PureCN - Quick Start

This tutorial provides a quick overview of the command line tools shipping with *PureCN*. For the R package and more detailed information, see the main vignette.

Prepare environment and files

- Get the path to command line scripts in R:

```
system.file("extdata", package="PureCN")

## [1] "/tmp/RtmpCJZ9Ur/Rinst26f74ee6d0c/PureCN/extdata"
```

- Store this path in an environment variable, for example in BASH:

```
$ export PURECN="/path/to/PureCN/extdata"
$ Rscript $PURECN/PureCN.R --help
Usage: /path/to/PureCN/inst/extdata/PureCN.R [-[-help|h]] ...
```

- Generate a basic interval file from a BED file containing target coordinates:

```
$ Rscript $PURECN/IntervalFile.R --infile baits_hg19.bed \
    --fasta hg19.fa --outfile baits_hg19_gcgene.txt
```

  Internally, this script uses *rtracklayer* to parse the `infile`. Make sure that the file format matches the file extension.

  See the main vignette how to add gene symbols to the interval file. Symbols are necessary to obtain gene-level copy number and LOH calls. For a test run, you will not need this.

Run PureCN with third-party segmentation

If you already have a segmentation from third-party tools (for example CNVkit, EXCAVATOR2). For a test run:

```
Rscript $PURECN/PureCN.R --outdir $OUT/$SAMPLEID  \
    --sampleid $SAMPLEID \
    --segfile $OUT/$SAMPLEID/${SAMPLEID}_cnvkit.seg \
    --vcf ${SAMPLEID}_mutect.vcf \
    --genome hg19 --gcgene baits_hg19_gcgene.txt
```

The main VCF (`-vcf`) is ideally created by *MuTect* 1.1.7. Support for *MuTect 2* and *FreeBayes* is available, but poorly tested and only very limited artifact filtering will be performed for these callers.

For a production pipeline run we provide a bit more information about the assay and genome:

```
Rscript $PURECN/PureCN.R --outdir $OUT/$SAMPLEID  \
    --sampleid $SAMPLEID \
    --segfile $OUT/$SAMPLEID/${SAMPLEID}_cnvkit.seg \
    --normal_panel $NORMAL_PANEL \
    --vcf ${SAMPLEID}_mutect.vcf \
```

```
    --statsfile ${SAMPLEID}_mutect_stats.txt \
    --snpblacklist hg19_simpleRepeats.bed \
    --genome hg19 --gcgene baits_hg19_gcgene.txt \
#    --funsegmentation none \
    --force --postoptimize
```

The normal panel VCF file is useful for mapping bias correction and especially recommended without matched normals. See the FAQ of the main vignette how to generate this file. It is not essential for test runs. The *MuTect* 1.1.7 stats file (the main output file besides the VCF) should be provided for better artifact filtering.

The `-funsegmentation` argument controls if the data should to be re-segmented using germline BAFs (default). Set this value to `none` if the provided segmentation should be used as is.

The `-postoptimize` flag defines that purity should be optimized using both variant allelic fractions and copy number instead of copy number only. This results in a significant runtime increase for whole-exome data.

Run PureCN with internal segmentation

The following describes *PureCN* runs with internal copy number normalization and segmentation. Provided are again minimal examples for test runs. See the main vignette how to get optimal results in production pipelines.

Coverage

For each sample, tumor and normal:

```
# From a BAM file
$ Rscript $PURECN/Coverage.R --outdir $OUT/$SAMPLEID \
    --bam ${SAMPLEID}.bam \
    --gcgene baits_hg19_gcgene.txt

# From a GATK DepthOfCoverage file
Rscript $PURECN/Coverage.R --outdir $OUT/$SAMPLEID \
    --gatkcoverage ${SAMPLEID}.coverage.sample_interval_summary \
    --gcgene  baits_hg19_gcgene.txt
```

NormalDB

To build a normal database, copy all GC-normalized normal coverage files in a single text file, line-by-line:

```
ls -a normal*loess.txt | cat > example_normal.list

# From already GC-normalized files
$ Rscript $PURECN/NormalDB.R --outdir $OUT \
    --coveragefiles example_normal.list \
    --genome hg19
```

PureCN

```
cd $OUT/$SAMPLEID
# From GC-normalized coverage data
$ Rscript $PURECN/PureCN.R --outdir . --tumor ${SAMPLEID}_coverage_loess.txt \
    --normal ${SAMPLEID_NORMAL}_coverage_loess.txt \
    --sampleid $SAMPLEID \
    --vcf ${SAMPLEID}_mutect.vcf \
    --genome hg19 \
    --gcgene baits_hg19_gcgene.txt

# Without a matched normal
$ Rscript $PURECN/PureCN.R --outdir . --tumor ${SAMPLEID}_coverage_loess.txt \
    --normaldb ../normalDB_hg19.rds \
    --sampleid $SAMPLEID \
    --vcf ${SAMPLEID}_mutect.vcf \
    --pool 5 \
    --genome hg19 --gcgene baits_hg19_gcgene.txt

# Recreate output after manual curation of Sample_purecn.csv
$ Rscript $PURECN/PureCN.R --rds ${SAMPLEID}_purecn.rds
```

Session Info

- R version 3.4.0 (2017-04-21), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Running under: `Ubuntu 16.04.2 LTS`
- Matrix products: default
- BLAS: `/home/biocbuild/bbs-3.5-bioc/R/lib/libRblas.so`
- LAPACK: `/home/biocbuild/bbs-3.5-bioc/R/lib/libRlapack.so`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.38.0, Biobase 2.36.2, BiocGenerics 0.22.0, Biostrings 2.44.0, DNAcopy 1.50.1, DelayedArray 0.2.4, GenomeInfoDb 1.12.1, GenomicFeatures 1.28.0, GenomicRanges 1.28.3, IRanges 2.10.2, PureCN 1.6.3, Rsamtools 1.28.0, S4Vectors 0.14.2, SummarizedExperiment 1.6.2, TxDb.Hsapiens.UCSC.hg19.knownGene 3.2.2, VariantAnnotation 1.22.1, XVector 0.16.0, matrixStats 0.52.2, org.Hs.eg.db 3.4.1
- Loaded via a namespace (and not attached): BSgenome 1.44.0, BiocParallel 1.10.1, BiocStyle 2.4.0, DBI 0.6-1, GenomeInfoDbData 0.99.0, GenomicAlignments 1.12.1, Matrix 1.2-10, RColorBrewer 1.1-2, RCurl 1.95-4.8, RSQLite 1.1-2, Rcpp 0.12.11, VGAM 1.0-3, XML 3.98-1.7, backports 1.1.0, biomaRt 2.32.0, bitops 1.0-6, colorspace 1.3-2, compiler 3.4.0, data.table 1.10.4, digest 0.6.12, edgeR 3.18.1,

evaluate 0.10, futile.logger 1.4.3, futile.options 1.0.0, ggplot2 2.2.1, grid 3.4.0, gtable 0.2.0, highr 0.6, htmltools 0.3.6, knitr 1.16, labeling 0.3, lambda.r 1.1.9, lattice 0.20-35, lazyeval 0.2.0, limma 3.32.2, locfit 1.5-9.1, magrittr 1.5, memoise 1.1.0, munsell 0.4.3, plyr 1.8.4, rlang 0.1.1, rmarkdown 1.5, rprojroot 1.2, rtracklayer 1.36.3, scales 0.4.1, splines 3.4.0, stringi 1.1.5, stringr 1.2.0, tibble 1.3.1, tools 3.4.0, yaml 2.1.14, zlibbioc 1.22.0