Structure determination of genomes and genomic domains by satisfaction of spatial restraints

a.k.a TADbit

Marc A. Marti-Renom
CNAG-CRG · ICREA

http://marciuslab.org
http://3DGenomes.org
http://cnag.crg.eu

cnag CRG ICREA

DISCLAIMER — 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).

DISCLAIMER — Many alternatives

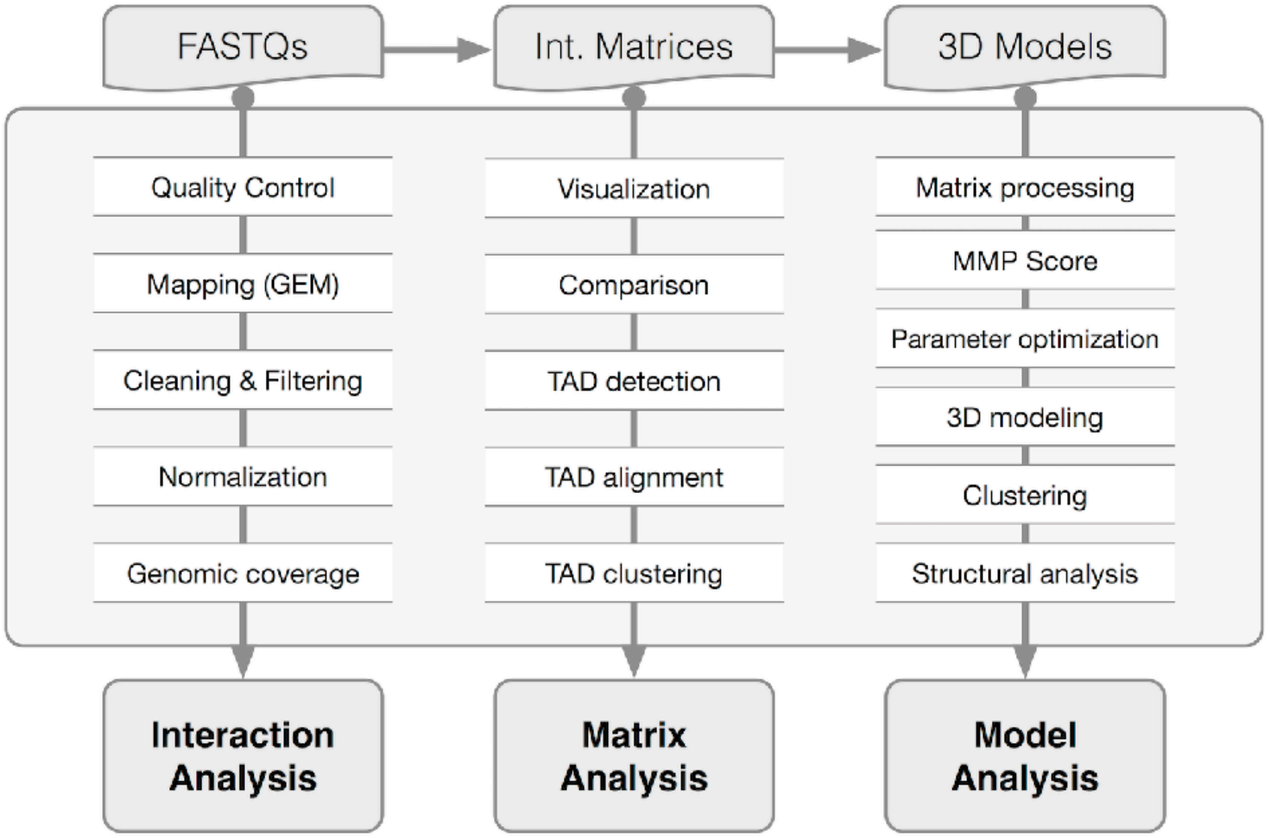
Method *available online	Representation	Scoring				Sampling	Models
		U_{3C}		U_{Biol}	U_{Phys}		
		$F_{ij} \rightarrow D_{ij}$ conversion	Functional form				
ChromSDE* [37]	Points	$D_{ij} = \begin{cases} (\frac{1}{F_{ij}})^{\alpha} \text{ if} F_{ij} > 0 \\ \infty \text{ if } F_{ij} = 0 \end{cases}$ α is optimized	$\sum_{(i,j)D_{ij}<\infty)} \frac{\langle r_{ij}^2 - D_{ij}^2 \rangle}{D_{ij}} - \lambda \sum_{(i,j)} r_{ij}^2$ 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

Serra, Baù, et al. (2017). PLOS CompBio
https://github.com/3DGenomes/tadbit
https://github.com/3DGenomes/MethodsMolBiol



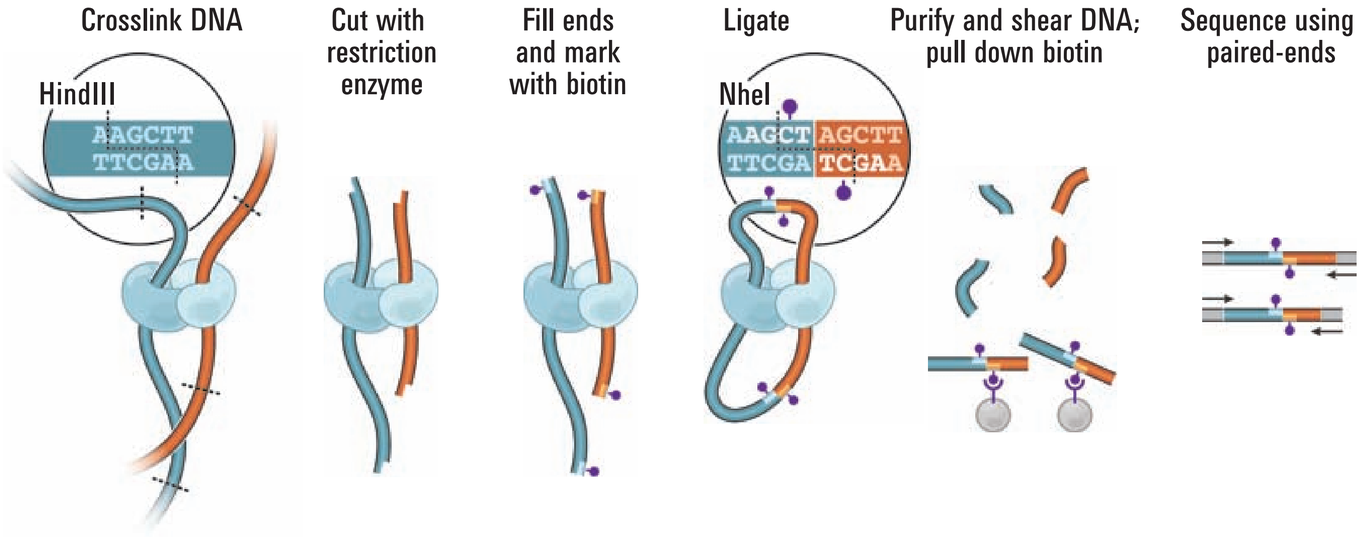
- Baù et al. Nat Struct Mol Biol (2011)
- Umbarger et al. Mol Cell (2011)
- Le Dily et al. Genes & Dev (2014)
- Belton et al. Cell Reports (2015)
- Trussart et al. Nature Communication (2017)
- Cattoni et al. Nature Communication (2017)
- Stadhouders et al. Nature Genetics (2018)
- Kojic, Cuadrado et al. Nat Struct Mol Biol (2018)
- Beekman et al. Nature Medicine (2018)
- Mas et al. Nature Genetics (2018)
- Pascual-Reguant et al. Nature Communication (2018)
- Nir, Farabella, Perez-Estrada, et al. PLOS Genetics (2018)
- Cuadrado, Giménez-Llorente et al. Cell Reports (2019)
- Vara et al. Cell Reports (2019)
- Miguel-Escalada et al. Nature Genetics (2019)
- Morf et al. Nature Biotechnology (2019)
- Di Stefano et al. Genetics (2020)
- Nguyen, Chattoraj, Castillo, et al. Nature Methods (2020)
- Soler-Vila et al. NAR (2020)
- Stik et al. Nature Genetics (2020)
- Galan et al. Nature Genetics (2020)
- Vilarassa-Blasi, Soler-Vila et al. Nature Communications (2020)

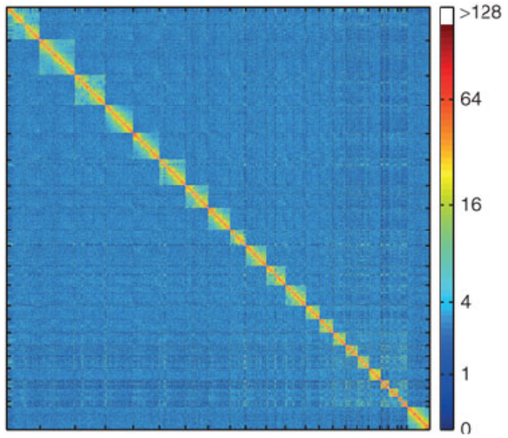
Nature Structural & Molecular Biology, 25(9), 766-777, 2018
Cell, 173(7), 1796-1809.e17, 2018
Structure, 26(6), 894-904.e2, 2018
Genome Research, 29(1), 29-39, 2019
Genome Research, 29(1), gr.238527.118, 2019
Cell Systems 9, 1–13.e1–e6, 2019
Nature Communications, 10(1), 5355, 2019
BMC Biology, 17(1), 55, 2019
Molecular Cell, 2019
Cell Systems, 9(5), 446-458.e6, 2019


Got FASTQ?

Hi-C experiment

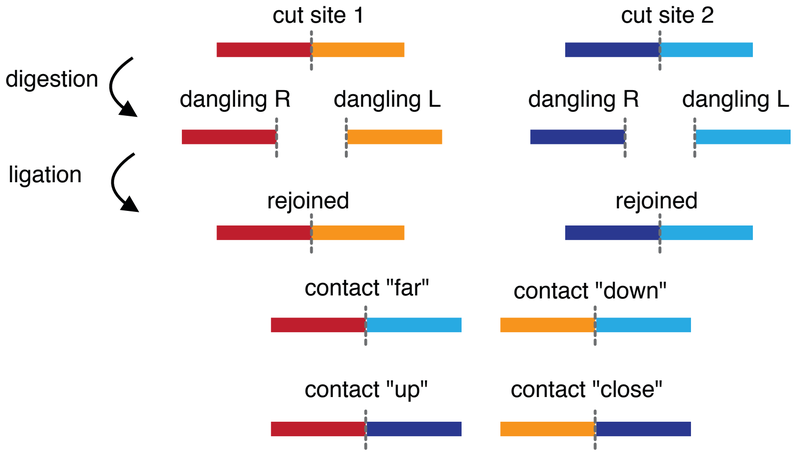
Lieberman-Aiden, E., et al. (2009). Science, 326(5950), 289–293.

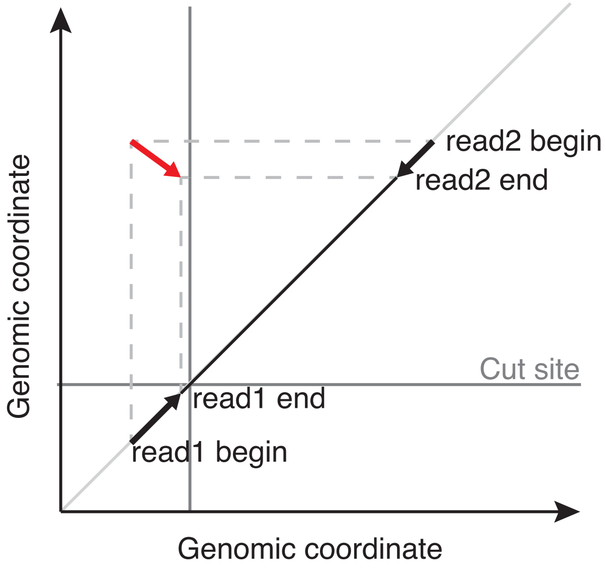


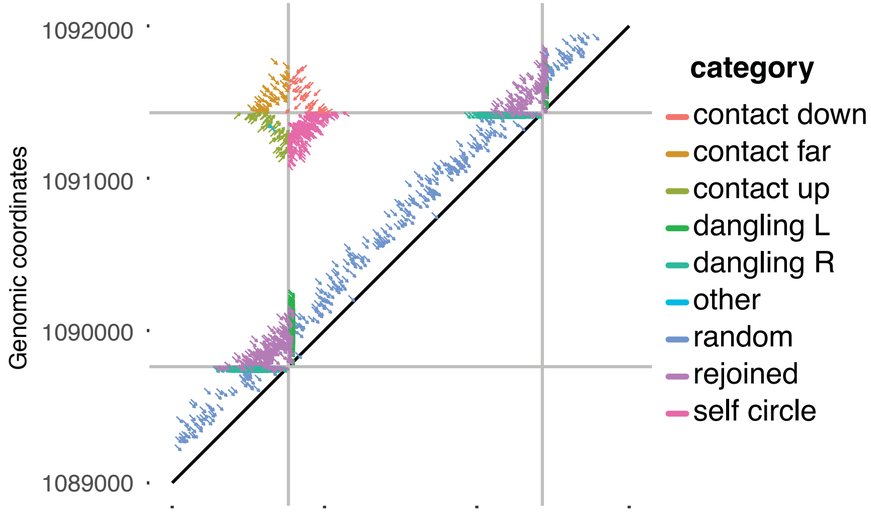


Mapping & Filtering

Imakaev, M. V et al. (2012). Nature Methods, 9(10), 999–1003.

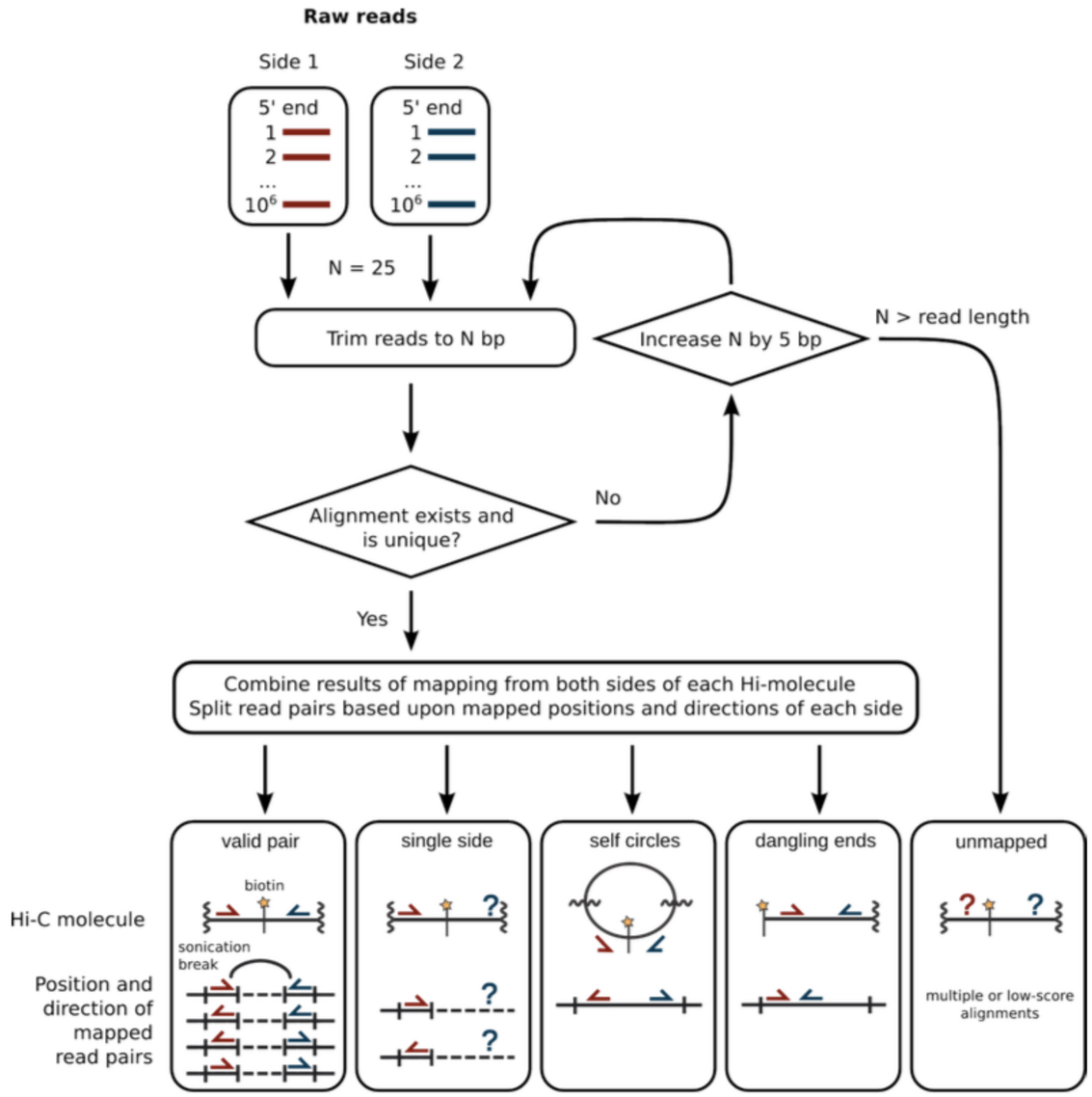






Mapping & Filtering

Imakaev, M. V et al. (2012). Nature Methods, 9(10), 999–1003.

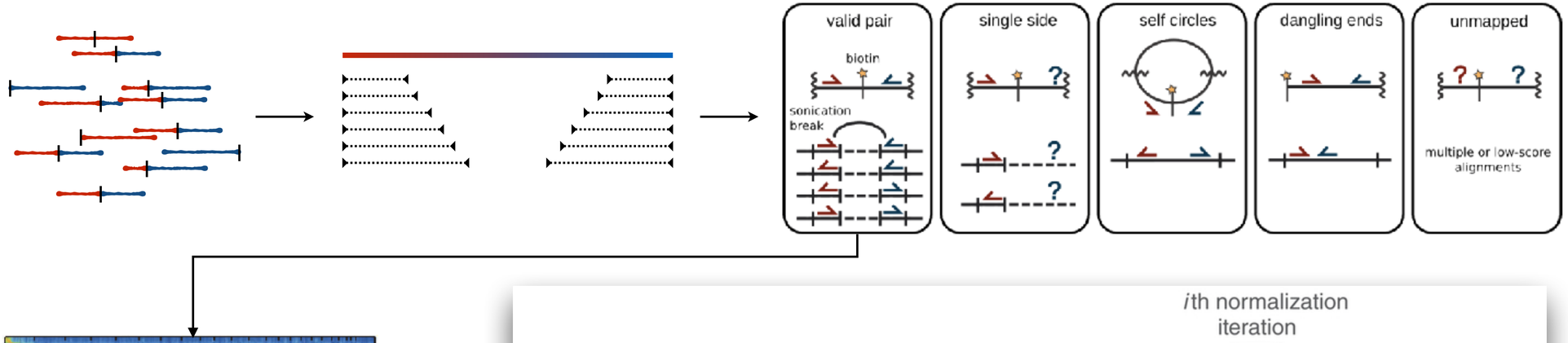


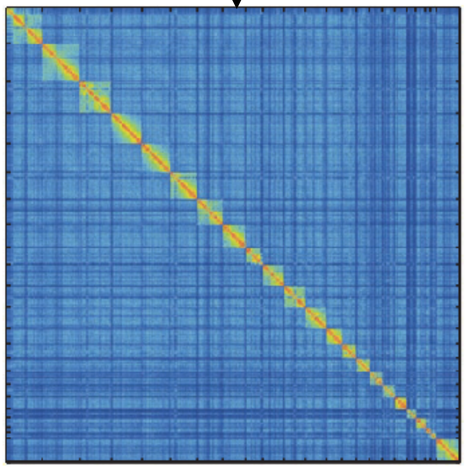
How much you normally map?

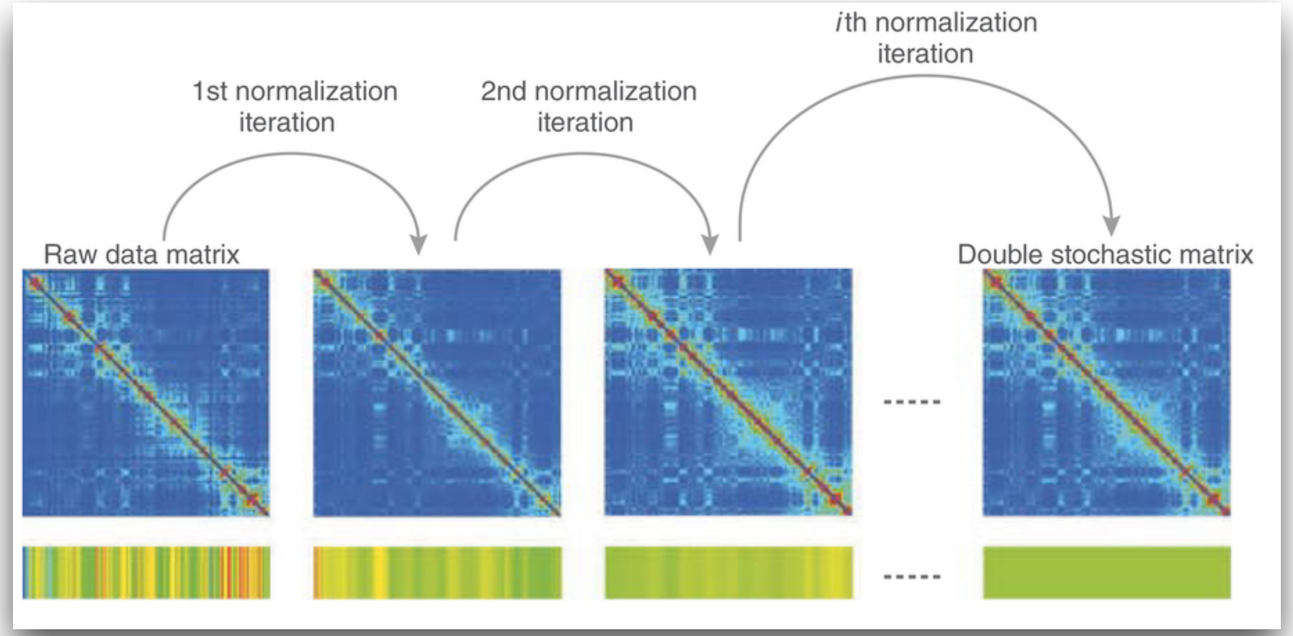
- 80-90% each end => 60-80% intersection
- ~1% multiple contacts
- Many of intersecting pairs will be lost in filtering…
- Final 40-60% of valid pairs
- One measure of quality is the CIS/TRANS ration (70-80% good)


Got mapped reads?

Interaction matrices

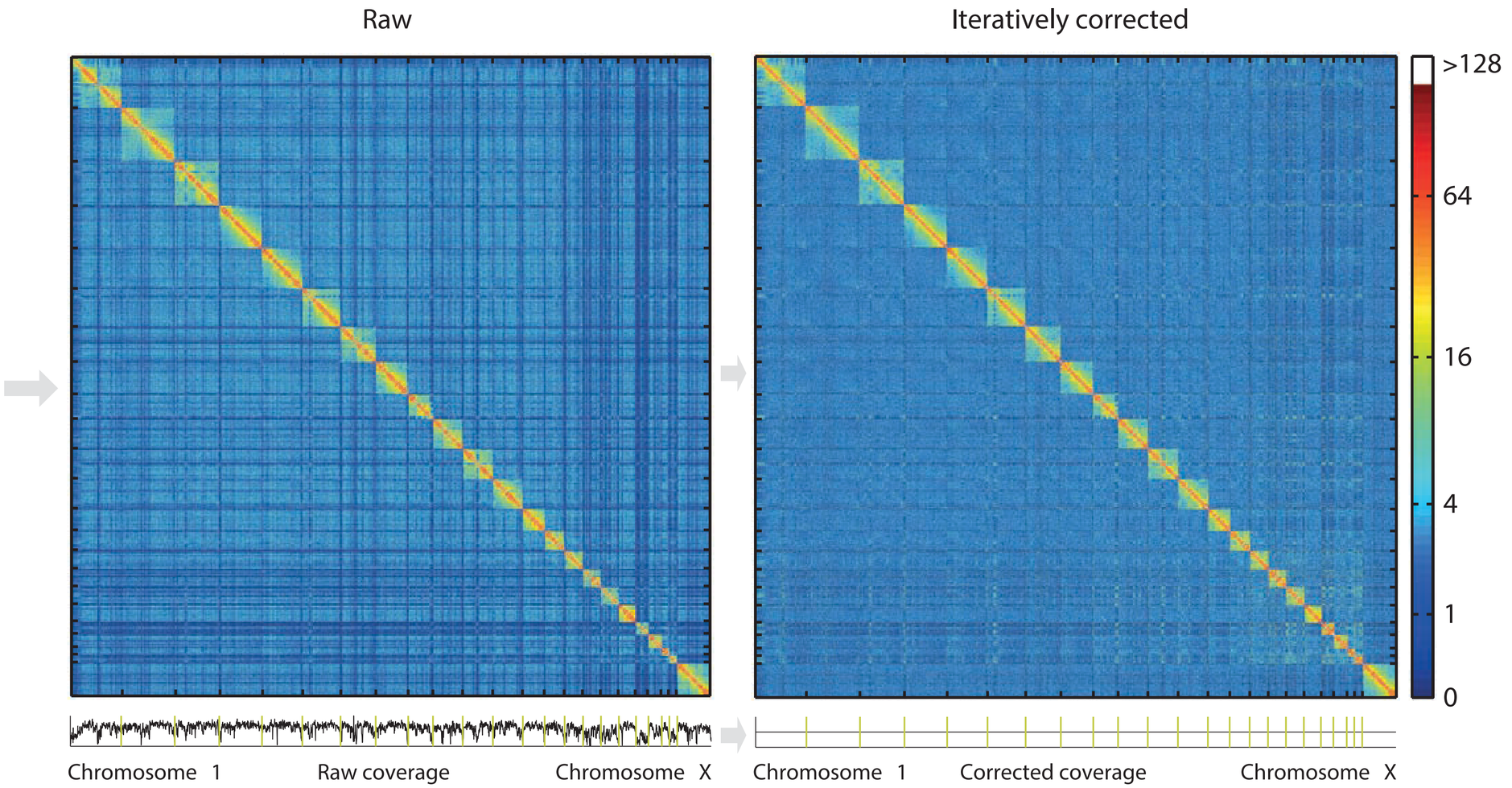






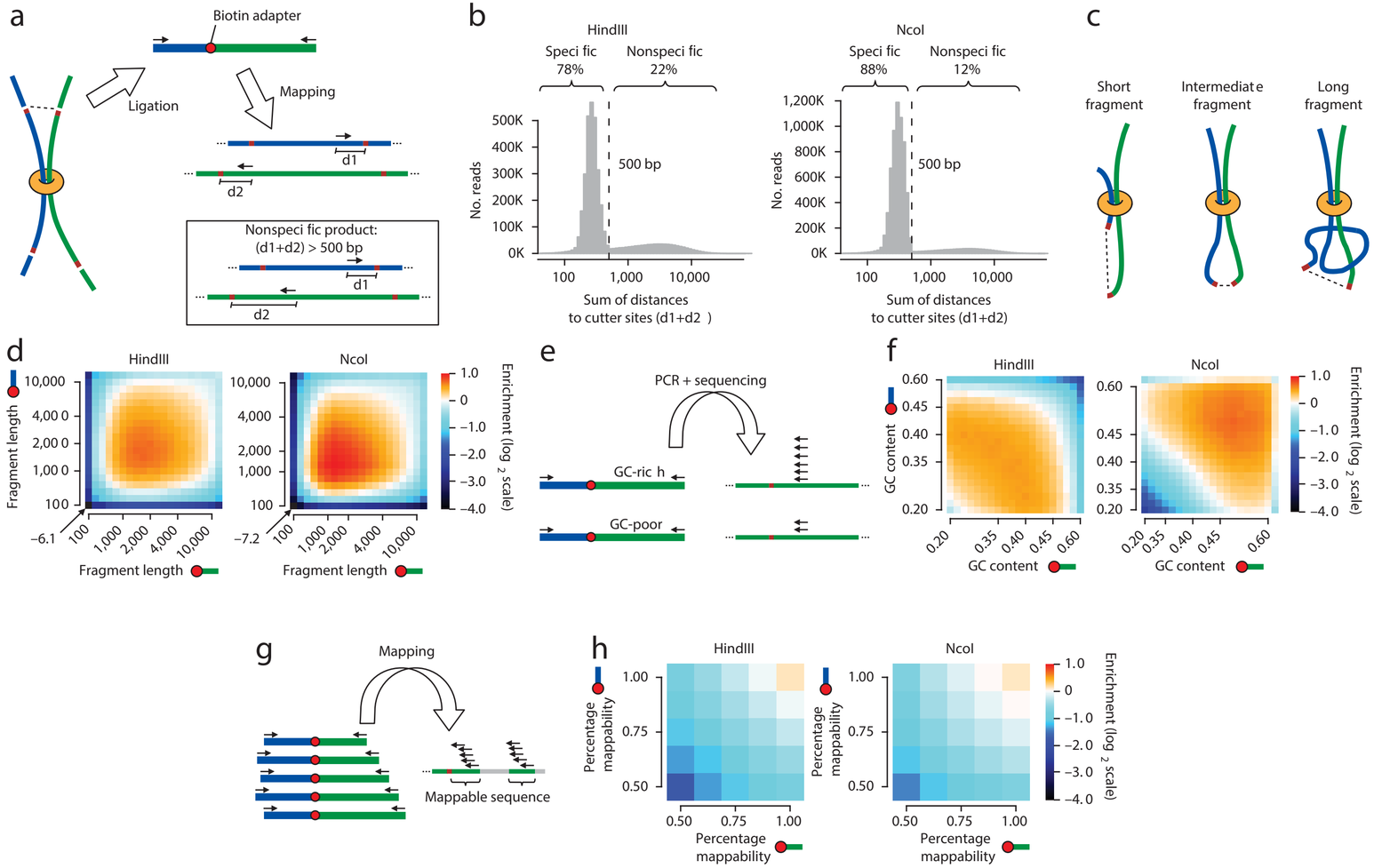
Zooming in on genome organization.
Zhou, X. J., & Alber, F. Nature Methods (2012)

Normalizing HiC data



Normalizing HiC data (a la Tanay)

Yaffe, E., & Tanay, A. (2011). Nature Genetics, 43(11), 1059–1065

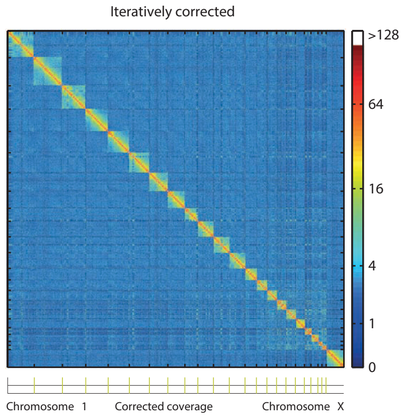


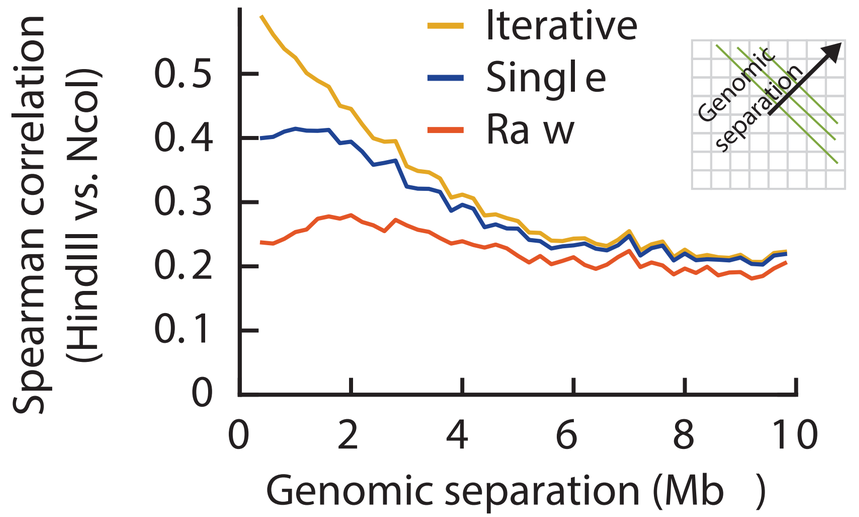
Normalizing HiC data (a la Mirny)

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., et al. (2012). Nature Methods, 9(10), 999–1003.

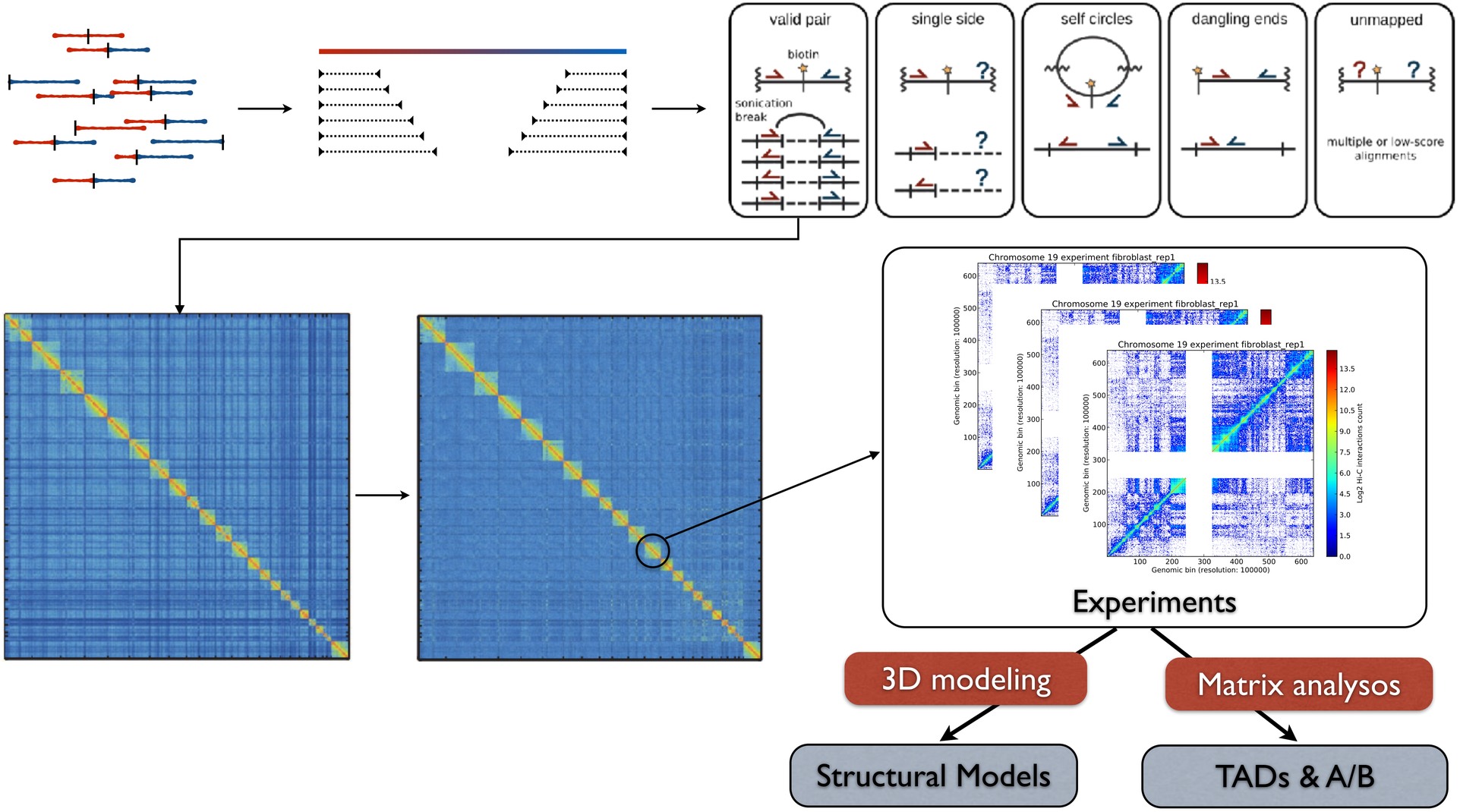
$$O_{ij} = B_i B_j T_{ij}$$

$$\Sigma_{i=1,|i-j|>1}^{N} T_{ij} = 1$$

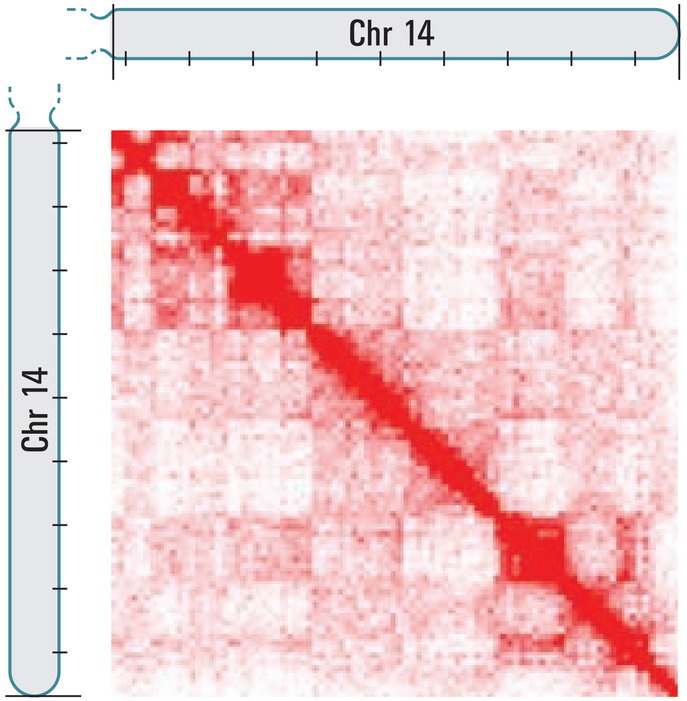


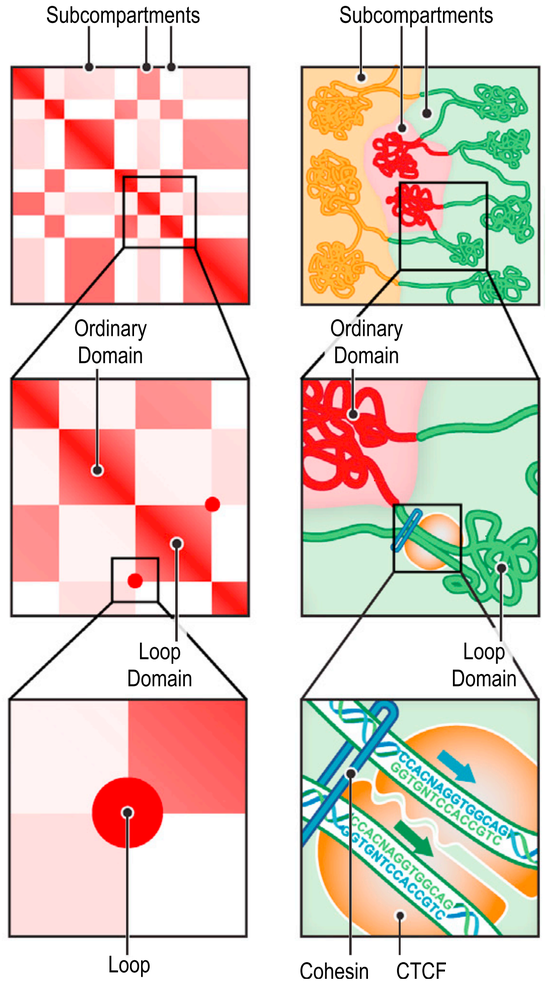


Interaction matrices



Hierarchical genome organisation

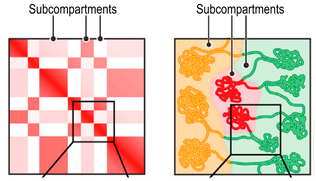


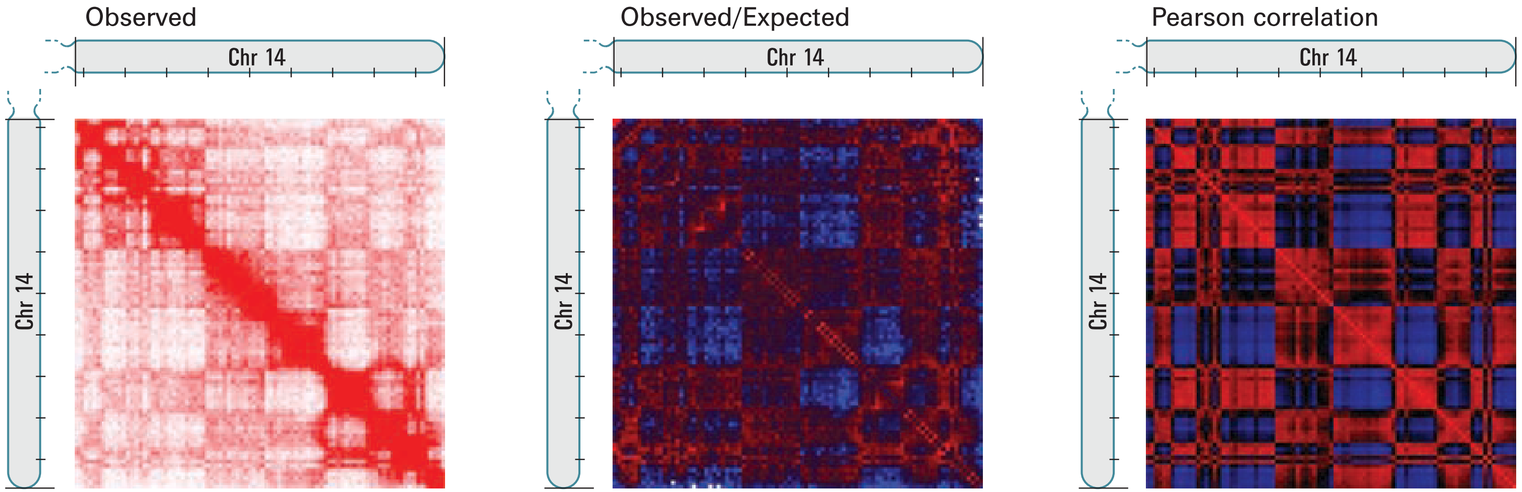


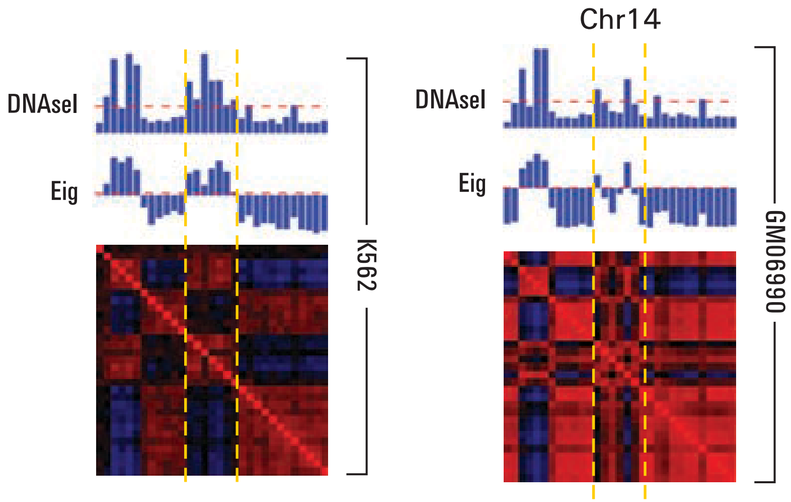
Lieberman-Aiden, E., et al. (2009). Science, 326(5950), 289–293.
Rao, S. S. P., et al. (2014). Cell, 1–29.

A/B Compartment

Human chromosome 14



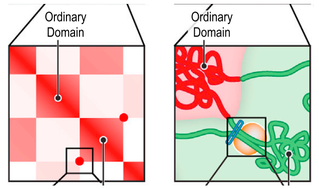


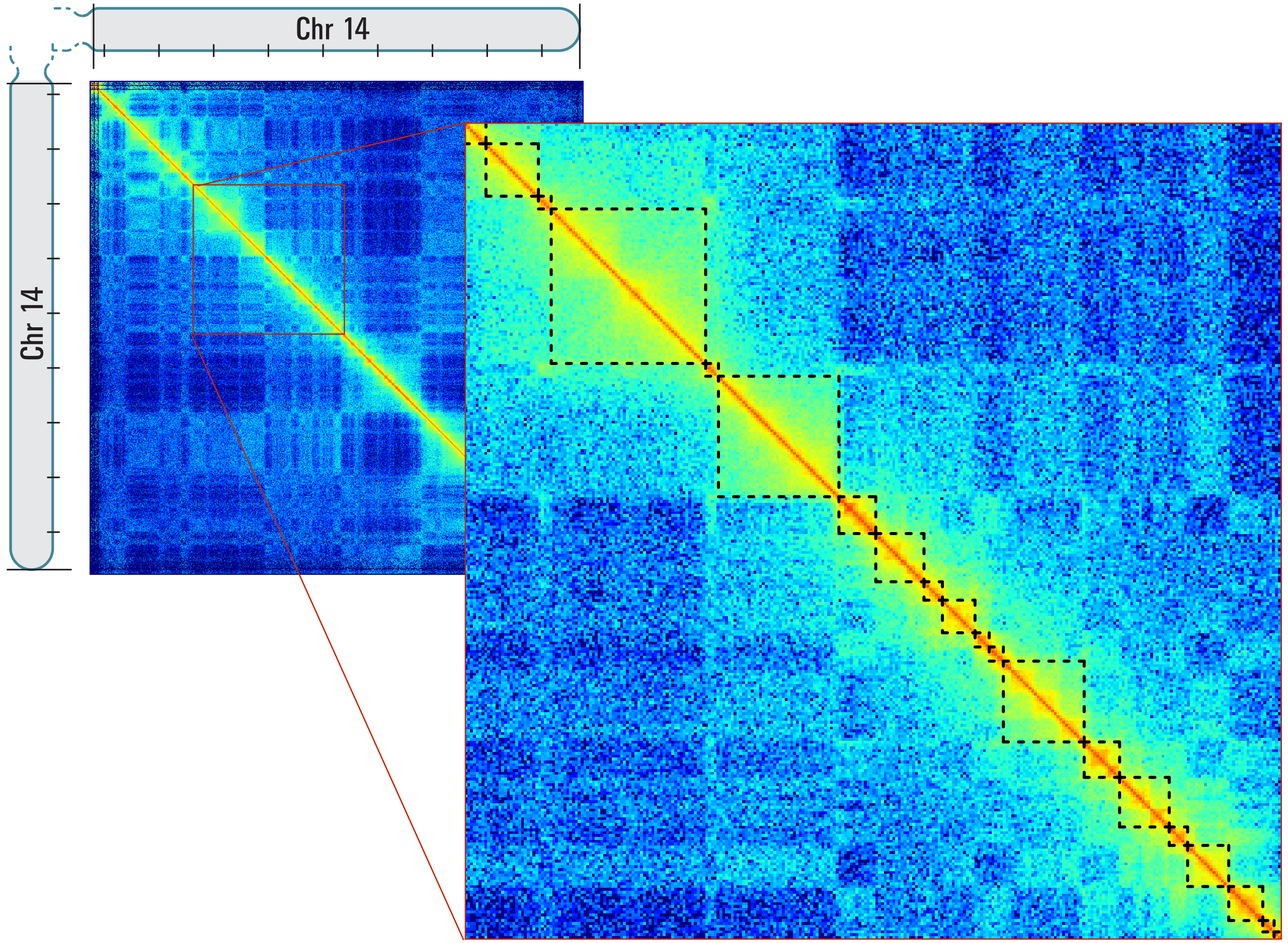


Lieberman-Aiden et al. Science. 2009

TADs

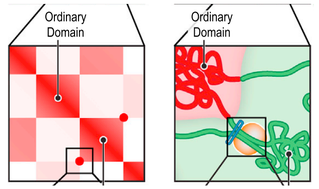
Chromosome 14

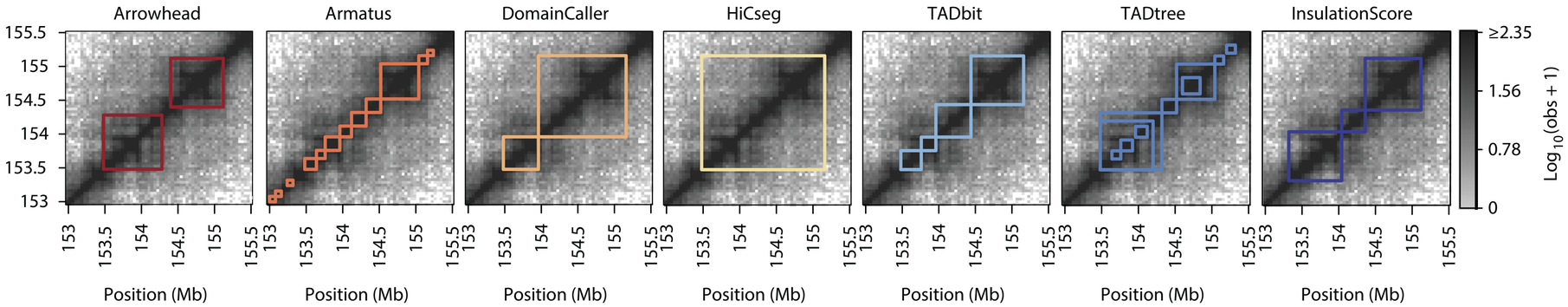


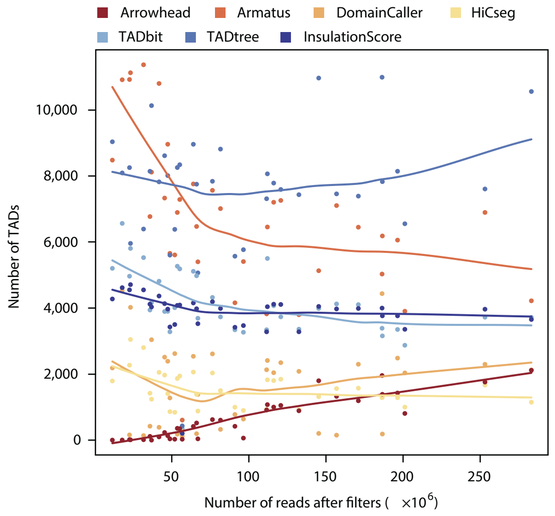


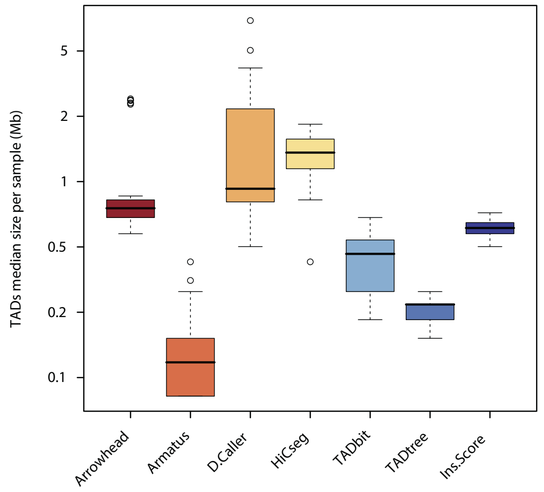
TADs

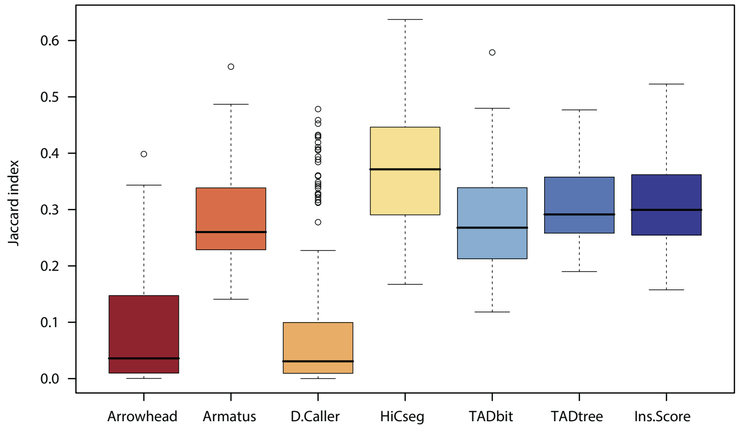
How well we do…







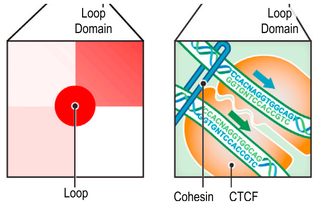


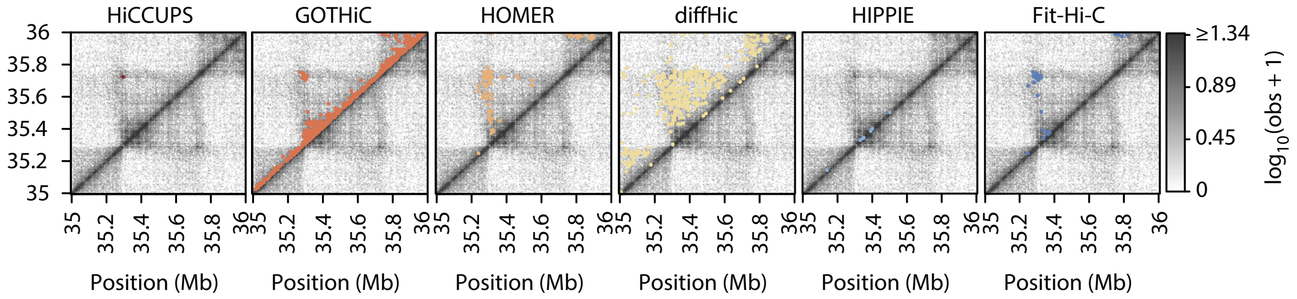


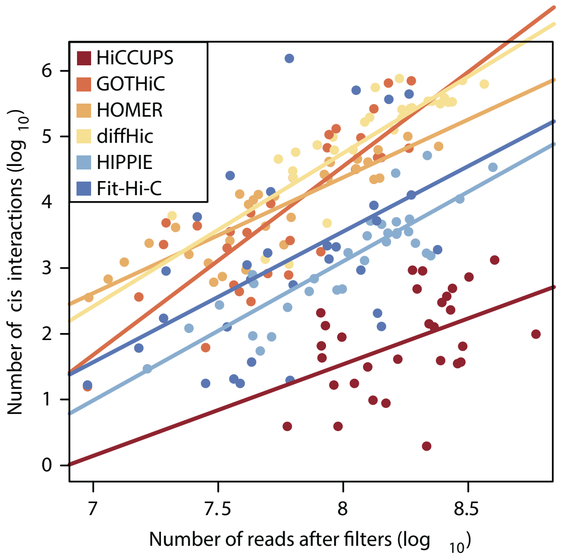
Forcato et al. Nat Methods. 2017 Jul; 14(7) 679-685

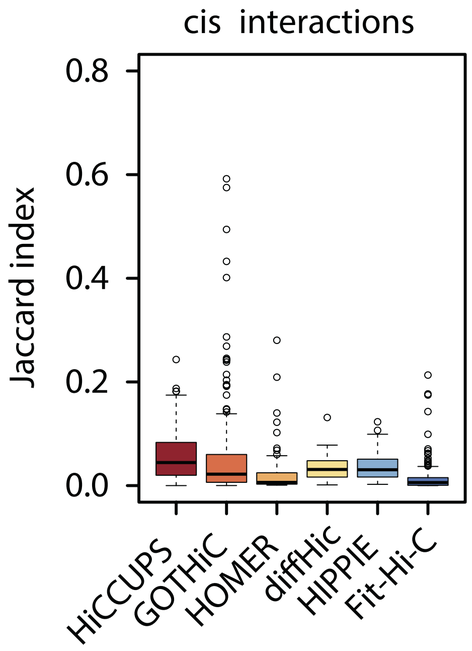
Loops

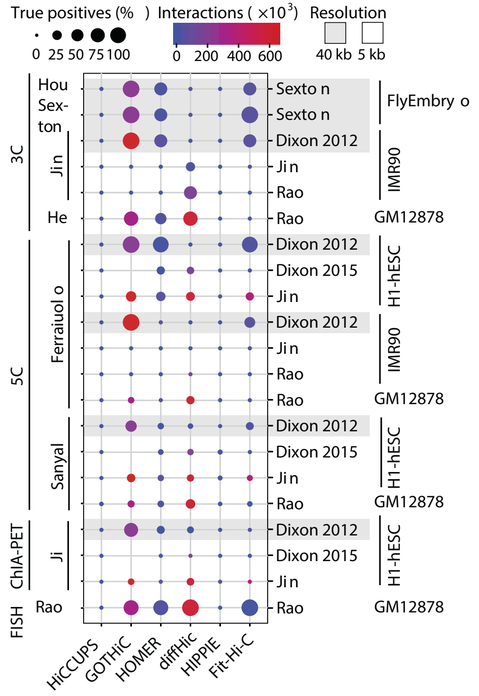
How well we do…





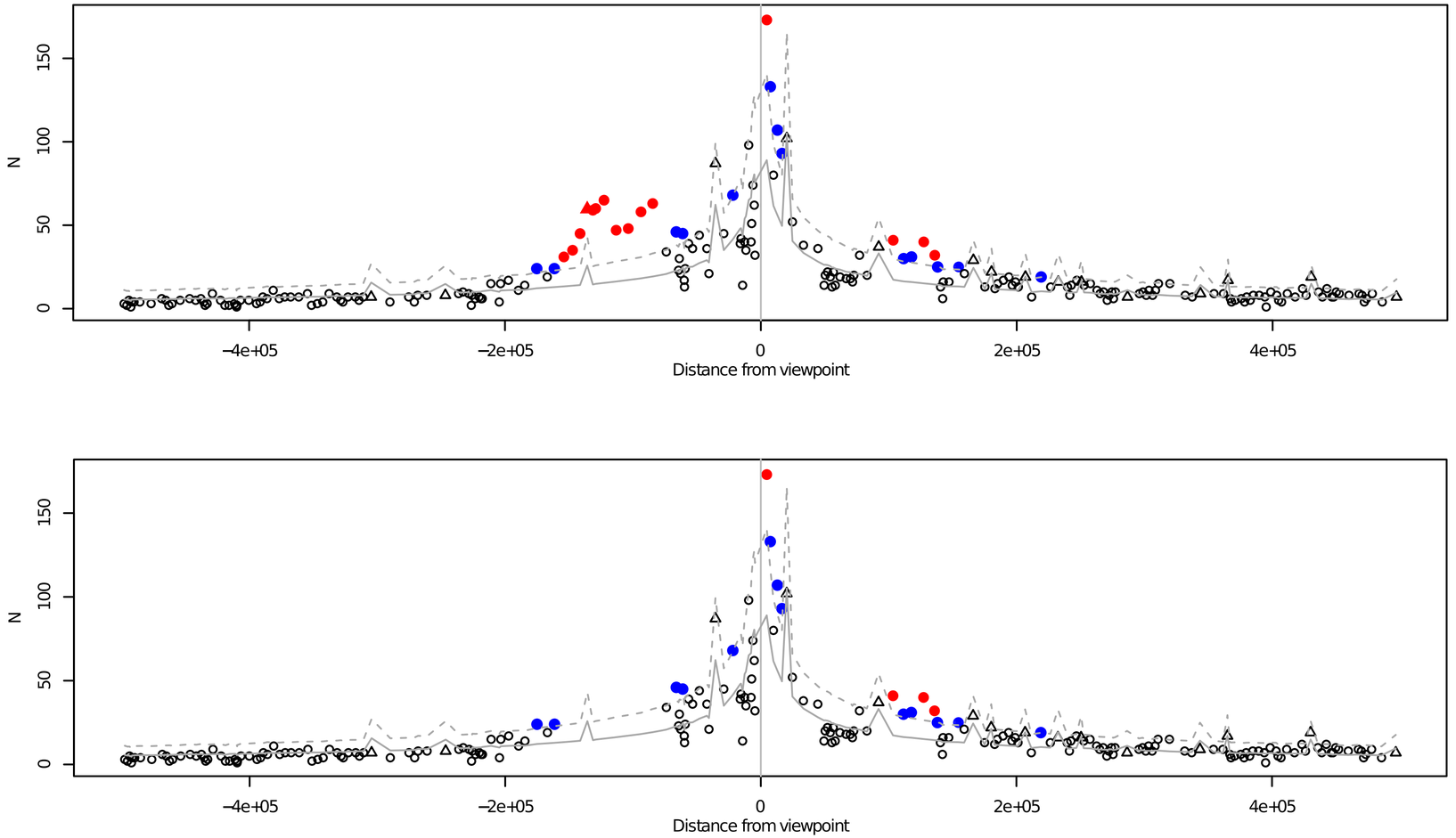




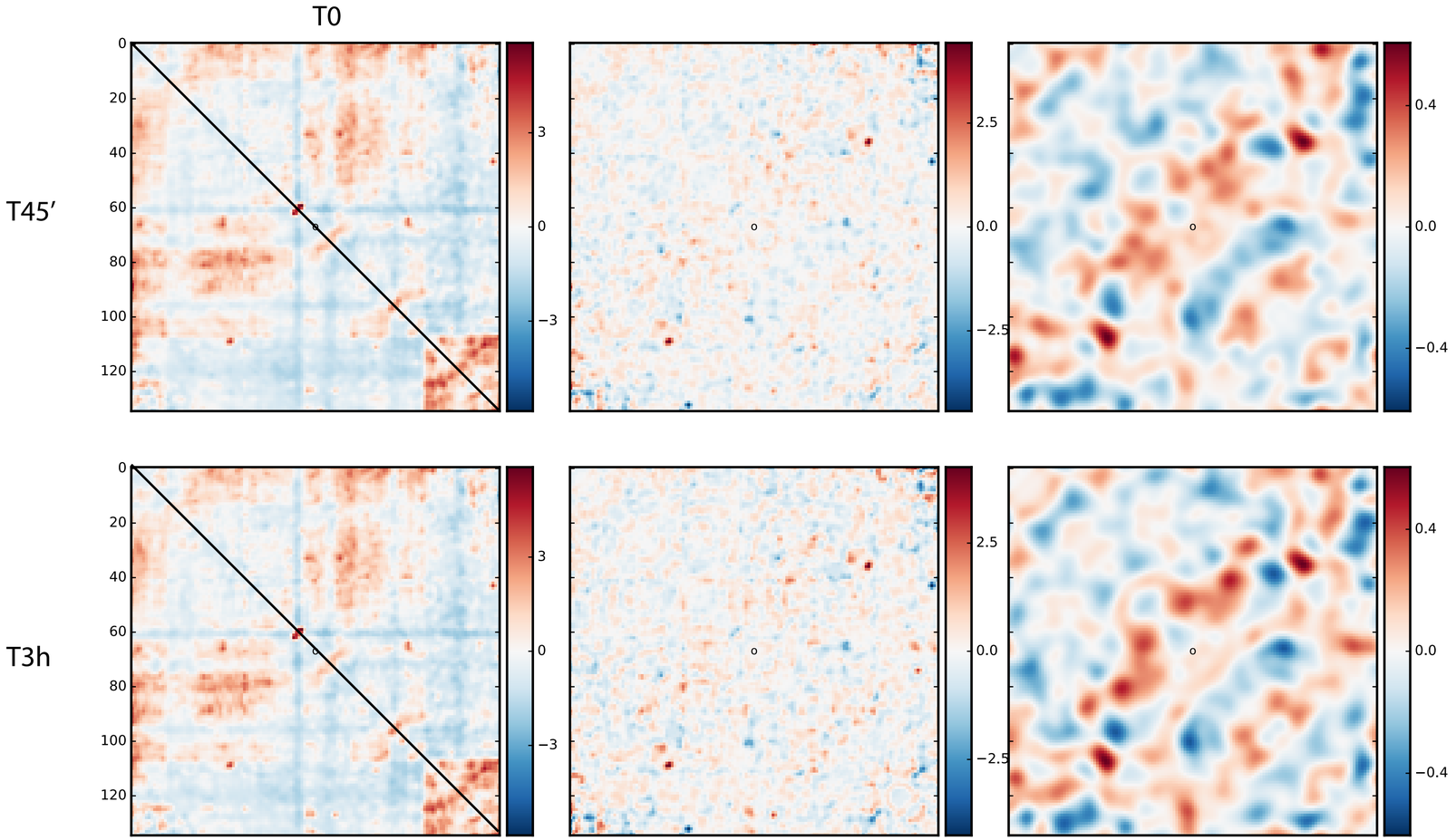


Forcato et al. Nat Methods. 2017 Jul; 14(7) 679-685

Comparing HiC data



Z-score differences (DekkerLab)



Comparing HiC data (GOTHIC)

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., et al. (2015). *Nature Genetics*, 1–12.

Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C

Borbala Mifsud[1,2,10], Filipe Tavares-Cadete[1,9], Alice N Young[3,10], Robert Sugar[1], Stefan Schoenfelder[3], Lauren Ferreira[3], Steven W Wingett[4], Simon Andrews[4], William Grey[5], Philip A Ewels[3], Bram Herman[6], Scott Happe[6], Andy Higgs[6], Emily LeProust[6,9], George A Follows[7], Peter Fraser[3], Nicholas M Luscombe[1,2,8] & Cameron S Osborne[3,5]

Transcriptional control in large genomes often requires looping interactions between distal DNA elements, such as enhancers and target promoters. Current chromosome conformation capture techniques do not offer sufficiently high resolution to interrogate these regulatory interactions on a genomic scale. Here we use Capture Hi-C (CHi-C), an adapted genome conformation assay, to examine the long-range interactions of almost 22,000 promoters in 2 human blood cell types. We identify over 1.6 million shared and cell type–restricted interactions spanning hundreds of kilobases between promoters and distal loci. Transcriptionally active genes contact enhancer-like elements, whereas transcriptionally inactive genes interact with previously uncharacterized elements marked by repressive features that may act as long-range silencers. Finally, we show that interacting loci are enriched for disease-associated SNPs, suggesting how distal mutations may disrupt the regulation of relevant genes. This study provides new insights and accessible tools to dissect the regulatory interactions that underlie normal and aberrant gene regulation.

Genome organization influences transcriptional regulation by facilitating interactions between gene promoters and distal regulatory elements. Many contacts have been identified using chromosome conformation capture methodologies[1–3]. For example, the ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) method has been used to map long-range interactions extending over hundreds of kilobases; however, these studies have only interrogated the subset of interactions involving highly transcriptionally active genes, whereas long-range interactions for weakly expressed and transcriptionally inactive genes remain unknown. Although the 5C (chromatin conformation capture carbon copy) method is not restricted by the nature of interactions, thus far, it has only been applied to a few small genomic regions. The Hi-C method simultaneously captures all genomic interactions, which provides a population-average snapshot of the genome conformation within a single experiment[4]; yet, owing to the enormous complexity of Hi-C libraries, it is costly to sequence to sufficient depth to provide enough spatial resolution to interrogate specific contacts between gene promoters and distal regulatory elements[5,6]. To circumvent these issues, we have used solution hybridization selection, originally developed for exon sequencing[7]—and recently used to capture the interactions of a few hundred promoters from 3C libraries[8]—to enrich Hi-C libraries for genome-wide, long-range contacts of both active and inactive promoters.

RESULTS

A genome-wide, long-range interaction capture assay

We prepared three HindIII-digested Hi-C libraries from GM12878 cells, a human Epstein-Barr virus (EBV)-transformed lymphoblastoid cell line that has been comprehensively assayed in the Encyclopedia of DNA Elements (ENCODE) Project, and two libraries from *ex vivo* $CD34^+$ hematopoietic progenitor cells. One Hi-C library from each cell type was sequenced to examine the di-tag (paired-end read) interaction distribution and depth of read coverage (**Supplementary Table 1**). As anticipated, we observed a higher density of di-tag interaction reads between restriction fragments in *cis* as compared with fragments in *trans*, with the highest density occurring between fragments separated by less than 20 kb (**Supplementary Fig. 1a**,**b**). We also observed demarcation of the genome into distinct contiguous, highly intraconnected topologically associated domains (TADs)[5] (**Supplementary Fig. 1c** and **Supplementary Table 2**). The distribution of read coverage was typical for a Hi-C experiment. In our initial comparison, we downsampled all data sets to 45 million unique sequencing reads. Each restriction fragment was represented by an average of 143 and 139 reads in the GM12878 and $CD34^+$ libraries, respectively (**Supplementary Fig. 1d**). We processed the reads using binomial statistics to identify ligation fragments that were significantly enriched ($q < 0.05$). This approach recognizes ligation products between

[1]The Francis Crick Institute, London, UK. [2]UCL Genetics Institute, University College London, London, UK. [3]Nuclear Dynamics Programme, Babraham Institute, Cambridge, UK. [4]Bioinformatics Group, Babraham Institute, Cambridge, UK. [5]Department of Medical and Molecular Genetics, King's College London School of Medicine, London, UK. [6]Diagnostics and Genomics Division, Agilent Technologies, Santa Clara, California, USA. [7]Department of Haematology, Cambridge University Hospitals National Health Service (NHS) Foundation Trust, Cambridge, UK. [8]Okinawa Institute of Science and Technology, Okinawa, Japan. [9]Present addresses: Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan (F.T.-C.) and Twist Bioscience, San Francisco, California, USA (E.L.). [10]These authors contributed equally to this work. Correspondence should be addressed to C.S.O. (cameron.osborne@kcl.ac.uk) or N.M.L. (nicholas.luscombe@ucl.ac.uk).

Received 5 December 2014; accepted 2 April 2015; published online 4 May 2015; doi:10.1038/ng.3286

Comparing HiC data (CHICAGO)

Cairns, J., Freire-Pritchett, P., Wingett, S. W., Várnai, C., Dimond, A., Plagnol, V., et al. (2016). *Genome Biology*, 1–17.

Cairns *et al. Genome Biology* (2016) 17:127
DOI 10.1186/s13059-016-0992-2

Genome Biology

METHOD Open Access

CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data

Jonathan Cairns[1†], Paula Freire-Pritchett[1†], Steven W. Wingett[1,2], Csilla Várnai[1], Andrew Dimond[1], Vincent Plagnol[3], Daniel Zerbino[4], Stefan Schoenfelder[1], Biola-Maria Javierre[1], Cameron Osborne[5], Peter Fraser[1] and Mikhail Spivakov[1*]

Abstract

Capture Hi-C (CHi-C) is a method for profiling chromosomal interactions involving targeted regions of interest, such as gene promoters, globally and at high resolution. Signal detection in CHi-C data involves a number of statistical challenges that are not observed when using other Hi-C-like techniques. We present a background model and algorithms for normalisation and multiple testing that are specifically adapted to CHi-C experiments. We implement these procedures in CHiCAGO (http://regulatorygenomicsgroup.org/chicago), an open-source package for robust interaction detection in CHi-C. We validate CHiCAGO by showing that promoter-interacting regions detected with this method are enriched for regulatory features and disease-associated SNPs.

Keywords: Gene regulation, Nuclear organisation, Promoter-enhancer interactions, Capture Hi-C, Convolution background model, *P* value weighting

Background

Chromosome conformation capture (3C) technology has revolutionised the analysis of nuclear organisation, leading to important insights into gene regulation [1]. While the original 3C protocol tested interactions between a single pair of candidate regions ("one vs one"), subsequent efforts focused on increasing the throughput of this technology (4C, "one vs all"; 5C, "many vs many"), culminating in the development of Hi-C, a method that interrogated the whole nuclear interactome ("all vs all") [1, 2]. The extremely large number of possible pairwise interactions in Hi-C samples, however, imposes limitations on the realistically achievable sequencing depth at individual interactions, leading to reduced sensitivity. The recently developed Capture Hi-C (CHi-C) technology uses sequence capture to enrich Hi-C material for multiple genomic regions of interest (hereafter referred to as "baits"), making it possible to profile the global interaction profiles of many thousands of regions globally ("many vs all") and at a high resolution (Fig. 1) [3–7].

CHi-C data possess statistical properties that set them apart from other 3C/4C/Hi-C-like methods. First, in contrast to traditional Hi-C or 5C, baits in CHi-C comprise a subset of restriction fragments, while any fragment in the genome can be detected on the "other end" of an interaction. This asymmetry of CHi-C interaction matrices is not accounted for by the normalisation procedures developed for traditional Hi-C and 5C [8–10]. Secondly, CHi-C baits, but not other ends, have a further source of bias associated with uneven capture efficiency. In addition, the need for detecting interactions globally and at a single-fragment resolution creates specific multiple testing challenges that are less pronounced with binned Hi-C data or the more focused 4C and 5C assays, which involve fewer interaction tests. Finally, CHi-C designs such as Promoter CHi-C and HiCap [3–5, 11] involve large numbers (many thousands) of spatially dispersed baits. This presents the opportunity to increase the robustness of signal detection by sharing information across baits. Such sharing is impossible in the analysis of 4C data that focuses on only a single bait and is of limited use in 4C-seq containing a small number of baits [12–14].

These distinct features of CHi-C data have prompted us to develop a bespoke statistical model and a

* Correspondence: mikhail.spivakov@babraham.ac.uk
†Equal contributors
[1]Nuclear Dynamics Programme, Babraham Institute, Cambridge, UK
Full list of author information is available at the end of the article

cnag CRG

Comparing HiC data (diffHiC)

Lun, A. T. L., & Smyth, G. K. (2015). *BMC Bioinformatics*, 1–11.

Lun and Smyth *BMC Bioinformatics* (2015) 16:258
DOI 10.1186/s12859-015-0683-0

BMC Bioinformatics

SOFTWARE Open Access

diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data

Aaron T.L. Lun[1,2] and Gordon K. Smyth[1,3*]

Abstract

Background: Chromatin conformation capture with high-throughput sequencing (Hi-C) is a technique that measures the *in vivo* intensity of interactions between all pairs of loci in the genome. Most conventional analyses of Hi-C data focus on the detection of statistically significant interactions. However, an alternative strategy involves identifying significant changes in the interaction intensity (i.e., differential interactions) between two or more biological conditions. This is more statistically rigorous and may provide more biologically relevant results.

Results: Here, we present the diffHic software package for the detection of differential interactions from Hi-C data. diffHic provides methods for read pair alignment and processing, counting into bin pairs, filtering out low-abundance events and normalization of trended or CNV-driven biases. It uses the statistical framework of the edgeR package to model biological variability and to test for significant differences between conditions. Several options for the visualization of results are also included. The use of diffHic is demonstrated with real Hi-C data sets. Performance against existing methods is also evaluated with simulated data.

Conclusions: On real data, diffHic is able to successfully detect interactions with significant differences in intensity between biological conditions. It also compares favourably to existing software tools on simulated data sets. These results suggest that diffHic is a viable approach for differential analyses of Hi-C data.

Keywords: Hi-C, Genomic interaction, Differential analysis

Background

Chromatin conformation capture with high-throughput sequencing (Hi-C) is a technique that is widely used to study global chromatin organization *in vivo* [1]. Briefly, samples of nuclear DNA are cross-linked and digested with a restriction enzyme to release chromatin complexes into solution (Fig. 1). Each complex may contain multiple restriction fragments, corresponding to an interaction between the associated genomic loci. After some processing, proximity ligation is performed between the ends of the restriction fragments. This favours ligation between restriction fragments in the same complex. The ligated DNA is sheared and purified for high-throughput paired-end sequencing. Each sequencing fragment represents a ligation product, such that each read in the pair originates from a different genomic locus. The intensity of an interaction between a pair of genomic loci can be quantified as the number of read pairs with one read mapped to each locus. The output from the Hi-C procedure spans the genome-by-genome "interaction space" whereby all pairwise interactions between loci can potentially be detected. As such, careful analysis is required to draw meaningful biological conclusions from this type of data.

Most analyses of Hi-C data have focused on identifying "significant" interactions from a single sample [2, 3]. This is challenging because non-specific ligation and apparent interactions can arise from a variety of uninteresting technical causes and rigorous analysis requires a precise quantitative understanding of these artifacts. Identifying biologically interesting interactions from a single sample requires elaborate modeling of the background signal in Hi-C experiments in order to correct for systematic biases due to GC content, mappability and fragment length [3]. Such modeling inevitably involves assumptions and approximations. Furthermore, the interaction space

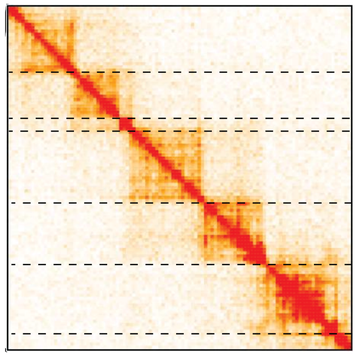
*Correspondence: smyth@wehi.edu.au
[1]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Melbourne, Australia
[3]Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Melbourne, Australia
Full list of author information is available at the end of the article

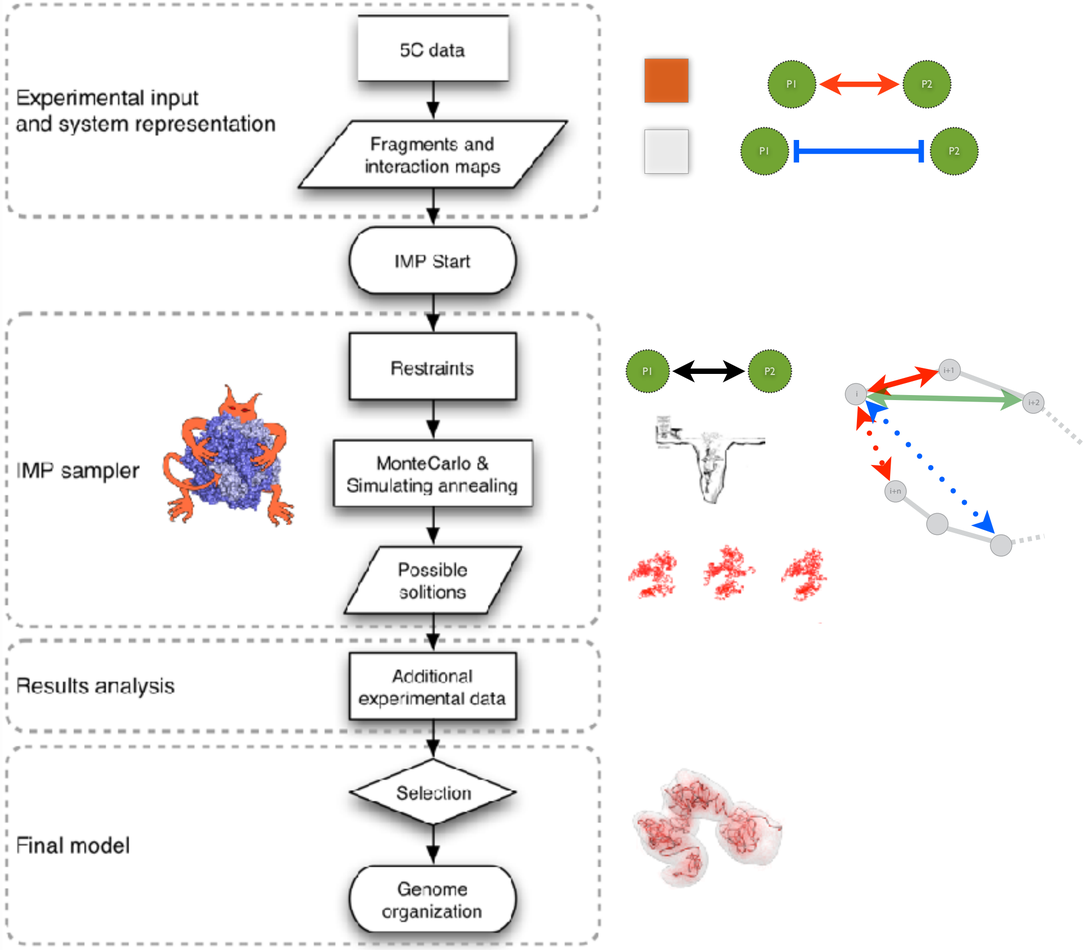
cnag CRG


Got normalized
Hi-C maps?

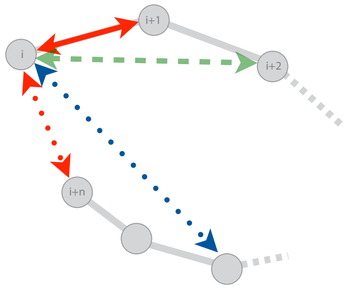
TADbit

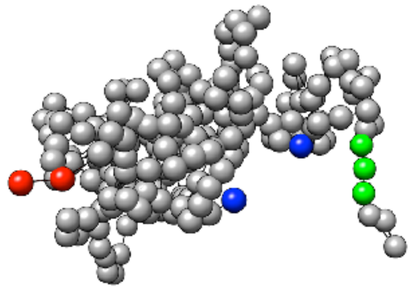
http://3DGenomes.org
http://www.integrativemodeling.org





Model representation and scoring

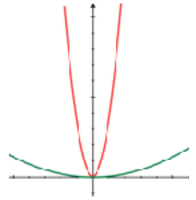




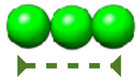
$d = d_0$

Harmonic

$$H_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2$$

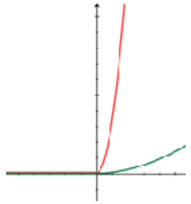


$d < d_0$



Harmonic Upper Bound

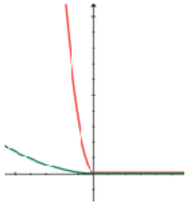
$$\begin{cases} if\ \ d_{i,j} \geq d^0_{i,j}; & ubH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} < d^0_{i,j}; & ubH_{i,j} = 0 \end{cases}$$



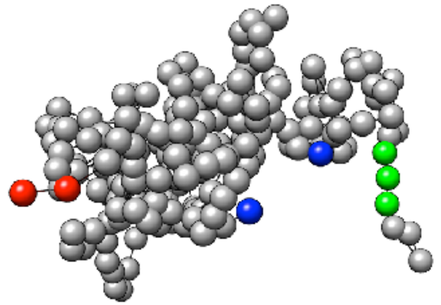
$d > d_0$

Harmonic Lower Bound

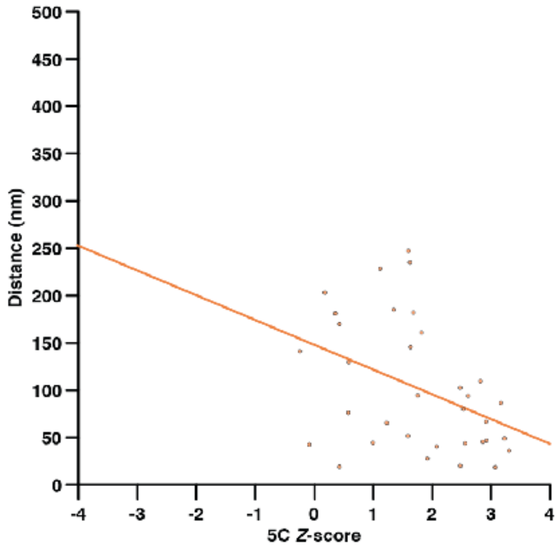
$$\begin{cases} if\ \ d_{i,j} \leq d^0_{i,j}; & lbH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} > d^0_{i,j}; & lbH_{i,j} = 0 \end{cases}$$



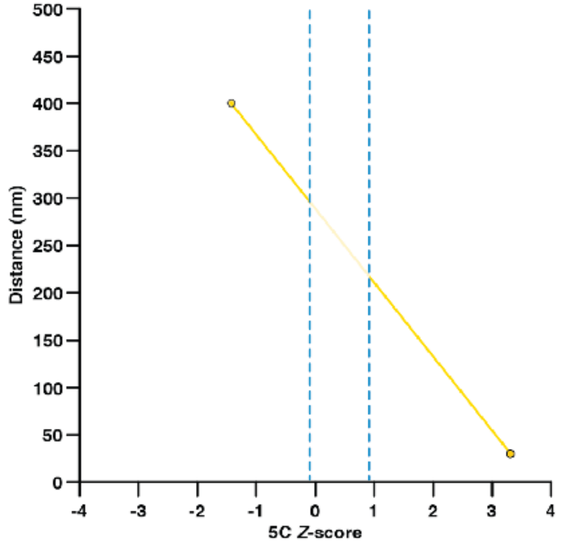
From 3C data to spatial distances



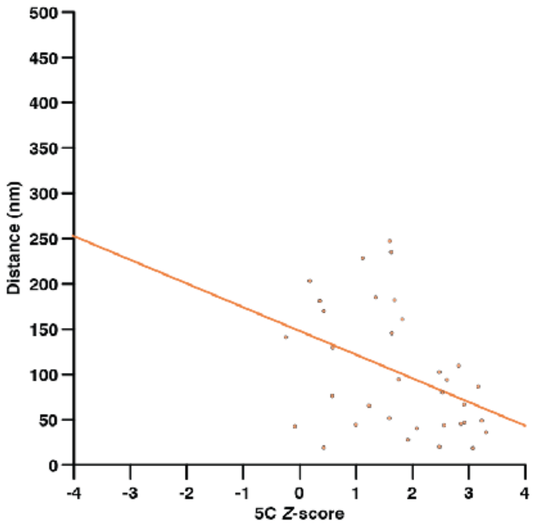
Neighbor fragments

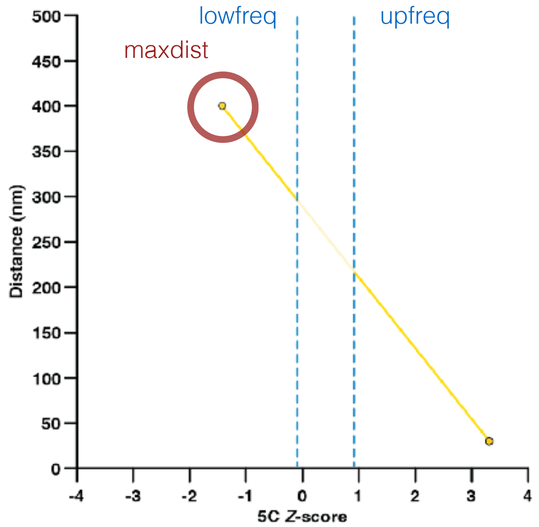


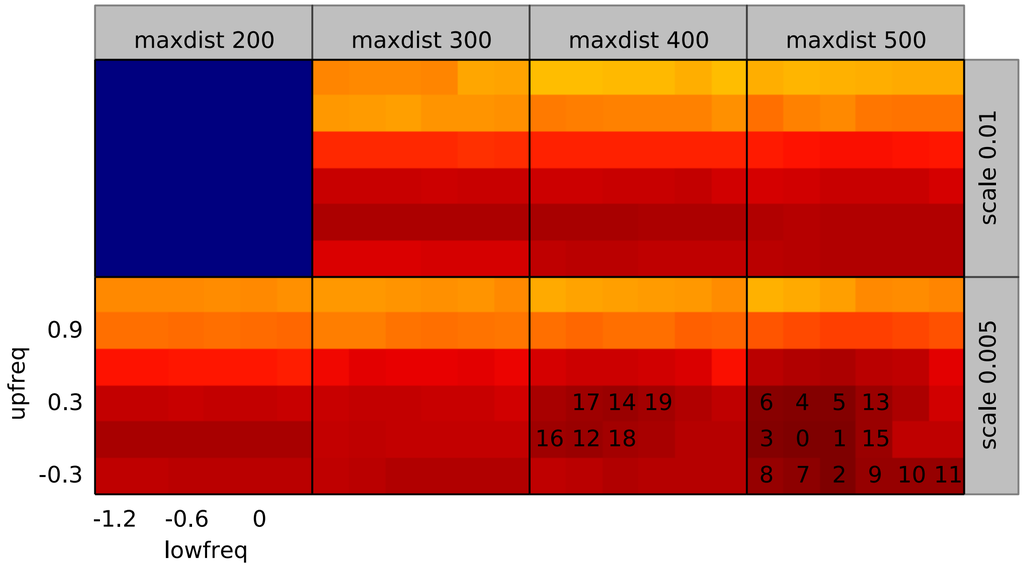
Non-Neighbor fragments



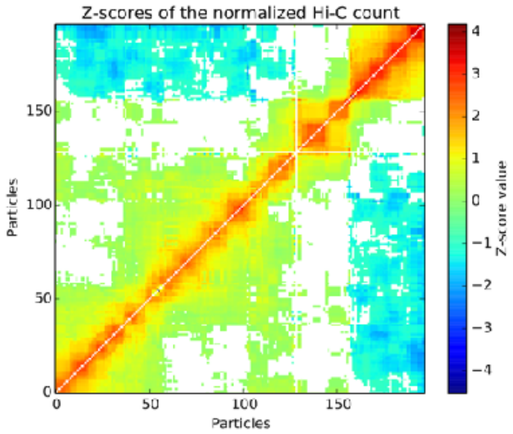
Parameter optimization

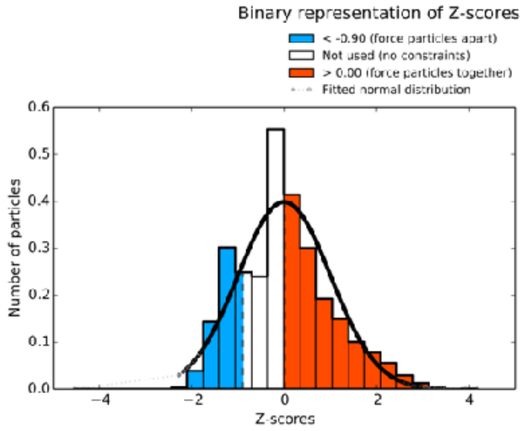


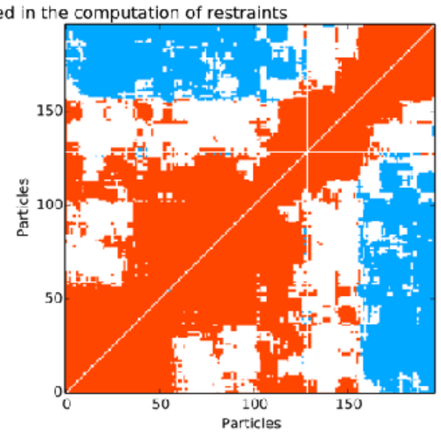


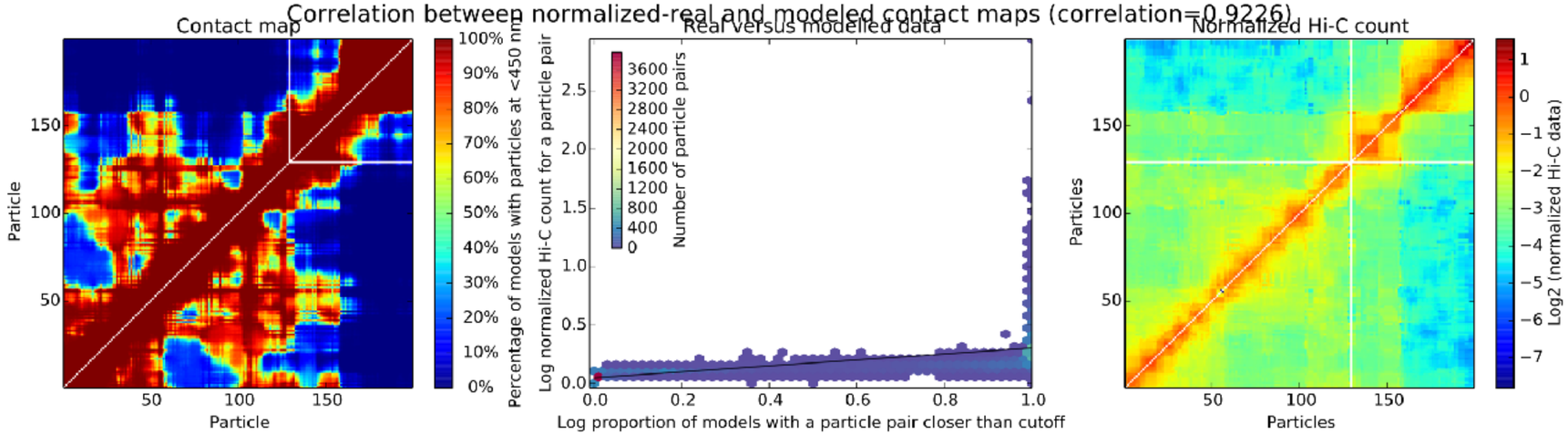


Parameter optimization

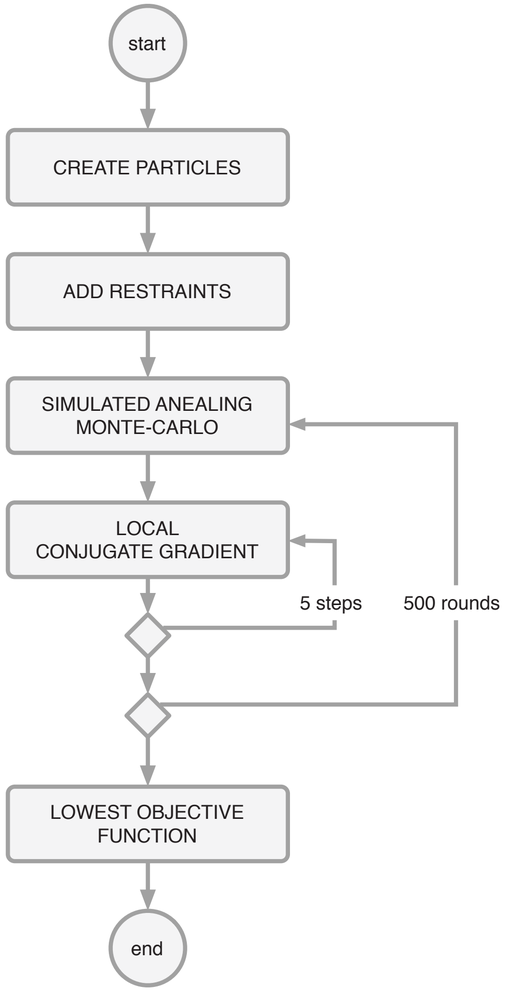


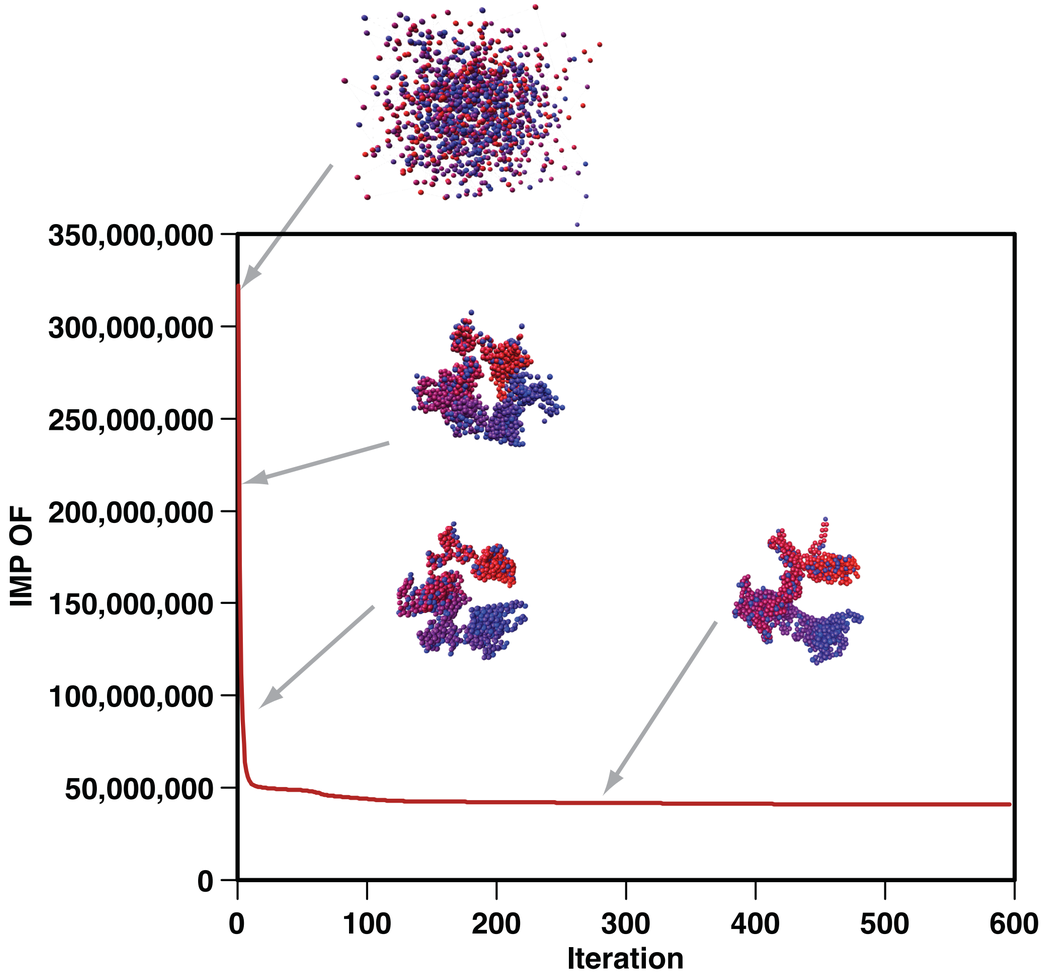




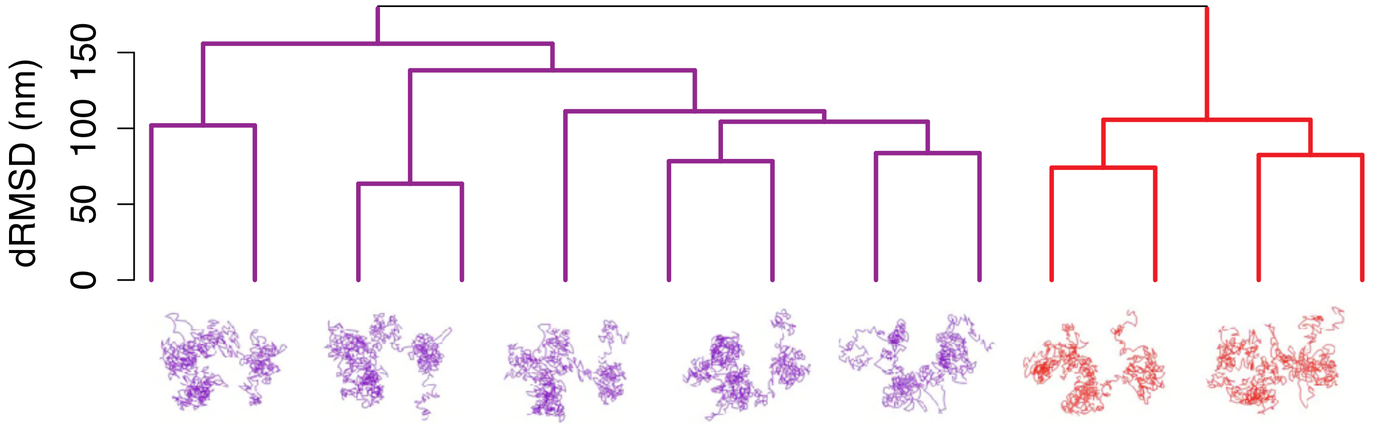


Optimization of the scoring function

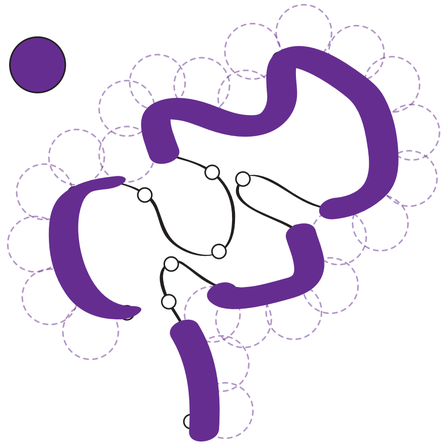




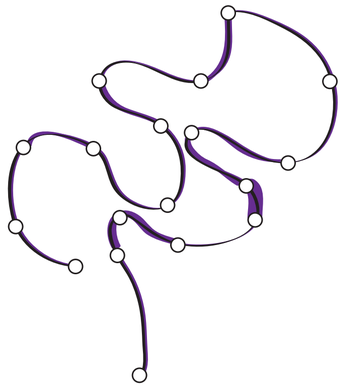
Model analysis: clustering and structural features



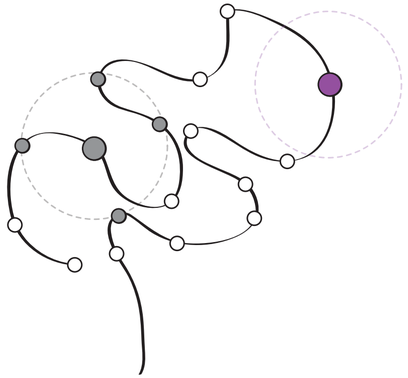
Accessibility (%)



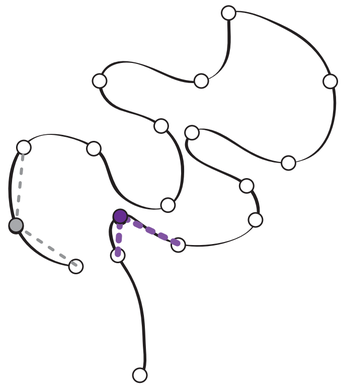
Density (bp/nm)



Interactions

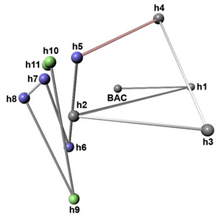


Angle

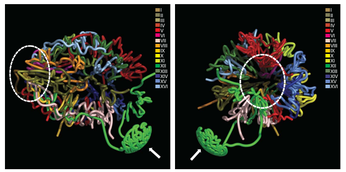


Are the models correct?

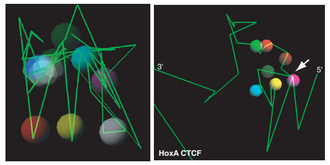
Trussart et al. NAR (2015)



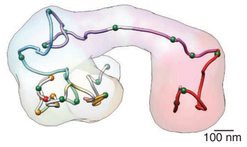
Jhunjhunwala (2008) Cell



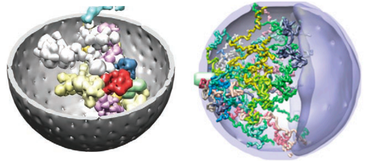
Duan (2010) Nature



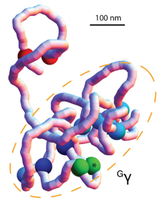
Fraser (2009) Genome Biology
Ferraiuolo (2010) Nucleic Acids Research



Baù (2011) Nature Structural & Molecular Biology



Kalhor (2011) Nature Biotechnology
Tjong (2012) Genome Research



Junier (2012) Nucleic Acids Research

Hu (2013) PLoS Computational Biology

Giorgetti, (2014) Cell

Nucleic Acids Research Advance Access published March 23, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv221

Assessing the limits of restraint-based 3D modeling of genomes and genomic domains

Marie Trussart[1,2], François Serra[3,4], Davide Baù[3,4], Ivan Junier[2,3], Luís Serrano[1,2,5] and Marc A. Marti-Renom[3,4,5,*]

[1]EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Barcelona, Spain, [4]Genome Biology Group, Centre Nacional d'Anàlisi Genòmica (CNAG), Barcelona, Spain and [5]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Received January 16, 2015; Revised February 16, 2015; Accepted February 22, 2015

ABSTRACT

Restraint-based modeling of genomes has been recently explored with the advent of Chromosome Conformation Capture (3C-based) experiments. We previously developed a reconstruction method to resolve the 3D architecture of both prokaryotic and eukaryotic genomes using 3C-based data. These models were congruent with fluorescent imaging validation. However, the limits of such methods have not systematically been assessed. Here we propose the first evaluation of a mean-field restraint-based reconstruction of genomes by considering diverse chromosome architectures and different levels of data noise and structural variability. The results show that: first, current scoring functions for 3D reconstruction correlate with the accuracy of the models; second, reconstructed models are robust to noise but sensitive to structural variability; third, the local structure organization of genomes, such as Topologically Associating Domains, results in more accurate models; fourth, to a certain extent, the models capture the intrinsic structural variability in the input matrices and fifth, the accuracy of the models can be *a priori* predicted by analyzing the properties of the interaction matrices. In summary, our work provides a systematic analysis of the limitations of a mean-field restrain-based method, which could be taken into consideration in further development of methods as well as their applications.

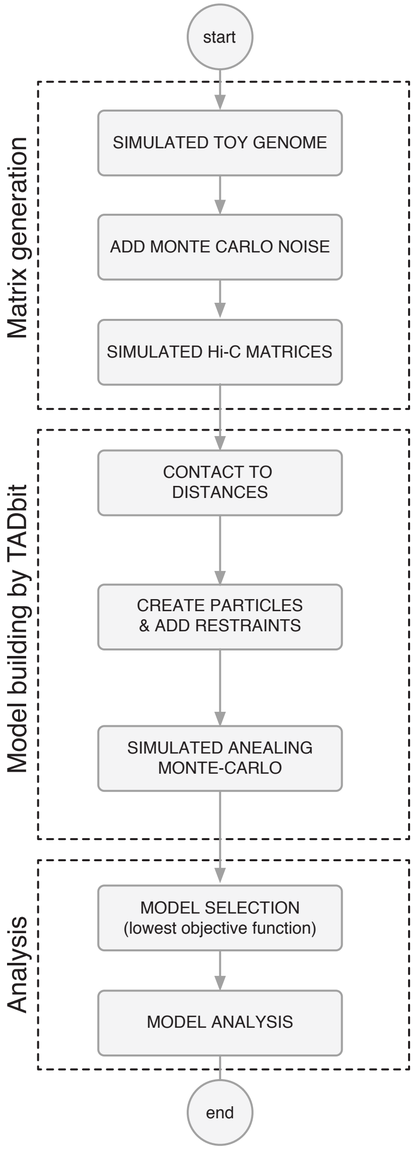
INTRODUCTION

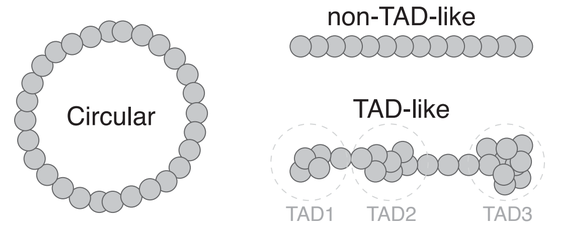
Recent studies of the three-dimensional (3D) conformation of genomes are revealing insights into the organization and the regulation of biological processes, such as gene expression regulation and replication (1–6). The advent of the so-called Chromosome Conformation Capture (3C) assays (7), which allowed identifying chromatin-looping interactions between pairs of loci, helped deciphering some of the key elements organizing the genomes. High-throughput derivations of genome-wide 3C-based assays were established with Hi-C technologies (8) for an unbiased identification of chromatin interactions. The resulting genome interaction matrices from Hi-C experiments have been extensively used for computationally analyzing the organization of genomes and genomic domains (5). In particular, a significant number of new approaches for modeling the 3D organization of genomes have recently flourished (9–14). The main goal of such approaches is to provide an accurate 3D representation of the bi-dimensional interaction matrices, which can then be more easily explored to extract biological insights. One type of methods for building 3D models from interaction matrices relies on the existence of a limited number of conformational states in the cell. Such methods are regarded as mean-field approaches and are able to capture, to a certain degree, the structural variability around these mean structures (15).

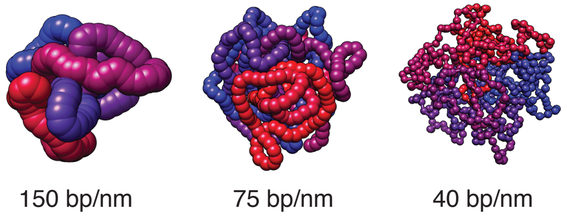
We recently developed a mean-field method for modeling 3D structures of genomes and genomic domains based on 3C interaction data (9). Our approach, called TADbit, was developed around the Integrative Modeling Platform (IMP, http://integrativemodeing.org), a general framework for restraint-based modeling of 3D bio-molecular structures (16). Briefly, our method uses chromatin interaction frequencies derived from experiments as a proxy of spatial proximity between the ligation products of the 3C libraries. Two fragments of DNA that interact with high frequency are dynamically placed close in space in our models while two fragments that do not interact as often will be kept apart. Our method has been successfully applied to model the structures of genomes and genomic domains in eukaryote and prokaryote organisms (17–19). In all of our studies, the final models were partially validated by assessing their

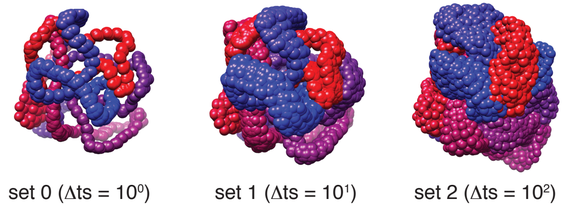
*To whom correspondence should be addressed. Tel: +34 934 020 542; Fax: +34 934 037 279; Email: mmarti@pcb.ub.cat

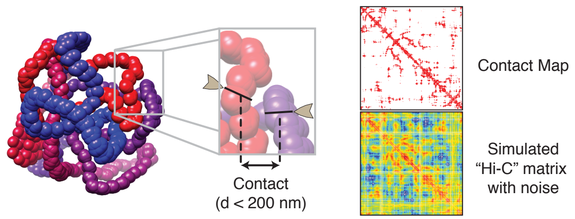
Toy models





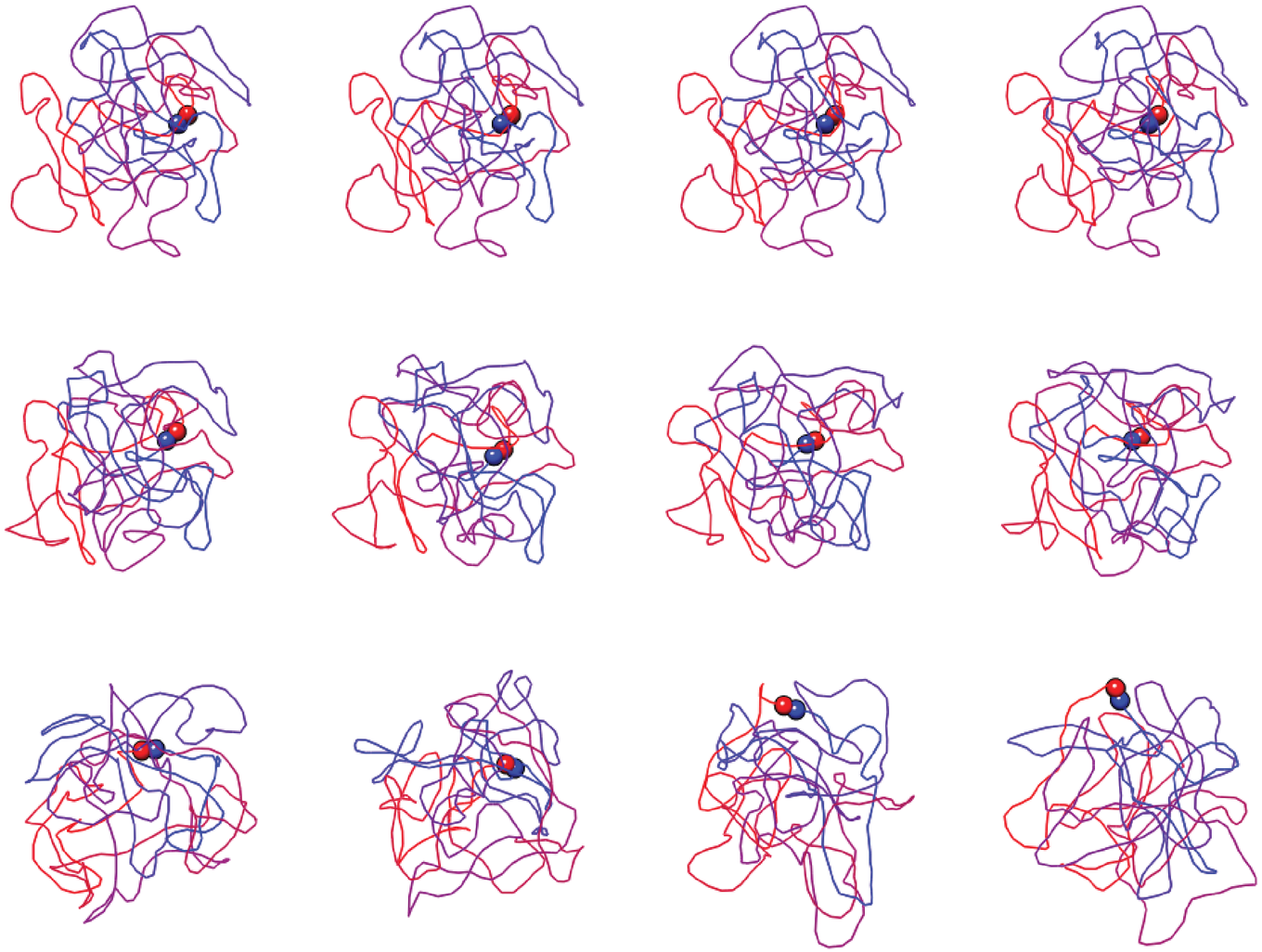




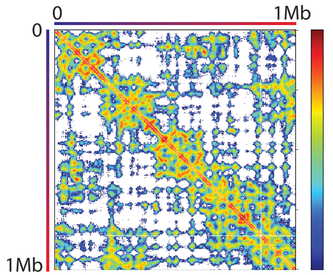


by Ivan Junier

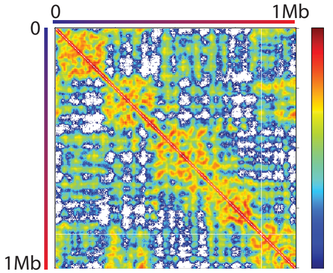
Toy interaction matrices



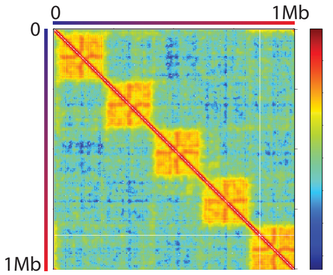
set 0 ($\Delta ts=10^0$)



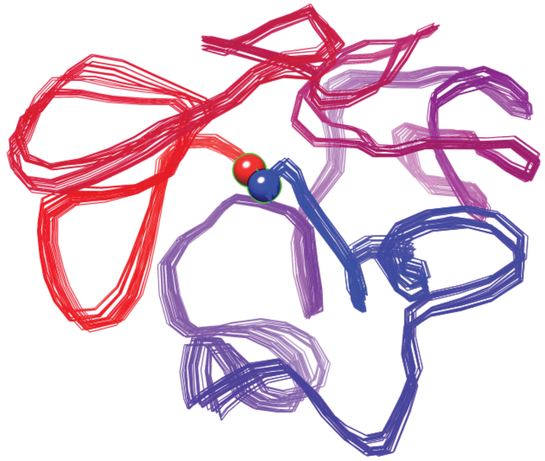
set 4 ($\Delta ts=10^4$)

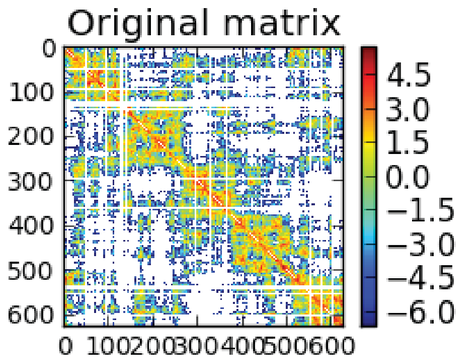


set 6 ($\Delta ts=10^6$)



Reconstructing toy models





chr40_TAD

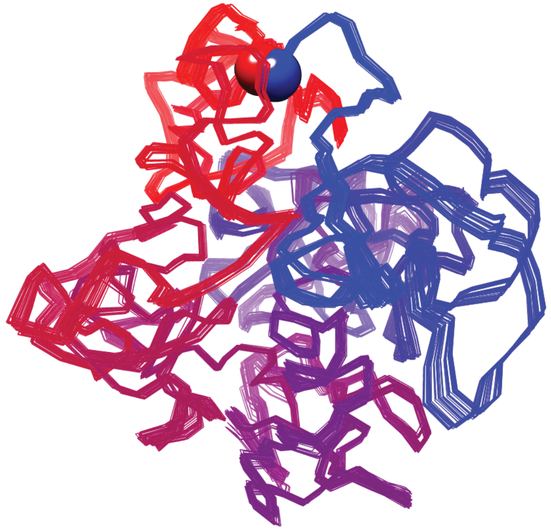
α=100

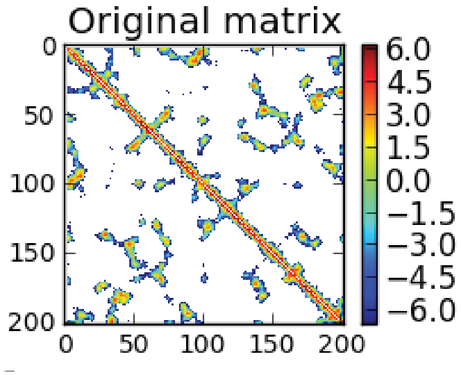
Δts=10

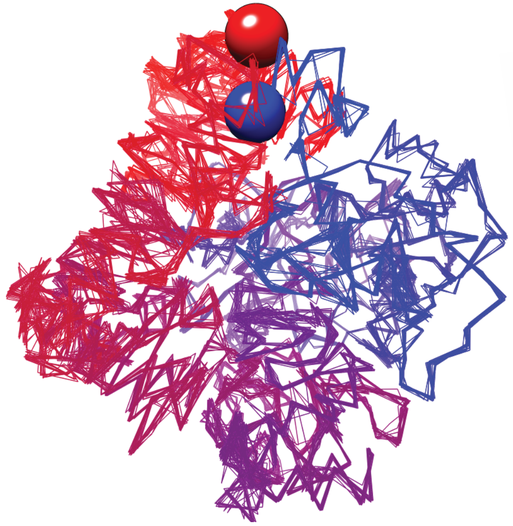
TADbit-SCC: 0.91

<dRMSD>: 32.7 nm

<dSCC>: 0.94







chr150_TAD

α=50

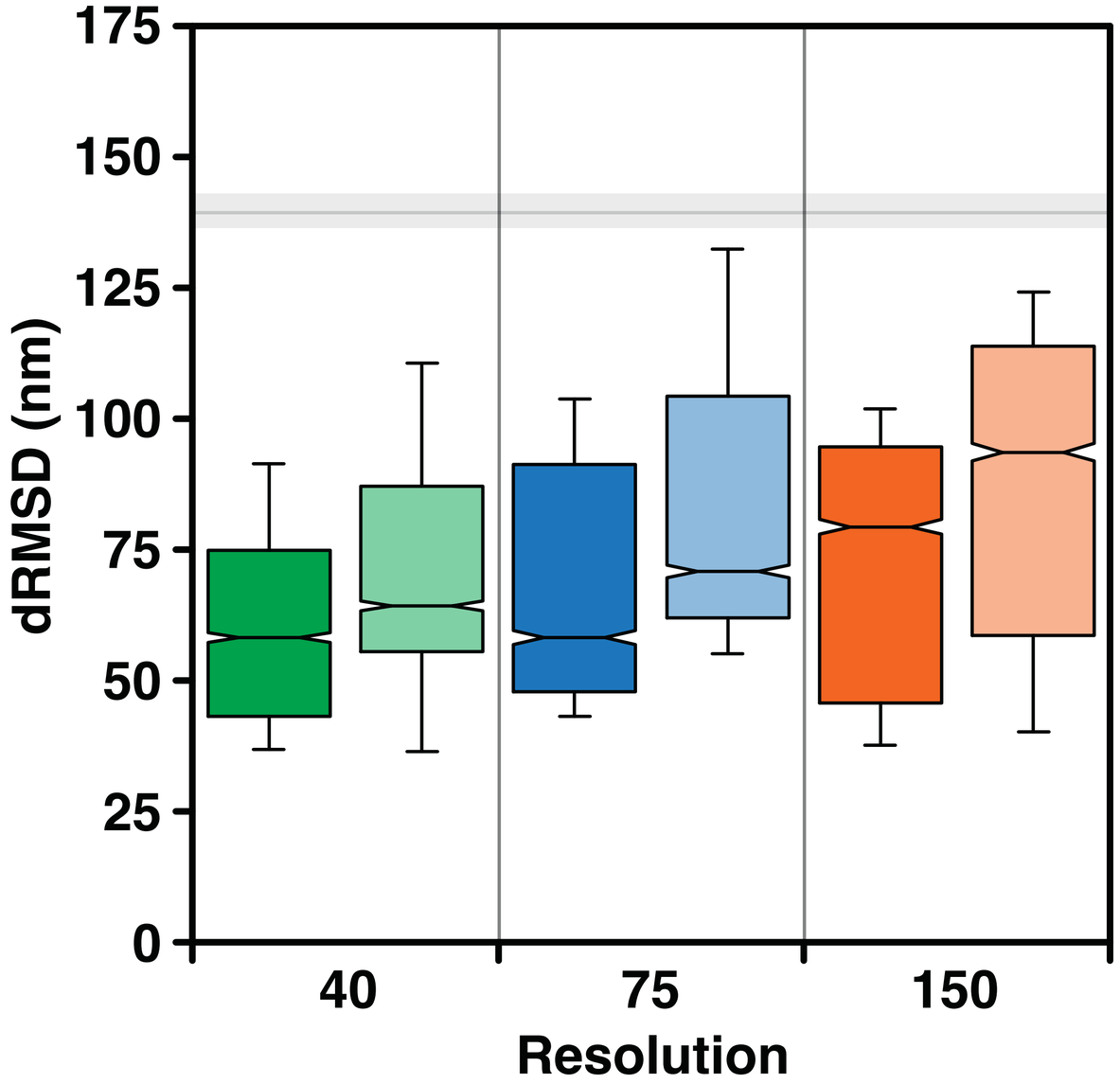
Δts=1

TADbit-SCC: 0.82

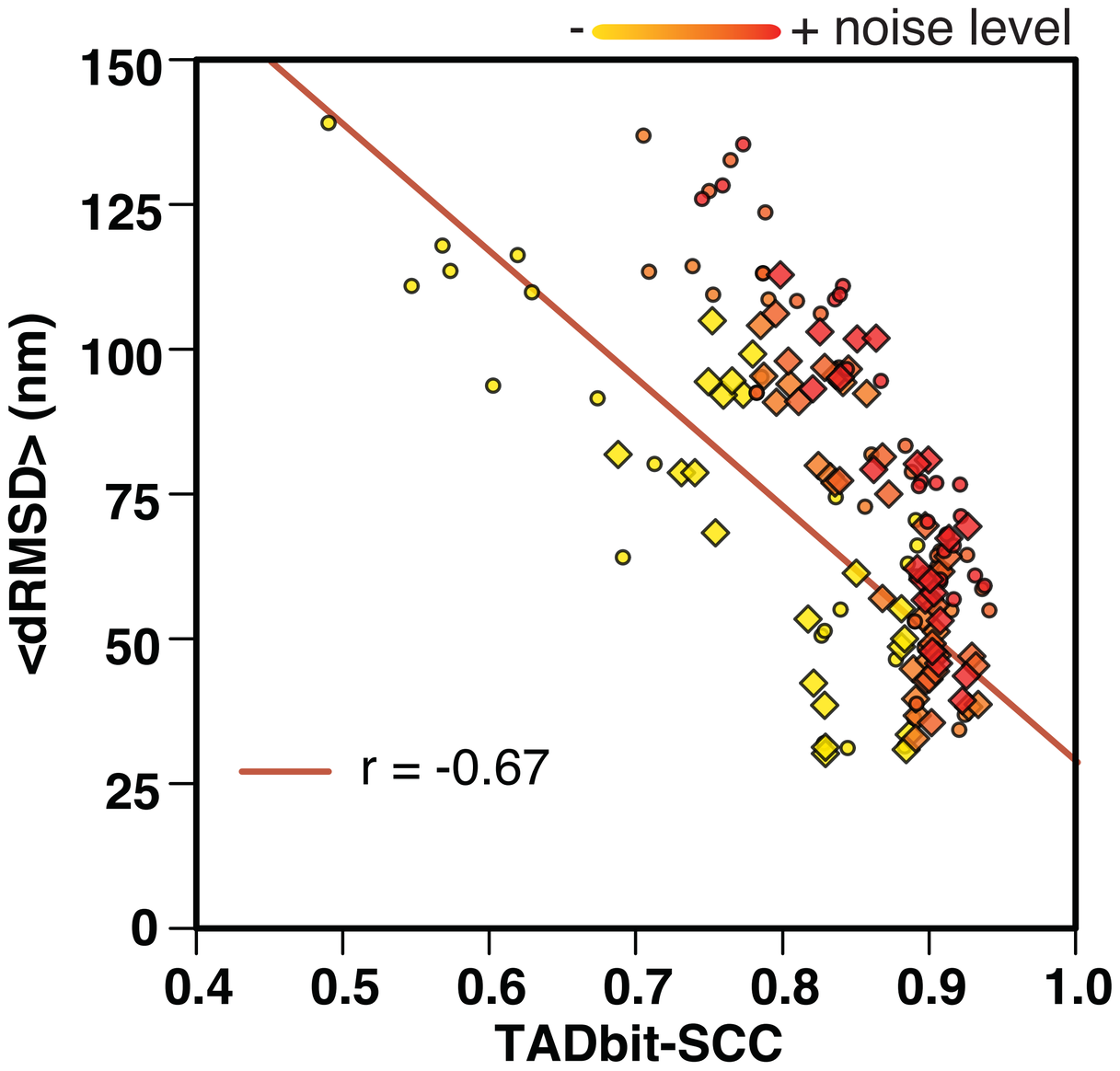
<dRMSD>: 45.4 nm

<dSCC>: 0.86

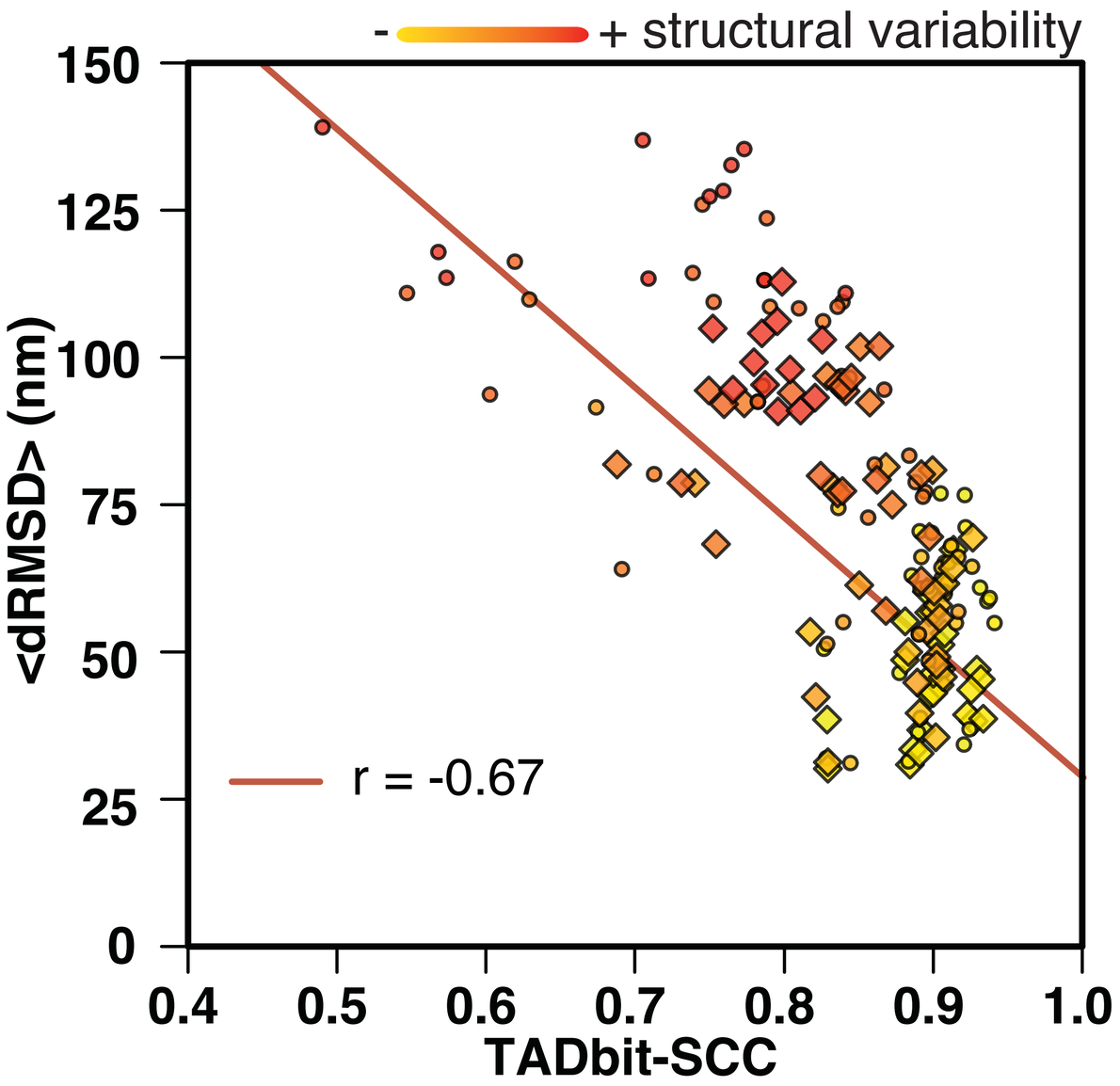
TADs & higher-res are "good"



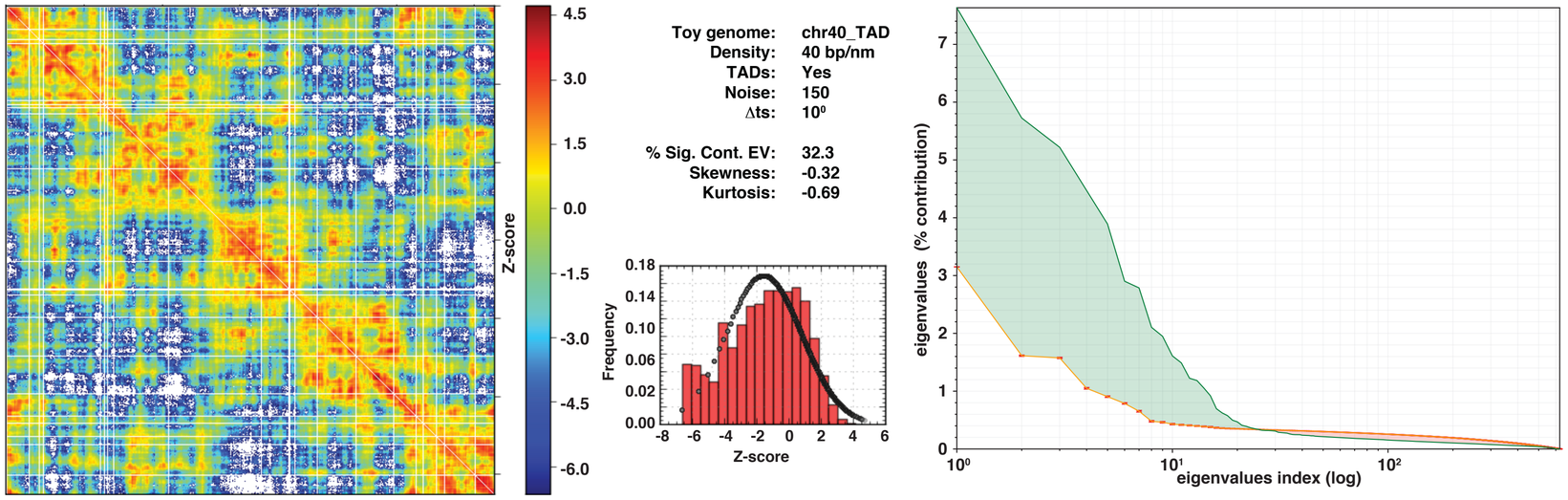
Noise is "OK"

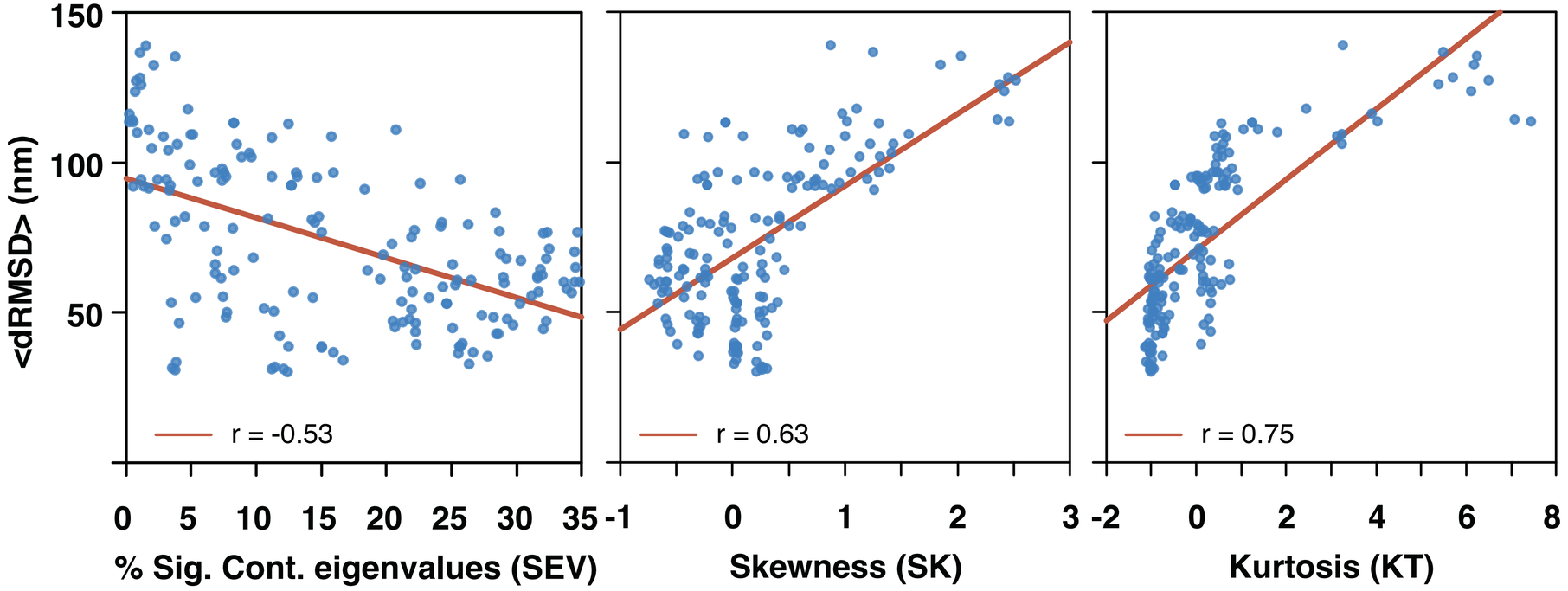

- + noise level
150
125
100
75
50
25
0
<dRMSD> (nm)
r = -0.67
0.4
0.5
0.6
0.7
0.8
0.9
1.0
TADbit-SCC

Structural variability is "NOT OK"

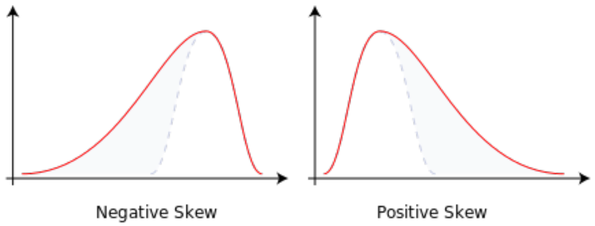


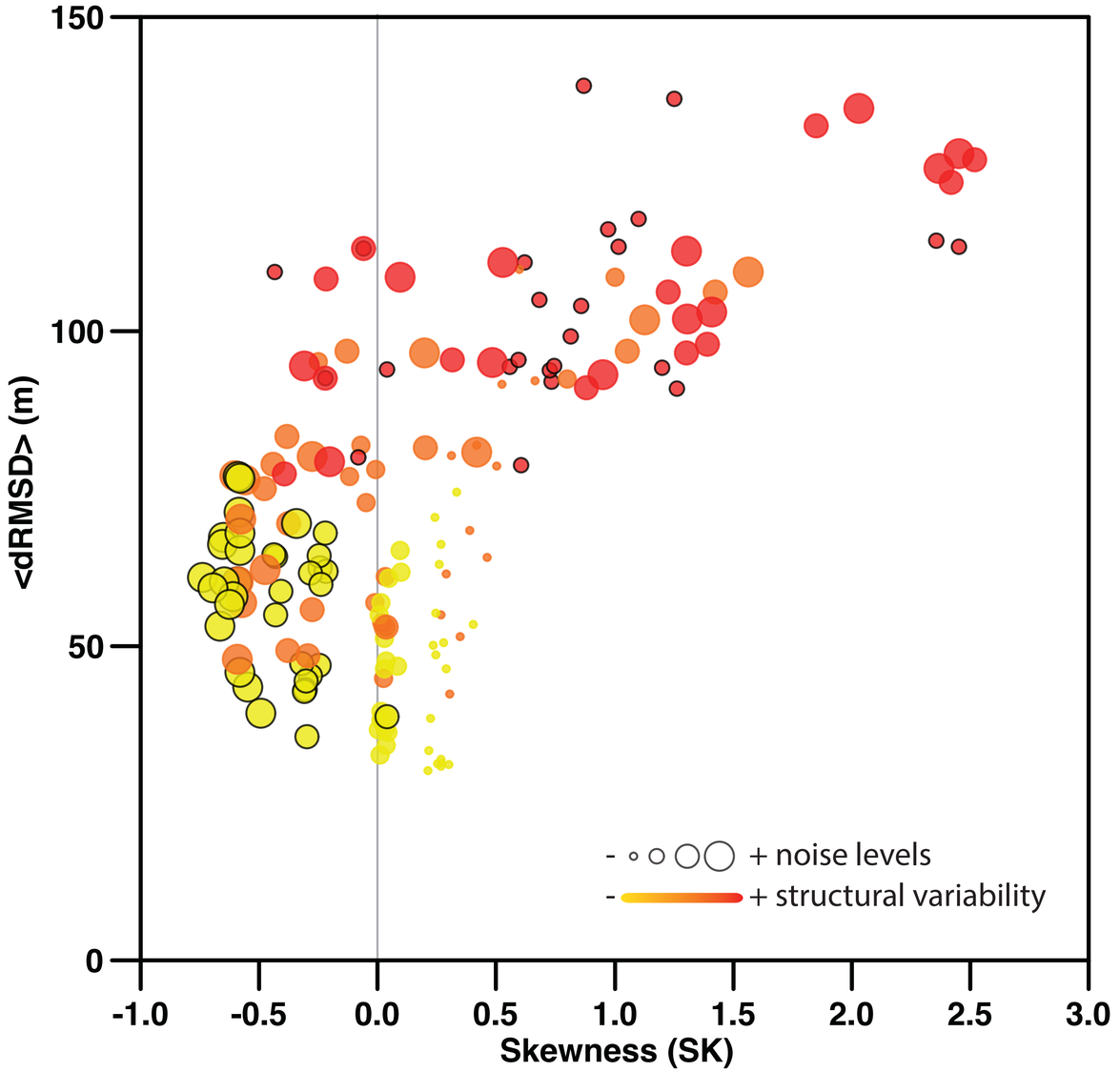
Can we predict the accuracy of the models?





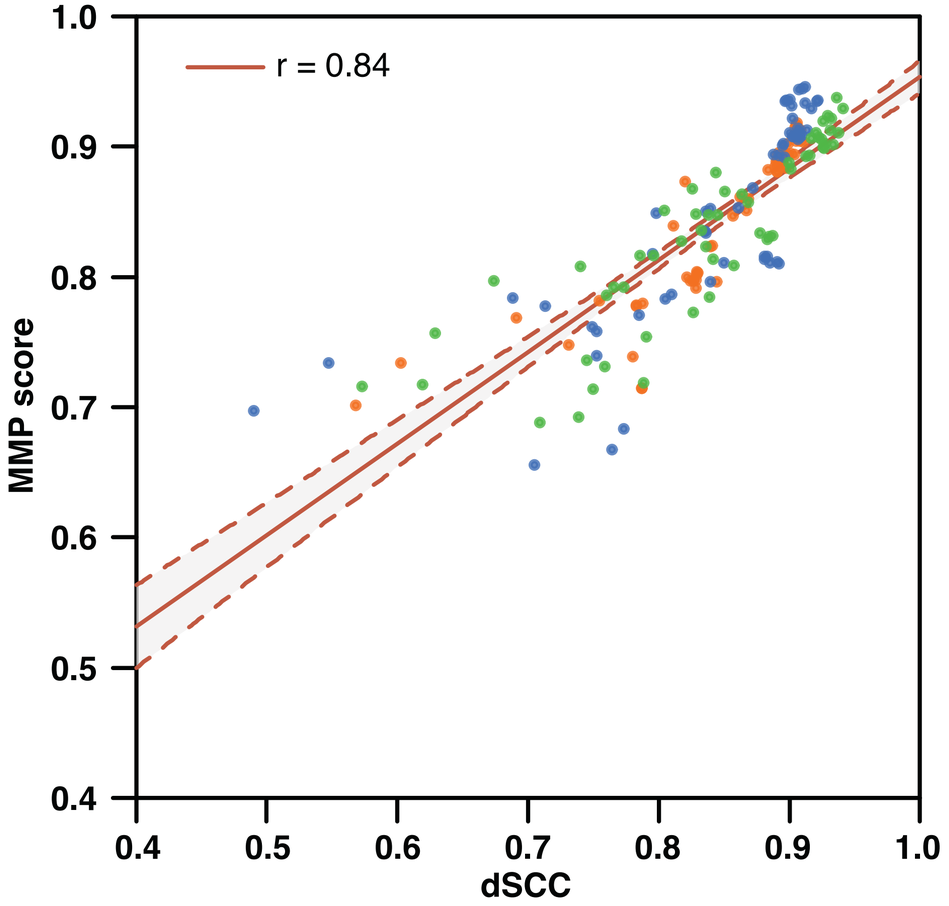
Skewness "side effect"





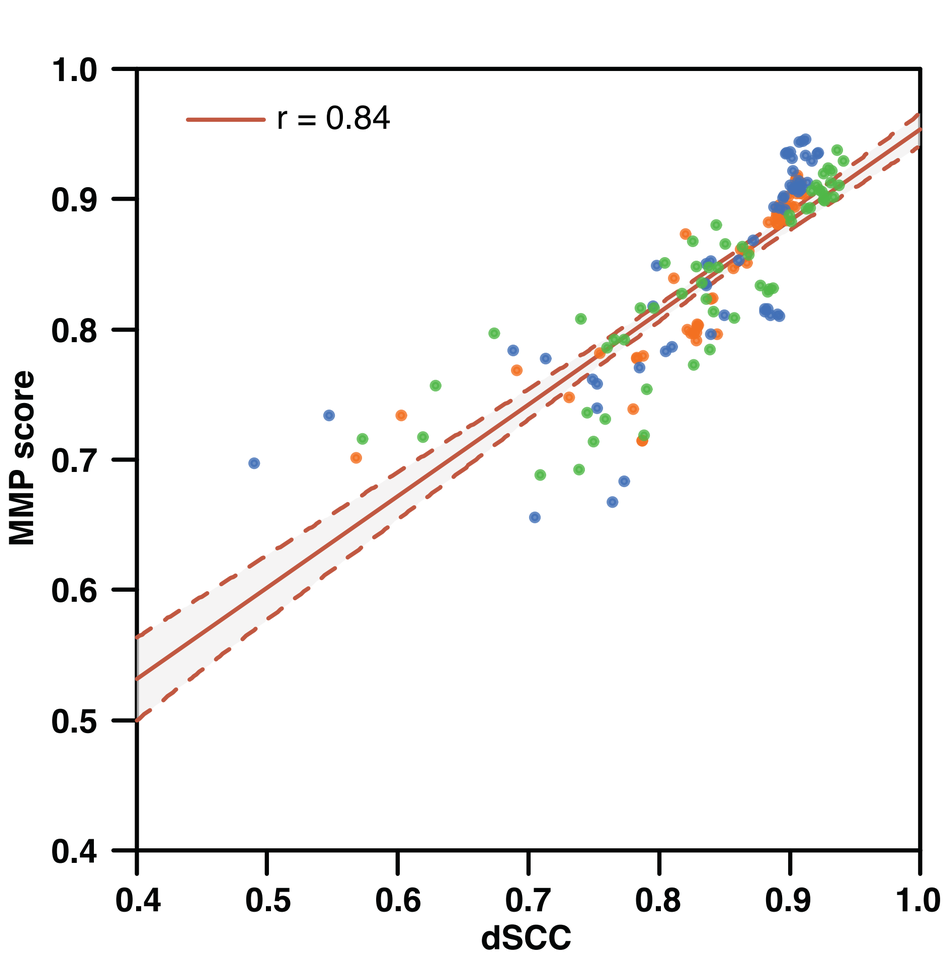
Can we predict the accuracy of the models?

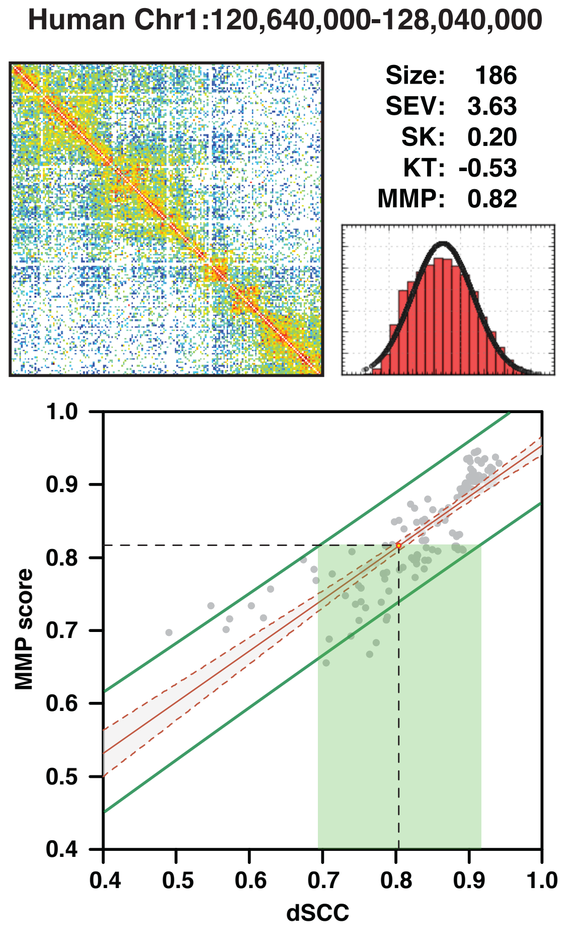
$$\mathrm{MMP} = -0.0002 * \mathrm{Size} + 0.0335 * \mathrm{SK} - 0.0229 * \mathrm{KU} + 0.0069 * \mathrm{SEV} + 0.8126$$



Can we predict the accuracy of the models?

$$MMP = -0.0002 * Size + 0.0335 * SK - 0.0229 * KU + 0.0069 * SEV + 0.8126$$





Higher-res is "good"

put your $$ in sequencing

Noise is "OK"

no need to worry much

Structural variability is "NOT OK"

homogenize your cell population!

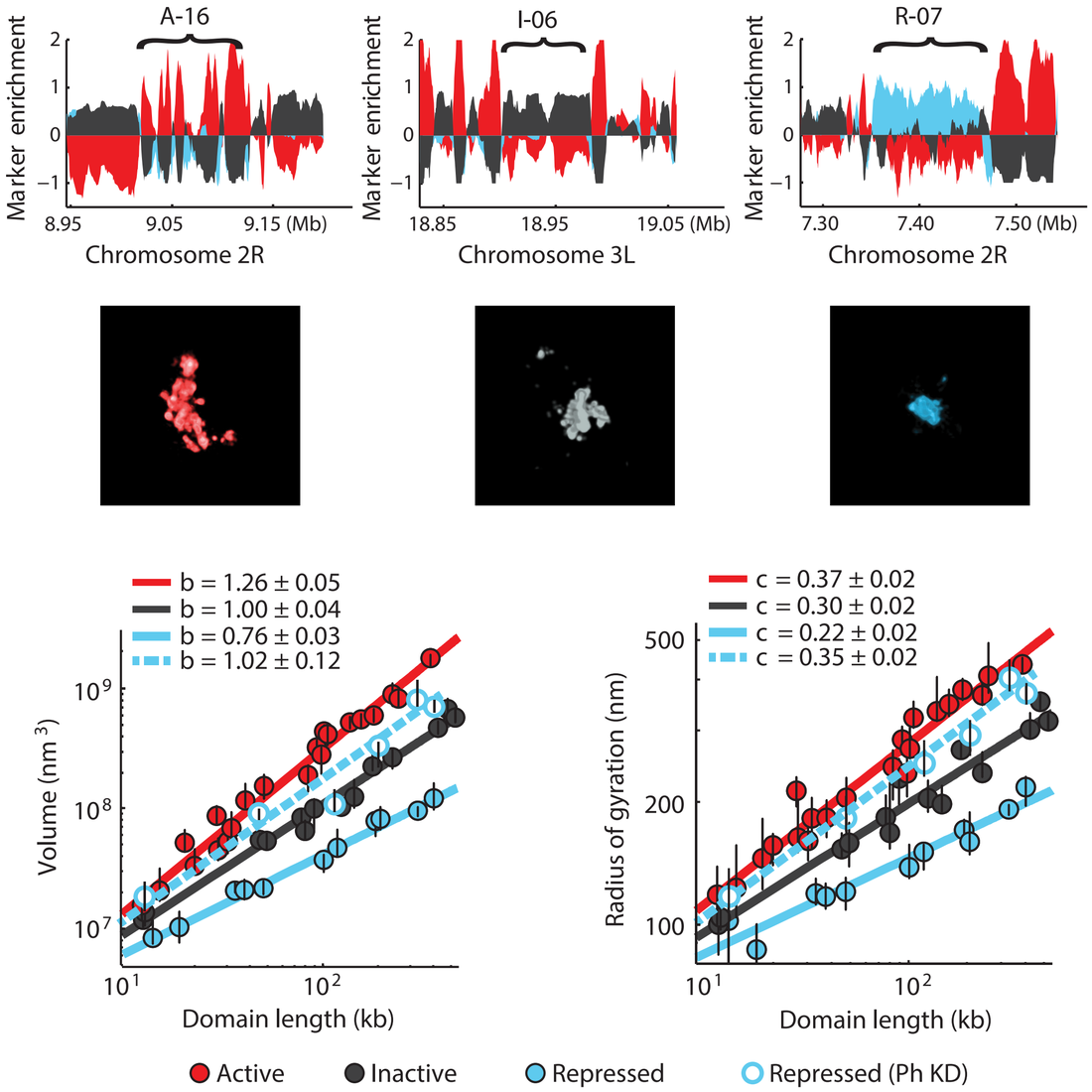
…but we can differentiate between noise and structural variability

and we can a priori predict the accuracy of the models

But… what about direct validation of models?

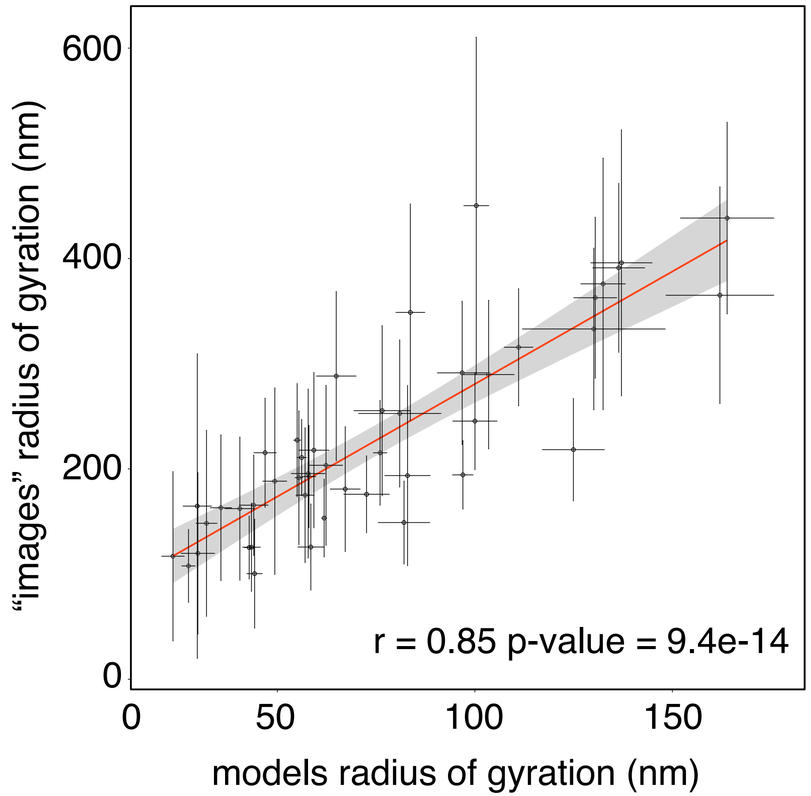
Model accuracy

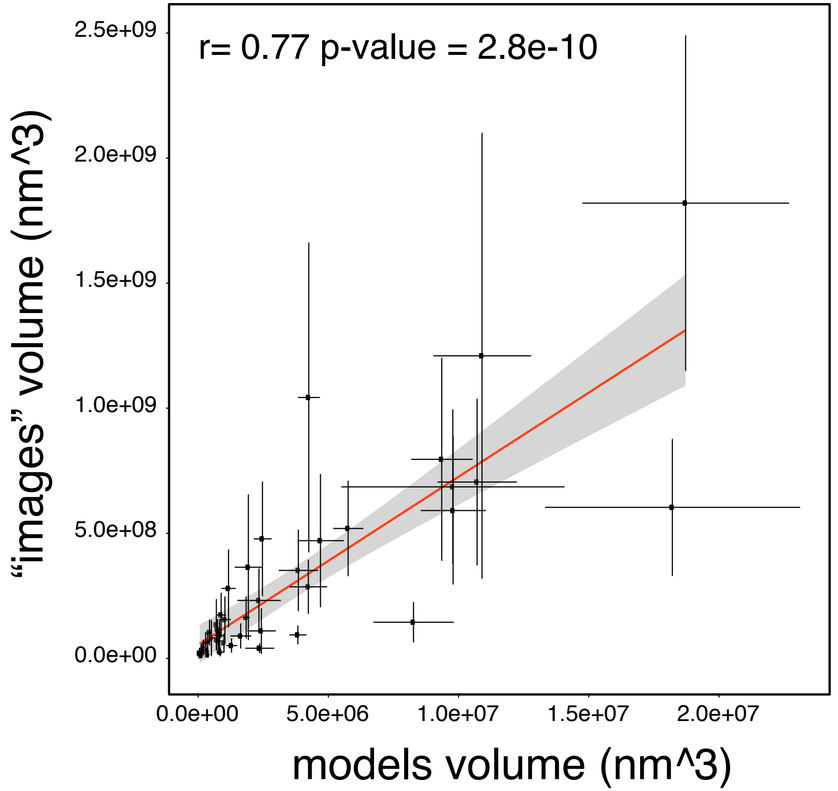
Boettiger, A. N., et al. (2016). Nature, 529, 418–422.



Model accuracy (fly@2Kb)

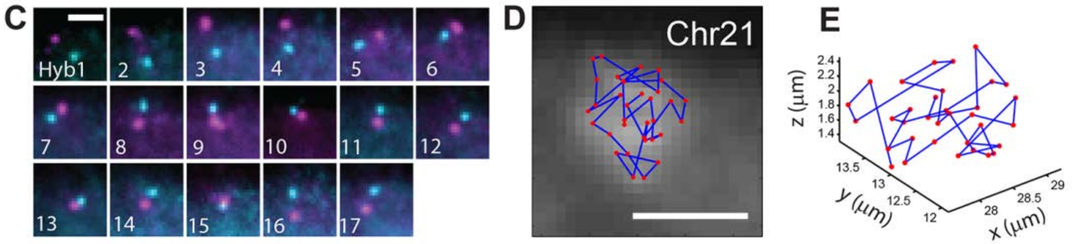
Boettiger, A. N., et al. (2016). Nature, 529, 418–422.

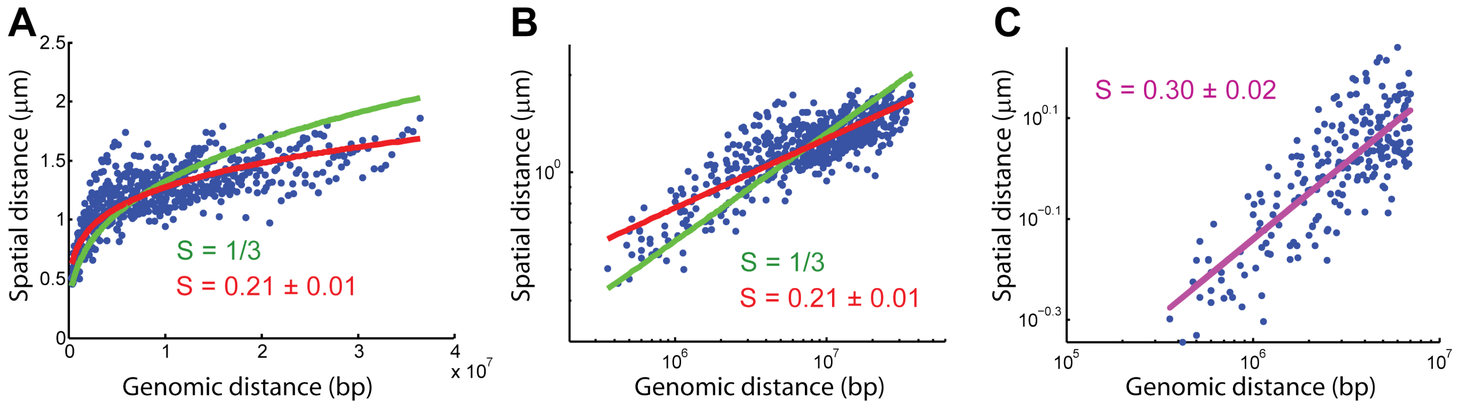


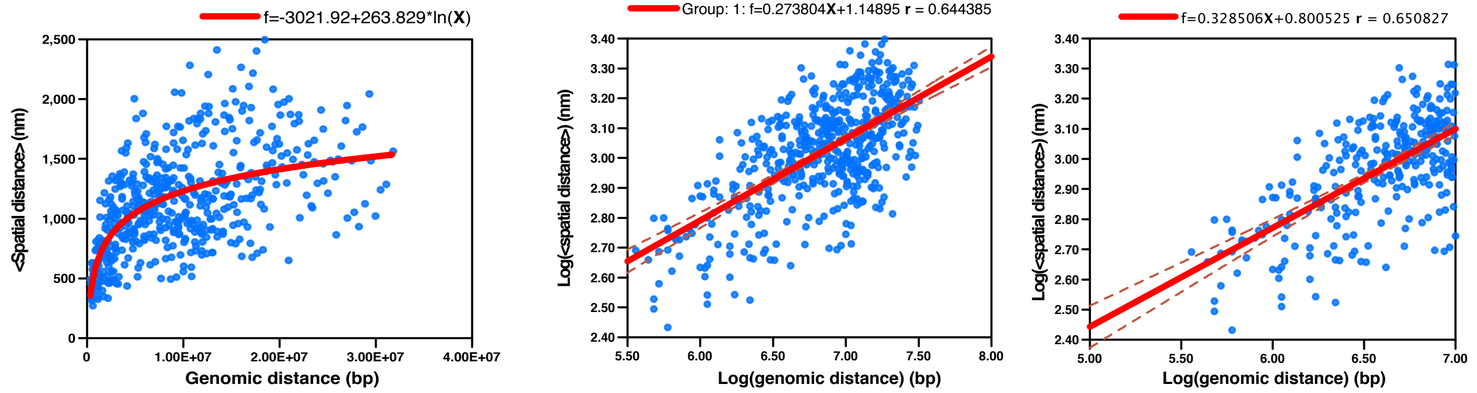


Model accuracy (Human Chr21@40Kb)

Wang, S., et al. (2016). Science 353, 598–602.







Model accuracy (Human Chr21@40Kb)

Wang, S., et al. (2016). Science 353(6299), 598–602.