**Experience in genome annotation**

We have substantial experience in genome annotation, which we can leverage in interpreting functional genomics data and relating it to the reference. In collaboration with the UCSC and Sanger groups, we have developed a variety of methods to identify pseudogenes[1-3]. Also, as a member of the GENCODE project, we used the pipelines to identify pseudogenes in human, mouse, worm, fly, and other model organisms. Moreover, using data from the 1000 Genomes Project in addition to the pseudogene annotation resulting from our pipelines, we were able to study the impact of pseudogene in human population variation[4]. We were also able to associate their origin mechanism to cell division [5].

**Experience in addressing questions related to the reference genome definition**

The first wave of personal genomes documents how no single individual genome contains the full complement of functional genes. We have previously described the extent of variation in gene numbers between individuals arising from inactivation events such as premature termination or aberrant splicing due to single-nucleotide polymorphisms[6]. This highlights the inadequacy of the current reference sequence and gene set. We also presented and discussed extensively five models for selecting the reference allele: (1) the allele present in the current reference, (2) the allele present in a chosen high-quality individual genome, (3) the most common human allele, (4) the ancestral allele, and (5) the maximum coding potential allele. We proposed to define a reference gene set that will remain stable as more individuals are sequenced. In particular, we recommended that the ancestral allele be used to define the reference sequence from which a core human reference gene annotation set can be derived. Our discussion and conclusions were based on a survey of the sequence variations in 21 published personal genomes and exomes for nonsense SNPs and SNPs in canonical splice sites using a uniform gene annotation[6].

**Experience in large scale variant calling**. This experience is through our membership in the 1000 Genomes Project, particularly from our participation in the analysis working group and the structural variant and functional interpretation subgroups of the consortium, where the majority of the variant calling tools were developed, deployed and interpreted [7-9].

****Structural variations are important contributors to human polymorphism, have great functional impact and are often implicated in diseases including cancer. We have developed a number of SV calling algorithms, including BreakSeq which compares raw reads with a breakpoint library (junction mapping)[10], CNVnator (Fig 1) which measures read depth[11], AGE which refines local alignment[12], and PEMer which uses paired end reads[13]. We have also developed array-based approaches[14] and a sequencing-based Bayesian model[15]. Furthermore, we have studied the distinct features of SVs that originate from different mechanisms and showed how creation processes may have potentially divergent functional impacts[16, 17].

**Fig 1. Read depth-based identification of copy number variation by CNVnator.**

**Experience in building personal genomes and performing allelic calculations**

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 this haploid human reference sequence since 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.

**Fig. 2.** Personal genome construction. Each haplotype in the diploid personal genome is derived by incorporating phased variants (SNVs, indels and SVs) into the human reference genome. The coordinates can be mapped back to the human reference coordinates to facilitate comparisons with other reference-based resources, such as gene annotations.

For personal genome construction, we have developed a computational tool, vcf2diploid[18]. The tool integrates an individual’s genomic variation data (SNVs, indels, and SVs) into the reference genome (Fig. 2). Phase information of heterozygous variants is also incorporated, producing maternal and paternal haplotypes. 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 mapping of genomic annotated regions (e. g. gene or peak coordinates for RNA-seq and ChIP-seq, respectively) between the genomes using available tools, such as the UCSC LiftOver tool[19].

We have previously constructed the personal diploid genome, splice-junction libraries and personalized gene annotations for NA12878. We have made this assembly available as a resource at alleleseq.gersteinlab.org and have been updating it as new versions of the human reference genome, genomic annotations, and NA12878 genetic 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[20] where we built 382 personal genomes using the variant call sets from the 1000 Genomes Project.

Allele-specific analyses are particularly sensitive to these biases. For this purpose, the initial step of our AlleleSeq pipeline[18] involves construction of the personal diploid genome. We have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project[16, 21, 22]. We annotated variants associated with allele-specific expression (ASE) and binding (ASB) in a large pool of individuals from the 1000 Genomes Project. For this analysis[20], we integrated matching functional datasets (955 RNA-seq and 165 ChIP-seq in total), which include ChIP-seq datasets from 14 lymphoblastoid cell lines in ENCODE[23]. Overall, we detected more than 6K and 63K SNVs associated with ASB and ASE, respectively. These results were made available as an online resource, AlleleDB (alleledb.gersteinlab.org). Most recently, we constructed 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 donors[24].

1. Zhang, Z., et al., *PseudoPipe: an automated pseudogene identification pipeline.* Bioinformatics, 2006. **22**(12): p. 1437-9.

2. Zheng, D. and M.B. Gerstein, *A computational approach for identifying pseudogenes in the ENCODE regions.* Genome Biol, 2006. **7 Suppl 1**: p. S13.1-10.

3. Navarro, F.C. and P.A. Galante, *RCPedia: a database of retrocopied genes.* Bioinformatics, 2013. **29**(9): p. 1235-7.

4. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.

5. Abyzov, A., et al., *Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division.* Genome Res, 2013. **23**(12): p. 2042-52.

6. Balasubramanian, S., et al., *Gene inactivation and its implications for annotation in the era of personal genomics.* Genes Dev, 2011. **25**(1): p. 1-10.

7. Mills, R.E., et al., *Mapping copy number variation by population-scale genome sequencing.* Nature, 2011. **470**(7332): p. 59-65.

8. Abecasis, G.R., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

9. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

10. Lam, H.Y., et al., *Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library.* Nat Biotechnol, 2010. **28**(1): p. 47-55.

11. Abyzov, A., et al., *CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing.* Genome Res, 2011. **21**(6): p. 974-84.

12. Abyzov, A. and M. Gerstein, *AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision.* Bioinformatics, 2011. **27**(5): p. 595-603.

13. Korbel, J.O., et al., *PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data.* Genome Biol, 2009. **10**(2): p. R23.

14. Wang, L.Y., et al., *MSB: a mean-shift-based approach for the analysis of structural variation in the genome.* Genome Res, 2009. **19**(1): p. 106-17.

15. Zhang, Z.D. and M.B. Gerstein, *Detection of copy number variation from array intensity and sequencing read depth using a stepwise Bayesian model.* BMC Bioinformatics, 2010. **11**: p. 539.

16. Khurana, E., et al., *Integrative annotation of variants from 1092 humans: application to cancer genomics.* Science, 2013. **342**(6154): p. 1235587.

17. Abyzov, A., et al., *Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms.* Nat Commun, 2015. **6**: p. 7256.

18. Rozowsky, J., et al., *AlleleSeq: analysis of allele-specific expression and binding in a network framework.* Mol Syst Biol, 2011. **7**: p. 522.

19. Hinrichs, A.S., et al., *The UCSC Genome Browser Database: update 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D590-8.

20. Chen, J., et al., *A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals.* Nat Commun, 2016. **7**: p. 11101.

21. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

22. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.

23. *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

24. Onuchic, V., et al., *Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci.* Science, 2018. **361**(6409).