WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Probabilistic generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies

Sündüz Keleş

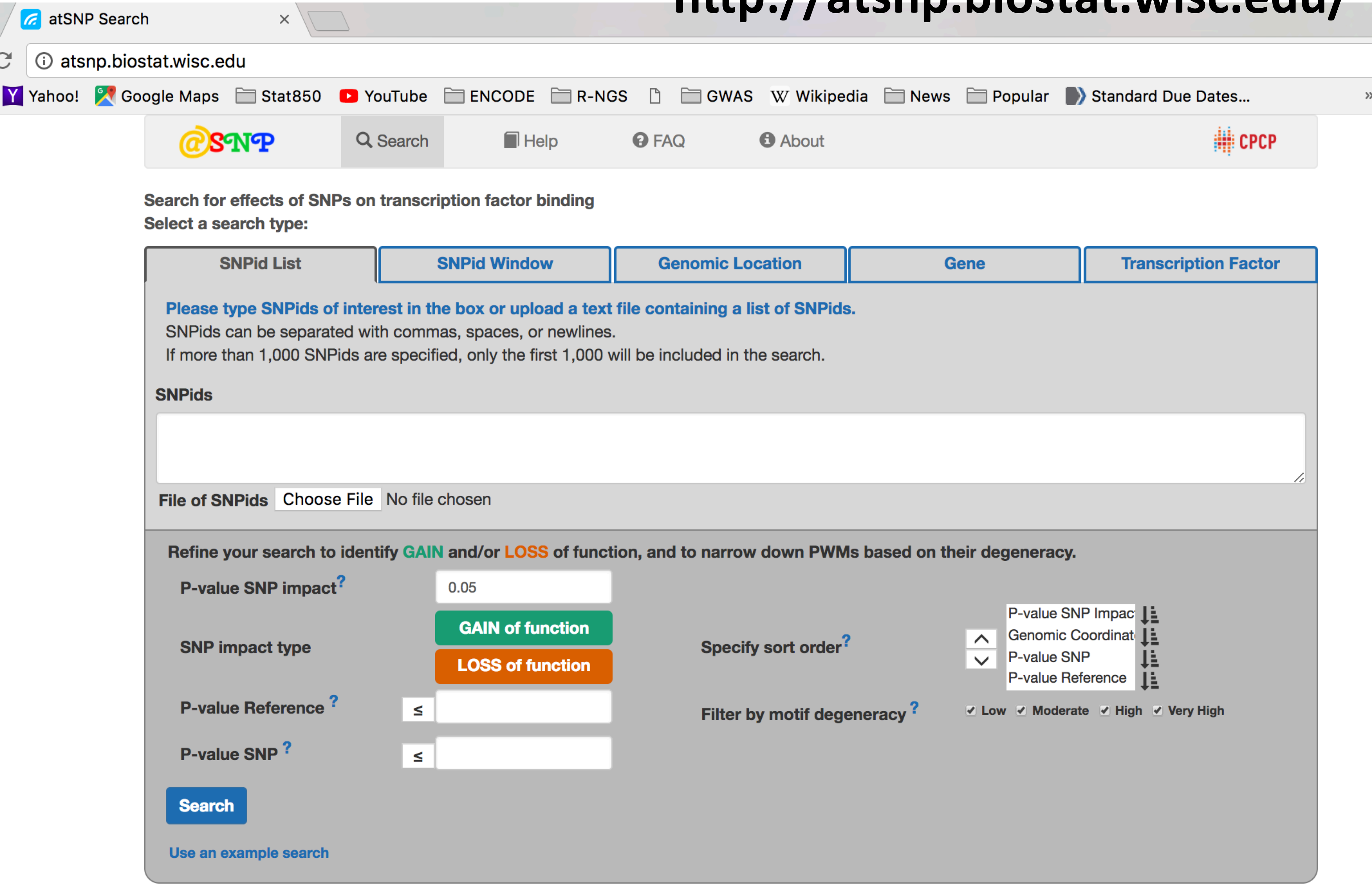
Department of Biostatistics and Medical Informatics

Department of Statistics

University of Wisconsin, Madison

Google atsnp search

http://atsnp.biostat.wisc.edu/



Yahoo! Google Maps Stat850 YouTube ENCODE R-NGS GWAS Wikipedia News Popular Standard Due Dates...

Search for effects of SNPs on transcription factor binding

Select a search type:

SNPid List | SNPid Window | Genomic Location | Gene | Transcription Factor

Please type SNPids of interest in the box or upload a text file containing a list of SNPids.

SNPids can be separated with commas, spaces, or newlines.

If more than 1,000 SNPids are specified, only the first 1,000 will be included in the search.

SNPids

File of SNPids Choose File No file chosen

Refine your search to identify GAIN and/or LOSS of function, and to narrow down PWMs based on their degeneracy.

P-value SNP impact? 0.05

SNP impact type GAIN of function LOSS of function

P-value Reference ? ≤

P-value SNP ? ≤

Specify sort order? P-value SNP Impac, Genomic Coordinat, P-value SNP, P-value Reference

Filter by motif degeneracy ? ☑ Low ☑ Moderate ☑ High ☑ Very High

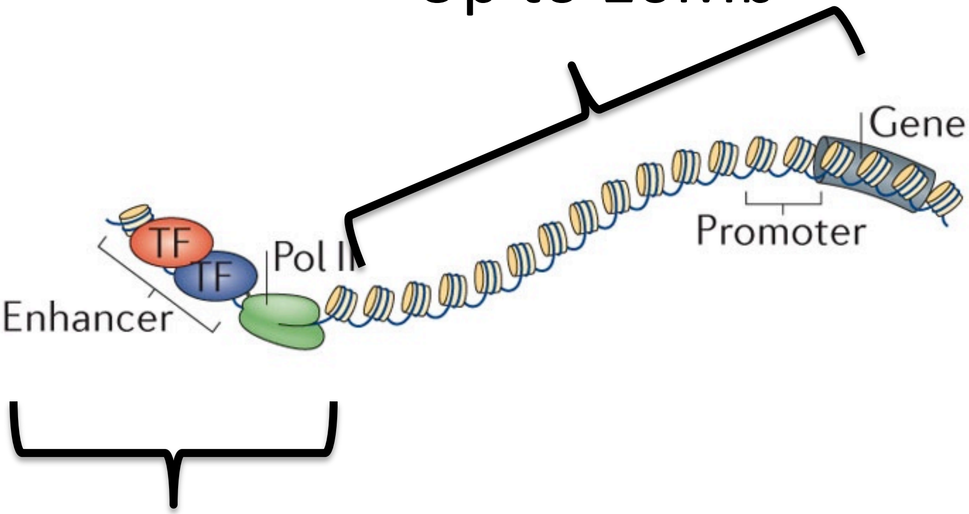
Search

Use an example search

Google atsnp search

High throughput chromatin conformation capture (Hi-C) for studying long-range interactions

Up to 10Mb

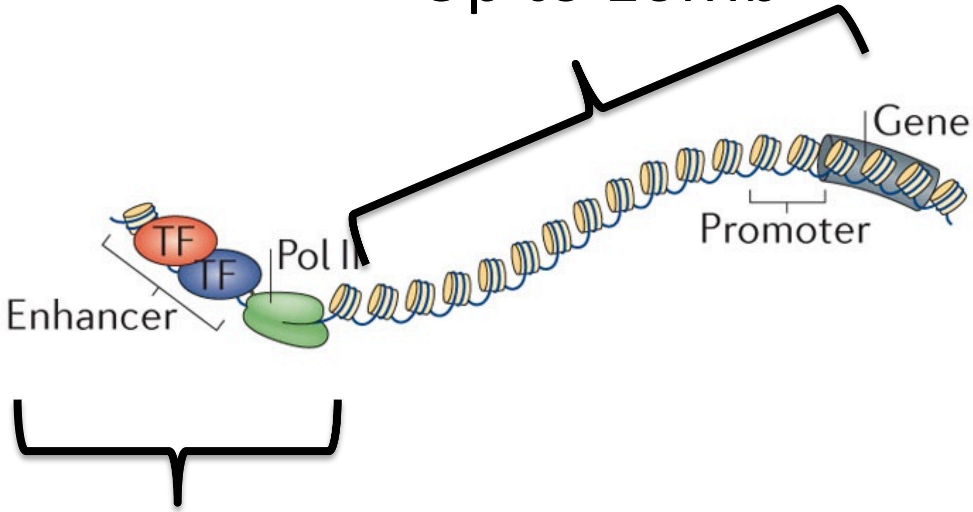


ENCODE project generated catalogs of enhancers.

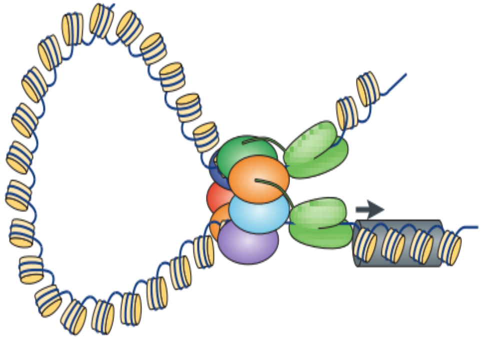
Pombo & Dillon, *Nature Reviews Molecular Cellular Biology*, 2015

Hi-C for studying long-range interactions

Up to 10Mb



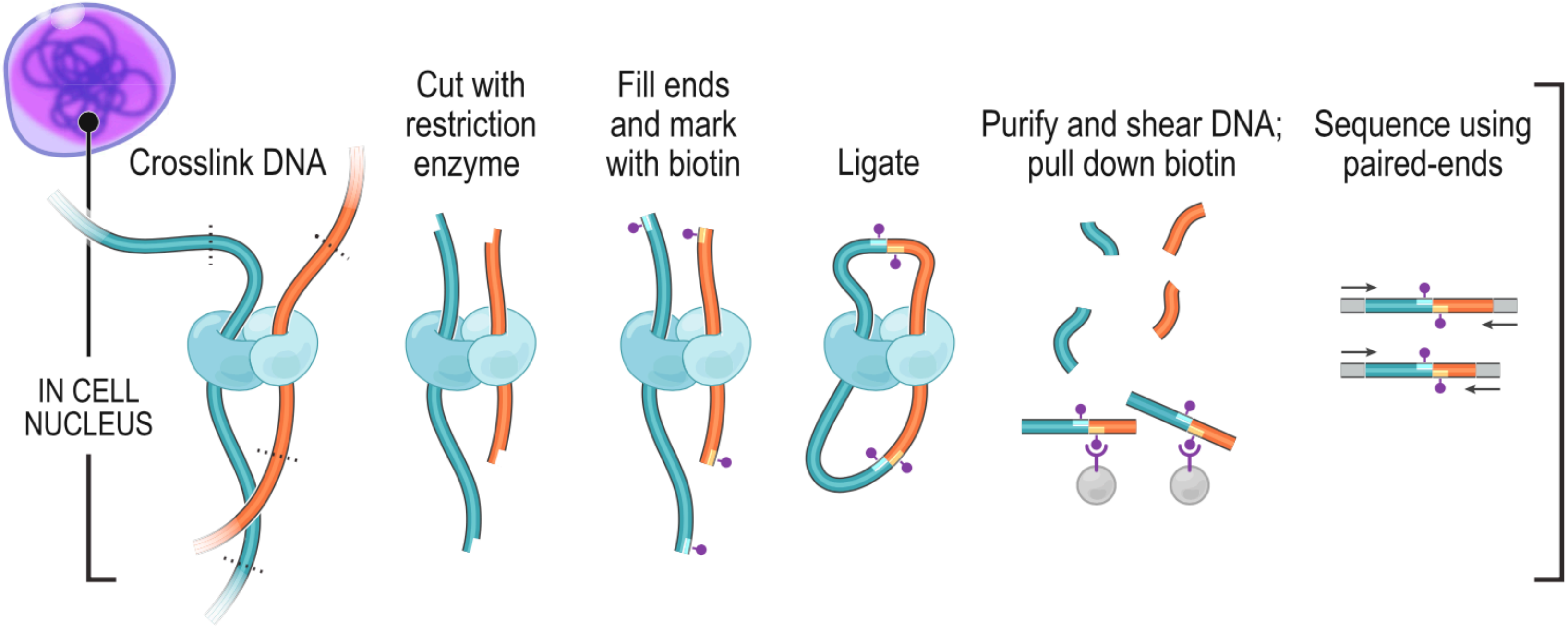
Looping of DNA



ENCODE project generated catalogs of enhancers.

Pombo & Dillon, *Nature Reviews Molecular Cellular Biology*, 2015

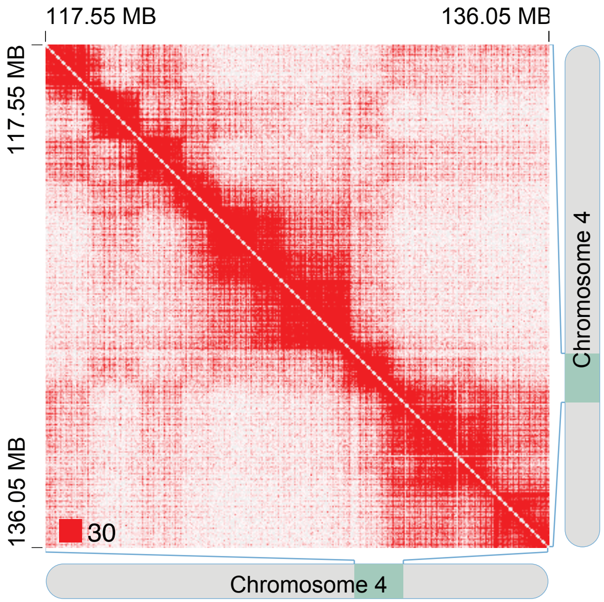
Hi-C experimental protocol

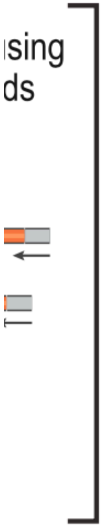

Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
IN CELL NUCLEUS

Rao *et al*., *Cell*, 2014

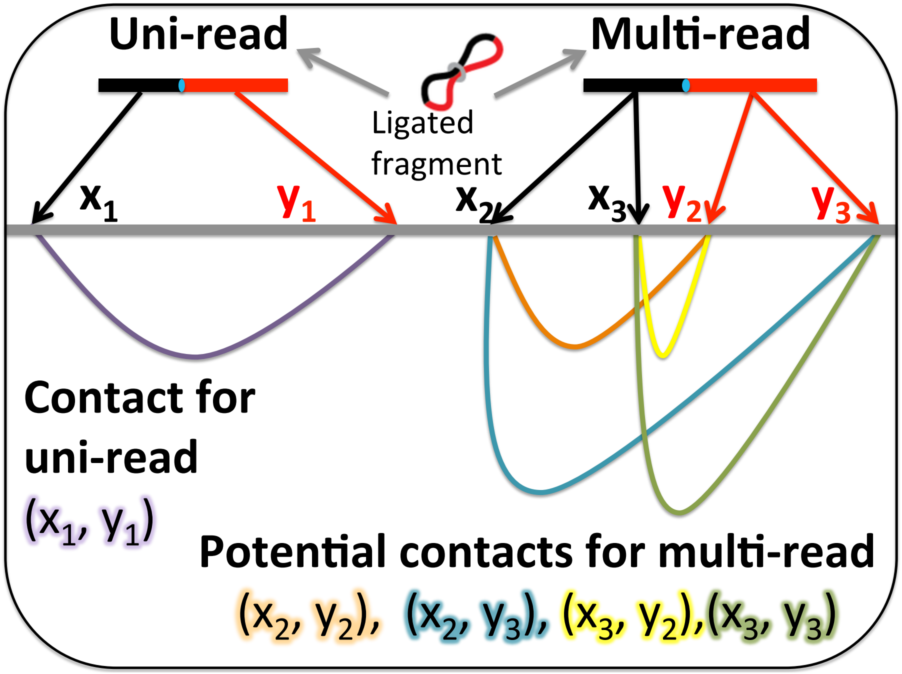
Hi-C experimental protocol

Data gets summarized as a contact count matrix.



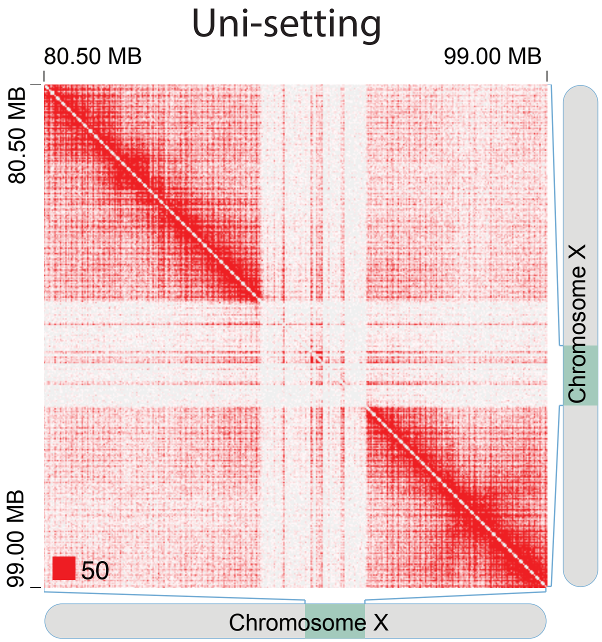


Just like any sequencing dataset, Hi-C analysis start with read alignment



Signals from repetitive regions are under-represented

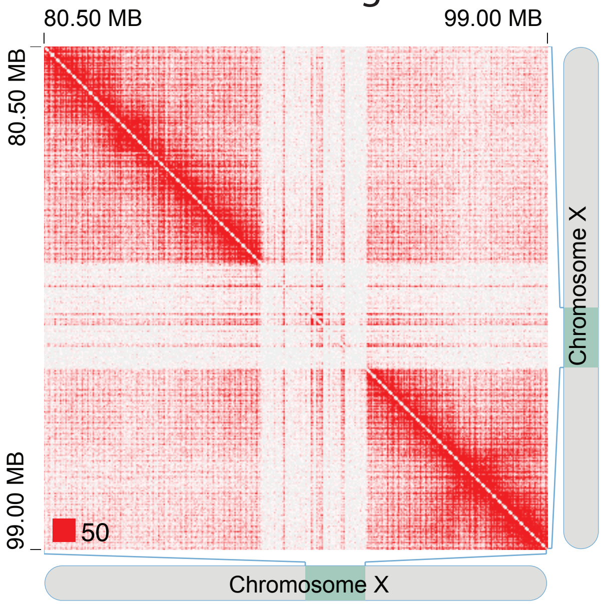
Before mHi-C



Signals from repetitive regions are under-represented

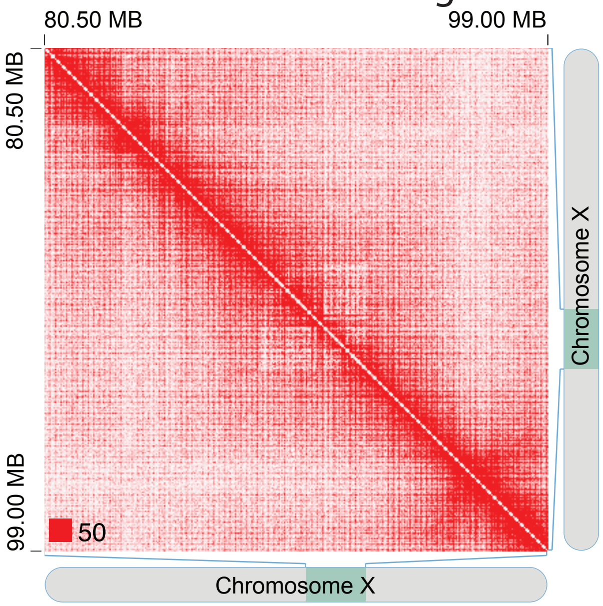
Before mHi-C

Uni-setting



After mHi-C

Uni&Multi-setting



Evaluation: 6 independent studies, with 8 datasets, and multiple replicates per dataset

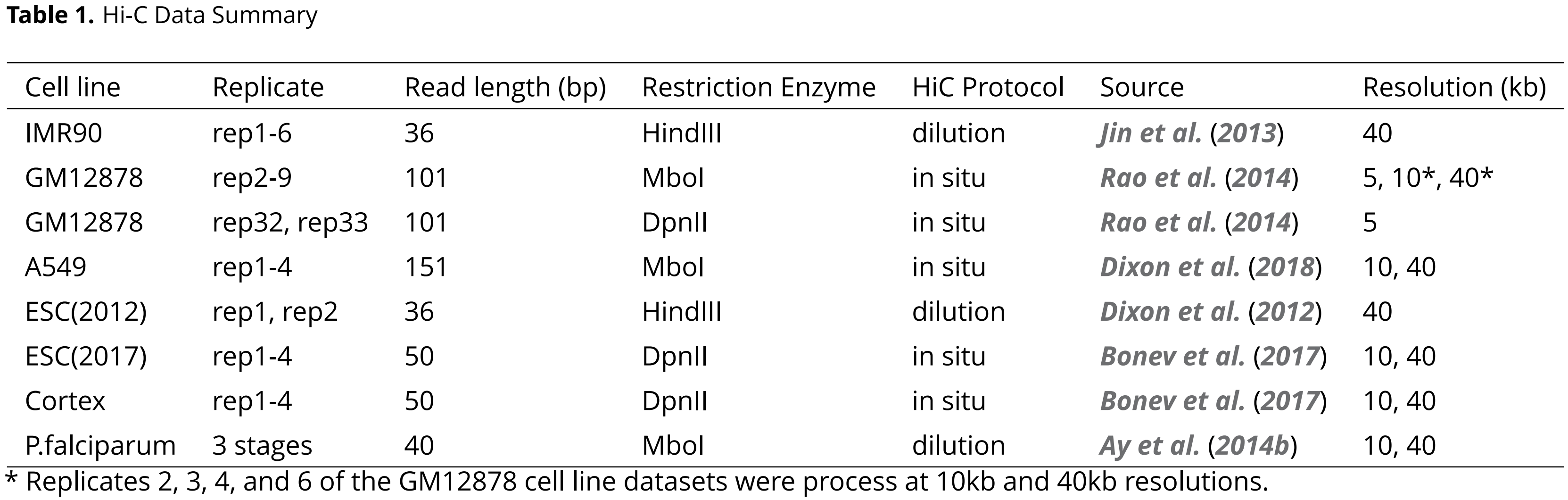


Table 1. Hi-C Data Summary

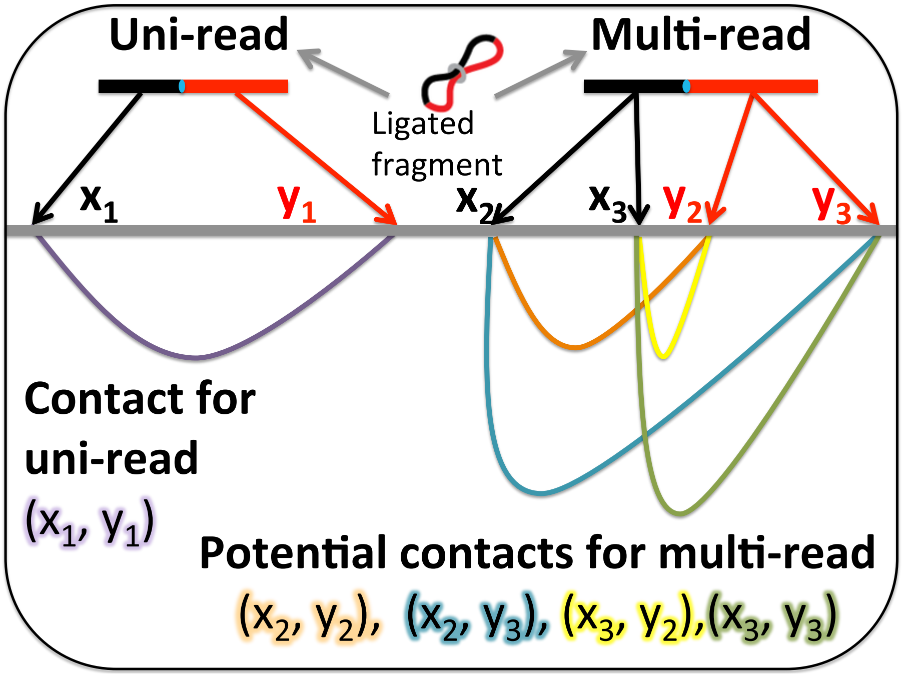
Cell line	Replicate	Read length (bp)	Restriction Enzyme	HiC Protocol	Source	Resolution (kb)
IMR90	rep1-6	36	HindIII	dilution	***Jin et al.*** (***2013***)	40
GM12878	rep2-9	101	MboI	in situ	***Rao et al.*** (***2014***)	5, 10*, 40*
GM12878	rep32, rep33	101	DpnII	in situ	***Rao et al.*** (***2014***)	5
A549	rep1-4	151	MboI	in situ	***Dixon et al.*** (***2018***)	10, 40
ESC(2012)	rep1, rep2	36	HindIII	dilution	***Dixon et al.*** (***2012***)	40
ESC(2017)	rep1-4	50	DpnII	in situ	***Bonev et al.*** (***2017***)	10, 40
Cortex	rep1-4	50	DpnII	in situ	***Bonev et al.*** (***2017***)	10, 40
P.falciparum	3 stages	40	MboI	dilution	***Ay et al.*** (***2014b***)	10, 40

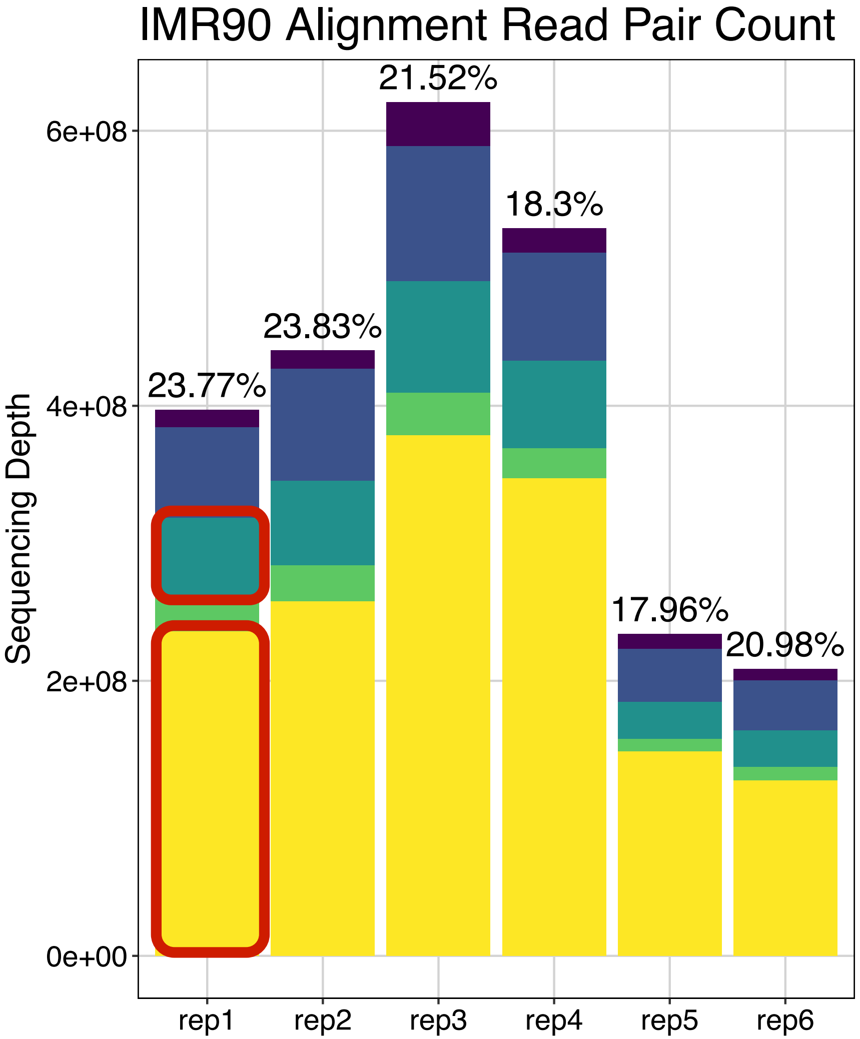
* Replicates 2, 3, 4, and 6 of the GM12878 cell line datasets were process at 10kb and 40kb resolutions.

Criteria for selection

- Genome size (large, small)
- Sequencing depth, coverage
- *Cis*-to-*Trans* ratio
- Proportion of mappable and valid reads

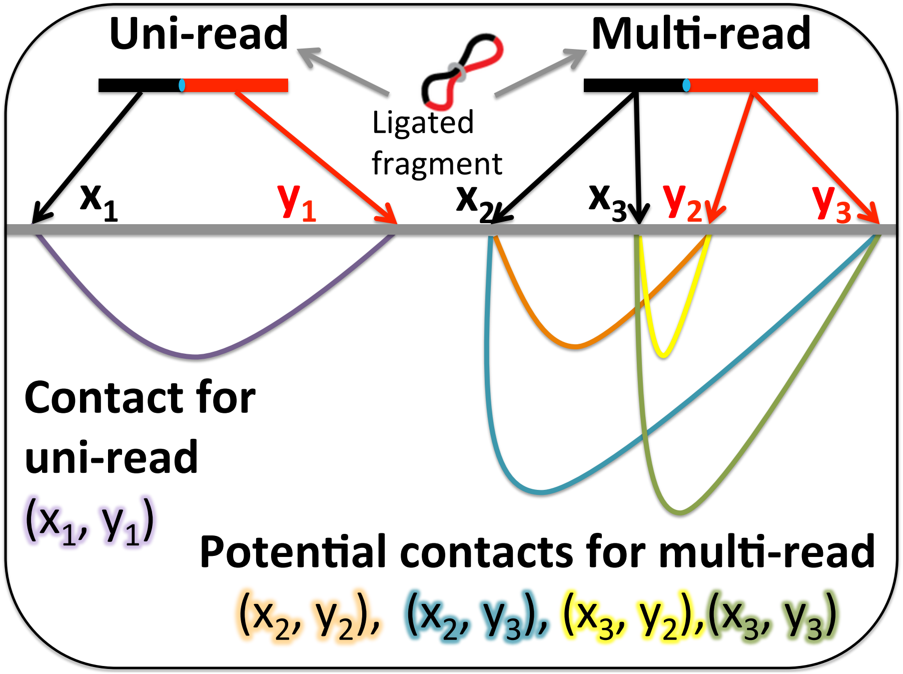
Multi-reads are abundant





🟪	Unmapped
🟦	Singleton
🟩	Multi–reads (High Quality)
🟩	Multi–reads (Low Quality)
🟨	Uni–reads

Multi-reads are abundant

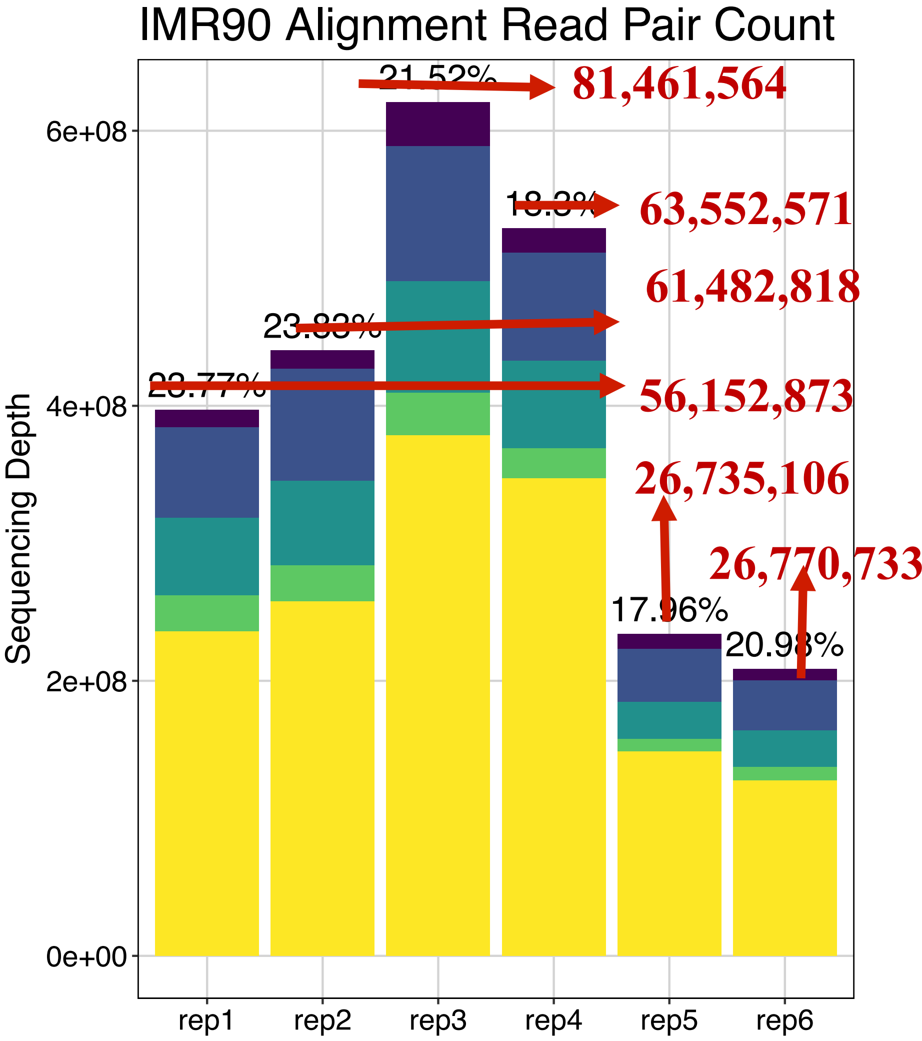


Contact for uni-read

(x_1, y_1)

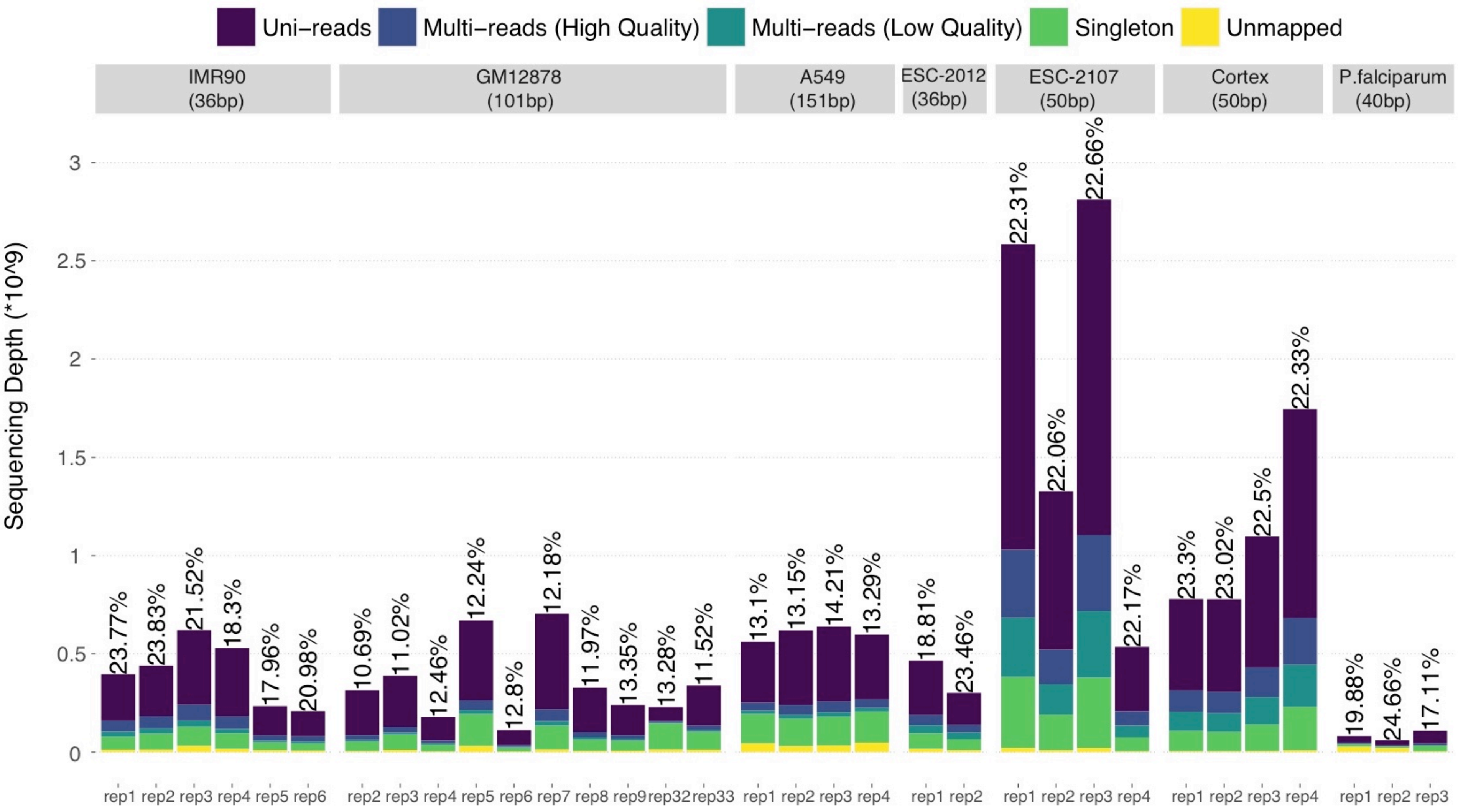
Potential contacts for multi-read

(x_2, y_2), (x_2, y_3), (x_3, y_2),(x_3, y_3)

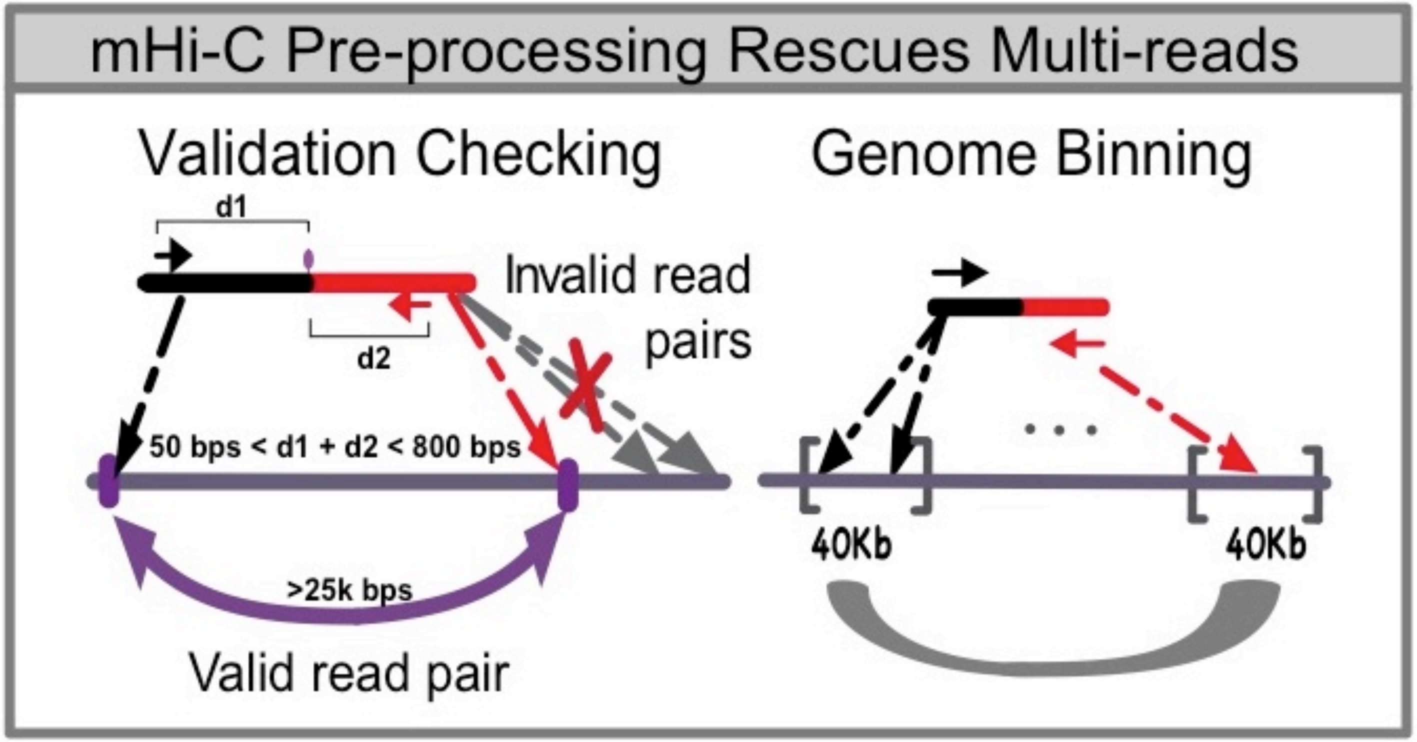


- ■ Unmapped
- ■ Singleton
- ■ Multi–reads (High Quality)
- ■ Multi–reads (Low Quality)
- ■ Uni–reads

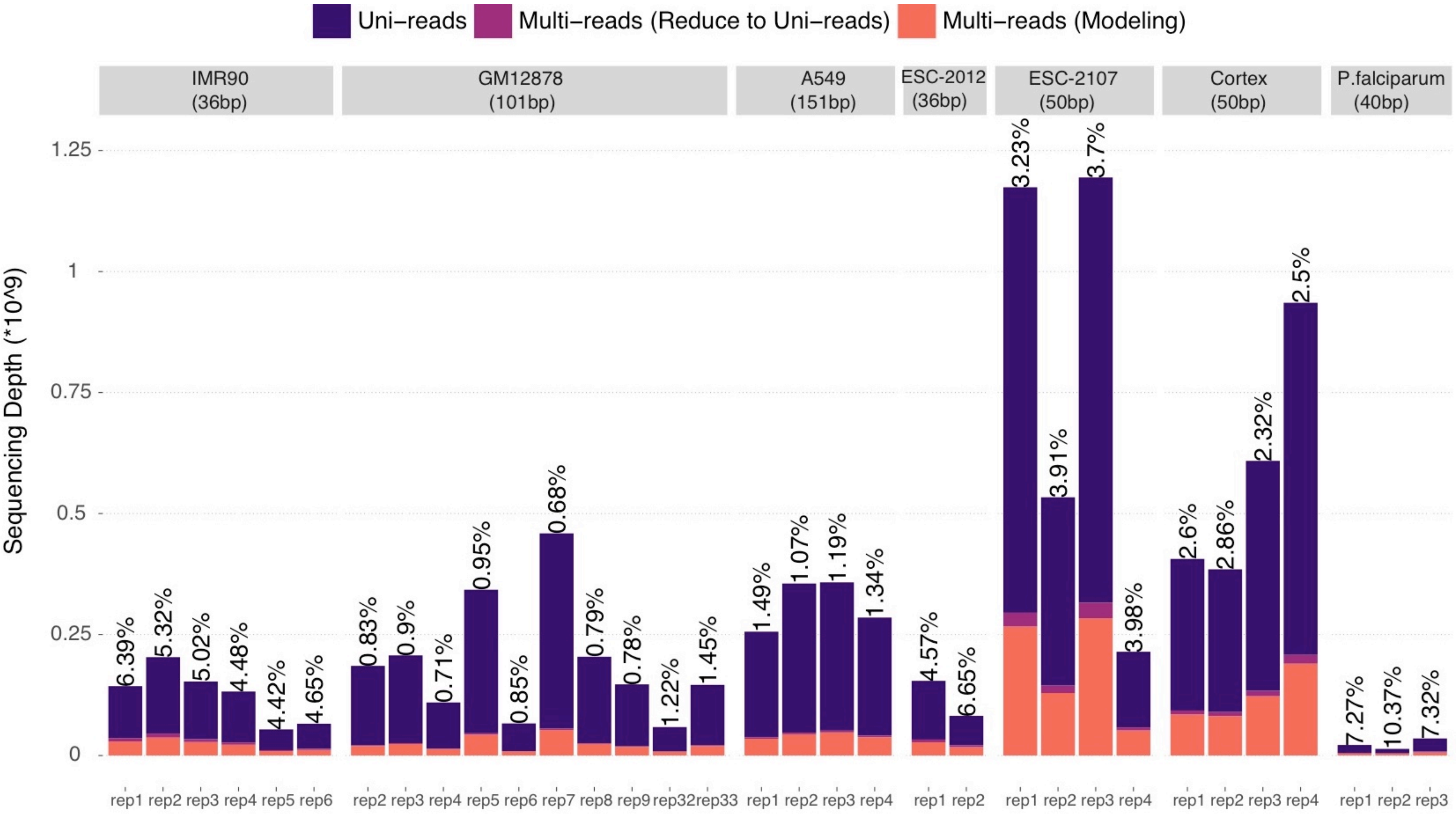
Results across eight studies



Sometimes, there is free lunch

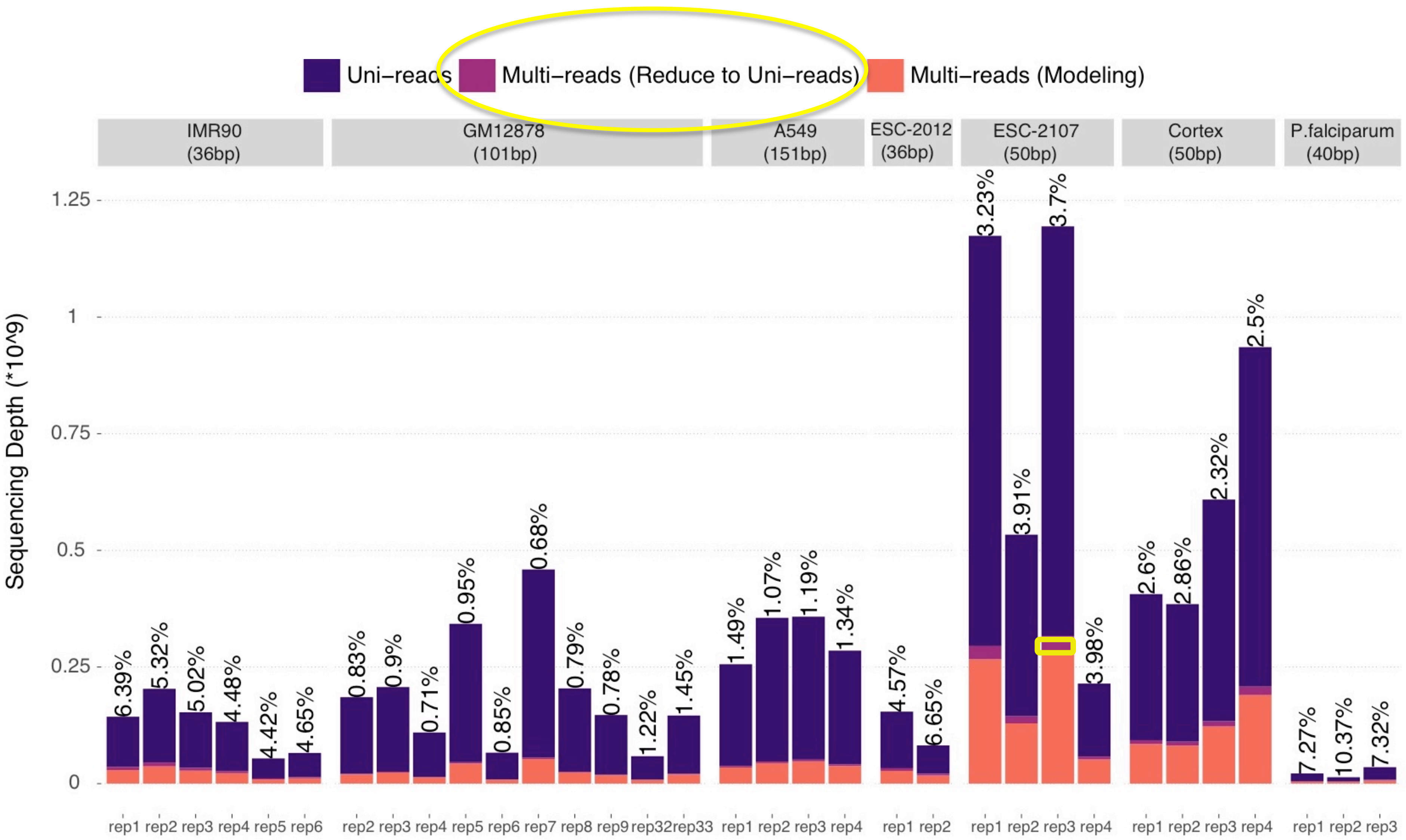


Sometimes, there is free lunch



Sometimes, there is free lunch

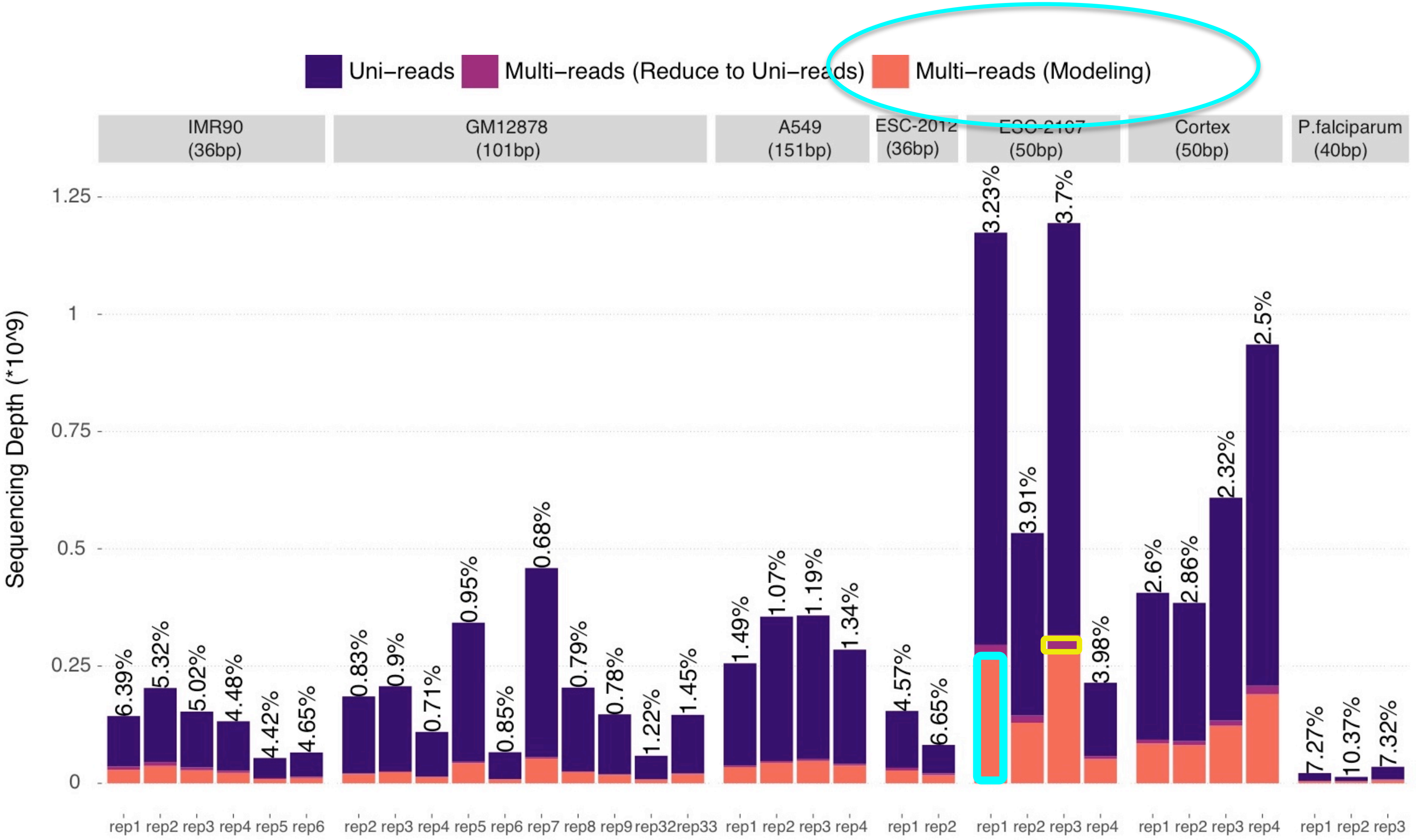
No-cost multi-reads: add ~5%



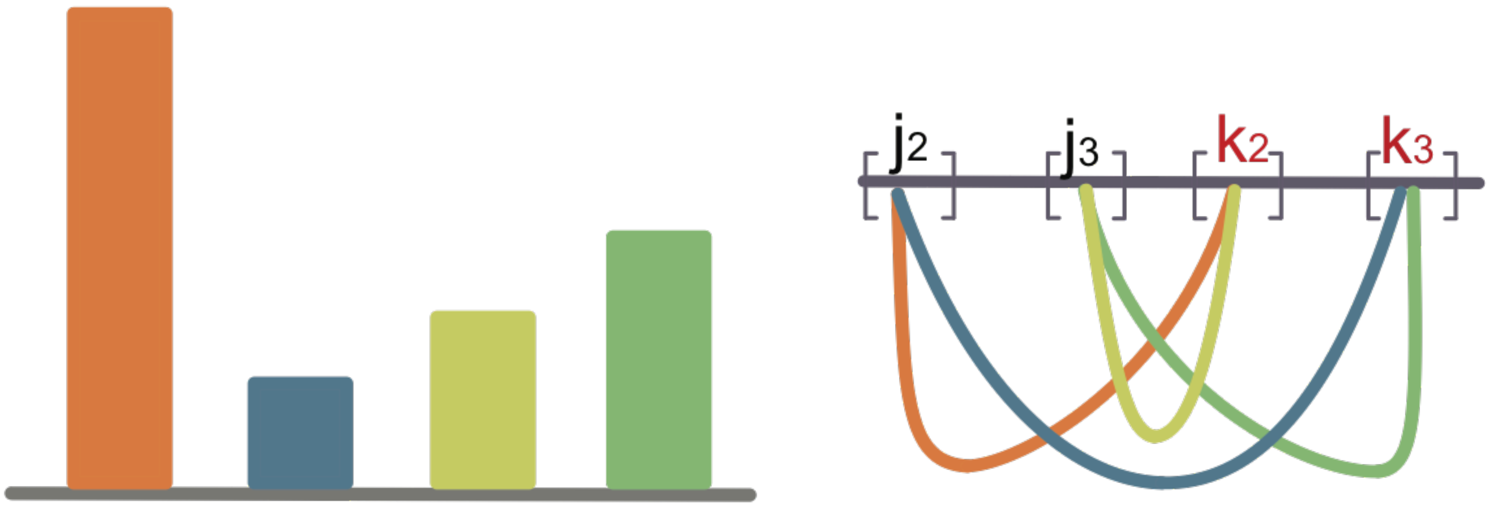
Sometimes, there is free lunch

No-cost multi-reads: add ~5%

Multi-reads need rescuing: add ~ 23%



mHi-C: multi-read allocation for Hi-C



Local Bin-pair Contact Counts

mHi-C model

Observed: $Y_{i,(j,k)} = 1.$

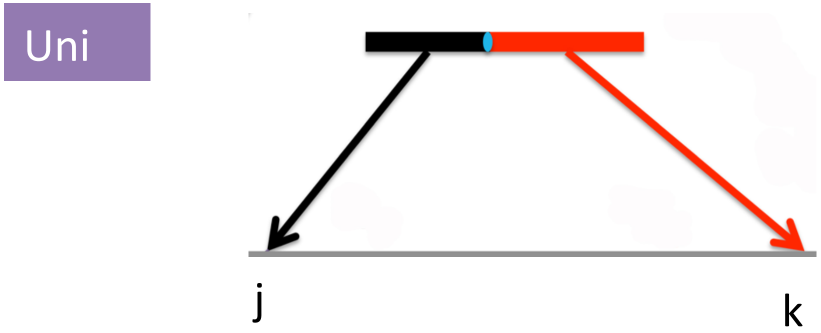
Valid read pair i aligned

to contact unit (j, k).

mHi-C model

Observed: $Y_{i,(j,k)} = 1.$

Valid read pair i aligned to contact unit (j, k).



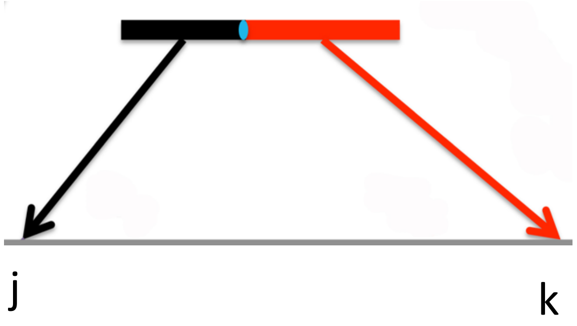
$$\sum_{j,k} Y_{i,(j,k)} = 1, \quad i = 1, \cdots, N$$

mHi-C model

Observed: $Y_{i,(j,k)} = 1.$

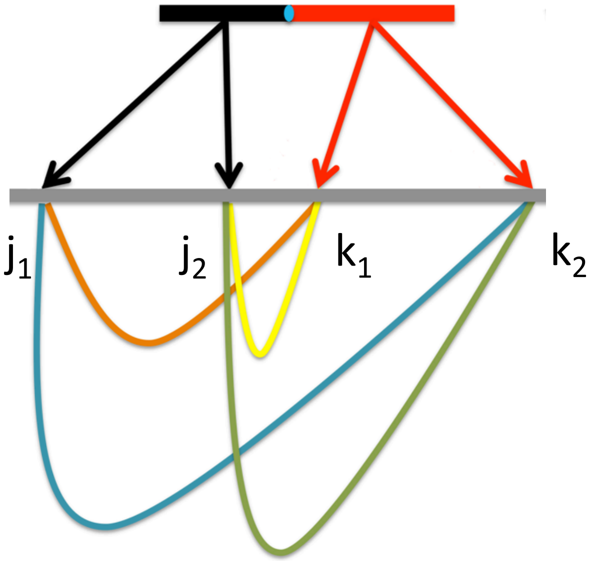
Valid read pair i aligned to contact unit (j,k).

Uni



$$\sum_{j,k} Y_{i,(j,k)} = 1, \quad i = 1, \cdots, N$$

Multi



e.g.

$$\sum_{j,k} Y_{i,(j,k)} = 4$$

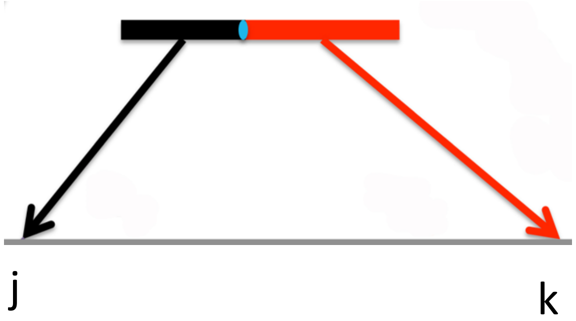
$$\sum_{j,k} Y_{i,(j,k)} \geq 1 \quad , i = 1, \cdots, N$$

mHi-C model

Observed: $Y_{i,(j,k)} = 1.$

Valid read pair i aligned to contact unit (j, k).

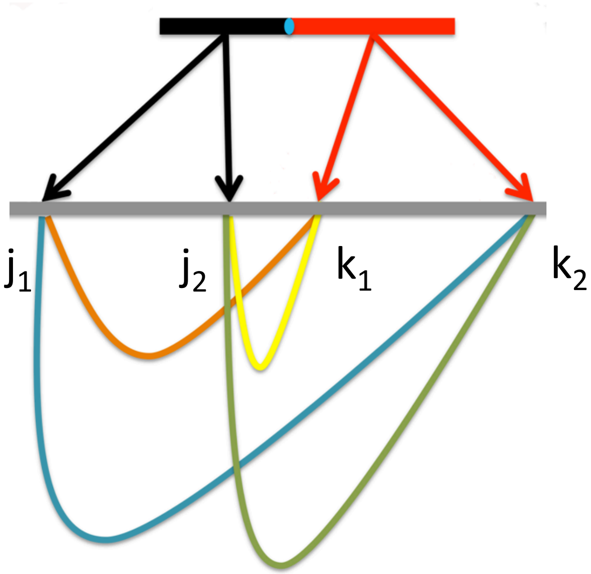
Uni



Hidden: $Z_{i,(j,k)} = 1,$

Valid read pair i originated from contact unit (j, k).

Multi

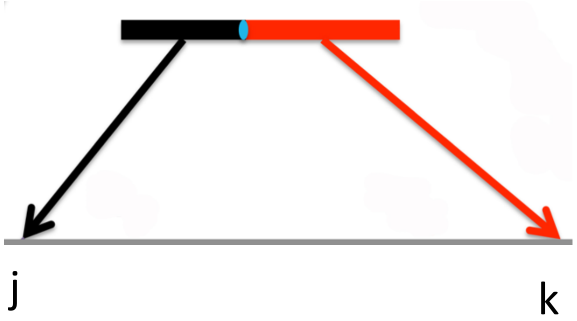


mHi-C model

Observed: $Y_{i,(j,k)} = 1.$

Valid read pair i aligned to contact unit (j, k).

Uni

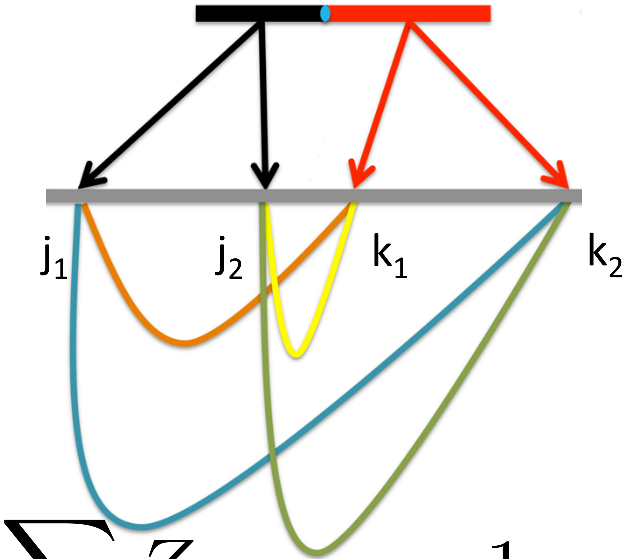


$$\sum_{j,k} Z_{i,(j,k)} = 1$$

Hidden: $Z_{i,(j,k)} = 1,$

Valid read pair i originated from contact unit (j, k).

Multi



$$\sum_{j,k} Z_{i,(j,k)} = 1$$

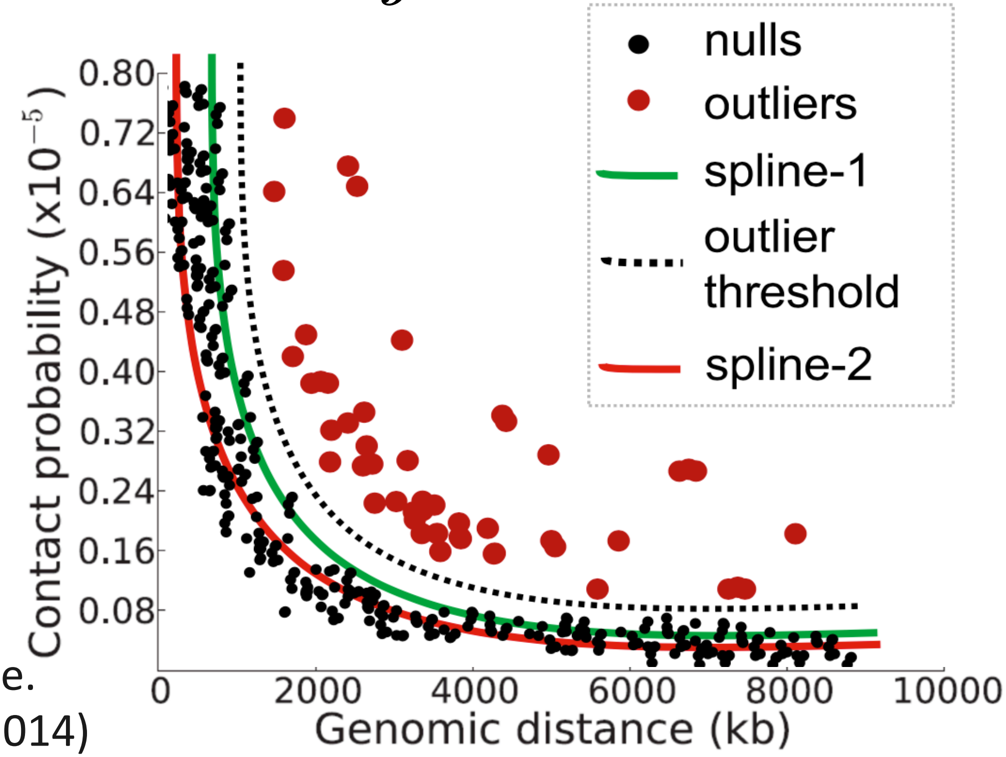
mHi-C model

$$Z_i \sim \text{Multinomial}(\pi_{(1,2)}, \pi_{(j,k)}, \cdots, \pi_{(M,M-1)})$$

$$\pi \sim \text{Dirichlet}(\gamma_{(1,2)}, \cdots, \gamma_{(j,k)}, \cdots, \gamma_{(M,M-1)})$$

$\gamma_{(j,k)}$ is modeled as a function of the distance between contact units j and k

$\gamma_{(j,k)}$ play the role of **pseudo-counts** in the Dirichlet-Multinomial framework.

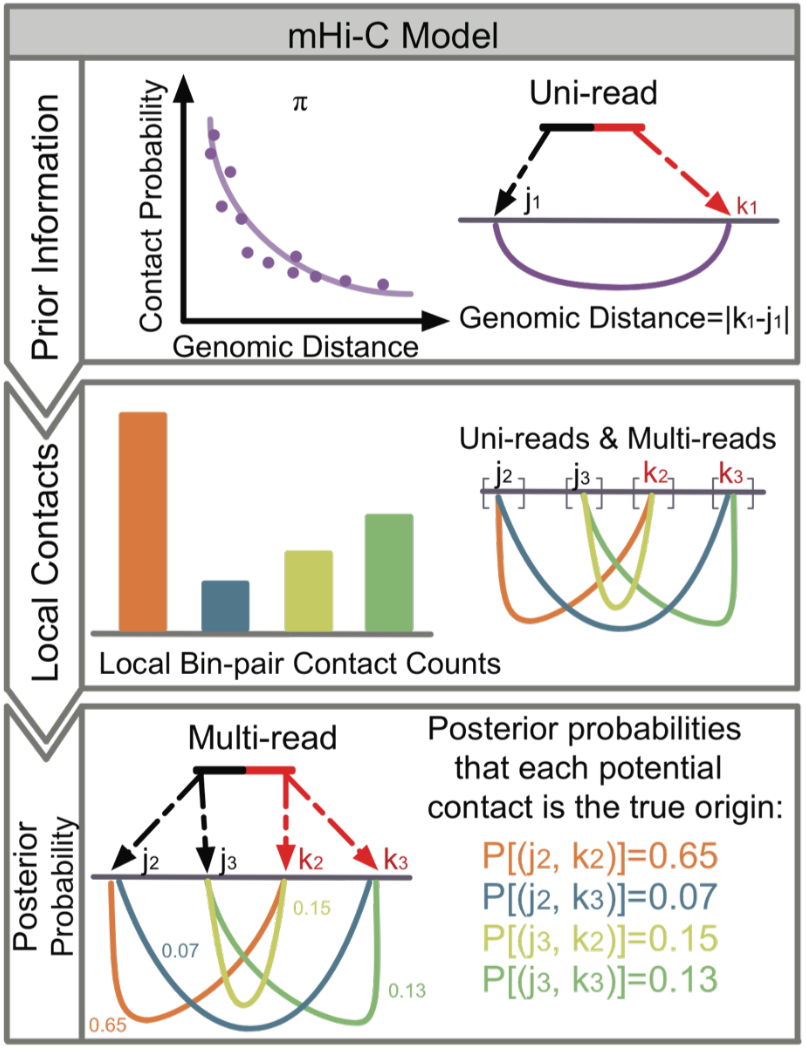


Ay, Bailey, and Noble. *Genome research* (2014)

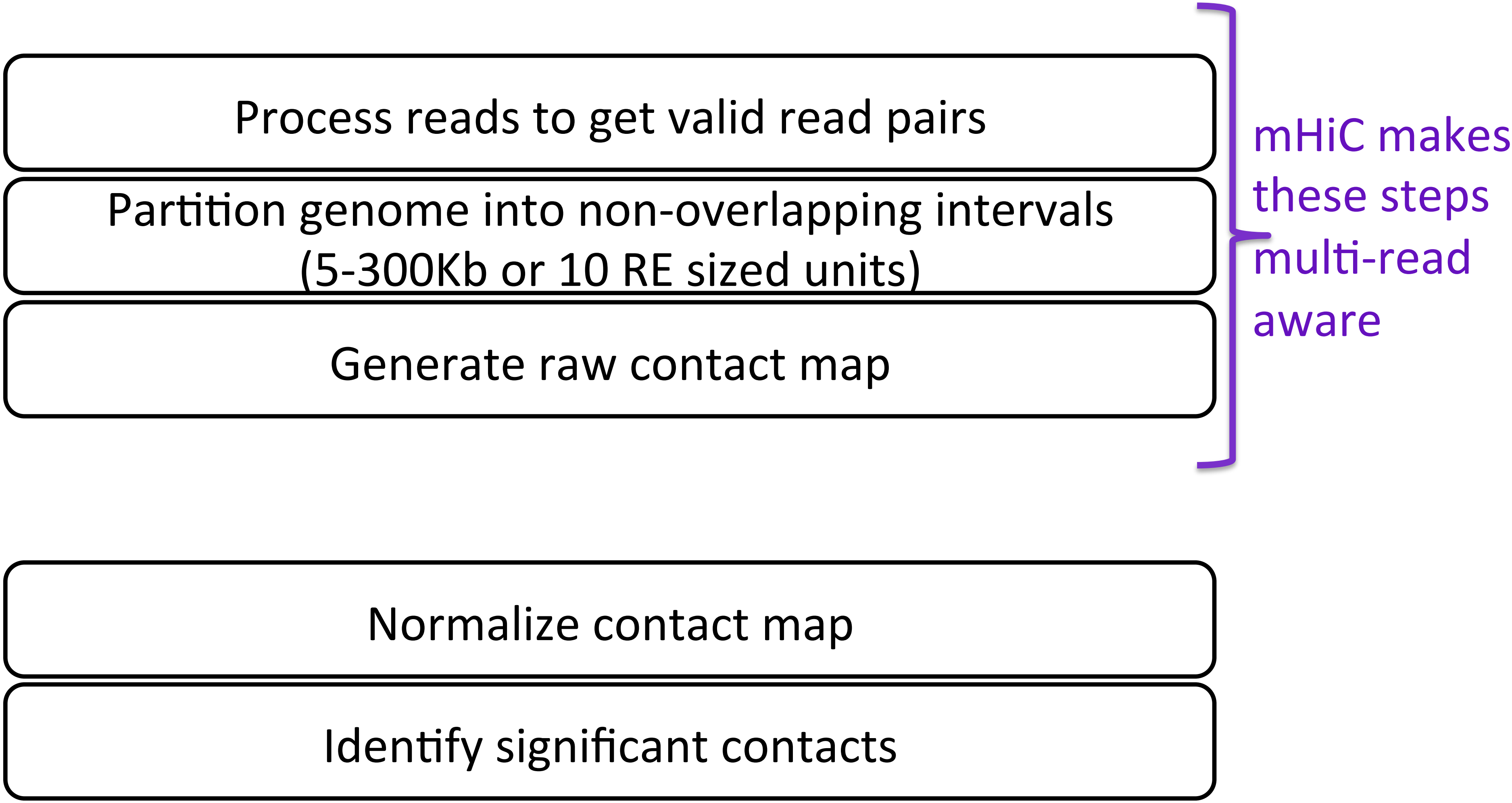
mHi-C

$$P(Z_{i,(j,k)} = 1 \mid Y_{i,(j',k')}, \forall j', k')$$

Threshold posterior probabilities to use resulting alignments with existing significant contact identification methods (e.g., fit-HiC).



mHi-C: from read-pairs to significant contacts



Process reads to get valid read pairs

Partition genome into non-overlapping intervals (5-300Kb or 10 RE sized units)

Generate raw contact map

mHiC makes these steps multi-read aware

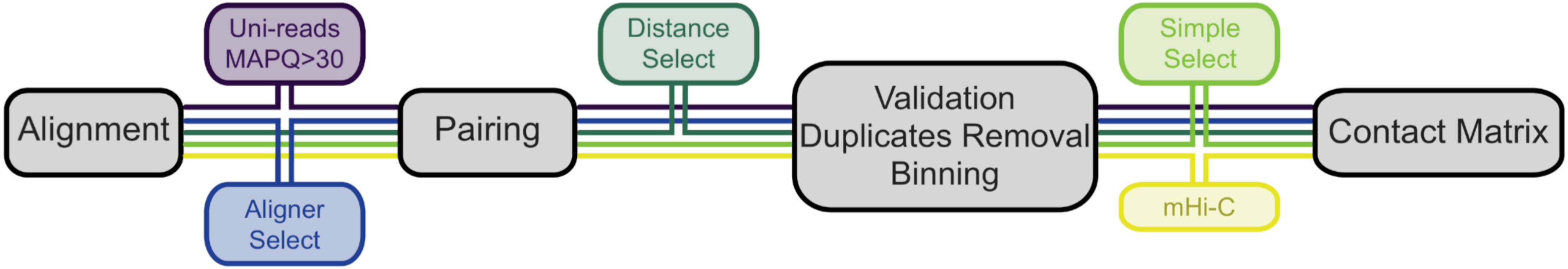
Normalize contact map

Identify significant contacts

Evaluation

A. Sequencing depth	✔
B. Accuracy of multi-read assignment by trimming experiments	
C. Impact on coverage	
D. Reproducibility across replicates: both raw contact count matrix and also identified contacts	
E. Biological impact: Novel promoter-enhancer interactions	
F. Biological impact: TAD inference	

B. Alternative read rescue



Uni-reads MAPQ>30

Alignment

Aligner Select

Pairing

Distance Select

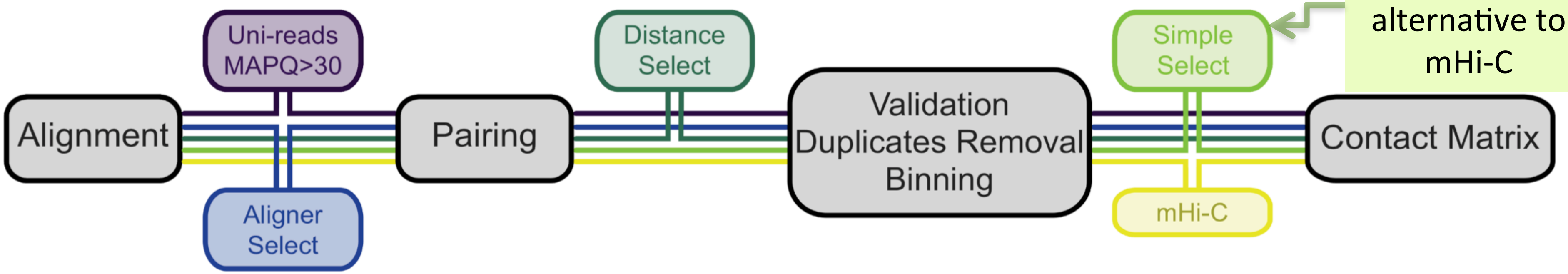
Validation Duplicates Removal Binning

Simple Select

mHi-C

Contact Matrix

B. Accuracy



A nonparametric alternative to mHi-C

Trimming experiments:

Start with long read datasets (e.g., ≥100 bp).

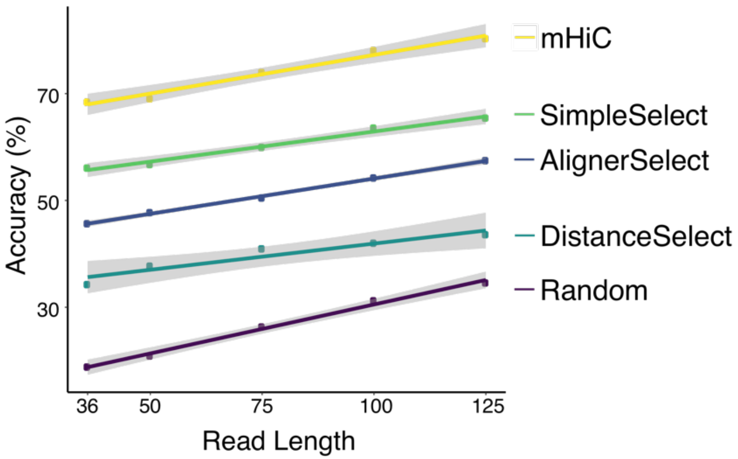
Align and get uni-reads (long uni-reads).

Trim the long uni-reads to generate short reads.

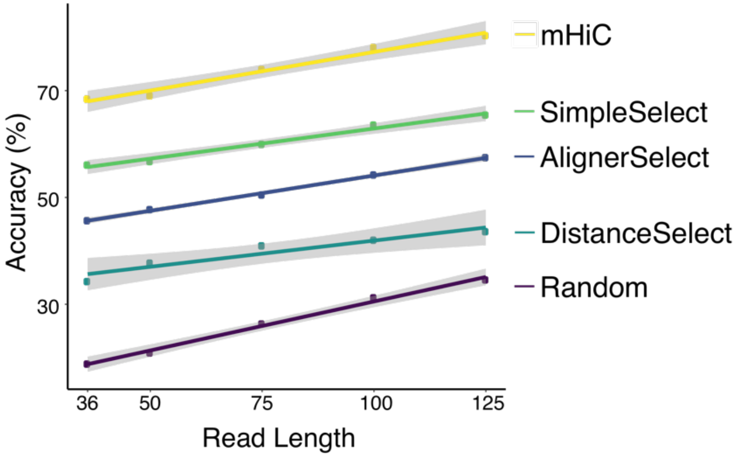
Align trimmed reads, some of which are now multi-reads.

Evaluate them against their true alignment positions from the longer uni-read set.

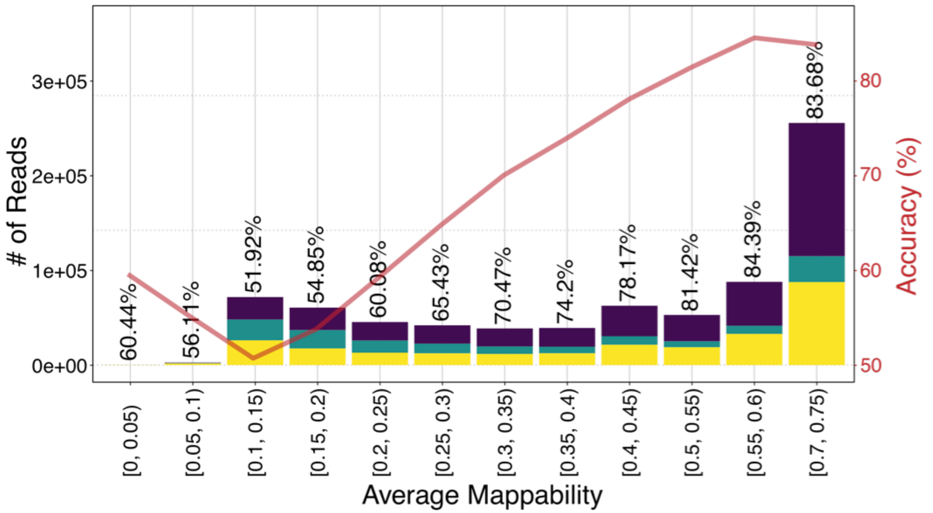
B. Accuracy


Accuracy (%)
70
50
30
36
50
75
100
125
Read Length
mHiC
SimpleSelect
AlignerSelect
DistanceSelect
Random

B. Accuracy

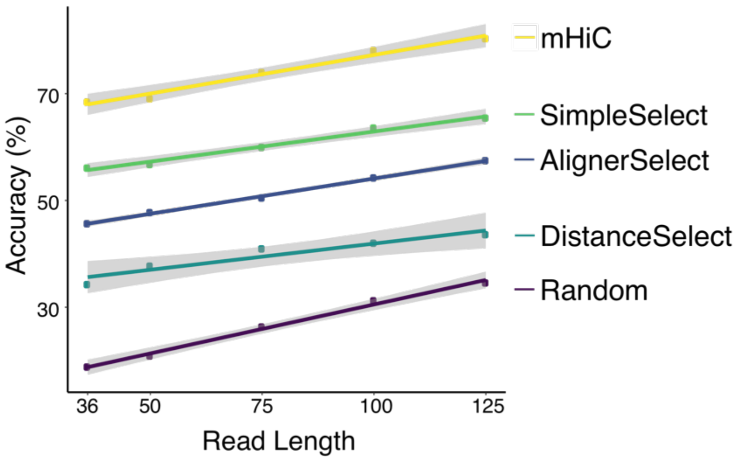


Accuracy (%)
70
50
30
36
50
75
100
125
Read Length
mHiC
SimpleSelect
AlignerSelect
DistanceSelect
Random

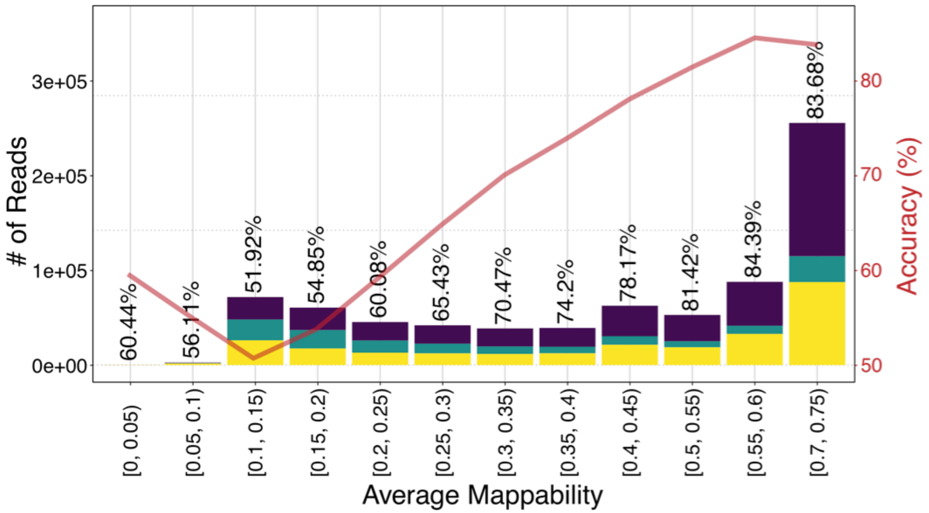


Correctly Assigned
Falsely Assigned
Not Assigned
# of Reads
3e+05
2e+05
1e+05
0e+00
Accuracy (%)
80
70
60
50
60.44%
56.11%
51.92%
54.85%
60.08%
65.43%
70.47%
74.2%
78.17%
81.42%
84.39%
83.68%
[0, 0.05)
[0.05, 0.1)
[0.1, 0.15)
[0.15, 0.2)
[0.2, 0.25)
[0.25, 0.3)
[0.3, 0.35)
[0.35, 0.4)
[0.4, 0.45)
[0.5, 0.55)
[0.55, 0.6)
[0.7, 0.75)
Average Mappability

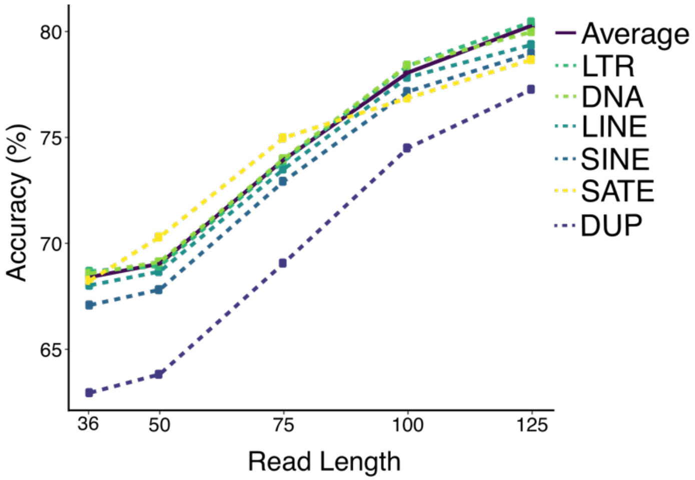
B. Accuracy



Accuracy (%)
70
50
30
36
50
75
100
125
Read Length
mHiC
SimpleSelect
AlignerSelect
DistanceSelect
Random

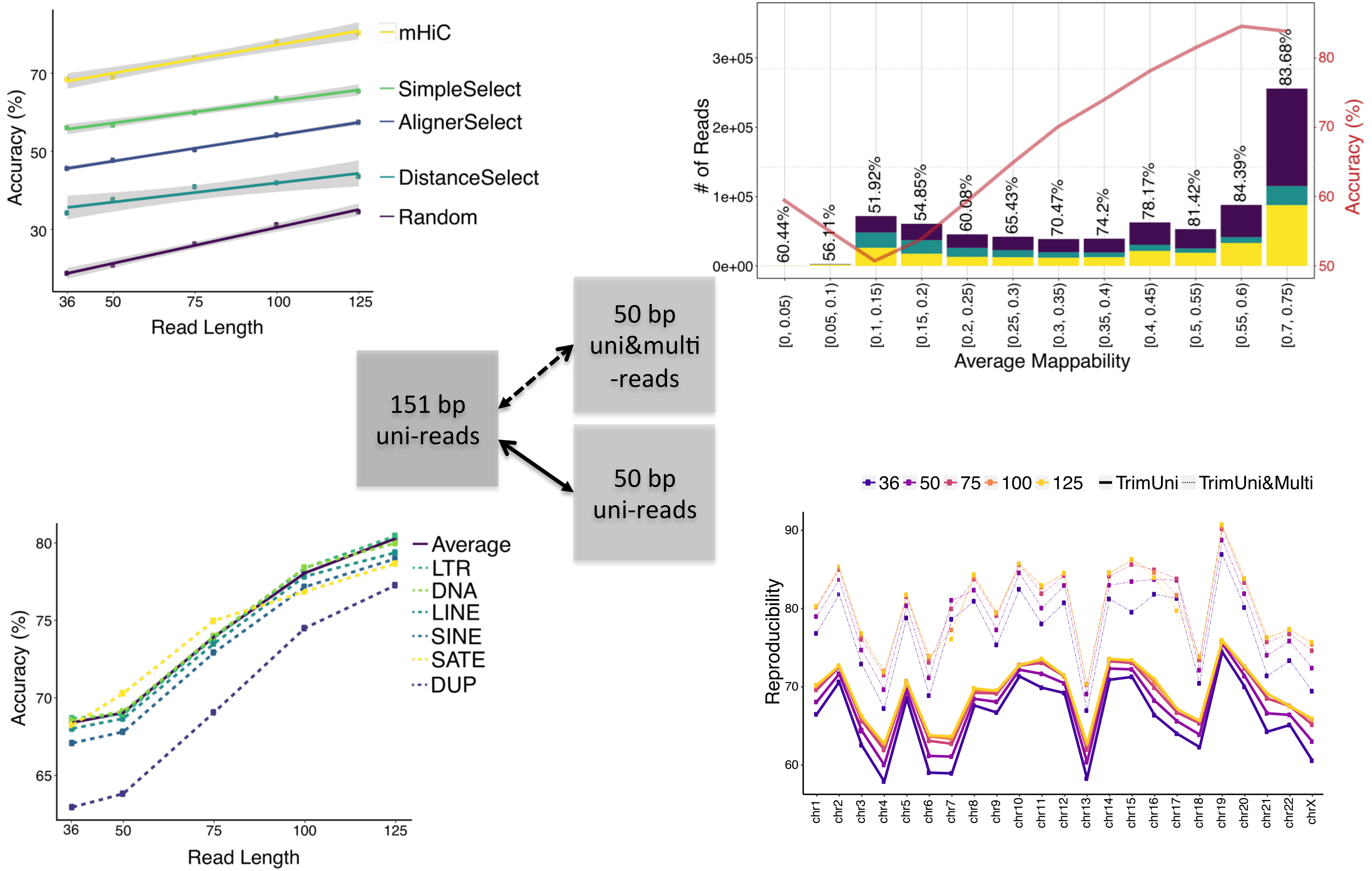


# of Reads
3e+05
2e+05
1e+05
0e+00
Accuracy (%)
80
70
60
50
60.44%
56.11%
51.92%
54.85%
60.08%
65.43%
70.47%
74.2%
78.17%
81.42%
84.39%
83.68%
[0, 0.05)
[0.05, 0.1)
[0.1, 0.15)
[0.15, 0.2)
[0.2, 0.25)
[0.25, 0.3)
[0.3, 0.35)
[0.35, 0.4)
[0.4, 0.45)
[0.5, 0.55)
[0.55, 0.6)
[0.7, 0.75)
Average Mappability



Accuracy (%)
80
75
70
65
36
50
75
100
125
Read Length
Average
LTR
DNA
LINE
SINE
SATE
DUP

B. Recovering the full length contact matrix

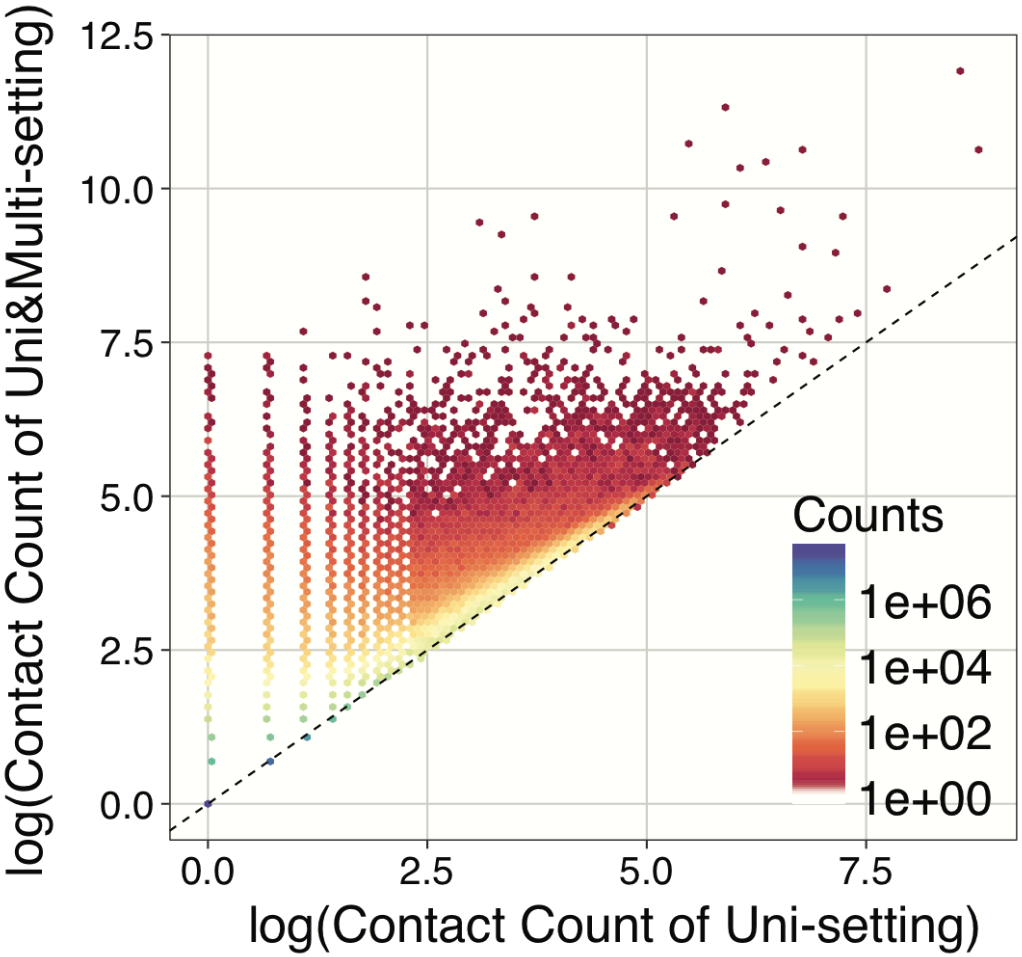


151 bp uni-reads

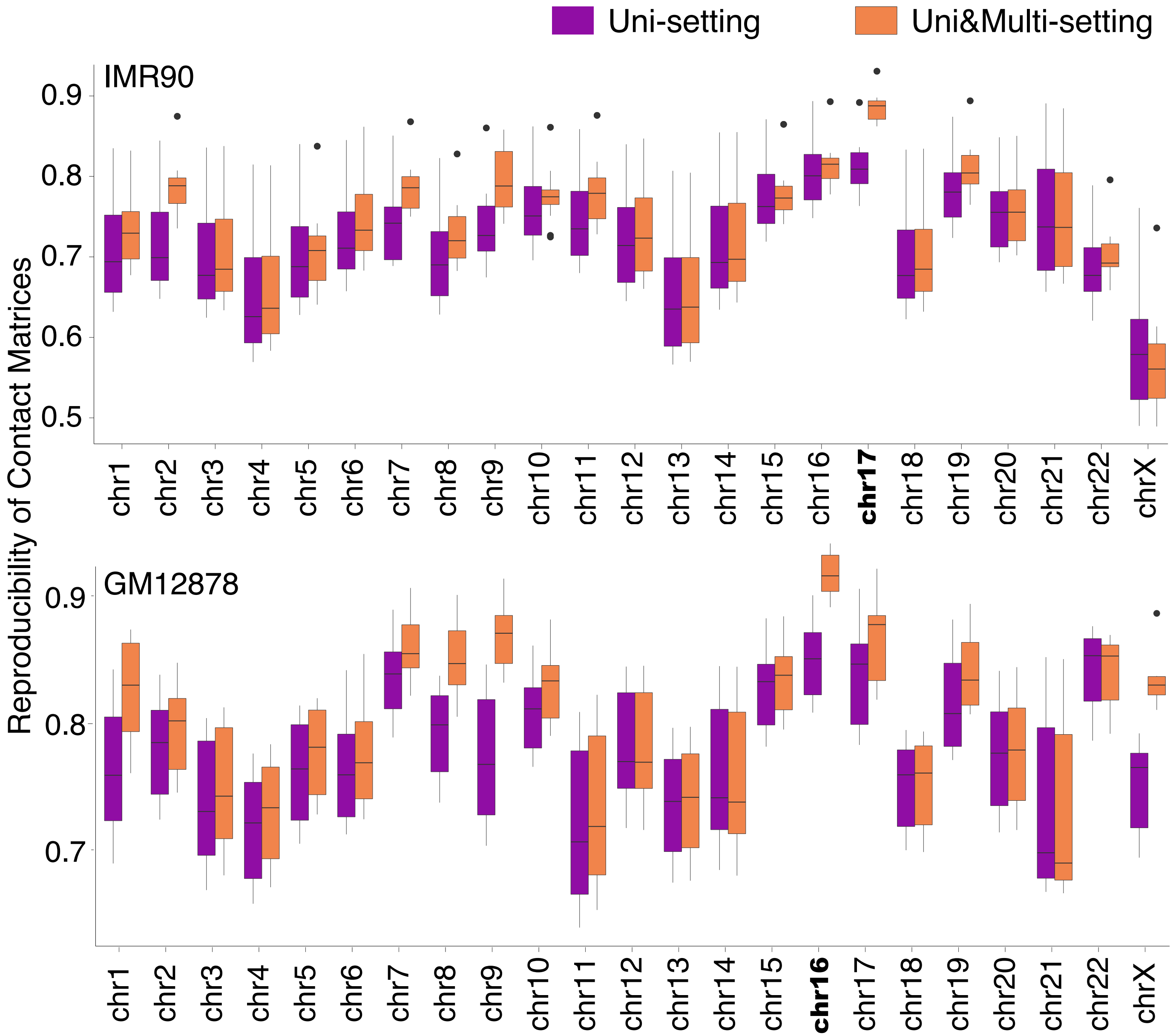
50 bp uni&multi -reads

50 bp uni-reads

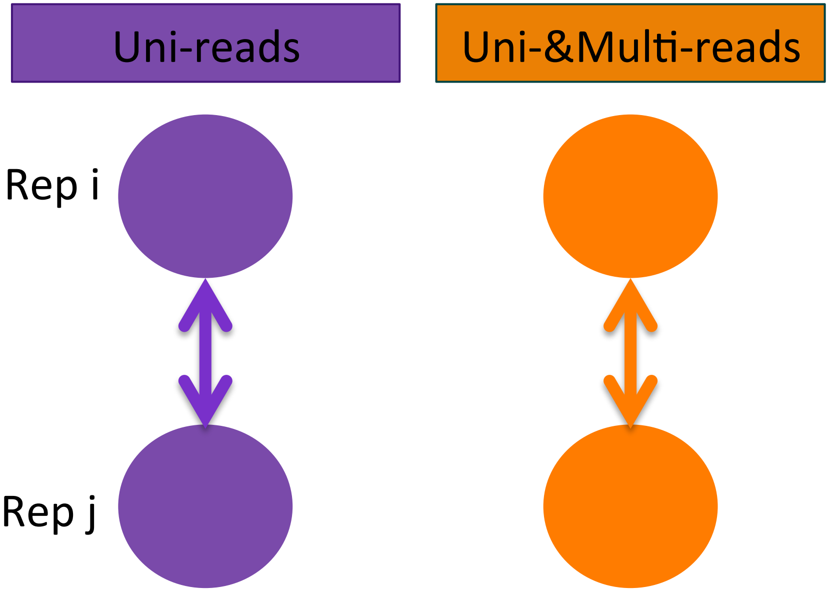
C. Major improvement in coverage



D. Reproducibility of the contact matrix



D. Reproducibility of the significant interactions

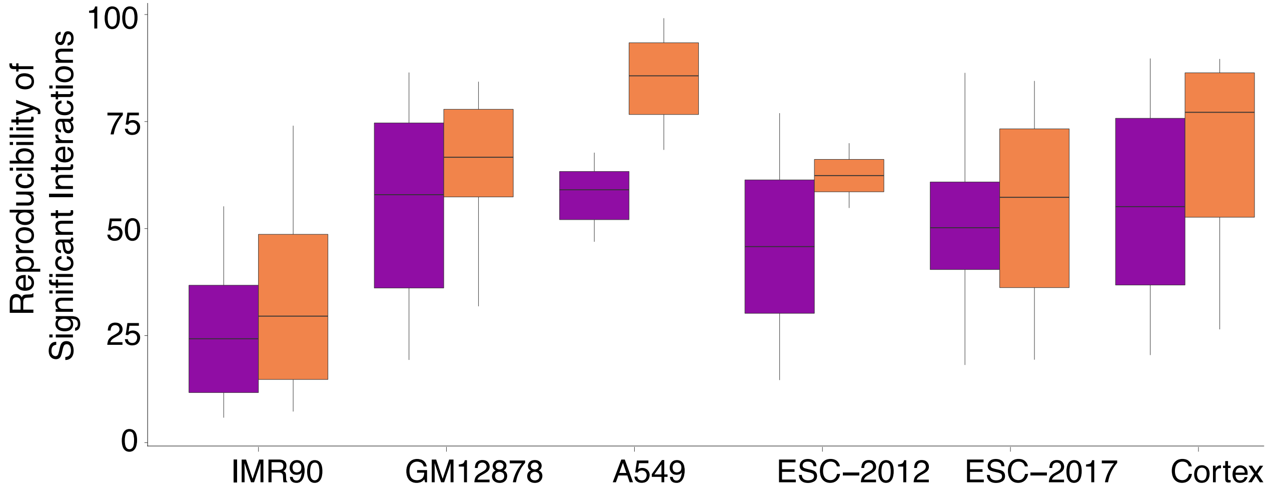


Uni-reads

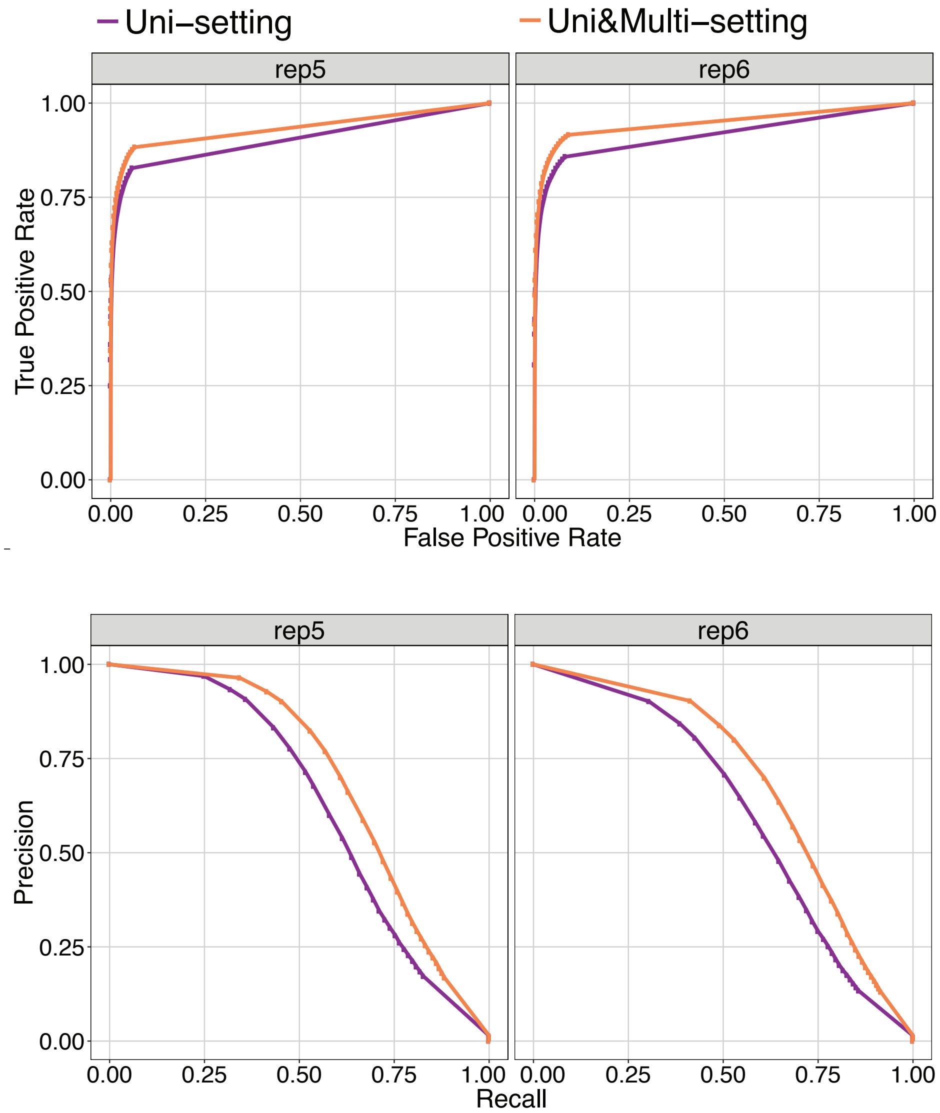
Uni-&Multi-reads

Rep i

Rep j



D. ROC- and PR-based on replicate gold standard

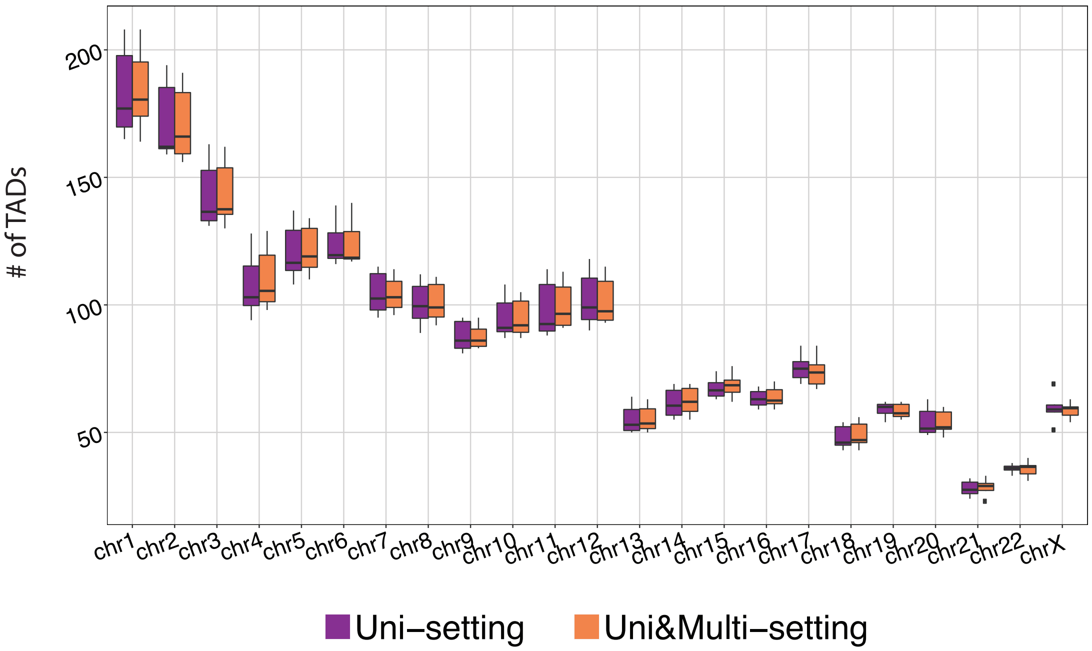


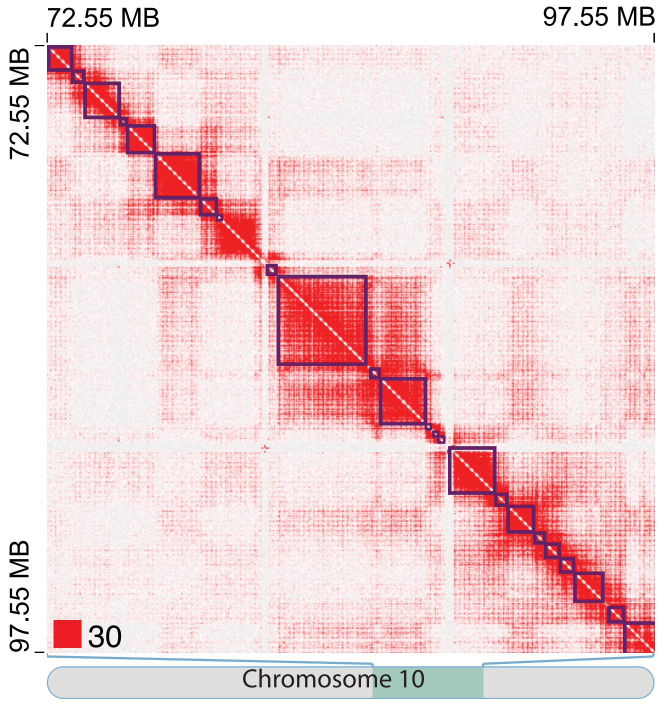
High depth replicates are used to define "true" positives and negatives.

E. Impact on TAD inference

TAD: Topologically associated domain

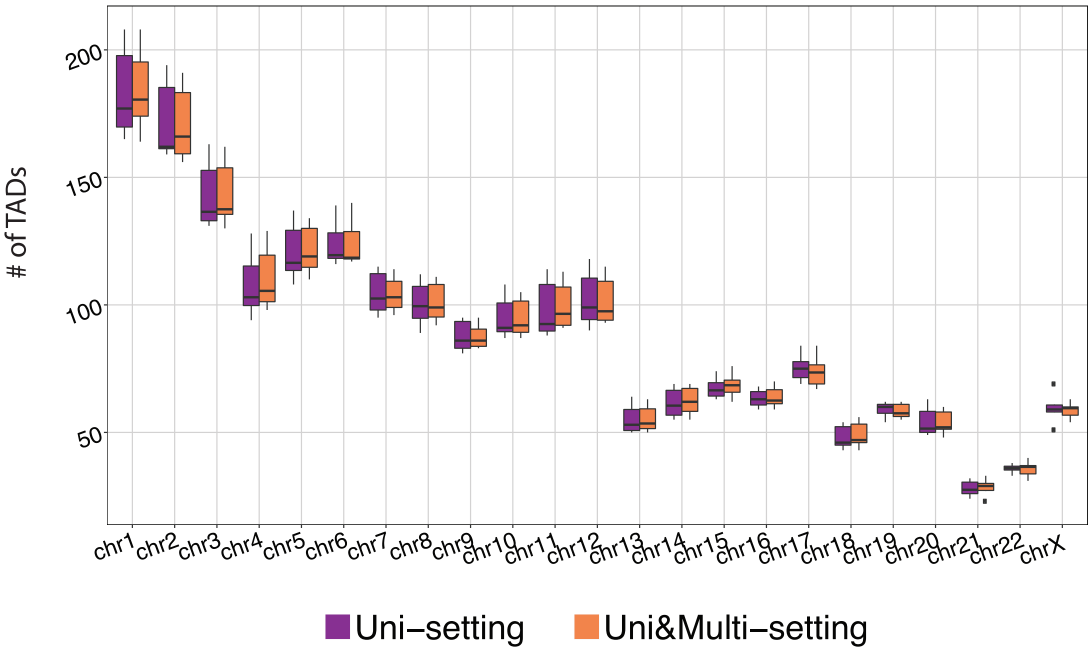
# of TADs detected do not change significantly.

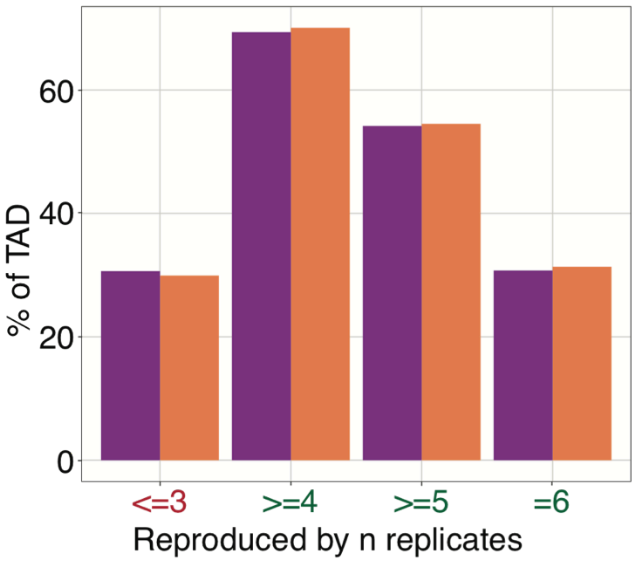




E. Impact on TAD inference

# of TADs detected do not change significantly.

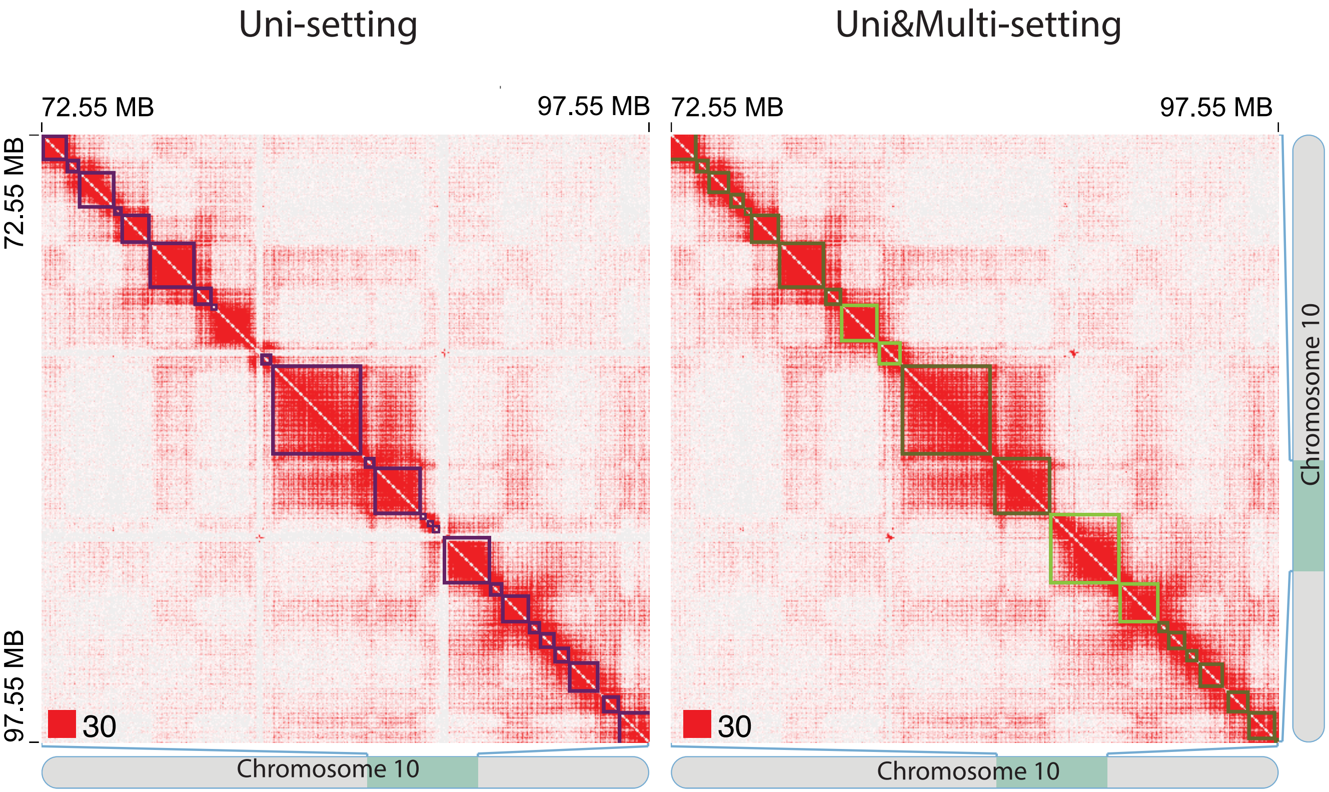




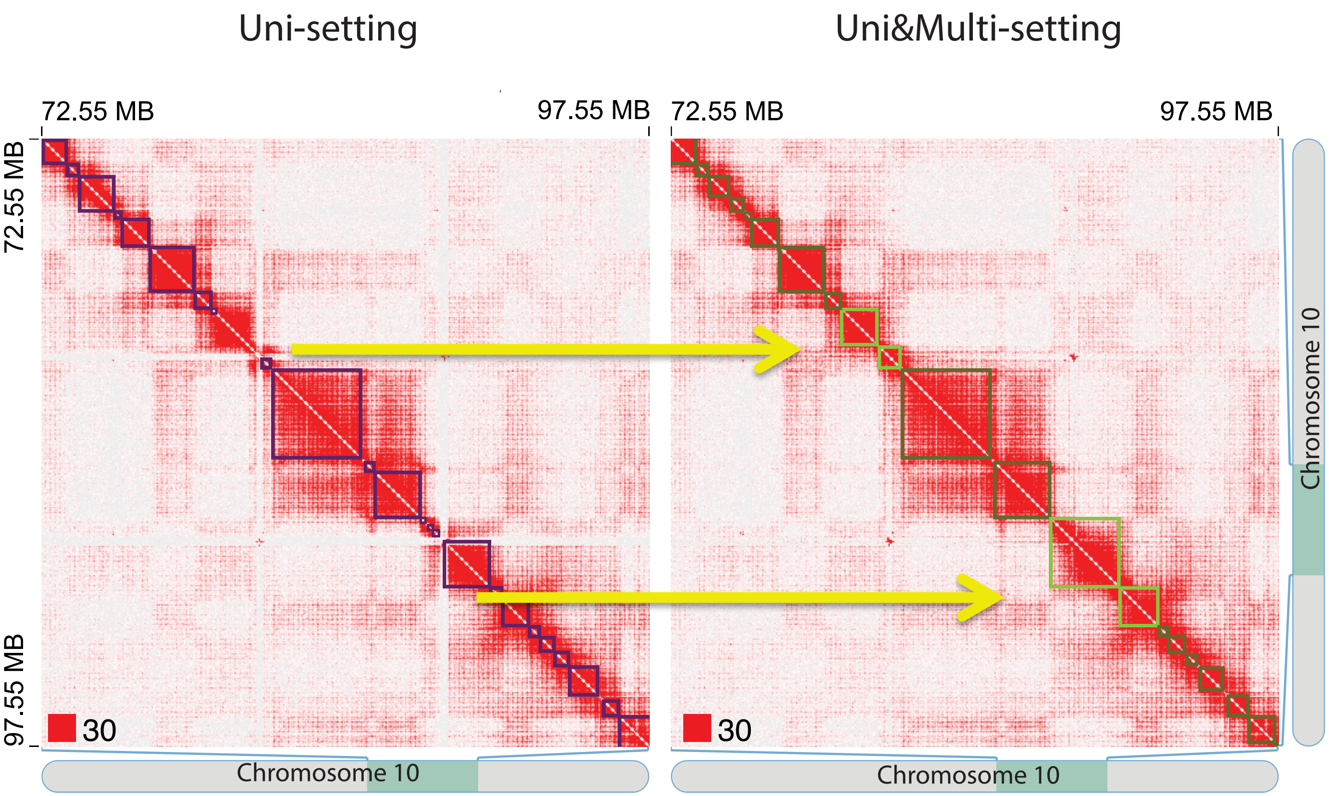
# of reproducible TADs increases by 2.01%.

# of irreproducible TADs decreases by 2.36%.

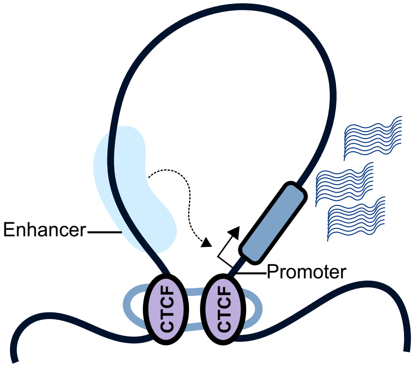
E. Impact on TAD inference

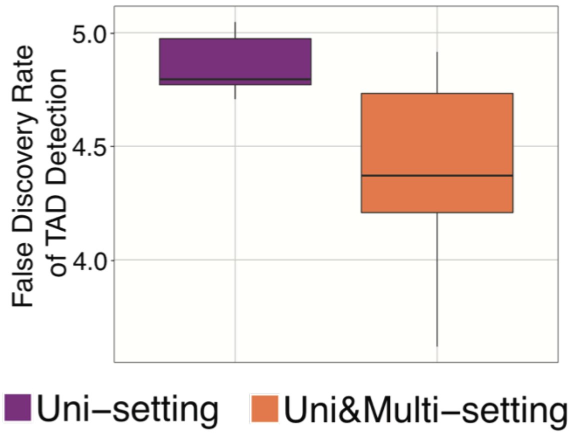


E. Impact on TAD inference



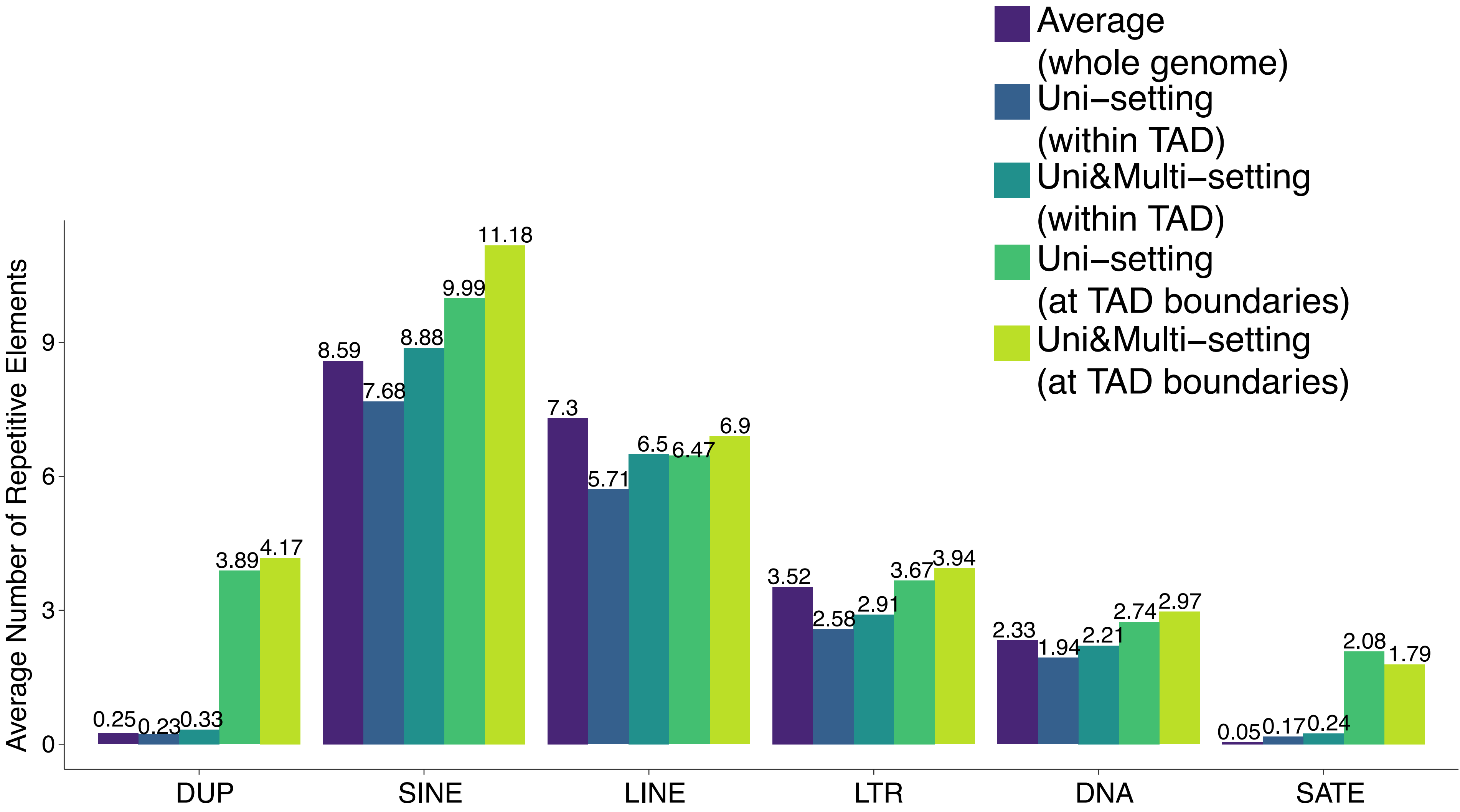
Arzate-Mejia *et al.*, 2018



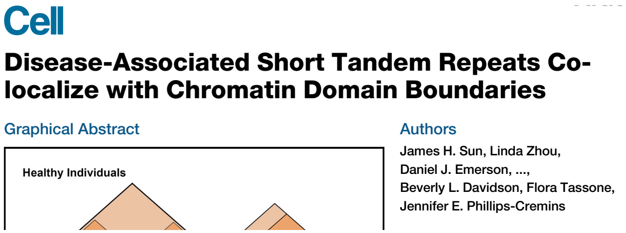


TADs that are not reproducible and lack CTCF peaks at the TAD boundaries are labeled as false positives

F. Repetitive elements at the boundaries of reproducible TADs



F. Disease-Associated short tandem repeats co-localize with domain boundaries



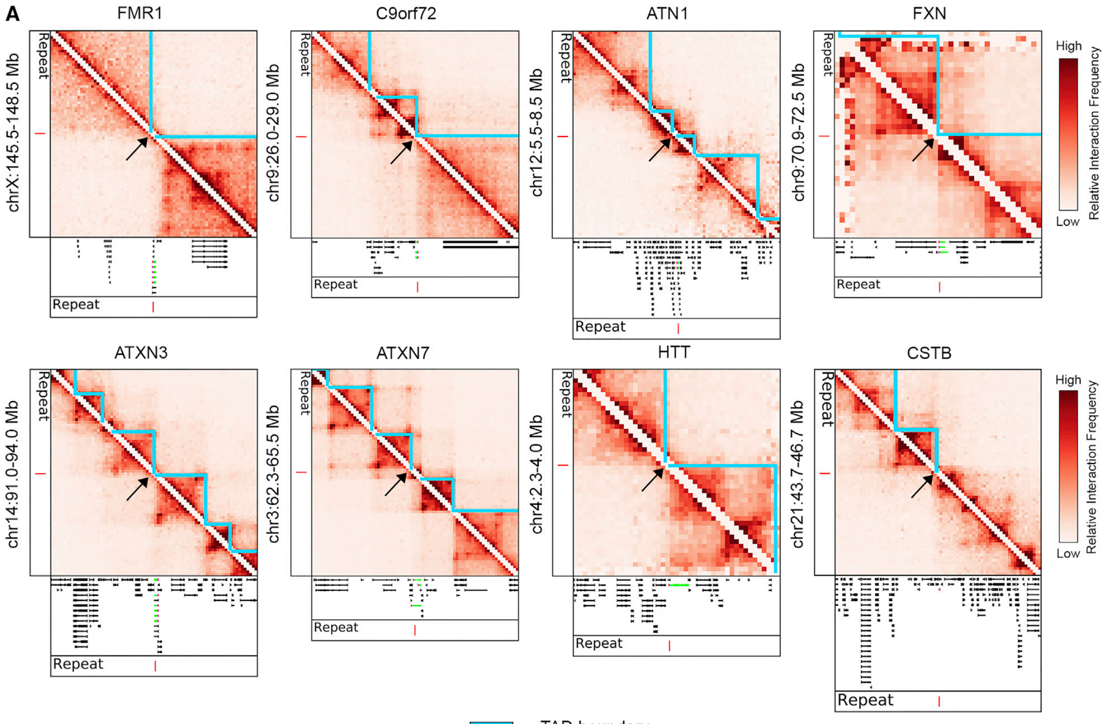
Cell

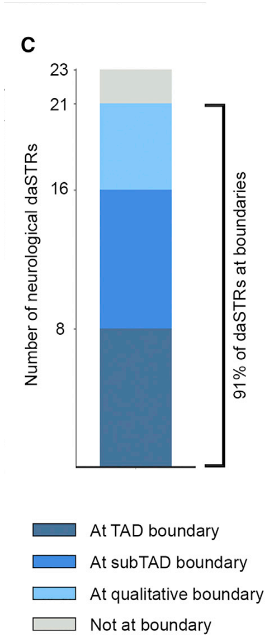
Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries

Graphical Abstract

Authors

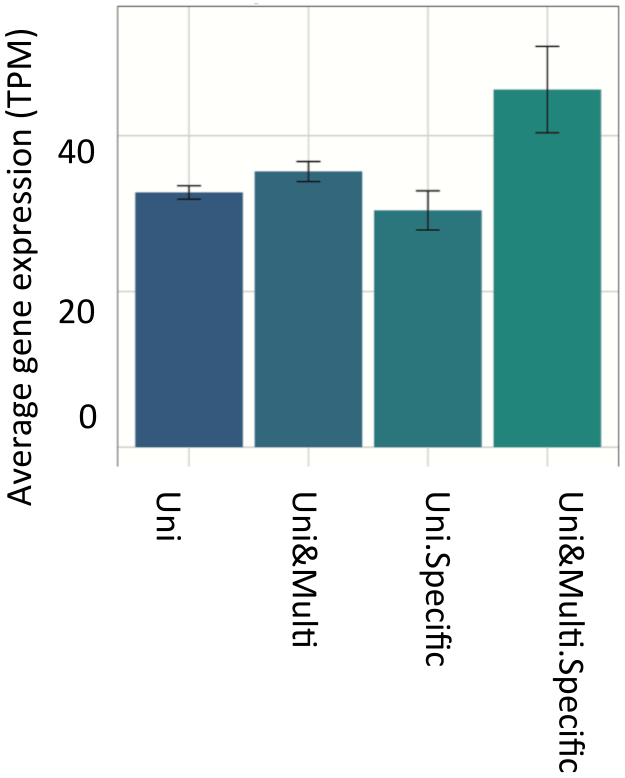
James H. Sun, Linda Zhou, Daniel J. Emerson, ..., Beverly L. Davidson, Flora Tassone, Jennifer E. Phillips-Cremins





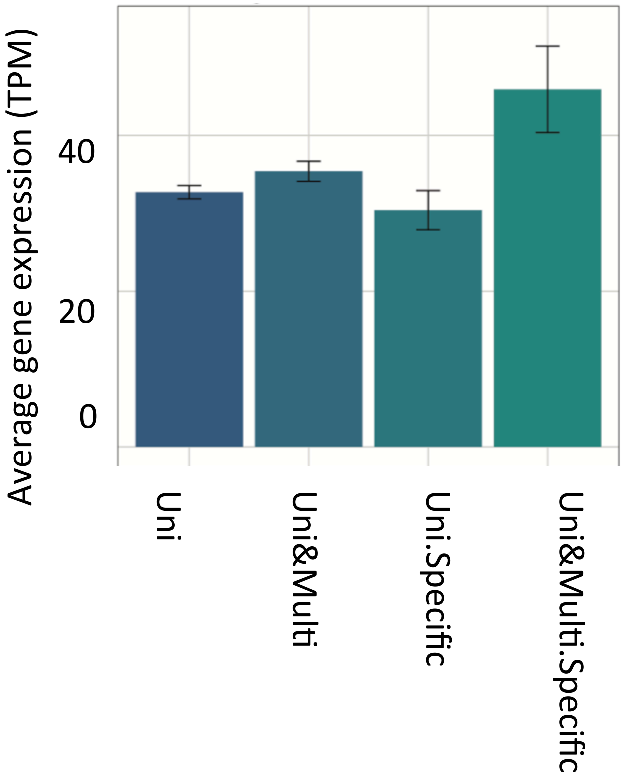
F. Novel promoter-enhancer interactions

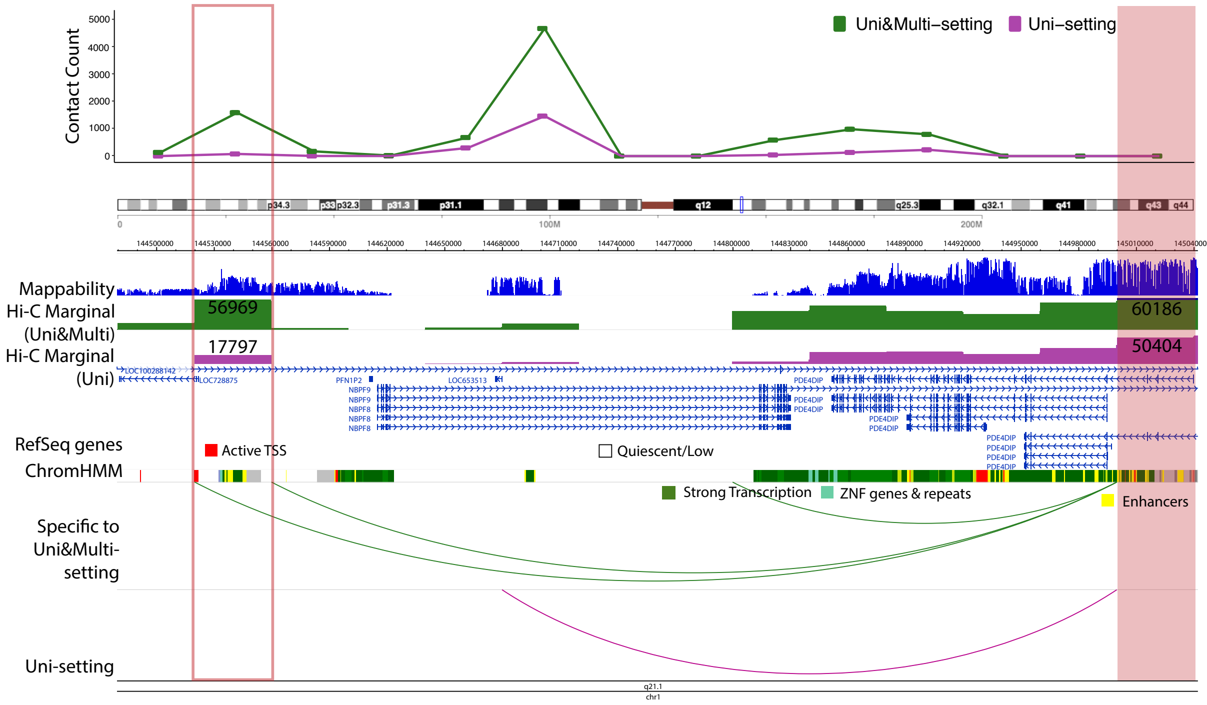
15.8% more promoter-enhancer interactions that are reproducible in at least 2 replicates.



F. Novel promoter-enhancer interactions

15.8% more promoter-enhancer interactions that are reproducible in at least 2 replicates.





Summary

- Software

  https://github.com/keleslab/mhic

- Paper

  https://www.biorxiv.org/content/early/2018/10/03/301705

- More results on chimeric reads, impact on differential Hi-C analysis are available in the manuscript.

Acknowledgements

Keleș Group

Ye Zheng

Collaborators

Ferhat Ay (La Jolla Institute for Allergy & Immunology)

Center for Predictive Computational Phenotyping

U54 AI117924

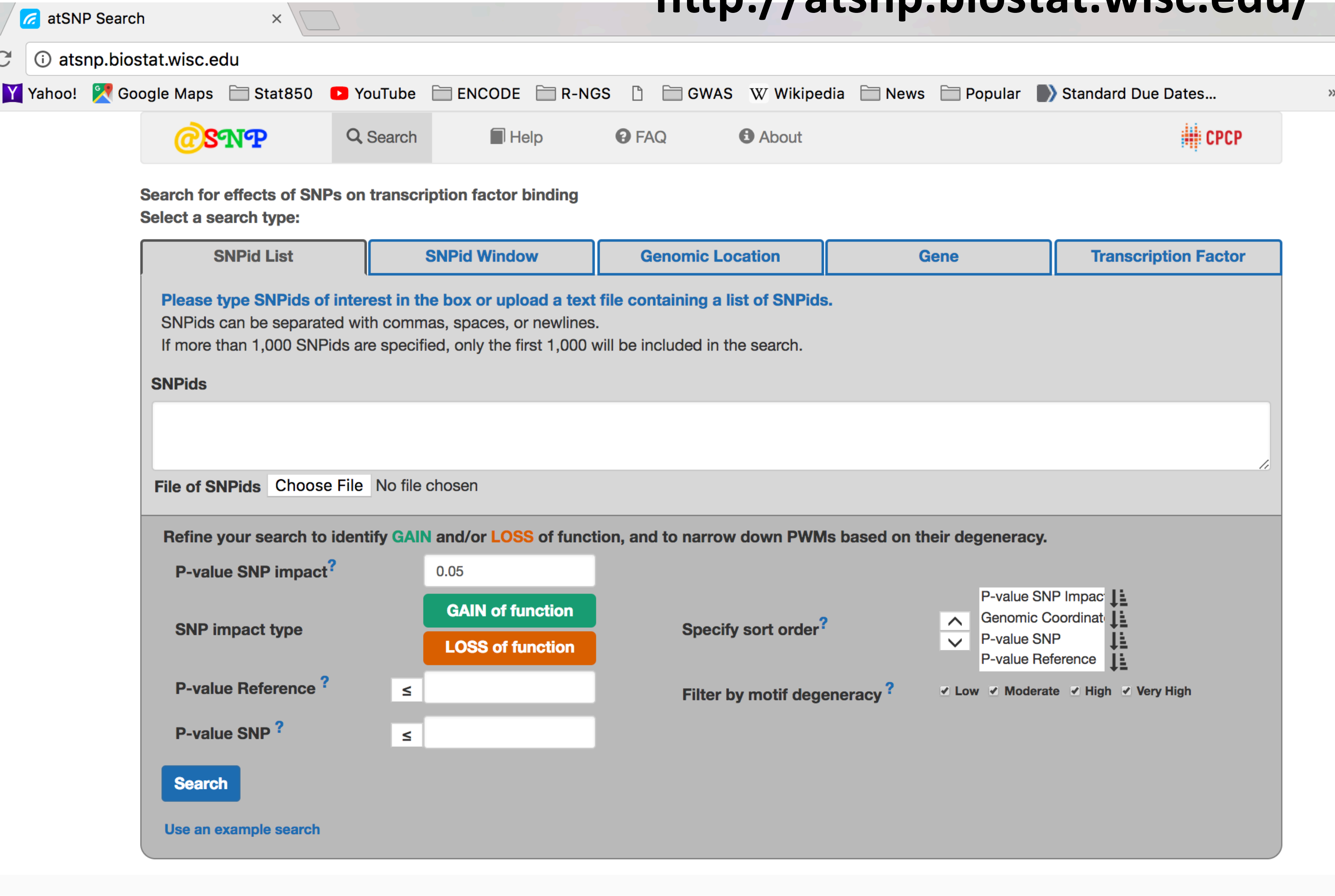
ENCODE

U01 HG007019

Software: https://github.com/keleslab

Thanks to NIH: RO1 HG003747, R21 HG009744

http://atsnp.biostat.wisc.edu/



Search for effects of SNPs on transcription factor binding

Select a search type:

SNPid List | SNPid Window | Genomic Location | Gene | Transcription Factor

Please type SNPids of interest in the box or upload a text file containing a list of SNPids.

SNPids can be separated with commas, spaces, or newlines.

If more than 1,000 SNPids are specified, only the first 1,000 will be included in the search.

SNPids

File of SNPids Choose File No file chosen

Refine your search to identify GAIN and/or LOSS of function, and to narrow down PWMs based on their degeneracy.

P-value SNP impact? 0.05

SNP impact type GAIN of function LOSS of function

P-value Reference ? ≤

P-value SNP ? ≤

Specify sort order? P-value SNP Impac / Genomic Coordinat / P-value SNP / P-value Reference

Filter by motif degeneracy ? ☑ Low ☑ Moderate ☑ High ☑ Very High

Search

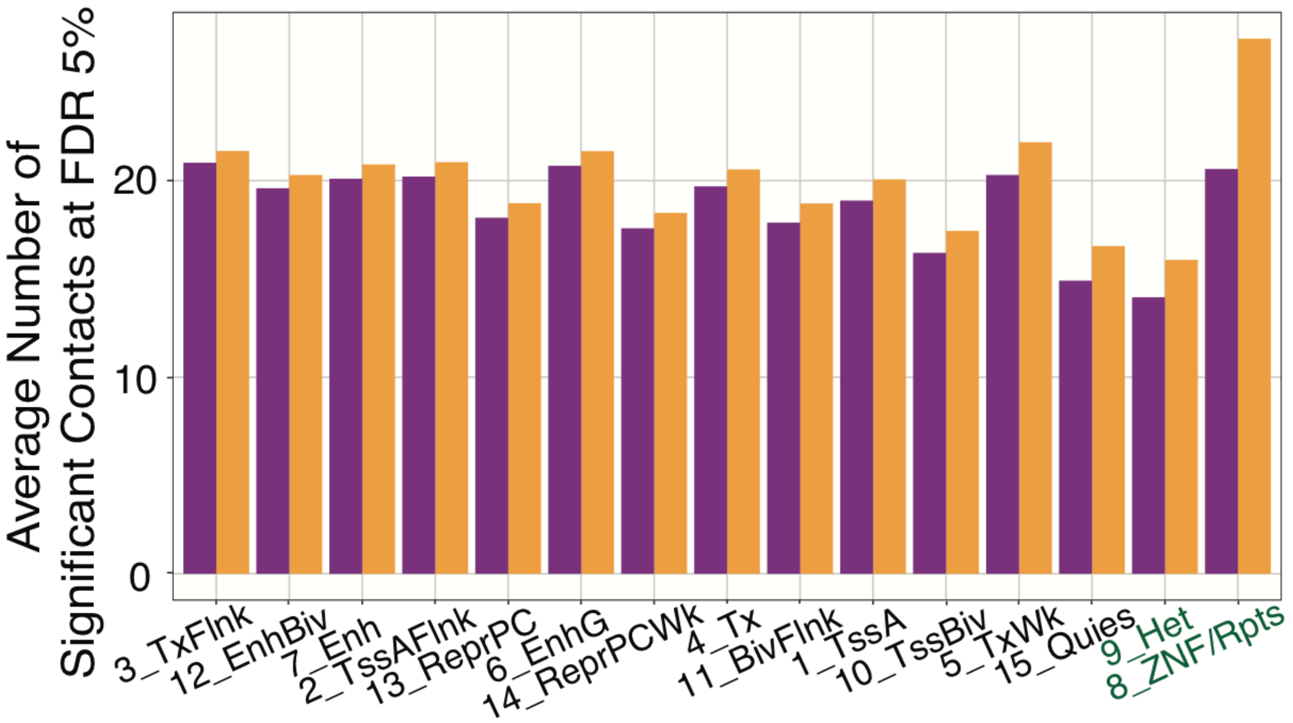
Use an example search

Available positions

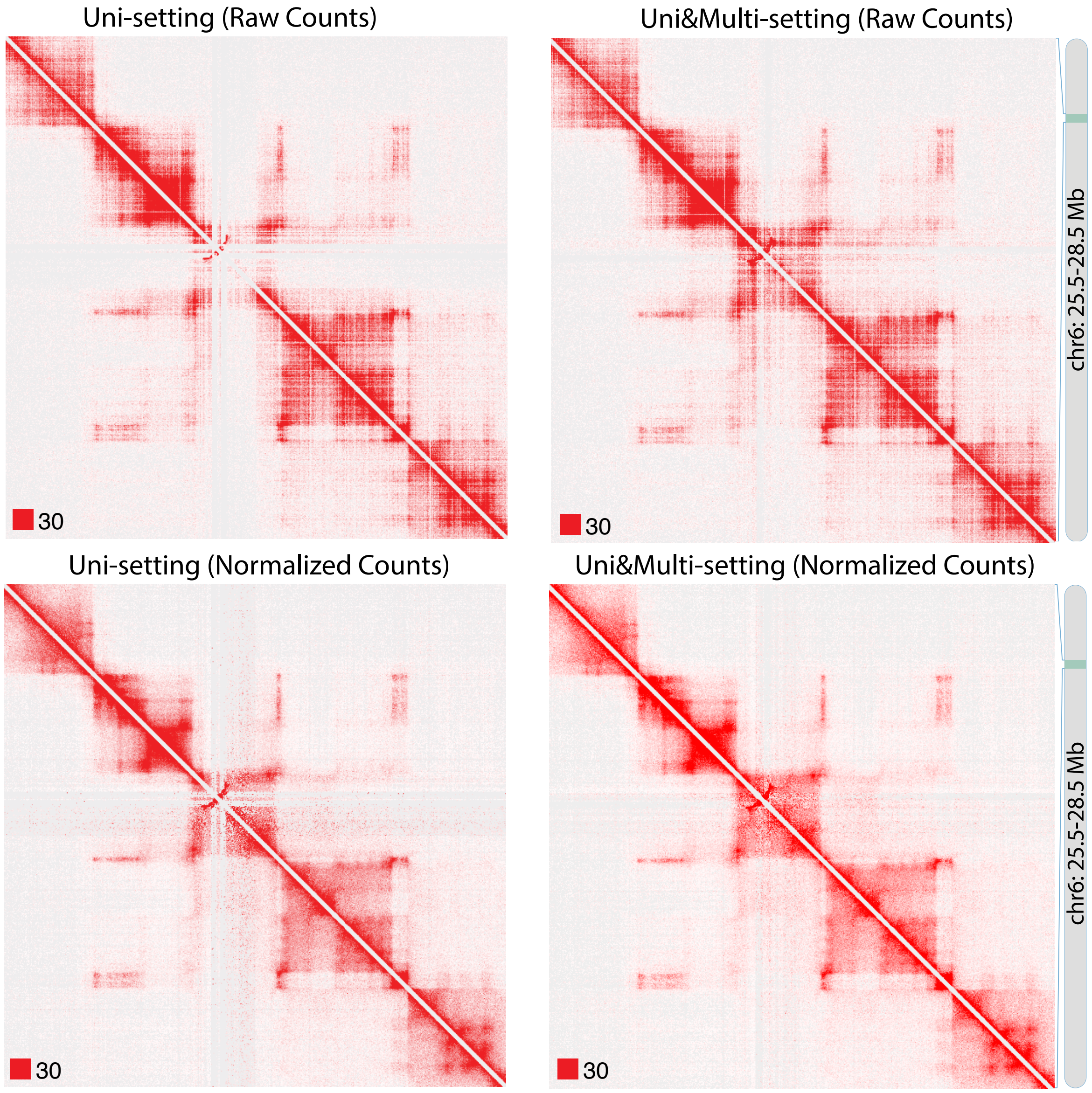
1-2 postdoctoral researcher positions in statistical genomics.

If interested, send CV to keles@stat.wisc.edu

F. Genomic characteristics



C. Count matrices



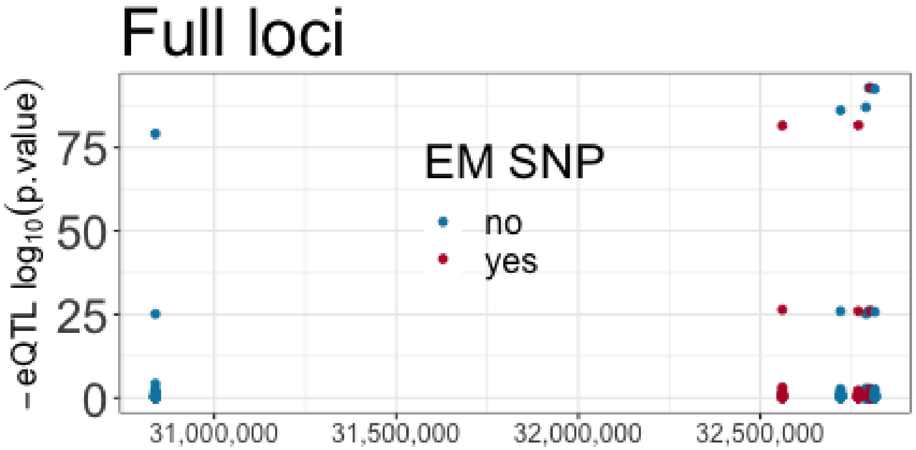
Uni-setting (Raw Counts)

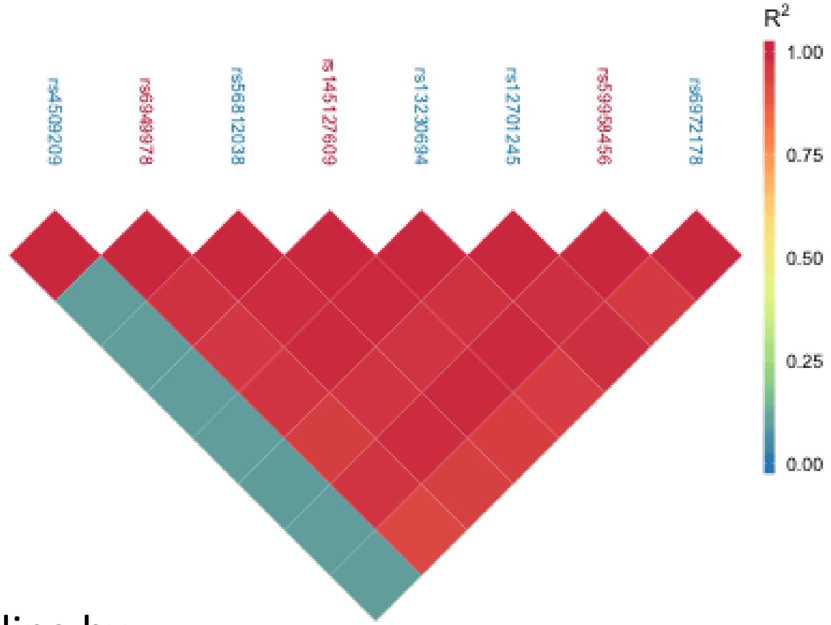
Uni&Multi-setting (Raw Counts)

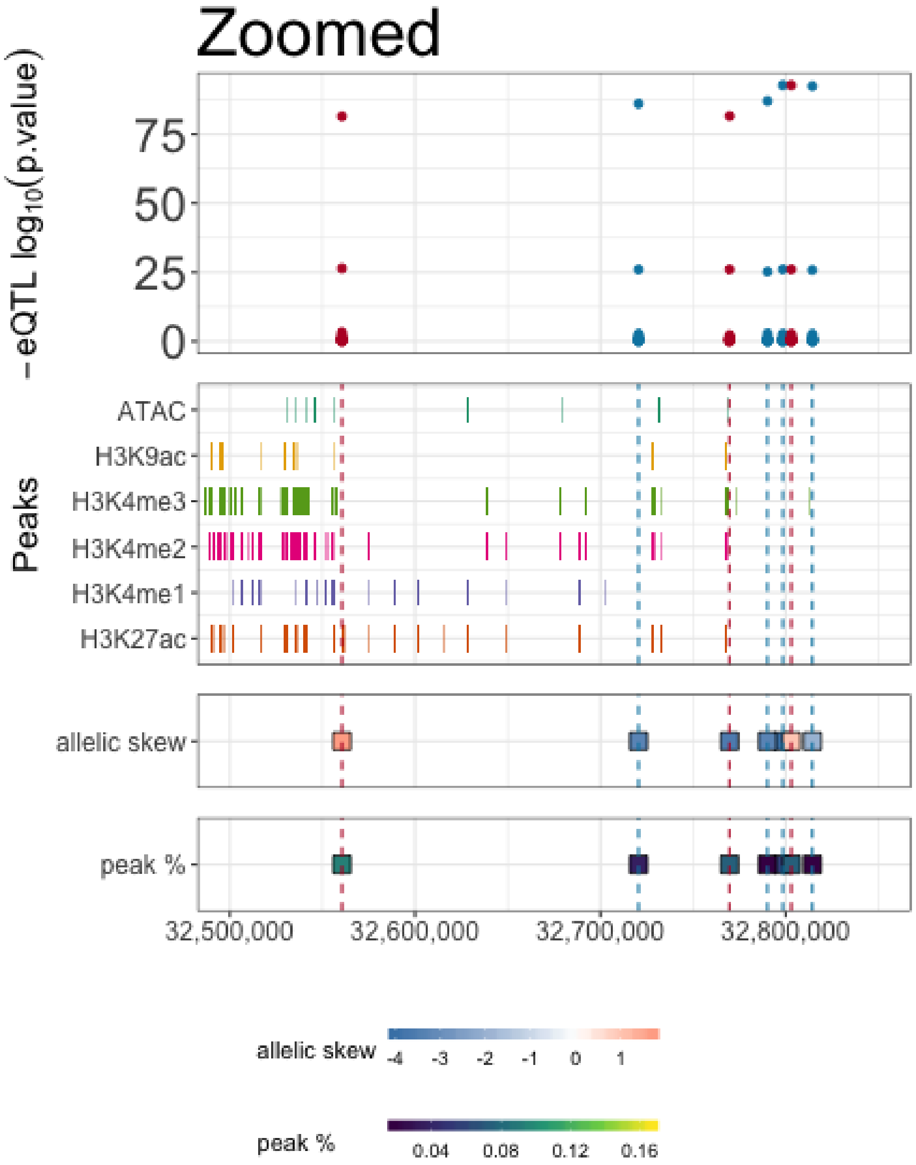
Uni-setting (Normalized Counts)

Uni&Multi-setting (Normalized Counts)

POSTER
SNPs in high LD: a formidable challenge







Labeling by
Massively parallel reporter assays (MPRA)