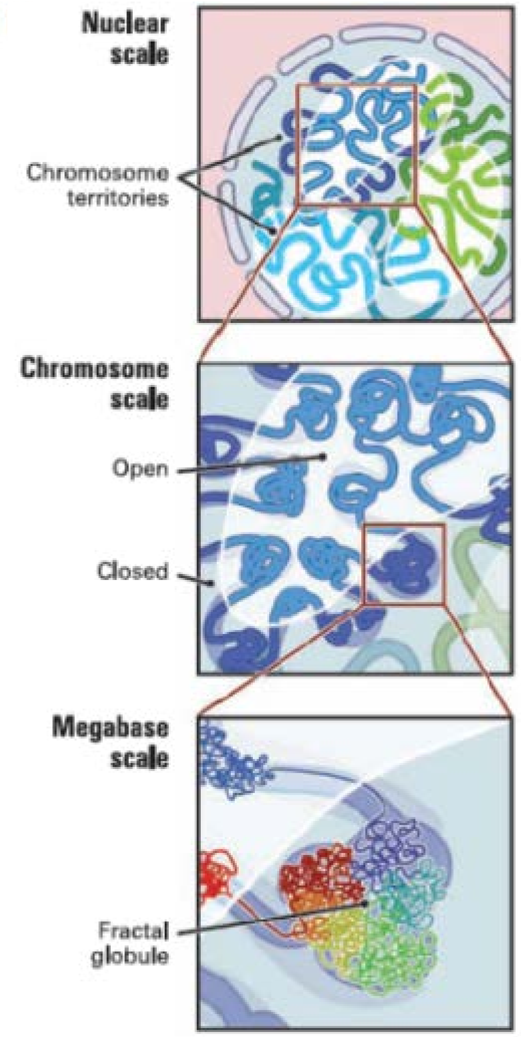
Chromosome Conformation Capture & Application to Cancer

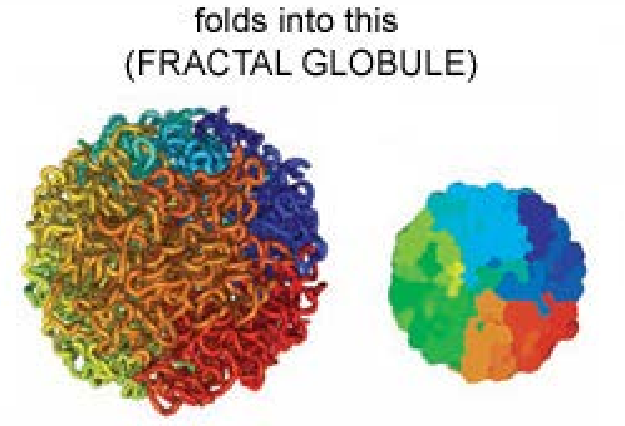
Nicolas Servant
Institut Curie, INSERM U900, Mines ParisTech

« La diversité tumorale à travers les plateformes technologiques franciliennes »
Paris, 14th of January 2016

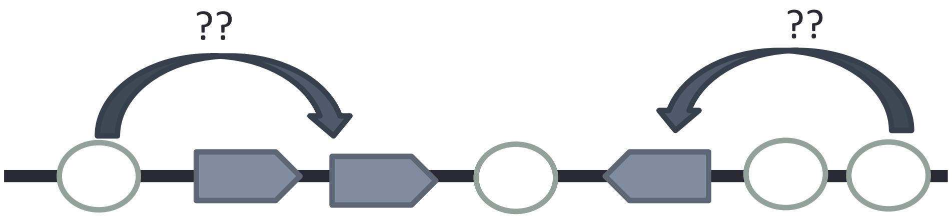
Measuring physical interactions



How is the genome organized ?



Which element regulates which genes ?
What is the impact of chromatin conformation on gene expression ?



Dekker et al., Liberman-Aiden et al. 2009

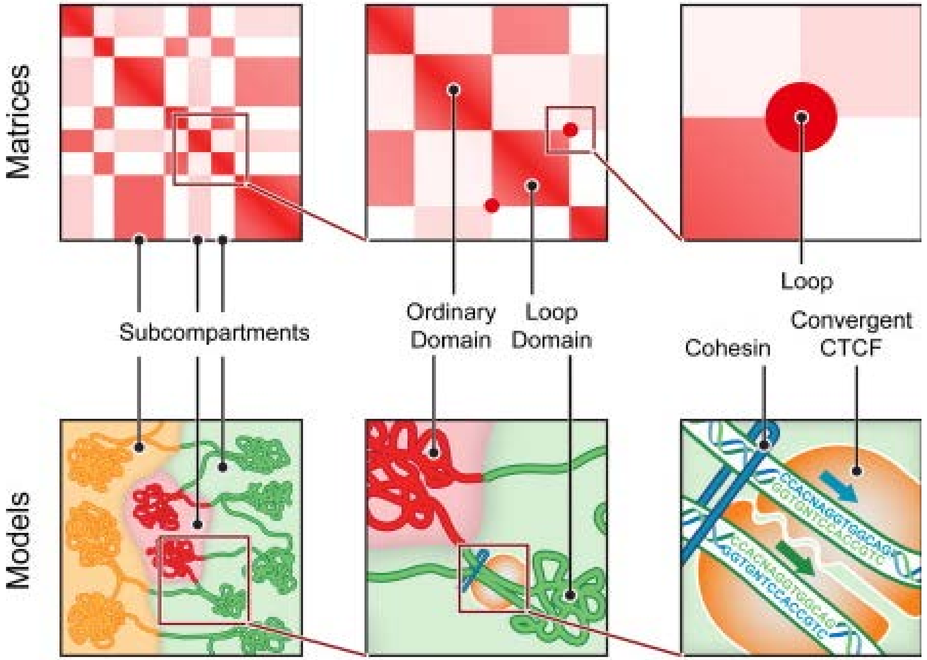
Hi-C and Genome Organization

Overview of features revealed by Hi-C

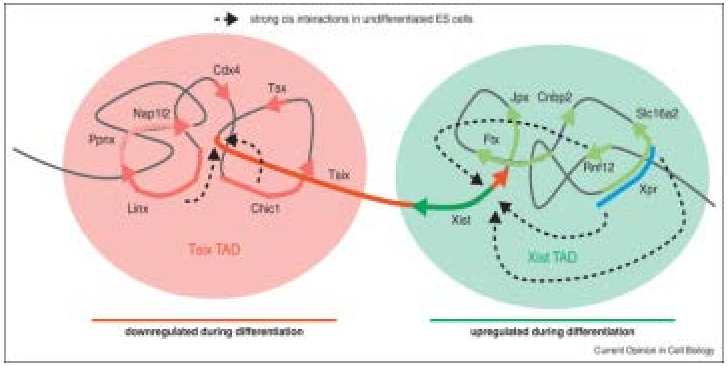
Liberman-Aiden et al. 2009
Genome organization and chromosomes compartments each bearing a distinctive pattern of epigenetic features

Dixon et al. 2012, Nora et al. 2012
Detection of topological domains (1Mb scale on average)

Rao et al. 2014
CTC/cohesin loop structures

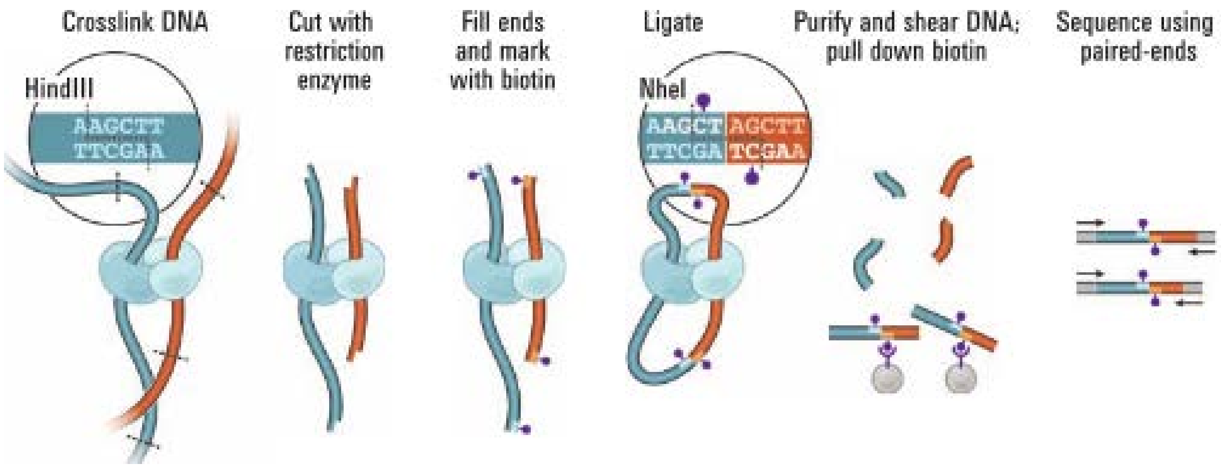


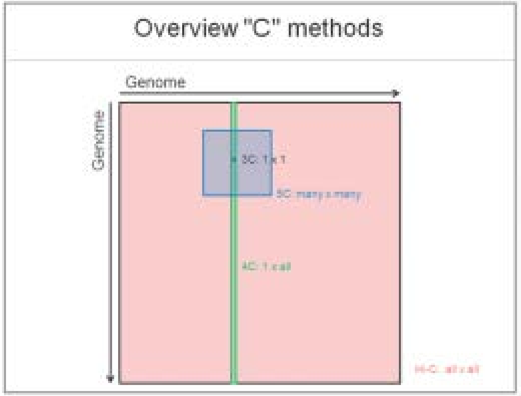
The chromatin conformation is an important factor of epigenetics regulation


strong cis interactions in undifferentiated ES cells
Cdx4
Tsx
Nap1l2
Ppnx
Tsix
Linx
Chic1
Tsix TAD
Jpx
Cnbp2
Slc16a2
Ftx
Rnf12
Xpr
Xist
Xist TAD
downregulated during differentiation
upregulated during differentiation
Current Opinion in Cell Biology

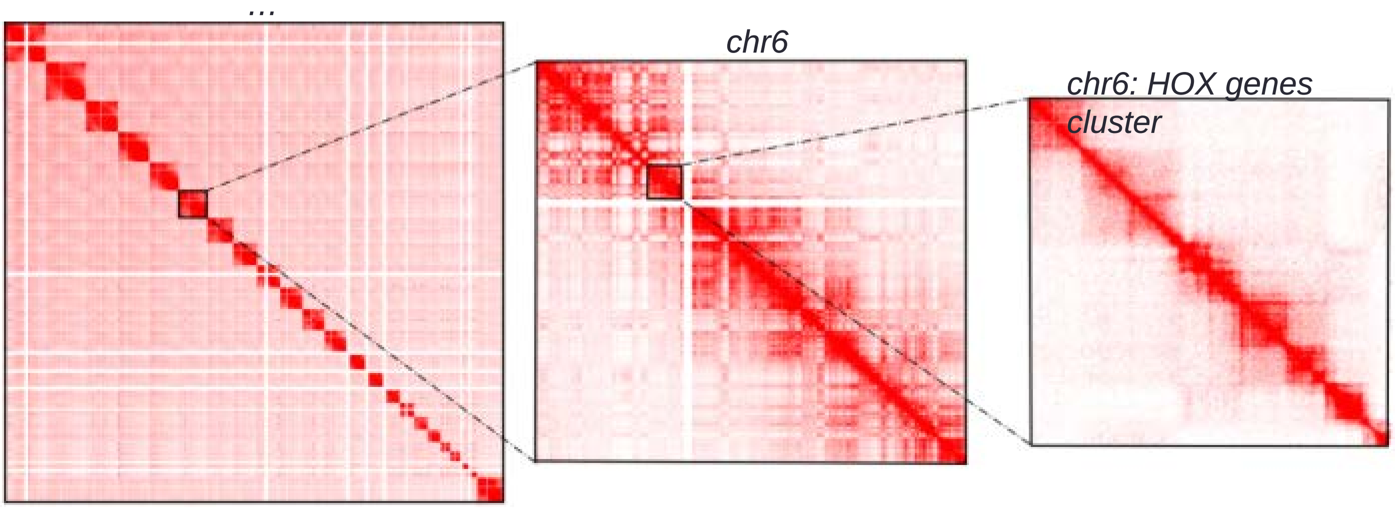
Lieberman-Aiden et al. 2009, Rao et al. 2014

Genome-wide ‘C’ – Hi-C


Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends



Lieberman-Aiden et al. 2009



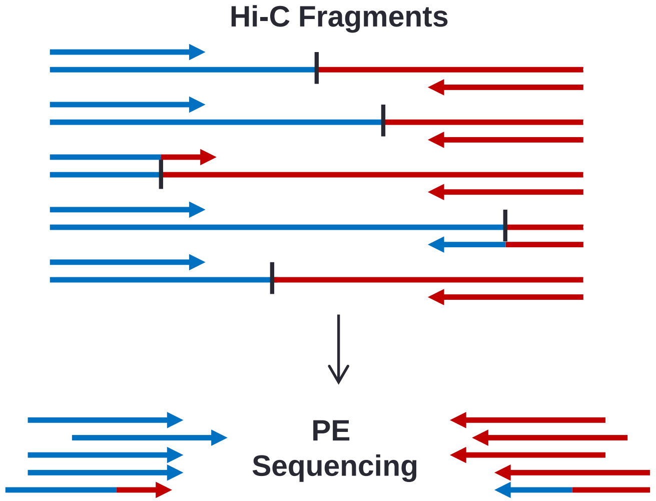
Whole genome map

Intrachromosomal contact maps

Topological domains

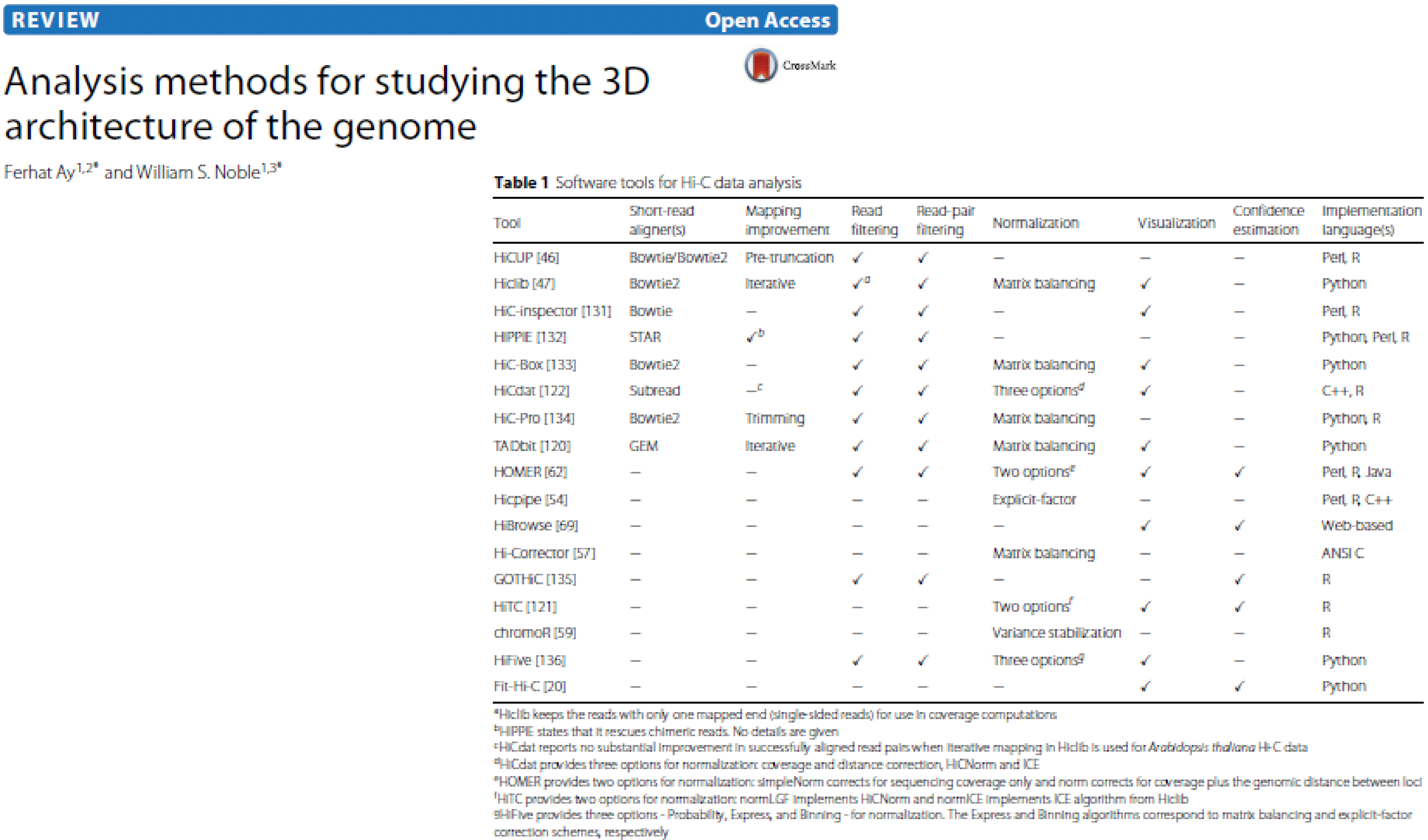
What does Hi-C data look like ?

Illumina paired-end sequencing



	# read pairs (M)/sample	Disk size (Go)	Resolution (kb)	Genome matrix size (bins)
Dixon et al. 2012	400	172	20-40	150 000
Jin et al. 2013	1 200	-	5-10	600 000
Rao et al. 2014	1 500	1 200	1-5	3 000 000

How to process Hi-C data ?



REVIEW **Open Access**

Analysis methods for studying the 3D architecture of the genome

Ferhat Ay[1,2*] and William S. Noble[1,3*]

Table 1 Software tools for Hi-C data analysis

Tool	Short-read aligner(s)	Mapping improvement	Read filtering	Read-pair filtering	Normalization	Visualization	Confidence estimation	Implementation language(s)
HiCUP [46]	Bowtie/Bowtie2	Pre-truncation	✓	✓	–	–	–	Perl, R
Hiclib [47]	Bowtie2	Iterative	✓[a]	✓	Matrix balancing	✓	–	Python
HiC-inspector [131]	Bowtie	–	✓	✓	–	✓	–	Perl, R
HIPPIE [132]	STAR	✓[b]	✓	✓	–	–	–	Python, Perl, R
HiC-Box [133]	Bowtie2	–	✓	✓	Matrix balancing	✓	–	Python
HiCdat [122]	Subread	–[c]	✓	✓	Three options[d]	✓	–	C++, R
HiC-Pro [134]	Bowtie2	Trimming	✓	✓	Matrix balancing	–	–	Python, R
TADbit [120]	GEM	Iterative	✓	✓	Matrix balancing	✓	–	Python
HOMER [62]	–	–	✓	✓	Two options[e]	✓	✓	Perl, R, Java
Hicpipe [54]	–	–	–	–	Explicit-factor	–	–	Perl, R, C++
HiBrowse [69]	–	–	–	–	–	✓	✓	Web-based
Hi-Corrector [57]	–	–	–	–	Matrix balancing	–	–	ANSI C
GOTHiC [135]	–	–	✓	✓	–	–	✓	R
HiTC [121]	–	–	–	–	Two options[f]	✓	✓	R
chromoR [59]	–	–	–	–	Variance stabilization	–	–	R
HiFive [136]	–	–	✓	✓	Three options[g]	✓	–	Python
Fit-Hi-C [20]	–	–	–	–	–	✓	✓	Python

[a] Hiclib keeps the reads with only one mapped end (single-sided reads) for use in coverage computations

[b] HIPPIE states that it rescues chimeric reads. No details are given

[c] HiCdat reports no substantial improvement in successfully aligned read pairs when iterative mapping in Hiclib is used for *Arabidopsis thaliana* Hi-C data

[d] HiCdat provides three options for normalization: coverage and distance correction, HiCNorm and ICE

[e] HOMER provides two options for normalization: simpleNorm corrects for sequencing coverage only and norm corrects for coverage plus the genomic distance between loci

[f] HiTC provides two options for normalization: normLGF implements HiCNorm and normICE implements ICE algorithm from Hiclib

[g] HiFive provides three options - Probability, Express, and Binning - for normalization. The Express and Binning algorithms correspond to matrix balancing and explicit-factor correction schemes, respectively

HiC-Pro

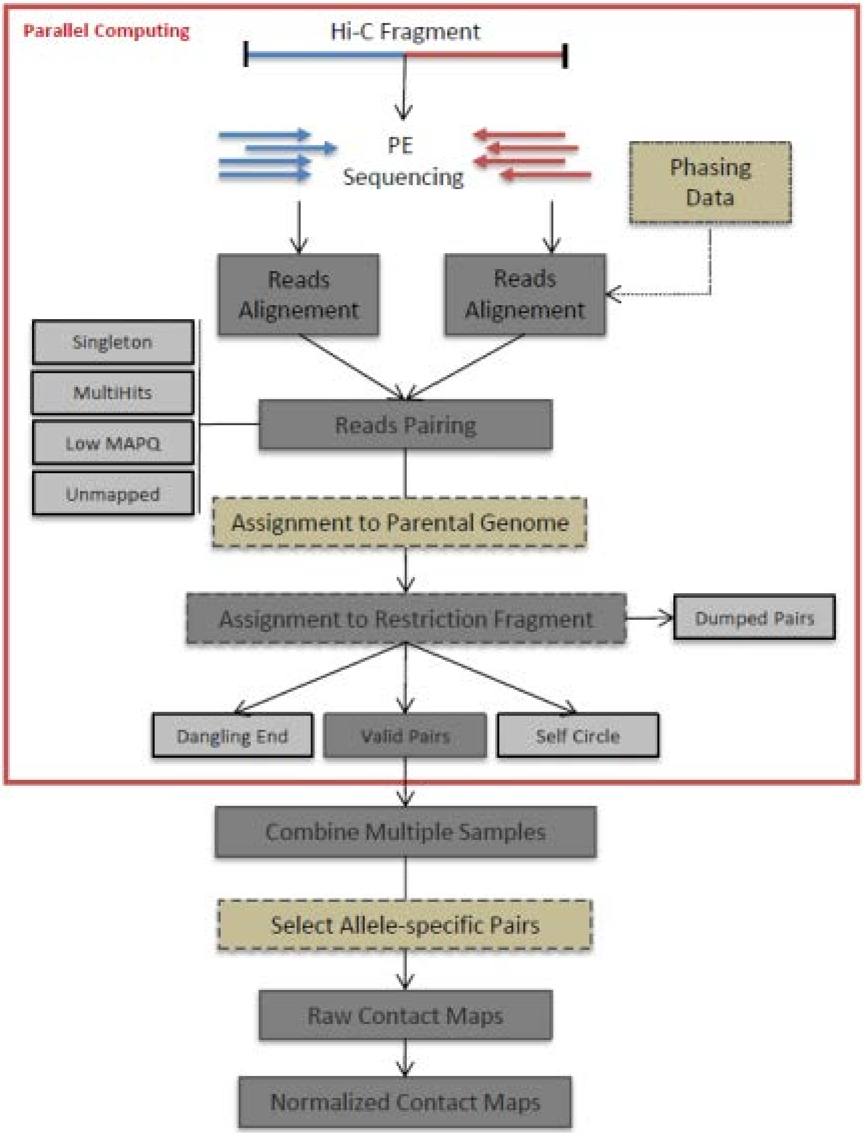
Easy-to-use
- i.e. one command line
- Only a few dependencies

Optimized
- Python/C++/R

Scalable
- Memory efficient, fast and parallelized
- Genome-wide ICE normalization at high resolution

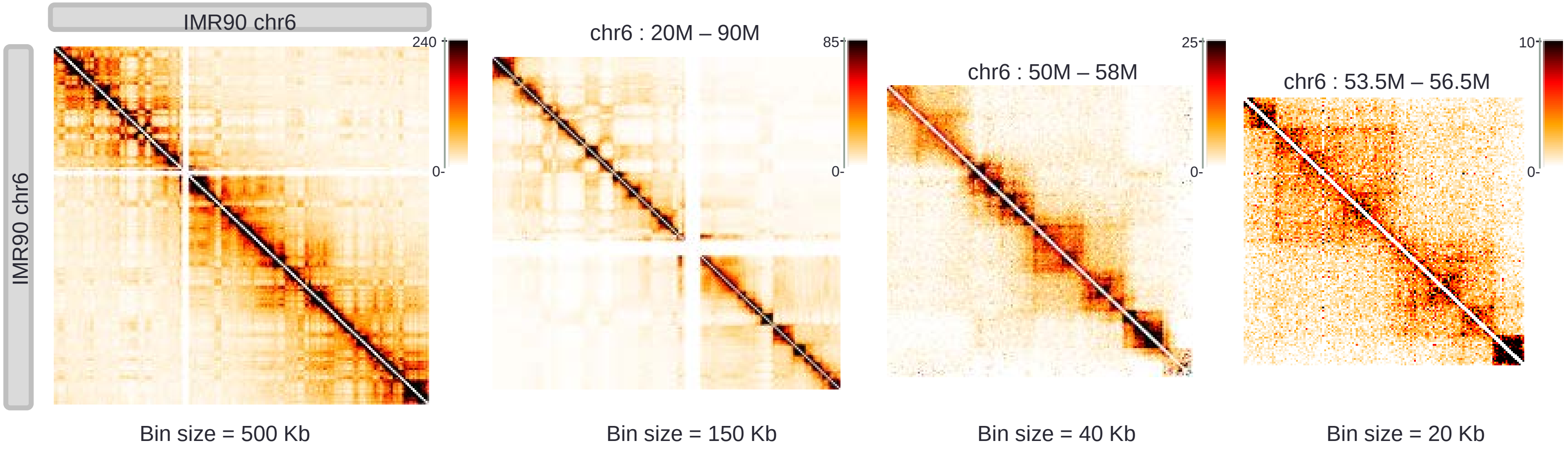
Flexible
- Collaborative project
- New functionalities can be added

Complete
- From raw reads to normalized contact maps
- Can analyse Hi-C data not based on restriction enzyme digestion such as Dnase Hi-C
- Can perform allele-specific analysis

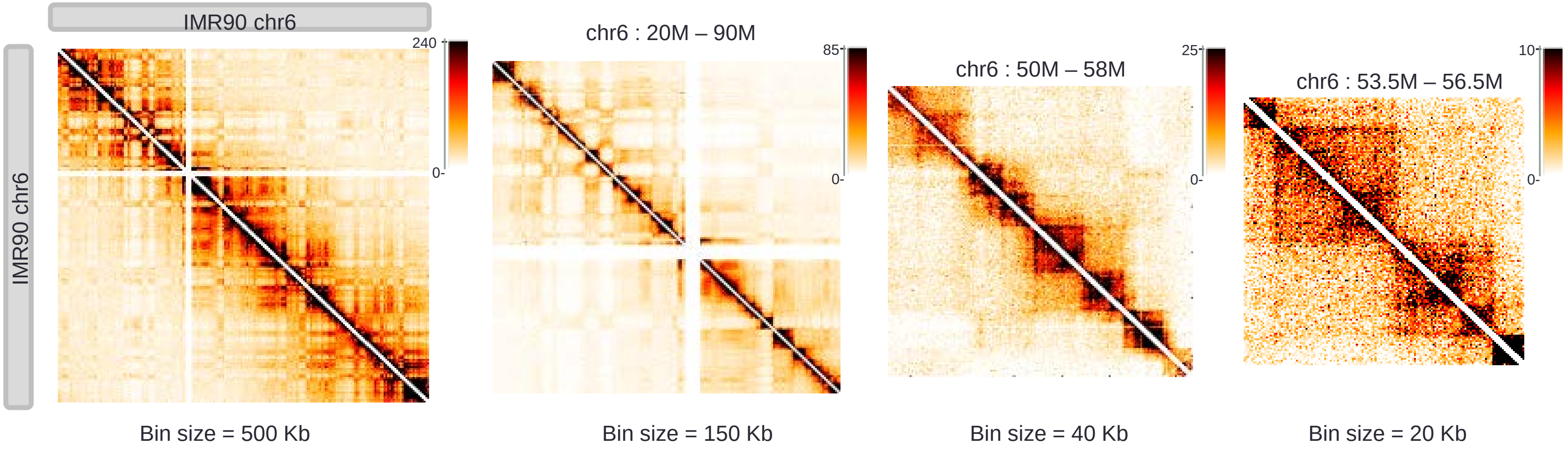


Does the pipeline work ?

Chromosome 6 contact map (hiclib)

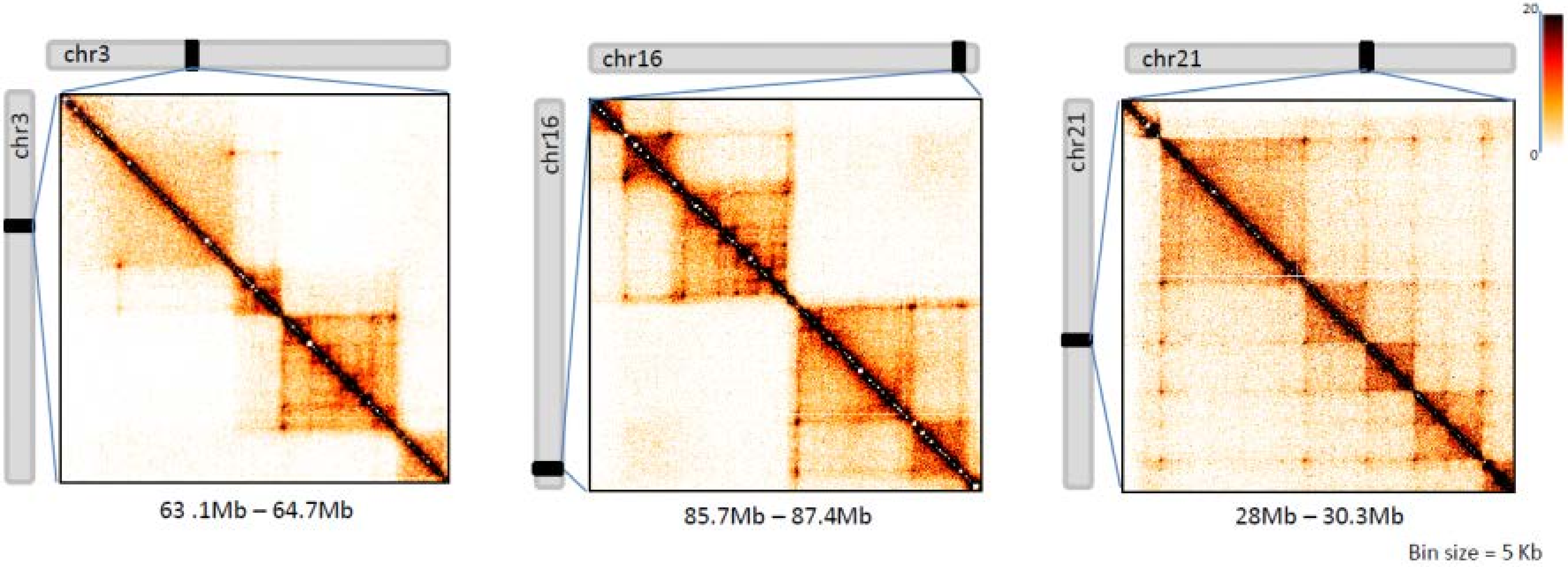


Chromosome 6 contact map (HiC-Pro)



Does the pipeline work ?

Rao et al. IMR90 5kb maps generated with HiC-Pro



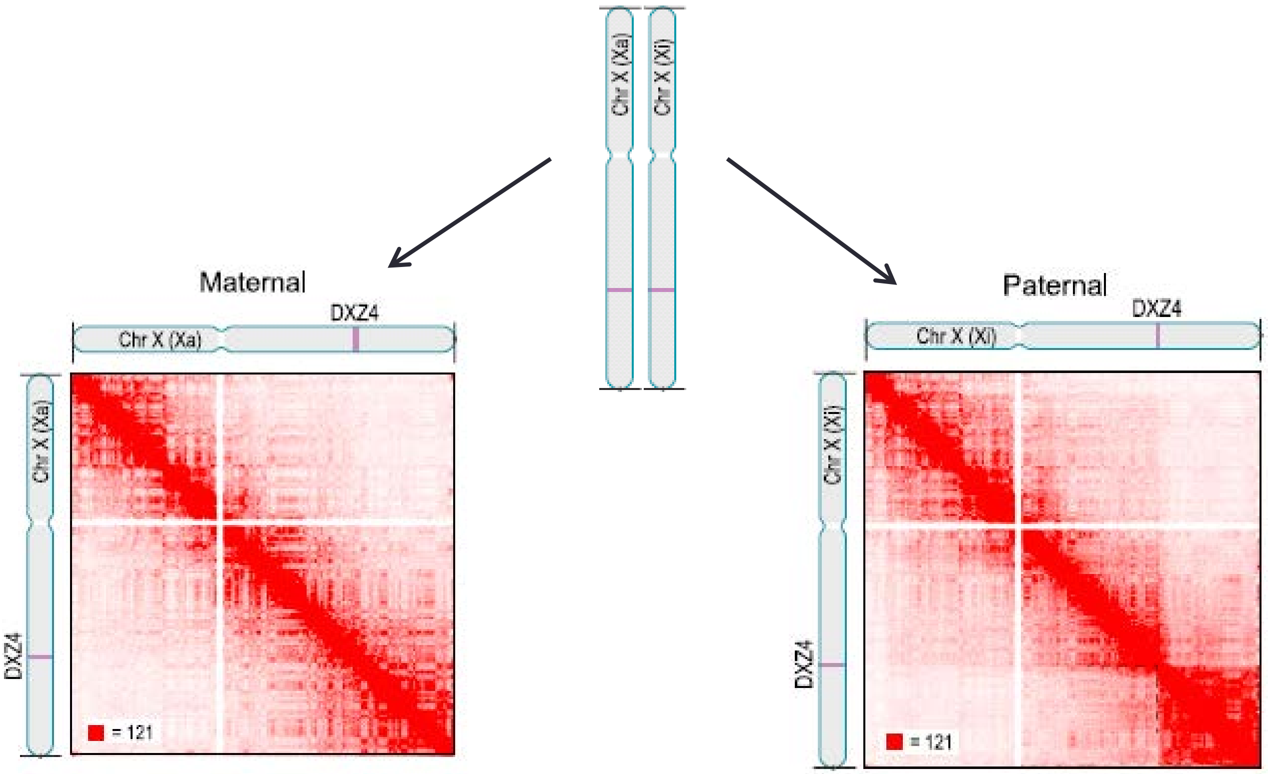
HiC-Pro : Pipeline Implementation

Complete workflow

	hiclib	**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
Wall Time	**28:24**	**17:56**	**02:08**	**11:41**
-- 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

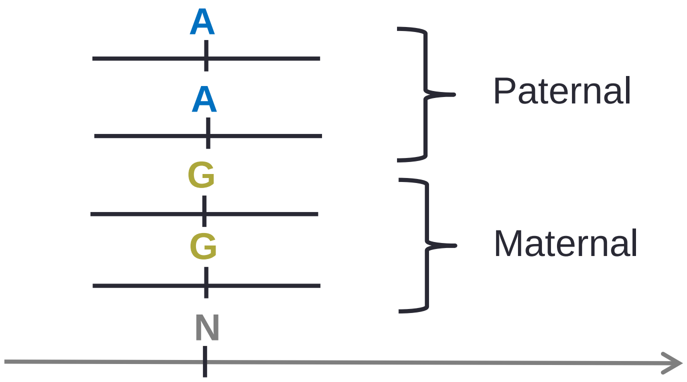
Allele-specific contact maps

How to assign contacts to specific chromosomal homologs using phasing data ?



Rao et al 2014

Allele-specific contact maps



HiC-Pro : Allele specific mode

Input :

- raw sequencing reads **+ phasing data** (.VCF)

Mapping :

- mask all SNPs on the reference genome and align reads
- Assign each reads to a parental genome

Read pairs classification :

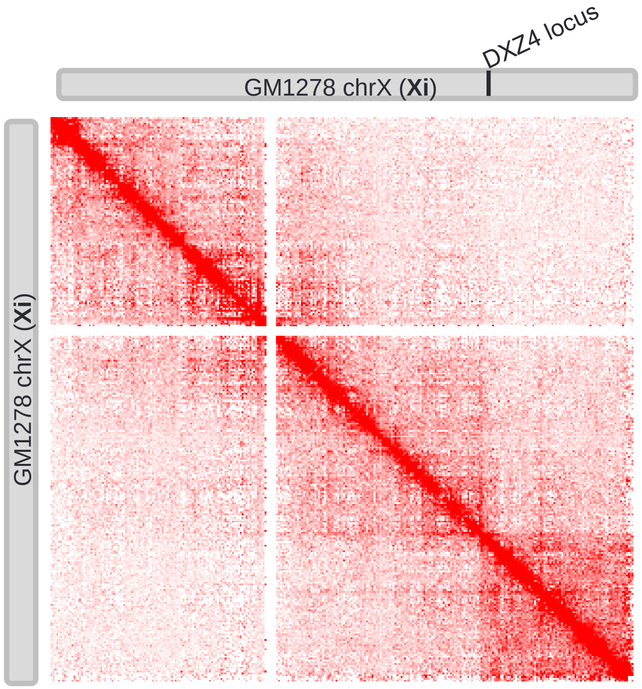
- Classify each valid interaction pairs as allele specific (paternal or maternal), uninformative (U) or ambiguous (A)

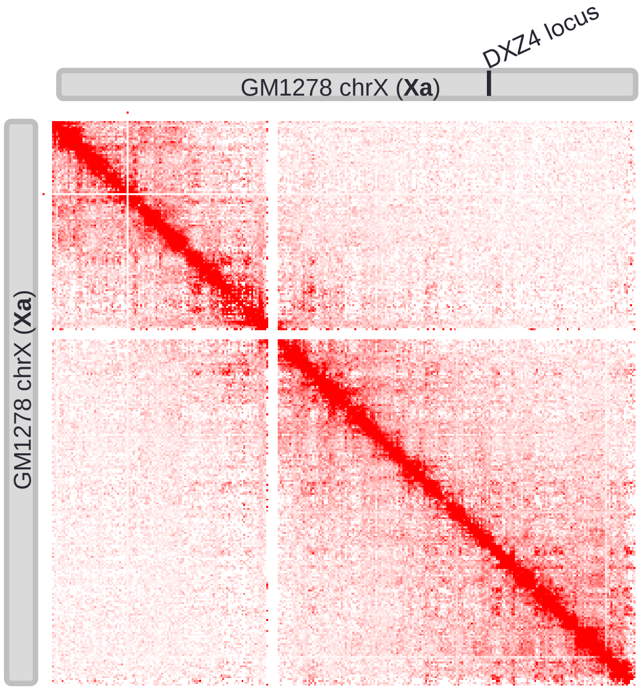
Contact maps and normalization :

- Build and normalize allele specific interaction maps of paternal and maternal valid pairs

Example of Selvaraj et al. GM1278

Total number of read pairs	826 414 879
Total number of valid pairs	503 536 186 (100%)
Number of pairs assigned to G1	28 391 258 (5.64%)
Number of pairs assigned to G2	28 308 925 (5.62%)
Number of trans G1/G2 pairs	603 213 (0.12%)
Number of unassigned reads	446 171 241 (88.60%)
Number of conflicting reads	61 549 (0.01%)





HiC-Pro Summary

Able to process Hi-C data from raw sequencing reads to iced contact maps

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

- Freely available and open to contribution
- Can be applied on any organism (with a reference genome)
- Automatic installation process and a few dependencies
- **Easy-to-use**, i.e one command line and step-by-step procedure
- **Fast,** scalable
- Time and memory efficient
- Based on an efficient contact maps format
- **Alllele-specific analysis**
- Can process **Dnase** Hi-C samples

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

Genome **Biology**

SOFTWARE **Open Access**

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

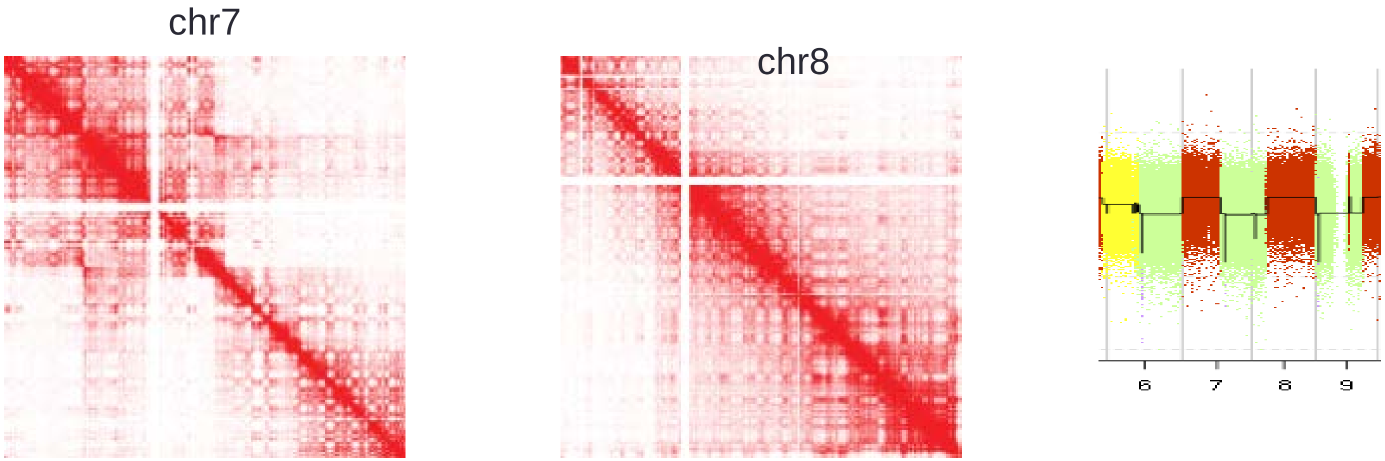
Nicolas Servant[1,2,3*], Nelle Varoquaux[1,2,3], Bryan R. Lajoie[4], Eric Viara[5], Chong-Jian Chen[1,2,3,6,7,8], Jean-Philippe Vert[1,2,3], Edith Heard[1,6,7], Job Dekker[9] and Emmanuel Barillot[1,2,3]

Application of Hi-C to cancer

- How the cancer genome is organized ?
- Can we detect new enhancer/promoter loop ?
- Do you see any changes in the chromosome compartments / topological domains ?
- Are these changes correlated with gene expression or any histone modification ?

But working with Cancer data also open new challenges :

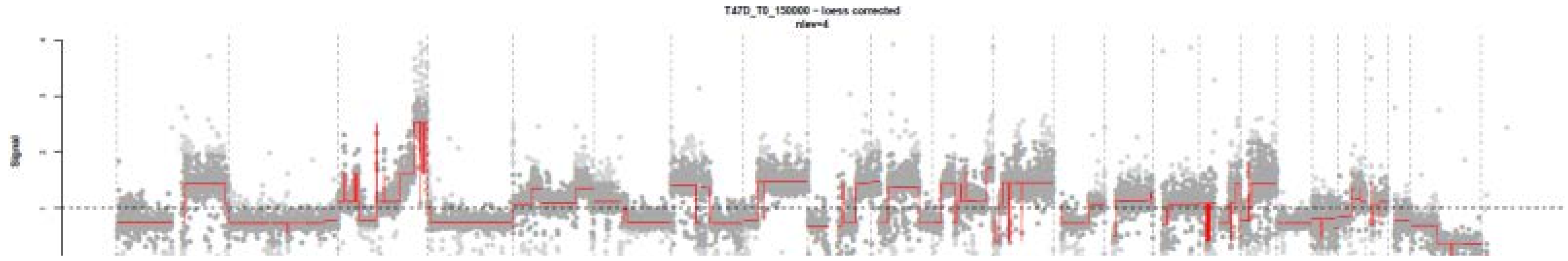
- Effect of CNVs on the Hi-C maps ?
- Can we use the same normalization approach ?
- How to compare samples ?
- …

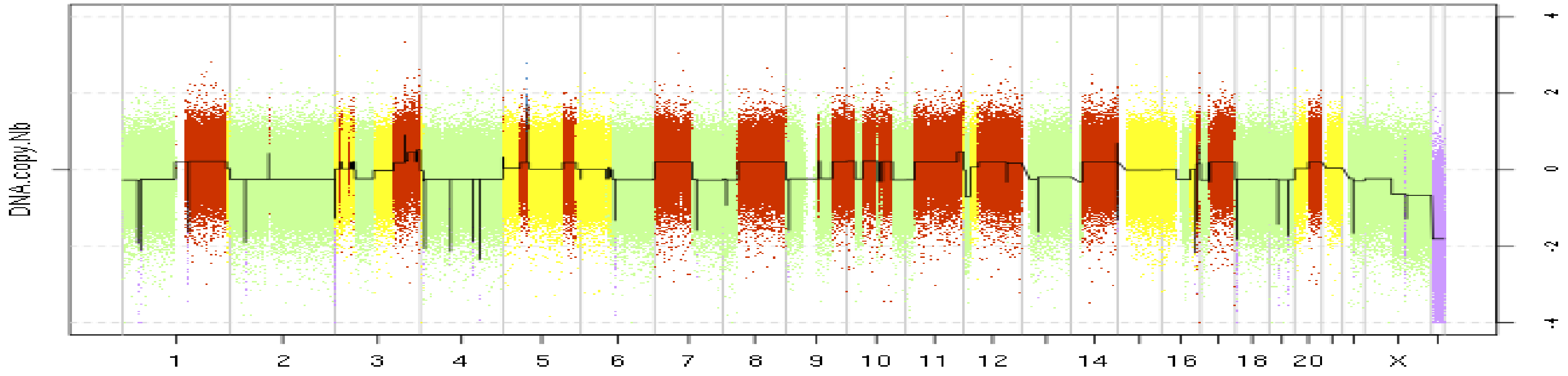


T47D Hi-C data, Le Dily et al.

Hi-C and CNV
CNV estimation - T47D data

Processing of public raw data using HiC-Pro
CNV estimation and comparison with SNP6.0 profile





Many Thanks

INSERM U900
Nelle Varoqaux
Eric Viara
Jean-Philippe Vert
Emmanuel Barillot

Collaborators
Edith Heard (Institut Curie)
David Gentien (Institut Curie)
Bryan Lajoie (UMASS)
Job Dekker (UMASS)
Felix Kruger (Babraham Institute)