PeakSegJoint: fast supervised peak detection via joint segmentation of count data samples

Toby Dylan Hocking
toby.hocking@mail.mcgill.ca
joint work with Guillaume Bourque

June 17, 2015

ChIP-seq data and previous work on peak detection

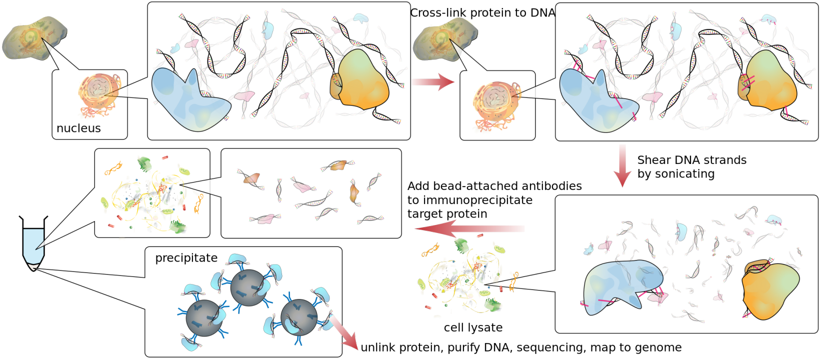
The PeakSeg and PeakSegJoint models

Speed and test error results

Conclusions

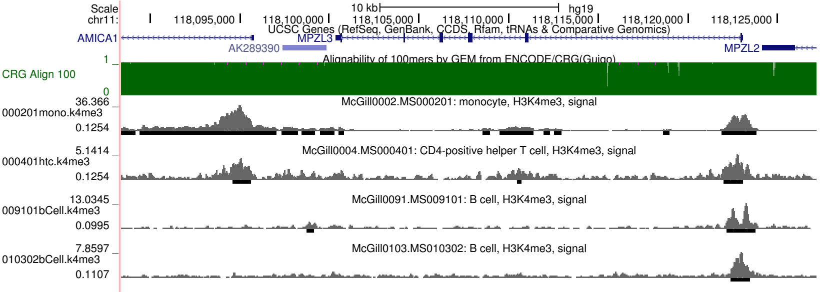
Chromatin immunoprecipitation sequencing (ChIP-seq)

Analysis of DNA-protein interactions.



Source: “ChIP-sequencing,” Wikipedia.

Problem: find peaks in each of several samples



Grey profiles are normalized aligned read count signals.
Black bars are "peaks" called by MACS2 (Zhang et al, 2008):

- many false positives.
- overlapping peaks have different start/end positions.

Existing peak detection algorithms

- Model-based analysis of ChIP-Seq (MACS), Zhang et al, 2008.
- SICER, Zang et al, 2009.
- HOMER findPeaks, Heinz et al, 2010.
- RSEG, Song and Smith, 2011.
- Histone modifications in cancer (HMCan), Ashoor et al, 2013.
- ... dozens of others.

Two big questions: how to choose the best...

- ...algorithm?
- ...parameters?

How to choose model parameters?

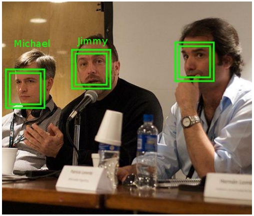
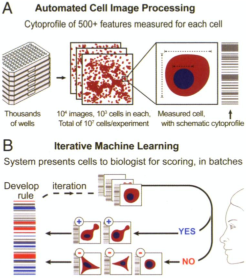
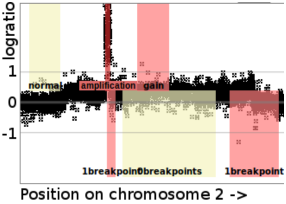
19 parameters for Model-based analysis of ChIP-Seq (MACS), Zhang et al, 2008.

```
[-g GSIZE]
[-s TSIZE] [--bw BW] [-m MFOLD MFOLD] [--fix-bimodal]
[--nomodel] [--extsize EXTSIZE | --shiftsize SHIFTSIZE]
[-q QVALUE | -p PVALUE | -F FOLDENRICHMENT] [--to-large]
[--down-sample] [--seed SEED] [--nolambda]
[--slocal SMALLLOCAL] [--llocal LARGELOCAL]
[--shift-control] [--half-ext] [--broad]
[--broad-cutoff BROADCUTOFF] [--call-summits]
```

10 parameters for Histone modifications in cancer (HMCan), Ashoor et al, 2013.

```
minLength 145
medLength 150
maxLength 155
smallBinLength 50
largeBinLength 100000
pvalueThreshold 0.01
mergeDistance 200
iterationThreshold 5
finalThreshold 0
maxIter 20
```

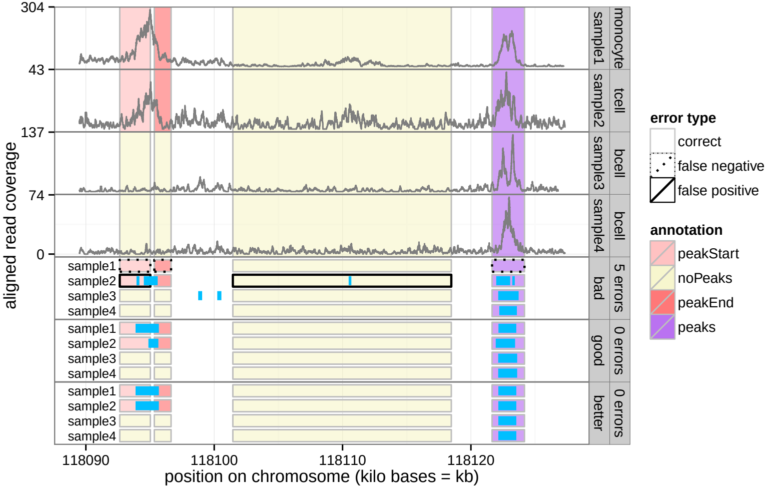
Previous work in computer vision: look and add labels to...

Photos	Cell images	Copy number profiles
		
Labels: names	phenotypes	alterations
CVPR 2013 246 papers	CellProfiler 873 citations	SegAnnDB Hocking et al, 2014.

Sources: `http://en.wikipedia.org/wiki/Face_detection`
Jones et al PNAS 2009. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning.

Labels indicate presence/absence of peaks

False negative is too few peaks, false positive is too many peaks.



Goals: peaks in same positions across samples,
with minimal number of incorrect regions.

Goal: minimize number of incorrect labels in test data

- $S = 4$ samples.
- $B = 50,000$ base positions.
- $\mathbf{Z} \in \mathbb{Z}_+^{B \times S}$ matrix of count data.
- Set of labels L (peaks, noPeaks, peakStart, peakEnd).
- Goal: find a peak caller $c : \mathbb{Z}_+^{B \times S} \rightarrow \{0, 1\}^{B \times S}$

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

where E is the number of incorrect labels
(false positives + false negatives).

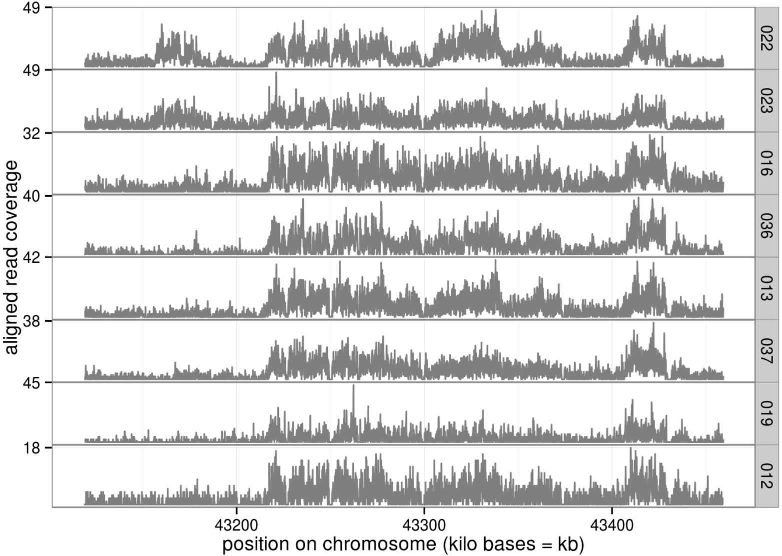
ChIP-seq data and previous work on peak detection

The PeakSeg and PeakSegJoint models

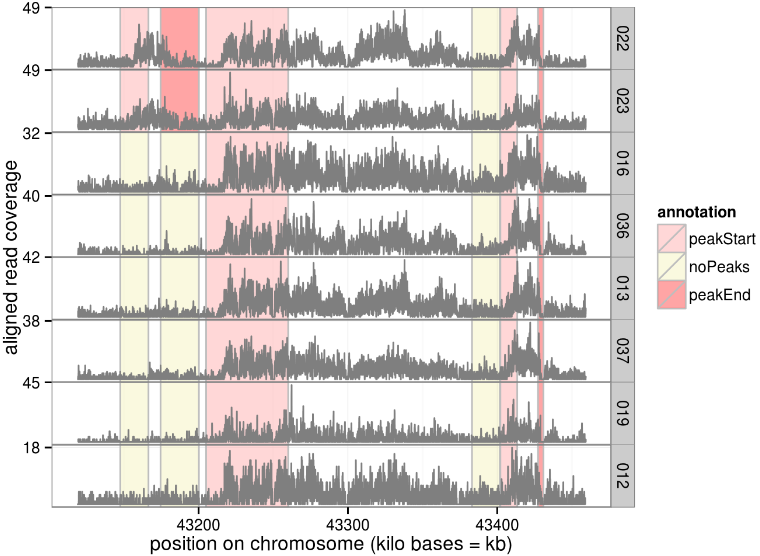
Speed and test error results

Conclusions

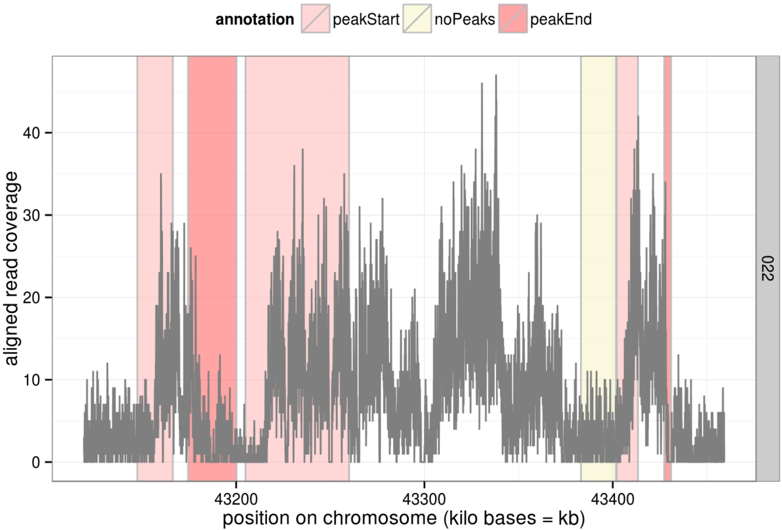
Peaks visually obvious in H3K36me3 data



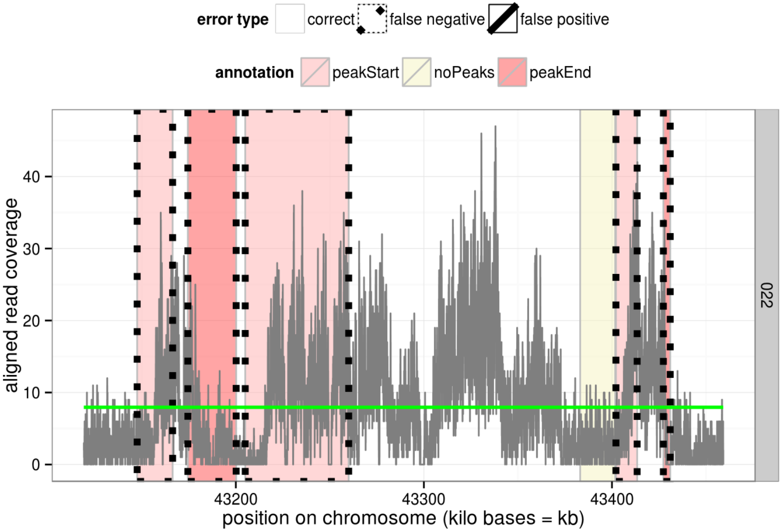
H3K36me3 data and visually determined labels



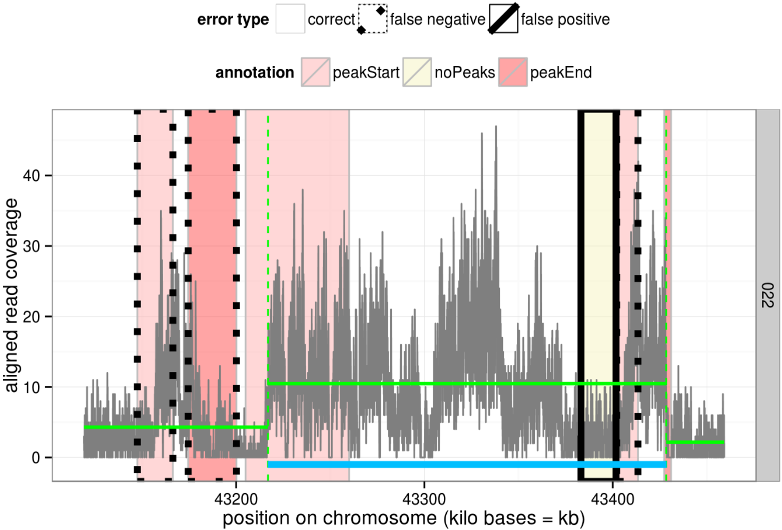
H3K36me3 data and labels (zoom to one sample)



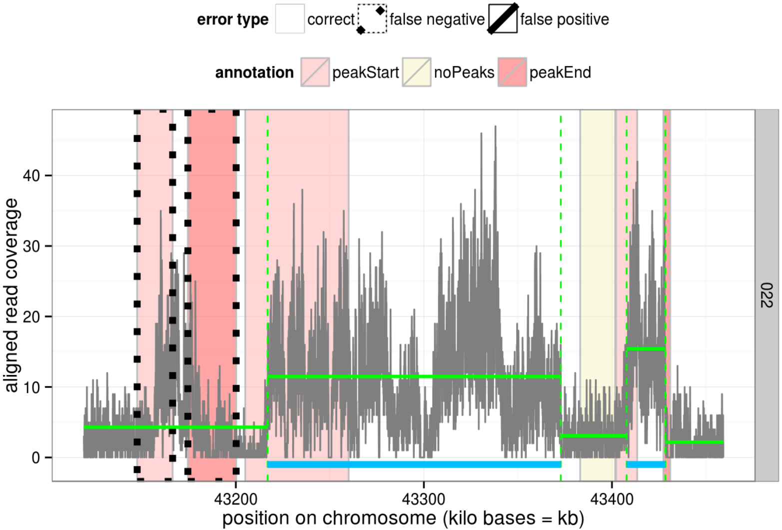
PeakSeg model with 0 peaks



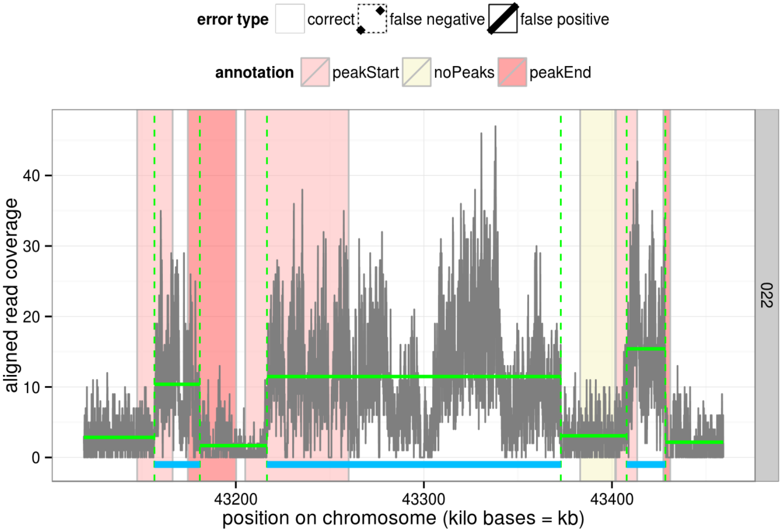
PeakSeg model with 1 peak



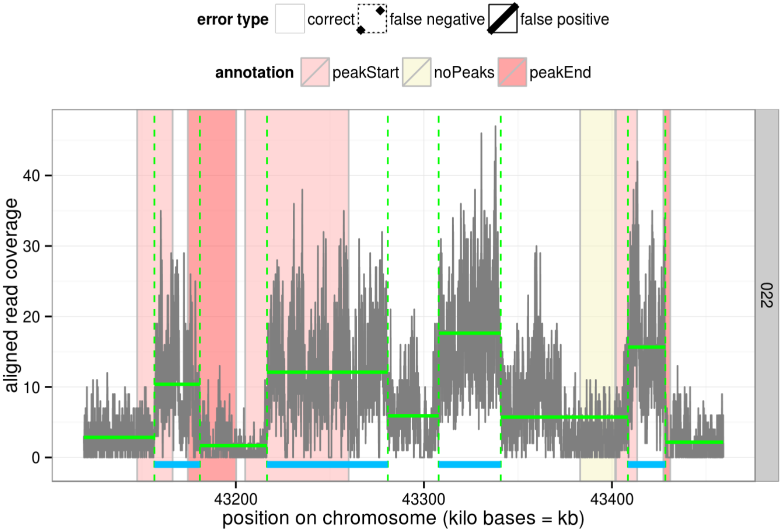
PeakSeg model with 2 peaks



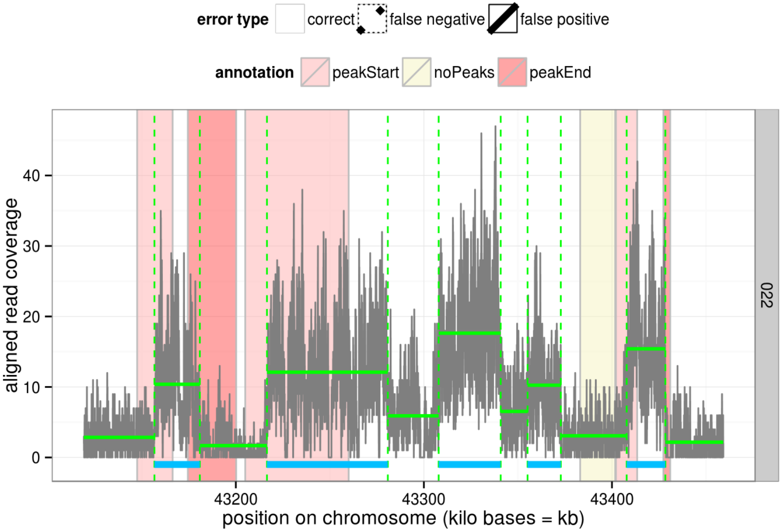
PeakSeg model with 3 peaks



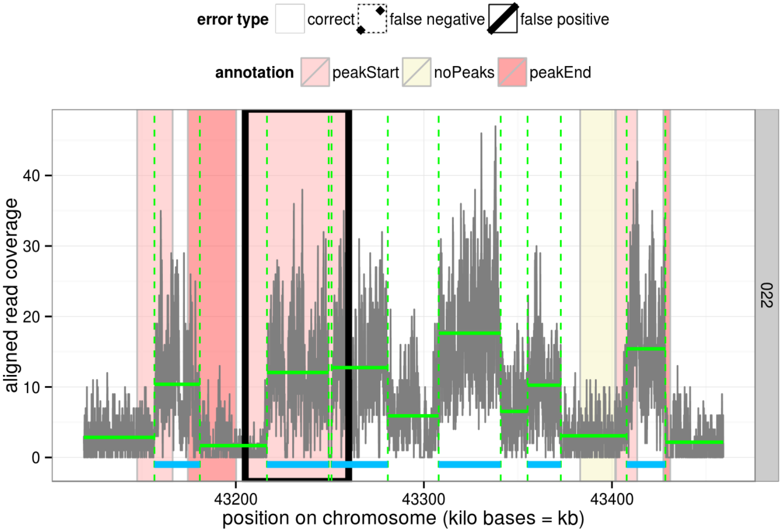
PeakSeg model with 4 peaks



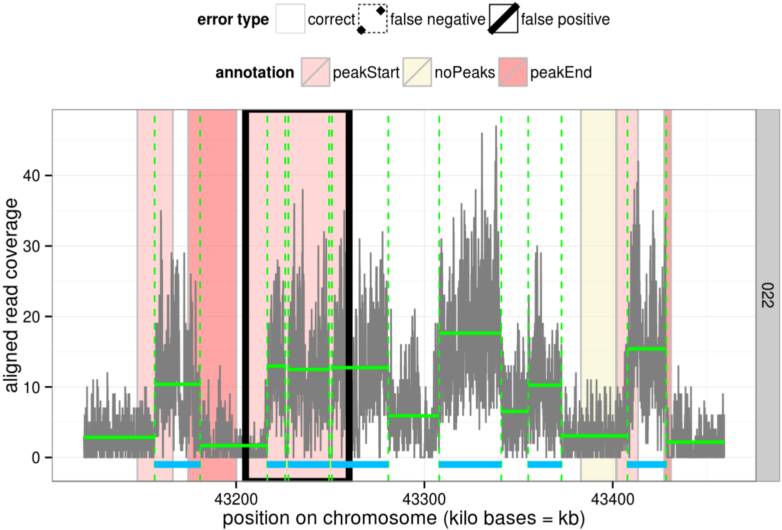
PeakSeg model with 5 peaks



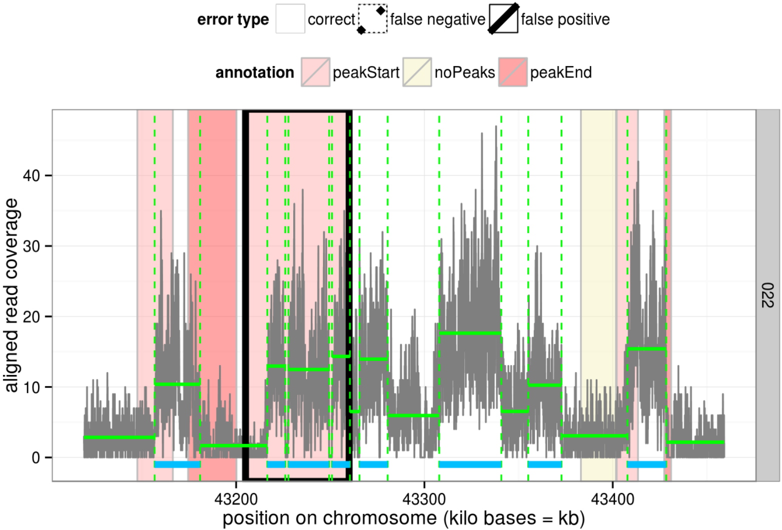
PeakSeg model with 6 peaks



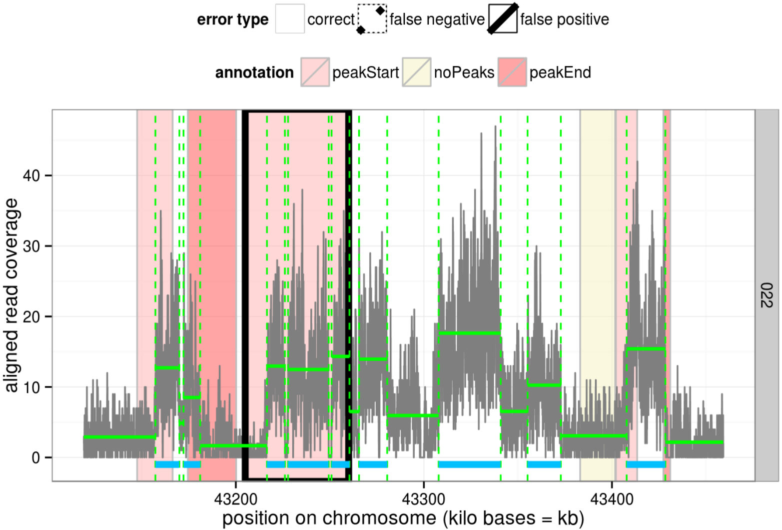
PeakSeg model with 7 peaks



PeakSeg model with 8 peaks



PeakSeg model with 9 peaks



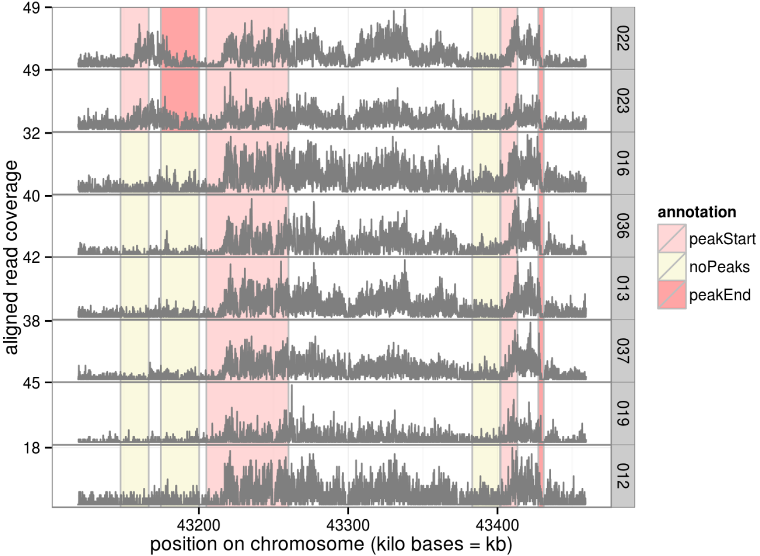
PeakSeg: most likely $0, \dots, p_{\max}$ peaks in a single sample

- Count data $\mathbf{Z} = \left[\begin{array}{ccc} \mathbf{z}_1 & \cdots & \mathbf{z}_S \end{array} \right] \in \mathbb{Z}_+^{B \times S}$ for S samples and B bases.
- For $p \in \{0, \dots, p_{\max}\}$ peaks, and for each sample $\mathbf{z} \in \mathbb{Z}_+^B$, compute the piecewise constant mean vector:

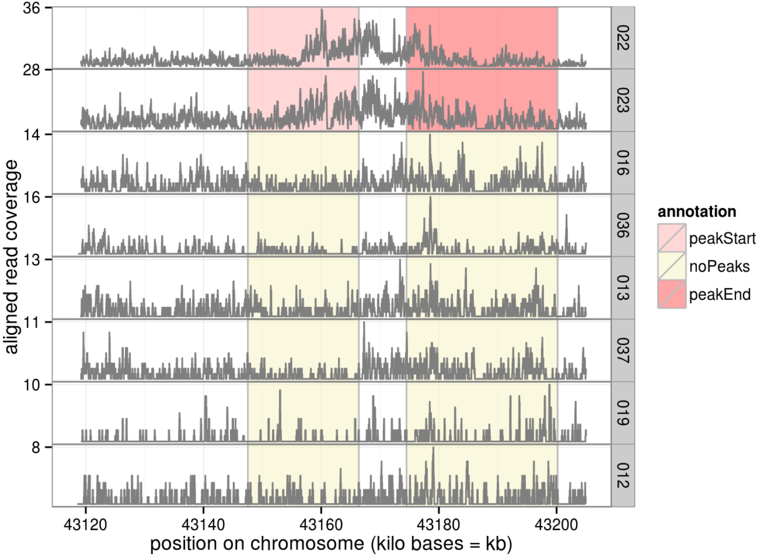
$$\begin{aligned} \tilde{\mathbf{m}}^p(\mathbf{z}) = \operatorname*{arg\,min}_{\mathbf{m} \in \mathbb{R}^B} \quad & \text{PoissonLoss}(\mathbf{m}, \mathbf{z}) \\ \text{such that} \quad & \text{Peaks}(\mathbf{m}) = p, \\ & \forall j \in \{2, \dots, B\}, \quad P_j(\mathbf{m}) \in \{0, 1\}. \\ & \text{up, down, up, down constraint.} \end{aligned}$$

- Peak indicator: $P_j(\mathbf{m}) = \sum_{k=2}^{j} \operatorname{sign}(m_k - m_{k-1})$.
- Hyper-parameters to choose: genomic window size B, maximum number of peaks $p_{\max}$.
- $O(p_{\max} B^2)$ Constrained Dynamic Programming Algorithm (cDPA), Hocking et al, ICML 2015.

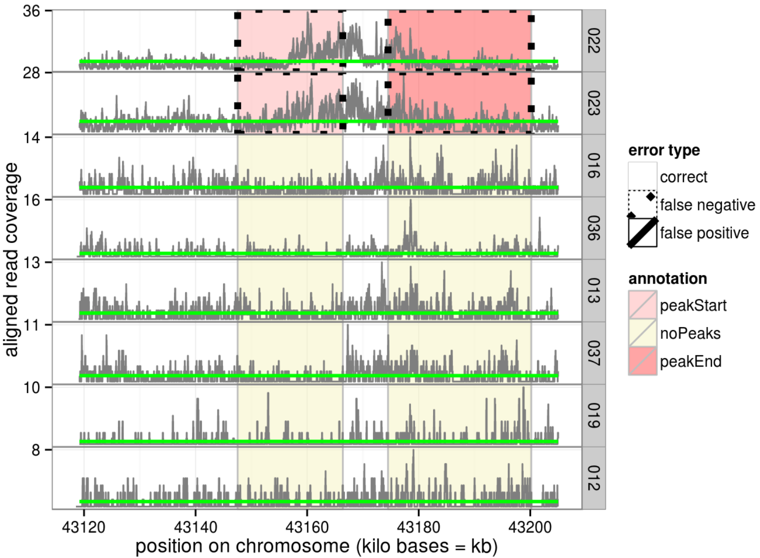
H3K36me3 data and visually determined labels



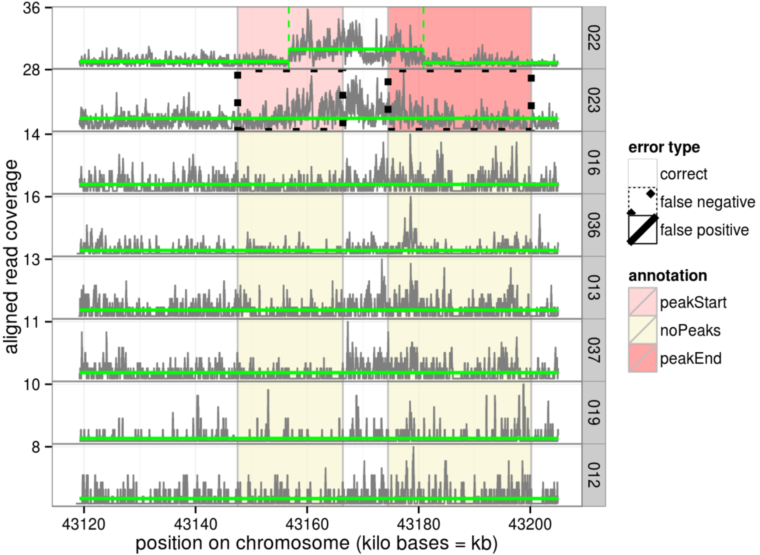
H3K36me3 data and labels (zoom to one peak)



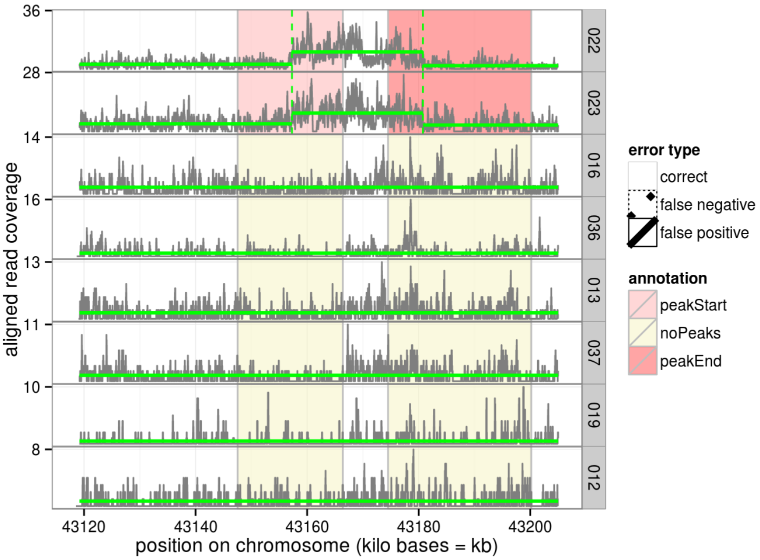
PeakSegJoint model with 0 peaks



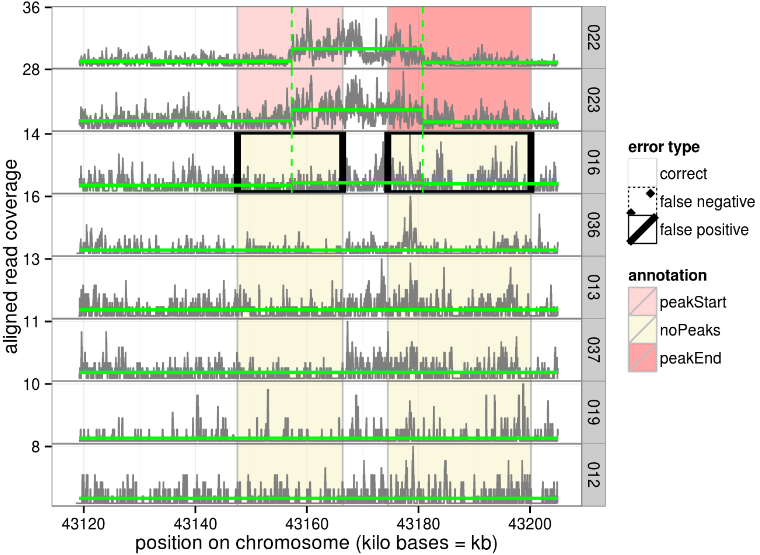
PeakSegJoint model with 1 peak



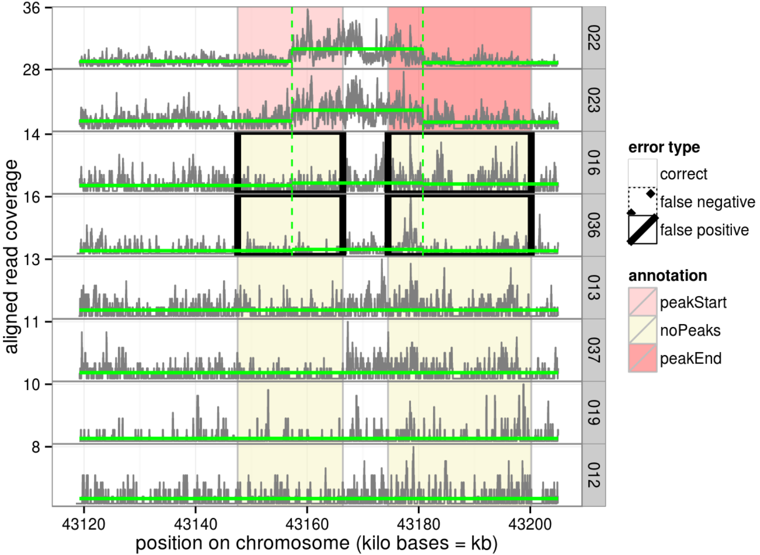
PeakSegJoint model with 2 peaks



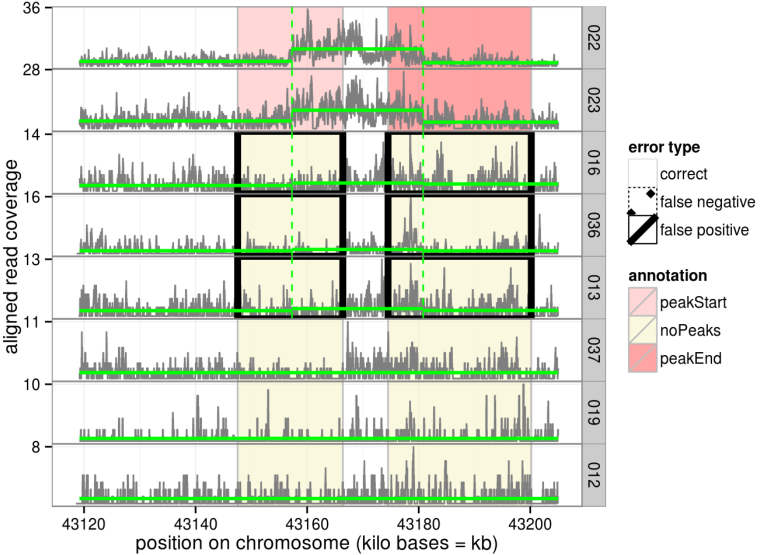
PeakSegJoint model with 3 peaks



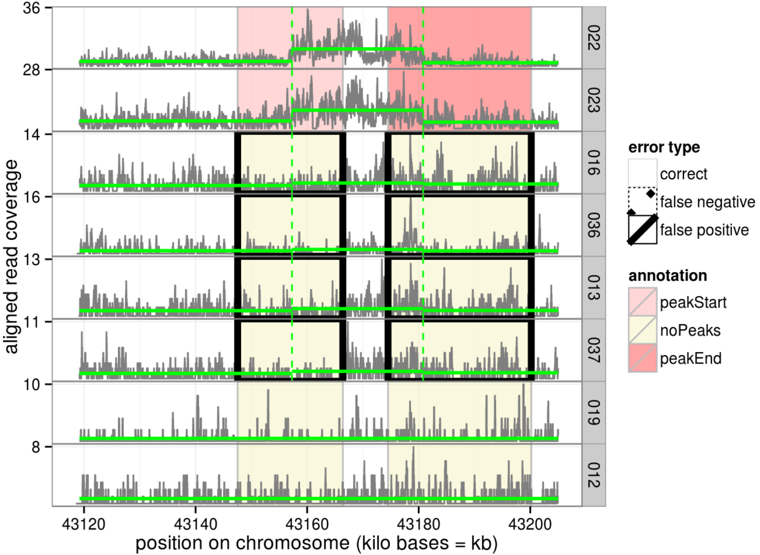
PeakSegJoint model with 4 peaks



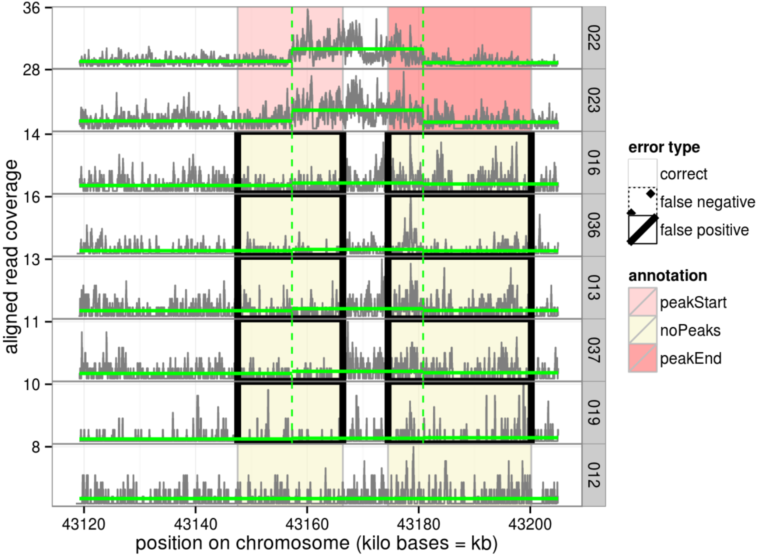
PeakSegJoint model with 5 peaks



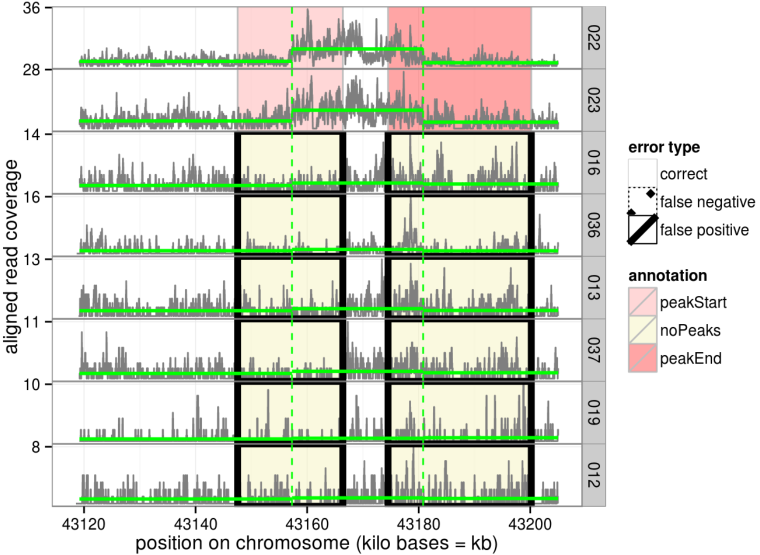
PeakSegJoint model with 6 peaks



PeakSegJoint model with 7 peaks



PeakSegJoint model with 8 peaks



PeakSegJoint: best common peak in $0, \ldots, S$ samples

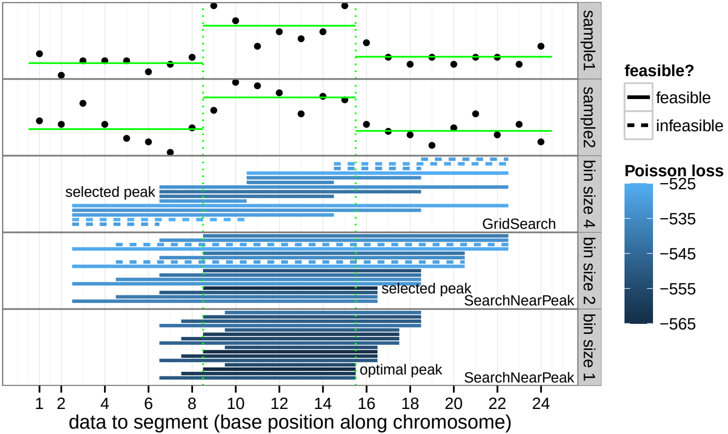
- $\mathbf{Z} = \left[\begin{array}{ccc} \mathbf{z}_1 & \cdots & \mathbf{z}_S \end{array} \right] \in \mathbb{Z}_+^{B \times S}$ for B bases and S samples.
- For $p \in \{0, \ldots, S\}$ samples each with 1 common peak, compute the mean matrix

$$\hat{\mathbf{M}}^p(\mathbf{Z}) = \operatorname*{arg\,min}_{\mathbf{M} \in \mathbb{R}^{B \times S}} \sum_{s=1}^{S} \text{PoissonLoss}(\mathbf{m}_s, \mathbf{z}_s)$$

$$\begin{aligned}
\text{up, down, up, down: } & \forall s \in \{1, \ldots, S\}, \forall j \in \{2, \ldots, B\}, \\
& P_j(\mathbf{m}_s) \in \{0, 1\}, && (1) \\
\text{peaks per sample: } & \forall s \in \{1, \ldots, S\}, \text{Peaks}(\mathbf{m}_s) \in \{0, 1\}, && (2) \\
\text{total peaks: } & p = \sum_{s=1}^{S} \text{Peaks}(\mathbf{m}_s), && (3) \\
\text{same starts/ends: } & \forall s_1 \neq s_2 \mid \text{Peaks}(\mathbf{m}_{s_1}) = \text{Peaks}(\mathbf{m}_{s_2}) = 1, \\
& \forall j \in \{1, \ldots, B\}, P_j(\mathbf{m}_{s_1}) = P_j(\mathbf{m}_{s_2}). && (4)
\end{aligned}$$

Hocking and Bourque, arXiv:1506.01286.

Demonstration of approximate JointZoom algorithm



Interactive figure at `http://bit.ly/1AA6TgK`

Example runs of approximate JointZoom algorithm

Previous slide: small data with $B = 24$ bases.

- Zoom in to a bin size of 4 bases.
- That gives $b = 7$ bins.
- Consider all peak starts/ends $= O(b^2) = 15$ models.
- Consider 16 models each at bin sizes 2 and 1.

Real data: $B = 85846$ bases.

- Zoom in to a bin size of 16384 bases.
- That gives $b = 6$ bins.
- Consider all peak starts/ends $= O(b^2) = 10$ models.
- Consider 16 models each at bin sizes 8192, 4096, ..., 4, 2, 1.

Zoom factor parameter fixed at $\beta = 2$.

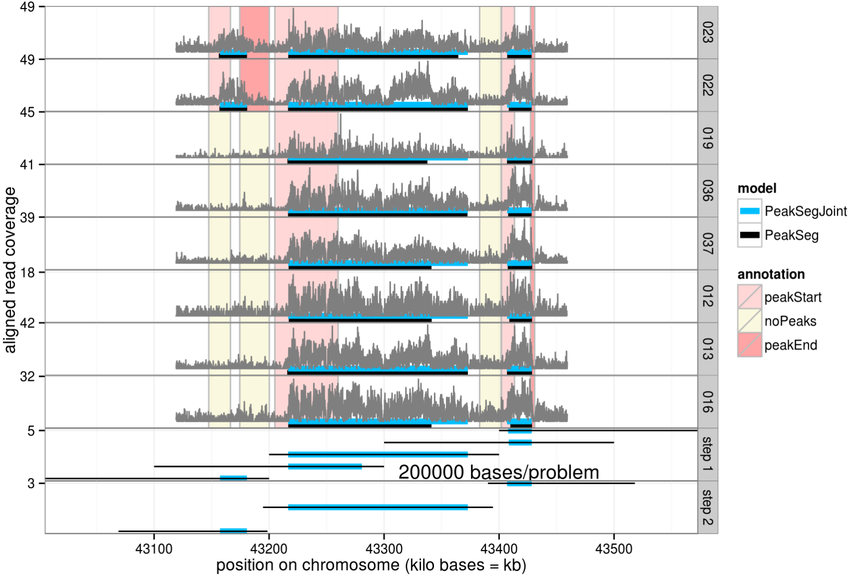
Time complexity of approximate JointZoom algorithm

Require: count data $\mathbf{Z} \in \mathbb{Z}_+^{B \times S}$, zoom factor $\beta \in \{2, 3, \dots\}$, number of samples with 1 peak $p \in \{0, \dots, S\}$.

$\text{BinSize} \leftarrow \textsc{MaxBinSize}(B, \beta)$.
$\text{Peak}, \text{Samples} \leftarrow \textsc{GridSearch}(\mathbf{Z}, p, \text{BinSize})$.
while $1 < \text{BinSize}$ **do**
    $\text{BinSize} \leftarrow \text{BinSize}/\beta$.
    $\text{Peak} \leftarrow \textsc{SearchNearPeak}(\mathbf{Z}, \text{Samples}, \text{BinSize}, \text{Peak})$
end while
return Peak, Samples.

- $\textsc{GridSearch}$ checks $O(1)$ models.
- Each $\textsc{SearchNearPeak}$ checks $O(\beta^2)$ models.
- While loop executed $O(\log B)$ times.
- Computing feasibility and maximum likelihood is $O(pB)$.
- Time for one model: $O(\beta^2 pB \log B)$.
- Time for $S + 1$ models: $O(\beta^2 SB \log B)$.

H3K36me3 data, PeakSeg and Joint model



http://bl.ocks.org/tdhock/raw/b77c1a7e4d6aee40bf6c/

ChIP-seq data and previous work on peak detection

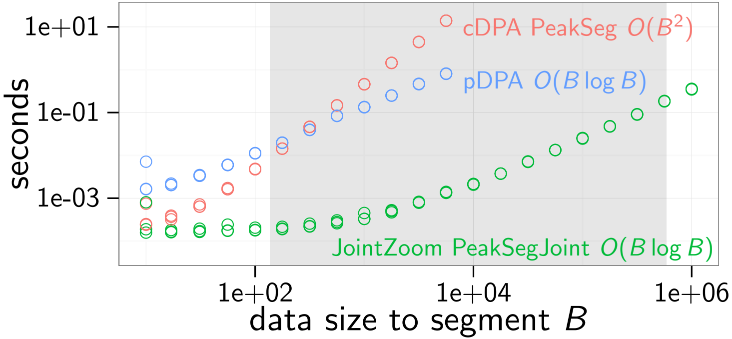
The PeakSeg and PeakSegJoint models

Speed and test error results

Conclusions

PeakSegJoint much faster than other Poisson segmentation algorithms

Data: simulated single-sample, single-peak.



pDPA from Segmentor3IsBack R package (Cleynen et al, 2014).
cDPA from PeakSegDP R package (Hocking et al, ICML 2015).

Timings on example H3K36me3 data

Find best 0,...,9 peaks in each of 8 samples (80 PeakSeg models):

sample.id	bases	data	seconds	seconds2000
McGill0023	340174	35507	157.70	0.73
McGill0022	340167	43291	218.96	0.73
McGill0019	340223	12109	28.88	0.72
McGill0036	340627	28001	106.42	0.73
McGill0037	340174	29338	114.90	0.73
McGill0012	340763	15673	27.78	0.72
McGill0013	340303	32784	193.26	0.73
McGill0016	340132	33321	117.41	0.72
total			965.32	5.81

Find best common peak in 0,...,8 samples in each of 5 genomic regions (45 PeakSegJoint models):

	data	seconds
chr21:43000000-43200000	22875	0.05
chr21:43100000-43300000	111333	0.08
chr21:43200000-43400000	165214	0.11
chr21:43300000-43500000	118699	0.09
chr21:43400000-43600000	41952	0.05
total		0.39

ChIP-seq data and previous work on peak detection

The PeakSeg and PeakSegJoint models

Speed and test error results

Conclusions

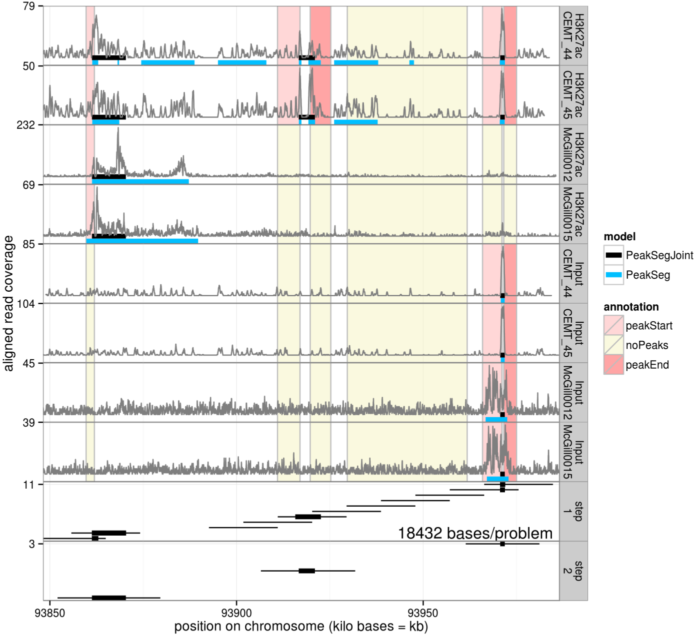
Thanks for your attention!

Write me at toby.hocking@mail.mcgill.ca to collaborate!

Source code for slides, figures, paper online!
https://github.com/tdhock/PeakSegJoint-paper

Supplementary slides appear after this one.

H3K27ac and Input data, PeakSeg and Joint model



Timings on example H3K27ac data

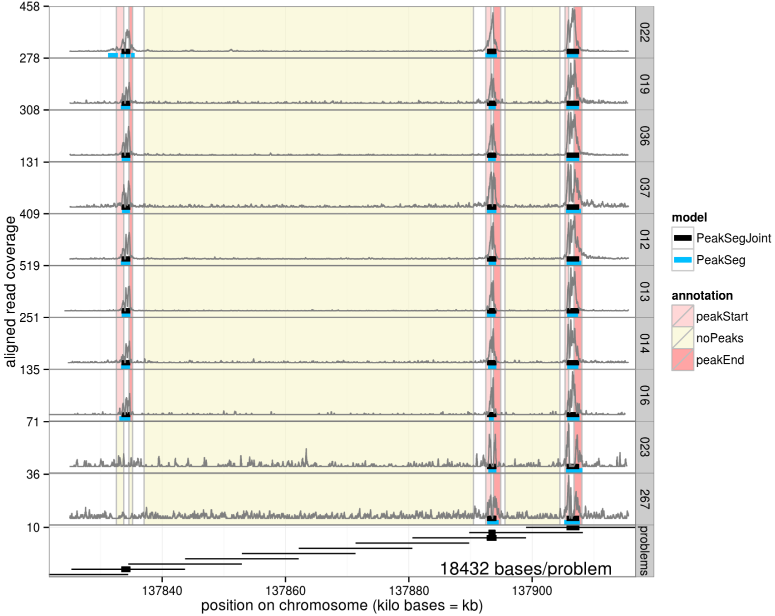
Find best
0,...,9 peaks
in each of 8 samples
(80 PeakSeg models)

seconds	sample.id
0.99	H3K27ac CEMT_44
0.96	H3K27ac CEMT_45
1.00	H3K27ac McGill0012
1.00	H3K27ac McGill0015
0.99	Input CEMT_44
1.00	Input CEMT_45
1.01	Input McGill0012
1.00	Input McGill0015
7.94	total

Find best common peak
in 0,...,8 samples
in each of 11 genomic regions
(99 PeakSegJoint models)

	data	seconds
chr11:93846528-93864960	7510	0.03
chr11:93855744-93874176	11675	0.03
chr11:93892608-93911040	5619	0.03
chr11:93901824-93920256	6236	0.03
chr11:93911040-93929472	5559	0.03
chr11:93920256-93938688	5149	0.04
chr11:93929472-93947904	4359	0.01
chr11:93938688-93957120	3071	0.03
chr11:93947904-93966336	3030	0.02
chr11:93957120-93975552	7184	0.04
chr11:93966336-93984768	7446	0.04
total		0.32

H3K4me3 data, PeakSeg and Joint model



Timings on example H3K4me3 data

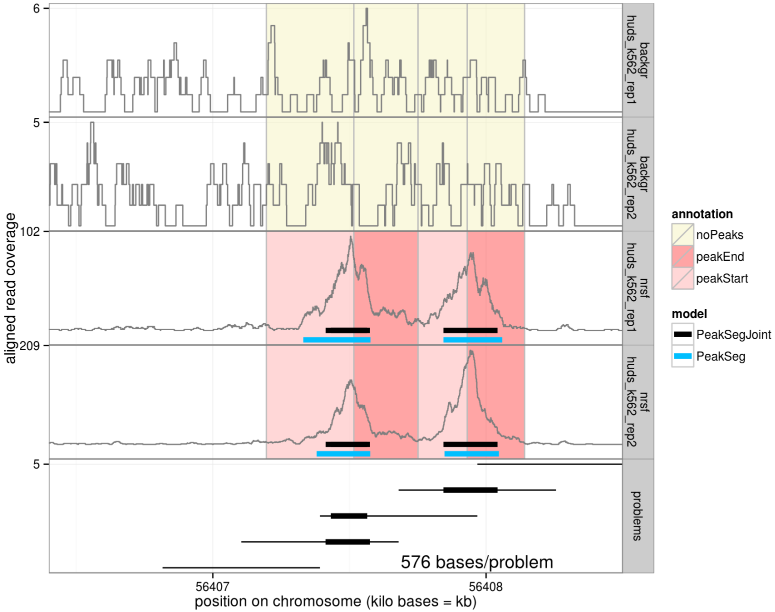
Find best
0,...,9 peaks
in each of 10 samples
(100 PeakSeg models)

seconds	sample.id
0.72	McGill0022
0.71	McGill0019
0.72	McGill0036
0.72	McGill0037
0.74	McGill0012
0.76	McGill0013
0.72	McGill0014
0.72	McGill0016
0.73	McGill0023
0.75	McGill0267
7.30	total

Find best common peak
in 0,...,10 samples
in each of 10 genomic regions
(110 PeakSegJoint models)

	data	seconds
chr21:137816064-137834496	7603	0.01
chr21:137825280-137843712	12420	0.05
chr21:137834496-137852928	7023	0.01
chr21:137843712-137862144	3915	0.04
chr21:137852928-137871360	3597	0.03
chr21:137862144-137880576	3588	0.03
chr21:137871360-137889792	4255	0.05
chr21:137880576-137899008	13317	0.05
chr21:137889792-137908224	26436	0.05
chr21:137899008-137917440	19644	0.05
total		0.36

NRSF transcription factor data, PeakSeg and Joint model



Timings on example transcription factor data

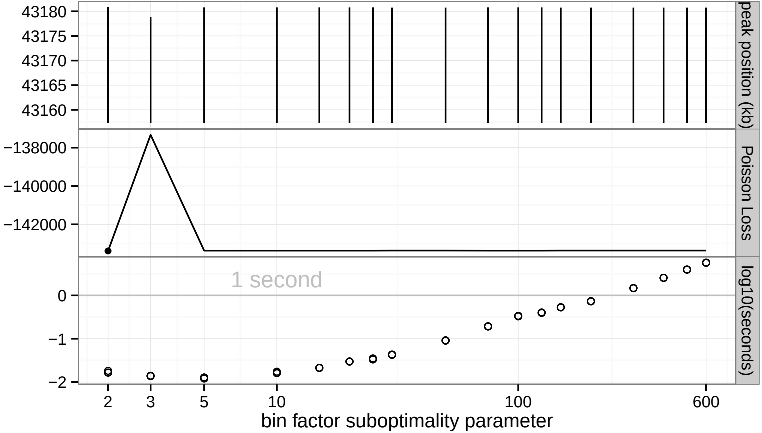
Find best
0,...,9 peaks
in each of 4 samples
(40 PeakSeg models)

seconds	sample.id
0.26	backgr huds_k562_rep1
0.24	backgr huds_k562_rep2
0.30	nrsf huds_k562_rep1
0.31	nrsf huds_k562_rep2
1.10	total

Find best common peak
in 0,...,4 samples
in each of 5 genomic regions
(25 PeakSegJoint models)

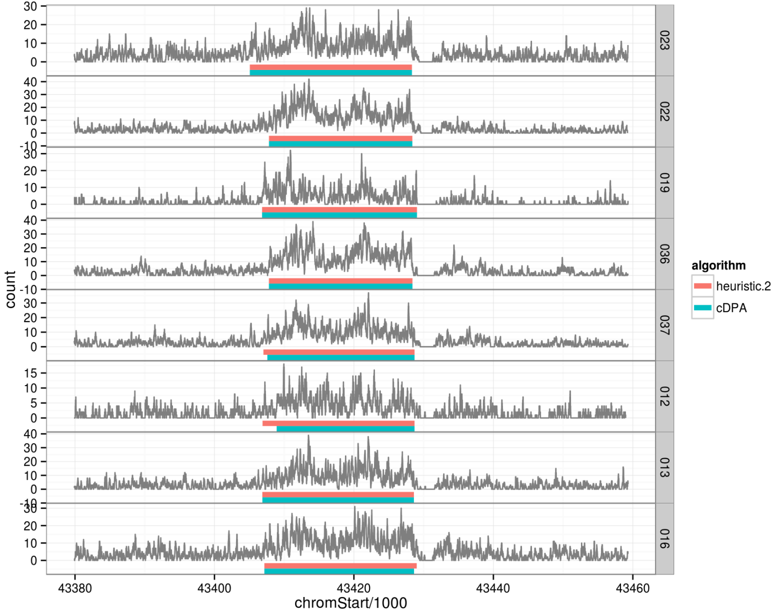
	data	seconds
chr21:56406816-56407392	345	0.01
chr21:56407104-56407680	761	0.02
chr21:56407392-56407968	975	0.01
chr21:56407680-56408256	709	0.02
chr21:56407968-56408544	298	0.01
total		0.07

Bin factor parameter controls optimality and speed

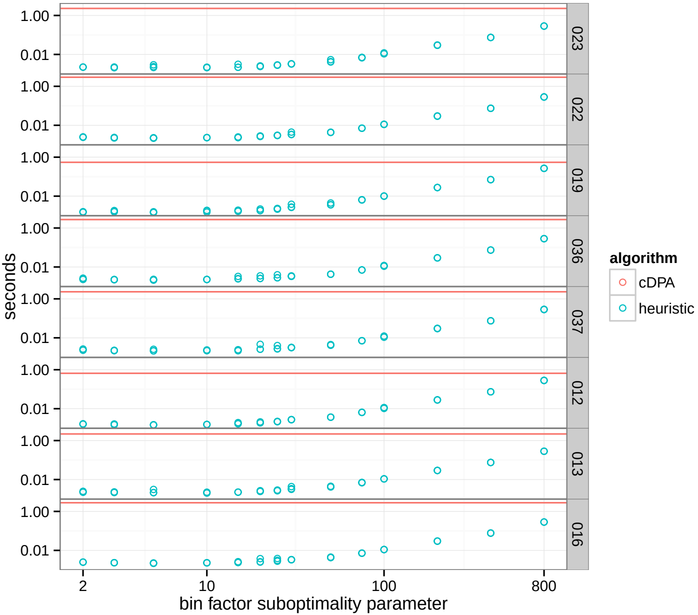


H3K36me3 example data set, PeakSegJoint model with 2 peaks.

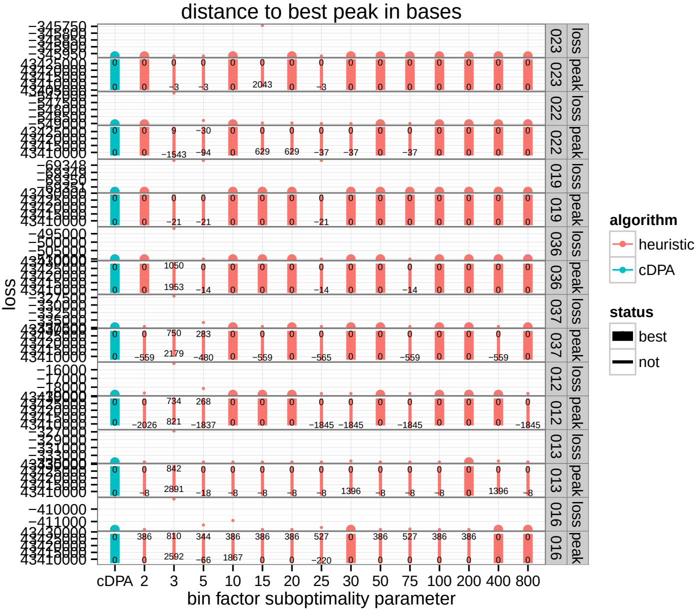
H3K36me3 data, cDPA and heuristic algorithms



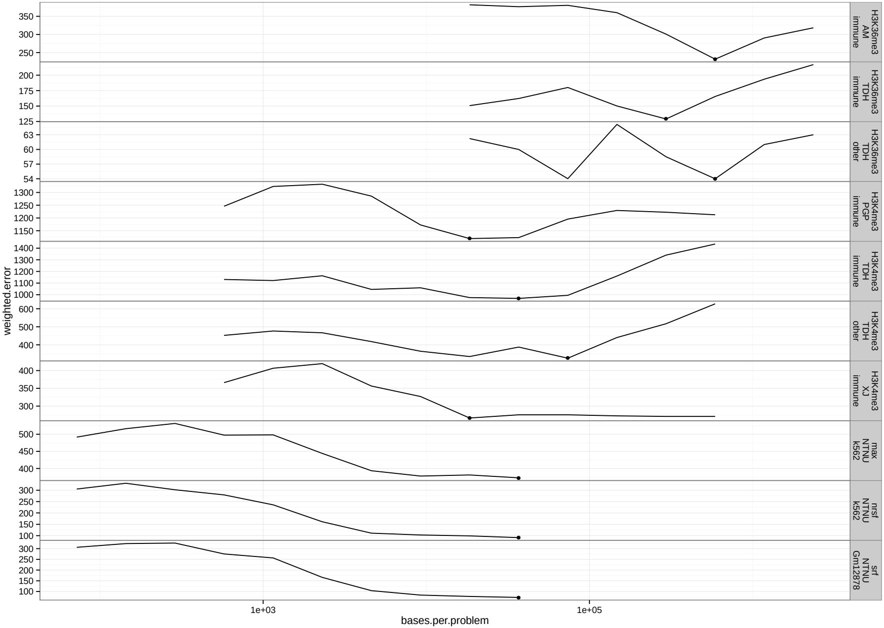
Heuristic is much faster than cDPA



Heuristic often as good as cDPA



Weighted train error not good for model selection



Select L1-regularized model with minimal validation error