ChIP-Seq Analysis with R and Bioconductor

Advanced R/Bioconductor Workshop on High-Throughput Genetic Analysis

-

Fred Hutchinson Cancer Research Center - Seattle, WA

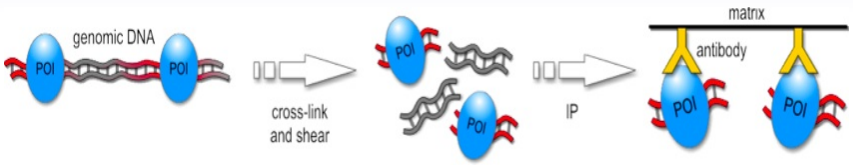
Thomas Girke

February 27-28, 2012

Outline

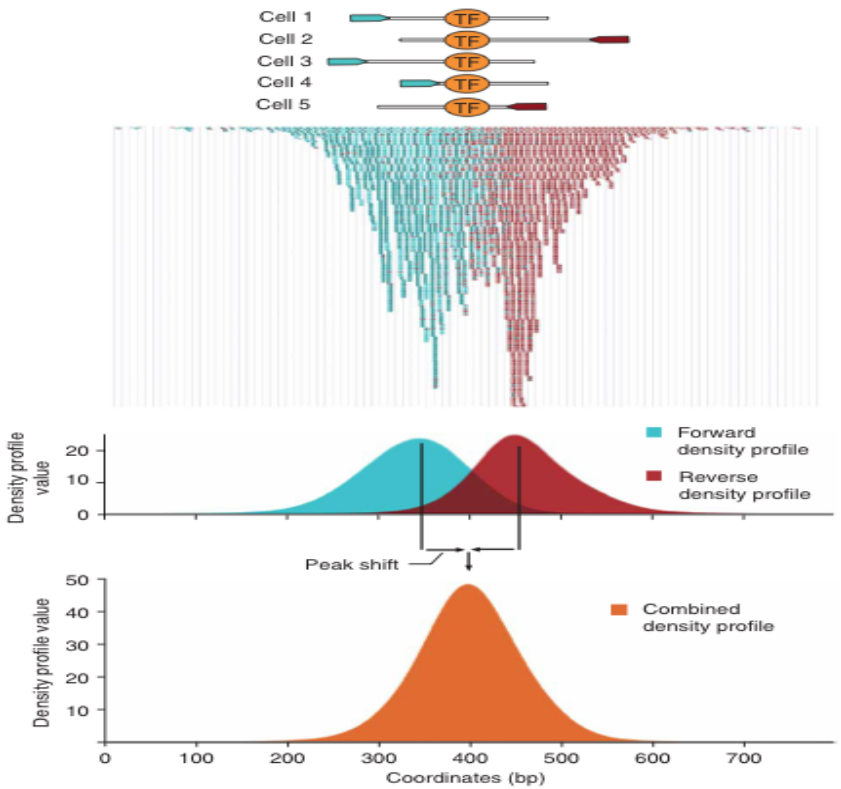
Outline

ChIP-Seq Technology



⇓

ChIP-Seq



ChIP-Seq Workflow

- Read mapping
- Peak calling
- Peak annotation/filtering
- Differential peak analysis
- Motif enrichment analysis in sequences under peaks

Peak Callers (Command-line Tools)

- CisGenome
- ERANGE
- FindPeaks
- F-Seq
- GLITR
- MACS
- PeakSeq
- QuEST
- SICER
- SiSSRs
- spp
- USeq
- ...

Pepke, S, Wold, B, Mortazavi, A (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6, 22-32. Link

Outline

General Purpose Resources for ChIP-Seq Analysis in R

- GenomicRanges Link: high-level infrastructure for range data
- Rsamtools Link: BAM support
- rtracklayer Link: Annotation imports, interface to online genome browsers
- DESeq Link: RNA-Seq analysis
- edgeR Link: RNA-Seq analysis
- chipseq Link: Utilities for ChIP-Seq analysis
- ChIPpeakAnno Link: Annotating peaks with genome context information
- ...

Peak Calling in R

- BayesPeak Link: hidden Markov models (HMM) and Bayesian statistics
- PICS Link: probabilistic inference
- DiffBind Link: Differential binding analysis of ChIP-Seq peak data
- MOSAiCS Link: model-based analysis of ChIP-Seq data
- iSeq Link: Hidden Ising Models
- ChIPseqR Link
- CSAR Link: tests based on Poisson distribution
- ChIP-Seq Link
- SPP Link

Outline

Outline

Data Sets and Experimental Variables

- To make the following sample code work, users can download the sample data into the directory of their current R session as shown below.
- It contains slimmed down versions of four FASTQ files from the ChIP-Seq experiment published by Kaufman et al (2010, GSE20176 Link), a shortened GFF3 annotation file Link and the corresponding reference genome from *Arabidopsis thaliana*.

For RStudio/AMI setup of workshop: create a symbolic link to data sets located under `"/R-pkgs/"`

```
> system("ln -s /R-pkgs/data data")
```

Users working on local Unix/Linux or OS X systems want to download the sample data from the Internet like this:

```
> system("wget http://biocluster.ucr.edu/~tgirke/HTML_Presentations/Manuals/Rngsapps/chipseqBioc2012/data.z
> system("unzip data.zip")
> download.file("http://faculty.ucr.edu/~tgirke/HTML_Presentations/Manuals/Rngsapps/chipseqBioc2012/Rchipse
> download.file("http://faculty.ucr.edu/~tgirke/HTML_Presentations/Manuals/Rngsapps/chipseqBioc2012/Rchipse
```

Windows users can download the same files manually from here: Slides Link, R code Link, Data Sets Link

Kaufmann et al (2010) Orchestration of floral initiation by APETALA1. Science 328, 85-89. Link

Outline

Align Reads and Output Indexed Bam Files

Note: Rsubread is Linux only. OS X/Windows users want to skip the mapping step and download the Bam files from here: Link

```
> library(Rsubread); library(Rsamtools)
> dir.create("results") # Note: all output data will be written to directory 'results'
> buildindex(basename="./results/tair10chr.fasta", reference="./data/tair10chr.fasta") # Build indexed refe
> targets <- read.delim("./data/targets.txt") # Import experiment design information
> targets

   Samples Factor           Fastq
1  AP1IND1   sig1 SRR038845.fastq
2  AP1IND2   sig1 SRR038846.fastq
3 AP1UIND1   bgr1 SRR038848.fastq
4 AP1UIND2   bgr1 SRR038850.fastq

> input <- paste("./data/", targets[,3], sep="")
> output <- paste("./results/", targets[,3], ".sam", sep="")
> reference <- "./results/tair10chr.fasta"
> for(i in seq(along=targets[,3])) {
+         align(index=reference, readfile1=input[i], output_file=output[i], nthreads=8, indels=1, TH1=2)
+         asBam(file=output[i], destination=gsub(".sam", "", output[i]), overwrite=TRUE, indexDestination=T
+         unlink(output[i])
+ }
```

Outline

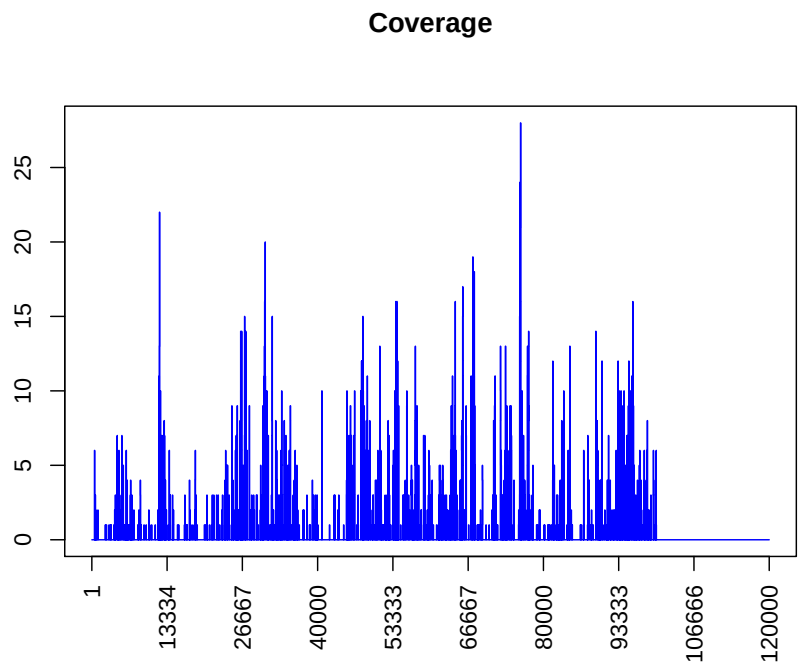
Important Resources for ChIP-Seq Analysis

Coverage and peak slicing

```
> library(rtracklayer); library(GenomicRanges)
> samples <- as.character(targets$Fastq)
> samplespath <- paste("./results/", samples, ".bam", sep="")
> aligns <- readBamGappedAlignments(samplespath[3])
> cov <- coverage(aligns)
> islands <- slice(cov, lower = 15)
> islands[[1]]

Views on a 30427671-length Rle subject
views:
        start      end width
 [1]    11998    12003     6 [22 22 22 22 22 22]
 [2]    27064    27077    14 [15 15 15 15 15 15 15 15 15 15 15 15 15 15]
 [3]    30619    30623     5 [16 16 16 16 16]
 [4]    30709    30725    17 [16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 20 20]
 [5]    31956    31975    20 [15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15]
 [6]    48015    48029    15 [15 15 15 15 15 15 15 15 15 15 15 15 15 15 15]
 [7]    53839    53844     6 [16 16 16 16 16 16]
 [8]    54031    54066    36 [16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16]
 [9]    64384    64394    11 [16 16 16 16 16 16 16 16 16 15 15]
 ...      ...      ...   ... ...
[43] 13972892 13972899     8 [15 15 15 15 15 15 15 15]
[44] 13973159 13973206    48 [15 15 15 15 15 15 15 15 15 15 27 27 27 30 30 30 26 23 23 24 24 24 24 24 24 36 3
[45] 13973859 13973894    36 [26 26 26 26 26 27 27 27 27 27 27 27 27 27 27 27 27 27 18 18 18 18 18 18 18 1
[46] 15562796 15562831    36 [15 15 15 15 15 15 24 24 24 24 24 24 24 29 29 29 29 29 28 28 28 28 28 28 28 2
[47] 15563233 15563262    30 [17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 1
[48] 15564861 15564887    27 [15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 1
[49] 15871554 15871618    65 [19 19 19 19 28 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 47 4
[50] 20477102 20477112    11 [17 17 17 17 17 17 17 17 17 17 17]
[51] 20558301 20558332    32 [20 20 20 20 20 20 20 20 19 19 19 19 19 19 19 19 18 18 17 17 17 17 17 17 17 1
```

Coverage Plot

```
> plotCov <- function(mycov=cov, mychr=1, mypos=c(1,1000), mymain="Coverage", ...) {
+     op <- par(mar=c(8,3,6,1))
+     plot(as.numeric(mycov[[mychr]][mypos[1]:mypos[2]]), type="l",
+     lwd=1, col="blue", ylab="", main=mymain, xlab="", xaxt="n", ...)
+     axis(1, las = 2, at=seq(1,mypos[2]-mypos[1], length.out= 10),
+     labels=as.integer(seq(mypos[1], mypos[2], length.out= 10)))
+     par(op)
+ }
> plotCov(mycov=cov, mychr="Chr1", mypos=c(1,120000)) # Remember: read data is trunctated to first 100kbp
```



Import Aligned Read Data

Import aligned reads (bam files) and extend to 200bp

```
> chip_signal_list <- sapply(samplespath, list)
> for(i in seq(along=samplespath)) {
+         aligns <- readBamGappedAlignments(samplespath[i])
+         chip_signal_list[[i]] <- as(aligns, "GRanges")
+ }
> chip_signal_list[["./results/SRR038845.fastq.bam"]][1:4,]

GRanges with 4 ranges and 0 elementMetadata cols:
      seqnames     ranges strand
         <Rle>  <IRanges>  <Rle>
  [1]     Chr1 [121, 156]      -
  [2]     Chr1 [121, 156]      -
  [3]     Chr1 [216, 251]      +
  [4]     Chr1 [295, 330]      +
  ---
  seqlengths:
       Chr1     Chr2     Chr3     Chr4     Chr5     ChrC     ChrM
   30427671 19698289 23459830 18585056 26975502   154478   366924

> chip_signal_list <- sapply(names(chip_signal_list), function(x) resize(chip_signal_list[[x]], width = 200
```

Outline

Naive Peak Calling by Coverage Value

Compute coverage and call peaks

```
> Nreads <- sapply(names(chip_signal_list), function(x)  length(chip_signal_list[[x]]))
> normfactor <- 10^6/Nreads
> chip_signal_list <- sapply(names(chip_signal_list), function(x) coverage(chip_signal_list[[x]]) * normfac
> chip_signal_list[["./results/SRR038845.fastq.bam"]][1:2,]

SimpleRleList of length 2
$Chr1
'numeric' Rle of length 30427671 with 4709 runs
  Lengths:              139               17               59               10               69
  Values : 6.42287571414349  12.845751428287 6.42287571414349 9.63431357121524 22.4800649995022 35.3258164
$Chr2
'numeric' Rle of length 19698289 with 21336 runs
  Lengths:              821                2               12                1                1
  Values :                0 6.42287571414349 9.63431357121524 16.0571892853587 19.2686271424305 22.480064

> chip_peak_list <- sapply(names(chip_signal_list), function(x) slice(chip_signal_list[[x]], lower=5))
> chip_peak_list[[1]][[1]][1:3]

Views on a 30427671-length Rle subject
views:
    start  end width
[1]     1  689   689 [ 6.422876  6.422876  6.422876  6.422876  6.422876  6.422876  6.422876  6.422876  6.4
[2]   764  963   200 [9.634314 9.634314 9.634314 9.634314 9.634314 9.634314 9.634314 9.634314 9.634314 9.6
[3]  1463 1697   235 [12.84575 12.84575 12.84575 12.84575 12.84575 12.84575 12.84575 12.84575 12.84575 12.
```

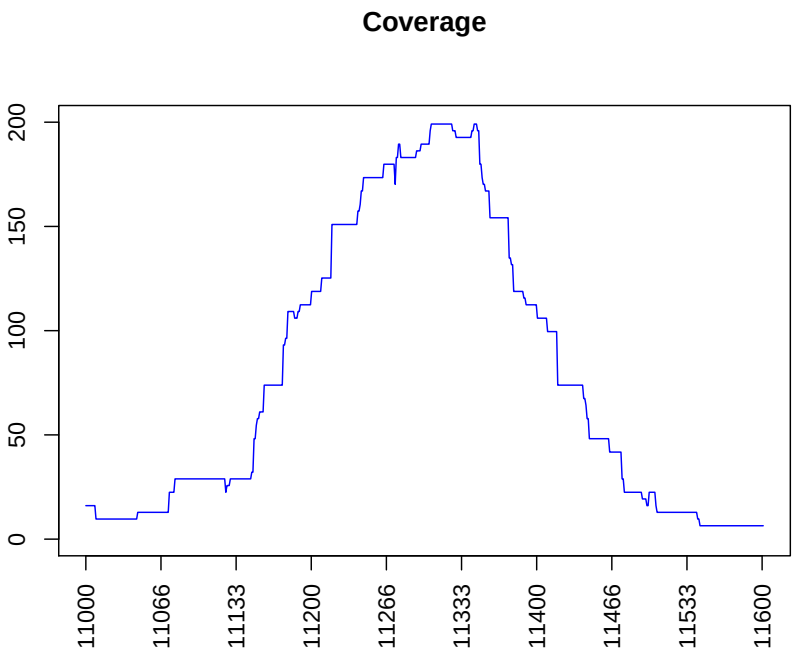
Peak Calling with BayesPeak

Compute coverage and call peaks

```
> library(BayesPeak)
> sig <- readBamGappedAlignments(samplespath[1])
> bgr <- readBamGappedAlignments(samplespath[3])
> sig <- as(as(sig, "GRanges"), "RangedData")
> bgr <- as(as(bgr, "GRanges"), "RangedData")
> raw.output <- bayespeak(treatment=sig, control=bgr, start = 1, end = 100000)

..............

> # unreliable.jobs <- log(raw.output$QC$lambda1) < 1.5 # Removal of false positives due to overfitting.
> # bpeaks <- as.data.frame(summarise.peaks(raw.output, method = "lowerbound", exclude.jobs = unreliable.jo
> bpeaks <- as.data.frame(summarise.peaks(raw.output, method = "lowerbound"))
> source("./data/Fct/chipseqFct.R") # Imports the rangeCoverage function.
> sigcovDF <- rangeCoverage(summaryFct=viewMeans, myname="sig_", peaksIR=bpeaks[,1:3], sig=sig, readextend=
> bgrcovDF <- rangeCoverage(summaryFct=viewMeans, myname="bgr_", peaksIR=bpeaks[,1:3], sig=bgr, readextend=
> bpeaksDF <- cbind(bpeaks, sigcovDF[,-1], bgrcovDF[,-1])
> bpeaksDF[1:4,]

  space start   end width        PP   sig_cov sig_cov.pos sig_cov.neg     bgr_cov bgr_cov.pos bgr_cov.neg
1  Chr1  8301  8401   101 0.5717232 0.1907785   0.1271857  0.06359283 0.039603881 0.039603881 0.000000000
2  Chr1 11151 11351   201 0.9998908 0.6390921   0.5911602  0.04793191 0.009950229 0.009950229 0.000000000
3  Chr1 13601 13751   151 0.9998793 0.4040882   0.2339458  0.17014240 0.092715046 0.079470040 0.013245007
4  Chr1 16601 17051   451 0.9999922 0.4699665   0.1922590  0.27770749 0.008869162 0.004434581 0.004434581
```

Coverage Plot

```
> plotCov(mycov=chip_signal_list[[1]], mychr="Chr1", mypos=c(11000, 11600), ylim=c(0,200))
```



Identify Common Peaks Among Two Methods

Compares results from simple cutoff method with BayesPeak results

```
> simple_peak <- as.data.frame(as(chip_peak_list[[1]], "IRangesList"))
>    # simple_peak <- as.data.frame(chip_peak_list[[1]])
> commonpeaks <- subsetByOverlaps(as(bpeaks, "RangedData"), as(simple_peak, "RangedData"), minoverlap=100)
> bpeaksDF[bpeaksDF$start %in% start(commonpeaks),][1:4,]
```

	space	start	end	width	PP	sig_cov	sig_cov.pos	sig_cov.neg	bgr_cov	bgr_cov.pos	bgr_cov.neg
1	Chr1	8301	8401	101	0.5717232	0.1907785	0.1271857	0.06359283	0.039603881	0.039603881	0.000000000
2	Chr1	11151	11351	201	0.9998908	0.6390921	0.5911602	0.04793191	0.009950229	0.009950229	0.000000000
3	Chr1	13601	13751	151	0.9998793	0.4040882	0.2339458	0.17014240	0.092715046	0.079470040	0.013245007
4	Chr1	16601	17051	451	0.9999922	0.4699665	0.1922590	0.27770749	0.008869162	0.004434581	0.004434581

Exercise 1: Compare Results with Published Peaks

Task 1 Import peaks predicted by Kaufmann et al (2010).

Task 2 Determine how many of the published peaks have at least a 50% length overlap with the results from the BayesPeak and the naive peak calling methods.

Required information:

```
> pubpeaks <- read.delim("./data/Kaufmann_peaks100k.txt") # Published peaks for first 100kbp on chromosomes
> pubpeaks <- pubpeaks[order(pubpeaks$space, pubpeaks$start),]
> pubpeaks[1:4,]

              PeakID space start   end score_position     score length
chr1_3132   chr1_3132  Chr1  3094  3172           3132  4.656515     79
chr1_8365   chr1_8365  Chr1  8222  8425           8365 11.046212    204
chr1_11298 chr1_11298  Chr1 11149 11440          11298 52.109592    292
chr1_13686 chr1_13686  Chr1 13602 13756          13686  5.341907    155

> # Import olRanges function, which accepts two GRranges (IRanges) objects
> source("./data/Fct/rangeoverlapper.R")
```

Outline

Import Annotation Data from GFF

Annotation data from GFF

```
> library(rtracklayer); library(GenomicRanges); library(Rsamtools)
> gff <- import.gff("./data/TAIR10_GFF3_trunc.gff", asRangedData=FALSE)
> seqlengths(gff) <- end(ranges(gff[which(elementMetadata(gff)[,"type"]=="chromosome"),]))
> subgene_index <- which(elementMetadata(gff)[,"type"] == "gene")
> gffsub <- gff[subgene_index,] # Returns only gene ranges
> strand(gffsub) <- "*" # For strand insensitive analysis
> gffsub[1:4,1:2]

GRanges with 4 ranges and 2 elementMetadata cols:
      seqnames         ranges strand |   source     type
         <Rle>      <IRanges>  <Rle> | <factor> <factor>
  [1]     Chr1 [ 3631,  5899]      * |   TAIR10     gene
  [2]     Chr1 [ 5928,  8737]      * |   TAIR10     gene
  [3]     Chr1 [11649, 13714]      * |   TAIR10     gene
  [4]     Chr1 [23146, 31227]      * |   TAIR10     gene
  ---
  seqlengths:
       Chr1     Chr2     Chr3     Chr4     Chr5     ChrC     ChrM
   30427671 19698289 23459830 18585056 26975502   154478   366924

> ids <- elementMetadata(gffsub)[, "group"]
> gffgene <- gffsub
> gffsub <- split(gffsub) # Coerce to GRangesList
```

Annotate Peaks with ChIPpeakAnno

```
> library(ChIPpeakAnno)
> annoRD <- unlist(gffsub)
> names(annoRD) <- gsub(".*=", "", elementMetadata(annoRD)[, "group"])
> annoRD <- as(annoRD, "RangedData")
> peaksRD <- RangedData(space=bpeaksDF$space, IRanges(bpeaksDF$start, bpeaksDF$end))
> annotatedPeak <- annotatePeakInBatch(peaksRD, AnnotationData = annoRD)
> as.data.frame(annotatedPeak)[1:4,1:11]

  space start   end width        names peak strand   feature start_position end_position insideFeature
1  Chr1  8301  8401   101 01 AT1G01020   01      + AT1G01020           5928         8737        inside
2  Chr1 11151 11351   201 02 AT1G01030   02      + AT1G01030          11649        13714      upstream
3  Chr1 13601 13751   151 03 AT1G01030   03      + AT1G01030          11649        13714    overlapEnd
4  Chr1 16601 17051   451 04 AT1G01030   04      + AT1G01030          11649        13714    downstream

> bpeaksDF[1:4,]

  space start   end width        PP   sig_cov sig_cov.pos sig_cov.neg     bgr_cov bgr_cov.pos bgr_cov.neg
1  Chr1  8301  8401   101 0.5717232 0.1907785   0.1271857  0.06359283 0.039603881 0.039603881 0.000000000
2  Chr1 11151 11351   201 0.9998908 0.6390921   0.5911602  0.04793191 0.009950229 0.009950229 0.000000000
3  Chr1 13601 13751   151 0.9998793 0.4040882   0.2339458  0.17014240 0.092715046 0.079470040 0.013245007
4  Chr1 16601 17051   451 0.9999922 0.4699665   0.1922590  0.27770749 0.008869162 0.004434581 0.004434581
```

Alternative Peak Annotation Approach

Alternative approach using `olRanges` function

```
> source("./data/Fct/rangeoverlapper.R")
> olRanges(query=gffgene, subject=as(as(bpeaks, "RangedData"), "GRanges"), output="df")[1:2,]

  space Qindex Sindex Qstart  Qend Sstart  Send OLstart OLend OLlength  OLpercQ   OLpercS OLtype
1  Chr1      2      1   5928  8737   8301  8401    8301  8401      101 3.594306 100.00000 inside
2  Chr1      3      3  11649 13714  13601 13751   13601 13714      114 5.517909  75.49669 oldown

> as.data.frame(annotatedPeak)[c(2,5),1:11] # Corresponding result from ChIPpeakAnno

  space start   end width        names peak strand   feature start_position end_position insideFeature
2  Chr1 11151 11351   201 02 AT1G01030   02      + AT1G01030          11649        13714      upstream
5  Chr1 20901 21151   251 05 AT1G01040   05      + AT1G01040          23146        31227      upstream
```

Outline

Import Annotation Data from GFF

Annotation data from GFF

```
> peakranges <- GRanges(seqnames = Rle(bpeaksDF$space), ranges = IRanges(bpeaksDF$start, bpeaksDF$end),
+                strand = Rle(strand("*")), peakIDs=paste("peak", seq(along=bpeaksDF[,1]), sep="_"))
> countDF <- data.frame(row.names=elementMetadata(peakranges)[,"peakIDs"])
> peakranges <- split(peakranges) # Coerce to GRangesList
> for(i in samplespath) {
+         aligns <- readBamGappedAlignments(i) # Substitute next two lines with this one.
+         counts <- countOverlaps(peakranges, aligns)
+         countDF <- cbind(countDF, counts)
+ }
> colnames(countDF) <- samples
> rownames(countDF) <- gsub(".*=", "", rownames(countDF))
> countDF[1:4,]
```

	SRR038845.fastq	SRR038846.fastq	SRR038848.fastq	SRR038850.fastq
peak_1	9	36	2	5
peak_2	54	69	1	9
peak_3	31	13	7	3
peak_4	75	101	2	13

```
> write.table(countDF, "./results/countDF", quote=FALSE, sep="\t", col.names = NA)
> countDF <- read.table("./results/countDF")
```

Simple RPKM Normalization

RPKM: here defined as reads per kilobase of sequence range per million mapped reads

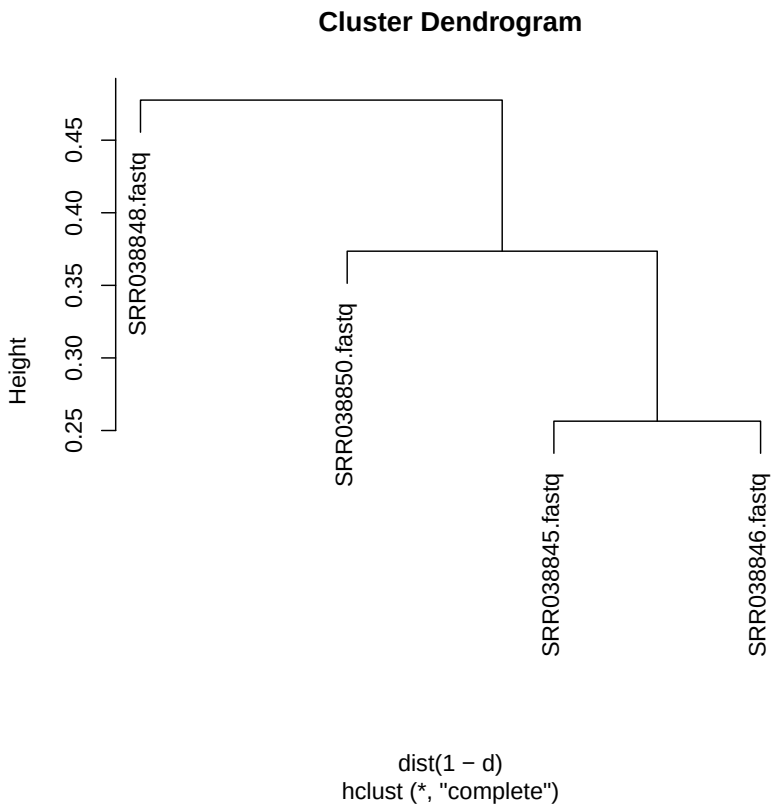
```
> returnRPKM <- function(counts, ranges) {
+         geneLengthsInKB <- sum(width(ranges))/1000 # Number of bases per sequence range in kbp
+         millionsMapped <- sum(counts)/1e+06 # Factor for converting to million of mapped reads.
+         rpm <- counts/millionsMapped # RPK: reads per kilobase of sequence range.
+         rpkm <- rpm/geneLengthsInKB # RPKM: reads per kilobase of sequence range per million mapped reads
+         return(rpkm)
+ }
> countDFrpkm <- apply(countDF, 2, function(x) returnRPKM(counts=x, ranges=peakranges))
> countDFrpkm[1:4,]
```

	SRR038845.fastq	SRR038846.fastq	SRR038848.fastq	SRR038850.fastq
peak_1	3761.298	12458.864	600.4239	1881.3160
peak_2	11340.033	11999.146	150.8528	1701.6082
peak_3	8665.654	3009.288	1405.6281	755.0182
peak_4	7019.422	7827.844	134.4630	1095.4182

QC Check

QC check by computing a sample correlating matrix and plotting it as a tree

```
> d <- cor(countDFrpkm, method="spearman")
> plot(hclust(dist(1-d))) # Sample tree
```



Identify DiffPeaks with Simple Fold Change Method

Compute mean values for replicates

```
> source("./data/Fct/colAg.R")
> countDFrpkm_mean <- colAg(myMA=countDFrpkm, group=c(1,1,2,2), myfct=mean)
> countDFrpkm_mean[1:4,]
```

```
       SRR038845.fastq_SRR038846.fastq SRR038848.fastq_SRR038850.fastq
peak_1                        8110.081                        1240.8700
peak_2                       11669.590                         926.2305
peak_3                        5837.471                        1080.3232
peak_4                        7423.633                         614.9406
```

Log2 fold changes

```
> countDFrpkm_mean <- cbind(countDFrpkm_mean, log2ratio=log2(countDFrpkm_mean[,1]/countDFrpkm_mean[,2]))
> countDFrpkm_mean <- countDFrpkm_mean[is.finite(countDFrpkm_mean[,3]), ]
> degs2fold <- countDFrpkm_mean[countDFrpkm_mean[,3] >= 1 | countDFrpkm_mean[,3] <= -1,]
> degs2fold[1:4,]
```

```
       SRR038845.fastq_SRR038846.fastq SRR038848.fastq_SRR038850.fastq log2ratio
peak_1                        8110.081                        1240.8700  2.708364
peak_2                       11669.590                         926.2305  3.655239
peak_3                        5837.471                        1080.3232  2.433881
peak_4                        7423.633                         614.9406  3.593606
```

```
> write.table(degs2fold, "./results/degs2fold", quote=FALSE, sep="\t", col.names = NA)
> degs2fold <- read.table("./results/degs2fold")
```

Identify DiffPeaks with DESeq Library

Raw count data are expected here!

```
> library(DESeq)
> countDF <- read.table("./results/countDF")
> conds <- targets$Factor
> cds <- newCountDataSet(countDF, conds) # Creates object of class CountDataSet derived from eSet class
> counts(cds)[1:4, ] # CountDataSet has similar accessor methods as eSet class.

       SRR038845.fastq SRR038846.fastq SRR038848.fastq SRR038850.fastq
peak_1               9              36               2               5
peak_2              54              69               1               9
peak_3              31              13               7               3
peak_4              75             101               2              13

> cds <- estimateSizeFactors(cds) # Estimates library size factors from count data. Alternatively, one can
> cds <- estimateDispersions(cds, fitType="local") # Estimates the variance within replicates
> res <- nbinomTest(cds, "bgr1", "sig1") # Calls DEGs with nbinomTest
> res <- na.omit(res)
> res2fold <- res[res$log2FoldChange >= 1 | res$log2FoldChange <= -1,]
> res2foldpadj <- res2fold[res2fold$padj <= 0.2, ] # Here padj set very high for demo purpose
> res2foldpadj[1:2,1:8]

      id baseMean baseMeanA baseMeanB foldChange log2FoldChange        pval        padj
2 peak_2 26.86951  6.748212  46.99082   6.963447       2.799802 0.0007686464 0.009838673
4 peak_4 38.65216 10.122757  67.18157   6.636687       2.730463 0.0005413709 0.008661935
```

Identify DiffPeaks with edgeR Library

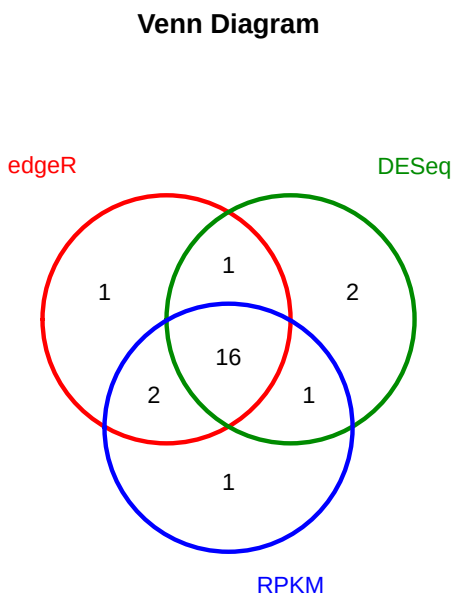
Raw count data are expected here!

```
> library(edgeR)
> countDF <- read.table("./results/countDF")
> y <- DGEList(counts=countDF, group=conds) # Constructs DGEList object
> y <- estimateCommonDisp(y) # Estimates common dispersion
> y <- estimateTagwiseDisp(y) # Estimates tagwise dispersion
> et <- exactTest(y, pair=c("bgr1", "sig1")) # Computes exact test for the negative binomial distribution.
> topTags(et, n=4)

Comparison of groups:  sig1-bgr1
          logFC   logCPM       PValue          FDR
peak_37 3.869453 10.52275 6.284620e-08 4.022157e-06
peak_48 3.910630 10.90523 8.975451e-07 2.085477e-05
peak_4  3.710208 10.81567 9.775674e-07 2.085477e-05
peak_34 3.268384 11.16464 4.043445e-05 6.469513e-04

> edge <- as.data.frame(topTags(et, n=50000))
> edge2fold <- edge[edge$logFC >= 1 | edge$logFC <= -1,]
> edge2foldpadj <- edge2fold[edge2fold$FDR <= 0.01, ]
```

Merge Results and Compute Overlaps Among Methods

Here overlaps for 20 best ranking peaks of each method!

```
> bothDF <- merge(res, countDFrpkm_mean, by.x=1, by.y=0, all=TRUE); bothDF <- na.omit(bothDF)
> cor(bothDF[,"log2FoldChange"], bothDF[,"log2ratio"], method="spearman")
[1] 0.9949634
> source("./data/Fct/overLapper.R")
> setlist <- list(edgeR=rownames(edge[order(edge$FDR),][1:20,]),
+                 DESeq=res[order(res$padj),][1:20,"id"],
+                 RPKM=rownames(degs2fold[order(-degs2fold$log2ratio),][1:20,]))
> OLlist <- overLapper(setlist=setlist, sep="_", type="vennsets")
> counts <- sapply(OLlist$Venn_List, length)
> vennPlot(counts=counts)
```



Outline

Inspect Results in IGV

View peak_3 in IGV

- Download and open IGV Link
- Select in menu in top left corner `A. thaliana (TAIR10)`
- Upload the following indexed/sorted Bam files with `File -> Load from URL...`

http://faculty.ucr.edu/~tgirke/HTML_Presentations/Manuals/Rngsapps/chipseqBioc2012/results/SRR038845.fastq.bam
http://faculty.ucr.edu/~tgirke/HTML_Presentations/Manuals/Rngsapps/chipseqBioc2012/results/SRR038846.fastq.bam
http://faculty.ucr.edu/~tgirke/HTML_Presentations/Manuals/Rngsapps/chipseqBioc2012/results/SRR038848.fastq.bam
http://faculty.ucr.edu/~tgirke/HTML_Presentations/Manuals/Rngsapps/chipseqBioc2012/results/SRR038850.fastq.bam

- To view peak_3, enter its coordinates `Chr1:16656-16956` in position menu on top.

Outline

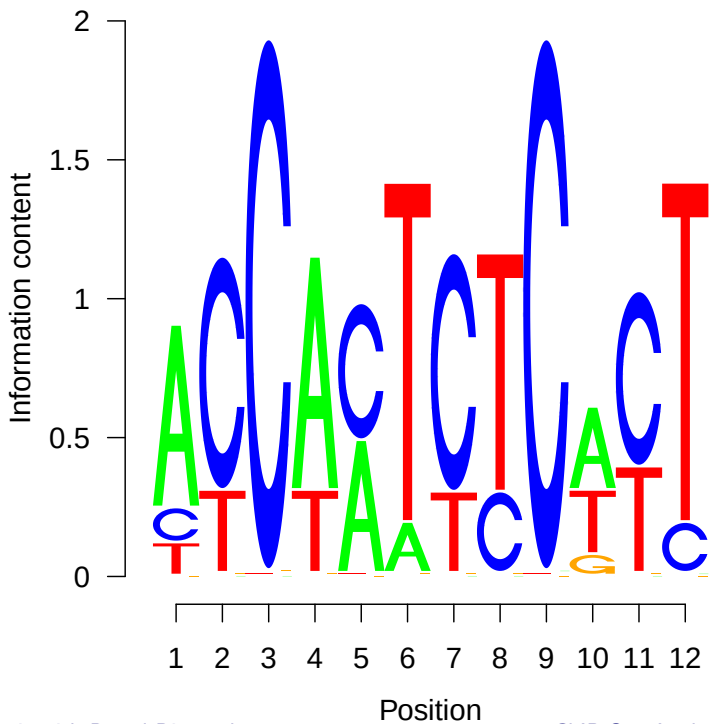
Sequence Motifs Enriched in Peak Sequences

Extract peak sequences and predict enriched motifs with COSMO library

```
> library(Biostrings); library(cosmo); library(seqLogo)
```

```
Welcome to cosmo version 1.22.0
cosmo is free for research purposes only. For more details, type
license.cosmo(). Type citation('cosmo') for details on how to cite
cosmo in publications.
```

```
> pseq <- getSeq(FaFile("./data/tair10chr.fasta"), as(as(bpeaksDF, "RangedData"), "GRanges"))
> names(pseq) <- paste(bpeaksDF$space, bpeaksDF$start, sep="_")
> write.XStringSet(pseq[1:8], "./results/pseq.fasta") # Note: reduced to 8 sequences to run quickly.
> res <- cosmo(seqs="./results/pseq.fasta", silent=TRUE)
> plot(res)
```



Exercise 2: Motif Enrichment Analysis

Task 1 Extract from the cosmo result stored in `res` the motif occurence patterns and generate with them a position weight matrix using the `PWM` function from Biostrings.

Task 2 Enumerate the motif matches in the peak sequences and the entire genome using Biostring's `countPWM` function.

Task 3 Determine which sequence type, peak or genome, shows more matches per 1kbp sequence for this motif.

Task 4 Homework: write a function for computing enrichment p-values for motif matches based on the hypergeometric distribution.

```
          seq pos orient        motif      prob
1  Chr1_8301  67     -1 AAAGAAATAGGG 1.0000000
2 Chr1_20901  63     -1 AAAGGGATTGGT 1.0000000
3 Chr1_37951  97     -1 AGTGAGTGTGGT 1.0000000
4 Chr1_54651 429      1 ATCACTTTCACC 1.0000000
5 Chr1_61351  62      1 ACCAATCCCATT 1.0000000
6 Chr1_11151 134     -1 AGAGAGAGAGAA 1.0000000
7 Chr1_16601  90      1 ACCACTCTCACT 0.9998101
8 Chr1_13601  70     -1 AGCGAGATTGGT 0.9359452
```

Session Information

```
> sessionInfo()

R version 2.15.0 (2012-03-30)
Platform: x86_64-unknown-linux-gnu (64-bit)
locale:
[1] C

attached base packages:
[1] grid      stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] cosmo_1.22.0                        seqLogo_1.22.0                      edgeR_2.6.1
 [7] limma_3.12.0                        org.Hs.eg.db_2.7.1                  GO.db_2.7.1
[13] BSgenome.Ecoli.NCBI.20080805_1.3.17 multtest_2.12.0                     Biobase_2.16.0
[19] caTools_1.12                        bitops_1.0-4.1                      gdata_2.8.2
[25] ShortRead_1.14.2                    latticeExtra_0.6-19                 RColorBrewer_1.0-5
[31] Rsamtools_1.8.4                     Biostrings_2.24.1                   GenomicRanges_1.8.4

loaded via a namespace (and not attached):
 [1] MASS_7.3-18        RCurl_1.91-1       XML_3.9-4          annotate_1.34.0    genefilter_1.38.0  genepl
[12] xtable_1.7-0       zlibbioc_1.2.0
```