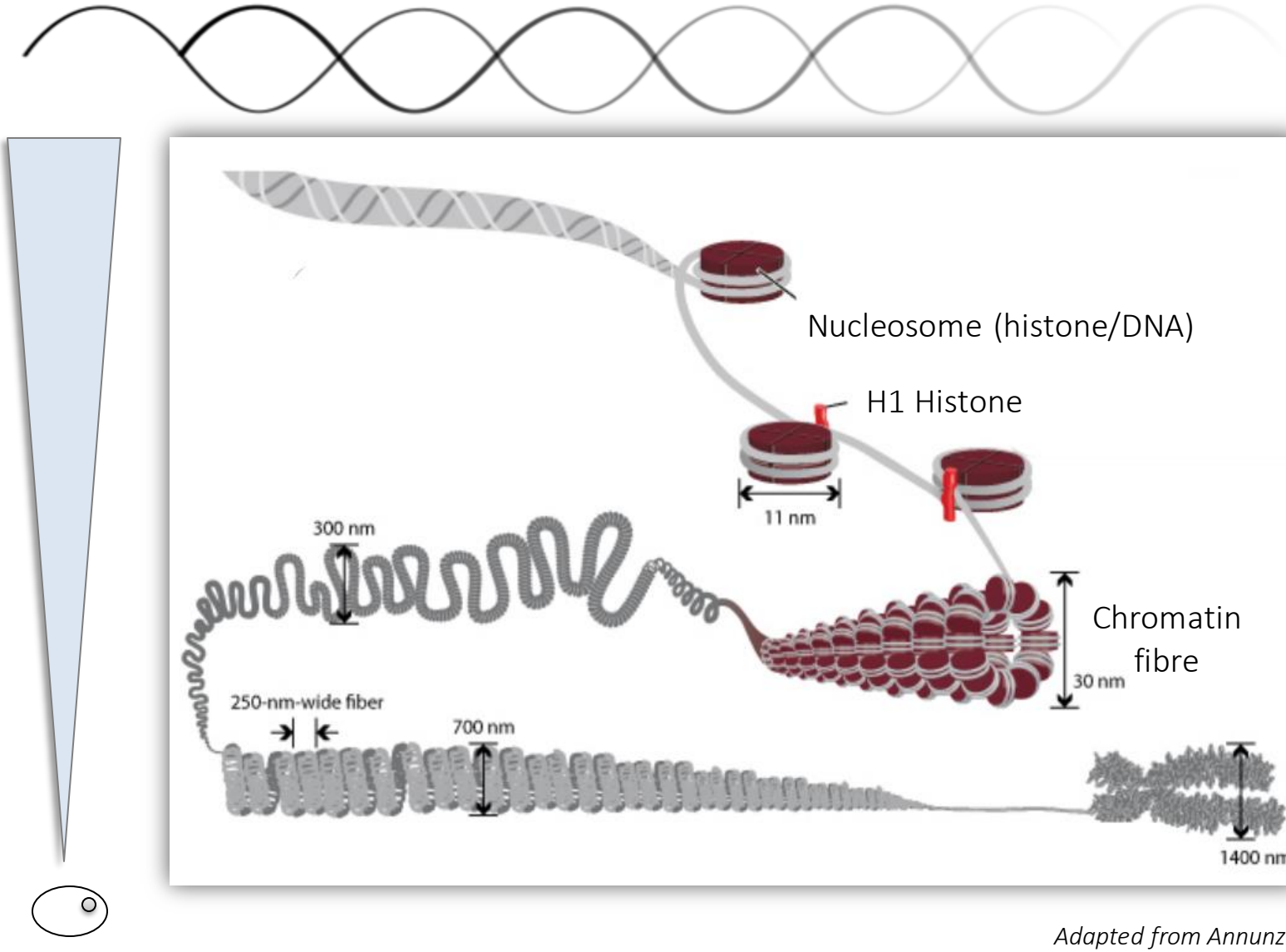
Efficient processing of Hi-C data and application to cancer

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

04th of Decembre 2019
Hi-C days, Toulouse

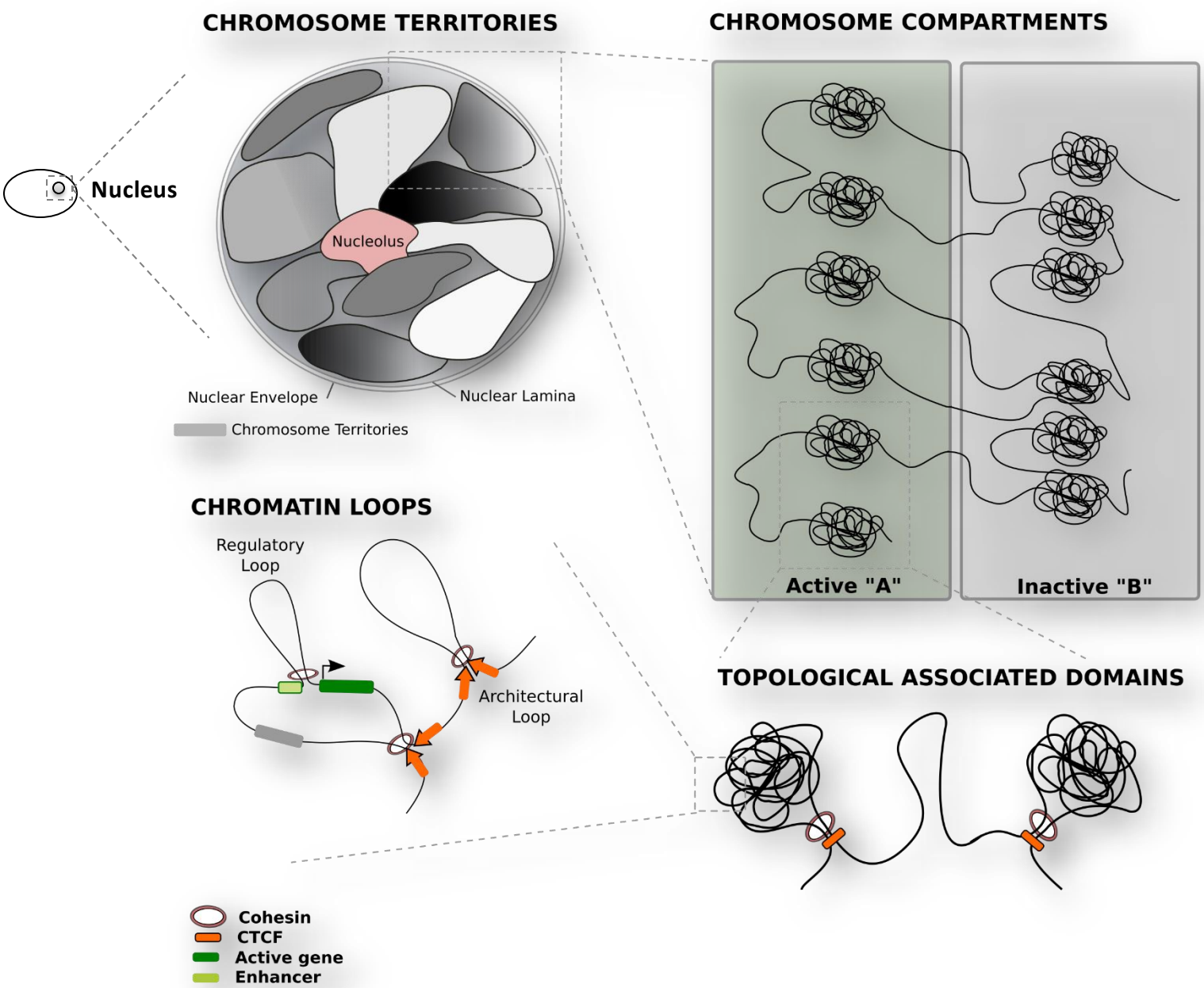
Spatial organization of the genome

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

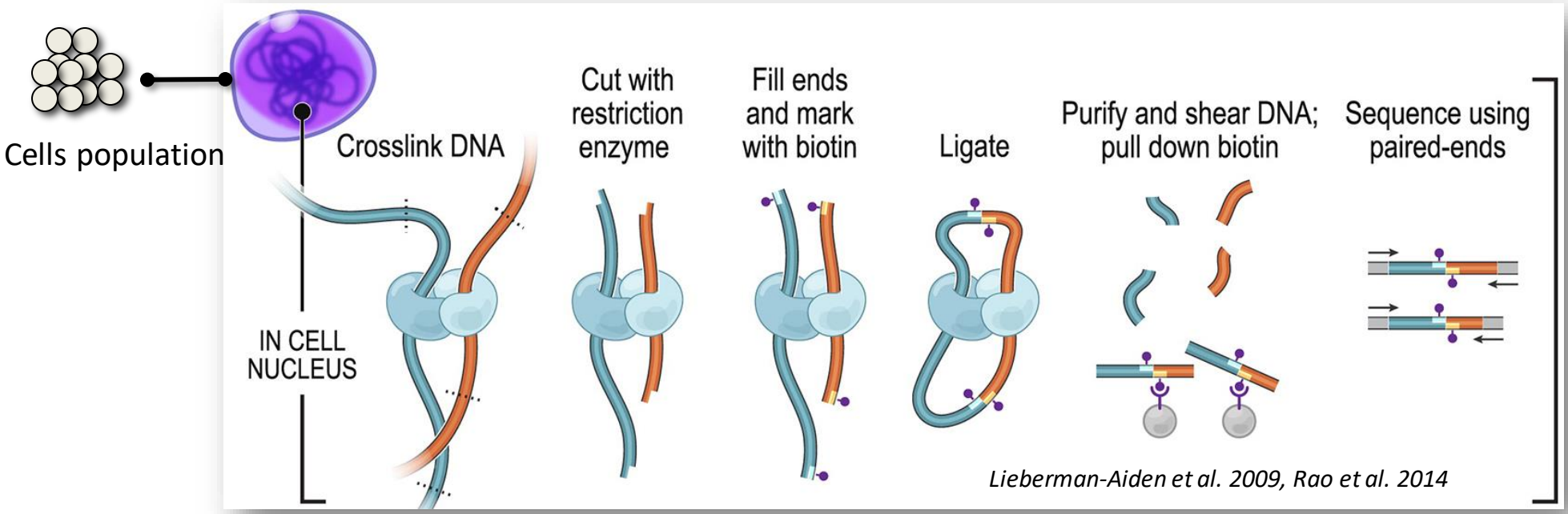


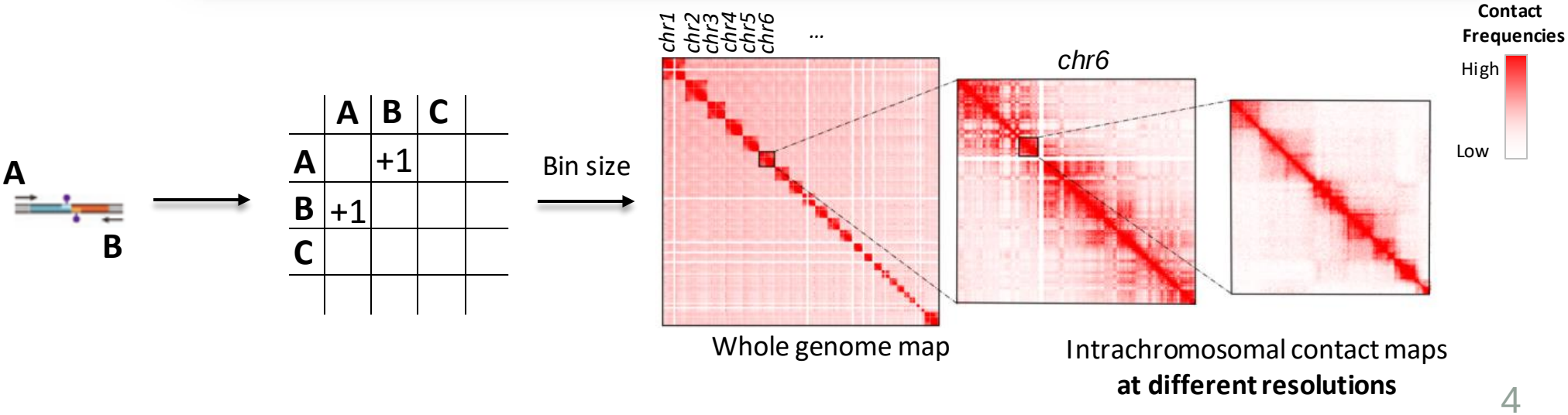
Adapted from Annunziato et al. 2008

Different levels of spatial organization

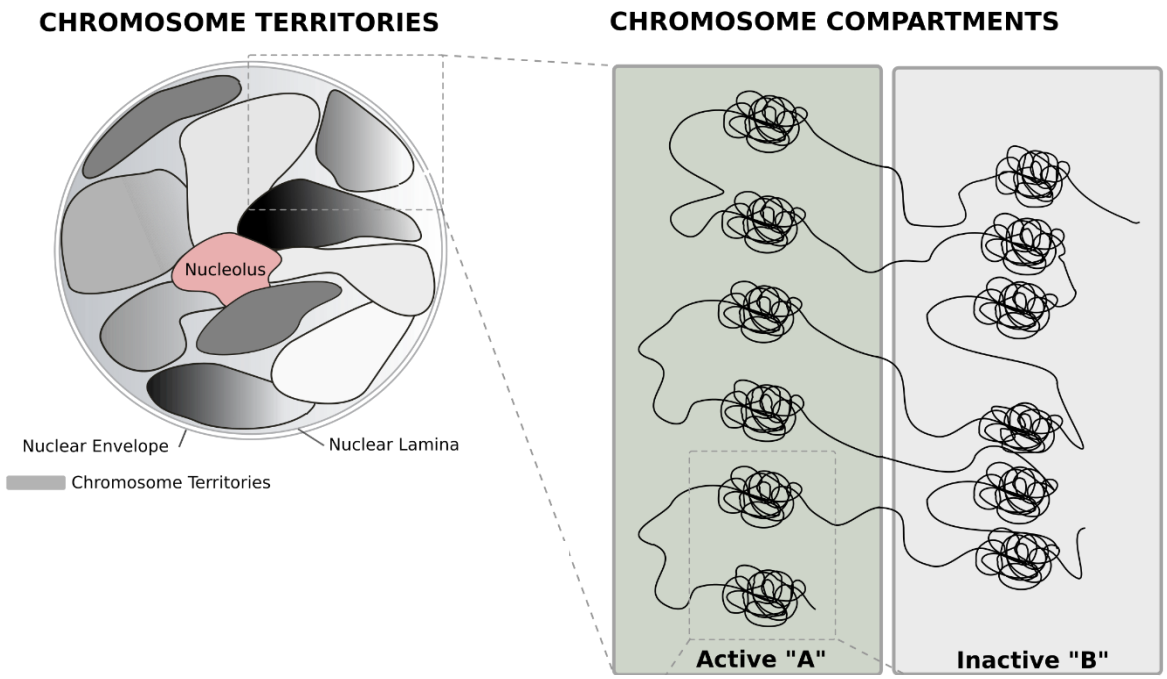


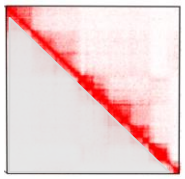
Hi-C captures the chromatin conformation within the nucleus

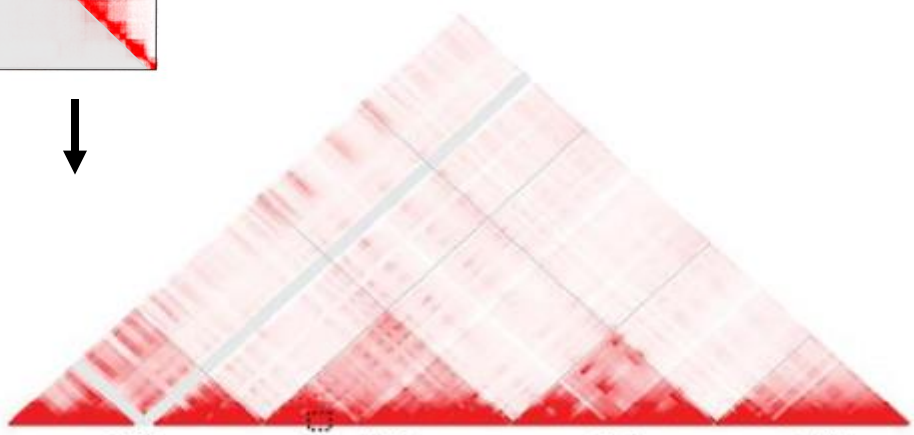




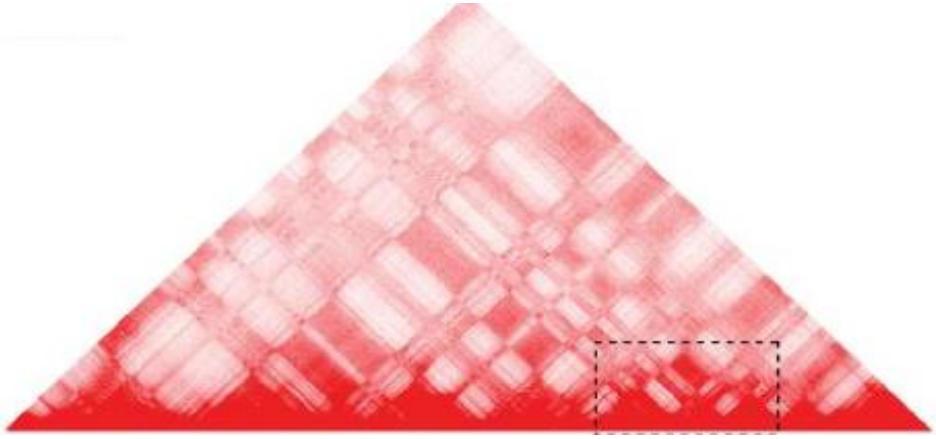
Genome organization and Hi-C







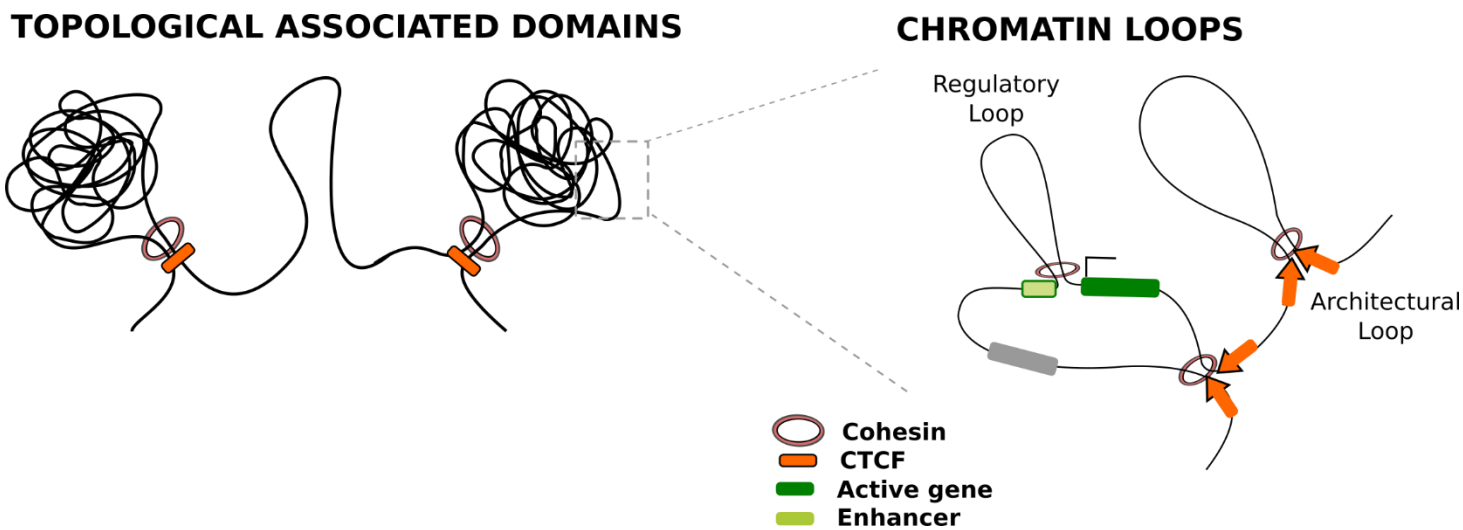
chr1 chr2 chr3 chr4

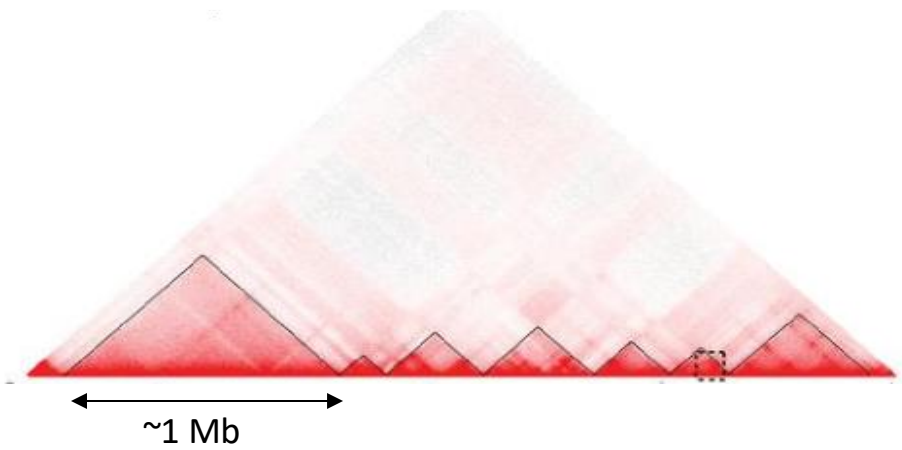


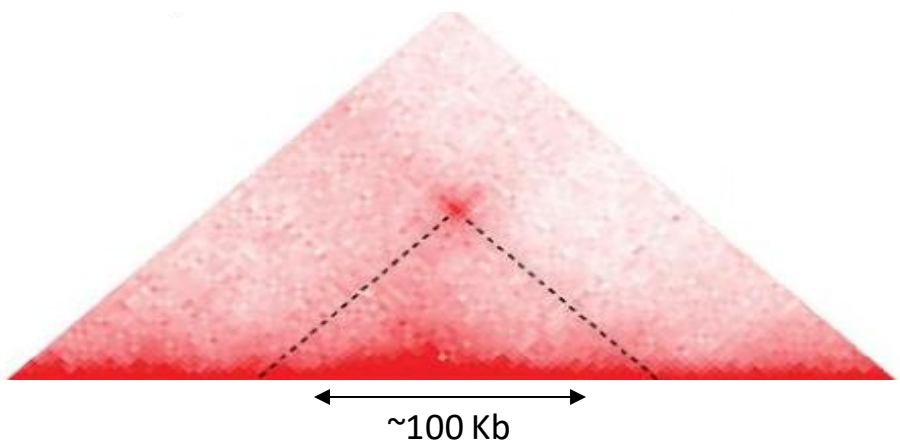
chr2

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

Genome organization and Hi-C



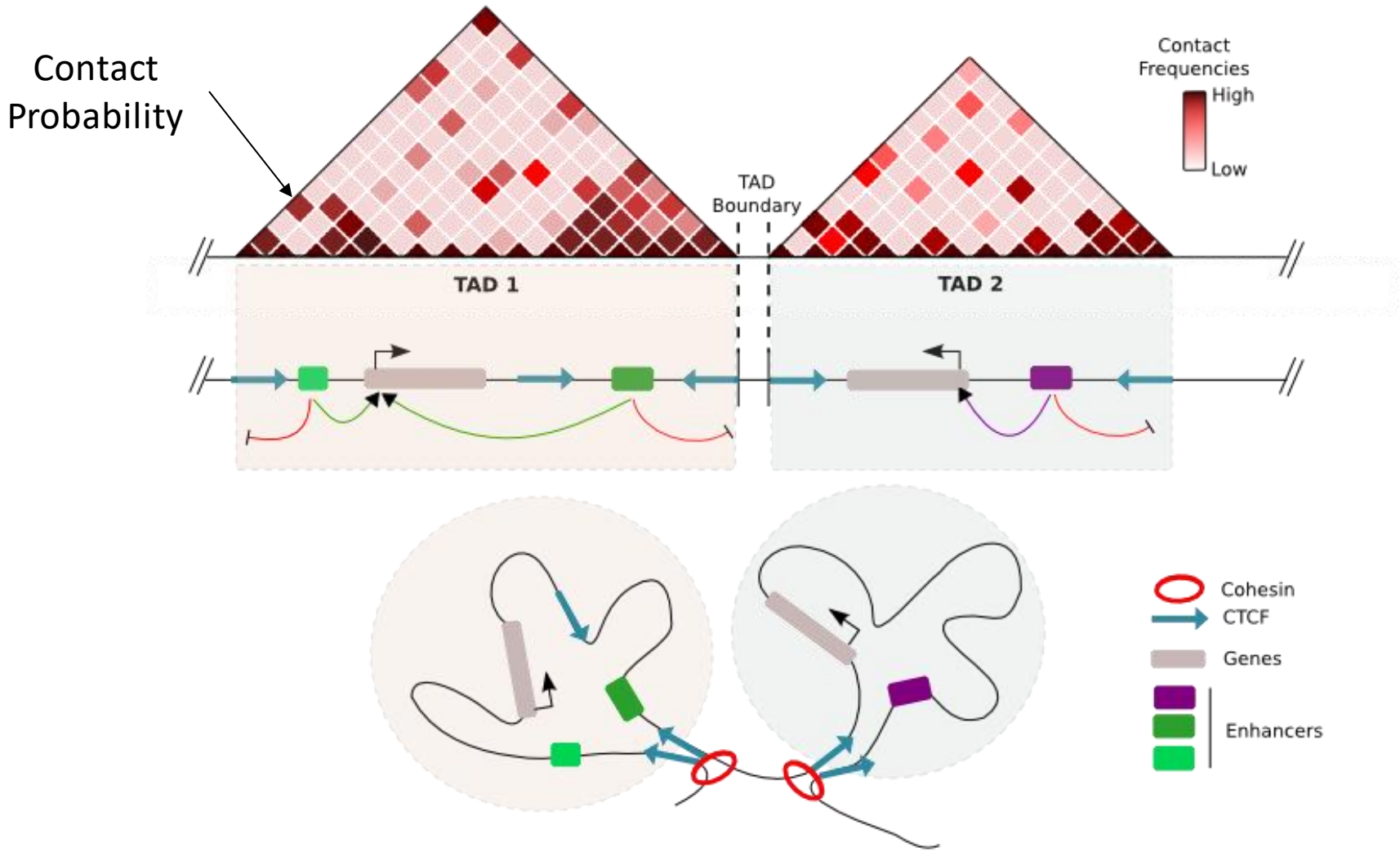




Lieberman-Aiden et al. 2009, Dixon et al. 2012, Rao et al. 2014

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.



'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)*

Ready-to-use Hi-C Kits

https://www.qiagen.com

https://arimagenomics.com/kit

https://www.phasegenomics.com

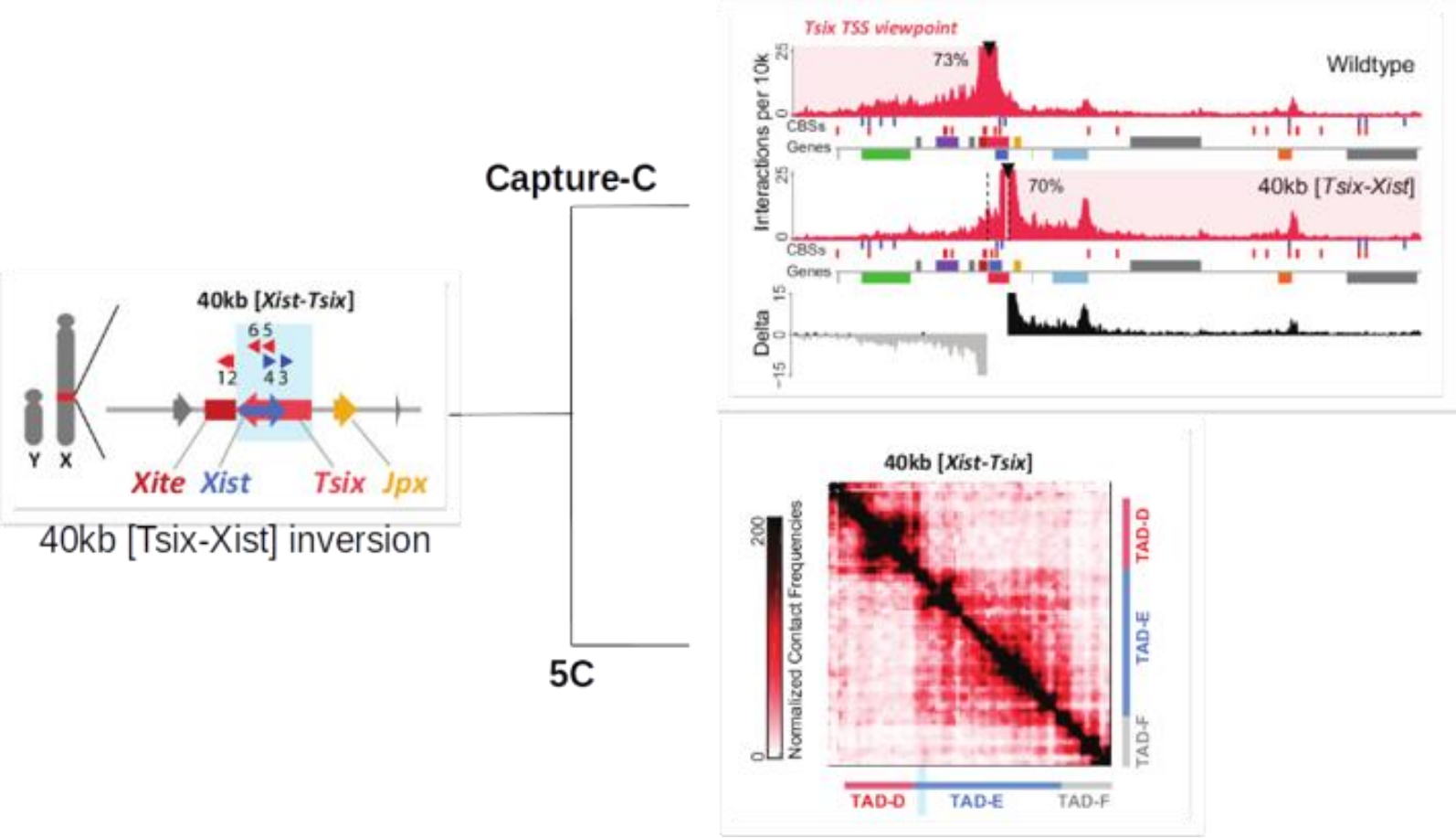
Which approach for which purpose ?

Nat Genet. 2019 Jun;51(6):1024-1034. doi: 10.1038/s41588-019-0412-0. Epub 2019 May 27.

The bipartite TAD organization of the X-inactivation center ensures opposing developmental regulation of Tsix and Xist.

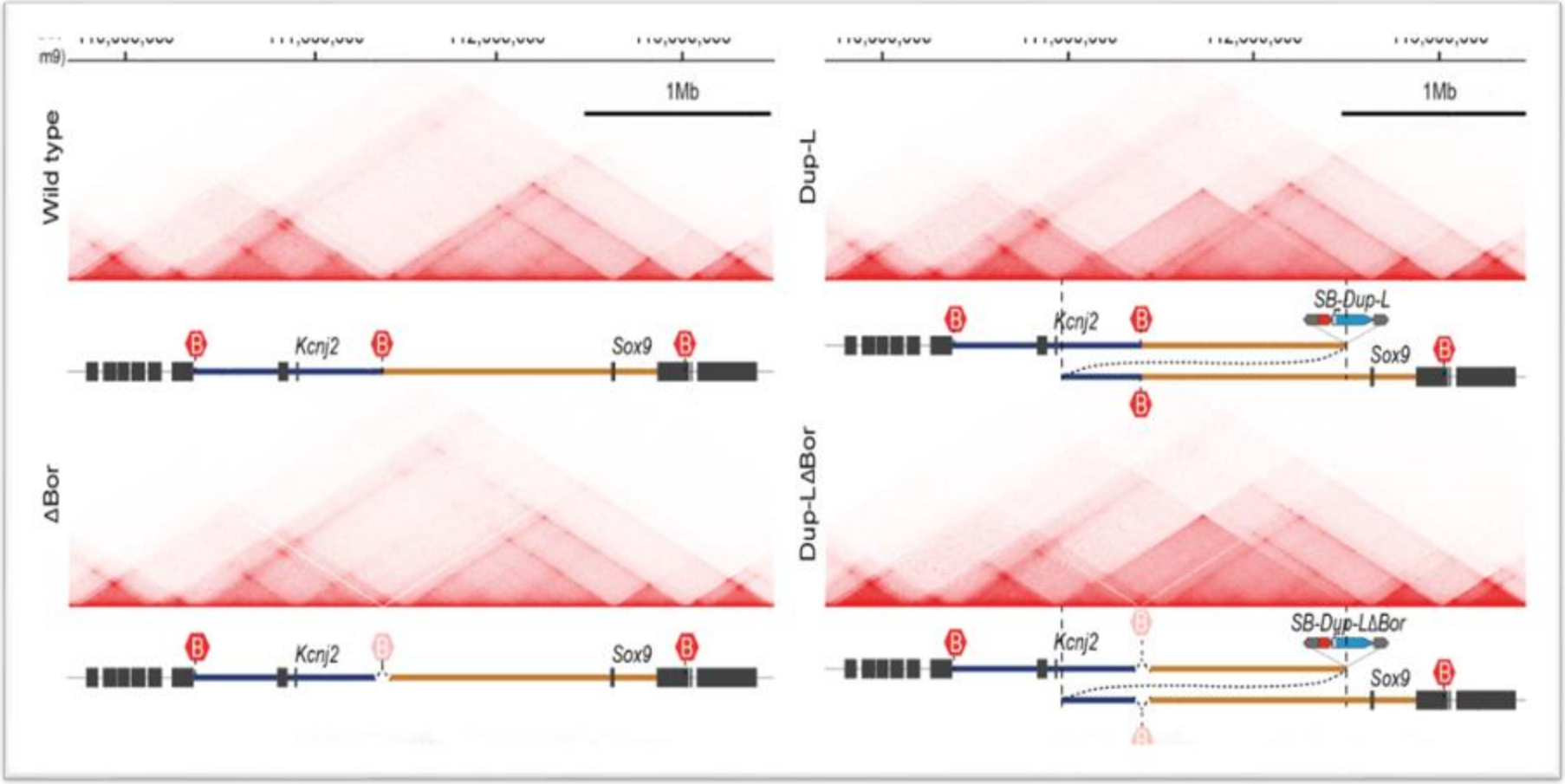
van Bemmel JG[1,2,3], Galupa R[4,5], Gard C[4], Servant N[6,7], Picard C[4], Davies J[8], Szempruch AJ[9], Zhan Y[10,11], Żylicz JJ[4,12], Nora EP[13], Lameiras S[14], de Wit E[15], Gentien D[16], Baulande S[14], Giorgetti L[10], Guttman M[9], Hughes JR[8], Higgs DR[8], Gribnau J[17], Heard E[18].

Author information



Which approach for which purpose ?

Capture Hi-C protocol (*Franck et al. 2016*)
i.e. Hi-C library combined with capture of a dedicated genomic region



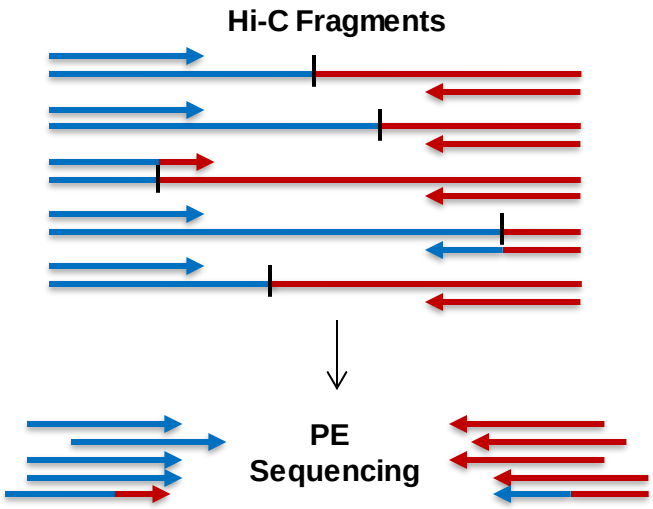
Questions ?

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

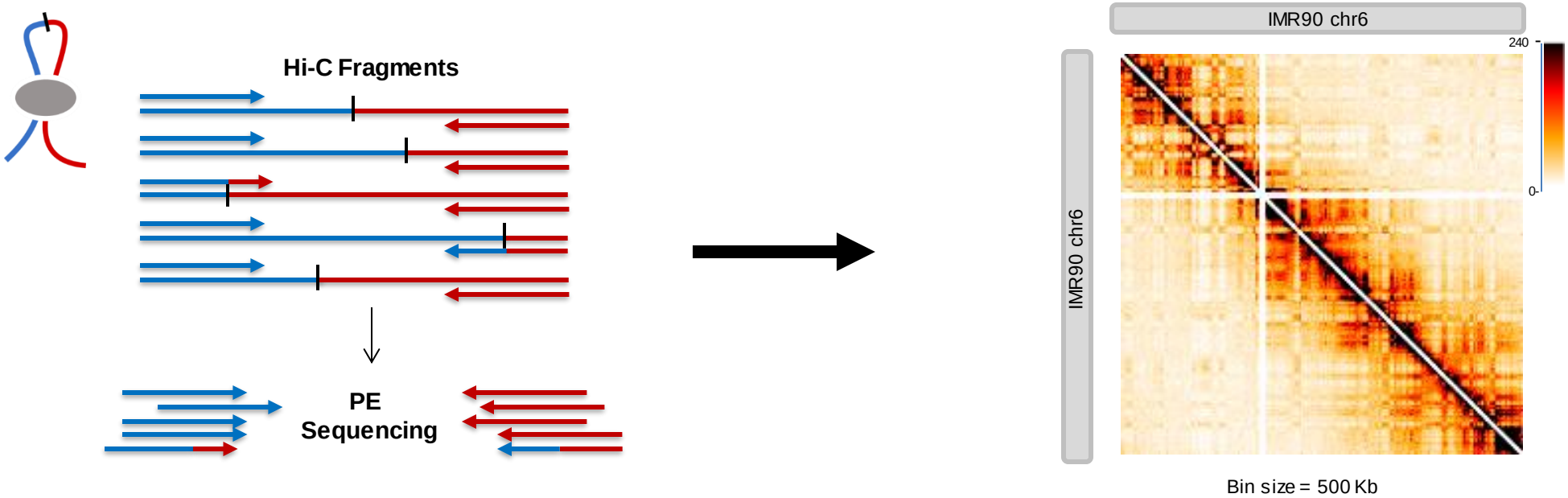


A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin

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

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

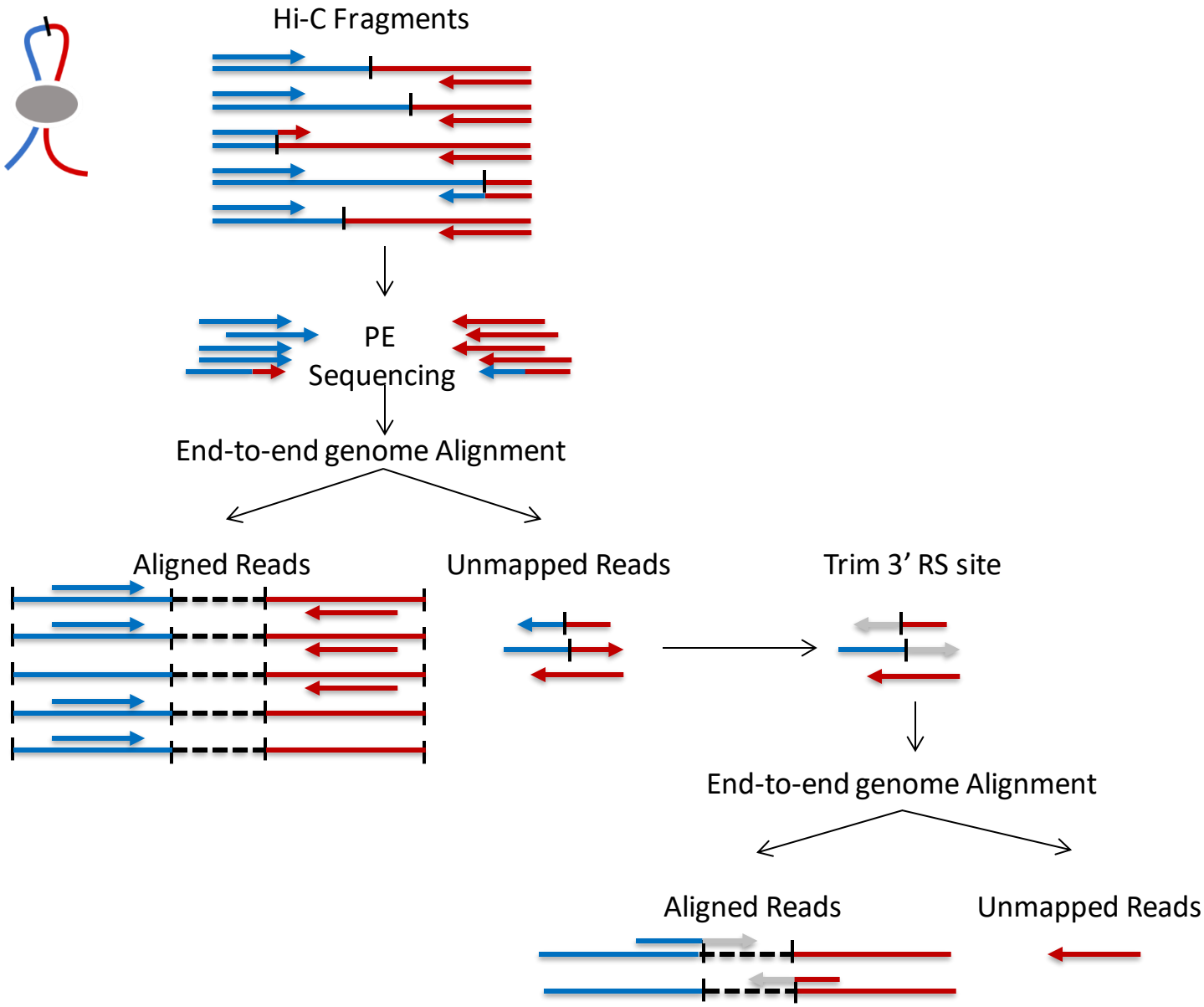
Challenges in Hi-C data processing



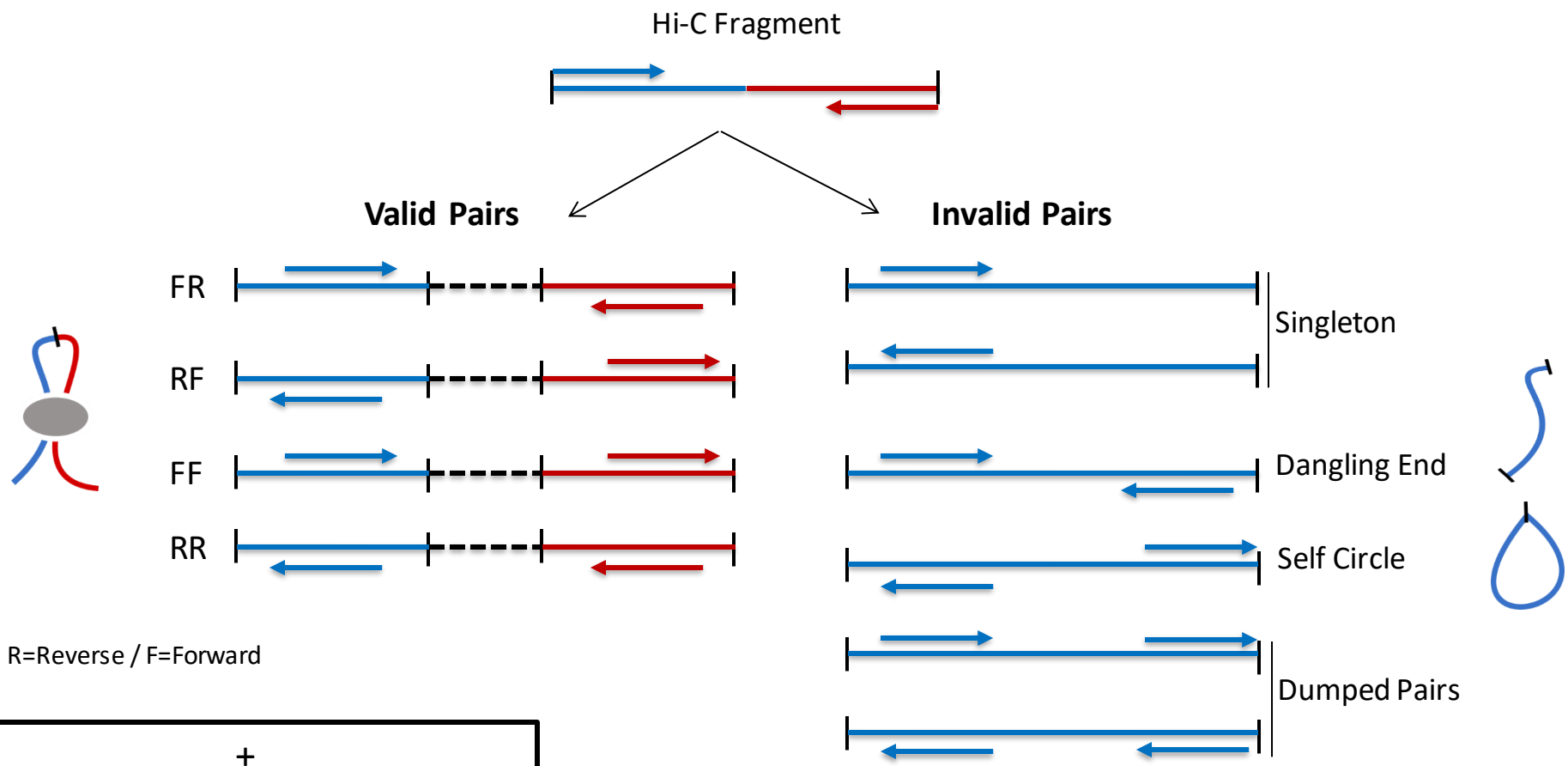
How to process Hi-C data in an **easy** and **efficient** way taking into account ;

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

Reads mapping strategy



Detection of valid interaction products



+

Filtering on :

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

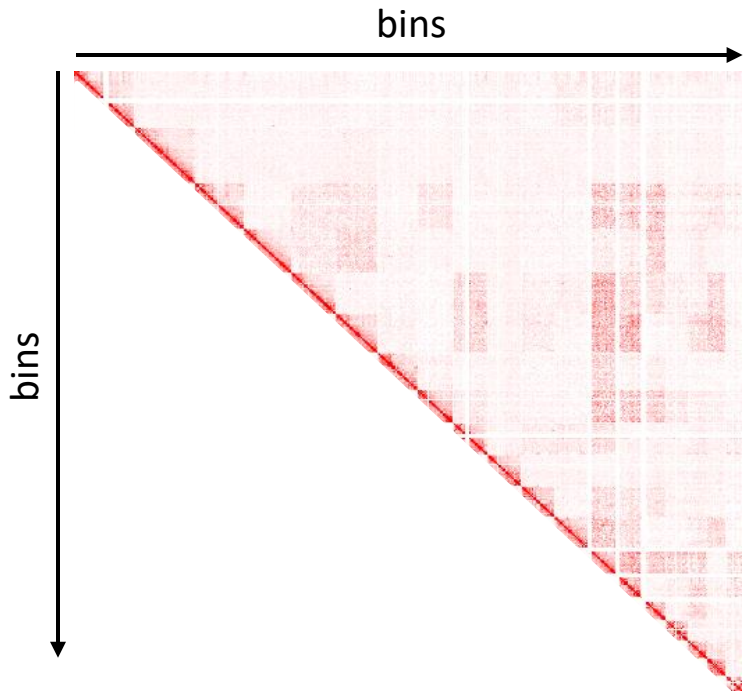
Building contact maps

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 formats

.hic files (Juicer, Juicebox)

- Contact matrices in multiple resolutions and summary statistics stored in one file
- Java and C bindings
- Command line tools
- Extant suite of analysis tools
- Extant visualization tool.

.cool files (cooler, higlass)

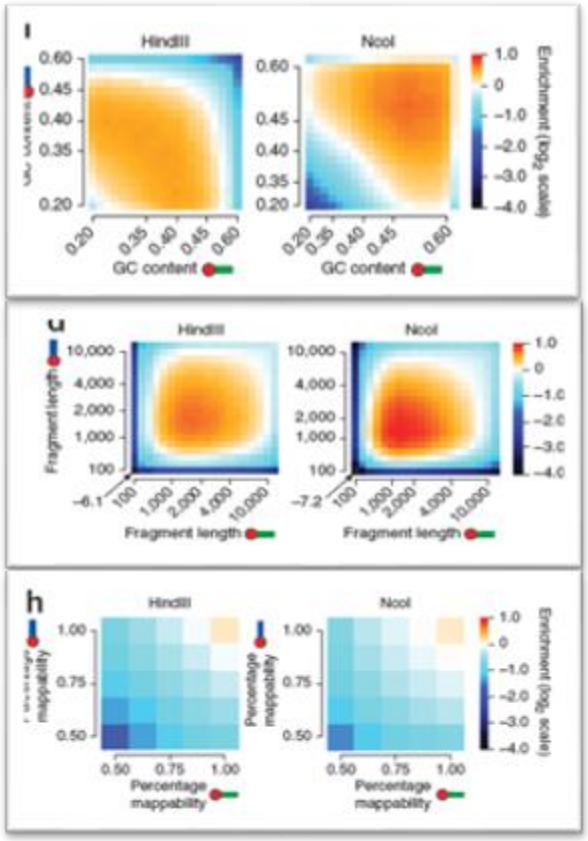
- Flexibility to store one or multiple matrices with varying bin sizes
- python library
- Command line tools
- HDF5, which has native bindings in practically all languages
- out of memory iterative matrix balancing, that can work on very large matrices.

Hi-C data normalization

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

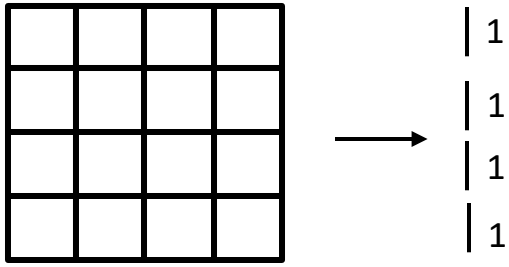
The iterative correction (ICE) method is **a widely used** approach for Hi-C data normalization.

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]



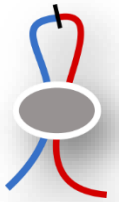
as $\Sigma_{i,\, i \neq j,\, j \pm 1} T_{ij} = 1$ for each region j.

HiC-Pro – processing of Hi-C/HiChiP data

Dedicated Reads Mapping Strategy

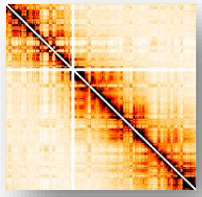
→

Detection of Valid Interaction Products



→

Generates raw and normalized contact maps



→

Quality Controls

MultiQC

- Easy-to-use
- Optimized and scalable
- Flexible
- Support most protocols
- Open to contribution
- Compatible with many downstream analysis software

Highly used in the last years

Available at https://github.com/nservant/HiC-Pro
Forum and discussion at
https://groups.google.com/forum/#!forum/hic-pro

Dimensions Badge

310

310	Total citations
248	Recent citations
149	Field Citation Ratio
7.48	Relative Citation Ratio

Building Efficient and Reproducible Workflows

For facilities
Highly optimised pipelines with excellent reporting.
Validated releases ensure reproducibility.

For users
Portable, documented and easy to use workflows.
Pipelines that you can trust.

For developers
Companion templates and tools help to validate your code and simplify common tasks.

nf-core

A community effort to collect a curated set of analysis pipelines built using Nextflow.

Analysis pipelines:

- Nextflow-based pipelines
- High level of reproducibily
- Strict Guidelines
- 17 released pipelines
- 19 under development

Community:

29 organisations over the world
More than 90 contributors

GitLab nextflow Jenkins

nf-core/hic

Analysis of Chromosome Conformation Capture data (Hi-C).

build passing | nextflow ≥19.04.0

install with bioconda | docker build manual | singularity available

DOI 10.5281/zenodo.2669513

Introduction

This pipeline is based on the HiC-Pro workflow. It was designed to process Hi-C data from raw fastq files (paired-end Illumina data) to normalized contact maps. The current version supports most protocols, including digestion protocols as well as protocols that do not require restriction enzymes such as DNase Hi-C. In practice, this workflow was successfully applied to many data-sets including dilution Hi-C, in situ Hi-C, DNase Hi-C, Micro-C, capture-C, capture Hi-C or HiChip data.

The pipeline is built using Nextflow, a workflow tool to run tasks across multiple compute infrastructures in a very portable manner. It comes with docker / singularity containers making installation trivial and results highly reproducible.

First version of nf-core Hi-C pipeline released !

V1.1.0 = Nextflow HiC-Pro version

- Automatic installation
- Natively support most schedulers
- Natively compatible with conda, docker, singularity
- Efficient tasks management
- Reads can be automaticaly splitted by chuncks to speed the processing

nf-core/hic

Analysis of Chromosome Conformation Capture data (Hi-C).

build passing | nextflow ≥19.04.0

install with bioconda | docker build manual | singularity available

DOI 10.5281/zenodo.2669513

Plans for the next versions :

- TADs calling (which methods ?)
- Compartment Calling
- Detection of significant contacts
- Specific pipelines for Hi-C based assembly ? Cancer Hi-C ?

Contribution is welcome !

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 Ballar
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

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,*]

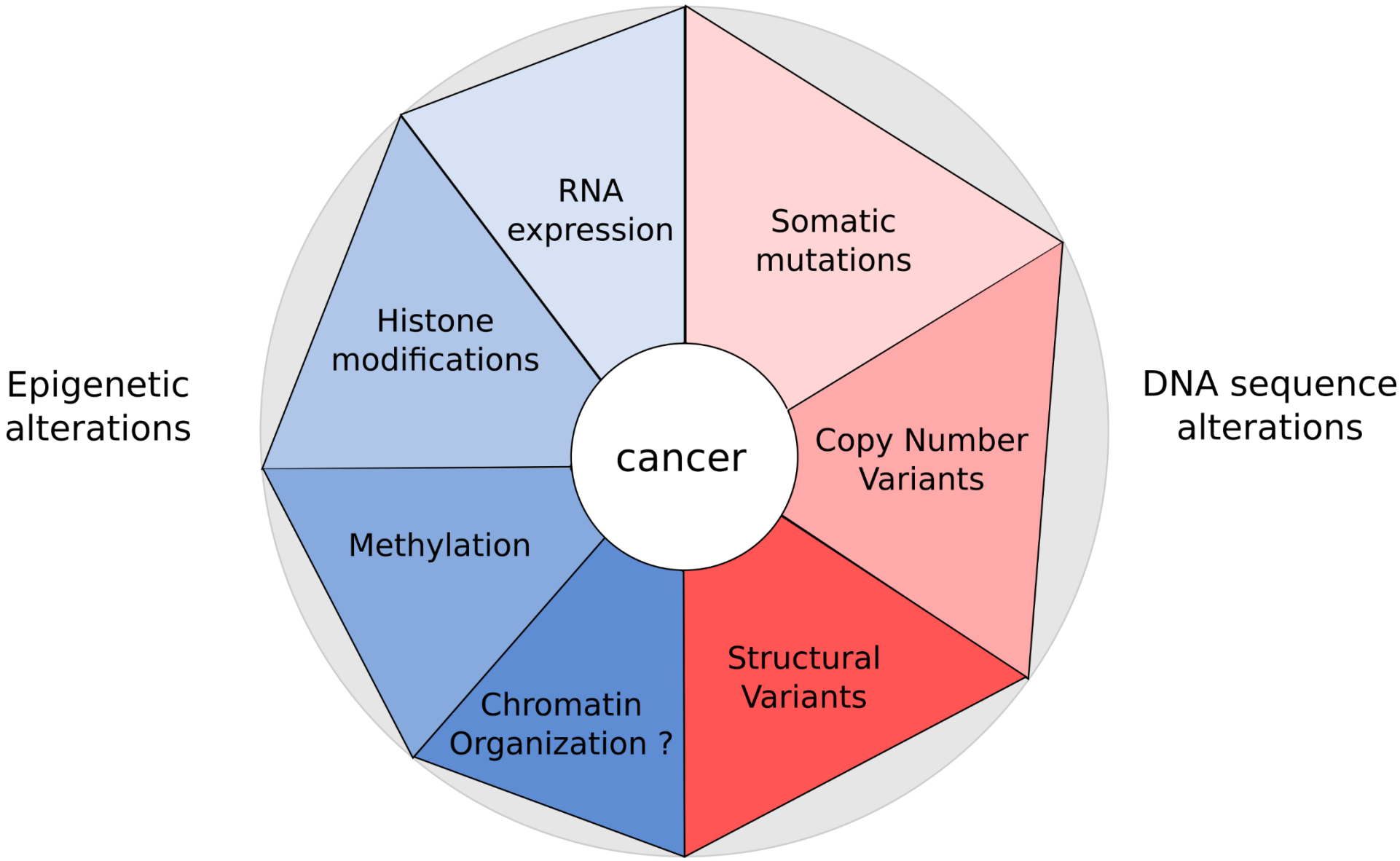
RESEARCH

CrossMark

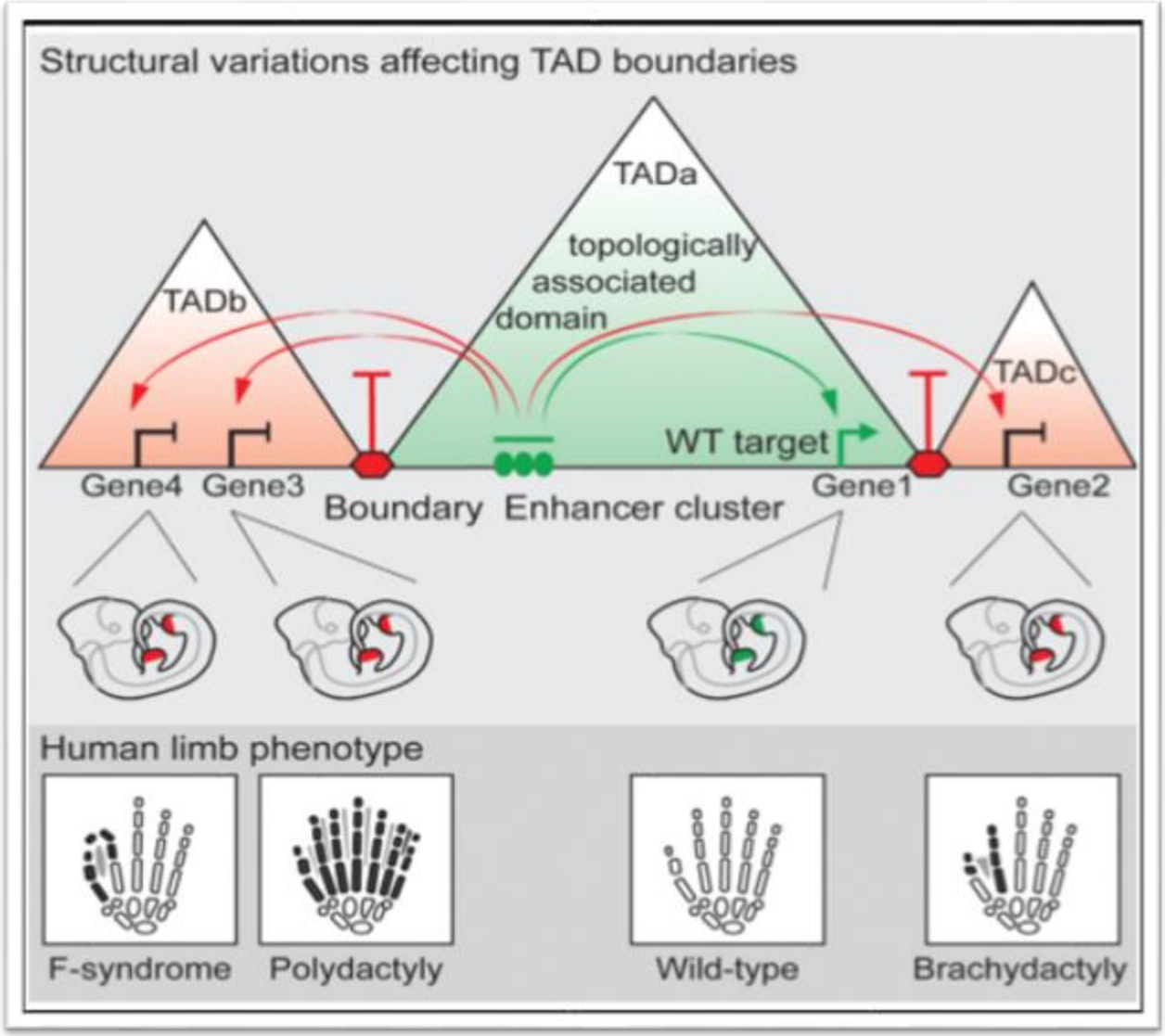
Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells

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



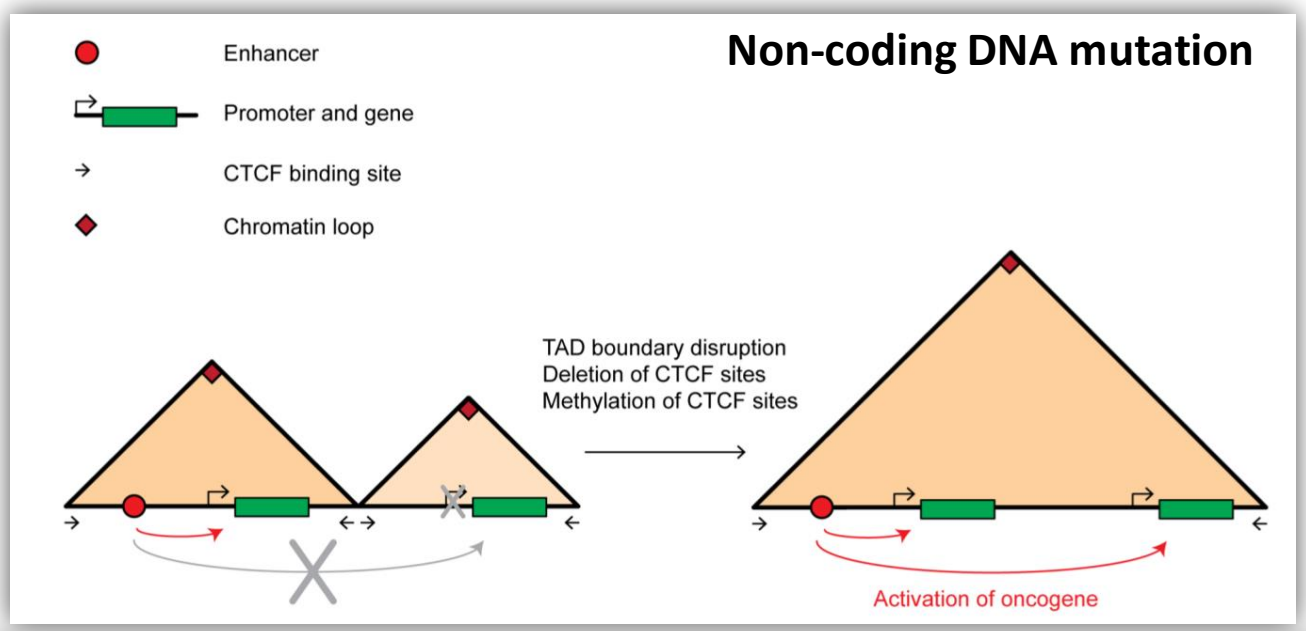
TADs are biologically relevant

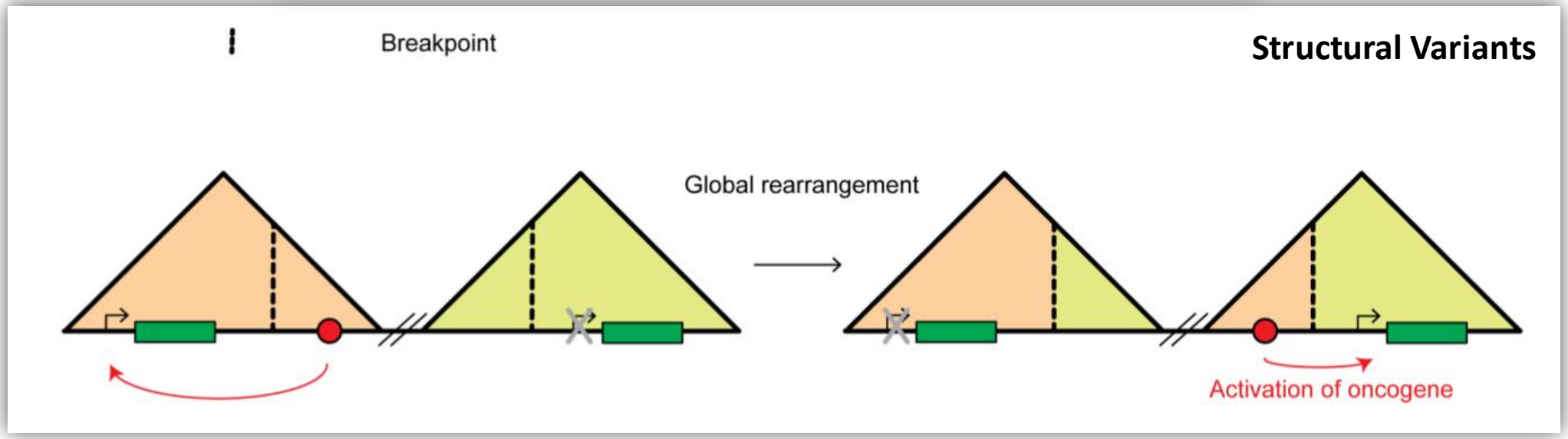


- TADs disruption leads to new enhancer/promoter contacts
- Abnormal enhancer/promoter contacts can have strong phenotypic impacts
- Structural variants can disrupt TADs structure

Luipanez et al. 2015, Franke et al. 2016

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
Received: 9 December 2016 | Accepted: 8 June 2017

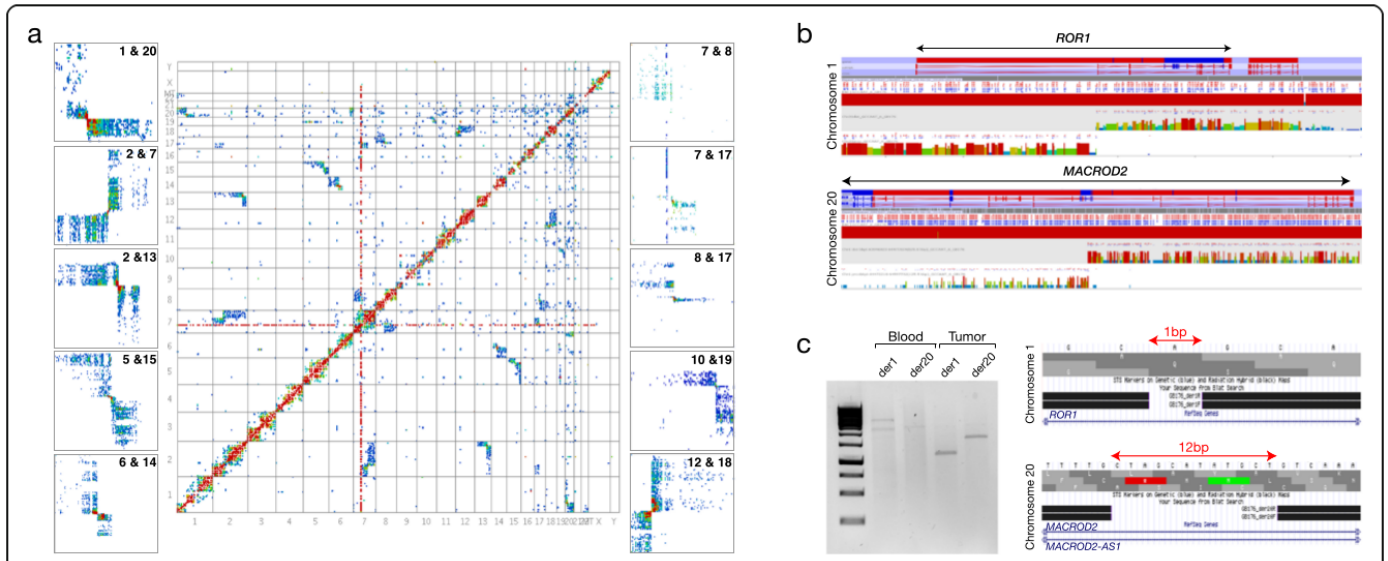
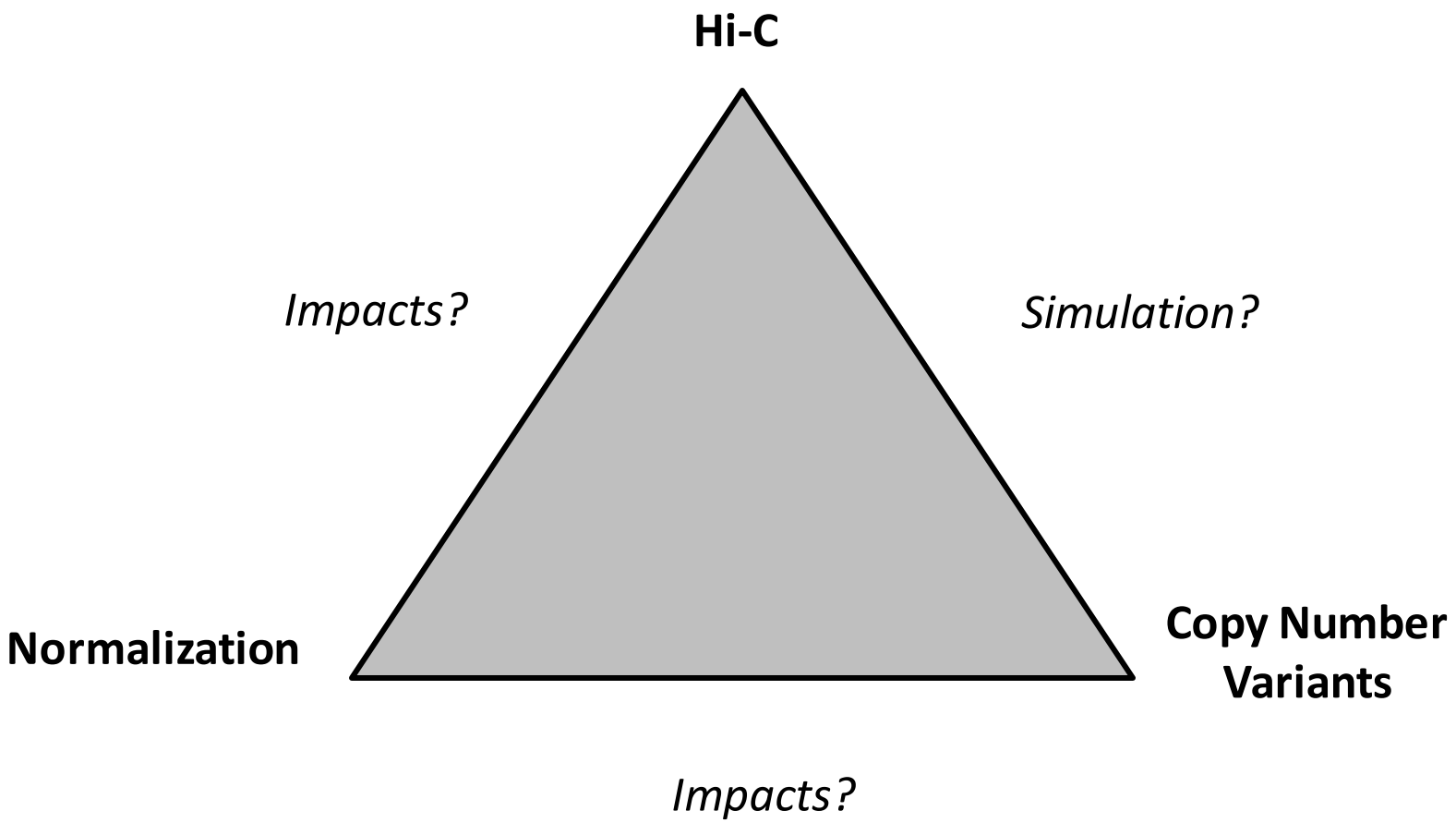
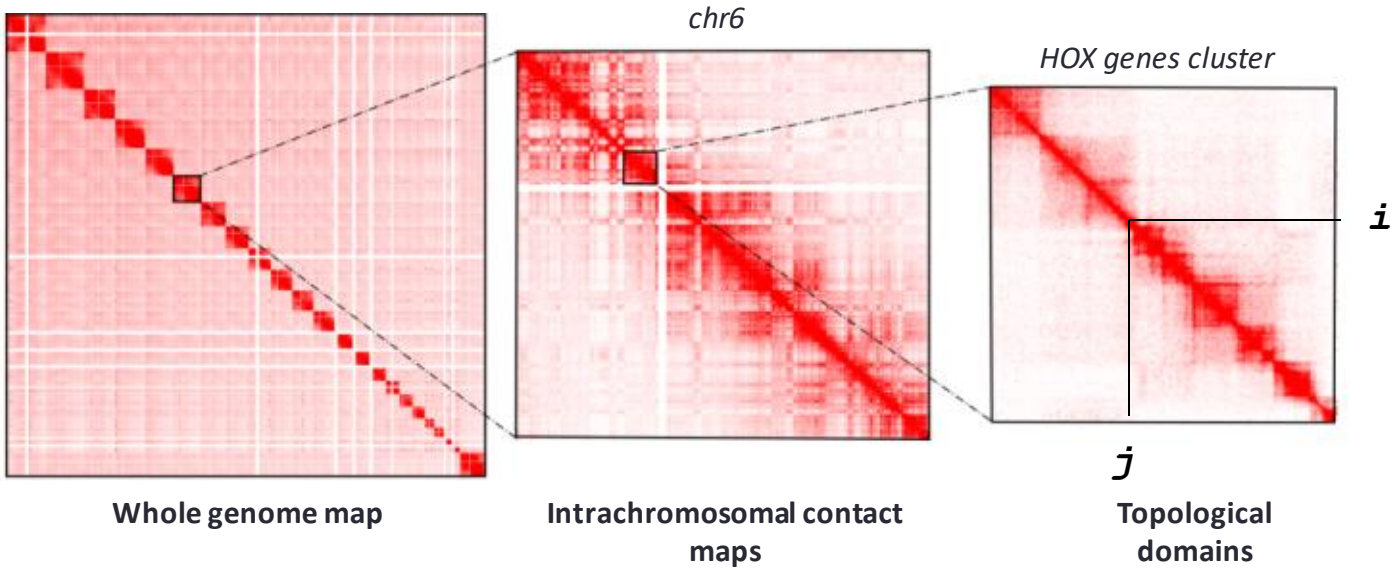


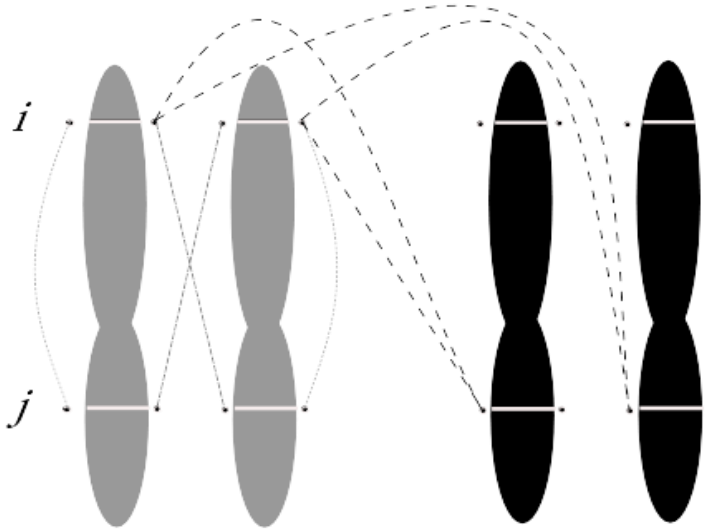
Fig. 3 Tumour GB176. **a** *Heatmap* and partial heatmaps of tumour GB176 showing some of the rearrangements present in this tumour. **b** Hi-C 'other ends' from regions distal and proximal to the suspected breakpoint on chromosome 1 (*top*) and chromosome 20 (*bottom*) showing the breakpoint regions. 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?



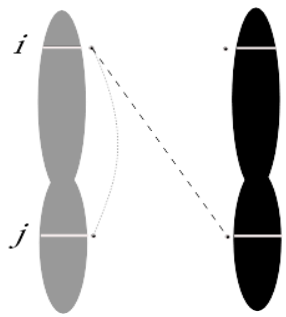


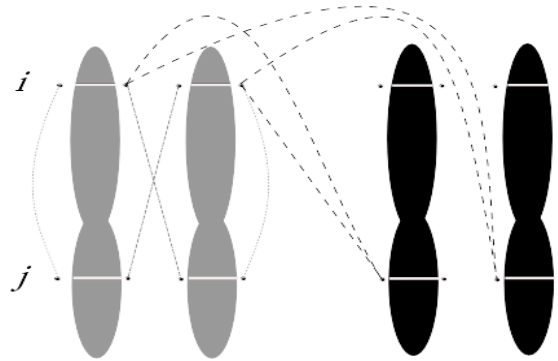
<u>In the context of a diploid genome</u>

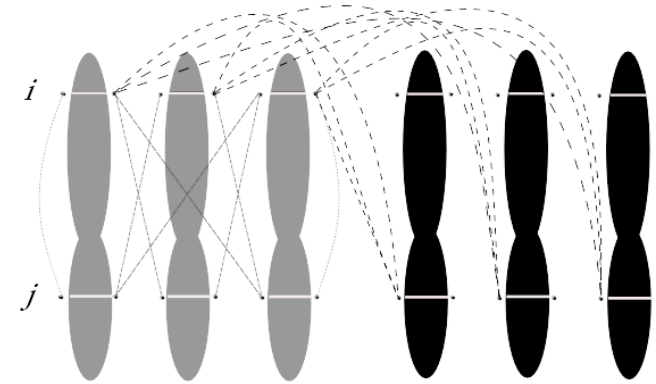
If *i* and *j* belong to the same chromosome
C_{ij}= *2 cis + 2 transH*

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

Generalization to polyploid genomes







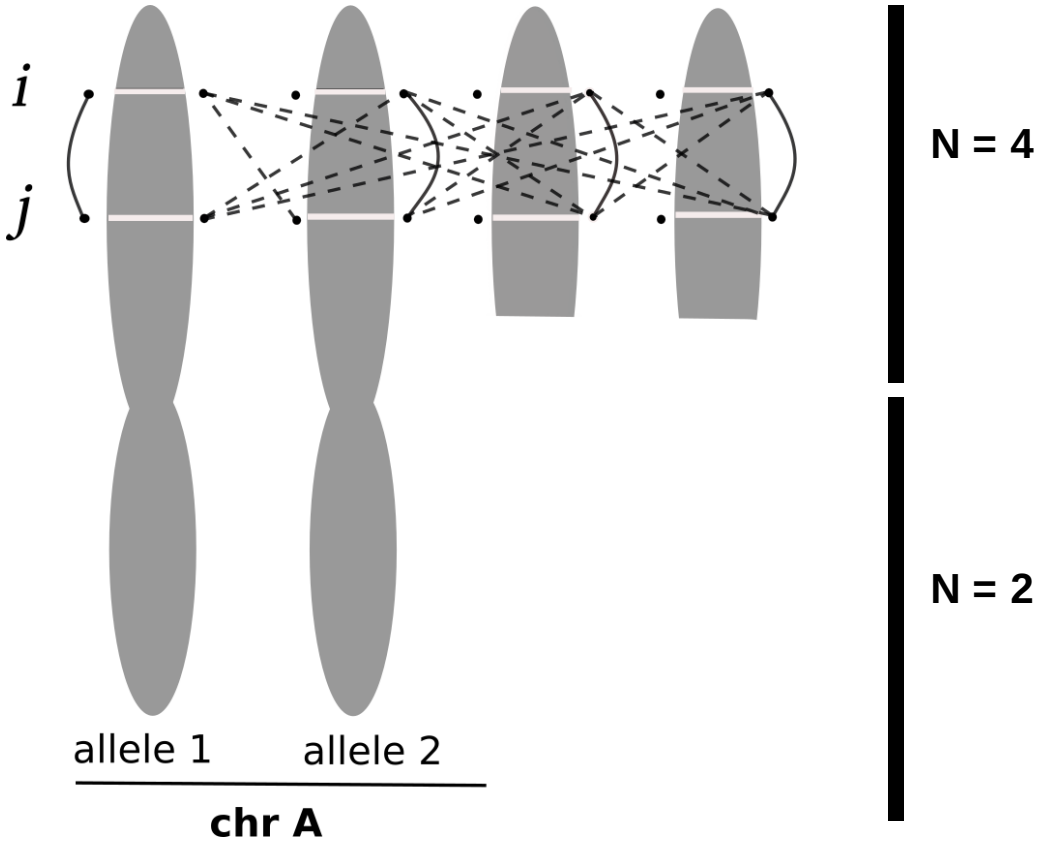
$N_i = N_j = 1$
If $chr_i = chr_j$, $C_{ij} = 1$ cis
If $chr_i \neq chr_j$, $C_{ij} = 1$ trans

$N_i = N_j = 2$
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 $chr_i = chr_j$, $\boldsymbol{C_{ij} = N_i\ cis + N_i\ (N_j - 1)\ transH}$
If $chr_i \neq chr_j$, $\boldsymbol{C_{ij} = N_i N_j\ trans}$

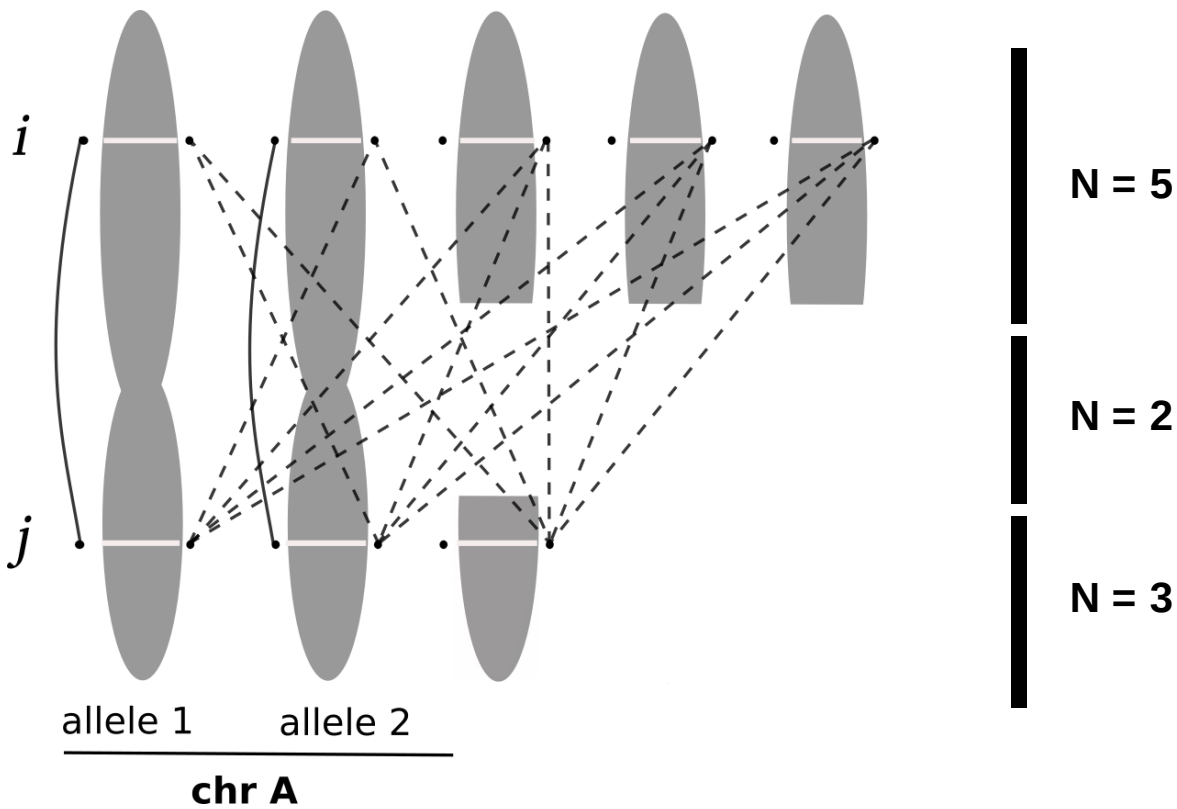
Extension to Cancer genome



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

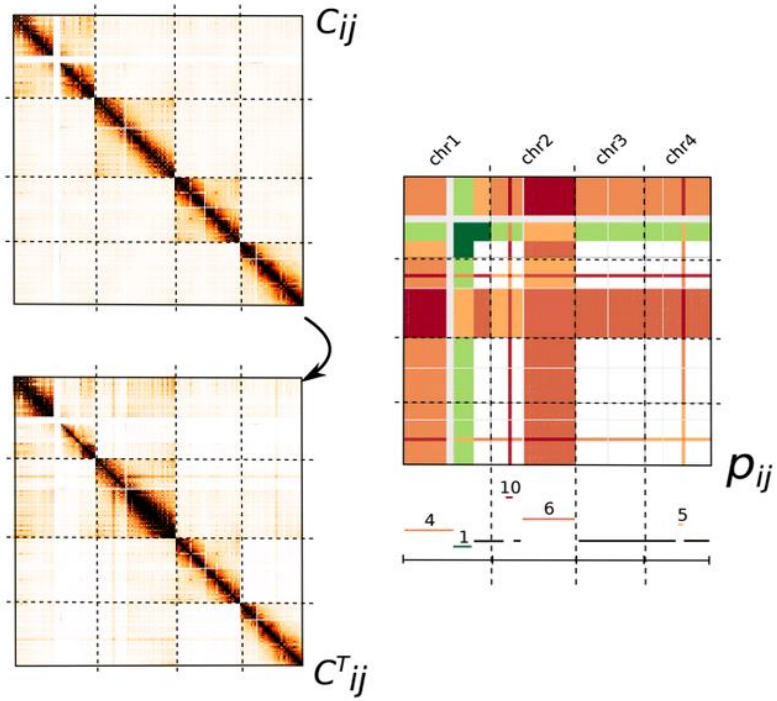
$$C_{ij} = 2\ cis_{ij} + 2\ transH$$

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

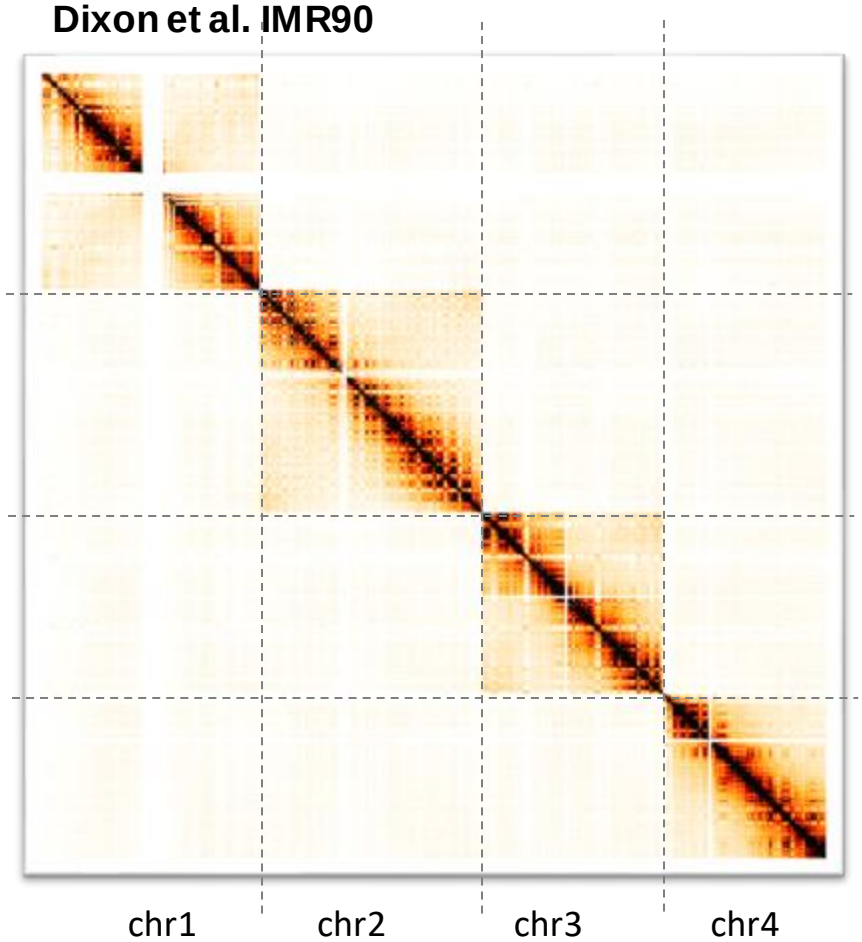
Calculate the expected factor of enrichment/depletion of interactions for the loci i and j matrix: $p_{ij} = E_{ij} / C_{ij}$

Estimate the simulated data using a binomial downsampling of parameter $C^T_{ij} \sim B(C_{ij}, p_{ij})$

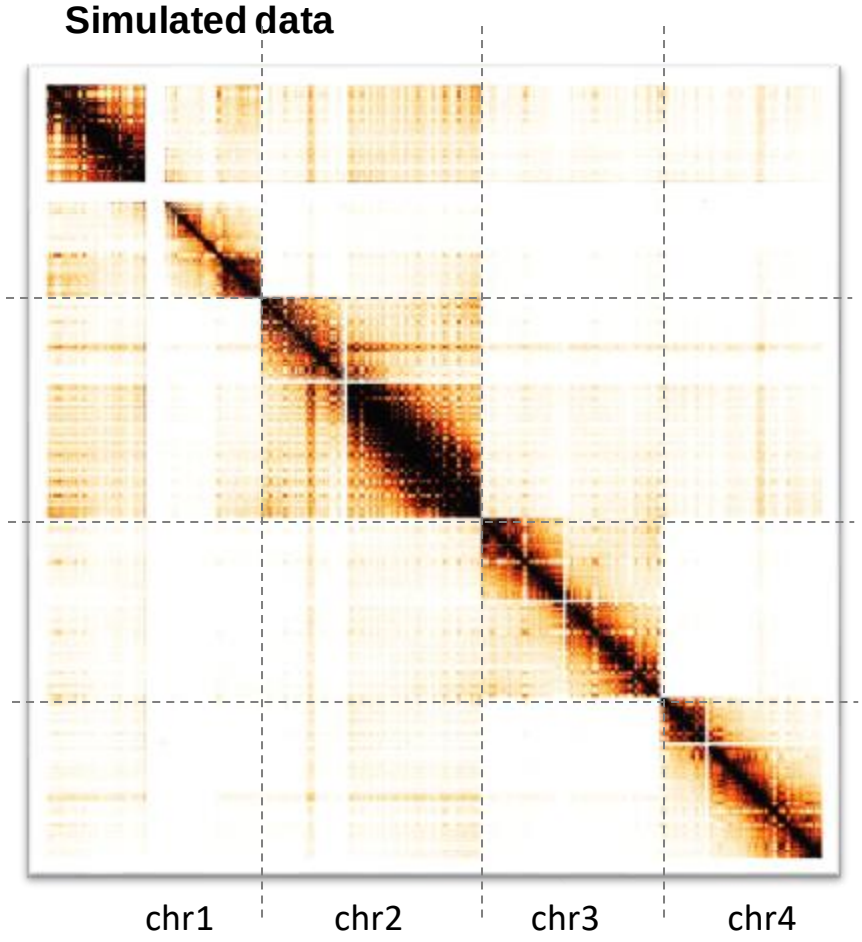


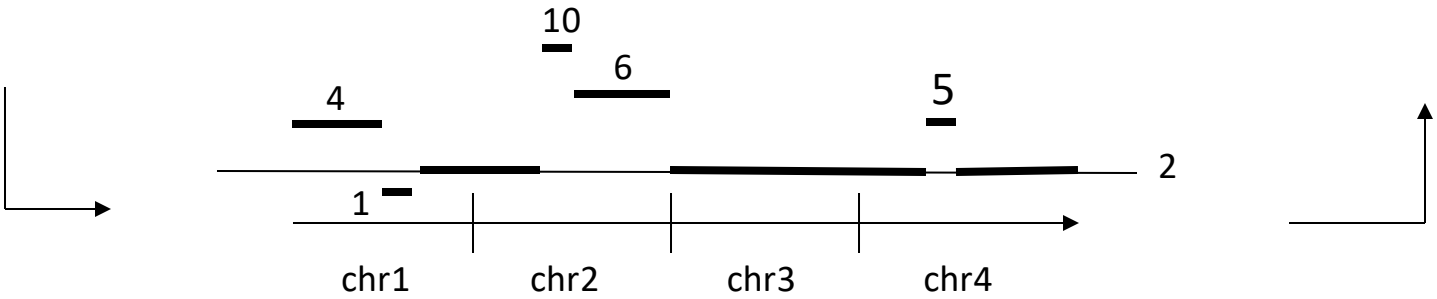
Simulation - Results

Dixon et al. IMR90



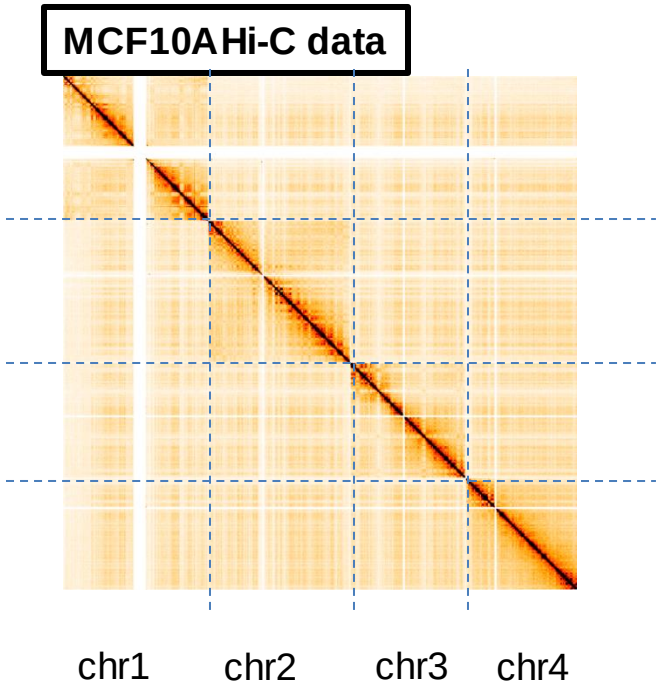
Simulated data



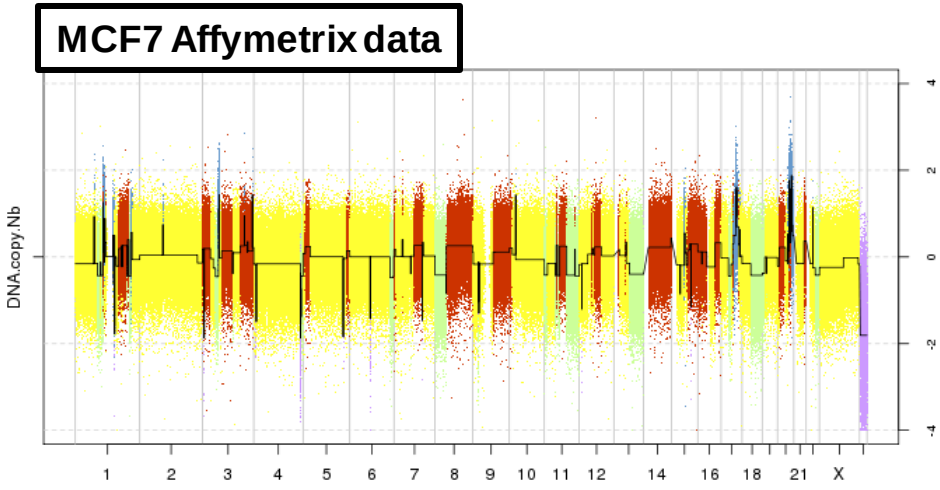


How to validate the simulation model ?

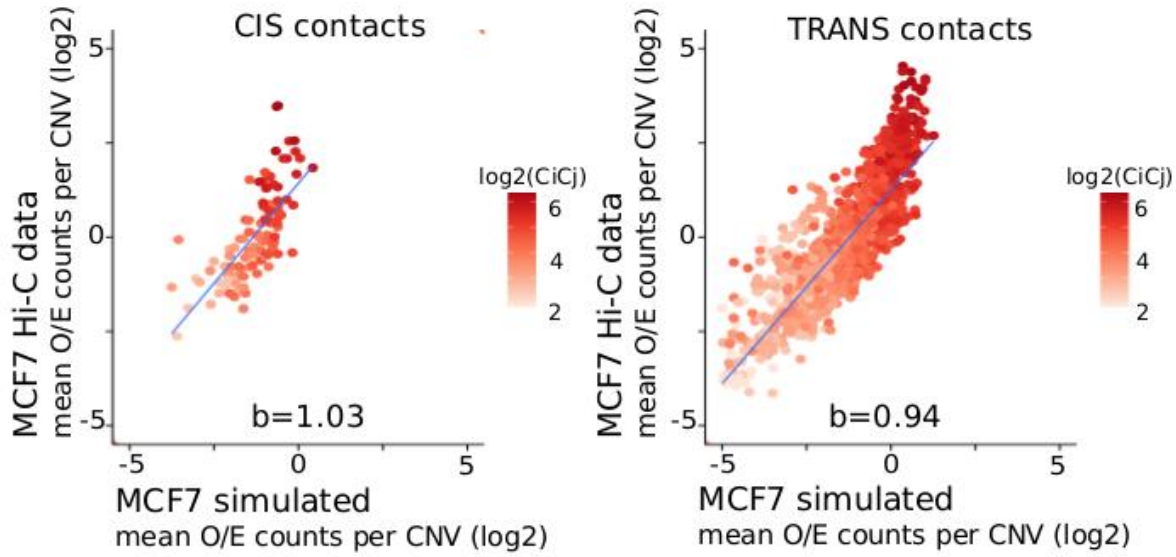
In order to validate our simulation model, we used Hi-C from MCF10 normal-like data, from which we simulated the MCF7 CNV profile

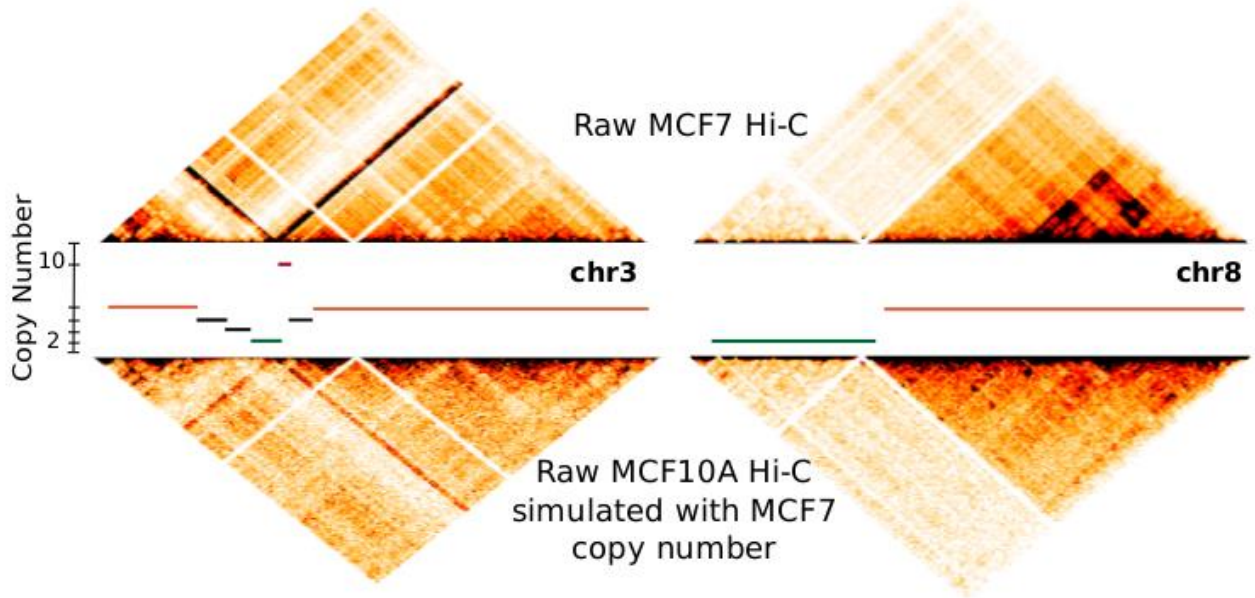


+



Simulation - Validation

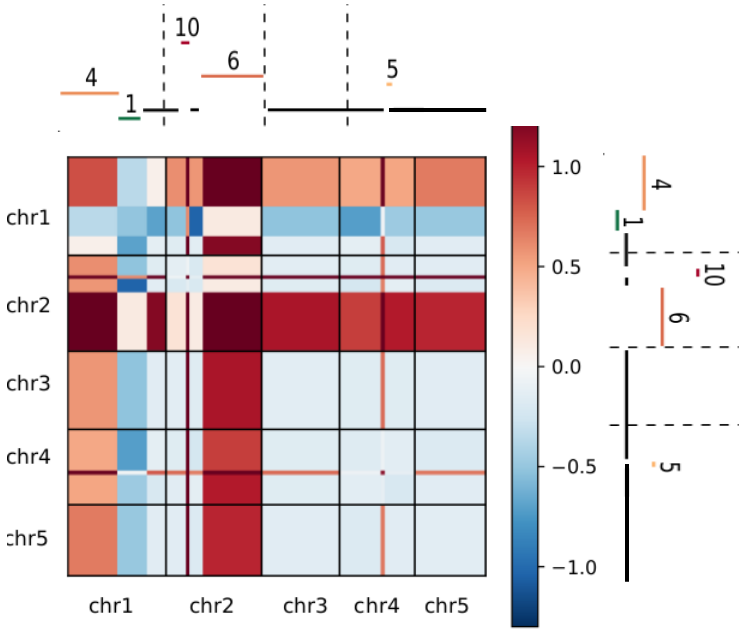




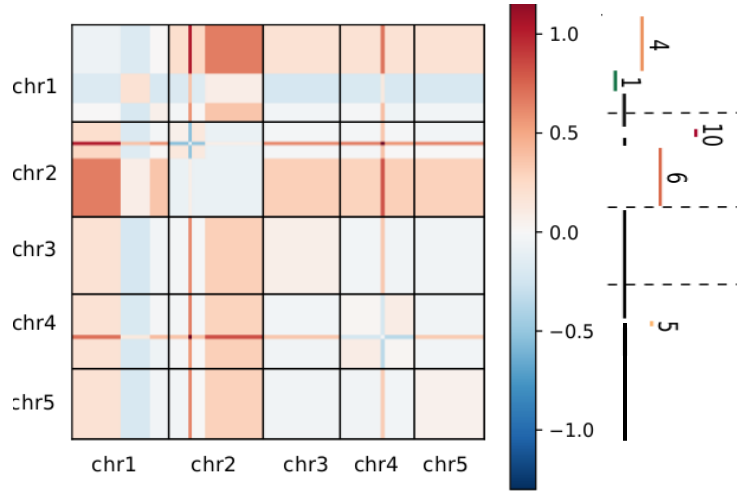
Effect of ICE normalization

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

RAW

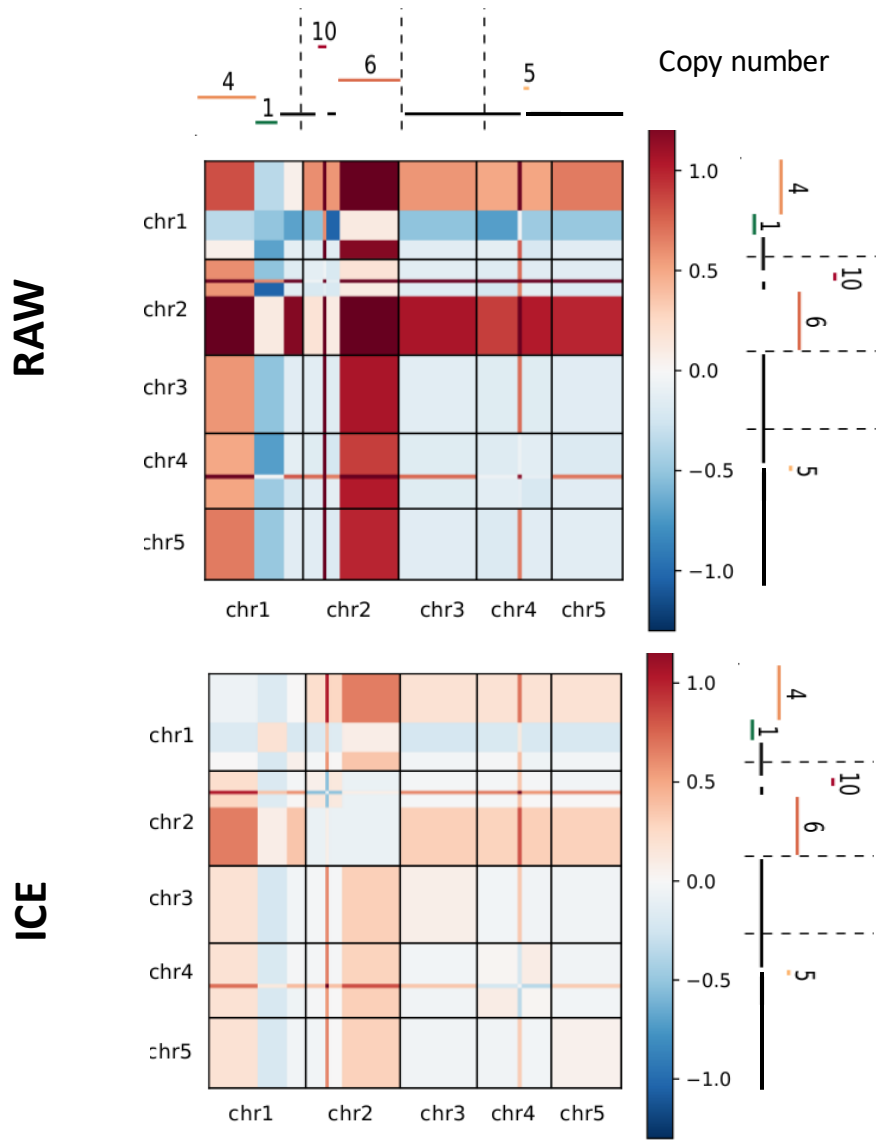


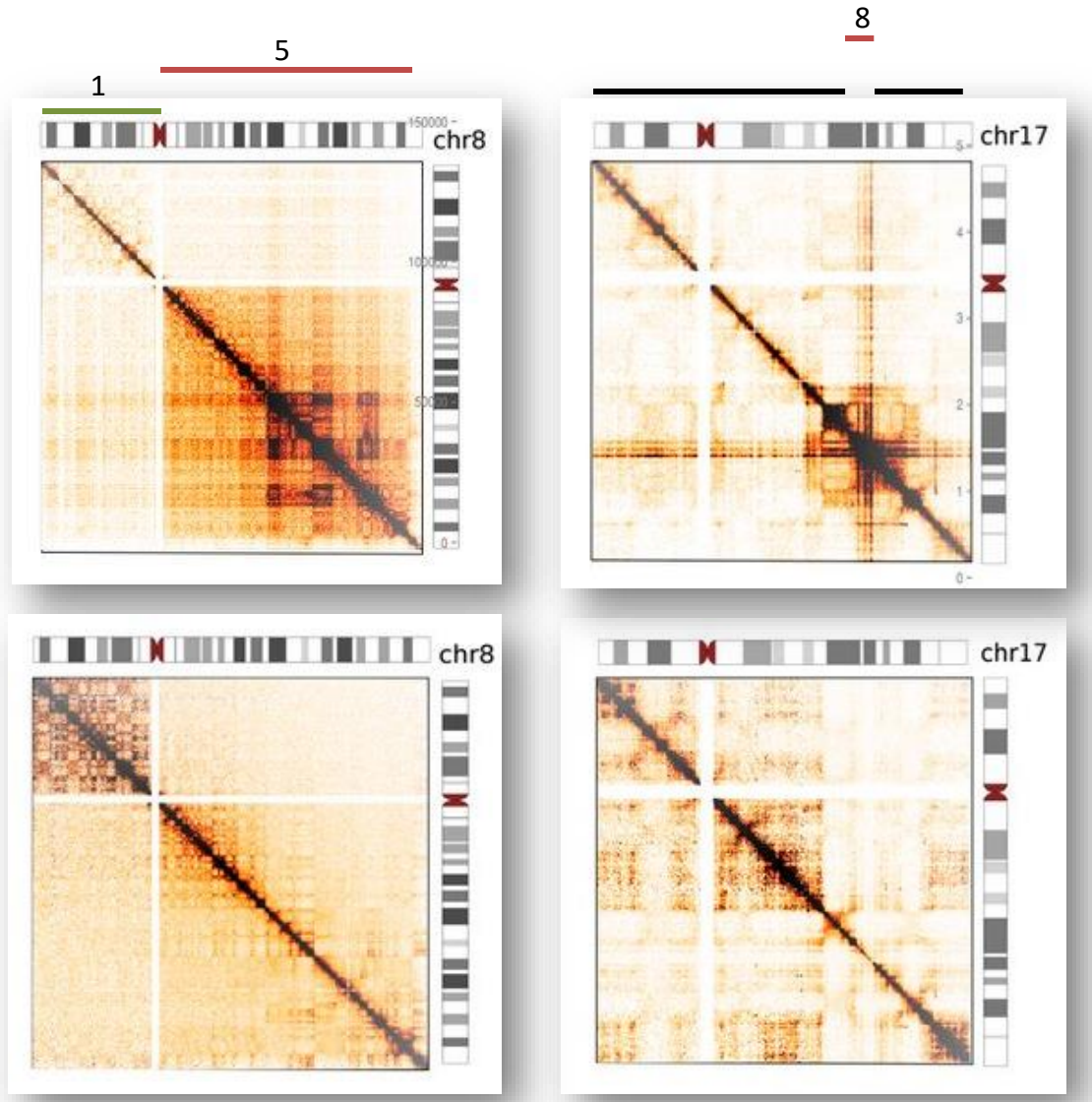
ICE



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**.





MCF7 Hi-C data

How to normalize cancer Hi-C data?

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

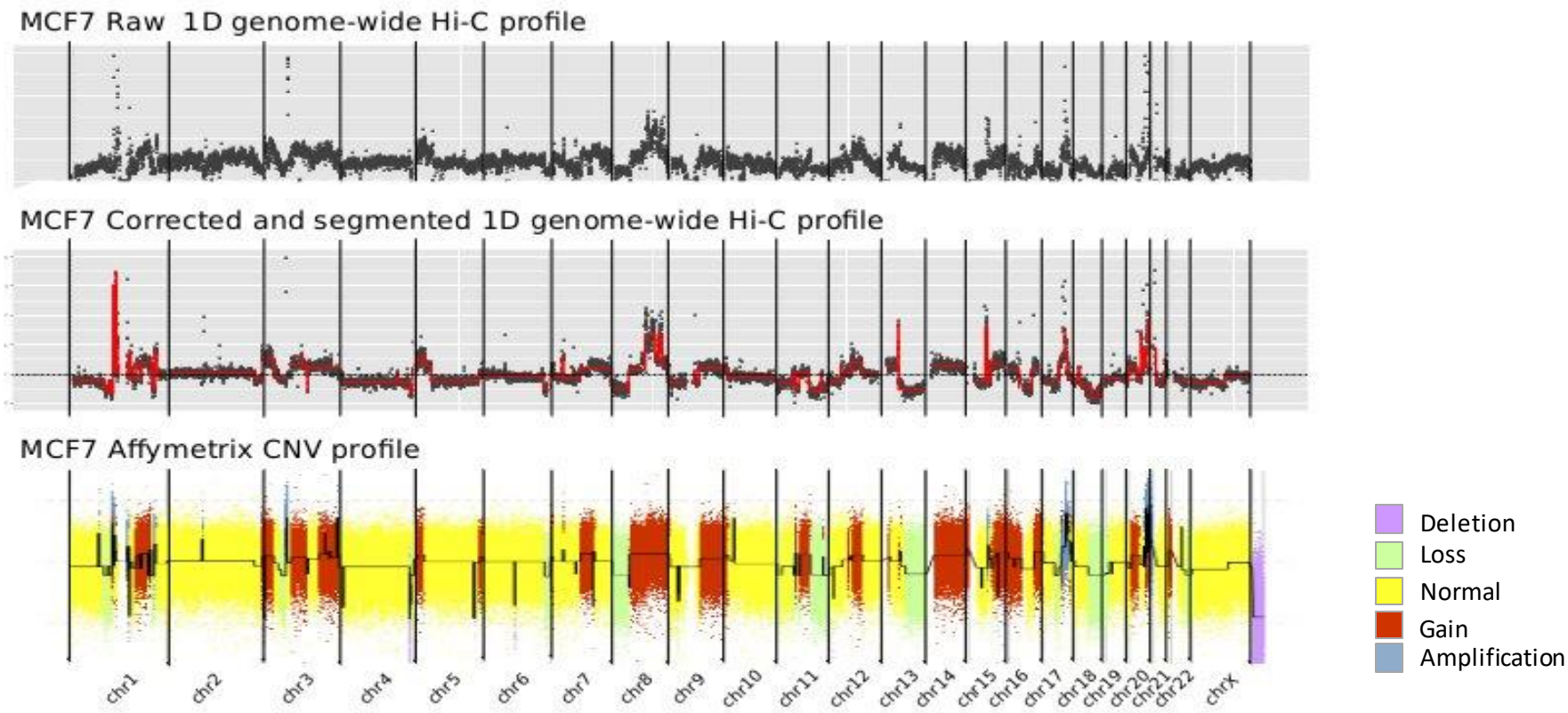
1. **Correct for systematic bias but not for the CNVs signal,** which can be useful for biological interpretation of cancer, for 3D modeling, genome reconstruction, contribution to CNVs to disease, *etc*.

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 :

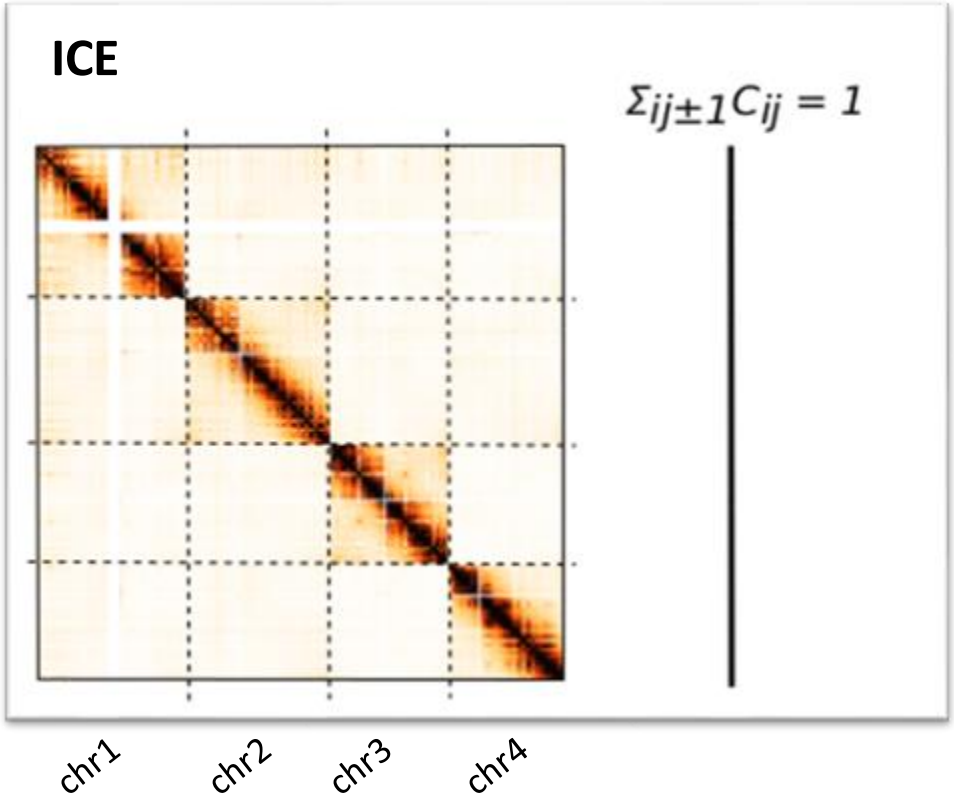
1. Generate the 1D Hi-C profile as the sum of contact per locus genome-wide
2. Remove systematic biases using a *Poisson* regression model
3. Segment the profile

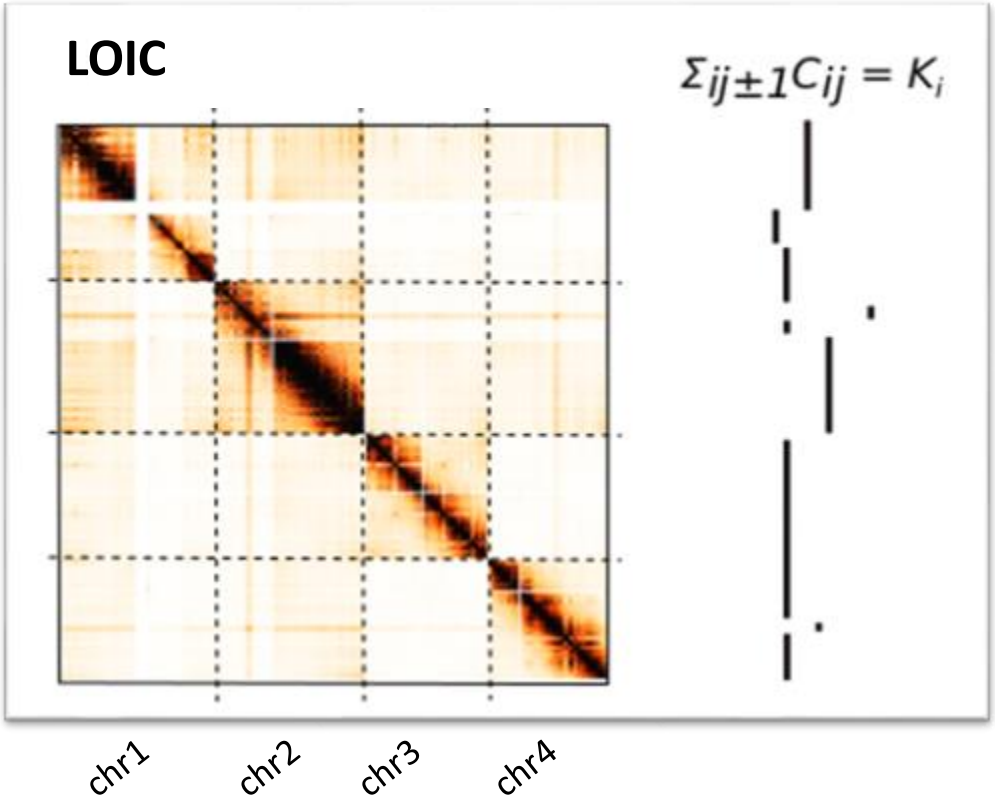


Validation on 100 simulated data-sets : 91% recall / 62.4% precision

CNV-based normalization of Hi-C cancer data

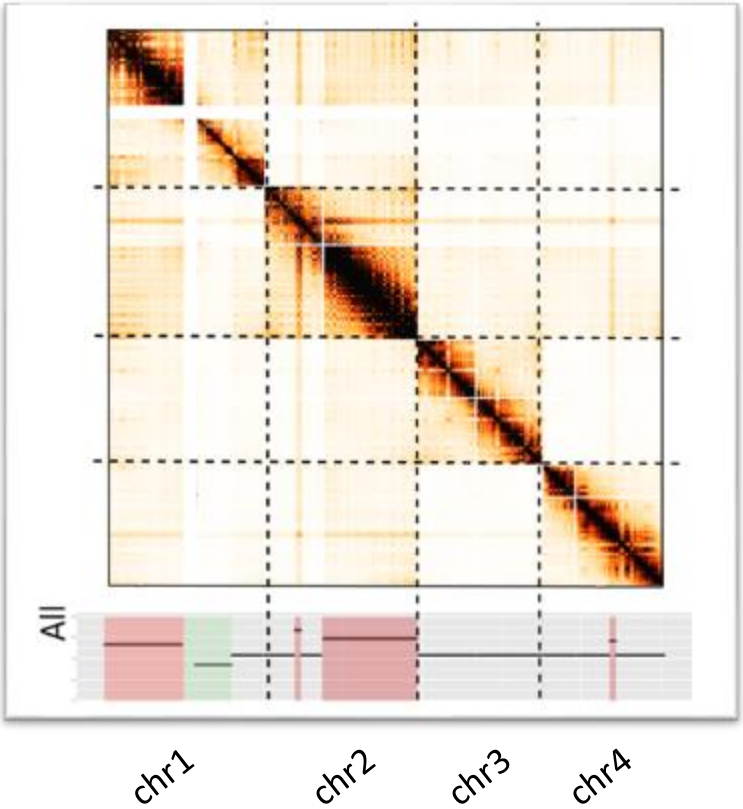
The Local Iterative correction (LOIC) normalization method extends the ICE model, making the assumption of local equal visibility per genomic segment





CNV-based normalization of Hi-C cancer data

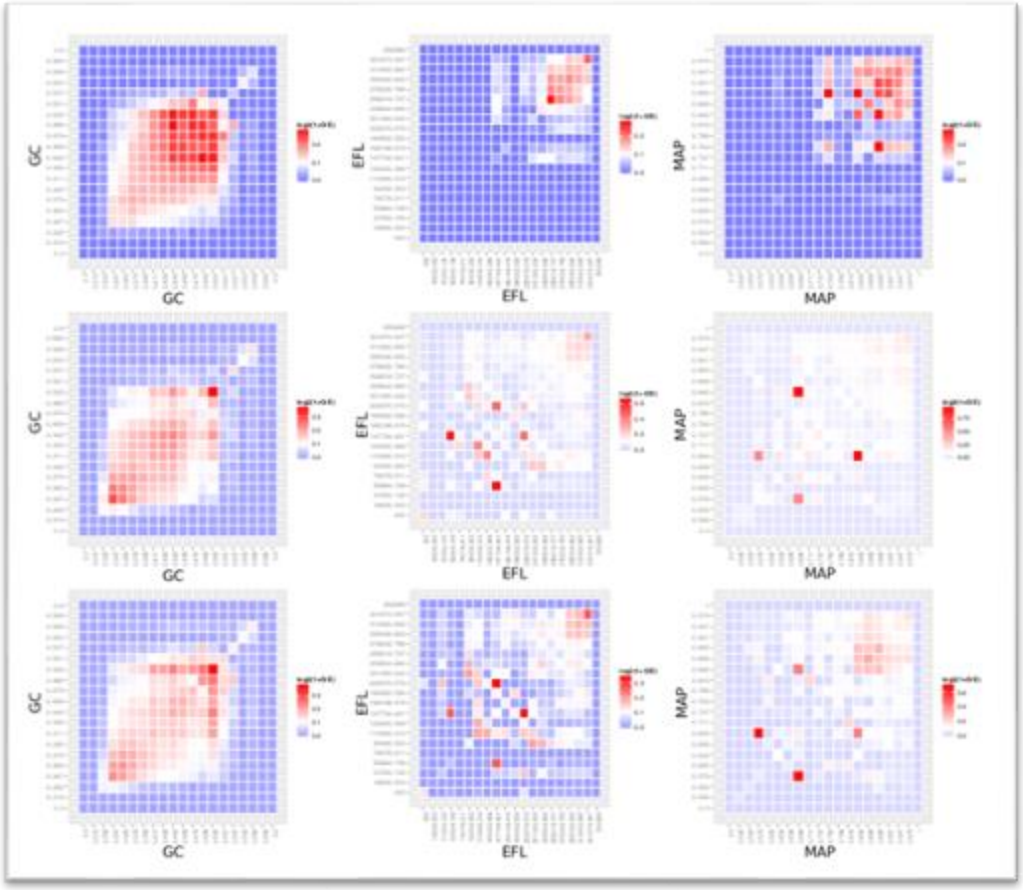
The Local Iterative correction (LOIC) normalization method extends the ICE model, making the assumption of local equal visibility per genomic segment



RAW

ICE

LOIC



GC content

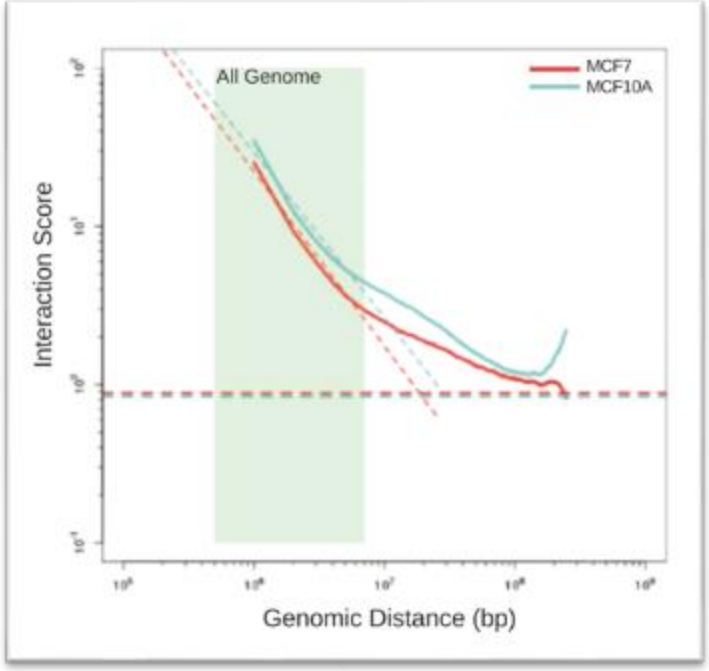
Fragment Length

Mappability

Removing CNVs from cancer Hi-C data

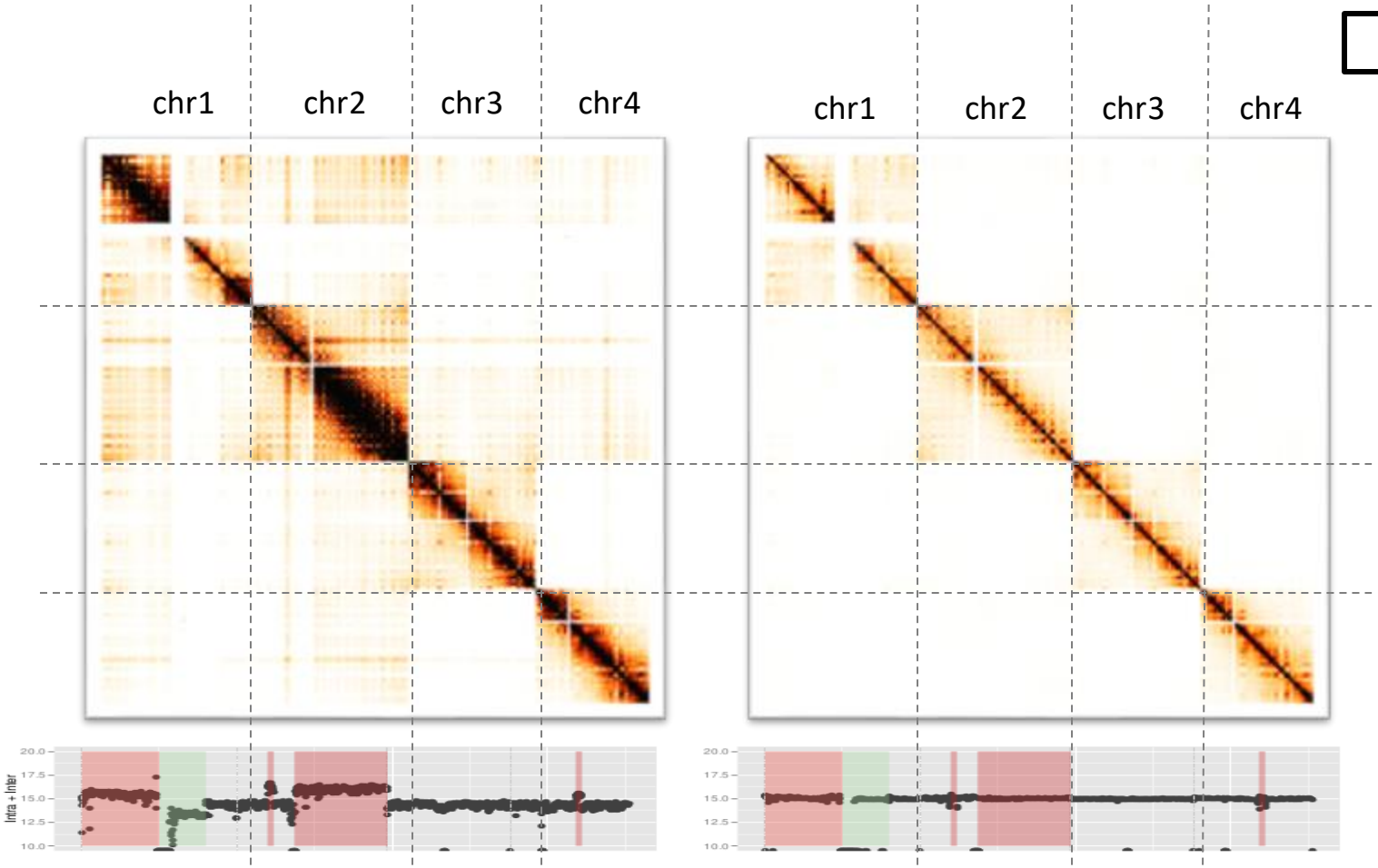
We assume that the copy number bias is constant per block and that the contact counts at a given genomic distance should be the same regardless the copy number status.

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



Raw data

CAIC data

Cancer Hi-C data normalization

METHODOLOGY ARTICLE Open Access

Effective normalization for copy number variation in Hi-C data

Nicolas Servant[1,2,3*†], Nelle Varoquaux[4,5†], Edith Heard[6], Emmanuel Barillot[1,2,3] and Jean-Philippe Vert[3,1,2,7]

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

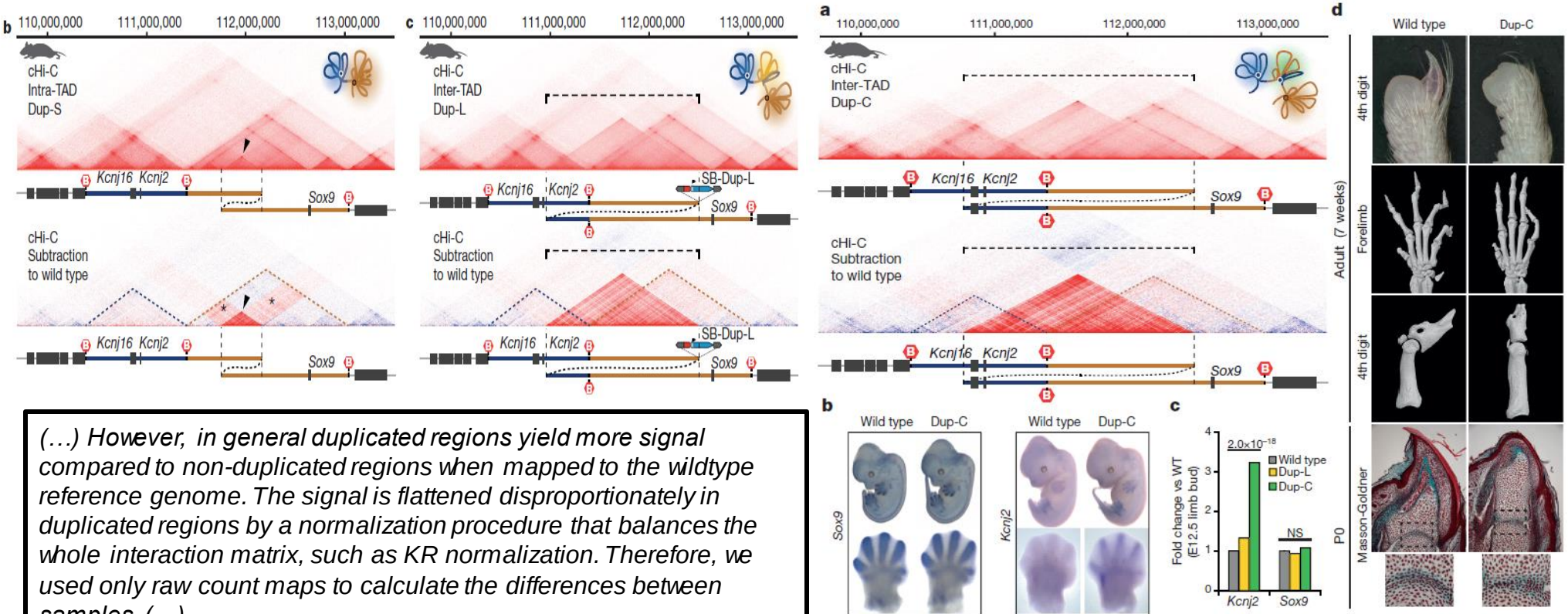
Available at https://github.com/nservant/cancer-hic-norm/

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

How useful is the LOIC method ?

Formation of new chromatin domains determines pathogenicity of genomic duplications

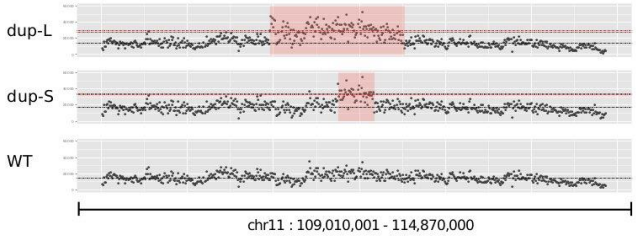
Martin Franke[1,2]*, Daniel M. Ibrahim[1,2,3]*, Guillaume Andrey[1], Wibke Schwarzer[4], Verena Heinrich[2,5], Robert Schöpflin[5], Katerina Kraft[1,2], Rieke Kempfer[1], Ivana Jerković[1,2], Wing-Lee Chan[2], Malte Spielmann[1,2], Bernd Timmermann[6], Lars Wittler[7], Ingo Kurth[8,9], Paola Cambiaso[10], 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]

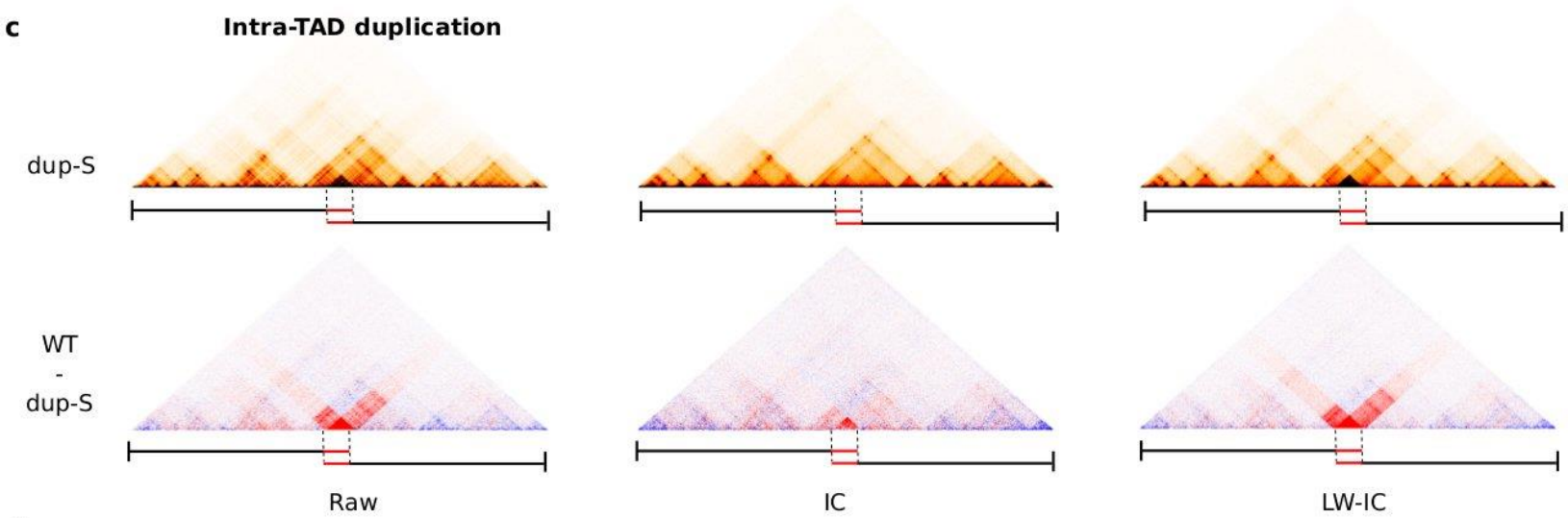


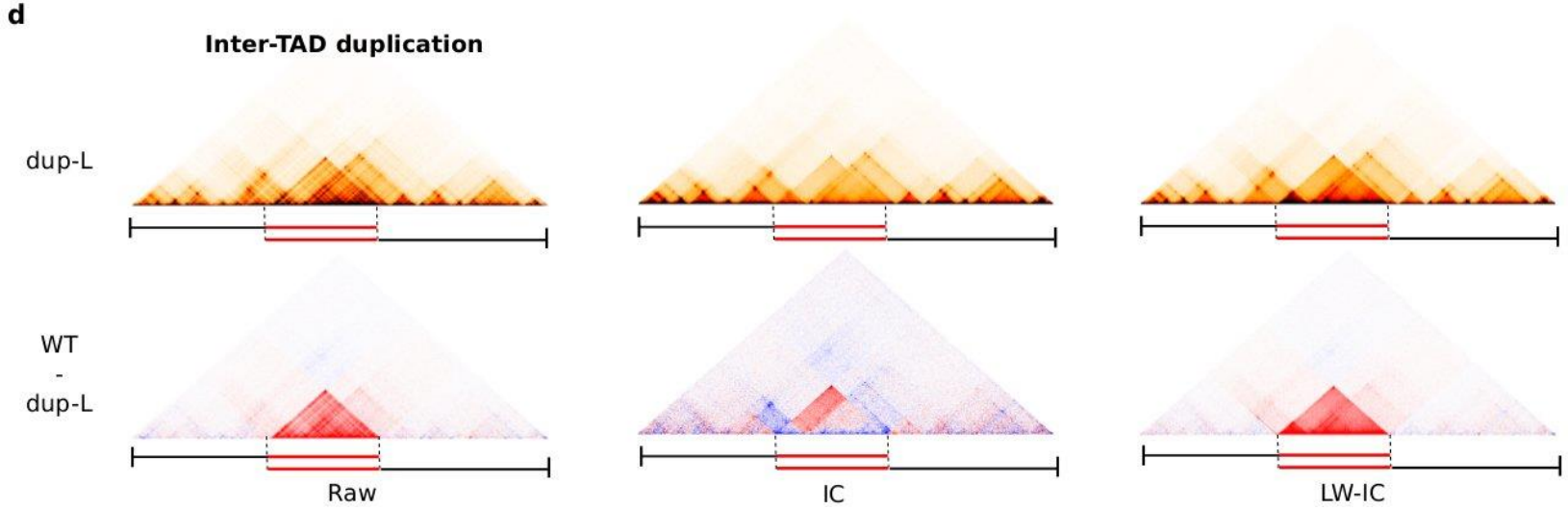
(…) However, in general duplicated regions yield more signal compared to non-duplicated regions when mapped to the wildtype reference genome. The signal is flattened disproportionately in duplicated regions by a normalization procedure that balances the whole interaction matrix, such as KR normalization.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

How useful is the LOIC method ?

Application of LW-IC on Franke et al. data





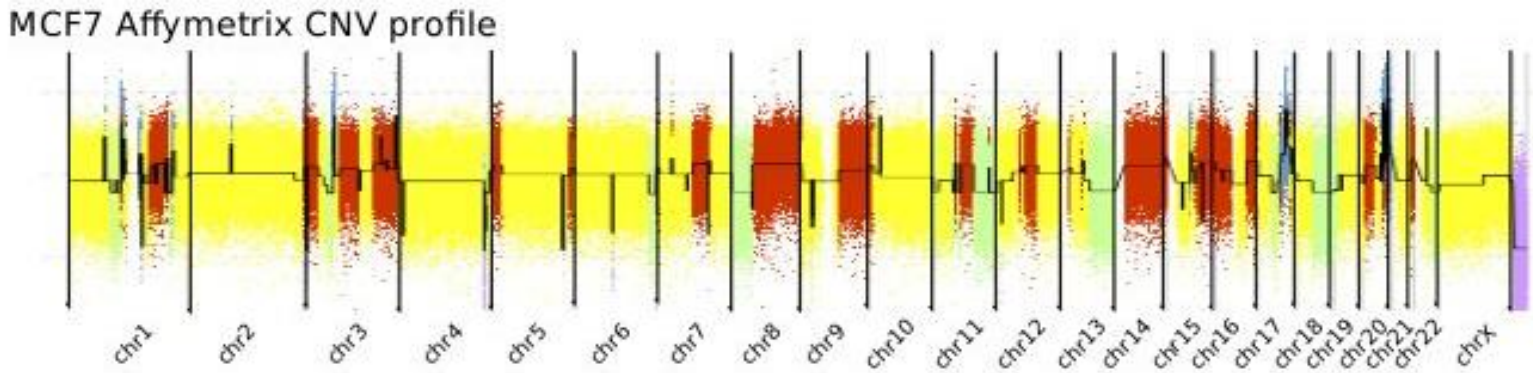


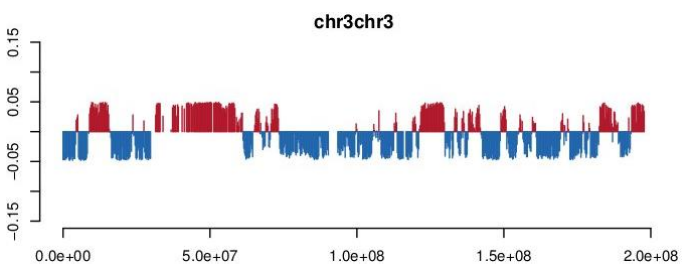
Going further with downstream analysis

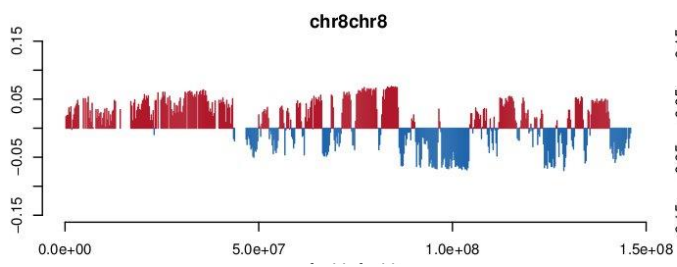
The detection of A/B chromosome compartments is usually based on PCA analysis of the intra-chromosomal maps correlation.

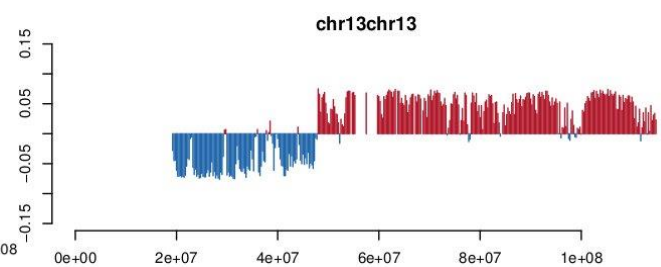
The methods is **surprisingly robust** to CNV variations

But for some chromosomes, the PC1 signal is biased toward the CNV profile



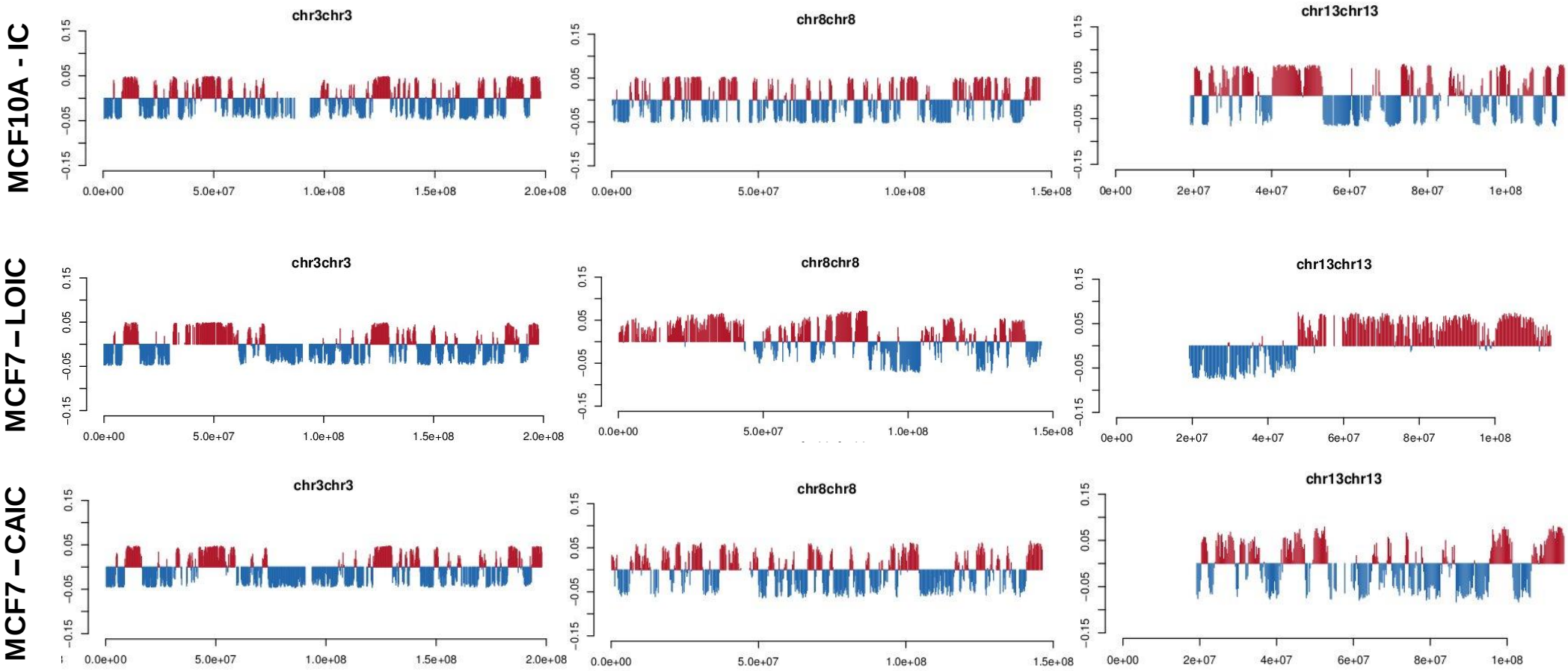






Removing CNVs from cancer Hi-C data

Detection of A/B chromosome compartments



Take Home Messages

HiC-Pro available at https://github.com/nservant/HiC-Pro
nf-core-hic is available at https://github.com/nf-core/hic

Both are collaborative projects, so do not hesitate to propose improvments or to report errors

In a copy number context, we demonstrate that the ICE normalization does not allow to correct for these effects 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

Many Thanks

Nelle Varoquaux, PhD

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

Agathe Neviere
Jean-Philippe Vert, PhD
Emmanuel Barillot, PhD

institutCurie

Edith Heard, PhD
Joke van Bemmel , PhD
Rafael Galupa , PhD
Agnese Loda , PhD
Elphege Nora , PhD

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR

FRANCE GÉNOMIQUE

Instituts thématiques
Inserm
Institut national
de la santé et de la recherche médicale

institutCurie