A quick introduction to aneufinder

Aaron Taudt*

September 19, 2015

Contents

1 Introduction

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

- CNV detection using a 6-state 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.

*aaron.taudt@gmail.com

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. It takes two compulsory arguments: A folder with BAM files containing your aligned sequencing data, and an output folder where all results and intermediate files will be stored. Running the function will look something like this:

```
library(aneufinder)
Aneufinder(inputfolder='<path-to-BAM-files>', outputfolder='<output-directory>')

## Using 1 CPUs
## Binning the data ...  458.29s
## Finding CNVs ...  493.36s
## Plotting distributions ...  45.02s
## Plotting heatmaps ...  95.61s
## Making arrayCGH plots ...  28.04s
## Plotting karyograms ...  418.52s
## Exporting browser files ...  62.05s
```

After the function has finished, you will find a folder <output-directory> containing all produced files and plots. Although in most cases the above command will produce good results, it can be worth to adjust the default parameters to improve performance and the quality of the results. You can get a description of all available parameters by typing

```
?Aneufinder
```

3 FAQ

4 Session Info

```
sessionInfo()

## R version 3.1.0 (2014-04-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
##
## 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
```

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

```
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] aneufinder_0.9.2     mclust_5.0.2         reshape2_1.4.1
##  [4] ggplot2_1.0.1        GenomicRanges_1.18.4 GenomeInfoDb_1.2.5
##  [7] IRanges_2.0.1        S4Vectors_0.4.0      BiocGenerics_0.12.1
## [10] knitr_1.10.5         devtools_1.8.0
##
## loaded via a namespace (and not attached):
##  [1] base64enc_0.1-2         BatchJobs_1.6
##  [3] BBmisc_1.9              BiocParallel_1.0.3
##  [5] Biostrings_2.34.1       bitops_1.0-6
##  [7] brew_1.0-6              checkmate_1.5.3
##  [9] codetools_0.2-11        colorspace_1.2-6
## [11] DBI_0.3.1               digest_0.6.8
## [13] doParallel_1.0.8        evaluate_0.7
## [15] fail_1.2                foreach_1.4.2
## [17] formatR_1.2             GenomicAlignments_1.2.2
## [19] ggdendro_0.1-17         git2r_0.10.1
## [21] grid_3.1.0              gtable_0.1.2
## [23] highr_0.5               iterators_1.0.7
## [25] magrittr_1.5            MASS_7.3-40
## [27] memoise_0.2.1           munsell_0.4.2
## [29] plyr_1.8.3              preseqR_1.2.1
## [31] proto_0.3-10            Rcpp_0.12.1
## [33] RCurl_1.95-4.6          Rsamtools_1.18.3
## [35] RSQLite_1.0.0           rversions_1.0.0
## [37] scales_0.3.0            sendmailR_1.2-1
## [39] stringi_0.5-5           stringr_1.0.0
## [41] tools_3.1.0             XML_3.98-1.1
## [43] XVector_0.6.0           zlibbioc_1.12.0

warnings()

## NULL
```