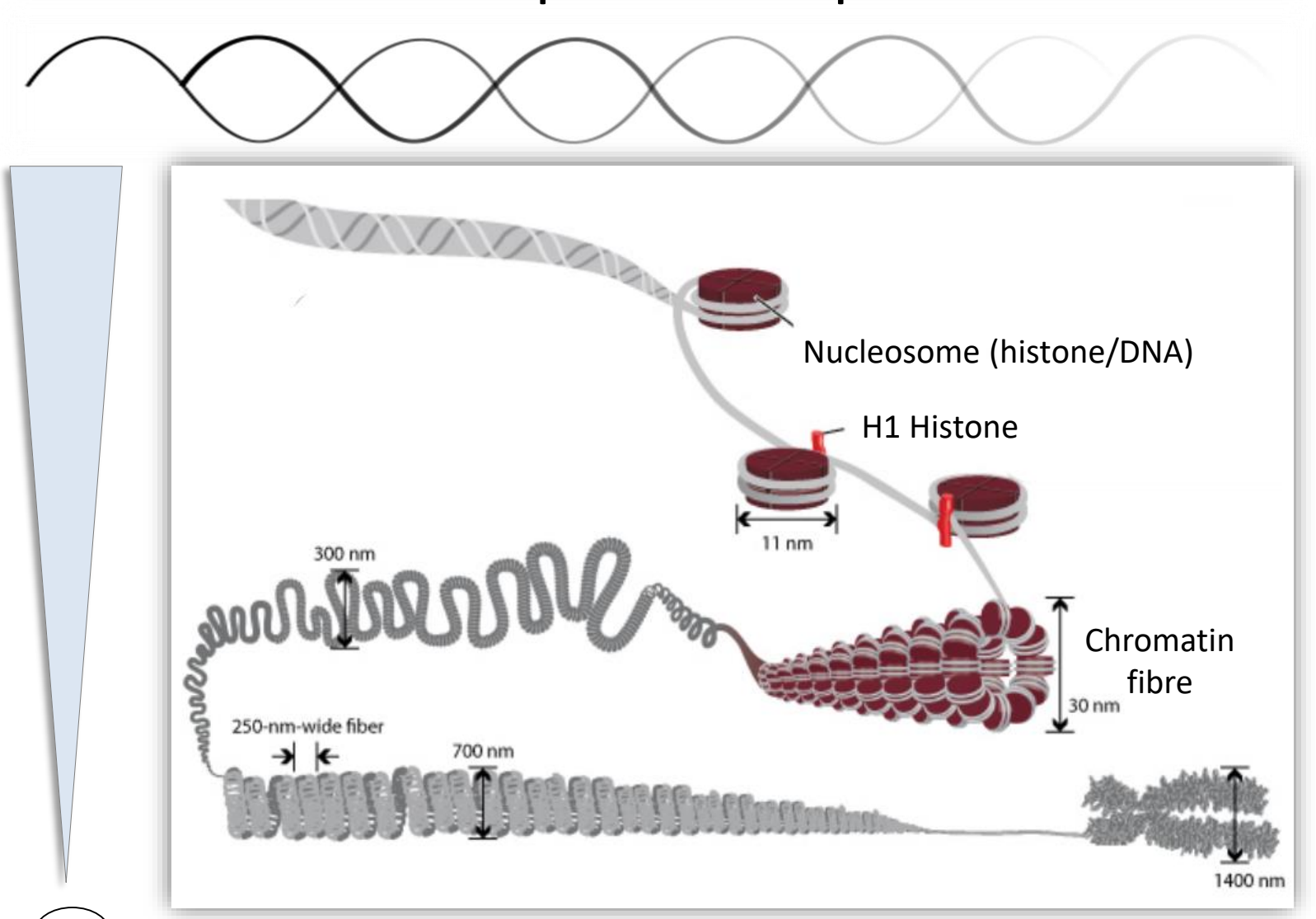
Analysis of chromosome conformation data and application to cancer

Nicolas Servant
Institut Curie, INSERM U900, Mines ParisTech, PSL-Research University

11th of April 2018
Math4Genomics, Evry

Spatial organization of the genome

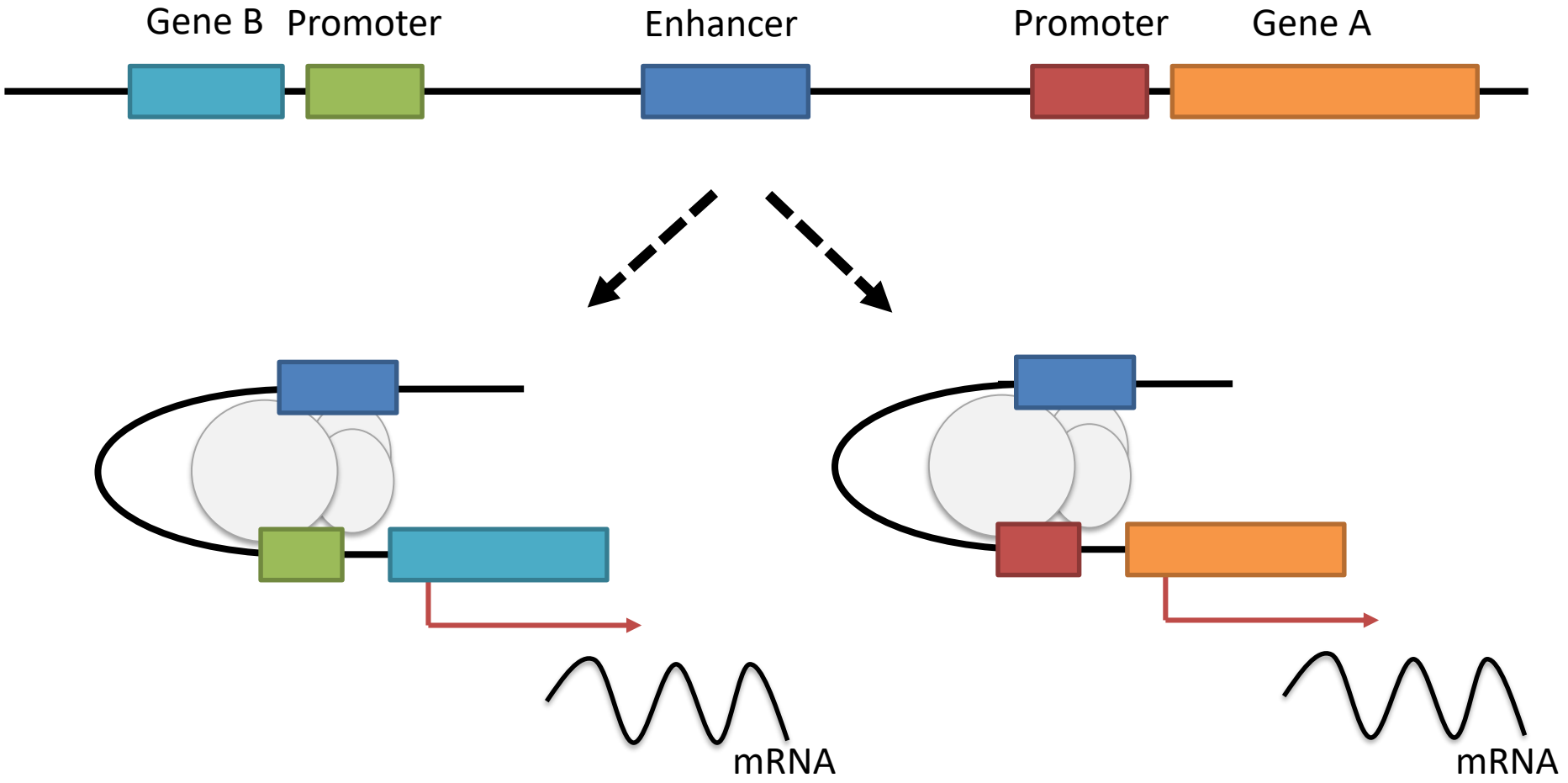
How are 2 meters of DNA packed into a 10µm diameter nucleus ?



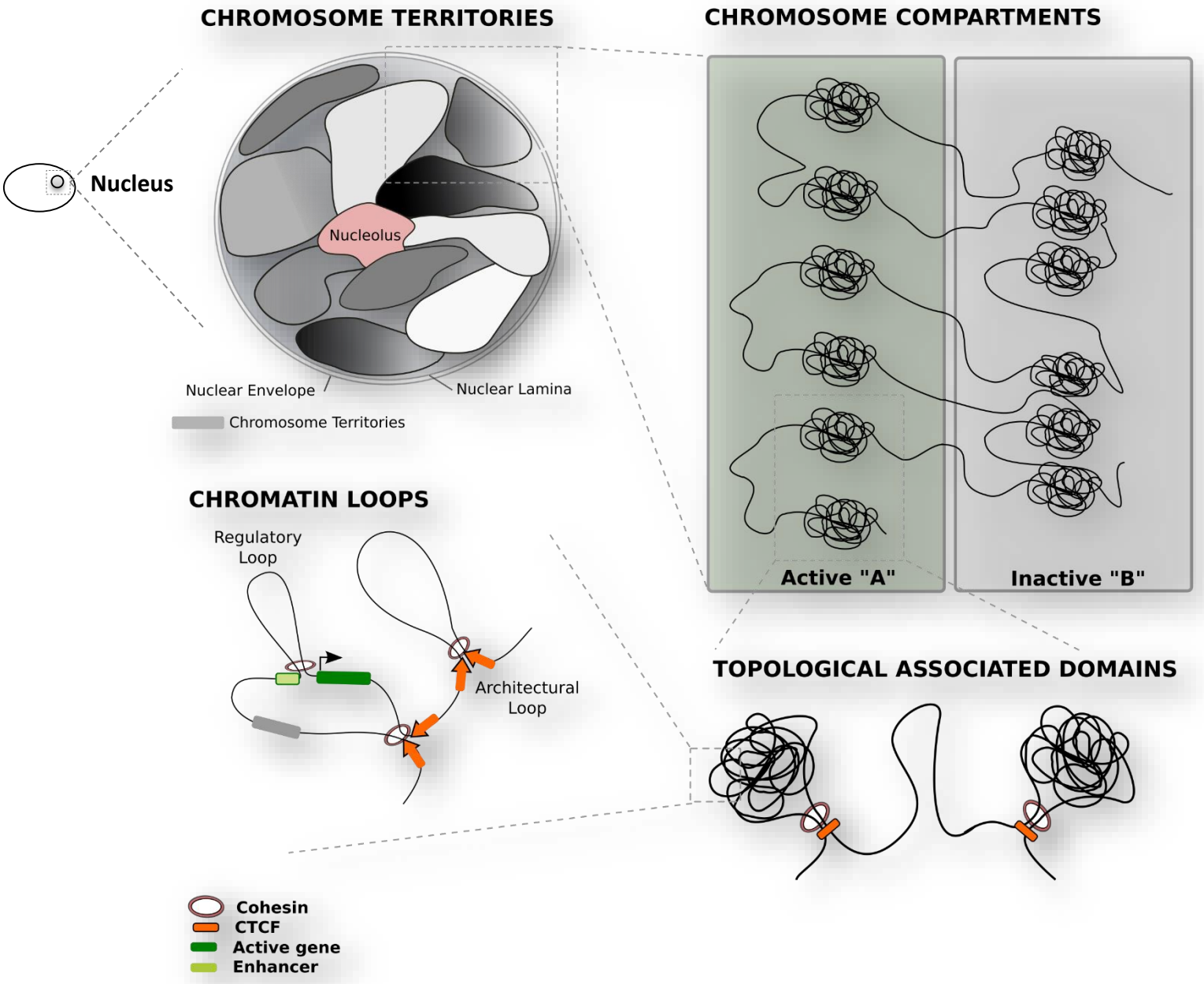
Adapted from Annunziato et al. 2008

Spatial organization and regulation

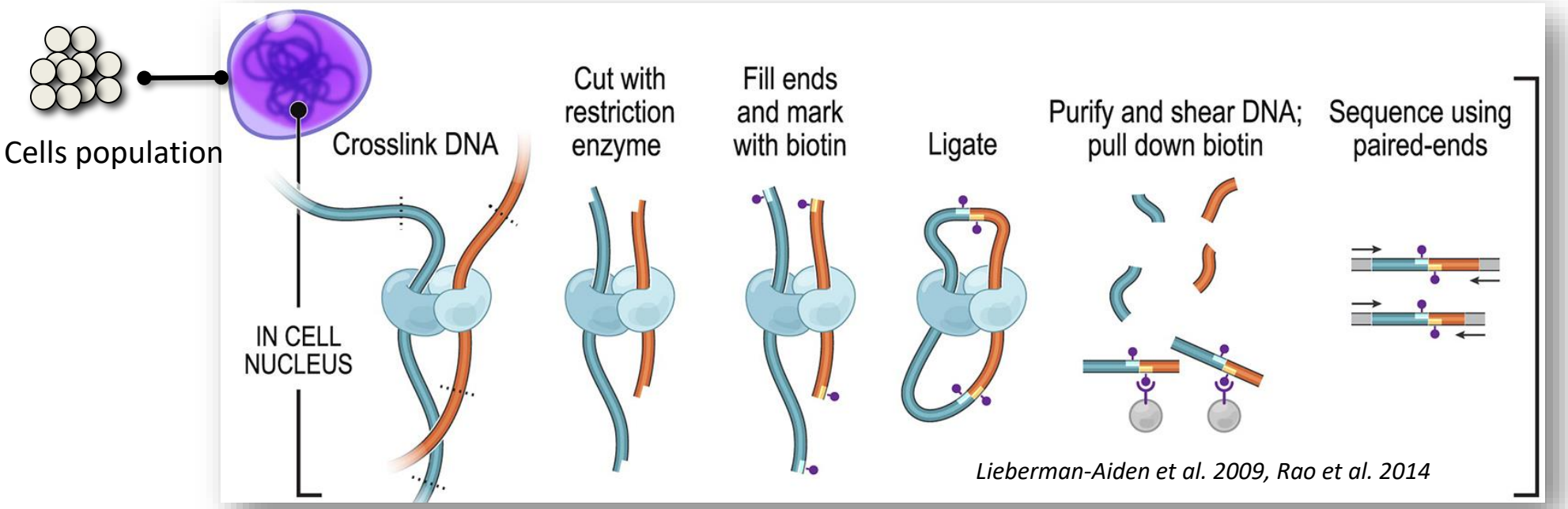
How does spatial organization influence gene regulation?

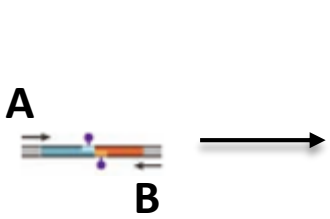


Different levels of spatial organization



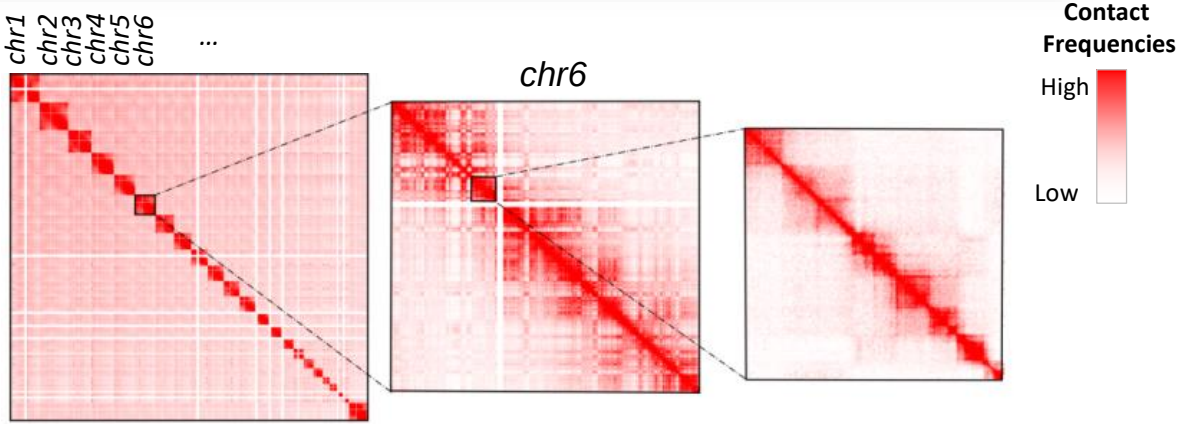
Hi-C captures the chromatin conformation within the nucleus





	A	B	C	
A		+1		
B	+1			
C				

Bin size



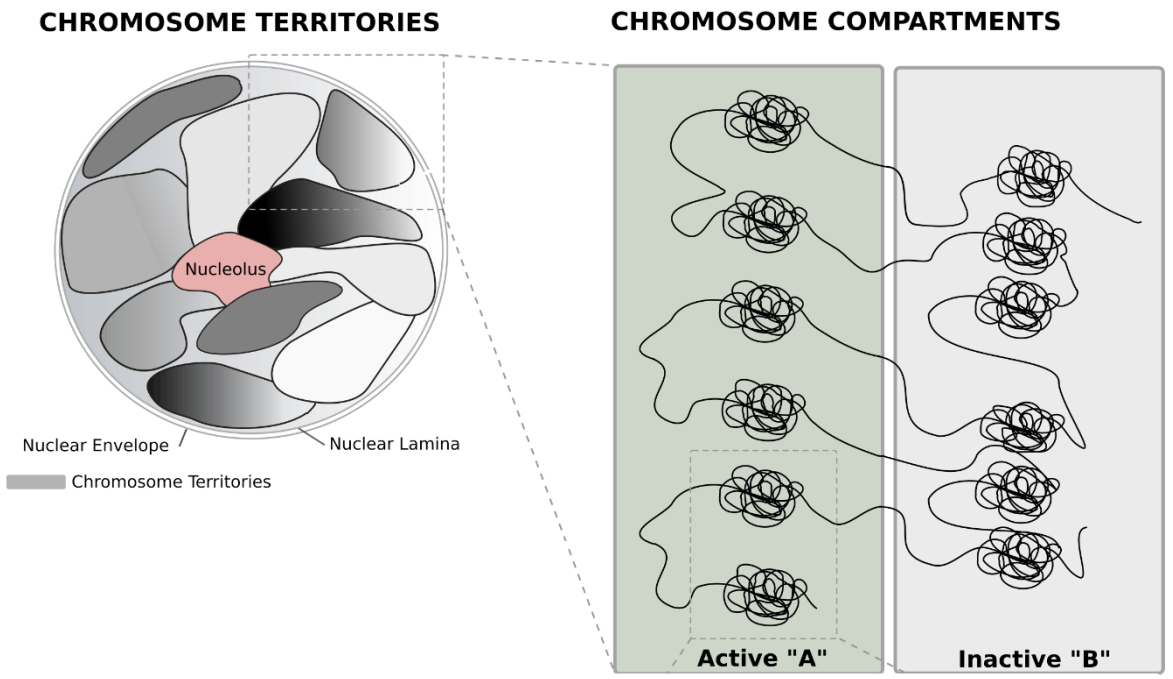
Whole genome map

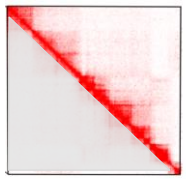
Intrachromosomal contact maps **at different resolutions**

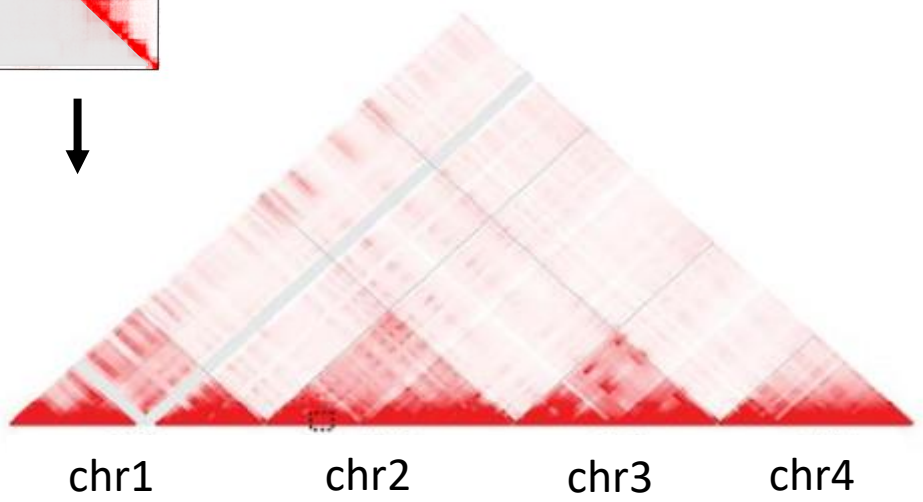
'Hi-C'-based experiments

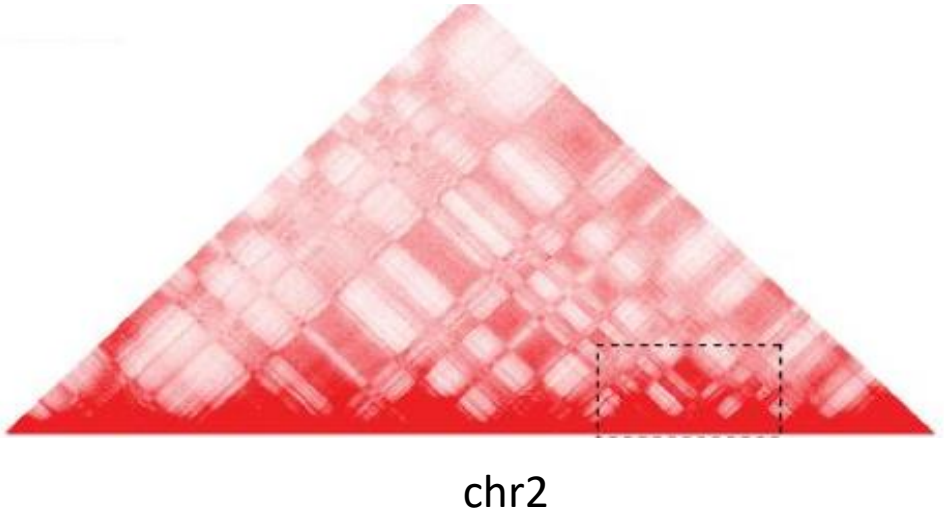
Method	Main features	References
Hi-C	For mapping whole-genome chromatin interaction in a cell population; proximity ligation is carried out in a large volume	*Lieberman-Aiden et al. (2009)*
TCC	Similar to Hi-C, except that proximity ligation is carried out on a solid phase-immobilized proteins	*Kalhor et al. (2011)*
Single-cell Hi-C	For mapping chromatin interactions at the single-cell level	*Nagano et al. (2013)*
In situ Hi-C	Proximity ligation is carried out in the intact nucleus	*Rao et al. (2014)*
Capture-C	Combines 3C with a DNA capture technology ; equivalent to high-throughput 4C	*Hughes et al. (2014)*
Dnase Hi-C	Chromatin is fragmented with Dnase I; proximity ligation is carried out on a solid gel	*Ma et al. (2015)*
Targeted Dnase Hi-C	Combine Dnase or in situ Dnase Hi-C with a capture technology	*Ma et al. (2015)*
Micro-C	Chromatin is fragmented with micrococcal nuclease	*Hsiech et al. (2015)*
In situ DNAse Hi-C	Chromatin is fragmented with Dnase I; proximity logation is carried out in the intact nucleus	*Deng et al. (2015)*
Capture-Hi-C	Combines 3C with a DNA capture technology ; equivalent to high-throughput 5C	*Mifsud et al. (2015)*
HiChIP	Detecting genome-wide chromatin interaction mediated by a particular protein ; equivalent to ChAI-PET	*Mumbach et al. (2016)*

Genome organization and Hi-C



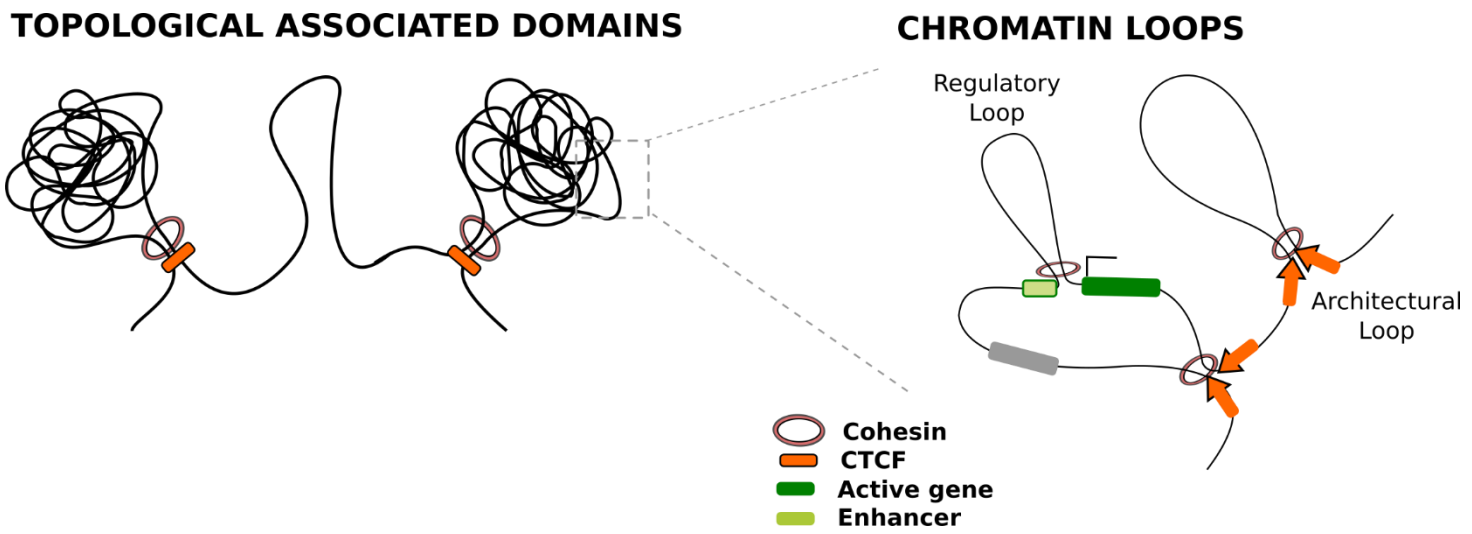


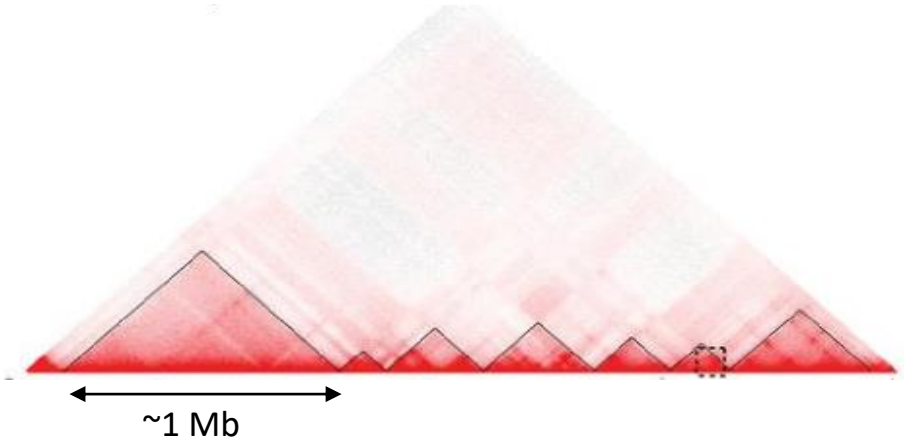


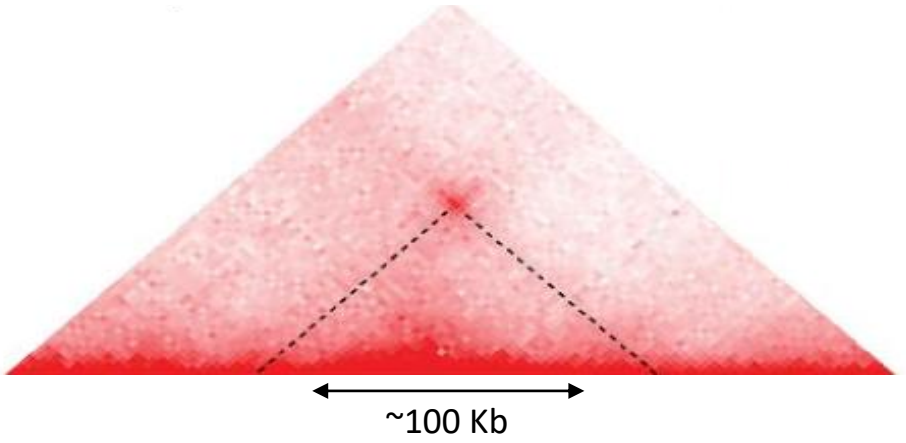


Lieberman-Aiden et al. 2009, Dixon et al. 2012, Rao et al. 2014

Genome organization and Hi-C



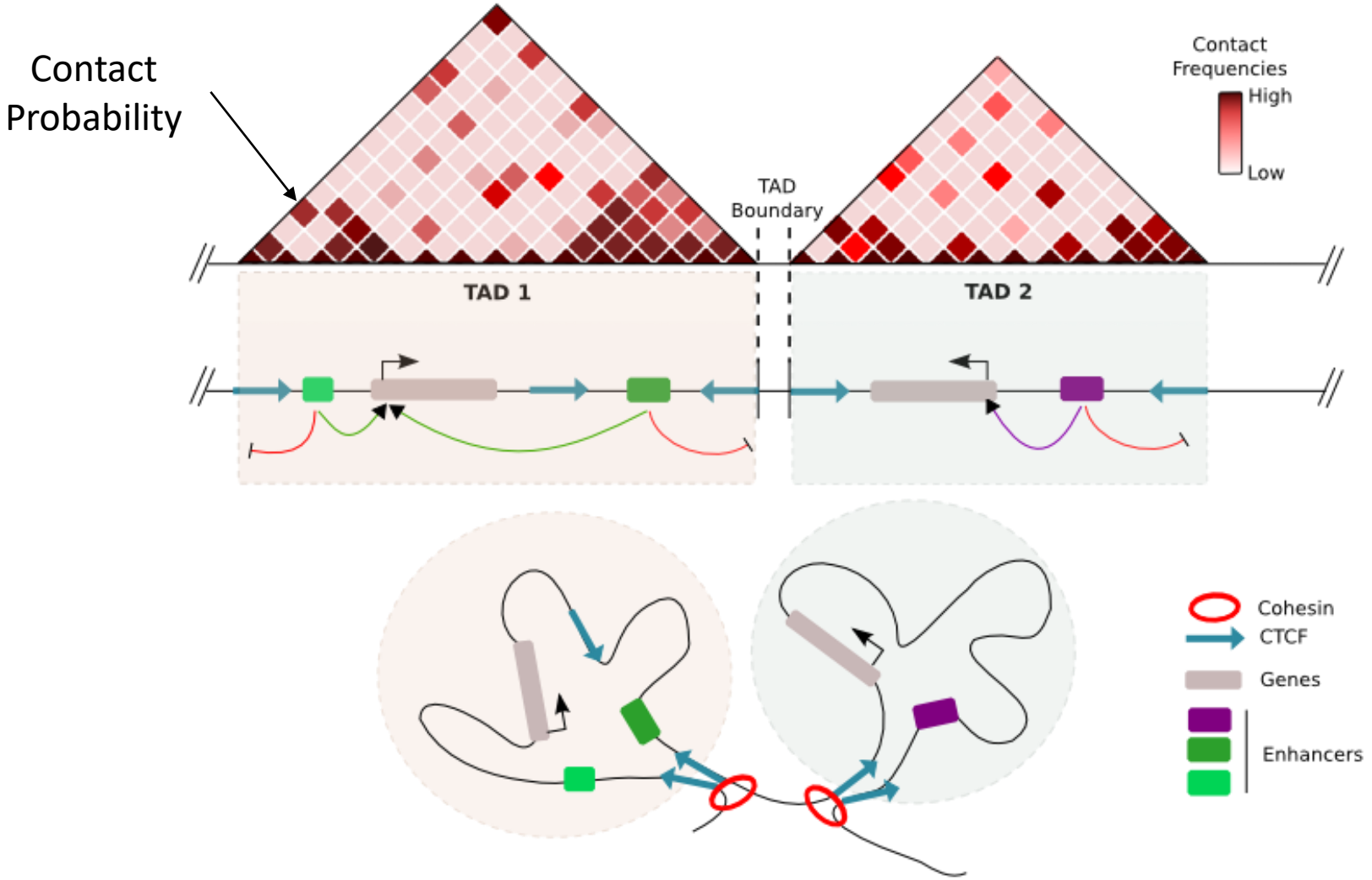




Lieberman-Aiden et al. 2009, Dixon et al. 2012, Rao et al. 2014

Topological Associated domains (TADs)

The topological domains (TADs) have been described as the functional units of the genome organization, able to promote enhancer/promoter interactions.



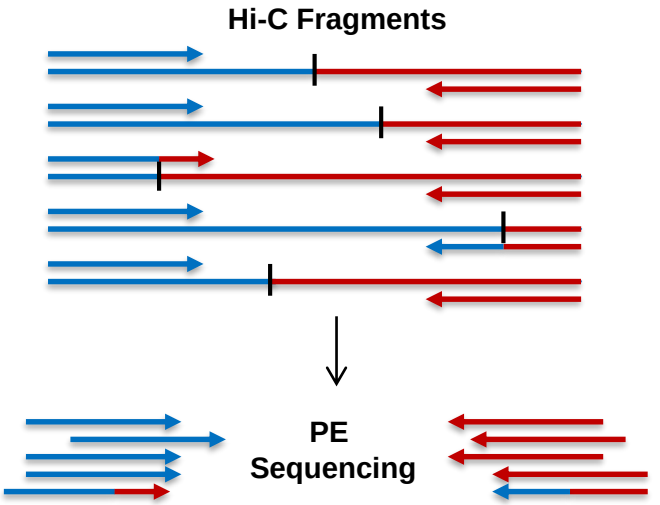
Questions ?

1. How to efficiently process Hi-C data?

2. Are there any specific computational challenges in analyzing Hi-C data from cancer samples ?

What does Hi-C data look like ?

Illumina paired-end sequencing

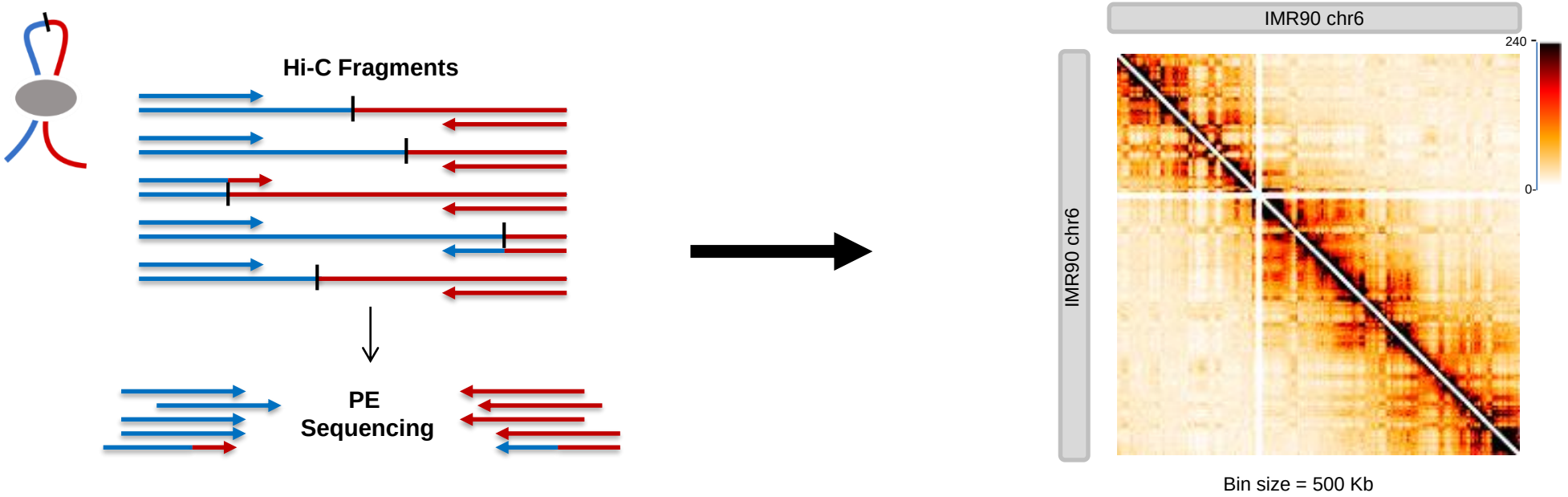


A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin I

Suhas S.P. Rao,[1,2,3,4,10] Miriam H. [illegible] Elena K. Stamenova,[1,2,3,4] Ivan D. Bochkov,[1,2,3] James [illegible] Ido Machol,[1,2,3] Arina D. Omer,[1,2,3] Eric S. Lander,[4,7,8,*] and Erez [illegible]

9 cell lines **242** Hi-C libraries
25 202 711 604 sequenced reads total
>1 500 000 reads per cell line in average

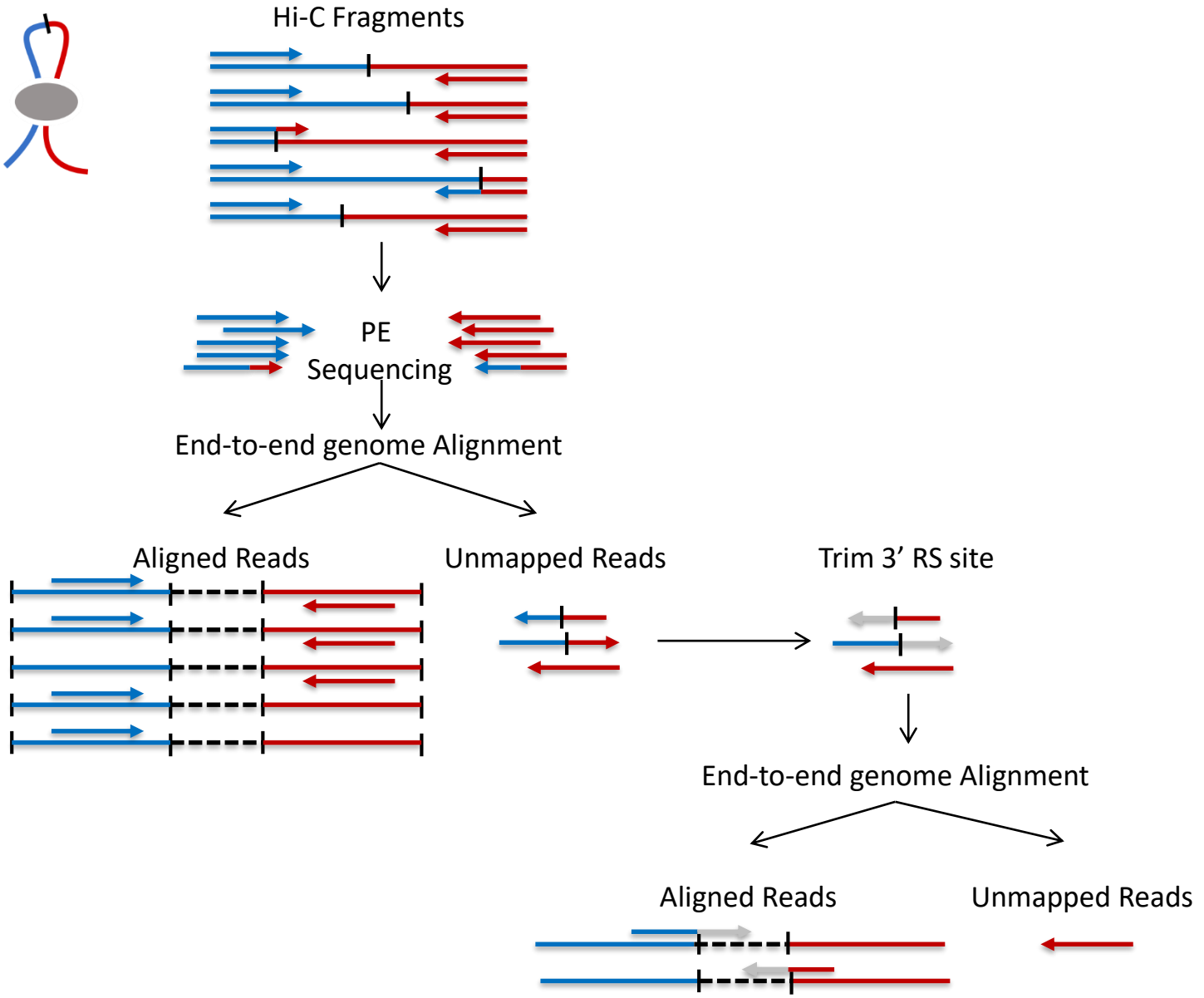
Challenges in Hi-C data processing



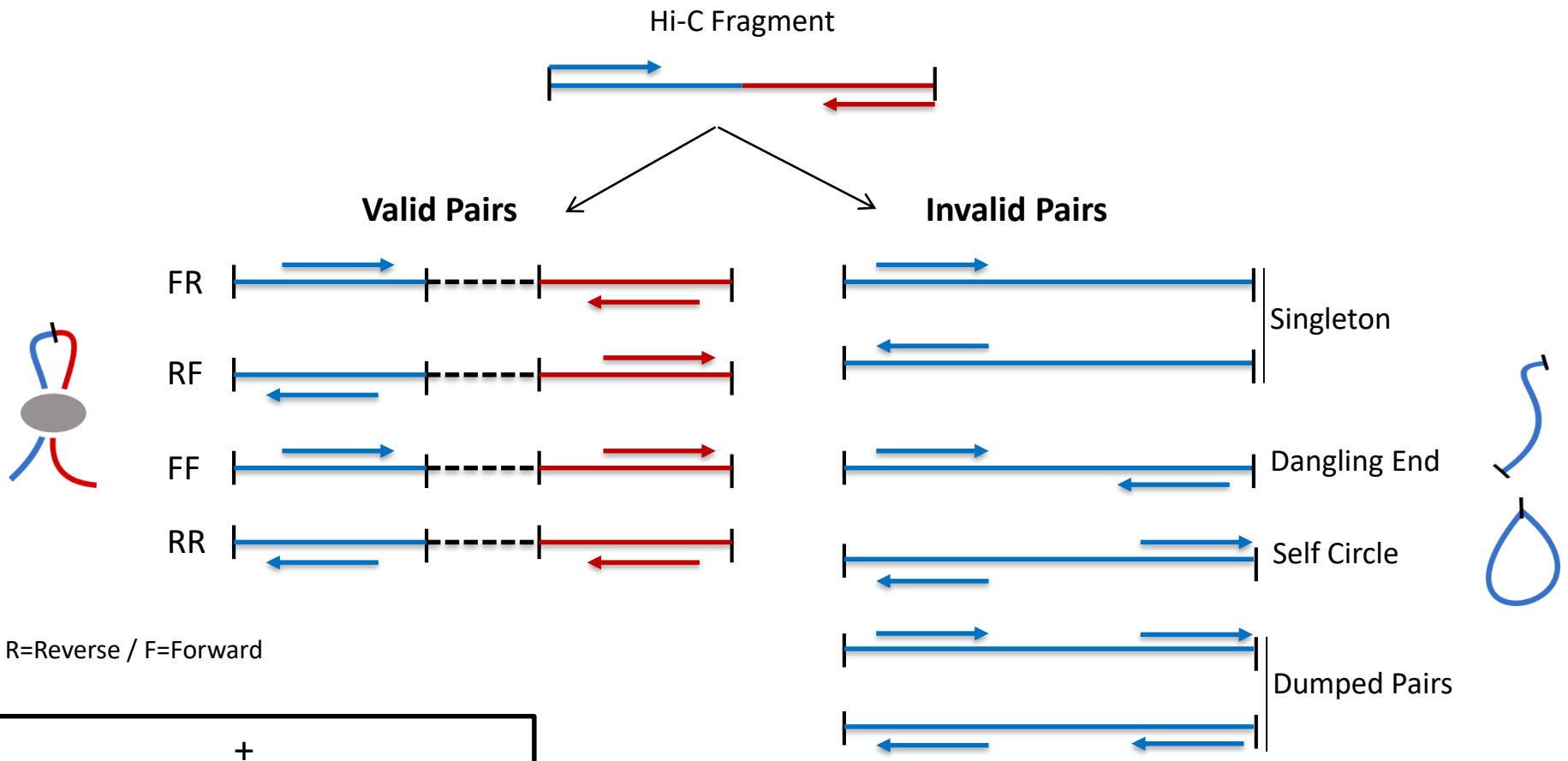
How to process Hi-C data in an **easy** and **efficient** way taking into account ;

- The huge amount of data
- The evolution of protocols
- The computational ressources

A dedicated mapping strategy



Detection of valid interaction products



+

Filtering on :

- Insert size
- Restriction fragment size
- MAPQ
- etc.

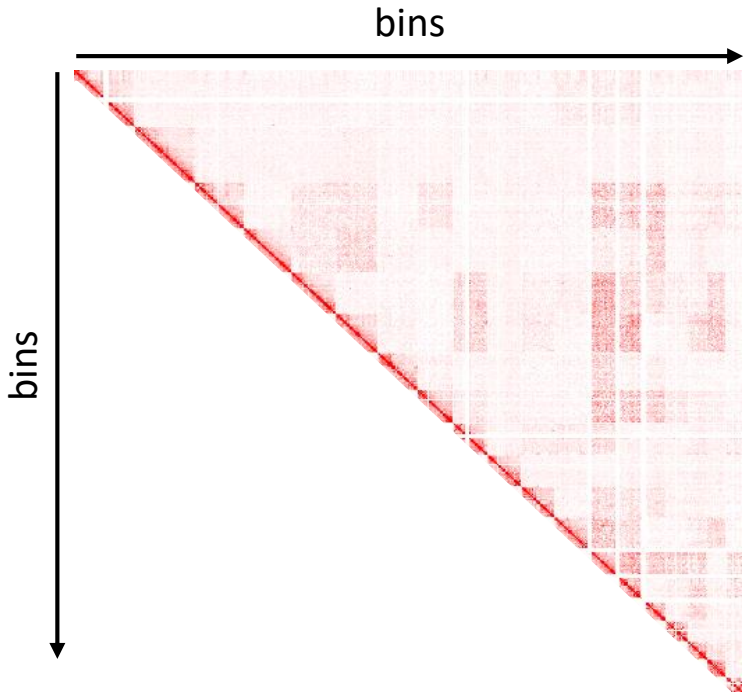
Building contact maps

There is currently no consensus about how to (efficiently) store the contact maps

A Hi-C contact map is :

- Usually very **sparse**
- **Symmetric**

We therefore propose to use a standard triplet sparse format to store only half of the non-zero contact values.



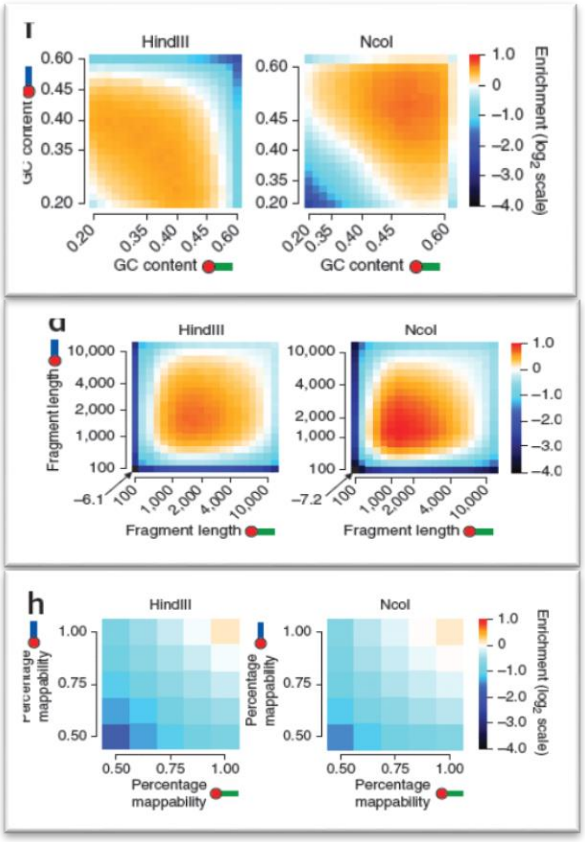
	Dense (MB)	Sparse Complete (MB)	Sparse Symmetric (MB)
1M	25	98	49
500Kb	77	363	182
150Kb	818	1 900	934
40Kb	12 000	3 800	1 900
20Kb	45 000	5 300	2 700
5Kb	>100 000 ??	8 600	4 300

Hi-C data normalization

All high-througthut techniques are subject to **technical and experimental biases**

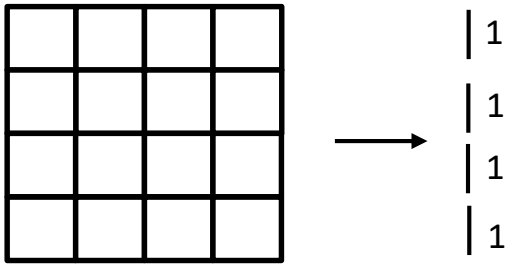
The iterative correction (ICE) method is **a widely used** approach for Hi-C data normalization.

This method is based on the assumption that **each locus should have the same probability of interaction genome-wide**, and is in theory able to correct for **any bias** in the contact maps.



Iterative correction of Hi-C data reveals hallmarks of chromosome organization

Maxim Imakaev[1,5], Geoffrey Fudenberg[2,5], Rachel Patton McCord[3], Natalia Naumova[3], Anton Goloborodko[1], Bryan R Lajoie[3], Job Dekker[3] & Leonid A Mirny[1,2,4]



as $\Sigma_{i,\, i\neq j,\, j\pm 1} T_{ij} = 1$ for each region j.

CPUs time and optimization

	Hiclib (*Imakaev et al.* 2012)	**HiC-Pro**		
	IMR90 GSE35156	IMR90 GSE35156	IMR90 GSE35156	IMR90_CCL186 GSE63525
#Read pairs	397 200 000	397 200 000	397 200 000	1 535 222 082
#Input Files	10	10	84	160
#Jobs in parallel	1	1	42	80
#CPU per Job	8	8	4	4
Max Memory (RAM) per Job	10 Gb	7 Gb	7 Gb	7 Gb
-- Mapping	22:03	12:53	00:21	05:56
-- Filtering	00:30	03:20	00:04	00:36
-- Merge multiple Inputs and remove duplicates		00:13	00:13	00:42
-- Contact maps builder	01:45	00:15	00:15	00:42
-- ICE normalization	04:06	01:15	01:15	03:49
Wall Time	**28:24**	**17:56**	**02:08**	**11:41**

HiC-Pro availibility

Servant *et al. Genome Biology* (2015) 16:259
DOI 10.1186/s13059-015-0831-x

SOFTWARE Open Access

HiC-Pro: an optimized and flexible pipeline for Hi-C data processing

Nicolas Servant[1,2,3*], Nelle Varoquaux[1,2,3], Bryan R. Lajoie[4], Eric Viara[5], Chong-Jian Chen[1,2,3,6,7,8], Jean-Philippe Vert[1,2,3], Edith Heard[1,6,7], Job Dekker[9] and Emmanuel Barillot[1,2,3]

- Fast, and simple to use
- Complete (from raw data to normalized contact maps)
- Open to the community

Available at https://github.com/nservant/HiC-Pro

Forum and discussion at https://groups.google.com/forum/#!forum/hic-pro

Two years later, HiC-pro ...

- is a **collaborative** project with contribution of several users
- is currently cited among the 3 most popular pipelines for Hi-C data processing with more than 70 citations
- is the only tool allowing **allele-specific Hi-C analysis** in an integrative manner
- supports **all** Hi-C based protocols (dilution Hi-C, in situ Hi-C, DNase Hi-C, Micro-C, Capture-C, Capture-Hi-C, HiChIP, etc...)
- is dedicated to Hi-C data processing but is now compatible with many downstream analysis software
- is still in active development !

Questions ?

1. How to efficiently process Hi-C data?

2. Are there any specific computational challenges in analyzing Hi-C data from cancer samples ?

Hi-C on cancer data

So far, most of the studies were dedicated to normal cell … and a few ones started to investigate chromatin structure of Breast and Prostate cancer using Hi-C

Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation

François Le Dily,[1,2,3] Davide Baù,[1,3] Andy Pohl,[1,2] Guillermo P. Vicent,[1,2] Françoi
Daniel Soronellas,[1,2] Giancarlo Castellano,[1,2,4] Roni H.G. Wright,[1,2] Cecilia Balla
Guillaume Filion,[1,2] Marc A. Marti-Renom,[1,3,5] and Miguel Beato[1,2]

[1]Gene Regulacion, Stem Cells, and Cancer Program, Centre de Regulació Genòmica (CRG), 08003 Barce

Barcu
Avar

Barutcu *et al. Genome Biology* (2015) 16:214
DOI 10.1186/s13059-015-0768-0

Genome

Three-dimensional disorganisation of the cancer genome occurs coincident with long range genetic and epigenetic alterations.

Phillippa C. Taberlay[1,2,#], Joanna Achinger-Kawecka[1,2,#], Aaron T.L. Lun[4,5], Fabian A. Buske[1], Kenneth Sabir[1], Cathryn M. Gould[1], Elena Zotenko[1,2], Saul A. Bert[1], Katherine A. Giles[1], Denis C. Bauer[3], Gordon K. Smyth[4,6], Clare Stirzaker[1,2], Sean I. O'Donoghue[1,3], Susan J. Clark[1,2,*]

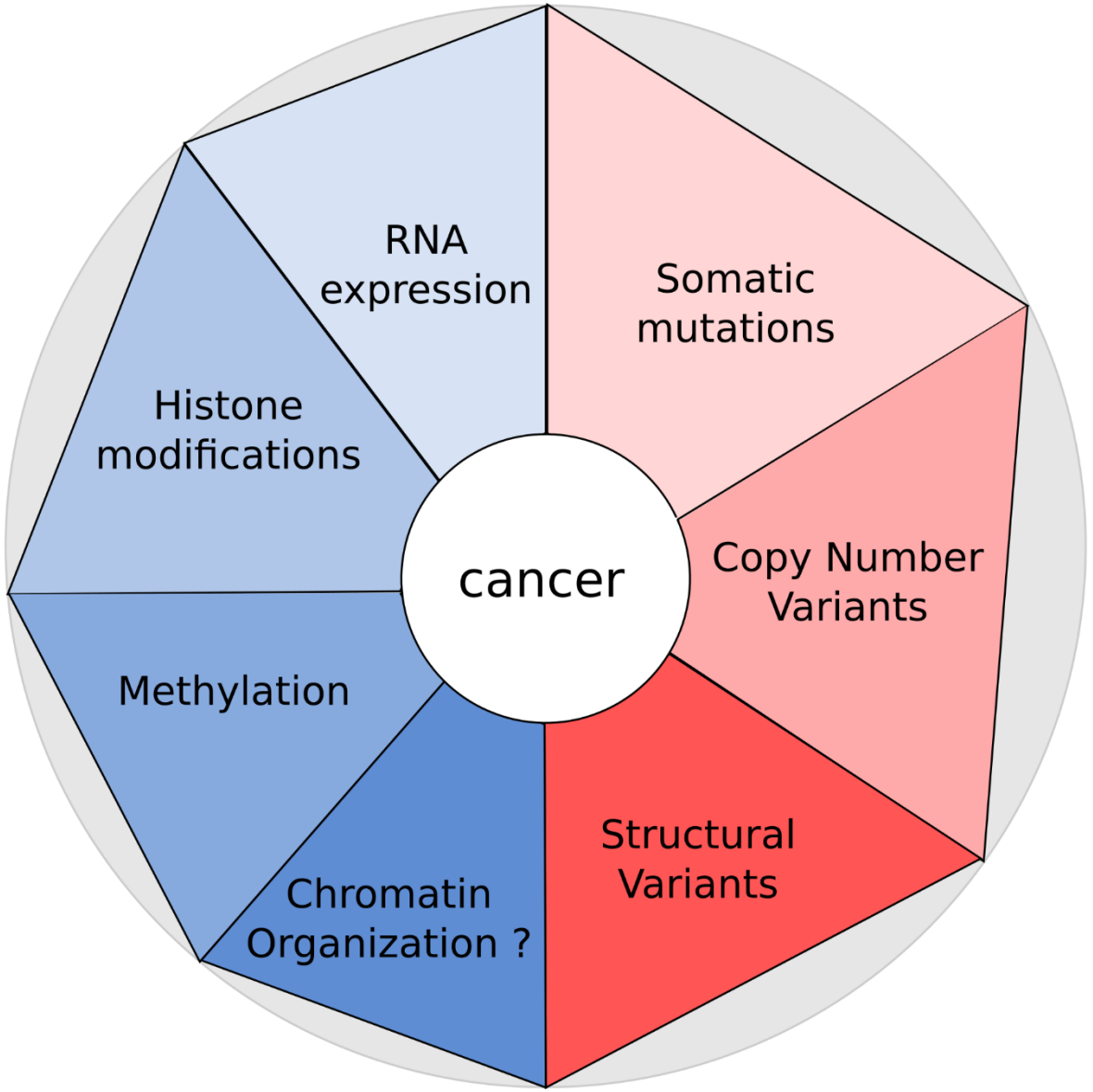
RESEARCH

CrossMark

Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells

A. Rasim Barutcu[1], Bryan R. Lajoie[2], Rachel P. McCord[2], Coralee E. Tye[5], Deli Hong[1,5], Terri L. Messier[5], Gillian Browne[5], Andre J. van Wijnen[4], Jane B. Lian[5], Janet L. Stein[5], Job Dekker[2,3], Anthony N. Imbalzano[1] and Gary S. Stein[5*]

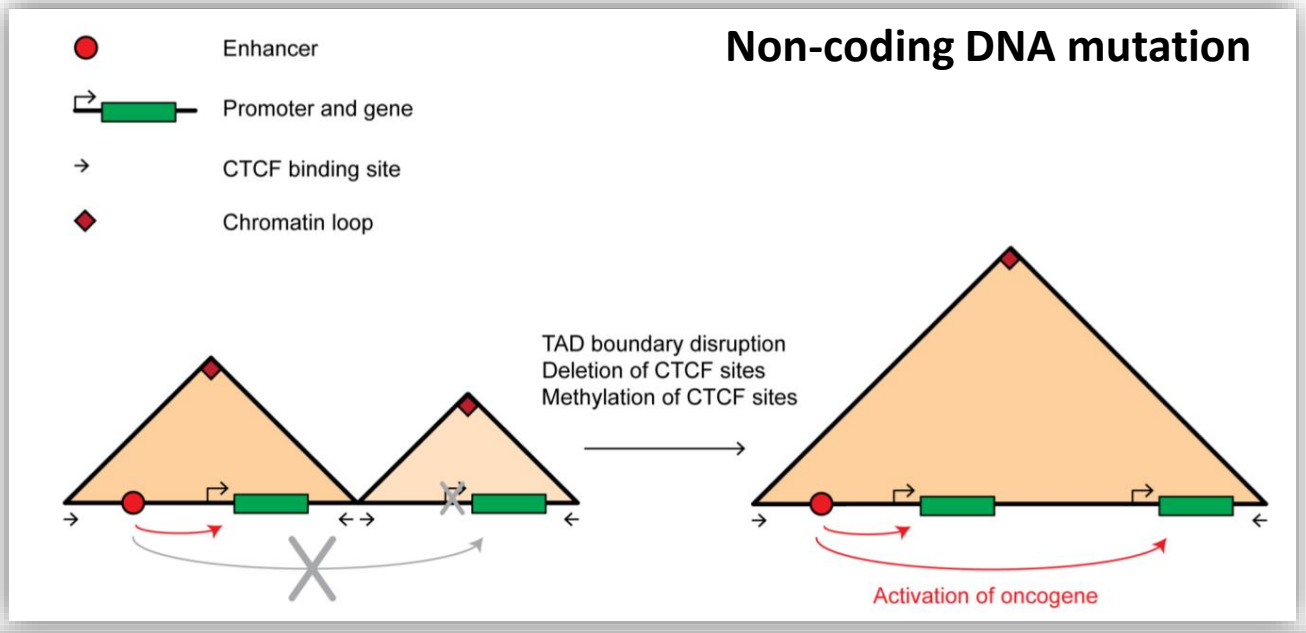
Alterations in cancer (epi)genomics

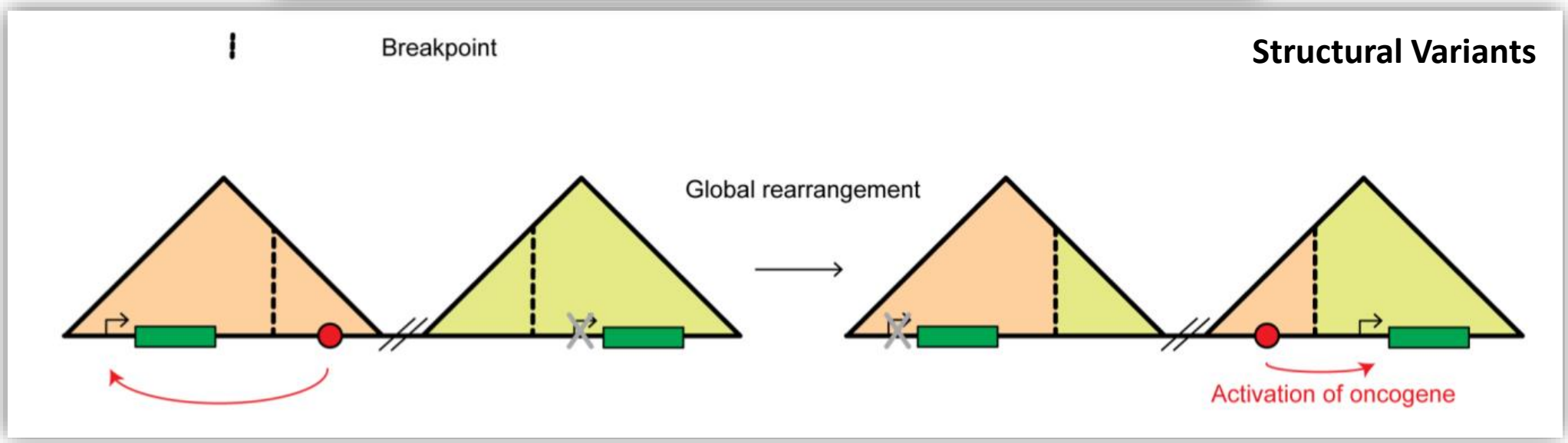


Epigenetic alterations

DNA sequence alterations

Organization of cancer genomes?





Valton & Dekker, 2016

Hi-C, a good tool to study CNVs ?

Method | Open Access

Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours

Louise Harewood, Kamal Kishore, Matthew D. Eldridge, Steven Wingett, Danita Pearson, Stefan Schoenfelder, V. Peter Collins and Peter Fraser

Genome Biology 2017 **18**:125
https://doi.org/10.1186/s13059-017-1253-8 | © The
Received: 9 December 2016 | Accepted: 8 June 2017

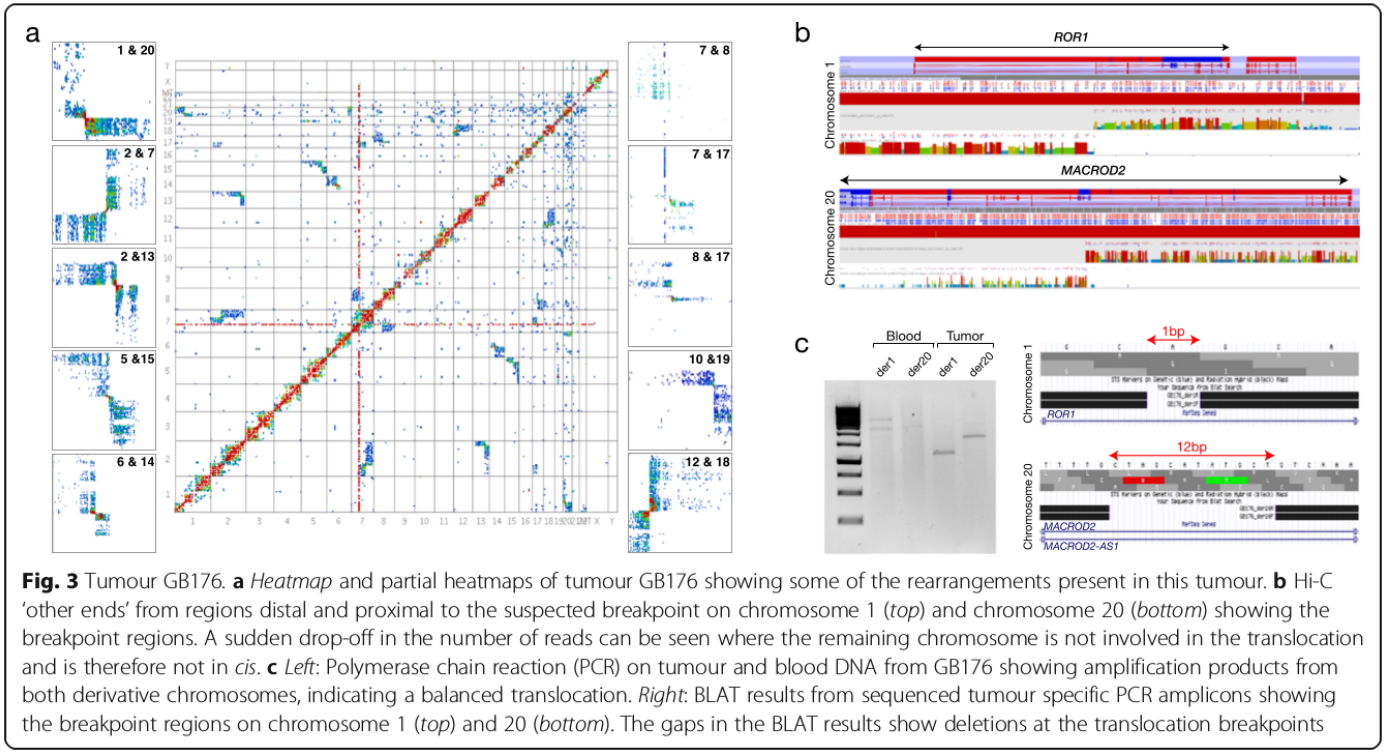
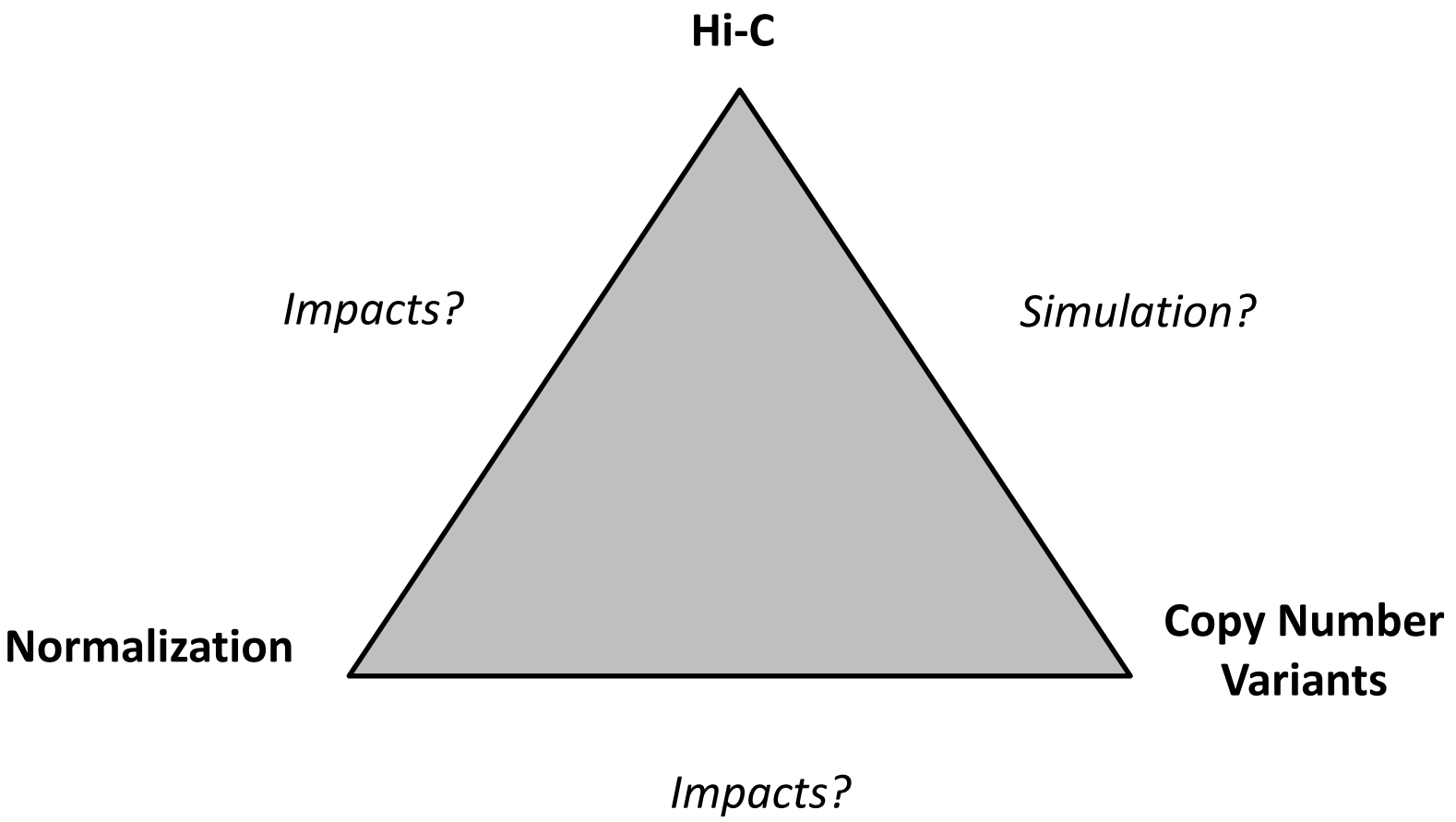
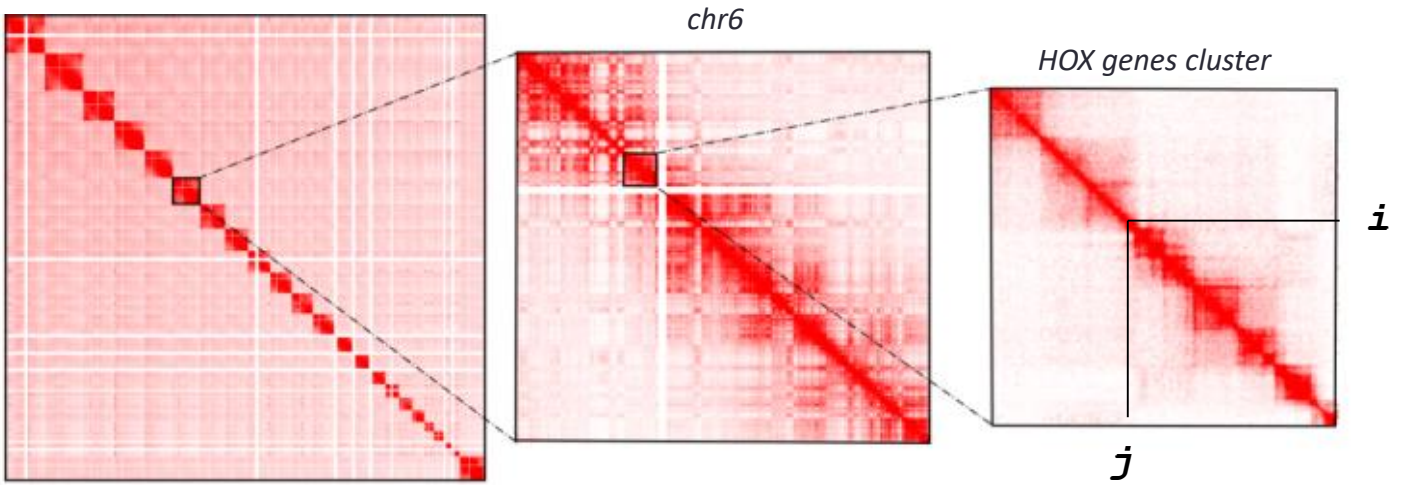


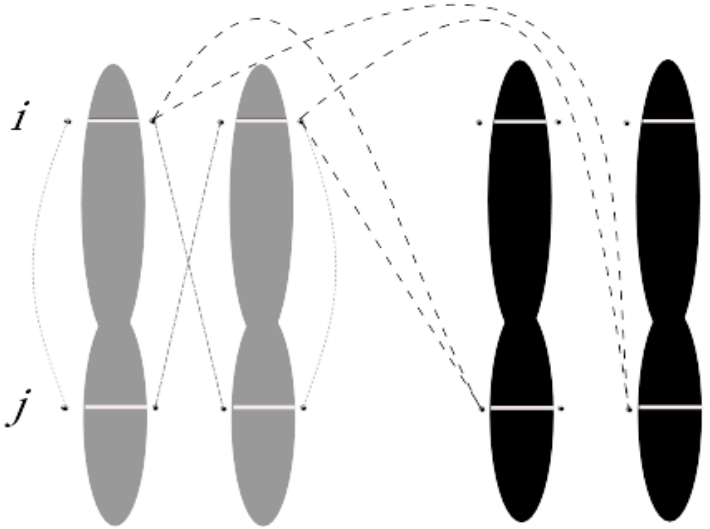
Fig. 3 Tumour GB176. **a** *Heatmap* and partial heatmaps of tumour GB176 showing some of the rearrangements present in this tumour. **b** Hi-C 'other ends' from regions distal and proximal to the suspected breakpoint on chromosome 1 (*top*) and chromosome 20 (*bottom*) showing the breakpoint regions. A sudden drop-off in the number of reads can be seen where the remaining chromosome is not involved in the translocation and is therefore not in *cis*. **c** *Left*: Polymerase chain reaction (PCR) on tumour and blood DNA from GB176 showing amplification products from both derivative chromosomes, indicating a balanced translocation. *Right*: BLAT results from sequenced tumour specific PCR amplicons showing the breakpoint regions on chromosome 1 (*top*) and 20 (*bottom*). The gaps in the BLAT results show deletions at the translocation breakpoints

Challenges in Hi-C cancer data?



Hi-C – What do we count?



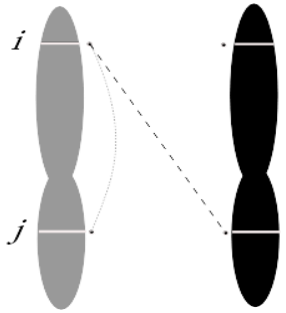


In the context of a diploid genome

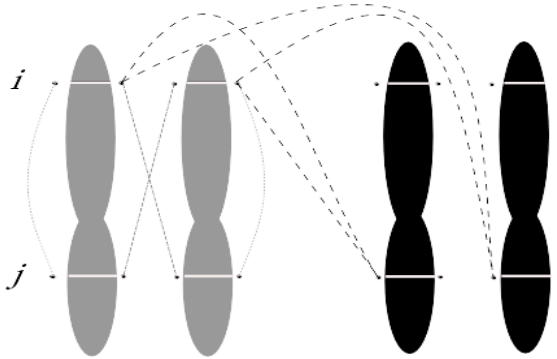
If *i* and *j* belong to the same chromosome
C_{ij} *= 2 cis + 2 transH*

If *i* and *j* belong to different chromosomes
C_{ij} *= 4 trans*

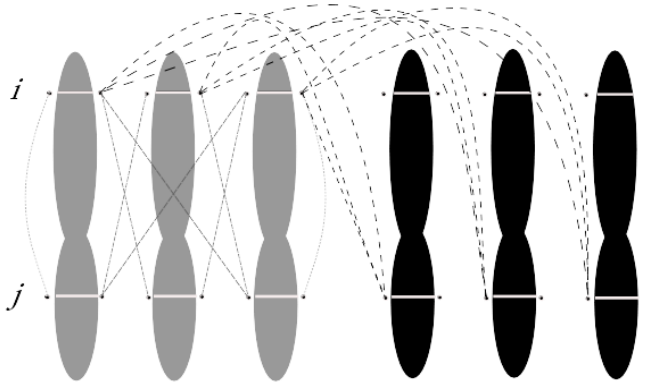
Generalization to polyploid genomes



$N_i = N_j = 1$
If $chr_i = chr_j$, C_{ij} = 1 cis
If $chr_i \neq chr_j$, C_{ij} = 1 trans



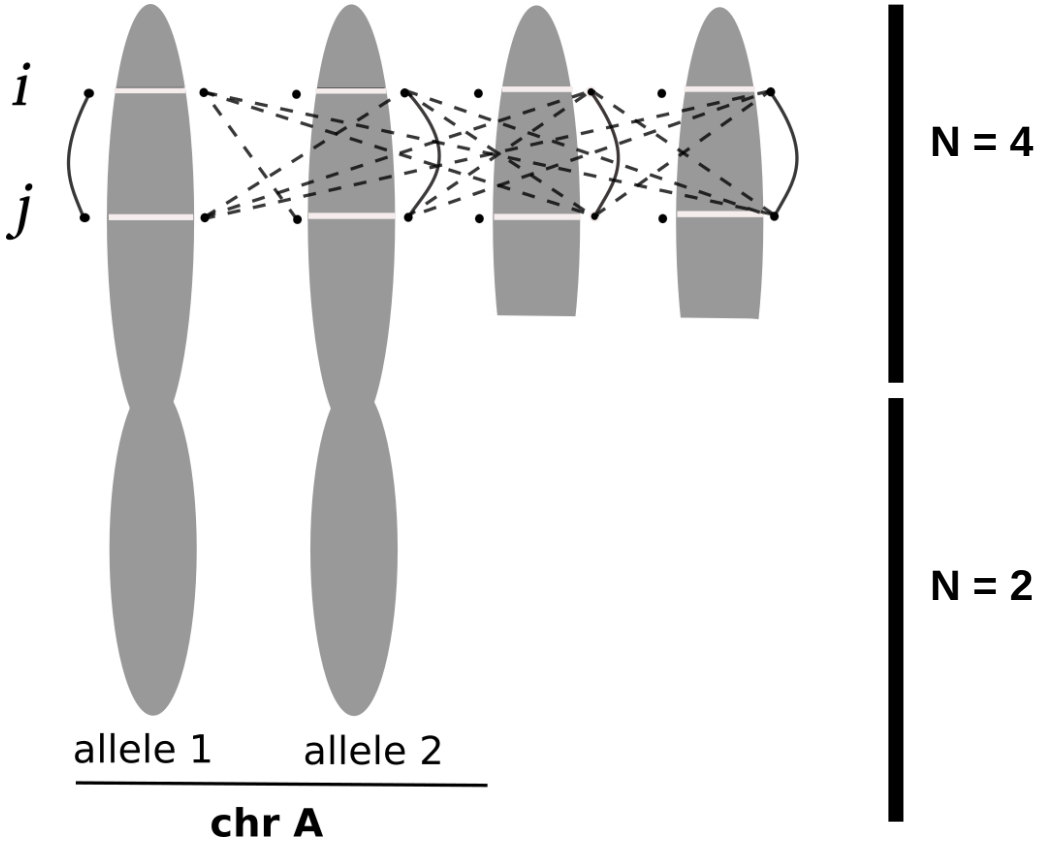
$N_i = N_j = 2$
If $chr_i = chr_j$, C_{ij} = 2 cis + 2 transH
If $chr_i \neq chr_j$, C_{ij} = 4 trans



$N_i = N_j = 3$
If $chr_i = chr_j$, C_{ij} = 3 cis + 6 transH
If $chr_i \neq chr_j$, C_{ij} = 9 trans

$N_i = N_j$
If $chr_i = chr_j$, ***$C_{ij} = N_i$ cis + N_i (N_j -1) transH***
If $chr_i \neq chr_j$, ***$C_{ij} = N_i N_j$ trans***

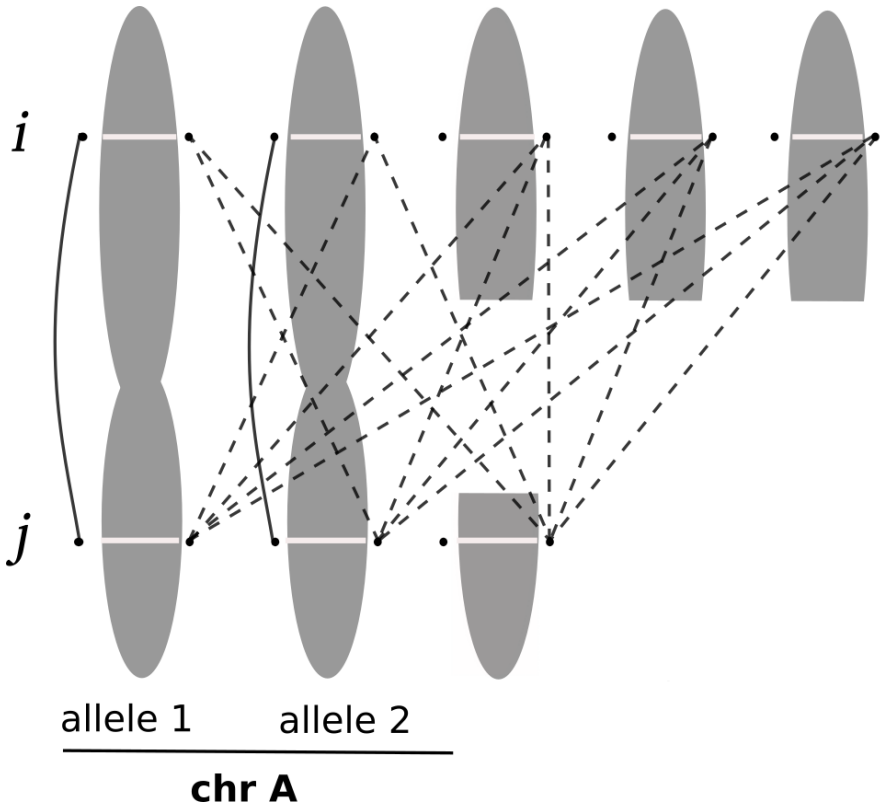
Extension to Cancer genome

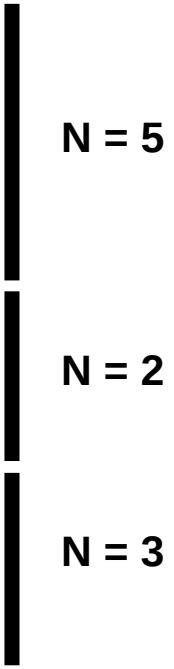


If *i* and *j* belong to the same chromosomal segment

$C_{ij} = N_i\ cis + N_i\ (N_j - 1)\ transH$

Extension to Cancer genome





C_{ij} = 2 cis + (2x4 + 5) transH

If *i* and *j* belong to different chromosomal segments

$C_{ij} = p\ cis + (N_i * N_j - p) * transH$

where *p* is the number of complete chromosomes

Simulation of cancer Hi-C data

1. Estimate the *cis*$_{ij}$ *and transH terms from a real diploid Hi-C dataset.*

Estimate *transH* under the assumption that the contact probability between homologuous chromosomes can be estimated using the observed trans contact between different chromosomes.

For each interaction C_{ij}, between the loci *i* and *j*, estimate the *cis* value using

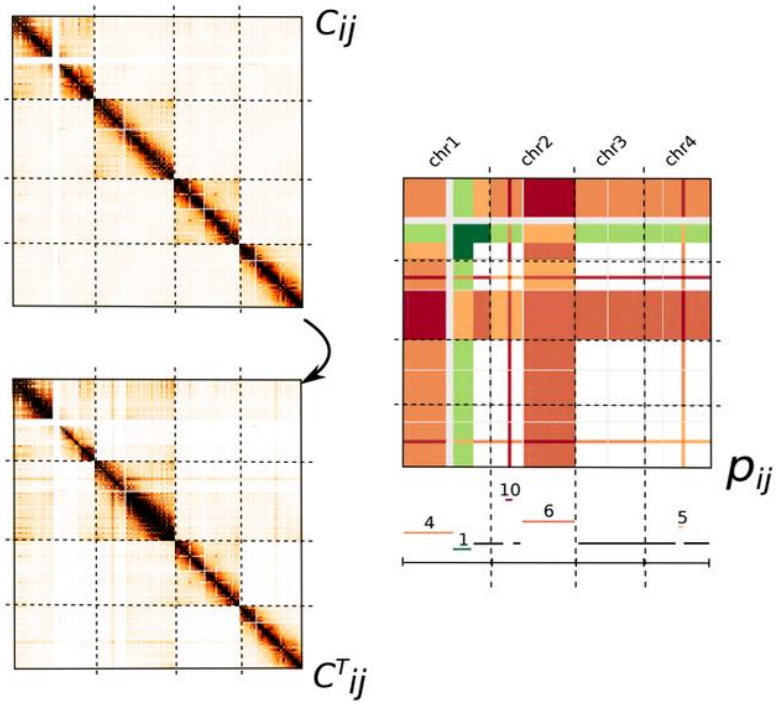
$$C_{ij} = 2\ cis_{ij} + 2\ transH$$

2. Simulate the effect of CNVs on the contact matrix

Given the *cis* and *transH* values for two loci *i* and *j*, calculate E_{ij}, the expected counts in the presence of CNVs

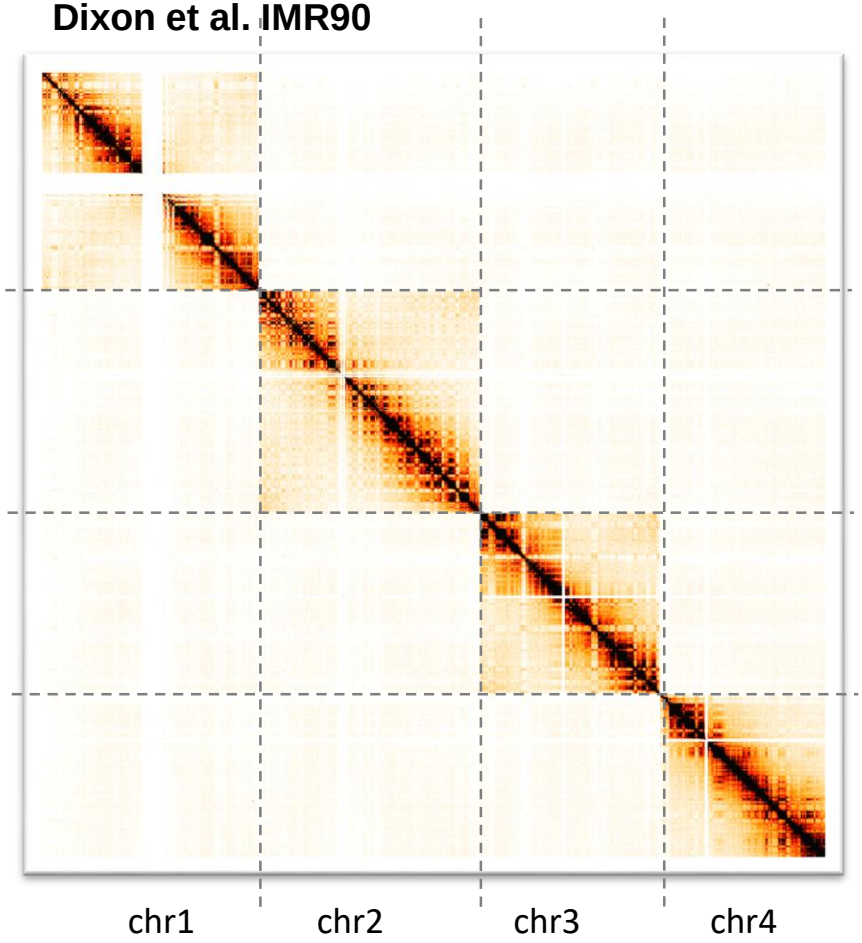
Calculate the scaling factor matrix $p_{ij} = E_{ij} / C_{ij}$

Estimate the simulated data using a binomial downsampling of parameter $p_{ij} / \max(p_{ij})$

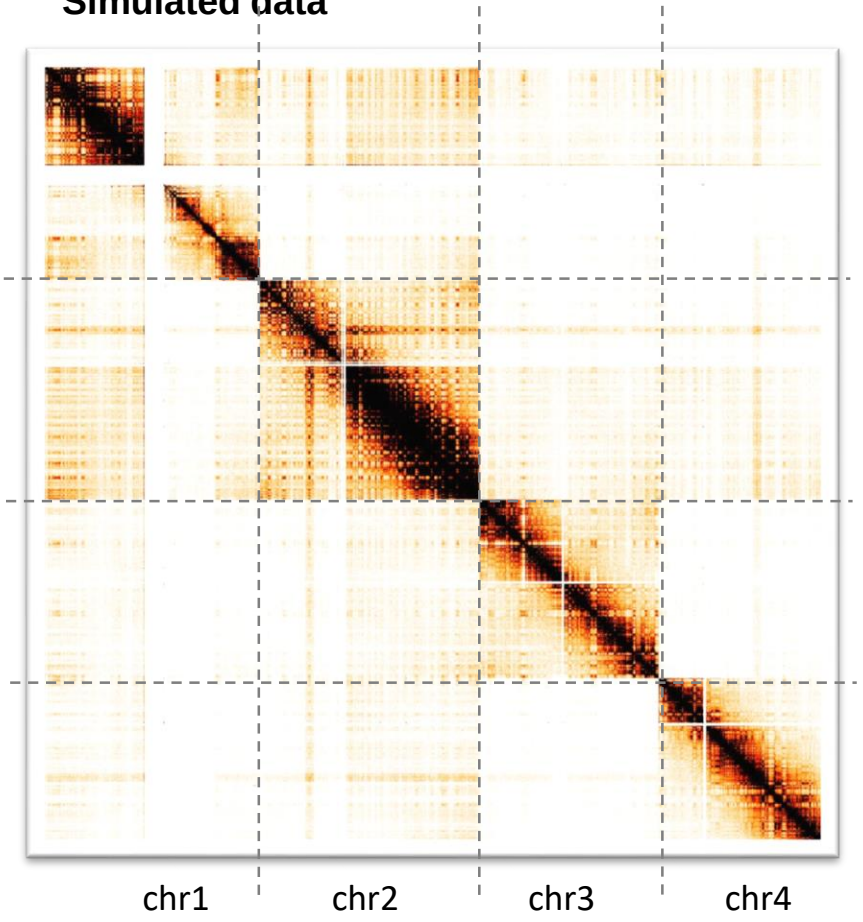


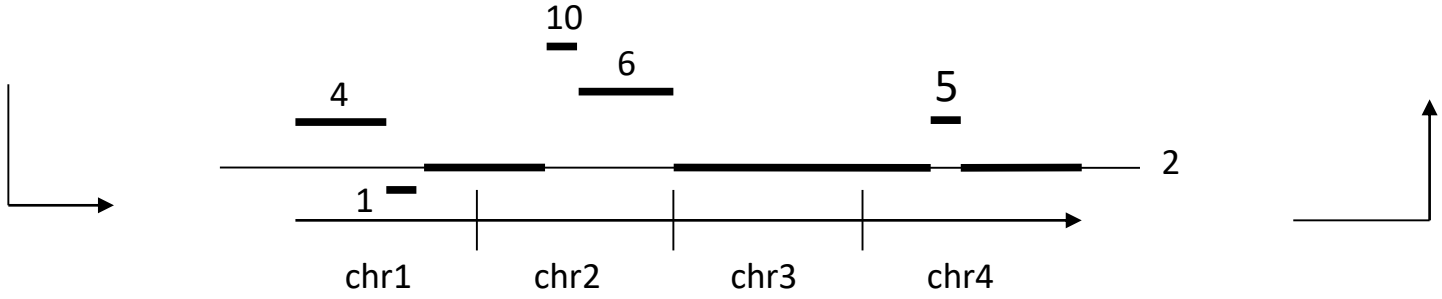
Simulation - Results

Dixon et al. IMR90



Simulated data

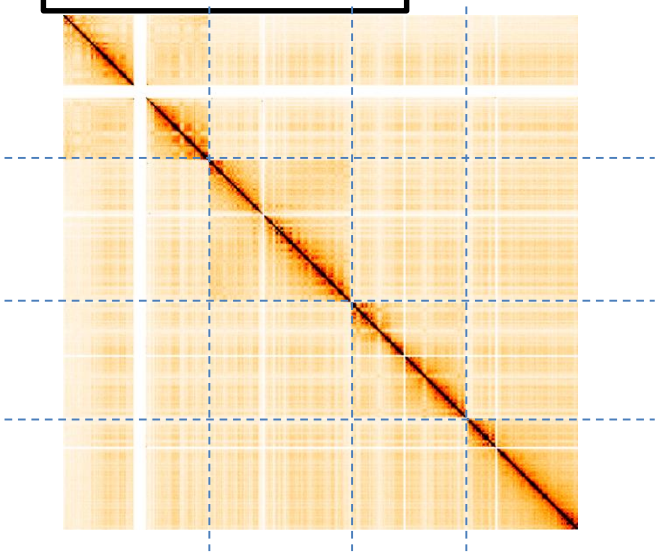




How to validate the simulation model ?

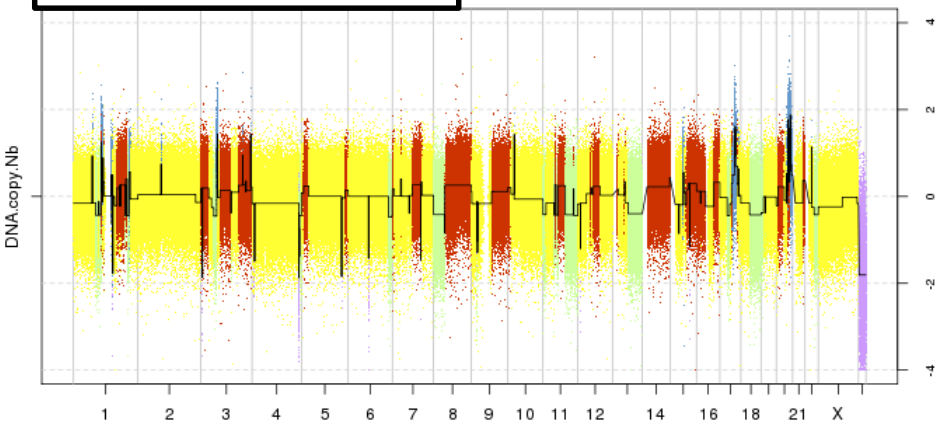
In order to validate our simulation model, we used Hi-C from MCF10 normal-like data, from which we simulated the MCF7 CNV profile

MCF10A Hi-C data

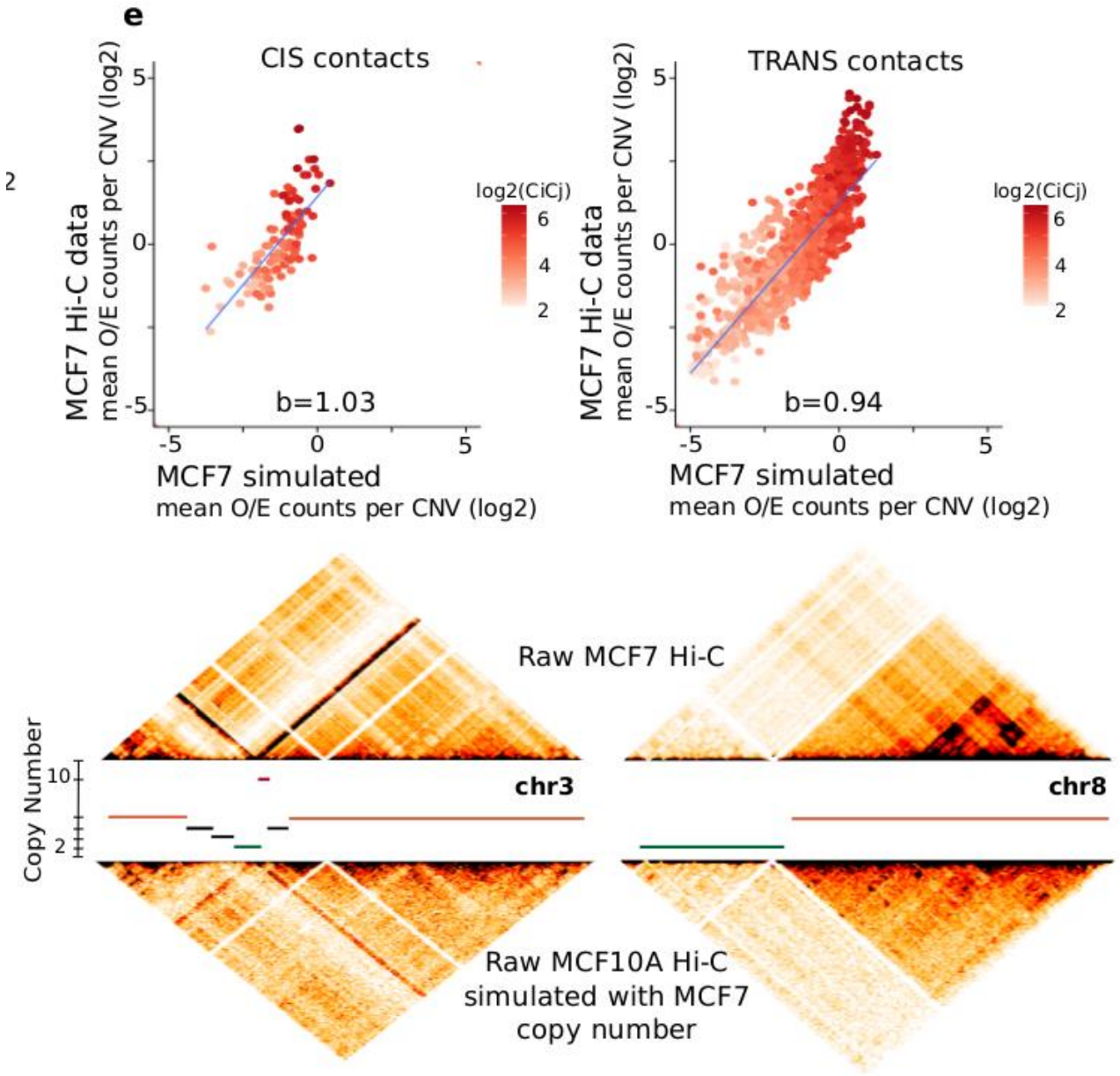


+

MCF7 Affymetrix data



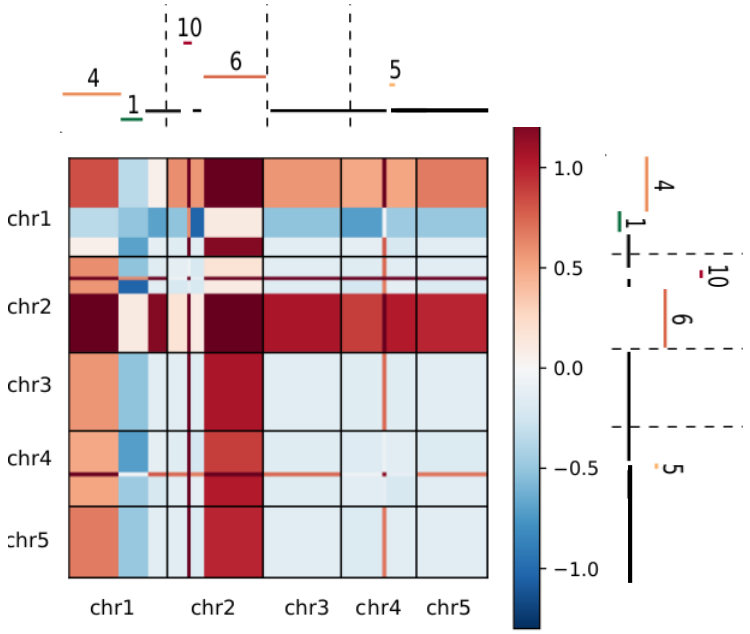
Simulation - Validation



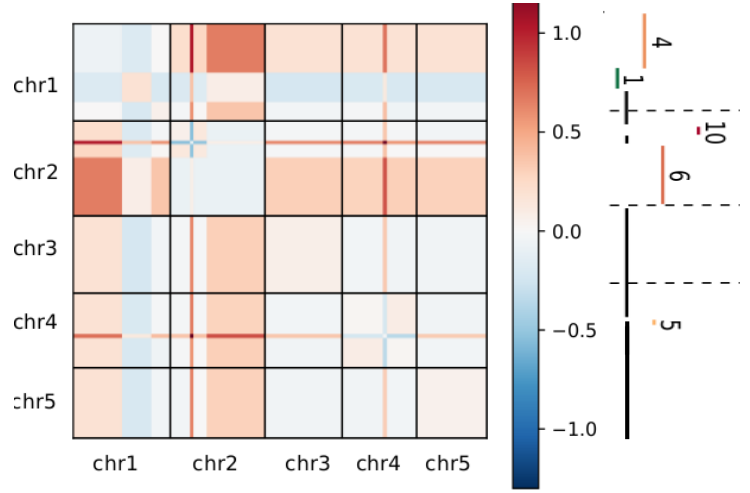
Effect of ICE normalization

The iterative correction (ICE) **does not** correct for CNV bias.

RAW

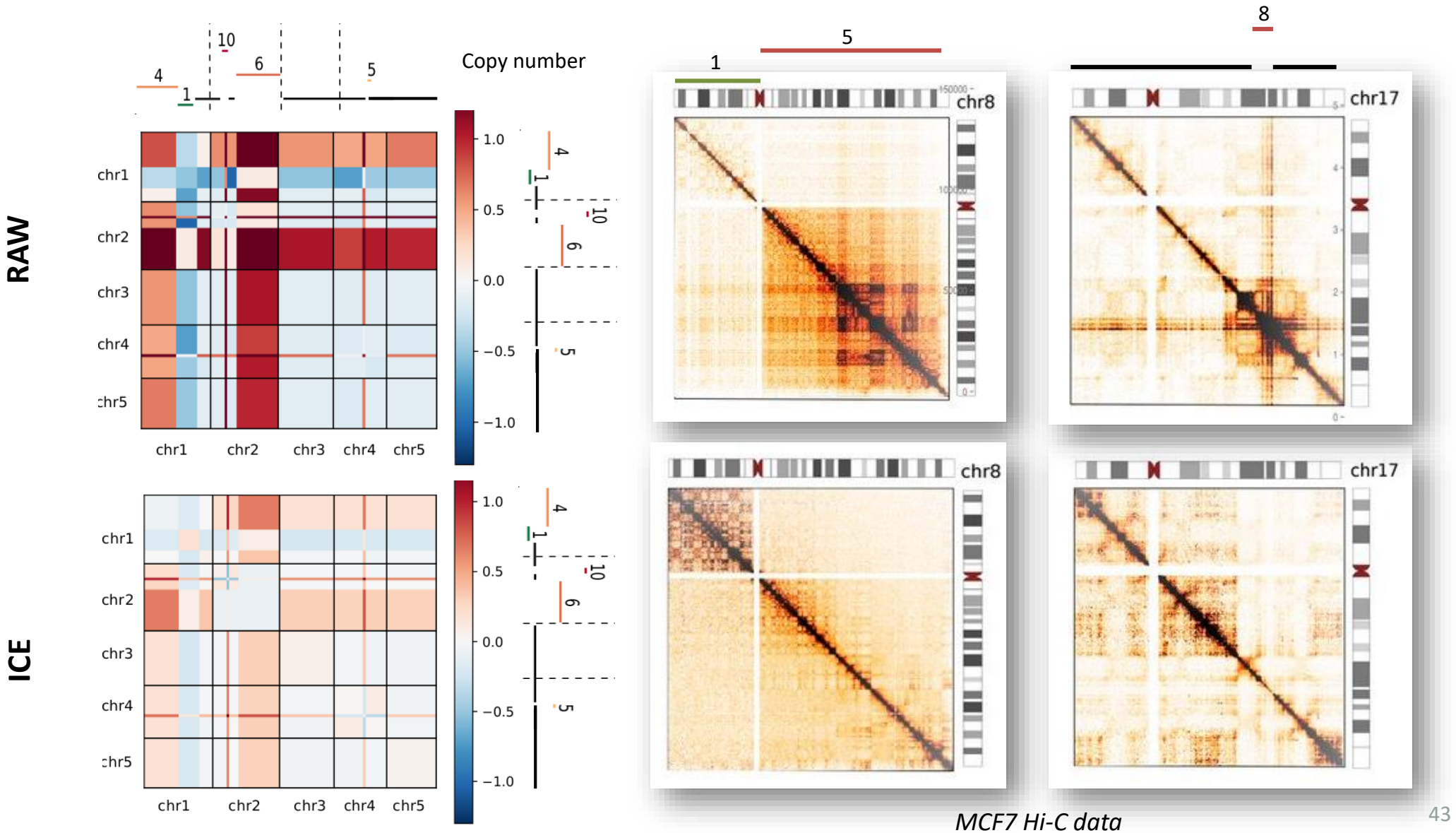


ICE



Effect of ICE normalization

The iterative correction (ICE) **does not** correct for CNV bias.
More importantly, it leads to **an inversion of the signal in cis**.



How to normalize cancer Hi-C data?

How to take into account the CNV signal into the normalization ?

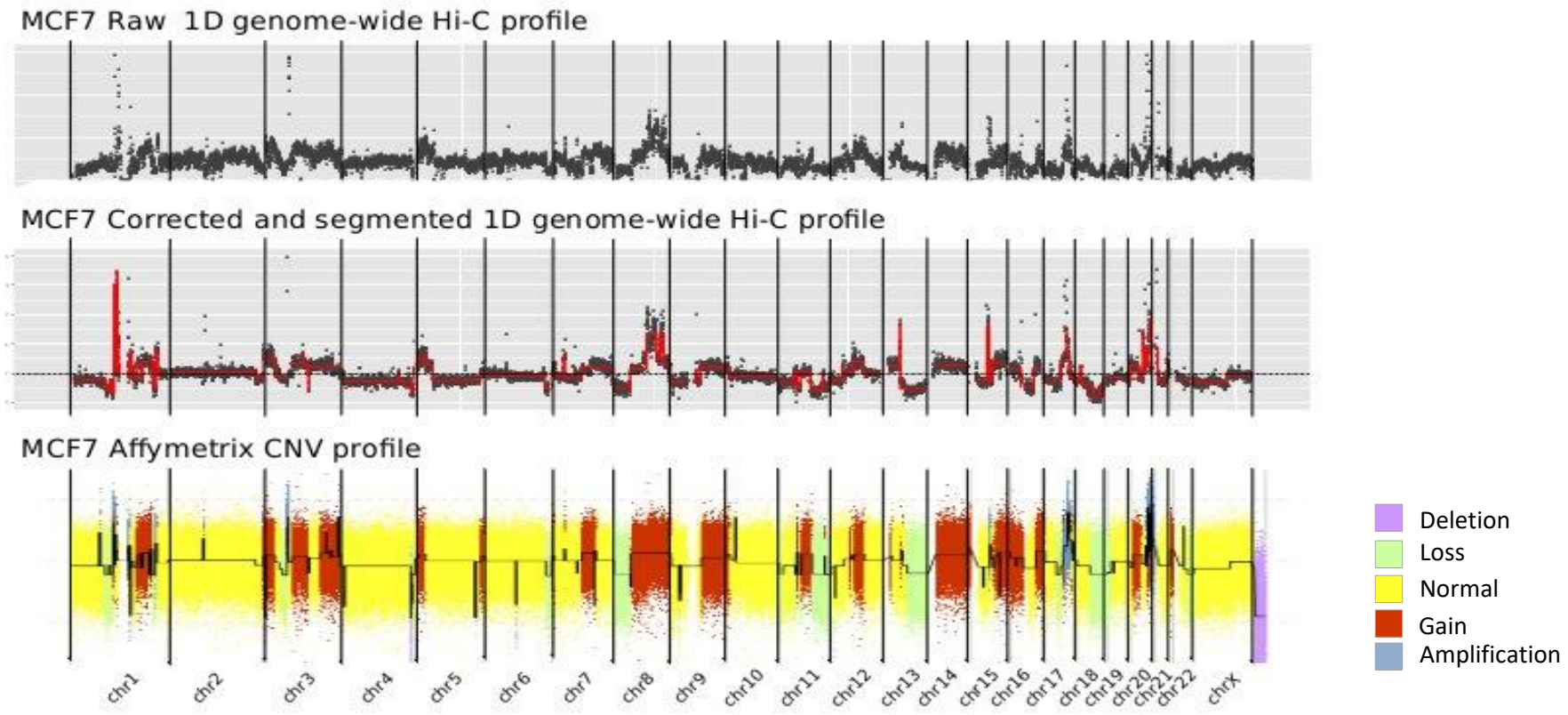
1. **Correct for systematic bias but not for the CNVs signal,** which can be useful for biological interpretation of cancer, for 3D modeling, genome reconstruction, contribution to CNVs to disease, *etc*.

2. **Correct for all bias including the CNVs** because it migth introduce a bias in my downstream analysis (differential contacts, detection of chromosome compartments, *etc*.)

Estimation of DNA breakpoints from Hi-C data

The segmentation of 1D Hi-C profile is performed as follow :

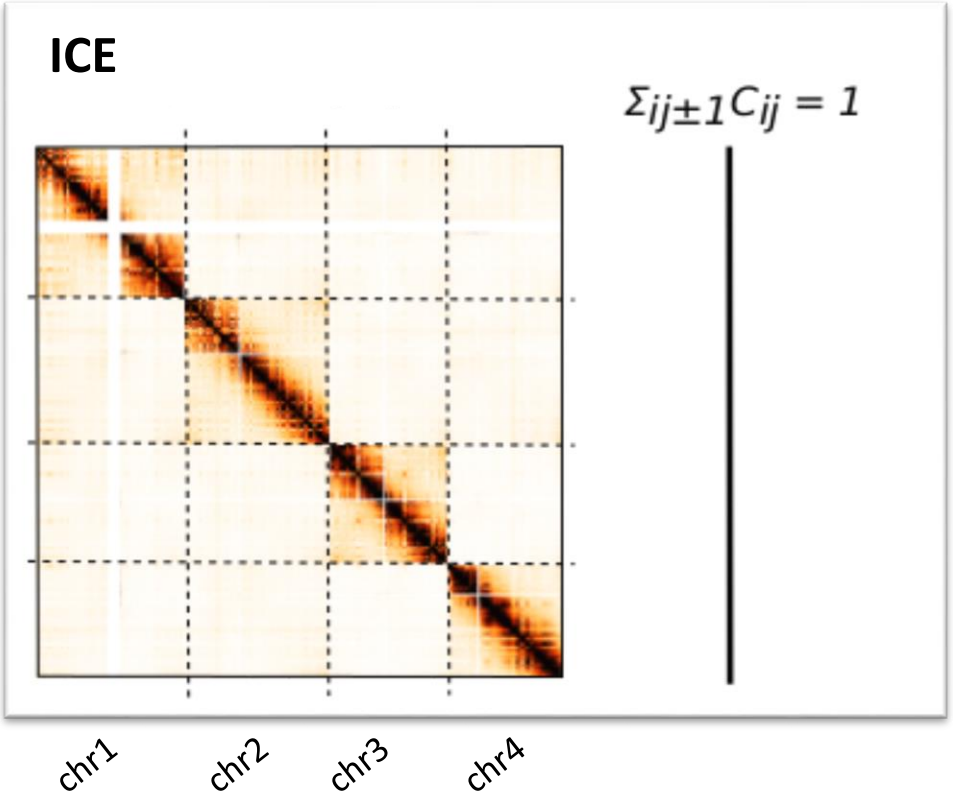
1. Generate the 1D Hi-C profile as the sum of contact per locus genome-wide
2. Remove systematic biases using a *Poisson* regression model
3. Segment the profile

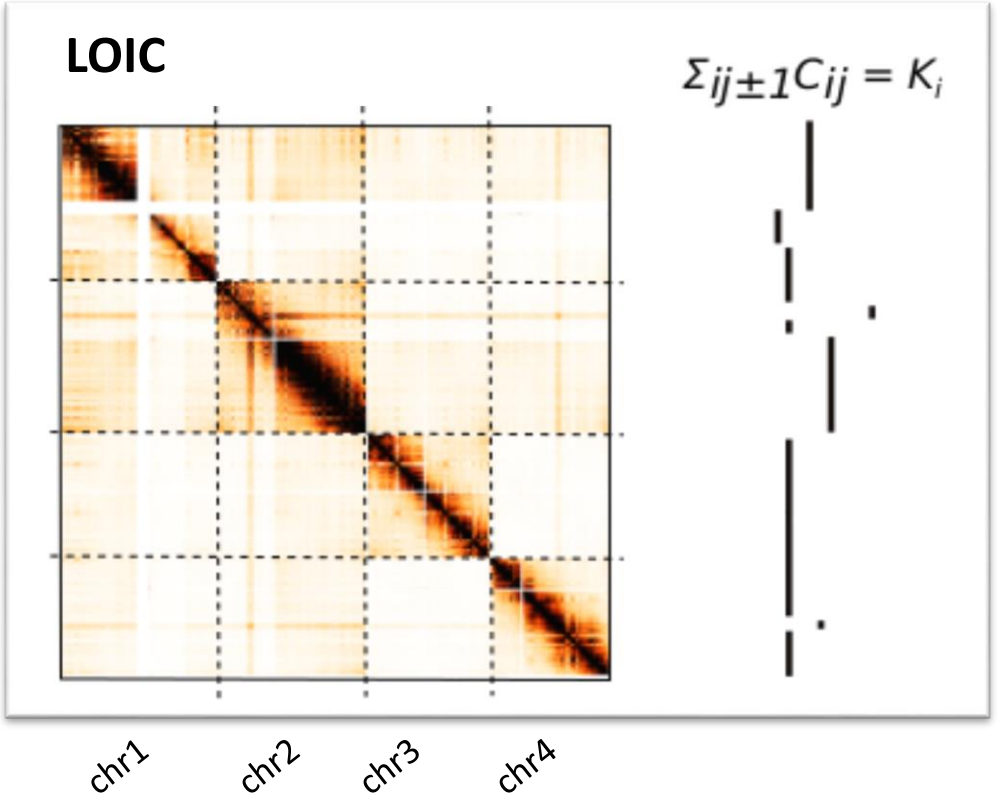


Validation on 100 simulated data-sets : 91% recall / 62.4% precision

CNV-based normalization of Hi-C cancer data

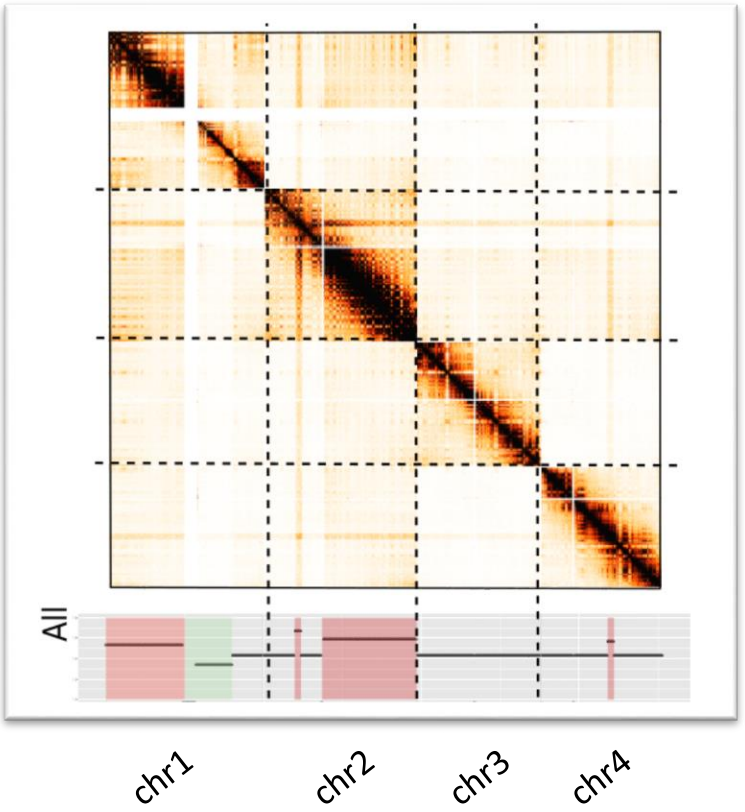
The Local Iterative correction (LOIC) normalization method extends the ICE model, making the assumption of local equal visibility per genomic segment





CNV-based normalization of Hi-C cancer data

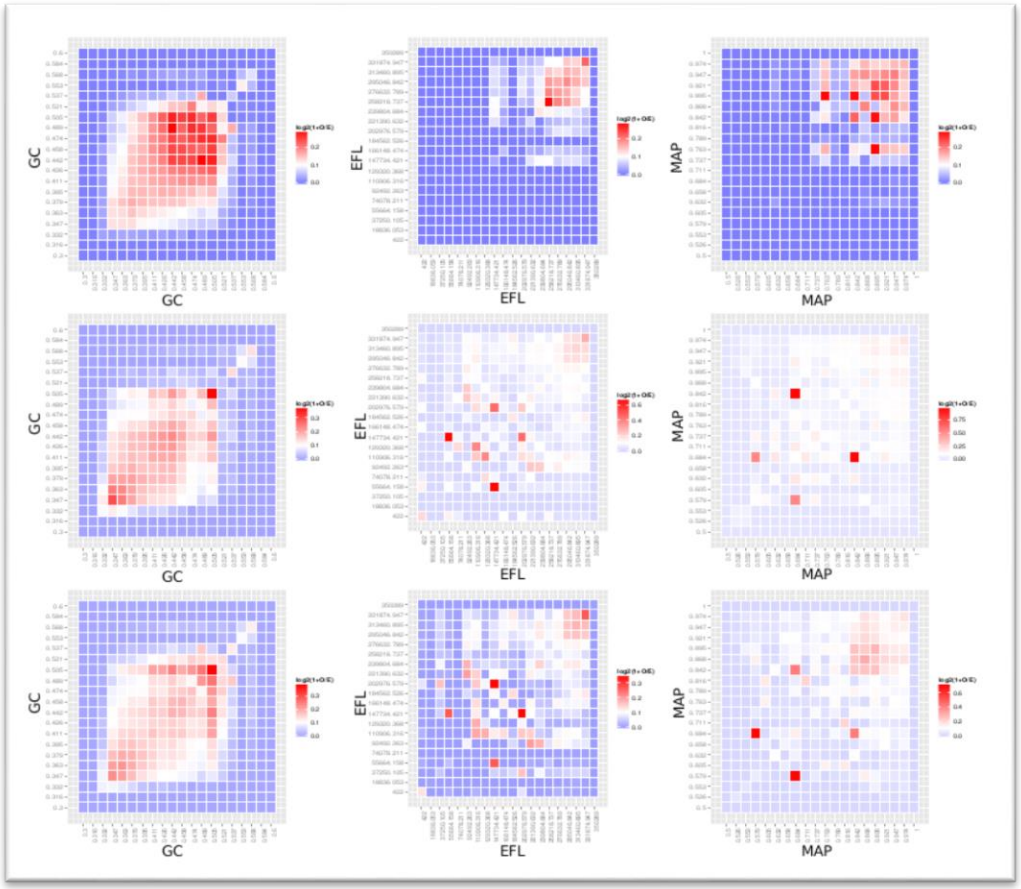
The Local Iterative correction (LOIC) normalization method extends the ICE model, making the assumption of local equal visibility per genomic segment



RAW

ICE

LOIC



GC content

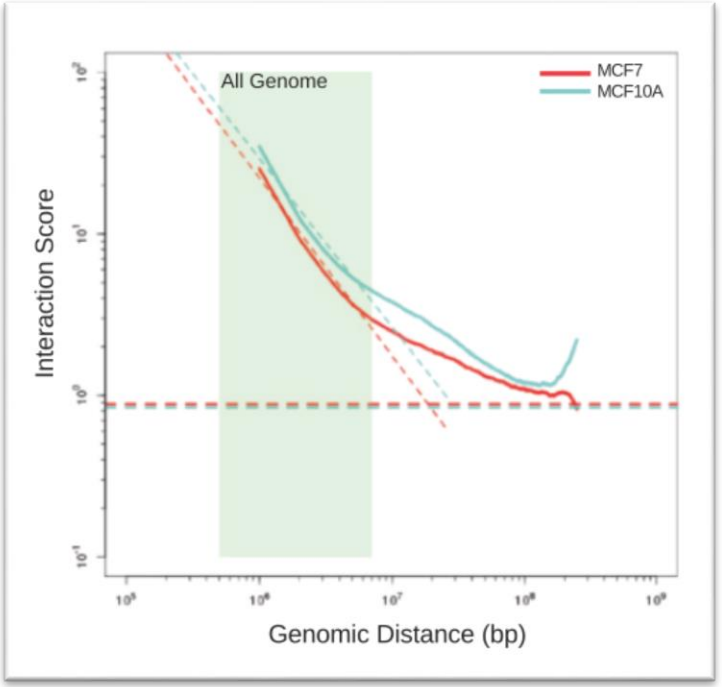
Fragment Length

Mappability

Removing CNVs from cancer Hi-C data

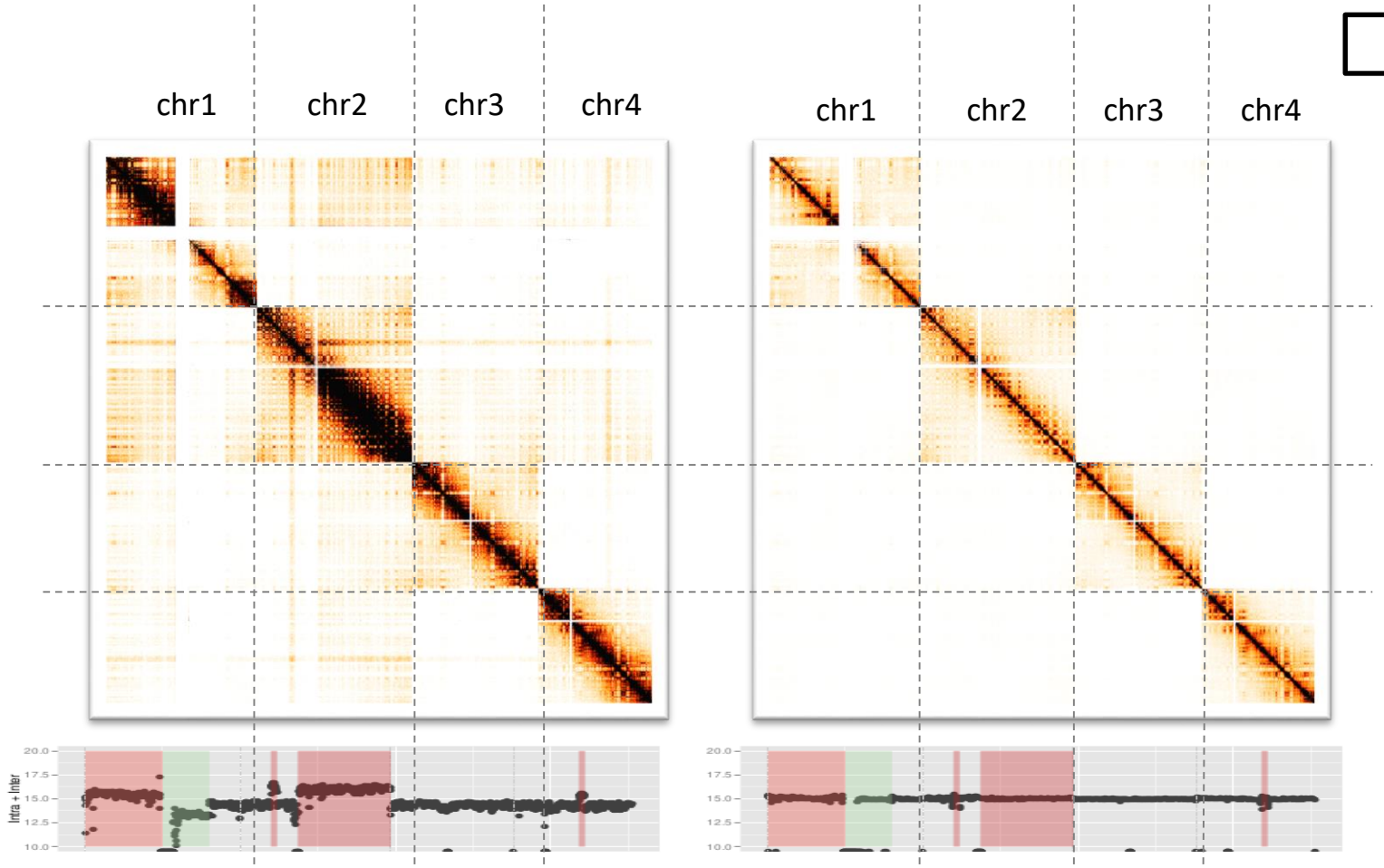
We assume that the copy number bias is constant per block and that the contact counts at a given genomic distance should be the same regardless the copy number status.

1- Run the ICE normalization
2- Estimate the average **counts ~ distance** signal on the genome-wide matrix
3- Based on the segmentation profile, rescale the counts ~ distance fit for each segmentation block

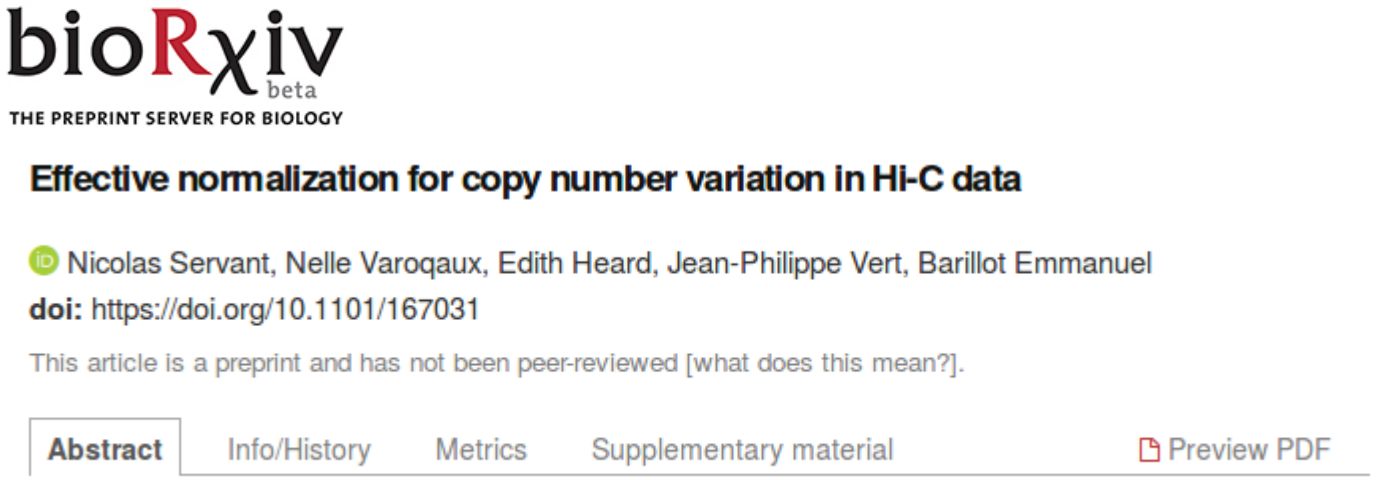


From Barutcu et al. 2015

Removing CNVs from cancer Hi-C data



Cancer Hi-C data normalization



- CNVs estimation from Hi-C data
- Cancer Hi-C data simulation
- Normalization of Hi-C cancer data

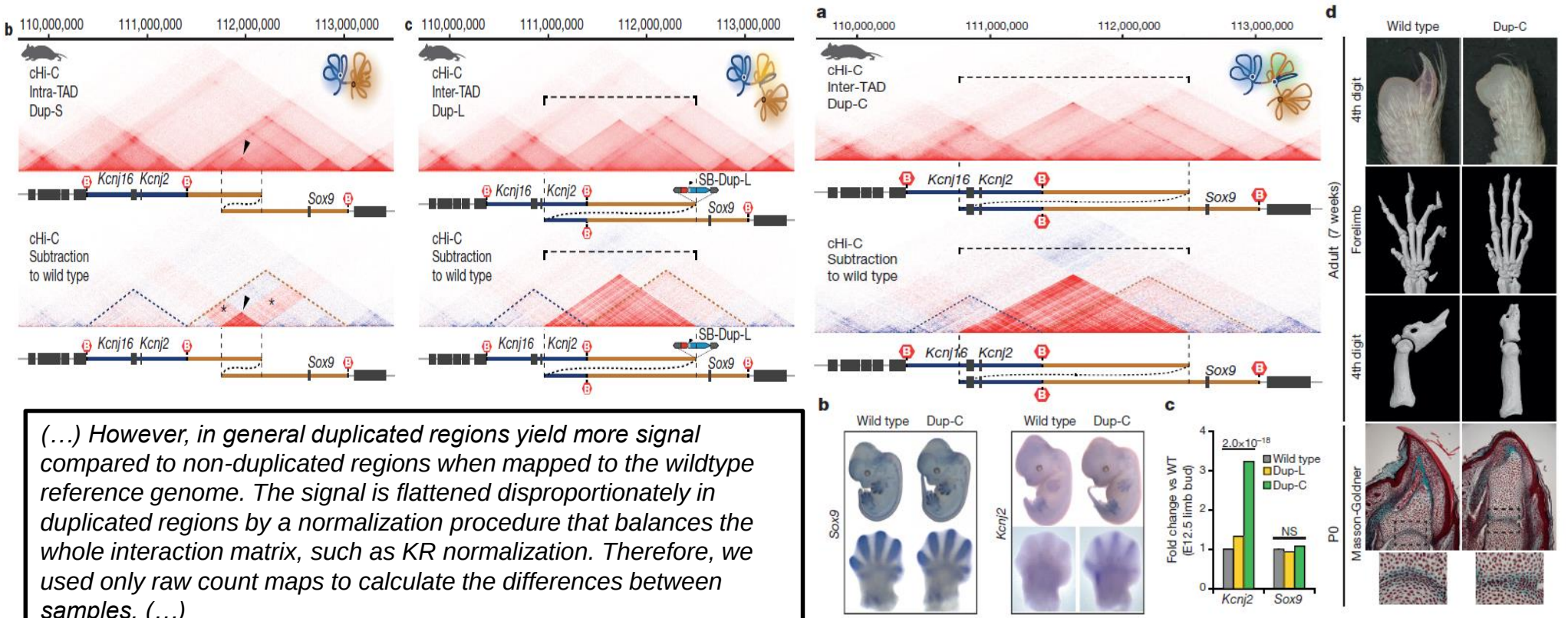
Available at https://github.com/nservant/cancer-hic-norm/

Normalization methods are included into the *iced* python module and available at https://github.com/hiclib/iced

How useful is the LW-IC method ?

Formation of new chromatin domains determines pathogenicity of genomic duplications

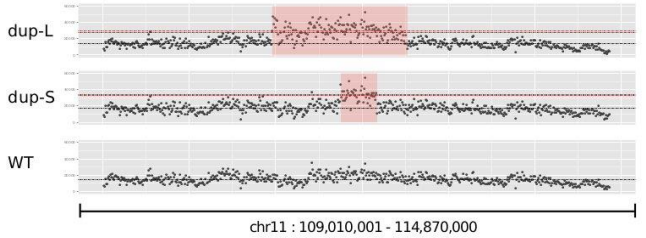
Martin Franke[1,2*], Daniel M. Ibrahim[1,2,3*], Guillaume Andrey[1], Wibke Schwarzer[4], Verena Heinrich[2,5], Robert Schöpflin[5], Katerina Kraft[1,2], Rieke Kempfer[1], Ivana Jerković[1,2], Wing-Lee Chan[2], Malte Spielmann[1,2], Bernd Timmermann[6], Lars Wittler[7], Ingo Kurth[8,9], Paola Cambiaso[10], Orsetta Zuffardi[11], Gunnar Houge[12], Lindsay Lambie[13], Francesco Brancati[14,15], Ana Pombo[3,16], Martin Vingron[5], Francois Spitz[4] & Stefan Mundlos[1,2,3,17]

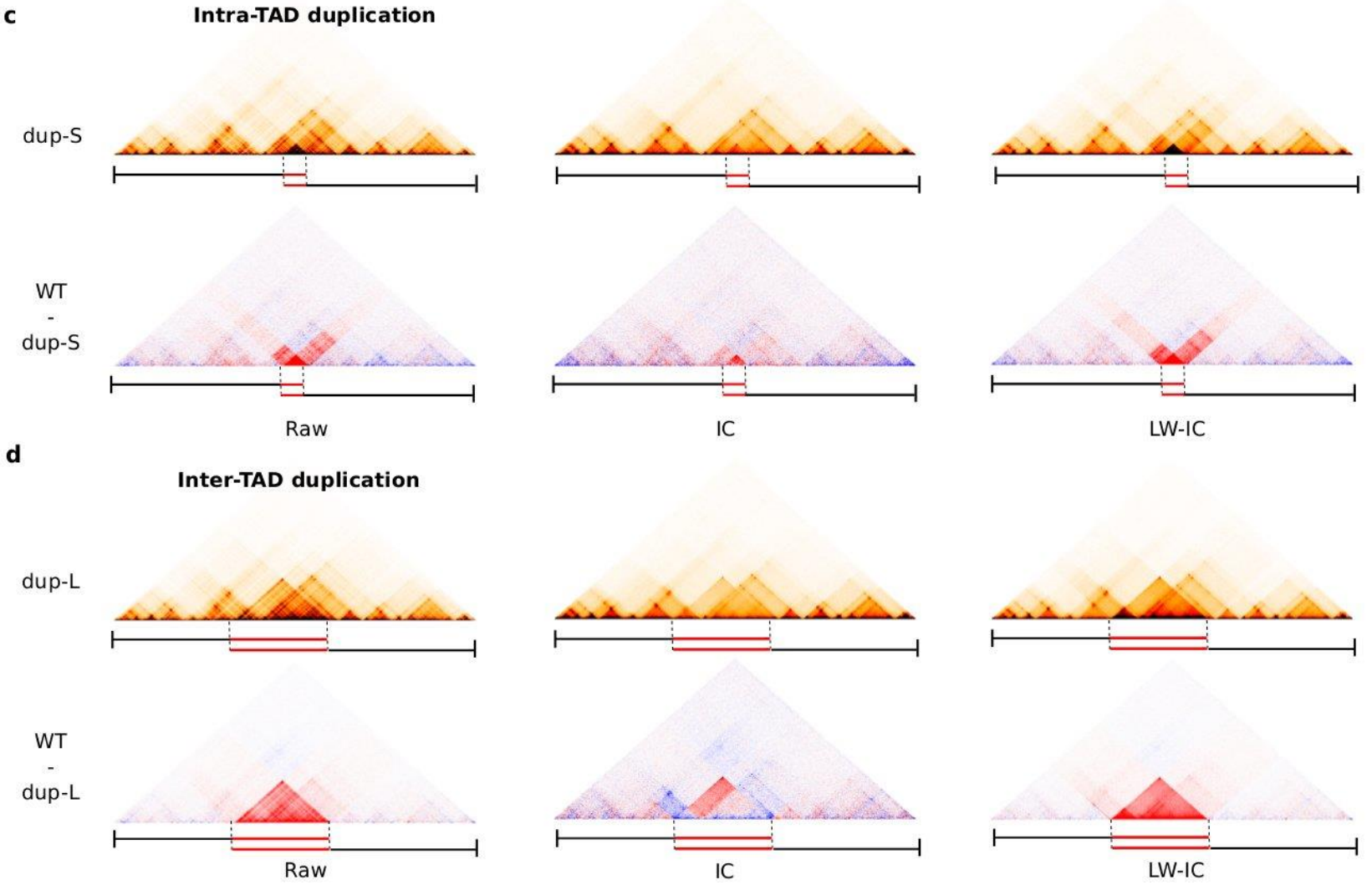


(…) However, in general duplicated regions yield more signal compared to non-duplicated regions when mapped to the wildtype reference genome. The signal is flattened disproportionately in duplicated regions by a normalization procedure that balances the whole interaction matrix, such as KR normalization. Therefore, we used only raw count maps to calculate the differences between samples. (…)

How useful is the LW-IC method ?

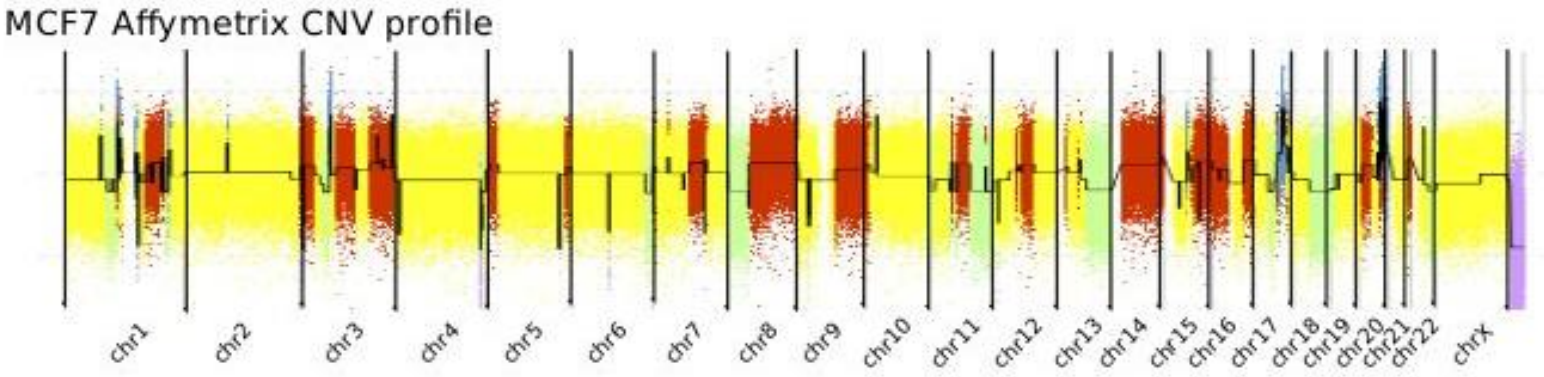
Application of LW-IC on Franke et al. data

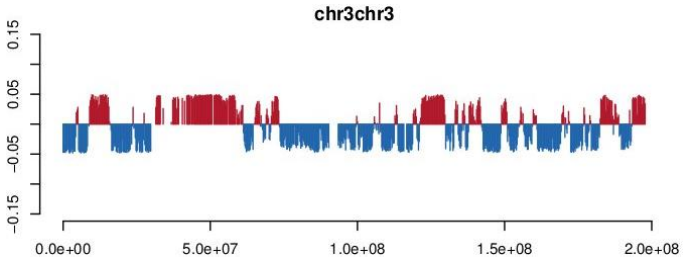


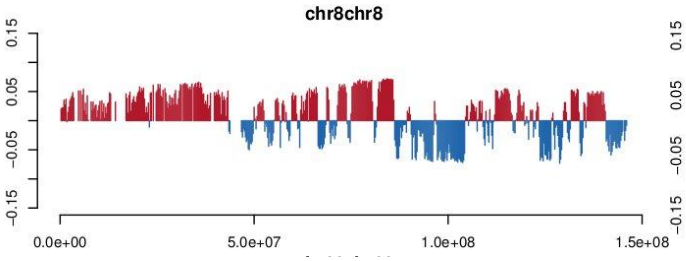


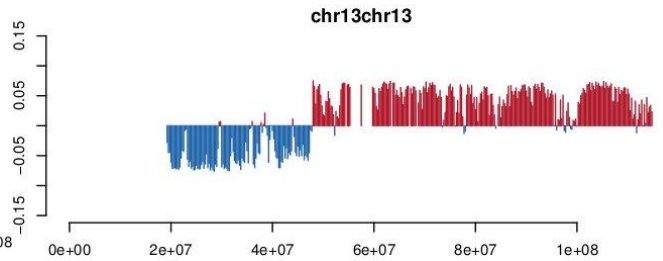
Going further with downstream analysis

- The detection of A/B chromosome compartments is usually based on PCA analysis of the intra-chromosomal maps correlation.
- The methods is surprisingly robust to CNV variations
- But for some chromosomes, the PC1 signal is biased toward the CNV profile



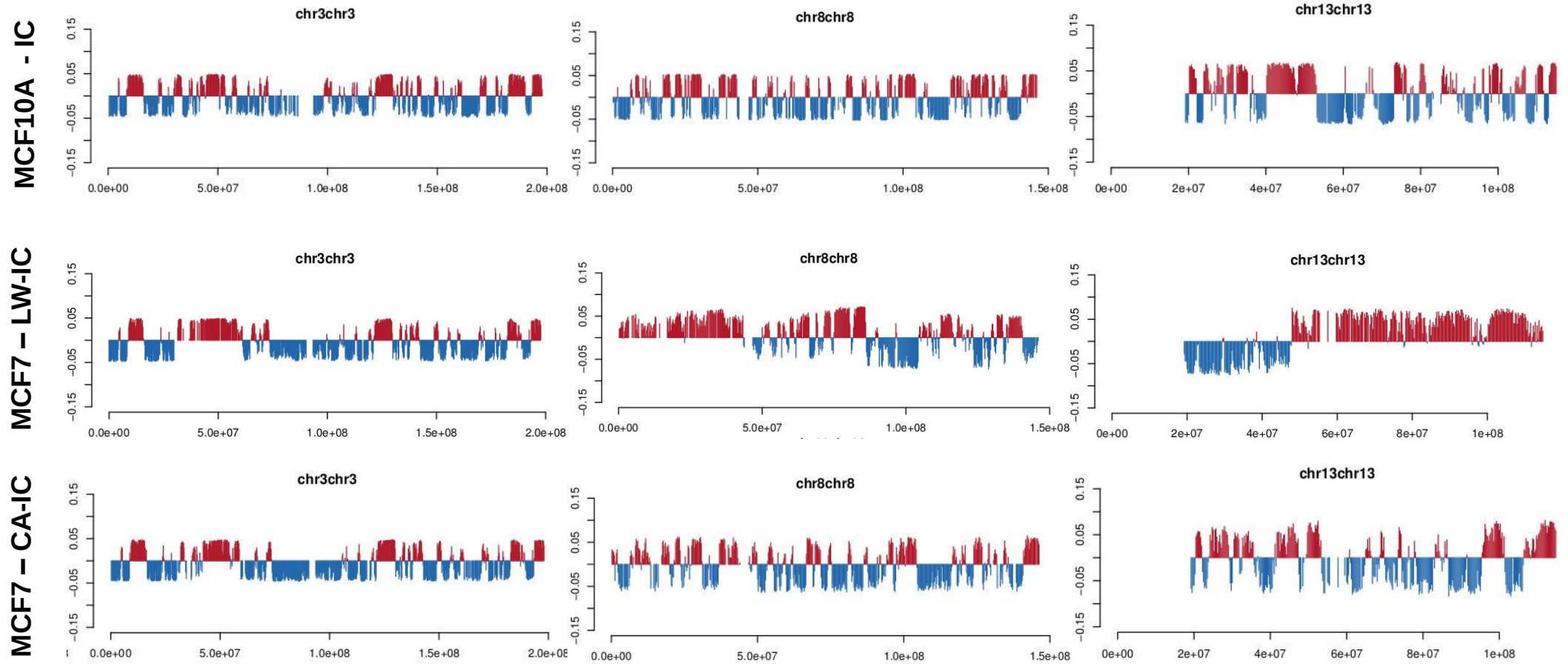






Removing CNVs from cancer Hi-C data

Detection of A/B chromosome compartments



Take Home Messages

- HiC-Pro available at https://github.com/nservant/HiC-Pro
- HiC-Pro is collaborative project, so do not hesitate to propose improvments or to report errors

- In a Cancer context, we demonstrate that the ICE normalization does not allow to correct for CNVs and that it results in a shift in contact probabilities between altered regions in cis
- We proposed a first simulation model to investigate the CNVs impact on Hi-C map

We then proposed two new methods for Cancer Hi-C data and applied it to different case studies

- LOIC to keep the CNVs information
- CAIC to remove the CNVs

Perspectives

- HiC-Pro is still under active development to answer the need of the community and to follow recent Hi-C protocols as capture Hi-C

- We are currently working to improve our normalization methods including the genomic distance in the model and updating the segmentation method

- These tools are currently applied to several cancer Hi-C projects

Many Thank

Nelle Varoquaux
Jean-Philippe Vert
Edith Heard
Emmanuel Barillot

And all HiC-pro contributors ...