Roslin Variants v2.2

Roslin is a **cancer informatics pipeline** maintained by the Platform Informatics group at the Center for Molecular Oncology (CMO). Its workflow for targeted-variants is capable of variant calling, annotation, and analysis of data from 341, 410, or 468 gene MSK-IMPACT assays [1], IMPACT+, HemePACT, and various exome capture kits. Additional workflows for xenograft, cell-free DNA, whole genome, and RNA-seq are planned for 2018.

Roslin builds on prior work by the Bioinformatics Core, Clinical Bioinformatics, and Computational Oncology groups, and continues to rely on their accumulated experience and expertise, with emphasis on these features:

Modular - Easily addon or replace sequence aligners, variant callers, false-positive filters, functional/clinical annotation, and analysis modules for manuscript-ready plots/tables.
Reproducible - Retain all older versions and documentation in sufficient detail to reproduce published results, with zero dependencies on proprietary software or obfuscated methods.
Portable - Install Roslin and process new datasets with minimal fuss on laptops, workstations, local compute clusters, or cloud compute servers.

Most of these goals are accomplished using UCSC's Toil [2], a cross-platform workflow management system that uses the Common Workflow Language (CWL), a workflow definition standard promoted by the Global Alliance for Genomics and Health (GA4GH).

Manifest of Roslin output files `v2.2.2`

The output of the Roslin pipeline consists of the following folders at the main project level. `$ROOT` is the path to the top of the output hierarchy and in that folder the output is organized as follows:

- `$ROOT/results`: Primary results directory, most users should look here first.
- `$ROOT/docs`: Documentation and pipeline parameters, settings versions.
- `$ROOT/bams`: BAM files with indices
- `$ROOT/output`: Detailed output of the pipeline

`$ROOT/docs`

This folder contains documentation about the output and pipeline and also the parameters used in this specific run of the pipeline. The input arguments/parameters are in the folder: `inputs`. A PDF copy of this file is in this folder `manifest.pdf`. This version of the documentation will match the pipeline version run.

The docs folder has a subfolder `qc` with a PDF of some of the core QC-metrics (full metric output is in the `output`) folder. The file is `${projNo}_QC_Report.pdf`.

The input folder has the following files:

- `inputs.yaml` - All inputs self-contained in a format that Roslin likes, including paths to reference genomes, assay target/bait sets, known somatic hotspots, etc.
- `settings` - Version information of Roslin pipeline and Roslin core
- `${projNo}_request.txt` - Relevant information from the original iLabs request
- `${projNo}_sample_mapping.txt` - Maps samples to the folders containing their FASTQs
- `${projNo}_sample_pairing.txt` - Pairs tumors to normals for somatic variant calling
- `${projNo}_sample_grouping.txt` - Groups samples that belong to the same patient
- `${projNo}_sample_data_clinical.txt` - Patient/sample data from the investigator

`$ROOT/results`

The files in this folder all start with a prefix which is the project number (e.g.; `Proj_01234`), here we will denote that with `$projNo`. There are three primary results types given: mutations, copy number (using the FACETS algorithm), and fusions. Each results type is in its own folder: `variants`, `copyNumber`, `rearrangements`

`$ROOT/results/variants`

There are two primary `MAF` files in the `$ROOT/results/variants` folder: The *portal* (DMP) MAF and the *analysis* MAF. The *portal* MAF contains a subset of the events that are in the *analysis* MAF. The steps to creating the *analysis* MAF are as follows:

```
- Merge all sample level MAFs (which are in `$ROOT/output/maf`)

- Remove events that are tag as false-positives and any from cmo_fillout.

- Remove splice region variants in non-coding genes, or those that are >3bp into introns.
    - For indels, use the closest distance to the nearby splice junction.

- Remove all non-coding events except interesting ones like TERT promoter mutations.
```

The *portal* MAF then applies the following additional filters.

```
- Remove silent aka synonymous mutations

- Remove genes without Entrez IDs, usually non-coding genes

- Remove intronic splice region mutations i.e. >2bp into introns

- For IMPACT/HemePACT runs, apply MSK-IMPACT cutoffs:
    - Total depth >= 20
    - Allele Depth `>=10` for non-hotspots, `>=8` for hotspots
    - VAF `>=5%` for non-hotspots, VAF `>=2%` for hotspots
    - Length f Indel or ONP < 30bps
```

Copy Number

Output from the FACETS copy number method.

- `${projNo}.hisens.gene.cna.txt`: Unified (all samples) gene level calls file.
- `${projNo}.hisens.seg.cna.txt`: Unified segmentation file in IGV format.

Facets was run in a two pass mode: Pass (1) was to option purity estimates with coarse grain parameters to get large features accurately and pass (2) with hi-sensitivity parameters to increase spatial resolution. In the results folder we have only the output from the hisens pass (the purity pass) is available in the full output folder.

- `${projNo}.purity.Parameters.out` - The run parameters used for the specific pass
- `${projNo}.purity.SamplesValues.out` - Sample level output from the Facets algorithm.
- `${projNo}.purity.CNCF.pdf` - Plots of the bi-segmentation profiles and integer copy number calls
- `${projNo}.purity.cncf.txt` - Segment level copy number and cell fraction and integer copy number.

Fusions

- `${projNo}.fusions.txt` - Filtered fusion calls. Calls were filtered to include fusions that are on a *white-list* derived from OncoKB fusion1 (http://oncokb.org/api/v1/variants/lookup?variant=fusion) and *also* have a precise breakpoint.

`$ROOT/bams`

The fully processed (markduplicated,realigned,recalibrated) BAM files for the project with indices. *N.B.* the two copies of the index files (`.bai` and `.bam.bai`) are *indentical*, both are provide because certian bioinformatics tools will only recognize one or the other.

`$ROOT/output`

The output folder contains a comprehensive set of output files from the pipeline; a verbose results folder.

`$ROOT/output/maf`

- `${sampleID}.svs.pass.vep.maf` - Structural variants (SVs) in MAF format (IMPACT only)
- `${sampleID}.muts.maf` - Small substitutions and indels in MAF format. False positives have a non-PASS tag in the FILTER column, and “fillout” rows (allele counts per event in other samples) are tagged “None” in the Mutation_Status column.

`$ROOT/output/vcf`

- `${sampleID}.vardict.vcf` - Substitutions and indels reported by VarDict in VCF format
- `${sampleID}.mutect.{vcf,txt}` - A MuTect VCF plus its more detailed tab-delimited format
- `${sampleID}.pindel.vcf` - Comprehensive VCF of indels reported by Pindel
- `${sampleID}.svs.vcf` - Comprehensive VCF of SVs detected by Delly (IMPACT only)
- `${sampleID}.svs.pass.vcf` - Shortlisted VCF of Delly SVs after some basic filtering

`$ROOT/output/facets`

- `${sampleID}_hisens.CNCF.png` - A plot for genome-wide integer copy-number (CN) per Facets
- `${sampleID}_hisens.cncf.txt` - Stats per segment with sufficient data for CN estimation
- `${sampleID}_hisens.seg` - Segmented copy-number data listing log-ratios
- `${sampleID}_purity.*` - All the files above, for the run of facets that estimated purity

`$ROOT/output/portal`

This directory contains the files used for upload to the portal and contain the exact same events displayed in the portal.

- `data_mutations_extended.txt` - The subset of mutations that the portal will display. It starts with the events from the main results maf (`${projNo}.muts.maf`) and then applies the following filters:
  - Remove silent aka synonymous muts
  - Remove genes without Entrez IDs, usually non-coding genes
  - Remove intronic splice region muts i.e. >2bp into introns
  - For IMPACT/HemePACT runs, apply DMP cutoffs:
    - Allele Depth `>=8`
    - VAF `>=5%` for non-hotspots, VAF `>=2%` for hotspots
- `data_clinical.txt` - Clinical data per patient/sample to display in the portal
- `data_CNA.txt` - Sample x Gene matrix listing discretized copy number alterations
- `data_fusions.txt` - Shortlist of somatic structural variants (IMPACT only)
- `${PI_UUID}_data_cna_hg19.seg` - Segmented copy-number data listing log-ratios. `${PI_UUID}` is the Roslin/Portal UUID for the project. *N.B.* this is differnt from $projNo
- `case_lists` - Folder containing lists of sample IDs with meta-data for the portal
- `meta_*.txt` - Metadata about the study and the files above that the portal needs

`$ROOT/output/qc`

- `${projNo}_CutAdaptStats.txt` - Stats on paired end reads that needed trimming
- `${projNo}_DiscordantHomAlleleFractions.txt` - Concordance of homozygous SNPs
- `${projNo}_FingerprintSummary.txt` - Concordance of heterozygous SNPs
- `${projNo}_UnexpectedMatches.txt` - Samples from different patients with concordance
- `${projNo}_UnexpectedMismatches.txt` - Samples from same patient with discordance
- `${projNo}_MajorContamination.txt` - Contamination check using heterozygous SNP VAFs
- `${projNo}_MinorContamination.txt` - Contamination check using homozygous SNP VAFs
- `${projNo}_GcBiasMetrics.txt` - Check if coverage varies much by GC content
- `${projNo}_HsMetrics.txt` - Targeted library hybridization metrics per Picard
- `${projNo}_InsertSizeMetrics_Histograms.txt` - Insert size distribution per sample
- `${projNo}_markDuplicatesMetrics.txt` - Fragment duplication rates per Picard
- `${projNo}_pre_recal_MeanQualityByCycle.txt` - quality scores before GATK BQSR
- `${projNo}_post_recal_MeanQualityByCycle.txt` - quality scores after GATK BQSR
- `${projNo}_ProjectSummary.txt` - Table of QC stats that have passed or failed

- `${projNo}_SampleSummary.txt` - Table of successes/failures per sample

`$ROOT/output/log`

- `cwltoil.log` - Records warnings and messages from the Toil workflow manager
- `output-meta.json` - Metadata on all files generated and used by Toil
- `run-results.json` - Indication of completed or failed steps of Roslin pipeline
- `stderr.log` - Records warnings and failures
- `stdout.log` - Records the stdout of Roslin pipeline progress

Version info

```
export ROSLIN_PIPELINE_DESCRIPTION="Roslin Variant Pipeline v2.2.0"

# Roslin pipeline name/version
export ROSLIN_PIPELINE_NAME="variant"
export ROSLIN_PIPELINE_VERSION="2.2.0"

# which version of Roslin Core is required?
export ROSLIN_CORE_MIN_VERSION="2.0.2"
export ROSLIN_CORE_MAX_VERSION="2.0.2"

# cmo
export ROSLIN_CMO_VERSION="1.9.8"
```

Modules in the Roslin Targeted-Variants Workflow

Roslin is currently driven by 6 modules.

Module 1 - Alignment

Demultiplexed fastq files are trimmed using Trim Galore v0.2.5mod [3] which removes adapters and short reads. These are then aligned to a human reference genome (ie GRCh37) using BWA-MEM v0.7.5a [4]. Picard Tools v2.9 [5] AddOrReplaceReadGroups is performed to annotate read groups and MarkDuplicates to mark PCR duplicates.

Module 2 - Recalibration & Realignment

Genomic regions are identified using FindCoveredIntervals from GATK Tool Kit v3.3-0 [6] and subjected to indel realignment using Assembly Based ReAligner (ABRA) v2.12 [7].
GATK BaseRecalibrator is used to detect systematic errors in base quality scores.

Module 3 - Variant Calling

Roslin uses multiple variant callers in combination to detect somatic mutations.

Variant calling was performed in paired tumor/normal mode using MuTect v1.1.4 [8] for single nucleotide variants (SNV). Pindel v0.2.5a7 [9] is used for small insertions and deletions (indels). Vardict v1.5.1 [10] is an ultra sensitive variant caller to report both SNVs and indels.

MuTect parameters:
–totaldepth 5 #tumor total depth threshold
–alleledepth 3 #tumor allele depth threshold
–tnRatio 5 # tumor-normal varint frequency ratio threshold
–variantfraction 0.01 # tumor variant frequency threshold

Vardict parameters:
-f 0.01
-Q 20
-q 20
-X 5
-x 2000

Pindel parameters:
–min_var_len 0 #min length of indels
–max_var_len 200 #max length of indels
–max_hom_len 5 #max length of micro-homology at indel breakpoint

In addition, copy-number variants including chromosomal instability (CIS) and whole-genome doubling (WGD) were called using FACETS [11]. Microsatellite instability is also detected using Msisensor msi [12].

The vcfs generated by MuTect, Vardict, and Pindel are combined (need to detail).

Module 4 - Variant Filtering & Annotation

Resulting variants were annotated using vcf2maf v1.6.14 [13] which uses Ensembl's Variant Effect Predictor v86. Additional filtering is done to make sure complex variants (substitution with >1 bps replaced/deleted/inserted by another >1 bps) are called correctly. Roslin also flags false-positive somatic calls using ngs_filters v1.2.1 [14].

Module 5 - QC Metrics

A project report is generated detailing the different quality calls used and measured in the analysis. Roslin uses the following Picard metrics:
CollectAlignmentSummaryMetrics
CollectHsMetrics
CollectInsertSizeMetrics
CollectMultipleMetrics
CollectGcBiasMetrics
DepthOfCoverage

These will provide in great detail samples with low coverage, duplication rates, sample mismatches, read quality, etc.

Module 6 - Structural Variants

Somatic structural aberrations were identified using DELLY v0.7.7 [15].

DELLY parameters:
-s 9 #insert size cutoff
-u 5 # min. mapping quality for genotyping
-a 0.04 # min. fractional ALT support
–minsize 500 #min. SV size
–maxsize 500000000 # max. SV size
–ratiogeno 0.0 #min. fraction of genotyped samples
–pass true #Filter sites for PAS
–coverage 10 #min. coverage in tumor
–controlcontamination 0 #max. fractional ALT support in control
–gq 15 #min. median GQ for carriers and non-carriers
–rddel 0.800000012 #max. read-depth ratio of carrier vs. non-carrier for a deletion
–rddup 1.20000005 # min. read-depth ratio of carrier vs. non-carrier for a duplication

References

1. Cheng, D. T., Mitchell, T. N., Zehir, A., Shah, R. H., Benayed, R., Syed, A., ... & Brannon, A. R. (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. The Journal of molecular diagnostics, 17(3), 251-264.
2. Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., … Paten, B. (2017). Toil enables reproducible, open source, big biomedical data analyses. Nature Biotechnology, 35(4), 314–316. http://doi.org/10.1038/nbt.3772
3. Krueger, F. (2015). Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.
4. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]
5. https://github.com/broadinstitute/picard
6. i. a. DePristo et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43, 491-498 (2011).
7. i. a. Mose, M. D. Wilkerson, D. N. Hayes, C. M. Perou, J. S. Parker, ABRA: improved coding indel detection via assembly-based realignment. Bioinformatics 30, 2813-2815 (2014).
8. i. Cibulskis et al., Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31, 213-219 (2013).
9. i. Ye, M. H. Schulz, Q. Long, R. Apweiler, Z. Ning, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865-2871 (2009).
10. https://github.com/AstraZeneca-NGS/VarDict
11. https://github.com/mskcc/facets-suite
12. Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., McLellan, M. D., ... & Ding, L. (2013). MSIsensor: microsatellite

instability detection using paired tumor-normal sequence data. Bioinformatics, 30(7), 1015-1016.

13. https://github.com/mskcc/vcf2maf
14. https://github.com/mskcc/ngs-filters
15. Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M. Stuetz, Vladimir Benes, Jan O. Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012 28: i333-i339.