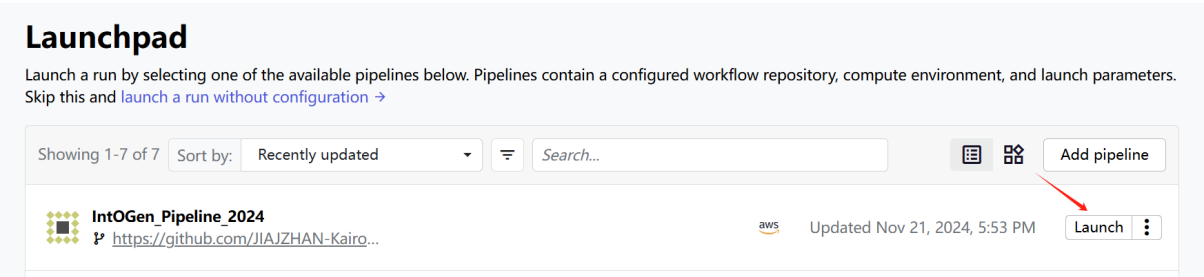
How to Use the IntOGen Pipeline on Seqera

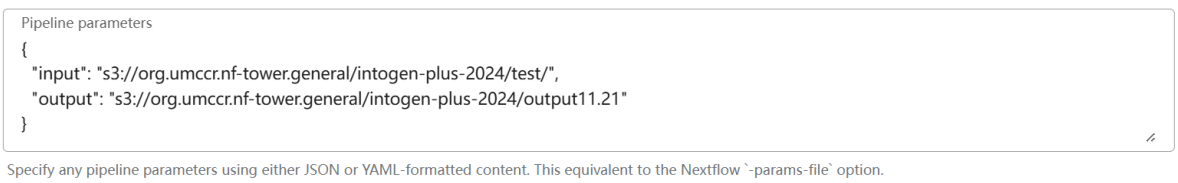
Contents

IntOGen have set up on Seqera, you don't need to install the pipeline and build the containers (Gonzalez-Perez, et al., 2013; Martinez-Jimenez, et al., 2020).

Usage



Firstly, navigate to Launchpad, find IntOGen_Pipeline_2024, click Launch.
The primary customization involves modifying the pipeline parameters:



The parameters need to follow JSON or YAML- format, it's better only edit the input and output path based on the original one.

You can add seed too. Seed to be used for reproducibility. This applies to 4 methods: smRegions, OncodriveCLUSTL, OncodriveFML, dNdScv.
An example configuration is shown below:

```
{

  "input": "s3://org.umccr.nf-tower.general/intogen-plus-2024/test/",

  "output": "s3://org.umccr.nf-tower.general/intogen-plus-2024/output/",

  "seed": 42

}
```

Input & output

Input

File Structure Requirements:

Although the pipeline does most of its computations at the cohort level, the pipeline is prepared to work with multiple cohorts at the same time.

Cohort files must contain:

- Chromosome
- Position
- Reference base
- Alternate base

Files are expected to be TSV files with a header line.

Important

All mutations should be mapped to the positive strand. The strand value is ignored.

Annotation File Configuration:

In addition, each cohort must be associated with:

- cohort ID (DATASET): a unique identifier for each cohort.
- a cancer type (CANCER): Any acronym in oncotree can be used here.

- a sequencing platform (PLATFORM): WXS for whole exome sequencing and WGS for whole genome sequencing
- a reference genome (GENOMEREF): only HG38 and HG19 are supported

Cohort file names, as well as the fields mentioned above must not contain dots.

The way to provide those values is through OpenVariant , a comprehensive Python package that provides different functionalities to read, parse and operate different multiple input file formats (e. g. tsv, csv, vcf, maf, bed). Whether you are planning to run single or multiple cohorts, you would need to provide an annotation file in yaml format to specify the above mentioned structure required by IntOGen.

What you need to do is copy the *s3://org.umccr.nf-tower.general/intogen-plus-2024/test/metadata.yaml* to your input folder (the metadata.yaml and your input data need to be in the same folder) and modify these parameters as necessary.

Example:

```yaml
pattern:
- 'combined-somatic-PASS.tsv' # The file name of your inpute data
recursive: false
format: tsv # The file format of your input data(e. g. tsv, csv, vcf, maf, bed)
```

...

```yaml
- type: static
  field: CANCER
  value: PAAD # The acronym cancer type of your cohort
```

...

```yaml
- type: static
  field: PLATFORM
  value: WXS # the platform of sequencing, WXS/WGS
- type: static
  field: GENOMEREF
  value: hg38 # The reference genome you used before, only hg38 or hg19.
```

```
- type: static
  field: DATASET
  value: # Edite to your cohort’s name
```

Once the parameters are configured, click 'Launch' to start the pipeline. You can monitor the task progress in the run channel.

Output

By default this pipeline outputs 4 files:

- *cohorts.tsv:* summary of the cohorts that have been analyzed
- *drivers.tsv:* summary of the results of the driver discovery by cohort
- *mutations.tsv:* summary of all the mutations analyzed by cohort
- *unique_drivers.tsv:* information on the genes reported as drivers (in any cohort)
- *unfiltered_drivers.tsv:* information on the filters applied to the post-processing step: from the output of the combination to the final set of driver genes.

Those files can be found in the path indicated with the --output options.
The output of each tools can be found in the step folder.

Output Explanation:

drivers.tsv:

1. **SYMBOL**
   The standard gene symbol.
2. **TRANSCRIPT**
   The specific transcript identifier for the gene (e.g., ENST numbers refer to Ensembl transcripts).
3. **COHORT**
   The name of the data cohort or dataset from which the samples are derived.
4. **CANCER_TYPE**
   The type of cancer being studied (e.g., PAAD stands for Pancreatic Adenocarcinoma).
5. **METHODS**
   The computational methods used to identify driver genes.
6. **MUTATIONS**
   The total number of mutations identified in the gene.

7. **SAMPLES**

   The number of samples in which these mutations were detected.

8. **%_SAMPLES_COHORT**

   The percentage of samples in the cohort carrying mutations in this gene (e.g., 0.0306 means ~3.06% of samples).

9. **QVALUE_COMBINATION**

   The FDR-adjusted p-value (False Discovery Rate) assessing statistical significance. Smaller values indicate higher significance.

10. **ROLE**

    The functional role of the gene:

    - LoF: Loss-of-function gene.
    - GoF: Gain-of-function gene.

11. **CGC_GENE**

    Indicates whether the gene is listed in the **Cancer Gene Census (CGC)**:

    - True: The gene is a known cancer-related gene.
    - False: The gene is not in the CGC.

12. **CGC_CANCER_GENE**

    Whether the gene is specifically linked to cancer in CGC.

13. **DOMAIN**

    Information on the functional domains of the gene's protein, if annotated.

14. **2D_CLUSTERS**

    The number of mutation clusters identified using 2D spatial analysis, highlighting potential mutation hotspots.

15. **3D_CLUSTERS**

    The number of mutation clusters identified using 3D structural analysis, indicating potential functional hotspots in the protein structure.

16. **EXCESS_MIS**

    Excess of missense mutations observed (compared to expected background mutations), often associated with gain-of-function events.

17. **EXCESS_NON**

    Excess of non-synonymous or truncating mutations (frameshift, nonsense, etc.), often linked to loss-of-function events.

18. **EXCESS_SPL**

    Excess of splice site mutations observed.

unfiltered_drivers.tsv:

1. **SYMBOL**
   The gene symbol.
2. **TRANSCRIPT**
   The Ensembl transcript ID representing a specific isoform of the gene.
3. **COHORT**
   The dataset or cohort name.
4. **CANCER_TYPE**
   The cancer type associated with the cohort (e.g., PAAD for Pancreatic Adenocarcinoma).
5. **MUTATIONS**
   The total number of mutations identified in this gene.
6. **SAMPLES_COHORT**
   The total number of samples in the cohort being analyzed.
7. **ALL_METHODS**
   All methods applied to identify potential driver genes for this gene (e.g., dndscv, cbase, mutpanning...).
8. **SIG_METHODS**
   Methods that identified the gene as statistically significant.
9. **QVALUE_COMBINATION**
   The combined FDR-adjusted p-value (False Discovery Rate), summarizing statistical significance across methods. Smaller values indicate higher significance.
10. **QVALUE_CGC_COMBINATION**
    FDR-adjusted p-value considering **Cancer Gene Census (CGC)** gene set. Smaller values indicate higher confidence if the gene is known to be cancer-related.
11. **RANKING**
    The gene's rank in terms of significance within the cohort.
12. **TIER**
    The assigned tier (or priority level) for the gene, often based on statistical and biological evidence:
    - 1: Strong evidence as a driver gene.

- 4: Weak evidence or no driver.

13. **ROLE**

    The functional classification of the gene:

    - LoF: Loss-of-function.
    - Act: Activating (gain-of-function).

14. **CGC_GENE**

    Whether the gene is listed in the **Cancer Gene Census (CGC)**:

    - True: Known cancer-related gene.
    - False: Not in CGC.

15. **TIER_CGC**

    Tier assignment based on CGC genes, prioritizing known cancer-related genes.

    "We annotated the catalog of highly confident driver genes according to their annotation level in CGC. Specifically, we created a three-level annotation: i) the first level included driver genes with a reported involvement in the source tumor type according to the CGC; ii) the second group included CGC genes lacking reported association with the tumor type; iii) the third group included genes that were not present in CGC."

16. **CGC_CANCER_GENE**

    Indicates whether the gene is included in the CGC.

17. **SIGNATURE9**

    Coding regions might be the target of mutations associated to COSMIC Signature 9 (https://cancer.sanger.ac.uk/cosmic/signatures) which is associated to non-canonical AID activity in lymphoid tumours.

    In those cancer types were Signature 9 is known to play a significant mutagenic role (i.e., LYMPHOID, CLLSLL, DLBCLNOS, NHL, ALL, LNM, TLL, BLL and HL) we discarded genes where more than 50% of mutations in a cohort of patients were associated with Signature 9.

18. **SIGNATURE10**

    The proportion of mutations in this gene attributed to Signature 10 from the COSMIC Mutational Signatures database, it is associated with mutations caused by defective DNA polymerase proofreading activity, specifically due to mutations in POLE (DNA polymerase epsilon).

19. **WARNING_EXPRESSION**

    Indicates if the gene expression analysis raised warnings for this gene. (We discarded non-

expressed genes using TCGA expression data. For tumor types directly mapping to cohorts from TCGA –including TCGA cohorts– we removed non-expressed genes in that tumor type. We used the following criterion for non-expressed genes: genes where at least 80% of the samples showed a negative log2 RSEM. For those tumor types which could not be mapped to TCGA cohorts this filtering step was not done.)

20. **WARNING_GERMLINE**
    Indicates if germline mutations raised warnings. (we discarded mutations overlapping with germline variants (germline count > 5) from a panel of normals (PON) from Hartwig Medical Foundation)

21. **SAMPLES_3MUTS**
    The number of samples with 3 or more mutations in this gene, potentially indicating artifact or noise. (We also removed non CGC genes with more than 2 mutations in one sample. This abnormally high number of mutations in a sample may be the result of either a local hypermutation process or cross contamination from germline variants.)

22. **OR_WARNING**
    Indicates if there were warnings related to the odds ratio calculation.

23. **WARNING_ARTIFACT**
    Indicates if the gene has potential mutation artifacts.

24. **KNOWN_ARTIFACT**
    Indicates if the gene is associated with known artifacts.

25. **NUM_PAPERS**
    The number of papers referencing this gene in the context of cancer research.

26. **WARNING_ENSEMBL_TRANSCRIPTS**
    Indicates whether there are potential issues with transcript annotation for this gene.

27. **DRIVER**
    Indicates whether this gene is classified as a **driver gene**:

    - True: Identified as a driver.
    - False: Not identified as a driver.

28. **FILTER**
    Final filter applied to the gene:

    - PASS: Gene passes all criteria for inclusion as a driver.
    - Other values (e.g., "No driver", "Lack of literature evidence", “Warning artifact”)

explain why the gene is excluded.

Strength & Limitation:

Strength:

1. Efficient Driver Gene Identification
    - IntOGen integrates multiple methods to analyze mutation enrichment and functional impact, enabling accurate identification of cancer-related driver genes.
2. Multi-Cohort Support
    - It supports analysis across multiple cancer cohorts, allowing for the integration and comparison of driver gene patterns across different cancer types and data sources.
3. Comprehensive Output
    - Provides multi-layered output files, including driver gene lists (drivers.tsv), mutation summaries (mutations.tsv), and filtering information (unfiltered_drivers.tsv), offering rich data for further research.
4. Flexible Input Data Support
    - Supports various file formats (e.g., TSV, VCF, MAF) and leverages the OpenVariant package for data parsing, enhancing tool adaptability.
5. Effective Filtering Mechanisms
    - Eliminates false positives with robust filtering criteria such as:
        - Filtering non-expressed genes.
        - Filtering mutation signatures (e.g., Signature 9 hypermutations).
        - Filtering high-frequency pseudo-mutations (e.g., genes with abnormally high mutations in single samples).
    - Ensures high accuracy and reliability of analysis results.
6. Integration with Cancer Gene Census (CGC)
    - Provides tiered annotations for driver genes through integration with the CGC, with a focus on high-confidence, cancer-related genes.
7. Focus on Functional Characteristics
    - Outputs include information on the functional roles of driver genes (e.g., LoF or GoF)

and mutation enrichment patterns (e.g., 2D and 3D clustering), providing biological insights for further research.

Limitation:

1. Limited to Known Mutation Signatures and Databases
    - The results heavily rely on existing COSMIC mutation signatures and the CGC database, which may lead to missed detections of rare cancers or novel mutation signatures.
2. Processing of Mutation Site Information
    - IntOGen requires mutation data files to include positional information (e.g., chromosome, position, reference base, and alternative base).
    - However, the output focuses on overall statistics for each gene (e.g., mutation frequency, sample coverage) rather than listing individual mutation sites.
3. Inability to Reveal Functional Mutations in Non-Coding Regions
    - IntOGen pipeline primarily focuses on coding regions, which is evident from the following:
        - Filtering Mechanism: Mutations in non-coding regions are excluded. For instance, based on TCGA expression data, non-expressed genes are filtered out to narrow the analysis scope further.
        - Supported COSMIC Mutation Signatures: Signatures like Signature 9 and Signature 10 are derived primarily from analyses of mutations in coding regions.
4. Lack of Functional Validation of Mutations
    - While IntOGen can predict driver genes, it does not directly validate the functional impact of mutations on gene activity, requiring additional functional experiments or data.
5. Complex Output Files
    - The output is detailed but complex, making it challenging for beginners to interpret the numerous fields and warning messages, thus requiring a learning curve.

Reference:

1. For more details, please refer to the IntOGen Documentation
2. Gonzalez-Perez, A.*, et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* 2013;10(11):1081-1082.
3. Martinez-Jimenez, F.*, et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* 2020;20(10):555-572.