

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 mutations according to synonymous, missense, nonsense or splice-site-disrupting changes[28]. We annotated variants from 1,092 humans in Phase 1 of the 1000GP[25] and observed that genes tolerant of loss-of-function (LoF) mutations are under the weakest selection and cancer-causal genes under the strongest 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[29]. 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[30].

To better understand nc regulatory regions, we developed tools to analyze ChIP-Seq data to identify genomic elements and interpret their regulatory potential. PeakSeq and MUSIC identify regions bound by TFs and chemically modified histones[31,32]. PeakSeq has been widely used in consortium projects such as ENCODE[31,33]. MUSIC 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[32]. 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[34-37]. In order 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[30 37,38]. We used metrics, such as diversity and fraction of rare variants, to characterize selection pressure on various classes and subclasses of functional annotations[30]. We have also defined variants that are disruptive to a TF-binding motif in a regulatory region[33].

Tools for helping annotate functional impact based on network and allelic expression analyses. We found that functionally significant and highly conserved genes tend to be more central in

various biological networks[39] and are positioned at the top of regulatory networks[38]. Further studies showed relationships between selection and protein network topology (e.g., quantifying selection in hubs relative to proteins on the network periphery[39,40]). Incorporating multiple network and evolutionary properties, we developed NetSNP[39] 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 analyzing network hierarchy[38]. To quantify the degree of hierarchy for a given hierarchical network, we defined a metric called hierarchical score maximization (HSM[41]). Finally, we also developed AlleleSeq[42] for the detection of candidate variants associated with allele-specific binding and allele-specific expression. These tools are based on the construction of a personal diploid genome sequence (and corresponding personalized gene annotation) using genomic sequence variants and can be used to prioritize variations disrupting allelic activity.

FunSeq: Tools for integrated functional prioritization. We recently developed a prioritization pipeline called FunSeq[25,43] that identifies annotations under strong selective pressure as determined using genomes from many individuals from diverse populations. FunSeq links each nc single-nucleotide mutation 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). We further enhanced FunSeq (FunSeq2) and identified ~100 nc candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer

samples[25].

Tools for identifying enrichment of variations in coding and non-coding regions. We have worked on statistical methods for analysis of nc regulatory regions. LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations) identifies significant mutation enrichments in nc elements by comparing observed mutation counts with expected counts under a whole genome background mutation model. LARVA also includes corrections for biases in mutation rate owing to DNA replication timing. For coding region analysis, we developed MuSiC[44] to analyze genetic changes using standardized sequence-based inputs, along with multiple types of clinical data, to establish correlations among variants, affected genes and pathways, and to ultimately separate commonly abundant passenger events from truly significant events.

REFRENCES.

. 17 Sudmant, P. H. An integrated map of structural variation in 2,504 human genomes. *Nature* **Accepted, in print** (2015).

. 25 Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587, doi:10.1126/science.1235587 (2013).

. 26 Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat Commun* **6**, 7256, doi:10.1038/ncomms8256 (2015).

. 27 Lam, H. Y. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature biotechnology* **28**, 47-55, doi:10.1038/nbt.1600 (2010).

. 28 Habegger, L. *et al.* VAT: a computational framework to functionally annotate variants in personal

genomes within a cloud-computing environment. *Bioinformatics* **28**, 2267-2269, doi:10.1093/bioinformatics/bts368 (2012).

. 29 Lu, Z. J. *et al.* Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome research* **21**, 276-285, doi:10.1101/gr.110189.110 (2011).

. 30 Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. & Gerstein, M. B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* **39**, 7058-7076, doi:10.1093/nar/gkr342 (2011).

. 31 Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology* **27**, 66-75, doi:10.1038/nbt.1518 (2009).

. 32 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* **15**, 474, doi:10.1186/s13059-014-0474-3 (2014).

. 33 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

. 34 Cheng, C. *et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* **12**, R15, doi:10.1186/gb-2011-12-2-r15 (2011).

. 35 Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome research* **22**, 1658-1667, doi:10.1101/gr.136838.111 (2012).

. 36 Gerstein, M. B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445-448, doi:10.1038/nature13424 (2014).

. 37 Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**, R48, doi:10.1186/gb-2012-13-9-r48 (2012).

. 38 Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).

. 39 Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* **9**, e1002886, doi:10.1371/journal.pcbi.1002886 (2013).

. 40 Kim, P. M., Korbel, J. O. & Gerstein, M. B. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* **104**, 20274-20279, doi:10.1073/pnas.0710183104 (2007).

. 41 Cheng, C. *et al.* An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. *Genome Biol* **16**, 63, doi:10.1186/s13059-015- 0624-2 (2015).

. 42 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522, doi:10.1038/msb.2011.54 (2011).

. 43 Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480, doi:10.1186/s13059-014-0480-5 (2014).

. 44 Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome research* **22**, 1589-1598, doi:10.1101/gr.134635.111 (2012).