CODEX vignette

Yuchao Jiang
yuchaoj@wharton.upenn.edu

May 3, 2016

This is a demo for using the `CODEX` package in R. `CODEX` is a normalization and copy number variation calling procedure for whole exome DNA sequencing data. `CODEX` relies on the availability of multiple samples processed using the same sequencing pipeline for normalization, and does not require matched controls. The normalization model in `CODEX` includes terms that specifically remove biases due to GC content, exon capture and amplification efficiency, and latent systemic artifacts. `CODEX` also includes a Poisson likelihood-based recursive segmentation procedure that explicitly models the count-based exome sequencing data. Below is an example on calling copy number variation using whole-exome sequencing data of 46 HapMap samples sequenced at the Washington University Genome Sequencing Center. Only the 401-500 exons from chromosome 22 are analysed for illustration purposes. R packages are available at Bioconductor for CODEX and the toy dataset WES.1KG.WUGSC.

1. Installation and online forum

Installation information is available here. The online Q&A forum is available here. If you've any questions regarding the software, please don't hesitate emailing us at codex_wes_cnv@googlegroups.com and/or yuchaoj@wharton.upenn.edu.

2. CODEX workflow:

2.1 Get directories of .bam files, read in exon target positions from .bed files, and get sample names.

The direct input of `CODEX` include: `bamdir`, which is a vector indicating the directories of all .bam files; `sampname`, which is a column vector with row entries of sample names; `bedFile`, which indicates the directory of the .bed file (WES bait file, no header, sorted by start and end positions); and `chr`, which specifies the chromosome. `CODEX` processes the entire genome chromosome by chromosome; make sure the chromosome formats are consistent between the .bed and the .bam files.

```
> library(CODEX)
> library(WES.1KG.WUGSC) # Load Toy data from the 1000 Genomes Project.
> dirPath <- system.file("extdata", package = "WES.1KG.WUGSC")
> bamFile <- list.files(dirPath, pattern = '*.bam$')
> bamdir <- file.path(dirPath, bamFile)
> sampname <- as.matrix(read.table(file.path(dirPath, "sampname")))
> bedFile <- file.path(dirPath, "chr22_400_to_500.bed")
> chr <- 22
> bambedObj <- getbambed(bamdir = bamdir, bedFile = bedFile,
+                        sampname = sampname, projectname = "CODEX_demo", chr)
> bamdir <- bambedObj$bamdir; sampname <- bambedObj$sampname
> ref <- bambedObj$ref; projectname <- bambedObj$projectname; chr <- bambedObj$chr
```

2.2 Get raw read depth from the .bam files. Read lengths across all samples are also returned.

```
> coverageObj <- getcoverage(bambedObj, mapqthres = 20)
> Y <- coverageObj$Y; readlength <- coverageObj$readlength
```

2.3 Compute GC content and mappability for each exon target.

```
> gc <- getgc(chr, ref)
> mapp <- getmapp(chr, ref)
```

2.4 Take a sample-wise and exon-wise quality control procedure on the depth of coverage matrix.

```
> qcObj <- qc(Y, sampname, chr, ref, mapp, gc, cov_thresh = c(20, 4000),
+     length_thresh = c(20, 2000), mapp_thresh = 0.9, gc_thresh = c(20, 80))
> Y_qc <- qcObj$Y_qc; sampname_qc <- qcObj$sampname_qc; gc_qc <- qcObj$gc_qc
> mapp_qc <- qcObj$mapp_qc; ref_qc <- qcObj$ref_qc; qcmat <- qcObj$qcmat
> write.table(qcmat, file = paste(projectname, '_', chr, '_qcmat', '.txt', sep=''),
+             sep='\t', quote=FALSE, row.names=FALSE)
```

2.5 Fit Poisson latent factor model for normalization of the read depth data.

```
> normObj <- normalize(Y_qc, gc_qc, K = 1:9)
> Yhat <- normObj$Yhat; AIC <- normObj$AIC; BIC <- normObj$BIC
> RSS <- normObj$RSS; K <- normObj$K
```

If the WES is designed under case-control setting, CODEX estimates the exon-wise Poisson latent factor using only the read depths in the control cohort, and then computes the sample-wise latent factor terms for the case samples by regression. `normal_index` specifies the indices of normal samples and the normalization function to use under this setting is `normalize2`.

```
> normObj <- normalize2(Y_qc, gc_qc, K = 1:9, normal_index=seq(1,45,2))
> Yhat <- normObj$Yhat; AIC <- normObj$AIC; BIC <- normObj$BIC
> RSS <- normObj$RSS; K <- normObj$K
```

2.6 Determine the number of latent factors. AIC, BIC, and deviance reduction plots are generated in a .pdf file.

CODEX reports all three statistical metrics (AIC, BIC, percent of Variance explained) and uses BIC as the default method to determine the number of Poisson factors. Since false positives can be screened out through a closer examination of the post-segmentation data, whereas CNV signals removed in the normalization step cannot be recovered, CODEX opts for a more conservative normalization that, when in doubt, uses a smaller value of K.

```
> choiceofK(AIC, BIC, RSS, K, filename = paste(projectname, "_", chr,
+     "_choiceofK", ".pdf", sep = ""))
```

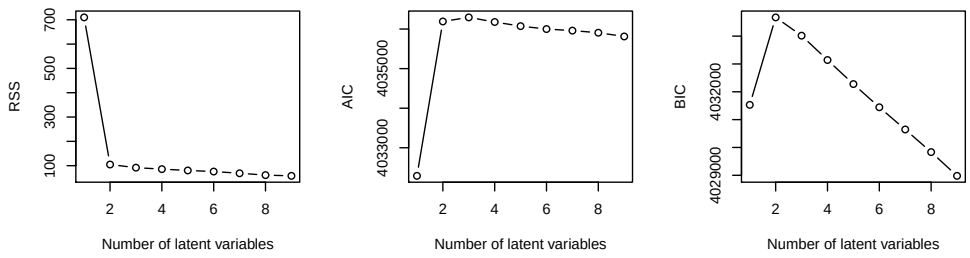

Figure 1: Determination of K via AIC, BIC, and deviance reduction. Optimal K is set at 2.

2.7 Fit the Poisson log-likelihood ratio based segmentation procedure to determine CNV regions across all samples.

For germline CNV detection, `CODEX` uses the "integer" mode; for CNV detection involving large recurrent chromosomal aberrations in mixture populations (e.g. somatic CNV detection in cancer), `CODEX` opts to use the "fraction" mode.
The output file is tab delimited and has 13 columns with rows corresponding to CNV events. The columns include sample_name (sample names), chr (chromosome), cnv (deletion or duplication), st_bp (cnv start position in base pair, the start position of the first exon in the cnv), ed_bp (cnv end position in base pair, the end position of the last exon in the cnv), length_kb (CNV length in kb), st_exon (the first exon after QC in the cnv, integer value numbered in qcObj$ref_qc), ed_exon (the last exon after QC in the cnv, integer value numbered in qcObj$ref_qc), raw_cov (raw coverage), norm_cov (normalized coverage), copy_no (copy number estimate), lratio (likelihood ratio of CNV event versus copy neutral event), mBIC (modified BIC value, used to determine the stop point of segmentation), pvalue (p-values computed by the Wilk's theorem from the likelihood ratio test).
For the "fraction" mode post segmentation thresholding is necessary to filter out long CNV events with factional copy numbers close to 2.

```
> optK = K[which.max(BIC)]
> finalcall <- segment(Y_qc, Yhat, optK = optK, K = K, sampname_qc,
+     ref_qc, chr, lmax = 200, mode = "integer")
> finalcall

 sample_name chr  cnv   st_bp      ed_bp      length_kb st_exon ed_exon raw_cov
 "NA18990"   "22" "dup" "22312814" "22326373" "13.56"   "60"    "72"    "1382"
 norm_cov copy_no lratio  mBIC
 "1000"   "3"     "60.33" "49.728"

> write.table(finalcall, file = paste(projectname, '_', chr, '_', optK,
+             '_CODEX_frac.txt', sep=''), sep='\t', quote=FALSE, row.names=FALSE)
> save.image(file = paste(projectname, '_', chr, '_image', '.rda', sep=''),
+      compress='xz')
```

3. Citation

CODEX: a normalization and copy number variation detection method for whole exome sequencing Yuchao Jiang; Derek A. Oldridge; Sharon J. Diskin; Nancy R. Zhang Nucleic Acids Research 2015; doi: 10.1093/nar/gku1363 (html, pdf).

4. Session information:

Output of sessionInfo on the system on which this document was compiled:

- R version 3.3.0 RC (2016-04-26 r70550), `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, parallel, stats, stats4, utils
- Other packages: BSgenome 1.40.0, BSgenome.Hsapiens.UCSC.hg19 1.4.0, BiocGenerics 0.18.0, Biostrings 2.40.0, CODEX 1.4.0, GenomeInfoDb 1.8.0, GenomicRanges 1.24.0, IRanges 2.6.0, Rsamtools 1.24.0, S4Vectors 0.10.0, WES.1KG.WUGSC 1.3.0, XVector 0.12.0, rtracklayer 1.32.0
- Loaded via a namespace (and not attached): Biobase 2.32.0, BiocParallel 1.6.0, GenomicAlignments 1.8.0, RCurl 1.95-4.8, SummarizedExperiment 1.2.0, XML 3.98-1.4, bitops 1.0-6, tools 3.3.0, zlibbioc 1.18.0