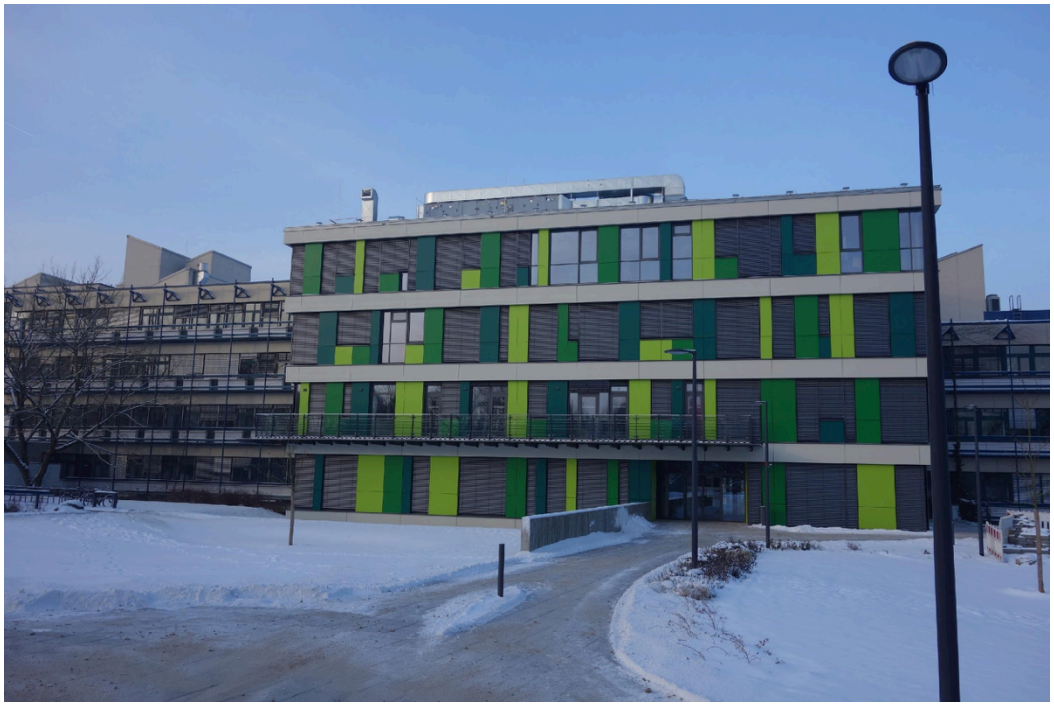
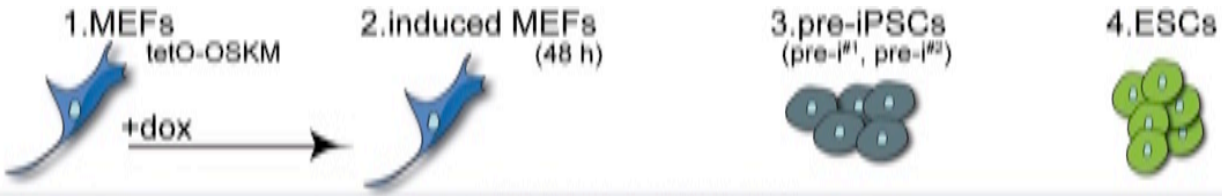
Functional Genomics
Integrative analysis of next generation sequencing data

Alena van Bömmel (Alena.vanBoemmel@molgen.mpg.de R 3.3.08)
Robert Schöpflin (schoepfl@molgen.mpg.de 3.3.15)
Max Planck Institute for Molecular Genetics



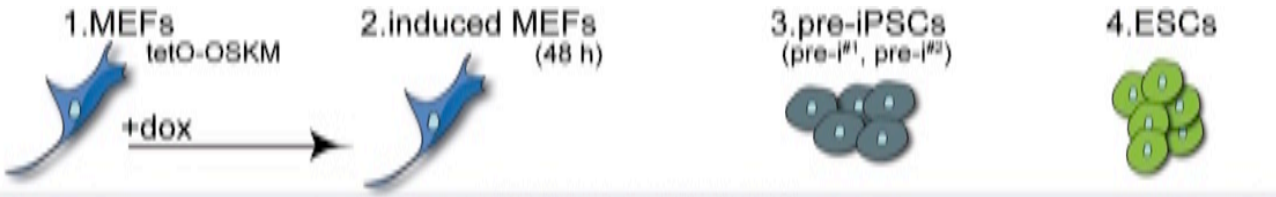
Data from Chronis, Fiziev et al. (2017)

4 reprogramming stages



Transcription factors		Epigenetic Regulators	Histone Modifications		Histones	Chromatin Accessibility	Expression
Oct4	Esrrb	p300	H3K27ac	H3K4me3	H3.3	ATAC-seq	RNA-seq
Sox2	Runx1	Hdac1	H3K27me3	H3K4me2	H3		
Klf4	Fra1	Brg1	H3K9me3	H3K4me1			
cMyc	Cebpa		H3K9ac	H3K36me3			
Nanog	Cebpb		H3K79me2				

4 reprogramming stages



Transcription factors		Epigenetic Regulators	Histone Modifications		Histones	Chromatin Accessibility	Expression
Oct4	Esrrb	p300	H3K27ac	H3K4me3	H3.3	ATAC-seq	RNA-seq
Sox2	Runx1	Hdac1	H3K27me3	H3K4me2	H3		
Klf4	Fra1	Brg1	H3K9me3	H3K4me1			
cMyc	Cebpa		H3K9ac	H3K36me3			
Nanog	Cebpb		H3K79me2				

Chronis, Fiziev et al. (2017)

- Each group will work with **2 different reprogramming stages**

- Group 1: pre-iPSCs vs ESCs (3 vs 4)
- Group 2: induced MEFs(48h) with OSKM vs ESCs (2 vs 4)
- Group 3: MEFs vs ESCs (1 vs 4)
- Group 4: induced MEFs(48h) with OSKM and Esrrb vs ESCs (2 vs 4)

Our project:
Analysis overview

Research questions

- What are the differences in transcription factor binding between somatic cells, reprogrammed intermediate states and ESCs?
- Does the promoter/enhancer activity change during reprogramming from somatic cells to ESCs?
- Which genes change their expression during reprogramming from somatic cells to ESCs?
- Can gene expression be predicted from epigenetic marks?

1. Univariate analysis

- binding events of pluripotent TFs (Oct4, Sox2, Klf4, cMyc) across the genome in both stages
- binding events of somatic TFs (Cebpa, Cebpb, Fra1, Runx1, Oct4, Sox2, Klf4, cMyc) across the genome in both stages

- histone modification patterns across the genome in both stages (H3K4me3, H3K4me1, H3K27ac)

- open chromatin across the genome in both stages (ATAC-seq)

- gene expression analysis (RNA-seq)

2. Genomic features and overlap analysis

- genomic location of the TF binding events (promoters, enhancers, gene body, introns) in both stages
- correlation between TF binding events and open chromatin
- histone modification patterns on promoters and on enhancers in both stages
- correlation between histone modification patterns and expression of the neighboring genes
- correlation between TF binding patterns and expression of the neighboring genes

3. Differential analysis

- differential binding events of pluripotent TFs between the two stages
- differential patterns of histone modifications between the two stages (which HMs change the most, which do not?)
- differential gene expression analysis between the two stages
- differential HM activity in promoters and enhancers between the two stages

4. Prediction model

- building a prediction model (linear regression, logistic regression, random forest) to predict gene expression levels from histone modification patterns on promoters
- improvement of the model with TF binding?
- can we predict expression in a different stage based on the model trained in a different developmental stage (e.g. trained on ESCs and predict for MEF)?

5. Sequence analysis

- motif analysis of the genomic sequences of the TF peaks
- do you find different motifs in TF peaks between the two stages?
- do you find consecutive binding motifs of the underlying TF in the TF peaks?

Our project:
First tasks

Literature survey

Prepare answers to these questions until the next meeting:

1. What is the function of Oct4, Sox2, Klf4, cMyc? Where do they usually bind in the genome, in which cell types? What are their known motifs?
2. What is the function of Runx1, Fra1, Cebpa, Cebpb? Where do they usually bind in the genome, in which cell types? What are their known motifs?
3. Summarize the information about the histone marks studied by Chronis et al. Which chromatin states do they usually mark (active/ inactive promoters, enhancers, gene body, repressed regions...)? Which of the histone marks correlate with open chromatin?
4. What is ChromHMM? Read the basic information from Ernst&Kellis (2012, 2017)

Literature survey

Prepare answers to these questions until the next meeting:

5. Read the paper from Chronis et al. Try to understand Figure 1 and Figure 2. Focus then on the Methods section.
6. Find the data source (note the relevant file numbers) and download the raw read files corresponding to the reprogramming stage of your group (do NOT download all datasets from the publication!!! >1 TB of data!!!) , only one group downloads the ESC data
   save all the data in:
   `/project/functional-genomics/2019/data/sra/`
5. Convert the data into the fastq format (SRA Toolkit)
6. Read the manuals of FASTQC and deepTools and run the quality control of your files. Which functions of deepTools do you need for it? Is the quality of the data sufficient? Why?
7. Read the paper from Karlić et al. What is the basic idea in the paper?

Literature survey

General points

- cite the publications and other sources that address the points
- use pubmed and/or scholar.google.com and other resources
- prepare slides with all answers and citation

References

- Chronis, Fiziev et al., *Cell* (2017)
- Ernst and Kellis, *Nature Methods* (2012)
- Ernst and Kellis, *Nature Protocols* (2017)
- Barth and Imhof, *Trends in Biochemical Sciences* (2010)
- Karlić et al., *PNAS* (2010)
- Lawrence at al., *Trends in Genetics* (2016)
- Takahashi and Yamanaka, *Cell* (2006)
- Bailey et al, PLOS Comp Biol (2013)

Univariate analysis

Next steps to do until March 21

1. Mapping the raw reads of ChIP-seq and ATAC-seq experiments to the reference genome mm9. Use **bowtie2** and try to set the parameters to obtain the same alignments as *Chronis et al.* Save the mapping statistics for each file. How many percent of the reads could be mapped with sufficient quality? Do you think it is a large/small proportion of reads?
2. You don't need to download and index the reference genome, it is already in: `/project/functional-genomics/2019/data/genome`
3. Create bigWig files from the bam files using bamCoverage in deeptools with bin size 25bp, use the normalization using RPKMs, ignore chr X for normalization and extend reads by 200bp

Univariate analysis

Next steps to do until March 21

4. Peak calling with MACS2 for the ChIP-seq and ATAC-seq data. Use the same parameters as Chronis et al.
5. Report the total number of significant peaks for all experiments (in both stages). How do these numbers change over different targets (TFs, HMs)? How do these number change in both stages?
6. What is the average width of the peaks in all experiments?
7. Use IGV to visualize the data. Load the bigWig files and the peak bed files into IGV. Color the tracks by the different features.
8. Find in the literature some known target genes of Oct4, Sox2, Klf4 and cMyc. Look up this genes in the IGV, including its promoters. Save the images. What do you observe?

Do all big calculations on servers (`ssh sympathyforthedevil, nyancat, holidayincambodia, bohemianrhapsody`). Check the available cores with `htop`. Use `screen` tools to be able to run commands, close the connection and to access the session later.

Mapping with bowtie2

- different parameter settings in bowtie2 than in bowtie(1)
- use the mapping quality score to filter reads that with low mapping quality
- remove duplicates (reads mapping to the same location)
- save the resulting file as a bam file
- see the manual recommendations: ***Mapping quality: higher=more unique***
- use samtools for the filtering
- helpful information about SAM flags:

https://broadinstitute.github.io/picard/explain-flags.html

Peak calling with MACS2

- find the peaks, i.e. the regions with a high density of reads, where the studied TF was bound
- **manual (read it!)**

1) call the peaks using the experiment (treatment) data vs. control

use the control experiment: Inputnative_MNase experiment

2) set the parameters e.g. fragment length, treatment of duplication reads
3) analyse the MACS results (BED file with peaks/summits)

RNA-seq data

- mapping with another tool, **STAR** recommended
- on top of the reference genome, an annotation file (gtf format) with gene locations is necessary
- we are interested in the mapping to the transcriptome only
- for differential analysis, we only need estimated read counts per gene
- **salmon** tool suitable for this
- additionally to the sequencing file, salmon needs a transcriptome reference file

https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

https://salmon.readthedocs.io/en/latest/