A quick introduction to AneuFinder

Aaron Taudt*

March 7, 2016

Contents

*aaron.taudt@gmail.com

1 Introduction

AneuFinder offers functionality for the study of copy number variations (CNV) in whole-genome single cell sequencing (WGSCS) data. Functionality implemented in this package includes:

- CNV detection using a Hidden Markov Model on binned read counts.
- Various plotting capabilities like genomewide heatmaps of CNV state and arrayCGH-like plots.
- Export of CNV calls in BED format for upload to the UCSC genome browser.
- Quality metrics.
- Measures for addressing karyotype heterogeneity.

2 Quickstart

The main function of this package is called `Aneufinder`[1] and performs all the necessary steps to get from aligned reads to interpretable output:

```
Aneufinder(inputfolder='folder-with-BAM', outputfolder='output-directory',
           format='bam', numCPU=2)
```

Although in most cases the above command will produce reasonably good results, it might be worthwile to adjust the default parameters to improve performance and the quality of the results (see section 3). You can get a description of all available parameters by typing

```
?Aneufinder
```

After the function has finished, you will find the folder <output-directory> containing all produced files and plots. This folder contains the following items and subfolders:

- AneuFinder.config: This file contains all the parameters that are necessary to reproduce your analysis. You can specify this file as

  ```
  Aneufinder(..., configfile='AneuFinder.config')
  ```

  to run another analysis with the same parameter settings.

[1] This function can also be run from command line, please see the INSTALL.md in the source package for details.

- binned: This folder contains the binned data. If you chose a correction method, you will also see a folder like 'binned-GC' in case of GC correction. You can load the data with

```
files <- list.files('output-directory/binned', full.names=TRUE)
binned.data <- loadGRangesFromFiles(files)
```

- browserfiles_data: This folder contains BED files with mapped reads that can be uploaded to the UCSC genome browser. The BED files contain the same reads as your input but filtered by mapping quality and other parameter settings that you can find in section [Binning] of the "AneuFinder.config" file.
- browserfiles: A folder which contains BED files with CNV calls that can be uploaded to the UCSC genome browser.
- data: This folder stores all the read data as RData objects. This exists mostly for internal usage.
- hmms: A folder with all produced Hidden Markov Models. You can load the results for further processing, such as quality control and customized plotting.

```
files <- list.files('output-directory/hmms', full.names=TRUE)
hmms <- loadHmmsFromFiles(files)
cl <- clusterByQuality(hmms)
heatmapGenomewide(cl$classification[[1]])
```

- plots: All plots that are produced by default will be stored here.

3 A detailed workflow

3.1 Mappability correction

The first step of your workflow should be the production of a reference file for mappability correction. Mappability correction is done via a variable-width binning approach (as compared to fixed-width bins) and requires a euploid reference. You can either simulate this reference file or take a real euploid reference. For optimal results we suggest to use a real reference, e.g. by merging BAM files of single cells from a euploid reference tissue. This can be achieved with the 'samtools merge' command (not part of R). Be careful: All CNVs that are present in the reference will lead to artifacts in the analysis later. This includes sex-chromosomes that are present in one copy only, so we advice to use a female reference and to exclude the Y-chromosome from the analysis. If you have no reference available, you can simulate one with the `simulateReads` command:

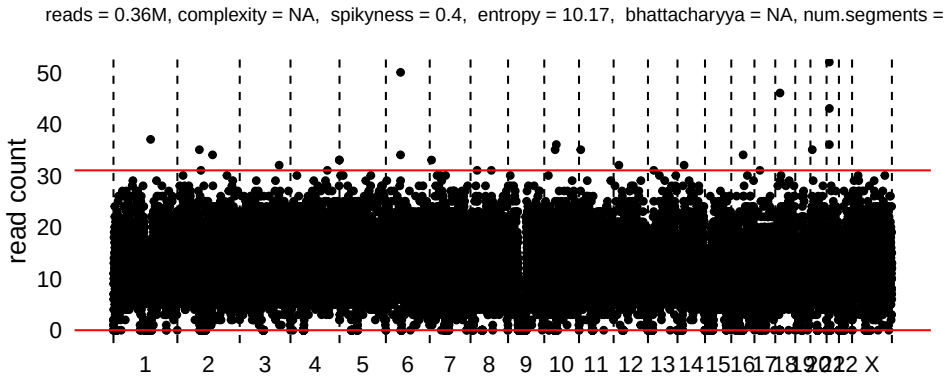
```
## Load human genome
library(BSgenome.Hsapiens.UCSC.hg19)

## Get a BAM file for the estimation of quality scores
bamfile <- system.file("extdata/example.bam", package="AneuFinder")

## Simulate reads of length 51bp for human genome
outputfile <- tempfile() # replace this with your destination file
simulateReads(BSgenome.Hsapiens.UCSC.hg19, readLength=51, bamfile=bamfile,
              file=outputfile, every.X.bp=500)
```

This simulated FASTQ file must then be aligned with your aligner of choice (ideally the same that you used for your other samples) and given as reference in the `Aneufinder` function (option `variable.width.reference`).

3.2 Blacklisting

To further improve the quality of the results and remove artifacts caused by high mappability repeat regions, e.g. near centromers, a blacklist can be used in option `blacklist` of the `Aneufinder` function. All reads falling into the regions specified by the blacklist will be discarded when importing the read files. You can either download a blacklist from the UCSC genome browser, e.g. the "DAC Blacklisted Regions from ENCODE/DAC(Kundaje)" mappability track, or make your own. For optimal results, we advice to make your own blacklist from a euploid reference. The following code chunck takes a euploid reference and makes fixed-width bins of 100kb. Bins with read count above and beloow the 0.999 and 0.05 quantile are taken as blacklist:

```
## Get a euploid reference
bedfile <- system.file("extdata/hg19_euploid.bam.bed.gz", package="AneuFinder")
## Make 100kb fixed-width bins
bins <- binReads(bedfile, format='bed', assembly='hg19',
                 binsizes=100e3, chromosomes=c(1:22,'X'))[[1]]
## Make a plot for visual inspection and get the blacklist
lcutoff <- quantile(bins$counts, 0.05)
ucutoff <- quantile(bins$counts, 0.999)
p <- plot(bins) + coord_cartesian(ylim=c(0,50))
p <- p + geom_hline(aes(yintercept=lcutoff), color='red')
p <- p + geom_hline(aes(yintercept=ucutoff), color='red')
print(p)
```




```
blacklist <- bins[bins$counts < ucutoff & bins$counts > lcutoff]
## Write blacklist to file
exportGRanges(blacklist, filename=tempfile(), header=FALSE,
              chromosome.format='NCBI')
```

3.3 Running Aneufinder

3.4 Quality control

3.5 Karyotype measures

4 Session Info

```
sessionInfo()

## R version 3.2.3 (2015-12-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.4 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=nl_NL.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=nl_NL.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=nl_NL.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=nl_NL.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] AneuFinder_0.99.0    cowplot_0.6.1        ggplot2_2.1.0
##  [4] GenomicRanges_1.22.4 GenomeInfoDb_1.6.3   IRanges_2.4.8
```

```
##  [7] S4Vectors_0.8.11     BiocGenerics_0.16.1  knitr_1.12.3
## [10] devtools_1.10.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.3                highr_0.5.1
##  [3] formatR_1.2.1              futile.logger_1.4.1
##  [5] plyr_1.8.3                 XVector_0.10.0
##  [7] futile.options_1.0.0       bitops_1.0-6
##  [9] iterators_1.0.8            tools_3.2.3
## [11] zlibbioc_1.16.0            mclust_5.1
## [13] digest_0.6.9               evaluate_0.8
## [15] memoise_1.0.0              gtable_0.2.0
## [17] foreach_1.4.3              polynom_1.3-8
## [19] ggdendro_0.1-18            preseqR_2.0.0
## [21] stringr_1.0.0              caTools_1.17.1
## [23] gtools_3.5.0               Biostrings_2.38.4
## [25] grid_3.2.3                 Biobase_2.30.0
## [27] BiocParallel_1.4.3         gdata_2.17.0
## [29] ReorderCluster_1.0         lambda.r_1.1.7
## [31] reshape2_1.4.1             magrittr_1.5
## [33] gplots_2.17.0              scales_0.4.0
## [35] Rsamtools_1.22.0           codetools_0.2-14
## [37] MASS_7.3-45                GenomicAlignments_1.6.3
## [39] SummarizedExperiment_1.0.2 colorspace_1.2-6
## [41] labeling_0.3               KernSmooth_2.23-15
## [43] stringi_1.0-1              munsell_0.4.3
## [45] doParallel_1.0.10
```

```
warnings()
```

```
## NULL
```