Colorectal tumoroid project - Example

RP Cardenas

04/08/2020

1 Overview

This document describes the analysis of colorectal cancer samples that were grown in 3D culture (tumouroid) that were either compared to colon-crypt samples derived straight from patients (crypt), or from colorectal crypt samples that were also grown in 3D culture (organoid). The aim for this analysis was to investigate mutations across samples using exome-sequencing and cutting-edge bioinformatics tools.

Samples were subjected to rigourous QC and then analysed for high confidence mutations using cgpWXS (CAVeMAN and Pindel) and GATK-mutect2 variant callers. The cancer genome interpeter was used to annotate variants, this allowed the identification of a number of driver (TIER1/2) mutations. In addition, these annotated driver mutations showed a high variant allele frequency in the sample indicating that these mutation are indeed being selected for within the clone and therefore are likely to be facilitating tumorigenicity.

# 2 Methods

2.1 QC and Alignment

FASTQ files were first trimmed for TruSeq adapter sequences (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT, GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG) and FastQC

analysis (Figure 1). Samples were then aligned using the docker image cgpMAP (https://github.com/cancerit/dockstore-cgpmap (https://github.com/cancerit/dockstore-cgpmap)) which utilises bwa-mem - genome GRCh37d5.

BAM files were sorted using SAMtools and then analysed using Picard tools for QC. QC analysis included remove duplicates erroneously generated through PCR and sequencing (optical duplicates). Exome-seq capture of targets were analysed using HS.metrics. A summary of the tools used is shown below:
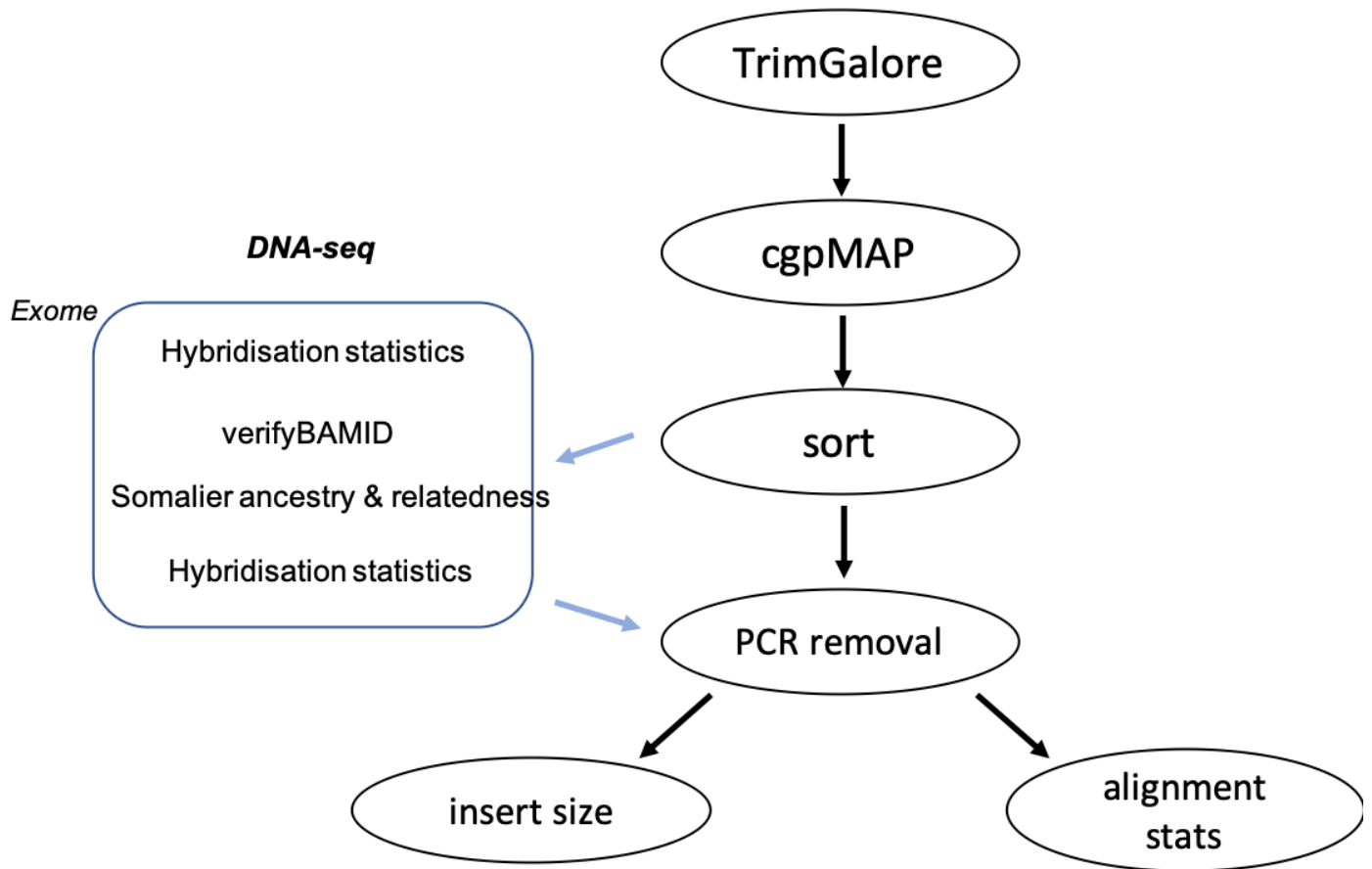

2.3 Variant analysis

Sanger cgpWXS and GATK-mutect2 were both indepdently run on the samples. The docker container cgpWXS (https://github.com/cancerit/dockstore-cgpwxs (https://github.com/cancerit/dockstore-cgpwxs)) utilises CAVeMAN and Pindel to identify single base substitutions (SBS) and indels, respectively. An in-house custom pipeline was developed to automate the reccomennded tools from the GATK mutect2 workflow (https://gatk.broadinstitute.org/hc/en-us/articles/360035531132--How-to-Call-somatic-mutations-using-GATK4-Mutect2 (https://gatk.broadinstitute.org/hc/en-us/articles/360035531132--How-to-Call-somatic-mutations-using-GATK4-Mutect2))

VCF files from both analyses were overlapped for SBS and indels, using bcftools and bedtools respectively, giving a consensus of variants identified using both methods. VCF files were annotated using the cancer genome interpreter tool (CGI - https://www.cancergenomeinterpreter.org/home (https://www.cancergenomeinterpreter.org/home)).

3 Results

## Mutect2 and cgpWXS overlap

Below shows the overlap between the two variant callers used.
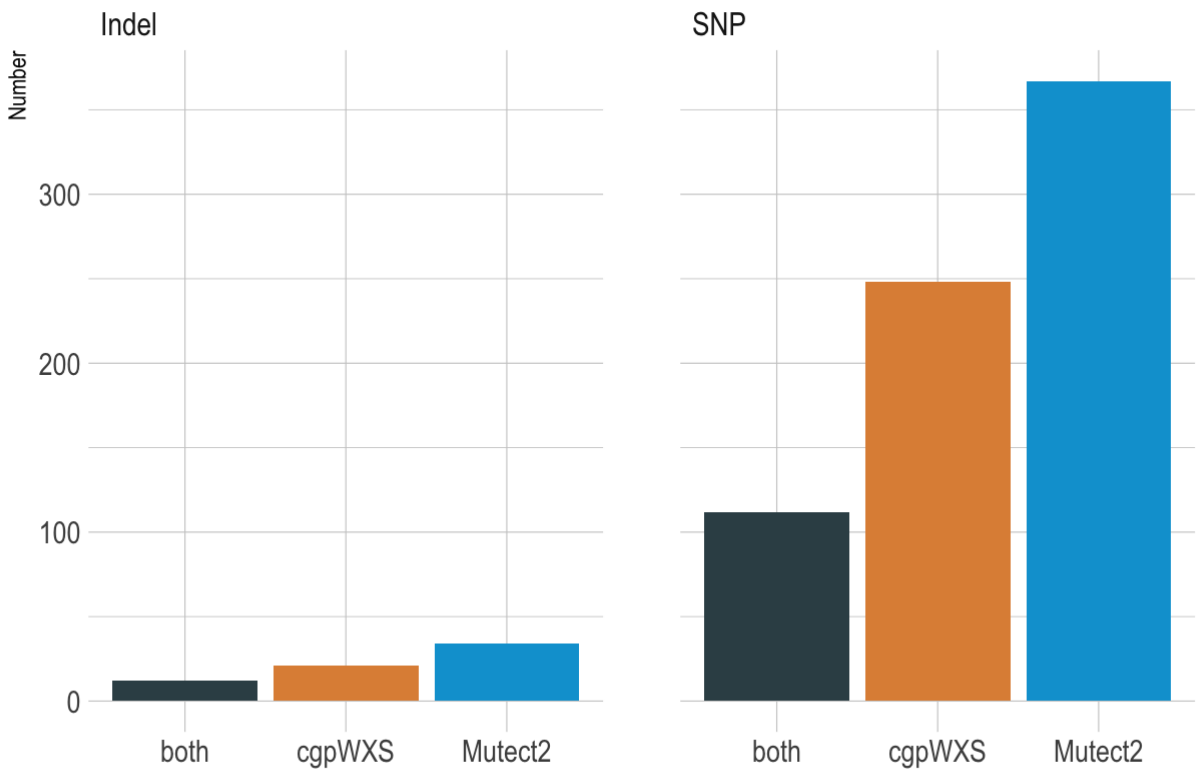
Cancer Genome Interpreter

The cancer genome interpreter is a tool that enables the annotation of variants, giving predictions of deleteriousness, prediction of tumour drivers and a list of published drugs that may be effective for samples.

Below shows variant types across all samples that have also been annotated based on known cancer driver genes (variant types tab); tier 1 and 2 represent higher and lower level of stringency of the driver prediction, respectively. Data is accessible using the CGI datatable tab - export of table is supported by selection of CSV, excel or pdf buttons. For TIER1/2 driver genes a table of published drugs known to be effective against varianst are given (CGI drug prediction tab).

| Variant types | CGI Datatable | CGI drug prediction |
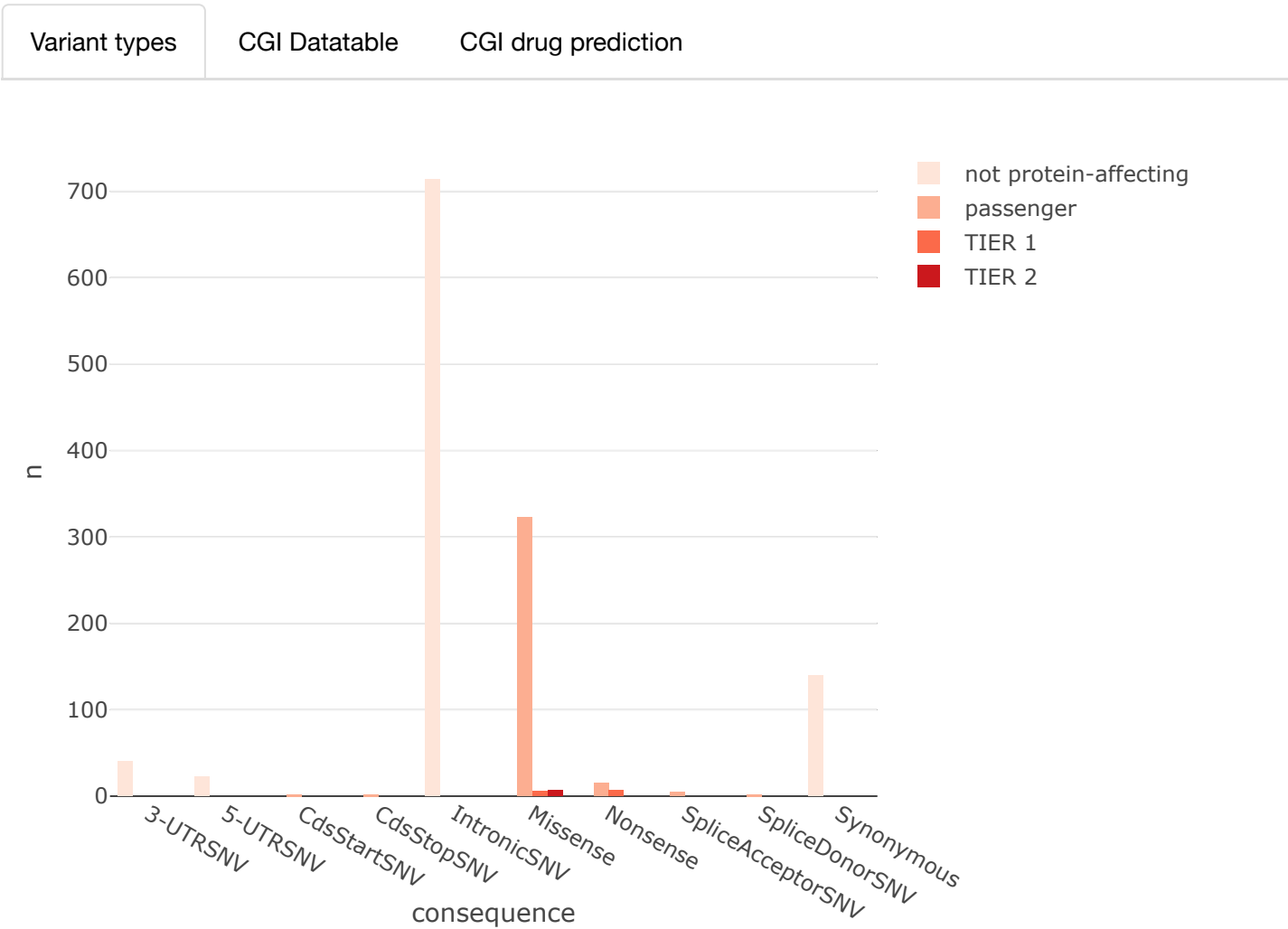
Variant allele frequency

Variant (mutant) allele frequency, VAF, are used commonly in cancer samples to estimate the intratumoral heterogeneity in samples. VAF shows the proportion of a particular variant in the samples. Mutations in cancer occur frequently, many of which most likely have no / little effect on tumoriogenecity - termed passenger

mutations. As these mutations are not selected for you would expect these genes to have low VAF. In contrast, if mutations occur in driver genes you would expect these mutations to be selected for and to drive clonal growth.

If the clones are completely pure, you would expect passenger mutation to have a VAF < 0.5 and driver mutation VAF > 0.5. Most cancers, however, are heterogeneous and change from sample to sample and is therefore difficult to select such a hardline cutoff value. Visualising the VAF for samples can help inform the selection of driver mutations.

Below shows the VAF for 'sample 1', TIER1/2 mutations also have high VAF values suggesting variants within these genes are being selected for and most likely driver mutations.
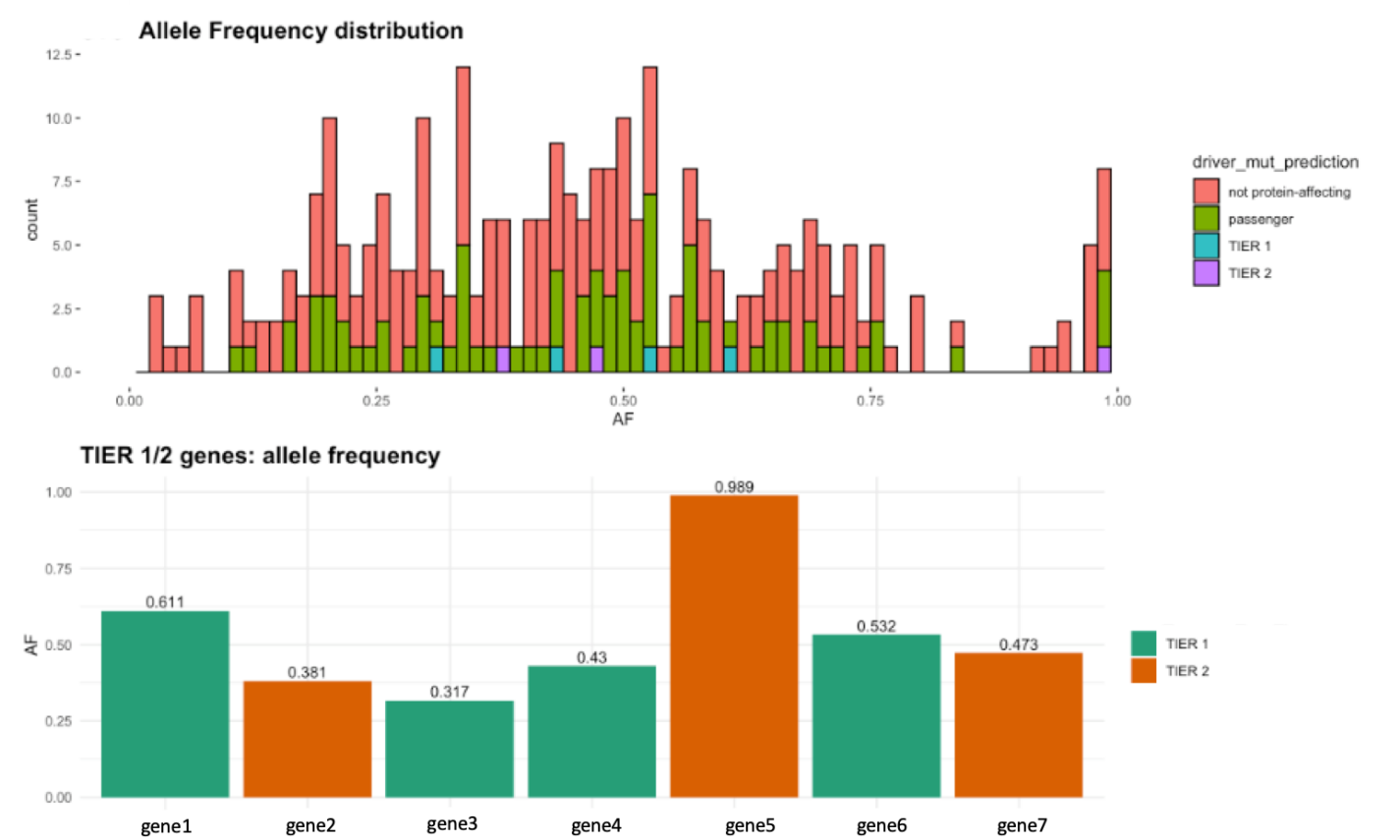
Summary

In summary, this analysis identified a number of consensus variants that were analysed using two independent software - GATK mutect2 and cgpWXS. For turmor samples, CGI identified that a number of these genes as putative driver mutations. In support of this, these genes showed a high VAF (>0.5) suggesting these genes have been selected for and are indeed driving tumorigenicity.