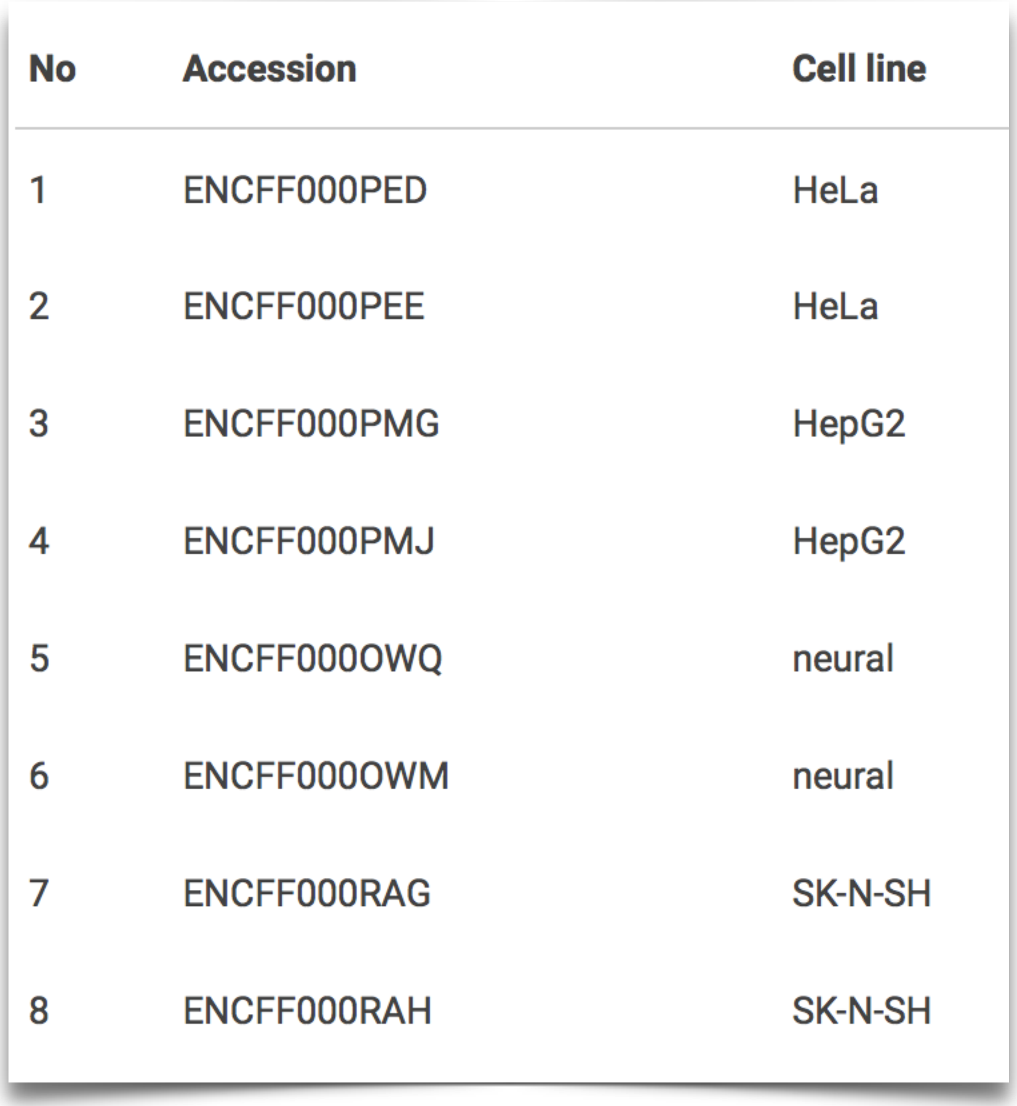
ChIP-seq data analysis

Introduction to practicals I: data processing

ENCODE Dataset

- REST is transcriptional repressor that represses neuronal genes in non-neuronal cells
- It represses transcription by binding a DNA sequence element called neuron-restrictive silencer element (NRSE)
- The protein is also found in undifferentiated neuronal progenitor cells, and REST may act as a master negative regulator of neurogenesis

chr 1 & chr 2



No	Accession	Cell line
1	ENCFF000PED	HeLa
2	ENCFF000PEE	HeLa
3	ENCFF000PMG	HepG2
4	ENCFF000PMJ	HepG2
5	ENCFF000OWQ	neural
6	ENCFF000OWM	neural
7	ENCFF000RAG	SK-N-SH
8	ENCFF000RAH	SK-N-SH

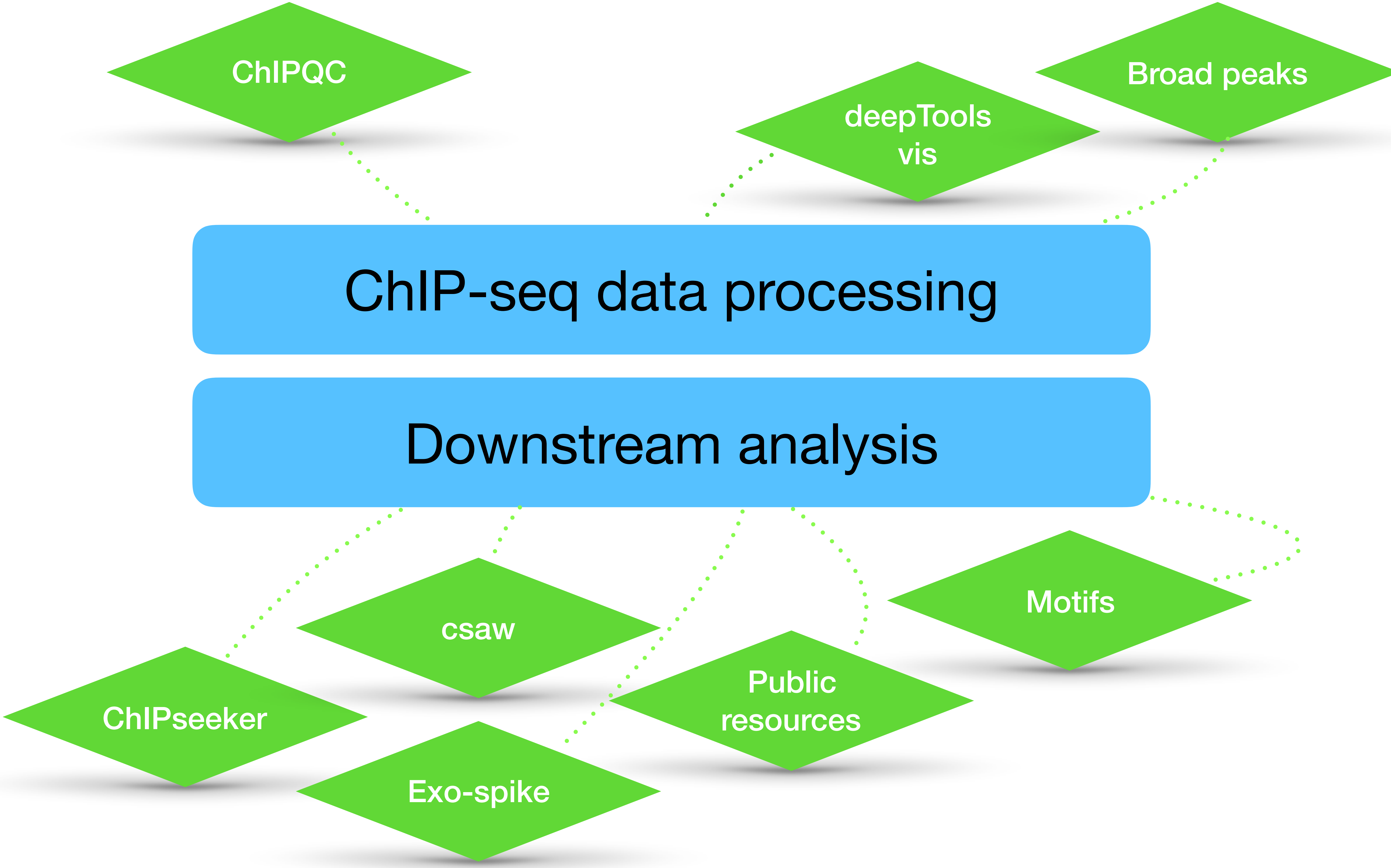
+ corresponding inputs

Practicals

ChIP-seq data processing

Downstream analysis

Practicals



ChIPQC

deepTools vis

Broad peaks

ChIP-seq data processing

Downstream analysis

csaw

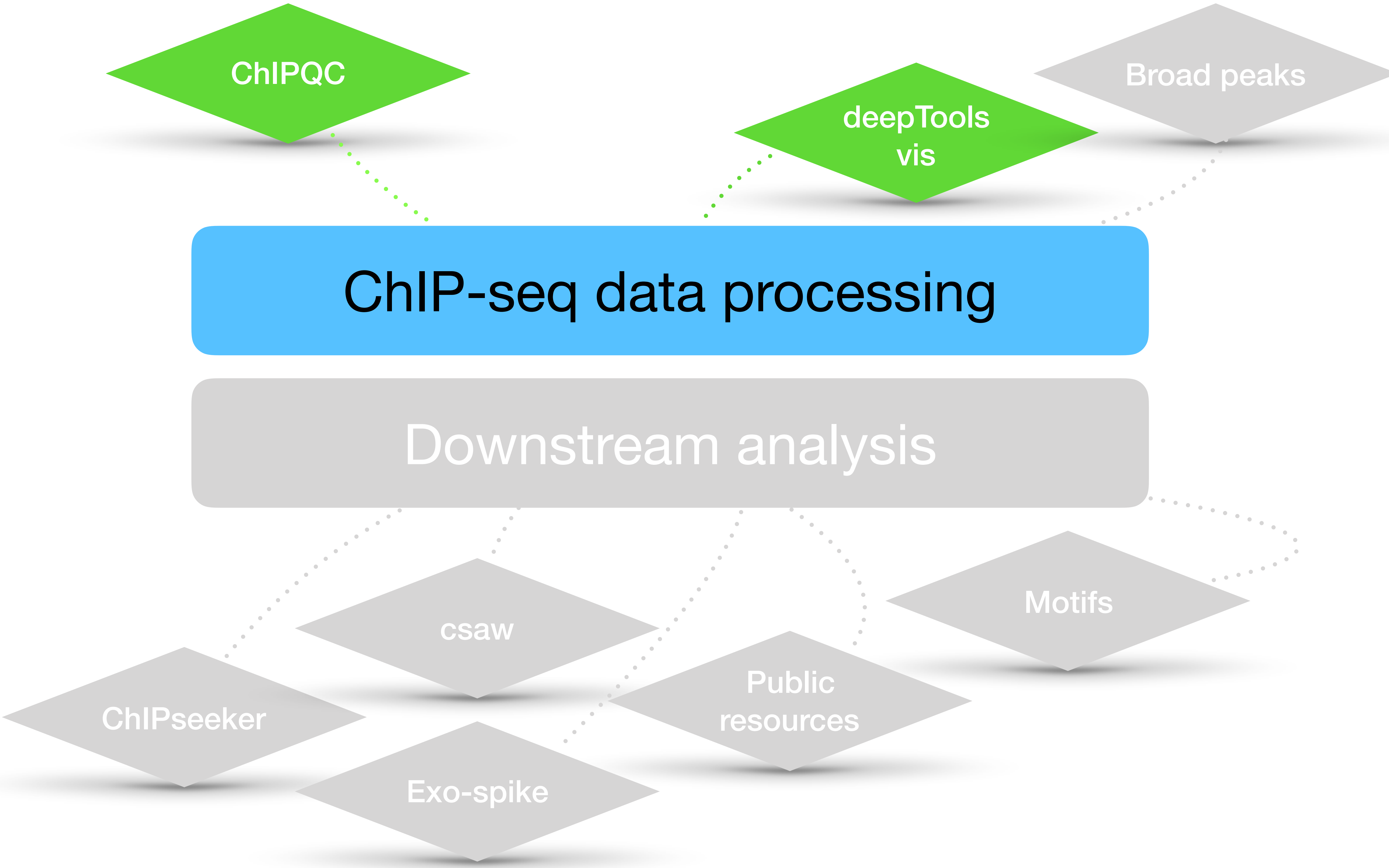
Motifs

ChIPseeker

Public resources

Exo-spike

Practicals



ChIPQC

deepTools
vis

Broad peaks

ChIP-seq data processing

Downstream analysis

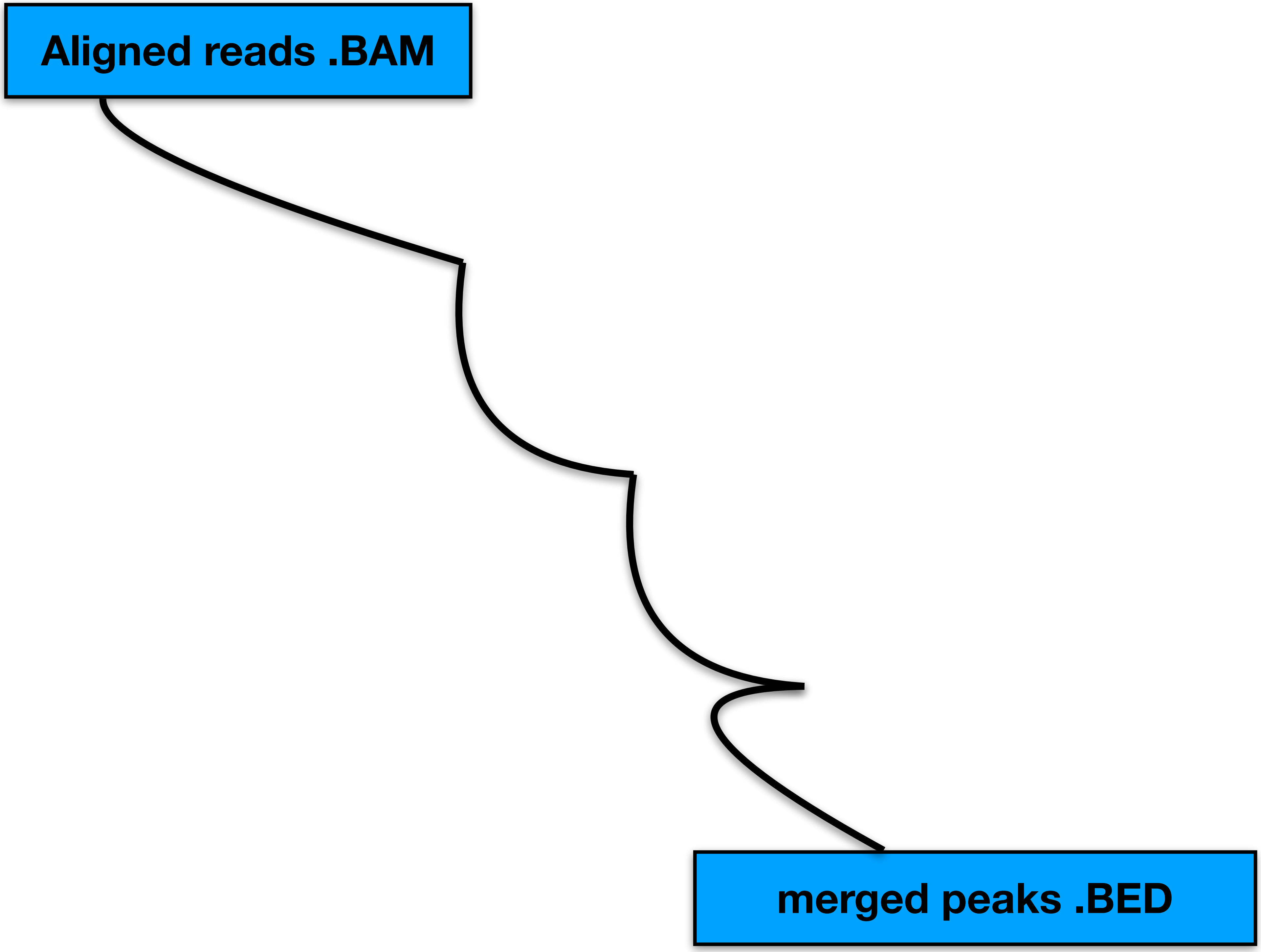
csaw

Motifs

Public
resources

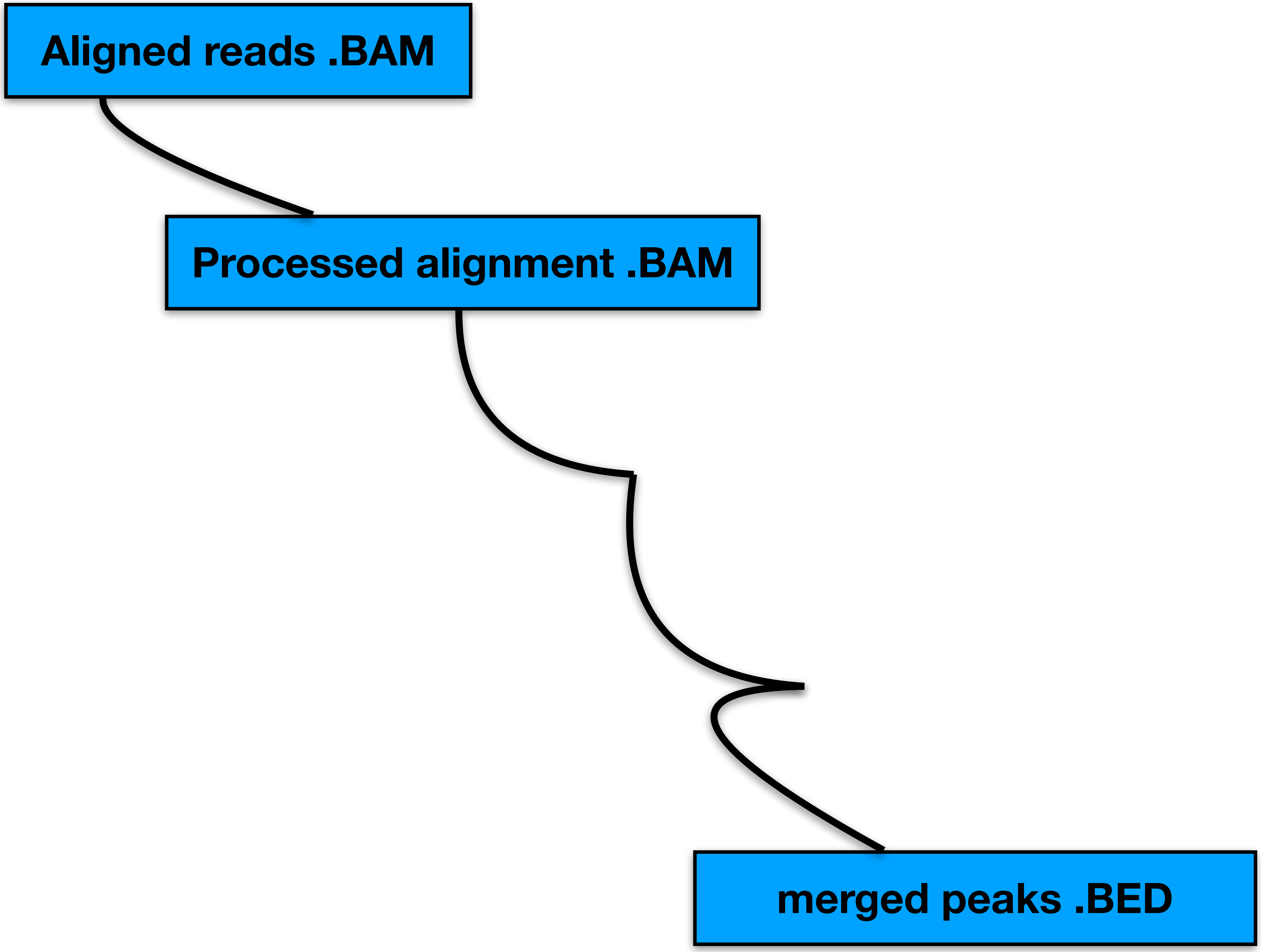
ChIPseeker

Exo-spike



Aligned reads .BAM

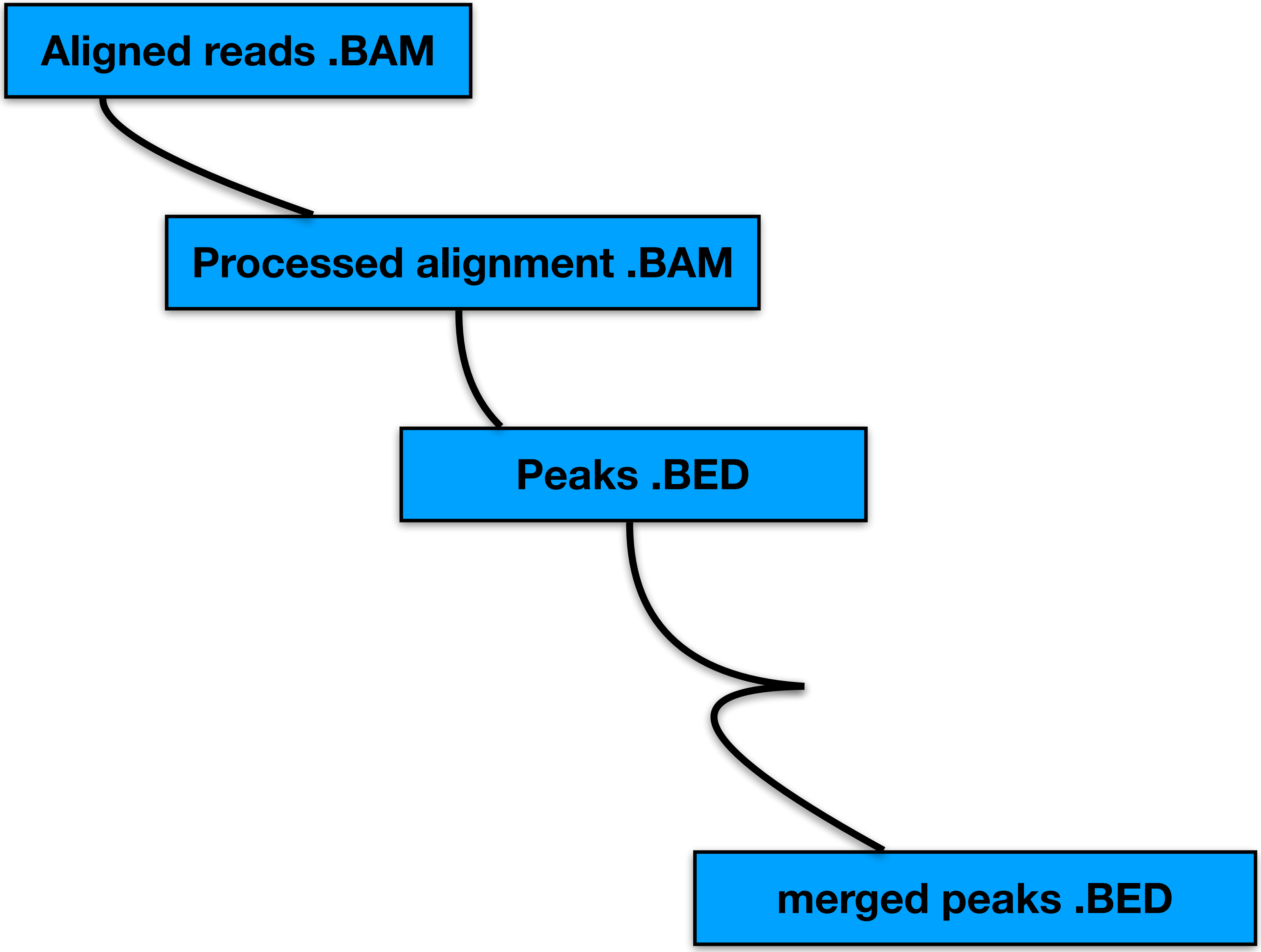
merged peaks .BED



Aligned reads .BAM

Processed alignment .BAM

merged peaks .BED

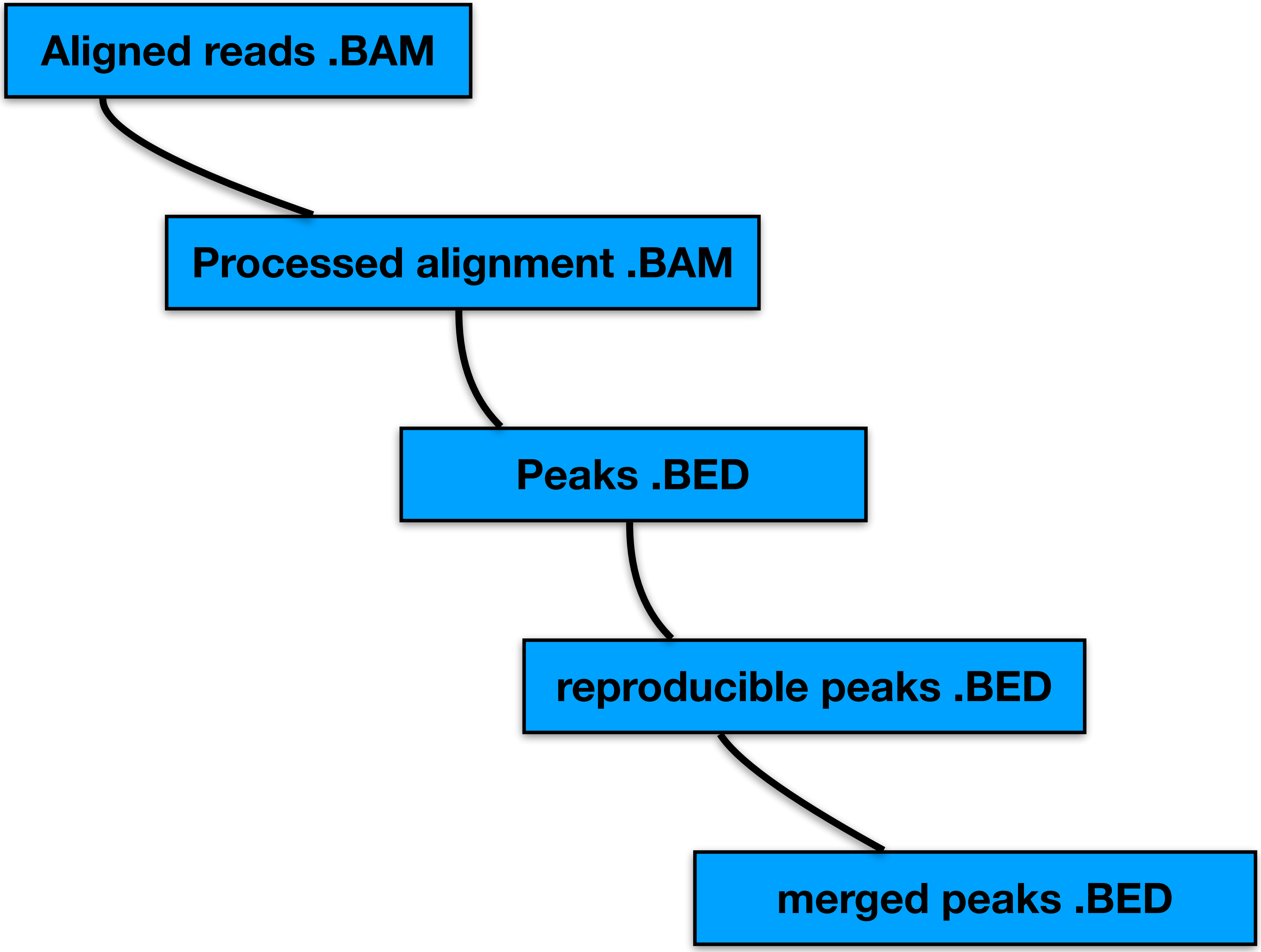


Aligned reads .BAM

Processed alignment .BAM

Peaks .BED

merged peaks .BED



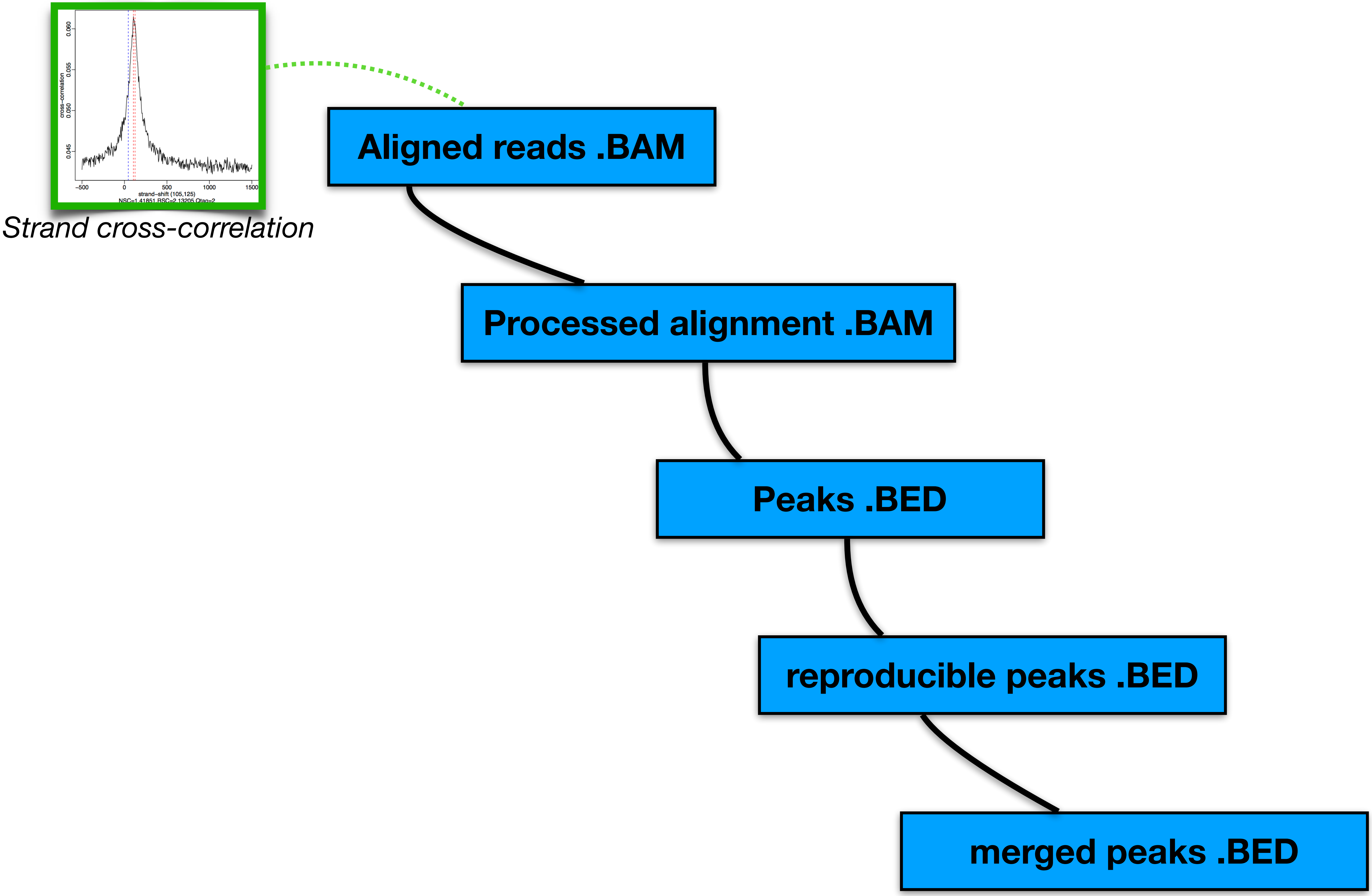
Aligned reads .BAM

Processed alignment .BAM

Peaks .BED

reproducible peaks .BED

merged peaks .BED



Strand cross-correlation

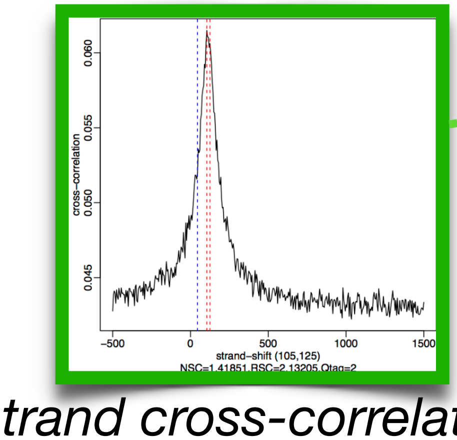
Aligned reads .BAM

Processed alignment .BAM

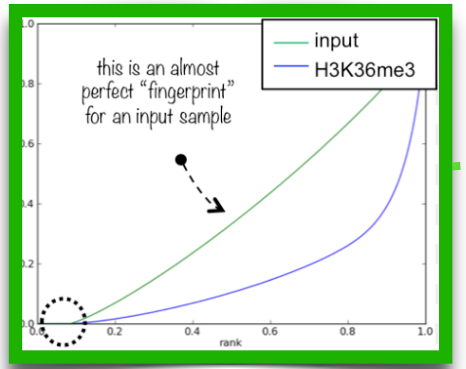
Peaks .BED

reproducible peaks .BED

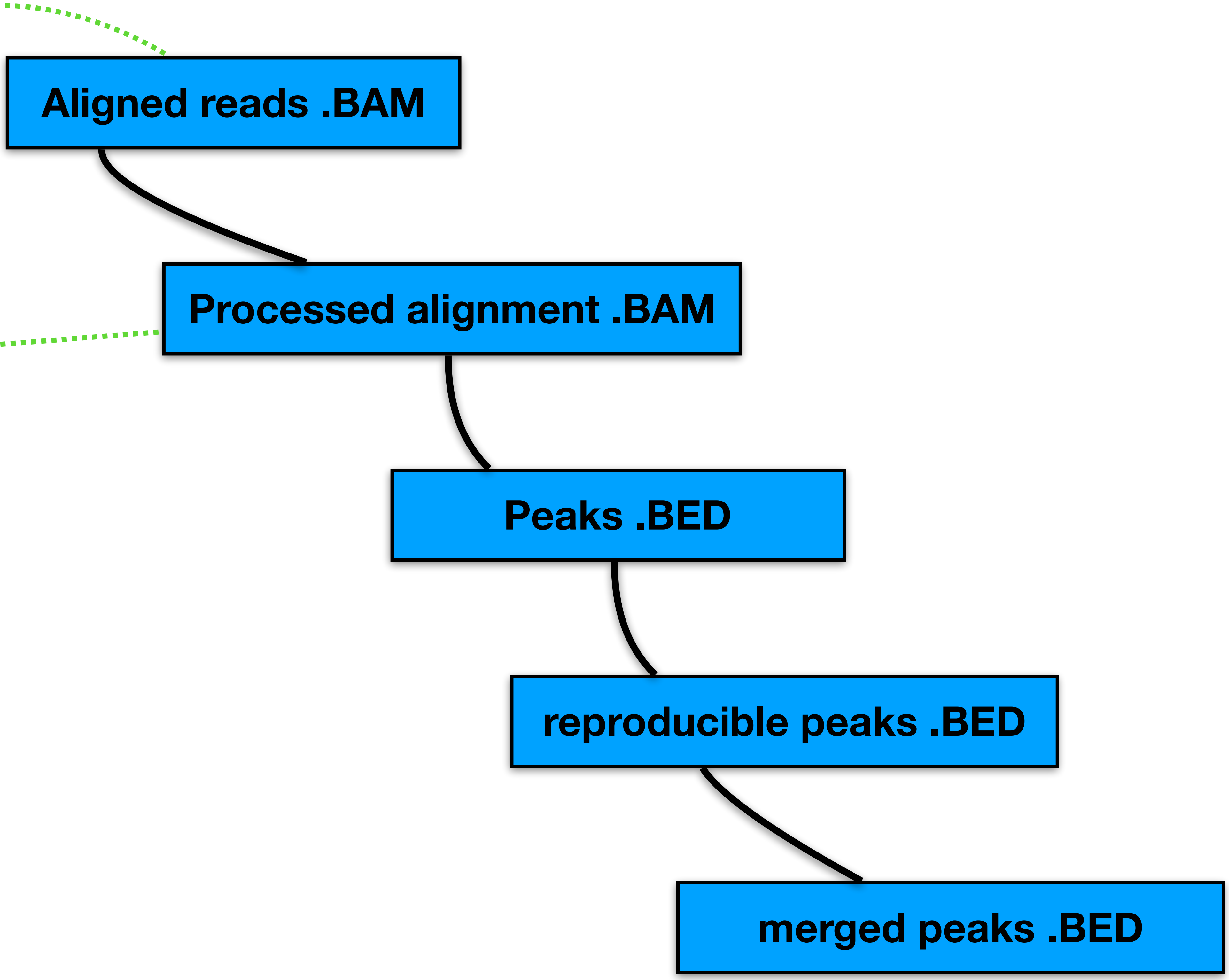
merged peaks .BED



Strand cross-correlation



Cumulative enrichment



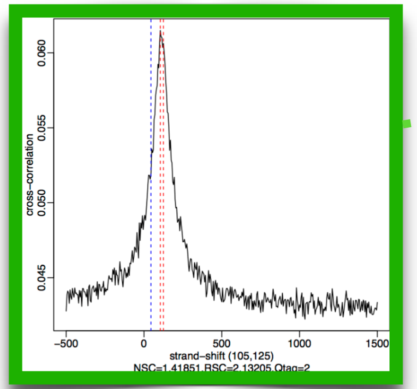
Aligned reads .BAM

Processed alignment .BAM

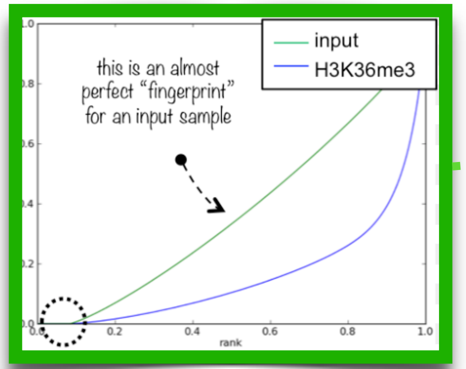
Peaks .BED

reproducible peaks .BED

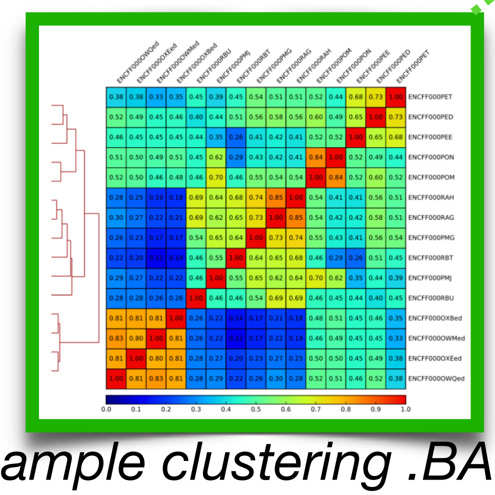
merged peaks .BED



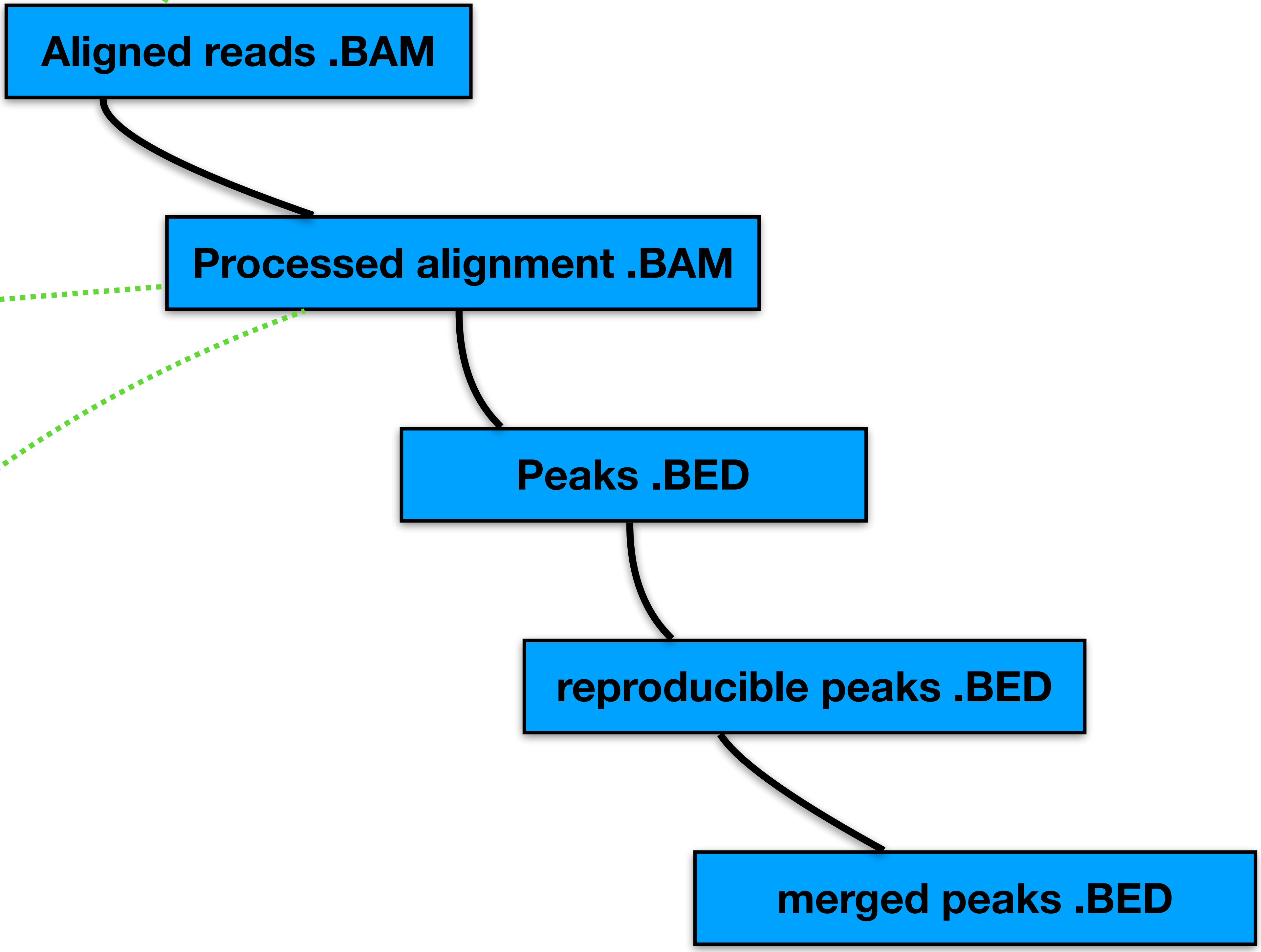
Strand cross-correlation



Cumulative enrichment



Sample clustering .BAM



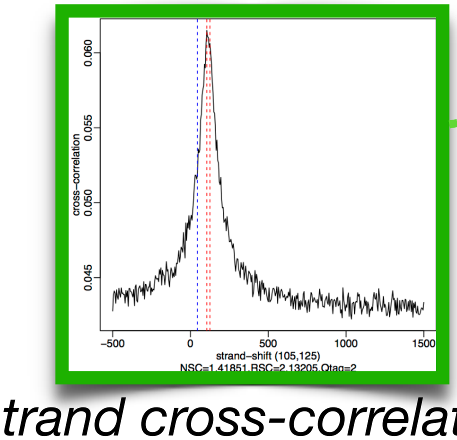
Aligned reads .BAM

Processed alignment .BAM

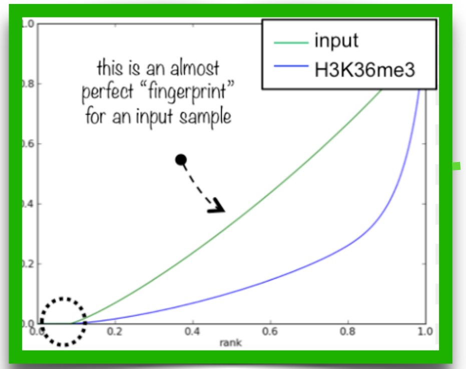
Peaks .BED

reproducible peaks .BED

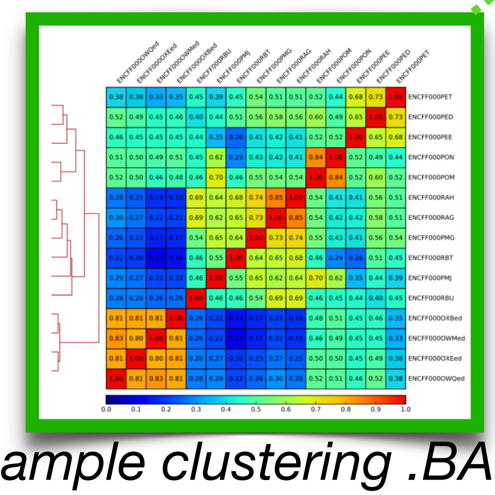
merged peaks .BED



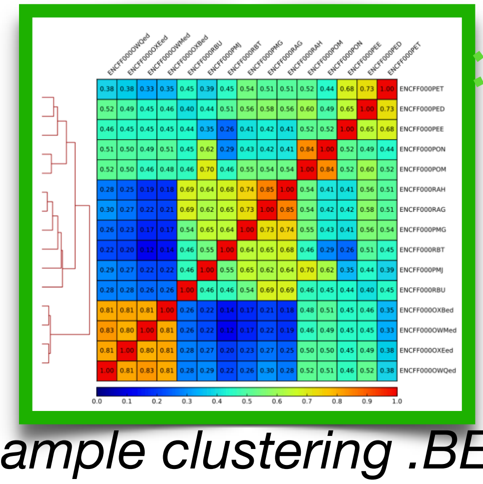
Strand cross-correlation



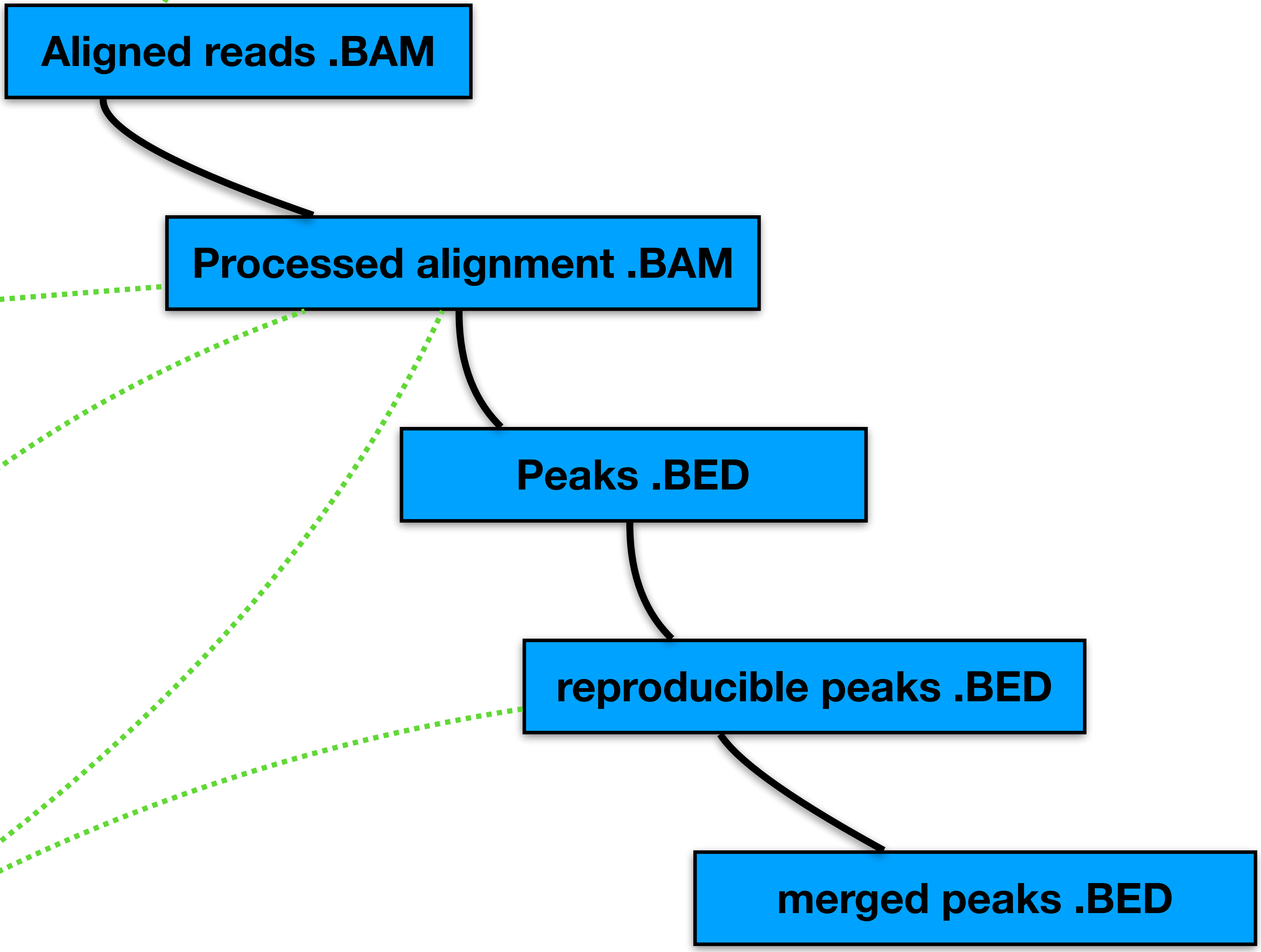
Cumulative enrichment

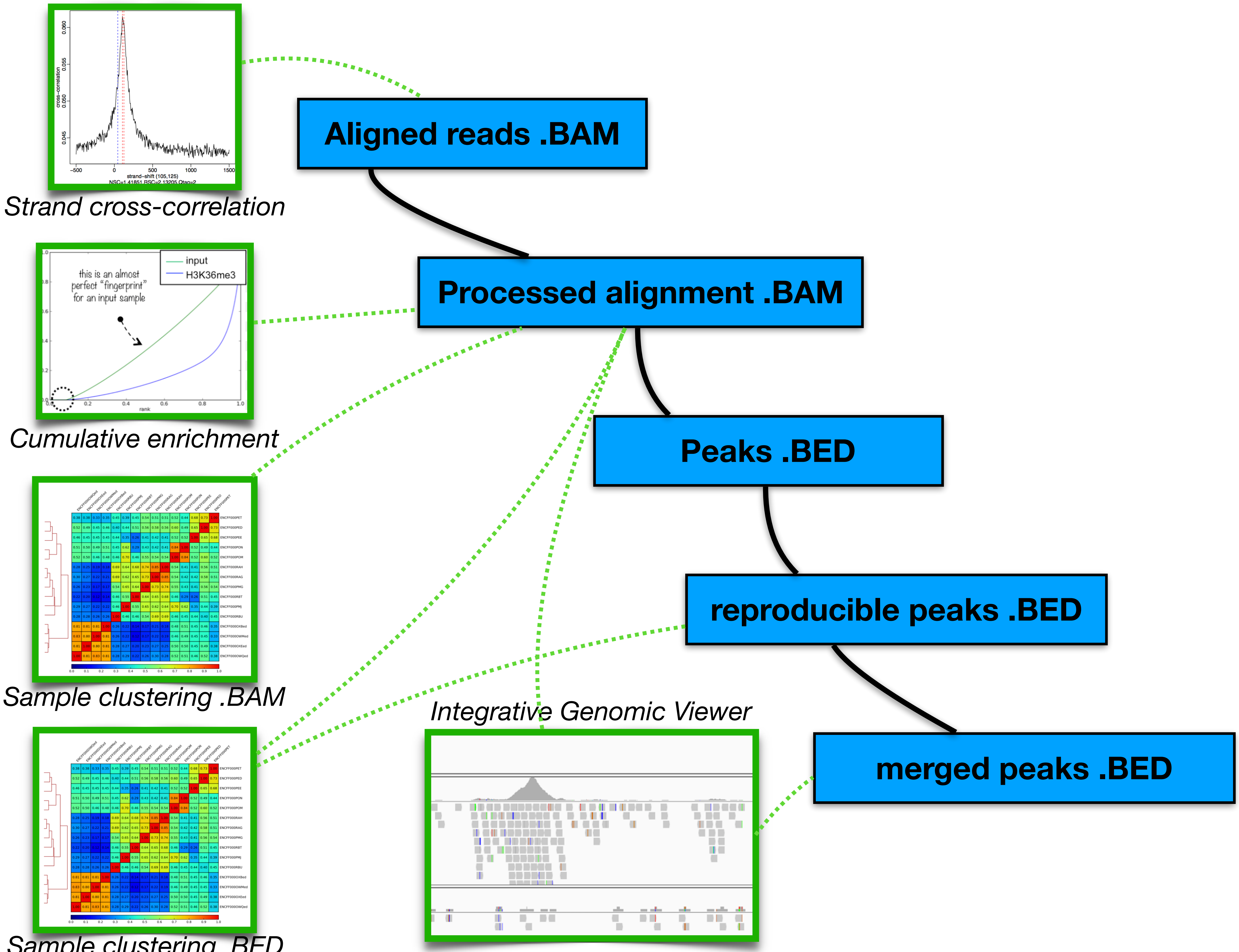


Sample clustering .BAM



Sample clustering .BED





Aligned reads .BAM

Processed alignment .BAM

Peaks .BED

reproducible peaks .BED

merged peaks .BED

Strand cross-correlation

Cumulative enrichment

Sample clustering .BAM

Sample clustering .BED

Integrative Genomic Viewer

Aligned reads .BAM

Processed alignment .BAM

Peaks .BED

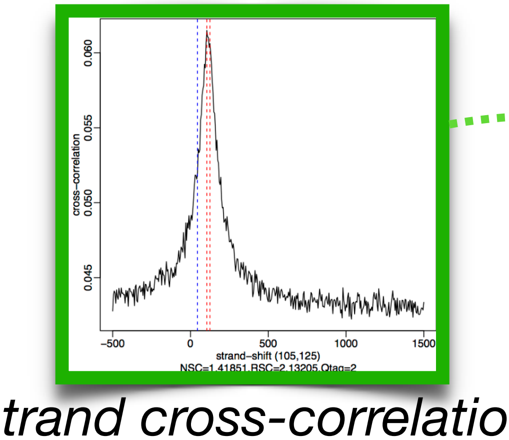
reproducible peaks .BED

merged peaks .BED

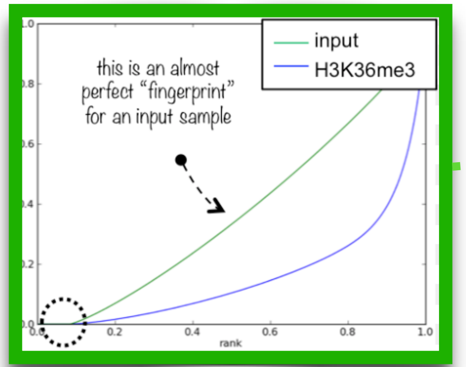
duplicated reads

black listed regions

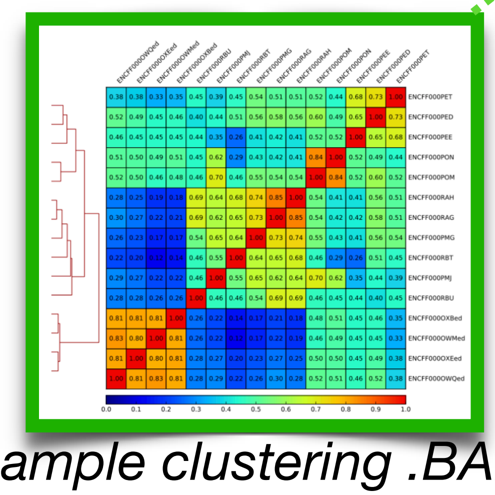
1 x read coverage



Strand cross-correlation

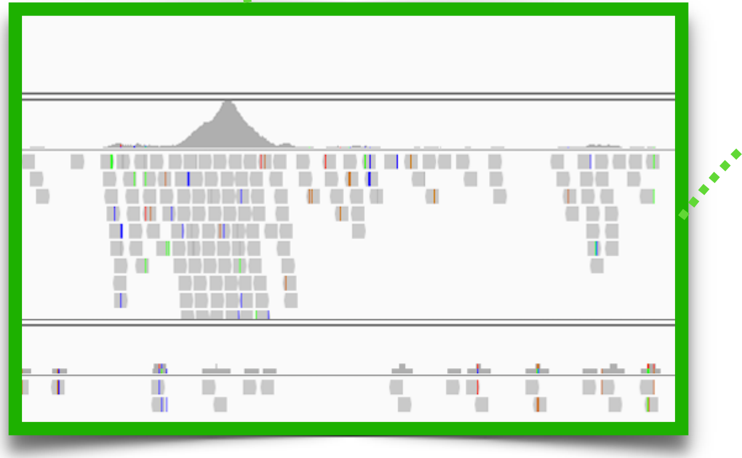


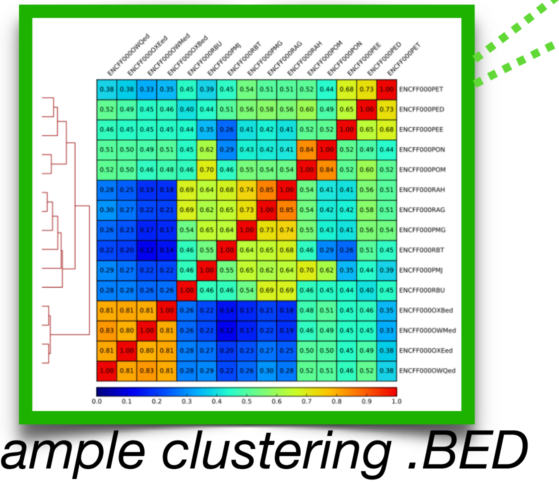
Cumulative enrichment



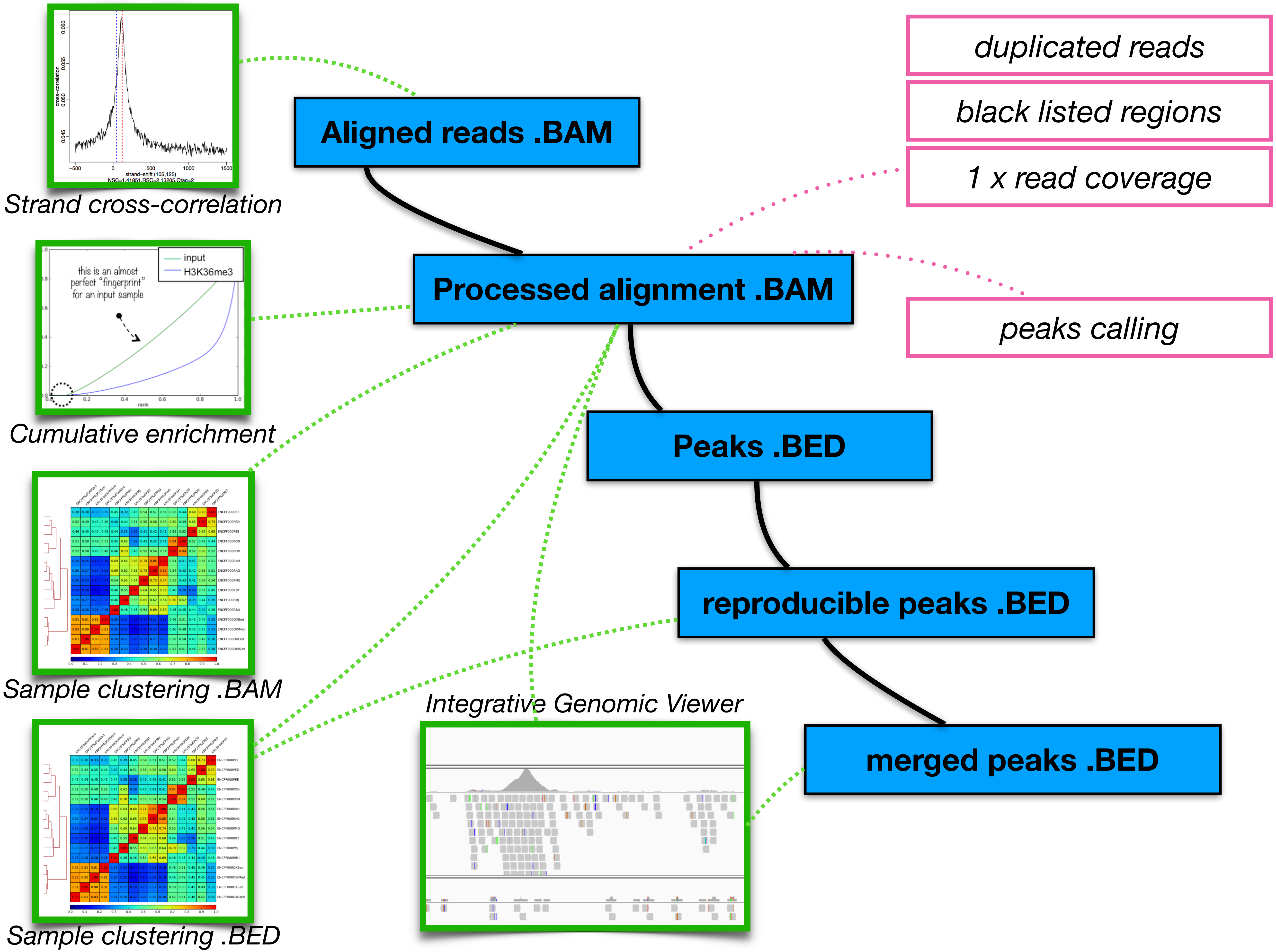
Sample clustering .BAM

Integrative Genomic Viewer





Sample clustering .BED



Aligned reads .BAM

Processed alignment .BAM

Peaks .BED

reproducible peaks .BED

merged peaks .BED

duplicated reads

black listed regions

1 x read coverage

peaks calling

Strand cross-correlation

Cumulative enrichment

Sample clustering .BAM

Sample clustering .BED

Integrative Genomic Viewer

Aligned reads .BAM

Processed alignment .BAM

Peaks .BED

reproducible peaks .BED

merged peaks .BED

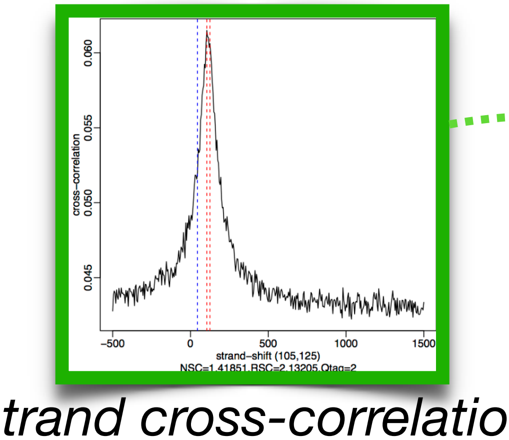
duplicated reads

black listed regions

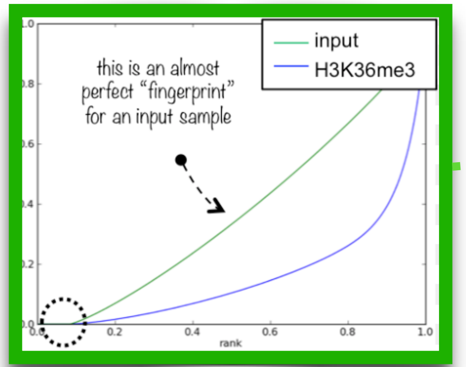
1 x read coverage

peaks calling

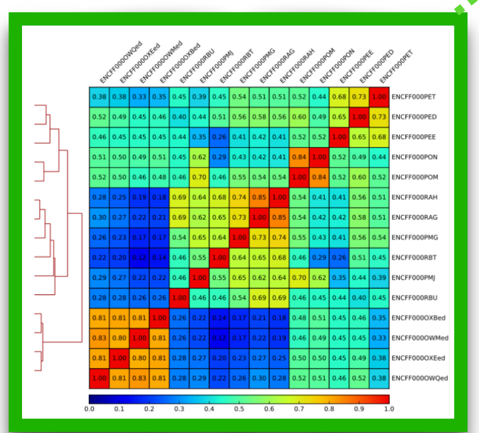
peaks intersecting



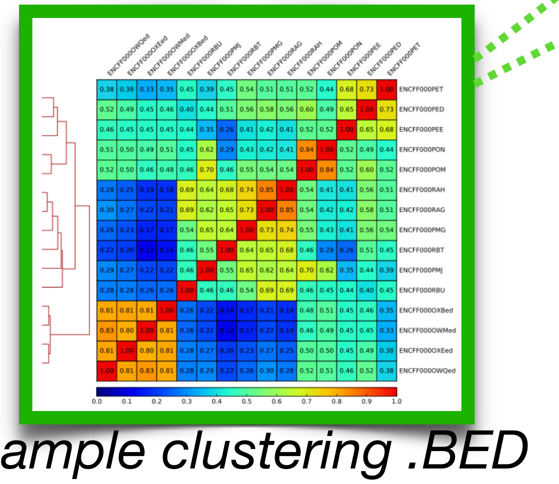
Strand cross-correlation



Cumulative enrichment

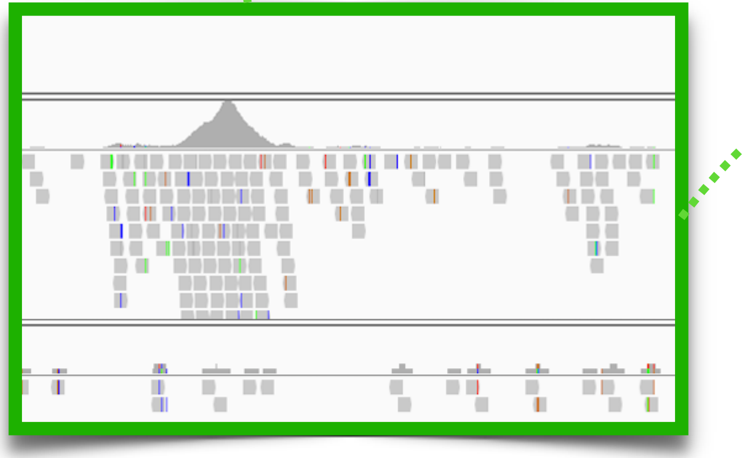


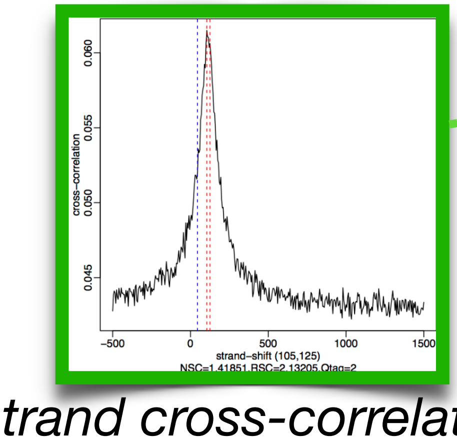
Sample clustering .BAM



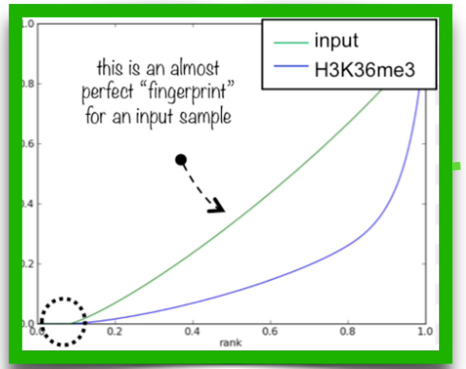
Sample clustering .BED

Integrative Genomic Viewer

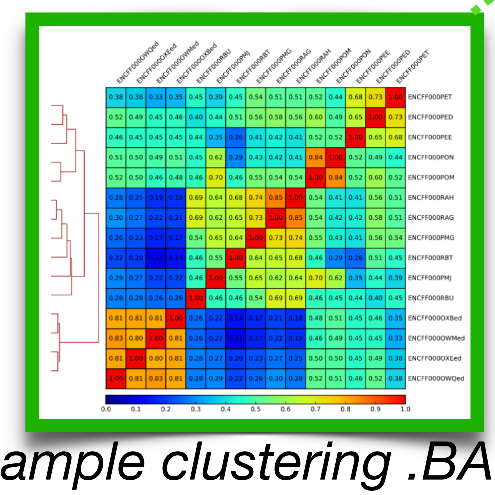




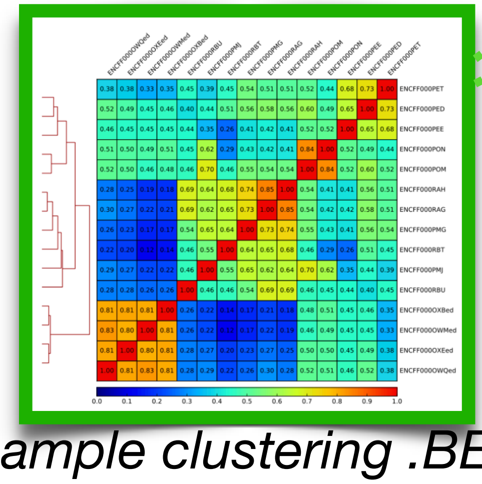
Strand cross-correlation



Cumulative enrichment

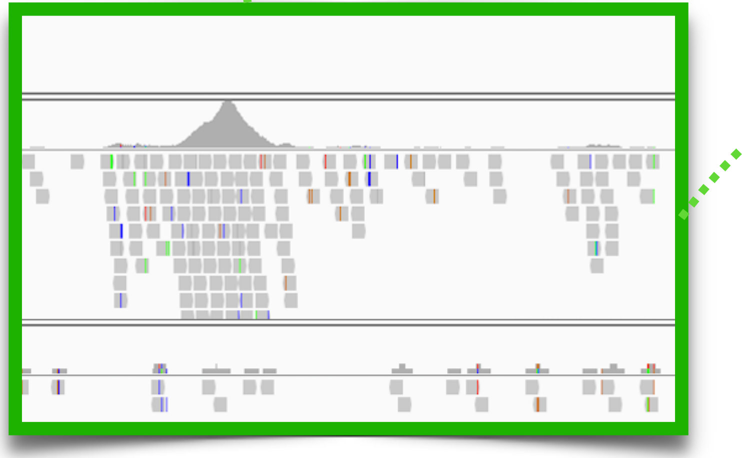


Sample clustering .BAM



Sample clustering .BED

Integrative Genomic Viewer



Aligned reads .BAM

duplicated reads

black listed regions

1 x read coverage

Processed alignment .BAM

peaks calling

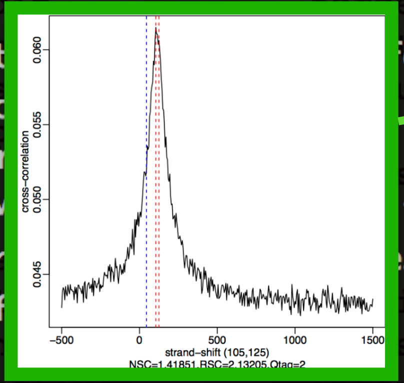
Peaks .BED

peaks intersecting

reproducible peaks .BED

peaks merging

merged peaks .BED



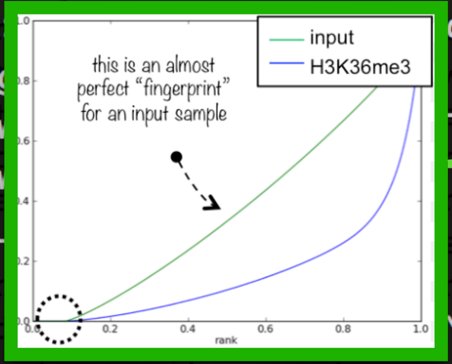
cross correlation

Aligned reads .BAM

duplicated reads

black listed regions

1 x read coverage

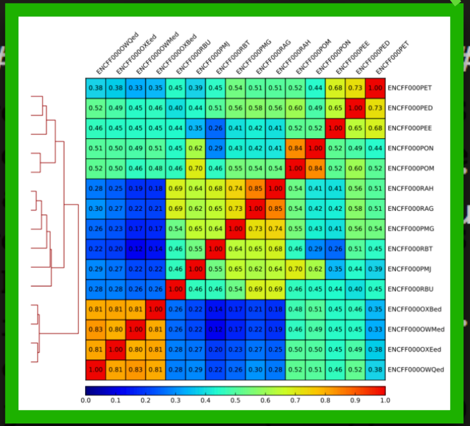


Cumulative enrichment

Processed alignment .BAM

peaks calling

Peaks .BED



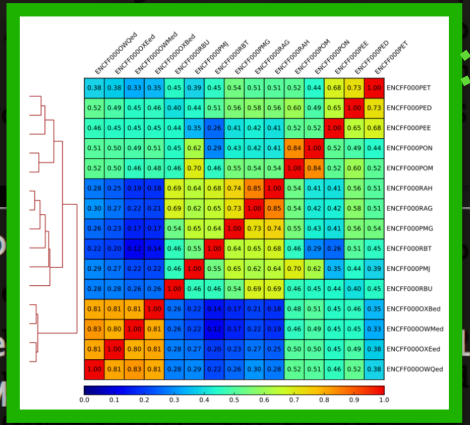
Sample clustering .BAM

peaks intersecting

reproducible peaks .BED

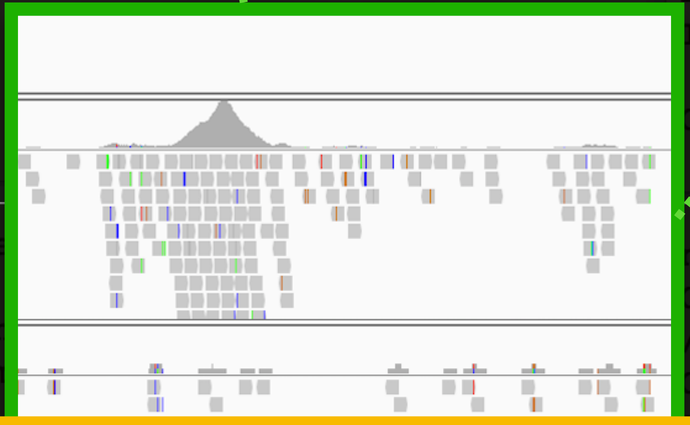
peaks merging

merged peaks .BED



Sample clustering .BED

Integrative Genomic Viewer



Using computational resources

We have booked half a node on Rackham per course participant. To run the tutorial in the interactive mode log to Rackham and run *interactive* command.

```
ssh -Y username@rackham.uppmax.uu.se
interactive -A g2018030 -p core -n 4 --reservation=g2018030_WED
interactive -A g2018030 -p core -n 4 --reservation=g2018030_THU (on Thursday)
interactive -A g2018030 -p core -n 4 --reservation=g2018030_FRI (on Friday)
```

Check which node you were assigned

```
$ squeue -u <username>
```

And connect to your node with

```
ssh -Y <nodename>
```

Files structure

There are many files which are part of the data set as well as there are additional files with annotations that are required to run various steps in this tutorial. Therefore saving files in a structured manner is essential to keep track of the analysis steps (and always a good practice). We have preset data access and environment for your. To use these settings run:

- `chiseq_data.sh` that sets up directory structure and creates symbolic links to data as well as copies smaller files **[RUN ONLY ONCE]**
- `chipseq_env.sh` that sets several environmental variables you will use in the exercise: **[RUN EVERY TIME when the connection to Uppmax has been broken, i.e. via logging out]**

Copy the scripts to your home directory and execute them:

```
cp /sw/share/compstore/courses/ngsintro/chipseq/scripts/setup/chipseq_data.sh ./
cp /sw/share/compstore/courses/ngsintro/chipseq/scripts/setup/chipseq_env.sh ./

source chipseq_data.sh
source chipseq_env.sh
```

You should see a directory named “chipseq”

```
ls ~
cd ~/chipseq/analysis
```