OICR Genomics
Whole Genome Accredited Assay
Analysis Pipeline Deliverables

Sequencing data generated by our Whole Genome and Transcriptome (WGTS) Assay is processed through the associated WG Analysis Pipeline. Each case consists of 2 samples, a tumour with a matched normal/reference. Whole genome (WG) libraries are generated from the tumour/normal pair. All libraries are sequenced on the illumina Novaseq X Plus platform.

Pipeline Steps

- **Sequence Data Generation (WG)**
- **Call Ready Alignments (WG)**
- **Mutation Calls (WG)**
- **Copy Number Calls (WG)**
- **Structural Variant Calls (WG)**
- **Homologous Recombination Deficiency (WG)**
- **Microsatellite Repeats (WG)**

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
Whole Genome Accredited Assay
Analysis Pipeline Deliverables

Sequence Data Generation (WG)

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
Whole Genome Accredited Assay
Analysis Pipeline Deliverables

Call Ready Alignments (WG)

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
Whole Genome Accredited Assay
Analysis Pipeline Deliverables

Mutation Calls (WG)

Call ready alignments from a tumour/normal pair are analyzed with mutect to generate somatic variants (snps and short indels). Records are modified with gatk FilterMutectCalls to apply various filters and identify pass records, then by variant effect predictor to annotate variants.

Output Files:

1. Somatic calls, with identified filters and annotation (.vcf.gz)
2. Index file (.tbi)
3. Somatic calls, maf format (.maf.gz)

Resources:

1. Workflow : https://github.com/oicr-gsi/mutect2
2. Workflow : https://github.com/oicr-gsi/variantEffectPredictor
3. Mutect2 : https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2
4. FilterMutectCalls : https://gatk.broadinstitute.org/hc/en-us/articles/360036856831-FilterMutectCalls
5. VariantEffectPredictor : https://useast.ensembl.org/info/docs/tools/vep/index.html
6. Vcf specification : https://samtools.github.io/hts-specs/VCFv4.2.pdf
7. MAF format (Default VEP output) : http://useast.ensembl.org/info/docs/tools/vep/vep_formats.html

OICR Genomics
Whole Genome Accredited Assay
Analysis Pipeline Deliverables

Copy Number Calls (WG)

Call ready alignments from a tumour/normal pair are analyzed with mutect2 to generate somatic variants (snps and short indels) and GRIDSS to generate copy number information. This provides input to Purple for evaluation of copy number. Purple generates a number of copy number solutions which require review to select the most appropriate set of results.

Output Files:

1. Purple Purity (.purple.purity.tsv)
2. Purple qc (.purple.qc)
3. Purple purity range (.purple.purity.range.tsv)
4. Purple somatic copy number profile (.purple.cnv.somatic.tsv)
5. Purple gene copy number file (purple.cnv.gene.tsv)
6. Purple Segmentation data (.purple.segment.tsv)
7. Purple primary solution files (.solPrimary.purple.zip)
8. Purple alternate solutions (.purple_alternates.zip)
9. Mutect2 somatic mutations with Purple purity info (.purple.somatic.vcf.gz)
10. GRIDSS somatic structural variants with Purple purity info (.purple.sv.vcf.gz)

Resources:

1. Workflow : https://github.com/oicr-gsi/gridss
2. Workflow : https://github.com/oicr-gsi/purple
3. GRIDSS : https://github.com/PapenfussLab/gridss
4. Purple : https://github.com/hartwigmedical/hmftools/tree/master/purple
5. Vcf specification : https://samtools.github.io/hts-specs/VCFv4.2.pdf

OICR Genomics
Whole Genome Accredited Assay
Analysis Pipeline Deliverables

Structural Variant Calls (WG, WT)

Call ready alignments from a tumour/normal pair are analyzed with delly to generate somatic structural variants (deletions, duplications, inversion, insertions, translocations. Delly calls (from whole genome libraries) are used as input to Mavis for validation, annotation and visualization of identified calls.

Output Files:

1. Delly somatic calls, all calls Pass and non-pass (.vcf.gz)for gz
2. Mavis : data table
3. Mavis : zipped file with drawings referenced in the data table

Resources:

1. Workflow : https://github.com/oicr-gsi/delly
2. Workflow : https://github.com/oicr-gsi/mavis
3. Delly : https://github.com/dellytools/delly
4. Mavis : https://github.com/bcgsc/mavis
5. Vcf specification : https://samtools.github.io/hts-specs/VCFv4.2.pdf

OICR Genomics
Whole Genome Accredited Assay
Analysis Pipeline Deliverables

Homologous Recombination Deficiency (WG)

Homologous Recombination Deficiency is detected with the mutational signature package sig-tools, using structural variants generated from the GRIDSS software. The HRDetect module assesses HRD using the following signatures

SNV3 : (alias SBS3,Signature3,RefSig3) : Defective homologous recombination-based DNA damage repair
SNV8 : (alias SBS8,Signature8,RefSig8)
SV3 : (alias RS3,RefSigR3) HR Deficiency
SV5 : (alias RS5, RefSigR5,RefsigR9)
del.mh : Deletions at micro-homology regions (Indels)
hrd = HRD-LOH index (CNV)

The Probability of HRD (BRCAness), and the weight that each of the signatures contributes to that assessment is calculated repeatedly using a bootstrapping process, based on a model defined in
https://pubmed.ncbi.nlm.nih.gov/28288110/
Davies et al : HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures
The reported output is the 5th, 50th and 95th percentiles of the generated values for each metric.

Output Files:

1. Signatures file (.signatures.json)

Resources:

1. Workflow : https://github.com/oicr-gsi/hrDetect
2. Sig-tools : https://github.com/Nik-Zainal-Group/signature.tools.lib
3. Sig-tools, hrDetect : https://github.com/Nik-Zainal-Group/signature.tools.lib/blob/master/userManuals/hrDetect_commandLine_userManual.pdf
4. Json format : https://www.w3schools.com/js/js_json_syntax.asp
5. Mutational Signatures : https://cancer.sanger.ac.uk/signatures/

OICR Genomics
Whole Genome Accredited Assay
Analysis Pipeline Deliverables

Microsatellite Instability Detection (WG)

Identification of Instability/Slippage in microsatellite regions, classifying the events as germline or somatic.

Output Files:

1. Microsatellites summary (.msi): msisensor call using all microsatellite sites in the genome, last column is MSI score
2. Germline microsatellites (.msi_germline): microsatellite sites found in the normal
3. Somatic microsatellites (.msi_somatic): somatic microsatellite sites
4. Bootstrap msi (.msi.booted): msisensor calls bootstrapped, ran 100 times with 500 random sites. File columns: bootstrap index, germline sites, somatic sites and the MSI score. Median score is used for reporting.

Resources:

6. Workflow : https://github.com/oicr-gsi/msisensor
7. MSISensor-pro : https://github.com/xjtu-omics/msisensor-pro