Updated on 02/07/2019

Detection of somatic mutations in paired tumor/normal data

Raw sequence data reads (WES/WGS) of the tumor or matching-normal samples will be aligned to the human reference genome (ex. hg19 or hg38) using the Burrows–Wheeler Aligner (BWA). After alignment, and deduplication of reads with Picard software, we sort and index the binary version of Sequence Alignment Map (SAM) file. We then use Genome Analysis Toolkit (GATK) v4.0 to perform left alignment of small insertions and deletions (indels), indel realignment, and base quality score recalibration. Somatic variant calling will be performed on matched tumor-normal pairs using MuTect2 (REF. 1).

With **matching tumor-normal data**, true somatic variants will be identified through comparison between tumor and matched-normal. MuTect2 can run on unmatched tumors but this produces high rates of false positives. We further filter variants using population-based databases (ex. 1000 Genomes, ExAc and gnomAD). Other false positive mutations, induced by the systemic artifacts such as due to Formalin-fixed paraffin-embedded (FFPE) or oxidative DNA damage, will also be removed (REF.2).

We use the following criteria to remove possible false positive mutations,

1. If a variant supports by one or more reads in matched normal sample.
2. If the read depth of a variant position is < 50, or the variant is supported by less than 10 reads.
3. C>T / G>A variants with a frequency less than 0.1 (possible FFPE artifacts).
4. C>A / G>T variants with a frequency less than 0.1 (possible artifactual mutations due to oxidative DNA damage during sample preparation).
5. If indels are supported only by forward or reverse reads.

For **tumor-only samples**, somatic variation will be assessed using population-based databases (ex. 1000 Genomes, ExAc and gnomAD). If a variant exists at a very low (e.g. <0.5%) or zero population frequency within those databases, it increases the likelihood that the variant is somatic. However, de novo and less common germline polymorphisms be reliably removed only by observation of a matched normal sample (REF. 3). Subsequent manual review of aligned read sequences (SAM/BAM file) using Integrative Genomics Viewer (IGV) will be performed to identify a high quality list of somatic variants.

Different mutational processes generate unique combinations of mutation types (mutational signatures). Each of the mutational signature patterns is associated with each cancer etiology in a tissue-specific manner. The proportion of known mutational signatures in the data set will be identified using COSMIC (Catalogue Of Somatic Mutations In Cancer) database and the R package *DeConstructSigs* (REF 4,5).

REQUIREMENTS:

Updated on 02/07/2019

1. Raw data files (fastq)
2. Metadata spreadsheet
3. Organism (ex. Human)

DELIVERABLES:

1. List of somatic mutations identified in each sample/patient.
2. Details of analysis workflow for your writing (manuscript)

REFERENCES:

1. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–303.
2. Do Hongdo, Dobrovic Alexander. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. Clin Chem. 2015; 61:64–71.
3. Barnell, E. K. et al. Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. Genet. Med. (2018).
4. Devecchi, Andrea, et al. The Genomics of Desmoplastic Small Round Cell Tumor Reveals the Deregulation of Genes Related to DNA Damage Response, Epithelial–Mesenchymal Transition, and Immune Response, Cancer Commun (Lond), 2018.
5. S.A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res, 45 (2017), pp. D777-D783.

Questions? Comments? If you have additional requirements or questions, please feel free to contact us at EICC@emory.edu