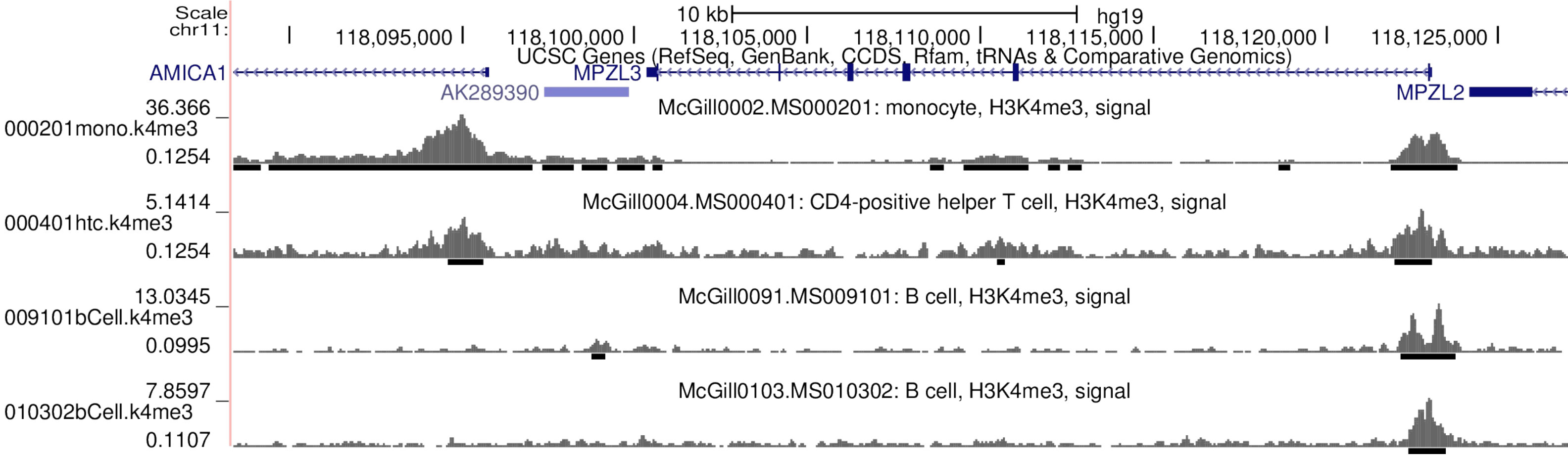


Toby Dylan Hocking, Guillem Rigail, Guillaume Bourque.

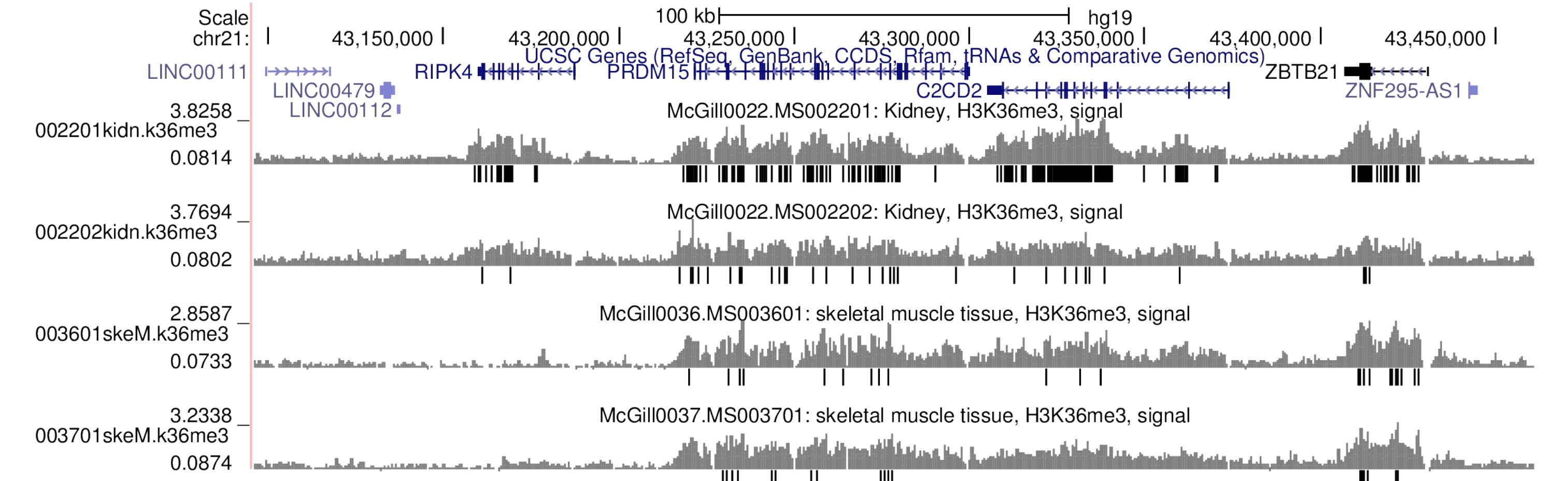
Introduction: ChIP-seq data characterize the genomic binding positions for proteins such as histones and transcription factors. The goal of a peak detection algorithm is to filter out background noise, and define the precise genomic positions of "peaks" with many aligned sequence reads.

Problem: unsupervised peak model does not match visually obvious peaks.

Data from <http://epigenomesportal.ca> H3K4me3 sharp peaks,



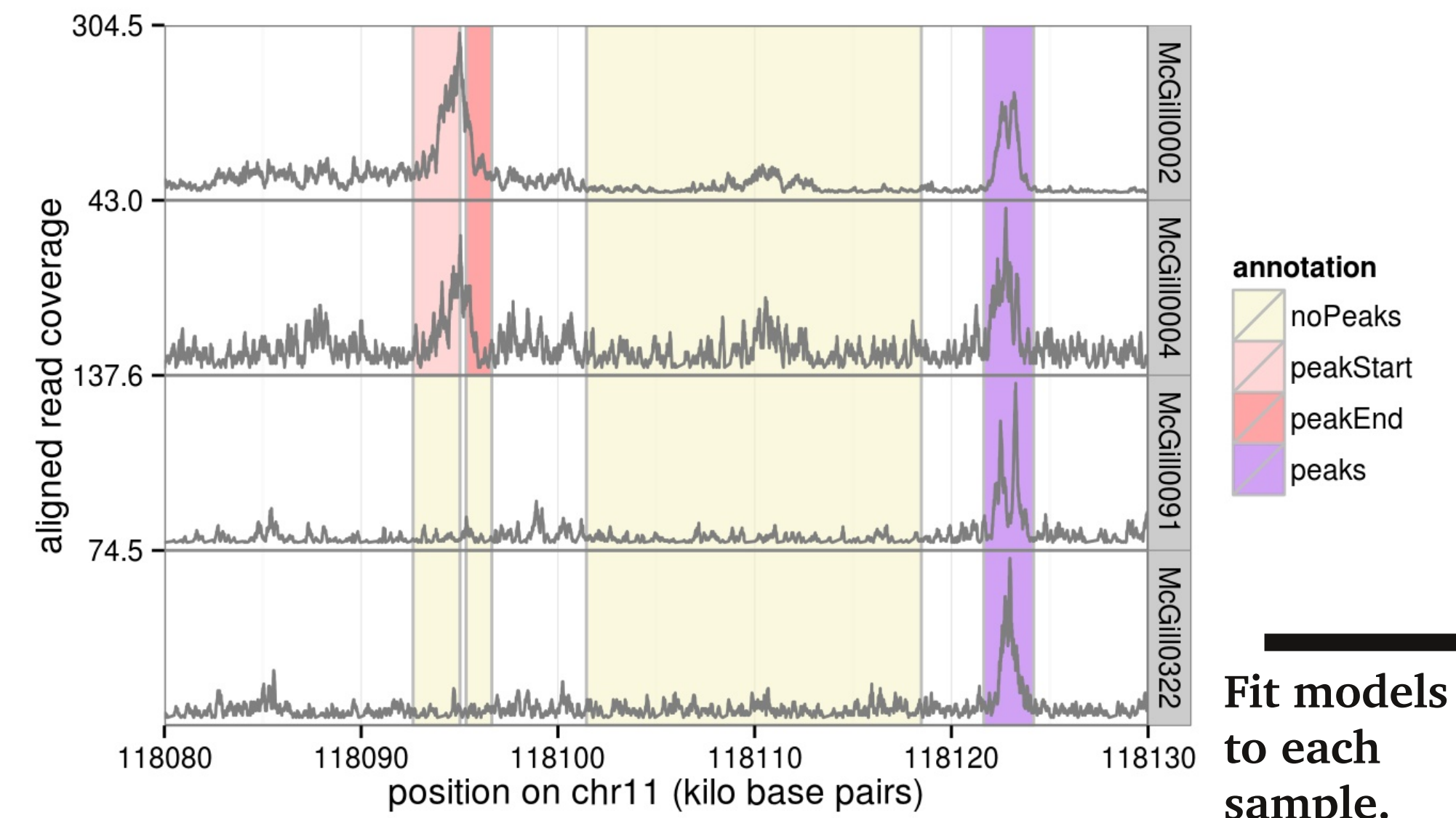
H3K36me3 broadly enriched domains, macs2 peak detector.



Benchmark data set: manually annotated regions containing peaks or no peaks

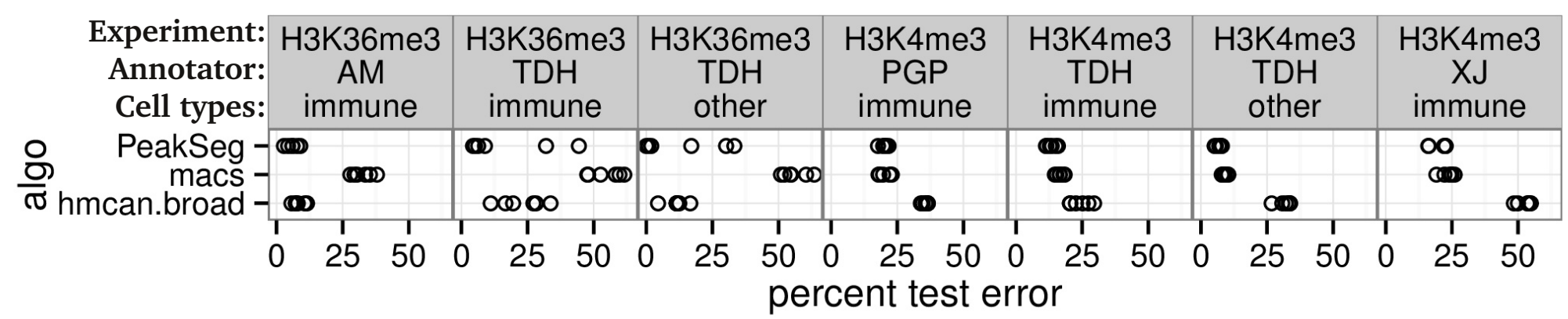
<http://cbio.ensmp.fr/~thocking/chip-seq-chunk-db/>

- Seven data sets containing 12,826 manually annotated regions across 37 samples, 8 cell types, from 4 annotators.
- peakStart/peakEnd regions should contain exactly 1 peak start/end.
- peaks regions should contain at least 1 overlapping peak.
- Below: 16 regions across 4 samples created by annotator TDH, 12 possible false positives, 8 possible false negatives.



Fit models
to each
sample.

Results: State-of-the-art peak detection for both sharp and broad peaks.



Proposed Method: constrained maximum likelihood segmentation

The Supervised Peak Detection Problem

- A ChIP-seq profile on a single chromosome with d base pairs is a vector $\mathbf{y} = [y_1 \ \cdots \ y_d] \in \mathbb{Z}_+^d$ of counts of aligned sequence reads.

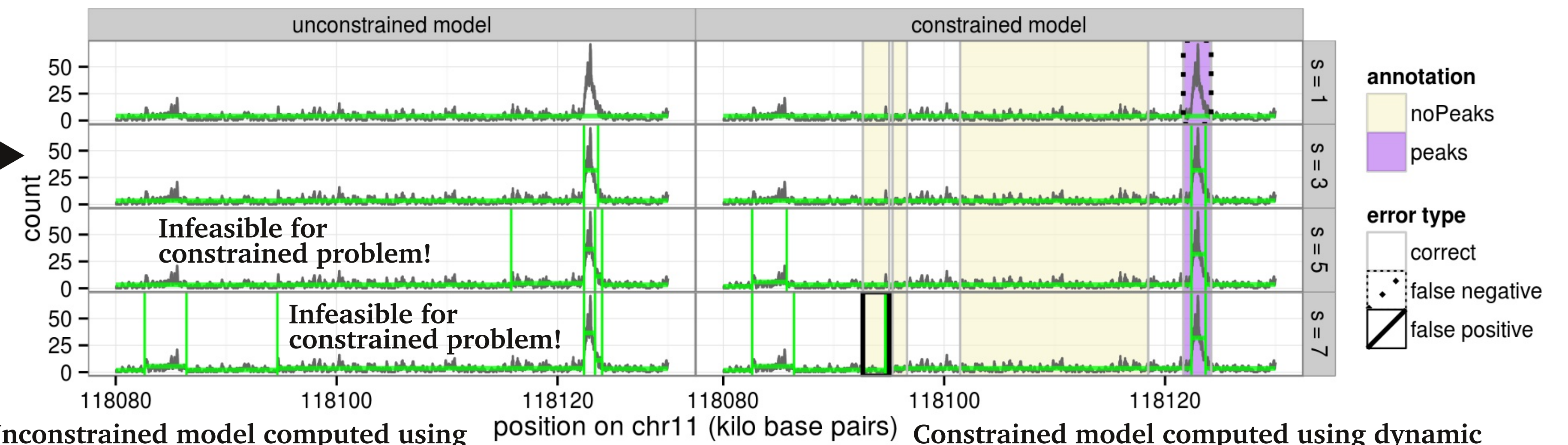
- A peak caller is a function $c : \mathbb{Z}_+^d \rightarrow \{0, 1\}^d$ which returns 0 for background noise and 1 for a peak.

- We are given $i \in \{1, \dots, n\}$ profiles \mathbf{y}_i and annotated regions R_i .

- The goal is to find a peak caller with minimal error on some test profiles:

$$\underset{c}{\text{minimize}} \sum_{i \in \text{test}} E[c(\mathbf{y}_i), R_i],$$

where E is the number of false positive and false negative regions R_i .



Unconstrained model computed using position on chr
Pruned Dynamic Programming Algorithm,
Segmentor3IsBack R package,
Cleynen et al. Algorithms for Molecular Biology 2014.

Unsupervised PeakSeg: constrained segmentation

$$\tilde{\mathbf{m}}^s(\mathbf{y}) = \arg \min_{\mathbf{m} \in \mathbb{R}^d} \rho(\mathbf{m}, \mathbf{y})$$

such that $\text{Segments}(\mathbf{m}) = s$,

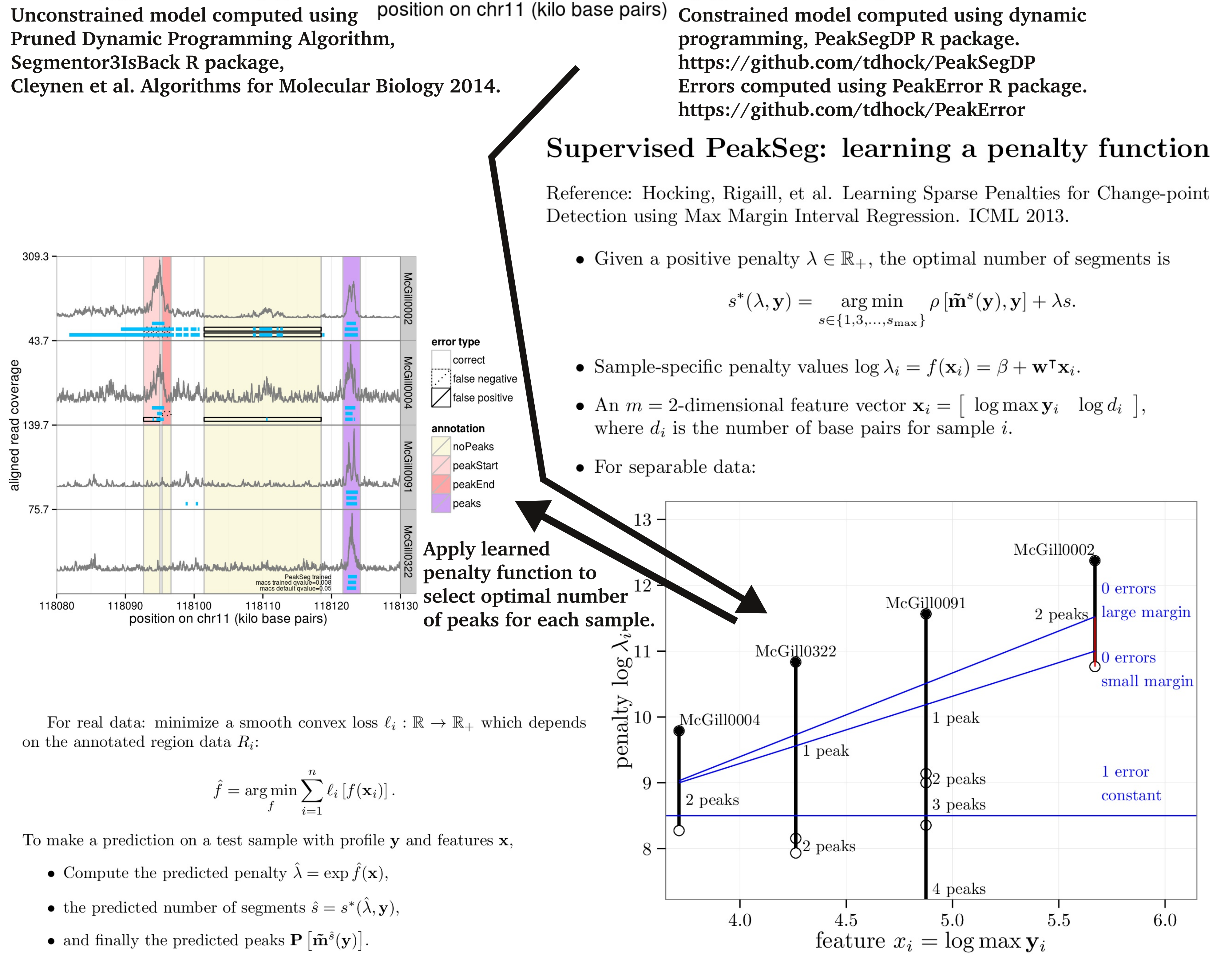
Peaks constraint: $P_j(\mathbf{m}) \in \{0, 1\}$ for all $j \in \{1, \dots, d\}$.

- The Poisson loss function is $\rho(\mathbf{m}, \mathbf{y}) = \sum_{j=1}^d m_j - y_j \log m_j$.

- $\text{Segments}(\mathbf{m}) = 1 + \sum_{j=2}^d I(m_j \neq m_{j-1})$, where I is the indicator function.

- The indicator for a peak at base j is $P_j(\mathbf{m}) = \sum_{k=1}^j \text{sign}(m_k - m_{k-1})$.

- Geometric interpretation of **Peaks constraint**: segment mean must change up, down, up, down.



Conclusions/future work:

- First supervised peak detector that learns from manually annotated regions with and without peaks.
- State-of-the-art peak detection on both sharp H3K4me3 and broad H3K36me3 profiles.
- Apply to other histone mark types and transcription factor ChIP data.
- Current dynamic programming algorithm has time complexity quadratic in number of data to segment (base pairs), but Pruned Dynamic Programming (Rigaill arXiv:1004.0887) is linear and could be used.
- Other features for penalty learning problem?
- Detecting peaks in the same genomic positions across several samples?