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, 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/mtmogan/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 `cat`enated 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 Saccer2.fa $f > aln/$g.sai
    ${BWA} samse Saccer2.fa aln/$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)) {
```

```
## transcript ranges
library(GenomicFeatures)
if (!file.exists(txdbFile)) {
    txdb <- makeTranscriptDbFromUCSC(genome="sacCer2",
                                     tablename="sgdGene")
    saveFeatures(txdb, txdbFile)
} else txdb <- loadFeatures(txdbFile)
exons <- exons(txdb, column="gene_id")
strand(exons) <- "*"         # protocol doesn't distinguish strand

## reads and counts
fls <- list.files("aln", pattern="fastq.sorted.bam$", full=TRUE)
cnt <- Map(function(fl, exons) {
    print(fl)
    allGenes <- as.character(values(exons)[["gene_id"]])
    uniqueGenes <- unique(allGenes)
    tmpl <- structure(numeric(length(uniqueGenes)), .Names=uniqueGenes)

    ga <- readGappedAlignments(fl)
    olap <- findOverlaps(ga, exons)
    ## divide reads amongst hits
    wt <- local({
        x <- tabulate(queryHits(olap))
        ifelse(x, 1 / x, 0)[queryHits(olap)]
    })
    hits <- sapply(split(wt, allGenes[subjectHits(olap)]), sum)
    tmpl[names(hits)] <- hits
    tmpl
}, fls, MoreArgs=list(exons=exons))
hitspergene <- as(cnt, "DataFrame")
dimnames(hitspergene) <-
    list(names(cnt[[1]]), sub(".fastq.sorted.bam", "", basename(fls)))

## sample annotations
df <- DataFrame(Sample=rep(c("RH", "dT"), each=3),
    Replicate=rep(c("Biological", "Original", "Technical"), 2),
    SRR=c("SRR002058", "SRR002059", "SRR002061",
        "SRR002062", "SRR002051", "SRR002064"))
elementMetadata(hitspergene) <-
    df[match(colnames(hitspergene), df$SRR),]
o <- with(elementMetadata(hitspergene),
          order(Sample, Replicate))
hitspergene <- hitspergene[,o]

save(hitspergene, file=countsFile)
```

```
} else load(countsFile)

## SRR002051.pluscvg and SRR002051.minuscvg
bamFile <- file.path("aln", "SRR002051.fastq.bam")
pluscvgFile <- file.path(pkgroot, "data", "SRR002051.pluscvg.rda")
minuscvgFile <- file.path(pkgroot, "data", "SRR002051.minuscvg.rda")
if (!file.exists(pluscvgFile) || !file.exists(minuscvgFile)) {
    bam <- readGappedAlignments(bamFile)
    library(BSgenome.Scerevisiae.UCSC.sacCer2)
    bam@seqlengths <- seqlengths(Scerevisiae)
    SRR002051.pluscvg <- coverage(grg(bam)[strand(bam) == "+"])
    SRR002051.minuscvg <- coverage(grg(bam)[strand(bam) == "-"])
    save(SRR002051.pluscvg, file=pluscvgFile)
    save(SRR002051.minuscvg, file=minuscvgFile)
} else {
    load(pluscvgFile)
    load(minuscvgFile)
}

```

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 2.12.1 beta (2010-12-07 r53813)
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

locale:
[1] C/en_US.UTF-8/C/C/en_US.UTF-8/en_US.UTF-8

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_1.4-3  XML_3.2-0    tools_2.12.1
```

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.