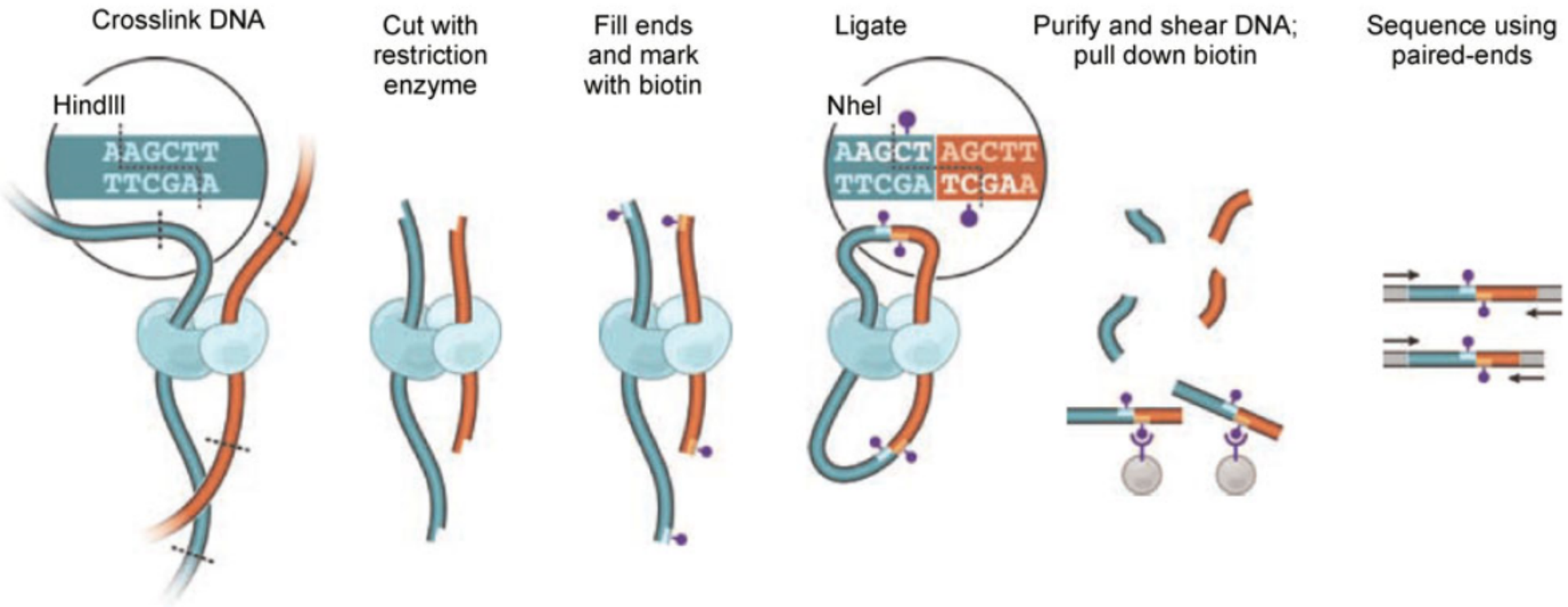
TADbit

a bioinformatic framework to analyse Hi-C experiments

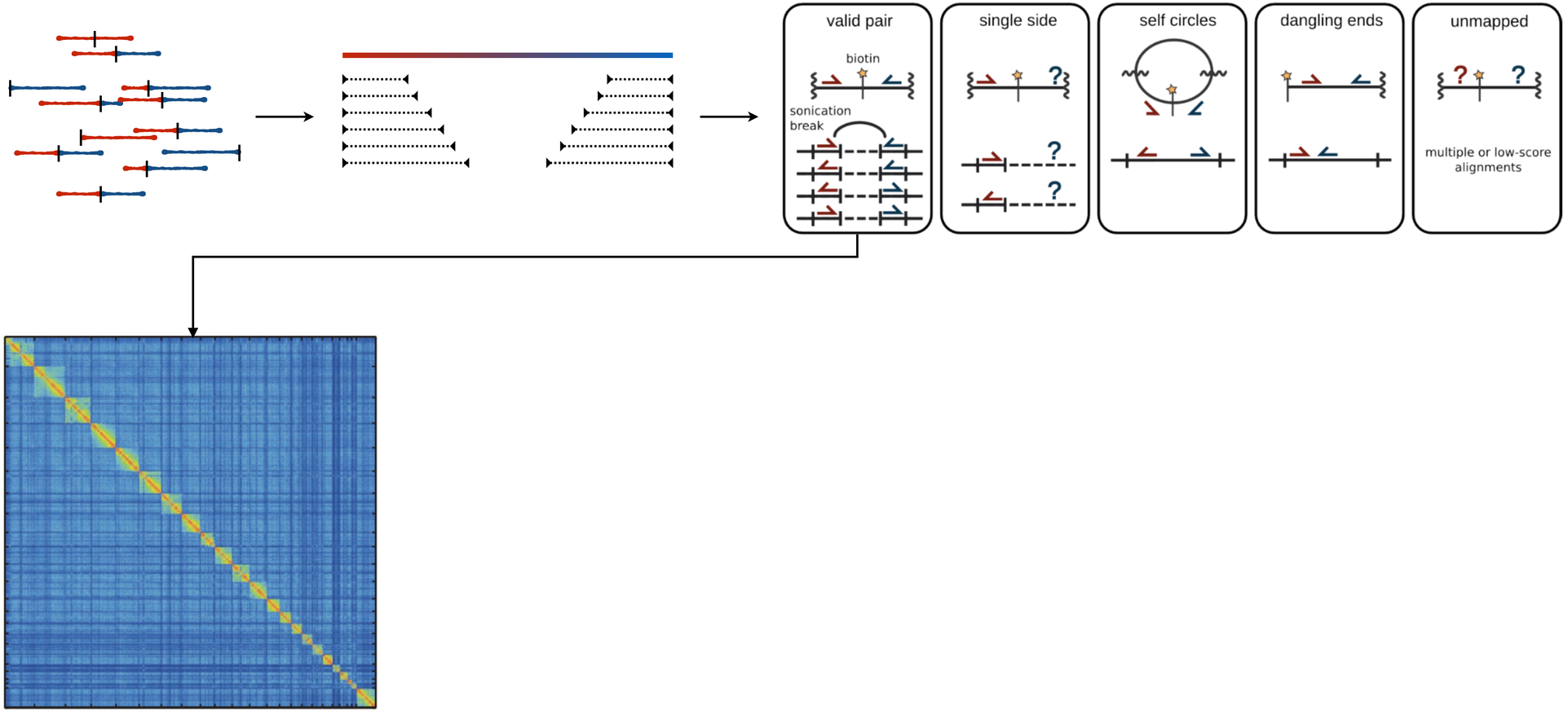
Marco Di Stefano, David Castillo & Marc A. Marti-Renom
Structural Genomics Group (CNAG-CRG)

Hi-C experiment



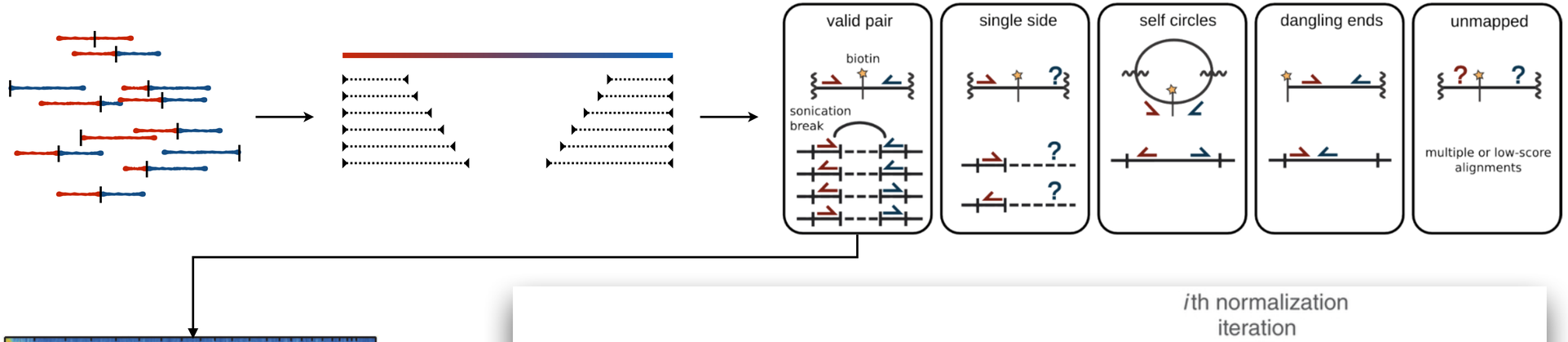
Lieberman-Aiden, E. … Dekker, J. (2009). Science, 326(5950),

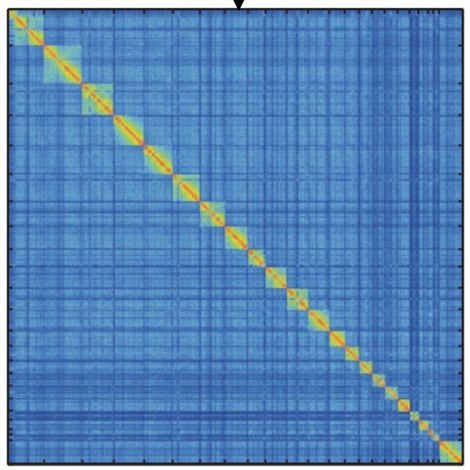
From FASTQ to interaction matrices

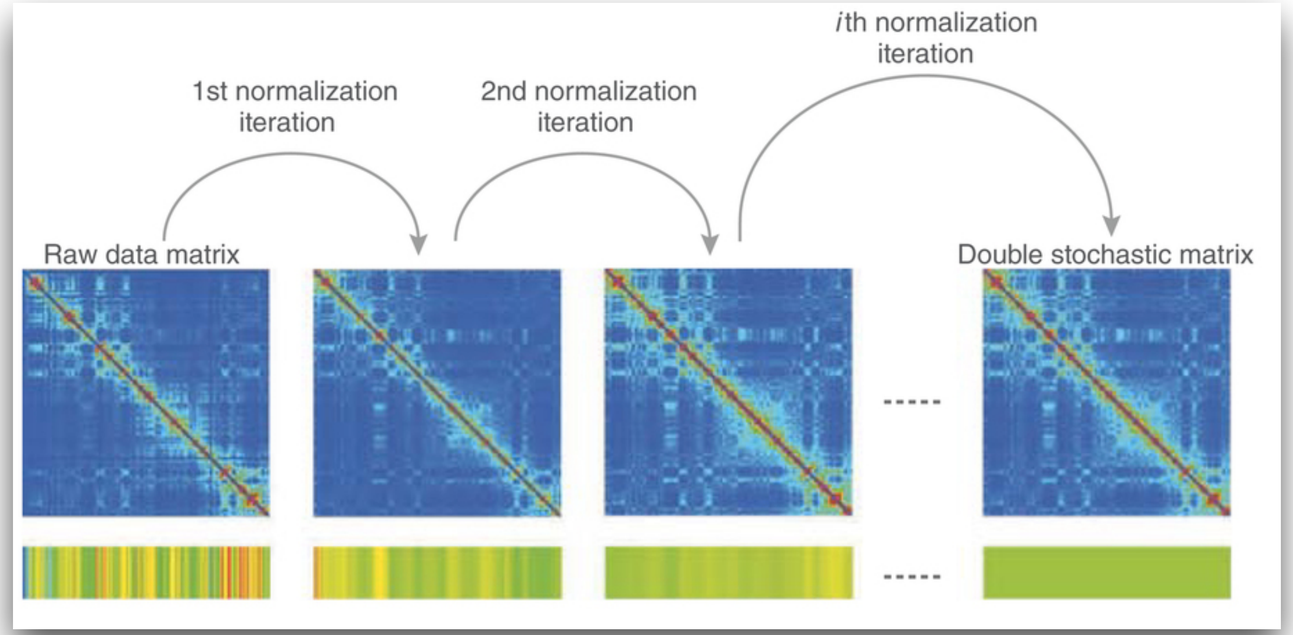


Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

From FASTQ to interaction matrices



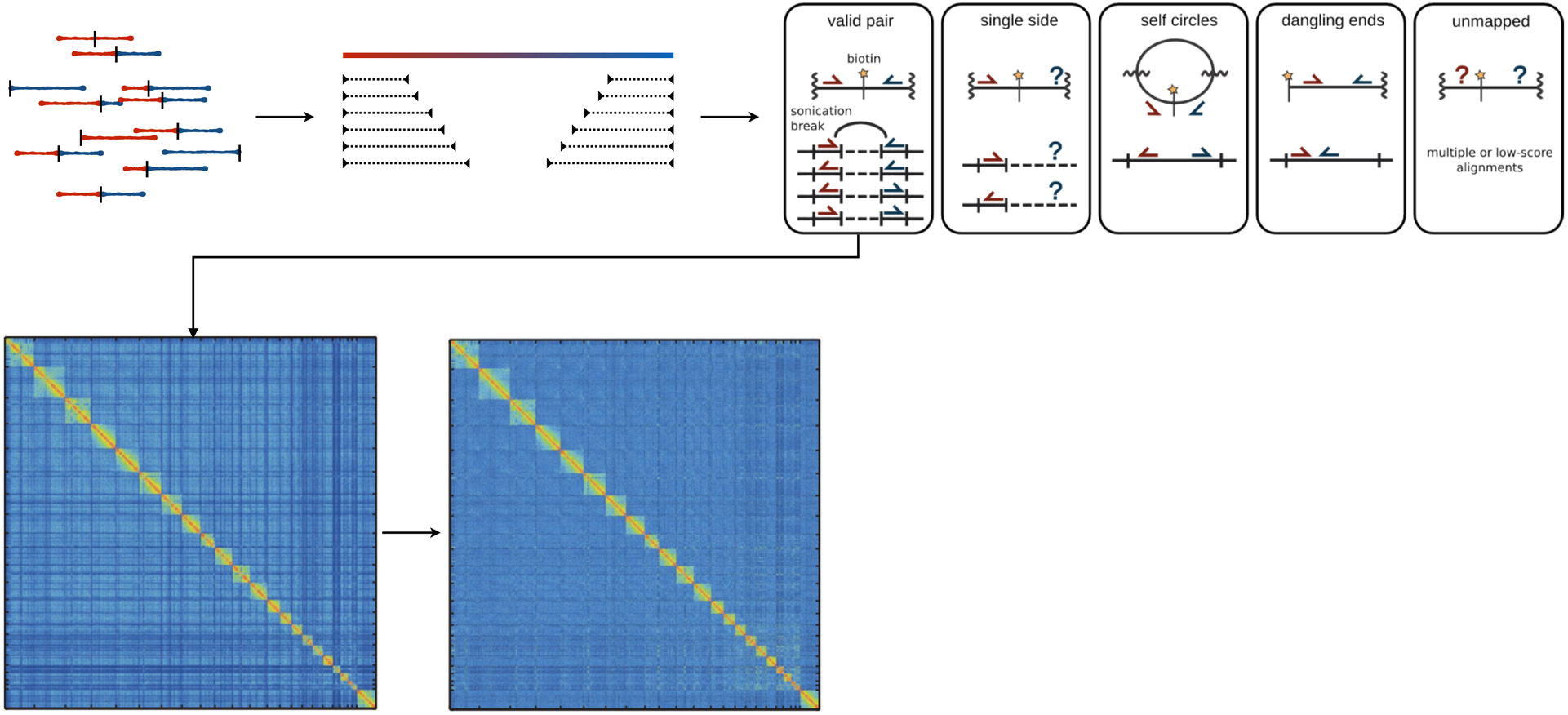




Zooming in on genome organization.
Zhou, X. J., & Alber, F. Nature Methods (2012)

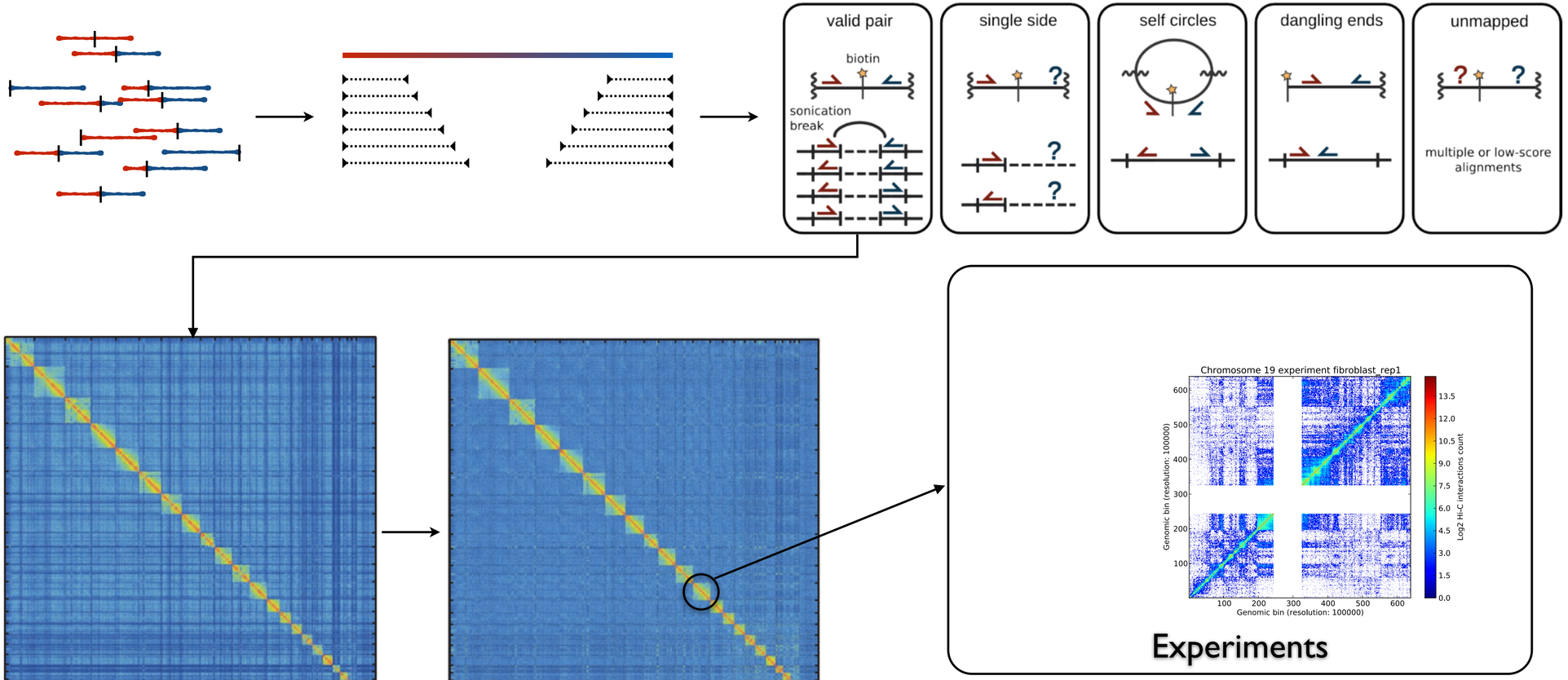
Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

From FASTQ to interaction matrices



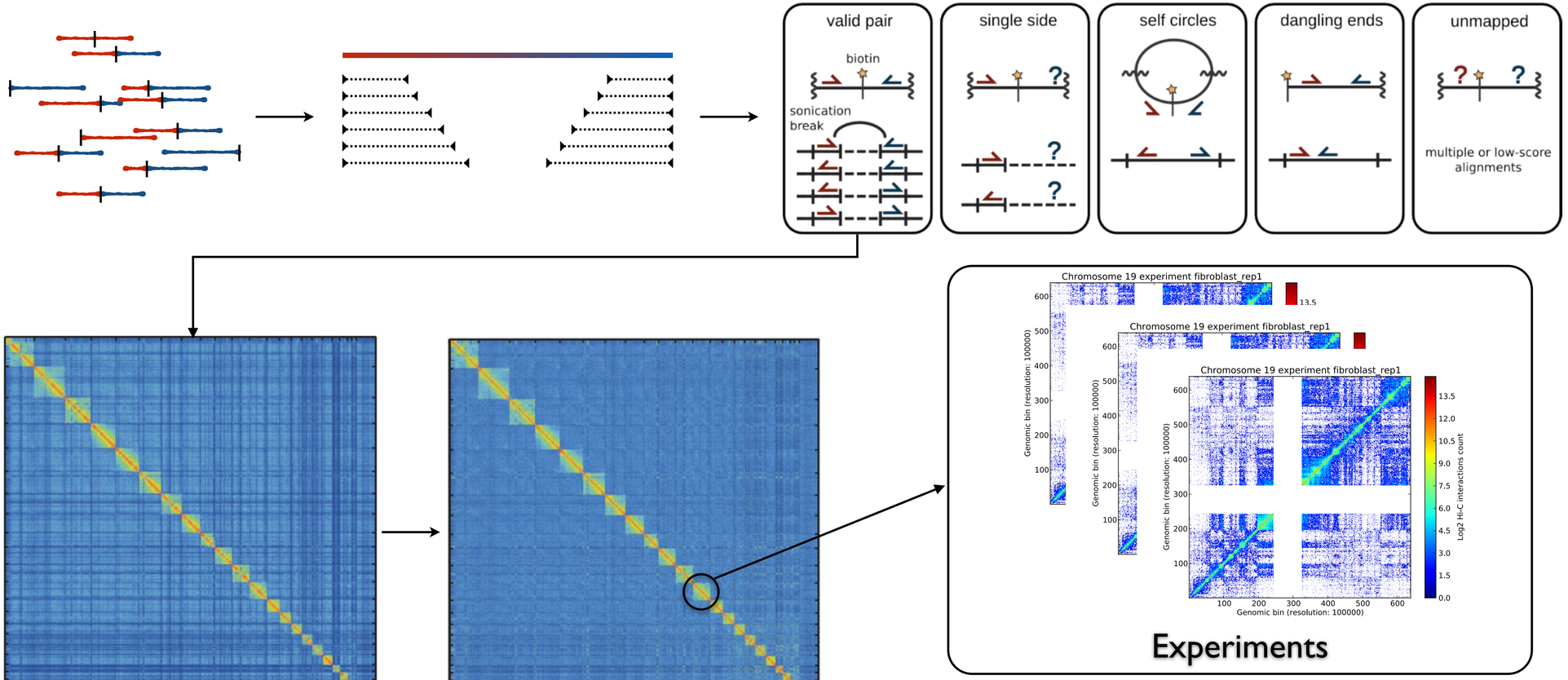
Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

From FASTQ to interaction matrices



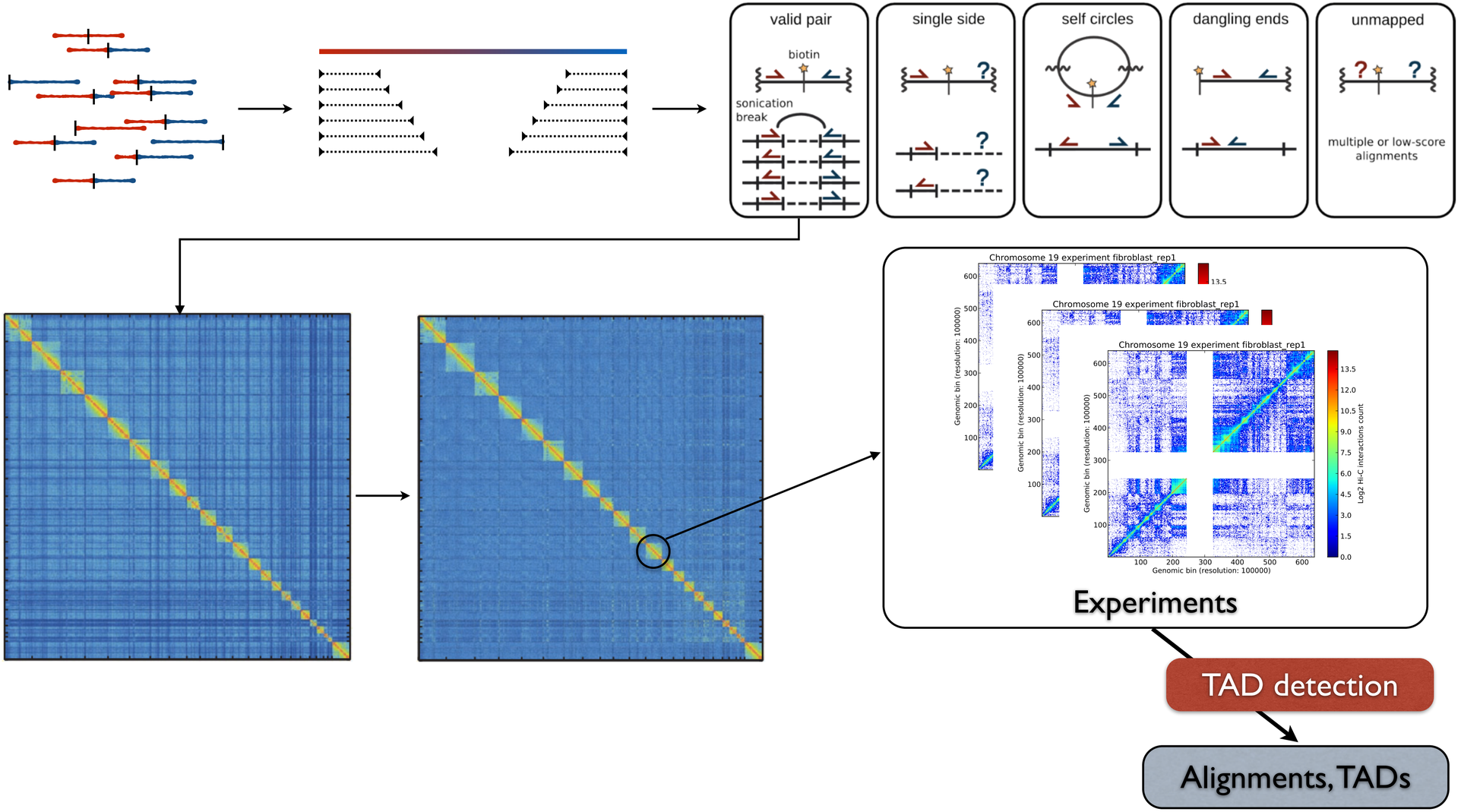
Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

From FASTQ to interaction matrices



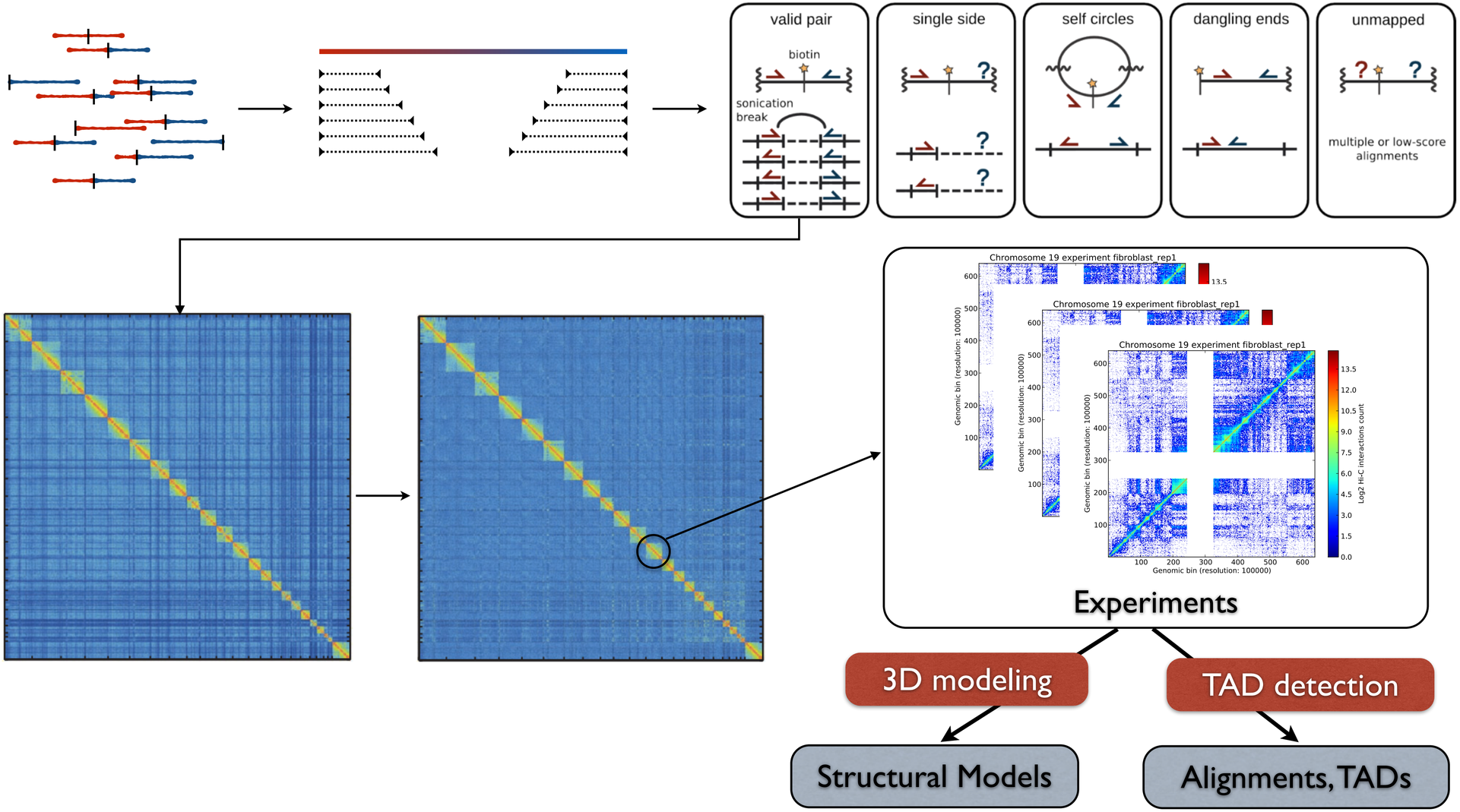
Iterative correction of Hi-C data reveals hallmarks of chromosome organization.
Imakaev et al. Nature Methods (2012)

From FASTQ to interaction matrices



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)

Many alternatives

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

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

Many alternatives

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

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

Many alternatives

Method *available online	Representation	Scoring				Sampling	Models
		U_{3C}		U_{Biol}	U_{Phys}		
		$F_{ij} \rightarrow D_{ij}$ conversion	Functional form				
ChromSDE* [37]	Points	$D_{ij} = \begin{cases} (\frac{1}{F_{ij}})^{\alpha} \text{ if} F_{ij} > 0 \\ \infty \text{ if } F_{ij} = 0 \end{cases}$ α is optimized	$\sum_{(i,j)D_{ij}<\infty)} \frac{(r_{ij}^2 - D_{ij}^2)}{D_{ij}} - \lambda \sum_{(i,j)} r_{ij}^2$ where λ is set to 0.01	N/A	N/A	Deterministic semidefinite programming to find the coordinates	Consensus
ShRec3D* [38]	Points	$D_{ij} = \begin{cases} \left(\frac{1}{F'_{ij}}\right)^{\alpha} & \text{if } F'_{ij} > 0 \\ \frac{N^2}{\sum_{(i,j)} F_{ij}} & \text{if } F'_{ij} = 0 \end{cases}$ F'_{ij} is the original F_{ij} corrected to satisfy all triangular inequalities with the shortest path reconstruction	N/A	N/A	N/A	Deterministic transformations of D_{ij} into coordinates	Consensus
TADbit* [43]	Spheres	$D_{ij} \propto \begin{cases} \alpha F_{ij} + \beta & \text{if } F_{ij} < \gamma' \text{ or } F_{ij} > \gamma \\ \frac{s_i + s_j}{2} & \text{if } \lvert i-j \rvert = 1 \end{cases}$ α and β are estimated from the max and the min F_{ij}, from the optimized max distance and from the resolution. $\gamma' < \gamma$ are optimized too. s_i is the radius of particle i	$\sum_{(i,j)} k_{ij}(r_{ij} - D_{ij})^2$ where $k_{ij} = 5$ if $\lvert i-j \rvert = 1$ or proportional to F_{ij} otherwise	Yes	U_{excl} and U_{bond} have harmonic forms	Monte Carlo (MC) sampling with Simulated annealing and Metropolis scheme	Resampling
BACH* [45]	Points	$D_{ij} \propto \frac{B_i B_j}{F_{ij}^{\alpha}}$. The biases B_i and B_j and α are optimized	$b_{ij} D_{ij}^{1/\alpha} + c_{ij} \log(D_{ij})$ where b_{ij} and c_{ij} are optimized parameters	No	No	Sequential importance and Gibbs sampling with hybrid MC and adaptive rejection	Population
Giorgetti et al. [40]	Spheres	Particles interact with pair-wise well potentials of depths B_{ij} and contact radius a, which is larger than a hard-core radius and smaller than a maximum contact radius. The parameters are optimized over all the population of models		No	N/A	MC sampling with metropolis scheme	Population
Duan et al. [41]	Spheres	$\overline{F_{\lvert i-j \rvert}} = \frac{\sum_{k=0}^{N-i-j} F_{(i,k+(i-j))}}{N-\lvert i-j \rvert}$ is the average of F_{ij} at genomic distance $\lvert i-j \rvert$ expressed in kb. $D_{ij} = \overline{F_{\lvert i-j \rvert}} \times 7.7 \times \lvert i-j \rvert$ assuming that α 1 kb maps onto 7.7 nm	$\sum_{(i,j)} (r_{ij} - D_{ij})^2$	Yes	U_{excl} and U_{bond} have harmonic forms	Interior-point gradient-based method	Resampling
MCMC5C* [49]	Points	$D_{ij} \propto \frac{1}{F_{ij}^{\alpha}}$ where is optimized	$\sum_{(i,j)} (F_{ij} - r_{ij}^{-1/\alpha})^2$	N/A	N/A	MC sampling with Markov chain based algorithm	Resampling
PASTIS* [47]	Points	$D_{ij} \propto \frac{1}{F_{ij}^{\alpha}}$ where α is optimized	$b_{ij} D_{ij}^{1/\alpha} + c_{ij} \log(D_{ij})$ where b_{ij} and c_{ij} are optimized parameters	No	No	Interior point and isotonic regression algorithms	Resampling
Meluzzi and Arya [48]	Spheres	$\sum_{(i,j)} k_{ij} r_{ij}^2$ where k_{ij} are adjusted such that the contact probabilities computed on the models match the F_{ij}		No	U_{excl} is a pure repulsive LJ potential. U_{bond} and U_{bend} have harmonic forms	Brownian dynamics	Resampling
AutoChrom3D* [44]	Points	$D_{ij} \propto \begin{cases} \alpha F_{ij} + \beta & \text{if } F_{min} < F_{ij} < F_{\gamma} \\ \alpha' F_{ij} + \beta' & \text{if } F_{\gamma} < F_{ij} < F_{max} \end{cases}$ where F_{min} (F_{max}) are the min(max) of F_{ij}. The parameters (α, β), (α', β') and F_{γ} are found using the nuclear size, the resolution and the decay of F_{ij} with $\lvert i-j \rvert$	$\sum_{(i,j)} \frac{(r_{ij} - D_{ij})^2}{D_{ij}^2}$	Yes	N/A	Non-linear constrained	Consensus
Kalhor et al. [14]	Spheres	$D_{ij} = R_{contact}$ to enforce the pair contact, if the normalized contact frequency F_{ij} is higher than 0.25. Otherwise the contact is not enforced	$\sum_{models} \sum_{(i,j)} k_{ij}(r_{ij} - D_{ij})^2$ where k_{ij} is different for pairs of particles, on different chromosomes, on the same chromosome, or connected	Yes	U_{excl} and U_{bond} have harmonic forms	Conjugate gradients sampling with Simulated annealing scheme	Population

* These methods are publicly available.

Restraint-based three-dimensional modeling of genomes and genomic domains.
Serra F, Di Stefano M, Spill YG, Cuartero Y, Goodstadt M, Baù D, Marti-Renom MA. FEBS Lett 589: 2987–2995 (2015)

TADbit

https://github.com/3DGenomes/tadbit

Current version: v1.0 | build passing | coverage 47% | license GPL

TADbit is a complete Python library to deal with all steps to analyze, model and explore 3C-based data. With TADbit the user can map FASTsQ files to obtain raw interaction binned matrices (Hi-C like matrices), normalize and correct interaction matrices, identify and compare the Topologically Associating Domains (TADs), build 3D models from the interaction matrices, and finally, extract structural properties from the models. TADbit is complemented by TADkit for visualizing 3D models.

François Serra
BSC

David Castillo
CNAG

PLOS COMPUTATIONAL BIOLOGY

advanced search

OPEN ACCESS PEER-REVIEWED

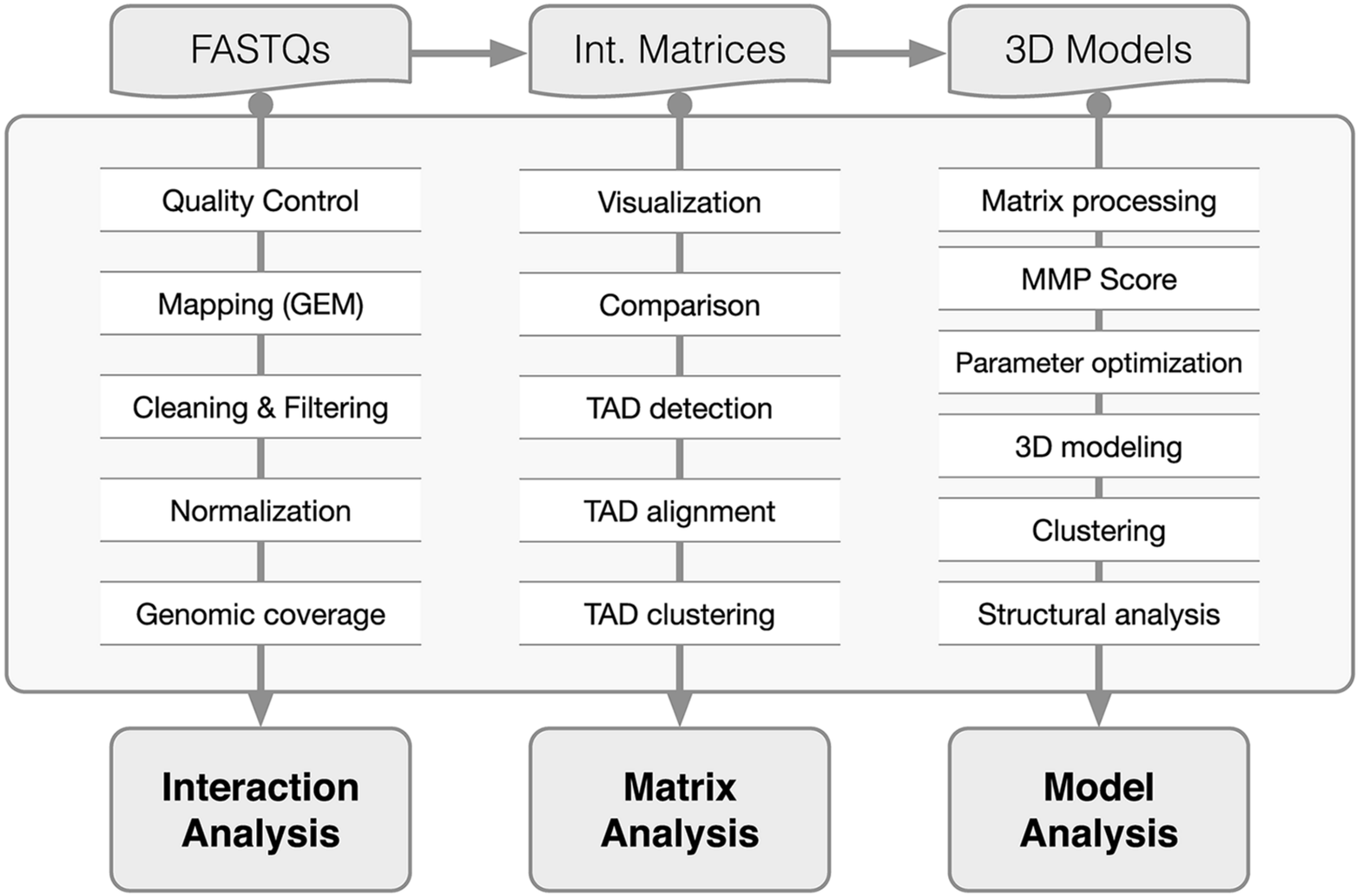
RESEARCH ARTICLE

Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors

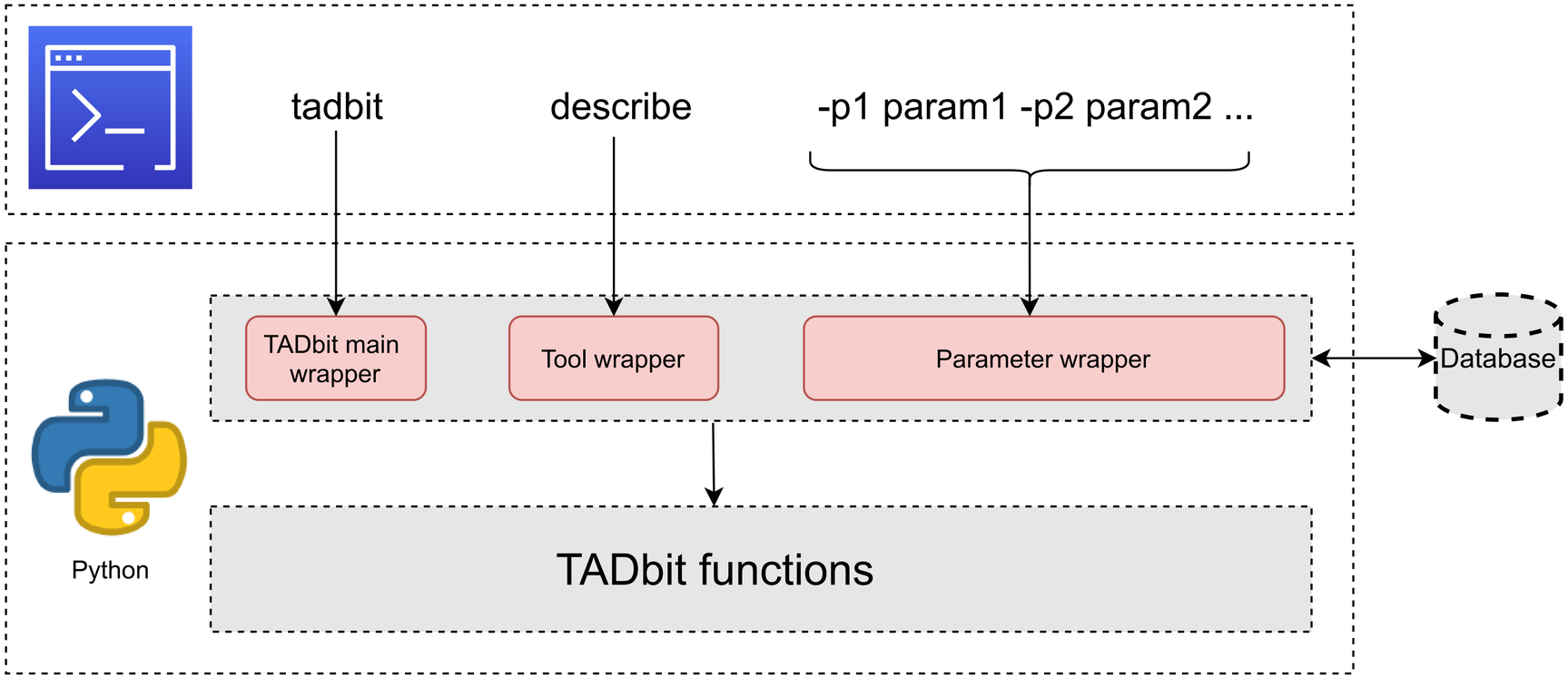
François Serra, Davide Baù, Mike Goodstadt, David Castillo, Guillaume J. Filion, Marc A. Marti-Renom

154 Save	143 Citation
8,474 View	31 Share

TADbit


FASTQs
Int. Matrices
3D Models
Quality Control
Mapping (GEM)
Cleaning & Filtering
Normalization
Genomic coverage
Visualization
Comparison
TAD detection
TAD alignment
TAD clustering
Matrix processing
MMP Score
Parameter optimization
3D modeling
Clustering
Structural analysis
Interaction Analysis
Matrix Analysis
Model Analysis

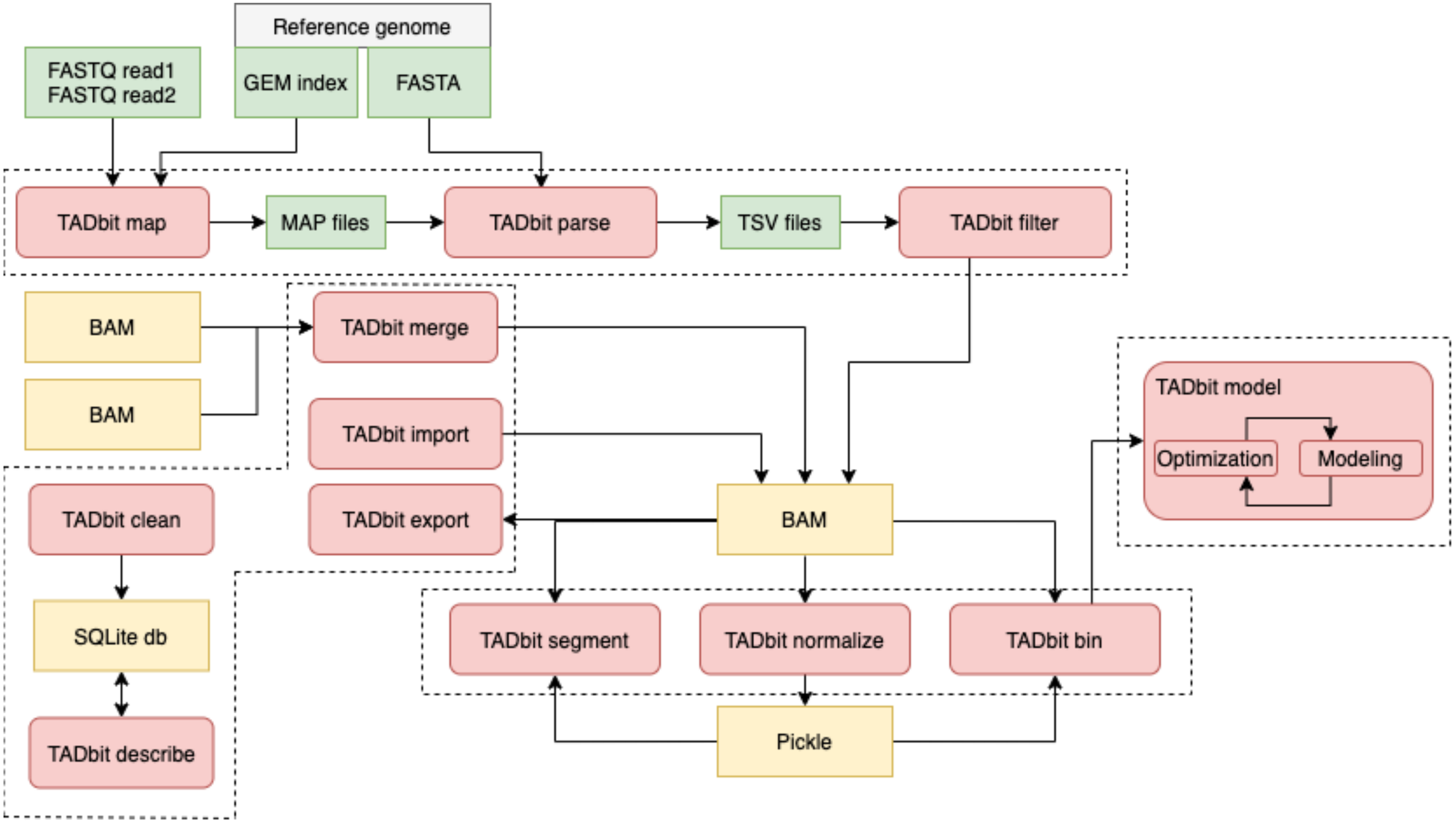
TADbit tools



Why TADbit tools?

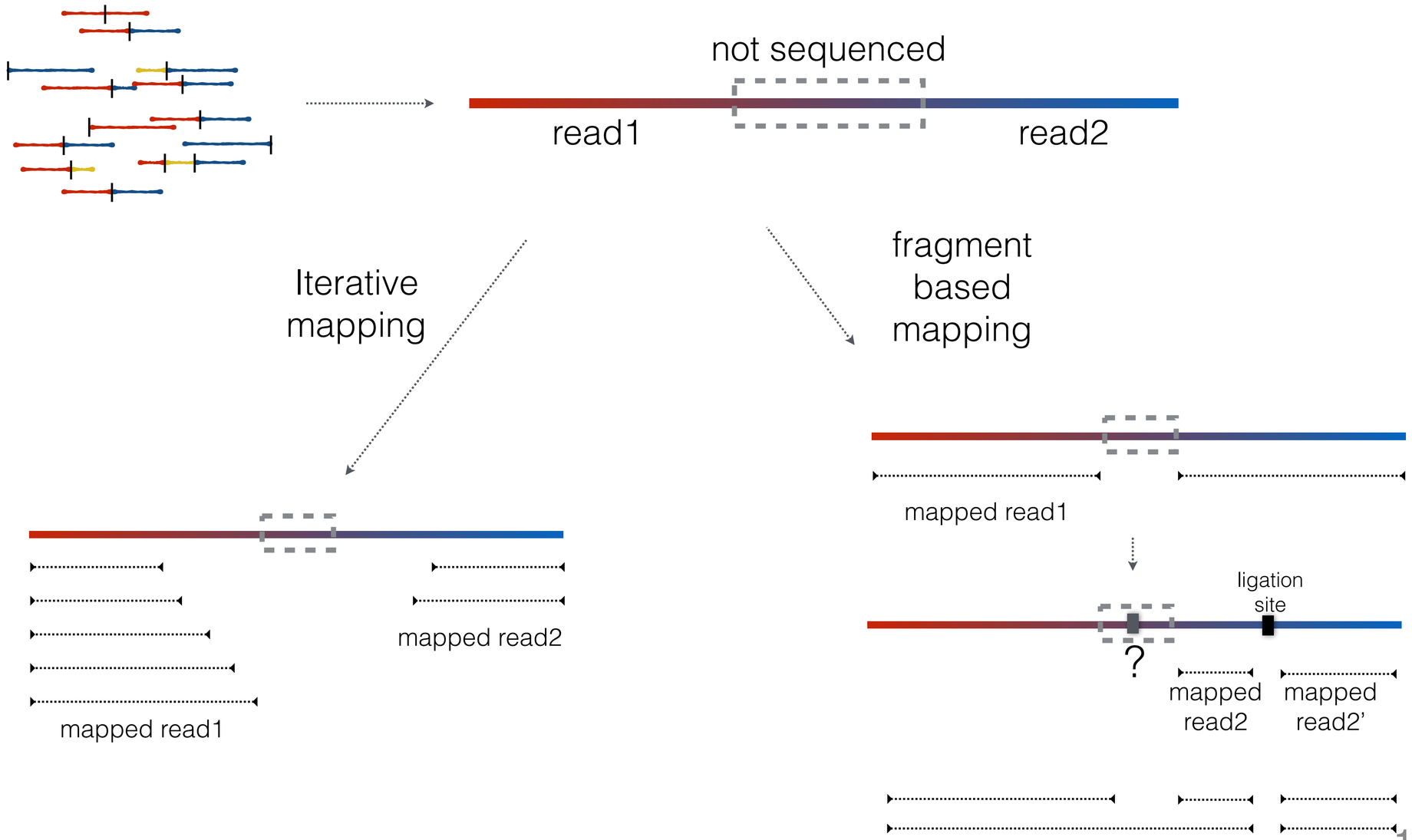
- Appearance of simplicity
- Bioinformaticians are familiar with command line
- The commands can be easily integrated in batch files and pipelines
- The folder structure created automatically when you run the tools is consistent and helps you maintaining an organized environment
- The database helps in the traceability and reproducibility

TADbit tools



TADbit map

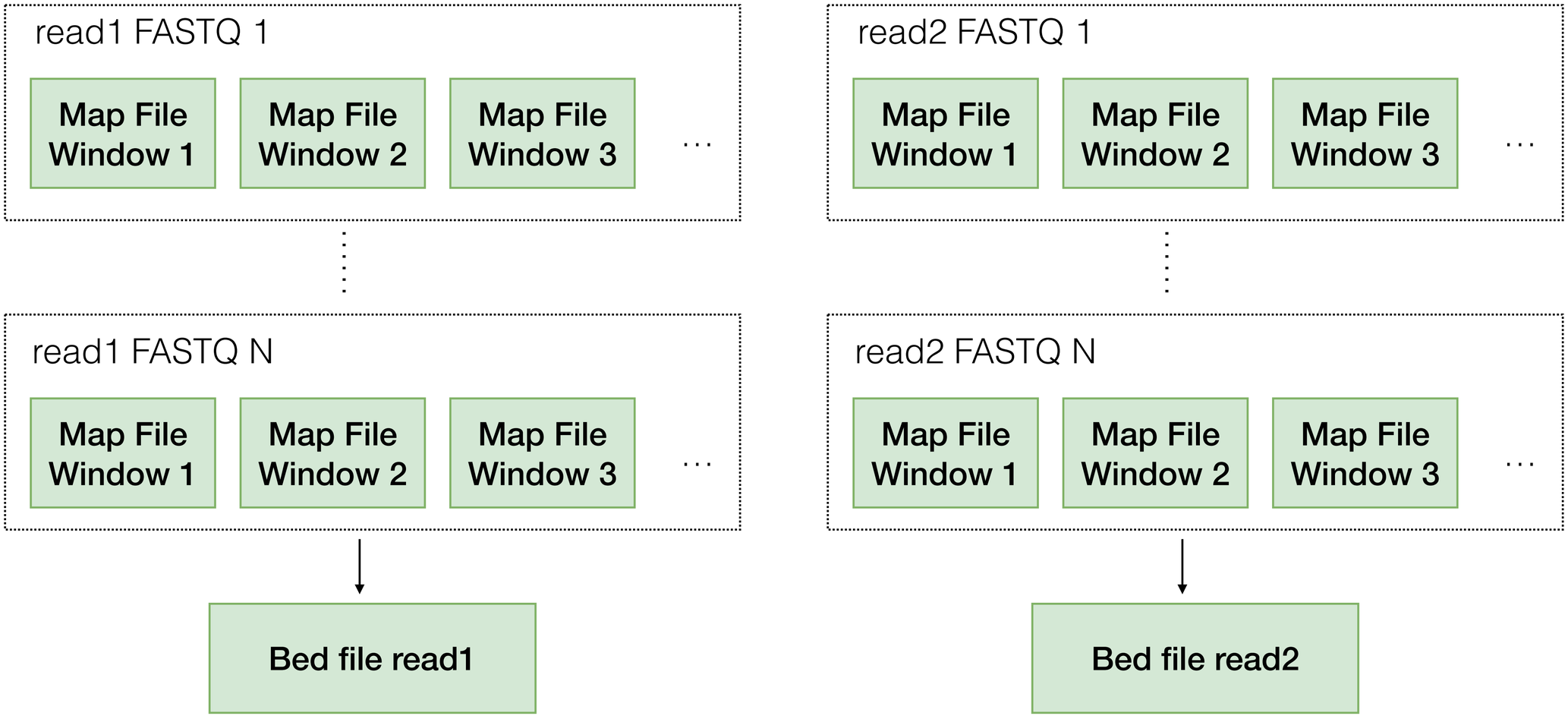
GEMv2, GEMv3, bowtie2, hisat2



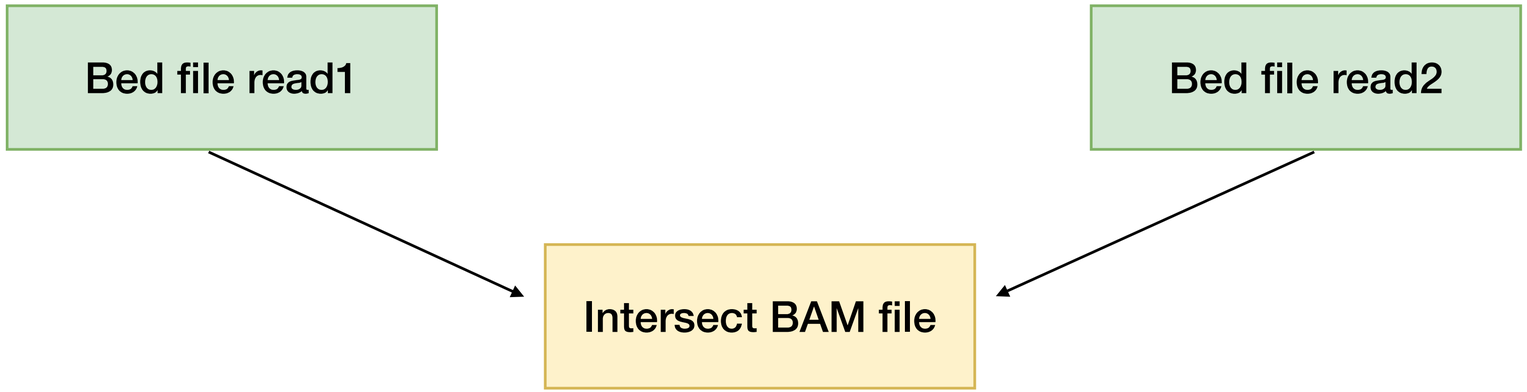
TADbit map

- Time consuming
- Slightly different results depending on the selected mapper
  - Allowed mismatch
- Requires lot of temporary disk space
  - Trim reads
  - Split reads
  - …

TADbit parse



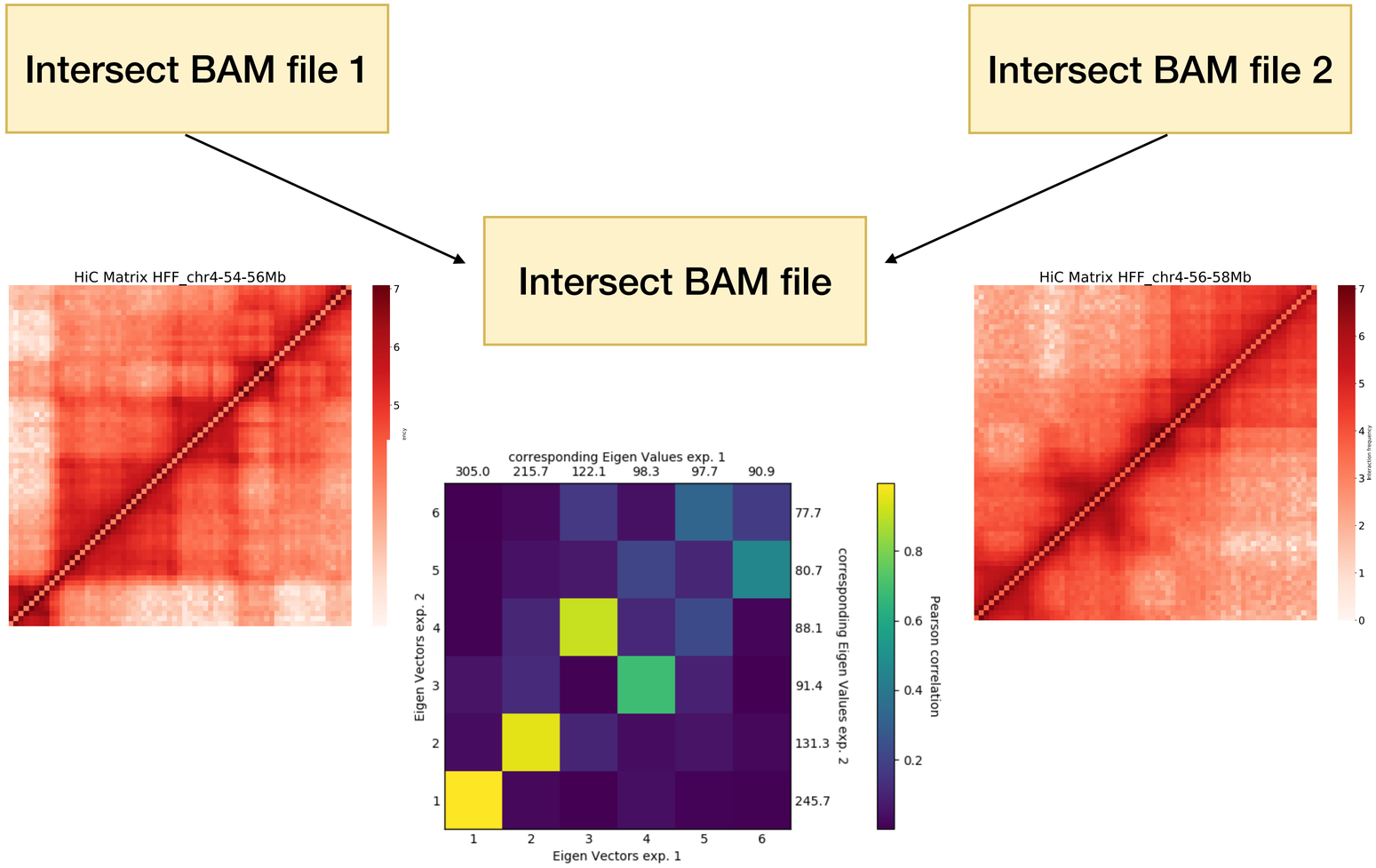
TADbit filter



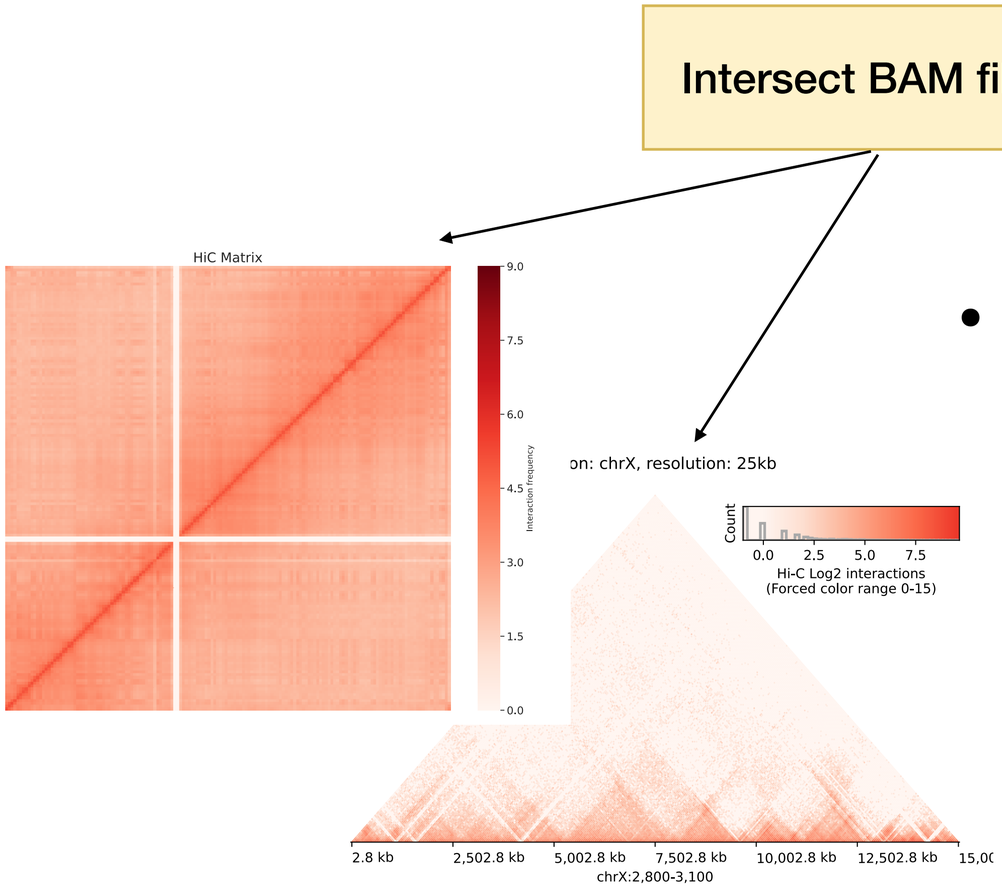
- Paired pseudo BAM (compressed and sorted)
- Each pair is categorised (tagged)
- Mirrored for fast access read1-read2 or read2-read1

TADbit merge

Merge and compare experiments

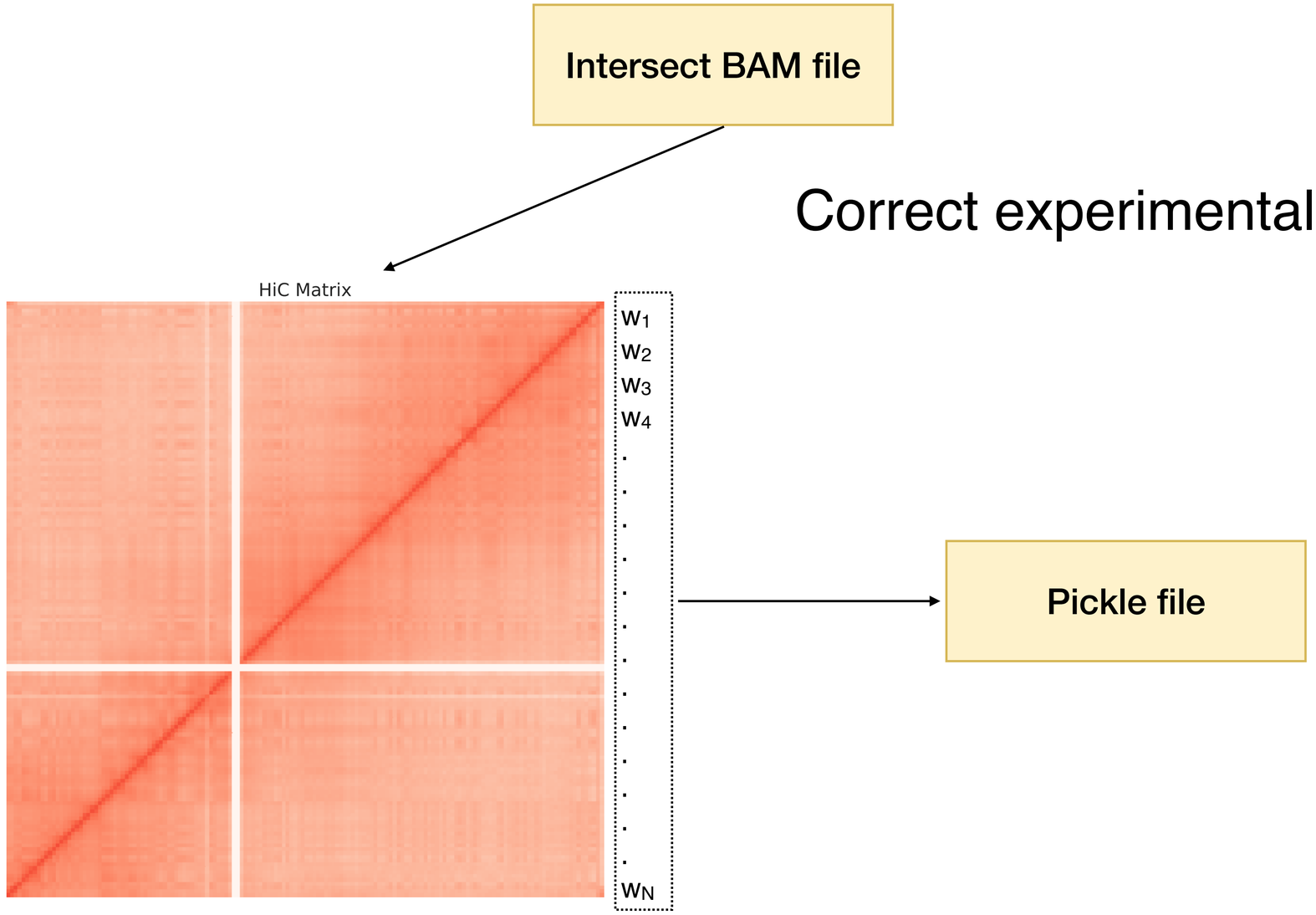


TADbit bin



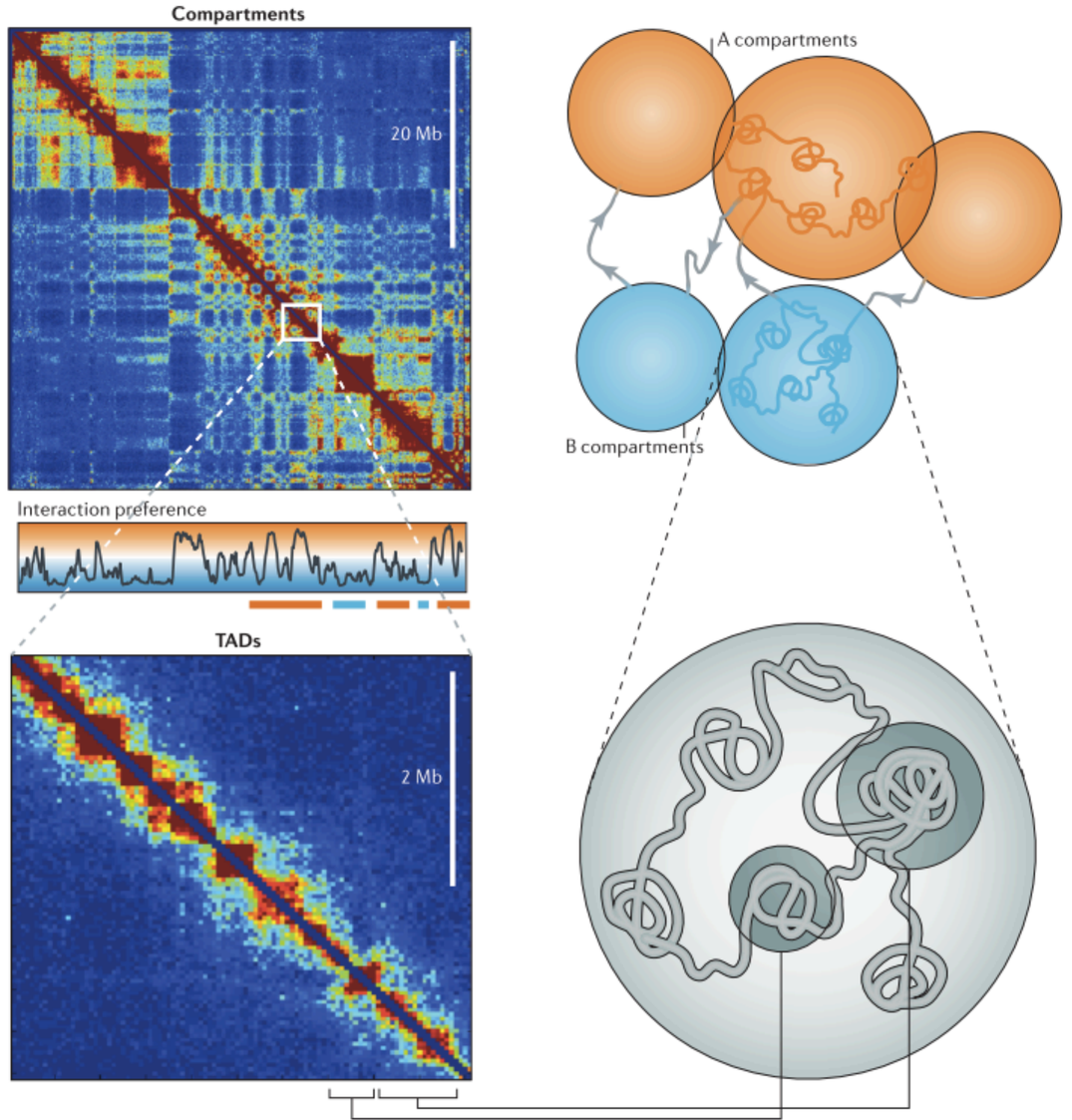
- Produce interaction matrices
  - Images
  - Text
  - Cooler

TADbit normalize



TADbit segment

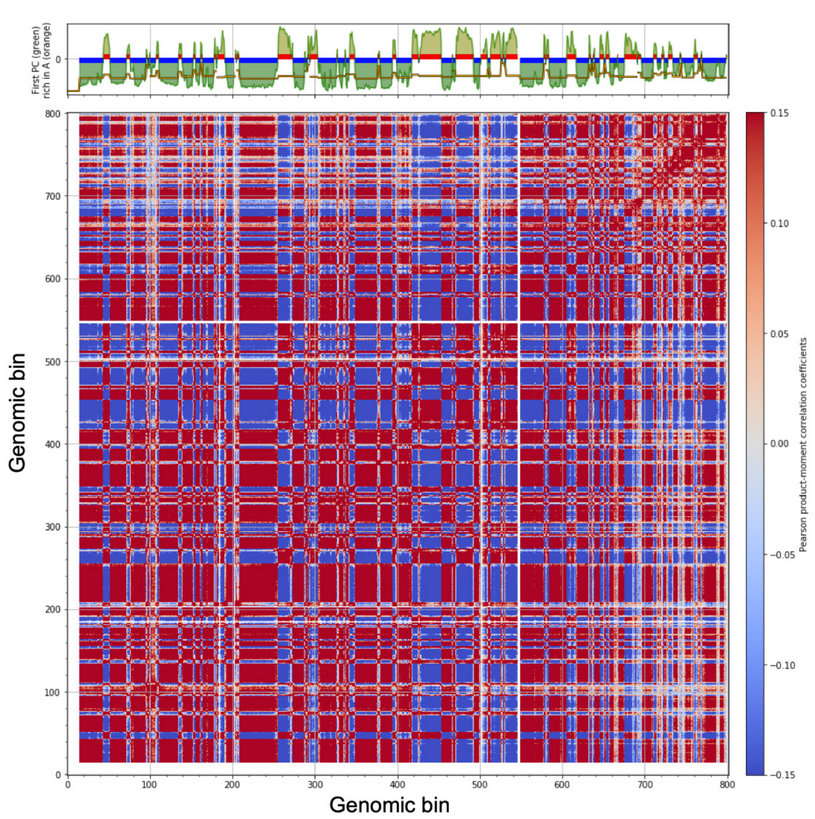
Find TADS and Compartments

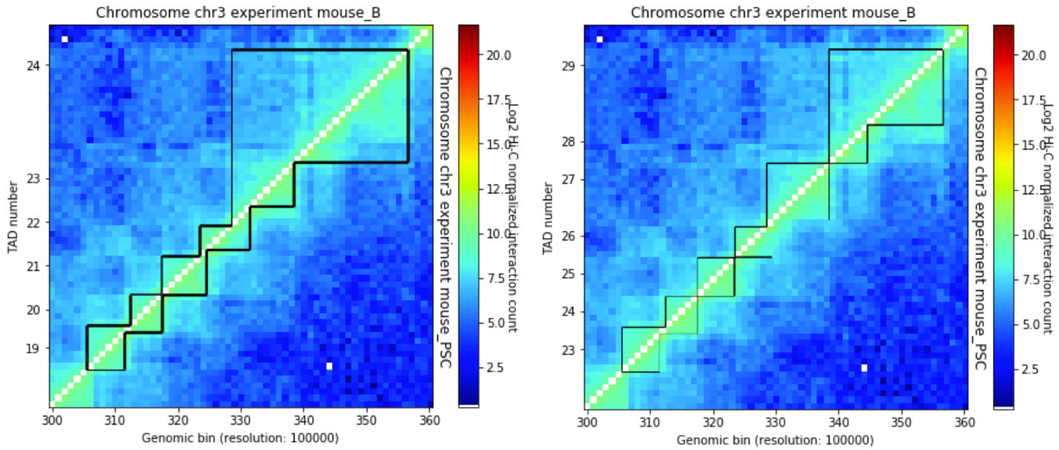


Dekker, J., Marti-Renom, M. A., & Mirny, L. a. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews. Genetics, 14(6), 390–403. http://doi.org/10.1038/nrg3454

TADbit segment

Intersect BAM file





Text file

TADbit model

Intersect BAM file

Build chromatin 3D models

Hi-C Matrix IMR90_chr21-28-30Mb