Analysis of broad occupancy patterns

Epigenomics Data Analysis Workshop

Stockholm, 24 November 2020

Agata Smialowska

NBIS, SciLifeLab, Stockholm University

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

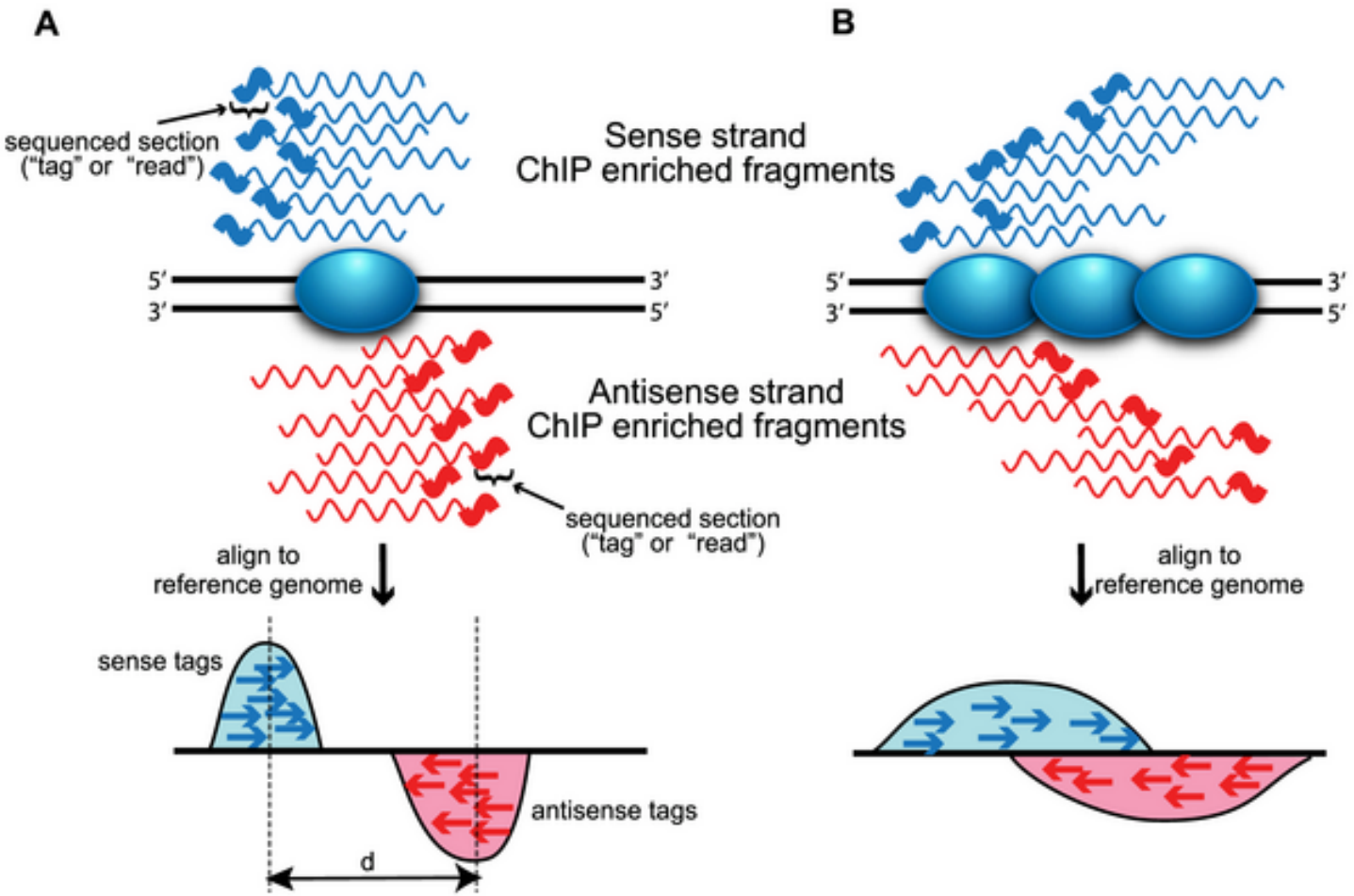
SciLifeLab

UNIVERSITET STOCKHOLMS

Point-source vs. broad peak detection

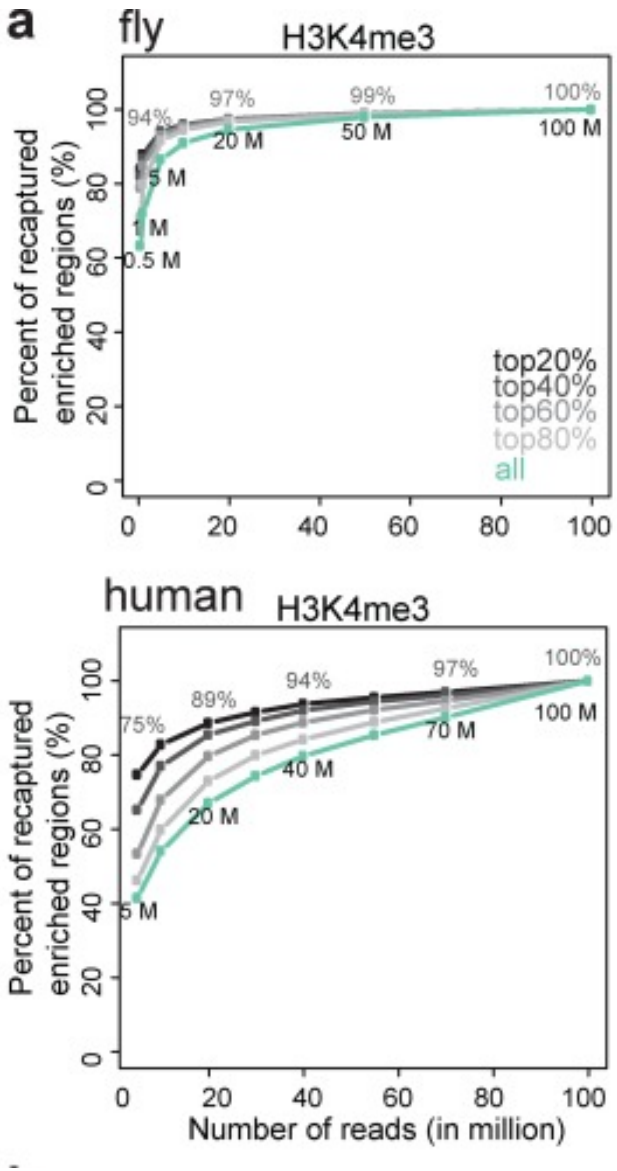
Sequence-specific binding (TFs)

Distributed binding (histones, RNApol2)



Wilbanks 2010

How to define sufficient read depth?



detection sensitivity - recovery of true enrichment region

↓

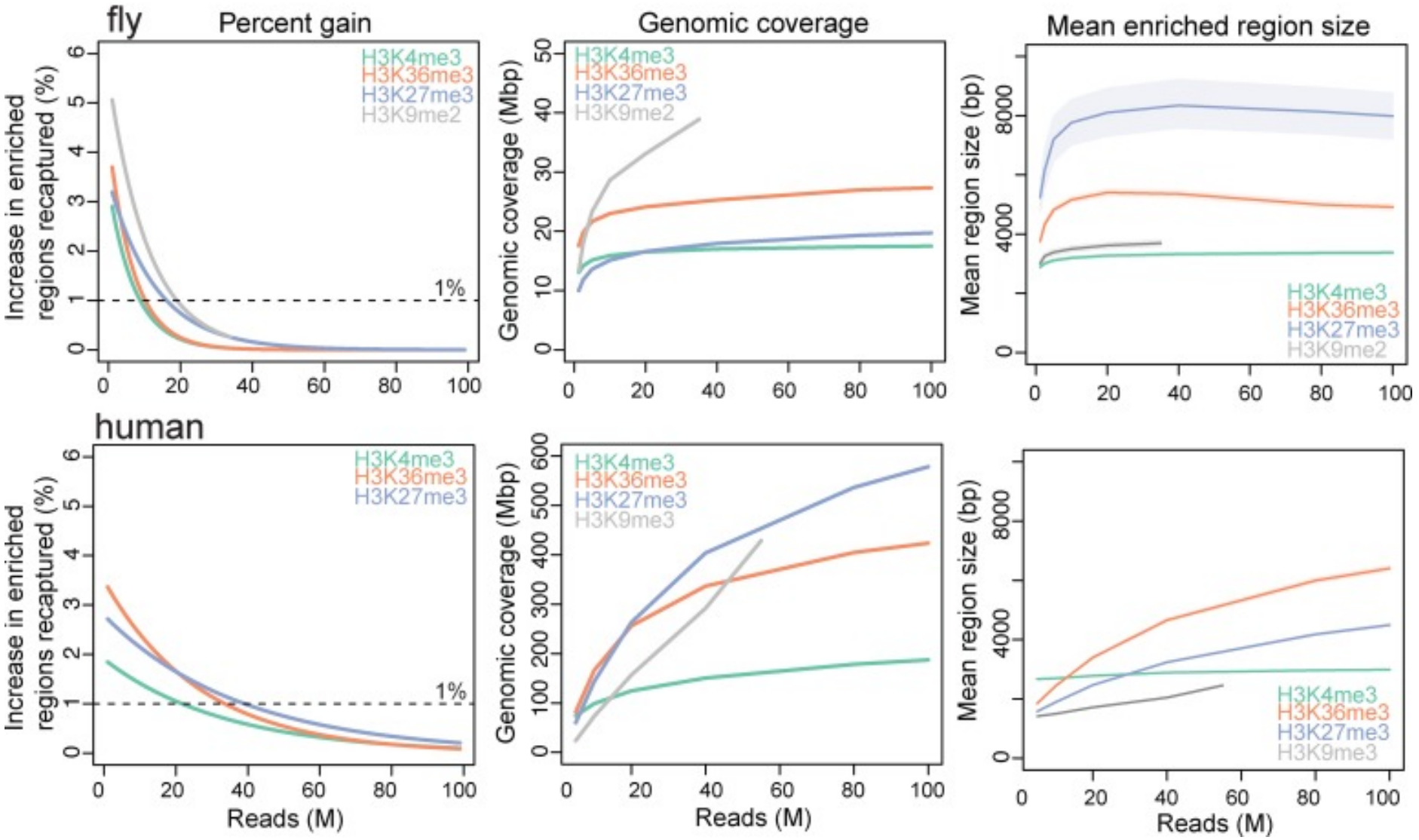
percent increase in enriched regions recaptured when an additional 1 million reads are sequenced

↓

'sufficient depth' - the sequencing depth at which the percent gain per 1 million additional sequence reads falls below 1%

Jung et al, NAR 2014

Effect of sequencing depth on site recovery



Sequencing depth

Chromatin Remodellers

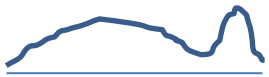
Histone marks

mixed signal

Chromatin Remodellers

Histone marks

RNA polymerase II



broad signal

Human: 3.2Gb

H3K36me3: 25 M

H3K27me3: 35 M

H3K9me3: >55 M

Drosophila: 300Mb

H3K36me3: 11 M

H3K27me3: 20M

H3K9me3: 20 M

No clear guidelines for mixed and broad type of peaks

Source: The ENCODE consortium; Jung et al, NAR 2014

Peak calling

Chromatin Remodellers

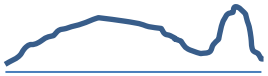
Histone marks

mixed signal

Chromatin Remodellers

Histone marks

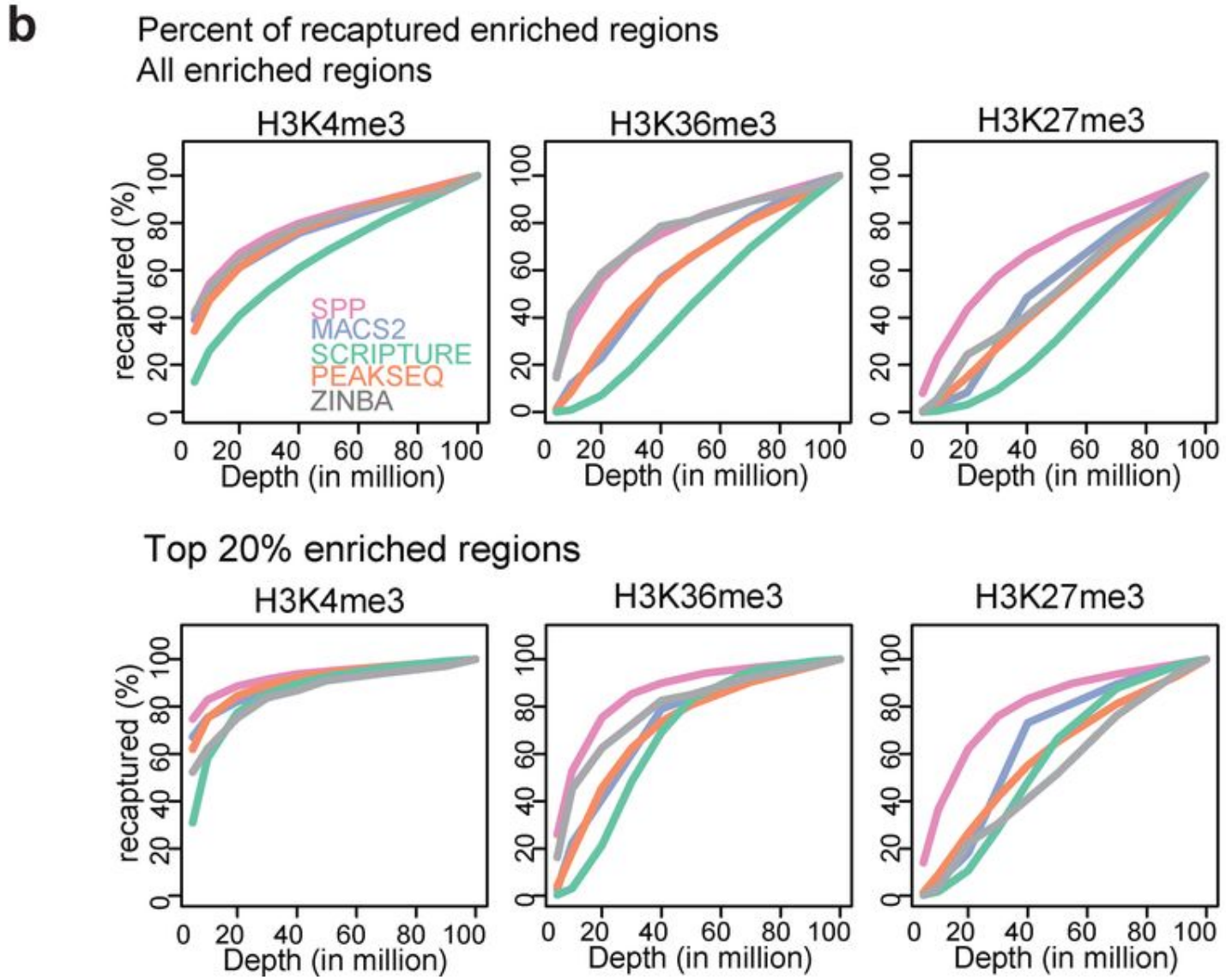
RNA polymerase II



broad signal

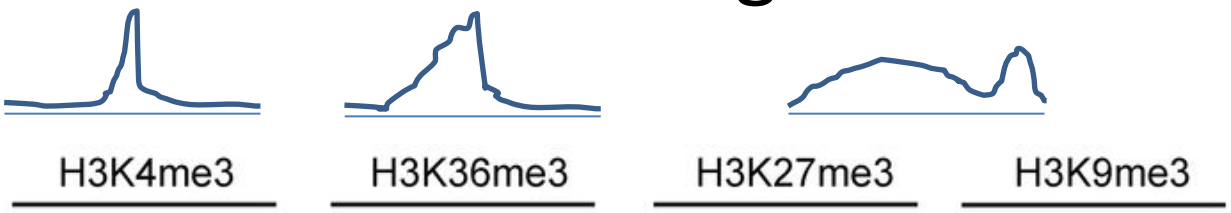
- **MACS2** in broad mode: window-based method using local statistics
- **csaw:** Scoring in moving windows
- **Epic2** (SICER): tendency of histone modifications to cluster to form the domains. This method identifies islands as *clusters* of enriched windows. Islands, rather than individual windows of fixed length, are the fundamental units of interest

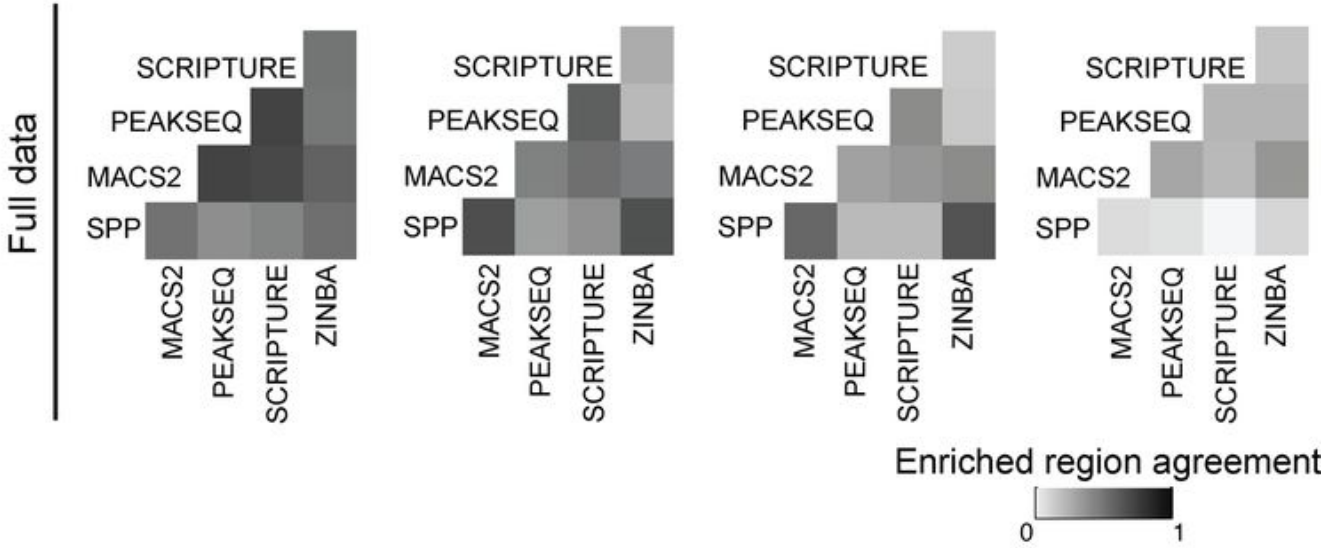
Effect of sequencing depth on regions detected by various algorithms



Jung 2014

Comparison of enriched regions detected by various algorithms

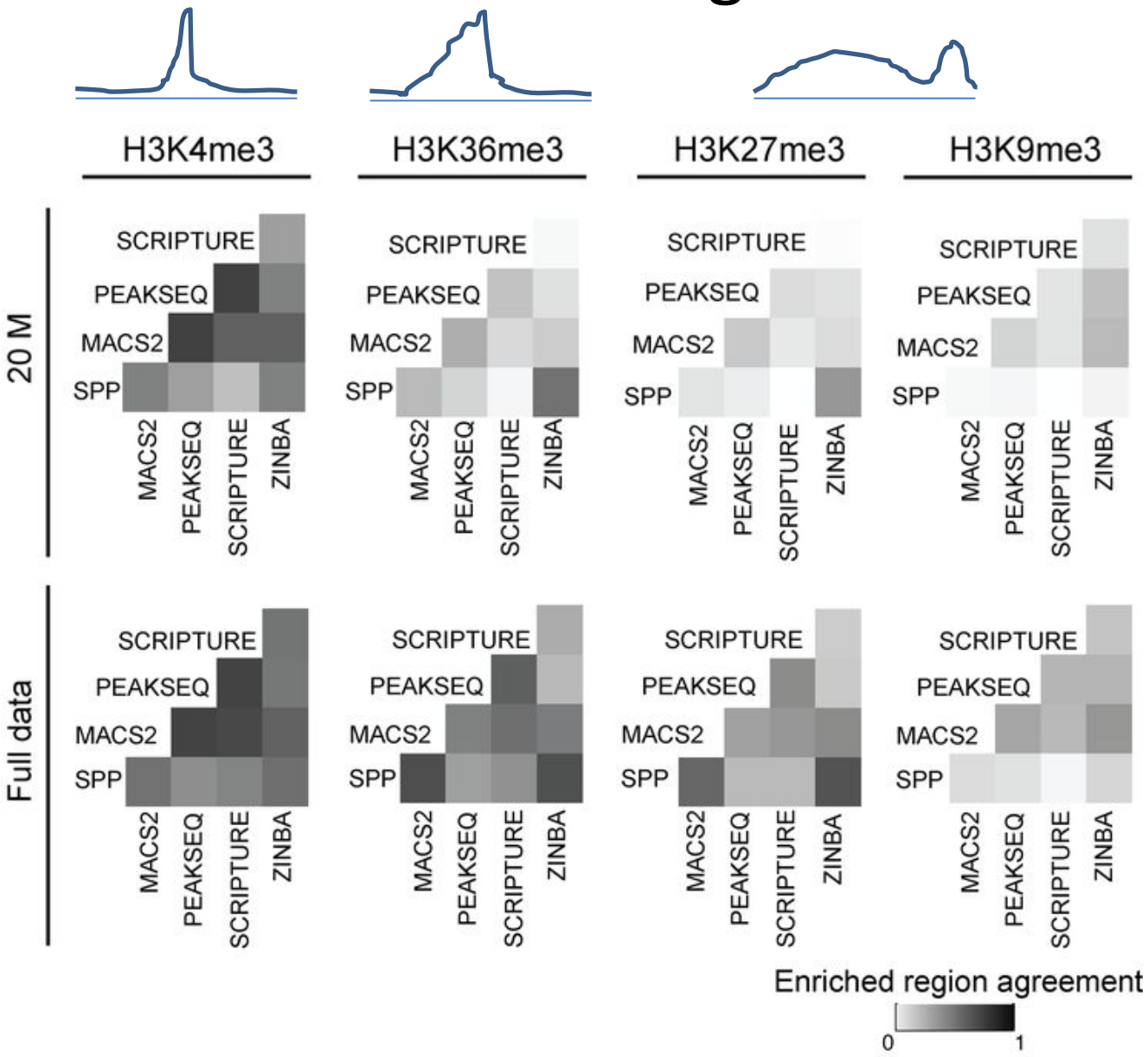




55M human

Jung 2014

Comparison of enriched regions detected by various algorithms



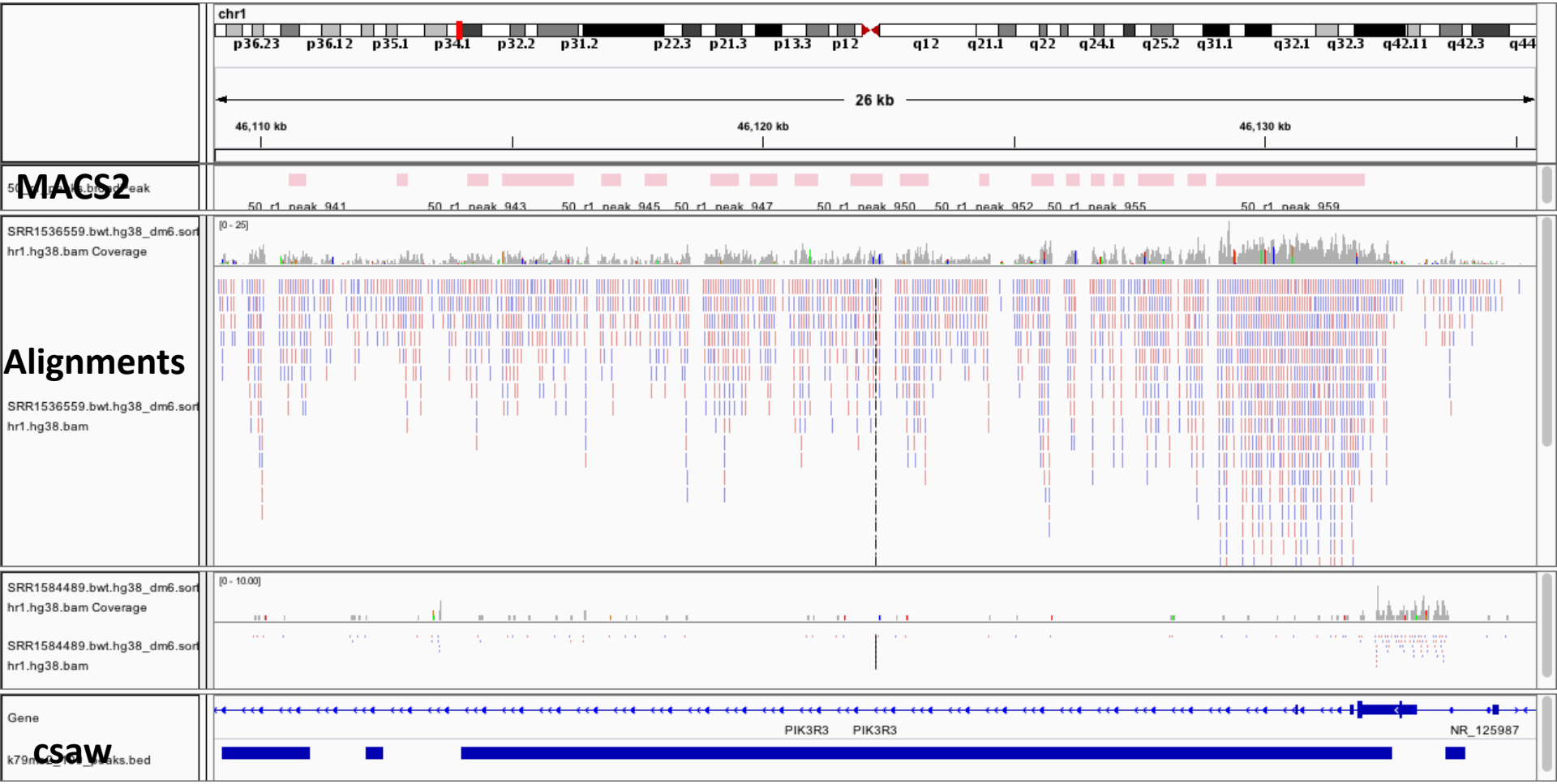
55M human

Jung 2014

Exercise

- MACS2 in `broad peak` mode;
- csaw: Detection of differentially bound regions in ChIP-seq data with sliding windows, with methods for normalization and proper FDR control;
- Low sequencing depth data; H3K79me2 (transcribed regions of active genes)

MACS2 vs. csaw



Resources for broad region analysis

- https://omictools.com/peak-calling-category
- https://www.encodeproject.org/chip-seq/histone/