

Preparing Data from Nagalakshmi et al.

Martin Morgan

Fred Hutchinson Cancer Research Center

9-10 December, 2010

This data set is based on Nagalakshmi et al. [1]. The data was retrieved from the [sequence read archive](http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP000227), aligned with [bwa](#), and converted to BAM format with [samtools](#).

Data were downloaded from the sequence read archive by visiting <http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP000227> and downloading reads for the entire study. This results in a directory structure

```
SRP000227
|---SRR002051
|   |-- SRR002051.sra
|---SRR002058
|   |-- SRR002058.sra
|---SRR002059
|   |-- SRR002059.sra
|---SRR002061
|   |-- SRR002061.sra
|---SRR002062
|   |-- SRR002062.sra
|---SRR002064
|   |-- SRR002064.sra
```

Fastq files were extracted using [NCBI SRA SDK](#).

```
#!/usr/bin/env sh
SRA=/home/mtmorgan/bin/sratoolkit.2.0b4-3-suse_linux32
for f in `ls */*.sra`
do
    ${SRA}/sratoolkit.2.0b4-3-suse_linux32/fastq-dump $f
done
```

Reference sequences were retrieved from [UCSC](#) golden path, unzipped, and catenated into a single file. The file was indexed for use by the [bwa](#) aligner with `bwa index sacCer2.fa`. Reads were aligned and converted to BAM as

```

#!/usr/bin/env sh
BWA=/home/mtmorgan/bin/bwa/bwa
SAMTOOLS=/home/mtmorgan/bin/samtools/samtools
for f in `ls ./SRP000227/*fastq`
do
    echo "processing: $f"
    g=`basename $f`
    ${BWA} aln -t 2 Sacer2.fa $f > aln/$g.sai
    ${BWA} samse Sacer2.fa aln/$g.sai $f > aln/$g.sam
    ${SAMTOOLS} view -Sb aln/$g.sam > aln/$g.bam
    rm aln/$g.sai aln/$g.sam
done

```

Descriptive information about samples (e.g., protocol and replicate) were extracted by hand from web pages at the SRA.

Objects used in the lab were processed using the following *R* script. This script requires files and a directory structure that are not distributed with this package.

```

> library(SeattleIntro2010)
> readScript("create-objects.R")

[1] ## http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP000227
[2] library(SeattleIntro2010)
[3] datasrc <- "/home/mtmorgan/SeattleIntro2010/NagalakshmiEtAl"
[4] pkgroot <- "/home/mtmorgan/SeattleIntro2010"
[5] setwd(datasrc)
[6]
[7] ## qa
[8] qaFile <- file.path(pkgroot, "data", "qa.rda")
[9] if (!file.exists(qaFile)) {
[10]     ## better: qa on the aligned reads?
[11]     fls <- list.files("SRP000227", pattern="fastq", full=TRUE)
[12]     qalst <- Map(function(fl) {
[13]         fq <- readFastq(fl)
[14]         qa(fq, lane=basename(fl))
[15]     }, fls)
[16]     qa <- do.call(rbind, qalst)
[17]     save(qa, file=qaFile)
[18] } else load(qaFile)
[19] if (interactive())
[20]     browseURL(report(qa))
[21]
[22] ## hitspergene
[23] countsFile <- file.path(pkgroot, "data", "hitspergene.rda")
[24] txdbFile <- file.path(pkgroot, "inst", "extdata", "sacCer2_sgdGene.sqlite")
[25] if (!file.exists(countsFile)) {

```

```

[26]
[27] ## transcript ranges
[28] library(GenomicFeatures)
[29] if (!file.exists(txdbFile)) {
[30]     txdb <- makeTranscriptDbFromUCSC(genome="sacCer2",
[31]                                     tablename="sgdGene")
[32]     saveFeatures(txdb, txdbFile)
[33] } else txdb <- loadFeatures(txdbFile)
[34] exons <- exons(txdb, column="gene_id")
[35] strand(exons) <- "*" # protocol doesn't distinguish strand
[36]
[37] ## reads and counts
[38] fls <- list.files("aln", pattern="fastq.sorted.bam$", full=TRUE)
[39] cnt <- Map(function(fl, exons) {
[40]     print(fl)
[41]     allGenes <- as.character(values(exons)[["gene_id"]])
[42]     uniqueGenes <- unique(allGenes)
[43]     tmp1 <- structure(numeric(length(uniqueGenes))), .Names=uniqueGenes)
[44]
[45]     ga <- readGappedAlignments(fl)
[46]     olap <- findOverlaps(ga, exons)
[47]     ## divide reads amongst hits
[48]     wt <- local({
[49]         x <- tabulate(queryHits(olap))
[50]         ifelse(x, 1 / x, 0)[queryHits(olap)]
[51]     })
[52]     hits <- sapply(split(wt, allGenes[subjectHits(olap)]), sum)
[53]     tmp1[names(hits)] <- hits
[54]     tmp1
[55] }, fls, MoreArgs=list(exons=exons))
[56] hitspergene <- as(cnt, "DataFrame")
[57] dimnames(hitspergene) <-
[58]     list(names(cnt[[1]]), sub(".fastq.sorted.bam", "", basename(fls)))
[59]
[60] ## sample annotations
[61] df <- DataFrame(Sample=rep(c("RH", "dT"), each=3),
[62]                 Replicate=rep(c("Biological", "Original", "Technical"), 2),
[63]                 SRR=c("SRR002058", "SRR002059", "SRR002061",
[64]                       "SRR002062", "SRR002051", "SRR002064"))
[65] elementMetadata(hitspergene) <-
[66]     df[match(colnames(hitspergene), df$SRR),]
[67] o <- with(elementMetadata(hitspergene),
[68]           order(Sample, Replicate))
[69] hitspergene <- hitspergene[,o]
[70]
[71] save(hitspergene, file=countsFile)

```

```

[72] } else load(countsFile)
[73]
[74] ## SRR002051.pluscvrg and SRR002051.minuscvrg
[75] bamFile <- file.path("aln", "SRR002051.fastq.bam")
[76] pluscvrgFile <- file.path(pkgroot, "data", "SRR002051.pluscvrg.rda")
[77] minuscvrgFile <- file.path(pkgroot, "data", "SRR002051.minuscvrg.rda")
[78] if (!file.exists(pluscvrgFile) || !file.exists(minuscvrgFile)) {
[79]     bam <- readGappedAlignments(bamFile)
[80]     library(BSgenome.Scerevisiae.UCSC.sacCer2)
[81]     bam@seqlengths <- seqlengths(Scerevisiae)
[82]     SRR002051.pluscvrg <- coverage(grg(bam)[strand(bam) == "+"])
[83]     SRR002051.minuscvrg <- coverage(grg(bam)[strand(bam) == "-"])
[84]     save(SRR002051.pluscvrg, file=pluscvrgFile)
[85]     save(SRR002051.minuscvrg, file=minuscvrgFile)
[86] } else {
[87]     load(pluscvrgFile)
[88]     load(minuscvrgFile)
[89] }
[90]

```

Relevant software versions are

Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.11 (r851)

Program: bwa (alignment via Burrows-Wheeler transformation)
Version: 0.5.8c (r1536)
Contact: Heng Li <lh3@sanger.ac.uk>

sratoolkit.2.0b4-3-suse_linux32/fastq-dump
Version: 2.0.0

> sessionInfo()

R version 2.12.0 Patched (2010-11-28 r53696)
Platform: i386-pc-mingw32/i386 (32-bit)

locale:

```

[1] LC_COLLATE=C
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

```

attached base packages:

```

[1] stats      graphics  grDevices  utils      datasets  methods
[7] base

```

other attached packages:

```
[1] SeattleIntro2010_0.0.41 biomaRt_2.6.0
[3] GO.db_2.4.5             hgu95av2.db_2.4.5
[5] org.Hs.eg.db_2.4.6      RSQLite_0.9-4
[7] DBI_0.2-5               AnnotationDbi_1.12.0
[9] Biobase_2.10.0
```

loaded via a namespace (and not attached):

```
[1] RCurl_1.5-0.1 XML_3.2-0.2  tools_2.12.0
```

References

- [1] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320:1344–1349, Jun 2008.