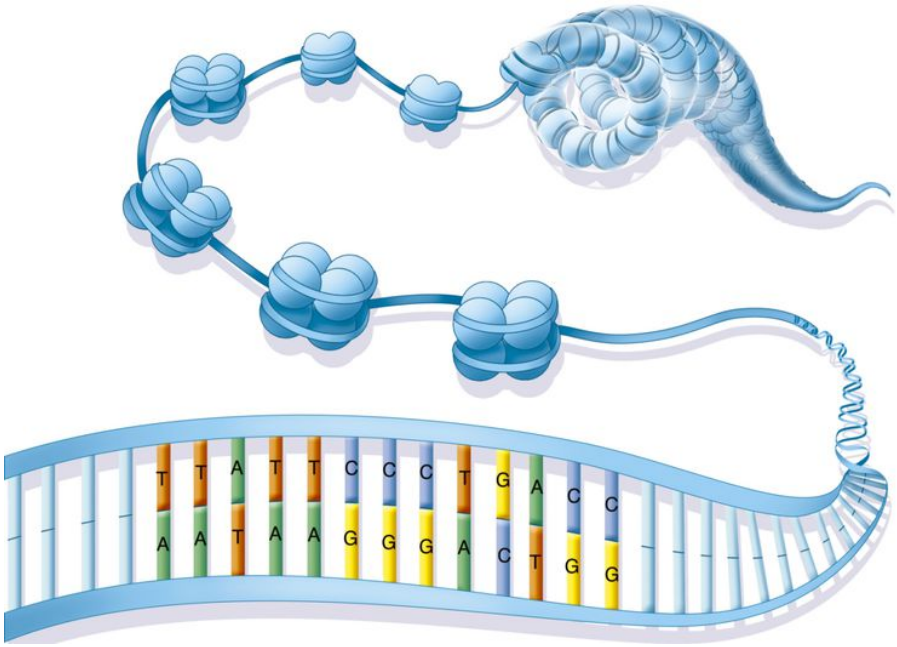
Exploratory analysis of ATAC-seq data from dopaminergic neurons

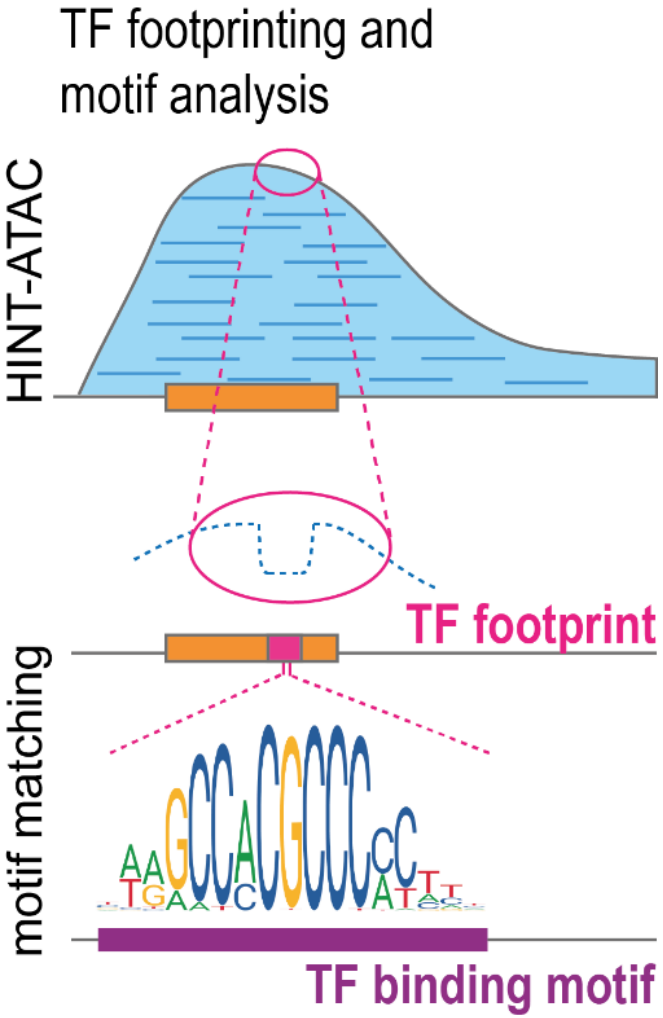
HPC School User Story

Nikola de Lange

Research objective

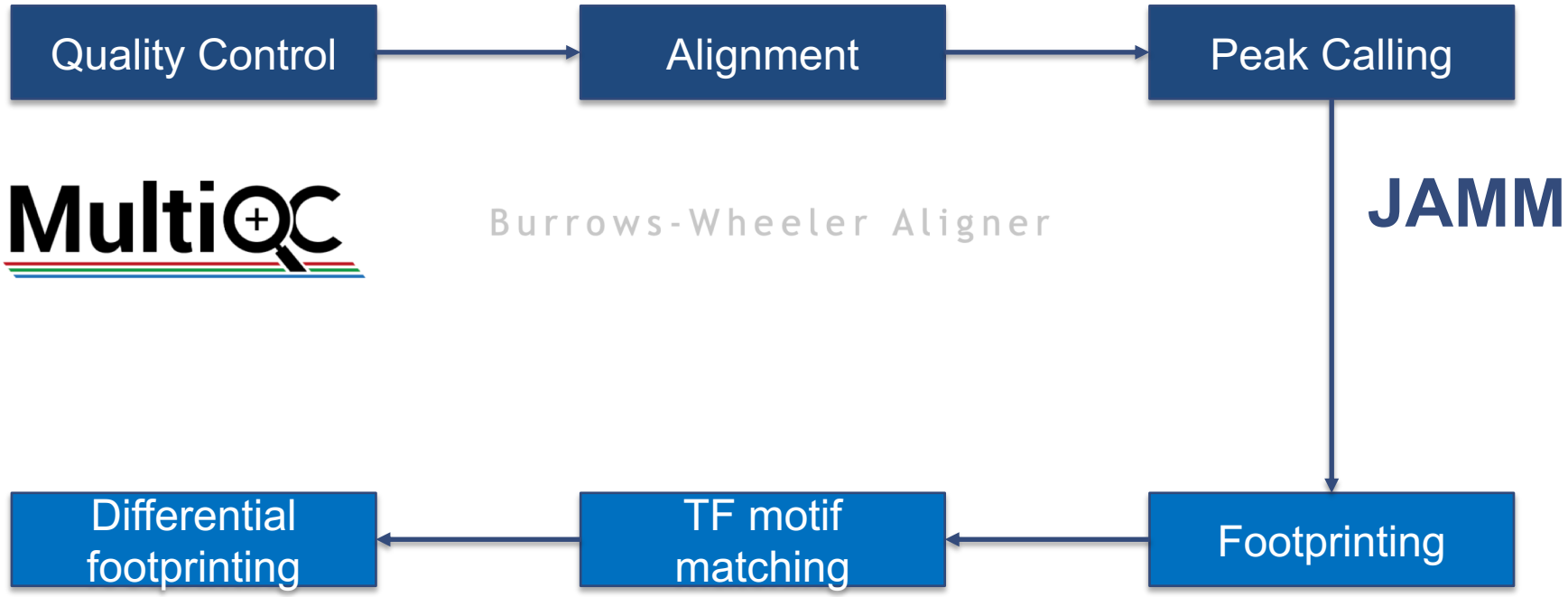


https://www.thoughtco.com/chromatin-373461



- ATAC-seq data from dopaminergic neurons (DA)
- Identification of transcription factor (TF) binding sites (footprints)
- Comparison of activity of TFs between different maturing states of DA

HPC workflow



Regulatory Genomics Toolbox

Ewels et al. Bioinformatics 2016, Li et al. Bioinformatics 2009, Ibrahim et al. Bioinformatics 2015, Li et al. Genome Biology 2019

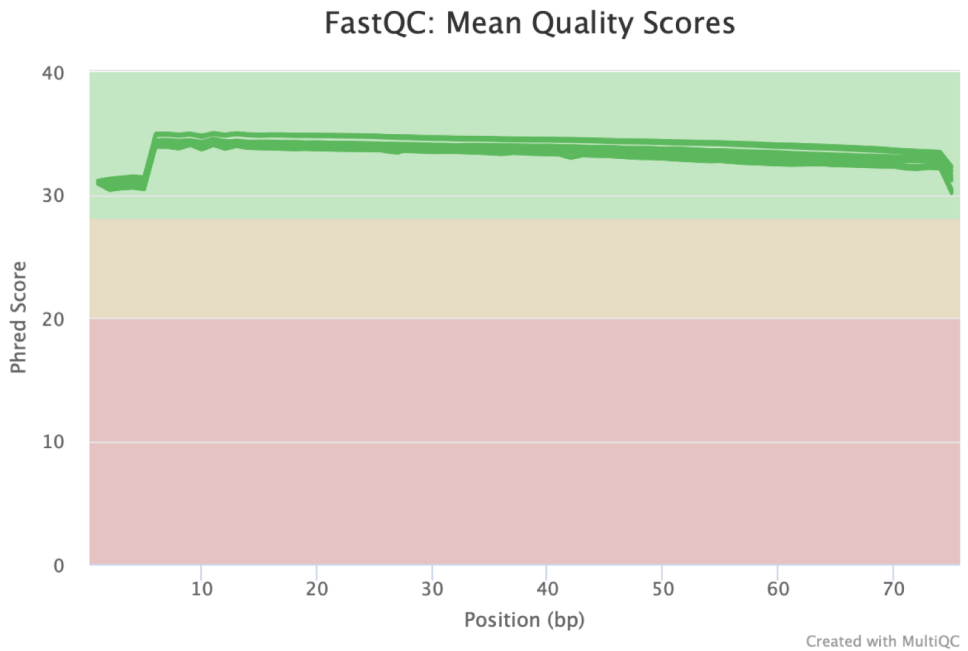
Data management and quality control

Data management

- Write protected raw data in project directory (/work/projects/…)
- Backup

Quality Control

- Typically run as interactive job
- parallel
- bio/FastQC
- Summary
  - MultiQC



Alignment

- Reference genome indexing and read alignment
  - Batch job
  - bio/BWA
  - bio/SAMtools
  - ~37GB memory
  - ~5h run time

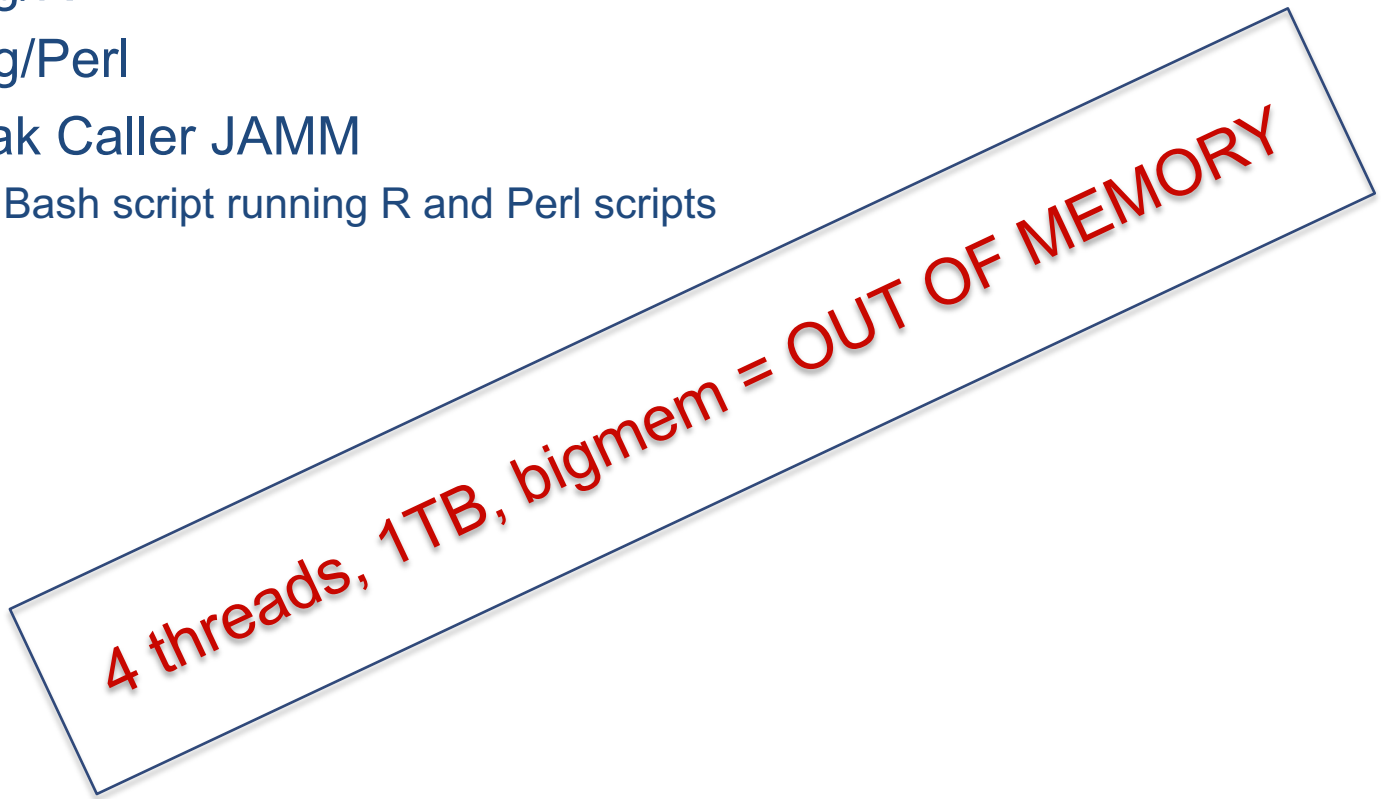
```
#!/usr/bin/bash -l
#SBATCH -J bwa
#SBATCH --mail-type=end,fail
#SBATCH --mail-user=nikola.delange@uni.lu
#SBATCH -N 1
#SBATCH -c 12
#SBATCH --ntasks-per-node=1
# #SBATCH --mem=48GB
#SBATCH -p batch
```

Peak Calling

- Batch job
- bio/BEDTools
- lang/R
- lang/Perl
- Peak Caller JAMM
  - Bash script running R and Perl scripts

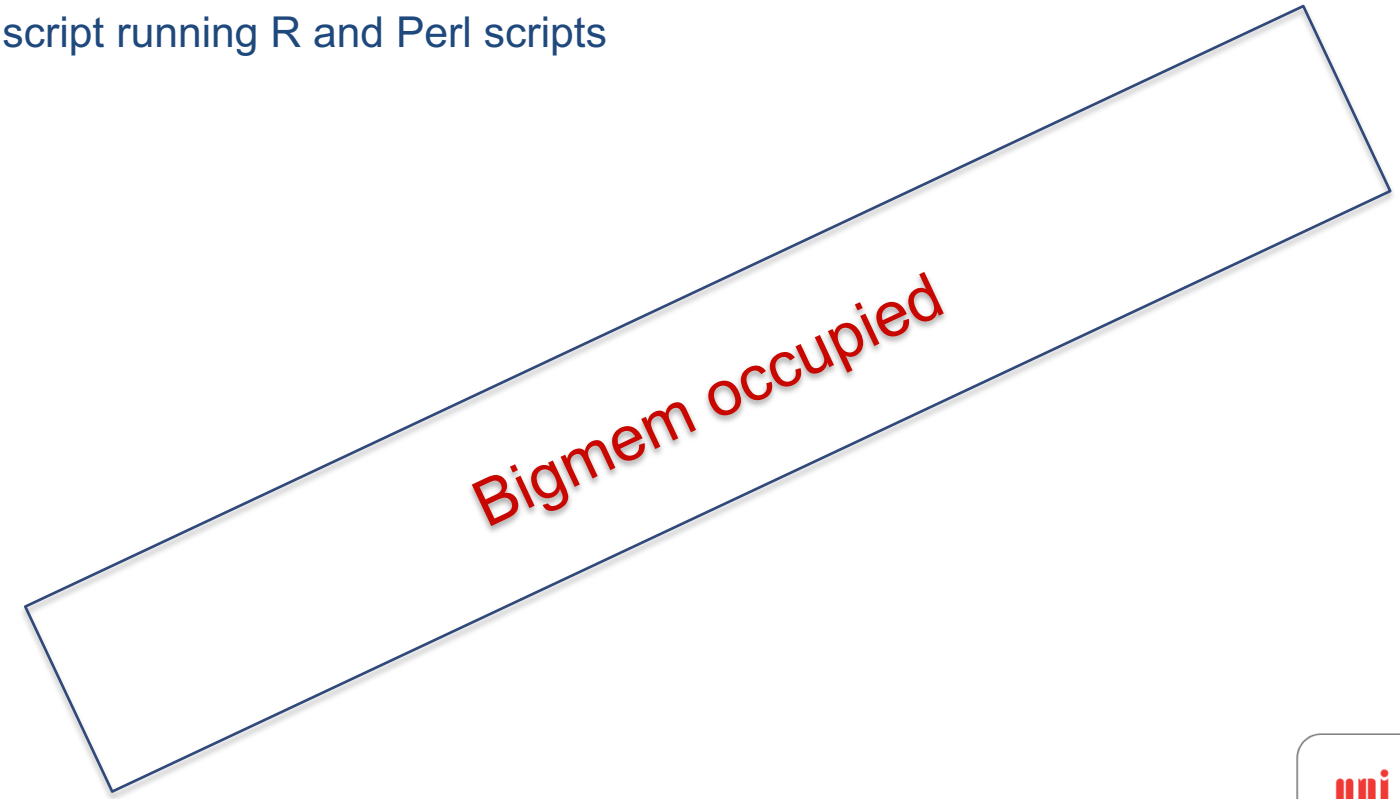
Peak Calling

- Batch job
- bio/BEDTools
- lang/R
- lang/Perl
- Peak Caller JAMM
  - Bash script running R and Perl scripts



Peak Calling

- Batch job
- bio/BEDTools
- lang/R
- lang/Perl
- Peak Caller JAMM
  - Bash script running R and Perl scripts



Peak Calling

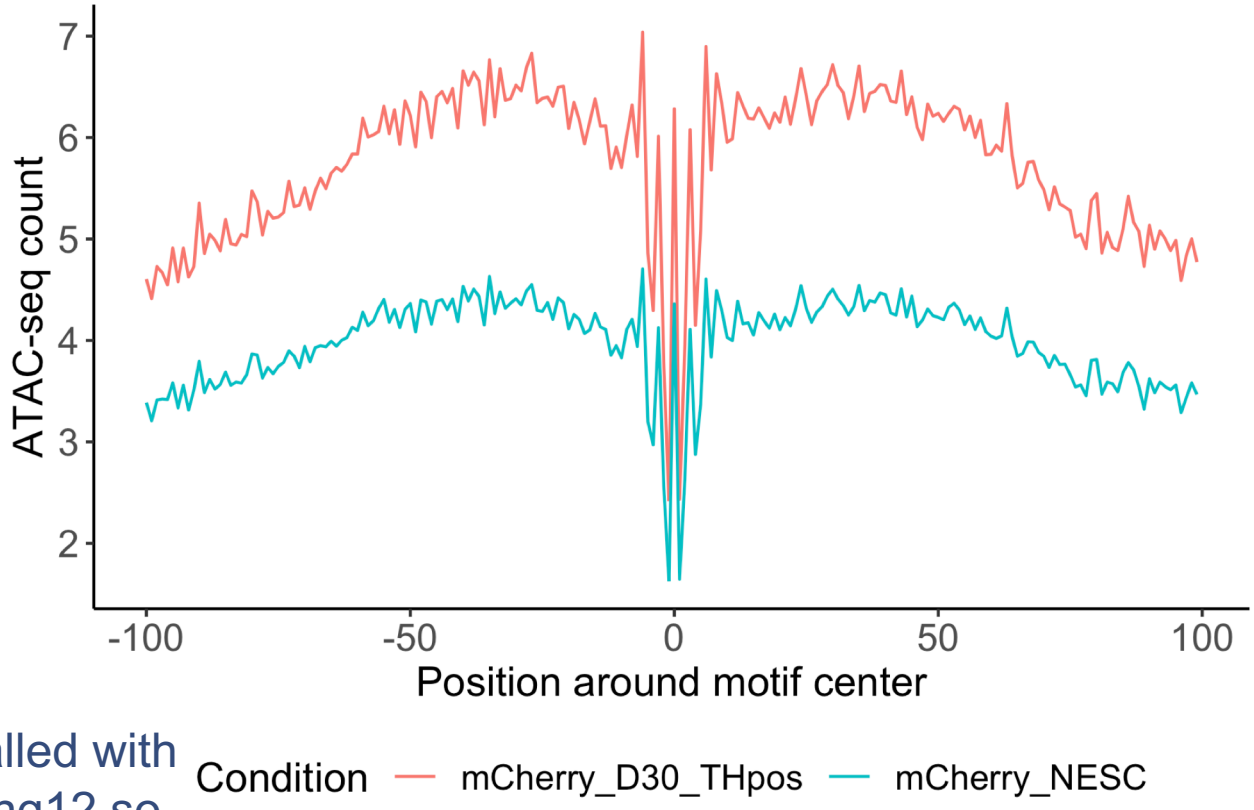
- Batch job
- bio/BEDTools
- lang/R
- lang/Perl
- Peak Caller JAMM
  - Bash script running R and Perl scripts

Number of threads	Memory	Run time
1	13 GB	18:34:50
2	385 GB	5:22:33

```
#!/usr/bin/bash -l
#SBATCH -J jamm
#SBATCH --mail-type=end,fail
#SBATCH --mail-user=nikola.delange@uni.lu
#SBATCH -N 1
#SBATCH -c 28
#SBATCH --ntasks-per-node=1
#SBATCH --mem=112GB
#SBATCH --time=0-24:00:00
#SBATCH -p batch
#SBATCH --qos=qos-batch
```

Regulatory Genomics Toolbox

- Batch job
- RGT
  - Python tool
- lang/Python
- lib/libpng
- Footprinting:
  - ~9h
  - ~2.5 GB memory
- Diff. footprinting:
  - ~3h
  - ~30 GB memory



Issue:

- UCSC tools installed with RGT require libpng12.so
- libpng16.so installed on cluster
- Manual replacement of required tools

Experience and best practices

Experience

- Different steps with different memory and CPU requirements
  - Several bash scripts
  - Tedious to coordinate
- Time limited training at a different lab
  - Needs planning in advance

Best practices

- Snakemake workflow engine
- Prevent idle resources
- Check use of resources (Ganglia)
- Set walltime

Acknowledgements

Roland Krause
Reinhard Schneider
Marcel Schulz
Patrick May
Lasse Sinkkonen
Jochen Ohnmacht
Borja Gomez Ramos
Sarah Peters

Thank you!

The BioCore
PARK-QC DTU

Luxembourg National Research Fund

This project has received funding from the Luxembourg National Research Fund (FNR) within the PARK-QC DTU (**PRIDE17/12244779/PARK-QC**).