Introduction to QDNAseq

Ilari Scheinin

October 8, 2016

Contents

1 Running QDNAseq

This is a short tutorial on how to use the *QDNAseq* package. It covers an example run using the included data set of chromosomes 7–10 of a low grade glioma (LGG) sample. First step is naturally to load the package.

```
> library(QDNAseq)
```

1.1 Bin annotations

Then we need to obtain bin annotations. These are available pre-calculated for genome build hg19 and bin sizes 1, 5, 10, 15, 30, 50, 100, 500, and 1000 kbp. They can be downloaded for example with:

```
> bins <- getBinAnnotations(binSize=15)
Downloading bin annotations for genome hg19, bin size 15kbp,
and experiment type SR50 ...
> bins
QDNAseq bin annotations for Hsapiens, build hg19.
Created by Ilari Scheinin with QDNAseq 0.7.5, 2014-02-06 12:48:04.
An object of class 'AnnotatedDataFrame'
  rowNames: 1:1-15000 1:15001-30000 ... Y:59370001-59373566 (206391
    total)
  varLabels: chromosome start ... use (9 total)
  varMetadata: labelDescription
```

After downloading, the bin annotations can be saved locally with `saveRDS()`, and in the future be read from the local file with `loadRDS()` instead of relying on downloading.

If you are working with another genome build (or another species), see the section on generating the bin annotations.

1.2 Processing bam files

Next step is to load the sequencing data from bam files. This can be done for example with one of the commands below.

```
> readCounts <- binReadCounts(bins)
> # all files ending in .bam from the current working directory
>
> # or
>
> readCounts <- binReadCounts(bins, bamfiles='tumor.bam')
```

```
> # file 'tumor.bam' from the current working directory
>
> # or
>
> readCounts <- binReadCounts(bins, path='tumors')
> # all files ending in .bam from the subdirectory 'tumors'
```

This will return an object of class *QDNAseqReadCounts*. If the same bam files will be used as input in future *R* sessions, option cache=TRUE can be used to cache intermediate files, which will speed up future analyses. Caching is done with package *R.cache*.

For the purpose of this tutorial, we load an example data set of chromosomes 7–10 of low grade glioma sample LGG150.

```
> data(LGG150)
> readCounts <- LGG150
> readCounts
```

```
QDNAseqReadCounts (storageMode: lockedEnvironment)
assayData: 38819 features, 1 samples
  element names: counts
protocolData: none
phenoData
  sampleNames: LGG150
  varLabels: name reads used.reads
    expected.variance
  varMetadata: labelDescription
featureData
  featureNames: 7:1-15000
    7:15001-30000 ...
    10:135525001-135534747 (38819
    total)
  fvarLabels: chromosome start ...
    use (9 total)
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:
```

Plot a raw copy number profile (read counts across the genome), and highlight bins that will be removed with default filtering (Figure **??**).

```
> plot(readCounts, logTransform=FALSE, ylim=c(-50, 200))
```

```
Plotting sample LGG150 (1 of 1) ...
```

```
> highlightFilters(readCounts, logTransform=FALSE,
+   residual=TRUE, blacklist=TRUE)
```

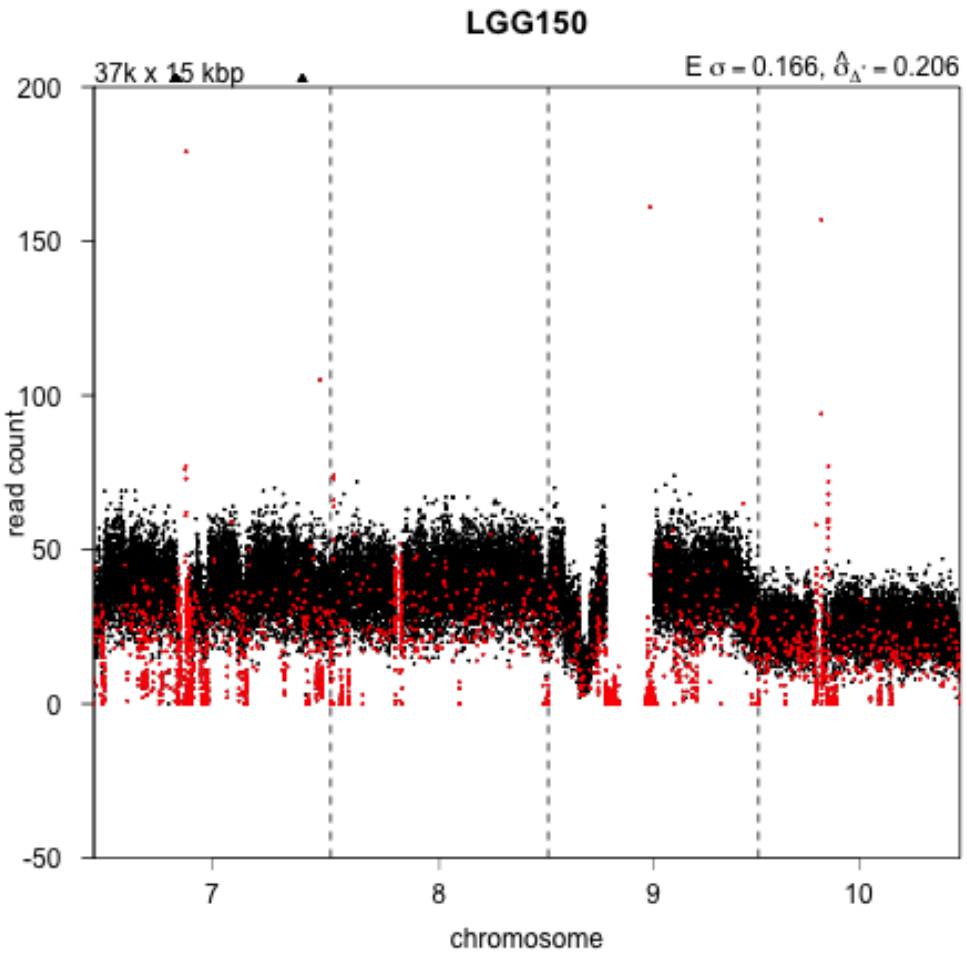
```
Highlighted 3,375 bins.
```

Apply filters and plot median read counts as a function of GC content and mappability (Figure **??**). As the example data set only contains a subset of the chromosomes, the distribution looks slithly less smooth than expected for the entire genome.

```
> readCountsFiltered <- applyFilters(readCounts,
+   residual=TRUE, blacklist=TRUE)
```

```
38,819        total bins
38,819        of which in selected chromosomes
36,722        of which with reference sequence
33,347        final bins
```

Figure 1: Read counts per bins. Highlighted with red are bins that will be filtered out.




```
> isobarPlot(readCountsFiltered)
```

```
Plotting sample LGG150 median read counts
```

Estimate the correction for GC content and mappability, and make a plot for the relationship between the observed standard deviation in the data and its read depth (Figure **??**). The theoretical expectation is a linear relationship, which is shown in the plot with a black line. Samples with low-quality DNA will be noisier than expected and appear further above the line than good-quality samples.

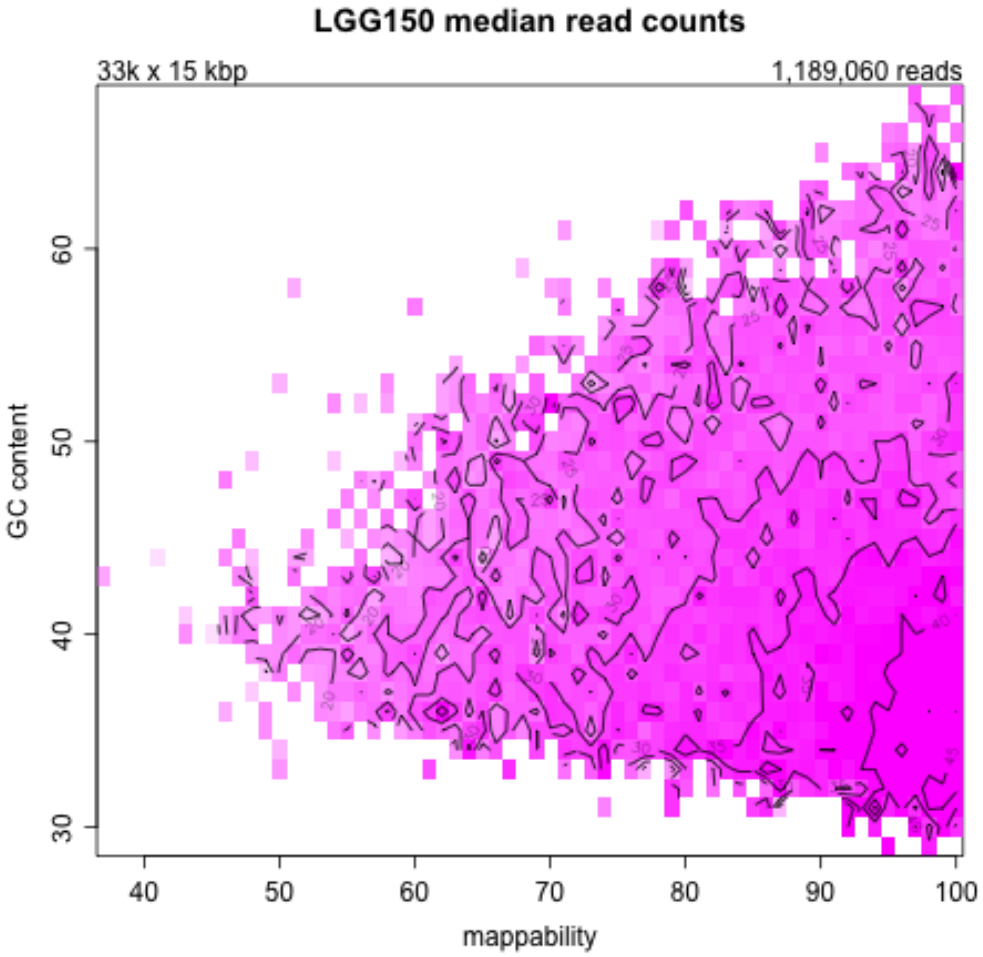
```
> readCountsFiltered <- estimateCorrection(readCountsFiltered)
```

```
Calculating correction for GC content and mappability
    Calculating fit for sample LGG150 (1 of 1) ...
Done.
```

```
> noisePlot(readCountsFiltered)
```

Next, we apply the correction for GC content and mappability. This will return a *QDNAseqCopyNumbers* object, which we then normalize, smooth outliers, and plot the copy number profile (Figure **??**).

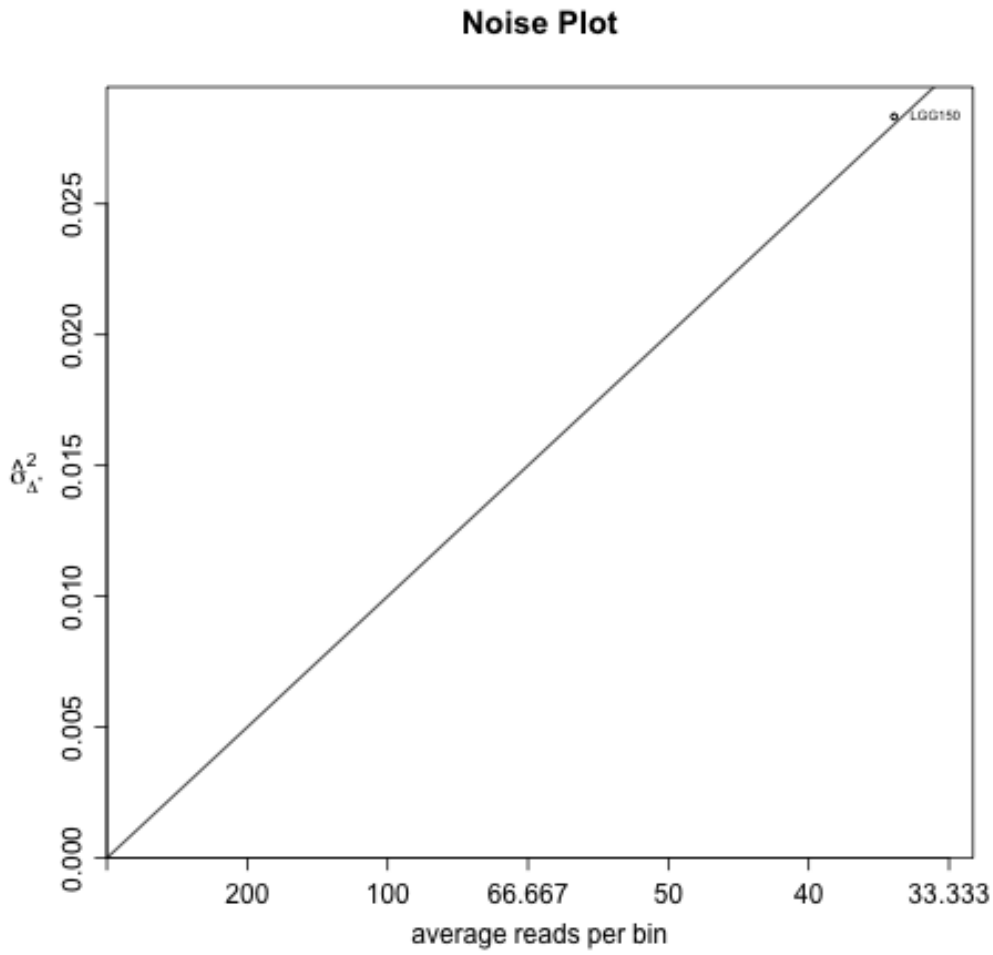
Figure 2: Median read counts per bin shown as a function of GC content and mappability.




```
> copyNumbers <- correctBins(readCountsFiltered)
> copyNumbers
```

```
QDNAseqCopyNumbers (storageMode: lockedEnvironment)
assayData: 38819 features, 1 samples
  element names: copynumber
protocolData: none
phenoData
  sampleNames: LGG150
  varLabels: name reads ...
    loess.family (6 total)
  varMetadata: labelDescription
featureData
  featureNames: 7:1-15000
    7:15001-30000 ...
    10:135525001-135534747 (38819
    total)
  fvarLabels: chromosome start ...
```

Figure 3: The relationship between sequence depth and noise.




```
    use (9 total)
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:
```

```
> copyNumbersNormalized <- normalizeBins(copyNumbers)
```

```
Applying median normalization ...
```

```
> copyNumbersSmooth <- smoothOutlierBins(copyNumbersNormalized)
```

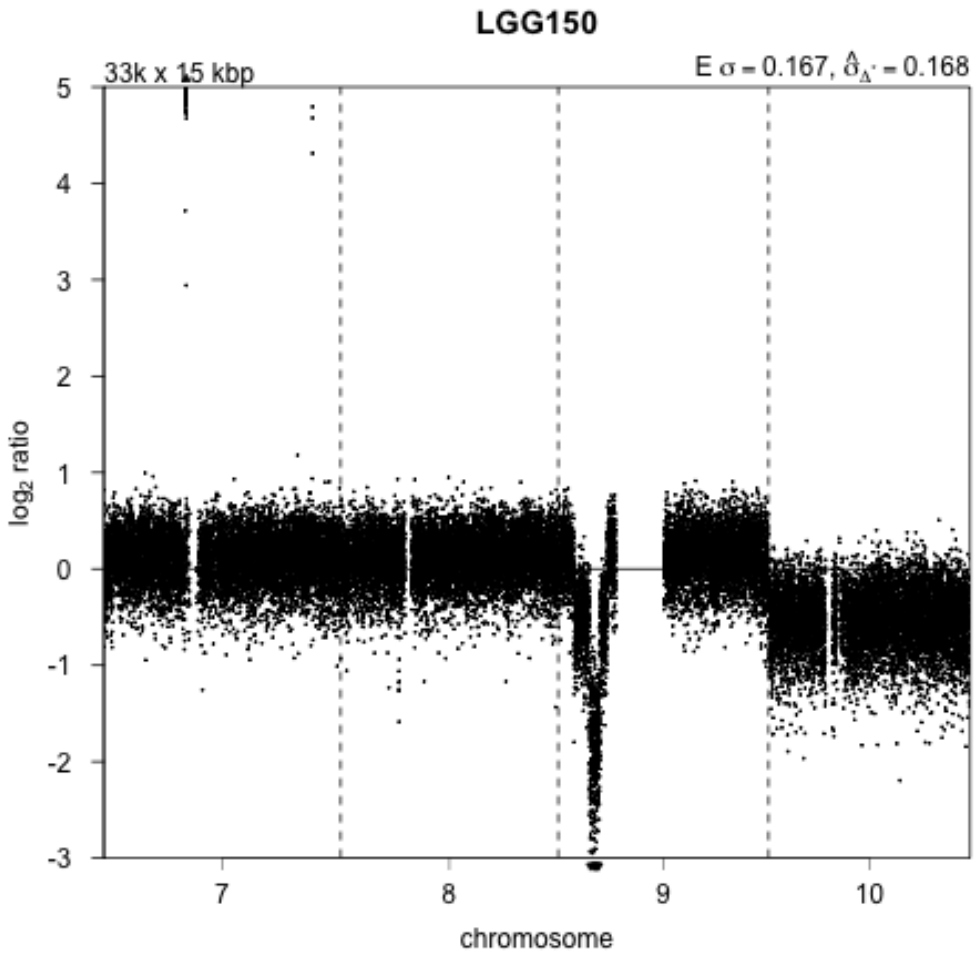
```
Smoothing outliers ...
```

```
> plot(copyNumbersSmooth)
```

```
Plotting sample LGG150 (1 of 1) ...
```

Data is now ready to be analyzed with a downstream package of choice. For analysis with an external program or for visualizations in *IGV*, the data can be exported to a file.

Figure 4: Copy number profile after correcting for GC content and mappability.




```
> exportBins(copyNumbersSmooth, file="LGG150.txt")
> exportBins(copyNumbersSmooth, file="LGG150.igv", format="igv")
> exportBins(copyNumbersSmooth, file="LGG150.bed", format="bed")
```

1.3 Downstream analyses

Segmentation with the *CBS* algorithm from *DNAcopy*, and calling copy number aberrations with *CGHcall* or cutoffs have been implemented for convenience.

By default, segmentation uses a $\log_2$-transformation, but a sqrt(x + 3/8) can also be used as it stabilizes the variance of a Poisson distribution (Anscombe transform):

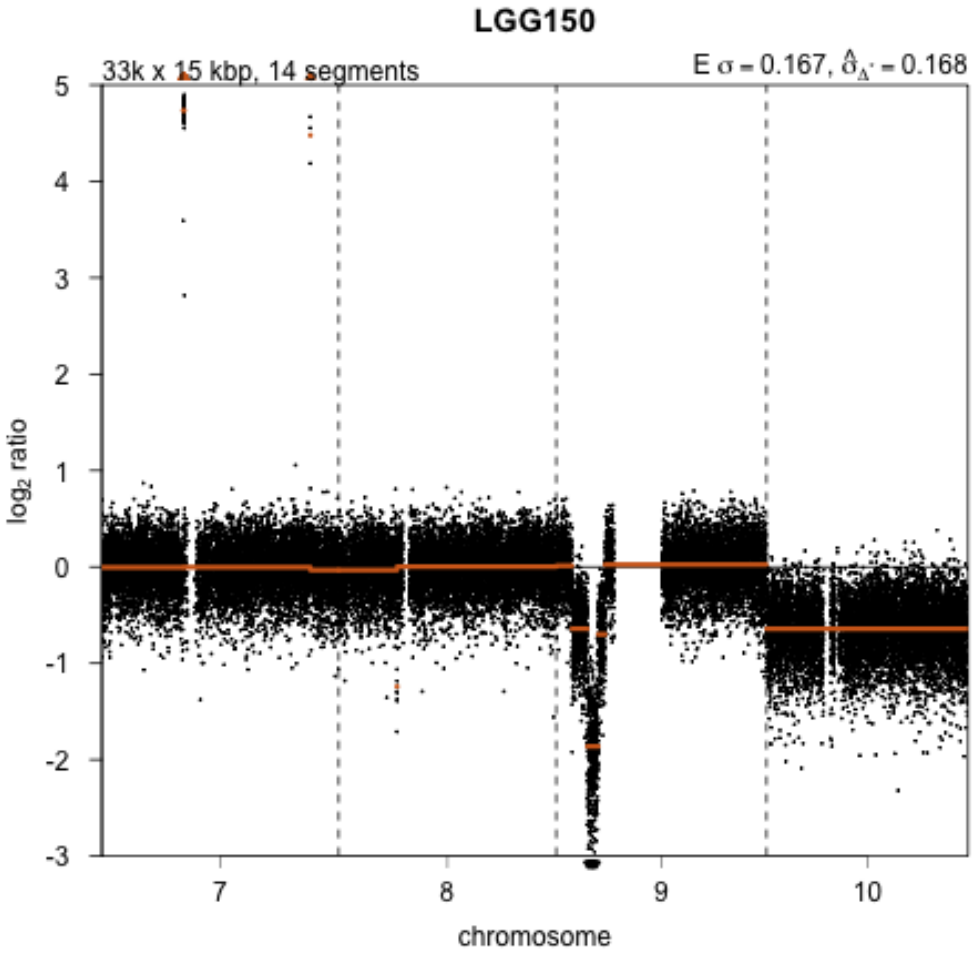
```
> copyNumbersSegmented <- segmentBins(copyNumbersSmooth, transformFun="sqrt")
```

```
Performing segmentation:
    Segmenting: LGG150 (1 of 1) ...
```

```
> copyNumbersSegmented <- normalizeSegmentedBins(copyNumbersSegmented)
```

```
> plot(copyNumbersSegmented)
Plotting sample LGG150 (1 of 1) ...
```

Figure 5: Copy number profile after segmenting.



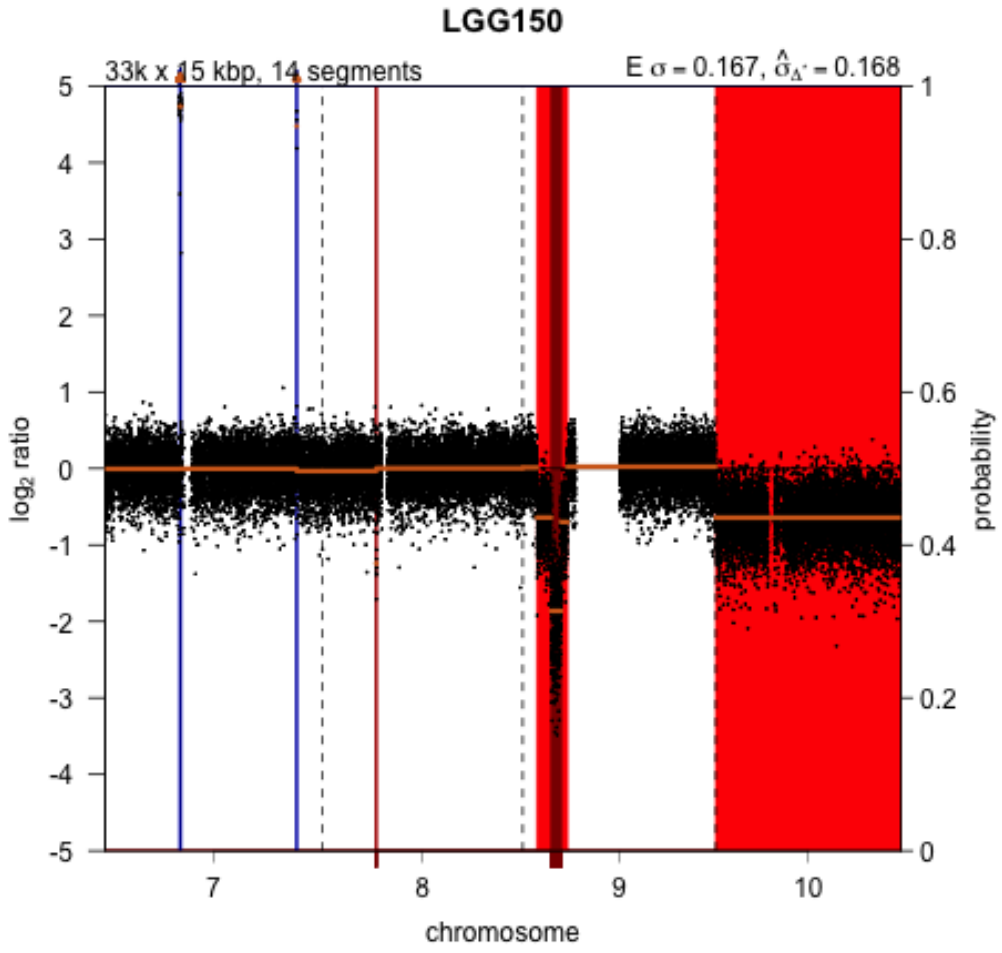
Tune segmentation parameters and iterate until satisfied. Next, call aberrations, and plot the final results.

```
> copyNumbersCalled <- callBins(copyNumbersSegmented)
[1] "Total number of segments present in the data: 14"
[1] "Number of segments used for fitting the model: 11"
> plot(copyNumbersCalled)
Plotting sample LGG150 (1 of 1) ...
```

It should be noted that *CGHcall* (which `callBins()` uses by default) was developed for the analysis of sets of cancer samples. It is based on a mixture model, and when there are not enough aberrations present in the data, model fitting can fail. This can happen especially with non-cancer samples, and/or when analyzing individual cases instead of larger data sets.

If *CGHcall* fails, `callBins()` can also perform simple cutoff-based calling by setting parameter `method="cutoff"`. The default cutoff values are based on the assumption of uniform cell populations, and in case of cancer samples will most

Figure 6: Copy number profile after calling gains and losses.



likely need calibration by adjusting parameter `cutoffs`.

Finally, for other downstream analyses, such as running *CGHregions*, it might be useful to convert to a *cghCall* object.

```
> cgh <- makeCgh(copyNumbersCalled)
> cgh
```

```
cghCall (storageMode: lockedEnvironment)
assayData: 33347 features, 1 samples
  element names: calls, copynumber, probamp, probdloss, probgain, probloss, probnorm, segmented
protocolData: none
phenoData
  sampleNames: LGG150
  varLabels: name reads ...
    loess.family (6 total)
  varMetadata: labelDescription
featureData
  featureNames: 7:45001-60000
    7:60001-75000 ...
```

```
    10:135420001-135435000 (33347
    total)
  fvarLabels: Chromosome Start ...
    use (9 total)
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:
```

This command can also be used to generate *cghRaw* or *cghSeg* objects by running it before segmentation or calling.

2 Parallel computation

QDNAseq supports parallel computing via the *future* package. After installing it with `install.packages("future")`, all that is required is to select an appropriate *plan*. The recommended way to do that is to include it in your '~/.Rprofile'.

The instructions below apply to all of *QDNAseq*'s own functions that support parallel processing. At the moment these include `estimateCorrection()`, `segmentBins()`, `createBins()`, and `calculateBlacklist()`.

However, when argument `method="CGHcall"` (which is the default), function `callBins()` calls function `CGHcall()` from package *CGHcall*, which uses another mechanism for parallel computation. For that, the number of processes to use should be specified with argument `ncpus`, with something along the lines of:

```
> copyNumbers <- callBins(..., ncpus=4L)
```

2.1 Non-parallel processing

The default is to use single-core processing via "eager" futures. This can be set explicitly with:

```
> future::plan("eager")
```

2.2 Parallel processing on the current machine

To process data in parallel using multiple processes on the current machine, use the following:

```
> future::plan("multiprocess")
```

After that, all functions that support parallel processing will automatically use it.

On Mac OS X and Linux, this will use multiple *forked R* processes (`"multicore"`). On Windows, which does not support *forking*, multiple *background R* processes (`"multisession"`) will be used instead.

The number of parallel processes to use can be defined via option `mc.cores`, for example with:

```
> options(mc.cores=4L)
```

2.3 Parallel processing on a cluster

To process data using multiple *R* sessions running on different machines, use something along the lines of:

```
> cl <- parallel::makeCluster(...)
> future::plan("cluster", cluster=cl)
```

See package *parallel* for more details.

3 Sex chromosomes

By default, *QDNAseq* ignores sex chromosomes. In order to include them in the analysis, function `applyFilters()` should be run with argument `chromosomes=NA` to include both X and Y, or `chromosomes="Y"` to include X only.

However, this will also affect which chromosomes are used when calculating the loess correction with `estimateCorrection()`. Unless the data set consists of only females, this could be undesirable. The solution is to first filter out the sex chromosomes, run `estimateCorrection()`, and then reverse the filtering of sex chromosomes:

```
> readCounts <- binReadCounts(getBinAnnotations(15))
> readCounts <- applyFilters(readCounts)
> readCounts <- estimateCorrection(readCounts)
> readCounts <- applyFilters(readCounts, chromosomes=NA)
> copyNumbers <- correctBins(readCounts)
```

Running `estimateCorrection()` and `correctBins()` with a different set of bins can have one side effect. This is caused by the fact that there can be bins in the sex chromosomes with a combination of GC content and mappability that is not found anywhere else in the genome. This will cause those bins to miss a correction estimate altogether, and these bins will be filtered out from subsequent steps by `correctBins()`. If this happens, it will print out a message speficying the number of bins affected.

Another possible approach is to allow extrapolation while calculating the loess correction. But please do note that the effect of extrapolation has not been properly evaluated.

```
> readCounts <- estimateCorrection(readCounts,
+   control=loess.control(surface="direct"))
```

4 Generating bin annotations

This section describes how bin annotations have been created for the hg19 build of the human reference genome, and can be applied for other genome builds and species. The first step is to create the bins based on chromosome sizes, and calculate their GC content and proportion of characterized nucletodies (non-N bases in the reference sequence). For this, the corresponding *BSgenome* package is needed.

```
> # load required packages for human reference genome build hg19
> library(QDNAseq)
> library(Biobase)
> library(BSgenome.Hsapiens.UCSC.hg19)
> # set the bin size
> binSize <- 15
> # create bins from the reference genome
> bins <- createBins(bsgenome=BSgenome.Hsapiens.UCSC.hg19, binSize=binSize)
```

The result is a *data.frame* with columns `chromosome`, `start`, `end`, `gc`, and `bases`. Next step is to calculate the average mappabilities, which requires a mappability file in the `bigWig` format and the *bigWigAverageOverBed* binary. The mappability file can be generated with *GEM library* from the reference genome sequence. Or it might be available directly, as was the case for hg19, and file 'wgEncodeCrgMapabilityAlign50mer.bigWig' downloaded from ENCODE's download section of the UCSC Genome Browser. The *bigWigAverageOverBed* binary can also be downloaded from UCSC Genome Browser's Other utilities section.

```
> # calculate mappabilites per bin from ENCODE mapability tracks
> bins$mappability <- calculateMappability(bins,
+   bigWigFile='/path/to/wgEncodeCrgMapabilityAlign50mer.bigWig',
+   bigWigAverageOverBed='/path/to/bigWigAverageOverBed')
```

If there are genomic regions that should excluded from analyses, such as ENCODE's Blacklisted Regions, the percentage overlap between the generated bins and these regions can be calculated as follows. The regions to be excluded need to be in the `BED` format, like files 'wgEncodeDacMapabilityConsensusExcludable.bed' and 'wgEncodeDukeMapabilityRegionsExcludable.bed' that were downloaded from ENCODE's download section of the UCSC Genome Browser for hg19.

```
> # calculate overlap with ENCODE blacklisted regions
> bins$blacklist <- calculateBlacklist(bins,
+   bedFiles=c('/path/to/wgEncodeDacMapabilityConsensusExcludable.bed',
+   '/path/to/wgEncodeDukeMapabilityRegionsExcludable.bed'))
```

To calculate median residuals of the LOESS fit from a control dataset, the following command can be used. For the pre-generated annotations, the control set used is 38 samples from the 1000 Genomes project. See the next section on how those were downloaded.

```
> # load data for the 1000 Genomes (or similar) data set, and generate residuals
> ctrl <- binReadCounts(bins,
+   path='/path/to/control-set/bam/files')
> ctrl <- applyFilters(ctrl, residual=FALSE, blacklist=FALSE,
+   mappability=FALSE, bases=FALSE)
> bins$residual <- iterateResiduals(ctrl)
```

The column `use` specifies whether each bin should be used for subsequent analyses by default. The command `applyFilters()` will change its value accordingly. By default, bins in the sex chromosomes, or with only uncharacterized nucleotides (N's) in their reference sequence, are flagged for exlusion.

```
> # by default, use all autosomal bins that have a reference sequence
> # (i.e. not only N's)
> bins$use <- bins$chromosome %in% as.character(1:22) & bins$bases > 0
```

Optionally, the resulting *data.frame* can be converted to an *AnnotateDataFrame* and metadata added for the columns.

```
> # convert to AnnotatedDataFrame and add metadata
> bins <- AnnotatedDataFrame(bins,
+   varMetadata=data.frame(labelDescription=c(
+   'Chromosome name',
+   'Base pair start position',
+   'Base pair end position',
+   'Percentage of non-N nucleotides (of full bin size)',
+   'Percentage of C and G nucleotides (of non-N nucleotides)',
+   'Average mappability of 50mers with a maximum of 2 mismatches',
+   'Percent overlap with ENCODE blacklisted regions',
+   'Median loess residual from 1000 Genomes (50mers)',
+   'Whether the bin should be used in subsequent analysis steps'),
+   row.names=colnames(bins)))
```

For the pre-generated annotations, some additional descriptive metadata has also been added.

```
> attr(bins, "QDNAseq") <- list(
+   author="Ilari Scheinin",
+   date=Sys.time(),
+   organism="Hsapiens",
+   build="hg19",
+   version=packageVersion("QDNAseq"),
+   md5=digest::digest(bins@data),
+   sessionInfo=sessionInfo())
```

5 Downloading 1000 Genomes samples

This section defines the criteria that were used to download samples from the 1000 Genomes project for the pre-generated bin annotations.

```
> # download table of samples
> server <- "ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/"
> g1k <- read.table(paste0(server, "sequence.index"),
+   header=TRUE, sep="\t", as.is=TRUE, fill=TRUE)
> # keep cases that are Illumina, low coverage, single-read, and not withdrawn
> g1k <- g1k[g1k$INSTRUMENT_PLATFORM == 'ILLUMINA', ]
> g1k <- g1k[g1k$ANALYSIS_GROUP == 'low coverage', ]
> g1k <- g1k[g1k$LIBRARY_LAYOUT == 'SINGLE', ]
> g1k <- g1k[g1k$WITHDRAWN == 0, ]
> # keep cases with read lengths of at least 50 bp
> g1k <- g1k[!g1k$BASE_COUNT %in% c("not available", ""), ]
> g1k$BASE_COUNT <- as.numeric(g1k$BASE_COUNT)
> g1k$READ_COUNT <- as.integer(g1k$READ_COUNT)
> g1k$readLength <- g1k$BASE_COUNT / g1k$READ_COUNT
> g1k <- g1k[g1k$readLength > 50, ]
> # keep samples with a minimum of one million reads
> readCountPerSample <- aggregate(g1k$READ_COUNT,
+   by=list(sample=g1k$SAMPLE_NAME), sum)
> g1k <- g1k[g1k$SAMPLE_NAME %in%
+   readCountPerSample$sample[readCountPerSample$x >= 1e6], ]
> g1k$fileName <- basename(g1k$FASTQ_FILE)
> # download fastq files
> for (i in rownames(g1k)) {
+   sourceFile <- paste0(server, g1k[i, "FASTQ_FILE"])
+   destFile <- g1k[i, "fileName"]
+   if (!file.exists(destFile))
+     download.file(sourceFile, destFile)
+ }
```

Next, reads were trimmed to 50 bp, and the multiple files for each sample (as defined by column SAMPLE_NAME) were combined by concatenating the FASTQ files together. Finally, they were aligned with *BWA* allowing two mismatches and end-trimming of bases with qualities below 40 (options `-n 2 -q 40`).

6 Session information

The version number of *R* and packages loaded for generating the vignette were:

- R version 3.3.1 (2016-06-21), `x86_64-apple-darwin13.4.0`
- Locale: `C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: QDNAseq 1.8.1
- Loaded via a namespace (and not attached): Biobase 2.32.0, BiocGenerics 0.18.0, BiocParallel 1.6.6, BiocStyle 2.0.3, Biostrings 2.40.2, CGHbase 1.32.0, CGHcall 2.34.1, DNAcopy 1.46.0, GenomeInfoDb 1.8.7, GenomicRanges 1.24.3, IRanges 2.6.1, R.methodsS3 1.7.1, R.oo 1.20.0, R.utils 2.4.0, Rsamtools 1.24.0, S4Vectors 0.10.3, XVector 0.12.1, bitops 1.0-6, impute 1.46.0, limma 3.28.21, marray 1.50.0, matrixStats 0.50.2, parallel 3.3.1, stats4 3.3.1, tools 3.3.1, zlibbioc 1.18.0