UOC-BC: mutation - Illumina Nextera Rapid Capture Exome, WES

Molecular Methods Description:

The PDTX samples were characterized at several passages using Whole Exome Sequencing (WES) for single nucleotide variations (SNVs). he sequencing was performed using 125bp paired-end reads

Analysis Description:

Data output is from the Caldas Laboratory at the University of Cambridge . Full details can be found in the paper Bruna et al., Cell 167, 260 - 274 (2016) .

The sequencing was performed using 125bp paired-end reads. Short reads were aligned using novoalign (Novocraft) with our custom pipeline to remove mouse contamination. Bam files were merged, sorted and indexed using samtools. Duplicates were marked using Picard tools and insertions and deletions (indels) were realigned using GATK.

For quality control purposes and to check for sample labeling mistakes, all samples were genotyped using GATK HaplotypeCaller and a few errors were identified and corrected. HaplotypeCaller was also employed for variant calling, and after that several filters were applied using the Bioconductor package VariantAnnotation: for single nucleotide variants (SNVs), a minimum genotyping quality of 20, at least 5 reads at the variant position, a strand bias Phred-scale p value smaller than 40 and no presence of homopolymers in the surrounding region. For indels, we increased the width of the region to detect nearby homopolymers. Genotypes and variant allele frequencies (VAFs) were computed from these calls.

All variants were annotated using annovar version March 2015 for gene/exon annotation, 1000 genomes version Oct 2014, dbSNP version138, repetitive regions genomicSuperDups database, SIFT, Polyphen 2, MutationTaster and MetaLR, all versions ljb26.

Variants in intergenic, intronic or ncRNA intronic positions were discarded.

In order to quantify variability in matched tumor/PDTX, different passages of PDTXs or matched PDTX/PDTC, all variants detected in at least one sample of each model were obtained. For those samples where those variants had not been detected GATK HaplotypeCaller was run again on those positions to see if this was a consequence of no reads in that region for that sample or a real absence of the variant.

Normal contamination estimates were obtained for each tumor sample combining copy number calls from shallow sequencing and SNV calls. First, we selected heterozygous SNPs in the matched normal sample (if available, otherwise all heterozygous variants in the tumor were considered), and then looked at only those in regions of copy number loss in the tumor, as defined by a segmented mean copy number log ratio smaller than –0.1. In these regions, as a loss of heterozygosity has occurred, it is expected that all variants will have a VAF of 0 or 1 (if no normal contamination is present). For the AB SNPs that are left to B genotype, the expected value of contamination would be 2 ^*^ VAF – 1, assuming that all tumor cells acquired the deletion. As this is a downward estimate, we chose as the tumor content estimate the maximum of the density function of the VAF of those variants. These estimates were used to correct tumor VAFs or copy numbers where needed in the rest of the analyses.

All variants that were present in the 1000 Genomes database or in any of our normal samples were labeled as germline. Regions marked as repetitive were also filtered, and insertions that represented a segmental duplication were removed if they were not present in at least three-fourths of all the samples for a given model or in 3 of them. Somatic variants that were not filtered were compiled for each model. Some manual curation was needed for genes like PI3KCA, where variants from a region of segmental duplication were included after manual inspection.