MRC | Cancer Unit

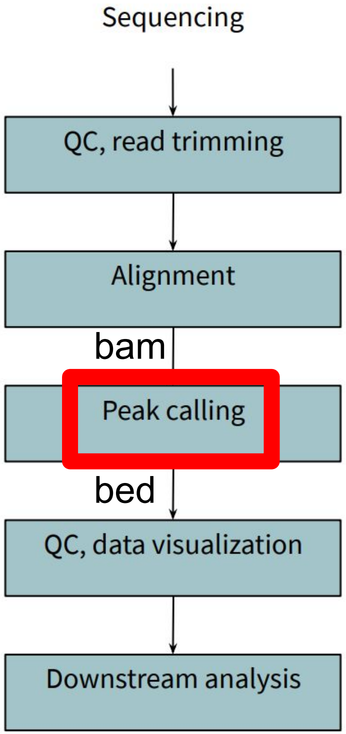
UNIVERSITY OF CAMBRIDGE

Peak Calling

Shoko Hirosue
MRC Cancer Unit, University of Cambridge

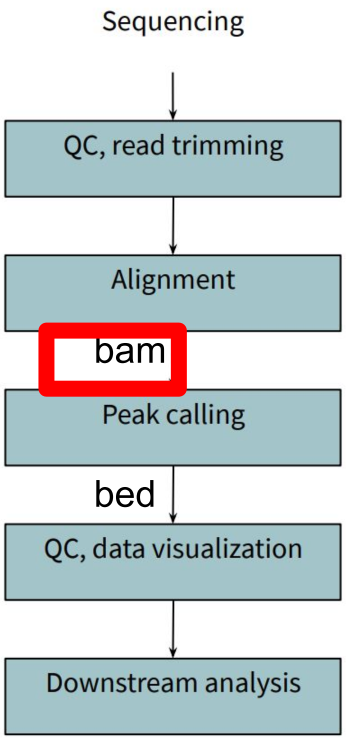
CRUK CI Bioinformatics Summer School July 2020

Peak calling



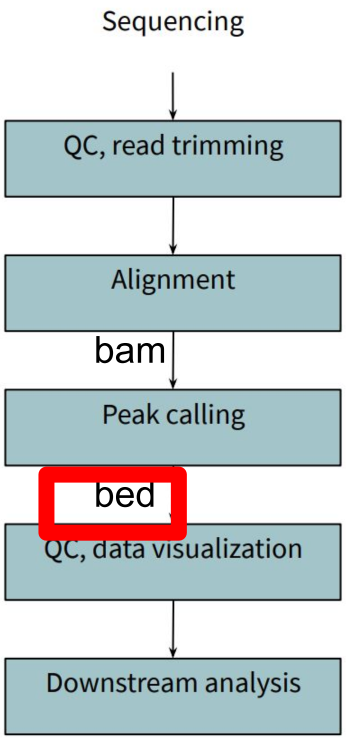
Adapted from Dora Bihary’s slides

Peak calling



Adapted from Dora Bihary’s slides

Peak calling



Sequencing

QC, read trimming

Alignment

bam

Peak calling

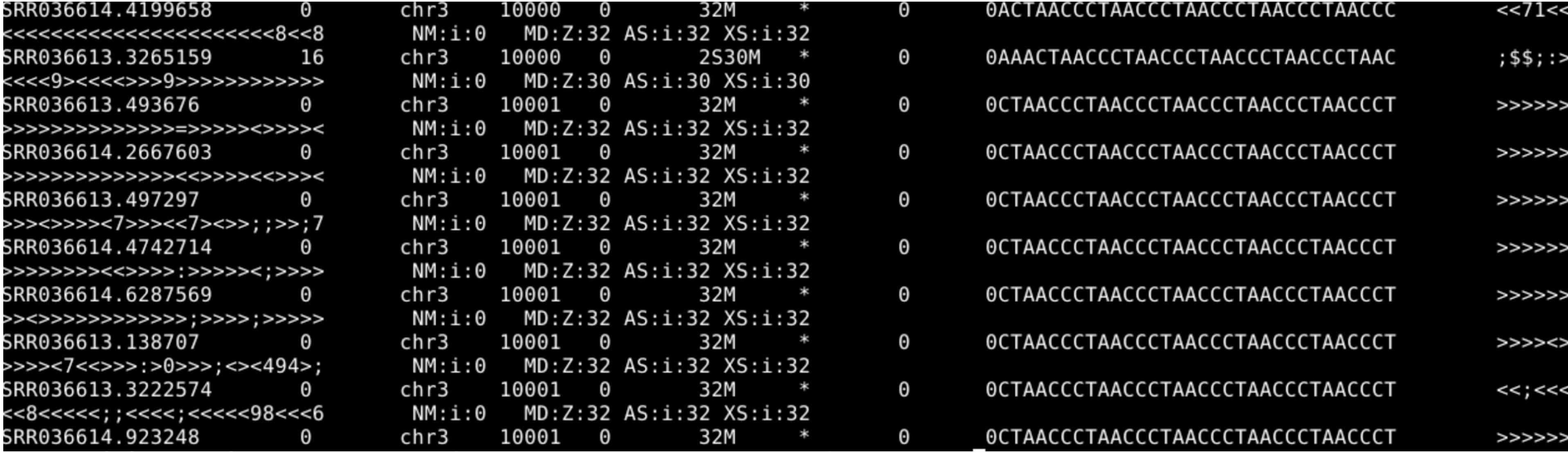
bed

QC, data visualization

Downstream analysis

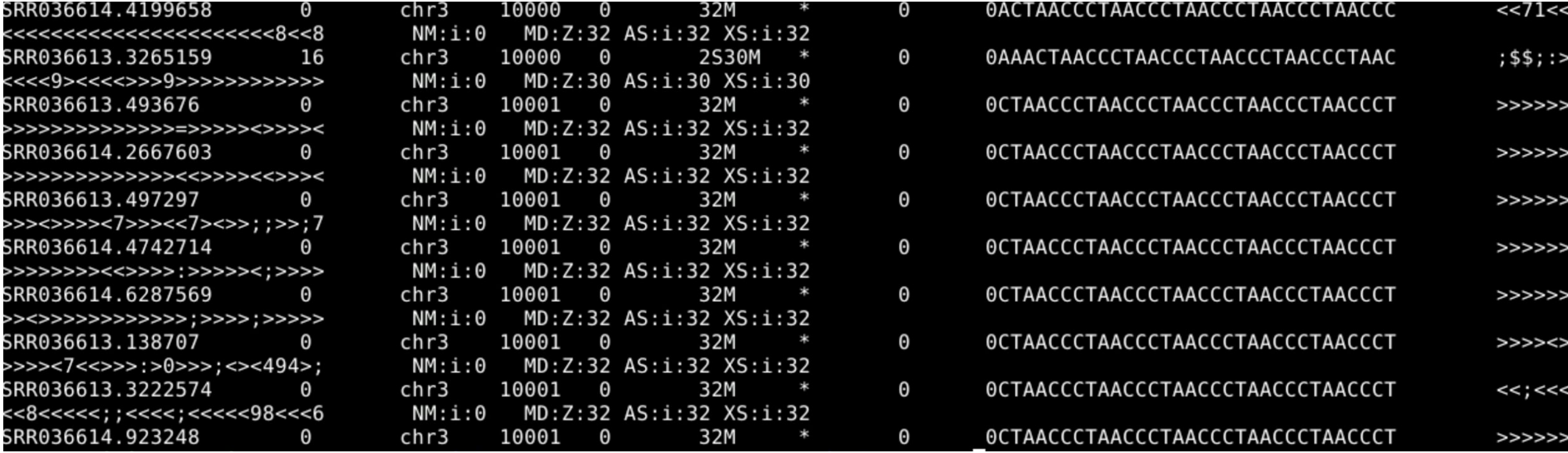
Adapted from Dora Bihary’s slides

Bam file (reads)

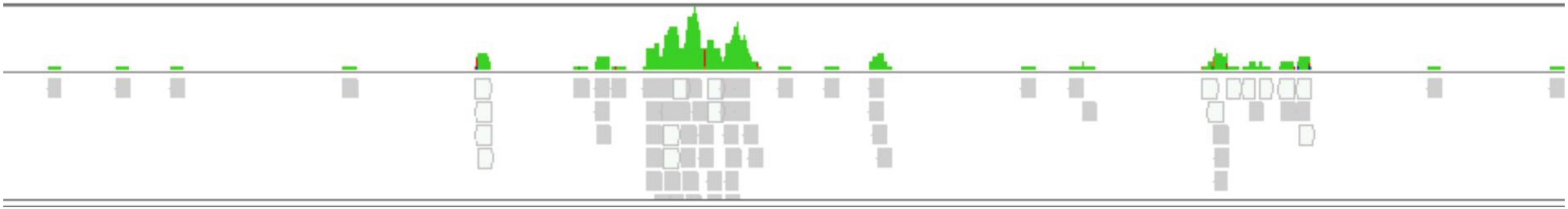


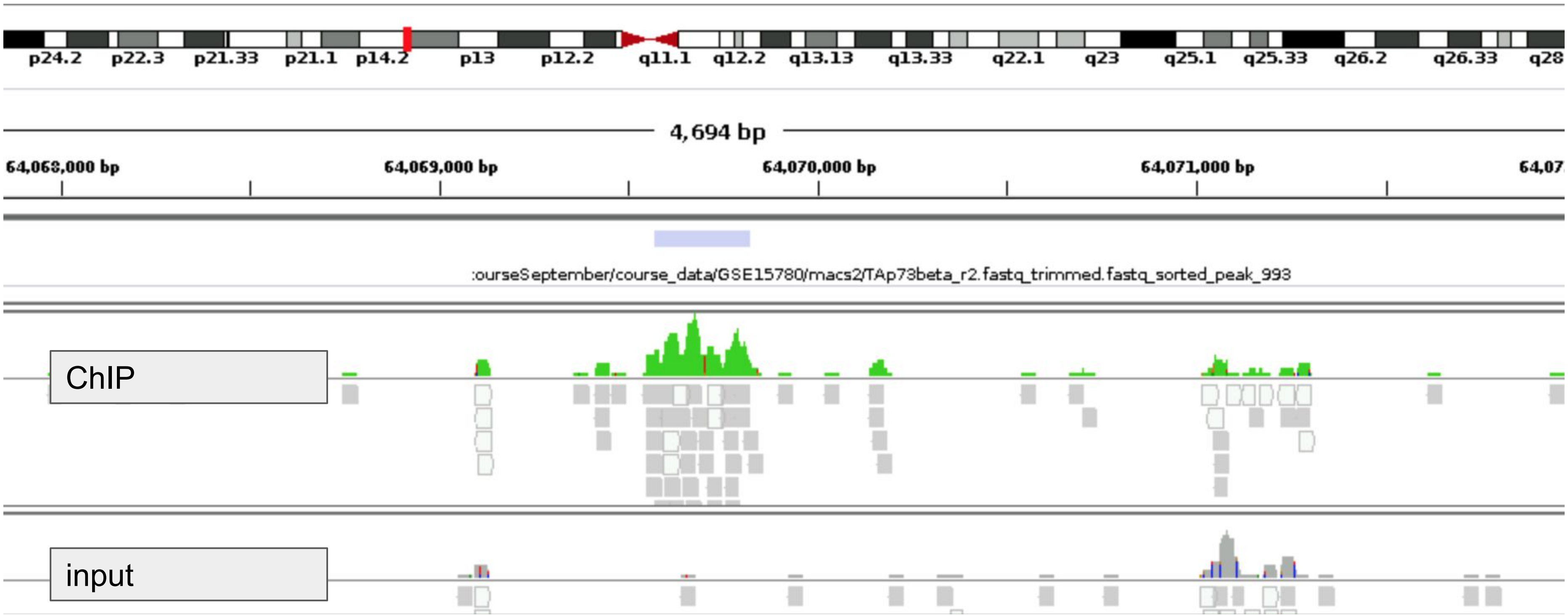
```
SRR036614.4199658       0       chr3    10000   0       32M     *       0       0ACTAACCCTAACCCTAACCCTAACCCTAACCC       <<71<<
<<<<<<<<<<<<<<<<<<<<<<8<<8       NM:i:0   MD:Z:32 AS:i:32 XS:i:32
SRR036613.3265159       16      chr3    10000   0       2S30M   *       0       0AAACTAACCCTAACCCTAACCCTAACCCTAAC       ;$$;:>
<<<<9><<<<>>>9>>>>>>>>>>>>       NM:i:0   MD:Z:30 AS:i:30 XS:i:30
SRR036613.493676        0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>>>>>>>>>>>>>>=>>>>><>>>><       NM:i:0   MD:Z:32 AS:i:32 XS:i:32
SRR036614.2667603       0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>>>>>>>>>>>>>><<>>>><<>>><       NM:i:0   MD:Z:32 AS:i:32 XS:i:32
SRR036613.497297        0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>>><>>>><7>>><<7><>>;;>>;7       NM:i:0   MD:Z:32 AS:i:32 XS:i:32
SRR036614.4742714       0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>>>>>>>><<>>>>:>>>>><;>>>>       NM:i:0   MD:Z:32 AS:i:32 XS:i:32
SRR036614.6287569       0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>><>>>>>>>>>>>>;>>>>;>>>>>       NM:i:0   MD:Z:32 AS:i:32 XS:i:32
SRR036613.138707        0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>><>
>>>><7<<>>>:>0>>>;<><494>;       NM:i:0   MD:Z:32 AS:i:32 XS:i:32
SRR036613.3222574       0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       <<;<<<
<<8<<<<<;;<<<<;<<<<<98<<<6       NM:i:0   MD:Z:32 AS:i:32 XS:i:32
SRR036614.923248        0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
```

Bam file (reads)



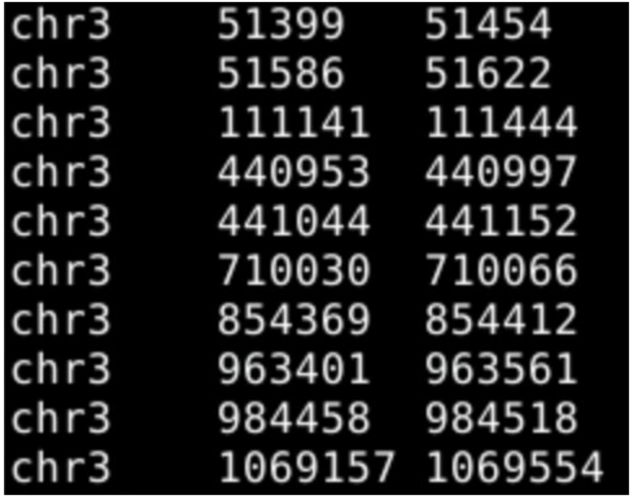
```
SRR036614.4199658       0       chr3    10000   0       32M     *       0       0ACTAACCCTAACCCTAACCCTAACCCTAACCC       <<71<<
<<<<<<<<<<<<<<<<<<<<<<<8<<8      NM:i:0  MD:Z:32 AS:i:32 XS:i:32
SRR036613.3265159       16      chr3    10000   0       2S30M   *       0       0AAACTAACCCTAACCCTAACCCTAACCCTAAC       ;$$;:>
<<<<9><<<<>>>9>>>>>>>>>>>>>      NM:i:0  MD:Z:30 AS:i:30 XS:i:30
SRR036613.493676        0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>>>>>>>>>>>>>>>=>>>>>><>>>><      NM:i:0  MD:Z:32 AS:i:32 XS:i:32
SRR036614.2667603       0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>>>>>>>>>>>>>><<>>>>><<>>><      NM:i:0  MD:Z:32 AS:i:32 XS:i:32
SRR036613.497297        0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>>><>>>><7>>><<7><>>;;>>;7      NM:i:0  MD:Z:32 AS:i:32 XS:i:32
SRR036614.4742714       0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>>>>>>>><<>>>>:>>>>>><;>>>>      NM:i:0  MD:Z:32 AS:i:32 XS:i:32
SRR036614.6287569       0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
>><>>>>>>>>>>>>;>>>>;>>>>>      NM:i:0  MD:Z:32 AS:i:32 XS:i:32
SRR036613.138707        0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>><>
>>>><7<<>>>:>0>>>;<><494>;      NM:i:0  MD:Z:32 AS:i:32 XS:i:32
SRR036613.3222574       0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       <<;<<<
<<8<<<<<;;<<<<;<<<<<98<<<6      NM:i:0  MD:Z:32 AS:i:32 XS:i:32
SRR036614.923248        0       chr3    10001   0       32M     *       0       0CTAACCCTAACCCTAACCCTAACCCTAACCCT       >>>>>>
```





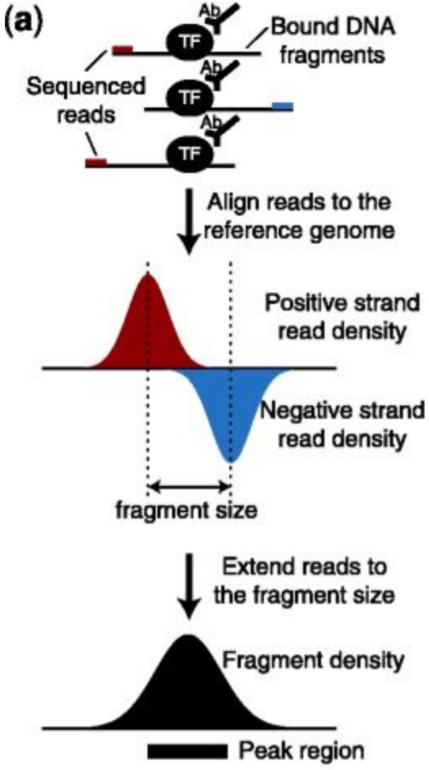
Adapted from Dora Bihary's slides

Bed file (peaks)



```
chr3    51399   51454
chr3    51586   51622
chr3    111141  111444
chr3    440953  440997
chr3    441044  441152
chr3    710030  710066
chr3    854369  854412
chr3    963401  963561
chr3    984458  984518
chr3    1069157 1069554
```

Reads to peaks

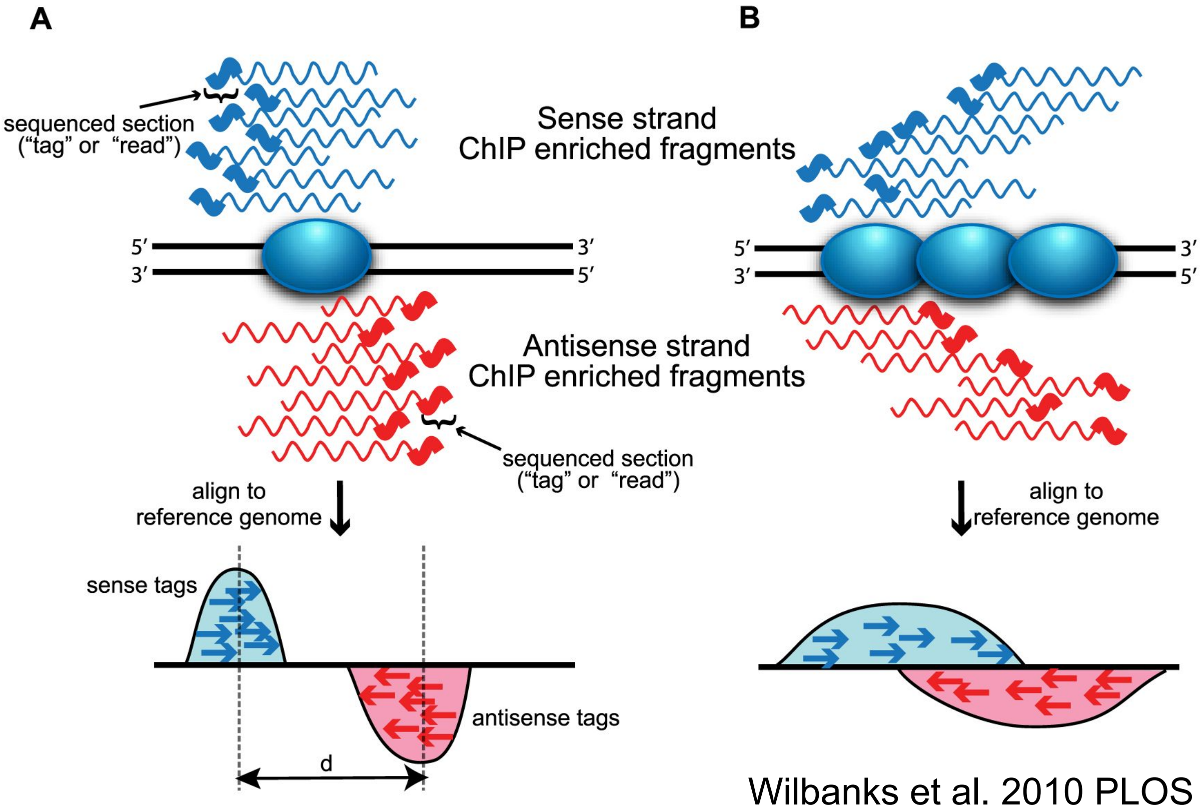

(a)
Ab
TF
Bound DNA fragments
Sequenced reads
Align reads to the reference genome
Positive strand read density
Negative strand read density
fragment size
Extend reads to the fragment size
Fragment density
Peak region

Bardet et al. Bioinformatics, 2013.

+ive and -ive strand reads do not represent true binding sites (Strand dependent bimodality)

Fragment length d needs to be estimated (if not known) from strand asymmetry in data

Difference in peak shapes

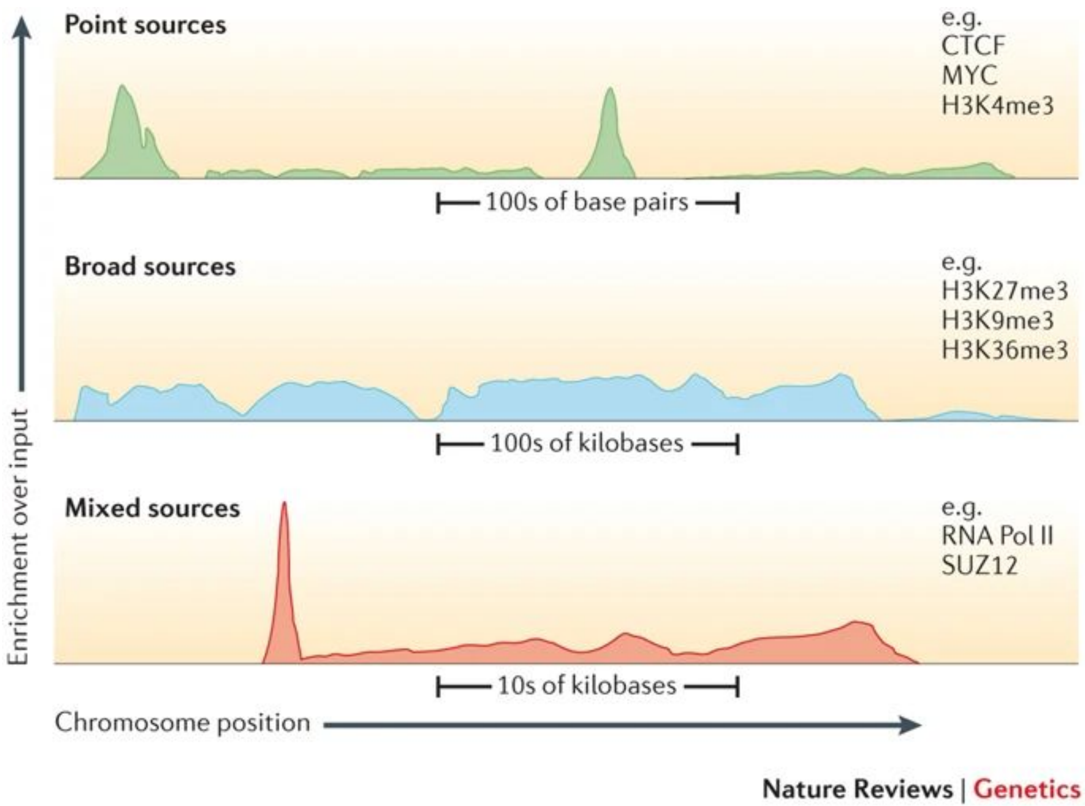


Wilbanks et al. 2010 PLOS One

A. For sequence-specific binding events the signal is sharp and shows strong strand dependent bimodality.

B. Distributed binding events produce a broader pattern. For most histone marks the signal is expected to be broad with less defined bimodal pattern.

Difference in peak shapes



Nature Reviews | Genetics

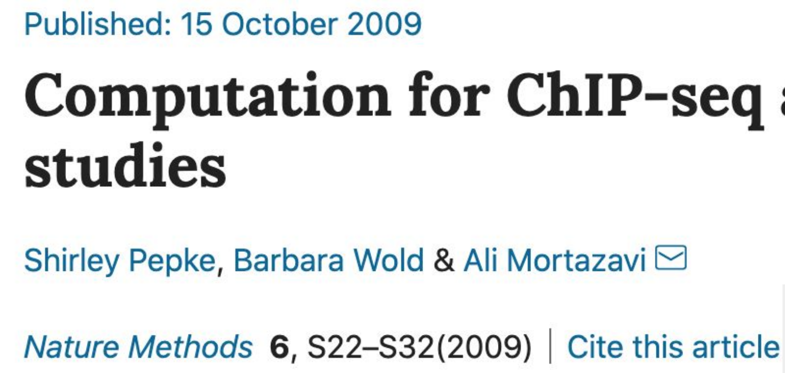
- Most TF peaks are narrow

- ChIP-seq peaks from epigenomic data can be narrow, broad or gapped. Histone marks such as H3K9me3 or H3K27me3 are broad while others such as H3K4me3 and proteins such as CTCF are narrow

- Other DNA binding proteins such as HP1 , Lamins (Lamin A or B), HMGA etc. form broad peaks or domains.

- PolII peaks can be narrow or broad depending on whether its detecting transcription initiation at the TSS or propagation along the gene body.

Sims et al., 2014 Nat Rev Genet.

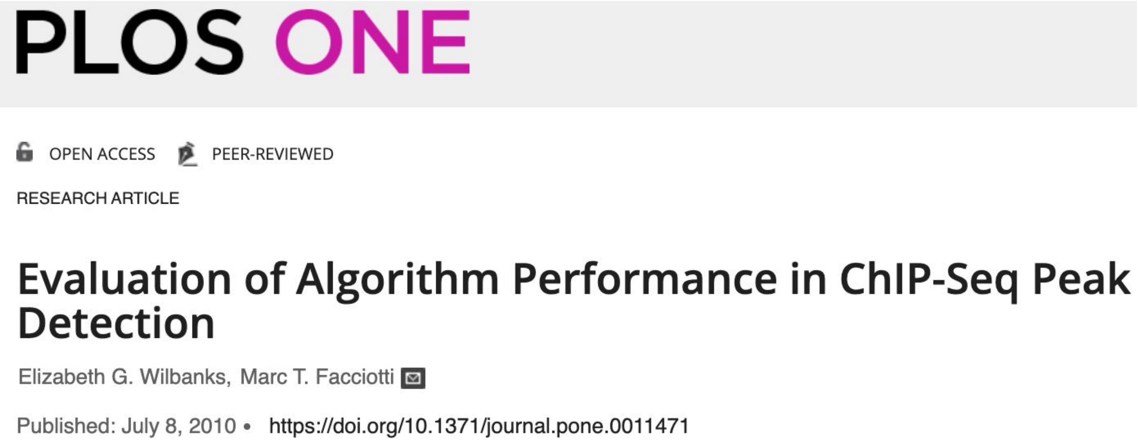


Published: 15 October 2009

Computation for ChIP-seq and RNA-seq studies

Shirley Pepke, Barbara Wold & Ali Mortazavi

Nature Methods **6**, S22–S32(2009) | Cite this article



PLOS ONE

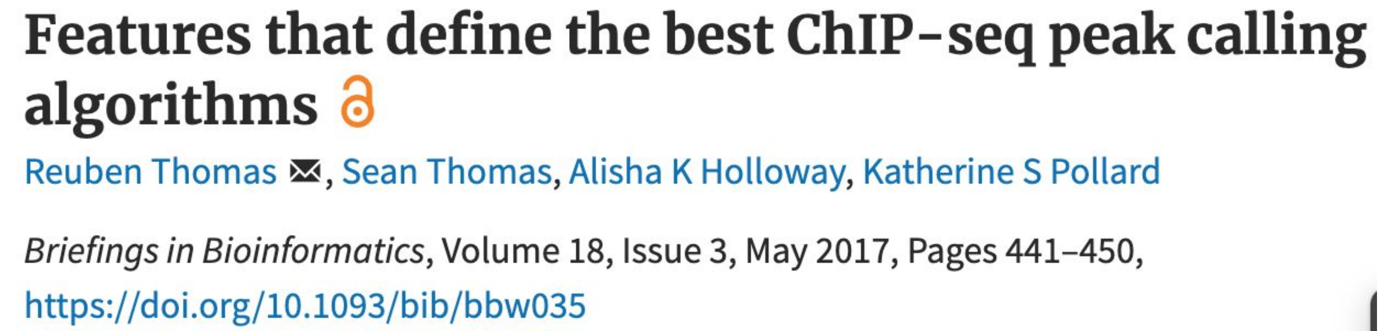
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Evaluation of Algorithm Performance in ChIP-Seq Peak Detection

Elizabeth G. Wilbanks, Marc T. Facciotti

Published: July 8, 2010 • https://doi.org/10.1371/journal.pone.0011471



Features that define the best ChIP-seq peak calling algorithms

Reuben Thomas, Sean Thomas, Alisha K Holloway, Katherine S Pollard

Briefings in Bioinformatics, Volume 18, Issue 3, May 2017, Pages 441–450,

https://doi.org/10.1093/bib/bbw035

Peak calling software: MACS2

- **M**odel-based **A**nalysis for **C**hIP **S**eq
- Most widely used peak caller
- It was developed for TF bindings, but also suitable for broader regions

Useful tutorials:

- MACS project github page
  https://github.com/macs3-project/MACS/wiki/Advanced%3A-Call-peaks-using-MACS2-subcommands
- Introduction to ChIP-seq using high performance computing
  https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

Step1: Filter duplicates

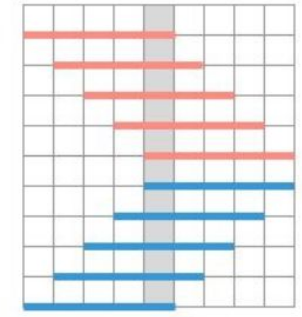
‘MACS2 filterdup’

Duplicate reads: reads at the same coordination on the same strand

Duplicates - What do we do with them?

- Duplicates can be artefacts:
  - PCR bias: certain genomic regions are preferentially amplified
  - Low initial starting material can introduce artificially enriched regions with overamplification
- Duplicates can also be “legitimate”:
  - It is unavoidable in highly enriched experiments and deeply sequenced ChIPs since it is naturally increasing with the sequencing depth
- Removing duplicates limits the dynamic range of ChIP signal:
  - Maximum signal/base: one fragment on each strand in each possible position of the read

$$Signal_{max} = 2*readLength$$



Adapted from Dora Bihary’s slides

Duplicates - What do we do with them?

Some approaches:

- Remove all duplicates
- Don’t remove duplicates as long as it has a reasonable rate
- Remove duplicates for some analysis:
  - Remove duplicates before peak-calling
  - Keep duplicates for differential binding analysis
- htSeqTools:
  - Estimate duplicate numbers expected taking into account the sequencing depth and using negative binomial model
  - Attempt to identify significantly outstanding duplicate numbers

Adapted from Dora Bihary’s slides

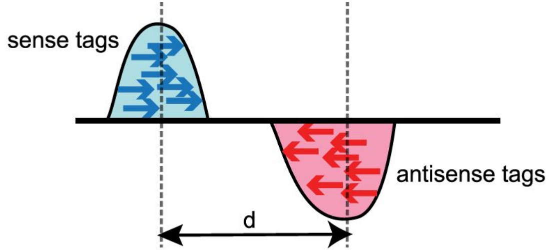
Duplicates - What do we do with them?

Some approaches:

- Remove all duplicates
- Don’t remove duplicates as long as it has a reasonable rate
- Remove duplicates for some analysis:
  - Remove duplicates before peak-calling
  - Keep duplicates for differential binding analysis
- htSeqTools:
  - Estimate duplicate numbers expected taking into account the sequencing depth and using negative binomial model
  - Attempt to identify significantly outstanding duplicate numbers

Adapted from Dora Bihary’s slides

Step2: Decide the fragment length d



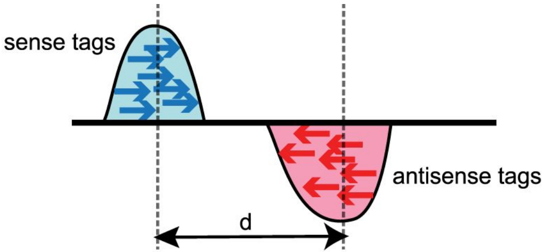
‘MACS2 predictd’

Find treatment regions more than ‘--mfold’ enriched relative to the background

MACS randomly samples 1,000 of these high-quality peaks, separates their positive and negative strand reads, and aligns them by the midpoint between their centers.

The distance between the two peaks in the alignment (d) is the estimated fragment length.

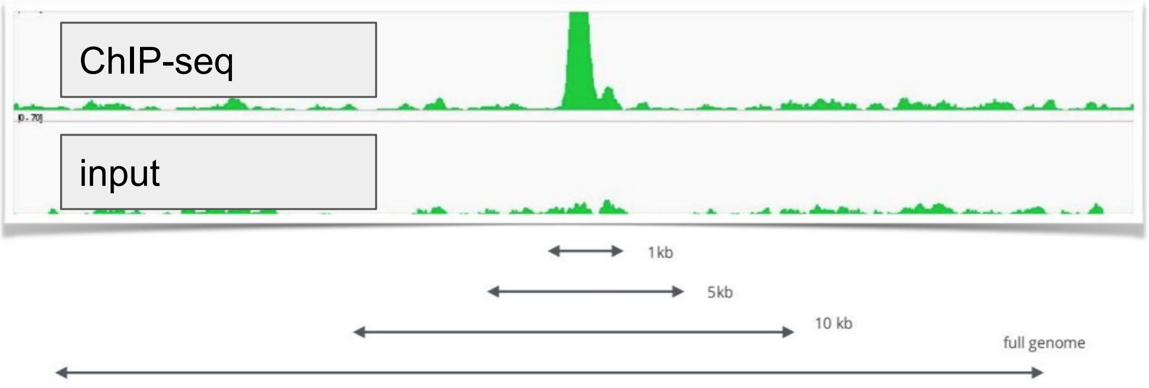
Step3: Extend ChIP sample



Extend reads by d (fragment length) in 5’ to 3’ direction

Step4: Build local bias track from control

λ is the expected number of reads in that window. (parameter of Poisson distribution)

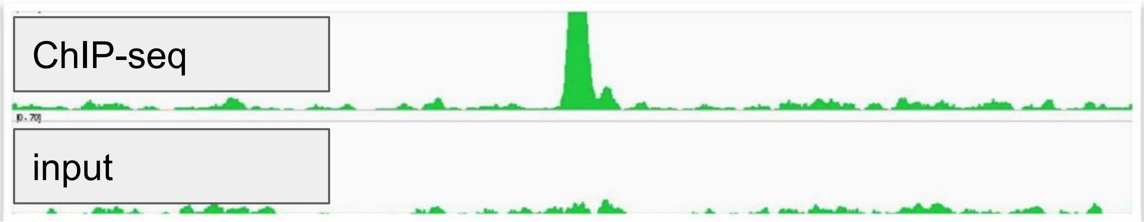


estimate parameter λ_{local} over different ranges, take max.

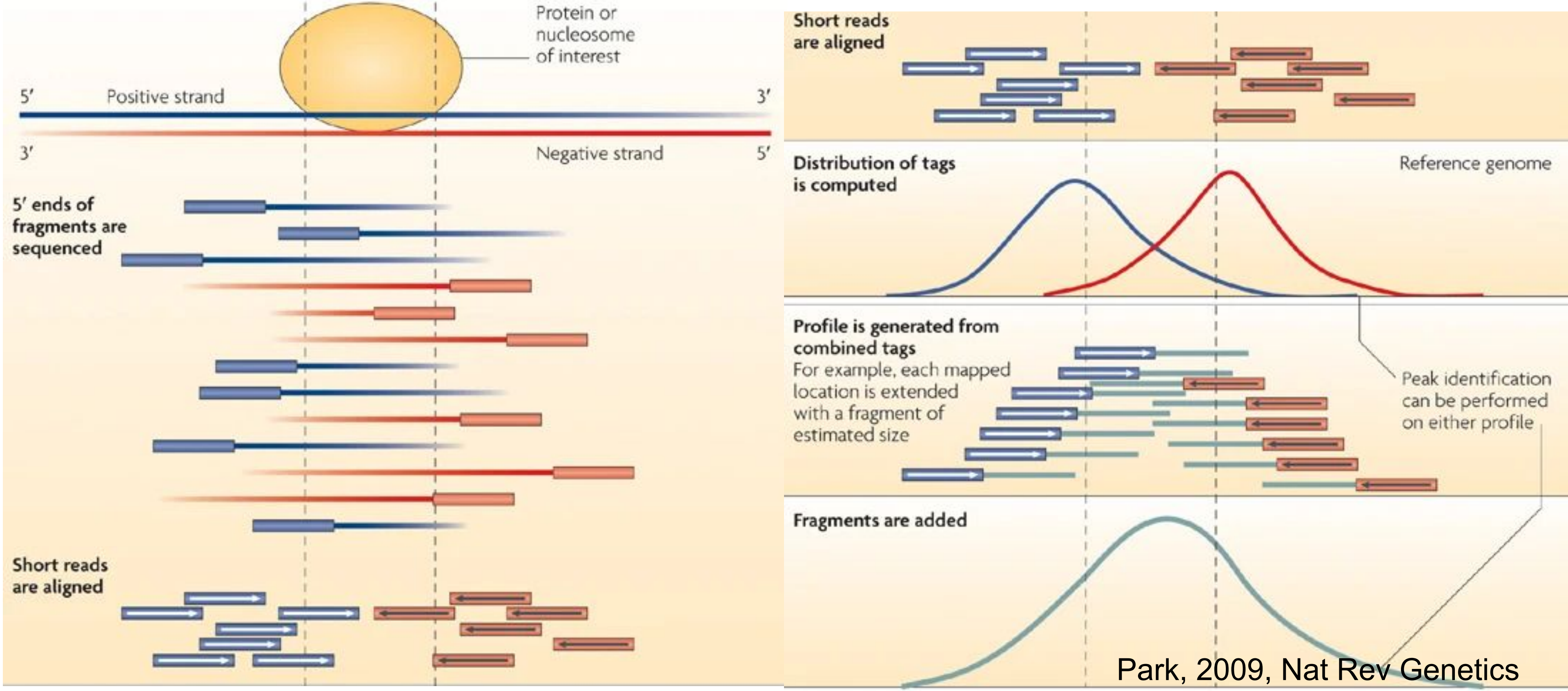
Adapted from Shamith Samarajiwa's slides

Step5: Identify enriched peak regions

1. Scale the ChIP and control to the same sequencing depth
2. Determine regions with ‘--pvalue’ threshold (Poisson distribution p-value based on **λ**) i.e. peaks
3. Overlapping enriched peaks are merged. The location in the peak with the highest fragment pileup (summit) is predicted as the precise binding location. The ration between the ChIP-seq tag count and **λ** is reported as the fold enrichment.



Any questions?


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

Park, 2009, Nat Rev Genetics

References

- CRUK summer school 2019 materials (https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2019/)
- Bardet et al. Bioinformatics, 2013. “Identification of transcription factor binding sites from ChIP-seq data at high resolution”
- Wilbanks et al. 2010 PLOS One. “Evaluation of Algorithm Performance in ChIP-Seq Peak Detection”
- Sims et al., 2014 Nat Rev Genet. “Sequencing depth and coverage: key considerations in genomic analyses”
- Park, 2009, Nat Rev Genetics. “ChIP–seq: advantages and challenges of a maturing technology”