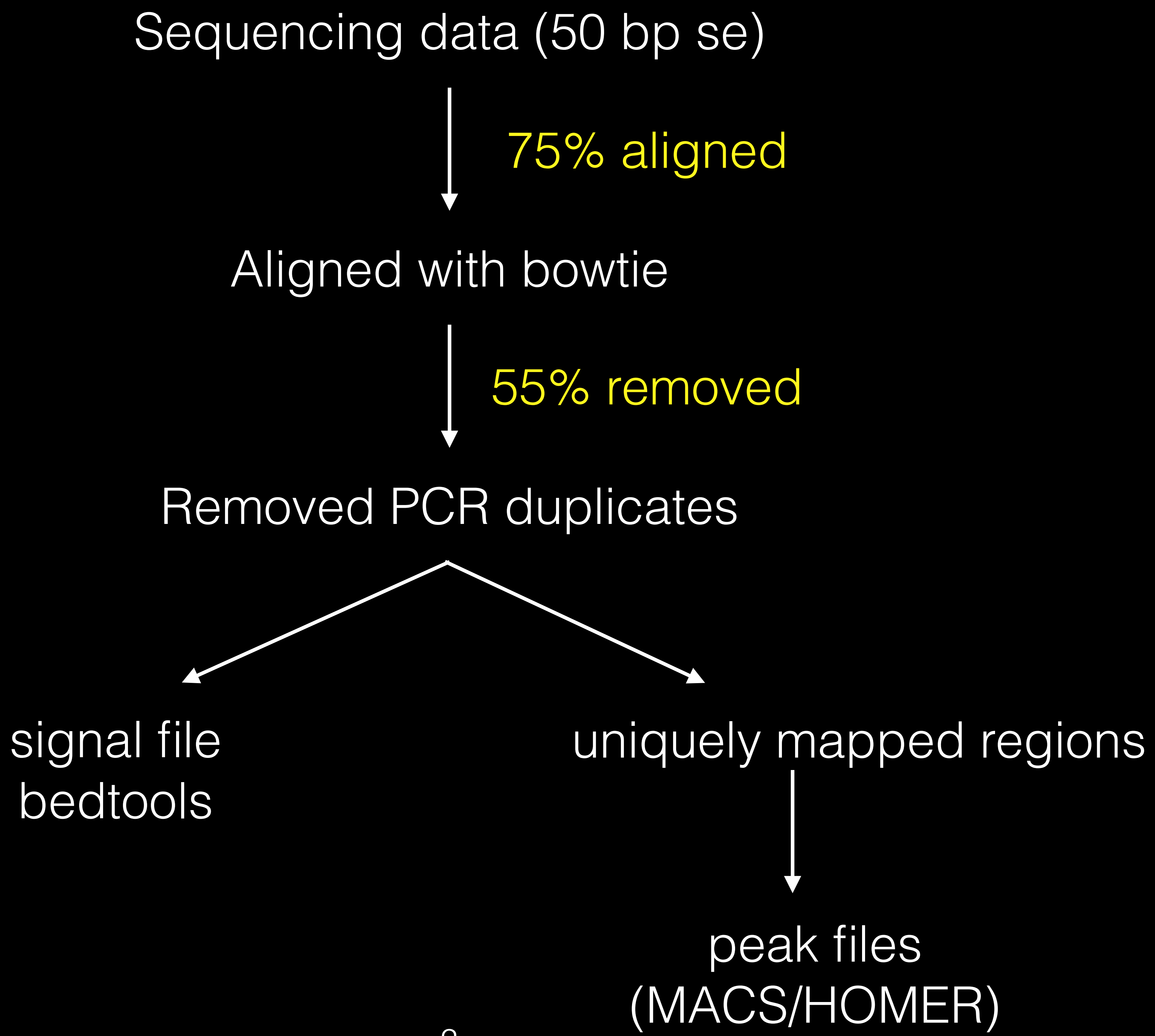
Peak calling for Sutton Assay

Anurag Sethi
May 2015

Pipeline for analyzing each pool



Sequencing data (50 bp se)

75% aligned

Aligned with bowtie

55% removed

Removed PCR duplicates

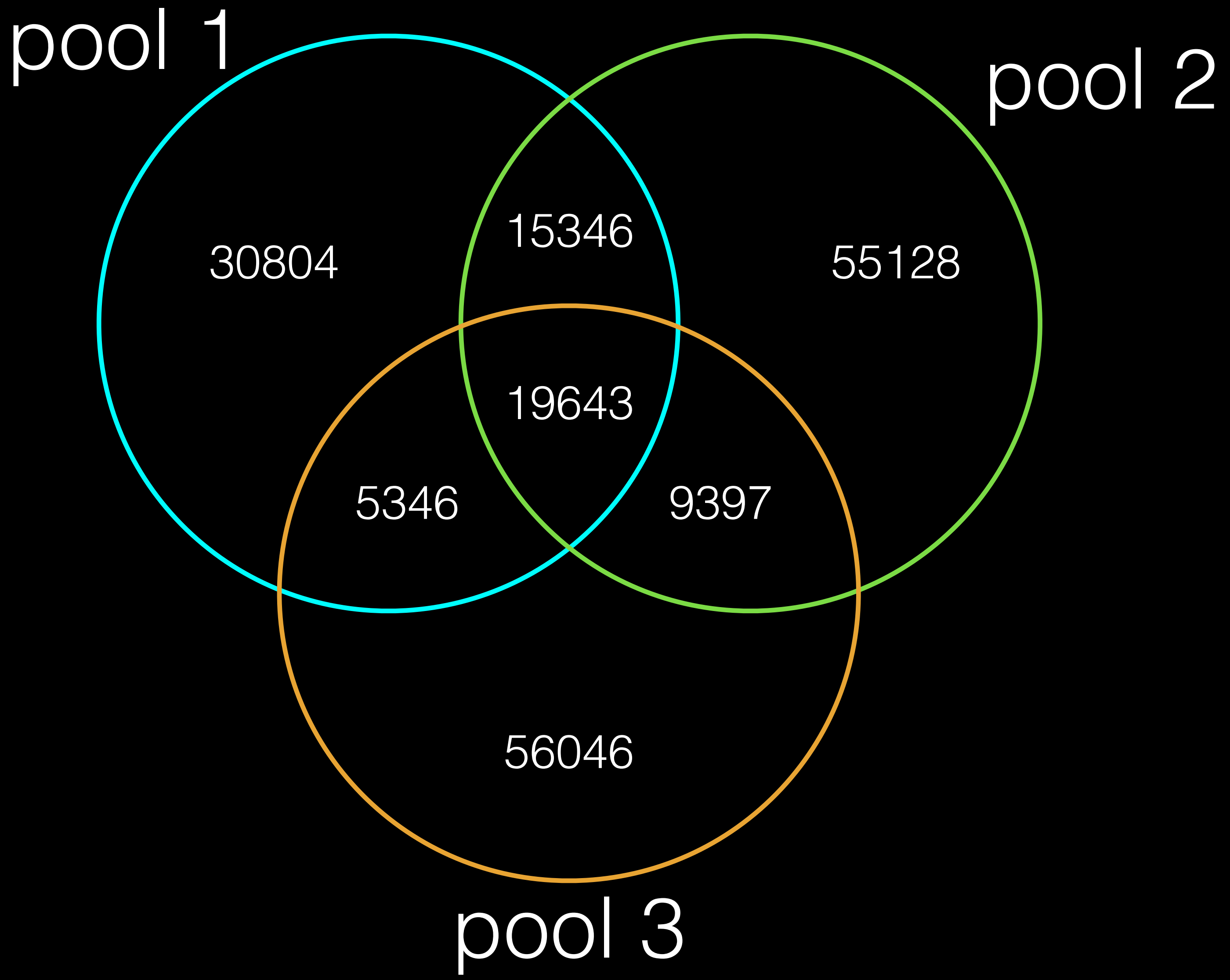
signal file
bedtools

uniquely mapped regions

peak files
(MACS/HOMER)

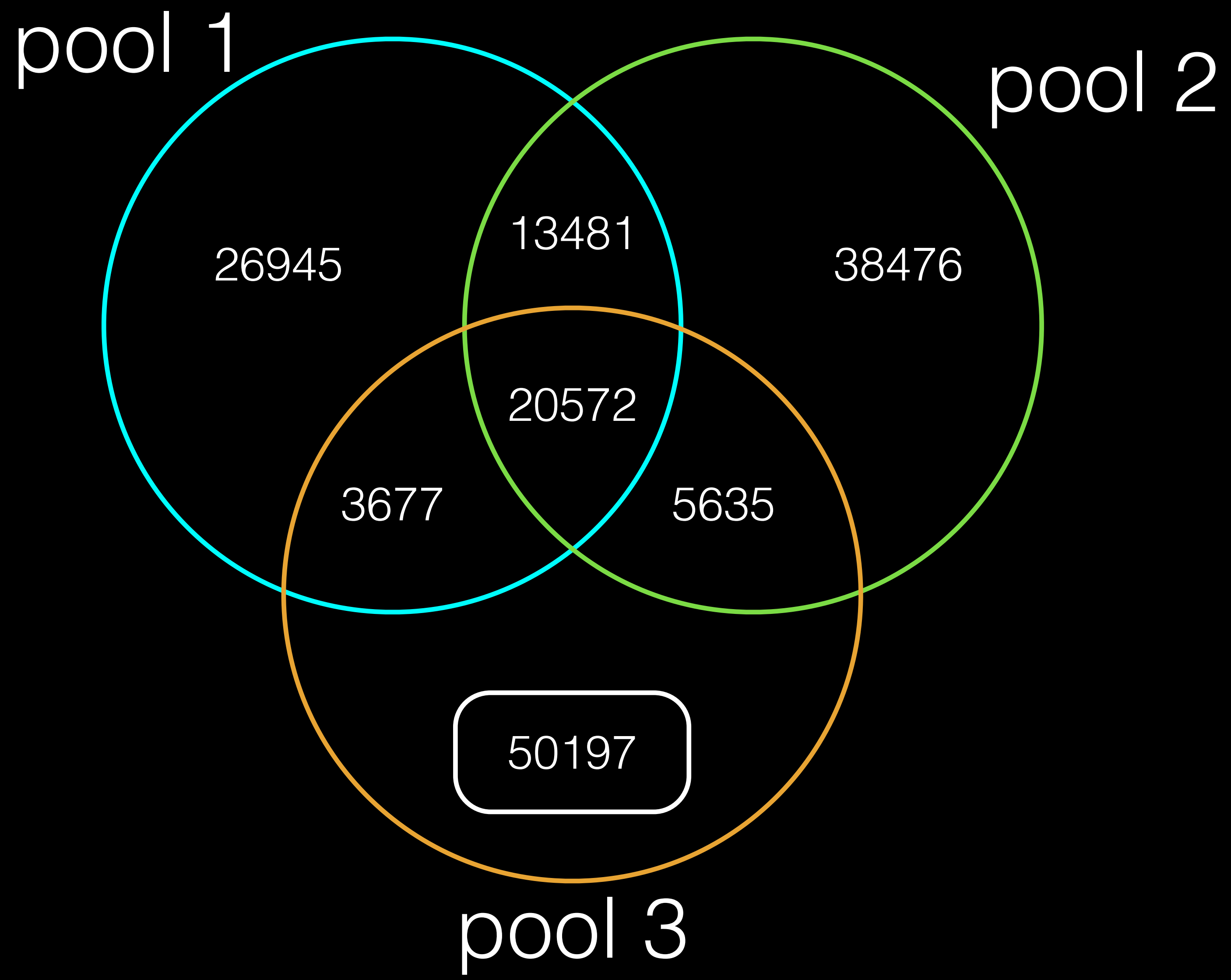
Peak calling might change based on simulations

Comparison of peaks (MACS)



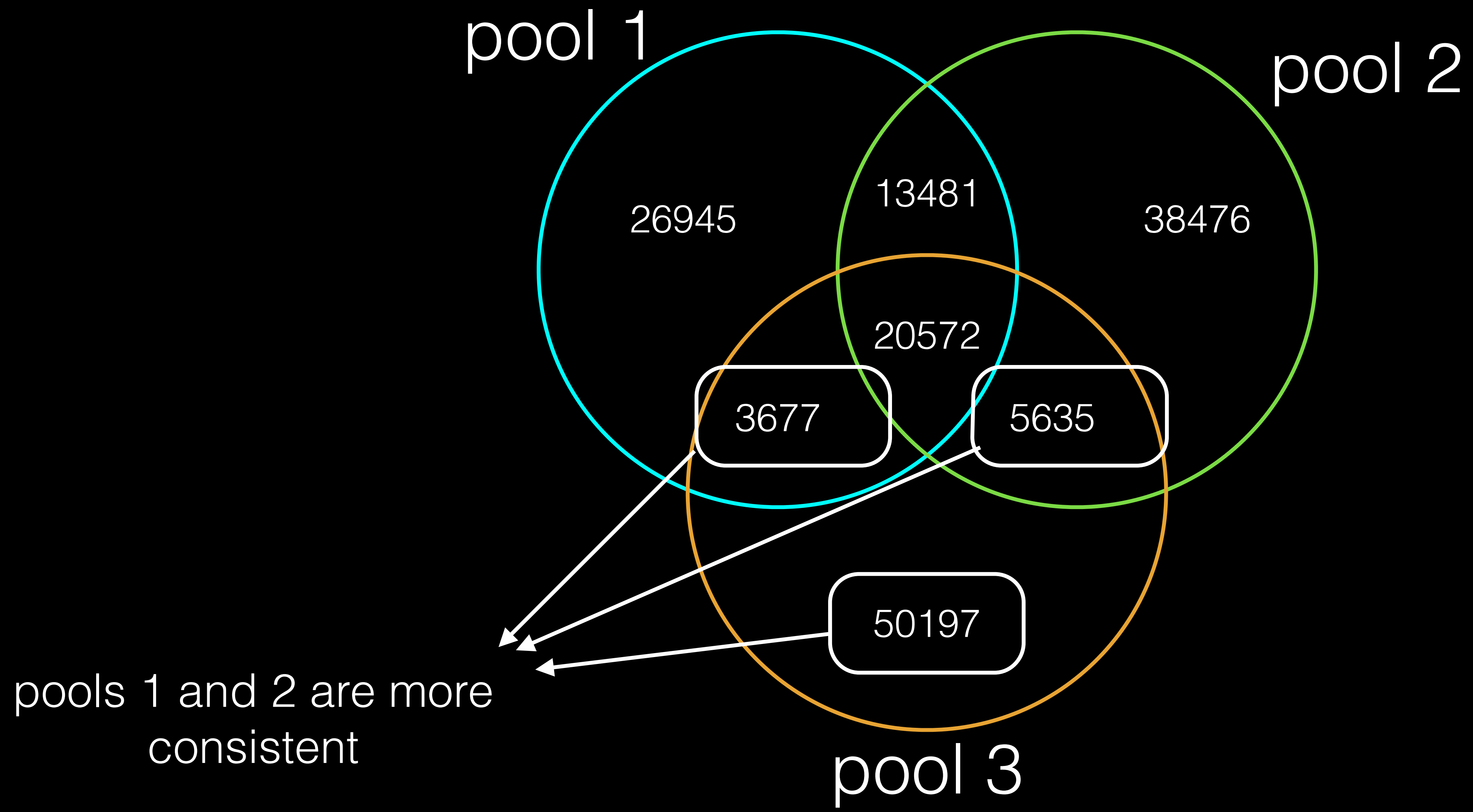
At least 20% overlap of peaks

Comparison of peaks (HOMER)



At least 20% overlap of peaks

Comparison of peaks (HOMER)



At least 20% overlap of peaks

Comparison of peaks (HOMER vs MACS)

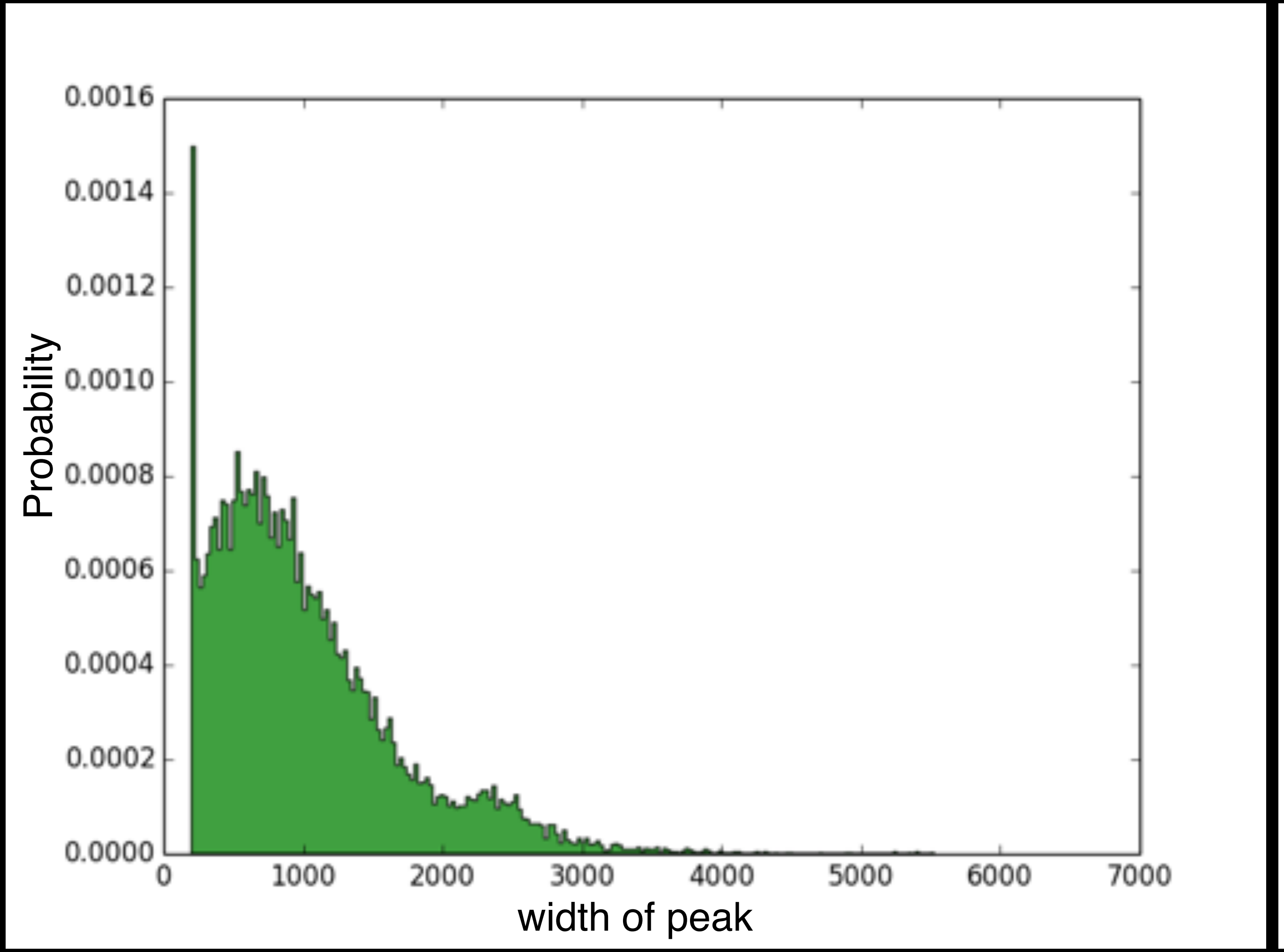
>90% peaks overlap in pools 1 and 2.
67% of peaks overlap in pool 3.

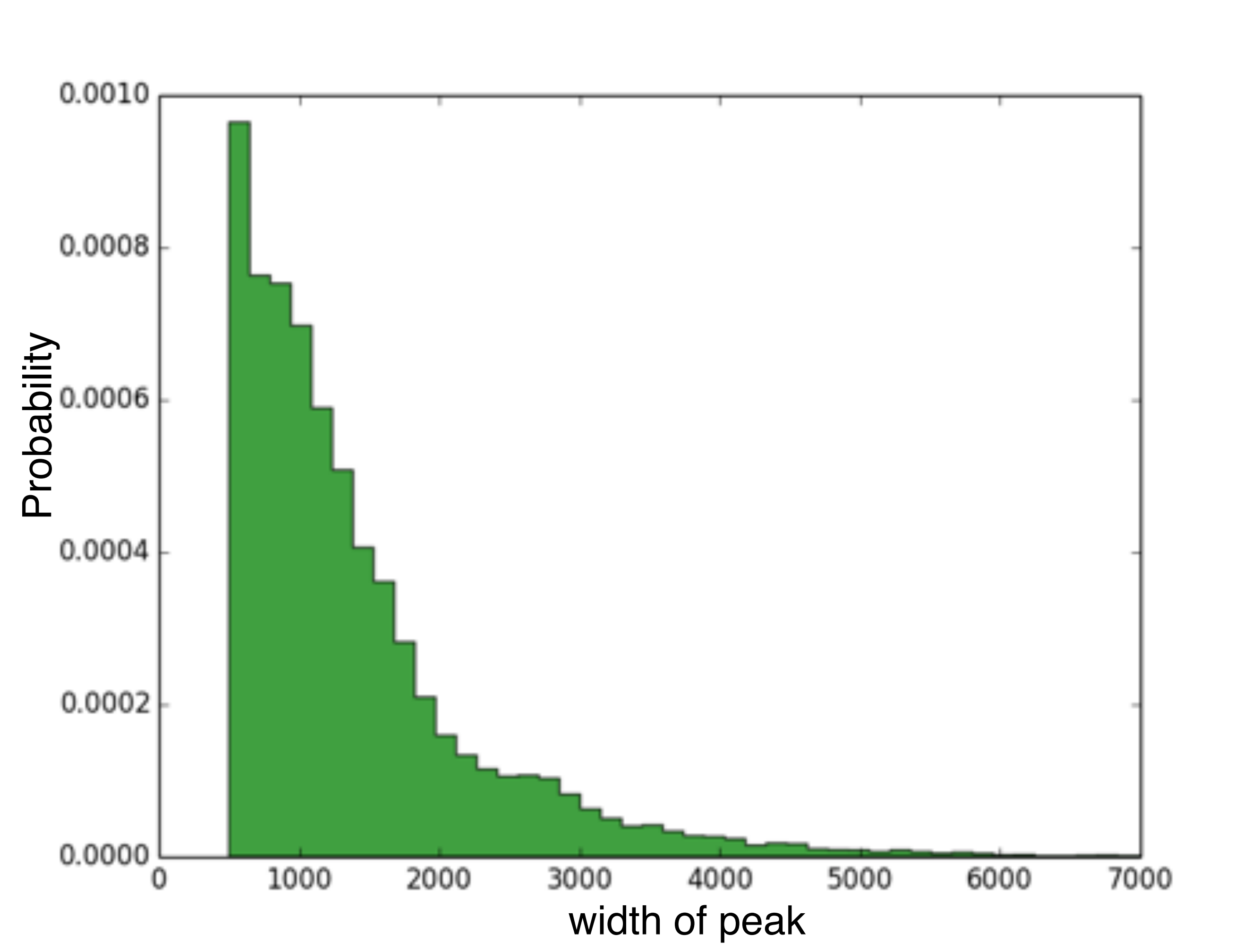
14648/19643 peaks overlap between shared peaks.

Comparison of peaks (new vs old)

10696/20572 HOMER peaks overlap with old peaks.
8683/19643 MACS peaks overlap with old peaks.

Length distribution of peaks



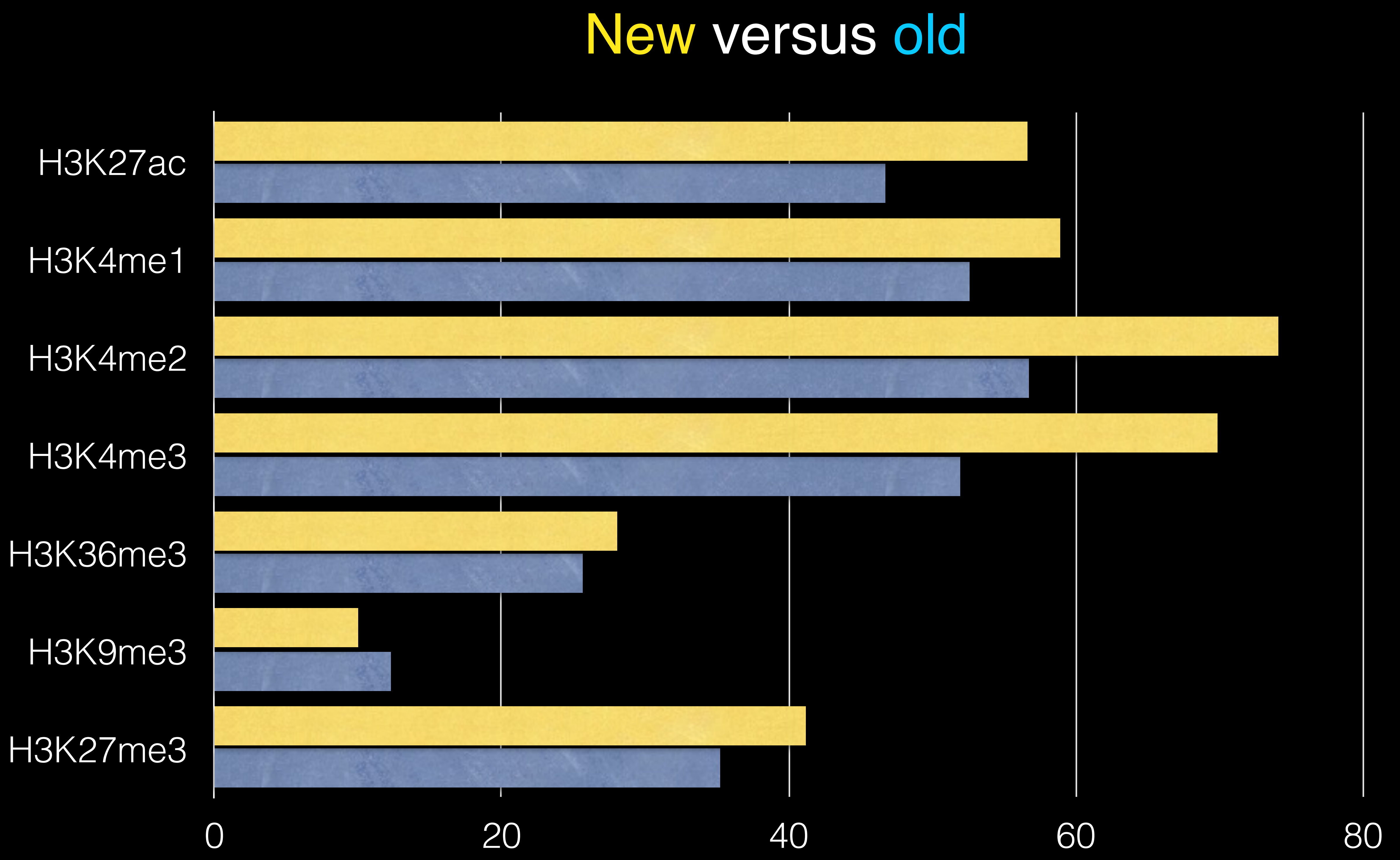


MACS peaks

HOMER peaks

Mean size close to 900-1200 bp width.
HOMER peaks are larger on average (1 peak of 30 kb width)

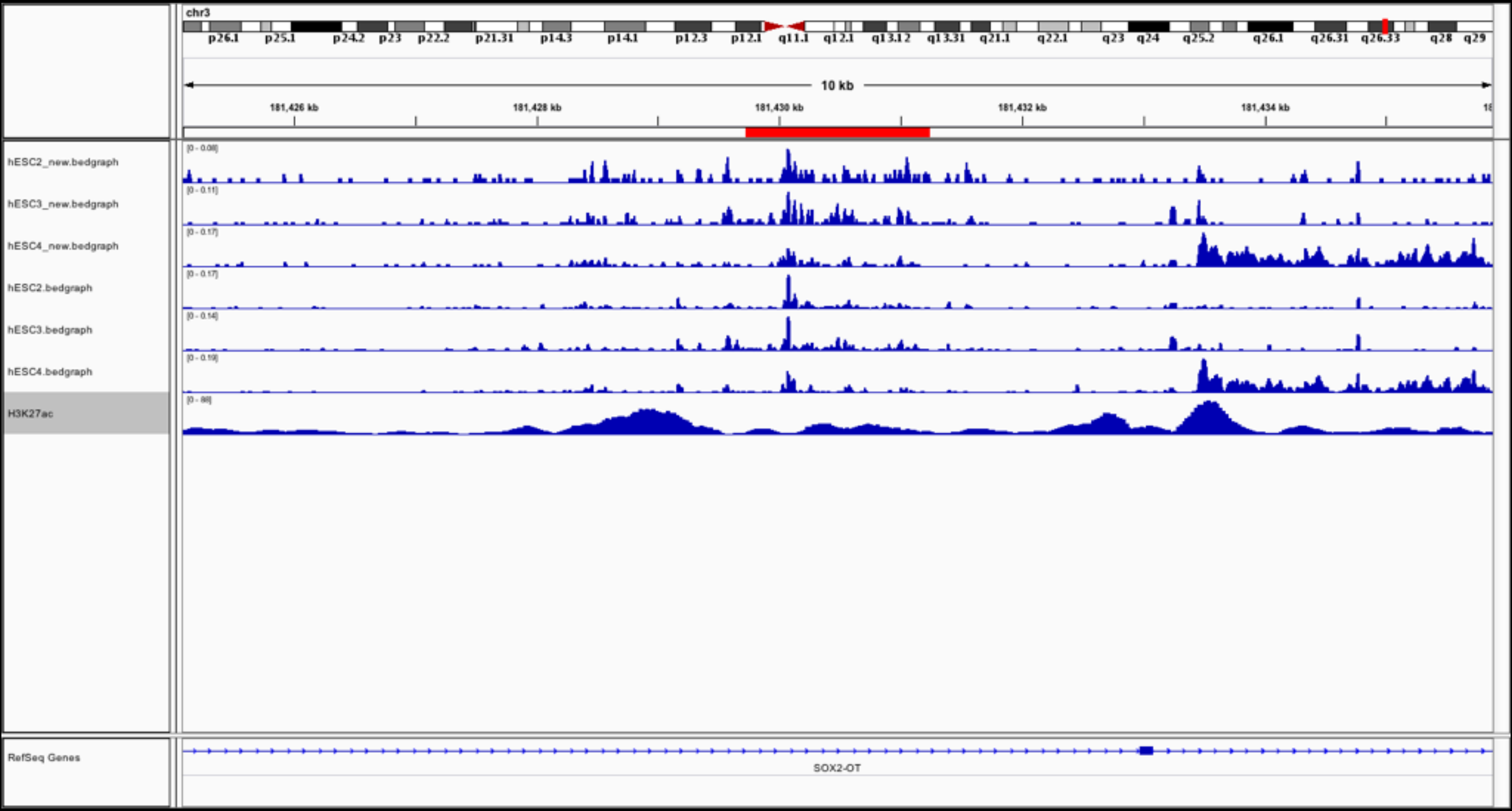
Overlap of shared peaks with histone marks



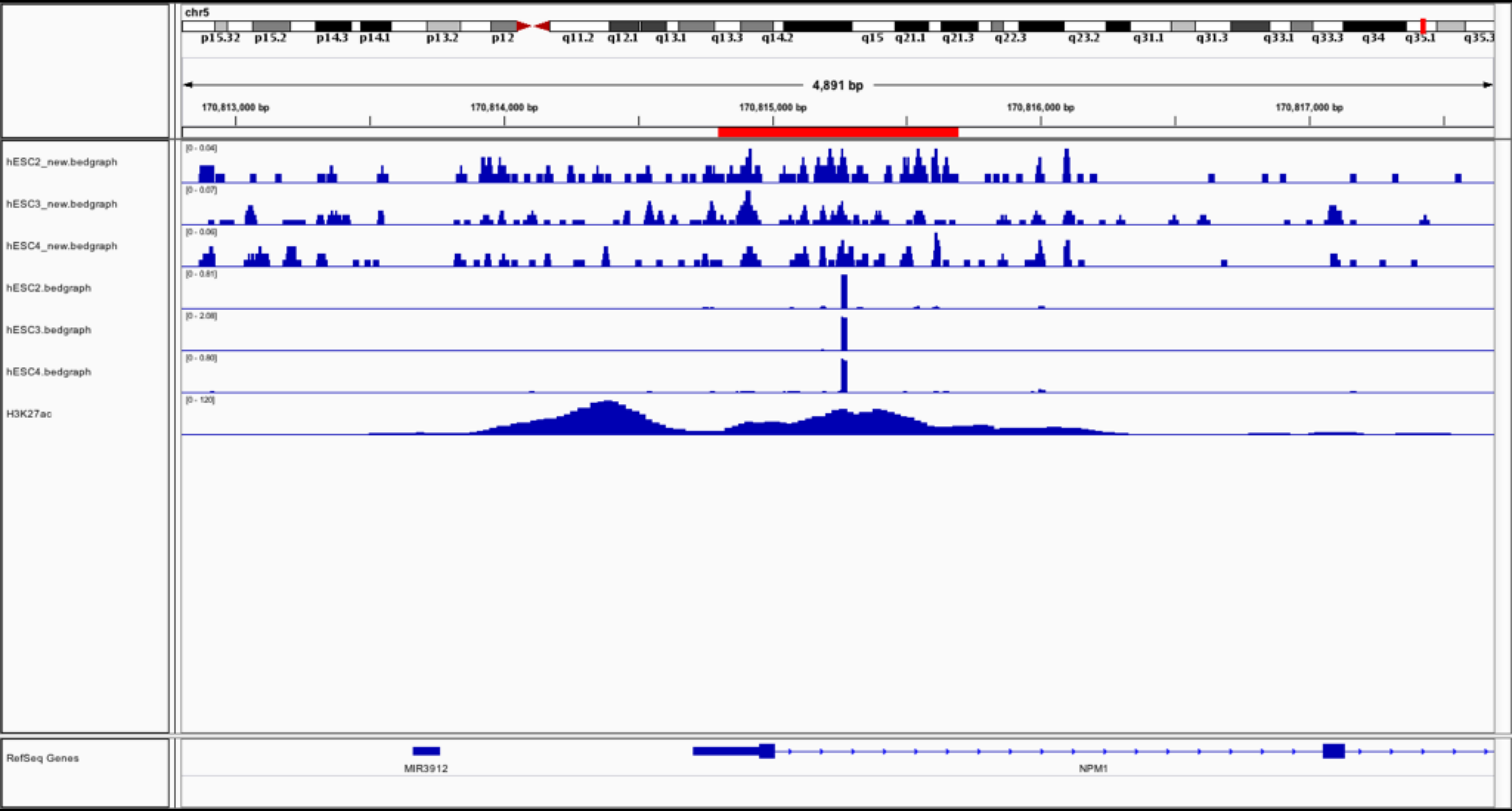
10.3% (21.4%) of peaks outside these histone peaks

Comparison of Signals

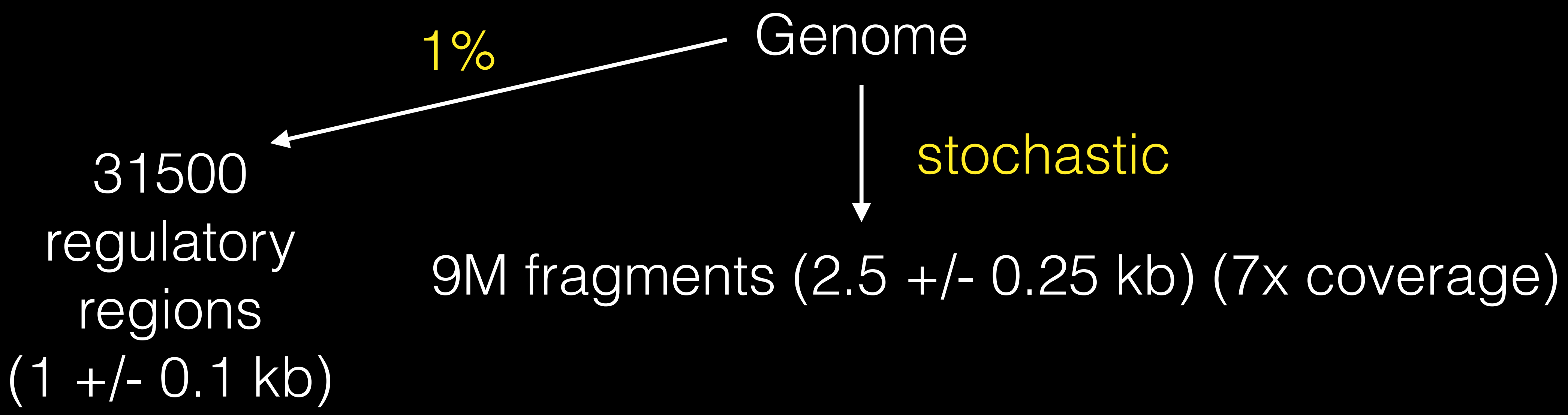
Effect of PCR duplicates on signal



Effect of PCR duplicates on signal



Simulation pipeline (what I have done so far)



Genome

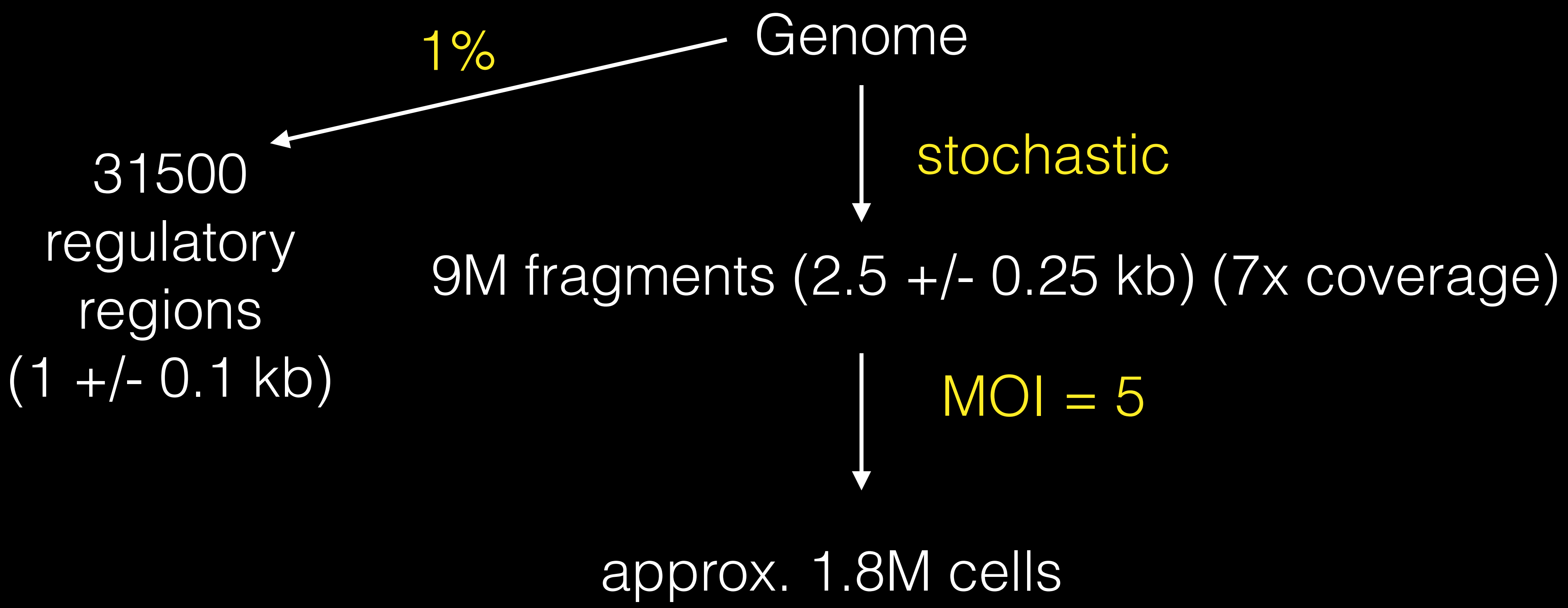
1%

31500 regulatory regions (1 +/- 0.1 kb)

stochastic

9M fragments (2.5 +/- 0.25 kb) (7x coverage)

Simulation pipeline (what I have done so far)



Genome

1%

31500
regulatory
regions
(1 +/- 0.1 kb)

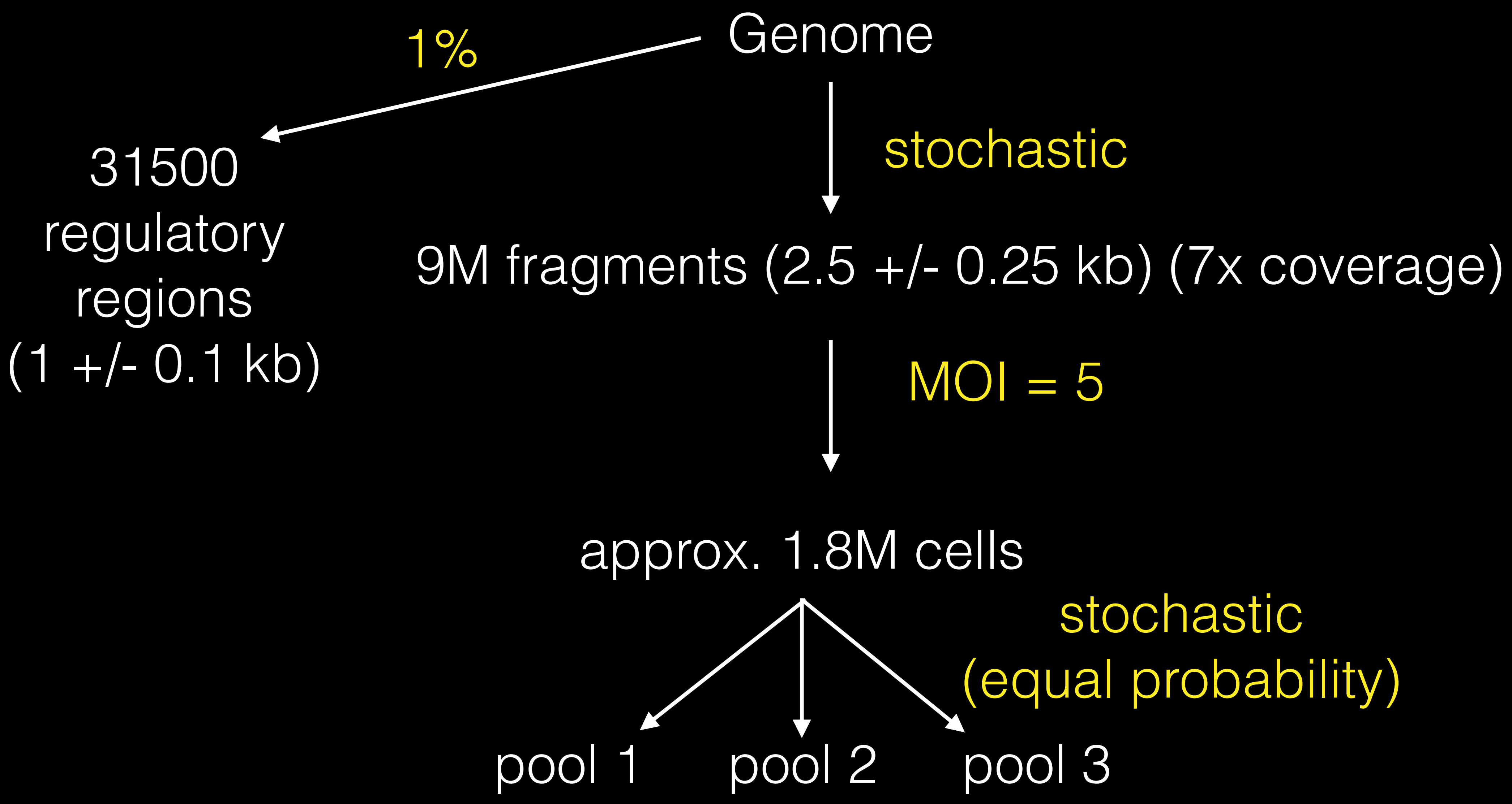
stochastic

9M fragments (2.5 +/- 0.25 kb) (7x coverage)

MOI = 5

approx. 1.8M cells

Simulation pipeline (what I have done so far)



Genome

1%

31500 regulatory regions (1 +/- 0.1 kb)

stochastic

9M fragments (2.5 +/- 0.25 kb) (7x coverage)

MOI = 5

approx. 1.8M cells

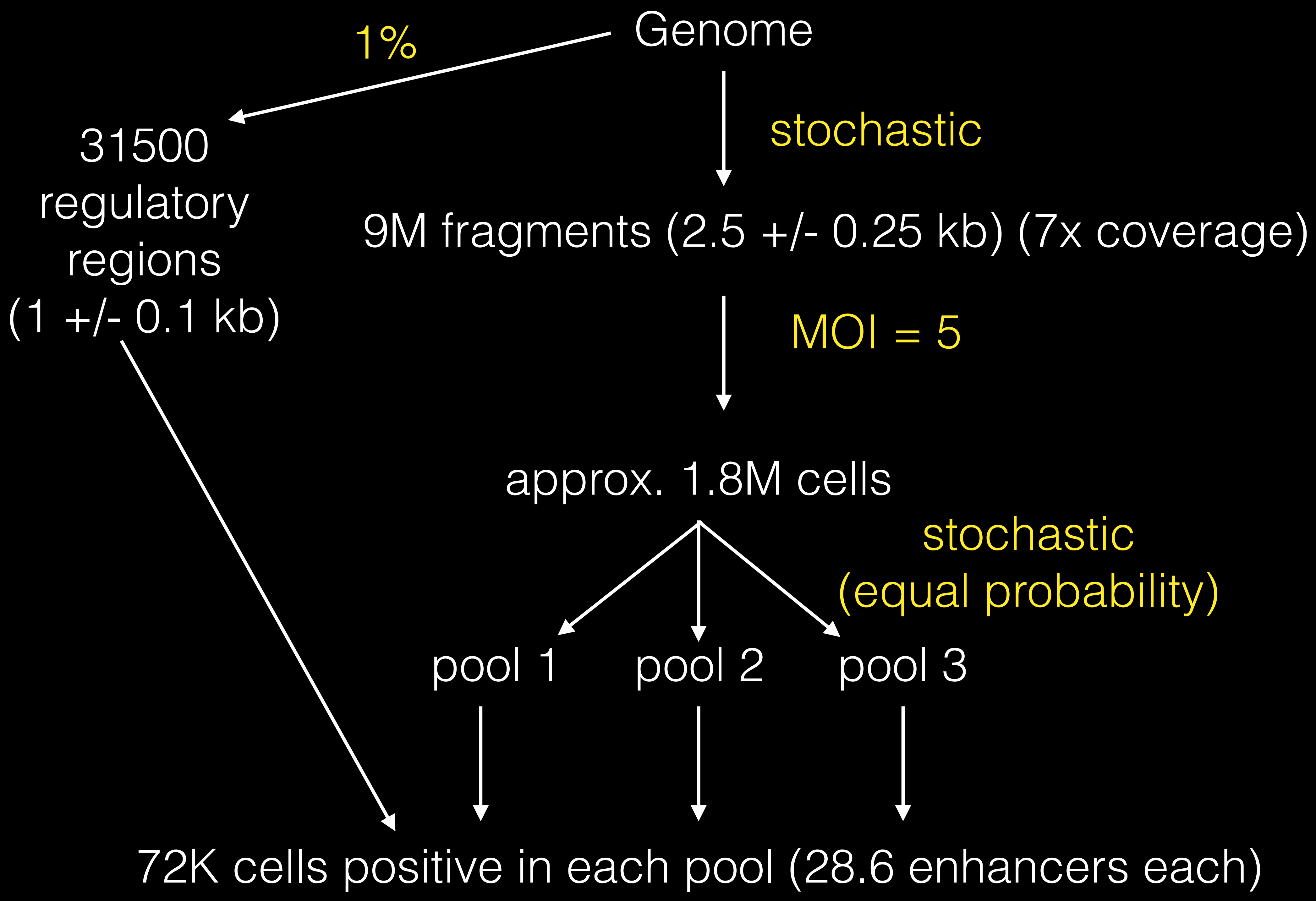
stochastic (equal probability)

pool 1

pool 2

pool 3

Simulation pipeline (what I have done so far)



Genome

1%

31500 regulatory regions (1 +/- 0.1 kb)

stochastic

9M fragments (2.5 +/- 0.25 kb) (7x coverage)

MOI = 5

approx. 1.8M cells

stochastic (equal probability)

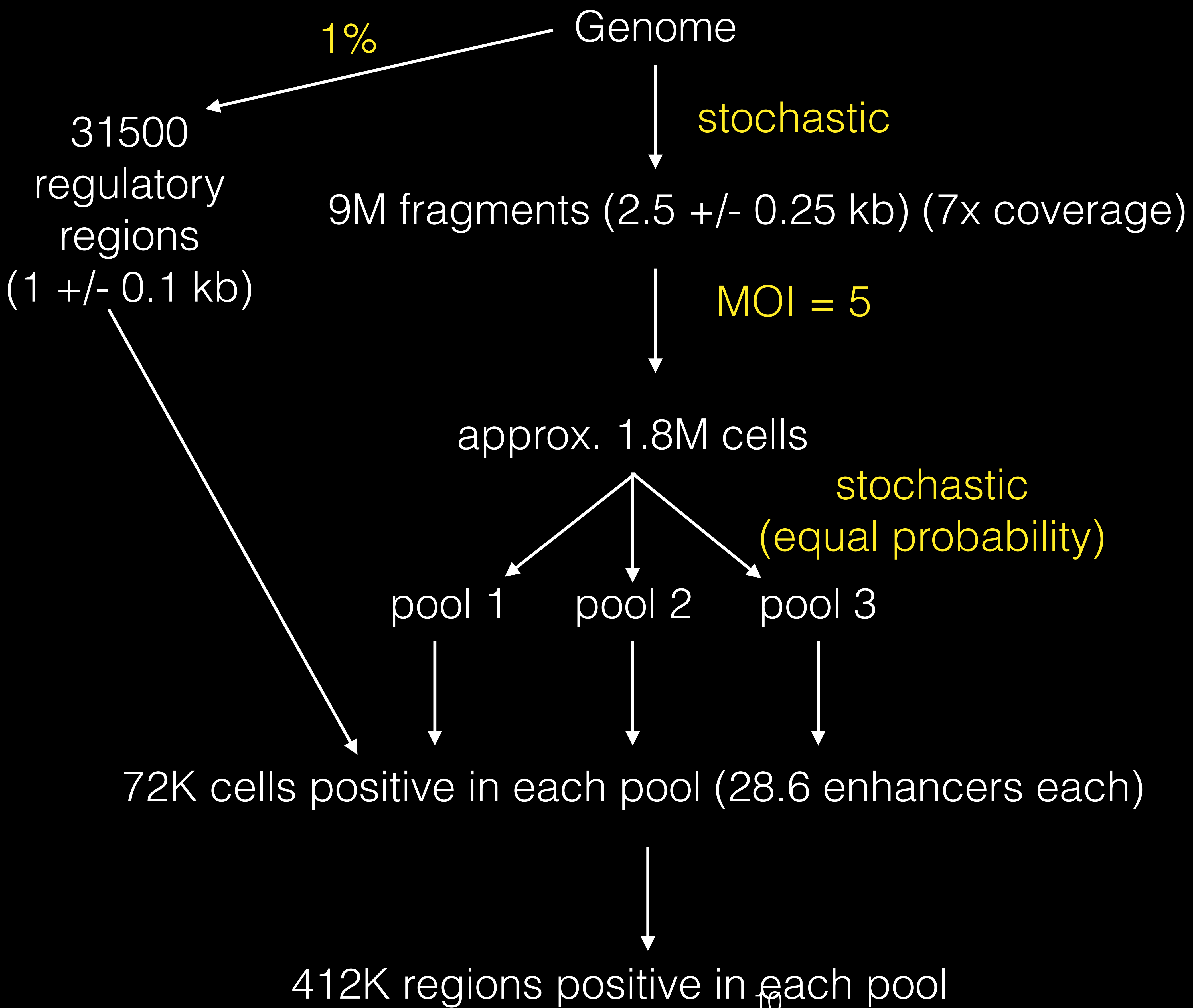
pool 1

pool 2

pool 3

72K cells positive in each pool (28.6 enhancers each)

Simulation pipeline (what I have done so far)



Genome

1%

31500 regulatory regions (1 +/- 0.1 kb)

stochastic

9M fragments (2.5 +/- 0.25 kb) (7x coverage)

MOI = 5

approx. 1.8M cells

stochastic (equal probability)

pool 1

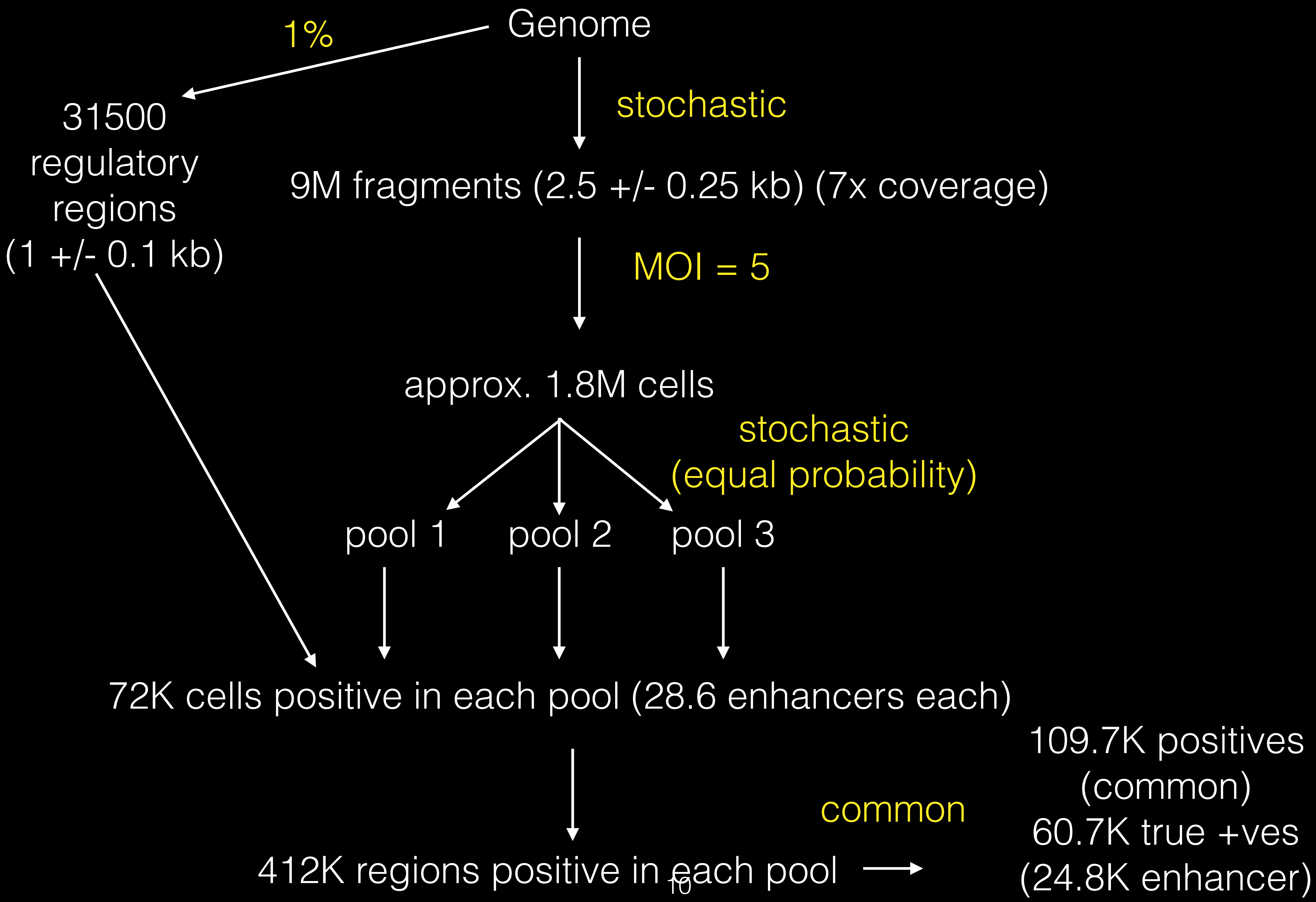
pool 2

pool 3

72K cells positive in each pool (28.6 enhancers each)

412K regions positive in each pool

Simulation pipeline (what I have done so far)



Genome

1%

31500 regulatory regions (1 +/- 0.1 kb)

stochastic

9M fragments (2.5 +/- 0.25 kb) (7x coverage)

MOI = 5

approx. 1.8M cells

stochastic (equal probability)

pool 1

pool 2

pool 3

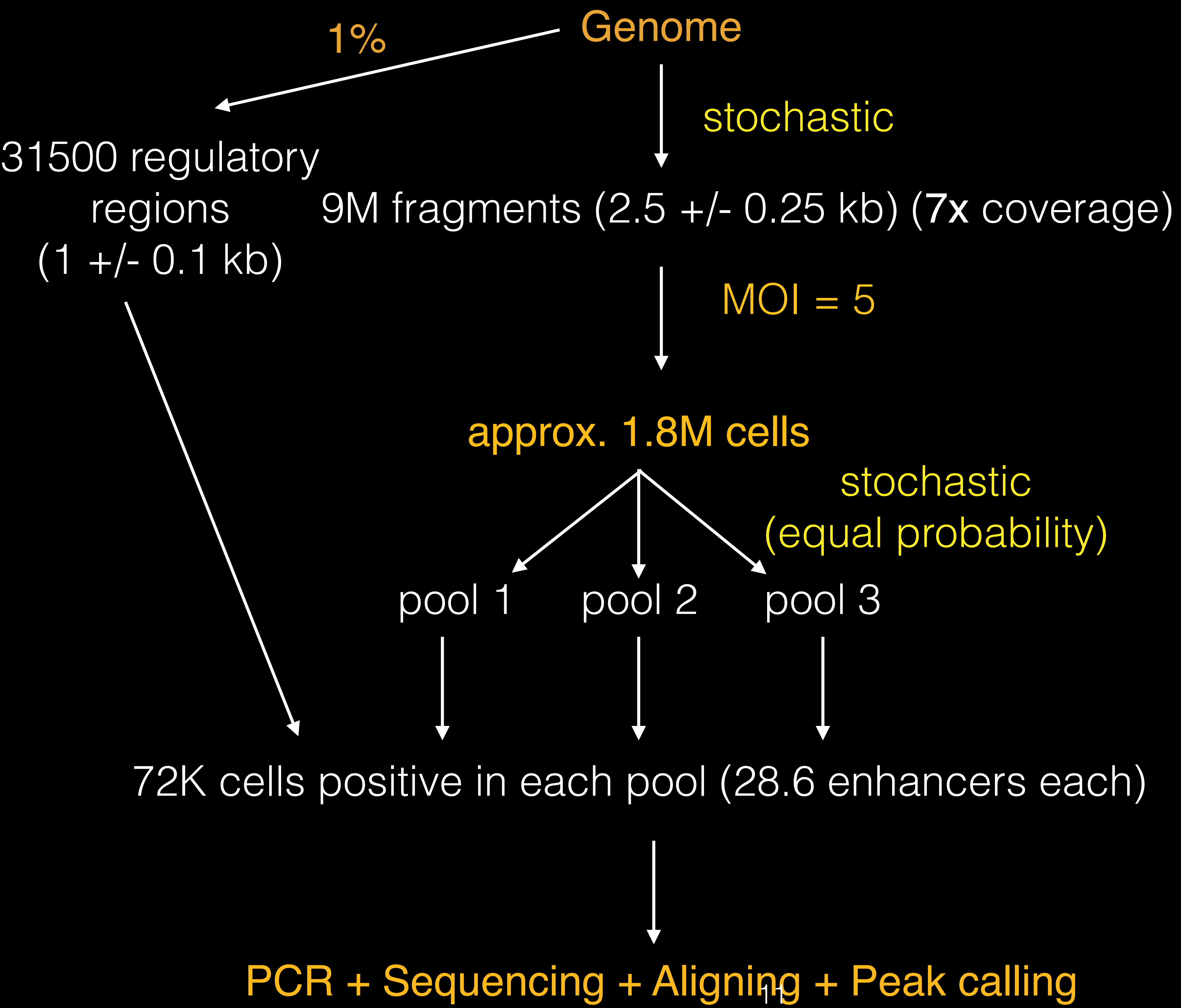
72K cells positive in each pool (28.6 enhancers each)

412K regions positive in each pool

common

109.7K positives (common)
60.7K true +ves
(24.8K enhancer)

Improvements in pipeline



Genome

1%

31500 regulatory regions (1 +/- 0.1 kb)

stochastic

9M fragments (2.5 +/- 0.25 kb) (**7x** coverage)

MOI = 5

approx. 1.8M cells

stochastic (equal probability)

pool 1

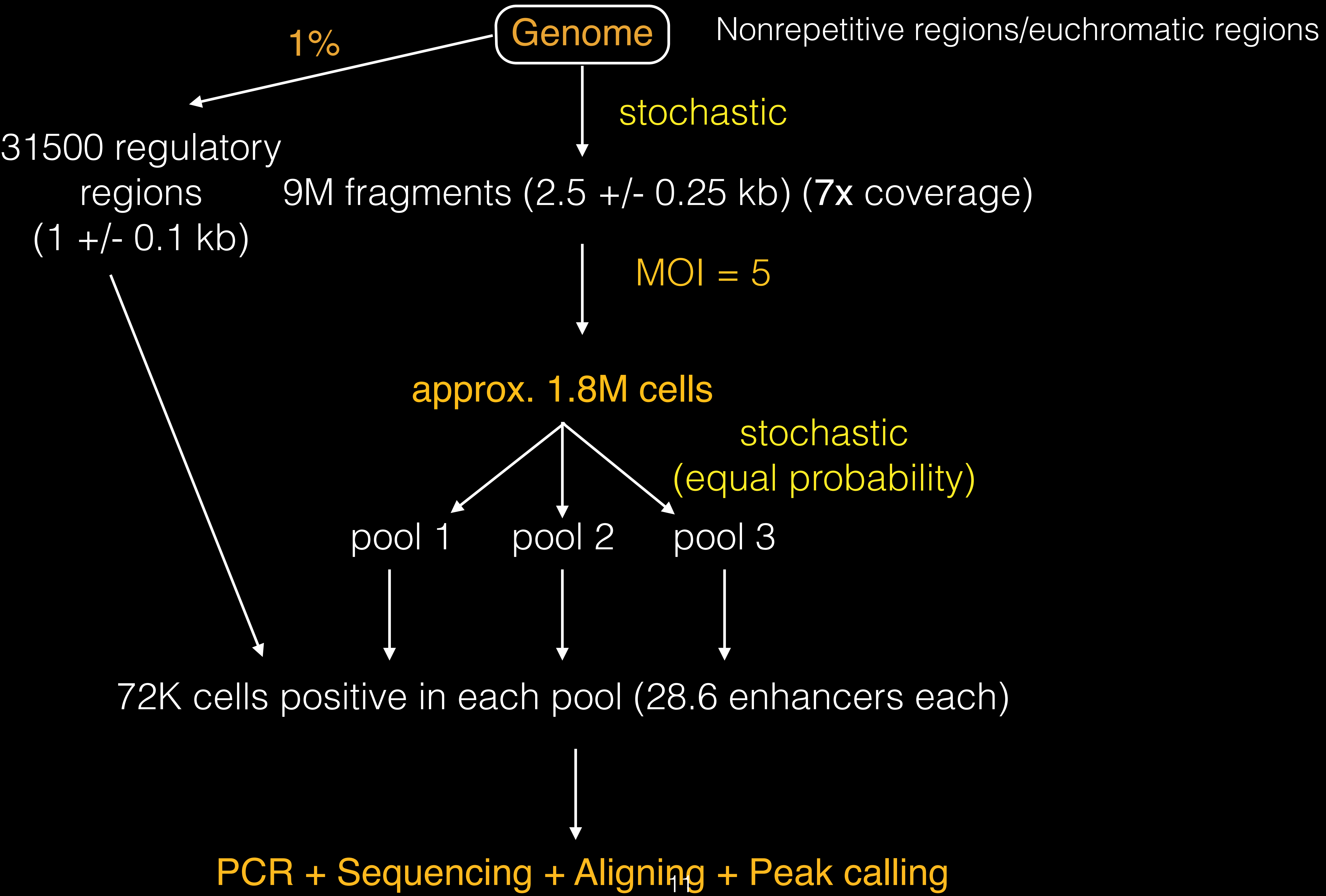
pool 2

pool 3

72K cells positive in each pool (28.6 enhancers each)

PCR + Sequencing + Aligning + Peak calling

Improvements in pipeline



Genome

1%

Nonrepetitive regions/euchromatic regions

stochastic

31500 regulatory regions (1 +/- 0.1 kb)

9M fragments (2.5 +/- 0.25 kb) (**7x** coverage)

MOI = 5

approx. 1.8M cells

stochastic (equal probability)

pool 1

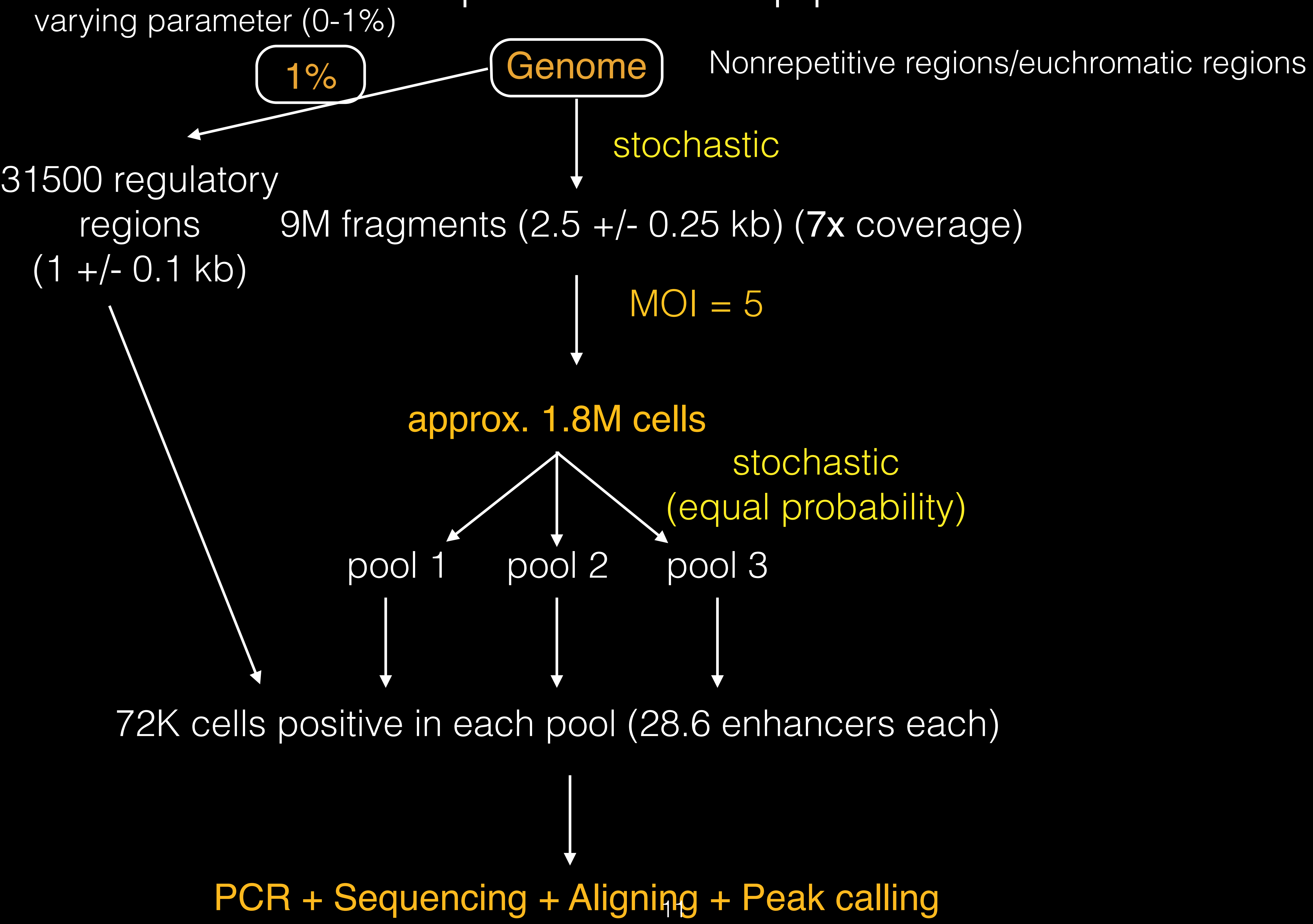
pool 2

pool 3

72K cells positive in each pool (28.6 enhancers each)

PCR + Sequencing + Aligning + Peak calling

Improvements in pipeline



varying parameter (0-1%)

1%

Genome

Nonrepetitive regions/euchromatic regions

stochastic

31500 regulatory
regions
(1 +/- 0.1 kb)

9M fragments (2.5 +/- 0.25 kb) (**7x** coverage)

MOI = 5

approx. 1.8M cells

stochastic
(equal probability)

pool 1

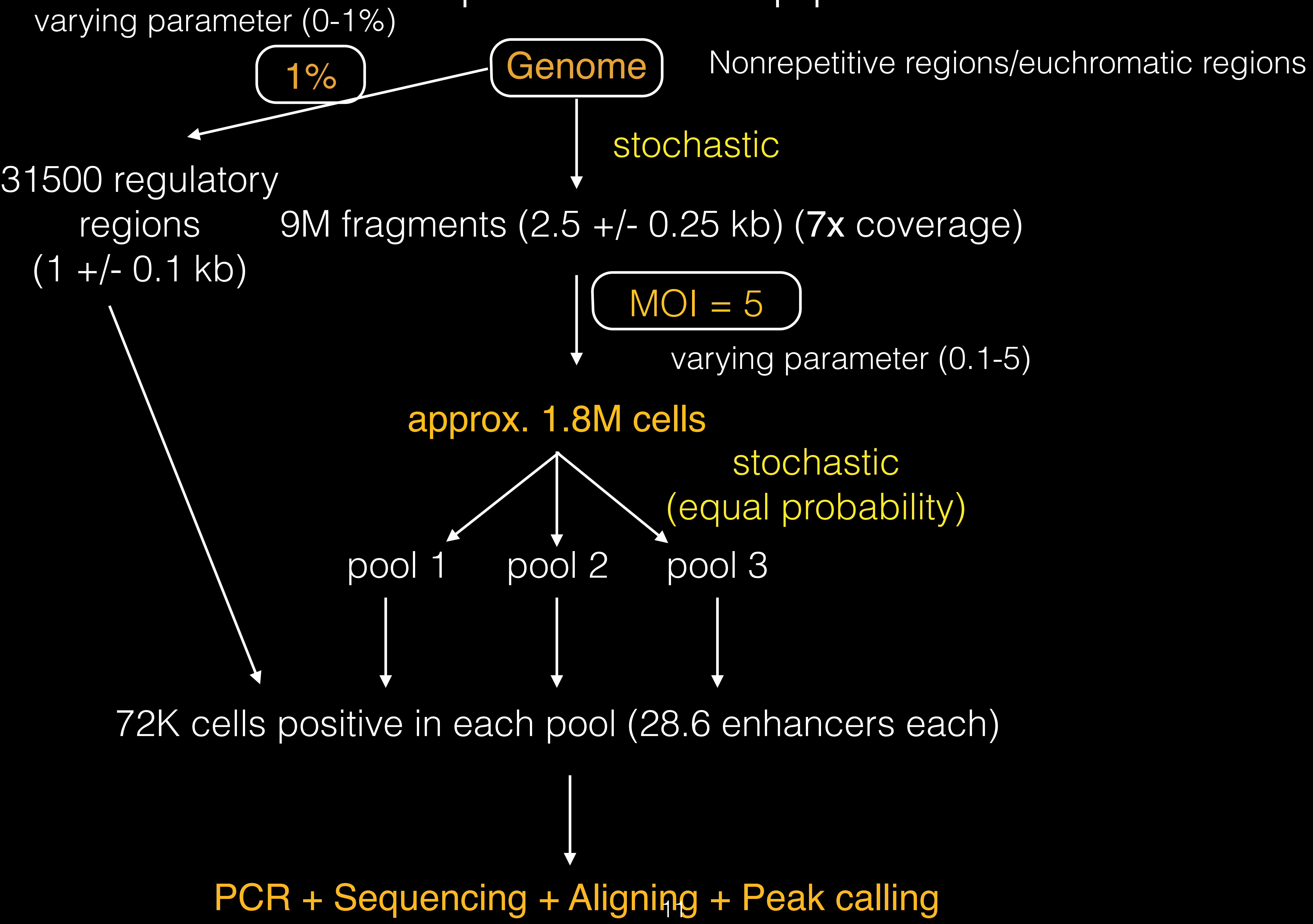
pool 2

pool 3

72K cells positive in each pool (28.6 enhancers each)

PCR + Sequencing + Aligning + Peak calling

Improvements in pipeline



varying parameter (0-1%)

1%

Genome

Nonrepetitive regions/euchromatic regions

stochastic

31500 regulatory regions (1 +/- 0.1 kb)

9M fragments (2.5 +/- 0.25 kb) (**7x** coverage)

MOI = 5

varying parameter (0.1-5)

approx. 1.8M cells

stochastic (equal probability)

pool 1

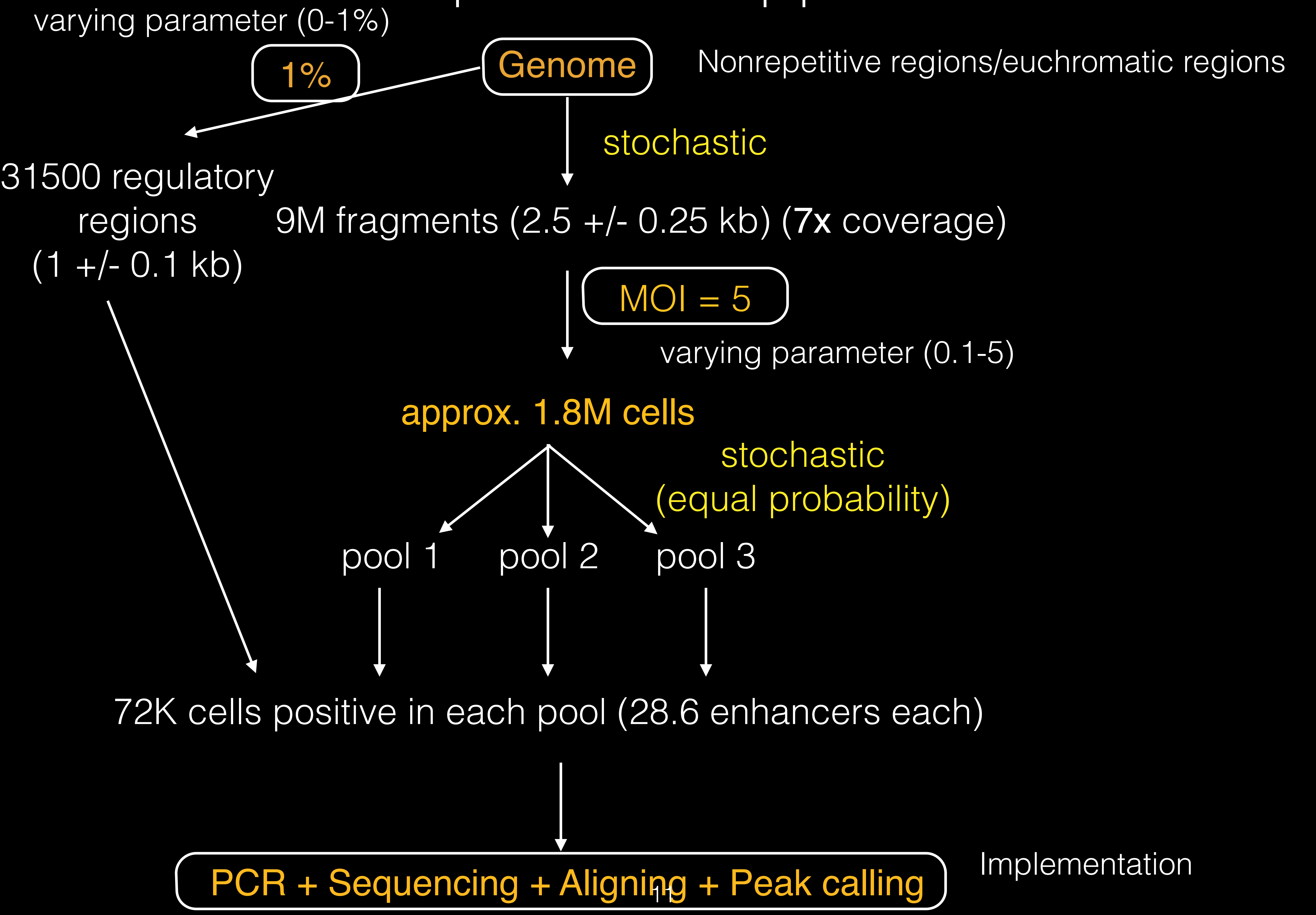
pool 2

pool 3

72K cells positive in each pool (28.6 enhancers each)

PCR + Sequencing + Aligning + Peak calling

Improvements in pipeline



varying parameter (0-1%)

1%

Genome

Nonrepetitive regions/euchromatic regions

stochastic

31500 regulatory regions (1 +/- 0.1 kb)

9M fragments (2.5 +/- 0.25 kb) (**7x** coverage)

MOI = 5

varying parameter (0.1-5)

approx. 1.8M cells

stochastic (equal probability)

pool 1

pool 2

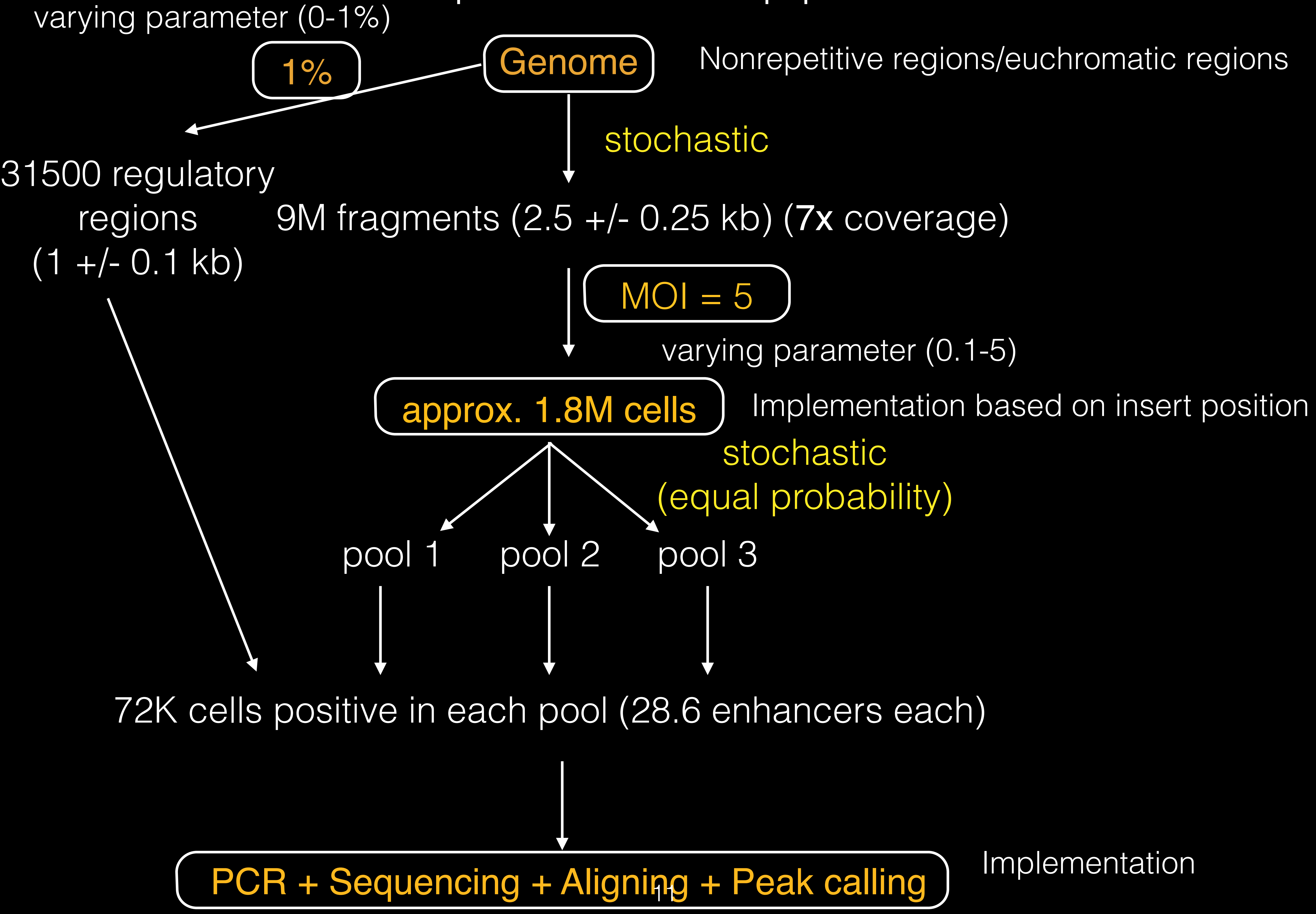
pool 3

72K cells positive in each pool (28.6 enhancers each)

PCR + Sequencing + Aligning + Peak calling

Implementation

Improvements in pipeline



varying parameter (0-1%)

1%

Genome

Nonrepetitive regions/euchromatic regions

stochastic

31500 regulatory regions (1 +/- 0.1 kb)

9M fragments (2.5 +/- 0.25 kb) (**7x** coverage)

MOI = 5

varying parameter (0.1-5)

approx. 1.8M cells

Implementation based on insert position

stochastic (equal probability)

pool 1

pool 2

pool 3

72K cells positive in each pool (28.6 enhancers each)

PCR + Sequencing + Aligning + Peak calling

Implementation

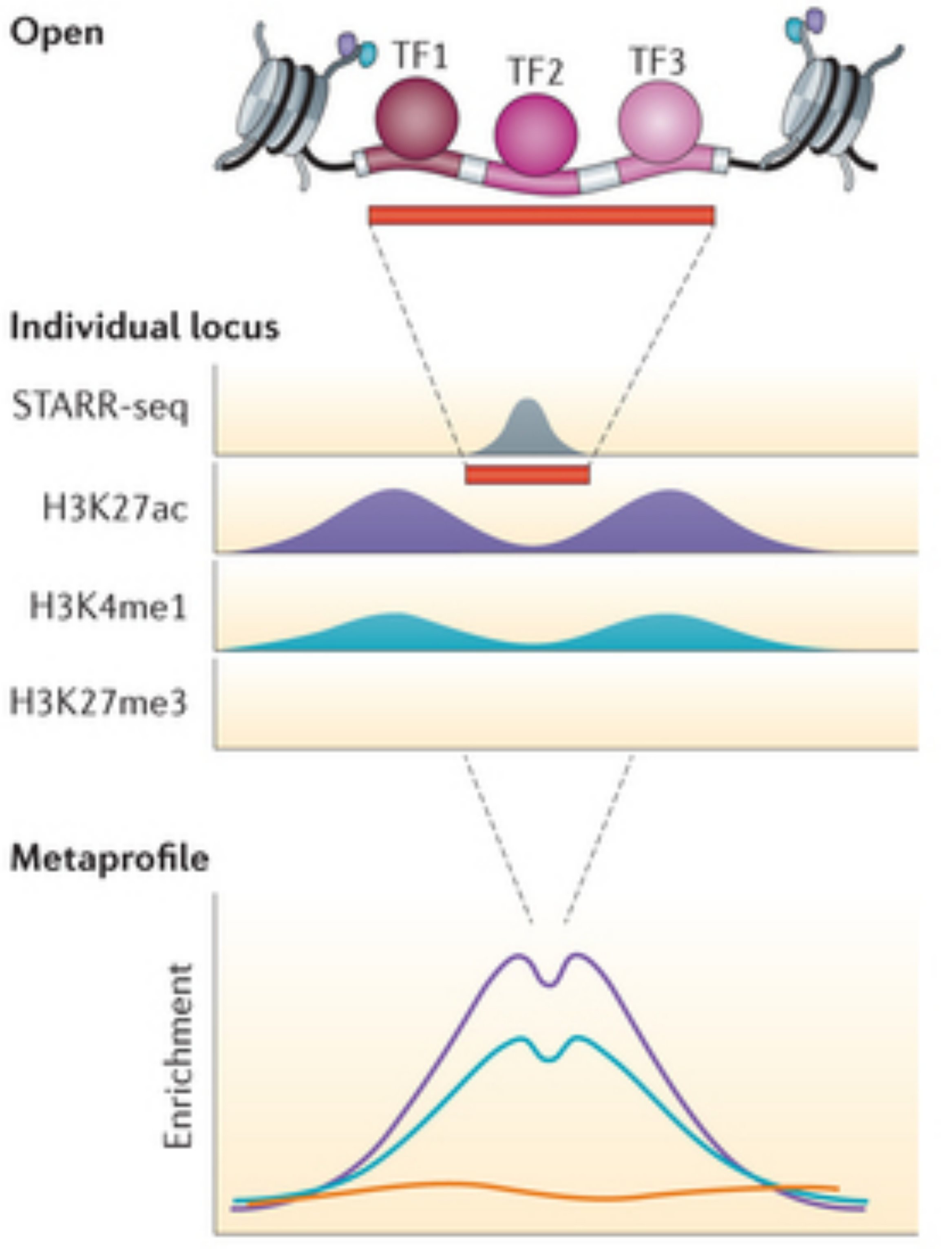
Future Work

Most of the work is related to the simulation method.
This can probably help decide threshold (5% FDR currently) for peak calling.

Postprocessing Analysis:
TF binding peaks on enhancers/promoters to better define enhancer elements.
TF motifs on enhancers/promoters.
Does this assay work better with certain kinds of promoters/enhancers?

Predicting Enhances using Signal Processing

Epigenetic signatures associated with active enhancers

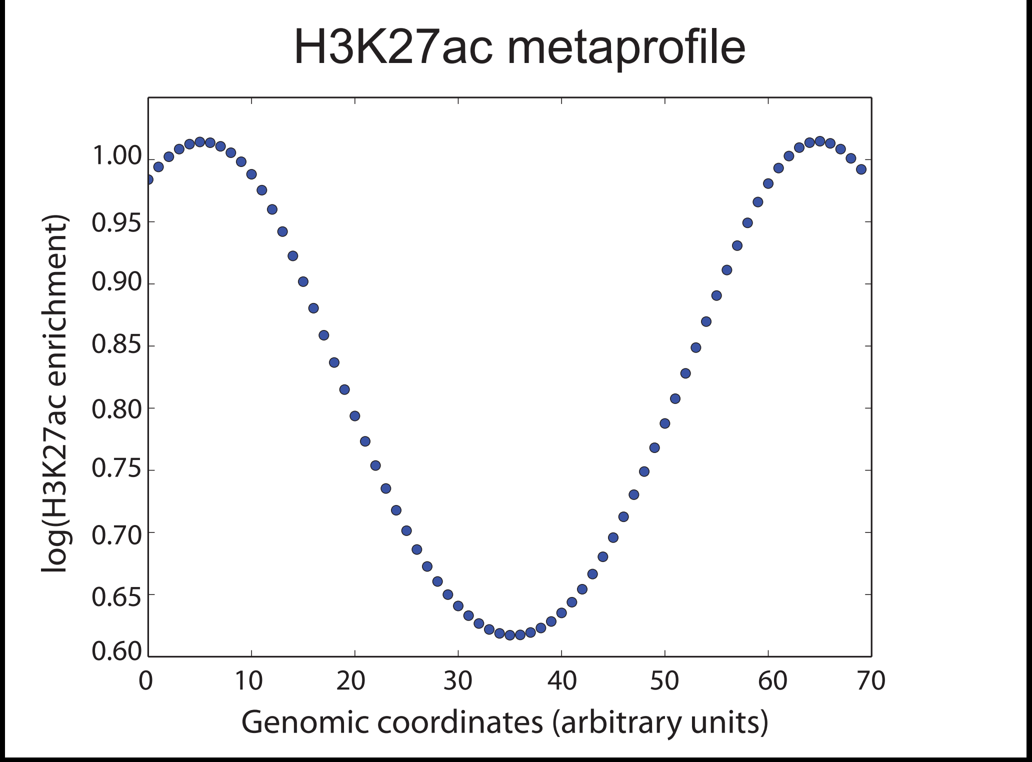


Development of massively parallel assays for enhancer activity have identified the epigenetic signatures associated with active enhancers.

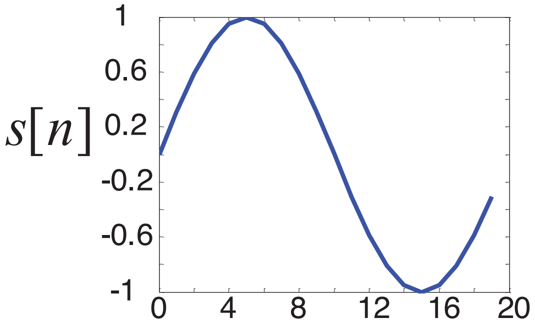
The trough in the histone modification peaks are due to open chromatin.

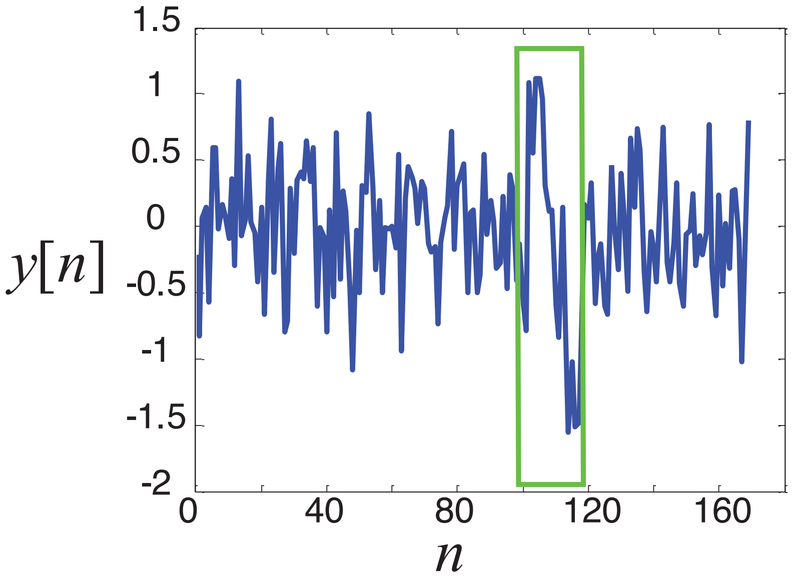
Shyueva, Stampfel, and Stark, *Nat. Rev. Genet.*, 2014

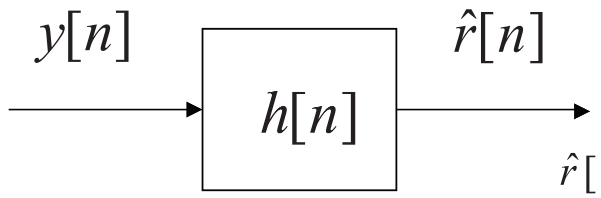
Matching pattern of histone signal could be used to predict enhancers

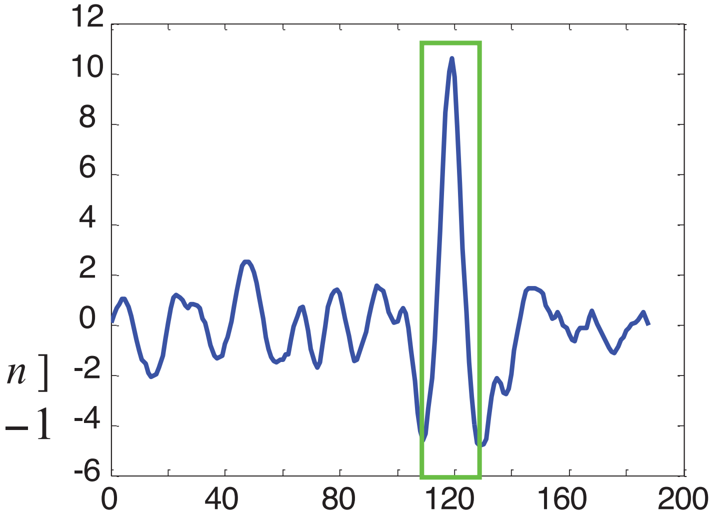


Match filter can be used to identify the occurrence of the chromatin pattern in the genome.







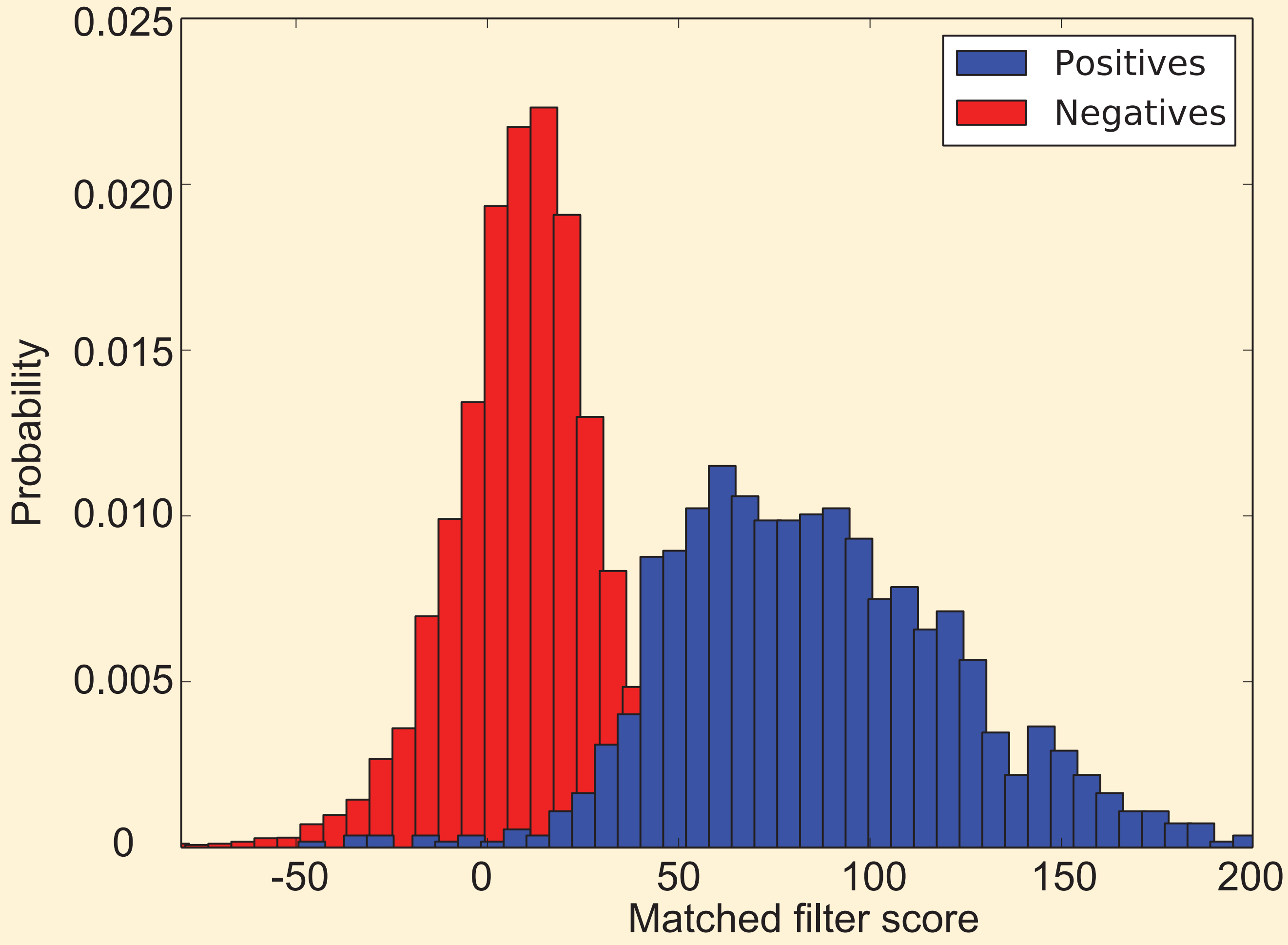


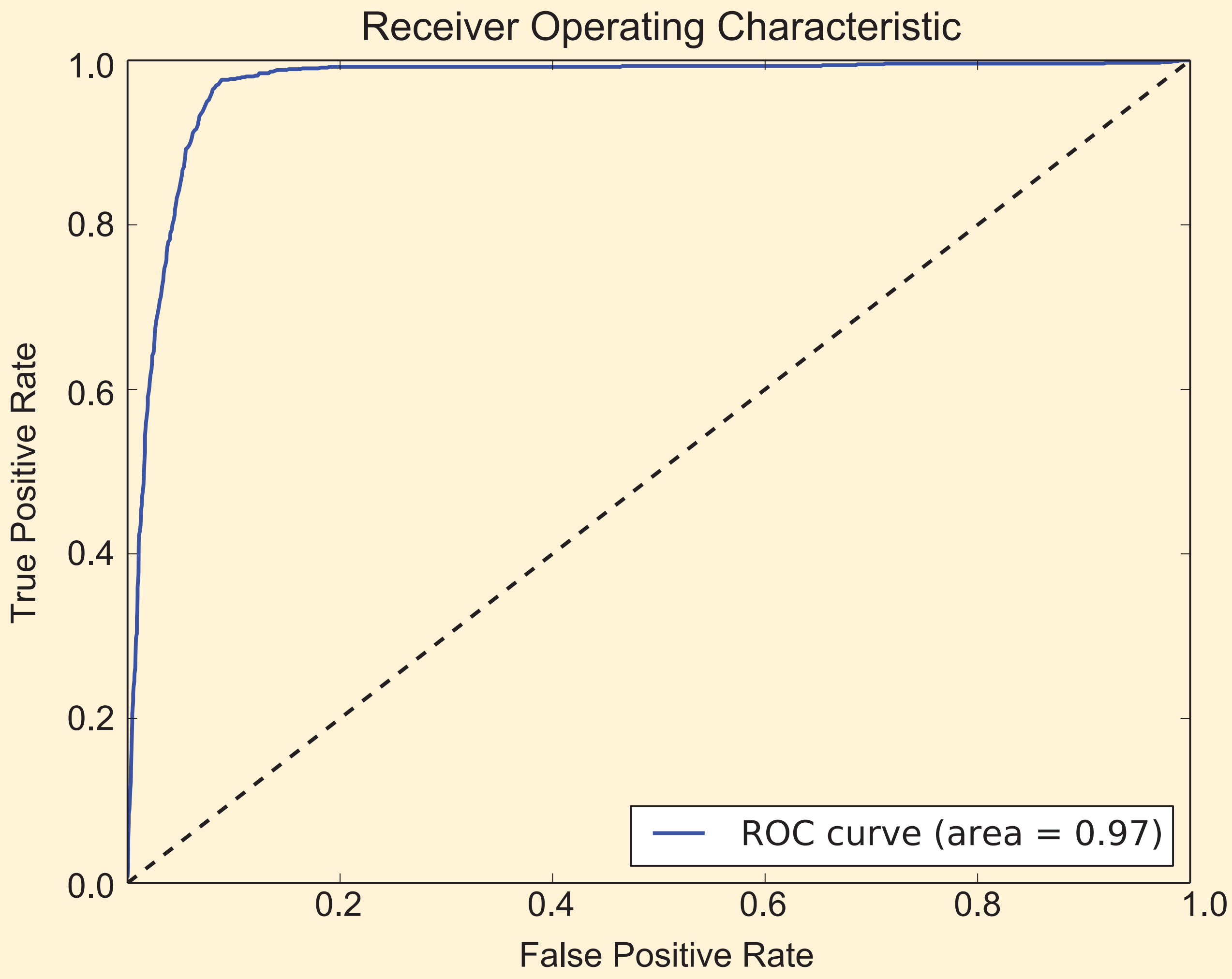
$$h[n] = s^*[N-1-n], \quad n = 0,\ldots,N-1$$

$$r[n] = \sum_{n=0}^{N-1} y[n+\ell]s^*[\ell]$$

$$= s^*[N-1]y[n+N-1] + \ldots + s^*[1]y[n+1] + s^*[0]y[n]$$

Accuracy of Matched Filter

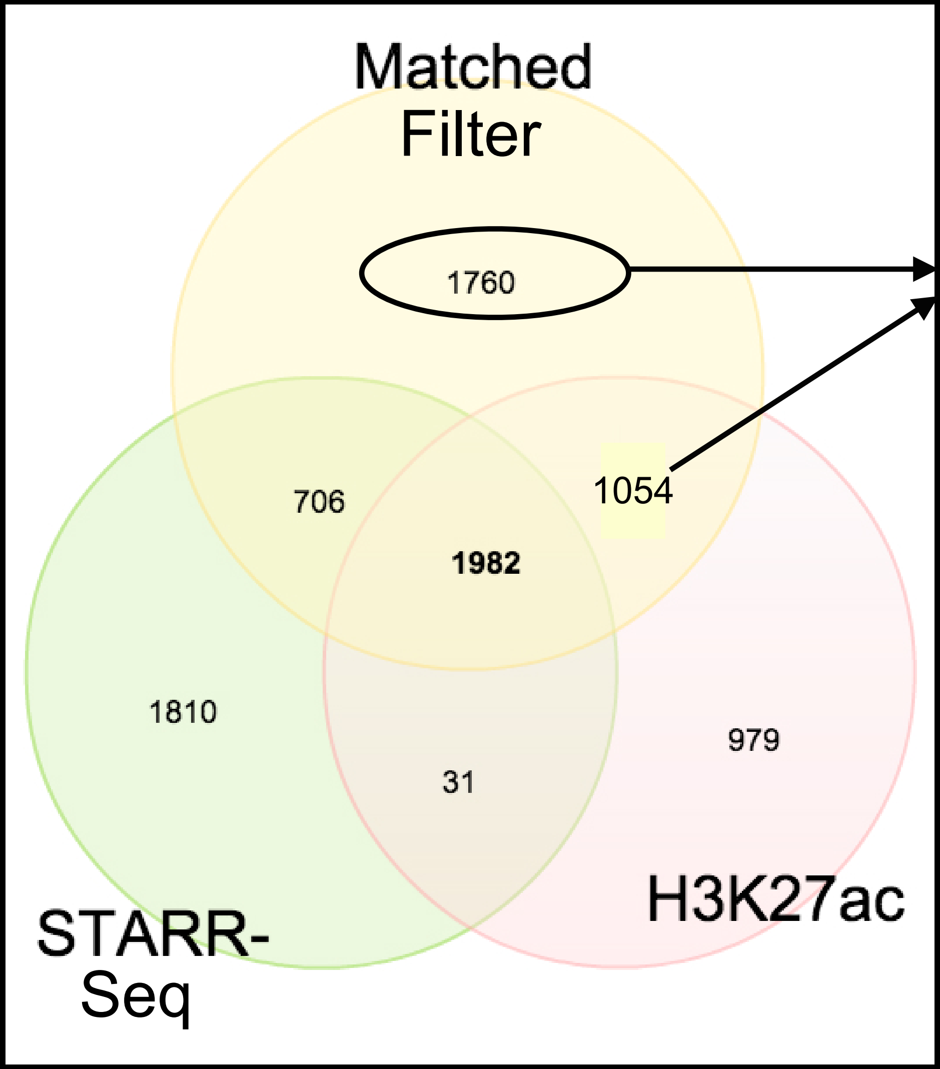




The occurrence of the pattern is very accurate for predicting potential enhancers in the genome.

Genome positives comparison across marks/cell-lines

H3K27ac Matched Filter



1964 on promoters

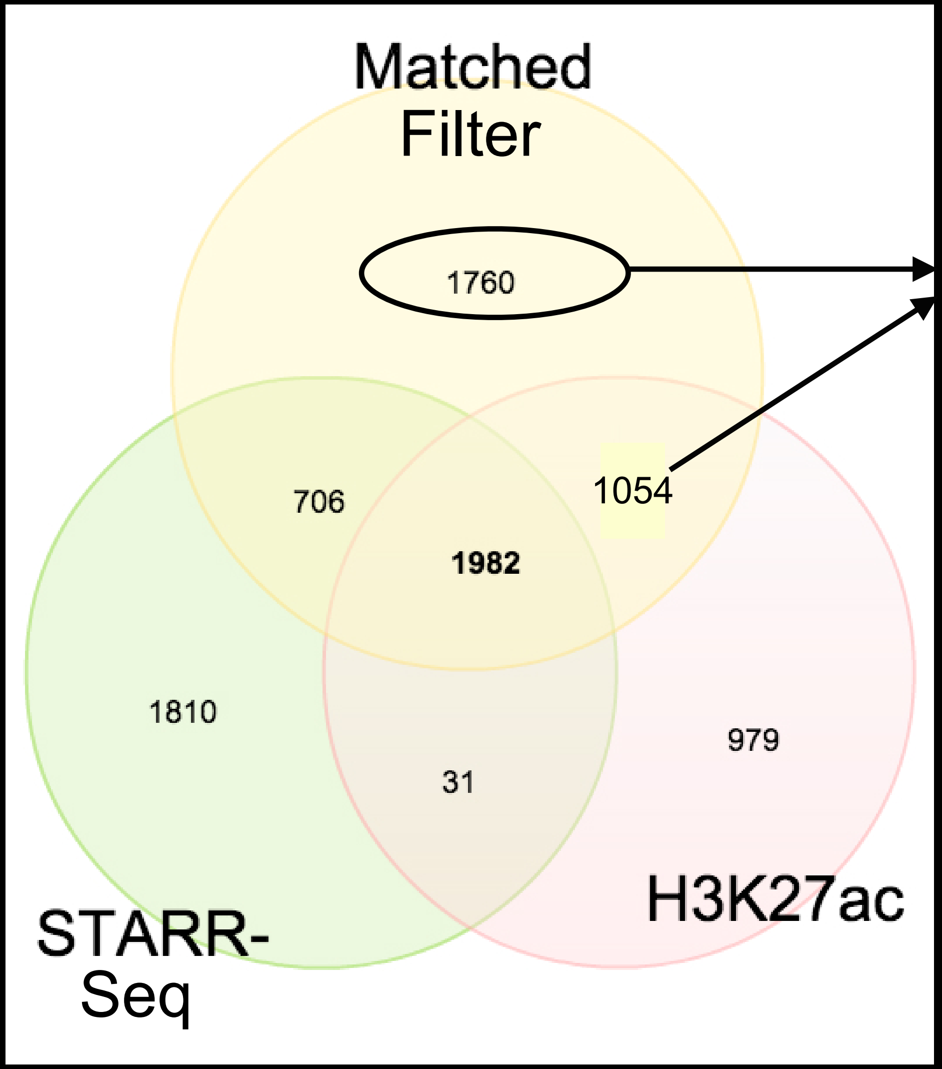
Need to show these promoters are active

1885 (1490) of STARR-seq peaks are positive in at least two (all three) filters

Most of the matched filter positives are positive on multiple histone marks. H3K9ac and H3K4me3 - most different (promoters).

Genome positives comparison across marks/cell-lines

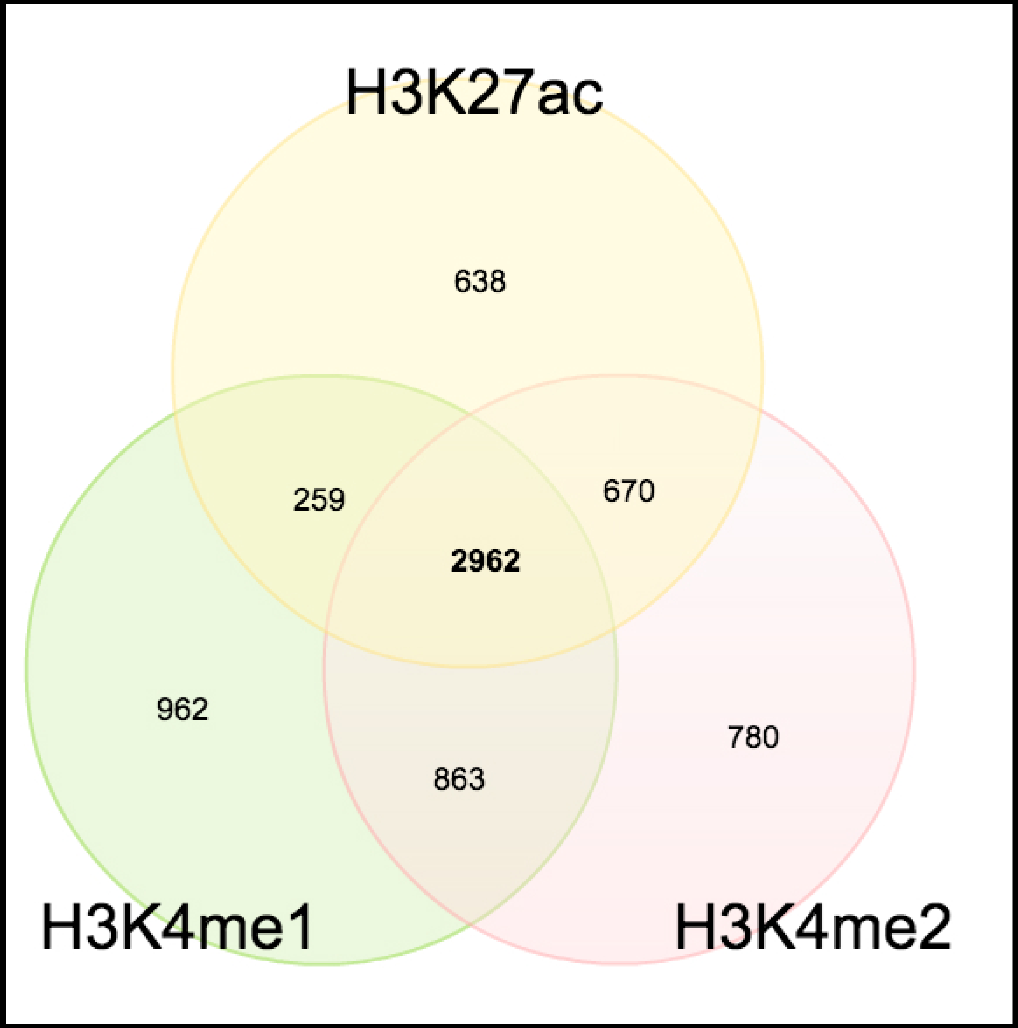
H3K27ac Matched Filter



1964 on promoters

Need to show these promoters are active

MF comparison



1885 (1490) of STARR-seq peaks are positive in at least two (all three) filters

Most of the matched filter positives are positive on multiple histone marks. H3K9ac and H3K4me3 - most different (promoters).

Stability of marks across cell lines

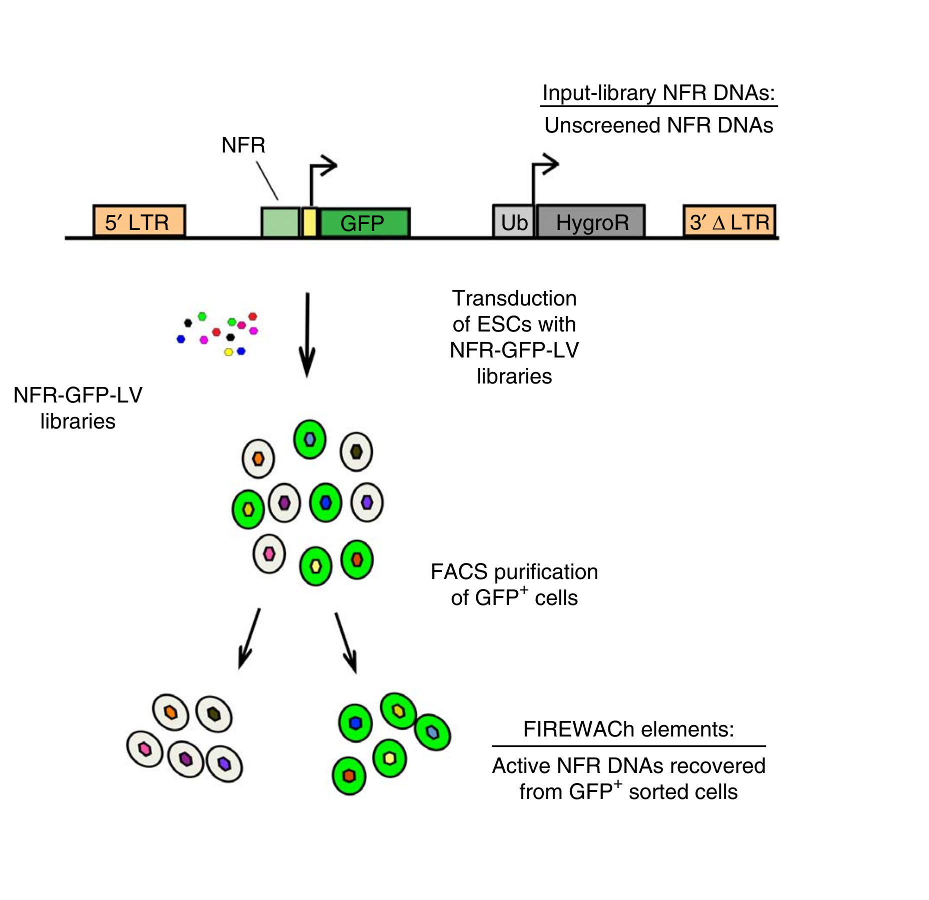
Histone Mark (double peak)	**AUC (ROC) S2 cell-line**	**AUC (ROC) BG3 cell line**
H3K27ac	0.88	0.97
H3K4me1	0.85	0.87
H3K4me2	0.85	0.86
H3K4me3	0.71	0.76
H3K9ac	0.88	0.75

Currently, extending the matched filter approach to mammalian enhanc
predictions.

Moving on to mammals

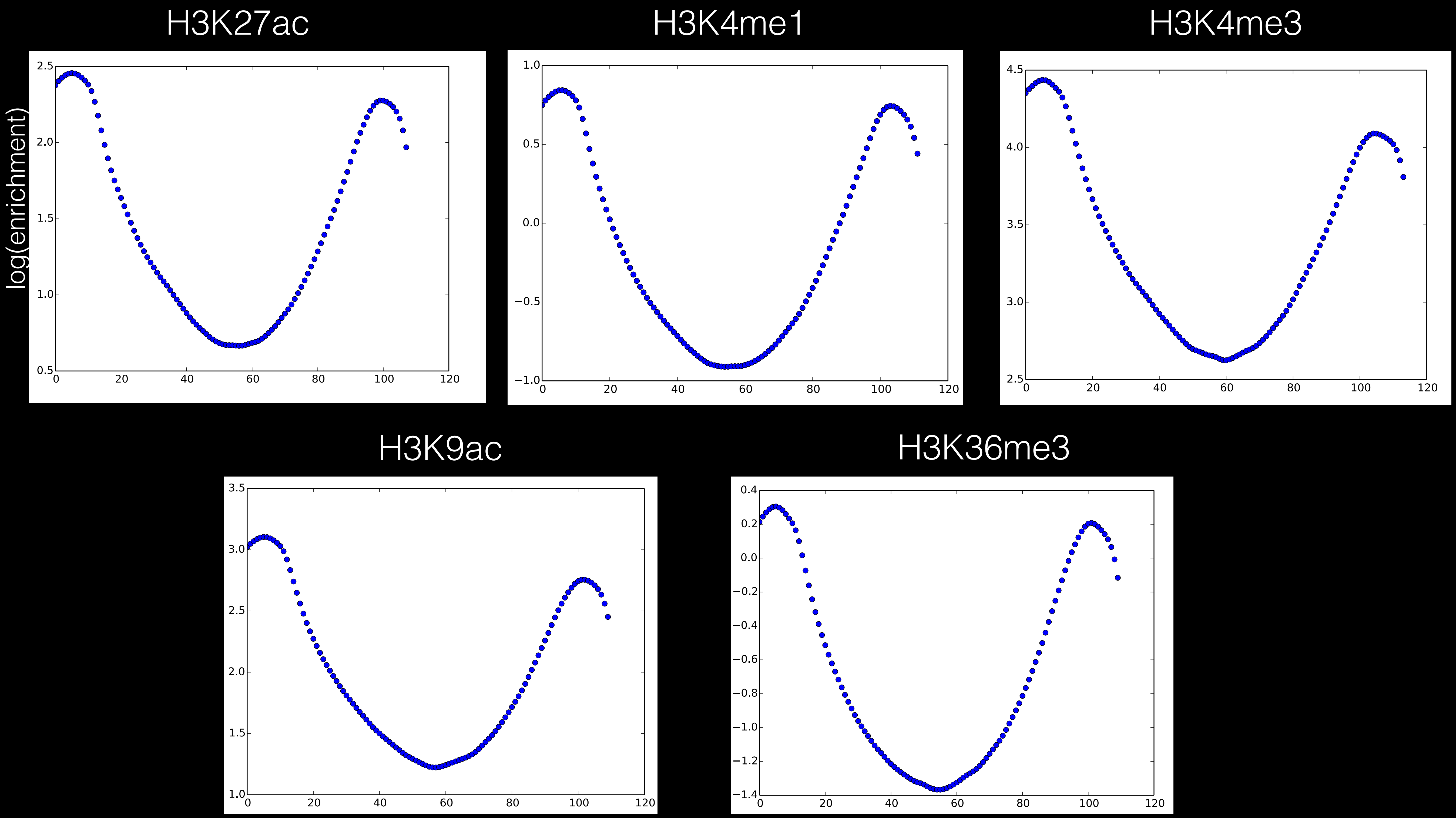
FIREWACh assay

- Enhancer candidates chosen based on open DNA in cell-line (murine ESC).
- Integrated into virus particles close to a minimal promoter and GFP.
- **Integrated into genome randomly with 1 clone per cell (H1-hESC).**
- **One potential enhancer of length 100-300 bp** per cell.
- FACS to sort cells expressing GFP.
- Small population of cells show positive enhancer activity.
- Amplified positive enhancer sequences with PCR using primers recognizing the flanking sequences.
- Tested enhancer activity using traditional assays.


Input-library NFR DNAs:
Unscreened NFR DNAs
NFR
5′ LTR
GFP
Ub
HygroR
3′ Δ LTR
Transduction of ESCs with NFR-GFP-LV libraries
NFR-GFP-LV libraries
FACS purification of GFP+ cells
FIREWACh elements:
Active NFR DNAs recovered from GFP+ sorted cells

Pro:
Chromatin context.
Con:
100-300 bp length.

Metaprofiles from FIREWACh



H3K27ac

H3K4me1

H3K4me3

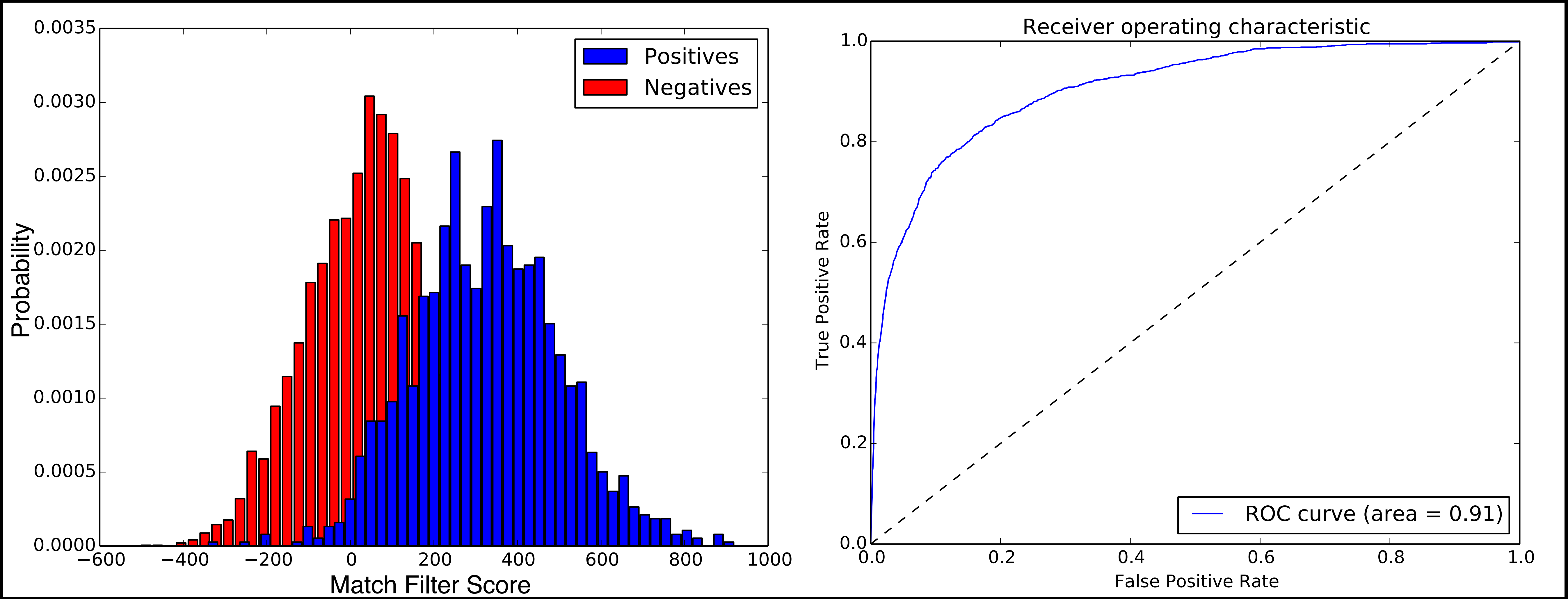
H3K9ac

H3K36me3

Heterogeneity in the metaprofiles close to regulatory regions - Anshul’s paper.

The metaprofile can be used to identify enhancers from random regions in the genome.

Performance of H3K27ac metaprofile



Accuracy of Predictions

Histone Mark (double peak)	**AUC (ROC)** **mESC cell-line**
H3K27ac	0.91
H3K4me1	0.70!!
H3K4me3	0.87
H3K9ac	0.88
H3K36me3	0.67