ChIP-Seq survival skills

Gene Expression Nanocourse

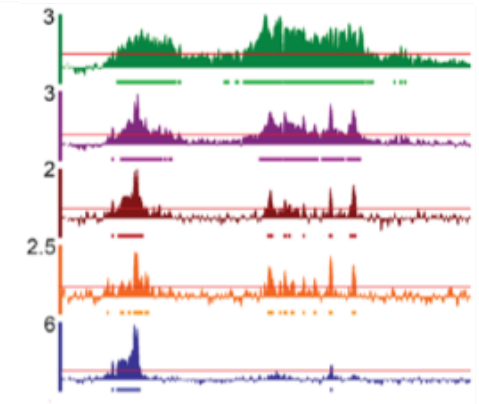
Gary Hon

Picture this…

- You:
  -  have this awesome hypothesis
  -  spend a few <u>years/weeks</u> to collect specimens
  -  spend a few <u>days</u> to make some ChIP-Seq libraries
  -  send them off for sequencing
  -  get the data back a <u>month</u> later
  -  have a bioinformatician … on vacation
  - need results, <u>now</u>

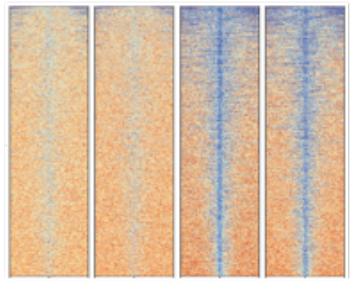
By the end of this class, you will be able to…

- Process raw ChIP-Seq data
- Visualize data
- Find peaks
- Find motifs
- Make heatmaps
- Perform differential analysis



browser

motifs



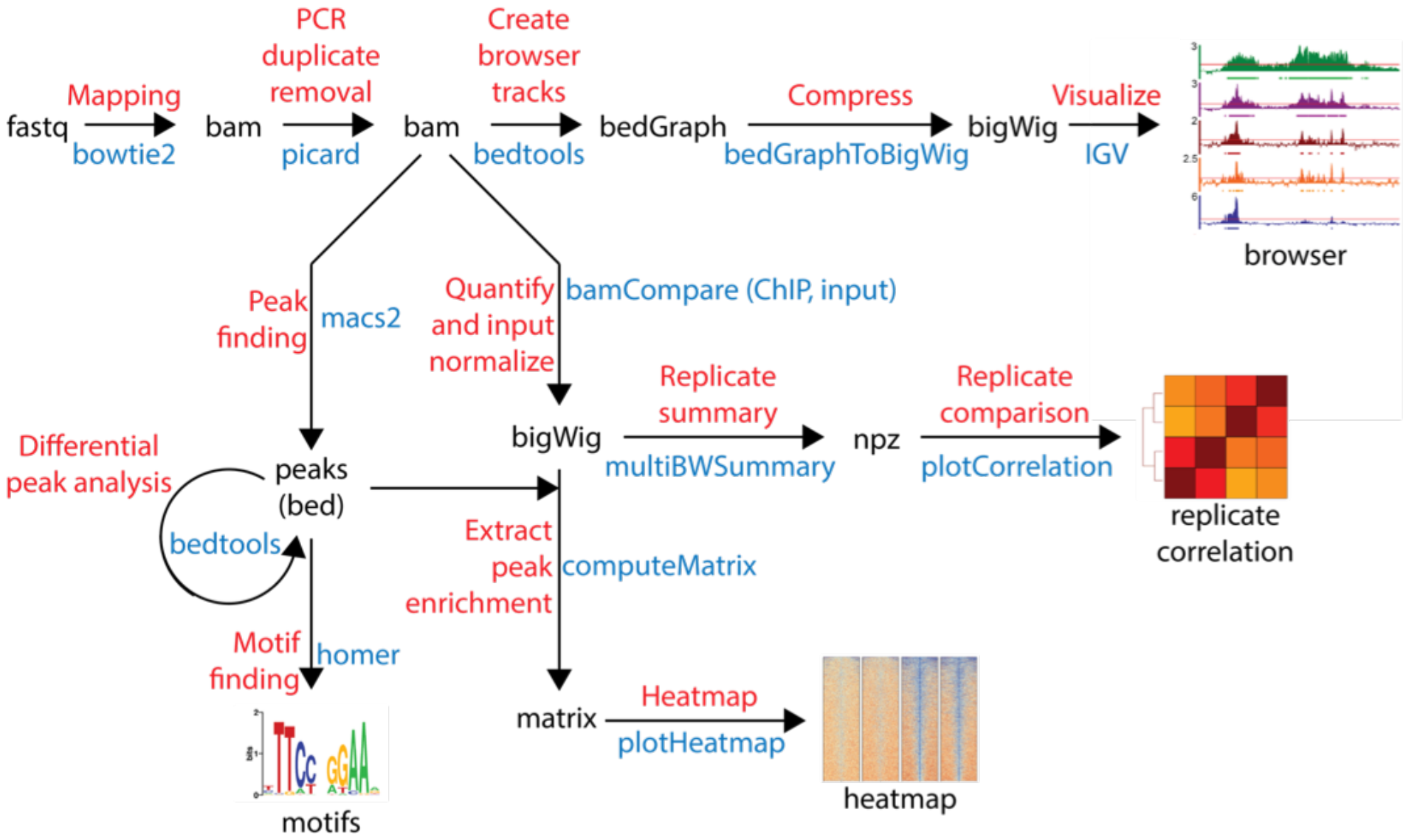
heatmap

Today's overview

- **AM (9-12pm)**
  1. ChIP-Seq experiment
  2. Data (+ QC)
  3. Mapping (+QC)
  4. Duplicate removal (+QC)
  5. Visualization (+QC)
  6. Peak finding

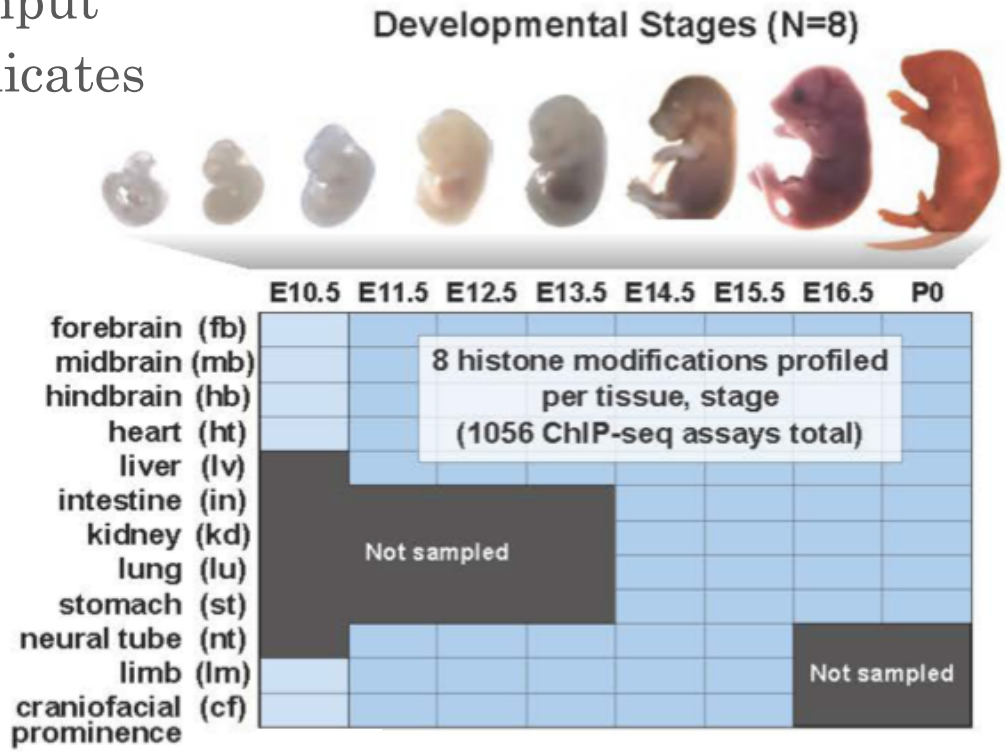
- **PM (1-5pm)**
  1. Quantification
  2. Heatmaps
  3. Meta plots
  4. Differential analysis
  5. Motif finding

What you'll accomplish today



Case study: Heart development

- ENCODE Project
- https://www.biorxiv.org/content/early/2017/07/21/166652
- Heart: e11.5 vs P0
  - H3K27ac and input
  - 2 biological replicates



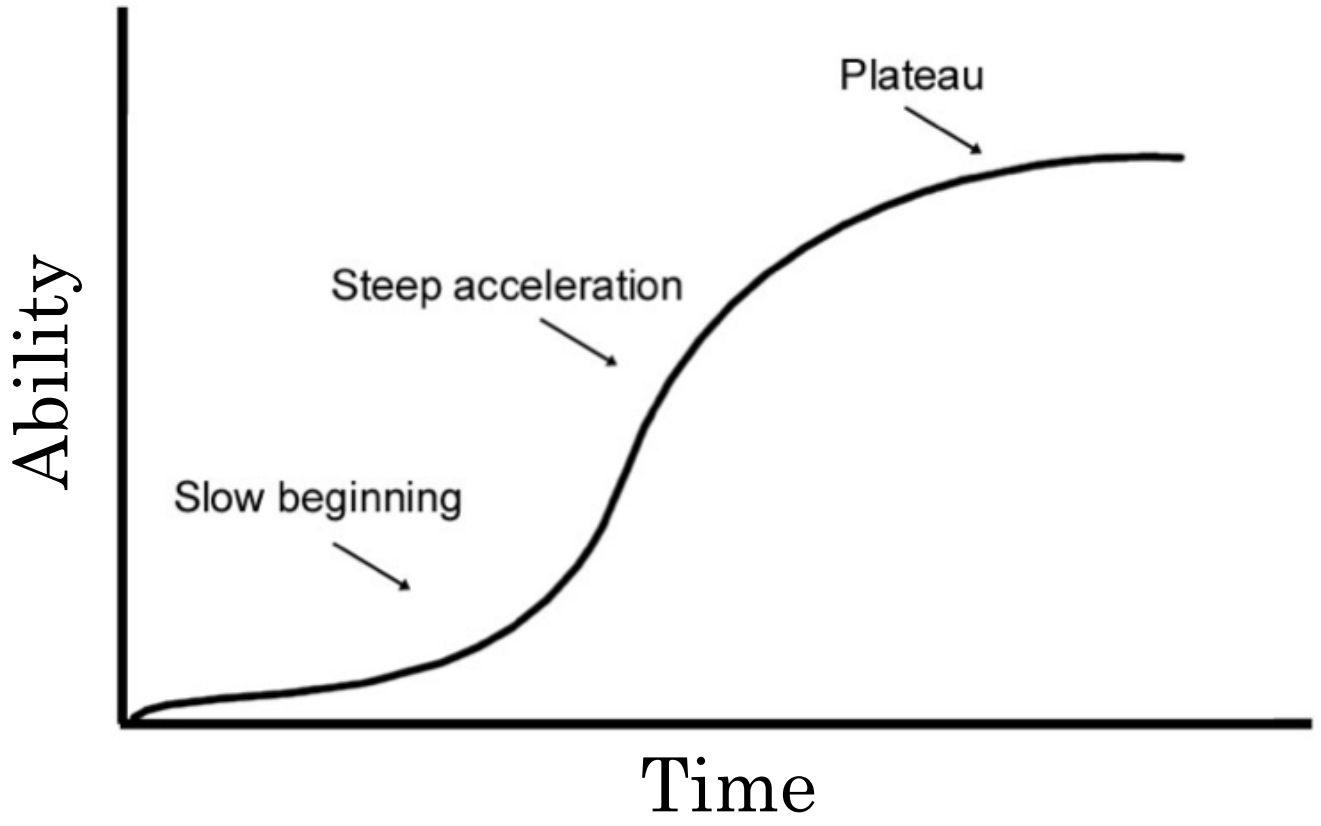
Case study: Questions we'll answer

1. Did the experiment work (technically)?
   - Does it map?
   - Is there ChIP enrichment?
   - Do the biological replicates agree?

2. How do H3K27ac peaks change during heart development?
   - Are more peaks gained or lost?

3. Which transcription factors drive development?

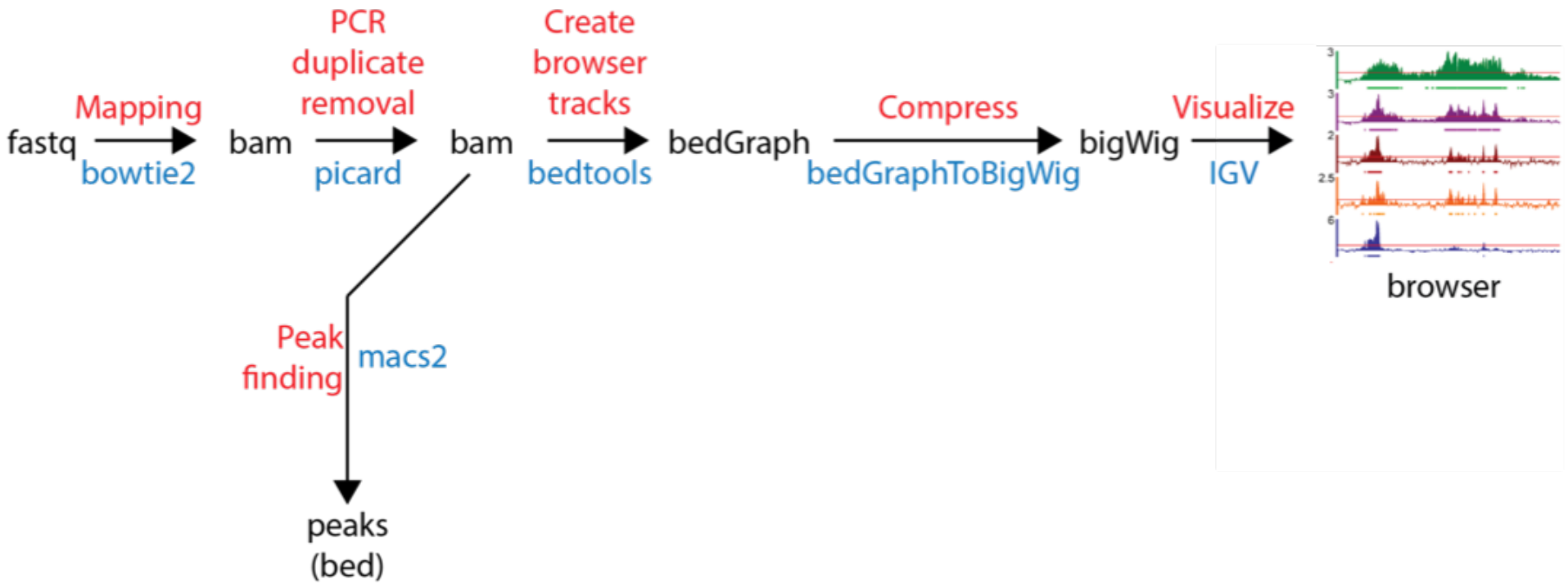
Bioinformatics: steep learning curve



Tips

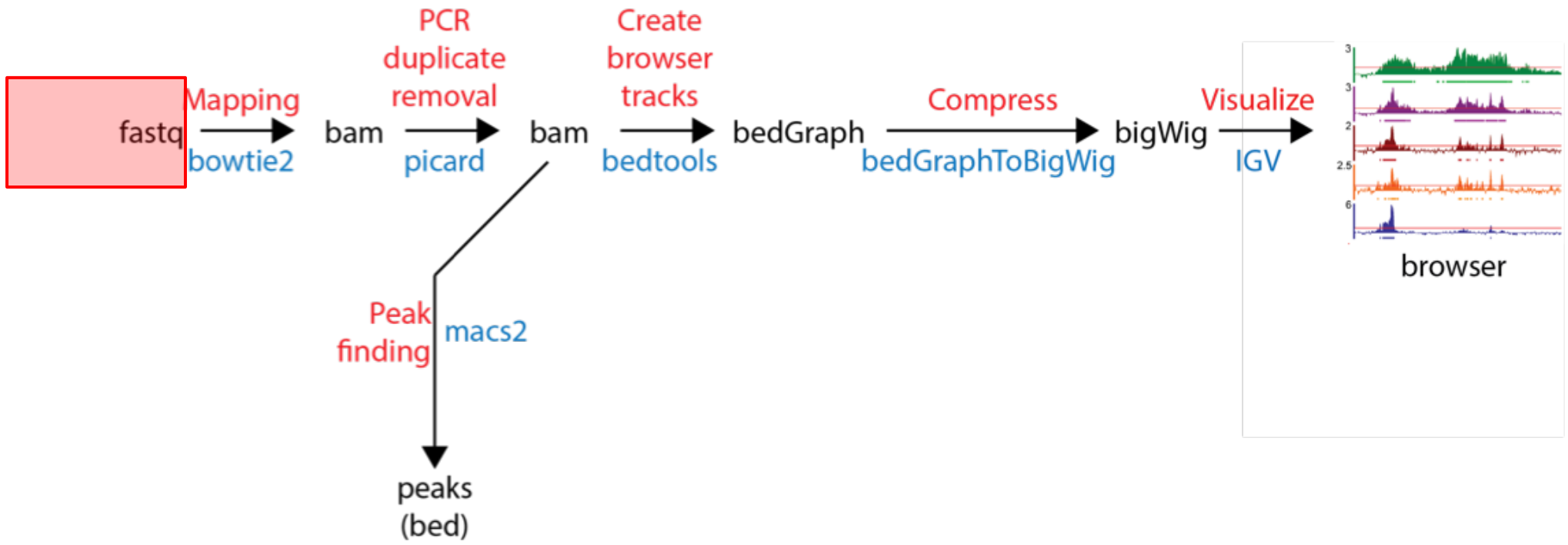
- It will be frustrating – keep trying!
- Ask questions!
- Work with a partner.
- Does it work? Try it!
- Tell us how we can improve.

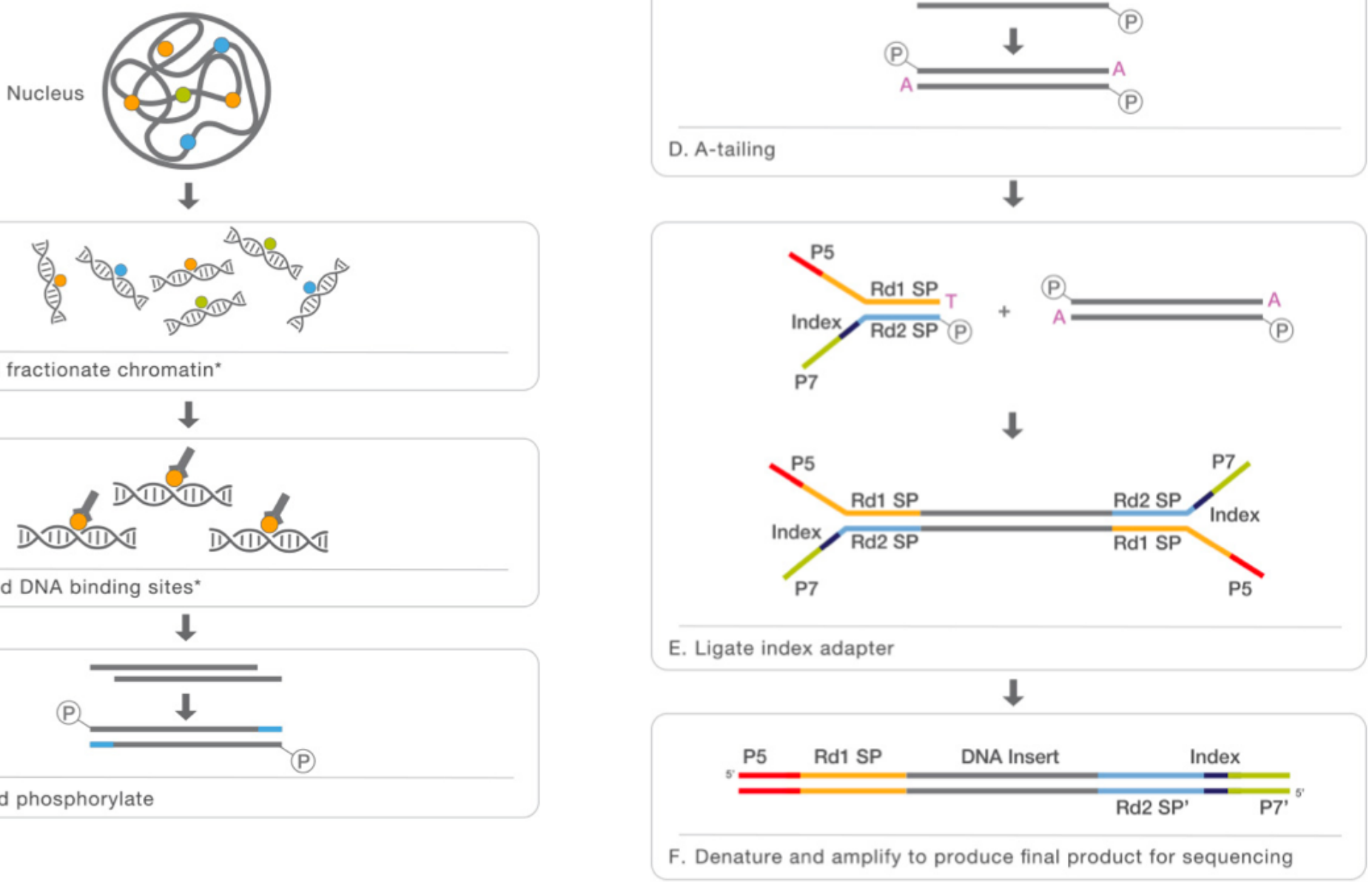
Overview: Morning



ChIP-Seq experiment

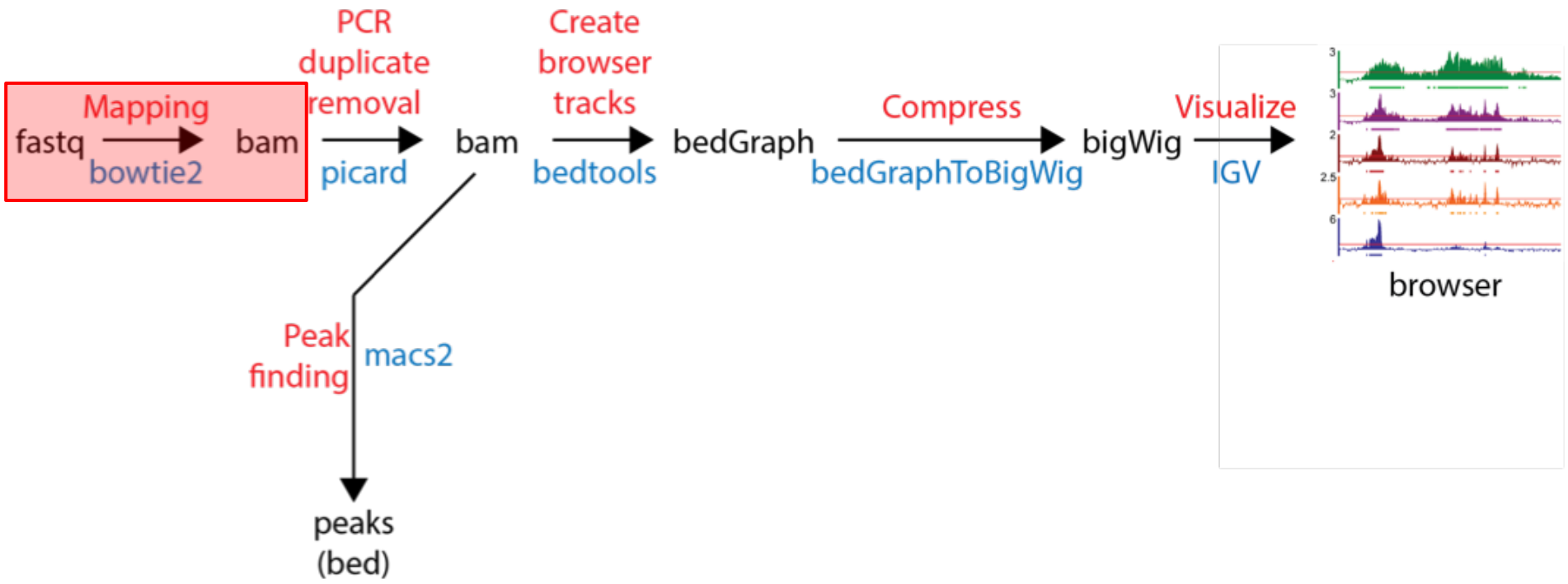
DNA ➔ fastq




Nucleus
A. Crosslink and fractionate chromatin*
B. ChIP: Enriched DNA binding sites*
C. End repair and phosphorylate
D. A-tailing
P5
Rd1 SP
T
Index
Rd2 SP
P7
A
E. Ligate index adapter
DNA Insert
Rd2 SP'
P7'
F. Denature and amplify to produce final product for sequencing

Mapping

Fastq ➔ BAM



Setup

```
mkdir part1

cd part1

cp ../../shared/class2/part1/* .
```

The command: bowtie2

```
module load bowtie2
module load samtools

# make a directory to hold your raw alignments
mkdir bam_files

# mapping
bowtie2 -p 32\
        -x [INDEX]\
        -U [FASTQ FILE] \
  | samtools view -q10 -bS -o - -\
  | samtools sort -o [OUTPUT FILE] -
```

Load the programs bowtie2 and samtools

Make a new folder

Call bowtie with 32 processors

Convert output to BAM

The command: bowtie2

```
module load bowtie2
module load samtools

# make a directory to hold your raw alignments
mkdir bam_files

# mapping
bowtie2 -p 32\
        -x /project/apps_database/iGenomes/Mus_musculus/UCSC/mm10/Sequence/Bowtie2Index/genome\
        -U /archive/nanocourse/June2018/shared/fastq.chr19/ChIP-Seq/heart.e11.5.rep1.ChIP-Seq.H3K27ac.fastq.gz\
  | samtools view -q10 -bS -o - -\
  | samtools sort -o bam_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.bam -
```

Open script1

- Applications -> Accessories -> gedit Text Editor
- File -> Open
- File System -> archive -> nanocourse -> June2018 -> trainXY -> part1 -> script1_mapping.sh

You try mapping (script1)

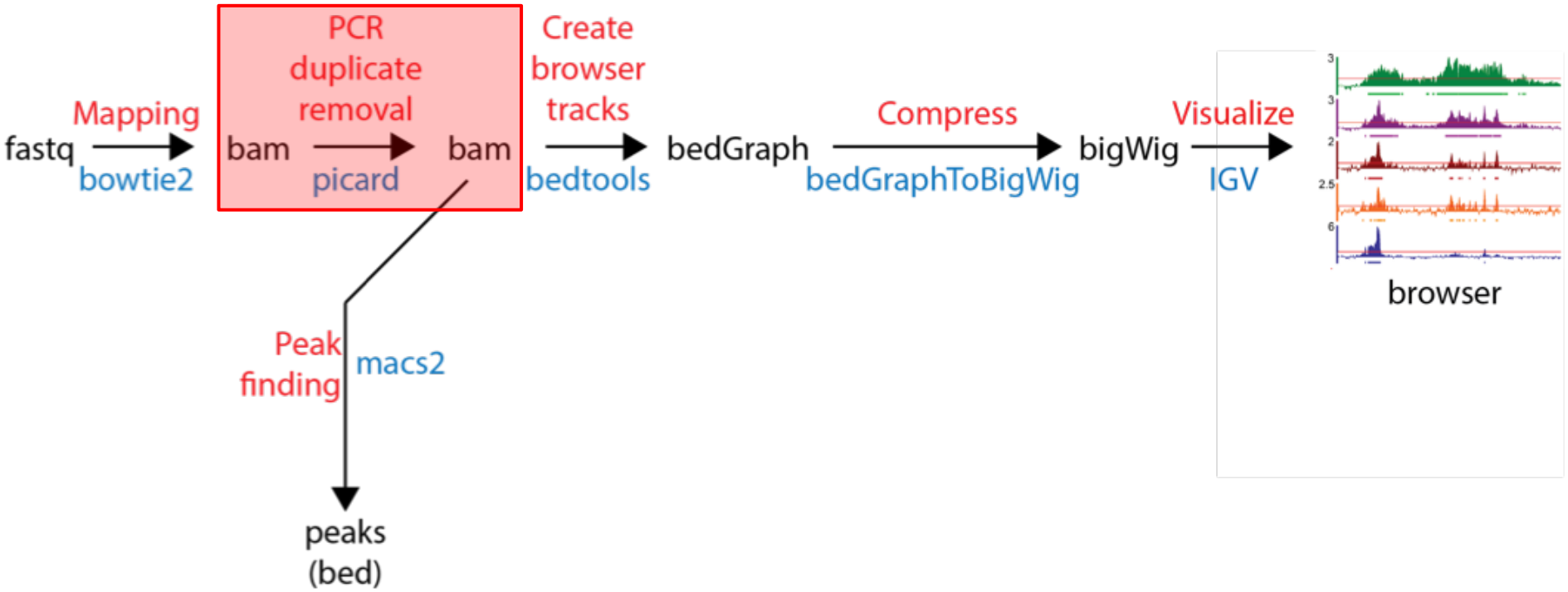
- Example: H3K27ac
- You do: input

Examining the output

- How many reads map to chr19?

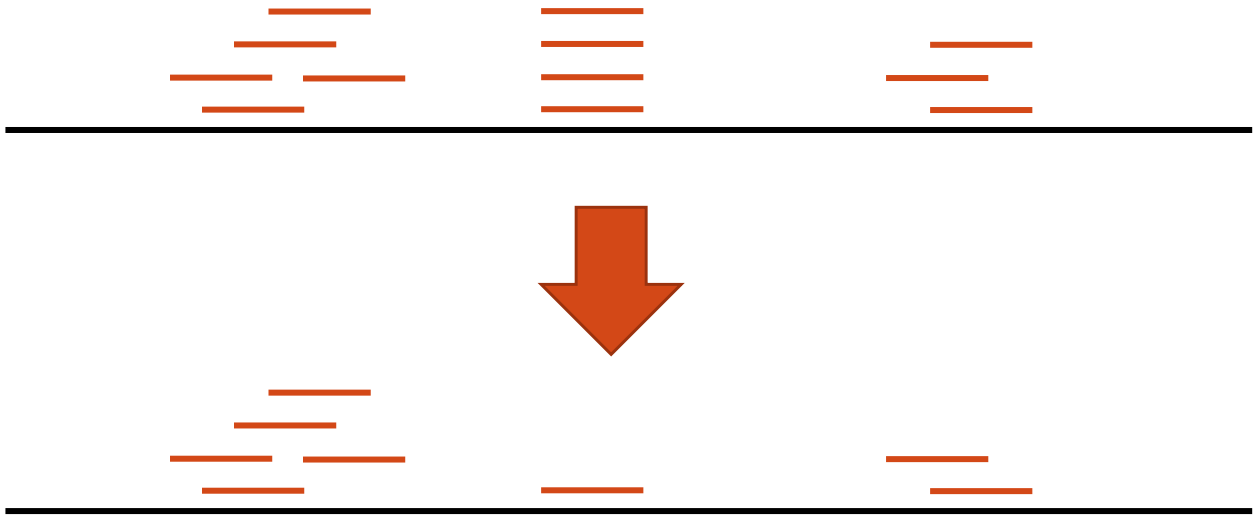
PCR duplicate removal

BAM ➔ nodup BAM



What is PCR duplicate removal?

- ChIP-Seq uses PCR
- Some fragments amplify better than others
- Remove bias: if 2+ reads map to same place, then keep 1



The command: picard

```
module load java
module load picard/2.10.3
module load samtools

# remove PCR duplicates
java -jar $PICARD_DIR/picard.jar MarkDuplicates \
      INPUT=[INPUT]\
      OUTPUT=[OUTPUT]\
      METRICS_FILE=metrics.txt\
      REMOVE_DUPLICATES=true\
      ASSUME_SORTED=true\
      TMP_DIR=temp

# index the final BAM file
samtools index [OUTPUT]
```

Load the programs java, picard and samtools

Call picard to remove duplicates.

Index the BAM file

The command: picard

```
module load java
module load picard/2.10.3
module load samtools

# remove PCR duplicates
java -jar $PICARD_DIR/picard.jar MarkDuplicates \
      INPUT=bam_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.bam\
      OUTPUT=bam_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.nodup.bam\
      METRICS_FILE=metrics.txt\
      REMOVE_DUPLICATES=true\
      ASSUME_SORTED=true\
      TMP_DIR=temp

# index the final BAM file
samtools index bam_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.nodup.bam
```

You try removing duplicates (script2)

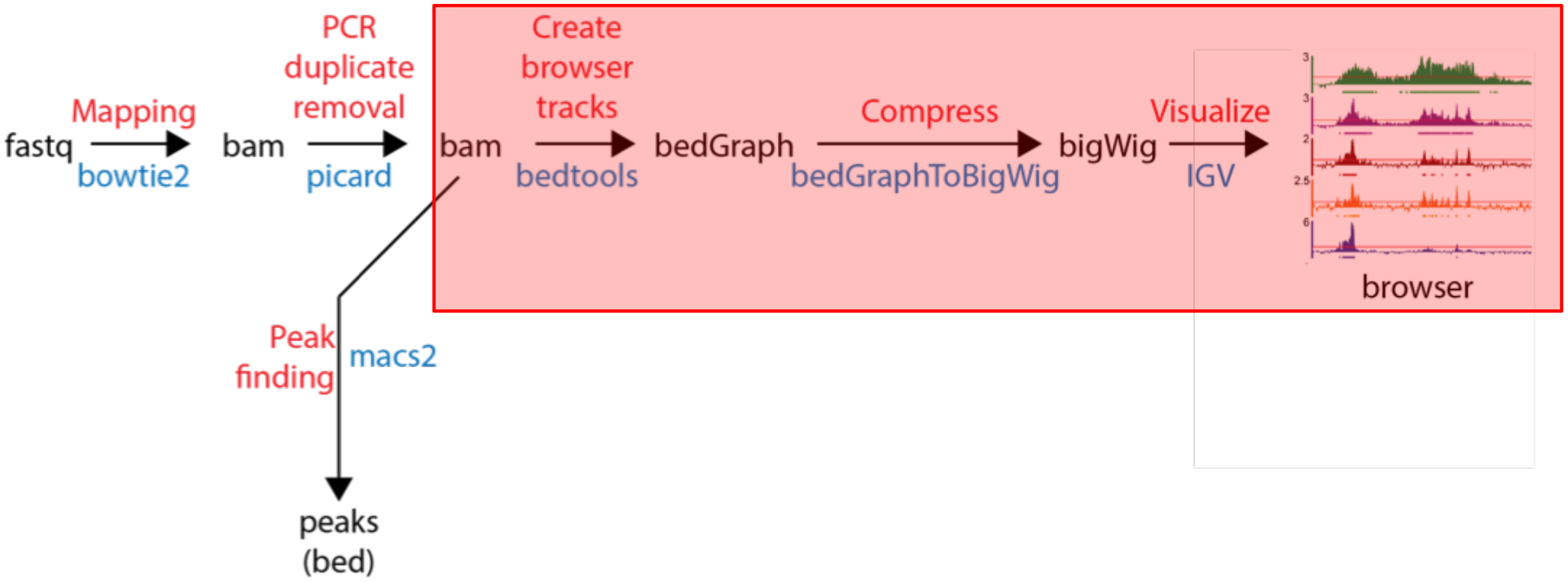
- Example: H3K27ac
- You do: input

Examining the output

- How many reads map to chr19?
- How many reads map to region:
  - chr19: 10,000,000 – 11,000,000?
- Compare with before PCR duplicate removal

Visualization

Nodup BAM ➔ bedGraph ➔ bigWig



The command: visualization

```
module load UCSC_userApps/v317
module load bedtools

# make a directory to hold your genome browser files
mkdir browser_files

# get genome browser track
bedtools genomecov -ibam\
                   [INPUT BAM FILE]\
                   -bga\
  > [OUTPUT BEDGRAPH FILE]

# compress genome browser track
bedSort [OUTPUT BEDGRAPH FILE] \
        [OUTPUT BEDGRAPH FILE]

fetchChromSizes mm10 > mm10.chrom.sizes
bedGraphToBigWig [OUTPUT BEDGRAPH FILE] \
                 mm10.chrom.sizes\
                 [OUTPUT BIGWIG FILE]
```

Load the programs UCSC tools and bedtools

Make a BedGraph file

Sort BedGraph file before compress

Get mouse chr size

Compress BedGraph file

The command: visualization

```
module load UCSC_userApps/v317
module load bedtools

# make a directory to hold your genome browser files
mkdir browser_files

# get genome browser track
bedtools genomecov -ibam\
                   bam_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.nodup.bam\
                   -bga\
  > browser_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.bedGraph

# compress genome browser track
bedSort browser_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.bedGraph\
        browser_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.bedGraph

fetchChromSizes mm10 > mm10.chrom.sizes
bedGraphToBigWig browser_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.bedGraph\
                 mm10.chrom.sizes\
                 browser_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.bw
```

You try creating browser files (script3)

- Example: H3K27ac
- You do: input

Examining the output

- What does a bedGraph file look like?
- Compare the size of a bedGraph file and a bigwig file

Let's see the data

- Use Integrated Genome Viewer

```
module load IGV

igv.sh

   (1. load mm10)

   (2. load BigWig files:
   /archive/nanocourse/June2018/trainXY/part1/browser_fil
   es)

   (3. go to chr19)
```

Case study: Questions we'll answer

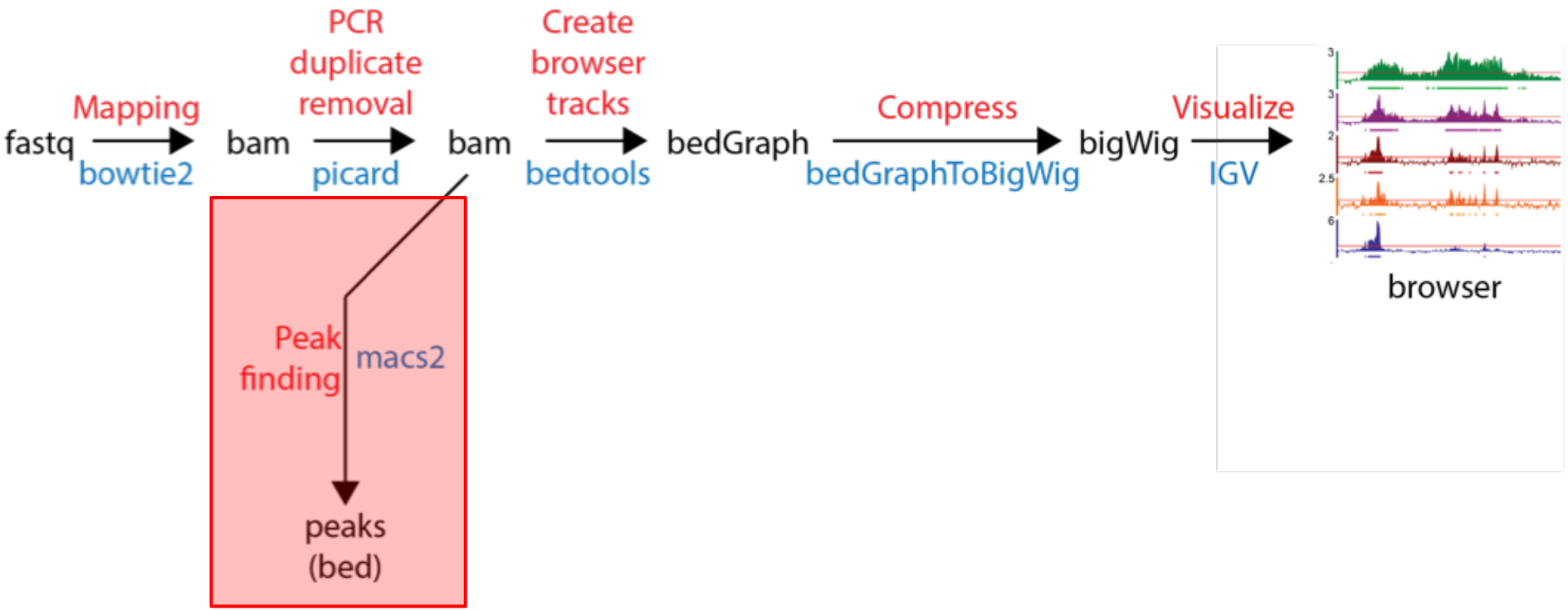
1. Did the experiment work (technically)?
   - Does it map?
   - Is there ChIP enrichment?
   - Do the biological replicates agree?

2. How do H3K27ac peaks change during heart development?
   - Are more peaks gained or lost?

3. Which transcription factors drive development?

Peak finding

Nodup BAM ➔ peaks (bed)



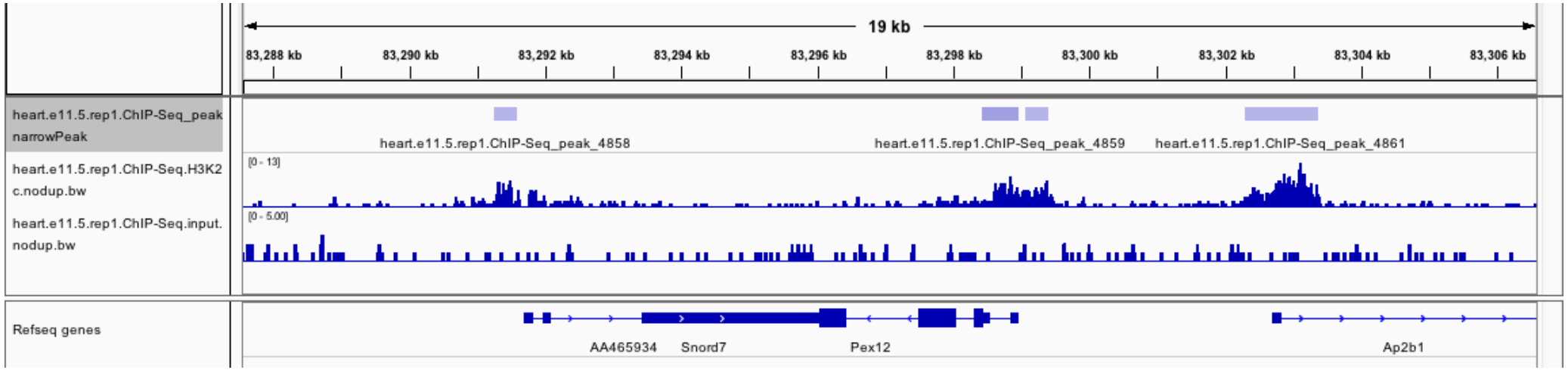
What's a ChIP-Seq peak?

- Enrichment of ChIP signal over input signal

peak

ChIP

input



The command: peak finding

```
module load macs

# make a new directory for peaks
mkdir peaks

# call peaks
macs2 callpeak -t [TREATMENT BAM FILE]\
               -c [CONTROL BAM FILE]\
               -f BAM\
               -g mm\
               --outdir peaks\
               -n [NAME OF RUN]
```

Load the program macs.

Call ChIP-Seq peaks.
mm = mouse

The command: peak finding

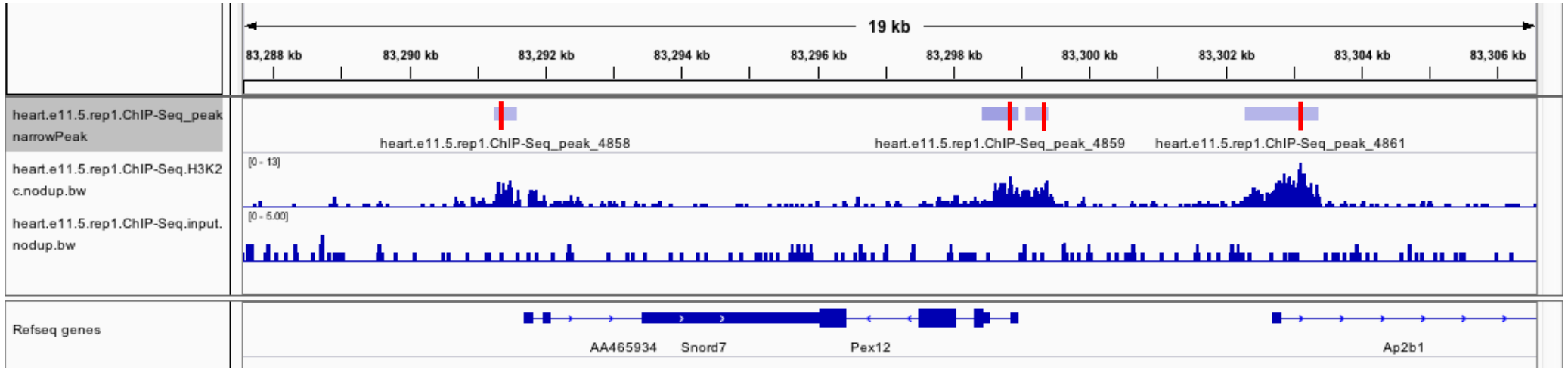
```
module load macs

# make a new directory for peaks
mkdir peaks

# call peaks
macs2 callpeak -t bam_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.nodup.bam\
               -c bam_files/heart.e11.5.rep1.ChIP-Seq.input.nodup.bam\
               -f BAM\
               -g mm\
               --outdir peaks\
               -n heart.e11.5.rep1.ChIP-Seq
```

You try calling peaks (script4)

- Example: H3K27ac

Examining the output

- narrowPeak vs summit



Examining the output

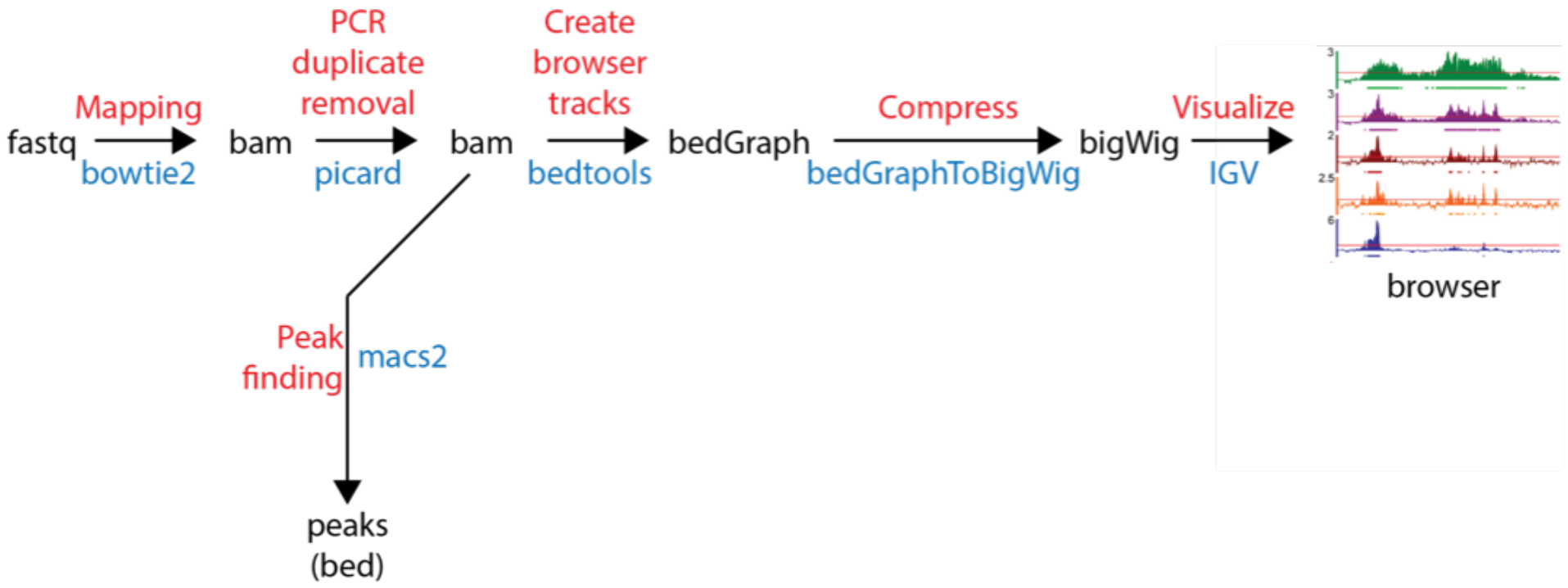
- narrowPeak

Chr	start	end	name	score	strand	signal	pval	qval	peak
chr1	3120366	3120936	name_1	86	.	5.79747	10.8519	8.64752	455
chr1	4427112	4427307	name_2	36	.	3.86498	5.64030	3.67179	104
chr1	4491729	4492010	name_3	43	.	3.96239	6.30988	4.31582	232
chr1	4492246	4493684	name_4	134	.	7.40788	15.8530	13.4946	1195
chr1	4494431	4494661	name_5	44	.	4.18706	6.43675	4.42406	57
chr1	4495649	4495910	name_6	58	.	4.52844	7.89264	5.81651	114
chr1	4495998	4496399	name_7	55	.	3.95847	7.62939	5.56343	159
chr1	4496464	4498153	name_8	364	.	13.1959	39.2158	36.4125	693
chr1	4543964	4544329	name_9	79	.	5.39833	10.1365	7.96145	182
chr1	4544635	4544829	name_10	22	.	3.11331	4.16014	2.29512	120

Visualize peaks

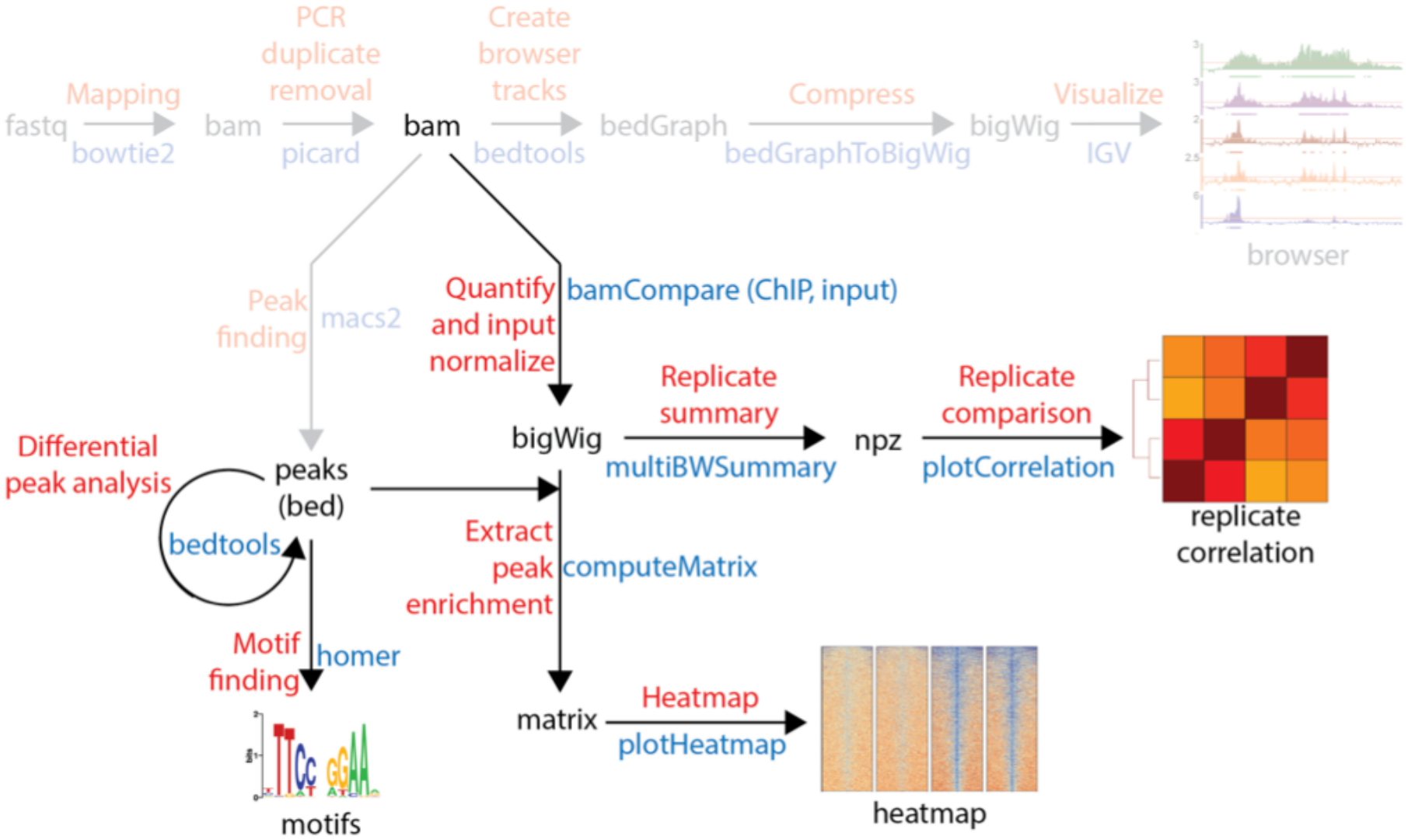
- Load the narrowPeak file onto IGV.
- Verify that peaks make sense.

Mission accomplished: Morning



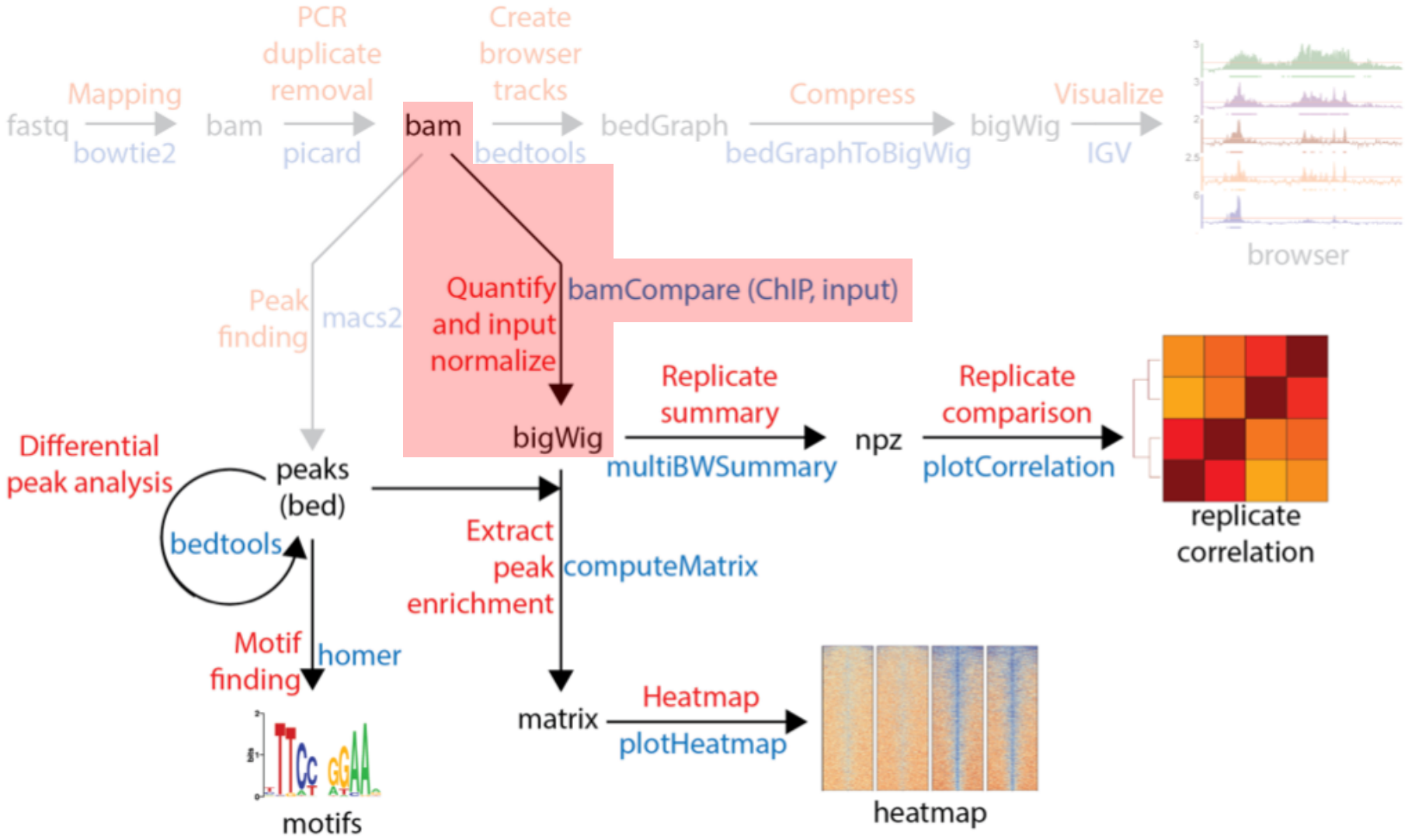
Lunch

Overview: Afternoon



Quantification

BAM ➔ log2 (ChIP / input)

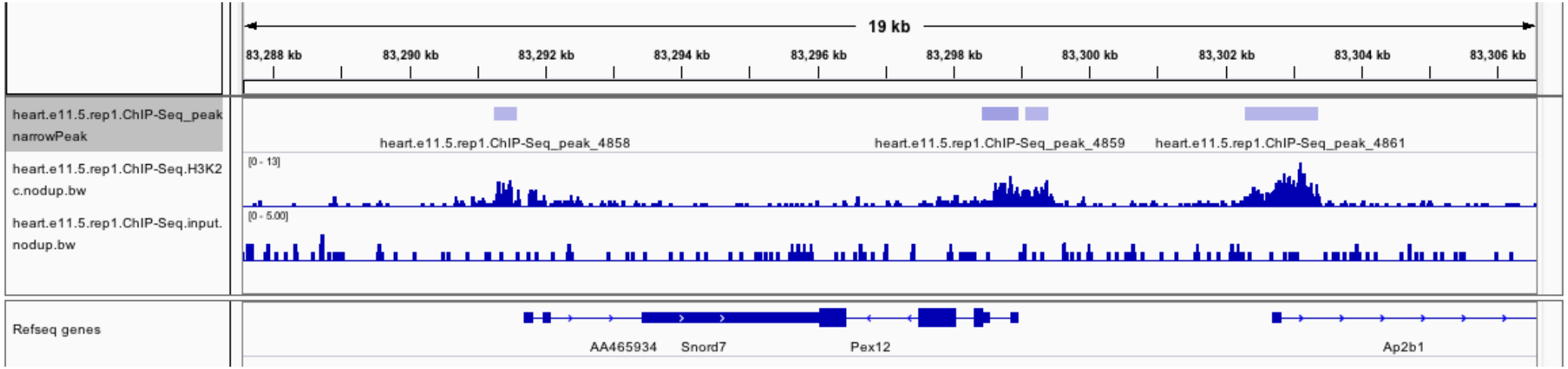


How do you quantify ChIP-Seq enrichment?

peak

ChIP

input



- ChIP enrichment = $\log_2$ (ChIP signal / input signal)
- = $\log_2$ (ChIP RPKM / input RPKM)
- RPKM = # reads per kilobase of genome per million reads sequenced

The command: bamCompare

```
module load deeptools/2.5.0.1

mkdir bw_log2ratio

bamCompare --normalizeUsingRPKM\
           -b1 [CHIP FILE]\
           -b2 [INPUT FILE]\
           -o [OUTPUT FILE]
```

Load the program deeptools.

Normalize ChIP and input.

The command: bamCompare

```
module load deeptools/2.5.0.1

mkdir bw_log2ratio

bamCompare --normalizeUsingRPKM\
           -b1 bam_files/heart.e11.5.rep1.ChIP-Seq.H3K27ac.nodup.bam\
           -b2 bam_files/heart.e11.5.rep1.ChIP-Seq.input.nodup.bam\
           -o bw_log2ratio/heart.e11.5.rep1.ChIP-Seq.H3K27ac.log2ratio.bw
```

You try normalization (script5)

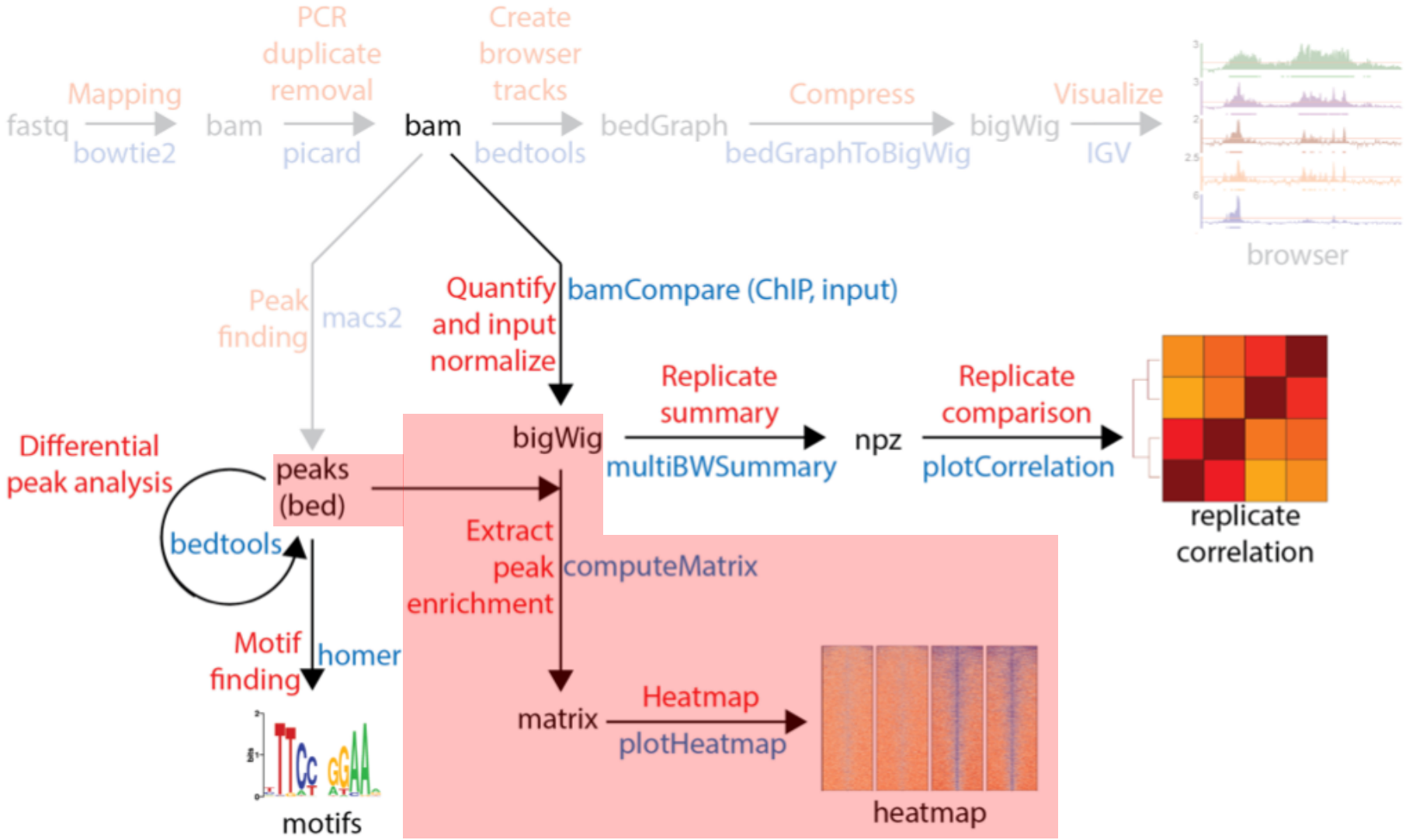
- Example: H3K27ac

Visualize the normalized data

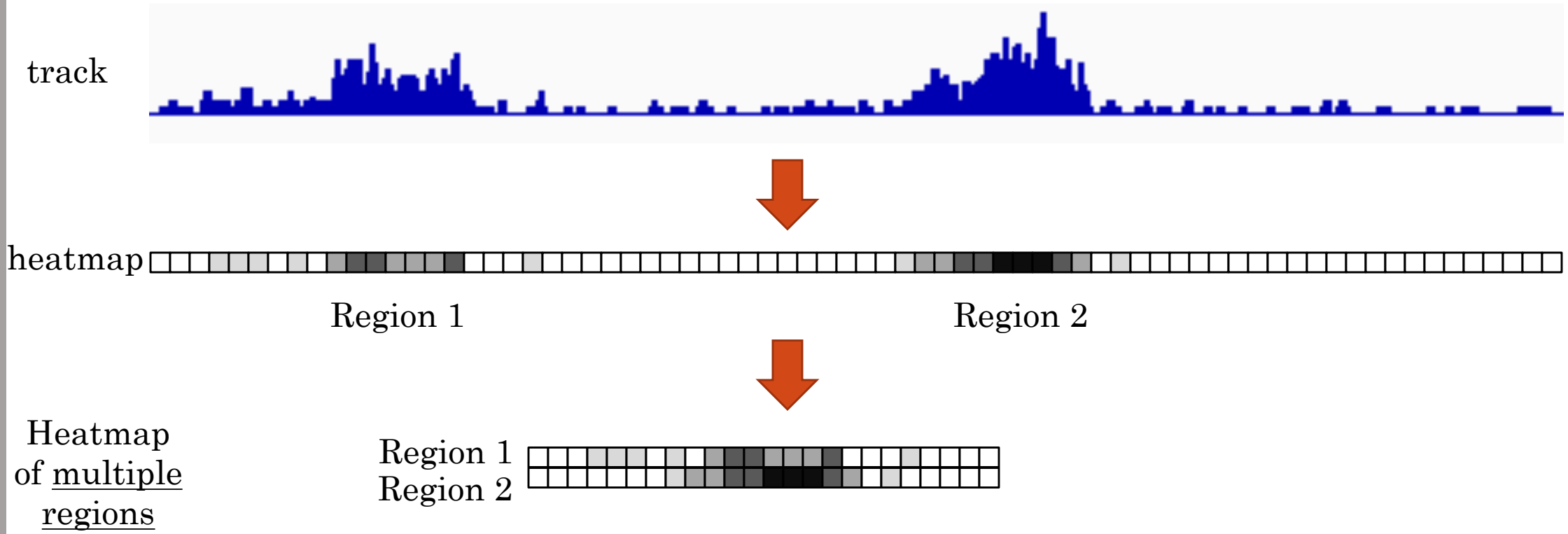
- Load the normalized bigWig file onto IGV.
- Verify that enrichment makes sense.

Heatmap of peaks

peaks + bigWig ➔ matrix ➔ heatmap



What is a heatmap?



- Need: A list of regions (peaks) and tracks to extract heatmaps

The command: computeMatrix and plotHeatmap

```
module load deeptools/2.5.0.1

mkdir matrix

computeMatrix reference-point
              -S [INPUT BIGWIG FILE]\
              -R [INPUT PEAKS SUMMITS FILE]\
              -a 3000\
              -b 3000\
              -out [OUTPUT MATRIX]

plotHeatmap --matrixFile [OUTPUT MATRIX]\
            --outFileName [OUTPUT HEATMAP]
```

Load the program deeptools.

Create a matrix of heatmap values with window -3kb to 3kb.

Use matrix to make a heatmap image.

The command: bamCompare

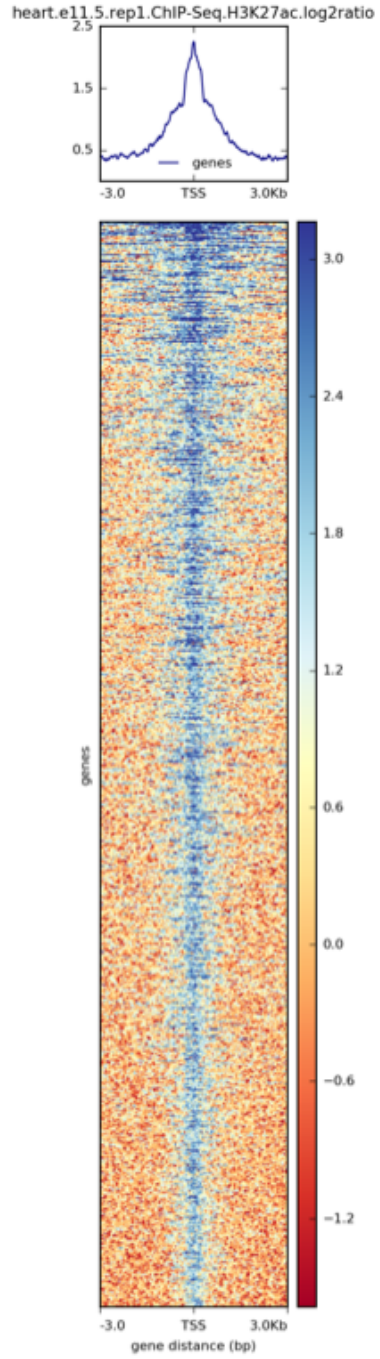
```
module load deeptools/2.5.0.1

mkdir matrix

computeMatrix reference-point
              -S bw_log2ratio/heart.e11.5.rep1.ChIP-Seq.H3K27ac.log2ratio.bw\
              -R peaks/heart.e11.5.rep1.ChIP-Seq_summits.bed\
              -a 3000\
              -b 3000\
              -out matrix/heart.e11.5.rep1.ChIP-Seq.peaks.H3K27ac.log2ratio.matrix

plotHeatmap --matrixFile matrix/heart.e11.5.rep1.ChIP-Seq.peaks.H3K27ac.log2ratio.matrix\
            --outFileName matrix/heart.e11.5.rep1.ChIP-Seq.peaks.H3K27ac.log2ratio.matrix.png
```

You try making a heatmap (script6)

- Example: H3K27ac on chr19

Examining the output

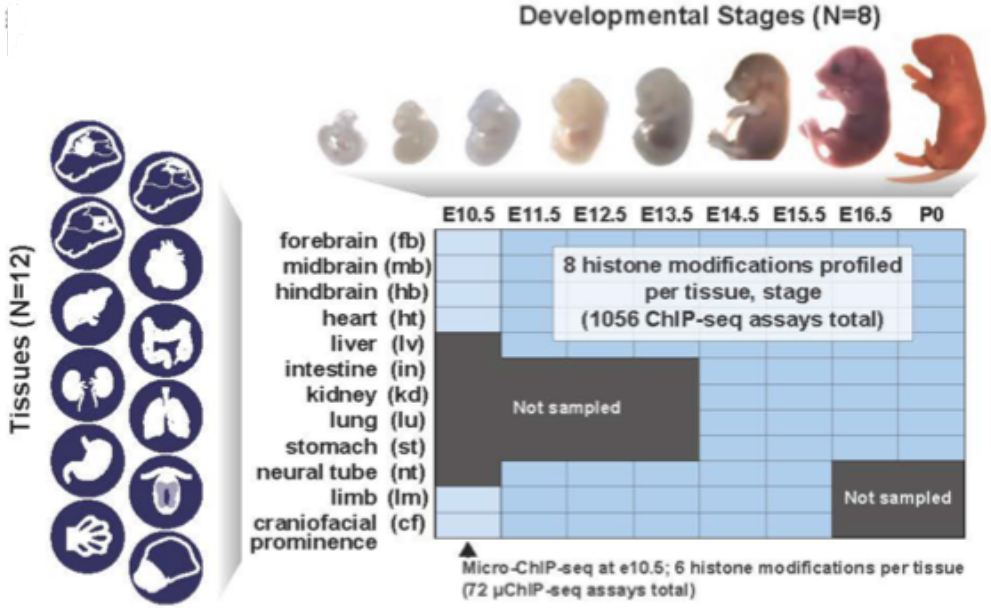
- See your heatmap (using the eog command)
- Verify that enrichment makes sense.



At this point

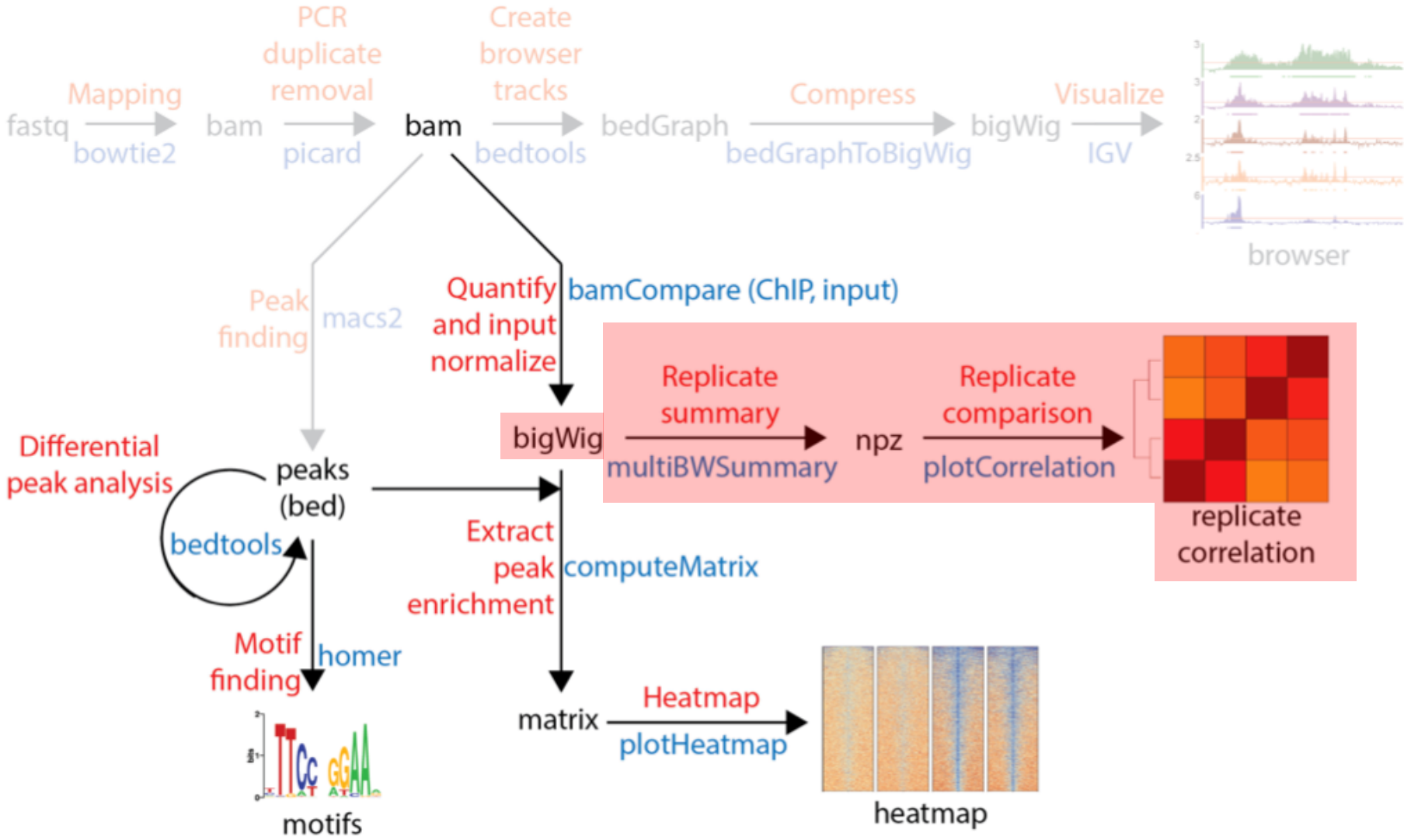
- You analyzed 1 ChIP-Seq experiment on chr19.
- Let's use more data on the full genome for more interesting analysis.
  - 2 cells: e11.5 and p0
  - 2 libraries per cell: H3K27ac and input
  - 2 replicates per library: rep1 and rep2
- Setup:

```
cd ..

mkdir part2

cd part2

cp ../../shared/class2/part2/* .
```



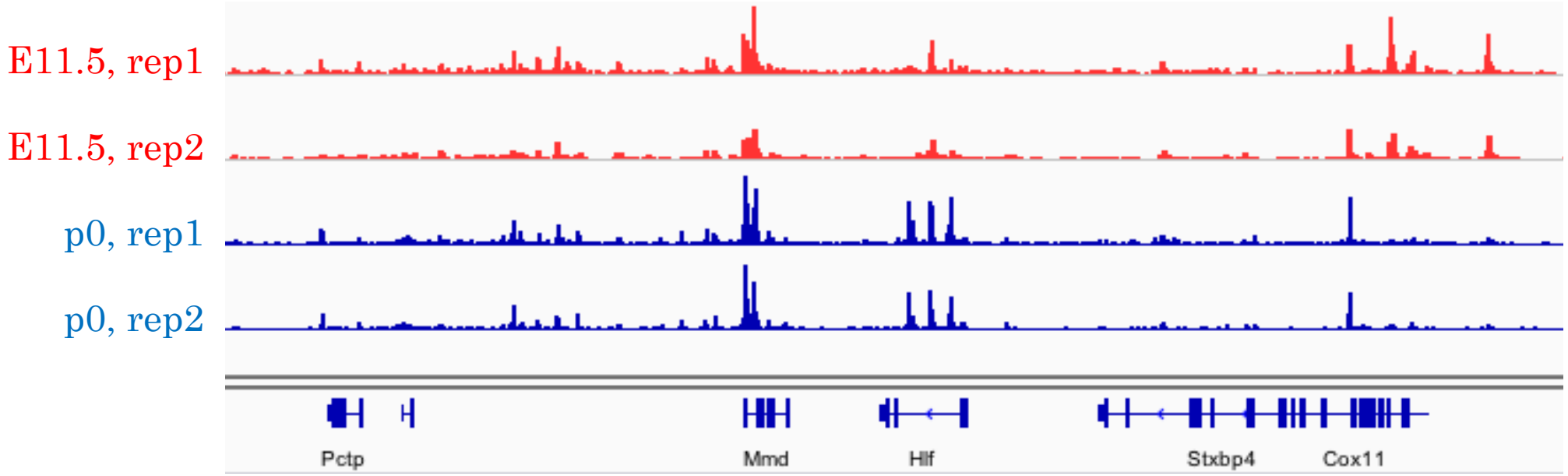
Compare replicates

Many bigWigs ➔ npz ➔ plot



How to QC multiple samples?

- Key idea:
  - Libraries from biological replicates are **more** similar
  - Libraries from different conditions are **less** similar



The command: multiBWSummary and plotCorrelation

```
module load deeptools/2.5.0.1

cp -rp ../../shared/processed/ChIP-Seq/bw_log2ratio .

multiBigwigSummary bins -b [BW FILES]\
                        -o [OUTPUT NPZ]

plotCorrelation -in [OUTPUT NPZ]\
                --corMethod pearson\
                --whatToPlot heatmap\
                -o [HEATMAP OUTPUT]

plotCorrelation -in [OUTPUT NPZ]\
                --corMethod pearson\
                --whatToPlot scatterplot\
                -o [SCATTERPLOT OUTPUT]
```

Load the program deeptools.

Copy bw files.

Summarize.

Plot correlation and scatterplot comparisons.

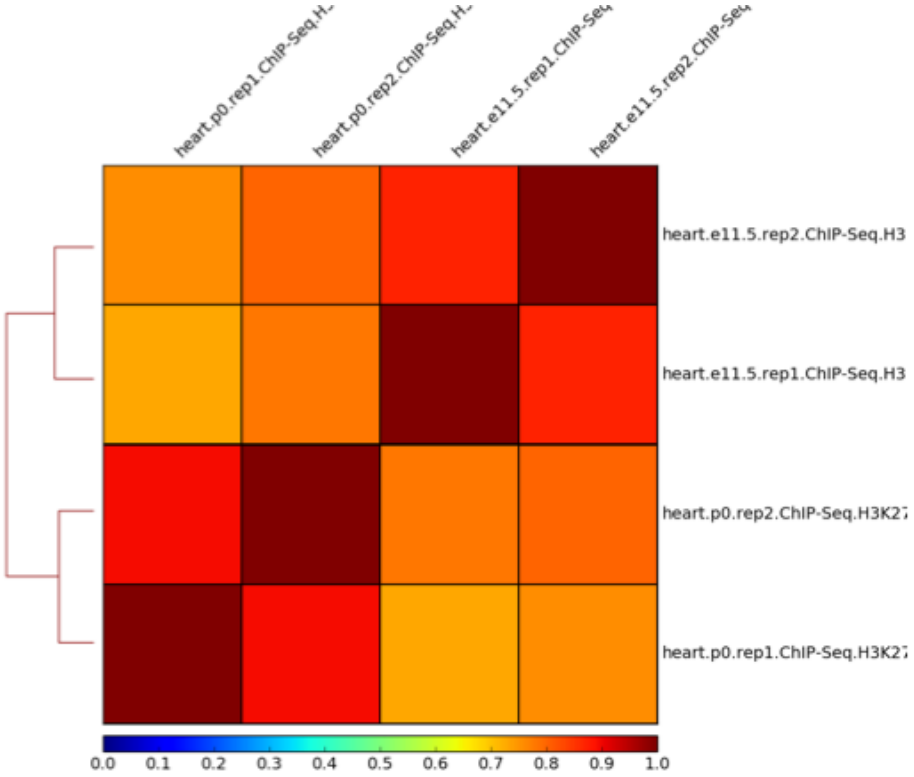
The command: multiBWSummary and plotCorrelation

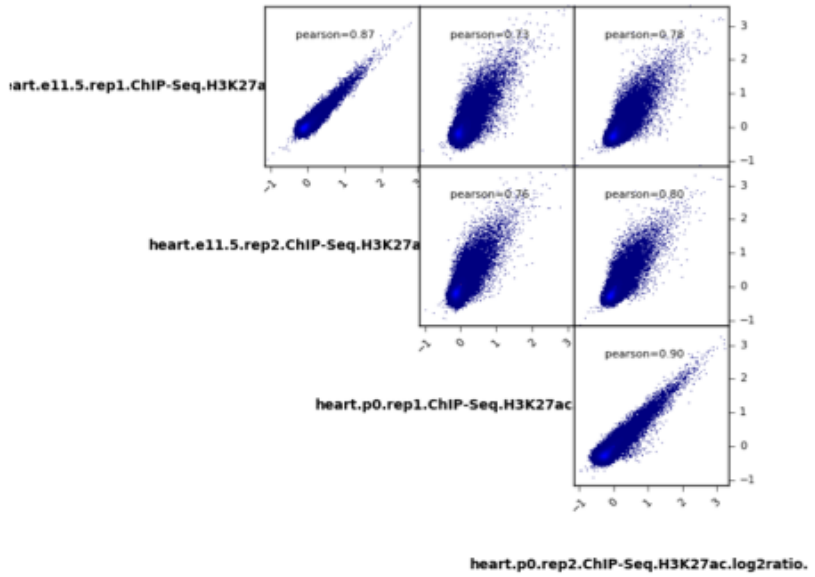
```
module load deeptools/2.5.0.1

cp -rp ../../shared/processed/ChIP-Seq/bw_log2ratio .

multiBigwigSummary bins -b bw_log2ratio/*bw\
                        -o all_samples.H3K27ac.npz

plotCorrelation -in all_samples.H3K27ac.npz\
                --corMethod pearson\
                --whatToPlot heatmap\
                -o all_samples.H3K27ac.heatmap.png

plotCorrelation -in all_samples.H3K27ac.npz\
                --corMethod pearson\
                --whatToPlot scatterplot\
                -o all_samples.H3K27ac.scatterplot.png
```

You try plotting correlation (script7)

Examining the output

- Do the replicates correlate?
- Verify that enrichment makes sense.





Case study: Questions we'll answer

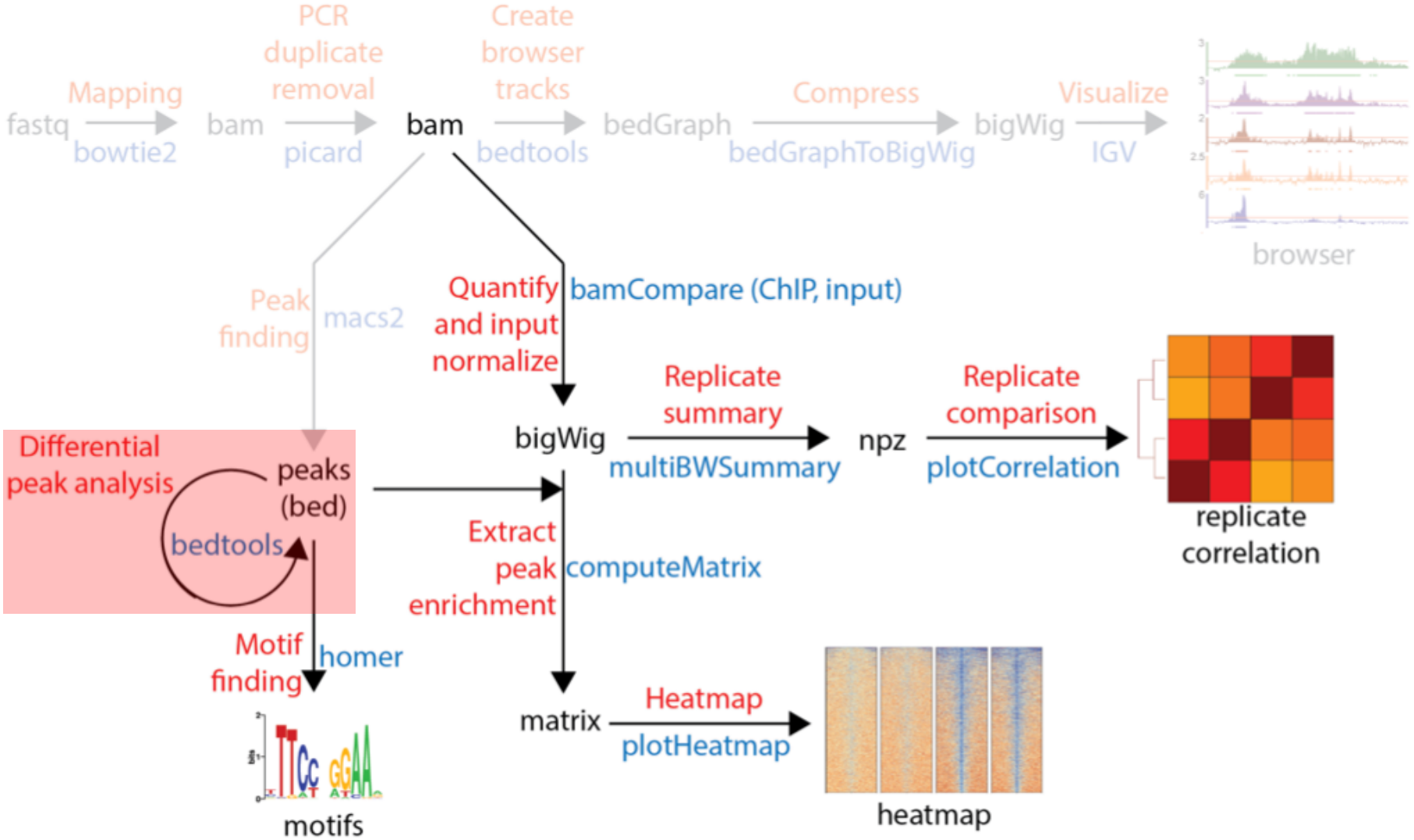
1. Did the experiment work (technically)?
   - Does it map?
   - Is there ChIP enrichment?
   - Do the biological replicates agree?

2. How do H3K27ac peaks change during heart development?
   - Are more peaks gained or lost?

3. Which transcription factors drive development?

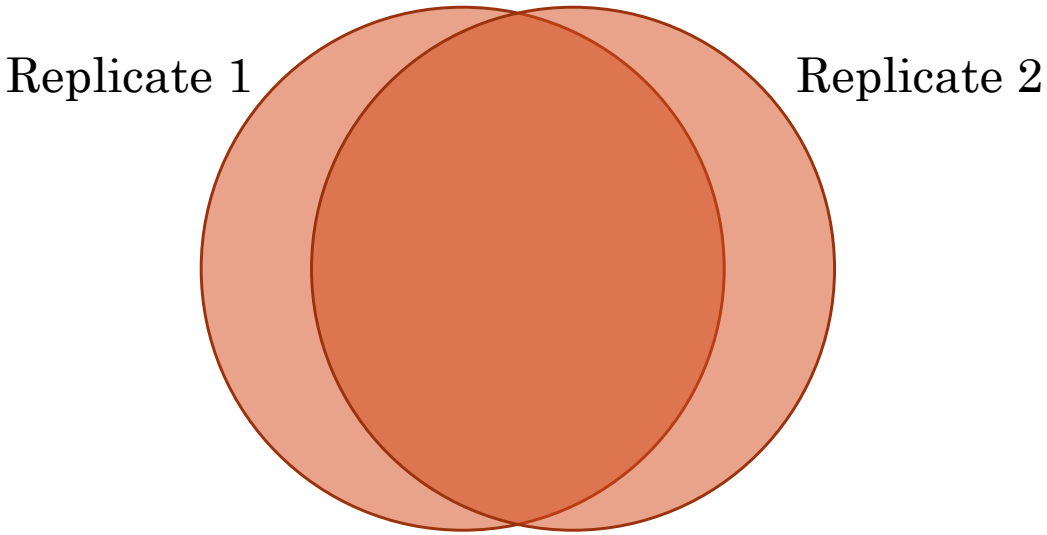
Differential peak analysis

Peaks ➔ peaks



Find confident peaks

- Key idea:
  - Confident peaks: peaks shared between replicates



The command: bedtools

```
module load bedtools

cp -rp ../../shared/processed/ChIP-Seq/peaks .

# get replicate common
bedtools window -w 2000\
                -u\
                -a [PEAKS1 INPUT BED]\
                -b [PEAKS2 INPUT BED]\
  > [OUTPUT PEAKS BED]
```

Load bedtools.

Copy peaks.

e11.5 peaks.

The command: bedtools

```
module load bedtools

cp -rp ../../shared/processed/ChIP-Seq/peaks .

# get replicate common
bedtools window -w 2000\
                -u\
                -a peaks/heart.e11.5.rep1.ChIP-Seq_summits.bed\
                -b peaks/heart.e11.5.rep2.ChIP-Seq_summits.bed\
  > peaks/heart.e11.5.common.ChIP-Seq_summits.bed
```

You try processing peaks (script8)

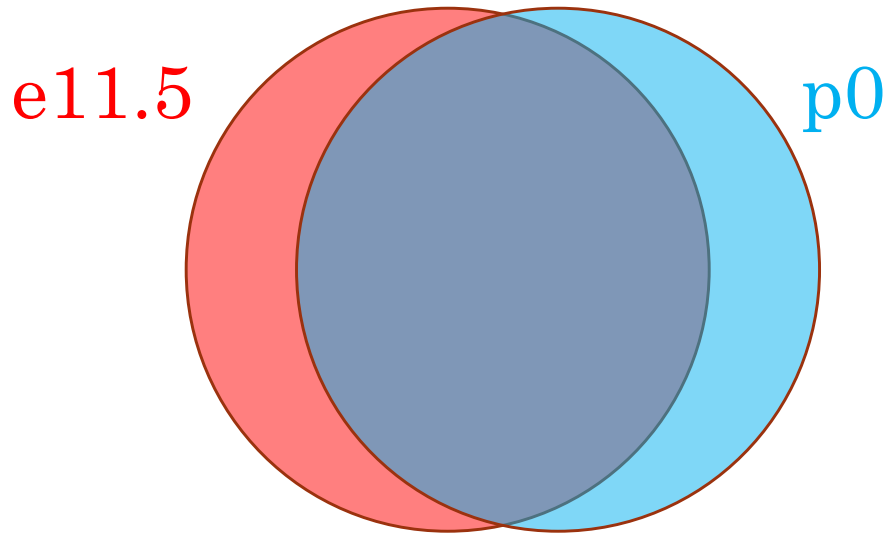
- Example: e11.5
- You do: p0

Examining the output

- How many peaks are there in each replicate?
- How many peaks are common between replicates?

Find biologically interesting peaks

- Unique to e11.5 heart
- Unique to p0 heart



The command: bedtools

```
module load bedtools

# get early unique
bedtools window -w 2000\
                -v\
                -a [INPUT PEAKS]\
                -b [INPUT PEAKS TO REMOVE]\
  > [OUTPUT PEAKS]
```

Load bedtools.

Get e11.5-unique.

The command: bedtools

```
module load bedtools

# get early unique
bedtools window -w 2000\
                -v\
                -a peaks/heart.e11.5.common.ChIP-Seq_summits.bed\
                -b peaks/heart.p0.common.ChIP-Seq_summits.bed\
  > peaks/heart.early_unique.ChIP-Seq_summits.bed
```

You try processing peaks (script8)

- Example: e11.5-unique
- You do: p0-unique

Examining the output

- How many peaks are unique to e11.5?
- How many peaks are unique to p0?

Case study: Questions we'll answer

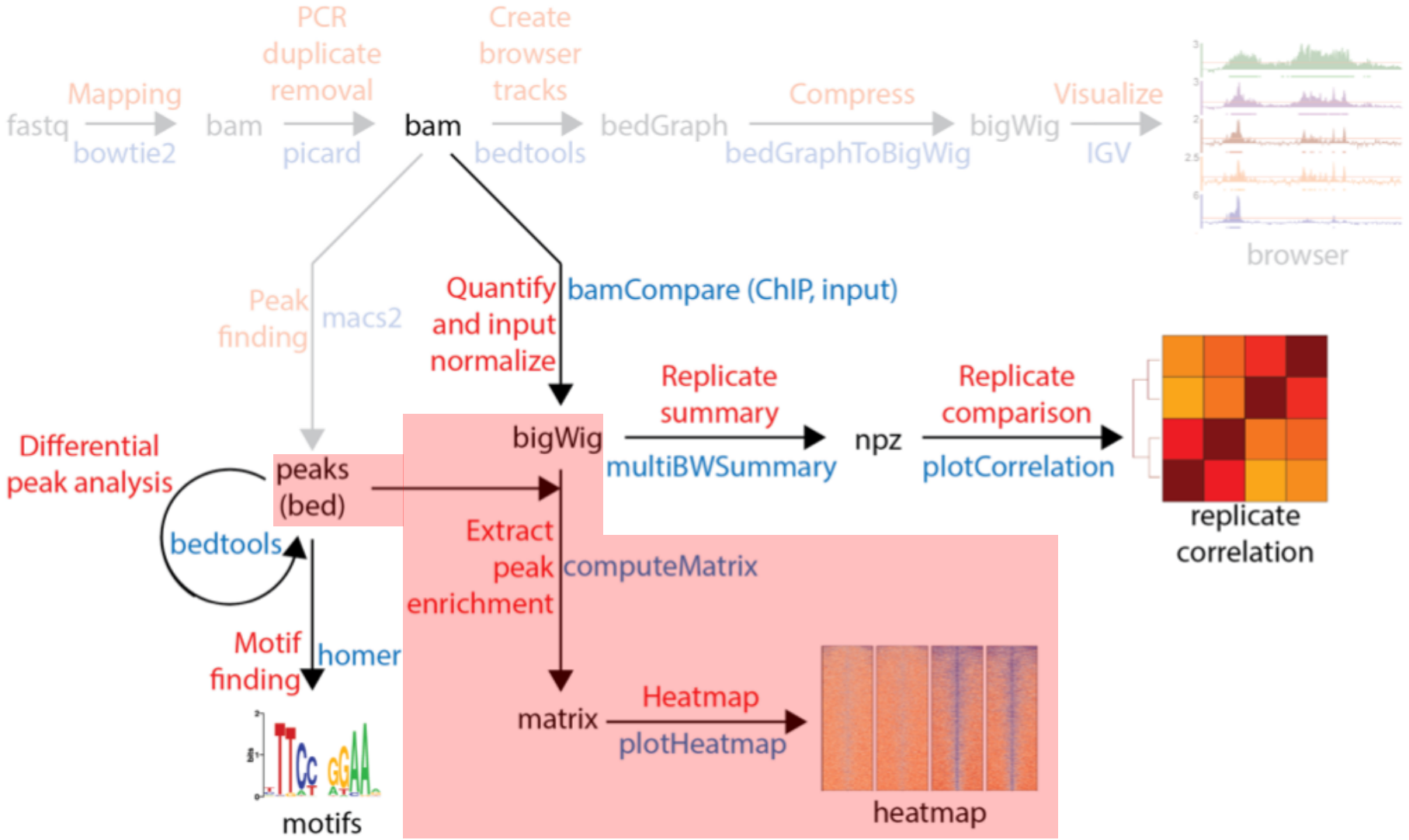
1. Did the experiment work (technically)?
   - Does it map?
   - Is there ChIP enrichment?
   - Do the biological replicates agree?

2. How do H3K27ac peaks change during heart development?
   - Are more peaks gained or lost?

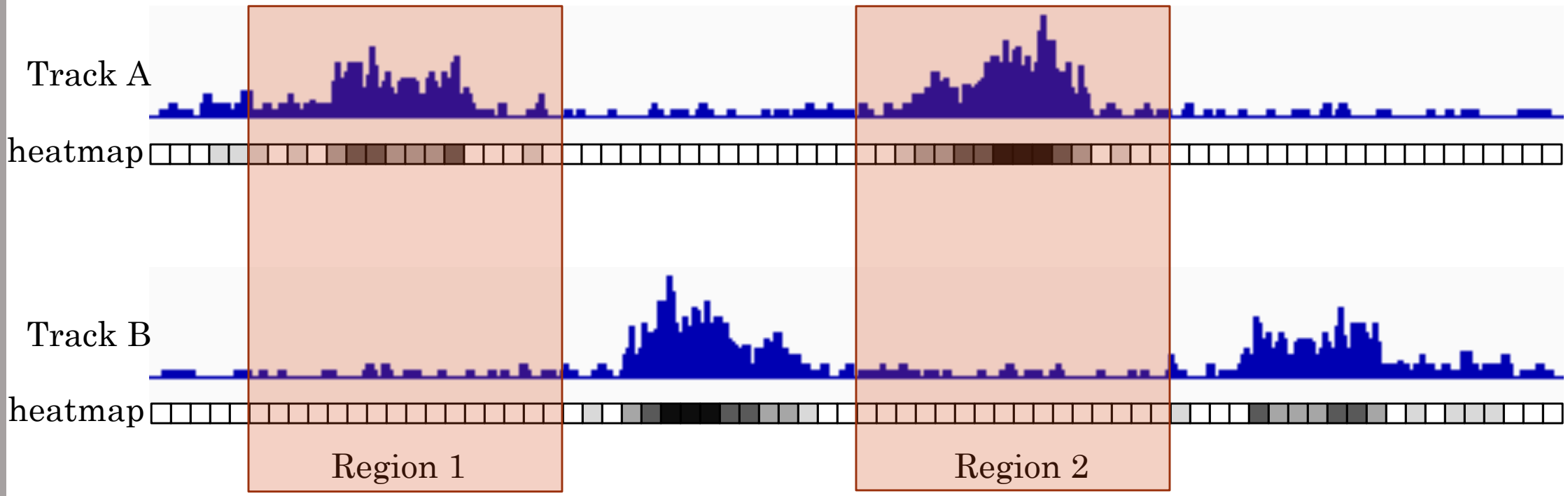
3. Which transcription factors drive development?

Heatmaps of peaks

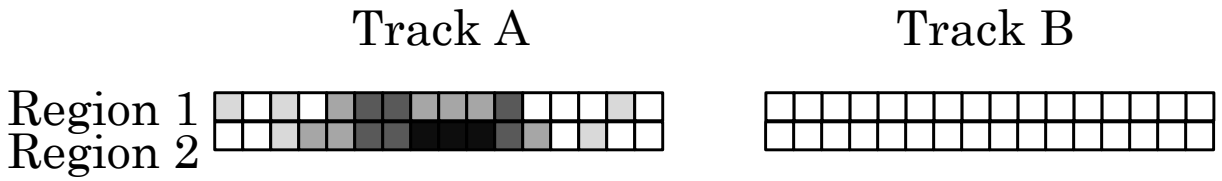
peaks + bigWig ➔ matrix ➔ heatmap



Side-by-side heatmaps



Heatmap of multiple regions



The command: computeMatrix and plotHeatmap

```
module load deeptools/2.5.0.1

mkdir matrix

computeMatrix reference-point
              -S [INPUT BIGWIG FILE1]\
                 [INPUT BIGWIG FILE2]\
                 [INPUT BIGWIG FILE3]\
                 [INPUT BIGWIG FILE4]\
              -R [INPUT PEAKS SUMMITS FILE]\
              -a 3000\
              -b 3000\
              -p max\
              -out [OUTPUT MATRIX]

plotHeatmap --matrixFile [OUTPUT MATRIX]\
            --outFileName [OUTPUT HEATMAP]
```

Load the program deeptools.

Create a matrix of heatmap values with window -3kb to 3kb.

Use matrix to make a heatmap image.

The command: computeMatrix and plotHeatmap

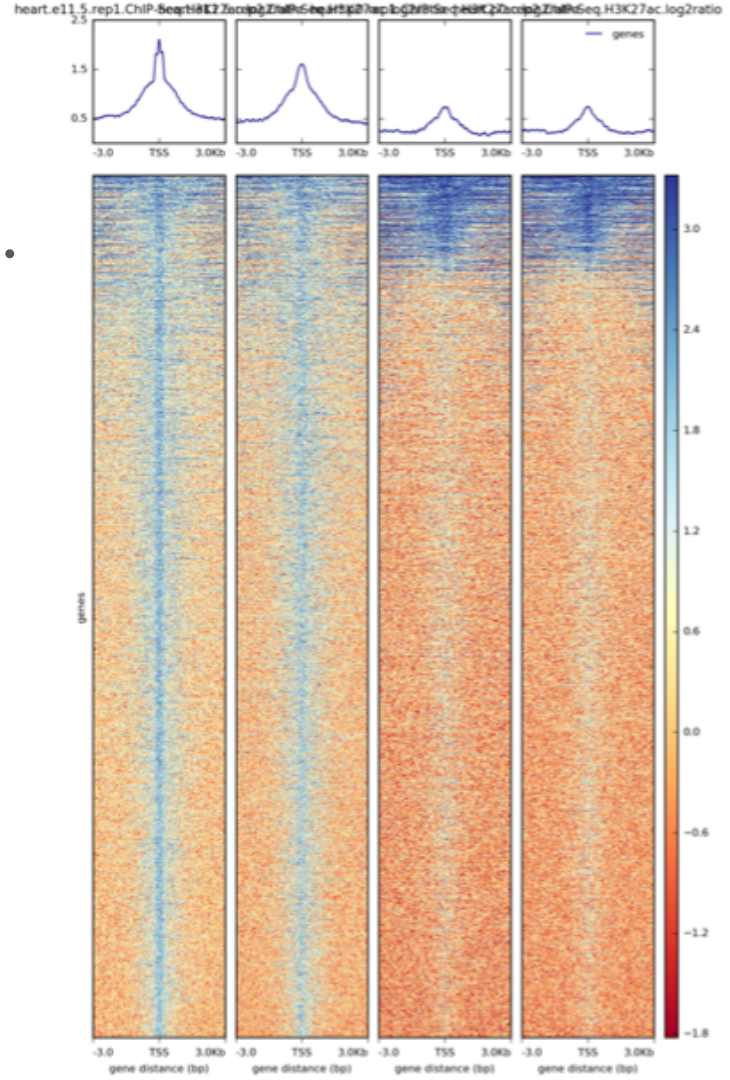
```
module load deeptools/2.5.0.1

mkdir matrix

computeMatrix reference-point
              -S bw_log2ratio/heart.e11.5.rep1.ChIP-Seq.H3K27ac.log2ratio.bw\
                 bw_log2ratio/heart.e11.5.rep2.ChIP-Seq.H3K27ac.log2ratio.bw\
                 bw_log2ratio/heart.p0.rep1.ChIP-Seq.H3K27ac.log2ratio.bw\
                 bw_log2ratio/heart.p0.rep2.ChIP-Seq.H3K27ac.log2ratio.bw\
              -R peaks/heart.early_unique.ChIP-Seq_summits.bed\
              -a 3000\
              -b 3000\
              -p max\
              -out matrix/all_samples.peaks_early_unique.H3K27ac.log2ratio.matrix

plotHeatmap --matrixFile matrix/all_samples.peaks_early_unique.H3K27ac.log2ratio.matrix\
            --outFileName matrix/all_samples.peaks_early_unique.H3K27ac.log2ratio.matrix.png
```

You try making a multi heatmap (script9)

- Example: e11.5-unique
- You try: p0-unique

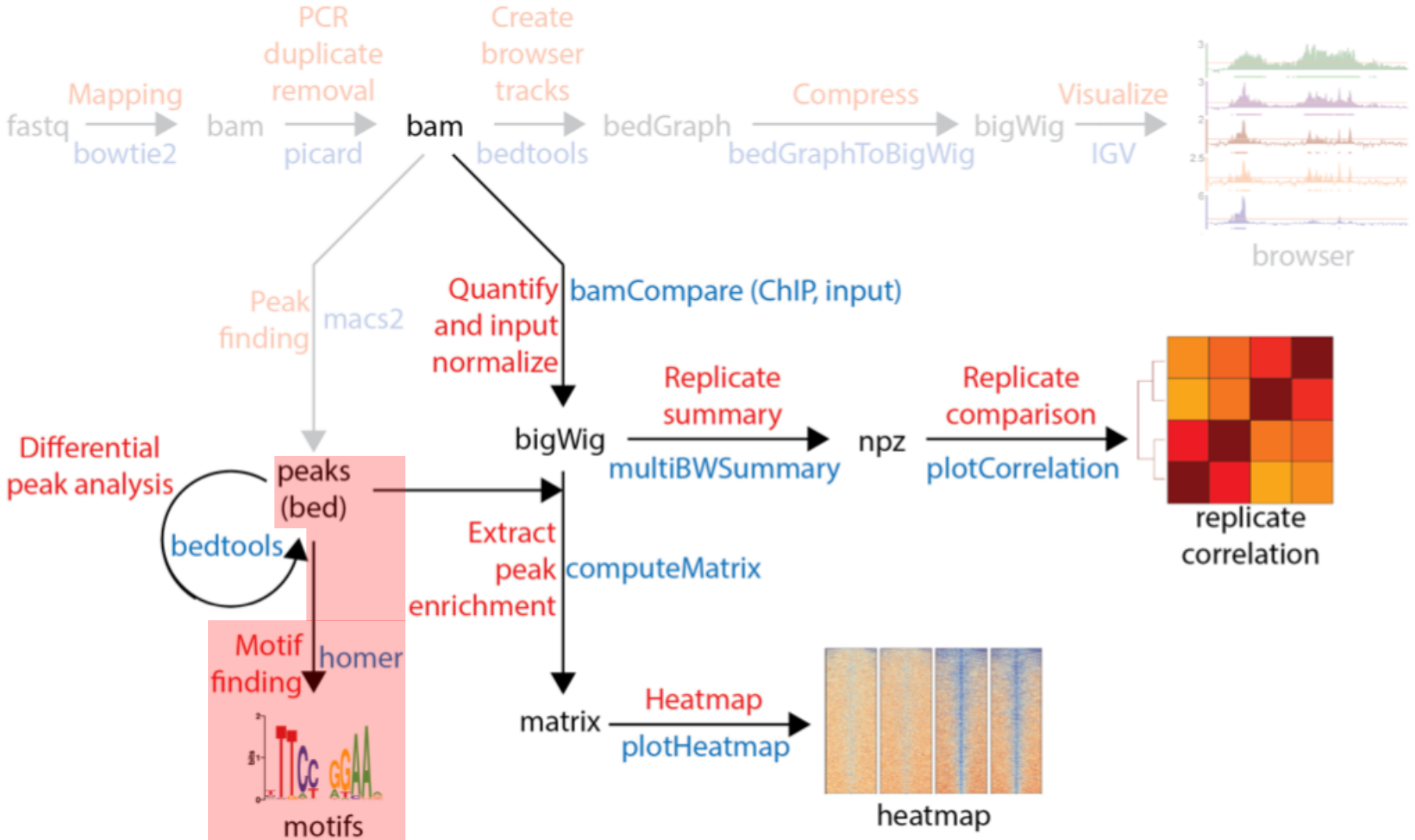
Examining the output

- See your heatmap
- Verify that enrichment makes sense.

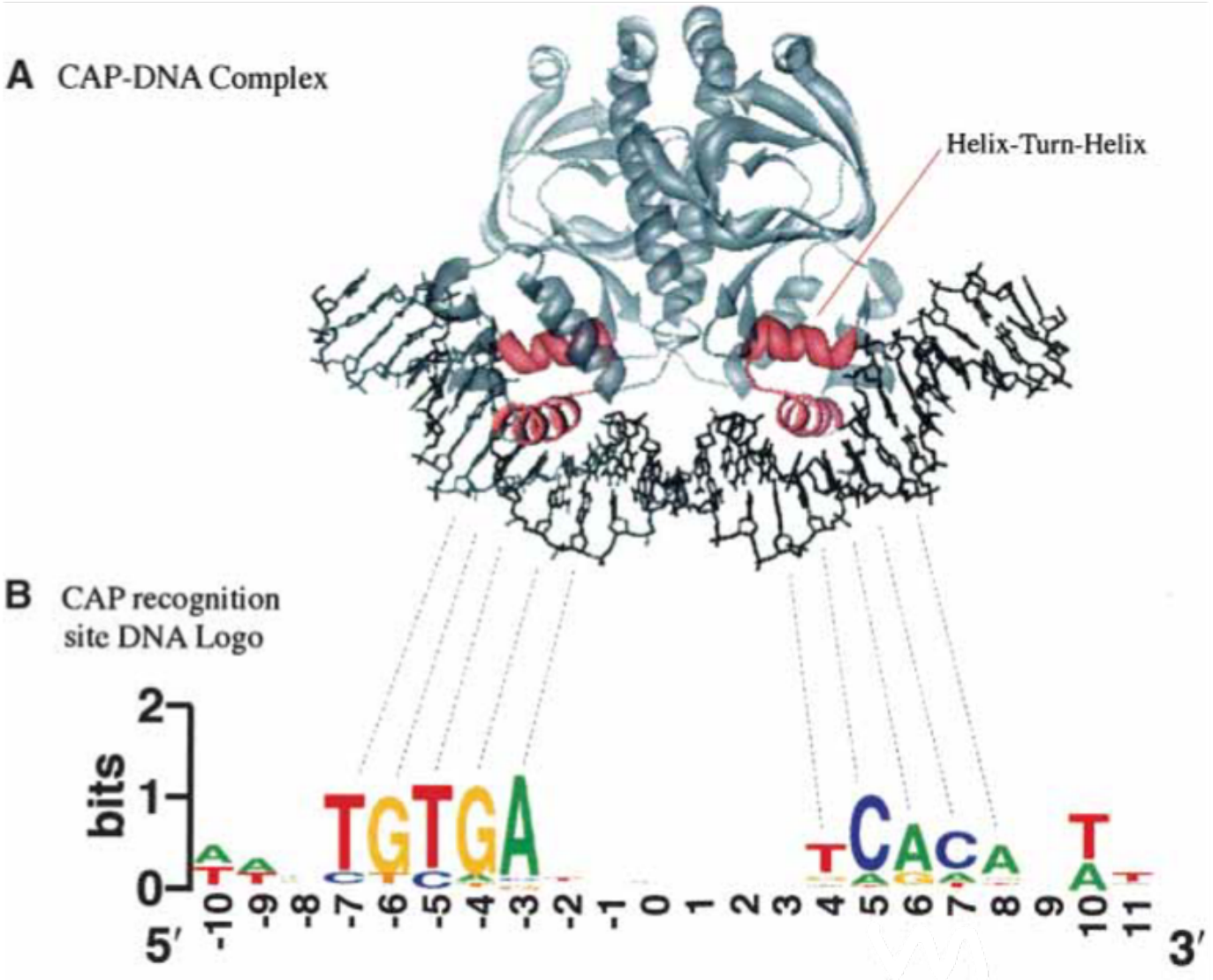


Finding motifs

peaks ➔ motifs



What's a sequence motif?



The command: homer

```
module load homer

mkdir motifs

findMotifsGenome.pl [INPUT PEAKS (BED) FOREGROUND]\
                    mm10\
                    [OUTPUT FOLDER]\
                    -size 500\
                    -p 32\
                    -bg [INPUT PEAKS (BED) BACKGROUND]\
                    -nomotif
```

Load the program deeptools.

Find known motifs.

The command: homer

```
module load homer

mkdir motifs

findMotifsGenome.pl peaks/heart.early_unique.ChIP-Seq_summits.bed\
                    mm10\
                    motifs/heart.early_unique\
                    -size 500\
                    -p 32\
                    -bg peaks/heart.late_unique.ChIP-Seq_summits.bed\
                    -nomotif
```

You try finding motifs (script10)

- Example: e11.5-unique
- You try: p0-unique

Examining the output

- Verify that the motifs makes sense.

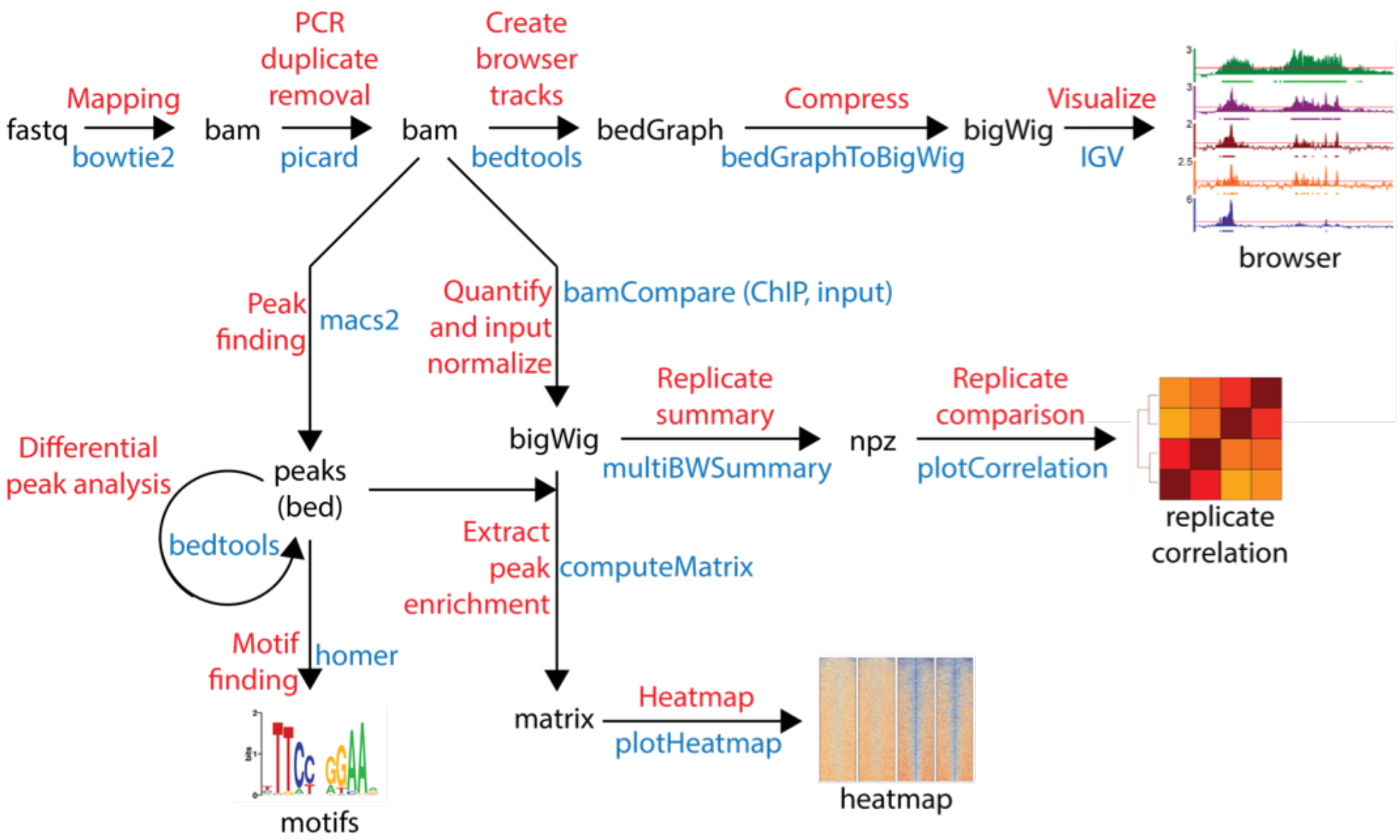
Case study: Questions we'll answer

1. Did the experiment work (technically)?
   - Does it map?
   - Is there ChIP enrichment?
   - Do the biological replicates agree?

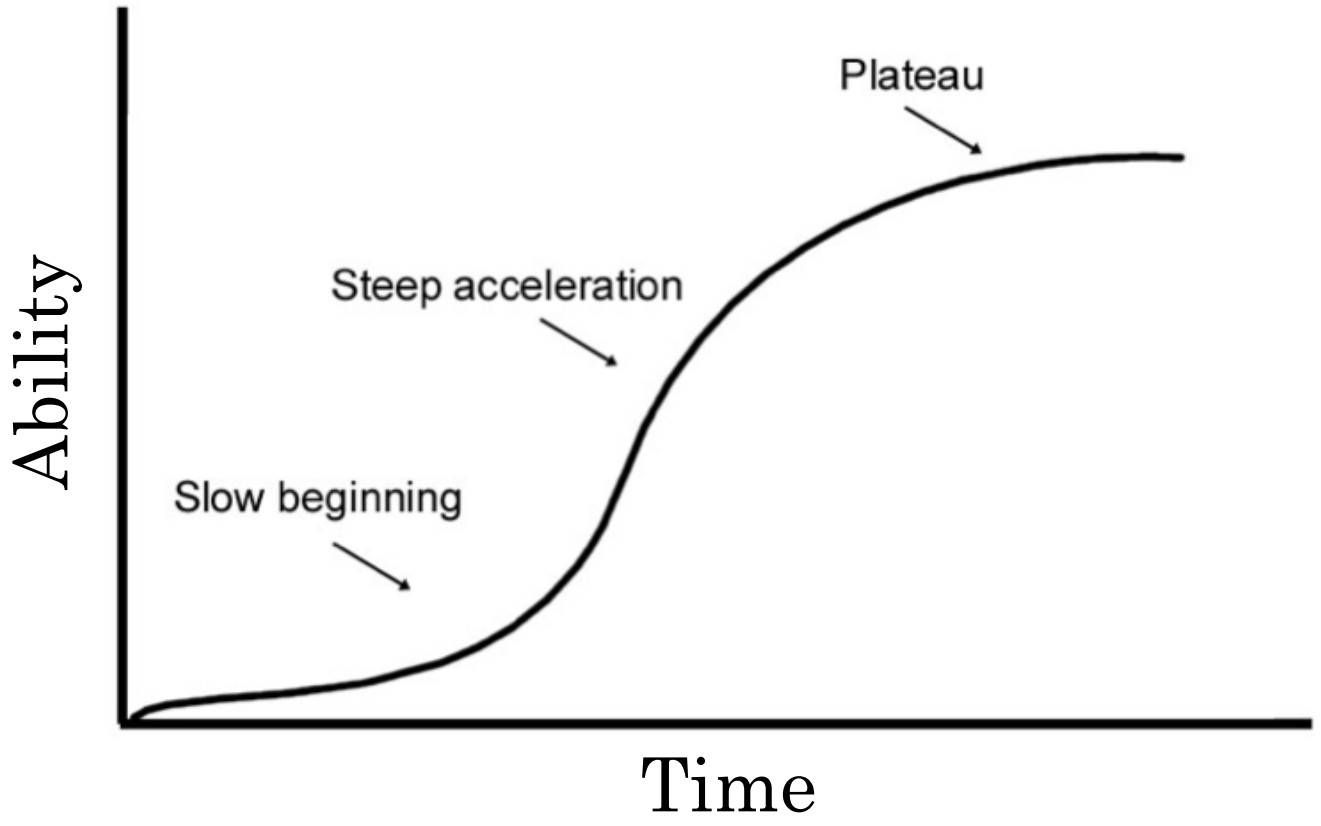
2. How do H3K27ac peaks change during heart development?
   - Are more peaks gained or lost?

3. Which transcription factors drive development?

Mission accomplished: Today



Bioinformatics: steep learning curve



Tips

- It will be frustrating – keep trying!
- Ask questions!
- Work with a partner!
- Does it work? Try it!
- Tell us how we can improve.


NOW YOU CAN
PICK UP ANYTHING !!!
ANALYZE