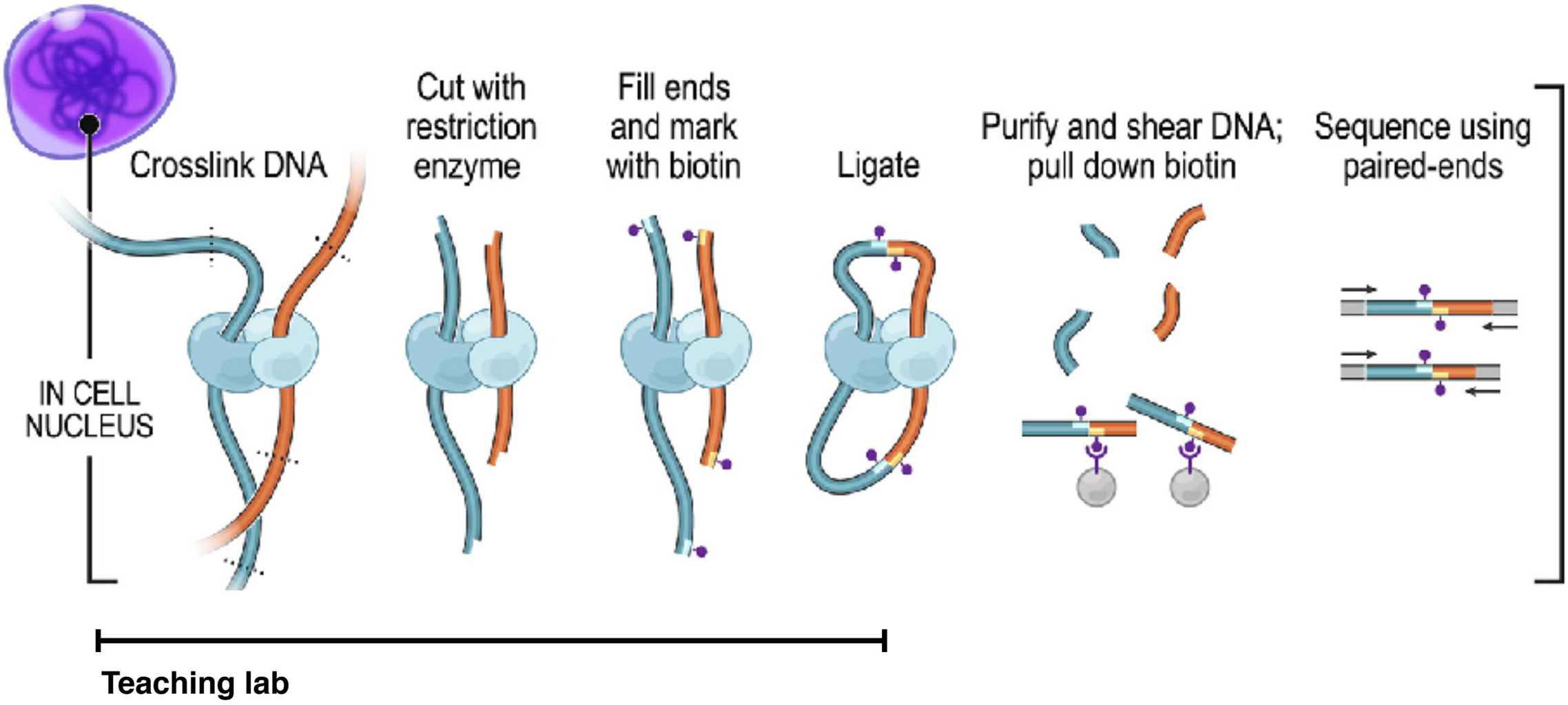
TADbit

a bioinformatic framework to analyse Hi-C experiments

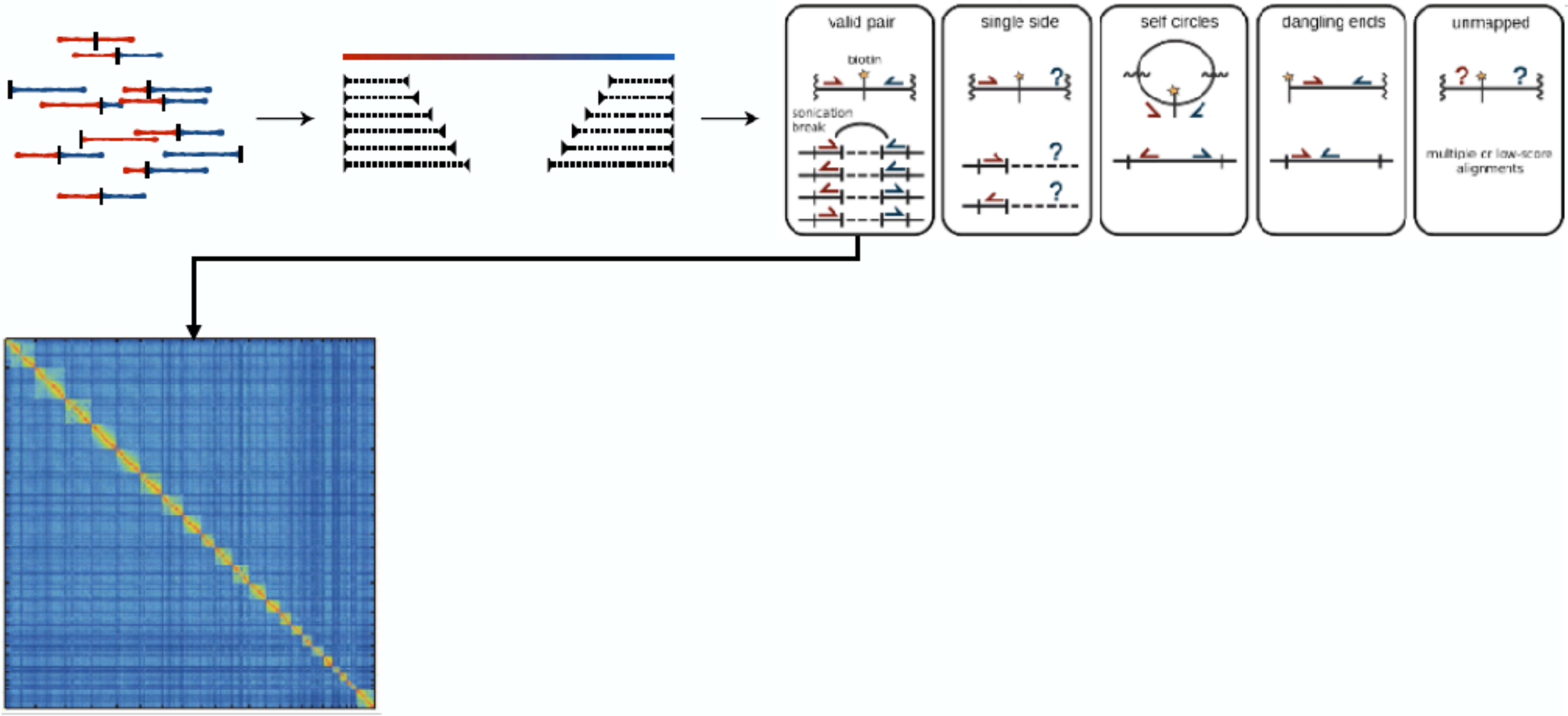
François Serra, Paula Soler & Marc A. Marti-Renom
Structural Genomics Group (CNAG-CRG)

HI-C: MAIN STEPS



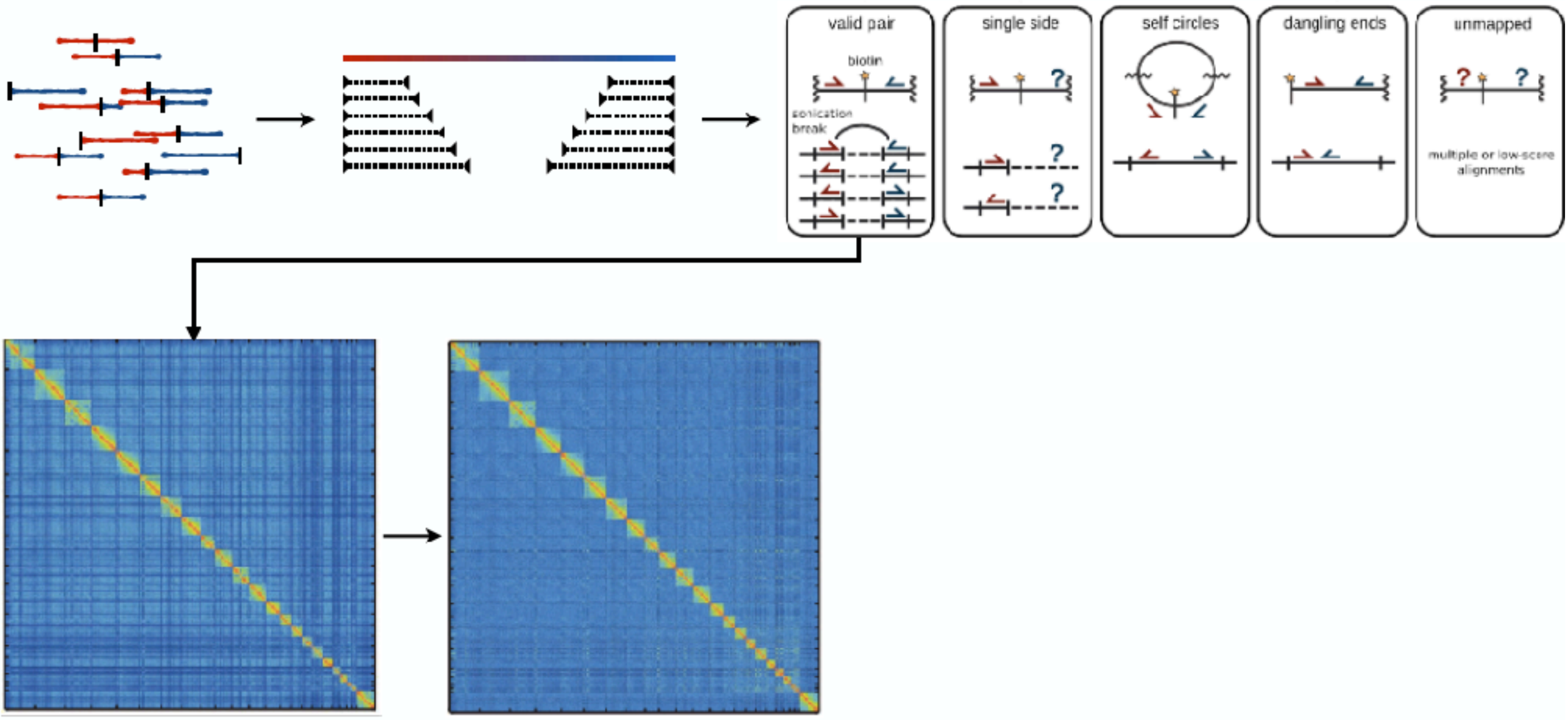
Lieberman-Aiden, E. … Dekker, J. (2009). Science, 326(5950),

From FASTQ to interaction matrices



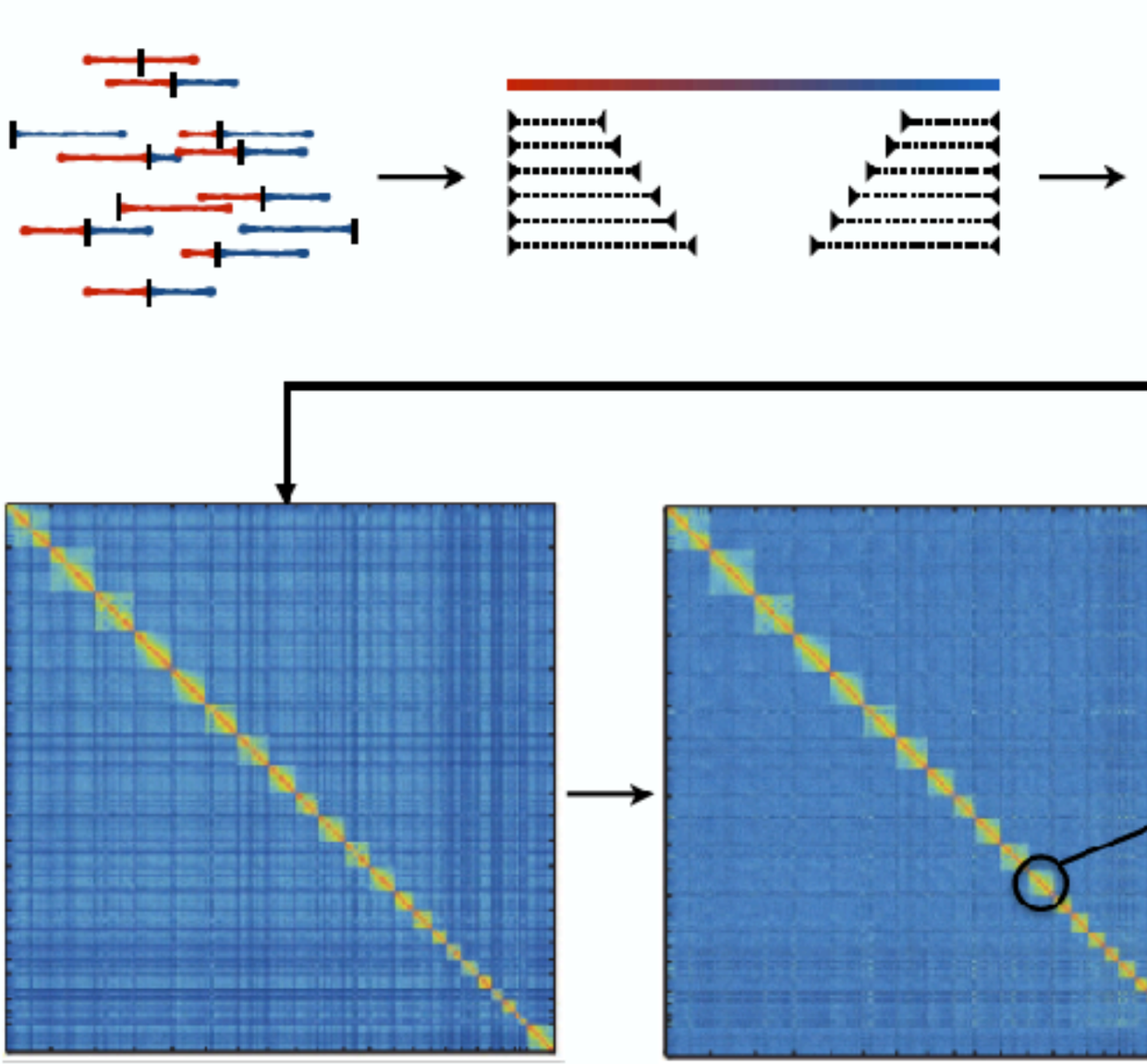
Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

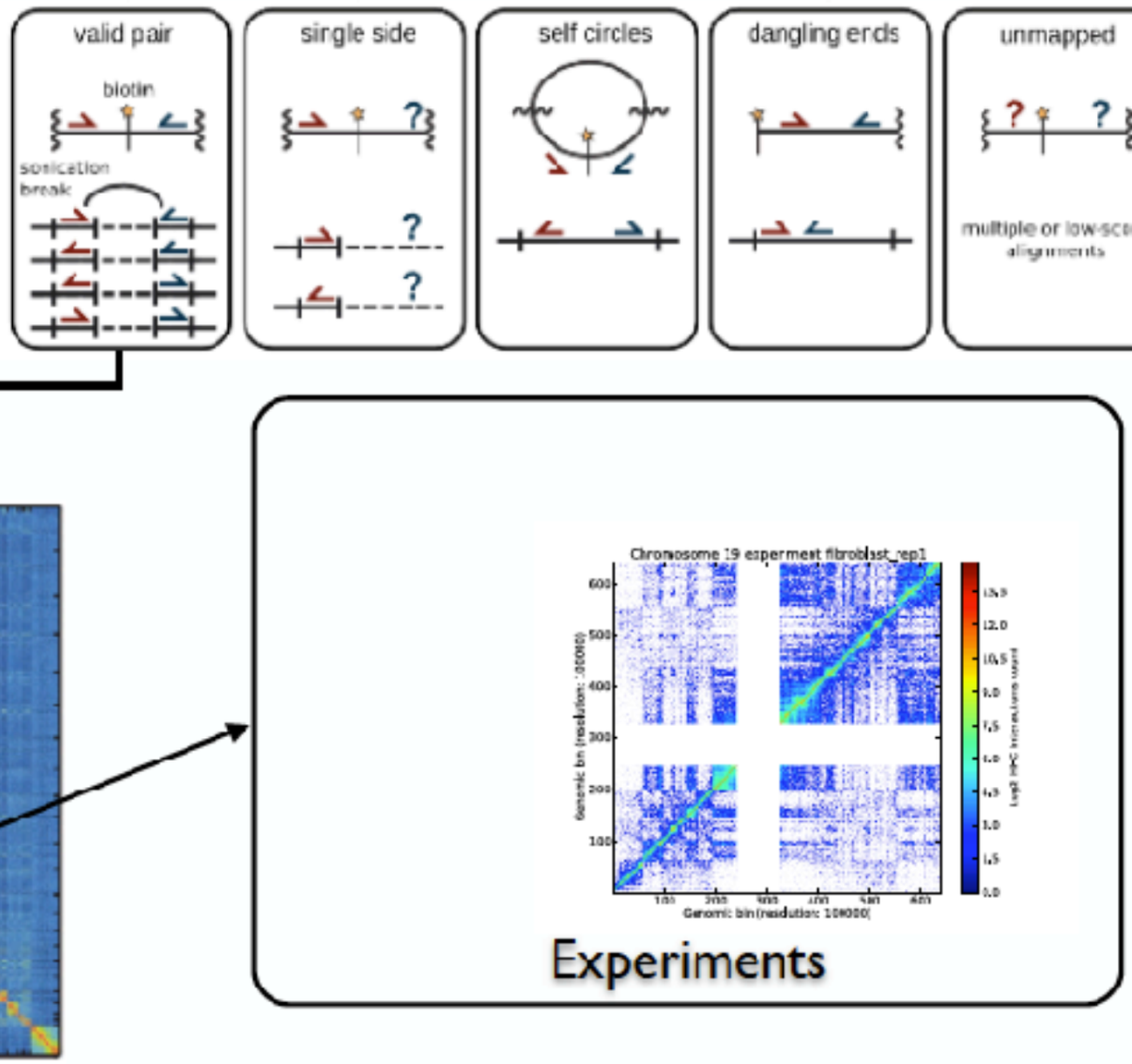
From FASTQ to interaction matrices



Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

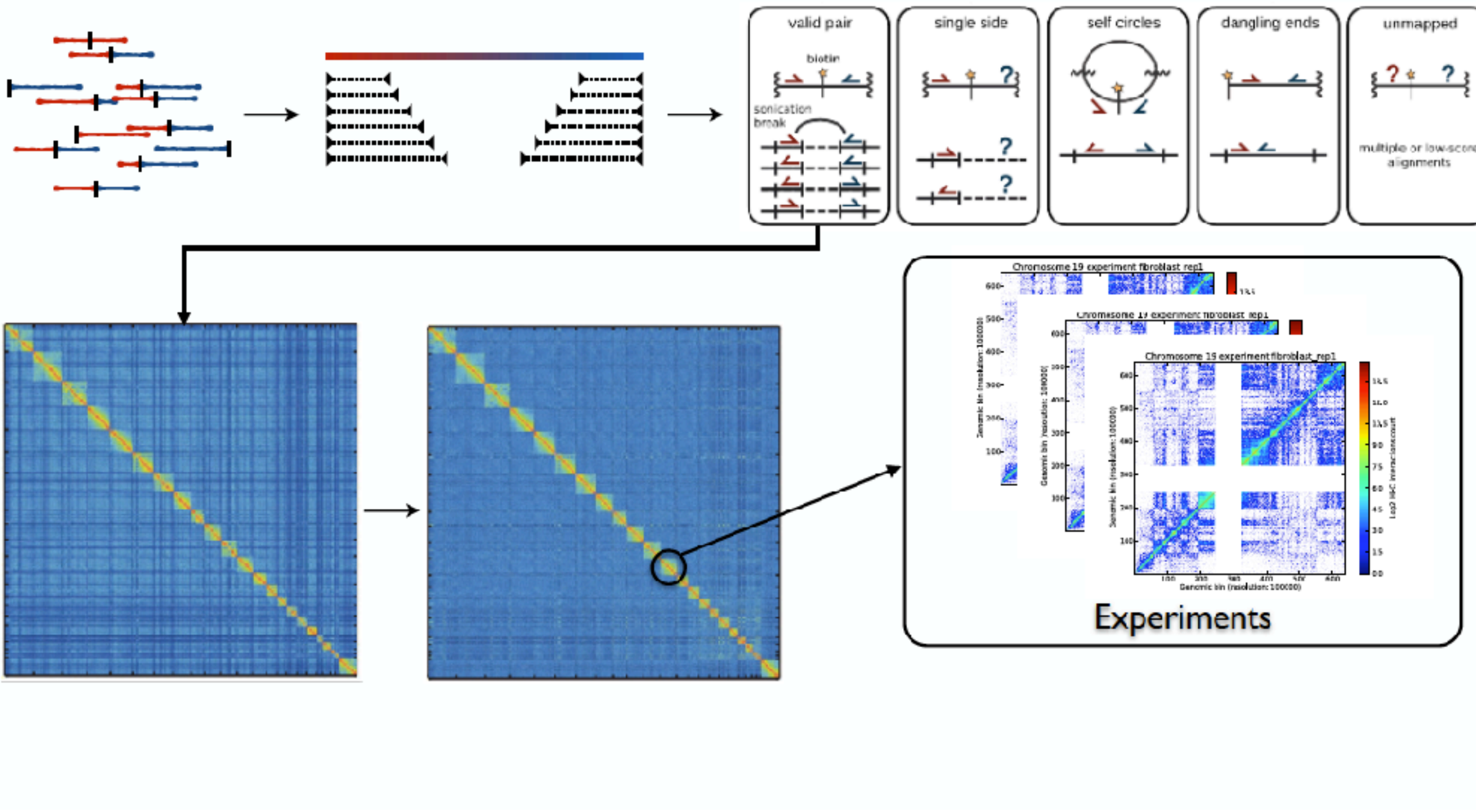
From FASTQ to interaction matrices





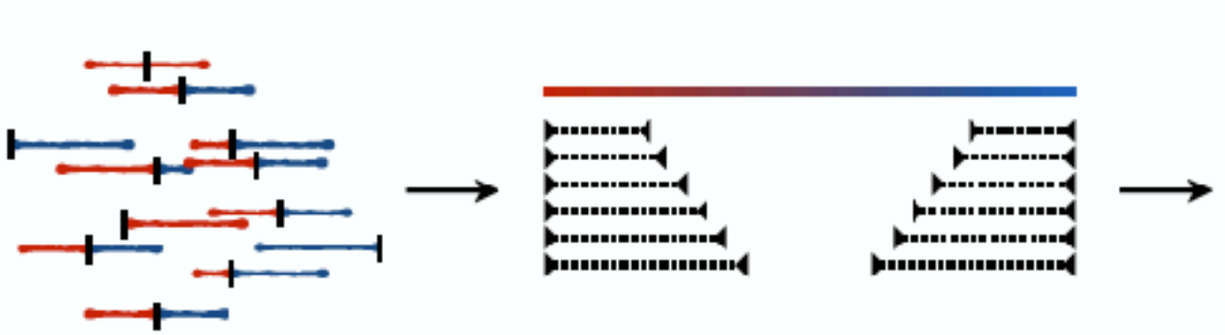
Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

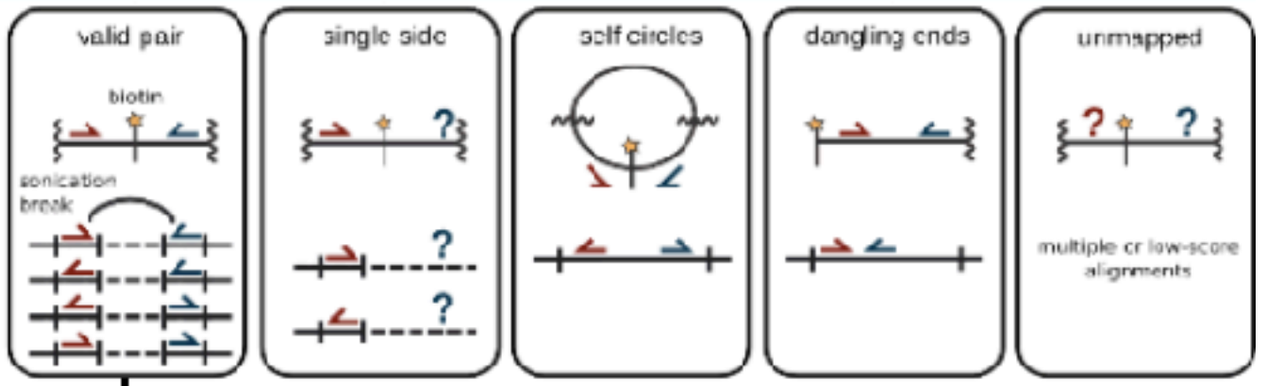
From FASTQ to interaction matrices

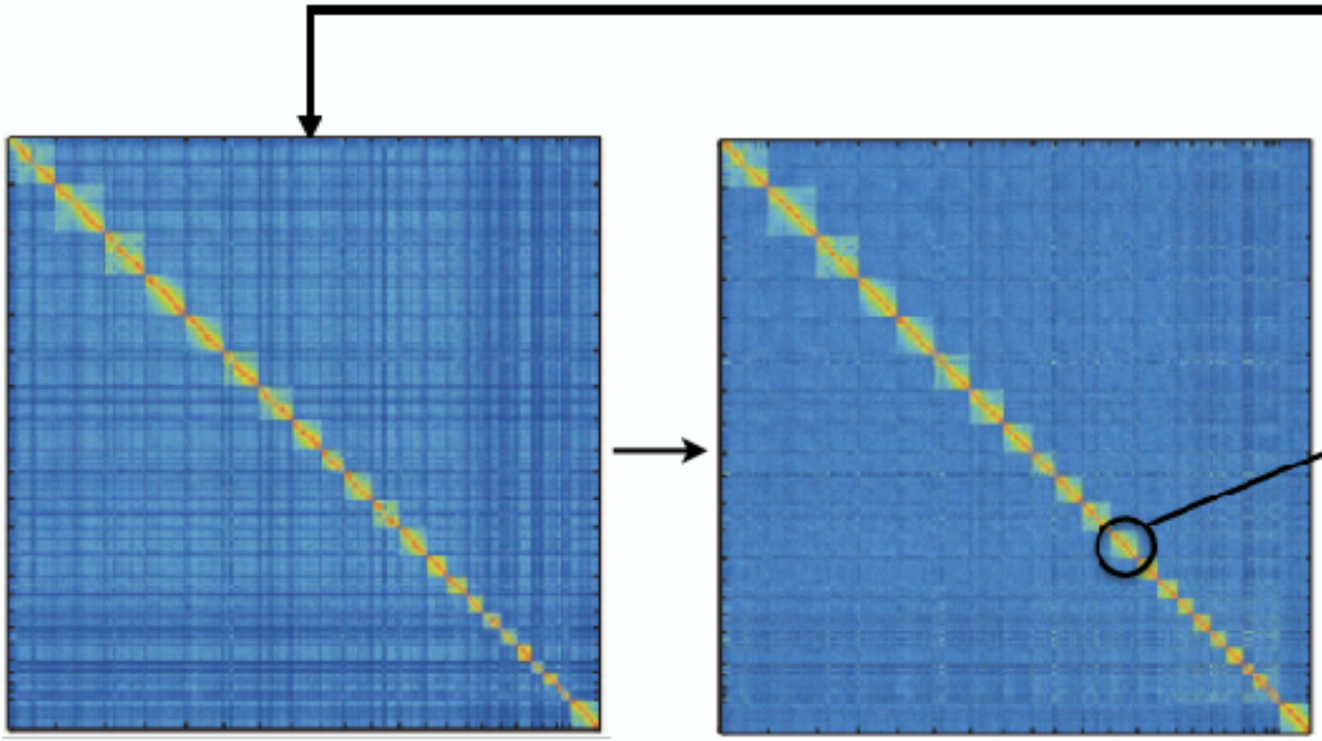


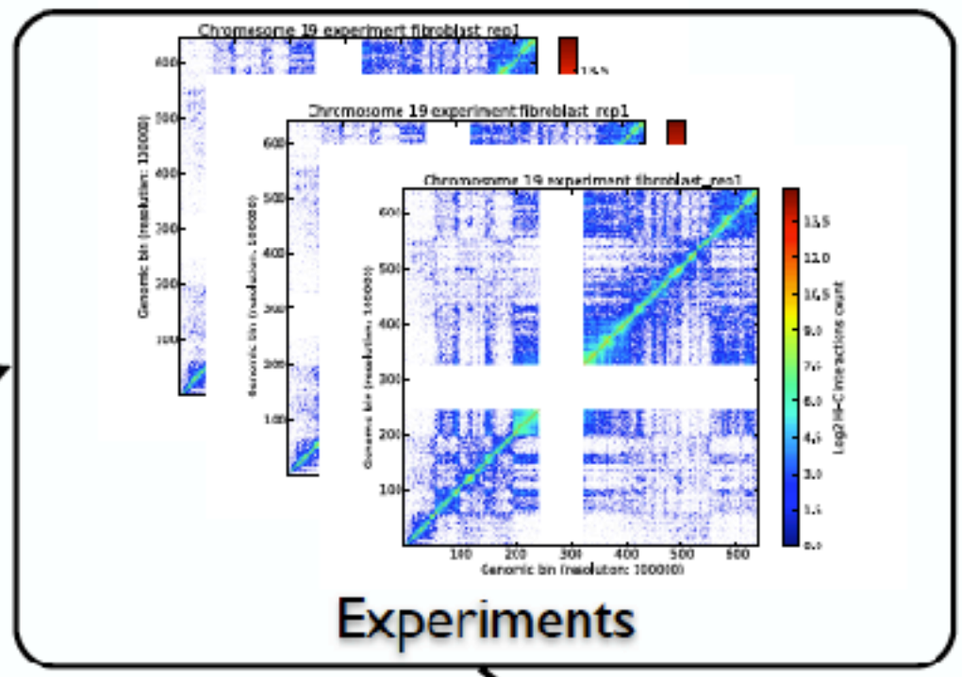
Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

From FASTQ to interaction matrices









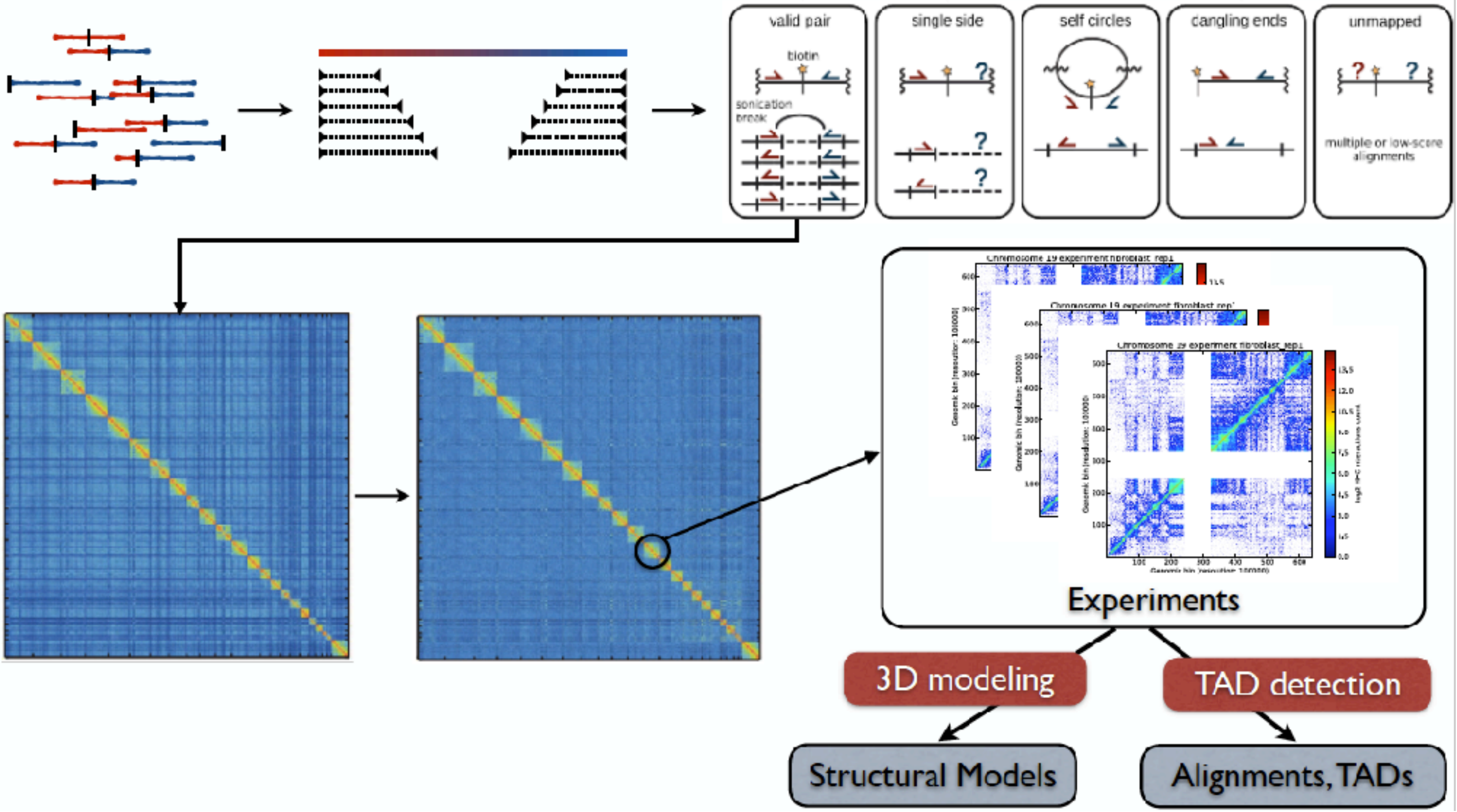
TAD detection

Alignments, TADs

Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

cnag CRG

From FASTQ to interaction matrices



Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

Many alternatives

Tool	Short-read aligner(s)	Mapping improvement	Read filtering	Read-pair filtering	Normalization	Visualization	Confidence estimation	Implementation language(s)
HCUP [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
HiCorrector [57]	–	–	–	–	Matrix balancing	–	–	ANSI C
GOTHiC [135]	–	–	✓	✓	–	–	✓	R
HTC [121]	–	–	–	–	Two options[f]	✓	✓	R
chromoR [59]	–	–	–	–	Variance stabilization	–	–	R
HiFive [136]	–	–	✓	✓	Three options[g]	✓	–	Python
Fit-Hi-C [20]	–	–	–	–	–	✓	✓	Python

Analysis methods for studying the 3D architecture of the genome
Ay, F. & Noble, W. S. Genome Biol. 16, 183 (2015).

cnag CRG

TADbit

a bioinformatic framework to analyse Hi-C experiments

François Serra, Paula Soler & Marc A. Marti-Renom
Structural Genomics Group (CNAG-CRG)

From FASTQ to interaction matrices

1. WHICH IS THE SIZE OF MY DATA?

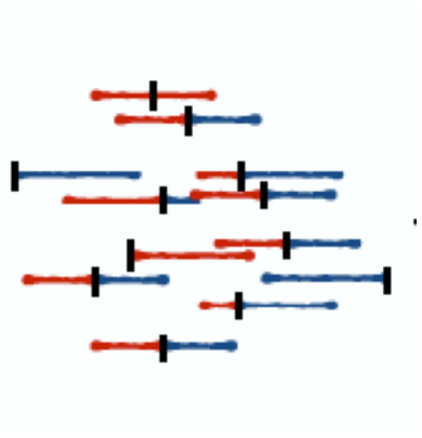
2. IS THE QUALITY OF MY SEQUENCED DATA OK?

From FASTQ to interaction matrices

1. WHICH IS THE SIZE OF MY DATA?

```
wc -l + name_file(fastq format)
```

The number of lines **/4** Number of reads



From FASTQ to interaction matrices

ID + description

Sequence

```
@NS500645:59:HCL32BGXY:1:11101:14163:1054 1:N:0:GCCAAT
CATTCNTAAAGAAAAGAATTTTCAACNCAGAATTTCATATCCAGCCAACTAAGCTAGCTTCAAGGAAATACATTT
+
AAAAA#EEAEEEEEEEEEEAE/EEEE#EEEEEEEEEEEEEEEAE</EE/EEEEAEEEEEEEEEEEAAEE</EEE
```

Phred Quality Score*

** Phred quality scores are defined as a property which is logarithmically related to the **base-calling error probabilities***

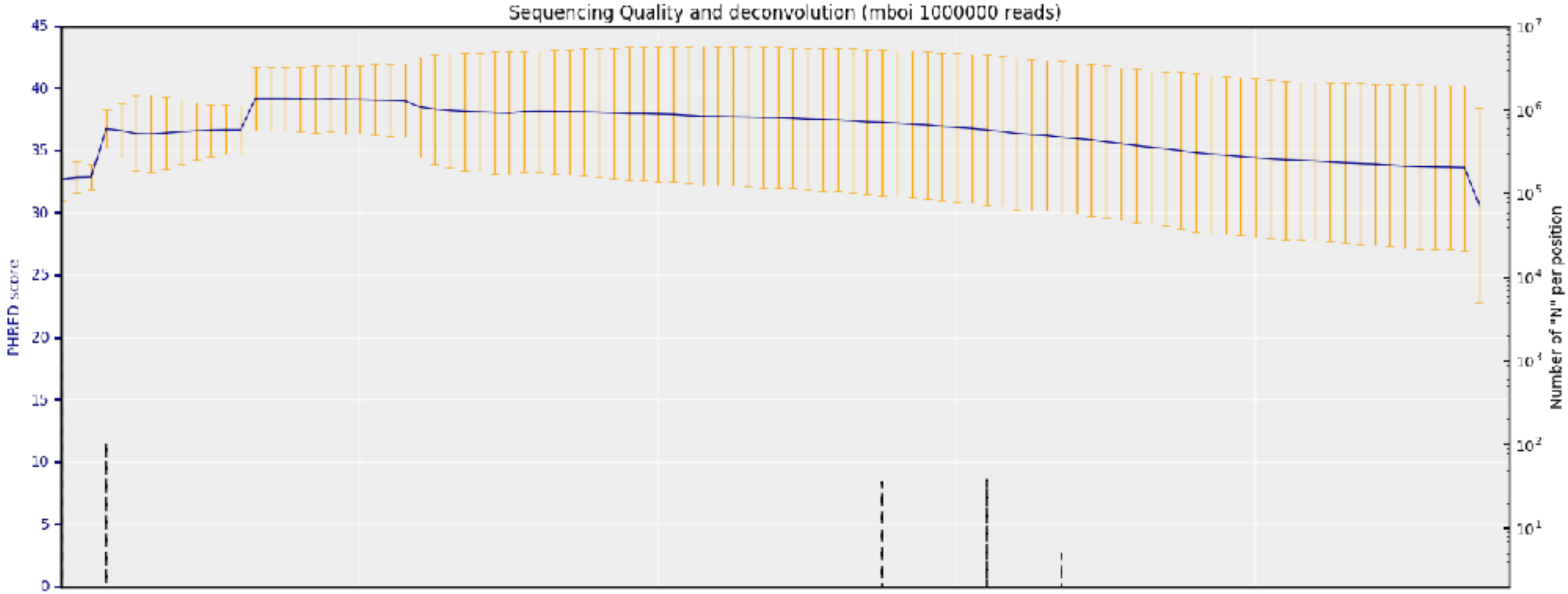
From FASTQ to interaction matrices

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

From FASTQ to interaction matrices

- Phred quality (QV) score string
- Number of “N” per position



From FASTQ to interaction matrices

HINDIII

5' N A|A G C T T N N N 3'
3' N T T C G|A A N N N 5'

MBOI

5' N N N|G A T C N N N 3'
3' N N N C T A G|N N N 5'

- **Percentage of dangling-ends** → *The number of time a digested site is found at the beginning of a read*
- Percentage of ligation sites
- Percentage of digestion sites
- **Percentage of undigested sites**

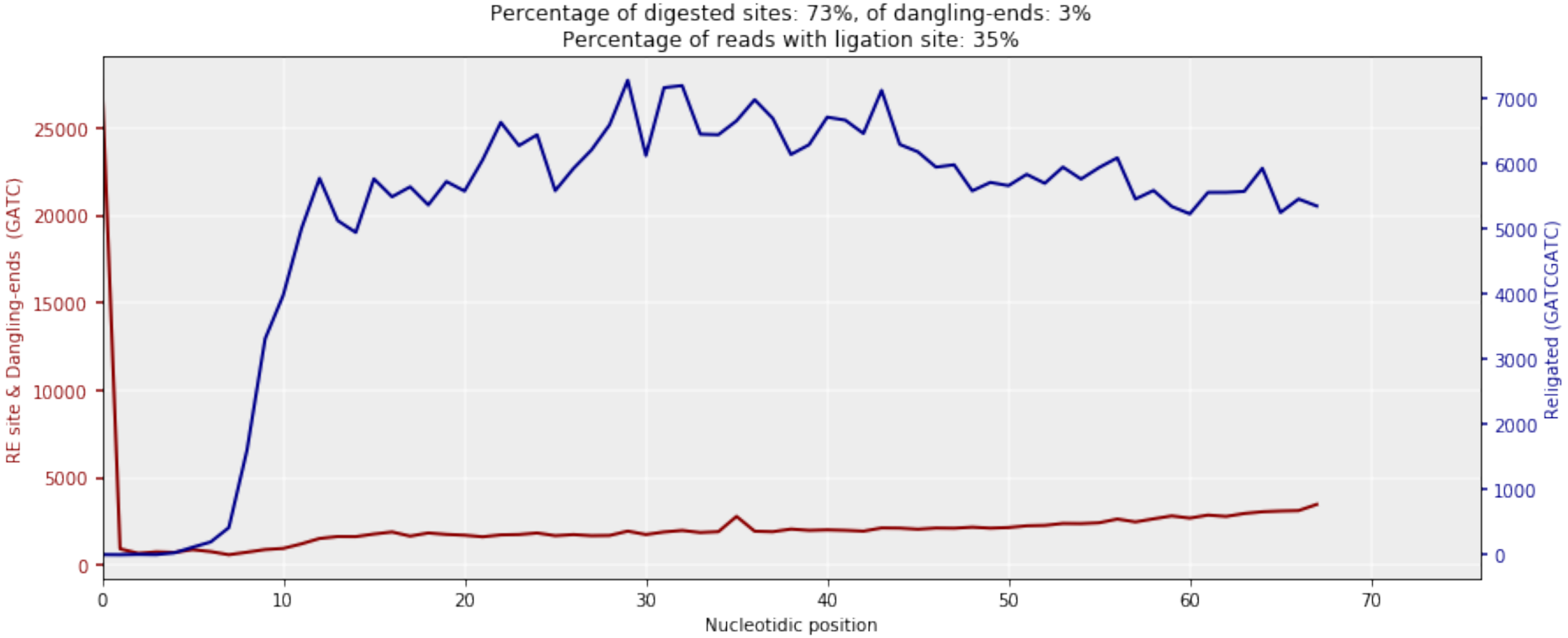
From FASTQ to interaction matrices

HINDIII

5’ N A|A G C T T N N N 3’
3’ N T T C G|A A N N N 5’

MBOI

5’ N N N|G A T C N N N 3’
3’ N N N C T A G|N N N 5’



From FASTQ to interaction matrices

HINDIII

5' N A|A G C T T N N N 3'
3' N T T C G|A A N N N 5'

MBOI

5' N N N|G A T C N N N 3'
3' N N N C T A G|N N N 5'