Mayo Clinic Brain Tumor Patient-Derived Xenograft National Resource

Whole-Exome Sequencing (WES)

1.1 WES data generation for brain tumor PDX models

Whole-exome sequencing was performed in Mayo Clinic's Medical Genome Facility. Paired-end libraries were prepared by using the manufacturer's protocol, and exome capture was carried out with SureSelect Human All Exon V5+UTRs (or V4+UTRs) kit from Agilent (Santa Clara, CA), or TGen's Strexome platform (https://tgen.org/).

The libraries were sequenced as 2X101 paired-end reads on the Illumina HiSeq 2000 platform according to the manufacturer's recommendation (Illumina Inc., San Diego, CA). Xenograft-derived raw sequencing reads were mapped to human (hg19) and mouse reference (mm10) simultaneously to remove potential reads from contaminating mouse cells with Xenome (version 1.0.1) before variant calling [1].

1.2 Identification of somatic alternations in PDXs

GenomeGPS is an internal comprehensive secondary analysis pipeline for WES data in the Mayo Clinic Bioinformatics Core. The pipeline integrates published variant detection methods for both germline and somatic variant calling.

In detail, FASTQ files were aligned to the hg19 reference genome using Novoalign (version 3.02.04) with the following options: -x 5 -i PE 425,80 -r Random --hdrhd off -v 120 (http://www.novocraft.com/).

Realignment and recalibration were then performed using GATK (version 3.3.0) by following the recommended Best Practices version 3 [2]. Variant calling (point mutation and small indels) from xenograft samples without matched blood were performed with GATK's HaplotypeCaller (version 3.3.0) [3]. For those PDXs with matched blood, somatic mutations were collectively called by SomaticSniper (version 1.0.4) ([-q 20 -Q 20 -F vcf) [4], JointSNVMix2 (version 0.8b2) [5] with the "--model snvmix2" parameter, or MuTect (version 1.1.4) (default parameters)[6], and somatic Indels by SomaticIndelDetector (GATK, version 1.6.9) (--window_size 1000)[3].

Variants were annotated using GATK variantannotator (v 3.3.0) for variant quality[2], BioR (version 4.1.2, Biological Annotation Data Repository), which consolidates the biological and clinical annotations (1000 Genome, ExAC and COSMIC databases) needed to interpret variants[7], and CAVA (vrersion1.2.0, Clinical Annotation of Variants), which stratifies variants into categories according to predicted severity of impact on protein function[8].

Raw variant sites were subject to a series of quality filtering, such as the allelic and overall depth of coverage, average mapping quality, base quality, proximity to homopolymer run, number of mapping-quality-zero reads, variant quality, strand bias and somatic score. Finally, common

variants were eliminated based on the minor allele frequencies (>0.01) available in the 1000 Genomes Project or EXAC.

References

1. Conway T, Wazny J, Bromage A, Tymms M, Sooraj D, et al. (2012) Xenome--a tool for classifying reads from xenograft samples. Bioinformatics 28: i172-i178 doi:10.1093/bioinformatics/bts236.

2. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline Van der Auwera GA, Carneiro M, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M, 2013 CURRENT PROTOCOLS IN BIOINFORMATICS 43:11.10.1-11.10.33

3. A framework for variation discovery and genotyping using next-generation DNA sequencing data DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M, 2011 NATURE GENETICS 43:491-498

4. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28: 311-317. doi: 10.1093/bioinformatics/btr665. pmid:22155872

5. Roth, A. et al. (2012) JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics, 28, 907-913.

6. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. NatBiotechnol. 2013; 31: 213-219. doi: 10.1038/nbt.2514.

7. Jean-Pierre A. Kocher, Daniel J. Quest, Patrick Duffy, Michael A. Meiners, Raymond M. Moore, David Rider, Asif Hossain, Steven N. Hart, Valentin Dinu; The Biological Reference Repository (BioR): a rapid and flexible system for genomics annotation. Bioinformatics 2014; 30 (13): 1920-1922. doi: 10.1093/bioinformatics/btu137

8. Márton Münz, Elise Ruark, Anthony Renwick, Emma Ramsay, Matthew Clarke, Shazia Mahamdallie, Victoria Cloke, Sheila Seal, Ann Strydom, Gerton Lunter, Nazneen Rahman. CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting. Genome Medicine 7:76, doi:10.1186/s13073-015-0195-6 (2015).