viral-ngs Documentation

Release v2.0.13

Broad Institute Viral Genomics

2019-11-07

CHAPTER 1

Contents

1.1 Description of the methods

This is a base module that provides basic utility functions, some short read aligners, and samtools and Picard.

1.2 Command line tools

1.2.1 reports.py - produce various metrics and reports

Functions to create reports from genomics pipeline data.

```
usage: reports.py subcommand
```

Sub-commands:

assembly_stats

Fetch assembly-level statistics for a given sample

```
usage: reports.py assembly_stats [-h]
                                 [--cov_thresholds COV_THRESHOLDS [COV_
↪THRESHOLDS ...]]
                                 [--assembly_dir ASSEMBLY_DIR]
                                 [--assembly_tmp ASSEMBLY_TMP]
                                 [--align_dir ALIGN_DIR]
                                 [--reads_dir READS_DIR]
                                 [--raw_reads_dir RAW_READS_DIR]
                                 [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                 [--version] [--tmp_dir TMP_DIR]
                                 [--tmp_dirKeep]
                                 samples [samples ...] outFile
```

Positional arguments:

samples	Sample names.
outFile	Output report file.

Options:

--cov_thresholds=(1, 5, 20, 100) Genome coverage thresholds to report on. (default: %(default)s)

--assembly_dir=data/02_assembly Directory with assembly outputs. (default: %(default)s)

--assembly_tmp=tmp/02_assembly Directory with assembly temp files. (default: %(default)s)

--align_dir=data/02_align_to_self Directory with reads aligned to own assembly. (default: %(default)s)

--reads_dir=data/01_per_sample Directory with unaligned filtered read BAMs. (default: %(default)s)

--raw_reads_dir=data/00_raw Directory with unaligned raw read BAMs. (default: %(default)s)

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

coverage_only

```
usage: reports.py coverage_only [-h]
                                [--cov_thresholds COV_THRESHOLDS [COV_
↪THRESHOLDS ...]]
                                [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                [--version] [--tmp_dir TMP_DIR]
                                [--tmp_dirKeep]
                               mapped_bams [mapped_bams ...] out_report
```

Positional arguments:

mapped_bams Aligned-to-self mapped bam files.

out_report Output report file.

Options:

--cov_thresholds=(1, 5, 20, 100) Genome coverage thresholds to report on. (default: %(default)s)

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

alignment_summary

Write or print pairwise alignment summary information for sequences in two FASTA files, including SNPs, ambiguous bases, and indels.

```
usage: reports.py alignment_summary [-h] [--outfileName OUTFILENAME]
                                    [--printCounts]
                                    [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                    [--version] [--tmp_dir TMP_DIR]
                                    [--tmp_dirKeep]
                                    inFastaFileOne inFastaFileTwo
```

Positional arguments:

inFastaFileOne	First fasta file for an alignment
inFastaFileTwo	First fasta file for an alignment

Options:

--outfileName	Output file for counts in TSV format
--printCounts=False	Undocumented
--loglevel=INFO	Verboseness of output. [default: %(default)s] Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION
--version, -V	show program's version number and exit
--tmp_dir=/tmp	Base directory for temp files. [default: %(default)s]
--tmp_dirKeep=False	Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

consolidate_fastqc

Consolidate multiple FASTQC reports into one.

```
usage: reports.py consolidate_fastqc [-h]
                                     [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                     [--version] [--tmp_dir TMP_DIR]
                                     [--tmp_dirKeep]
                                     inDirs [inDirs ...] outFile
```

Positional arguments:

inDirs	Input FASTQC directories.
outFile	Output report file.

Options:

--loglevel=INFO	Verboseness of output. [default: %(default)s] Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION
--version, -V	show program's version number and exit
--tmp_dir=/tmp	Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

consolidate_spike_count

Consolidate multiple spike count reports into one.

```
usage: reports.py consolidate_spike_count [-h]
                                          [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                          [--version] [--tmp_dir TMP_DIR]
                                          [--tmp_dirKeep]
                                          inDir outFile
```

Positional arguments:

in_dir Input spike count directory.

out_file Output report file.

Options:

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

aggregate_spike_count

aggregate multiple spike count reports into one.

```
usage: reports.py aggregate_spike_count [-h]
                                        [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                        [--version] [--tmp_dir TMP_DIR]
                                        [--tmp_dirKeep]
                                        inDir outFile
```

Positional arguments:

in_dir Input spike count directory.

out_file Output report file.

Options:

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

plot_coverage

Generate a coverage plot from an aligned bam file

```
usage: reports.py plot_coverage [-h] [--plotFormat] [--plotDataStyle]
                                [--plotStyle] [--plotWidth PLOT_WIDTH]
                                [--plotHeight PLOT_HEIGHT]
                                [--plotDPI PLOT_DPI] [--plotTitle PLOT_TITLE]
                                [--plotXLimits PLOT_X_LIMITS PLOT_X_LIMITS]
                                [--plotYLimits PLOT_Y_LIMITS PLOT_Y_LIMITS]
                                [-q BASE_Q_THRESHOLD] [-Q MAPPING_Q_THRESHOLD]
                                [-m MAX_COVERAGE_DEPTH]
                                [-l READ_LENGTH_THRESHOLD] [--binLargePlots]
                                [--binningSummaryStatistic {max,min}]
                                [--outSummary OUT_SUMMARY]
                                [--plotOnlyNonDuplicates]
                                [--loglevel
 ↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                [--version] [--tmp_dir TMP_DIR]
                                [--tmp_dirKeep]
                                in_bam out_plot_file
```

Positional arguments:

in_bam Input reads, BAM format.

out_plot_file The generated chart file

Options:

--plotFormat File format of the coverage plot. By default it is inferred from the file extension of out_plot_file, but it can be set explicitly via –plotFormat. Valid formats include: ps, eps, pdf, pgf, png, raw, rgba, svg, svgz, jpg, jpeg, tif, tiff

Possible choices: ps, eps, pdf, pgf, png, raw, rgba, svg, svgz, jpg, jpeg, tif, tiff

--plotDataStyle=filled The plot data display style. Valid options: filled, line, dots (default: %(default)s)

Possible choices: filled, line, dots

--plotStyle=ggplot The plot visual style. Valid options: seaborn-white, seaborn-pastel, seaborn-deep, seaborn-darkgrid, dark_background, seaborn-paper, grayscale, seaborn-muted, tableau-colorblind10, fivethirtyeight, seaborn-poster, seaborn, fast, seaborn-dark, seaborn-whitegrid, bmh, seaborn-bright, seaborn-dark-palette, _classic_test, ggplot, seaborn-notebook, seaborn-colorblind, Solarize_Light2, classic, seaborn-talk, seaborn-ticks (default: %(default)s)

Possible choices: seaborn-white, seaborn-pastel, seaborn-deep, seaborn-darkgrid, dark_background, seaborn-paper, grayscale, seaborn-muted, tableau-colorblind10, fivethirtyeight, seaborn-poster, seaborn, fast, seaborn-dark, seaborn-whitegrid, bmh, seaborn-bright, seaborn-dark-palette, _classic_test, ggplot, seaborn-notebook, seaborn-colorblind, Solarize_Light2, classic, seaborn-talk, seaborn-ticks

--plotWidth=880 Width of the plot in pixels (default: %(default)s)

--plotHeight=680 Width of the plot in pixels (default: %(default)s)

--plotDPI=100.0 dots per inch for rendered output, more useful for vector modes (default: %(default)s)

--plotTitle=Coverage Plot The title displayed on the coverage plot (default: '%(default)s')

--plotXLimits Limits on the x-axis of the coverage plot; args are '<min> <max>'

--plotYLimits Limits on the y-axis of the coverage plot; args are '<min> <max>'

-q The minimum base quality threshold

-Q The minimum mapping quality threshold

-m The max coverage depth (default: %(default)s)

-l Read length threshold

--binLargePlots=False Plot summary read depth in one-pixel-width bins for large plots.

--binningSummaryStatistic=max Statistic used to summarize each bin (max or min).

Possible choices: max, min

--outSummary Coverage summary TSV file. Default is to write to temp.

--plotOnlyNonDuplicates=False Plot only non-duplicates (samtools -F 1024), coverage counted by bedtools rather than samtools.

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

align_and_plot_coverage

Take reads, align to reference with BWA-MEM, and generate a coverage plot

```
usage: reports.py align_and_plot_coverage [-h] [--plotFormat]
                                          [--plotDataStyle] [--plotStyle]
                                          [--plotWidth PLOT_WIDTH]
                                          [--plotHeight PLOT_HEIGHT]
                                          [--plotDPI PLOT_DPI]
                                          [--plotTitle PLOT_TITLE]
                                          [--plotXLimits PLOT_X_LIMITS PLOT_X_
↪LIMITS]
                                          [--plotYLimits PLOT_Y_LIMITS PLOT_Y_
↪LIMITS]
                                          [-q BASE_Q_THRESHOLD]
                                          [-Q MAPPING_Q_THRESHOLD]
                                          [-m MAX_COVERAGE_DEPTH]
                                          [-l READ_LENGTH_THRESHOLD]
                                          [--binLargePlots]
```

```
                                          [--binningSummaryStatistic {max,min}
↪]
                                          [--outSummary OUT_SUMMARY]
                                          [--outBam OUT_BAM] [--sensitive]
                                          [--excludeDuplicates]
                                          [--JVMmemory JVMMEMORY]
                                          [--picardOptions [PICARDOPTIONS␣
↪[PICARDOPTIONS ...]]]
                                          [--minScoreToFilter MIN_SCORE_TO_
↪FILTER]
                                          [--aligner {novoalign,bwa}]
                                          [--aligner_options ALIGNER_OPTIONS]
                                          [--NOVOALIGN_LICENSE_PATH NOVOALIGN_
↪LICENSE_PATH]
                                          [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                          [--version] [--tmp_dir TMP_DIR]
                                          [--tmp_dirKeep]
                                          in_bam out_plot_file ref_fasta
```

Positional arguments:

in_bam Input reads, BAM format.

out_plot_file The generated chart file

ref_fasta Reference genome, FASTA format.

Options:

--plotFormat File format of the coverage plot. By default it is inferred from the file extension of out_plot_file, but it can be set explicitly via –plotFormat. Valid formats include: ps, eps, pdf, pgf, png, raw, rgba, svg, svgz, jpg, jpeg, tif, tiff

Possible choices: ps, eps, pdf, pgf, png, raw, rgba, svg, svgz, jpg, jpeg, tif, tiff

--plotDataStyle=filled The plot data display style. Valid options: filled, line, dots (default: %(default)s)

Possible choices: filled, line, dots

--plotStyle=ggplot The plot visual style. Valid options: seaborn-white, seaborn-pastel, seaborn-deep, seaborn-darkgrid, dark_background, seaborn-paper, grayscale, seaborn-muted, tableau-colorblind10, fivethirtyeight, seaborn-poster, seaborn, fast, seaborn-dark, seaborn-whitegrid, bmh, seaborn-bright, seaborn-dark-palette, _classic_test, ggplot, seaborn-notebook, seaborn-colorblind, Solarize_Light2, classic, seaborn-talk, seaborn-ticks (default: %(default)s)

Possible choices: seaborn-white, seaborn-pastel, seaborn-deep, seaborn-darkgrid, dark_background, seaborn-paper, grayscale, seaborn-muted, tableau-colorblind10, fivethirtyeight, seaborn-poster, seaborn, fast, seaborn-dark, seaborn-whitegrid, bmh, seaborn-bright, seaborn-dark-palette, _classic_test, ggplot, seaborn-notebook, seaborn-colorblind, Solarize_Light2, classic, seaborn-talk, seaborn-ticks

--plotWidth=880 Width of the plot in pixels (default: %(default)s)

--plotHeight=680 Width of the plot in pixels (default: %(default)s)

--plotDPI=100.0 dots per inch for rendered output, more useful for vector modes (default: %(default)s)

--plotTitle=Coverage Plot The title displayed on the coverage plot (default: '%(default)s')

--plotXLimits Limits on the x-axis of the coverage plot; args are '<min> <max>'

--plotYLimits Limits on the y-axis of the coverage plot; args are '<min> <max>'

-q The minimum base quality threshold

-Q The minimum mapping quality threshold

-m The max coverage depth (default: %(default)s)

-l Read length threshold

--binLargePlots=False Plot summary read depth in one-pixel-width bins for large plots.

--binningSummaryStatistic=max Statistic used to summarize each bin (max or min).

Possible choices: max, min

--outSummary Coverage summary TSV file. Default is to write to temp.

--outBam Output aligned, indexed BAM file. Default is to write to temp.

--sensitive=False Equivalent to giving bwa: '-k 12 -A 1 -B 1 -O 1 -E 1'. Only relevant if the bwa aligner is selected (the default).

--excludeDuplicates=False MarkDuplicates with Picard and only plot non-duplicates

--JVMmemory=2g JVM virtual memory size (default: %(default)s)

--picardOptions=[] Optional arguments to Picard's MarkDuplicates, OPTION-NAME=value ...

--minScoreToFilter Filter bwa alignments using this value as the minimum allowed alignment score. Specifically, sum the alignment scores across all alignments for each query (including reads in a pair, supplementary and secondary alignments) and then only include, in the output, queries whose summed alignment score is at least this value. This is only applied when the aligner is 'bwa'. The filtering on a summed alignment score is sensible for reads in a pair and supplementary alignments, but may not be reasonable if bwa outputs secondary alignments (i.e., if '-a' is in the aligner options). (default: not set - i.e., do not filter bwa's output)

--aligner=bwa aligner (default: %(default)s)

Possible choices: novoalign, bwa

--aligner_options aligner options (default for novoalign: "-r Random -l 40 -g 40 -x 20 -t 100 -k", bwa: bwa defaults

--NOVOALIGN_LICENSE_PATH A path to the novoalign.lic file. This overrides the NOVOALIGN_LICENSE_PATH environment variable. (default: %(default)s)

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

fastqc

```
usage: reports.py fastqc [-h] inBam outHtml
```

Positional arguments:

inBam Input reads, BAM format.

outHtml Output report, HTML format.

1.2.2 illumina.py - for raw Illumina outputs

Utilities for demultiplexing Illumina data.

```
usage: illumina.py subcommand
```

Sub-commands:

illumina_demux

Read Illumina runs & produce BAM files, demultiplexing to one bam per sample, or for simplex runs, a single bam will be produced bearing the flowcell ID. Wraps together Picard's ExtractBarcodes (for multiplexed samples) and IlluminaBasecallsToSam while handling the various required input formats. Also can read Illumina BCL directories, tar.gz BCL directories.

```
usage: illumina.py illumina_demux [-h] [--outMetrics OUTMETRICS]
                                  [--commonBarcodes COMMONBARCODES]
                                  [--sampleSheet SAMPLESHEET]
                                  [--runInfo RUNINFO] [--flowcell FLOWCELL]
                                  [--read_structure READ_STRUCTURE]
                                  [--max_mismatches MAX_MISMATCHES]
                                  [--minimum_base_quality MINIMUM_BASE_
↪QUALITY]
                                  [--min_mismatch_delta MIN_MISMATCH_DELTA]
                                  [--max_no_calls MAX_NO_CALLS]
                                  [--minimum_quality MINIMUM_QUALITY]
                                  [--compress_outputs COMPRESS_OUTPUTS]
                                  [--sequencing_center SEQUENCING_CENTER]
                                  [--adapters_to_check [ADAPTERS_TO_CHECK␣
↪[ADAPTERS_TO_CHECK ...]]]
                                  [--platform PLATFORM]
                                  [--max_reads_in_ram_per_tile MAX_READS_IN_
↪RAM_PER_TILE]
                                  [--max_records_in_ram MAX_RECORDS_IN_RAM]
```

```
                                    [--apply_eamss_filter APPLY_EAMSS_FILTER]
                                    [--force_gc FORCE_GC]
                                    [--first_tile FIRST_TILE]
                                    [--tile_limit TILE_LIMIT]
                                    [--include_non_pf_reads INCLUDE_NON_PF_
↪READS]
                                    [--run_start_date RUN_START_DATE]
                                    [--read_group_id READ_GROUP_ID]
                                    [--compression_level COMPRESSION_LEVEL]
                                    [--JVMmemory JVMMEMORY] [--threads THREADS]
                                    [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                    [--version] [--tmp_dir TMP_DIR]
                                    [--tmp_dirKeep]
                                    inDir lane outDir
```

Positional arguments:

inDir Illumina BCL directory (or tar.gz of BCL directory). This is the top-level run directory.

lane Lane number.

outDir Output directory for BAM files.

Options:

--outMetrics Output ExtractIlluminaBarcodes metrics file. Default is to dump to a temp file.

--commonBarcodes Write a TSV report of all barcode counts, in descending order. Only applicable for read structures containing "B"

--sampleSheet Override SampleSheet. Input tab or CSV file w/header and four named columns: barcode_name, library_name, barcode_sequence_1, barcode_sequence_2. Default is to look for a SampleSheet.csv in the inDir.

--runInfo Override RunInfo. Input xml file. Default is to look for a RunInfo.xml file in the inDir.

--flowcell Override flowcell ID (default: read from RunInfo.xml).

--read_structure Override read structure (default: read from RunInfo.xml).

--max_mismatches=0 Picard ExtractIlluminaBarcodes MAX_MISMATCHES (default: %(default)s)

--minimum_base_quality=20 Picard ExtractIlluminaBarcodes MINIMUM_BASE_QUALITY (default: %(default)s)

--min_mismatch_delta Picard ExtractIlluminaBarcodes MIN_MISMATCH_DELTA (default: %(default)s)

--max_no_calls Picard ExtractIlluminaBarcodes MAX_NO_CALLS (default: %(default)s)

--minimum_quality Picard ExtractIlluminaBarcodes MINIMUM_QUALITY (default: %(default)s)

--compress_outputs Picard ExtractIlluminaBarcodes COMPRESS_OUTPUTS (default: %(default)s)

--sequencing_center Picard IlluminaBasecallsToSam SEQUENCING_CENTER (default: %(default)s)

--adapters_to_check=('PAIRED_END', 'NEXTERA_V1', 'NEXTERA_V2') Picard IlluminaBasecallsToSam ADAPTERS_TO_CHECK (default: %(default)s)

--platform Picard IlluminaBasecallsToSam PLATFORM (default: %(default)s)

--max_reads_in_ram_per_tile=1000000 Picard IlluminaBasecallsToSam MAX_READS_IN_RAM_PER_TILE (default: %(default)s)

--max_records_in_ram=2000000 Picard IlluminaBasecallsToSam MAX_RECORDS_IN_RAM (default: %(default)s)

--apply_eamss_filter Picard IlluminaBasecallsToSam APPLY_EAMSS_FILTER (default: %(default)s)

--force_gc Picard IlluminaBasecallsToSam FORCE_GC (default: %(default)s)

--first_tile Picard IlluminaBasecallsToSam FIRST_TILE (default: %(default)s)

--tile_limit Picard IlluminaBasecallsToSam TILE_LIMIT (default: %(default)s)

--include_non_pf_reads=False Picard IlluminaBasecallsToSam INCLUDE_NON_PF_READS (default: %(default)s)

--run_start_date Picard IlluminaBasecallsToSam RUN_START_DATE (default: %(default)s)

--read_group_id Picard IlluminaBasecallsToSam READ_GROUP_ID (default: %(default)s)

--compression_level=7 Picard IlluminaBasecallsToSam COMPRESSION_LEVEL (default: %(default)s)

--JVMmemory=7g JVM virtual memory size (default: %(default)s)

--threads=0 Number of threads (default: 0)

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

lane_metrics

Write out lane metrics to a tsv file.

```
usage: illumina.py lane_metrics [-h] [--read_structure READ_STRUCTURE]
                                [--JVMmemory JVMMEMORY]
                                [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                [--version] [--tmp_dir TMP_DIR]
```

```
                                    [--tmp_dirKeep]
                                    inDir outPrefix
```

Positional arguments:

inDir Illumina BCL directory (or tar.gz of BCL directory). This is the top-level run directory.

outPrefix Prefix path to the *.illumina_lane_metrics and *.illumina_phasing_metrics files.

Options:

--read_structure Override read structure (default: read from RunInfo.xml).

--JVMmemory=8g JVM virtual memory size (default: %(default)s)

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

common_barcodes

Extract Illumina barcodes for a run and write a TSV report of the barcode counts in descending order

```
usage: illumina.py common_barcodes [-h] [--truncateToLength TRUNCATETOLENGTH]
                                   [--omitHeader] [--includeNoise]
                                   [--outMetrics OUTMETRICS]
                                   [--sampleSheet SAMPLESHEET]
                                   [--flowcell FLOWCELL]
                                   [--read_structure READ_STRUCTURE]
                                   [--max_mismatches MAX_MISMATCHES]
                                   [--minimum_base_quality MINIMUM_BASE_
 ↪QUALITY]
                                   [--min_mismatch_delta MIN_MISMATCH_DELTA]
                                   [--max_no_calls MAX_NO_CALLS]
                                   [--minimum_quality MINIMUM_QUALITY]
                                   [--compress_outputs COMPRESS_OUTPUTS]
                                   [--JVMmemory JVMMEMORY]
                                   [--loglevel
 ↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                   [--version] [--tmp_dir TMP_DIR]
                                   [--tmp_dirKeep]
                                   inDir lane outSummary
```

Positional arguments:

inDir Illumina BCL directory (or tar.gz of BCL directory). This is the top-level run directory.

lane Lane number.

outSummary Path to the summary file (.tsv format). It includes several columns: (barcode1, likely_index_name1, barcode2, likely_index_name2, count), where likely index names are either

the exact match index name for the barcode sequence, or those Hamming distance of 1 away.

Options:

--truncateToLength If specified, only this number of barcodes will be returned. Useful if you only want the top N barcodes.

--omitHeader=False If specified, a header will not be added to the outSummary tsv file.

--includeNoise=False If specified, barcodes with periods (".") will be included.

--outMetrics Output ExtractIlluminaBarcodes metrics file. Default is to dump to a temp file.

--sampleSheet Override SampleSheet. Input tab or CSV file w/header and four named columns: barcode_name, library_name, barcode_sequence_1, barcode_sequence_2. Default is to look for a SampleSheet.csv in the inDir.

--flowcell Override flowcell ID (default: read from RunInfo.xml).

--read_structure Override read structure (default: read from RunInfo.xml).

--max_mismatches=0 Picard ExtractIlluminaBarcodes MAX_MISMATCHES (default: %(default)s)

--minimum_base_quality=20 Picard ExtractIlluminaBarcodes MINIMUM_BASE_QUALITY (default: %(default)s)

--min_mismatch_delta Picard ExtractIlluminaBarcodes MIN_MISMATCH_DELTA (default: %(default)s)

--max_no_calls Picard ExtractIlluminaBarcodes MAX_NO_CALLS (default: %(default)s)

--minimum_quality Picard ExtractIlluminaBarcodes MINIMUM_QUALITY (default: %(default)s)

--compress_outputs Picard ExtractIlluminaBarcodes COMPRESS_OUTPUTS (default: %(default)s)

--JVMmemory=8g JVM virtual memory size (default: %(default)s)

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

guess_barcodes

Guess the barcode value for a sample name, based on the following: - a list is made of novel barcode pairs seen in the data, but not in the picard metrics - for the sample in question, get the most abundant novel barcode pair where one of the barcodes seen in the data matches one of the barcodes in the picard metrics (partial match) - if there are no partial matches, get the most abundant novel barcode pair

Limitations: - If multiple samples share a barcode with multiple novel barcodes, disentangling them is difficult or impossible

The names of samples to guess are selected: - explicitly by name, passed via argument, OR - explicitly by read count threshold, OR - automatically (if names or count threshold are omitted) based on basic outlier detection of deviation from an assumed-balanced pool with some number of negative controls

```
usage: illumina.py guess_barcodes [-h]
                                  [--readcount_threshold READCOUNT_THRESHOLD␣
↪| --sample_names [SAMPLE_NAMES [SAMPLE_NAMES ...]]]
                                  [--outlier_threshold OUTLIER_THRESHOLD]
                                  [--expected_assigned_fraction EXPECTED_
↪ASSIGNED_FRACTION]
                                  [--number_of_negative_controls NUMBER_OF_
↪NEGATIVE_CONTROLS]
                                  [--rows_limit ROWS_LIMIT]
                                  [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                  [--version] [--tmp_dir TMP_DIR]
                                  [--tmp_dirKeep]
                                  in_barcodes in_picard_metrics
                                  out_summary_tsv
```

Positional arguments:

in_barcodes The barcode counts file produced by common_barcodes.

in_picard_metrics The demultiplexing read metrics produced by Picard.

out_summary_tsv Path to the summary file (.tsv format). It includes several columns: (sample_name, expected_barcode_1, expected_barcode_2, expected_barcode_1_name, expected_barcode_2_name, expected_barcodes_read_count, guessed_barcode_1, guessed_barcode_2, guessed_barcode_1_name, guessed_barcode_2_name, guessed_barcodes_read_count, match_type), where the expected values are those used by Picard during demultiplexing and the guessed values are based on the barcodes seen among the data.

Options:

--readcount_threshold If specified, guess barcodes for samples with fewer than this many reads.

--sample_names If specified, only guess barcodes for these sample names.

--outlier_threshold=0.675 threshold of how far from unbalanced a sample must be to be considered an outlier.

--expected_assigned_fraction=0.7 The fraction of reads expected to be assigned. An exception is raised if fewer than this fraction are assigned.

--number_of_negative_controls=1 The number of negative controls in the pool, for calculating expected number of reads in the rest of the pool.

--rows_limit=1000 The number of rows to use from the in_barcodes.

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

miseq_fastq_to_bam

Convert fastq read files to a single bam file. Fastq file names must conform to patterns emitted by Miseq machines. Sample metadata must be provided in a SampleSheet.csv that corresponds to the fastq filename. Specifically, the _S##_ index in the fastq file name will be used to find the corresponding row in the SampleSheet

```
usage: illumina.py miseq_fastq_to_bam [-h] [--inFastq2 INFASTQ2]
                                      [--runInfo RUNINFO]
                                      [--sequencing_center SEQUENCING_CENTER]
                                      [--JVMmemory JVMMEMORY]
                                      [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                      [--version] [--tmp_dir TMP_DIR]
                                      [--tmp_dirKeep]
                                      outBam sampleSheet inFastq1
```

Positional arguments:

outBam Output BAM file.

sampleSheet Input SampleSheet.csv file.

inFastq1 Input fastq file; 1st end of paired-end reads if paired.

Options:

--inFastq2 Input fastq file; 2nd end of paired-end reads.

--runInfo Input RunInfo.xml file.

--sequencing_center Name of your sequencing center (default is the sequencing machine ID from the RunInfo.xml)

--JVMmemory=2g JVM virtual memory size (default: %(default)s)

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

extract_fc_metadata

Extract RunInfo.xml and SampleSheet.csv from the provided Illumina directory

```
usage: illumina.py extract_fc_metadata [-h]
                                       [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                       [--version] [--tmp_dir TMP_DIR]
                                       [--tmp_dirKeep]
                                       flowcell outRunInfo outSampleSheet
```

Positional arguments:

flowcell	Illumina directory (possibly tarball)
outRunInfo	Output RunInfo.xml file.
outSampleSheet	Output SampleSheet.csv file.

Options:

--loglevel=INFO	Verboseness of output. [default: %(default)s] Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION
--version, -V	show program's version number and exit
--tmp_dir=/tmp	Base directory for temp files. [default: %(default)s]
--tmp_dirKeep=False	Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.

1.2.3 broad_utils.py - for data generated at the Broad Institute

Utilities for getting sequences out of the Broad walk-up sequencing pipeline. These utilities are probably not of much use outside the Broad.

```
usage: broad_utils.py subcommand
```

Sub-commands:

get_bustard_dir

Find the basecalls directory from a Picard directory

```
usage: broad_utils.py get_bustard_dir [-h]
                                      [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                      inDir
```

Positional arguments:

inDir	Picard directory

Options:

--loglevel=ERROR	Verboseness of output. [default: %(default)s] Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

get_run_date

Find the sequencing run date from a Picard directory

```
usage: broad_utils.py get_run_date [-h]
                                   [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                   inDir
```

Positional arguments:

inDir	Picard directory

Options:

--loglevel=ERROR Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

get_all_names

Get all samples

```
usage: broad_utils.py get_all_names [-h]
                                    [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                    {samples,libraries,runs} runfile
```

Positional arguments:

type Type of name

Possible choices: samples, libraries, runs

runfile File with seq run information

Options:

--loglevel=ERROR Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

1.2.4 file_utils.py - utilities to perform various file manipulations

Utilities for dealing with files.

```
usage: file_utils.py subcommand
```

Sub-commands:

merge_tarballs

Merges separate tarballs into one tarball data can be piped in and/or out

```
usage: file_utils.py merge_tarballs [-h]
                                    [--extractToDiskPath EXTRACT_TO_DISK_PATH]
                                    [--pipeInHint PIPE_HINT_IN]
                                    [--pipeOutHint PIPE_HINT_OUT]
                                    [--threads THREADS]
                                    [--loglevel
↪{DEBUG,INFO,WARNING,ERROR,CRITICAL,EXCEPTION}]
                                    [--version] [--tmp_dir TMP_DIR]
                                    [--tmp_dirKeep]
                                    out_tarball in_tarballs [in_tarballs ...]
```

Positional arguments:

out_tarball output tarball (*.tar.gz|*.tar.lz4|*.tar.bz2|*.tar.zst|-); compression is inferred by the file extension. Note: if "-" is used, output will be written to stdout and –pipeOutHint must be provided to indicate compression type when compression type is not gzip (gzip is used by default).

in_tarballs input tarballs (*.tar.gz|*.tar.lz4|*.tar.bz2|*.tar.zst)

Options:

--extractToDiskPath If specified, the tar contents will also be extracted to a local directory.

--pipeInHint=gz If specified, the compression type used is used for piped input.

--pipeOutHint=gz If specified, the compression type used is used for piped output.

--threads Number of threads (default: all available cores)

--loglevel=INFO Verboseness of output. [default: %(default)s]

Possible choices: DEBUG, INFO, WARNING, ERROR, CRITICAL, EXCEPTION

--version, -V show program's version number and exit

--tmp_dir=/tmp Base directory for temp files. [default: %(default)s]

--tmp_dirKeep=False Keep the tmp_dir if an exception occurs while running. Default is to delete all temp files at the end, even if there's a failure.