OICR Genomics
Plasma WG Accredited Assay
Analysis Pipeline Deliverables

Sequencing data generated by our plasma Whole Genome (pWG) Assay is processed through the associated Analysis Pipeline. Each case consists of a single plasma sample and somatic variant calls from an associated tumour. Plasma whole genome (WG) libraries are generated from the tumour and sequenced on the illumina Novaseq X Plus platform. Variant calls are generated by our WG or WGTS assays.

Pipeline Steps

- **Sequence Data Generation**
- **Call Ready Alignments**
- **Minimal Residual Disease Assessment**

Details for each pipeline step are provided below including

- Description
- Output files
- Resources
    - GSI workflows repositories. This provides precise information on how the tools were run. Analysis is described and implemented in wdl (workflow description language) for processing under cromwell. Software and data resources are installed on our system as environmental modules.
    - Software pages. Links to manuals and help pages for the various software tools run in our workflows
    - File Type descriptions. Links to pages describing the format of regular file types. This information may also be available on pages for each particular tool

OICR Genomics
Plasma WG Accredited Assay
Analysis Pipeline Deliverables

Sequence Data Generation (pWG)

Generation of demultiplexed fastq records from illumina run folders for sequence data generated at OICR. If data had been generated elsewhere, then fastq files will be injected into our analysis system

Output Files:

1. Raw sequence data, paired end (.fastq.gz)

Resources:

1. Workflow : https://github.com/oicr-gsi/bcl2fastq
2. bcl2fastq software : https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html
3. Fastq format : https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html

OICR Genomics
Plasma WG Accredited Assay
Analysis Pipeline Deliverables

Call Ready Alignment

Raw sequence data (fastq) is trimmed to remove adapter sequence then to the hg38 genomic reference with bwa mem. This is done separately for each fastq pair. There may be multiple bam files per sample depending on how many lanes of sequence data had been generated.
Lane level alignments are merged and filtered to remove non-primary alignments (samtools flag -F 256). PCR duplicates are marked with picard MarkDuplicates. Base quality scores are recalibrated with GATK.

Output Files:

1. Aligned sequence (.bam), call ready
2. Index file (.bai)

Resources:

1. Workflow, lane level alignments : https://github.com/oicr-gsi/bwa
2. Workflow, merging and preprocessing : https://github.com/oicr-gsi/bam-merge-preprocessing
3. cutadapt : https://cutadapt.readthedocs.io/en/stable/
4. samtools : http://www.htslib.org/
5. bwa : **http://bio-bwa.sourceforge.net**
6. gatk : https://gatk.broadinstitute.org/hc/en-us
7. sam/bam specifications : https://samtools.github.io/hts-specs/SAMv1.pdf

OICR Genomics
Plasma WG Accredited Assay
Analysis Pipeline Deliverables

Minimal Residual Disease Assessment

MRD analysis looks for evidence of somatic variants in the sequence data generated from the plasma whole genome. The somatic variants of interest are derived from a tumour that had been obtained from the same subject. Background activity is assessed in a set of 22 curated normal control samples, and is used to generate a detection cutoff for the sample of interest.

Output Files:

1. MRD summary (.mrdtect.txt)
2. MRD detection, tumour + controls (.HBCs.csv)
3. MRD detection plot (.pWGS.svg)
4. MRD, pWG detected sites, variant allele frequency (.vaf.txt)
5. Somatic calls, associated tumour (.vcf)

Resources:

1. https://github.com/oicr-gsi/mrdetect
2. https://pmc.ncbi.nlm.nih.gov/articles/PMC8108131/
3. https://ctl.cornell.edu/industry/mrdetect-license-request/