Functional Genomics
Integrative analysis of next generation sequencing data

Alena van Bömmel (Alena.vanBoemmel@molgen.mpg.de R 3.3.83)
Robert Schöpflin (schoepfl@molgen.mpg.de 3.3.15)
Max Planck Institute for Molecular Genetics

Univariate and differential analysis steps

Next steps to do until March 28

1. Mapping the raw reads of RNA-seq experiments to the reference genome mm9. Use the tool `STAR`, save the output as a `bam` file and as a `wiggle` file, use the RPM normalization. The STAR index for the genome is already in:
   `/project/functional-genomics/2019/data/genome/STARindex/`

2. Convert the `wiggle` files to `BigWig` files using UCSC Tools. The genome version for STAR mapping is slightly different from the mm9 so there might be some problems with conversion of noncanonical chromosomes
3. Use the STAR mapper to produce a table with read counts per gene
4. Alternatively, you can use HTSeq tool to produce the read counts per gene. Use the gene annotation file in:
   `/project/functional-genomics/2019/data/genome/annotation`

Univariate and differential analysis steps

4. Compare the peaks of OSKM* in both stages. How many of them are unique for particular stage, how many are common? For an easier comparison, unify the peak widths: take the summit of the peaks and 100bp downstream and upstream of the summit
5. `bedtools` are very helpful for the comparison
6. Compare the peaks of OSKM* between the factors. How many peaks are shared among two, three or all four TFs? Determine for each TF the factor with the largest overlap.
7. Determine the genomic positions of the peaks (histone modifications, TFs of your choice). How many peaks are located in promoters, how many outside of promoters?
   You can download the promoter coordinates from UCSC genome browser (upstream2000.zip file) and use `bedtools` or you can use `Bioconductor` libraries (`biomaRt, GenomicRanges, GenomicFeatures`).

Differential and genomic analysis

Next steps

1. Count the reads of histone modifications experiments in all promoters. Use a normalized signal. Then calculate the correlation with the expression of the corresponding genes.
2. Use the following procedure for the normalization. Estimate first the slope of the correlation between the read counts of the sample (S) versus the read counts of the input control (C) (adding a pseudo-count of 1) by the median: m = median ((S+ 1)/ (C + 1))) of the ratio between the two over all promoters. The normalized read counts are then calculated by

   $$Snorm = (S+ 1)/ (C + 1) * 1/m$$

   Use log2(RPKM+1) values for the expression.

Differential and genomic analysis

Next steps

3. Which histone modifications are correlated with the expression?
4. Take the peaks of a transcription factor from BOTH stages. Split the peaks into 3 categories: shared, stage1-specific, stage2-specific. Plot a heatmap with read counts for the 3 categories with deepTools in both stages.
5. Find the binding motifs in these 3 categories of peaks with MEME (http://meme-suite.org/tools/meme) or with RSAT Tools (http://tagc.univ-mrs.fr/rsa-tools/peak-motifs_form.cgi). Do the found motifs differ?