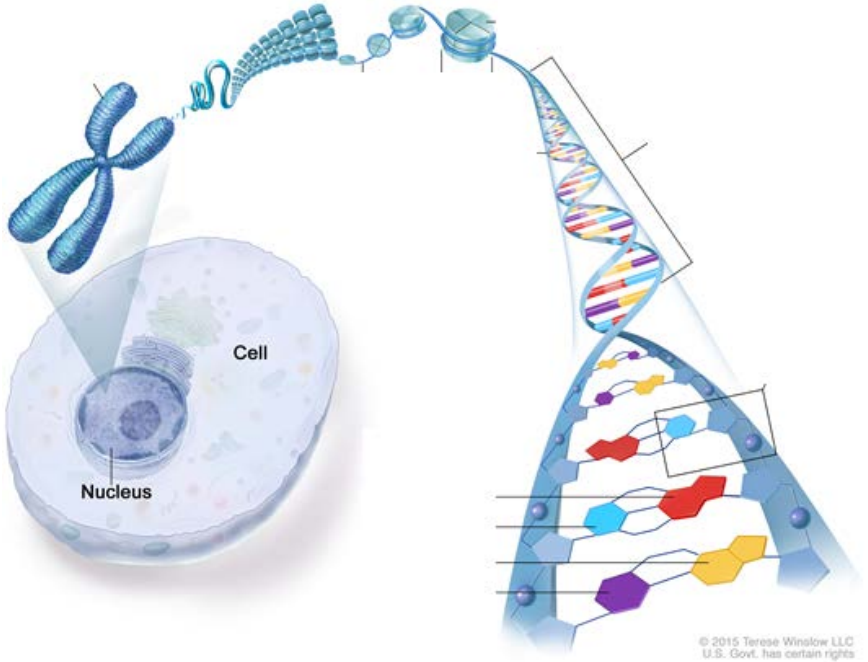
A set of tools for analysis of Hi-C data from normal and cancer genomes

Ferhat Ay, Ph.D.
Institute Leadership Assistant
Professor of Computational Biology
lji.org/faculty-research/labs/ay



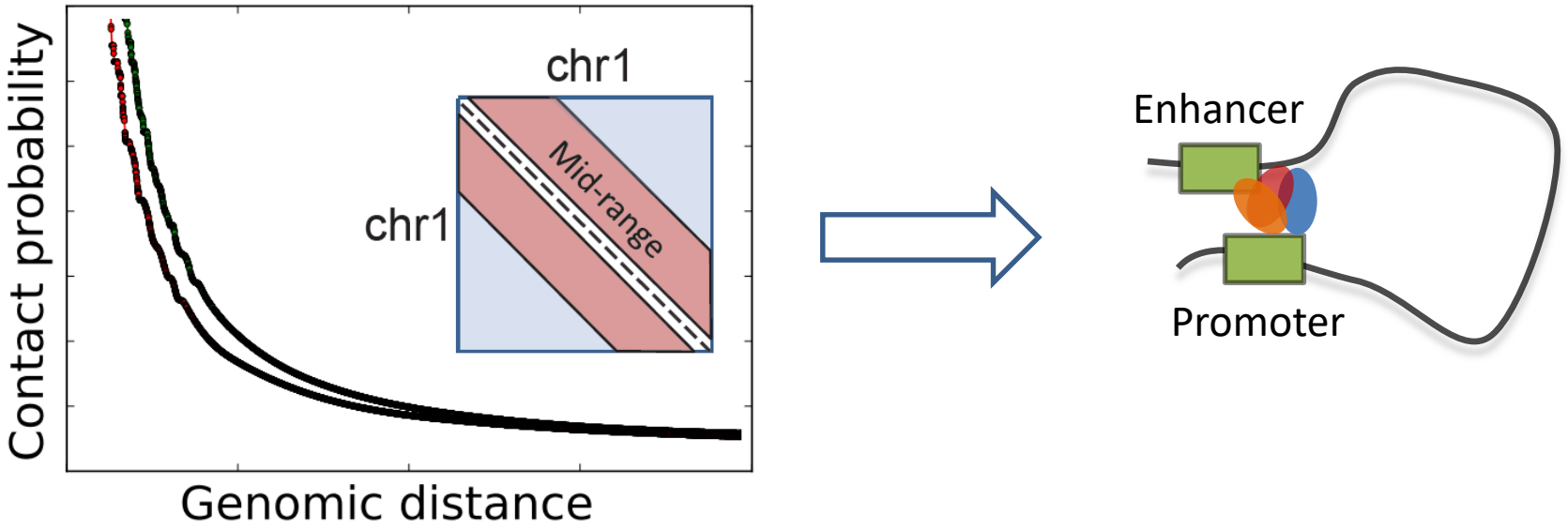
EN-TEx call

5/22/17

La Jolla Institute FOR ALLERGY AND IMMUNOLOGY | Life Without Disease.

Fit-Hi-C

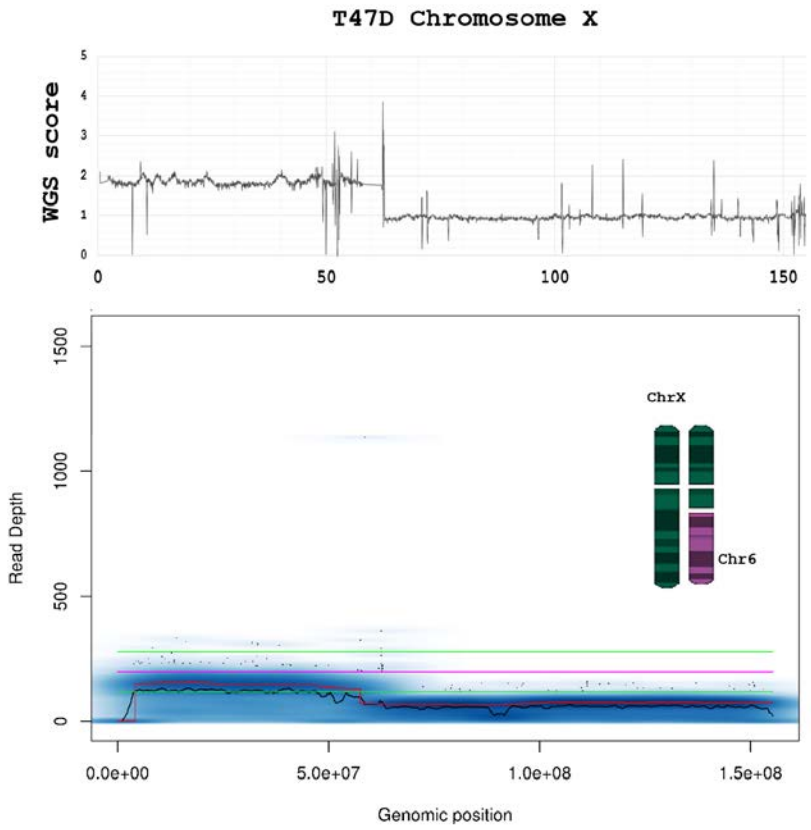
Assigning statistical confidence estimates to chromatin contact maps

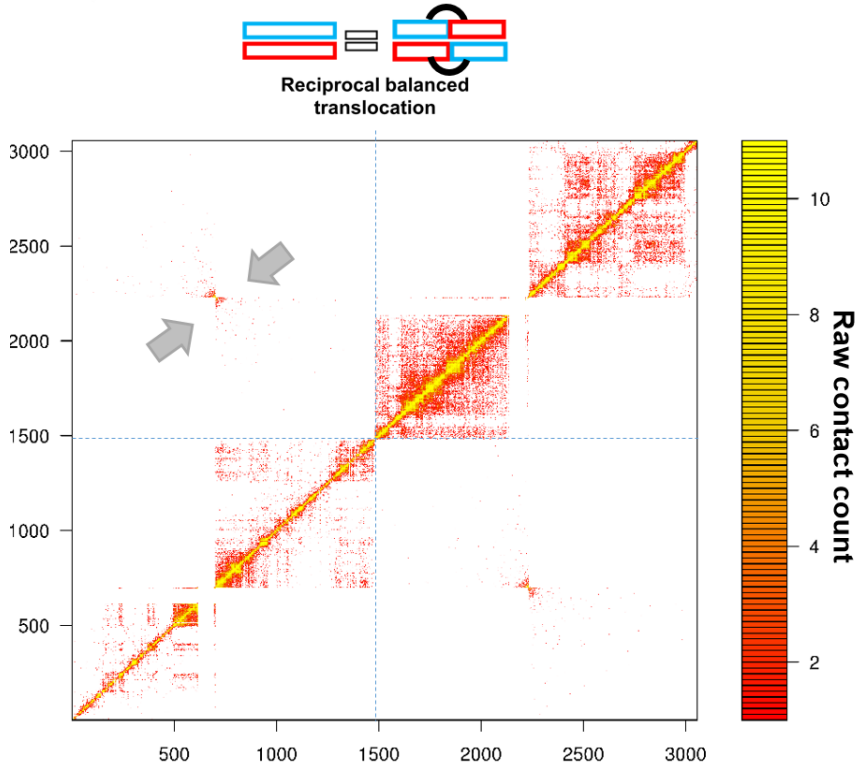


Python: http://noble.gs.washington.edu/proj/fit-hi-c
R: https://bioconductor.org/packages/release/bioc/html/FitHiC.html
Ay, Bailey & Noble. Genome Research, 2014.

HiCnv, HiCtrans & AveSim

Identification of copy number variations and translocations in cancer cells from Hi-C data





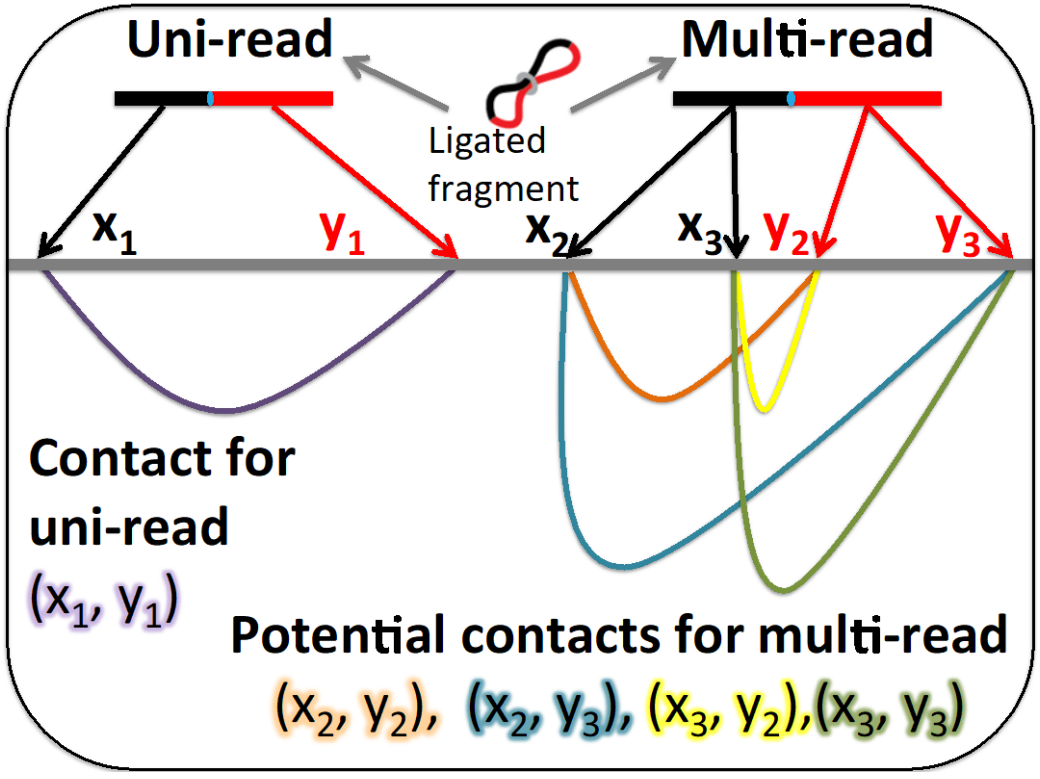
Abhijit Chakraborty

https://github.com/ay-lab

Chakraborty & Ay. Under review.

mHiC

Leveraging multi-mapping reads in Hi-C data

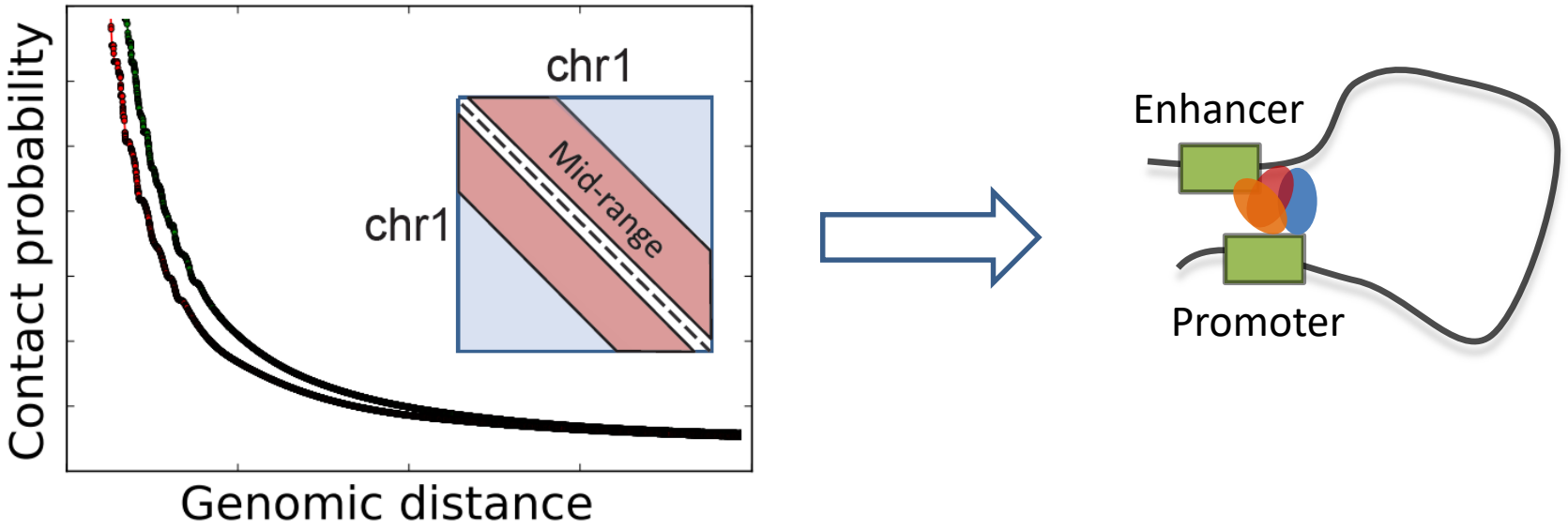


U. of Wisconsin - Madison
Sunduz Keles
Ye Zheng

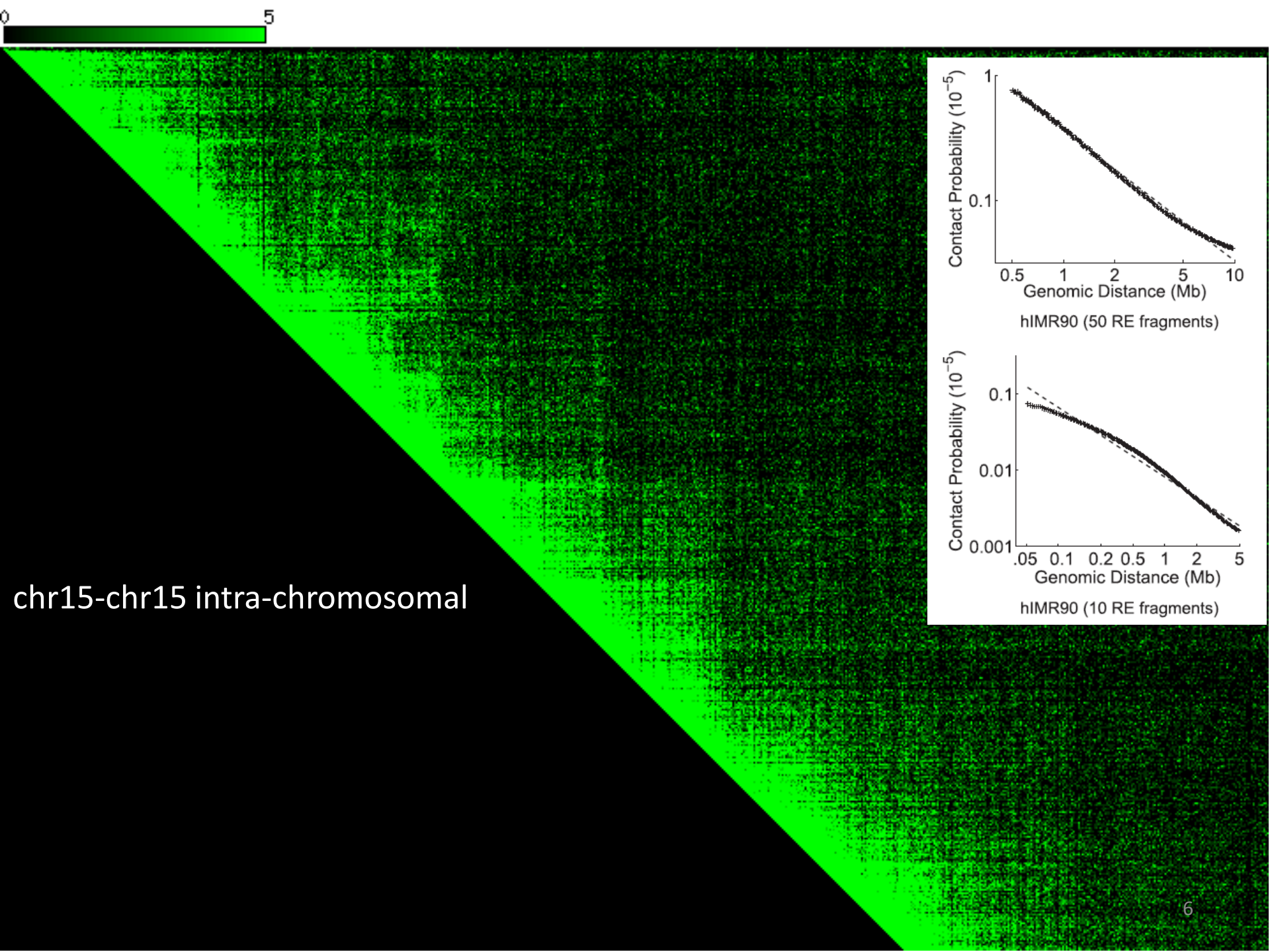
mHiC: a beta version is available from Ye Zheng yezheng@stat.wisc.edu

Fit-Hi-C

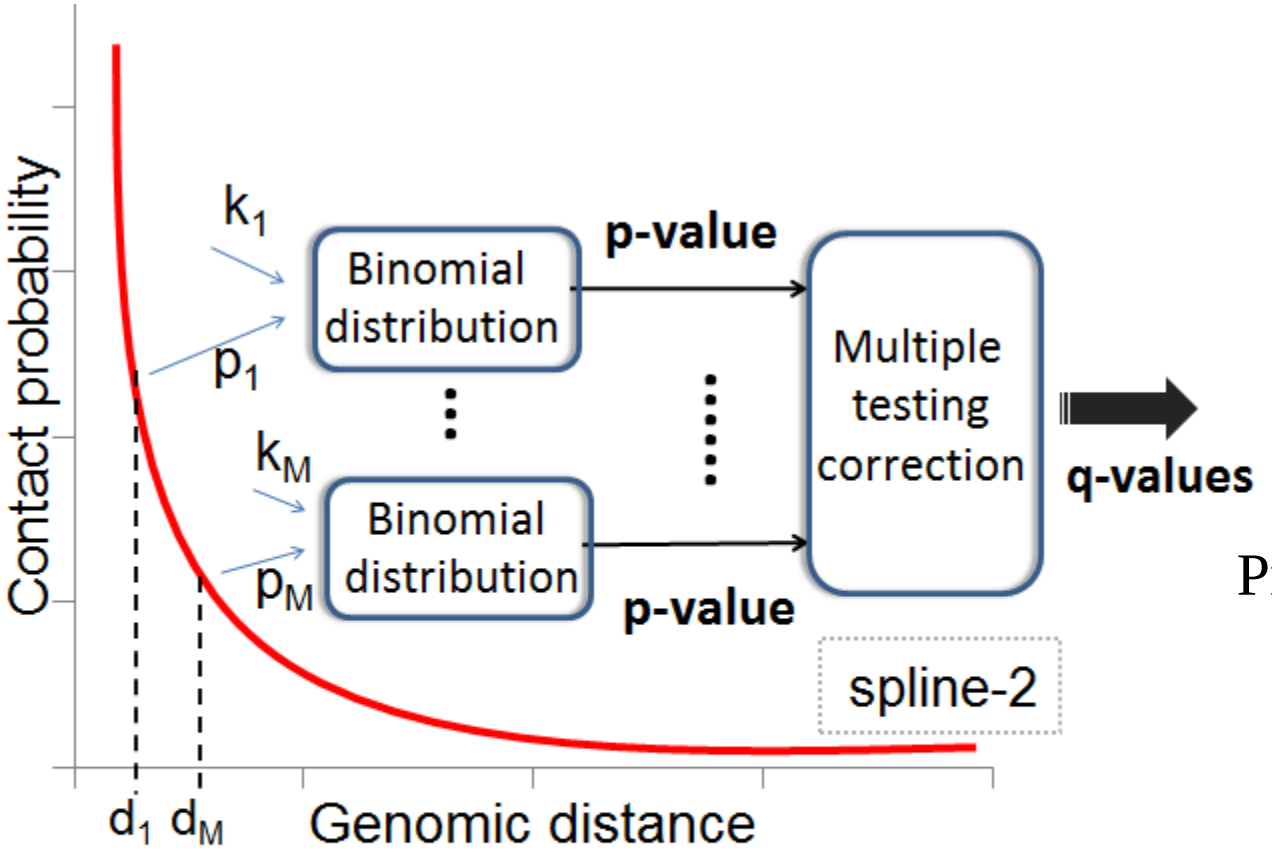
Assigning statistical confidence estimates to chromatin contact maps



Python: http://noble.gs.washington.edu/proj/fit-hi-c
R: https://bioconductor.org/packages/release/bioc/html/FitHiC.html
Ay, Bailey & Noble. Genome Research, 2014.


0
5
chr15-chr15 intra-chromosomal
Contact Probability (10^{-5})
1
0.1
0.5
1
2
5
10
Genomic Distance (Mb)
hIMR90 (50 RE fragments)
Contact Probability (10^{-5})
0.1
0.01
0.001
.05
0.1
0.2
0.5
1
2
5
Genomic Distance (Mb)
hIMR90 (10 RE fragments)

Statistical confidence estimation by Fit-Hi-C



Statistical confidence estimates for all mid-range locus pairs

We also incorporate biases learned from Hi-C data normalization

$$\overline{p} = p * b_1 * b_2$$

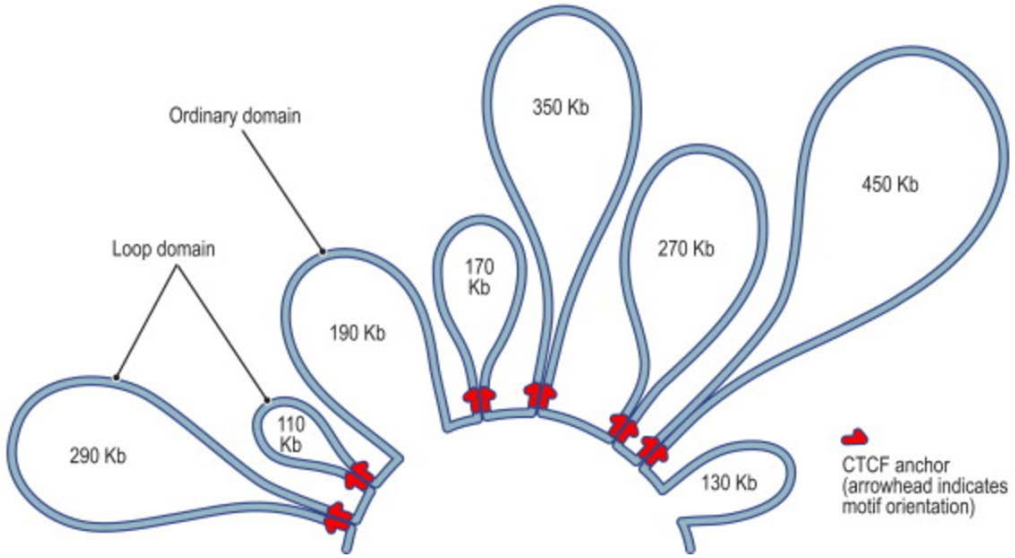
$$\Pr(X = k) = \binom{n}{k} \overline{p}^{k} (1 - \overline{p})^{n-k}$$

$$p - value = \Pr(X \geq k)$$

Ay, Bailey & Noble. Genome Research, 2014.

Erez's high resolution data from Rao et al 2014

- Six human cell lines: GM12878, HMEC, HUVEC, IMR90, K562, NHEK
- In situ Hi-C with a 4-bp cutter
- All cell lines have 5kb data, GM12878 also has 1kb data
- KR normalized contact maps gathered from [GSE63525]
- *_HiCCUPS_looplist.txt.gz files were used for comparison of loop calls



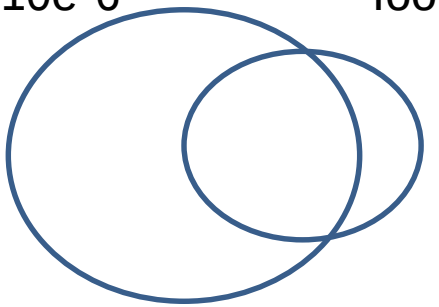
Fit-Hi-C vs HiCCUPS loop calls

- ✓ HICCUPS does NOT perform loop calls at 1kb resolution
- ✓ Out of 9448 HICCUPS loops for GM12878, 3132 are at 10kb and 6316 are at 5kb
- ✓ At 1kb Fit-Hi-C calls 142,264 FDR 0.01 loops (5kb, 500kb] (>1.52B possible pairs)

Analysis of 5kb data within (20kb, 2Mb] for six cell lines (~240M possible pairs)

Fit-Hi-C
p-val<10e-6

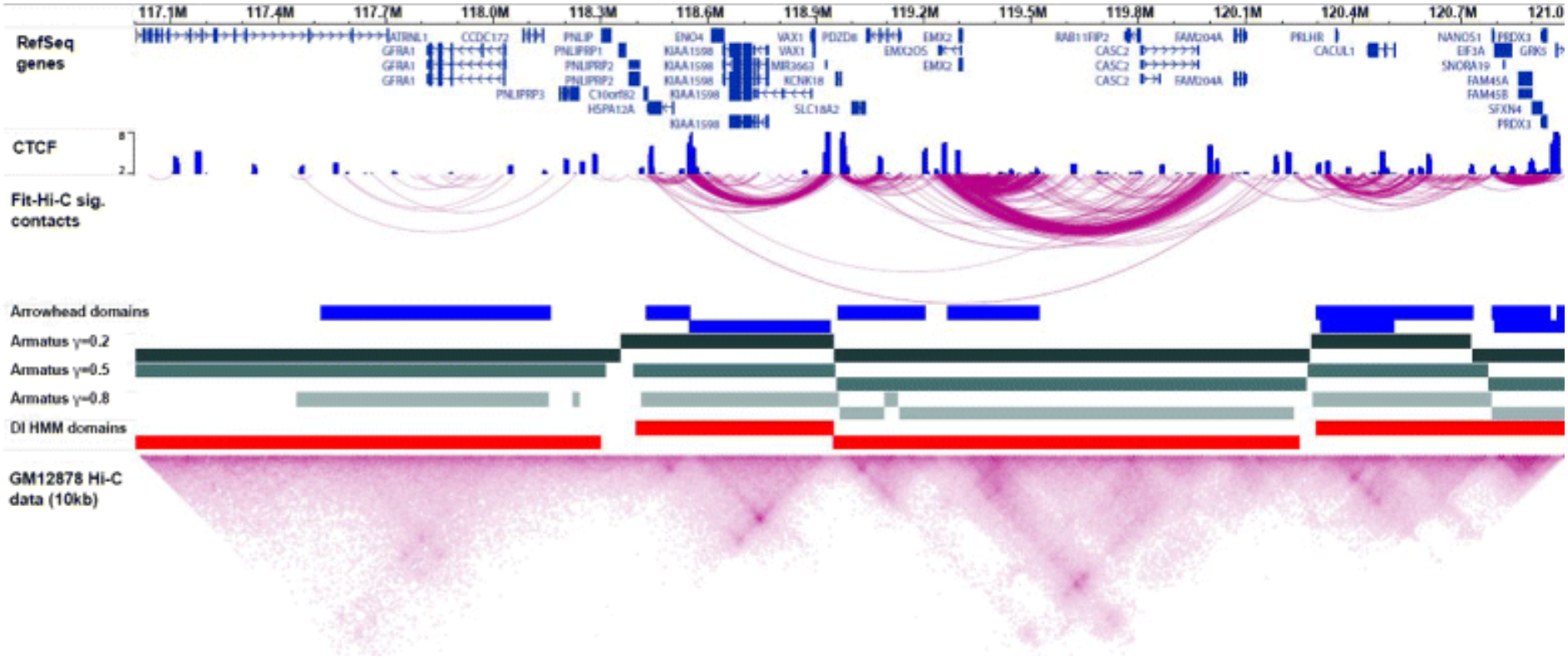
HICCUPS
loops



* p-val<10e-6 is more stringent than FDR<0.01 in general

CellLine / Loop calls	All HICCUPS	All Fit-Hi-C	Intersection - HICCUPS	Intersection - Fit-Hi-C	Percent covered
GM12878	9,270	1,521,610	8,674	13,700	93.6%
IMR90	7,992	300,707	7,128	12,016	89.2%
K562	5,938	152,779	4,038	8,490	68.0%
HMEC	5,152	27,808	3,516	4,568	68.2%
NHEK	4,913	14,054	2,197	3,576	44.7%
HUVEC	3,846	25,740	2,392	4,483	62.2%

Visualization of Fit-Hi-C contacts in WashU Epigenome Browser



Ay & Noble. *Genome Biology* 2015

Fit-Hi-C's statistical model works for a variety of conformation capture assays

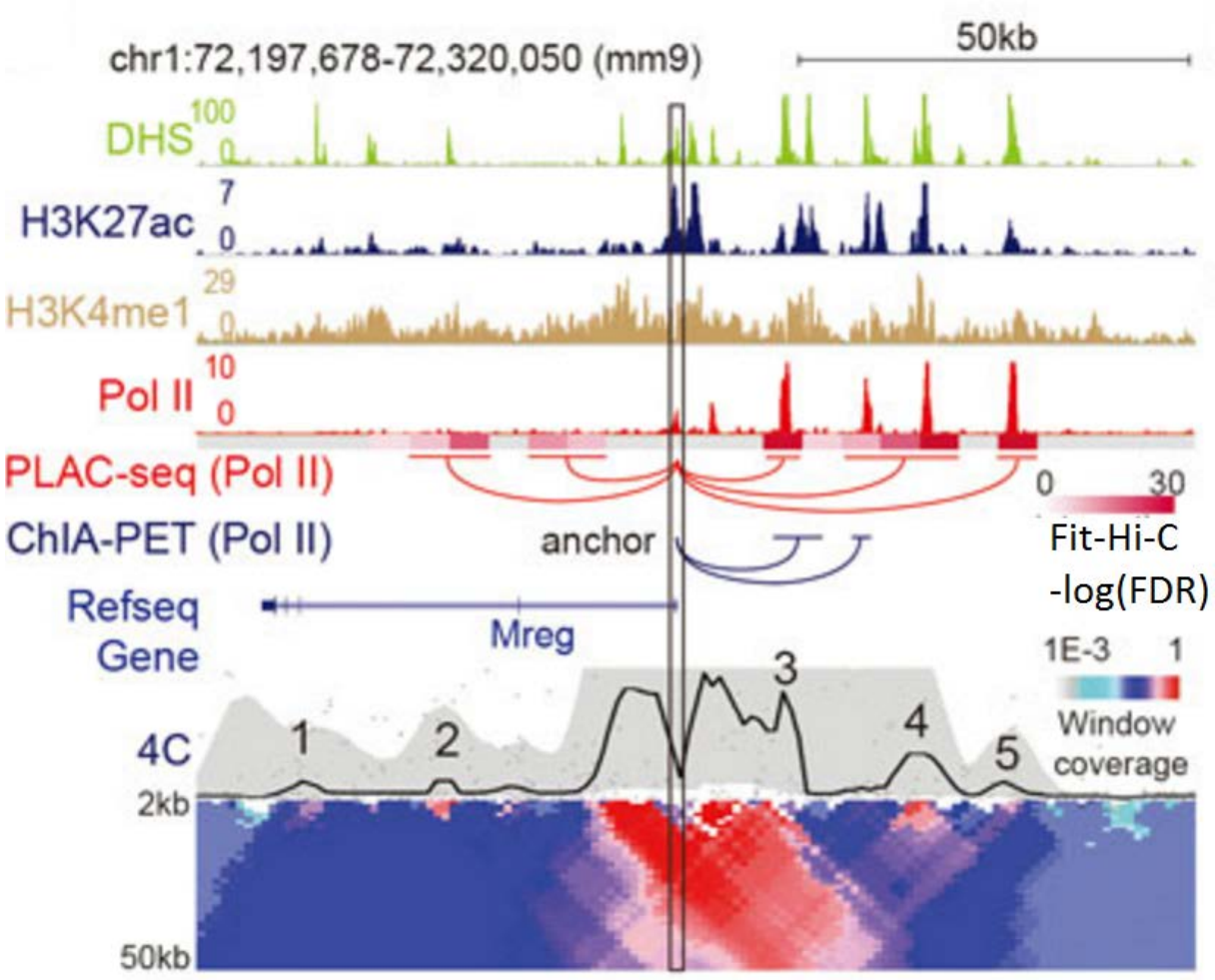


Fig 1C

PLAC-seq (Bing Ren Lab)
Fang et al. *Cell Research* 2016

Fit-Hi-C's statistical model works for a variety of conformation capture assays

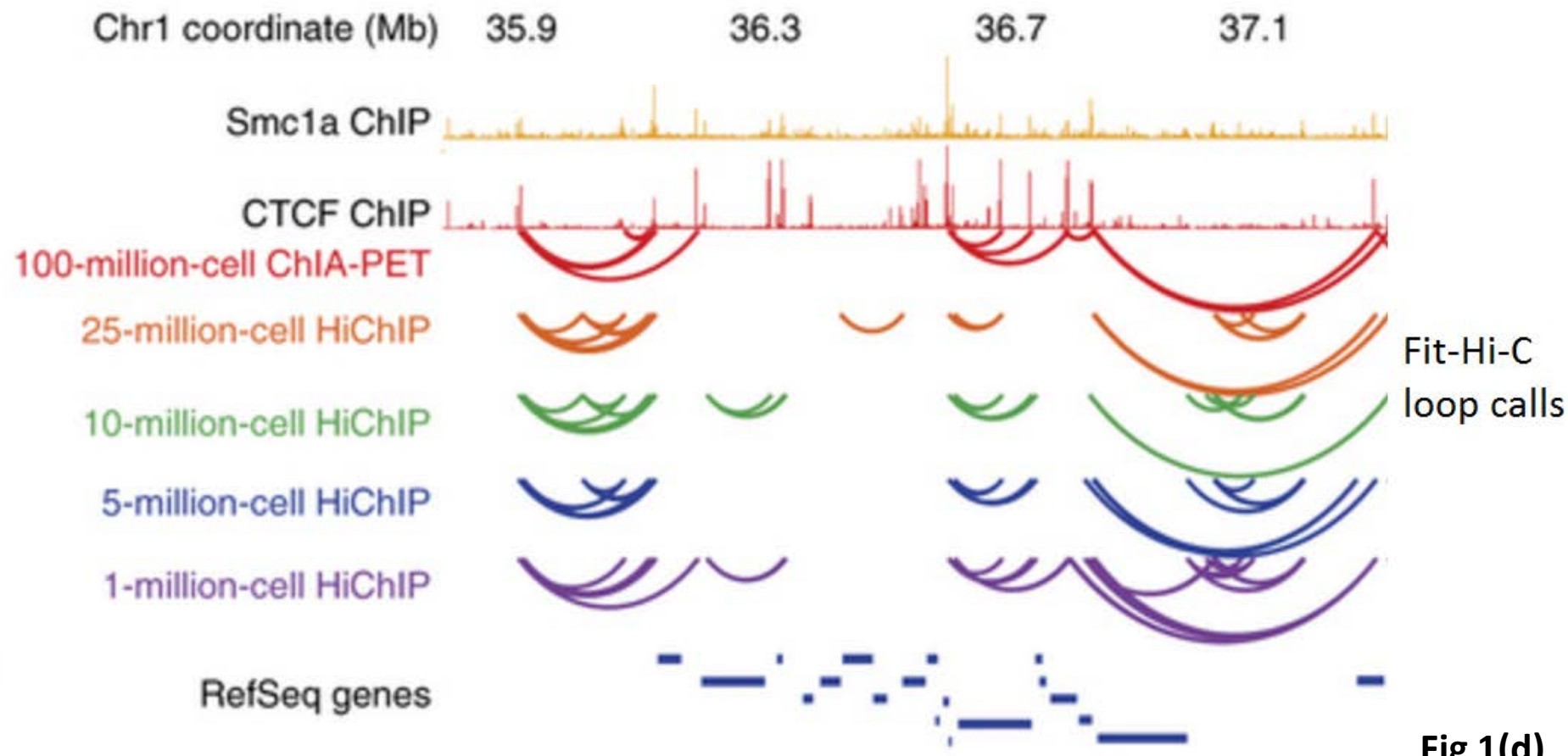


Fig 1(d)

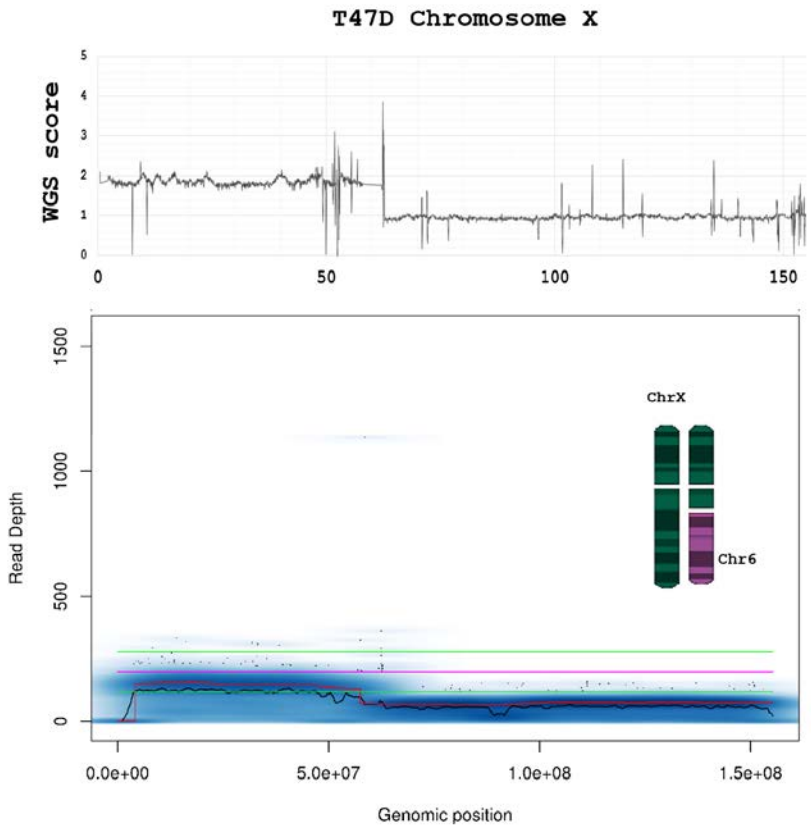
HiChIP (Greenleaf & Chang)
Mumbach et al. *Nature Methods* 2016

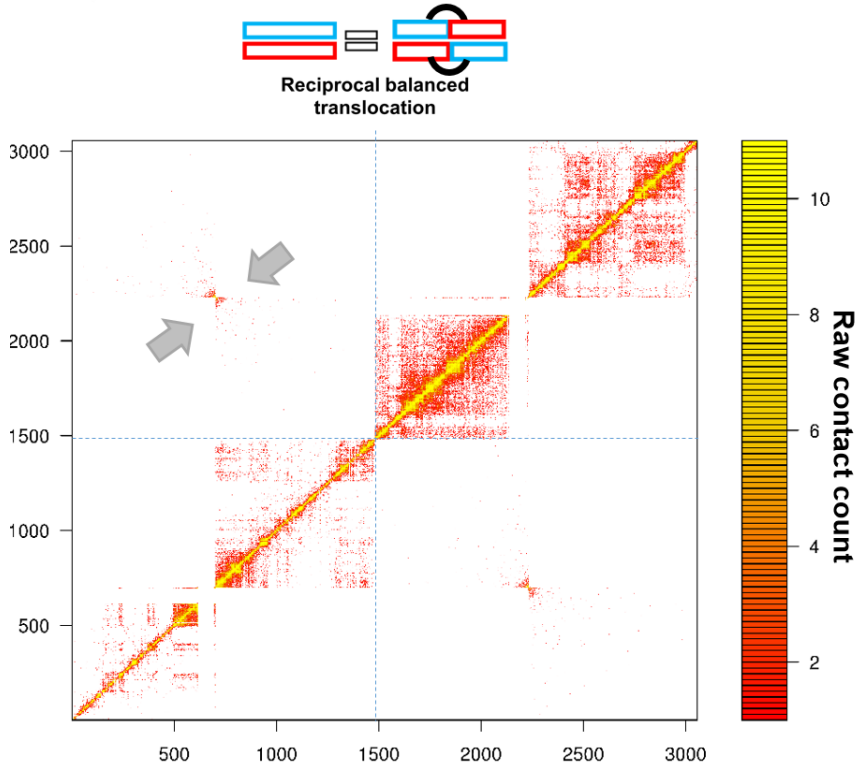
Fit-Hi-C result highlights

- ✓ Non-parametric spline fit flexible enough to work for any organism, any resolution and any sequencing depth
- ✓ Fast, robust and flexible statistical method for identifying loops from any genome-wide conformation capture data
- ✓ Fit-Hi-C can detect cell-type specific and validated contacts (3C, ChIA-PET)
- ✓ Significant interactions correlate with other functional genomics data
- ✓ Fit-Hi-C's statistical power depends on:
  - assay choice (traditional Hi-C vs ChIP-based methods),
  - sequencing depth,
  - resolution of contact maps,
  - genomic distance range of interest (multiple testing correction),
  - but not much on the amount of starting material.

HiCnv, HiCtrans & AveSim

Identification of copy number variations and translocations in cancer cells from Hi-C data



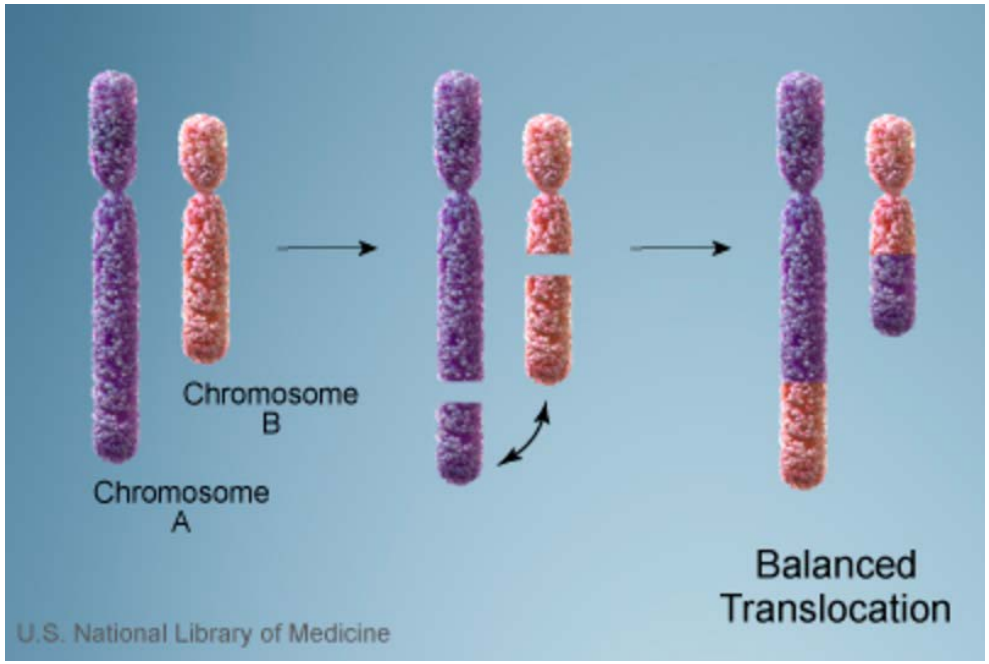


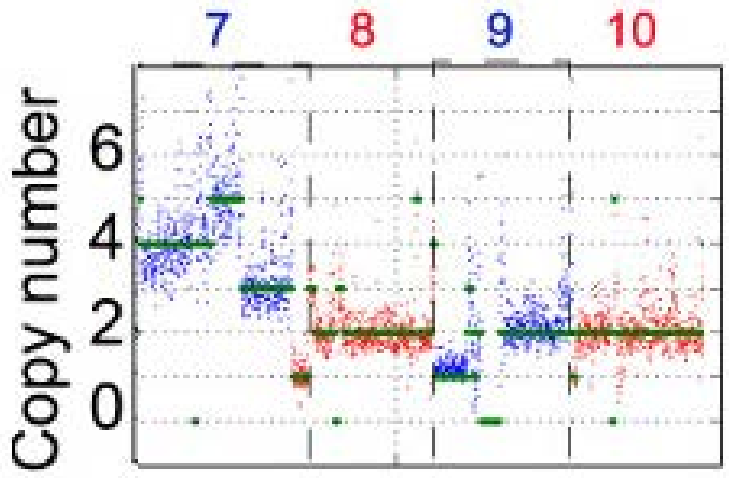
Abhijit Chakraborty

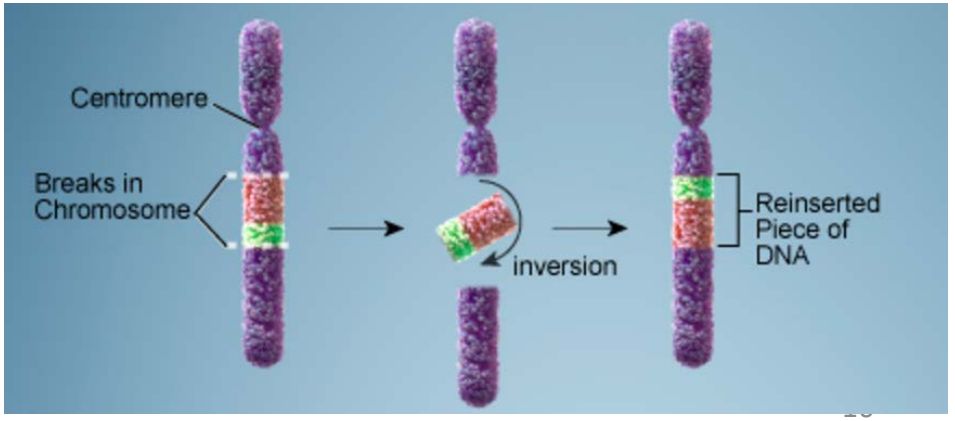
https://github.com/ay-lab

Chakraborty & Ay. Under review.

Chromosomal rearrangements are common in cancer







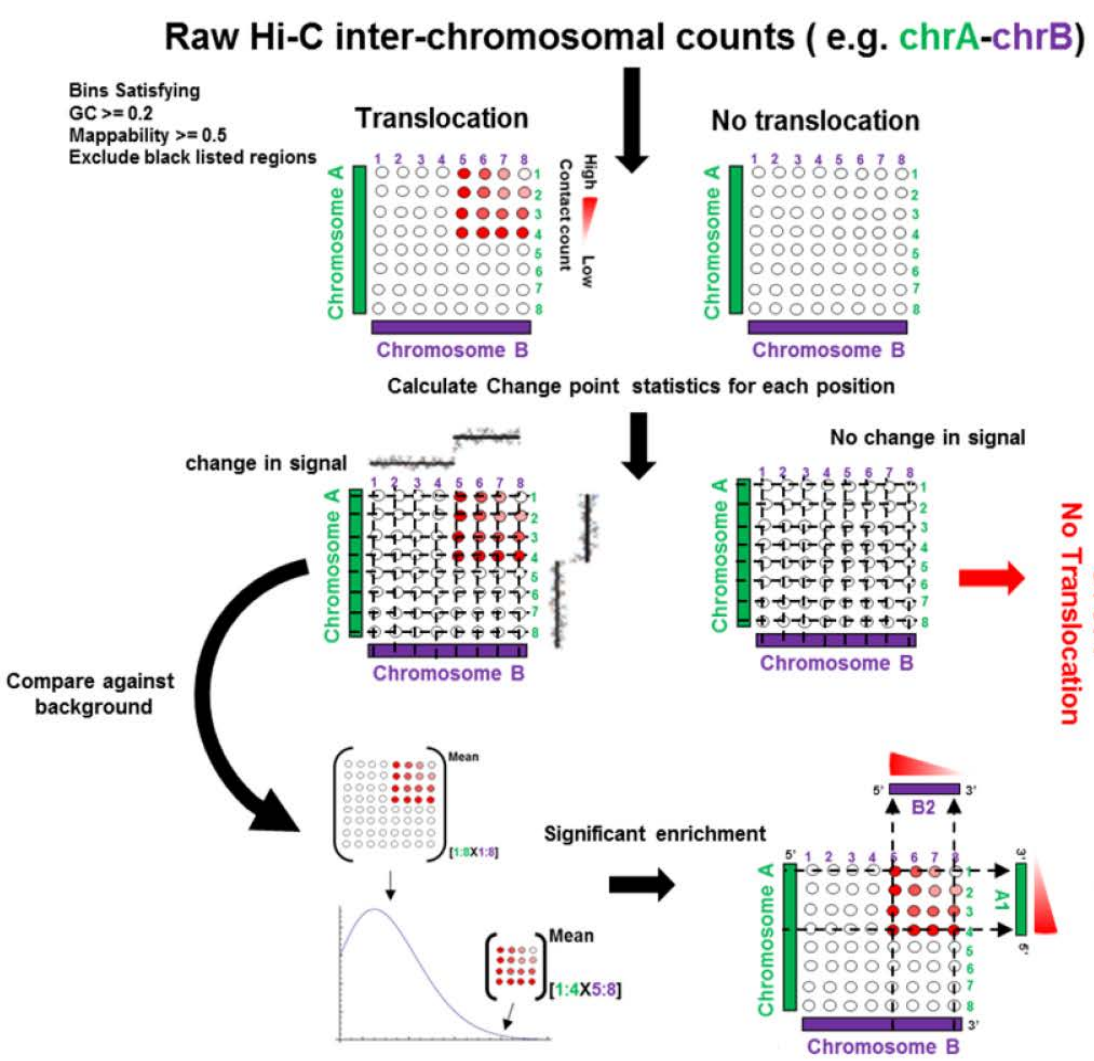
ENCODE-released Hi-C data from Job's lab

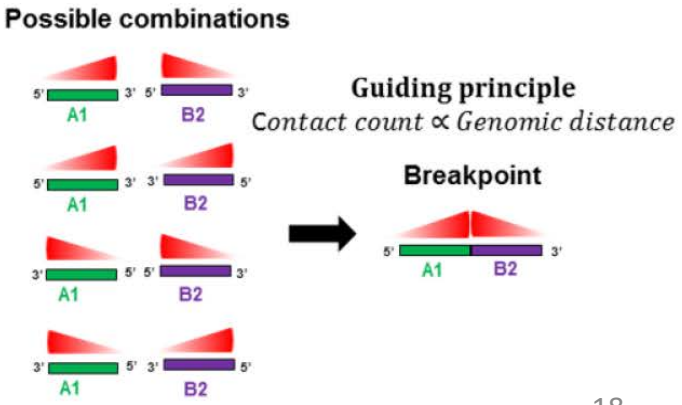
Cell line	HiCPro data summary		
	Raw pairs	Valid pairs	Percentage
A549	251,891,733	135,674,989	53.86%
CAKI2	323,731,060	168,096,814	51.92%
G401	340,927,844	174,130,474	51.08%
LNCaP	306,489,193	92,691,677	30.24%
NCIH460	313,205,689	162,906,364	52.01%
PANC1	288,978,052	160,552,758	55.56%
RPMI7951	335,883,359	189,765,014	56.50%
~~SJCRH30~~	~~152,235,750~~	~~6,432,592~~	~~4.23%~~
SKMEL5	303,482,692	133,713,968	44.06%
~~SKNDZ~~	~~291,853,821~~	~~59,307,125~~	~~20.32%~~
SKNMC	313,811,254	149,394,332	47.61%
T47D	247,702,528	133,681,534	53.97%

- HindIII digestion
- 150-350M 50bp paired-end
- HiCPro is used for mapping
- 10/12 analyzed further
- Dave Gilbert generated RT data for 8 cell lines
- Feng Yue generated WGS for 6 and Irys for 8
- http://biorxiv.org/content/early/2017/03/28/119651

https://www.encodeproject.org/search/?type=Experiment&assay_title=Hi-C&status=released

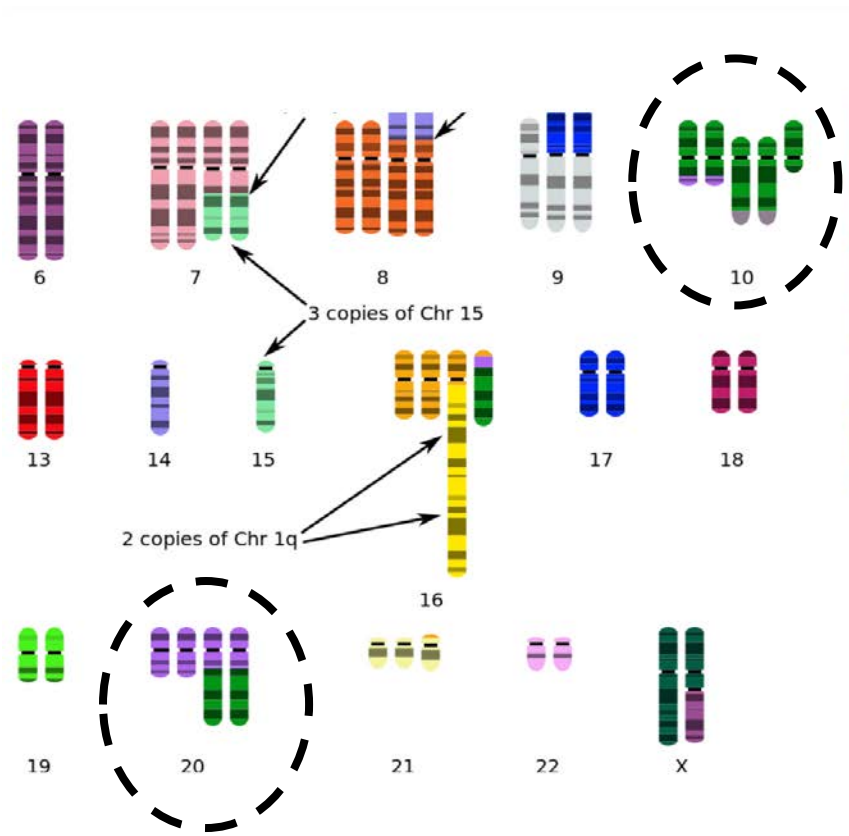
Detecting chromosomal translocations from Hi-C data (HiCtrans)

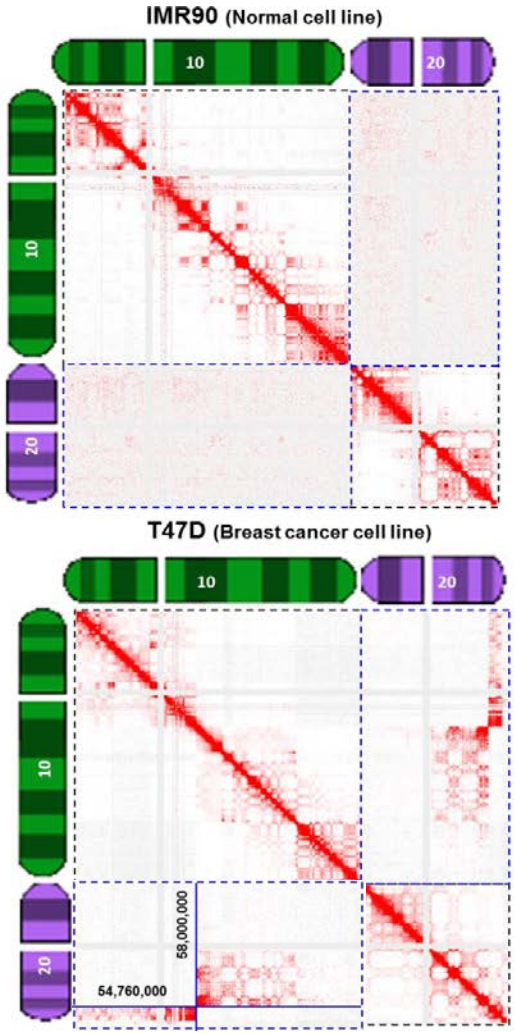




- Perform binary segmentation on each row and each column
- Find boxes of contact enrichment
- Test box mean vs overall mean
- Correct for multiple testing
- Find maximum raw count to determine translocation orientation

An example translocation identified by HiCtrans





An Integrative Framework For Detecting Structural Variations In Cancer Genomes

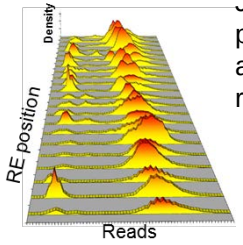
Jesse Dixon, Jie Xu, Vishnu Dileep, Ye Zhan, Fan Song, Victoria T. Le, Galip Gurkan Yardimci, Abhijit Chakraborty, Darrin V. Bann, Yanli Wang, Royden Clark, Lijun Zhang, Hongbo Yang, Tingting Liu, Sriranga Iyyanki, Lin An, Christopher Pool, Takayo Sasaki, Juan Carlos Rivera Mulia, Hakan Ozadam, Bryan R. Lajoie, Rajinder Kaul, Michael Buckley, Kristen Lee, Morgan Diegel, Dubravka Pezic, Christina Ernst, Suzana Hadjur, Duncan T. Odom, John A. Stamatoyannopoulos, James R. Broach, Ross Hardison, Ferhat Ay, William Stafford Noble, Job Dekker, David M Gilbert, Feng Yue

doi: https://doi.org/10.1101/119651

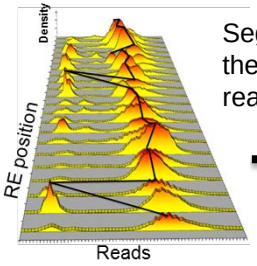
Detecting CNVs from Hi-C data (HiCnv)

Assign each read (including singletons and non-valid pairs) to the nearest RE

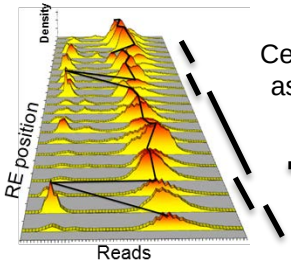
- Smooth the data by either using kernel density estimation or moving window average.
- Combine the smoothed counts and the kernel weights to obtain the approximation to the density estimate.



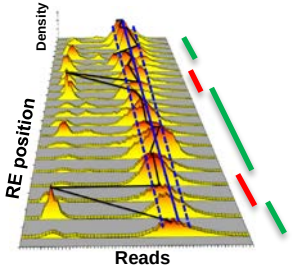
Join the highest points in the density as the representative read count.



Segmentation on the representative reads.



Cell line mean as reference



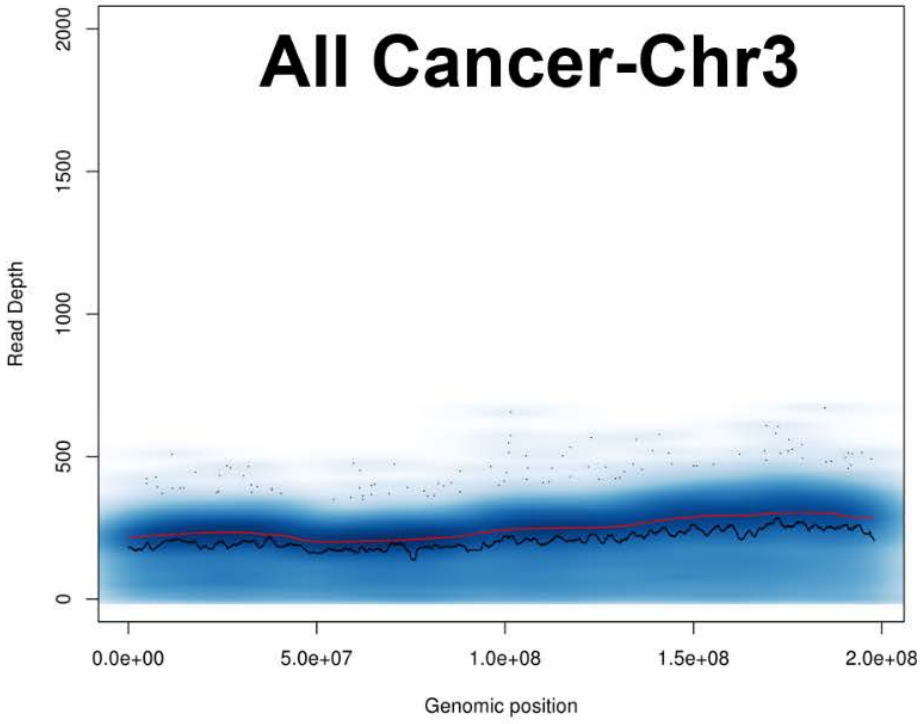
Normal (**Z-score ≥ -1 and ≤ 1**)
Amplified (**Z-score ≥ 1**)
Deleted (**Z-score ≤ -1**)

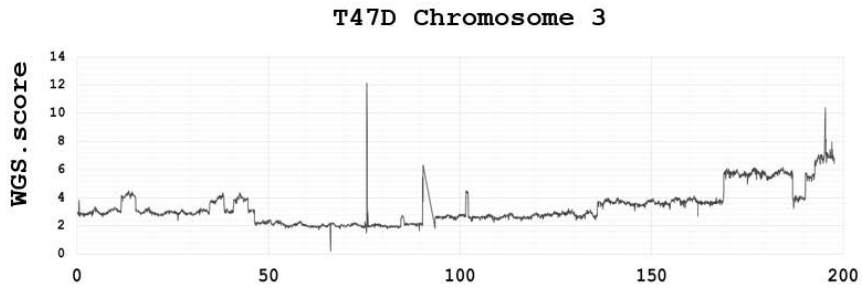
Discard the segments that are Z-score ≤ -1 in all combined cancer genome segmentation analysis.

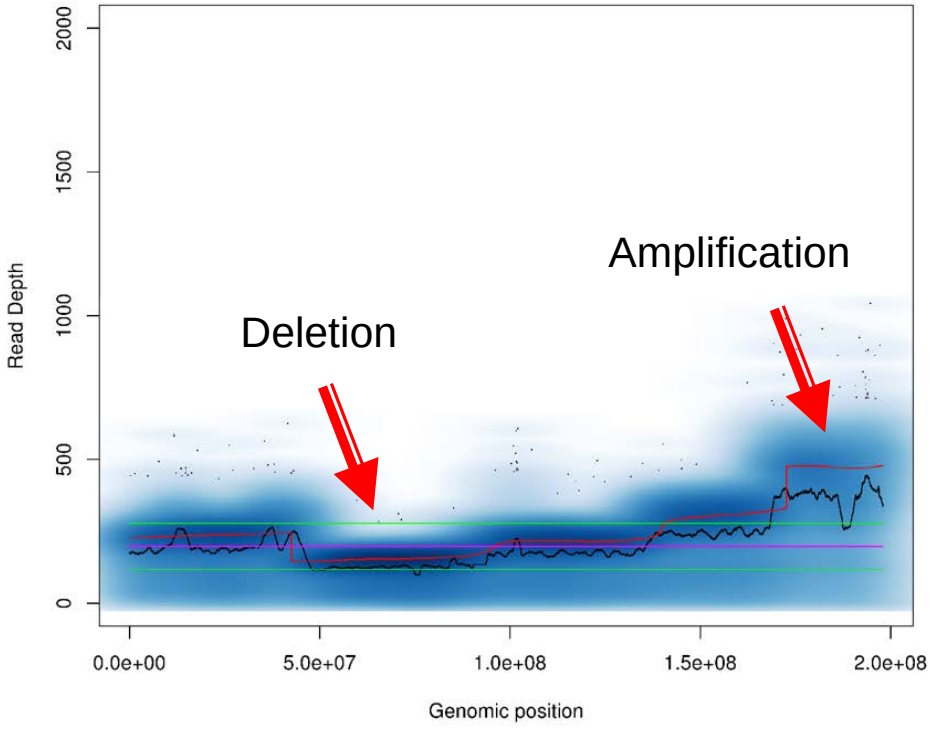
(If all the cancer cells have lower reads than it is possibly due to un-mappability or low GC content but not deletion)

Report Normal, Amplified and Deleted regions

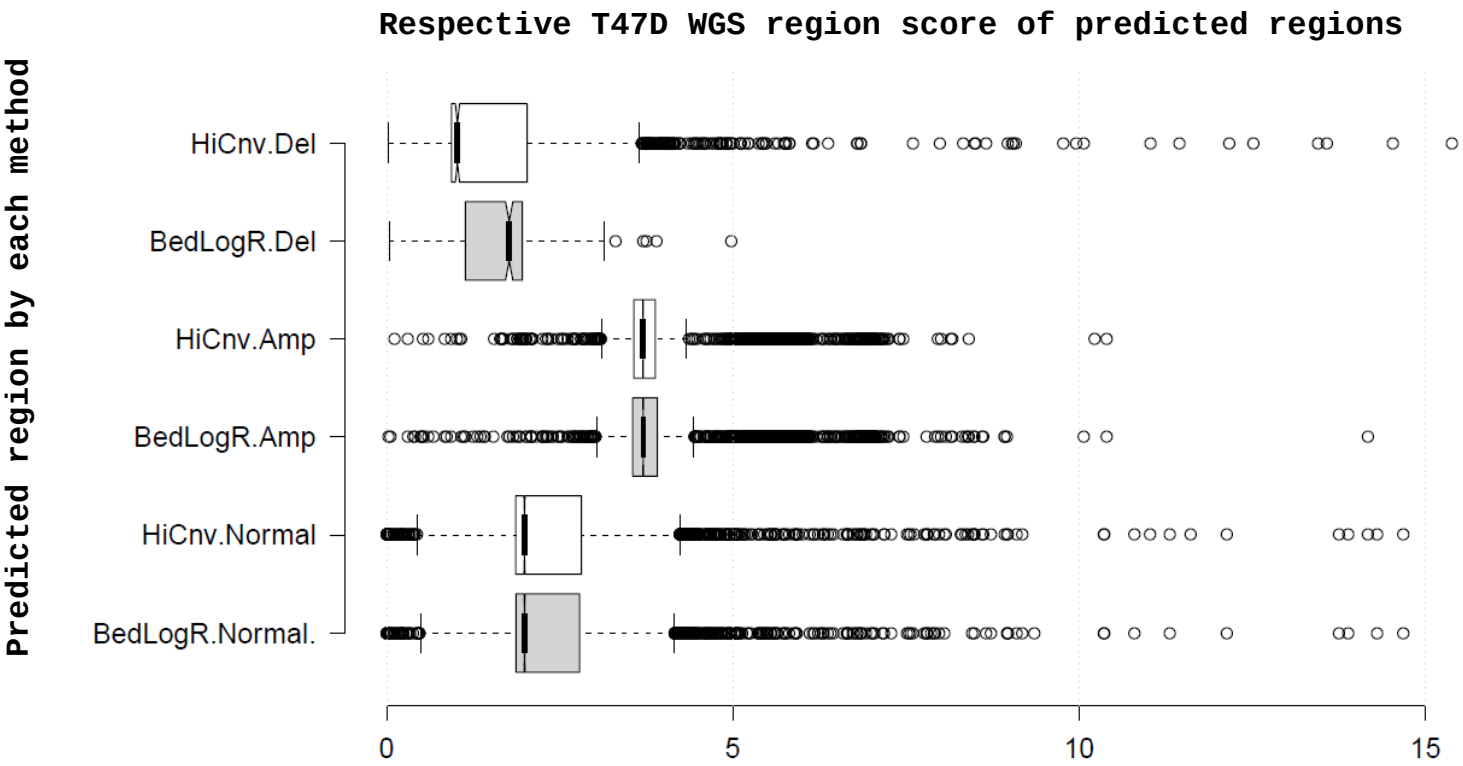
An example amplification and deletion identified by HiCnv







Comparison of HiCnv, BeadChip and WGS calls



Each point denotes a segment from WGS (~30x – Dixon et al, under review)

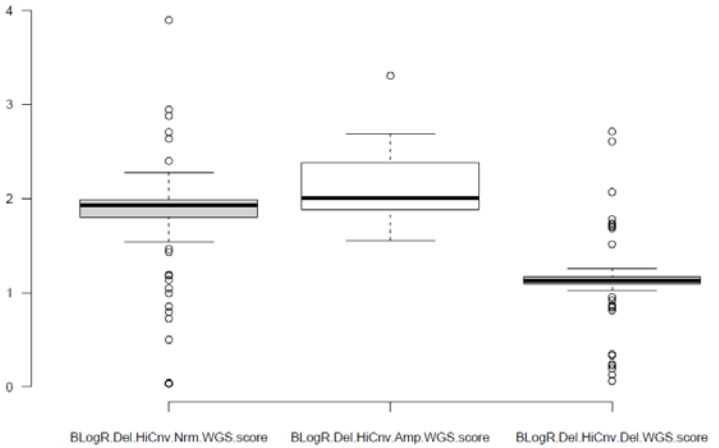
BedLogR values from HAIBGenotype (CNV and SNP) by Illumina 1M Duo and circular binary segmentation from ENCODE/Hudson Alpha)

	BedLogR.Normal	HiCnv.Normal	BedLogR.Amp	HiCnv.Amp	BedLogR.Del	HiCnv.Del
Median	1.99	1.99	3.70	3.70	1.77	1.02
No. of data points	34258	37139	11219	13310	614	3333

T47D deletions

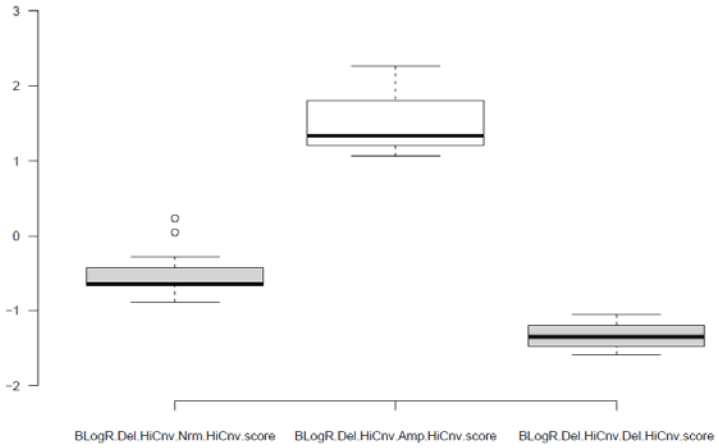
31 unique deletions in HAIB that are >= 10Kb in size.
8 reported as deletions by HiCnv.

	BLogR.Del.HiCnv.Nrm.WGS.score	BLogR.Del.HiCnv.Amp.WGS.score	BLogR.Del.HiCnv.Del.WGS.score
3rd quartile	1.99	2.39	1.17
Median	1.93	2.01	1.13
1st quartile	1.80	1.89	1.10



BedLogR Deleted But HiCnv Normal | BedLogR Deleted But HiCnv Ampllified | BedLogR Deleted And HiCnv Deleted

	BLogR.Del.HiCnv.Nrm.HiCnv.score	BLogR.Del.HiCnv.Amp.HiCnv.score	BLogR.Del.HiCnv.Del.HiCnv.score
3rd quartile	-0.42	1.80	-1.20
Median	-0.64	1.33	-1.35
1st quartile	-0.66	1.20	-1.47

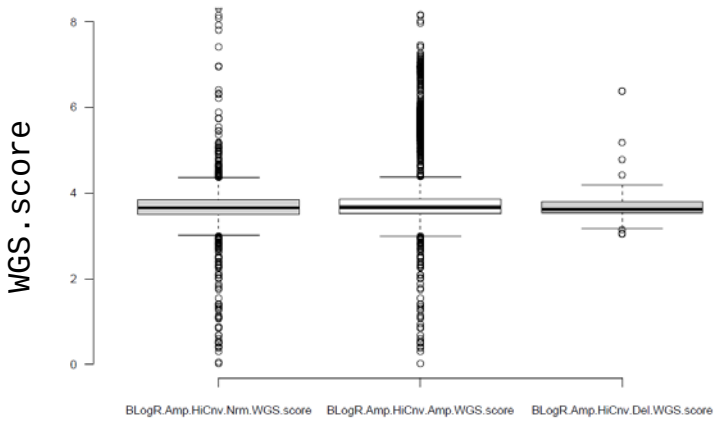


BedLogR Deleted But HiCnv Normal | BedLogR Deleted But HiCnv Ampllified | BedLogR Deleted And HiCnv Deleted

T47D amplifications

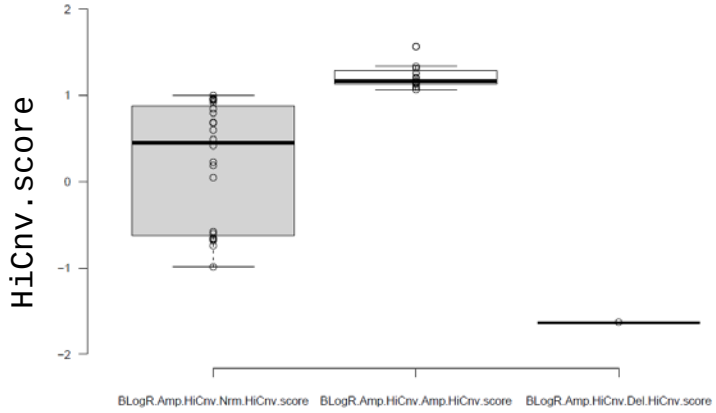
53 unique amplifications in HAIB that are >= 10Kb in size.
34 reported as amplifications by HiCnv.

	BLogR.Amp.HiCnv.Nrm.WGS.score	BLogR.Amp.HiCnv.Amp.WGS.score	BLogR.Amp.HiCnv.Del.WGS.score
3rd quartile	3.85	3.86	3.80
Median	3.66	3.67	3.62
1st quartile	3.51	3.52	3.54



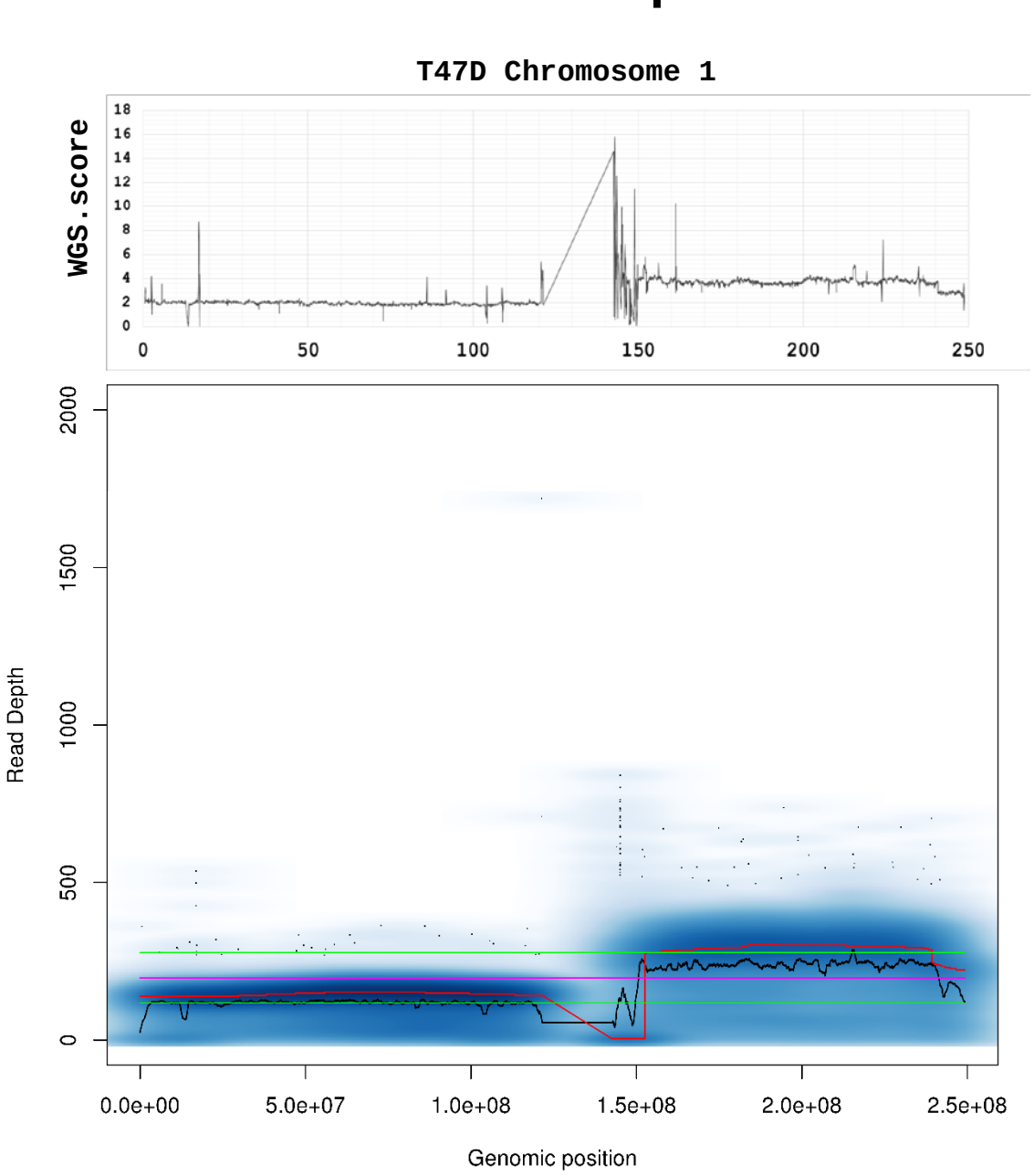
BedLogR Amplified But HiCnv Normal | BedLogR Amplified And HiCnv Ampllified | BedLogR Amplified But HiCnv Deleted

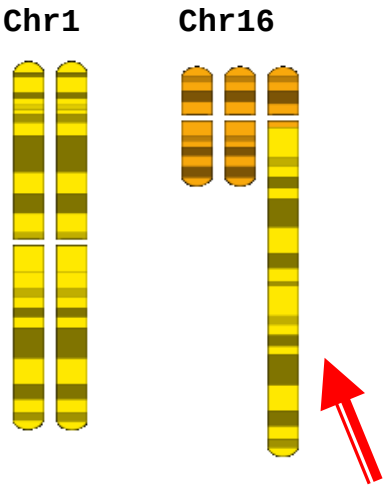
	BLogR.Amp.HiCnv.Nrm.HiCnv.score	BLogR.Amp.HiCnv.Amp.HiCnv.score	BLogR.Amp.HiCnv.Del.HiCnv.score
3rd quartile	0.88	1.28	-1.63
Median	0.45	1.17	-1.63
1st quartile	-0.63	1.13	-1.63



BedLogR Amplified But HiCnv Normal | BedLogR Amplified And HiCnv Ampllified | BedLogR Amplified And HiCnv Deleted

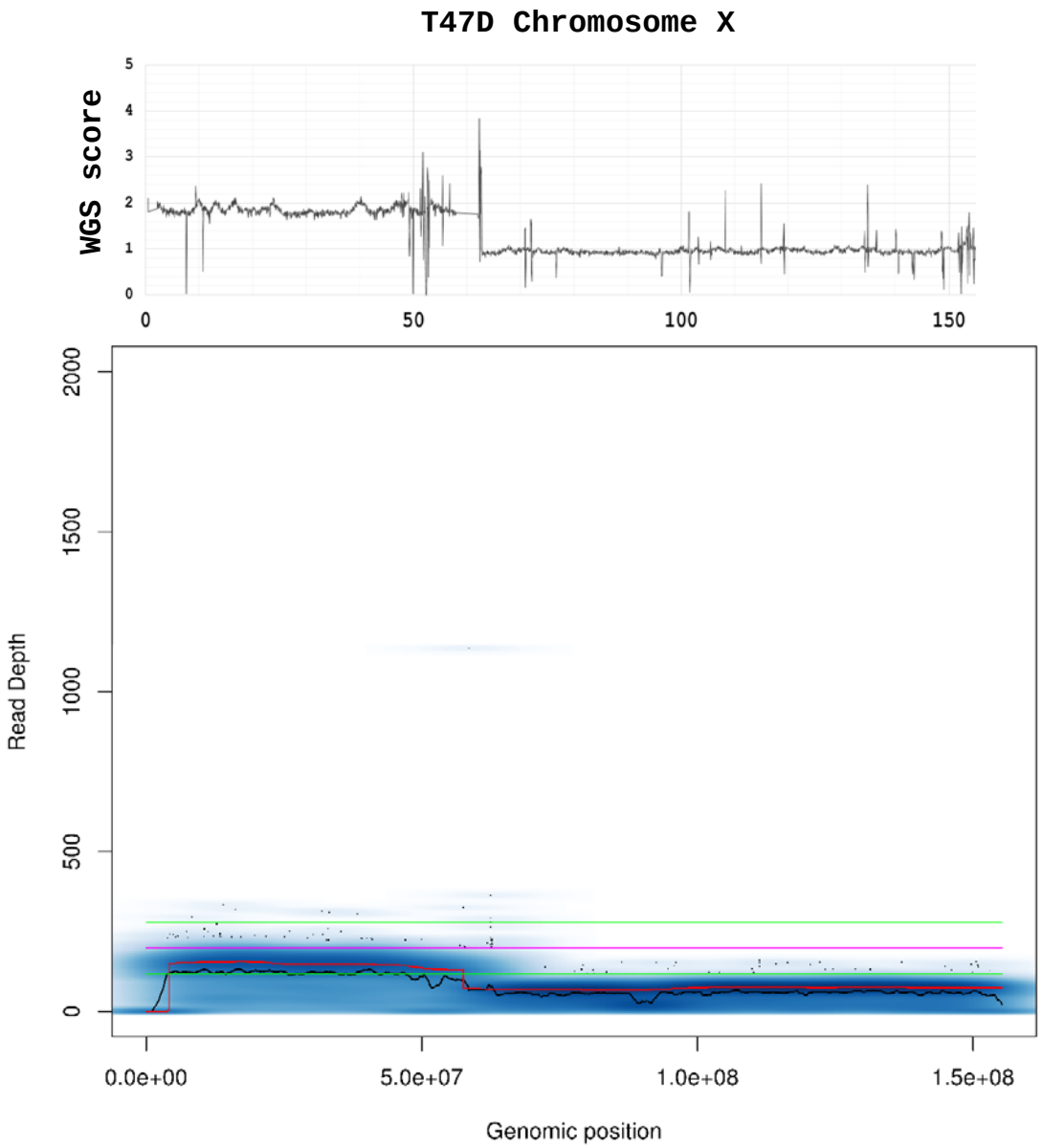
Example CNVs

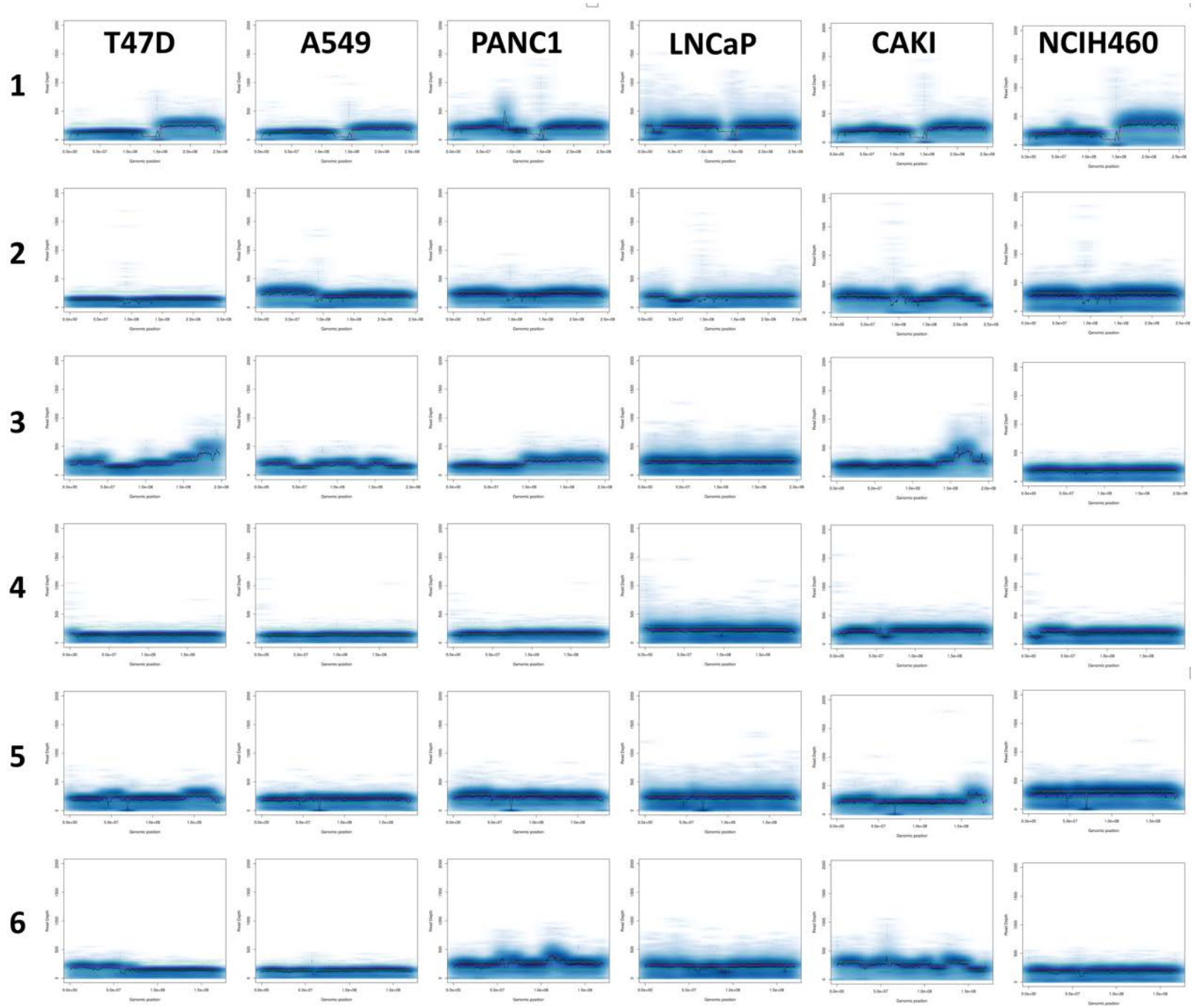




2 copies of Chr1 q arm

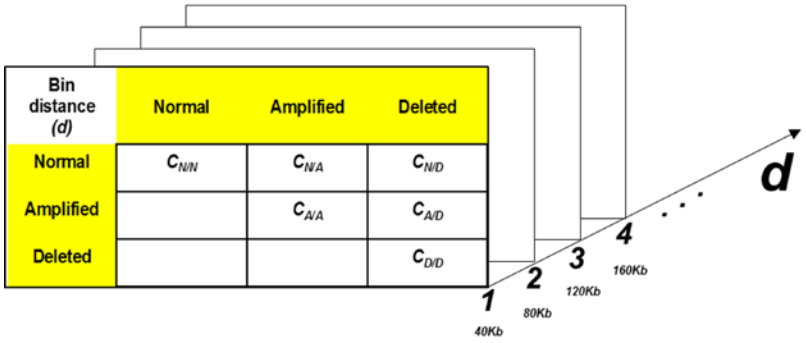
Example CNVs

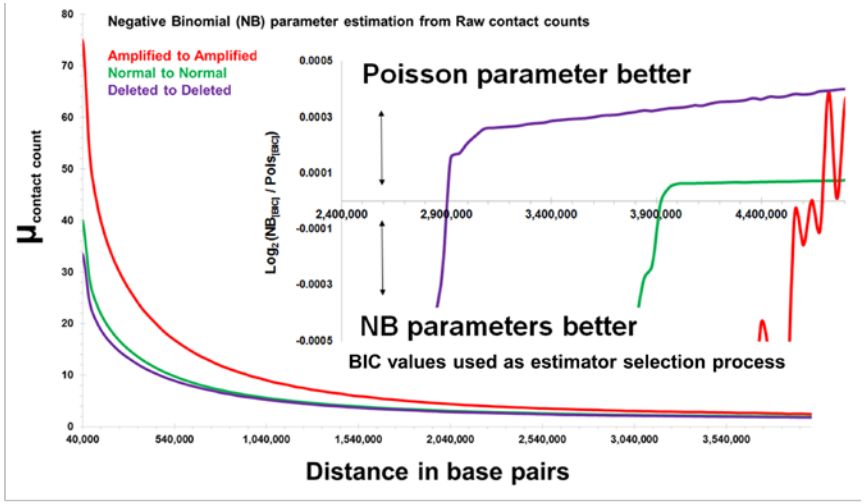



T47D
A549
PANC1
LNCaP
CAKI
NCIH460
1
2
3
4
5
6

Simulating Hi-C matrices with CNVs

- **Extracted the contact counts among all bin pairs with the same CNV label pair.**
- **Further categorize counts wrt genomic distance for each label pair.**





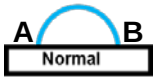
- **Fit distributions to predict expected counts given a bin distance and CNV label pair.**
- **Fitted the values up to 160 Mb of genomic distance (4000 bin) to both Poisson and Negative-Binomial distribution.**
- **For each bin distance, we selected either the negative binomial or the Poisson distribution as the best fit using Bayesian information criteria (BIC).**

Two alternative ways to simulate matrices

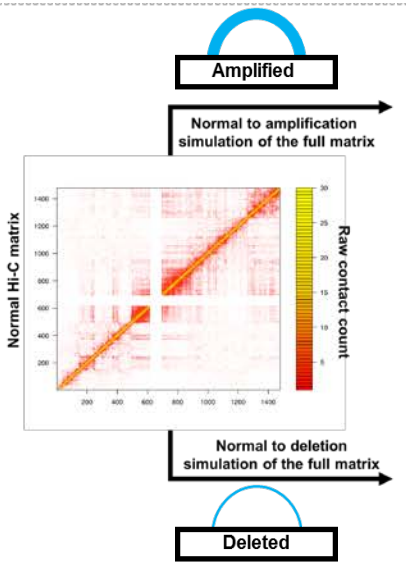
• Random count simulation:

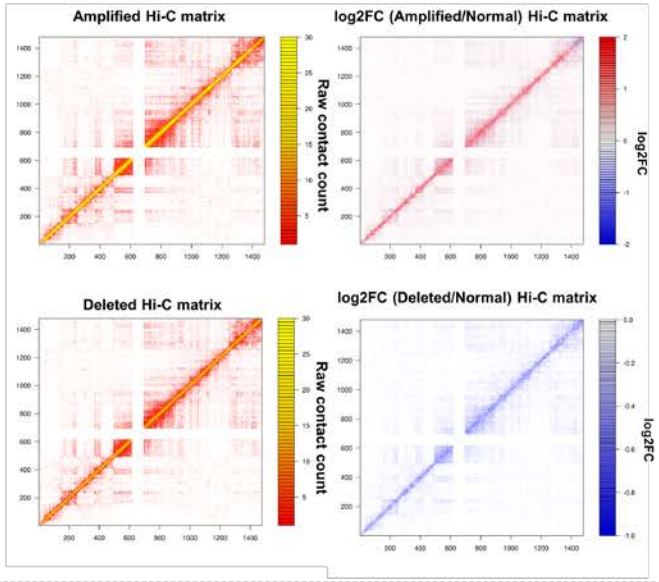
Bin1	Bin2	Bin distance	Simulated Count
A1	B1	d1	NB ($\mu_{N/N}$,θ)$_{d1}$
A2	B2	d2	NB ($\mu_{N/N}$,θ)$_{d2}$
A3	B3	d3	NB ($\mu_{N/N}$,θ)$_{d3}$
A4	B4	d4	NB ($\mu_{A/A}$,θ)$_{d4}$
A5	B5	d5	NB ($\mu_{D/D}$,θ)$_{d5}$

Bin connection same as that of original matrix



Assign the count randomly based on distance and CNV

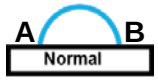




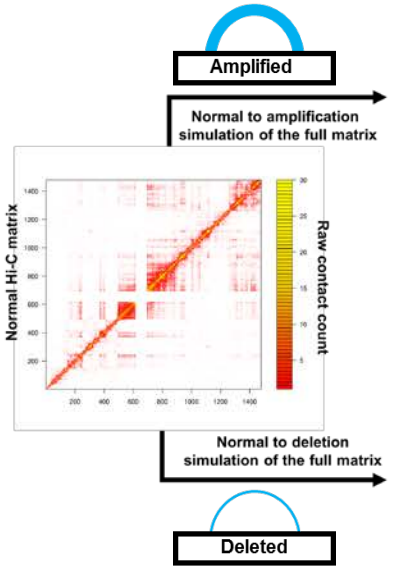
• Scaled observed count simulation:

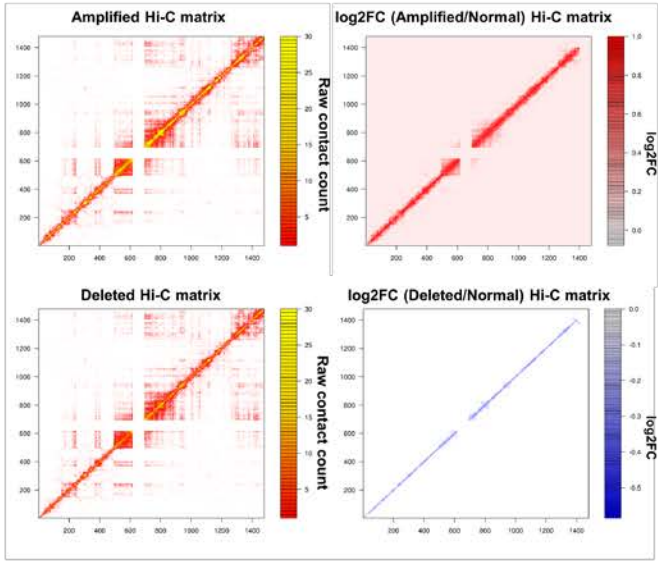
Bin1	Bin2	Bin distance	Original Count	Simulated Count
A1	B1	d1	C1	C1 X ($\mu_{A/A}$/$\mu_{N/N}$)$_{d1}$
A2	B2	d2	C2	C2 X ($\mu_{A/A}$/$\mu_{N/N}$)$_{d2}$
A3	B3	d3	C3	C3 X ($\mu_{D/D}$/$\mu_{N/N}$)$_{d3}$
A4	B4	d4	C4	C4 X ($\mu_{A/A}$/$\mu_{N/N}$)$_{d4}$
A5	B5	d5	C5	C5 X ($\mu_{D/D}$/$\mu_{N/N}$)$_{d5}$

Bin connection same as that of original matrix

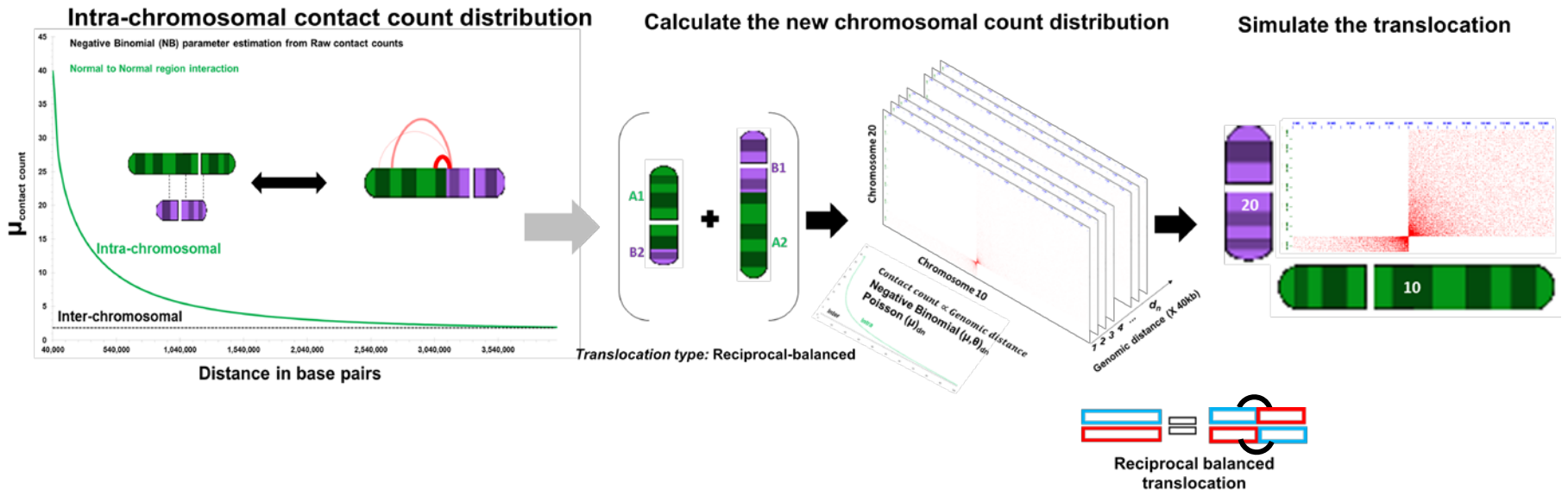


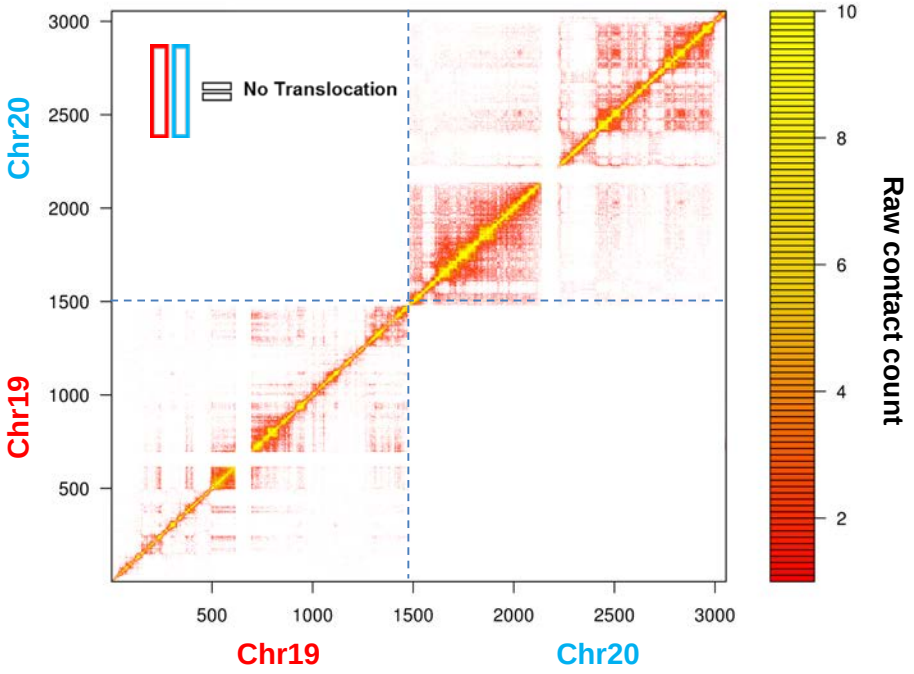
Multiply the count with the CNV ratio at that distance

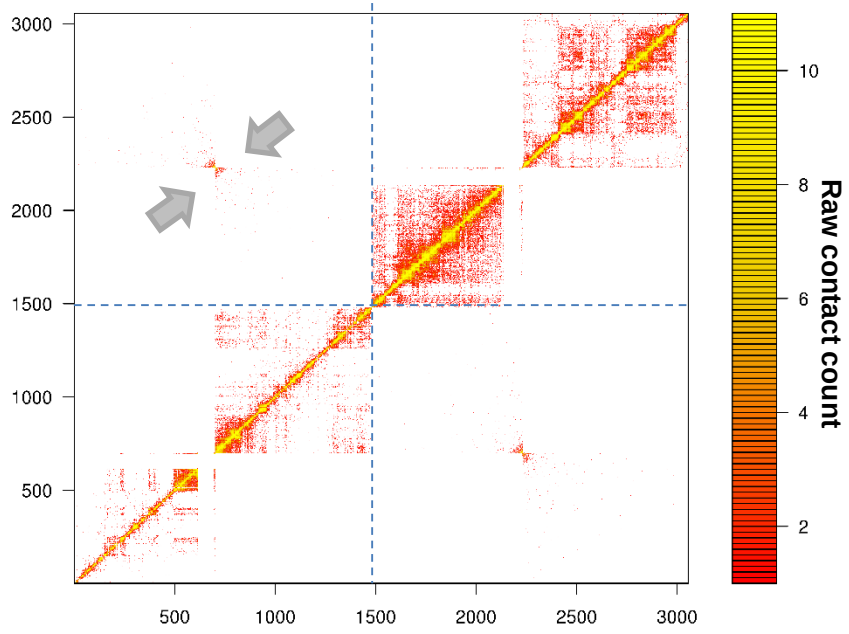




Simulating Hi-C matrices with translocations

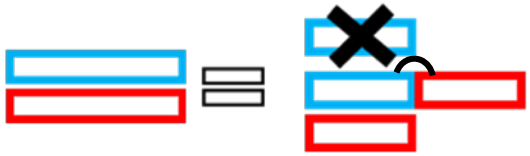




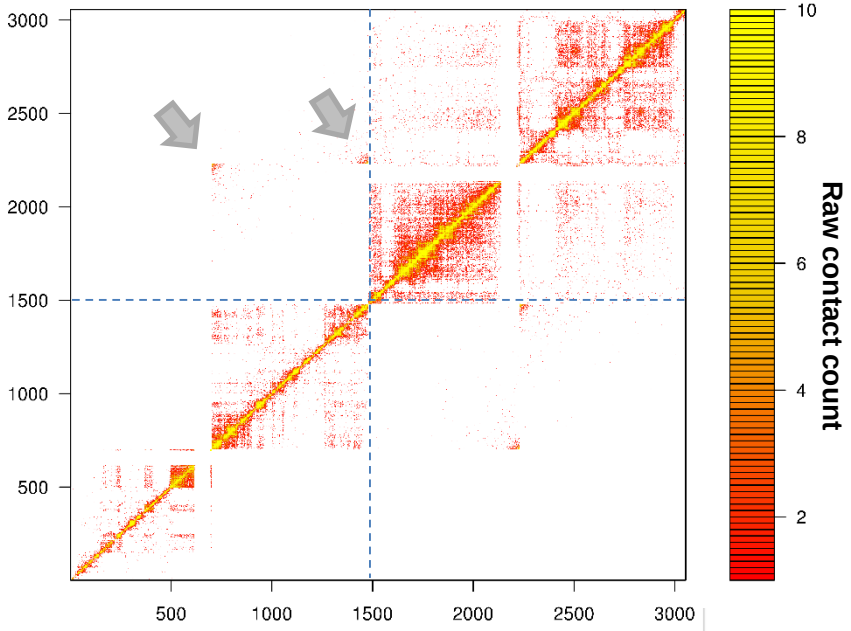


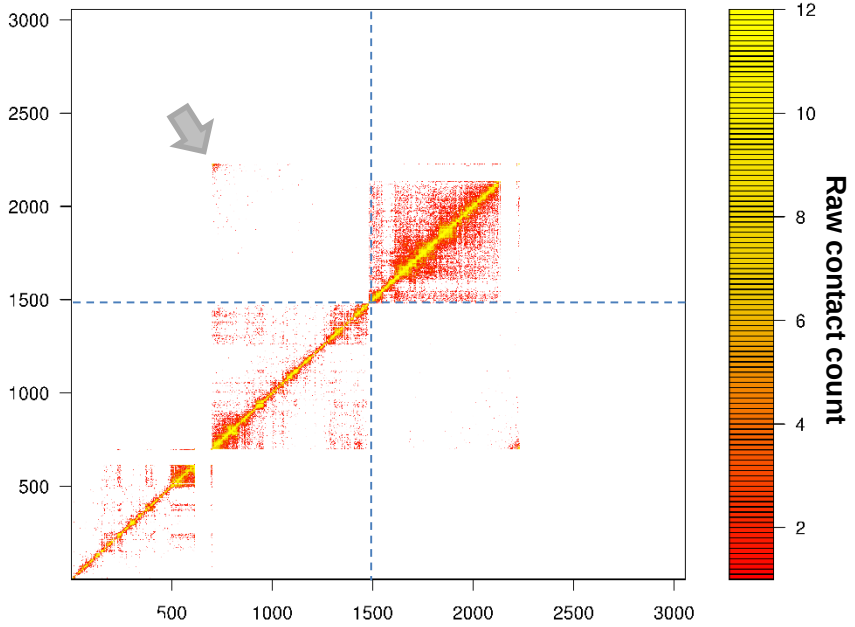
Simulating different types of translocations

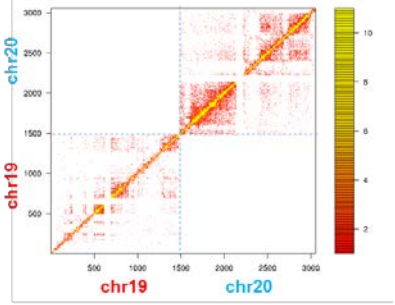
Non-reciprocal balanced translocation



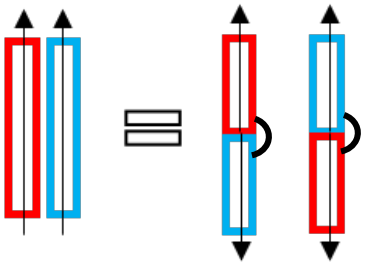
Unbalanced translocation

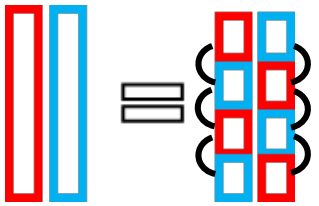




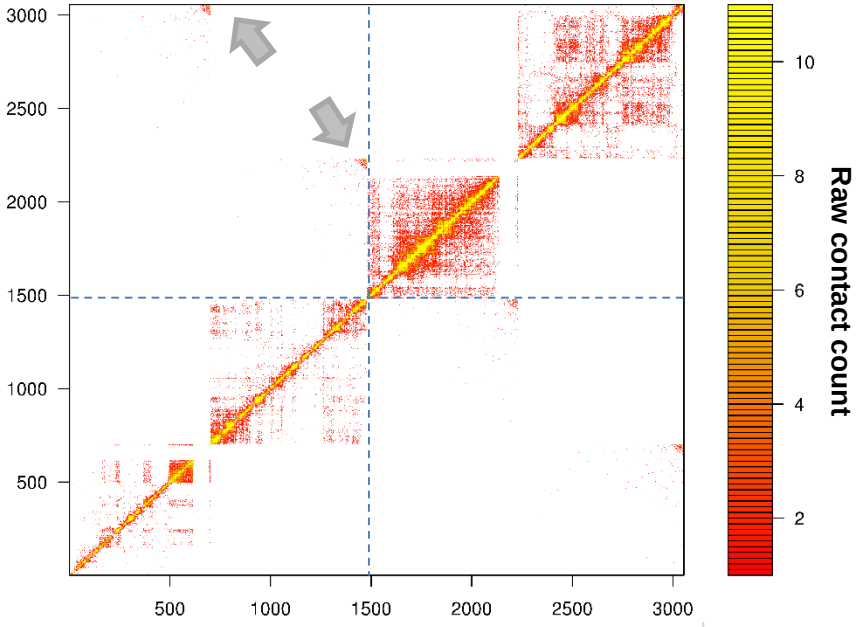


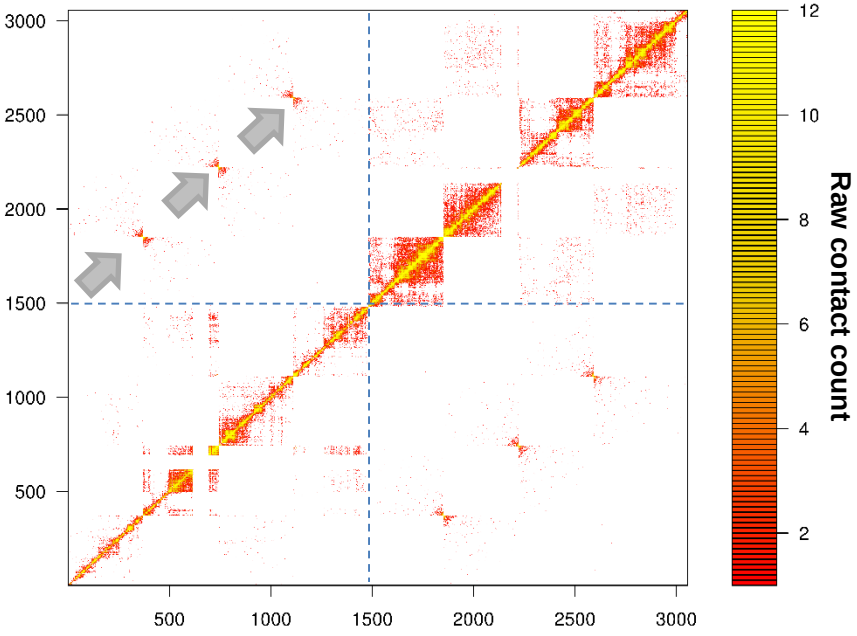
Simulating different types of translocations

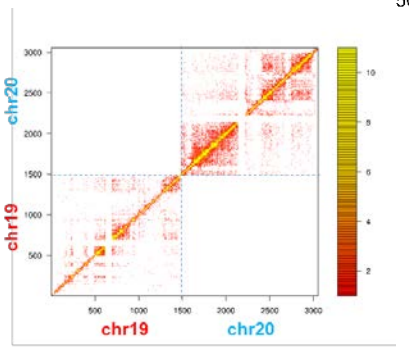




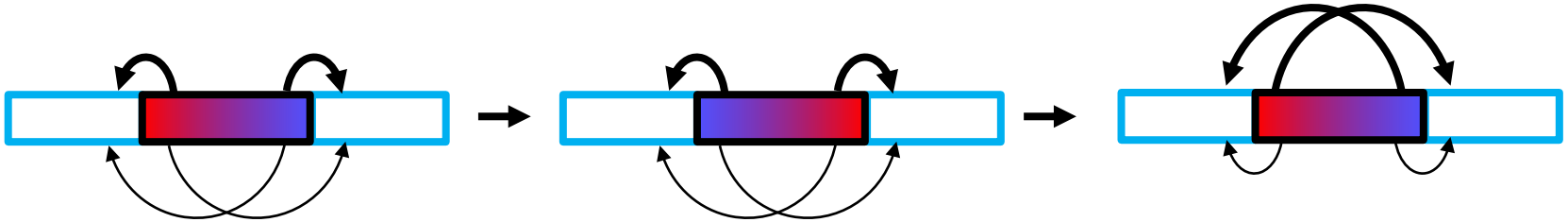
Translocation + inversion

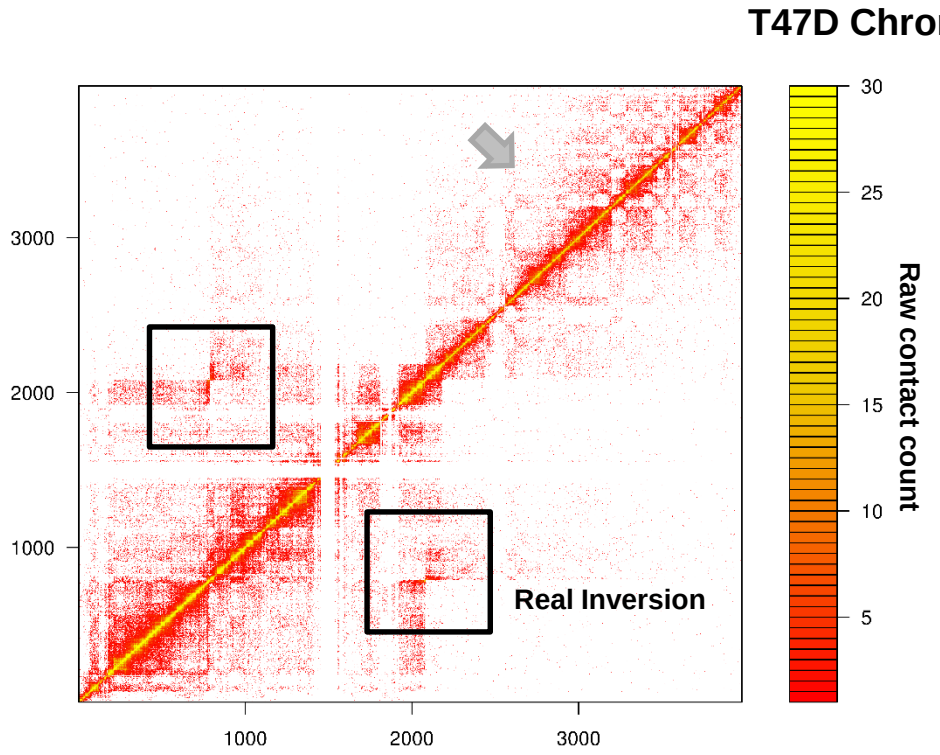


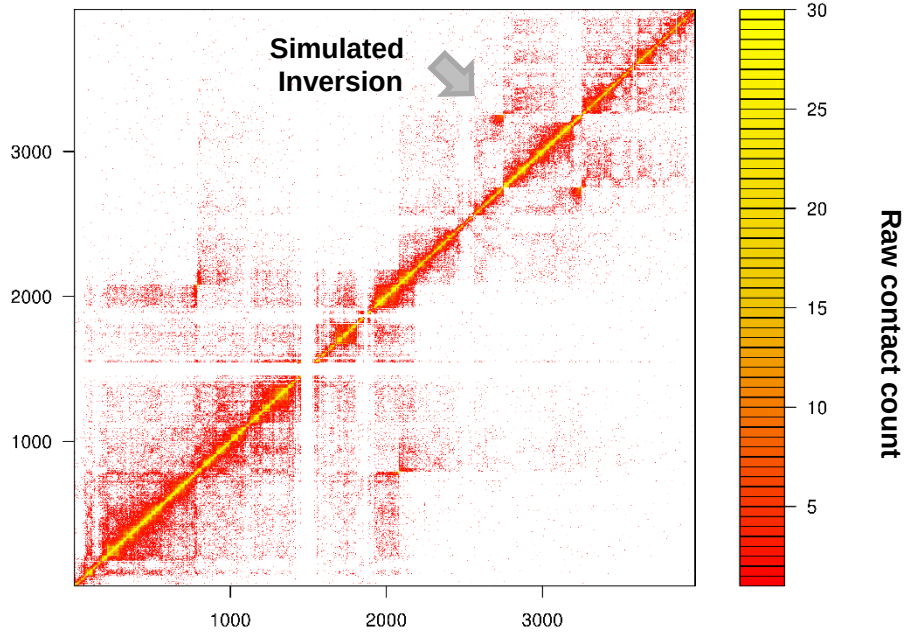




Simulating Hi-C matrices with inversions

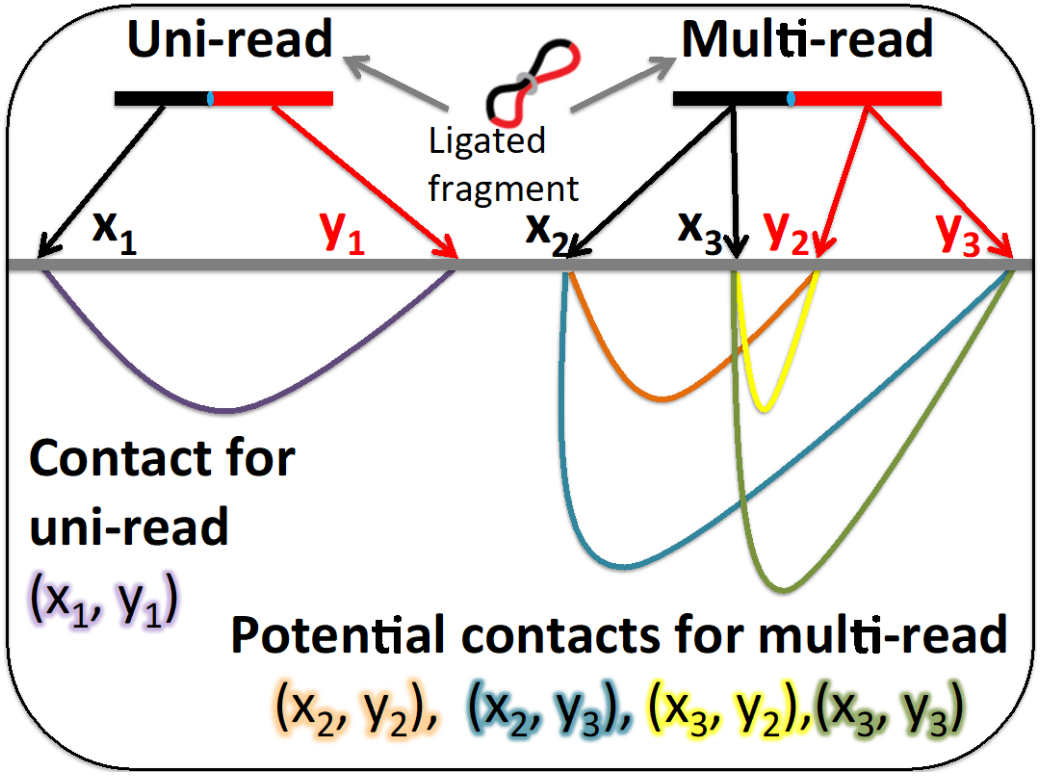






mHiC

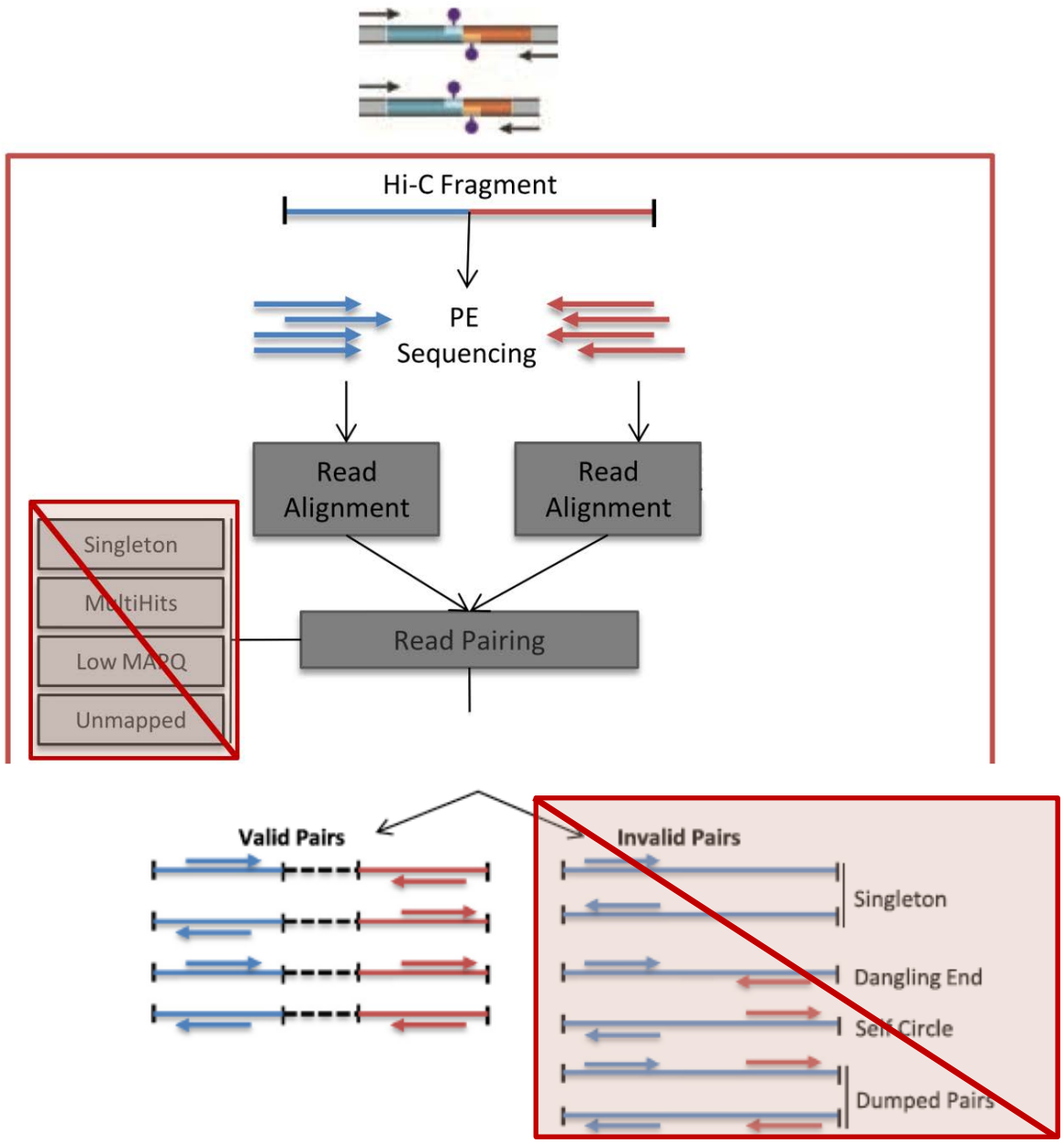
Leveraging multi-mapping reads in Hi-C data



U. of Wisconsin - Madison
Sunduz Keles
Ye Zheng

mHiC: a beta version is available from Ye Zheng yezheng@stat.wisc.edu

A typical Hi-C read processing pipeline



Multi-mapper aware read mapping

Read processing to get valid read pairs

Partition genome into fixed-size or RE-based bins

Generate raw contact map

mHiC makes these steps multi-read aware

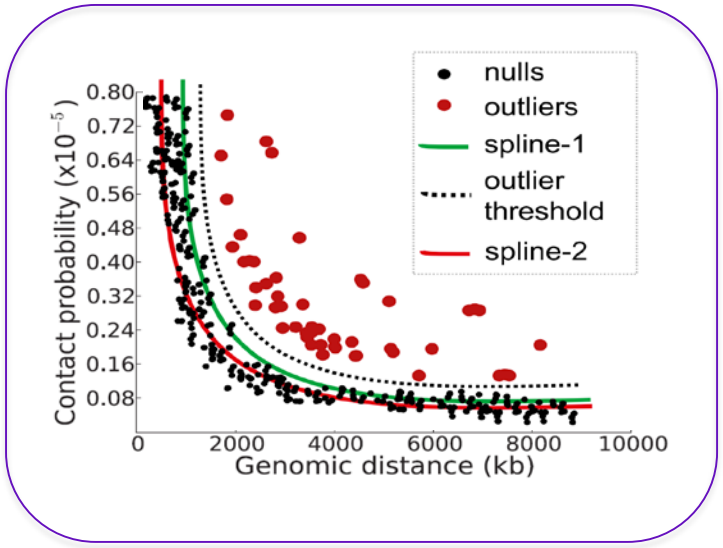
Normalize contact map

Identify significant contacts

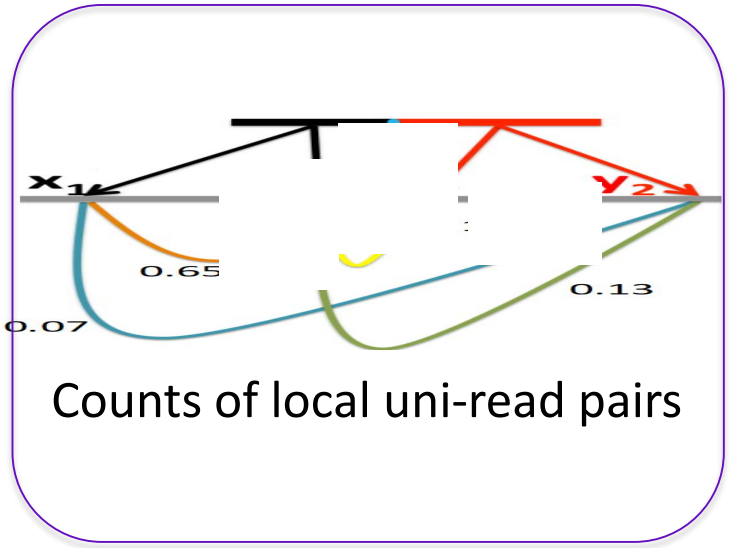
	# of conditions	# of reps	Genome size
Human: IMR90 (Jin et al., *Nature*, 2013)	1	4	3,234.83 Mb
P. falciparum (Ay et al., *GR*, 2014)	3	-	22.9 Mb

mHiC overview

1. Prior construction



2. Fit by EM

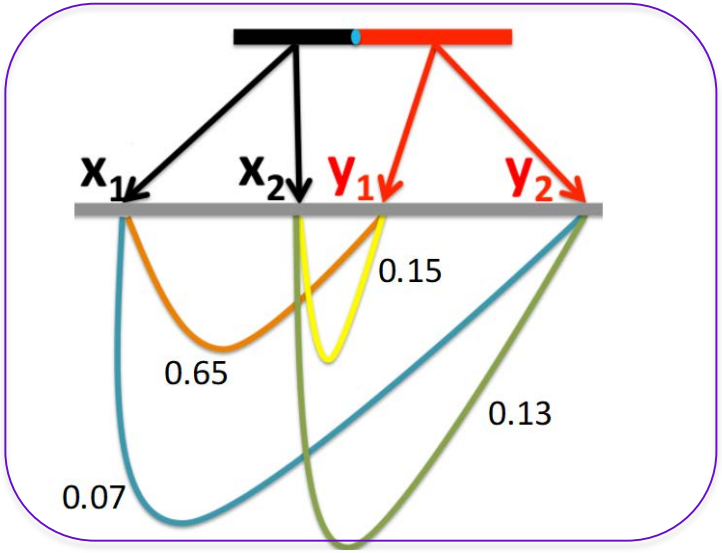


Counts of local uni-read pairs

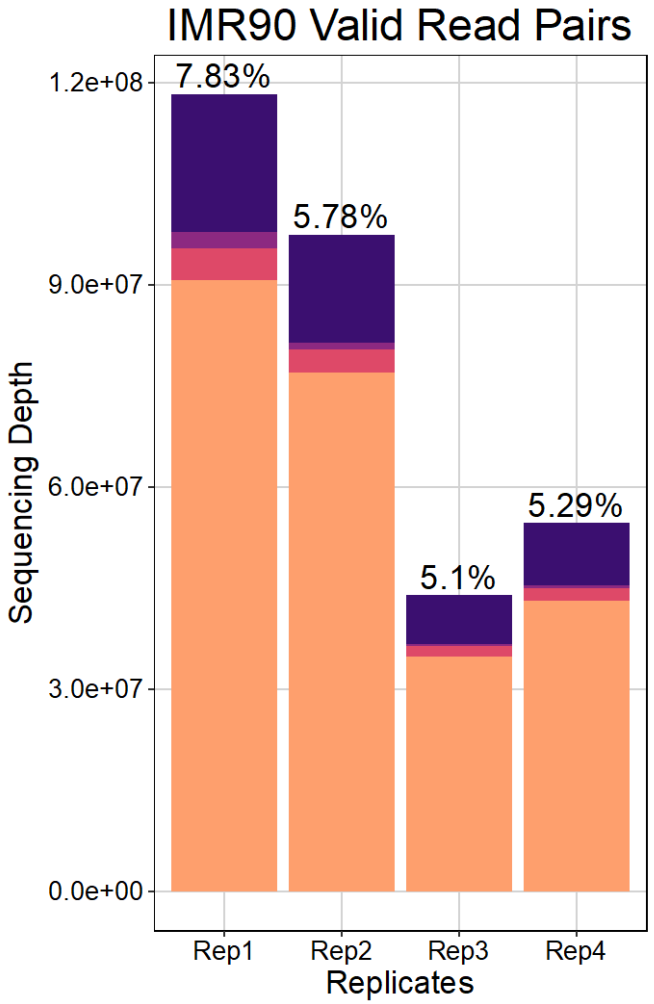
3. Posterior probabilities

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

Threshold posterior probabilities to use resulting alignments with existing significant contact identification pipelines (e.g.. Fit-Hi-C).



Improvement in sequencing depth using mHiC



A. Sequencing Depth	✓
B. Number of identified significant contacts	✓
C. Contact recovery at higher FDR	✓
D. Reproducibility across replicates	✓
E. Biological impact: Novel promoter-enhancer interactions	✓

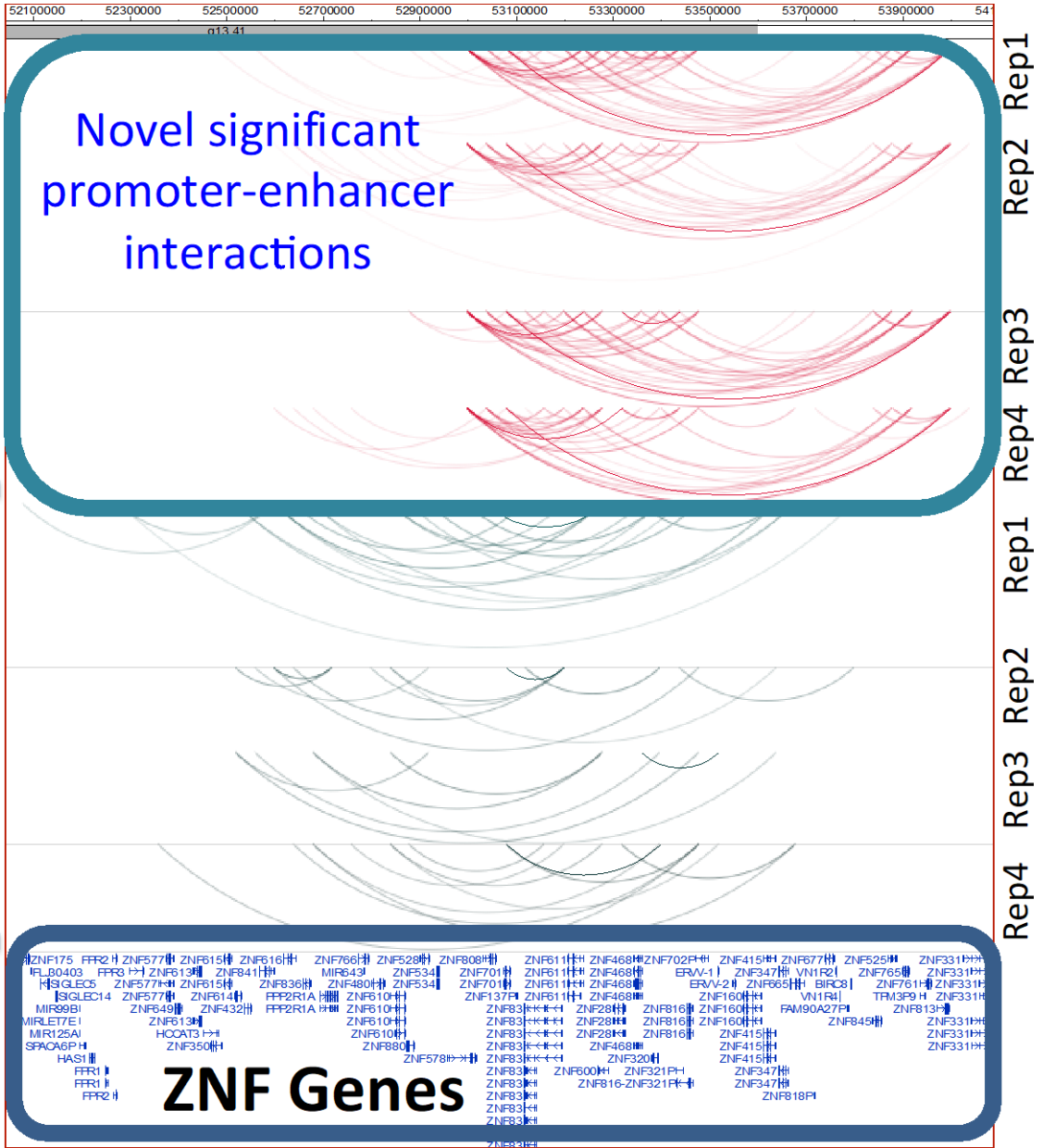
- ❑ Improves sequencing depth by 20-25%
- ❑ Only 5-8% are from EM modeling
- ❑ Substantial amount of gain from utilizing Hi-C read characteristics

mHiC identifies novel promoter-enhancer interactions

Chr19

20.4% (2,249) more promoter-enhancer interactions which are reproducible for at least 2 replicates.

Significant promoter-enhancer contacts identified even when only uni-reads are used



Big thanks to all my collaborators & funding

La Jolla Institute
P. Vijayanand
Benjamin Schmiedel
Greg Seumois
Anjana Rao lab

UC Riverside
Karine Le Roch lab

UT Health - San Antonio
Evelien Bunnik

UW- Genome Sciences
William Noble
Kate Cook
Wenxiu Ma
Max Libbrecht
Jay Shendure lab
Maitreya Dunham
Ivan Liachko

Stanford – Medicine
Andrew R. Hoffman lab

Ecole Des Mines, France
Jean-Philippe Vert
Nelle Varoquaux

Tel Aviv University
Judith Berman lab

Florida State – Biology
David M. Gilbert
Vishnu Dileep
Jiao Sima

Northwestern University
Jhumku Kohtz lab
Ramana Davuluri lab

U. Wisc.-Madison
Sushmita Roy lab
Sunduz Keles
Ye Zheng

U. Nevada - Reno
Timothy Bailey

ENCODE 3DN subgroup
Job Dekker
Jesse Dixon
Dave Gilbert
Ross Hardison
William Noble
Feng Yue

Funding

- NSF Computing Innovation Fellows (CIF-1136996)
- LJI Institute Leadership

La Jolla Institute FOR ALLERGY AND IMMUNOLOGY

Lab Members

- Abhijit Chakraborty (Postdoc)
- Jianlin Shao (Postdoc)
- Bharat Panwar (Postdoc – Ay & Vijay)
- Sourya B. (Postdoc – Ay & Vijay)
- Arya Kaul (UCSD)
- Rohan Paul (UCSD) → Stanford
- Ruyu Tan (UCSD) → U. of Colorado
- Lucas Patel (UCSD)
- Aarthi Venkat (UCSD)
- Jose Moreno (UCSD)

www.lji.org/faculty-research/labs/ay
ferhatay@lji.org