An analysis summary file provides processing details for every case in a batch. It includes information about what input data was used to generate each result, and lists all data to be submitted as pipeline results.

Note that the analysis summary file as described here, also known as the local analysis summary file, is different from the DCC analysis summary file used to report results to CPTAC3. The main distinction is that a local analysis summary file has local paths, while the DCC analysis summary has paths on DCC. The script submit.CPTAC3/src/stage_data.sh generates the DCC version from the local during the staging phase of data upload, and the resulting files can be seen in the CPTAC3.catalog/DCC_Analysis_Summary directory on GitHub.

Analysis summary file format

Typical filename: *batch_name*.analysis_summary.dat

Format:

- Tab separated text
- Comment lines start with #
- First row has column names, starts with #
- One row per output file

Columns will vary depending on the particular pipeline. Typical input may be,

- **Run name**
  - `Column name: run_name`
  - Unique name of run. When there is one run per case, run name is typically the same as case name. When there are multiple runs per case, this must be different for each run.
  - Run name column is new as of September 2021. See details below
- **Case name**
  - `Column name: case`
  - Note, this must be exact case name as in BamMap file
- **Disease**
  - `Column name: disease`
- **Output File Path**
  - `Column name: data_path`
- **Output File Format**
  - `Column name: file_format`
- Tumor sample name (e.g., “C3L-01032.WXS.N”, from column 1 of BamMap)
  - `Column name: tumor_name`
- Tumor BAM UUID
  - `Column name: tumor_uuid`
- Normal sample name (col 1 BamMap)

    - Column name: normal_name
- Normal BAM UUID
    - Column name: normal_uuid

For a run with just one input sample, use the following columns names, placed after "Output File Format":

- sample_name
- sample_uuid

Likewise, for a run which takes two FASTQ files, R1 and R2, the last four columns would be,

- R1_sample_name
- R1_sample_uuid
- R2_sample_name
- R2_sample_uuid

The first 5 fields (bold) are required. The remaining ones will depend on the pipeline, but all primary input data must be listed and include both the sample name and sample UUID. In addition, each row of the analysis summary file excluding the path must be unique, so if the same input files result in multiple output files with the same format, an additional field must be added which can distinguish the output in the two rows.

Example line:

```
# run_name  case  disease data    file_format tumor_name  tumor_uuid  normal_name normal_uuid
C3L-00081  C3L-00081  LSCC    /diskmnt/Projects/.../final.SV.WGS.vcf VCF C3L-00081.WGS.T.hg38
cb4885fd-11cd-4eca-a876-35c74daf9feb    C3L-00081.WGS.N.hg38    4571e0d7-4c43-4ff8-aad2-f39bc4964cdf
```

If an analysis produces multiple output files, add a column named “result_type”, and have one row for each file which is to be uploaded. Specify input files with both the sample_name and the UUID; do not include other details (e.g., sample type). Examples can be found in CPTAC3.catalog/DCC_Analysis_Summary

Complete example analysis summary file can be found on Katmai here:
`/home/mwyczalk_test/Projects/SomaticSV/somatic_sv_workflow/demo/task_call/katmai.C3/dat/analysis_summary.dat`

Automated scripts for generating analysis summary files can be found in SomaticSV project, specifically steps 1_make_yaml.sh and 3_make_analysis_summary.sh

The analysis summary file is required to submit pipeline analysis results (click link to submit)

Development note: the output file path must be unique, or else we can append CASE to it. However, for cases where one case has multiple results (e.g., tumor and normal), then it will still

not be unique. In this situation, CASE should be made unique, perhaps by having CASE-normal and CASE-tumor

Additional details needed

- Specify more precisely what is meant by file_format
    - Is .gz a different file format?
    - A VCF is a kind of TSV, which is a TXT or a DAT
        - It is not just the extension always
    - Importantly, when two or more results from single run, file_format is used to distinguish output
        - MSI outputs two files, one has file_format “DAT” the other “TSV”

Run Name details

This field is meant to distinguish runs where there is more than one tumor per case (e.g., heterogeneity samples). In the typical case, the run name is the same as the case name. In the case of a code like "DEEP_vrKGyn" in the sample name (which indicates a secondary tumor sample), add this code to the case name, e.g., C3L-01637.DEEP_vrKGyn

The important point is that each run name is different, even if the case is the same. It is OK to use the sample name of the tumor sample as the run name (e.g., C3L-01637.WXS.T.DEEP_vrKGyn.hg38 is OK).

Note that if a run generates more than result file, there will be multiple rows in the analysis summary, one for each result file, and these will have the same run name (indicating they are generated by the same run)

Batch names

Batches are used to identify groups of runs for assignments, reporting, and uploads. Batch names are of the format PIPELINE.BATCH_ID

Batch ID is a string which uniquely identifies this group of runs, and can take a variety of formats. Some suggestions:

- PDA_89 - run of 98 PDA cases
- Y3.b2 - Year 3 batch 2 analysis
- Y3.620 - Y3, analysis circa June 2020

Batch IDs are determined by the analyst

PIPELINE is one of the following canonical pipeline names:

- Methylation_Array
- miRNA-Seq
- RNA-Seq_Expression

- RNA-Seq_QC_SampleMatch
- RNA-Seq_Fusion
- RNA-Seq_Transcript
- WGS_QC_TN
- WGS_SV
- WGS_Somatic_Variant
- WGS_CNV_Somatic
- WXS_QC_SampleMatch
- WXS_QC_SMG
- WXS_QC_TN
- WXS_MSI
- WXS_Normal_Adjacent
- WXS_Somatic_Variant
- WXS_Germline