ampliQueso: multiple sequencing of amplicons analyzed RNAseq style

Alicja Szabelska, Marek Wiewiorka, Michal Okoniewski

May 3, 2016

Contents

1 Recent changes and updates

2 Introduction

ampliQueso is UNDER CONSTRUCTION :P

but we do our best to make it functional and useful :)

ampliQueso is intended to be a library that can analyze the data from small (and bigger) multiple- amplicon panels of RNA and DNA, in technologies such as AmpliSeq. In comparison to eg TaqMan and similar RT-PCR technologies - AmpliSeq is a technique that uses the sequencing library prepared with multiple primer pairs (eg up to 16000 in Comprehensive Cancer Panel) that get amplified in a normal PCR machine. It can be compared to other sequencing enrichment techniques (), the difference is that is purely PCR-based. That's why it is expected to perform well eg with the slightly degraded (eg. paraffin embedded) samples. AmpliSeq protocols are in RNA and DNA versions, the RNA one measures expression of the amplified region, and when coverage permits - allows to find SNPs and small indels too.

The number of amplicons in such kits may be too small to run DESeq or edgeR on count data in a way described eg in [**?**], and the amplicons may be designed eg in the critical regions of splicing, thus the basic analysis is based upon the coverage analysis described in [**?**] and implemented in the package rnaSeqMap [**?**].

ampliQueso adds the functionality of non-parametric tests based upon the coverage analysis (camel) measures from [**?**], and uses the external variant call software to add SNP and indel information.

The analyses are bundled in a wrapper runAQReport that produces pdfs in Beamer or standard LaTeX article format.

The schema of processing in ampliQueso is as follows:

3 Using ampliQueso non-parametric tests on single genomic regions

Single genomic regions or batches of genomic regions may be compared with the camel/coverage measures using functions:

```
> library(ampliQueso)
> data(ampliQueso)
> par(mfrow=c(1,2))
> simplePlot(ndMin,exps=1:2,xlab="genome coordinates \n RGS1:NM_002922",
+            ylab="coverage")
> simplePlot(ndMax,exps=1:2,xlab="genome coordinates \n PLEK:NM_002664",
+            ylab="coverage")
```

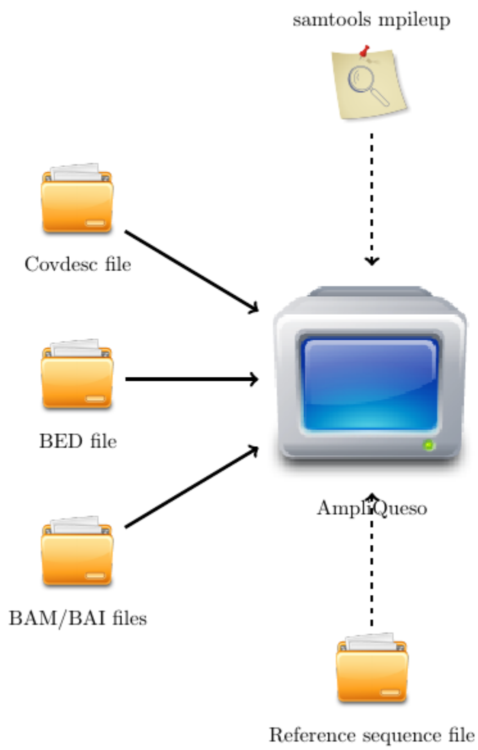
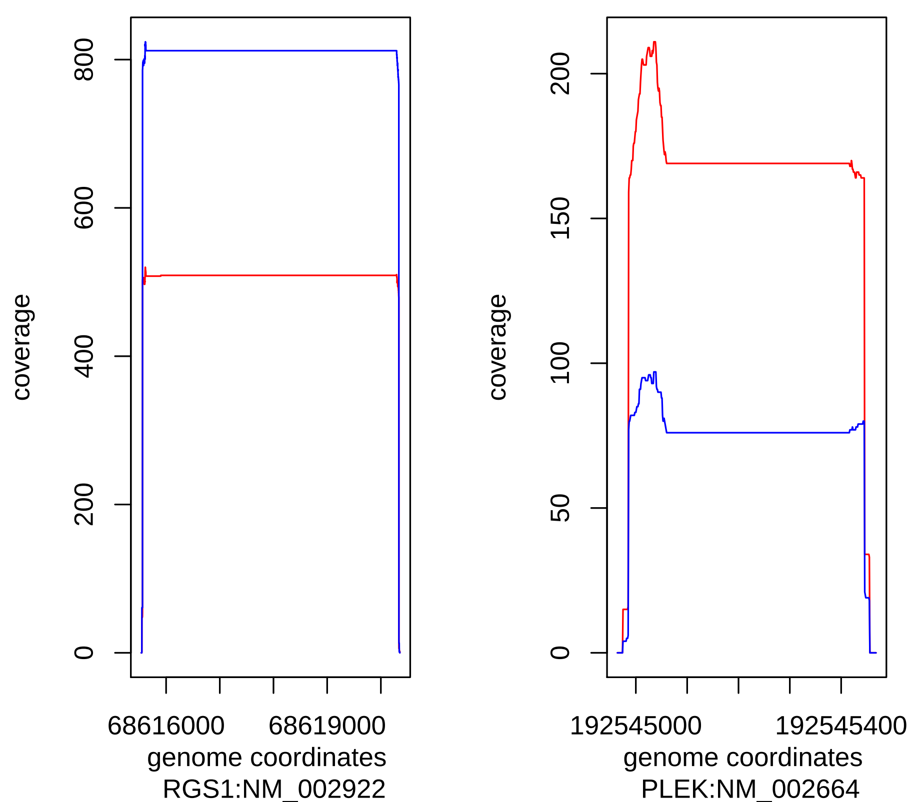


Figure 1: AmpliQueso high-level design, dashed arrows depict optional SNPs calling dataflow



Then, the measures from rnaSeqMap library can be used to generate p-values from non-paramertic tests that express how the coverage shapes are different, eg:

```
> library(xtable)
> curWd<-getwd()
> setwd(path.package("ampliQueso")) ##only for sample report
> iCovdesc=system.file("extdata","covdesc",package="ampliQueso")
> iBedFile=system.file("extdata","AQ.bed",package="ampliQueso")
> iT1="s"
> iT2="h"
> camelTestTable <- camelTest(iBedFile=iBedFile,iCovdesc=iCovdesc,
+                            iT1=iT1, iT2=iT2,iParallel=FALSE,iNPerm=5)
> #print sample p-values,not all of them
> camelTestTableDen<-camelTestTable[1:5,6:10]
> print(xtable(camelTestTableDen,caption="p-values from camel tests"))
```

	DA_density	QQ_density	PP_density	HD1_density	HD2_density
MMEL1:NM_033467	0.15	0.15	0.15	0.15	0.15
DDAH1:NM_012137	0.15	0.15	0.15	0.15	0.15
EVI5:NM_005665	1.00	1.00	0.82	1.00	1.00
SLC30A7:NM_133496	0.15	0.15	0.15	0.15	0.15
EXTL2:NR_048570	0.65	0.49	0.82	0.65	0.49

Table 1: p-values from camel tests

```
> setwd(curWd)
```

The p-values may be used to find the most significantly differential shapes, thus - most significant differential expression in amplicons. The values of camel measures can be used too, but normally they are less expressive. Still - can be compared between the regions:

```
> library(xtable)
> data(ampliQueso)
> print(xtable(camelSampleTable,
+              caption="Camel/coverage measures for two sample regions"))
```

	region	DA	QQ	PP	HD1	HD2
1	PLEK:NM_002664	0.138	65.4789	181890.6704	84.6894	73.5461
2	RGS1:NM_002922	0.0048	1.1178	185.1135	3.7024	3.5341

Table 2: Camel/coverage measures for two sample regions

```
>
```

The object loaded from the example data contains the coverages as NucleotideDistr objects, defined in rnaSeqMap library.

4 Classic read counting in amplicon regions

The simple analysis may include generating counts from BAM files, according to the amplicon description in the BED design file of the kit:

```
> library(ampliQueso)
> setwd(path.package("ampliQueso"))
> cc <- getCountTable(covdesc=system.file("extdata","covdesc",package="ampliQueso"),
+                      bedFile=system.file("extdata","AQ.bed",package="ampliQueso"))
```

```
....................
```

```
> cc[1:4,1:2]
```

```
                   ./extdata/sample_033_sort.bam ./extdata/sample_034_sort.bam
MMEL1:NM_033467                                0                             0
DDAH1:NM_012137                                0                             0
EVI5:NM_005665                               158                            99
SLC30A7:NM_133496                            120                           147
```

5 Using an external variant caller - samtools mpileup

In DNA amplicon kits and when the coverage in RNA ones is sufficient, the genomic variants can be found. The funcrtion `getSNP` encapsulates a system call to `samtools mpileup` with a reference genome:

```
> #in order to run this example you need provide reference sequence
> #in FASTA format and set refSeqFile parameter
> curWd<-getwd()
> setwd(path.package("ampliQueso")) ##only for sample report
> iCovdesc=system.file("extdata","covdesc",package="ampliQueso")
> iBedFile=system.file("extdata","AQ.bed",package="ampliQueso")
> snpList <- getSNP(covdesc=iCovdesc, minQual=10,
+                    refSeqFile="hg19.fa", bedFile = iBedFile)
> setwd(curWd)
```

6 The complete report on AmpliSeq experiment

The complete report of a two-group comparison on a given set of BAM files and given BED design desctiption includes all the parts of analysis described in the sections above and can be called as follows:

```
> #########Example##########################
> library(ampliQueso)
> curWd<-getwd()
> setwd(path.package("ampliQueso")) ##only for sample report
> iCovdesc=system.file("extdata","covdesc",package="ampliQueso")
```

```
> iBedFile=system.file("extdata","AQ.bed",package="ampliQueso")
> iRefSeqFile=NULL
> iGroup="group"
> iT1="s"
> iT2="h"
> iTopN=5
> iMinQual=NULL
> iReportFormat="pdf"
> iReportType="article"
> iReportPath=curWd
> iVerbose=FALSE
> iParallel=FALSE
> runAQReport(iCovdesc=iCovdesc,iBedFile=iBedFile,iRefSeqFile=iRefSeqFile,
+ iGroup=iGroup,iT1=iT1,iT2=iT2,iTopN=iTopN,iMinQual=iMinQual,
+ iReportFormat=iReportFormat,iReportType=iReportType,
+             iReportPath,iVerbose=iVerbose,iParallel=iParallel)
> setwd(curWd)
```

it produces the pdf output.

The report can be used also for the DNA kits, but then the fold change should be interpreted as a possible copy number difference. We are planning to differentiate the reports for RNA and DNA.

7 File formats

7.1 BED format

AmpliQueso supports BED files in the following format (see also **??**):

1. chromName
2. chromStart
3. chromEnd
4. strand
5. *unspecified*
6. name

```
chr1    2488068    2488201    .    TNFRSF14    AMPL242431688
chr1    2489144    2489273    .    TNFRSF14    AMPL262048751
chr1    2489772    2489907    .    TNFRSF14    AMPL241330530
chr1    2491241    2491331    .    TNFRSF14    AMPL242158034
chr1    2491314    2491444    .    TNFRSF14    AMPL242161604
```

Figure 2: Example BED file in the format supported by AmpliQueso

If a BED intened for use in AmpliQueso has a wrong column order one can easily rearrange them using for example `awk` tool:

```
awk '{print($1,"\t",$2,"\t",$3,"\t",$5,"\t",$6,"\t",$4)}' input.bed > output.bed
```

In the example above columns 4,5,6 positions are swapped.

8 Troubleshooting

8.1 Mac OS X

8.1.1 rgl package

Please make sure that `DISPLAY` environemt variable is not set prior to runnin R in terminal. Otherwise loading *rgl* package may hang without any obvious reason.

8.1.2 foreach package

It is not safe to use *foreach* package from R.app on Mac OS X. This is why, it is recommended to use *ampliQueso* from a terminal session, starting R from the command line.

8.2 Windows

8.2.1 foreach package

Depending on Windows firewall settings, it might be necessary to confirm firewall exceptions allowing launching R slave servers which necessary for using *foreach* package.

9 References

References

[1] Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD, Count-based differential expression analysis of RNA sequencing data using R and Bioconductor, pre-print: http://arxiv.org/abs/1302.3685, Nature Protocols, 2013 - accepted for publication

[2] Okoniewski, M. J., Lesniewska, A., Szabelska, A., Zyprych-Walczak, J., Ryan, M., Wachtel, M., et al. (2011). Preferred analysis methods for single genomic regions in RNA sequencing revealed by processing the shape of coverage. Nucleic acids research. doi:10.1093/nar/gkr1249

[3] Lesniewska, A., & Okoniewski, M. J. (2011). rnaSeqMap: a Bioconductor package for RNA sequencing data exploration. BMC bioinformatics, 12, 200. doi:10.1186/1471-2105-12-200