SUPPLEMENTARY METHODS

Somatic Variant Detection

Sequence data was aligned to reference sequence build GRCh37-lite-build37 using bwa mem[5] version 0.7.10 (params: -t 8::), then merged and deduplicated using picard version 1.113, (https://broadinstitute.github.io/picard/).

SNVs were detected using the union of four callers: 1) samtools[6] version r982 (params: mpileup -BuDs) intersected with Somatic Sniper[4] version 1.0.4 (params: -F vcf -G -L -q 1 -Q 15) and processed through false-positive filter v1 (params: --bam-readcount-version 0.4 --bam-readcount-min-base-quality 15 --min-mapping-quality 40 --min-somatic-score 40), 2) VarScan[3] version 2.3.6 filtered by varscan-high-confidence filter version v1 and processed through false-positive filter v1 (params: --bam-readcount-version 0.4 --bam-readcount-min-base-quality 15), 3) Strelka[10] version 1.0.11 (params: isSkipDepthFilters = 1), and 4) mutect[2] v1.1.4 (params: number-of-chunks=50).

Indels were detected using the union of 4 callers: 1) GATK[7] somatic-indel version 5336 2) pindel[11] version 0.5 filtered with pindel somatic calls and VAF filters (params: --variant-freq-cutoff=0.08), and pindel read support, 3) VarScan[3] version 2.3.6 filtered by varscan-high-confidence-indel version v1 and 4) Strelka[10] version 1.0.11 (params: isSkipDepthFilters = 1).

SNVs and Indels were further filtered by removing artifacts found in a panel of 905 normal exomes, removing sites that exceeded 0.1% frequency in the 1000 genomes or NHLBI exome sequencing projects, and then using a Bayesian classifier (https://github.com/genome/genome/blob/master/lib/perl/Genome/Model/Tools/Validation/IdentifyOutliers.pm) and retaining variants classified as somatic with a binomial log-likelihood of at least 10.

Copy number analysis was performed using Varscan 2.3.6 [24], and segmented with the DNAcopy package[8]. Data was recentered when necessary, then segments of less than 50 probes were filtered to remove noise, followed by merging of adjacent segments with absolute CN difference of less than 0.2.

Mutation signature

The mutational spectra of mutations of all samples were analyzed using deconstructSigs[9] to extract signatures based on the Wellcome Trust Sanger Institute Mutational Signature Framework[1] and statistically quantify the contribution of each signature for each tumor.

REFERENCES

1 Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale ALet al (2013) Signatures of mutational processes in human cancer. Nature 500: 415-421 Doi 10.1038/nature12477

2 Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31: 213-219 Doi 10.1038/nbt.2514

3 Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568-576 Doi 10.1101/gr.129684.111

4 Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28: 311-317 Doi 10.1093/bioinformatics/btr665

5 Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bioGN]:

6 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079 Doi 10.1093/bioinformatics/btp352

7 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly Met al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303 Doi 10.1101/gr.107524.110

8 Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5: 557-572 Doi 10.1093/biostatistics/kxh008

9 Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C (2016) DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol 17: 31 Doi 10.1186/s13059-016-0893-4

10 Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28: 1811-1817 Doi 10.1093/bioinformatics/bts271

11 Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865-2871 Doi 10.1093/bioinformatics/btp394