Identification of *Translocations*, *Amplifications* and *Deletions* in cancer genomes from HiC data

Abhijit Chakraborty (Postdoc)

Ay Lab

La Jolla Institute for Allergy and Immunology

09.16.2016

Immediate Objectives

1. A. Develop pipeline to identify possible translocations from HiC data.

   B. Identify the possible translocation boundaries in the two chromosomes.

2. Identify CNVs (amplifications and deletions) in the genomes from HiC data.

HiC data set used in the study

Cell line	All contacts (valid read pairs)		
	Combined	Repl1	Repl2
A549	136,384,017	66,335,366	70,054,115
LNCaP	149,387,648	81,257,702	68,130,231
PANC1	168,012,696	70,539,577	97,475,537
T47D	137,395,622	64,452,498	72,948,150
CAKI2	168,540,398	91,078,430	77,462,925
NCIH460	155,078,021	85,902,379	69,176,114

ENCODE/HAIB CNV information available

Cell Line	No. of Amp known	
	Total	≥ 40Kb
A549	12	11
LNCaP	55	10
PANC1	16	14
T47D	56	51
Total	139	86

Cell Line	No. of Del known	
	Total	≥ 40Kb
A549	88	20
LNCaP	87	43
PANC1	262	131
T47D	75	43
Total	512	237

Translocation calling pipeline

40Kb bins

Raw HiC inter-chromosomal counts (e.g. chrA-chrB)

Explicit normalization using HiCNorm

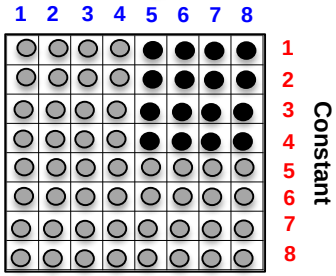
GC >= 0.2
Mappability >= 0.5
Exclude black listed regions

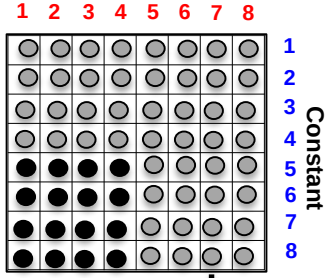
Normalized inter-chromosomal matrix

Fit Poisson dist. to calculate λ

Generate random matrix

Perform 1D binary segmentation on each chromosomal bin





Cluster chrB regions 5:8 × **Cluster chrA regions 1:4**

When multiple regions are present then every combination is compared

Detect changes in signal in chrA/bin

- 1 → 5:8
- 2 → 5:8
- 3 → 5:8
- 4 → 5:8

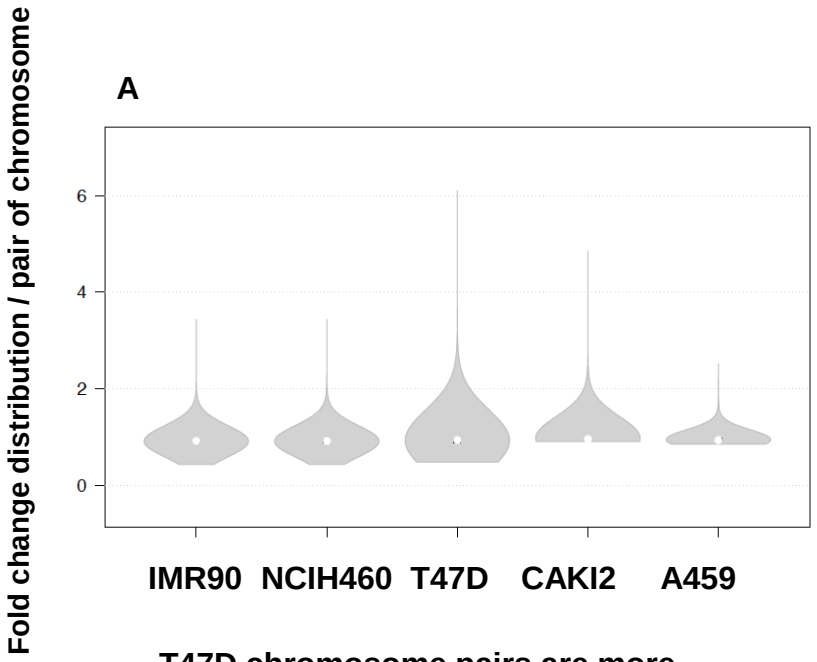
Detect changes in signal in chrB/bin

- 5 → 1:4
- 6 → 1:4
- 7 → 1:4
- 8 → 1:4

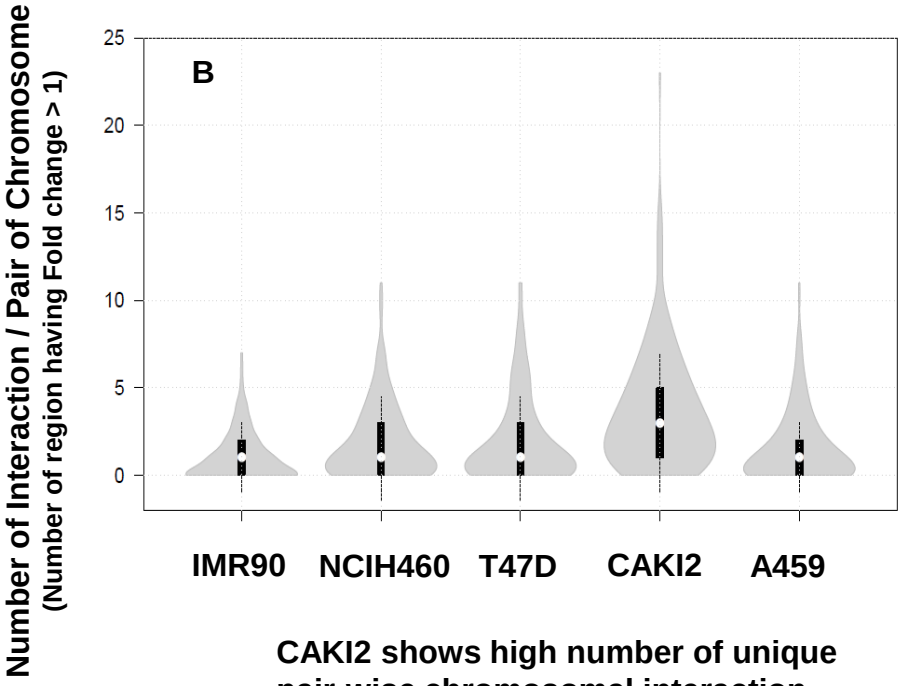
Extract the regions from matrix (M)

1. Each combination of "M" is then compared against a randomly generated similar sized matrix.
2. Calculate for each "M", the fold change = M^{norm} / RM^{norm} [expected = 1]
3. If fold change > 1 then report M(s) as regions having higher inter-chromosomal contact
4. Post-filtering on these regions.

Post-filtering and comparison with IMR90 cell line



T47D chromosome pairs are more enriched in interaction than random



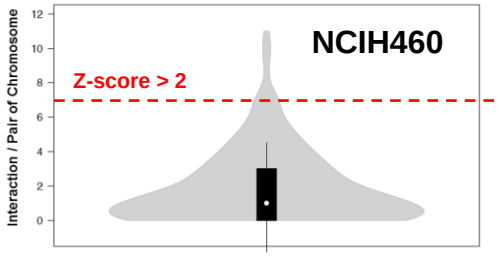
CAKI2 shows high number of unique pair-wise chromosomal interaction

Compared all unique chromosome pairs (253) and calculated the number of non-overlapped regions (fold change > 1) per pair of chromosome. The Y-axis in figure B plots the distribution of this "number of non-overlapped regions/per chromosome".

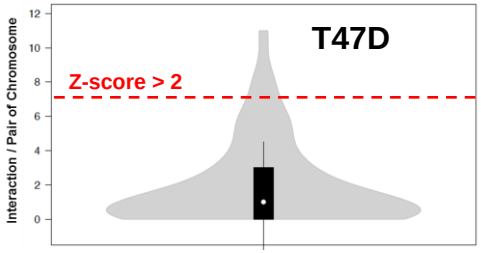
For e.g. IMR90 has a single pair with 7 interacting regions (all enrichment > 1) while CAKI2 has a single pair with 23 interacting regions. The assumption is - higher the number of interacting regions for a pair, more likely that pair contains translocation.

Overview of the translocation detection result in 4 cancer cell lines:

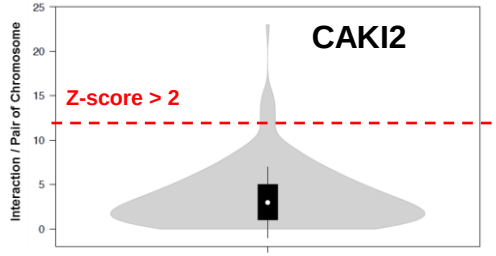
(A Z-score cutoff 2 is used as threshold for demonstration)



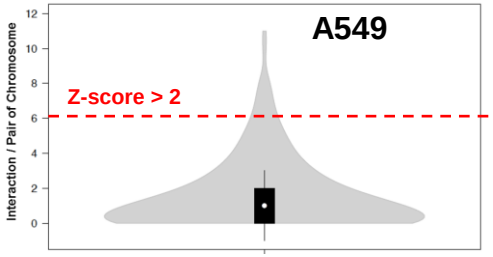
- Detected 10 translocations.
- 3 Known translocations and 1 RT detected breakpoint



- Detected 17 translocations.
- 4 Known translocations
- 3 RT predicted translocation is above Z-score 2
- The second RT predicted translocation not detected



- Detected 9 translocations.
- One RT predicted translocation is above Z-score 2



- Detected 15 translocations.
- 2 Known translocations
- One RT predicted translocation is above Z-score 2

Total in 4 cell line:

- 51 Translocations (Z-score > 2) [403 (Z-score > 0)]
- 9/15 Known translocation (Z-score > 2) [6 (Z-score > 0 & <= 2)]
- 6/9 RT predicted breakpoint (Z-score > 2) [RT: Replication Timing]
- 2/9 RT predicted breakpoint (Z-score >= 0 and <= 2)

Examples of identified Known translocations

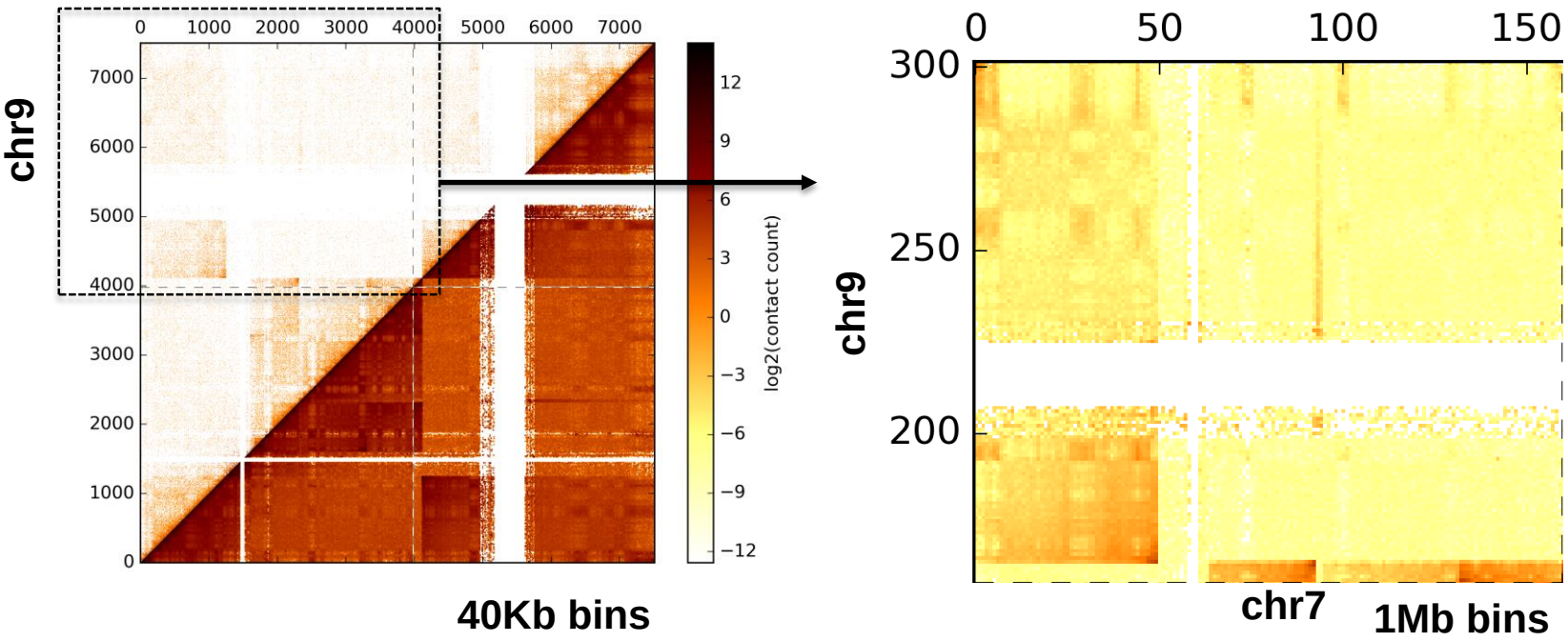
NCIH460 top translocations (Z score > 2)

	Rank	ChrA	ChrB	Interactions	Zscore
Known	1	chr7	chr9	11	4.29554649
Known	2	chr1	chr9	10	3.82415873
Known	3	chr7	chr16	10	3.82415873
	4	chr1	chr12	9	3.35277097
	5	chr5	chr17	7	2.40999544
	6	chr1	chr5	7	2.40999544
	7	chr2	chr9	7	2.40999544
	8	chr1	chr2	7	2.40999544
	9	chr12	chr21	7	2.40999544
	10	chr9	chr16	7	2.40999544

NCI-H460 [H460] (ATCC® HTB-177™)

Karyotype

modal numbr = 57; range = 53 to 65.This is a hypotriploid human cell line. The modal chromosome number is 57 although cells with 58 chromosomes occurred with a comparable frequency. The frequency of higher ploidies was 1.7%. Seven marker chromosomes, der(9)t(1;9)(q21;p24), der(9)t(7;9)(p11;p22), t(10q14q), der(16)t(7;16)(q11.23;q22), a small ring (about 1/2 the size of a G chromosome) and two others, were common to all cells. Three other markers were found in some cells only. The markers, t(7;9) and t(7;16) were mostly paired. Normal N9 was absent, and N7 and N16 had only a single copy per cell. Two copies each of the X and the Y were present in all cells.



Result :

Pair	Chr7		Chr9		Fold Change
	Start Index	End Index	Start Index	End Index	
1	3	32	2489	2632	1.63
2	3	32	1925	2485	1.19
3	31	74	2489	2632	1.39
4	31	74	822	1557	1.11
5	31	74	1925	2485	1.25
6	114	220	2489	2632	1.04
7	114	220	1925	2485	1.24
8	734	1193	822	1557	1.09
9	734	1193	77	536	2.07
10	1951	3544	77	536	1.27
11	1951	3544	1	58	1.35

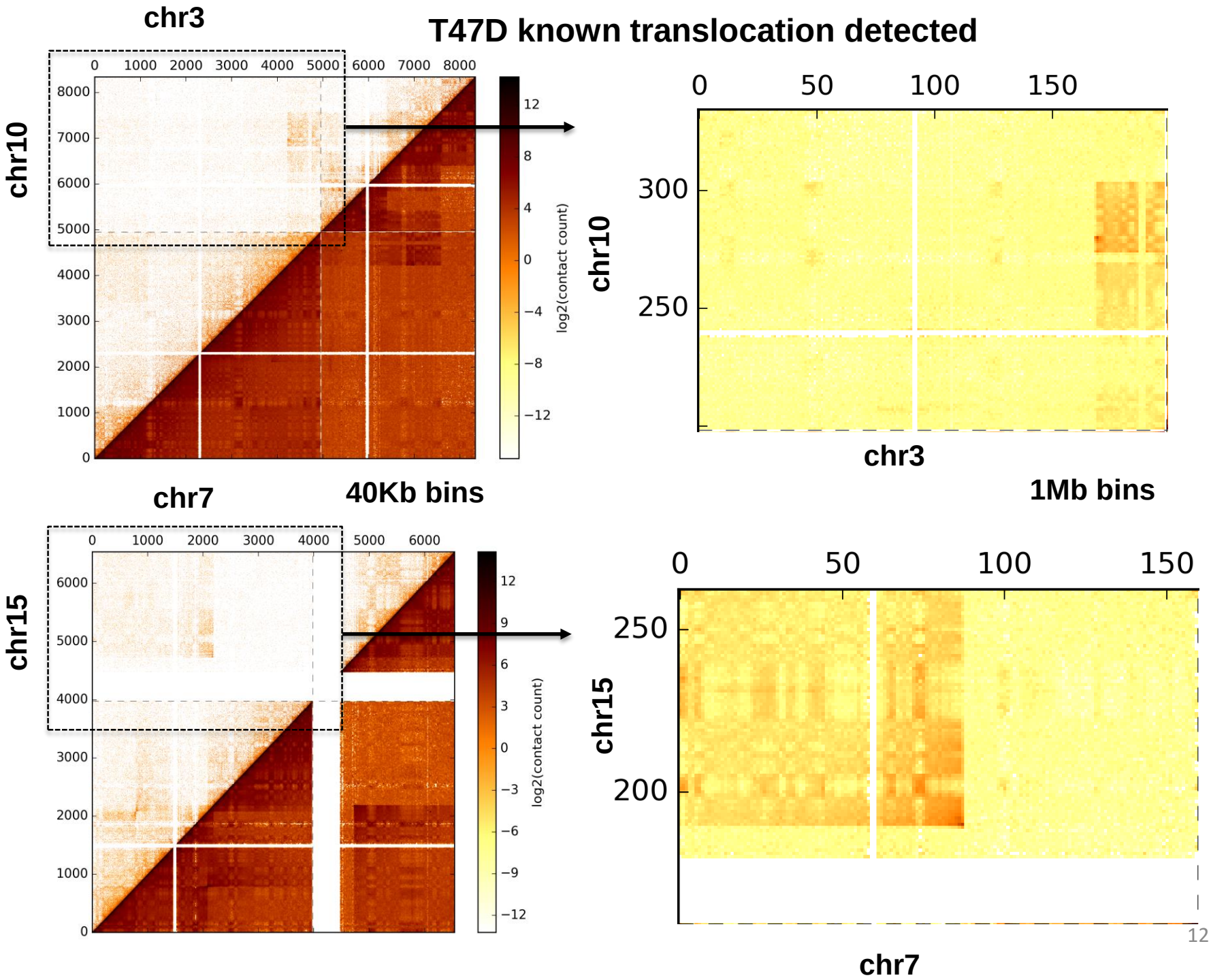
T47D top translocations (Z score > 2)

	Rank	ChrA	ChrB	Interactions	Zscore
Known	1	chr3	chr10	11	3.66737354
	2	chr3	chr5	11	3.66737354
	3	chr8	chr19	11	3.66737354
Known	4	chr8	chr14	10	3.26006647
	5	chr3	chr8	9	2.8527594
Known	6	chr7	chr15	9	2.8527594
	7	chr5	chr21	8	2.44545233
	8	chr3	chr21	8	2.44545233
	9	chr8	chr22	8	2.44545233
	10	chr8	chr21	8	2.44545233
	11	chr8	chr17	8	2.44545233
	12	chr7	chr21	8	2.44545233
	13	chr3	chr12	7	2.03814526
Known	14	chr10	chr20	7	2.03814526
	15	chr8	chr20	7	2.03814526
	16	chr8	chr9	7	2.03814526
	17	chr7	chr8	7	2.03814526

Differences and homologies of chromosomal alterations within and between breast cancer cell lines: a clustering analysis

Milena Rondón-Lagos[1,3], Ludovica Verdun Di Cantogno[2], Caterina Marchiò[1,2], Nelson Rangel[1], Cesar Payan-Gomez[3,4], Patrizia Gugliotta[1], Cristina Botta[1], Gianni Bussolati[1], Sandra R Ramírez-Clavijo[3], Barbara Pasini[1] and Anna Sapino[1,2*]

T47D known translocation detected



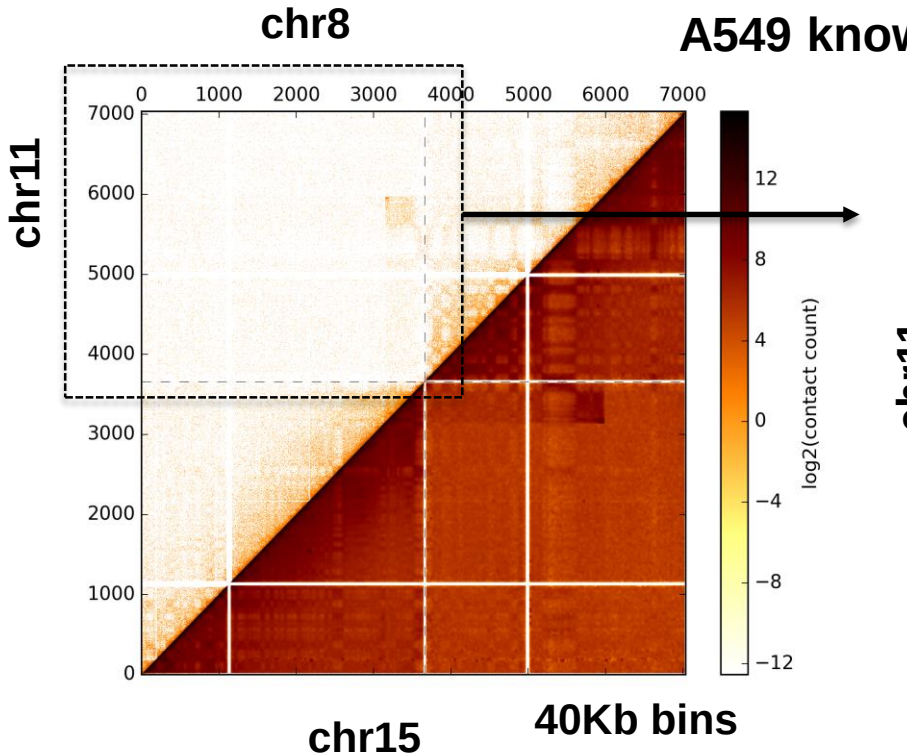
A549 top translocations (Z score > 2)

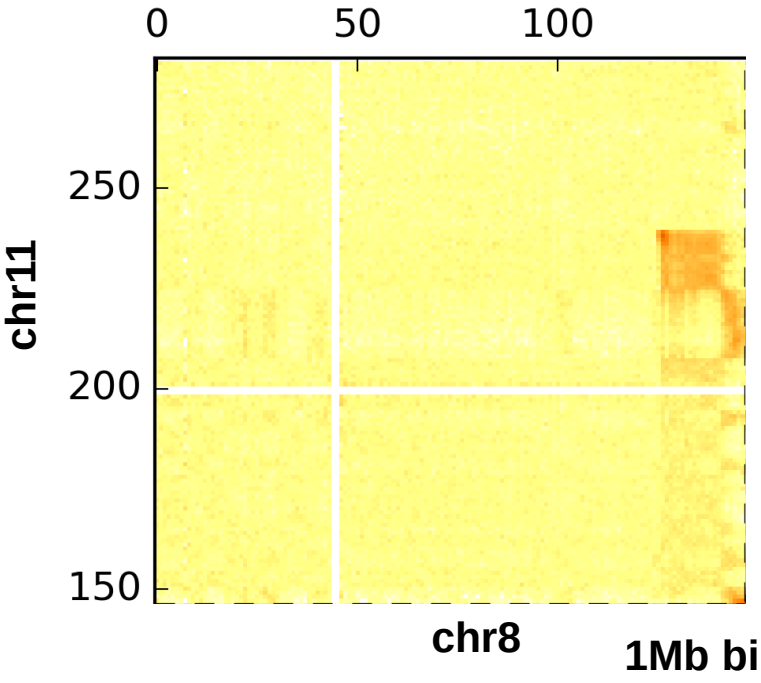
	Rank	ChrA	ChrB	Interactions	Zscore
Known	1	chr8	chr11	11	4.6516191
Known	2	chr15	chr19	10	4.1546345
	3	chr16	chr19	8	3.1606652
	4	chr9	chr19	8	3.1606652
	5	chr7	chr19	8	3.1606652
	6	chr4	chr19	7	2.6636806
	7	chr11	chr17	7	2.6636806
	8	chr11	chr12	7	2.6636806
	9	chr8	chr19	7	2.6636806
	10	chr2	chr12	6	2.1666959
	11	chr17	chr20	6	2.1666959
	12	chr11	chr20	6	2.1666959
	13	chr11	chr19	6	2.1666959
	14	chr8	chr21	6	2.1666959
	15	chr7	chr20	6	2.1666959
Known	27	chr3	chr20	4	1.1727266
Known	67	chr4	chr12	2	0.1787573

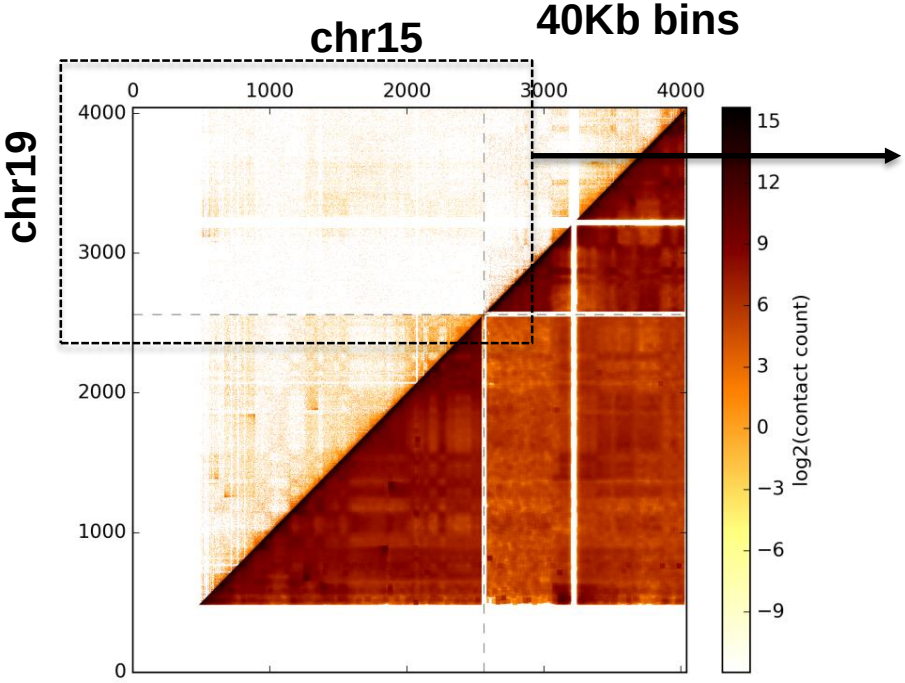
Characterization of Two Human Lung Adenocarcinoma Cell Lines by Reciprocal Chromosome Painting

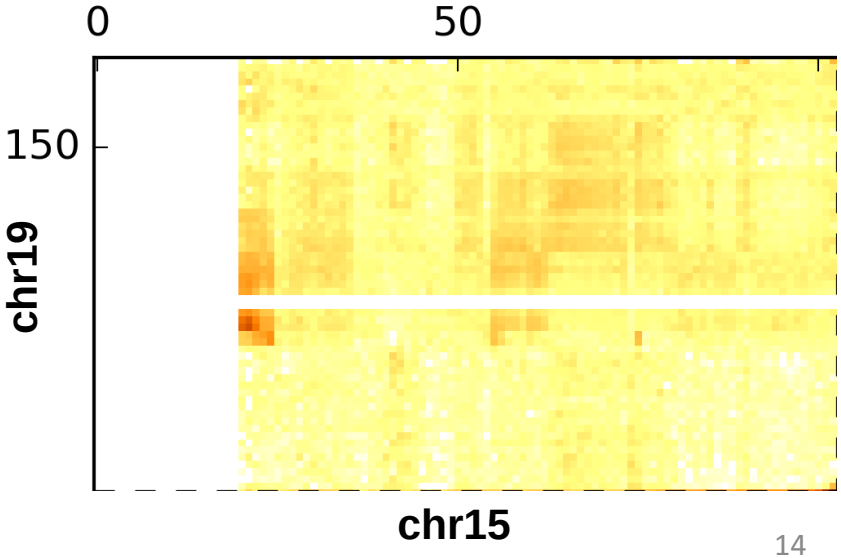
PENG Kun-Jing[1,4], WANG Jin-Huan[1], SU Wei-Ting[1], WANG Xi-Cai[2], YANG Feng-Tang[3], NIE Wen-Hui[1,*]

A549 known translocation detected

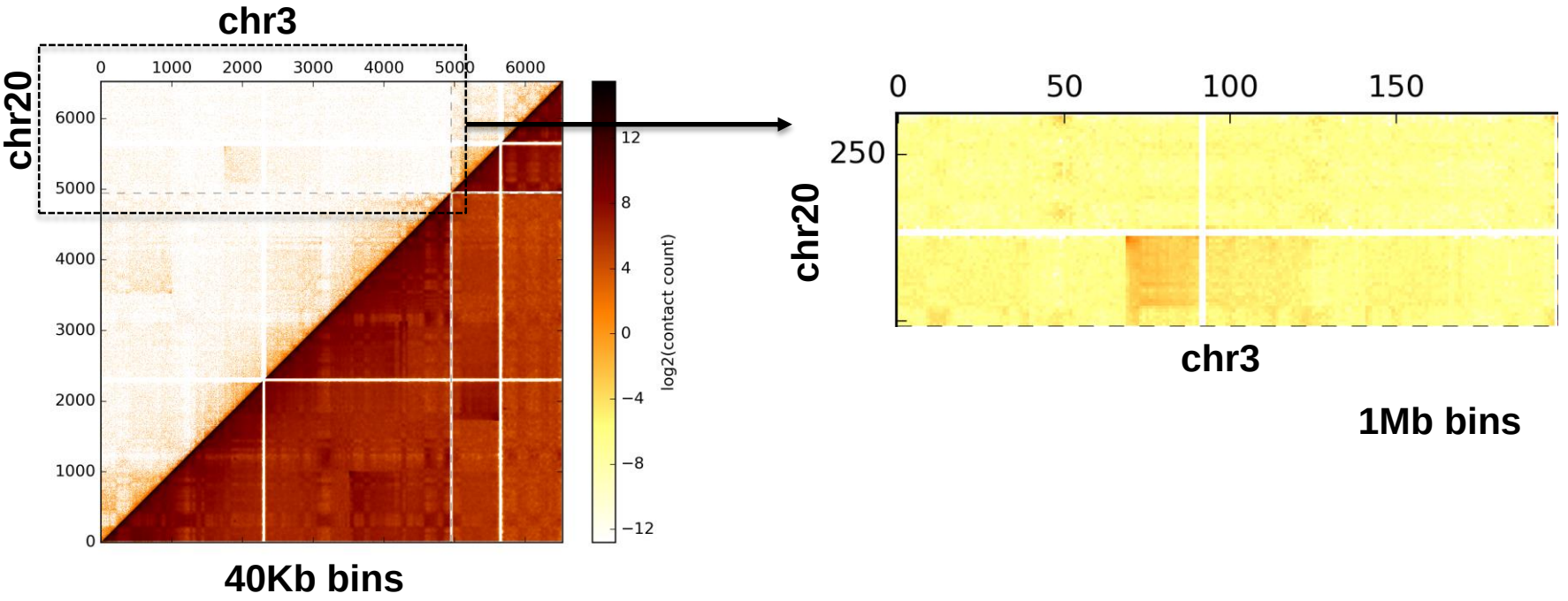








A549 known translocation detected



Breakpoints predicted from Replication Timing experiments

Breakpoints with the biggest shift in RT
Reported by Gilbert group.

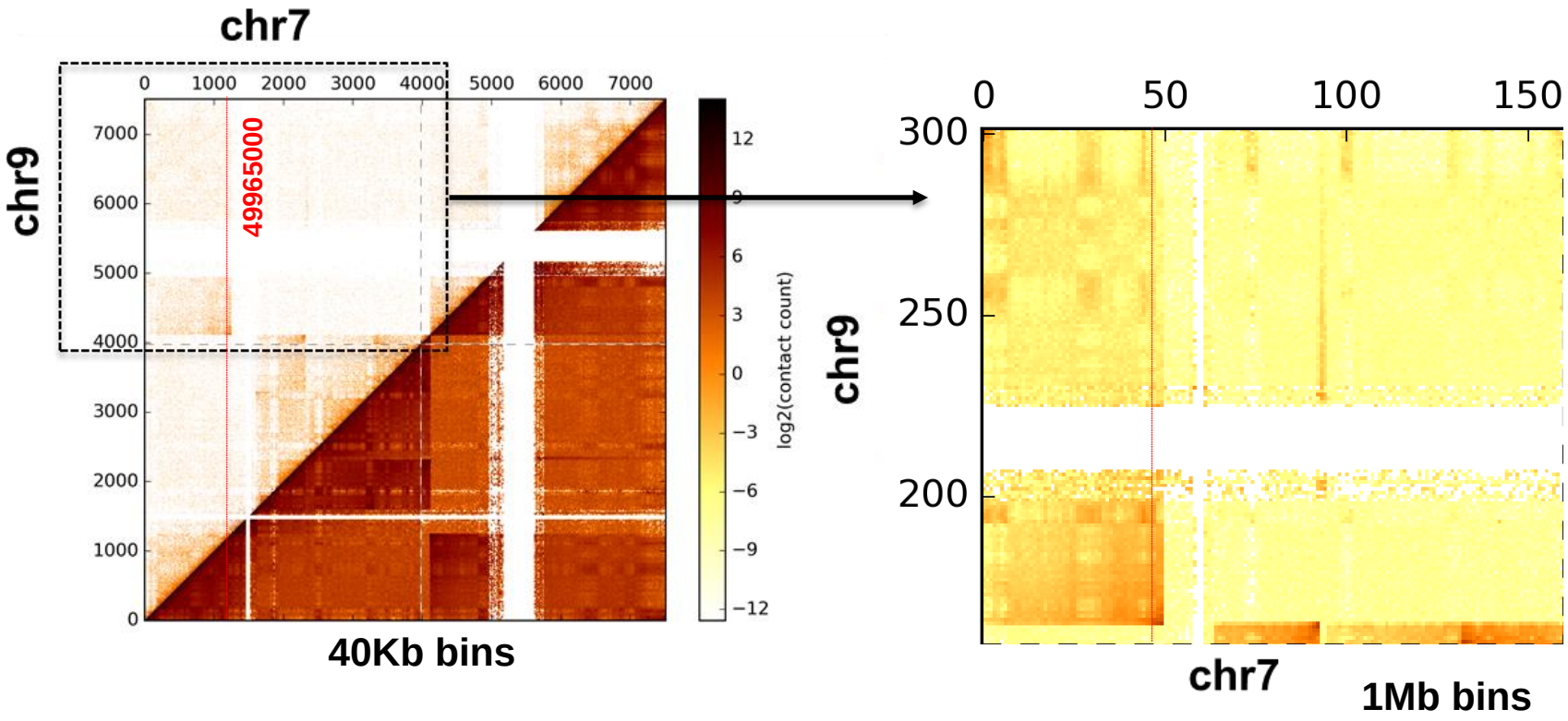
RT predicted breakpoints in NCIH460

- Chr7 : 49965000
- Chr19 : 37710000
- Chr19 : 36770000
- ChrX : 119185000 (Not detected)

Within NCIH460 top translocations (Z score > 2)

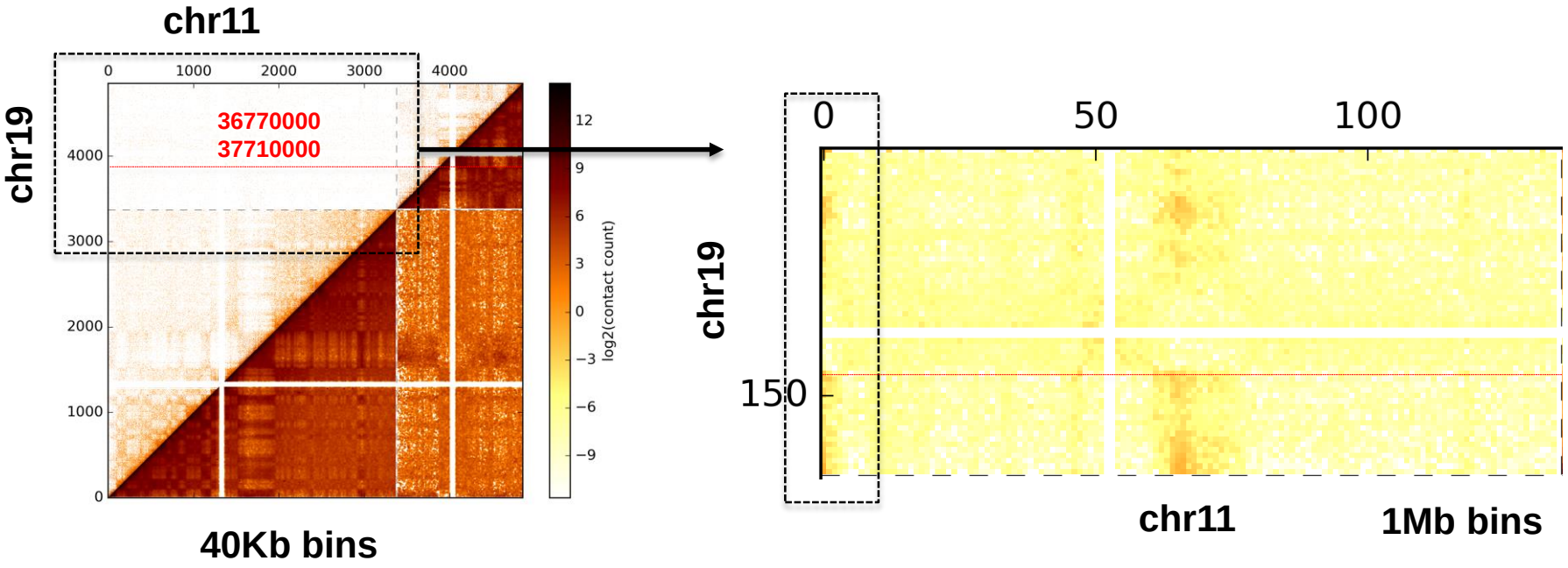
	Rank	ChrA	ChrB	Interactions	Zscore	
Known	1	chr7	chr9	11	4.29554649	RT detection
Known	2	chr1	chr9	10	3.82415873	
Known	3	chr7	chr16	10	3.82415873	
	4	chr1	chr12	9	3.35277097	
	5	chr5	chr17	7	2.40999544	
	6	chr1	chr5	7	2.40999544	
	7	chr2	chr9	7	2.40999544	
	8	chr1	chr2	7	2.40999544	
	9	chr12	chr21	7	2.40999544	
	10	chr9	chr16	7	2.40999544	

NCIH460 RT predicted breakpoint



NCIH460 RT predicted breakpoint

- **The result for Chr19 showed it has a positive Z-score (0.99) interaction with Chr11. The respective boundaries of translocation includes:**
  - **Chr11 : 280000-320000 To 760000-8e+05**
  - **Chr19 : 33680000-33720000 To 40800000 -40840000**



RT predicted breakpoints in T47D

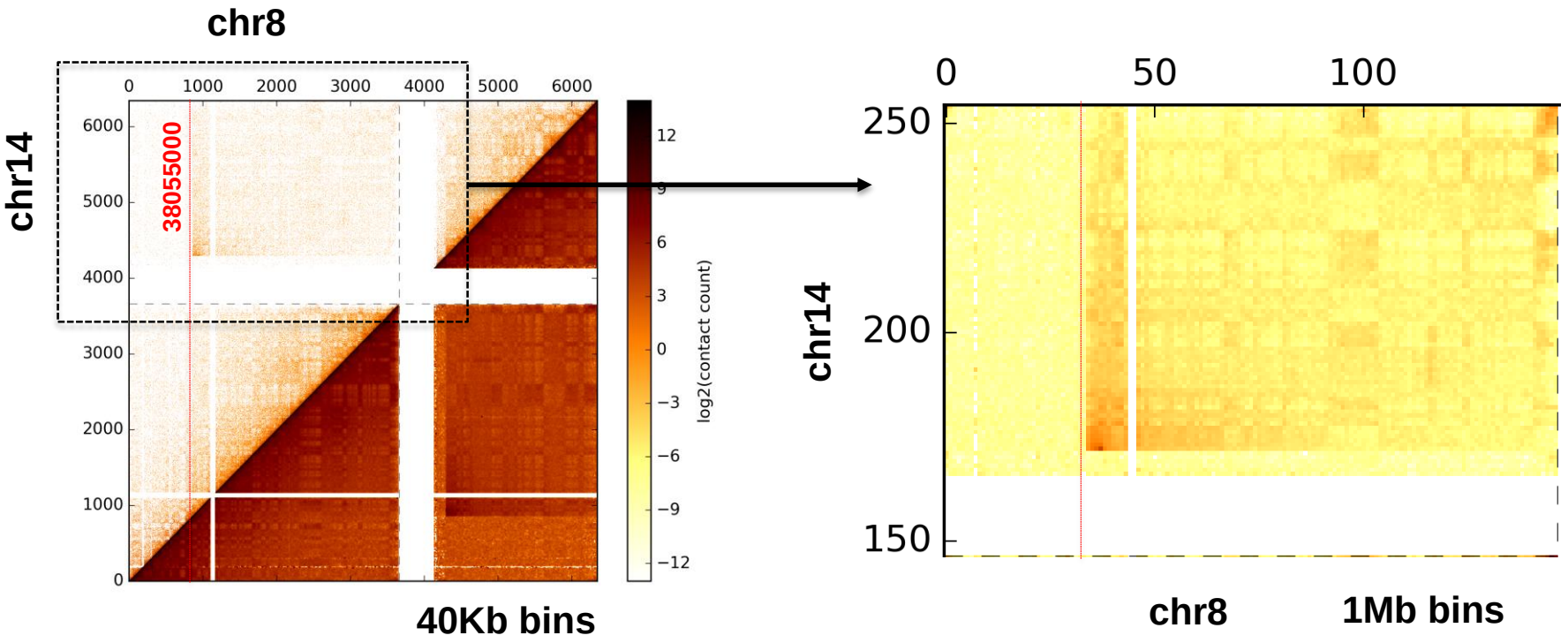
- Chr8 : 38055000
- Chr6 : 66935000 (Not detected)

T47D top translocations (Z score > 2)

	Rank	ChrA	ChrB	Interactions	Zscore	
Known	1	chr3	chr10	11	3.66737354	
	2	chr3	chr5	11	3.66737354	
	3	chr8	chr19	11	3.66737354	
Known	4	chr8	chr14	10	3.26006647	RT detection
	5	chr3	chr8	9	2.8527594	
Known	6	chr7	chr15	9	2.8527594	
	7	chr5	chr21	8	2.44545233	
	8	chr3	chr21	8	2.44545233	
	9	chr8	chr22	8	2.44545233	
	10	chr8	chr21	8	2.44545233	
	11	chr8	chr17	8	2.44545233	
	12	chr7	chr21	8	2.44545233	
	13	chr3	chr12	7	2.03814526	
Known	14	chr10	chr20	7	2.03814526	
	15	chr8	chr20	7	2.03814526	
	16	chr8	chr9	7	2.03814526	
	17	chr7	chr8	7	2.03814526	

T47D RT predicted breakpoint

- **The result for Chr8 showed it has a positive Z-score (3.26) interaction with Chr14. The respective boundaries of translocation includes:**
  - **Chr8 : 40880000-40920000 To 50280000-50320000**
  - **Chr14 : 25320000-25360000 To 31400000-31440000**



RT predicted breakpoints in A549

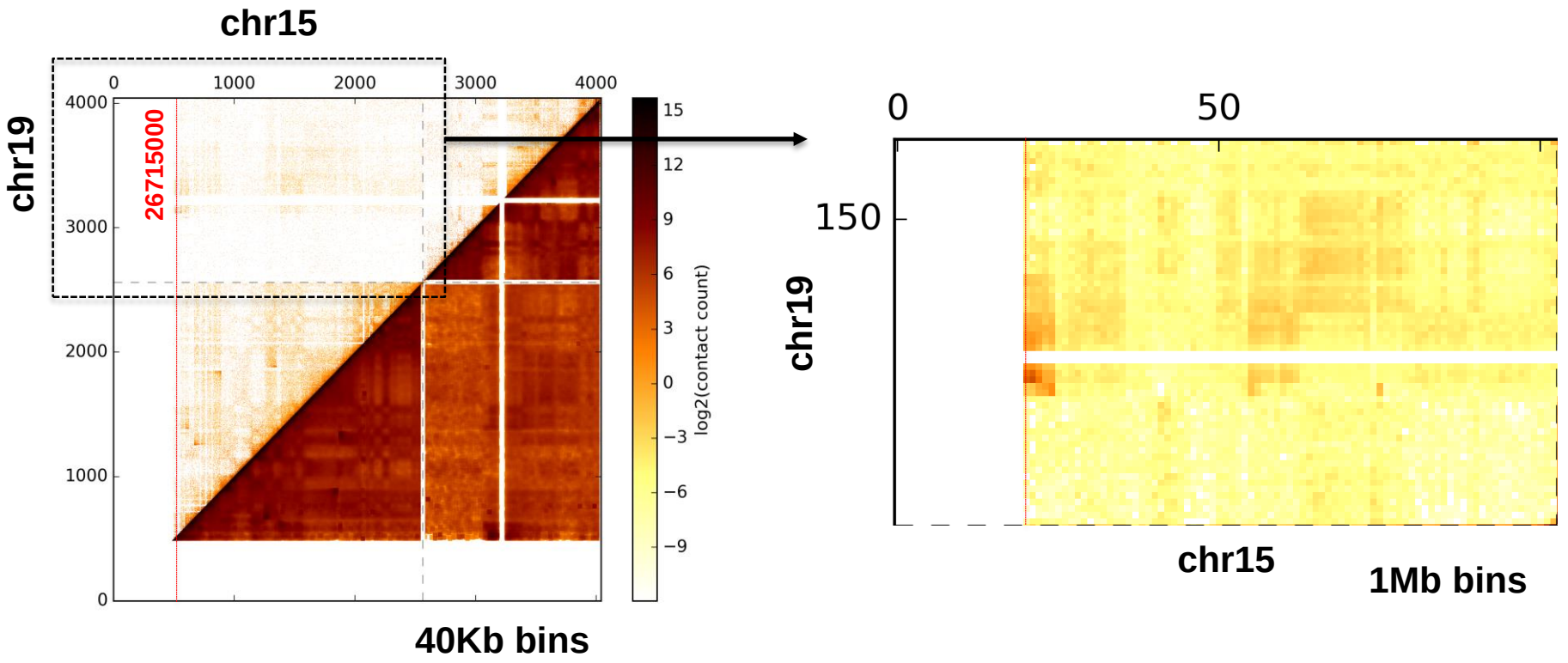
- Chr15 : 26715000

A549 top translocations (Z score > 2)

	Rank	ChrA	ChrB	Interactions	Zscore	
Known	1	chr8	chr11	11	4.6516191	
Known	2	chr15	chr19	10	4.1546345	RT detection
	3	chr16	chr19	8	3.1606652	
	4	chr9	chr19	8	3.1606652	
	5	chr7	chr19	8	3.1606652	
	6	chr4	chr19	7	2.6636806	
	7	chr11	chr17	7	2.6636806	
	8	chr11	chr12	7	2.6636806	
	9	chr8	chr19	7	2.6636806	
	10	chr2	chr12	6	2.1666959	
	11	chr17	chr20	6	2.1666959	
	12	chr11	chr20	6	2.1666959	
	13	chr11	chr19	6	2.1666959	
	14	chr8	chr21	6	2.1666959	
	15	chr7	chr20	6	2.1666959	
Known	27	chr3	chr20	4	1.1727266	
Known	67	chr4	chr12	2	0.1787573	

A549 RT predicted breakpoint

- **Chr15 : 26715000 (Z-score 4.12)**
  - **Chr15 : 20840000-20880000 To 26920000-26960000**
  - **Chr19 : 30120000-30160000 To 36520000-36560000**



RT predicted breakpoints in CAKI2

- Chr4 : 49085000
- ChrX : 114980000 (Not Detected)

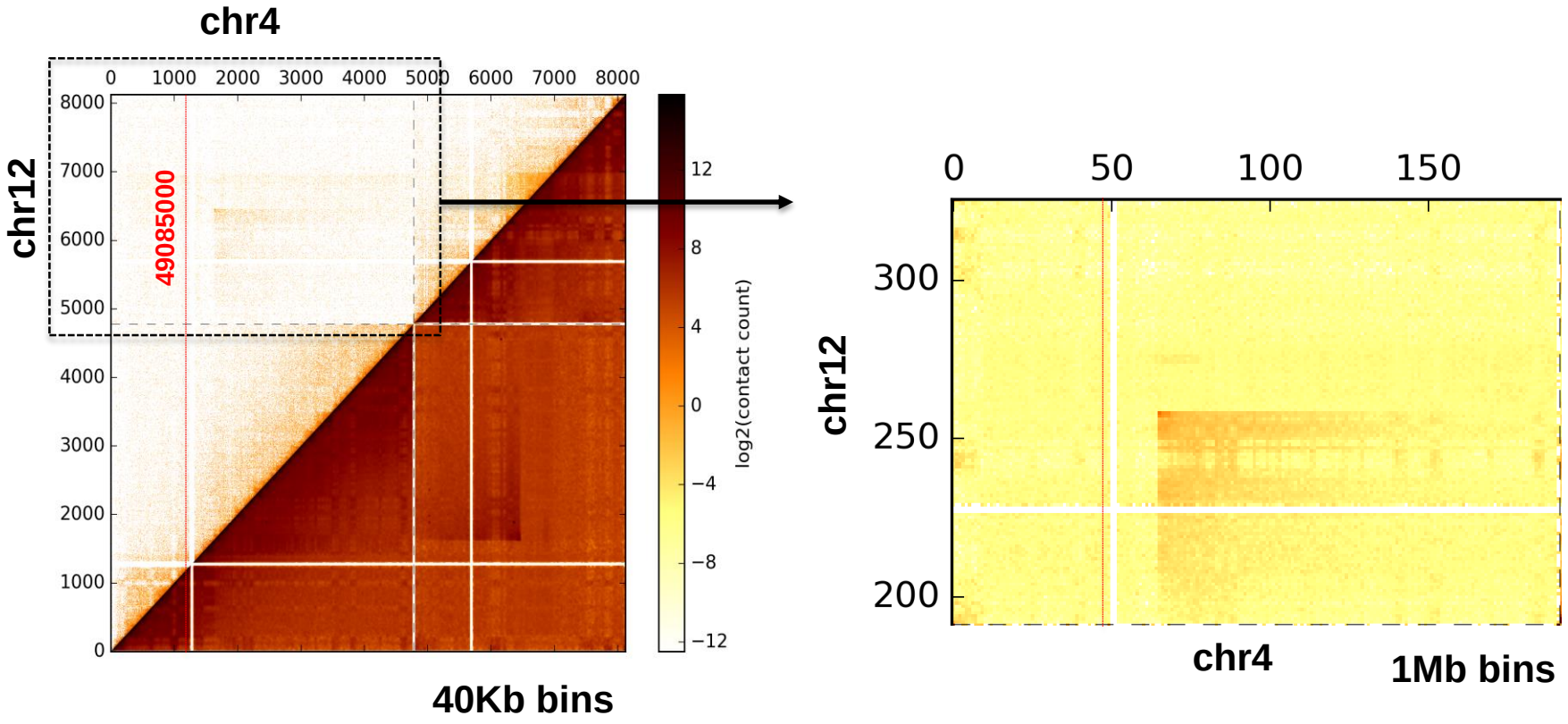
CAKI2 top translocations (Z score > 2)

Rank	ChrA	ChrB	Interactions	Zscore
1	chr3	chr12	23	5.8568246
2	chr2	chr12	15	3.4371036
3	chr6	chr12	15	3.4371036
4	chr2	chr3	14	3.1346385
5	chr1	chr12	14	3.1346385
6	chr7	chr12	13	2.8321734
7	chr5	chr12	12	2.5297083
8	chr4	chr12	12	2.5297083
9	chr11	chr12	12	2.5297083

RT detection

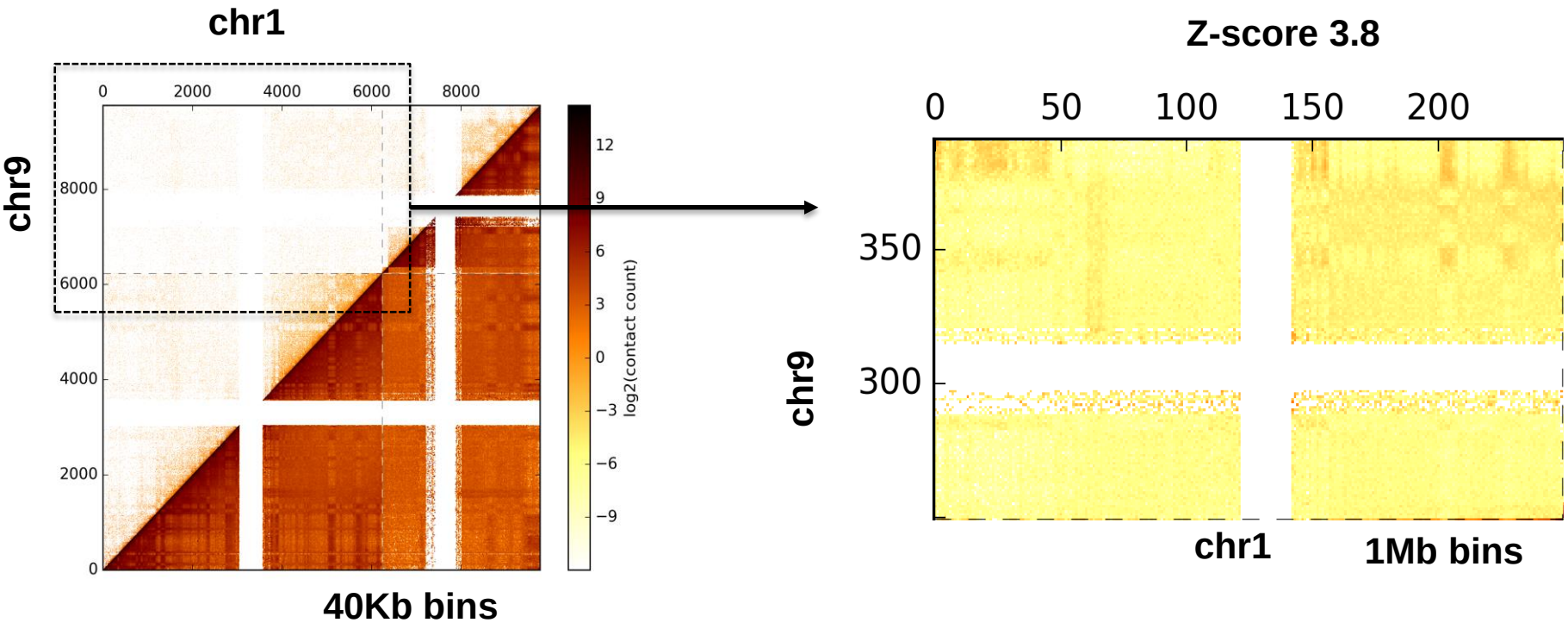
CAKI2 RT predicted breakpoint

- Chr4 : 49085000 (Z-score 2.53)
  - Chr4 : 22240000-22280000 To 57360000-57400000
  - Chr12: 31560000-31600000 To 58320000-58360000

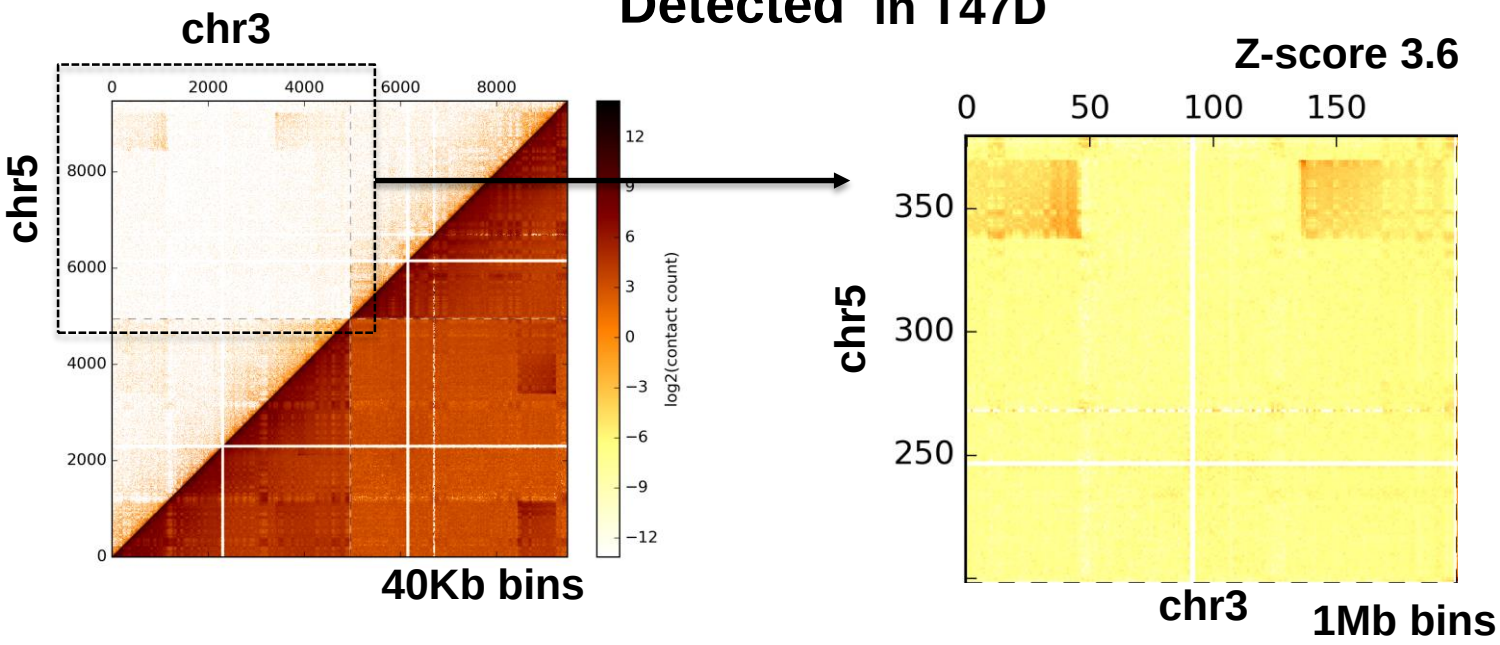


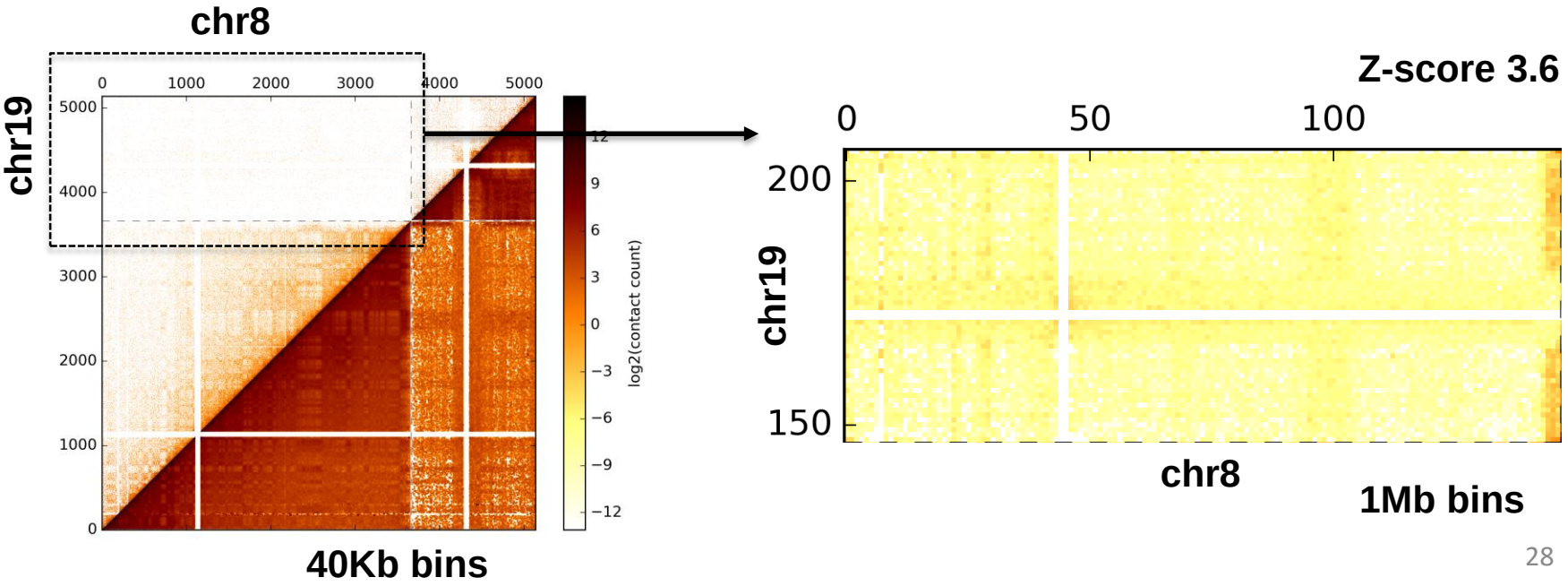
Examples of de novo translocation predictions

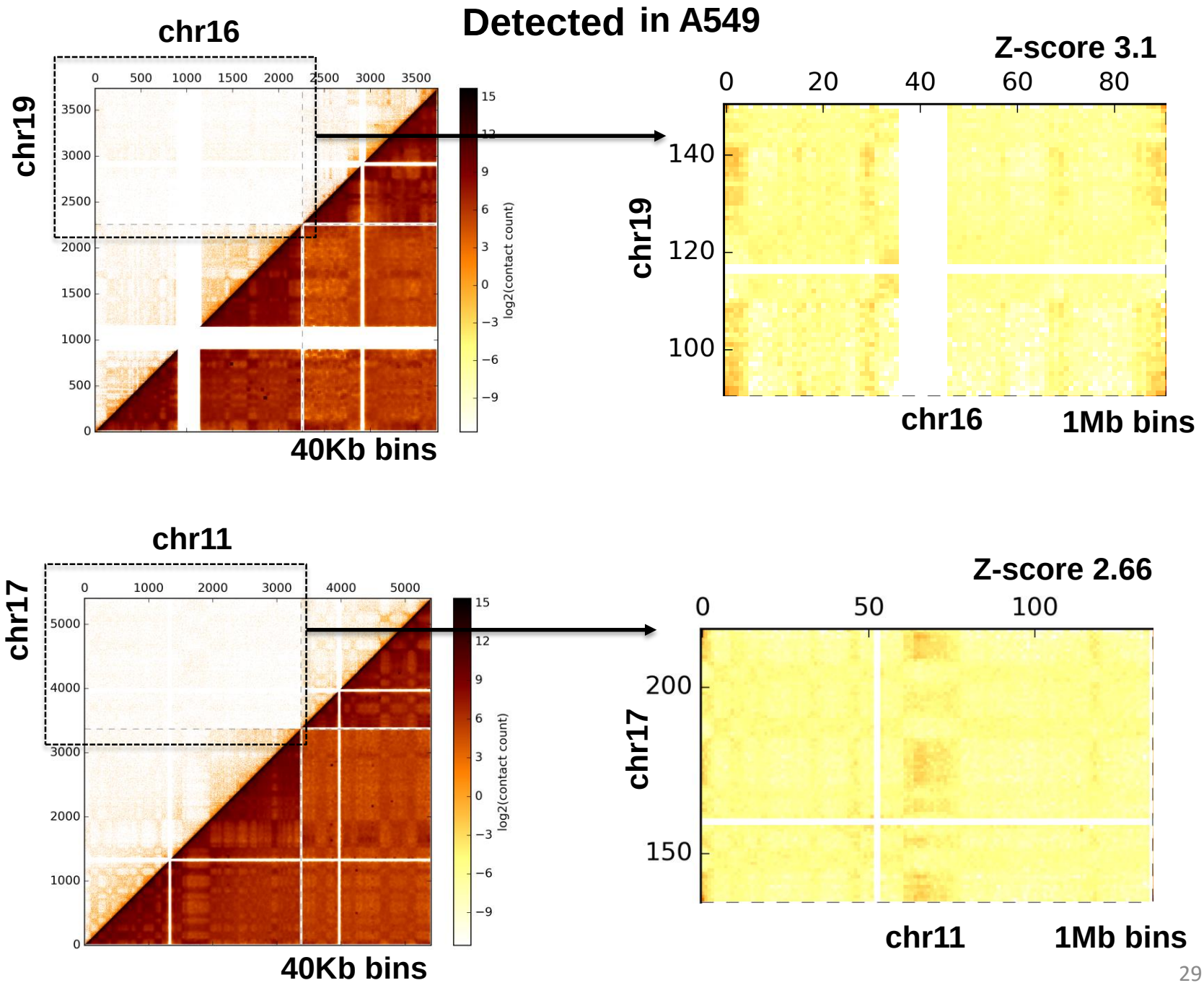
Detected in NCIH460



Detected in T47D



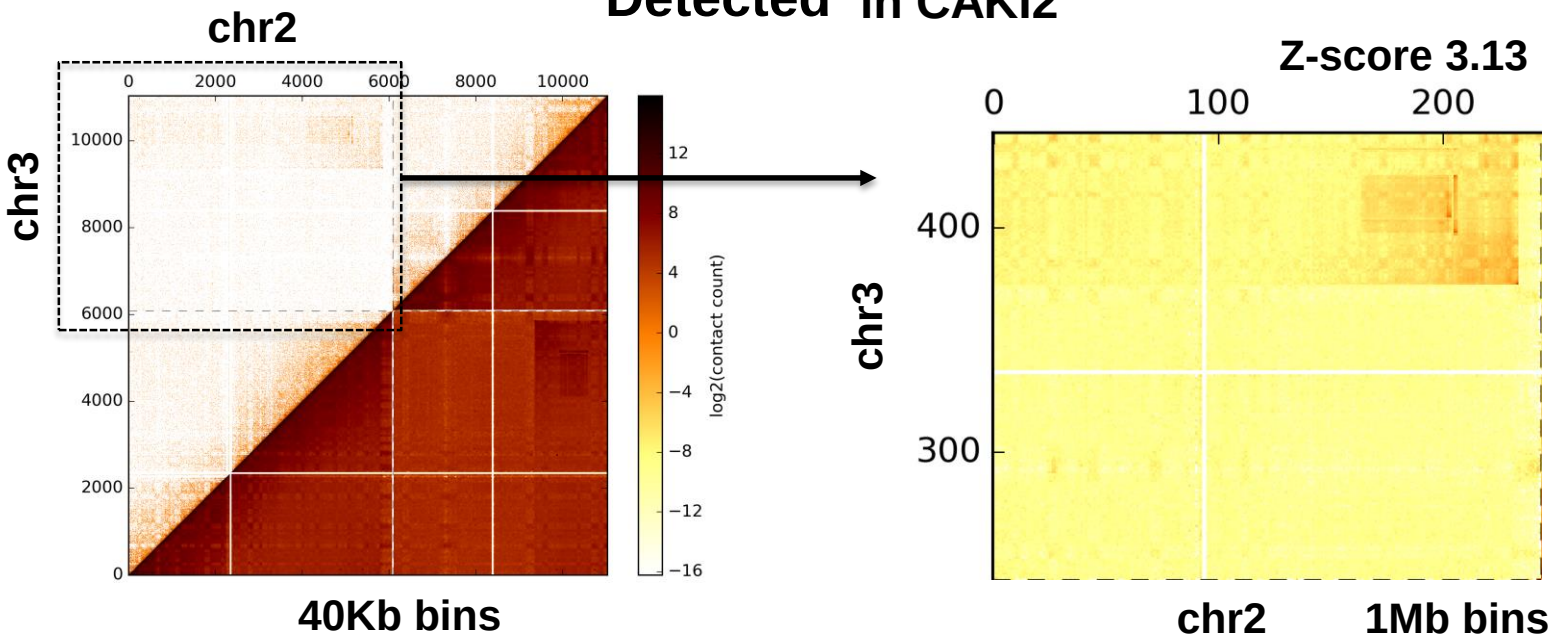



Detected in A549
chr16
chr19
40Kb bins
log2(contact count)
Z-score 3.1
chr19
chr16
1Mb bins
chr11
chr17
40Kb bins
log2(contact count)
Z-score 2.66
chr17
chr11
1Mb bins

CAKI2 top translocations (Z score > 2)
No known translocation

Rank	ChrA	ChrB	Interactions	Zscore
1	chr3	chr12	23	5.8568246
2	chr2	chr12	15	3.4371036
3	chr6	chr12	15	3.4371036
4	chr2	chr3	14	3.1346385
5	chr1	chr12	14	3.1346385
6	chr7	chr12	13	2.8321734
7	chr5	chr12	12	2.5297083
8	chr4	chr12	12	2.5297083
9	chr11	chr12	12	2.5297083

Detected in CAKI2



To-Do List:

- **Examine the missing translocations and identify the reasons to improve the translocation calling.**

- **Improve the clustering step.**

Amplification and Deletion detection

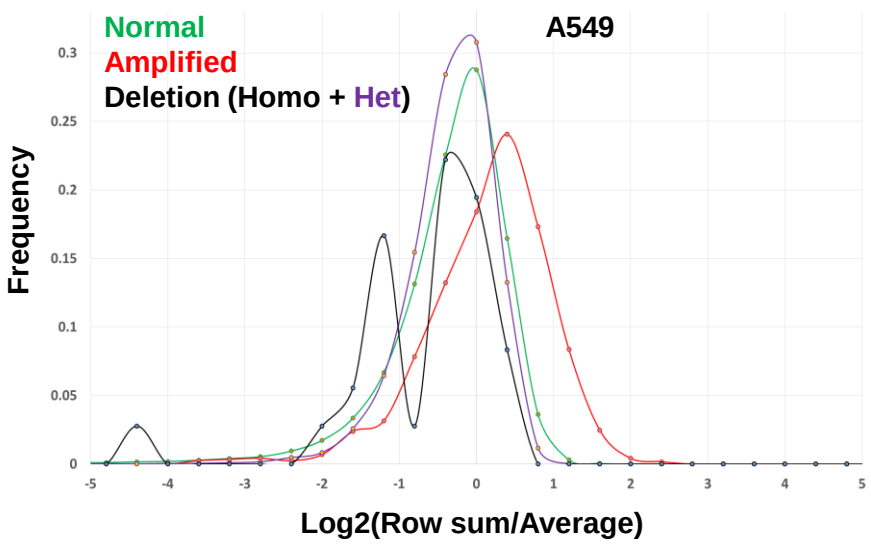
- **Collected the CNV information from HAIB Genotype track for A549, LNCaP, PANC1 and T47D cell lines.**

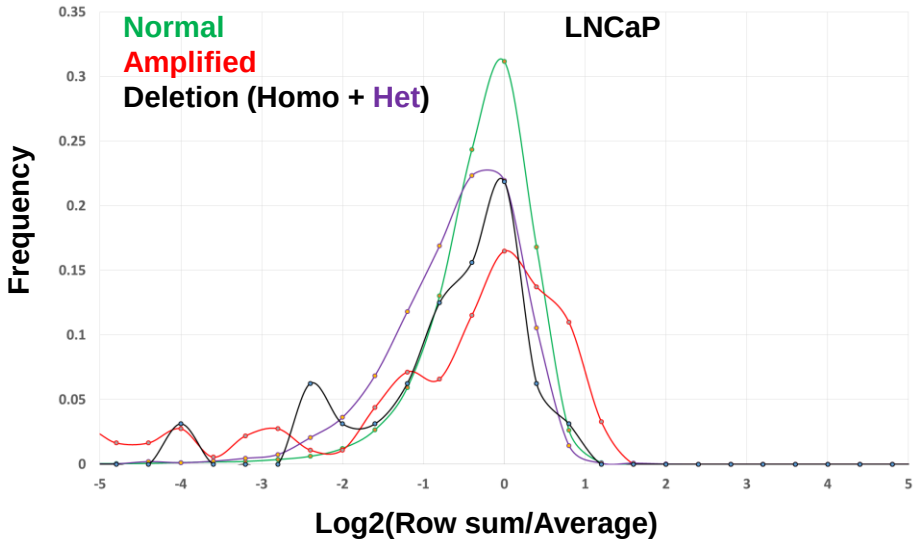
- **Extracted Normal, Amplified and Deletion regions for all the cell lines.**

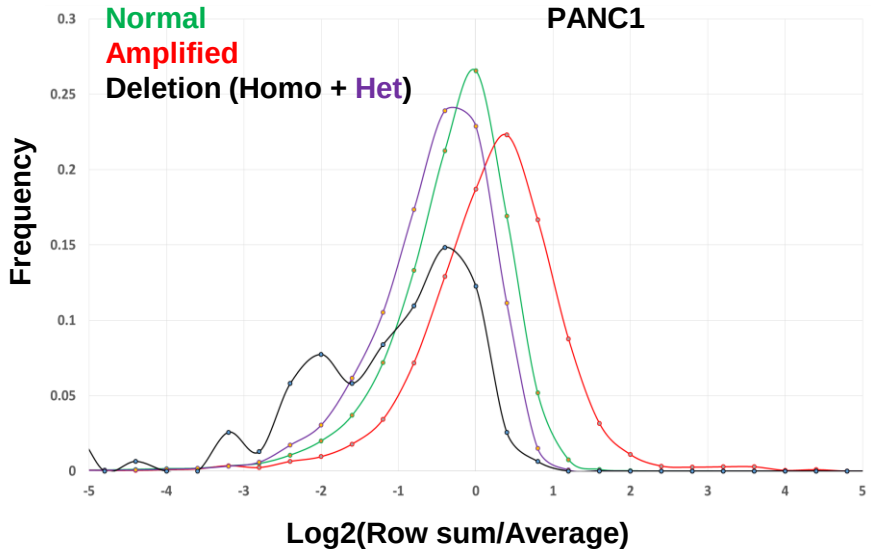
Cell Line	No. of Amp known	
	Total	≥ 40Kb
A549	12	11
LNCaP	55	10
PANC1	16	14
T47D	56	51
Total	139	86

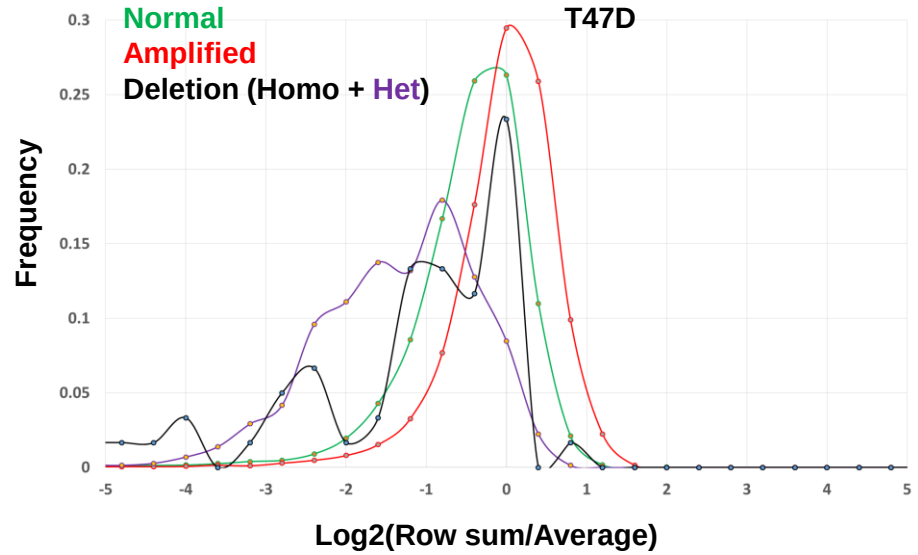
Cell Line	No. of Del known	
	Total	≥ 40Kb
A549	88	20
LNCaP	87	43
PANC1	262	131
T47D	75	43
Total	512	237

- **Amplified regions showed a higher interaction count followed by Normal region and then Deleted regions.**



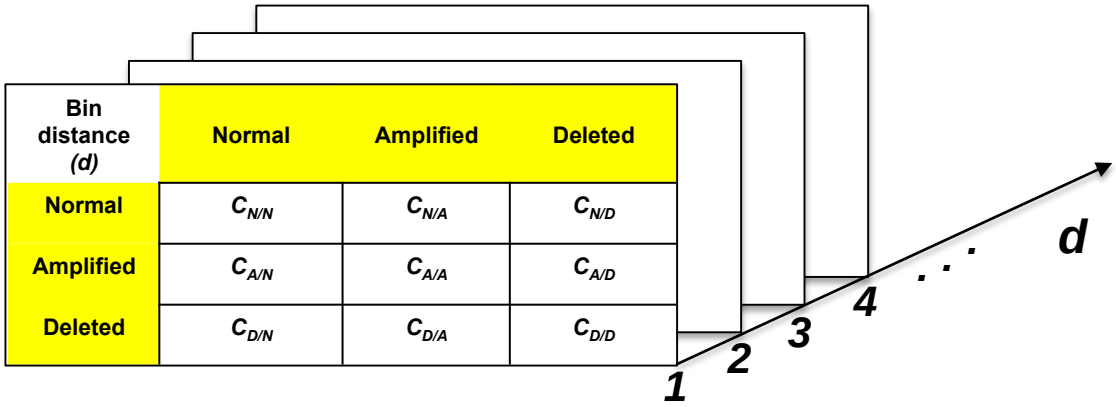


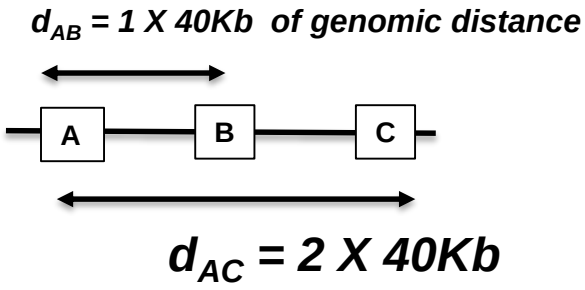




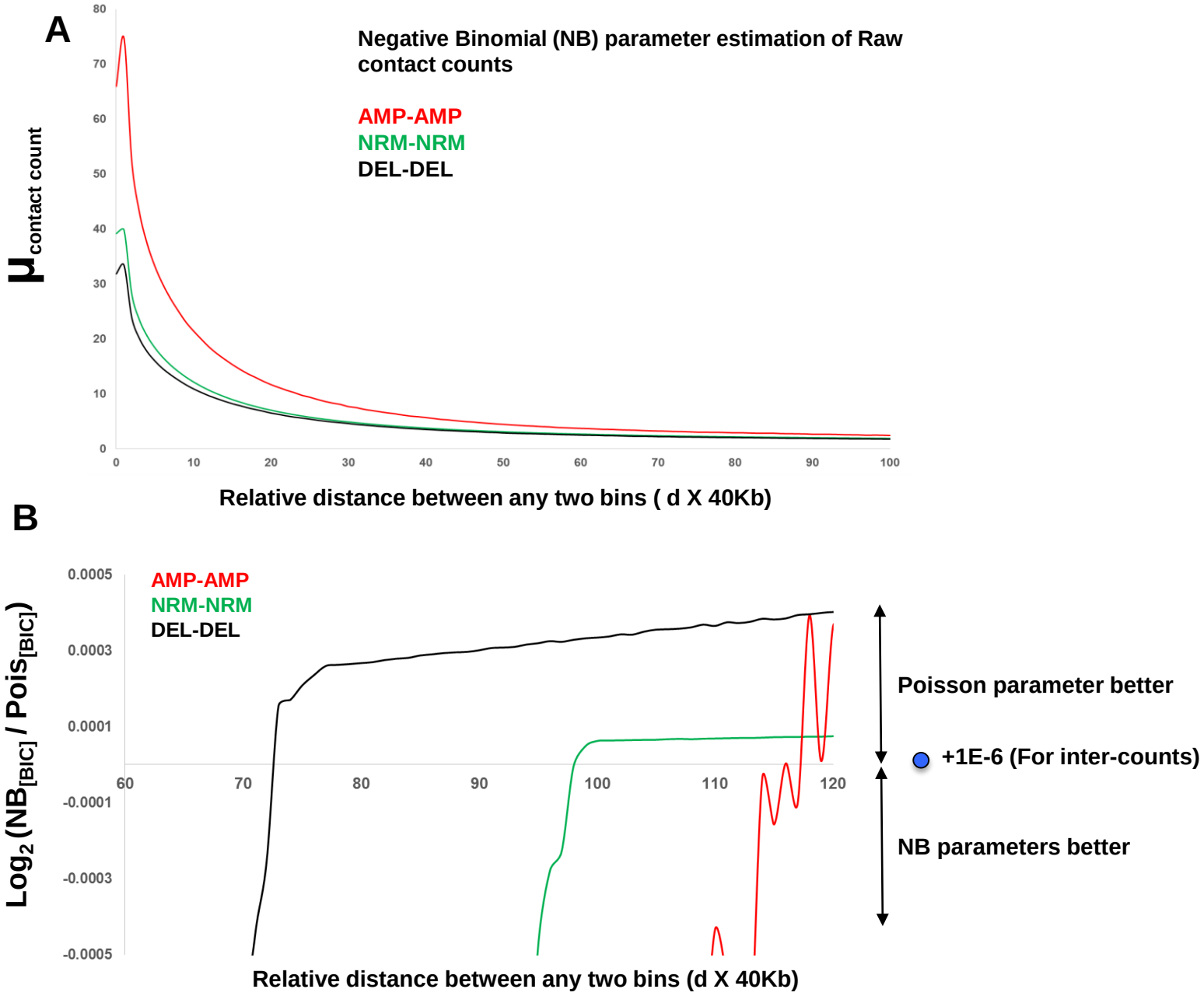
We used the CNV information to setup a HiC matrix simulation pipeline:

- Extracted the contact counts among all bin pairs with the same bin.
- Each bin distance was further categorized into following matrix.





- We used this distance wise contact values to fit distribution and then predict expected counts given a bin distance.
- Fitted the values upto 4000 bin distance (1 bin distance = 40Kb) to both Poisson and Negative-Binomial distribution.
- For each bin distance (1 to 4000) we selected either the negative binomial or the Poisson distribution as the best fit using Bayesian information criteria (BIC).

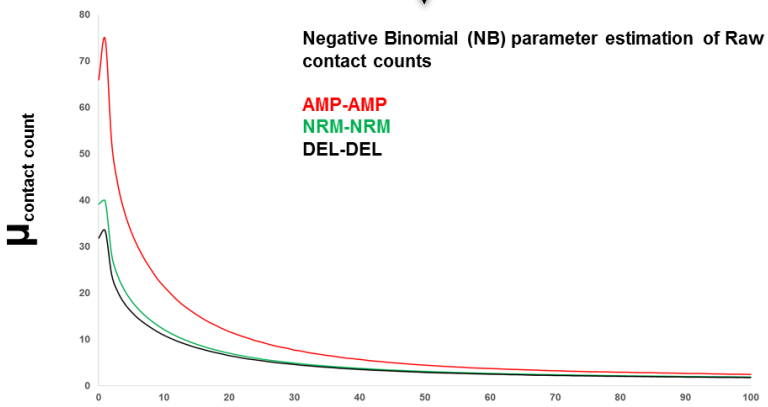


- **The similar analysis was carried for other combinations also.**

Two ways to generate simulated Hi-C matrices with or without CNVs

Bin distance (d)	Normal	Amplified	Deleted
Normal	$C_{N/N}$	$C_{N/A}$	$C_{N/D}$
Amplified	$C_{A/N}$	$C_{A/A}$	$C_{A/D}$
Deleted	$C_{D/N}$	$C_{D/A}$	$C_{D/D}$

1 2 3 4 . d



Ratio based simulation

In this case, we assume the existing HiC-bin as normal and multiply the observed count with the CNV ratio to which we like to convert.

Random simulation

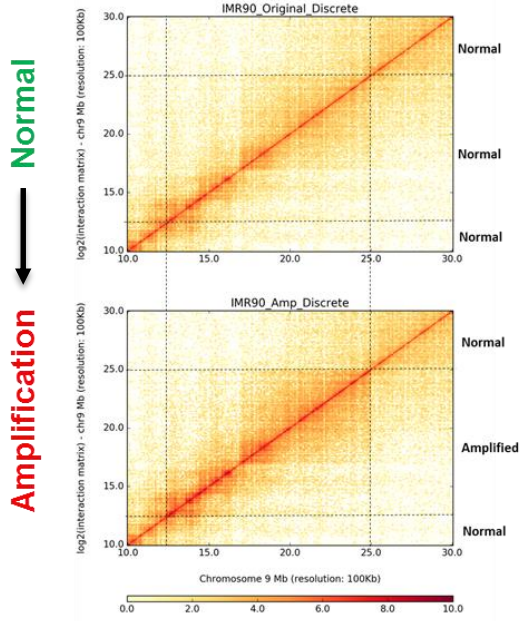
Here, we use the existing HiC bin connection information and assign a random raw count (from NB/Pois distribution) depending on the *d* value of a pair.

Examples simulated HiC matrix with Amplification

- **Ratio based 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 $(\mu_{D/D}/\mu_{N/N})_{d5}$

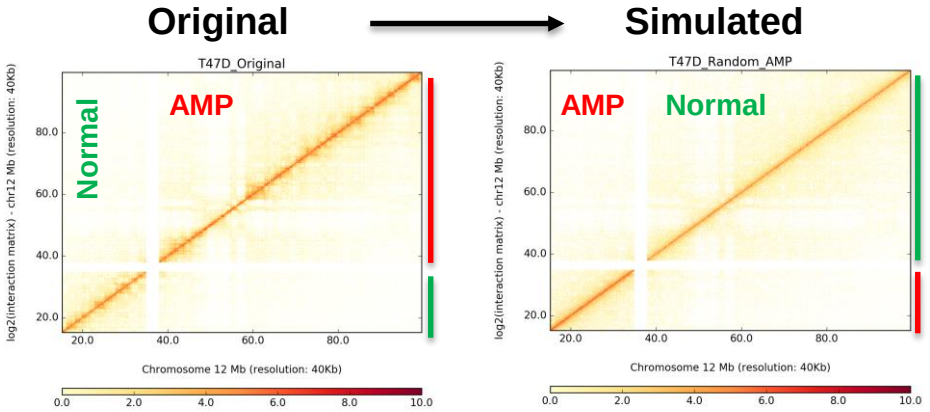
Bin connection same as that of original matrix*



- **Random simulation:**

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

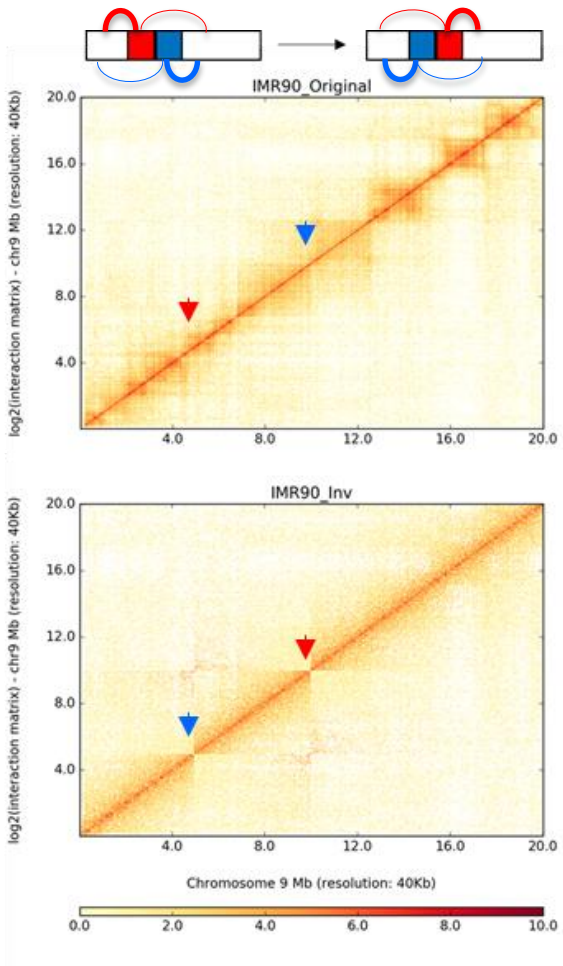
*



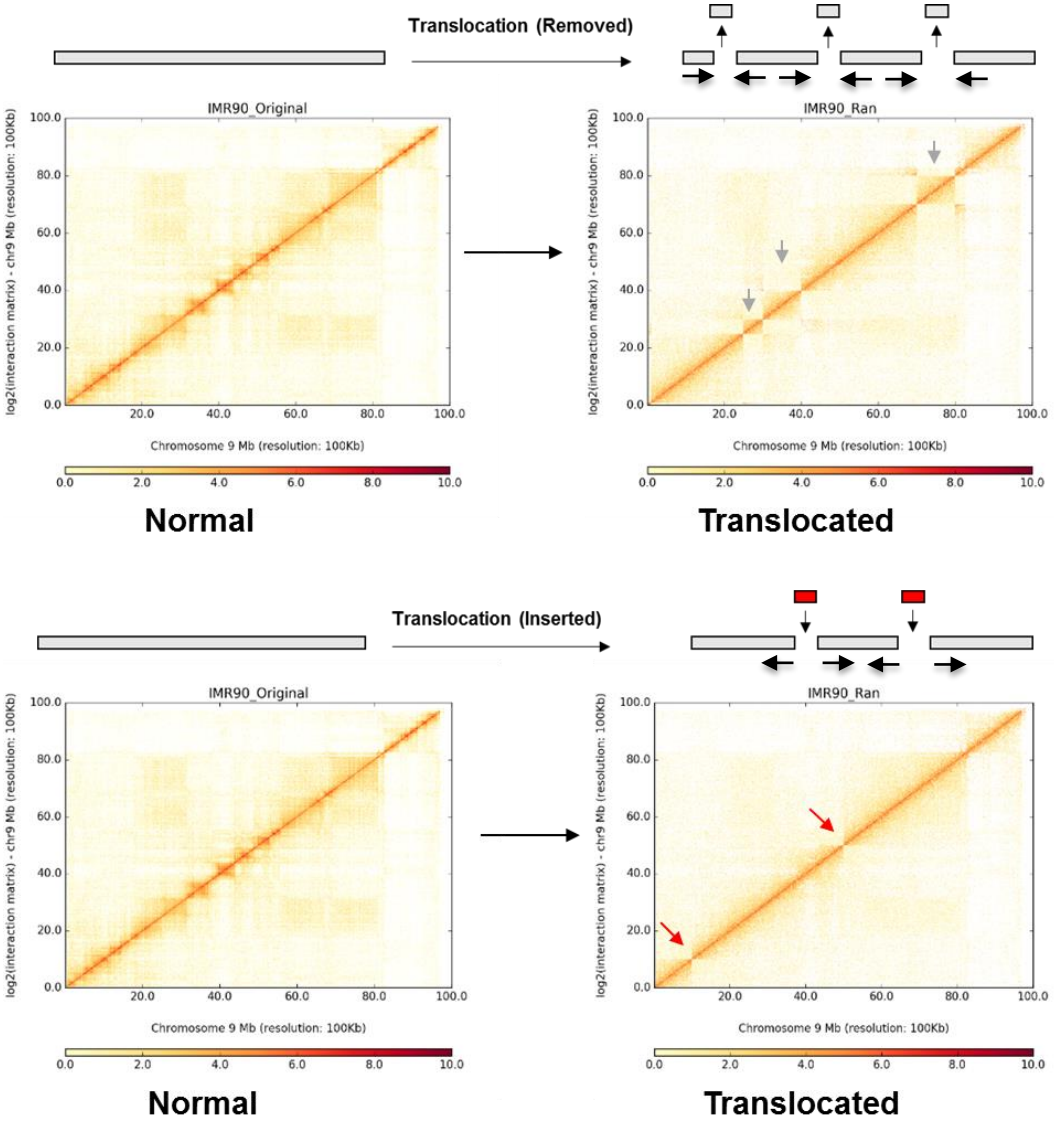
- **Using this pipeline we can insert and simulate any combination CNV incorporated HiC matrix.**

Examples simulated HiC matrix with Amplification with

1. Inversion



2. Translocation



Amplification and Deletion calling pipeline

CNV calling pipeline

Generate Randomly Simulated Normal HiC matrix

Raw HiC intra-chromosomal matrix

Explicit normalization using HiCNorm

Calculate the sum of contact count of a bin_i within its +2/-2 neighboring bins

Generate a 1D vector using the count values

Perform Loess smoothing on the vector

Normalized 1D Simulated vector

Normalize the 1D vector

S = Subtract 1D vectors (original-simulated)

Perform Binary segmentation on the "S"

Extract exact region of $segment^S_i$ values from simulated vector, $segment^R_i$

Extract $segment^S_i$ values for "S"

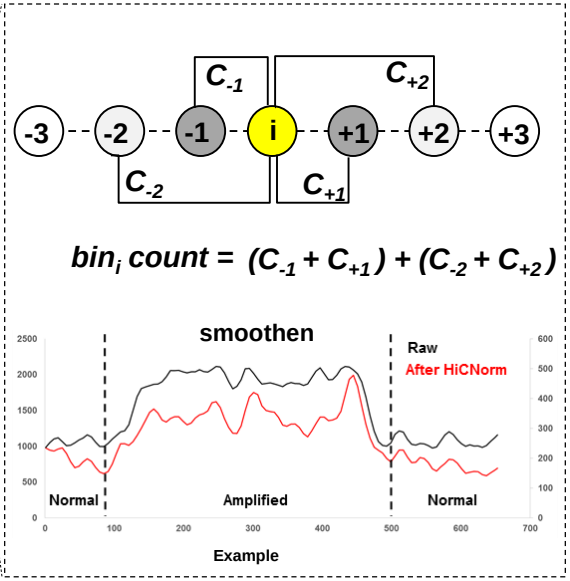
Perform Wilcoxon test to determine S_i has higher or lower population mean rank than $segment^R_i$

Adjusted p-value < 0.05

If high and significant then its *AMPLIFIED*

If low and significant then its *DELETED*

This is how it looks



Overview of the CNV detection result:

Each segment is associated with a p-value. The results shows predictions for all the chromosomes (ranked by adjusted p-value < 0.05) in each cell line and the number of CNV found

Cell Line	No. of Amp known		Detected (p.adj < 0.05)	
	Total	>= 40Kb	Total	>= 40Kb
T47D	56	51	39	37
LNCaP	55	10	14	7
PANC1	16	14	11	11
A549	12	11	8	7

Cell Line	No. of Del known		Detected (p.adj < 0.05)	
	Total	>= 40Kb	Total	>= 40Kb
T47D	75	20	67	19
LNCaP	87	43	74	37
PANC1	262	131	230	120
A549	88	43	71	37

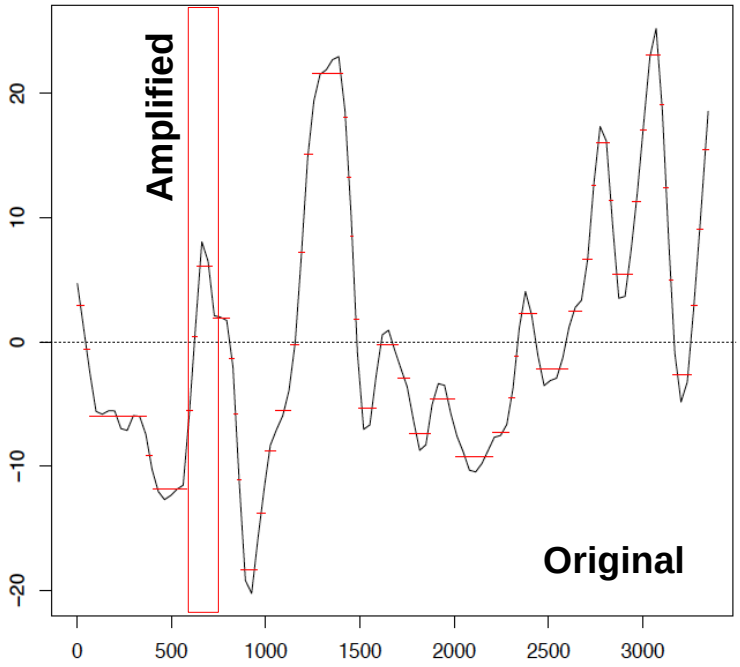
Top	Amplification Recall (Total)			
	T47D	LNCaP	PANC1	A549
25	0.308	0.143	0.273	0.500
50	0.667	0.214	0.636	0.625
100	0.872	0.714	0.909	0.875

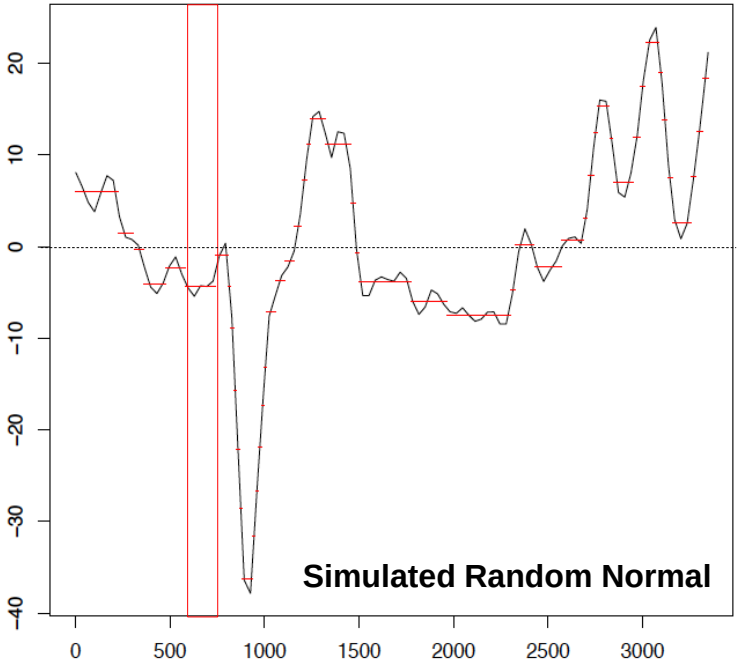
Top	Deletion Recall (Total)			
	T47D	LNCaP	PANC1	A549
25	0.075	0.095	0.183	0.197
50	0.209	0.243	0.257	0.352
100	0.478	0.405	0.443	0.648

Top	Amplification Recall (>= 40Kb)			
	T47D	LNCaP	PANC1	A549
25	0.297	0.857	0.273	0.500
50	0.676	0.857	0.636	0.625
100	0.892	1.000	0.909	0.875

Top	Deletion Recall (>= 40Kb)			
	T47D	LNCaP	PANC1	A549
25	0.211	0.054	0.242	0.351
50	0.368	0.351	0.333	0.514
100	0.579	0.486	0.542	0.784

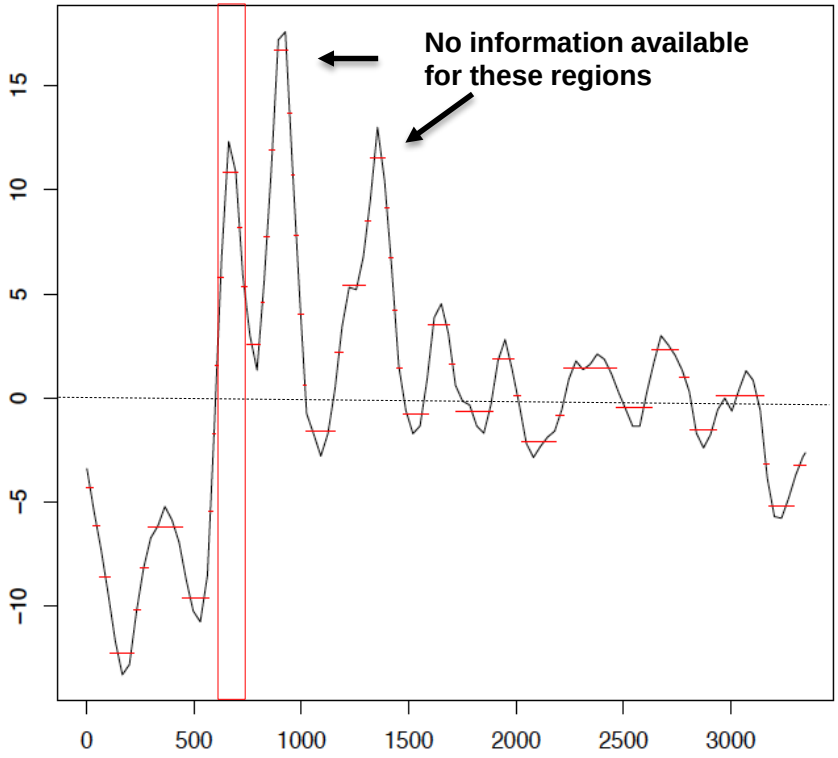
- **Lower recall value in deletion identification is due to HiC normalization process.**
- **Bins that are GC content < 0.2 and mappability < 0.5 are not normalized and thus after normalization, they are assigned as zero counts, leading to lower sum of contact counts. We will apply filtering process to remove those counts.**





PANC1 (Chr12)

Subtracted 1D vector (original-simulated)



Thank you