CNVPanelizer: Reliable CNV detection in targeted sequencing applications

Tuesday, October 28, 2014

Thomas Wolf,
Pathologie Institut

Cristiano Oliveira,
Pathologie Institut

Abstract

This paper describes the implementation of an R bioconductor package to use targeted sequencing data to reliably detect CNVs from clinical samples. To assess how reliable a change in reads counts in a specific region correlates with the presence of CNVs we implemented an algorithm which uses a subsampling strategy similar to Random Forest to predict the presence of reliable CNVs. We also introduce a novel method to correct for the background noise introduced by sequencing genes with a low number of amplicons. We describe the implementation of these models in the package **CNVPanelizer** and illustrate its usage to reliably detect CNVs on several simulation and real data examples including several code snippets when dealing with clinical data. For a more complete overview of the package's functionality and extensibility see Wolf and Cristiano (2015)....., the manual pages and the reference card.

Keywords: R, Random Forest, CNV, Bootstraping, Panel Sequencing, Ion Torrent.

Introduction

Targeted sequencing, over the last few years, has become a mainstay in the clinical use of next generation sequencing technologies [**?**]. For the detection of somatic and germline SNPs this has been proven to be a highly robust methodology [**?**]. One area of genomic analysis which is usually not covered by targeted sequencing, is the detection of copy number variations (CNVs). While a large number of available algorithms and software address the problem of CNV detection in whole genome or whole exome sequencing, there are no such established tools for targeted sequencing.

Methods

To assess how reliable a change in reads counts in a specific region correlates with the presence of CNVs. To this end we implemented an algorithm which uses a subsampling strategy similar to Random Forest to predict the presence of reliable CNVs. We also introduce a novel method to correct for the background noise introduced by sequencing genes with a low number of amplicons.

Implementation

We could reproduce the CNVs detected by Exome sequencing and the detected CNVs were also found on the Exome level. The CNVs detected by our approach but not by the Exome data was validated by a Taqman assay.

Using

This section provides gives an overview of the package functions, with selected illustrative simulation experiments which illustrate the typical usage of the package available functions.

Instalation Requirements

TODO Perl.. java.. and so on.. not even sure at this point

Installing and Loading the package

The package is available through the Bioconductor repository and could be installed using the following R command: TODO

```
library(AmpCNVrstudio)
```

Required parameters

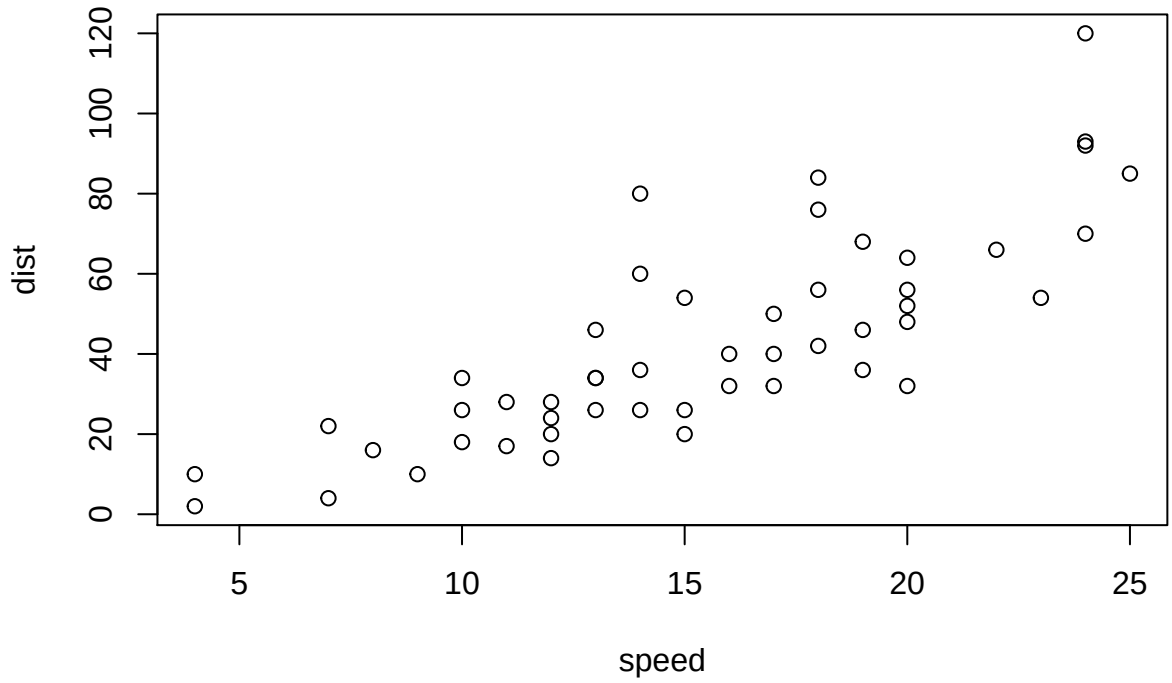
The package is available through the Bioconductor repository and could be installed using the following R command:

```
# Defining some required variables
bedFilepath <- "D:\\repository\\mixedJavaAndGroovy\\testData\\bed\\LCPv1(CNV).bed"
ampliconColumnNumber <- 4
sampleDirectory <- "D:\\repository\\mixedJavaAndGroovy\\testData\\samples"
referenceDirectory <- "D:\\repository\\mixedJavaAndGroovy\\testData\\reference"
outputDirectory <- "D:\\repository\\mixedJavaAndGroovy\\output10"
# Should duplicate reads (same start site,end site and strand be removed,keep one from each)
removePcrDuplicates <- TRUE
# Number of bootstrap replicates
replicates <- 10

referenceFilenames <- list.files(path = referenceDirectory,
                                 pattern = ".bam$",
                                 full.names = TRUE)
sampleFilenames <- list.files(path = sampleDirectory,
                              pattern = ".bam$",
                              full.names = TRUE)

#
#
# ##################################################################
# ## count the reads in the bedfile defined regions
```

```
# ######################################################################
#
# # Extract the information from a bed file
# genomicRangesFromBed <- BedToGenomicRanges(bedFilepath,
#                                            ampliconColumn = ampliconColumnNumber,
#                                            split = "_")
# metadataFromGenomicRanges <- elementMetadata(genomicRangesFromBed)
# geneNames = metadataFromGenomicRanges["geneNames"][, 1]
# ampliconNames = metadataFromGenomicRanges["ampliconNames"][, 1]
#
# # Read the Reference data set
# referenceReadCounts <- ReadCountsFromBam(referenceFilenames,
#                                          sampleNames = referenceFilenames,
#                                          genomicRangesFromBed,
#                                          ampliconNames = ampliconNames,
#                                          removeDup = removePcrDuplicates)
#
# # Read the sample data set
# sampleReadCounts <- ReadCountsFromBam(sampleFilenames,
#                                       sampleNames = sampleFilenames,
#                                       genomicRangesFromBed,
#                                       ampliconNames = ampliconNames,
#                                       removeDup = removePcrDuplicates)
#
# normalizedReadCounts <- CombinedNormalizedCounts(sampleReadCounts,
#                                                  referenceReadCounts,
#                                                  genomicRangesFromBed,
#                                                  amplicons = ampliconNames)
#
#
# ######################################################################
# ## perform the bootstrap based analysis
# ######################################################################
#
#
# samplesNormalizedReadCounts = normalizedReadCounts["samples"][[1]]
# referenceNormalizedReadCounts = normalizedReadCounts["reference"][[1]]
#
# ```
#
# ```{r warning=FALSE, message=FALSE}
# bootList <- BootList(geneNames,
#                      samplesNormalizedReadCounts,
#                      referenceNormalizedReadCounts,
#                      reps = replicates,
#                      refWeights = NULL)
# ```
#
# ```{r}
# ######################################################################
# ## estimate the background noise left after normalization
# ######################################################################
#
```

```
# sigList = CheckSignificance(bootList)
# backgroundNoise <- Background(geneNames,
#                                samplesNormalizedReadCounts,
#                                referenceNormalizedReadCounts,
#                                sigList,
#                                replicates = replicates)
#
# reportTables <- ReportTables(bootList,
#                               geneNames,
#                               backgroundNoise,
#                               referenceNormalizedReadCounts,
#                               samplesNormalizedReadCounts)
#
#
# reportOutputFilepath <- paste0(outputDirectory, "/", "report_tables.xlsx")
#
# print("<><><><><><><><><><><><><><><><><><><><><><><><><><><><><><><>")
#
# print(reportOutputFilepath)
#
# library(WriteXLS)
#
# #
# # writeReadCounts <- function(sampleReadCounts, referenceReadCounts, file = "") {
# #
# #   writeSampleReadCounts <- as.data.frame(sampleReadCounts)
# #   #rownames(writeSampleReadCounts) <- paste0(seq_along(ampliconNames), "_",ampliconNames)
# #
# #   writeReferenceReadCounts <- as.data.frame(referenceReadCounts)
# #
# #   writeList <- list(SampleReadCounts = writeSampleReadCounts, referenceReadCounts = writeReferenceR
# #
# #   WriteXLS("writeList", paste0(outputFolder, file), AdjWidth = TRUE, BoldHeaderRow = TRUE, row.name
# # }
# #
#
# WriteXLS("reportTables",
#           reportOutputFilepath,
#           AdjWidth = TRUE,
#           BoldHeaderRow = TRUE,
#           row.names = TRUE,
#           col.names = TRUE)
#
# PlotBootstrapDistributions(bootList, reportTables, outputDirectory)
```

You can also embed plots, for example:



Conclusions

We showed that targeted sequencing can be used to reliably detect CNVs from clinical samples. The methods presented in this article are available as an R Bioconductor package.

Reading data from files

bla bla reading data from files.. some sample code..

Saving data to files

bla bla Saving data to files..

Below is a Line spanning the entire width of the page

Below is a 2cm long line

Below is a 4cm long line

Below is a 8cm long line

Below is a 12cm long line

Texto centrado \end{center}

I think LaTeXis fun

Note the difference in right and left quotes in ‘single quotes’ and “double quotes”

X-rays are discussed in pages 221–225 of Volume 3—the volume on electromagnetic waves.

Something

Text. . .

Other Paragraph

blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa blbasldkfjsl sjlkfs dlkfa

e outro ainda

eheh ahahha hashha hha hshs hsaha h ashs ahs hahs a eheh ahahha hashha hha hshs hsaha h ashs ahs hahs aeheh ahahha hashha hha hshs hsaha h ashs ahs hahs aeheh ahahha hashha hha hshs hsaha h ashs ahs hahs aeheh ahahha hashha hha hshs hsaha h ashs ahs hahs aeheh ahahha hashha hha hshs hsaha h ashs ahs hahs aeheh ahahha hashha hha hshs hsaha h ashs ahs hahs aeheh ahahha hashha hha hshs hsaha h ashs ahs hahs a

e isto?! aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa aaaaaa aaaa aaaaaaa

This is the first line.

This is the second linesdk ksdhfkjsdahf kjsh sdkfjhsd?k jfhs?kj fhj skdhf ?kjshdf?kjsdh f?kkjsdh f?kj sk?jdfh?ksdjhfk?sdjhfksjhksjdhf ?kjshfdkjsdhfkjshksdjfh ?ksjdhfk?jsadhfkjsahdfkjhsdskjldfh lksdajhflksah The TeXnical Institute

Certificate

Preface

Academic literature uses the abstract to succinctly communicate complex research. An abstract may act as a stand-alone entity instead of a full paper. As such, an abstract is used by many organizations as the basis for selecting research that is proposed for presentation in the form of a poster, platform/oral presentation or workshop presentation at an academic conference. Most literature database search engines index only abstracts rather than providing the entire text of the paper. Full texts of scientific papers must often be purchased because of copyright and/or publisher fees and therefore the abstract is a significant selling point for the reprint or electronic form of the full text.

Some theory

blablalba

blobloblo

This is the first line.

This is the second line The TEXnical Institute

Certificate

This is the first line.

This is the second line The TEXnical Institute

Certificate

This is to certify that Mr. N. O. Vice has undergone a course at this institute and is qualified to be a TEXnician.

The Director
The TEXnical Institute

flushing left

The TEXnical Institute

Certificate

This is to certify that Mr. N. O. Vice has undergone a course at this institute and is qualified to be a TEXnical Expert.

The Director
The TEXnical Institute

Abstract

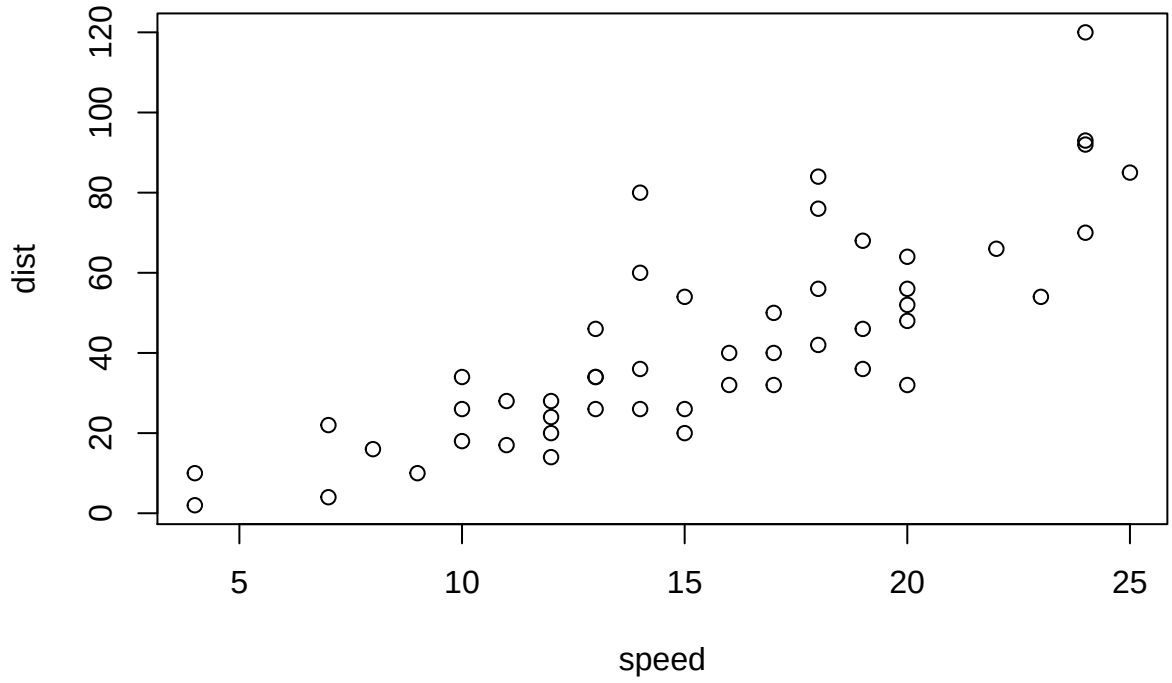
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
##black is also a color
#nice color
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

You can also embed plots, for example:



something..

```
## [1] 4
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.