Supplementary Methods

1. Sequence data analysis

The Whole Exome Sequencing fastqs were aligned to GRCh38 reference with BWA_kit. (0.7.1) [1] Samtools (1.12) was used to sort and index bam files and Genome Analysis Toolkit. (4.2.0.0) [2] Markduplicates was used to remove PCR duplicate reads. GATK BaseRecalibrator and applyBQSR were used to recalibrate base quality scores. GATK (3.8.1.0) DepthofCoverage and GATK (4.2.0.0) Flagstat were used to check quality. Somatic SNPs and INDELs were called with GATK Mutect2 using gnomAD [3] project population allele frequency information for filtering of likely germline variations since our samples did not have matched normals. GATK LearnReadOrientationModel was used for filtering of read orientation. GATK GetPileupSummaries and CalculateContamination were used to summarize and calculate the fraction of reads coming from cross-sample contamination. Then GATK FilterMutectCalls was performed to filter outputs from Mutect2 calling with tumor segmentation file from GetPileupSummaries and contamination table from CalculateContamination on minimum allele fraction 0.001 and minimum reads per strand 1.

Germline SNPs and indels were called with GATK HaplotypeCaller using for sites with East Asian population Variant Allele Frequency higher than 0.01 from in the 1000 Genomes Project [4]. GATK GenotypeGVCFs were performed for genotyping and the genotyping quality of less than 20 were filtered out using GATK VariantFiltration. Then GATK SelectVariant was used to remove unused alternates, exclude non-variants and filtered genotypes.

CNVkit (0.9.8) was performed to estimate copy number using flat reference because the samples did not have matched normal [5]. PureCN (2.0.2) was then performed to estimate tumor ploidy, copy number, and loss of heterozygosity and to classify SNVs by somatic status and clonality [6]. Manually curated best matching model was selected. Amplified pathogenic driver genes were obtained from filtering OncoKB Cancer Gene List annotated ‘Oncogene’ with copy number higher than 6. Genes gaining pathogenicity from deletion were selected from filtering OncoKB Cancer Gene List annotation ‘Tumor suppressor gene’ with copy number 0.

RNA sequencing data was processed following the TOPMED v9 pipeline. Briefly, STAR (2.6.1d) [7] was used to align reads to the human reference genome (GRCh38), and Picard (2.18.17) MarkDuplicates was used to remove PCR duplicate reads. RNA-SeQC (2.3.3) [8] was used to measure quality metrics and to calculate gene level read counts. RNA fusion transcripts were detected using STARfusion (1.9.0) [9] and FusionInspector (2.6.0).

All mutations were annotated with OncoKB-Annotator. For SNP and INDEL, mutations with the OncoKB-Annotator ‘Likely Oncogenic’ or ‘Oncogenic’ annotation and Mutect2 QC filter ‘PASS’ were assigned as driver mutations [10]. For copy number aberration and RNA fusion transcripts, aberrations with OncoKB-Annotator filtration ‘Likely Oncogenic’ or ‘Oncogenic’ were selected.

References

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754-60.
2. van der Auwera G, O'Connor BD. Genomics in the Cloud: using Docker, GATK, and WDL in Terra. Sebastopol, CA: O'Reilly Media, 2020.
3. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434-43.
4. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68-74.
5. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput Biol. 2016;12:e1004873.
6. Riester M, Singh AP, Brannon AR, Yu K, Campbell CD, Chiang DY, et al. PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med. 2016;11:13.
7. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.
8. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28:1530-2.
9. Haas B, Dobin A, Stransky N, Li B, Yang X, Tickle T, et al. STAR-Fusion: fast and accurate fusion transcript detection from RNA-Seq. Preprint at: https://doi.org/10.1101/120295 (2017).
10. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017:2017:PO.17.00011.