Methods for producing somatic Single Nucleotide Variant (sSNV) calls in tumors with Whole Genome Sequence (WGS) data

To identify noncoding single-nucleotide variants and predict which were likely to be functional, a high-confidence set of whole genome sSNV calls were generated.

Alignment. HiSeq paired-end reads were aligned to the hg19 human reference genome using bwa-mem, an implementation of BWA v0.7.3 (1) that permits gapped alignments. Output sam files were converted to bam, sorted, and indexed using samtools v0.1.17 (2). MarkDuplicates, part of Picard Tools v1.51 (3), was used to remove duplicate reads generated during the PCR amplification stage. Duplicate removal identifies all reads that have identical 5' coordinates and keeps only the read pair with the highest base quality sums. After duplicate removal, fine-tuning of the alignment was performed using GATK v2.1 (4) as outlined in (5) and summarized here: Local positions to target for realignment were called using RealignerTargetCreator and then realigned using IndelRealigner. Quality scores were then recalibrated using BaseRecalibrator and PrintReads, which bins reads based on the original quality score, the dinucleotide, and the position in the read.

Variant calling and filtering. After creating high-quality alignments for each tumor and normal sample, somatic single-nucleotide were called by comparing the tumor samples to their matched normal. Somatic single-nucleotide variants (sSNVs) were called using MuTect v1.1.5 (6) and Strelka v1.0.13 (7).

MuTect has high sensitivity and calls many variants even in regions of lower coverage. To reduce false positives, we performed three filtering steps: sample-level, dataset-level, and caller-level. In the sample-level filtering, which considered each sample independently, called sSNVs were discarded if they had fewer than 14 reads in the tumor, fewer than 10 reads in the normal, less than 10% variant reads in the tumor, or more than 2% variant reads in the normal. They were also discarded if they were suspected to be a single-nucleotide polymorphism (SNP). Our in-house SNP database includes all SNPs in dbSNP v134 (8) and those found by the NHLBI Exome Sequencing Project (downloaded 17Dec2012)(9,10). with the exception of the cancer-related variants found in COSMIC v60 (11).

The dataset-level filtering step took into consideration variants called across samples. By comparing calls across samples, we can identify and discard variants that are likely to be high sequencing error sites or common germline variants not found in the SNP database. Candidate sSNVs were discarded if insertions or deletions in the region prevented them from being reliably quantified in the majority of samples. They were also discarded if reads matching the variant were seen in greater than 10% of reads in a sample from another patient but were not called by MuTect. Lastly, to remove variants that had a low read frequency in many samples, the Binomial distribution was used to determine if the number of reads matching a called variant exceeded the background rate, which was estimated using the proportion of reads matching that variant in samples from the other patients. The variant was discarded if the Binomial p-value exceeded 1e-8. The combination of these dataset-level filtering steps is highly effective at removing false positive whole genome sequence sSNV calls while retaining true positives. We determined this two ways by comparing pre- and post-filtering MuTect whole genome calls to: 1. The higher-coverage higher-confidence whole exome calls from the same samples, and 2. The pre- and post-filtering whole genome calls using a secondary sSNV caller, Strelka. The same sample- and dataset-level filtering steps were applied to both the MuTect and Strelka calls. Finally, to produce the highest confidence sSNV call set for analysis of function, only the calls made by both the MuTect and Strelka pipelines were kept.

[[JAS]]
Finally, to rank these variants in terms of patiently functional impact we used the FunSeq2 algorithm (1) based on an earlier version (2). Breifly, we use an enrichment of rare nonsynonymous SNPs (derived allele frequency (DAV) < 0.5%) as a proxy for the existence of purifying selection. Regions with comparable fractions of rare variants are labeled sensitive (DAV=68.8%) and ultrasensitive (65.7%).

References

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–1760.

2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–2079.

3. Picard Tools [Internet]. Available from: http://picard.sourceforge.net/

4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–1303.

5. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 2011 May;43(5):491–8.

6. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 2013 Feb 10;31(3):213–219.

7. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28:1811–1817.

8. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308–311.

9. Exome Variant Server, (ESP), NHLBI GO Exome Sequencing Project [Internet]. Seattle, WA. Available from: http://evs.gs.washington.edu/EVS/

10. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson D a, Bamshad MJ, Akey JM. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493:216–20.

11. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA. COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2011;39.

12. Fu et al. Genome Biology 2014, 15:480

13. Khurana et el. Science. 2013 Oct 4;342(6154):1235587