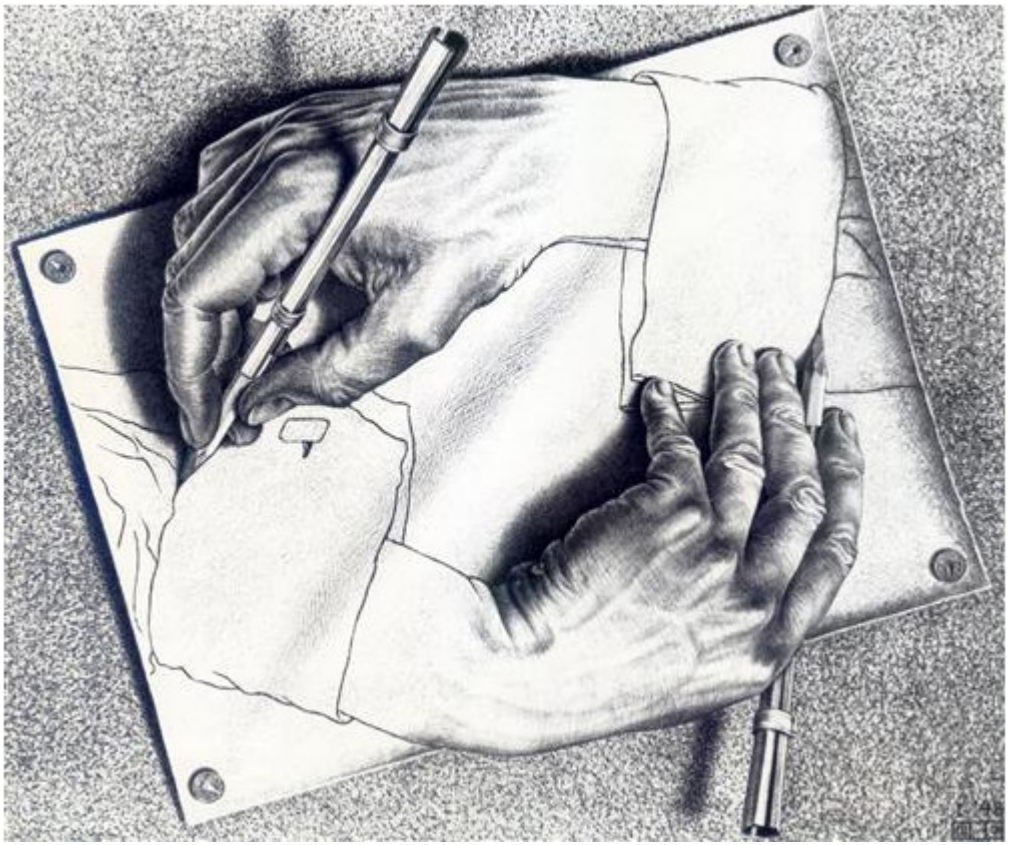
Hi-C data analysis - Exploring the 3D structure of the chromatin by processing DNA sequences

Sylvain Foissac, INRA Toulouse

NETBIO Paris - September 2015

Outline

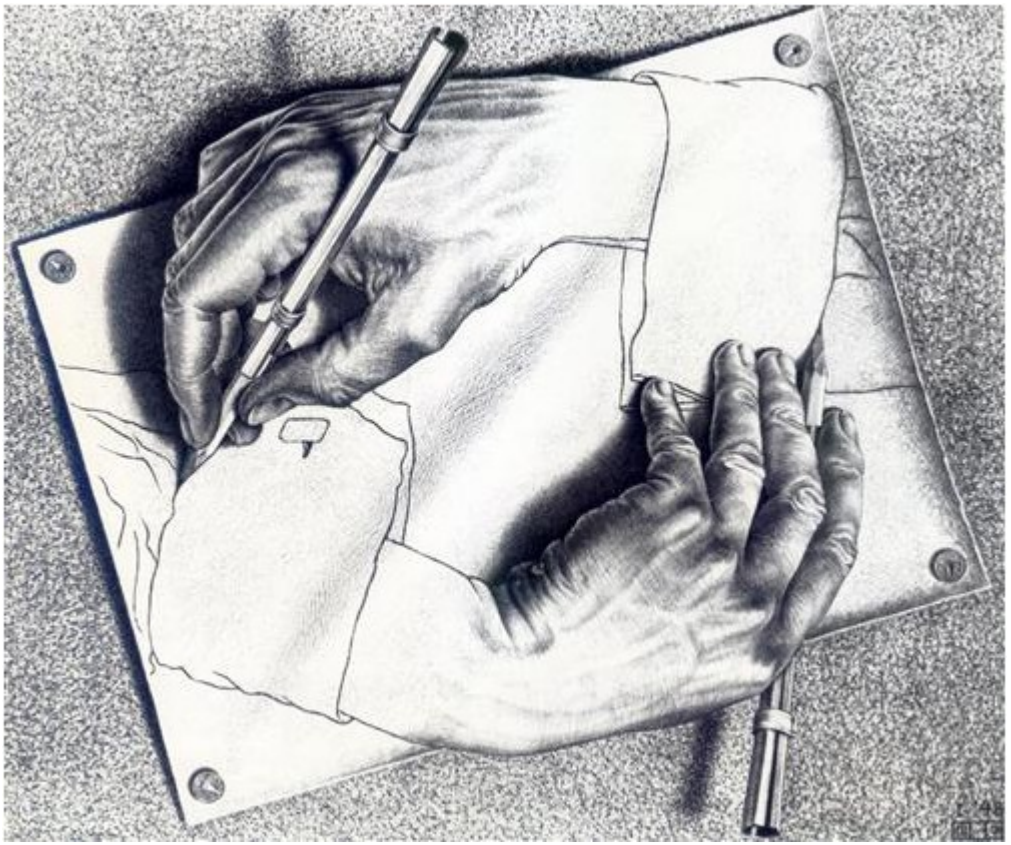
- Biological context
- More biological context
- Hi-C data processing
  - map
  - filter
  - count
  - normalize
  - segment
  - compare
- Conclusion, discussion, NETBIO lunch



M.C. Escher, 1948

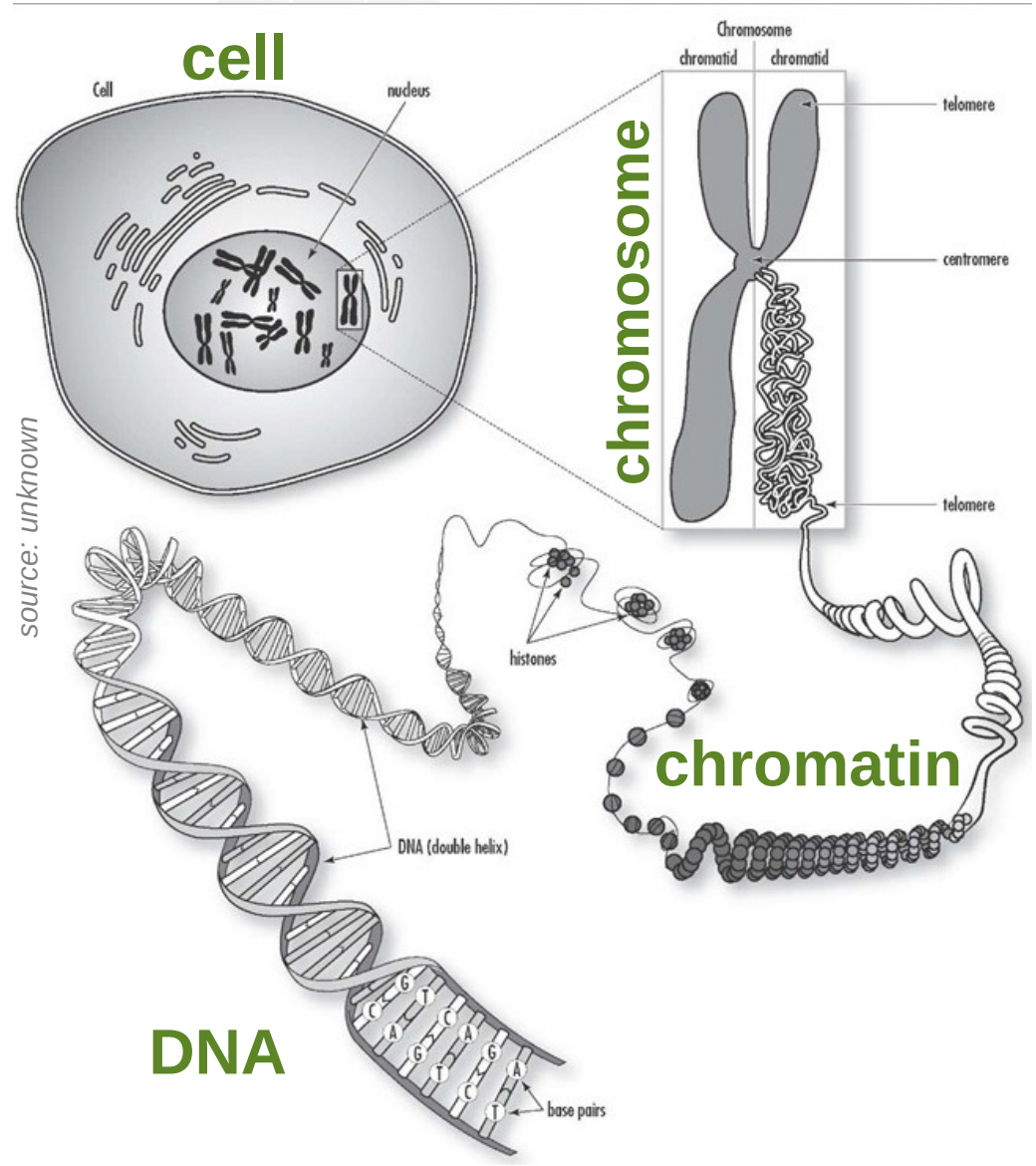
Outline

- Biological context
- More biological context
- Hi-C data processing
  - map
  - filter
  - count
  - normalize
  - segment
  - compare
- Conclusion, discussion, NETBIO lunch

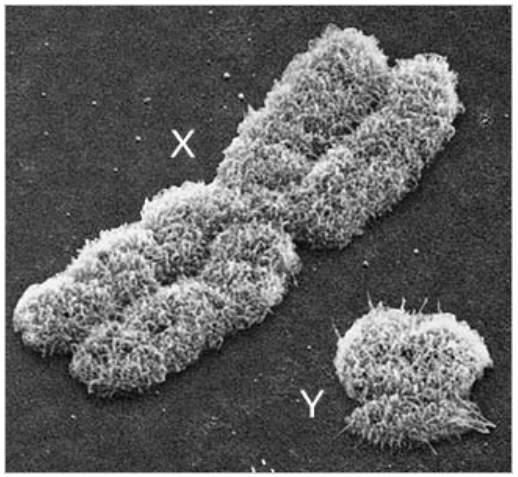


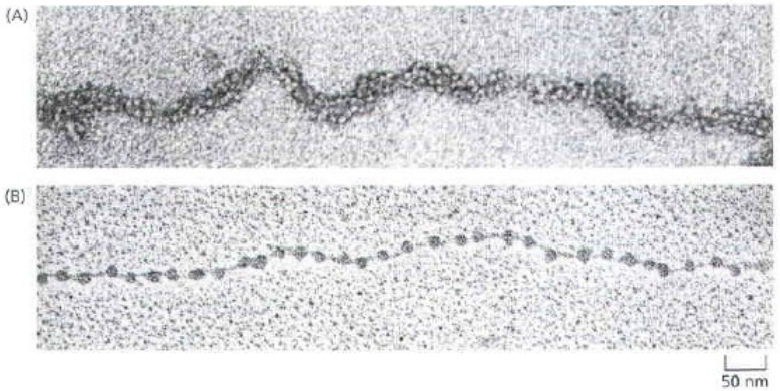
M.C. Escher, 1948

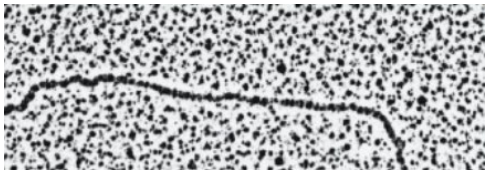
Life, cell, chromosome & DNA



source: unknown

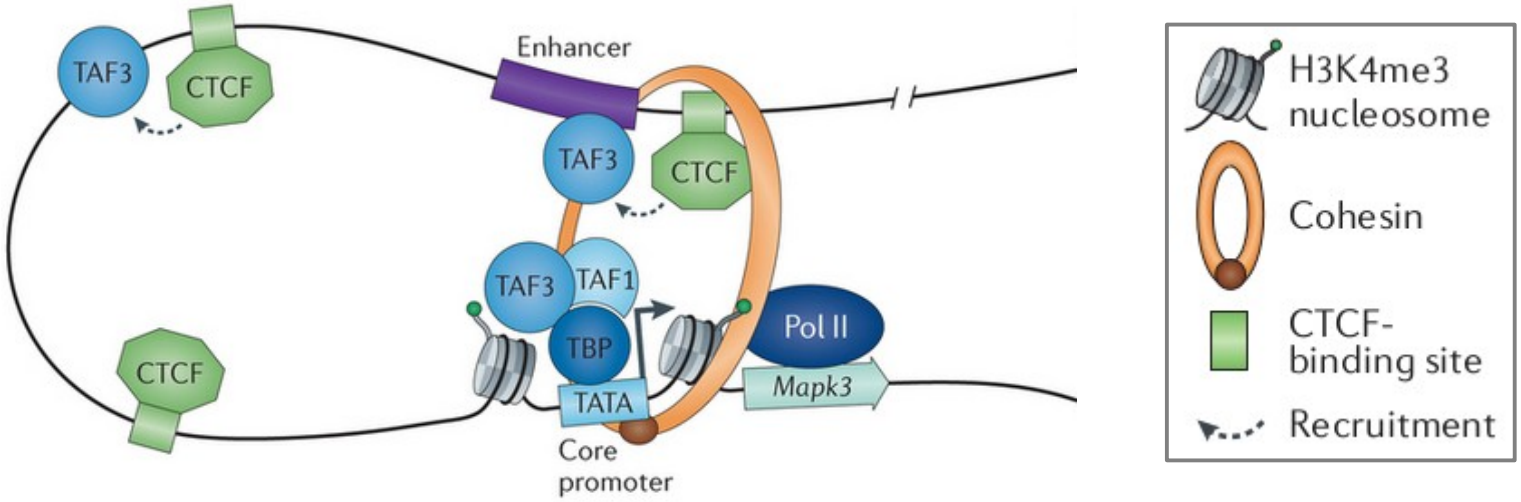






Life, cell, chromosome & DNA

From structure to function



Ong & Corces, Nat. Rev. Genet., 2014

3D structure impacts gene expression

From structure to function

Cell

Chromosomal Contact Permits Transcription between Coregulated Genes

Stephanie Fanucchi,[1] Youtaro Shibayama,[1] Shaun Burd,[1] Marc S. Weinberg,[3,4] and Musa M. Mhlanga[1,2,*]
[1]Gene Expression and Biophysics Group, Synthetic Biology Emerging Research Area, Biosciences Unit, Council for Scientific and Industrial Research, Pretoria, Gauteng 0001, South Africa
[2]Unidade de Biofísica e Expressão Genética, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, 1649-028 Portugal
[3]Antiviral Gene Therapy Research Unit, Department of Molecular Medicine and Haematology, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, Gauteng 2193, South Africa
[4]Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA
*Correspondence: yoda@mhlangalab.org
http://dx.doi.org/10.1016/j.cell.2013.09.051

SUMMARY

Transcription of coregulated genes occurs in the context of long-range chromosomal contacts that

2012). These highly sensitive assays ca
nascent mRNA and have revealed the
FISH foci in a fraction of the populatio
2010; Papantonis et al., 2010). This suc

CSH PRESS GENOME RESEARCH

Next Genera
USA Congre
27 - 28 October 2015,

HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CON

Institution: SWETS SUBSCRIPTIONSERVICE Sign In via User Nam

Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin

Elizabeth Ing-Simmons[1,2,7], Vlad C. Seitan[1,7], Andre J. Faure[3,8], Paul Flicek[3,4],

Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions

Darío G. Lupiáñez,[1,2] Katerina Kraft,[1,2] Verena Heinrich,[2] Peter Krawitz,[1,2] Francesco Brancati,[3] Eva Kl
Denise Horn,[2] Hülya Kayserili,[5] John M. Opitz,[6] Renata Laxova,[6] Fernando Santos-Simarro,[7,8]
Brigitte Gilbert-Dussardier,[9] Lars Wittler,[10] Marina Borschiwer,[1] Stefan A. Haas,[11] Marco Osterwalder,[12]
Bernd Timmermann,[13] Jochen Hecht,[1,14] Malte Spielmann,[1,2,14] Axel Visel,[12,15,16] and Stefan Mundlos[1]
[1]Max Planck Institute for Molecular Genetics, RG Development & Disease, 14195 Berlin, Germany
[2]Institute for Medical and Human Genetics, Charité Universitätsmedizin Berlin, 13353 Berlin, Germany
[3]Medical Genetics Unit, Policlinico Tor Vergata University Hospital, 00133 Rome, Italy

Nucleic Acids Research Advance Access published February 4, 2015

Nucleic Acids Research, 2015 1
doi: 10.1093/nar/gkv046

Spatial re-organization of myogenic regulato sequences temporally controls gene expres

Akihito Harada[1], Chandrashekara Mallappa[2], Seiji Okada[1], John T. Butler[2],
P. Baker[2,3], Jeanne B. Lawrence[2], Yasuyuki Ohkawa[1,2,*] and Anthony N. Im

[1]Department of Advanced Medical Initiatives, JST-CREST, Faculty of Medicine, Kyushu Univers
812-8582, Japan, [2]Department of Cell and Developmental Biology, University of Massachusetts

Cell

21. G. A. Wray, Nat. Rev. Genet. 8, 206–216 (2007).
22. S. B. Carroll, Cell 134, 25–36 (2008).

and K. Paszkiewicz and the Exeter Sequencing Service facilit
for genome sequencing services. This work was also support

TRANSCRIPTION

CTCF establishes discrete functional chromatin domains at the *Hox* clusters during differentiation

Varun Narendra,[1,2] Pedro P. Rocha,[3] Disi An,[4] Ramya Ravi
Esteban O. Mazzoni,[4,*] Danny Reinberg[1,2,*]

Polycomb and Trithorax group proteins encode the epigenetic
identity by establishing inheritable domains of repressive and active chromatin within
the *Hox* clusters. Here we demonstrate that the CCCTC-binding factor (CTCF) functions

Nuclear Aggregation of Olfactory Receptor Genes Gov Their Monogenic Ex

E. Josephine Clowney,[1] Mark A. LeGros,[2,4] Colleen P.
Eirene C. Markenskoff-Papadimitriou,[3] Markko Myllys,
and Stavros Lomvardas[1,2,3,*]
[1]Program in Biomedical Sciences
[*]Physical Biosciences Division, Lawrence Berkeley National La

Leading Edge
Previews

Cell

A CRISPR Connection between Chromatin Topology and Genetic Disorders

Bing Ren[1,*] and Jesse R. Dixon[1]
[1]Ludwig Institute for Cancer Research, University of California, San Diego, School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093-0653
*Correspondence: biren@ucsd.edu
http://dx.doi.org/10.1016/j.cell.2015.04.047

Structural variations are common in the human genome, but their contributions to human diseases
fine. Lupiáñez et al. demonstrate that some structural variants can interrupt
resulting in ectopic enhancer-promoter interactions, altered spatiotemporal
erns, and developmental disorders.

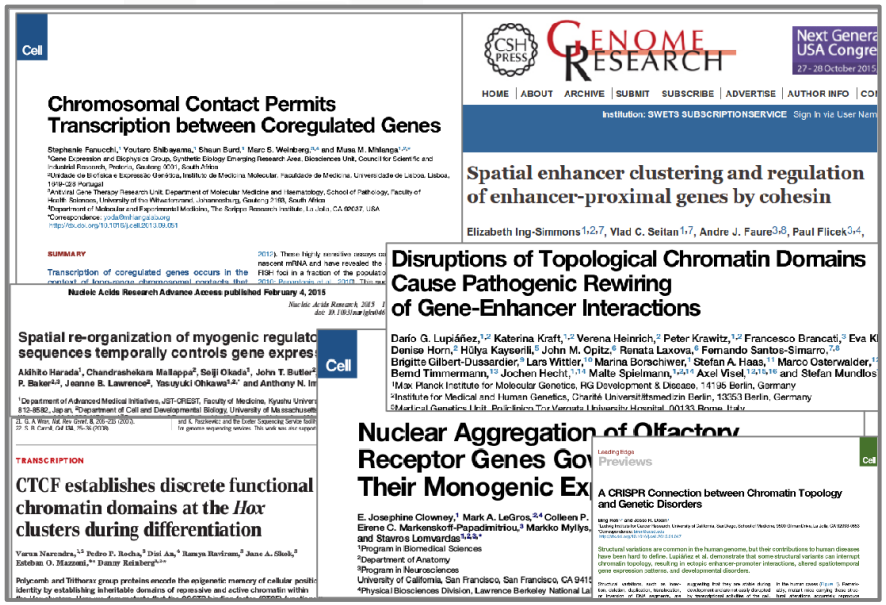
as inser-
tion, deletion, duplication, translocation,
or inversion of DNA segments, are

suggesting that they are stable during
development and are not easily disrupted
by transcriptional activities of the cell.

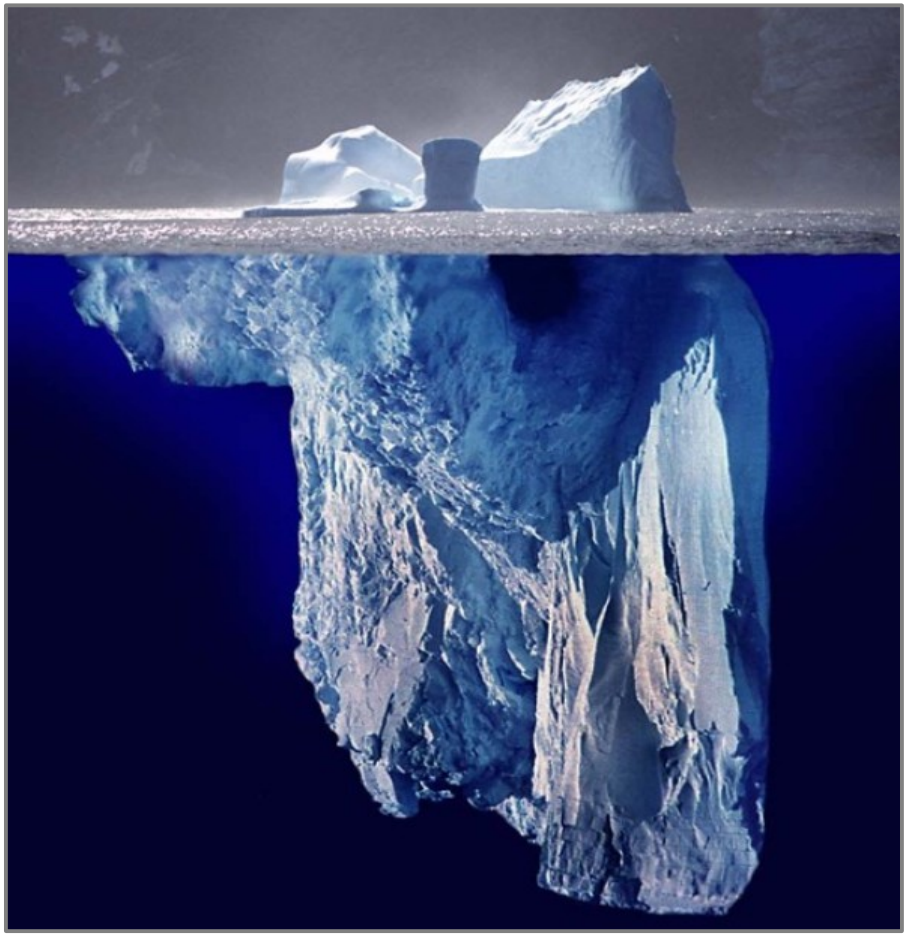
in the human cases (Figure 1). Remark-
ably, mutant mice carrying these struc-
tural alterations accurately reproduce

3D structure impacts gene expression

From structure to function



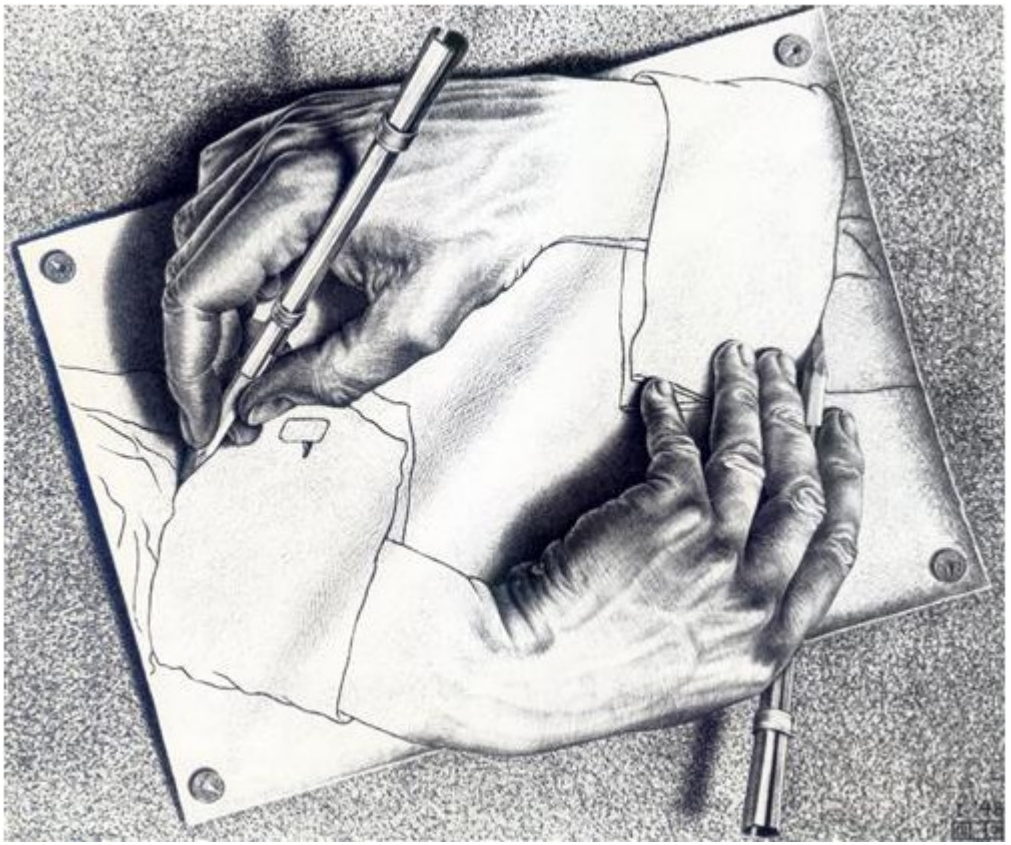
?



3D structure impacts gene expression

Outline

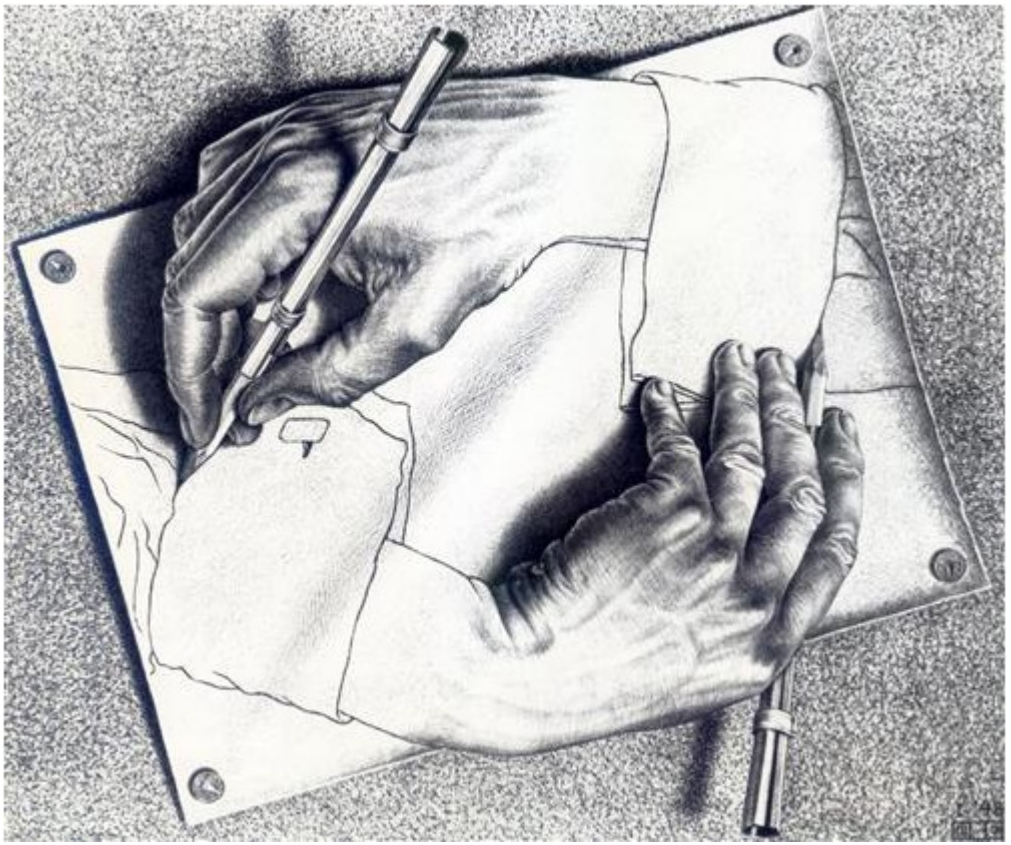
- Biological context
- More biological context
- Hi-C data processing
  - map
  - filter
  - count
  - normalize
  - segment
  - compare
- Conclusion, discussion, NETBIO lunch



M.C. Escher, 1948

Outline

- Biological context
- More biological context
- Hi-C data processing
  - map
  - filter
  - count
  - normalize
  - segment
  - compare
- Conclusion, discussion, NETBIO lunch

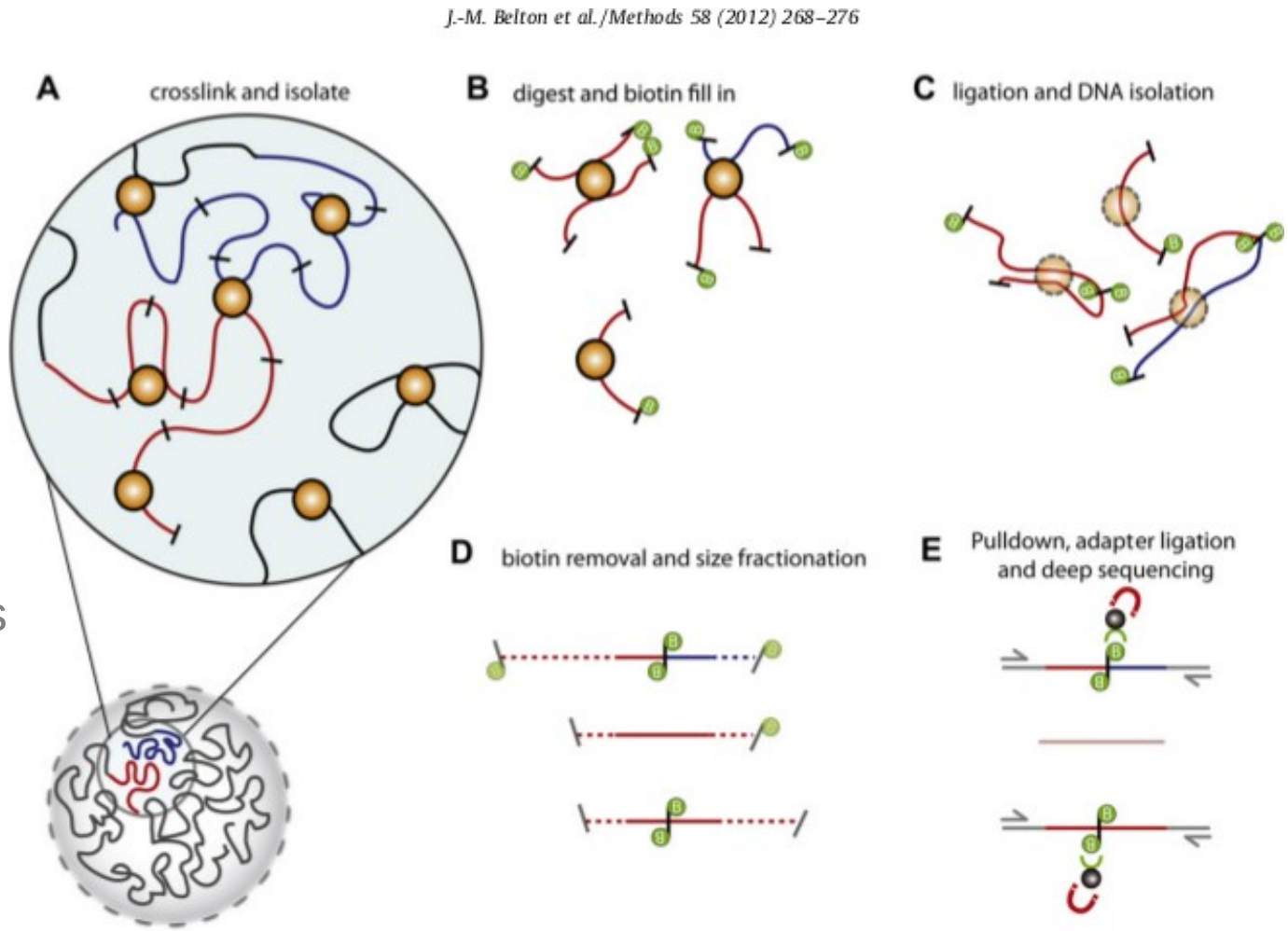


M.C. Escher, 1948

Hi-C: the experiment

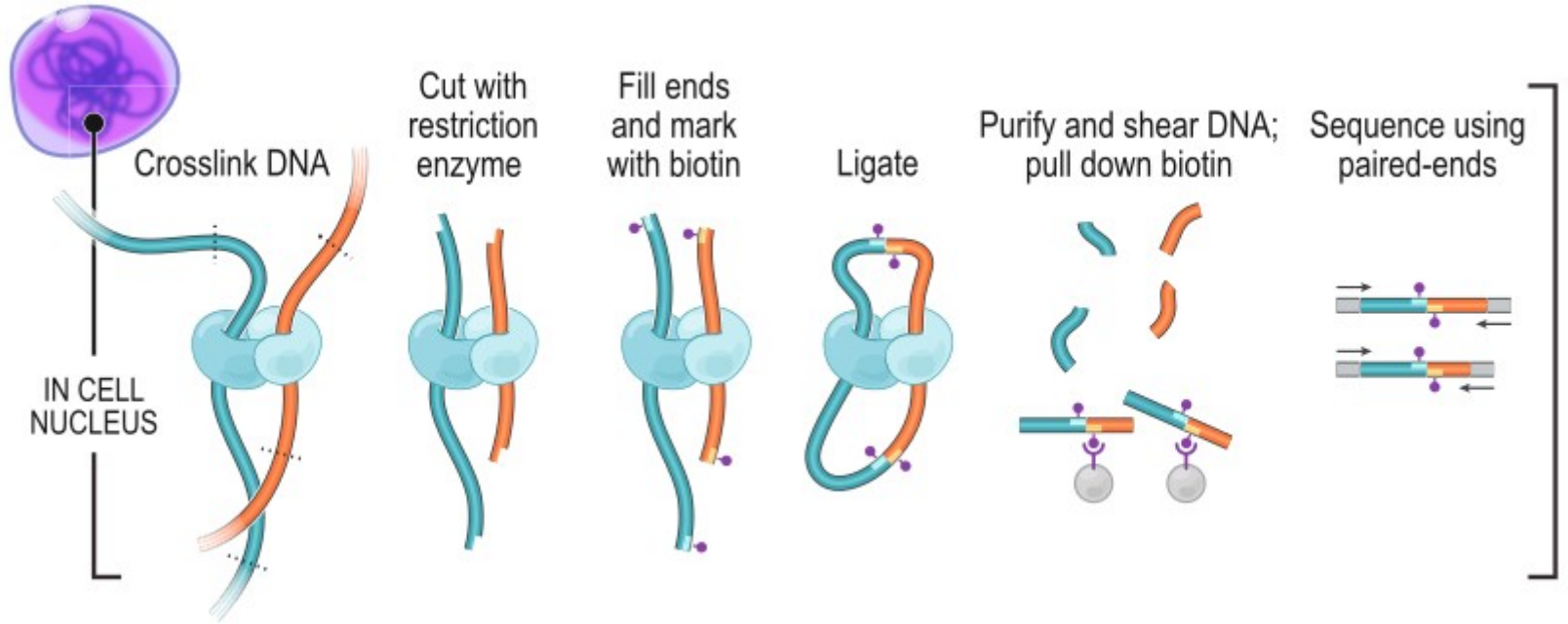
Hi-C: high-throughput chromatin conformation capture
(Lieberman-Aiden et al, Science, 2009, Rao et al, Cell, 2014)

- crosslink DNA ("fixation")
- cleave genome with restriction enzyme
- biotin-mark and ligate extremities
- fragment, select biotin-marked junctions
- sequence fragments (paired-ends)



Hi-C: the experiment

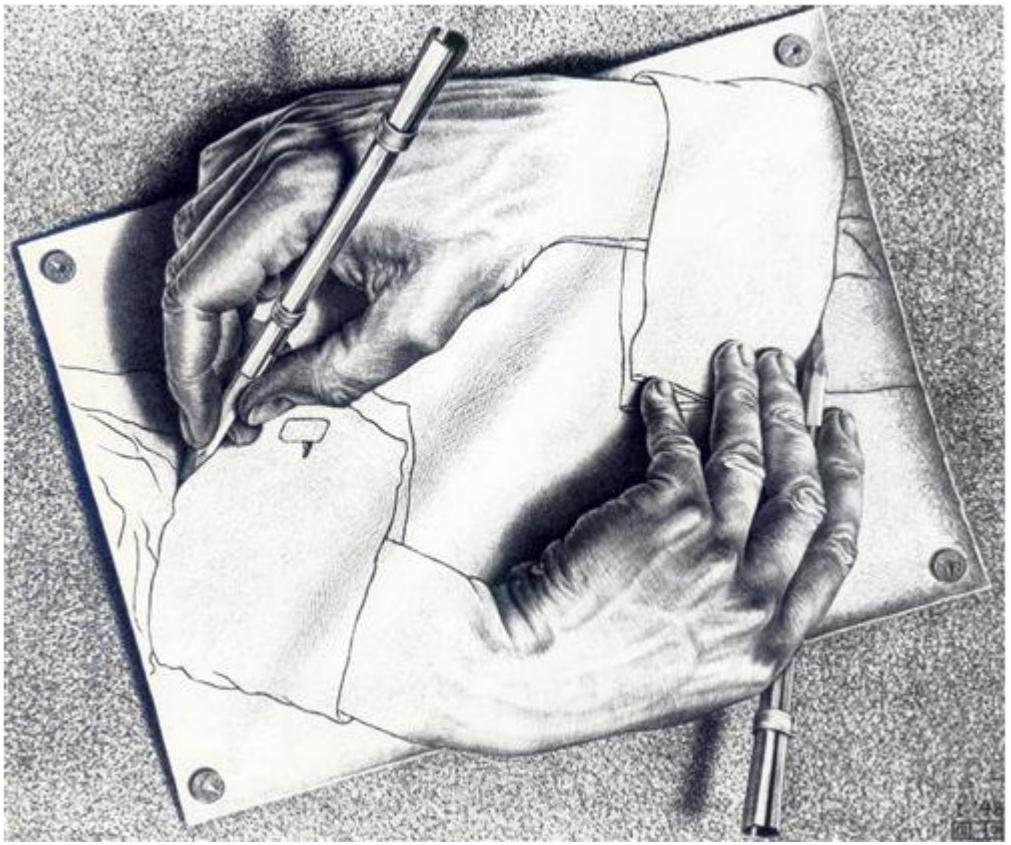
Hi-C: high-throughput chromatin conformation capture
(Lieberman-Aiden et al, Science, 2009, Rao et al, Cell, 2014)



Rao et al, Cell, 2014

Outline

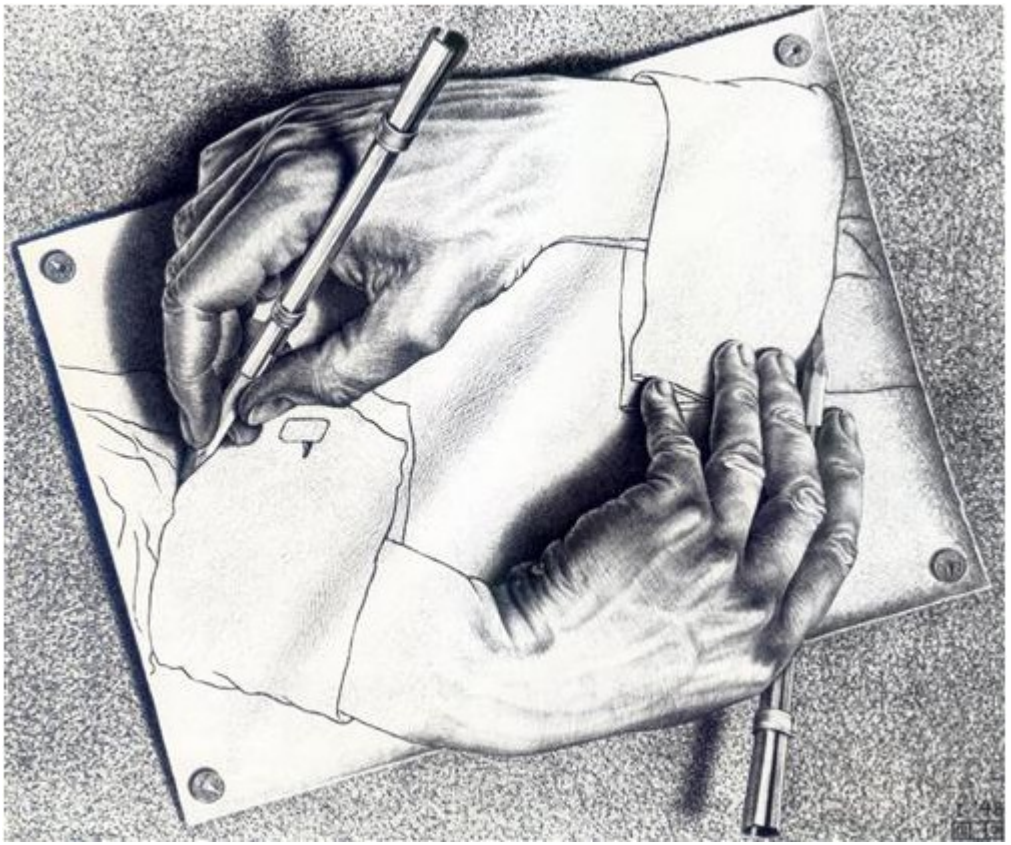
- Biological context
- More biological context
- Hi-C data processing
  - map
  - filter
  - count
  - normalize
  - segment
  - compare
- Conclusion, discussion, NETBIO lunch



M.C. Escher, 1948

Outline

- Biological context
- More biological context
- Hi-C data processing
  - map
  - filter
  - count
  - normalize
  - segment
  - compare
- Conclusion, discussion, NETBIO lunch

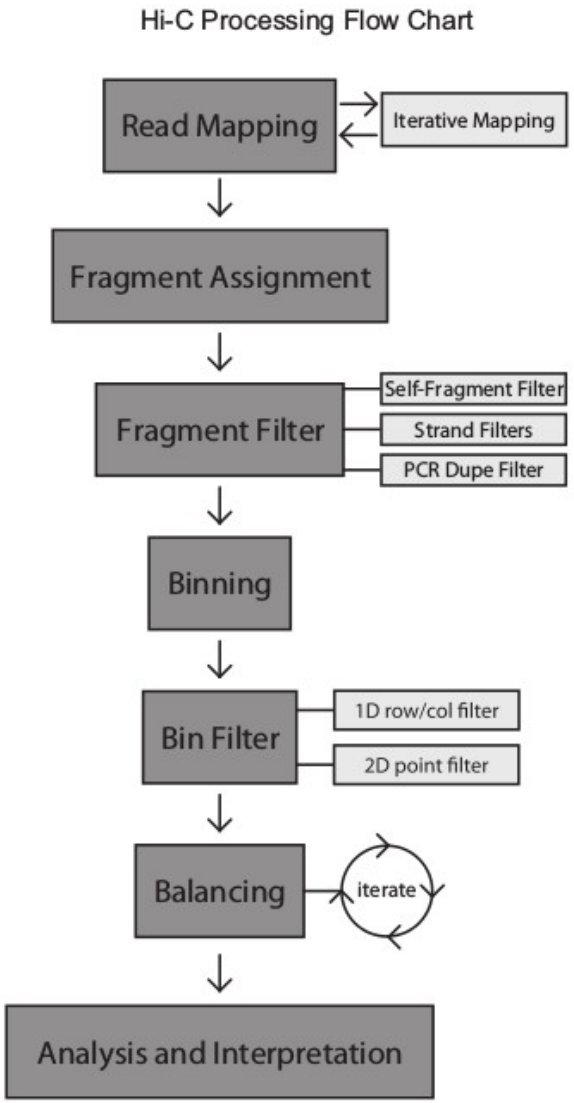


M.C. Escher, 1948

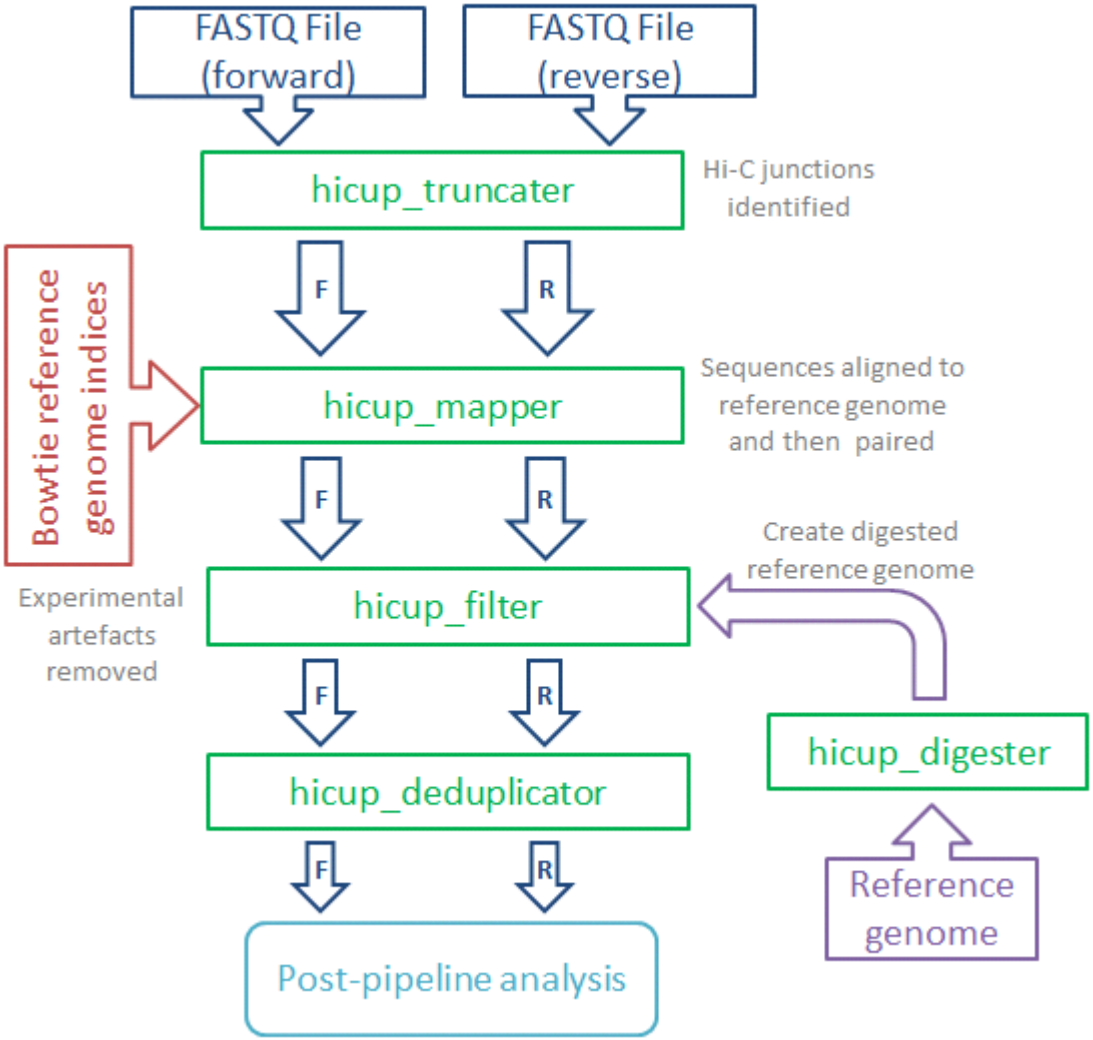
Hi-C data analysis: overview

- clean and trim the reads
- map the reads on the genomic reference
- filter bogus configurations
- count the reads per genomic bin => contact matrix
- normalize the matrix
- identify topological domains, cis- and trans- interactions
- comparative/integrative analysis

Hi-C data analysis: overview

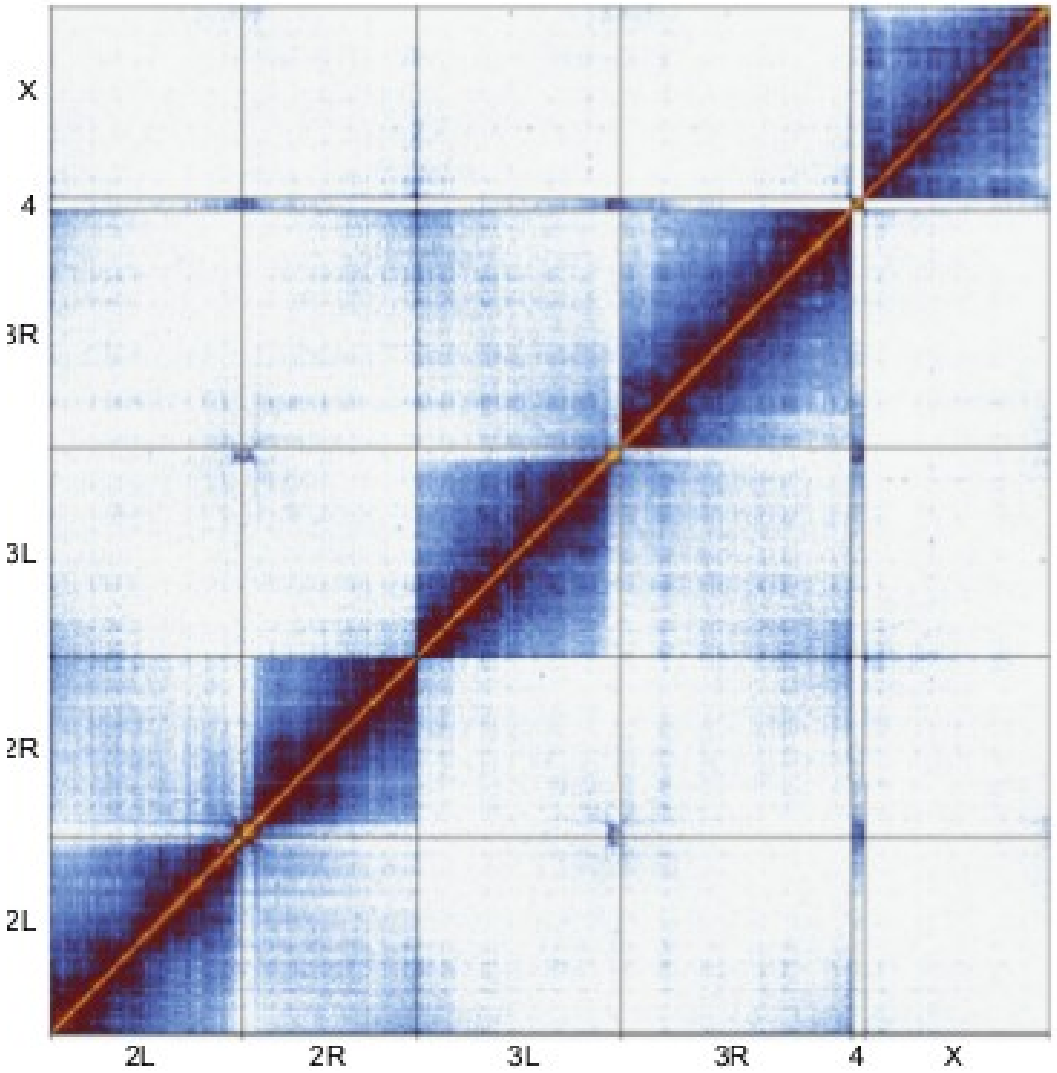


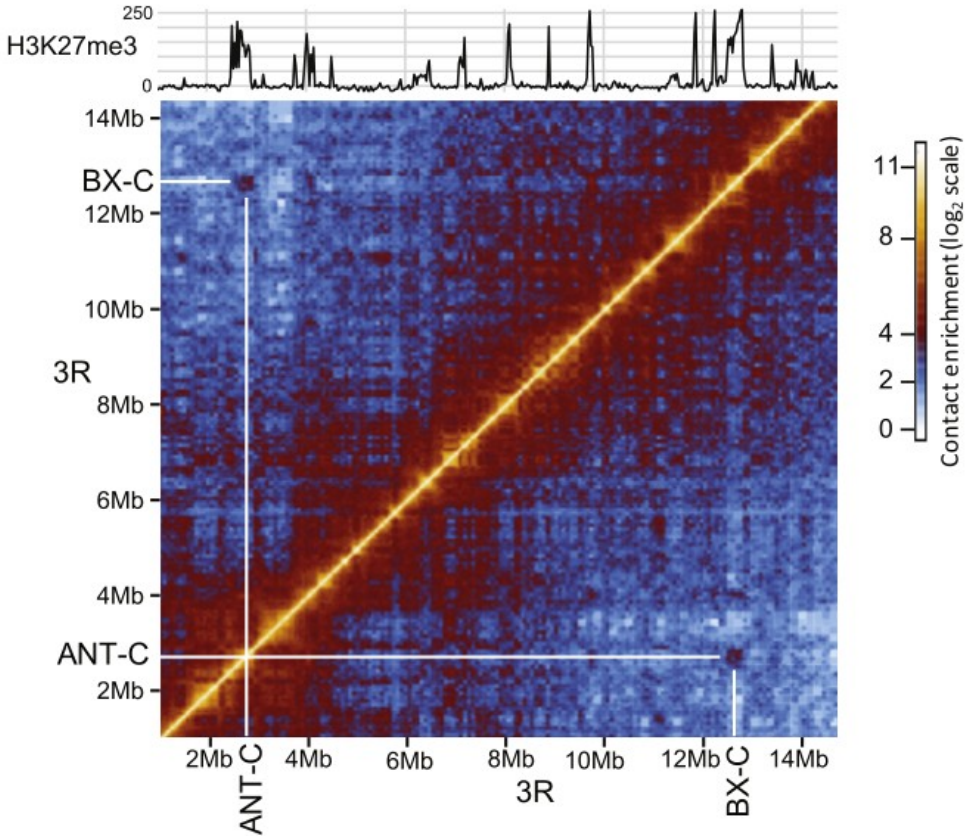
Lajoie et al, 2015



HiCUP, *www.bioinformatics.babraham.ac.uk/*

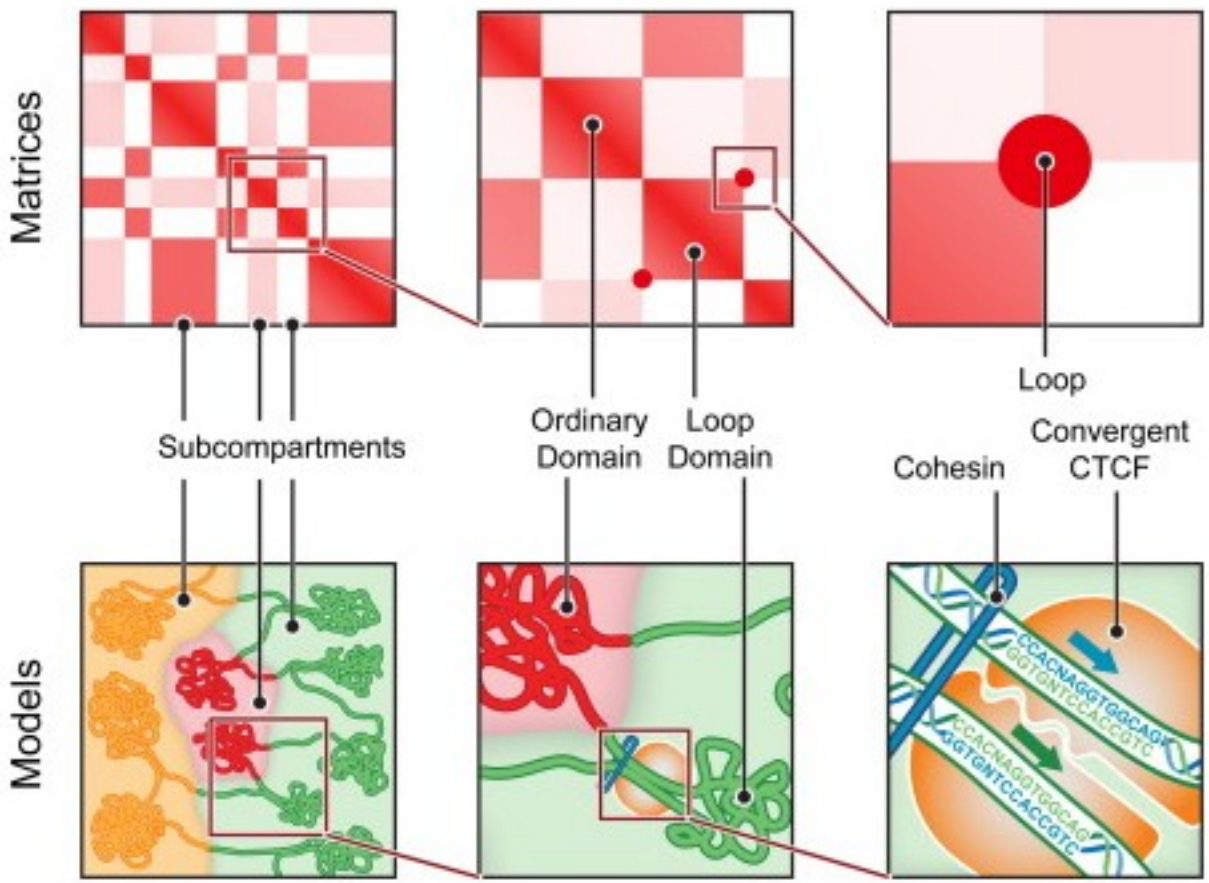
Hi-C data analysis: the contact matrix





Sexton et al 2012

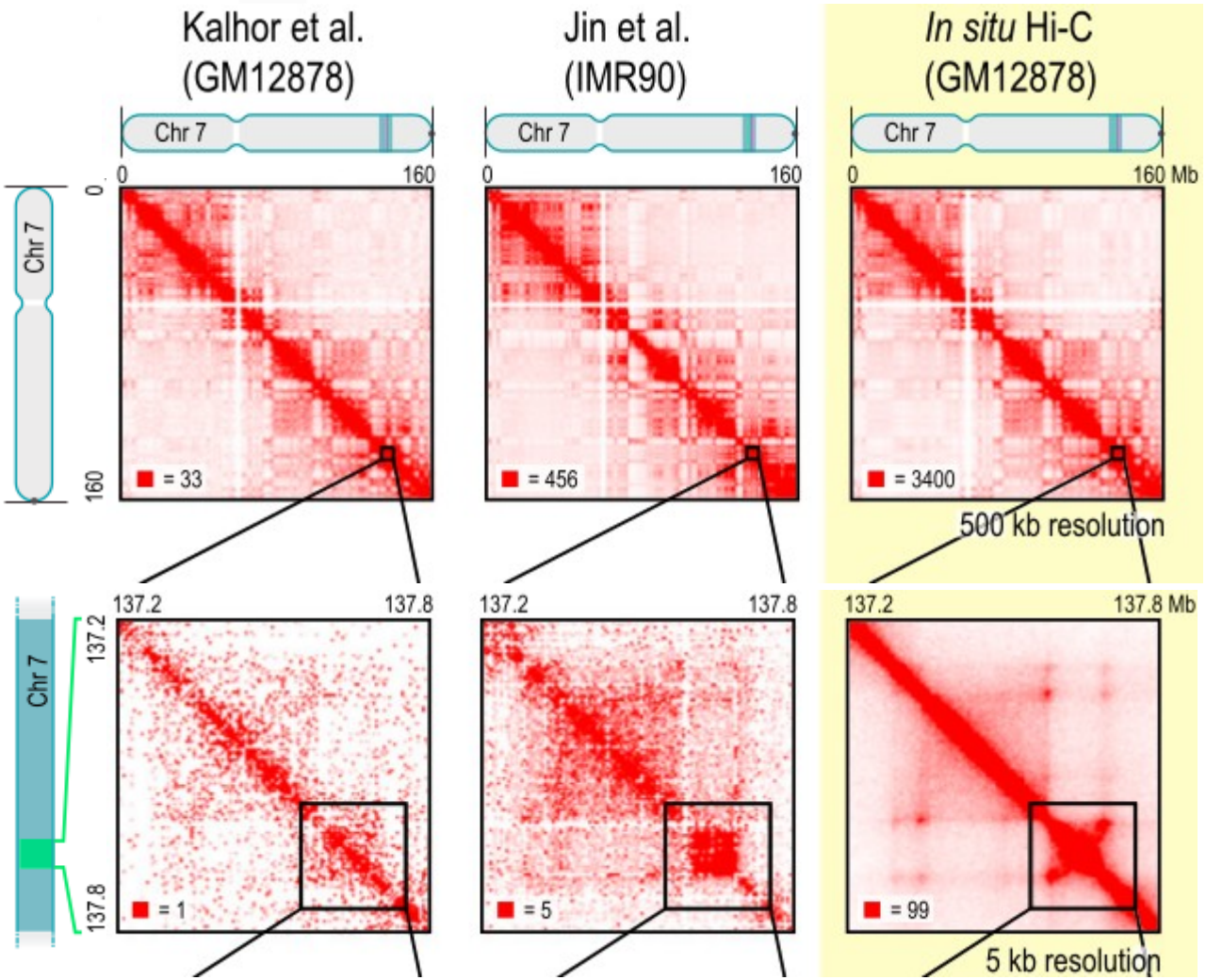
Hi-C data analysis: the contact matrix

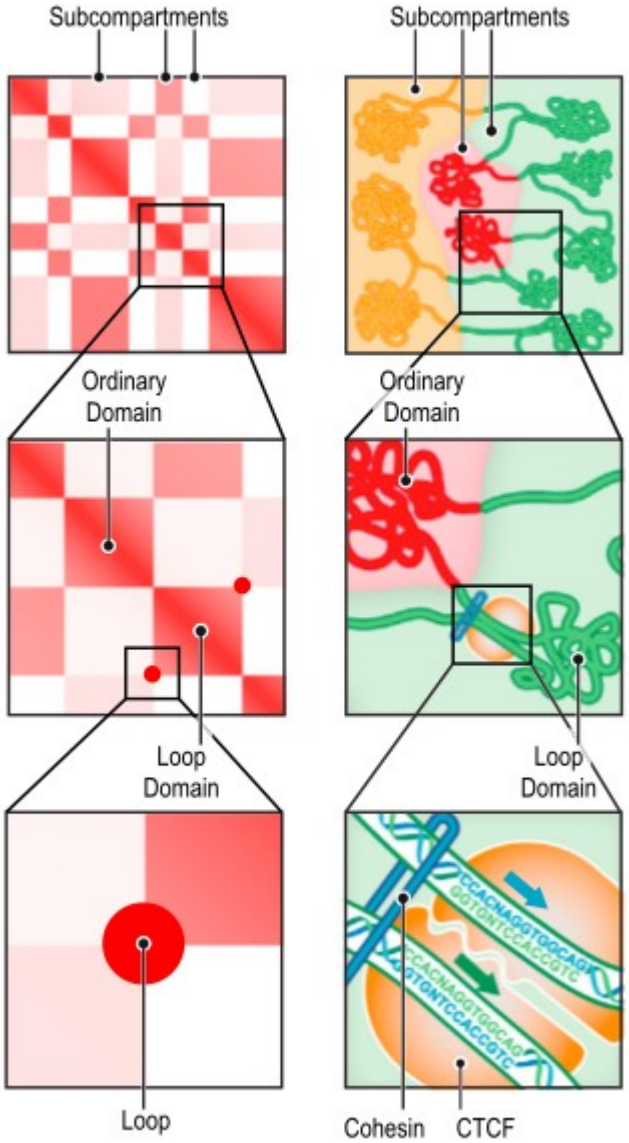


Rao et al, Cell, 2014

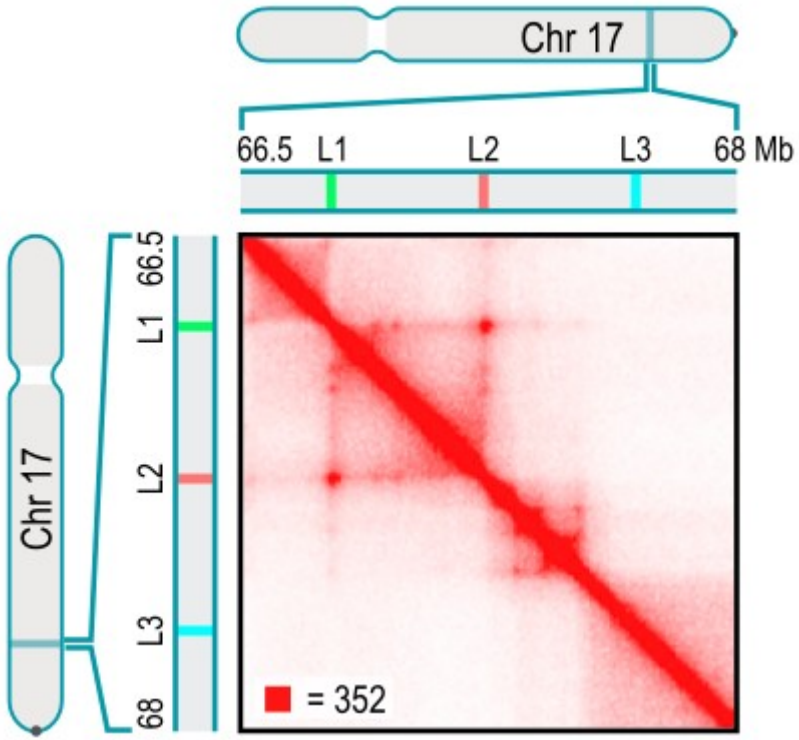
Hi-C data analysis: the contact matrix

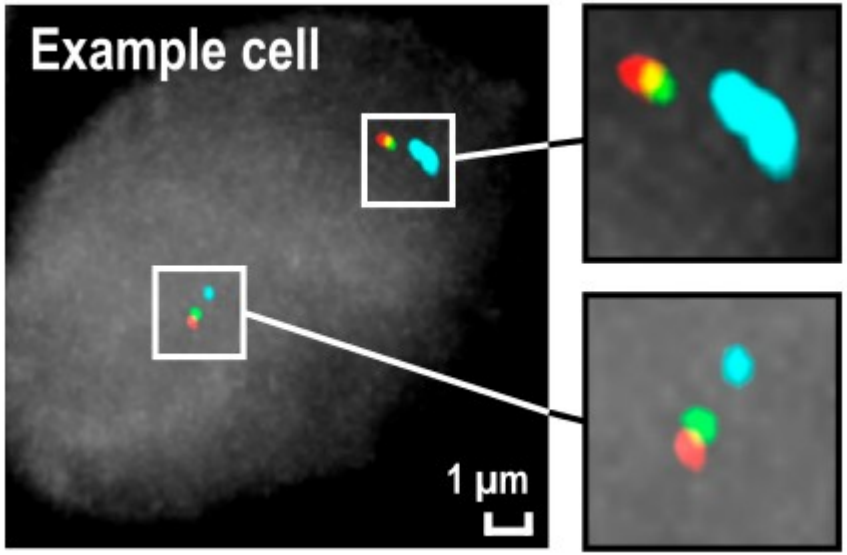
Rao et al, Cell, 2014





Hi-C data analysis: the contact matrix





Rao et al, Cell, 2014

Hi-C data analysis: matrix normalization

Number of reads per bin (coverage) depends on:

- GC%
- density of restriction sites
- repeats and “mappability”
- overall depth of coverage
- Others?

=> “Parametric” vs. “non-parametric” normalization

Hi-C data analysis: matrix normalization

Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture

Eitan Yaffe & Amos Tanay

Hi-C experiments measure the probability of physical proximity between pairs of chromosomal loci on a genomic

To fulfill this promise, 3C techniques and their derivations must become robust and quantitative. The complicated experimental

Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization

Maxim Imakaev[1,*], Geoffrey Fudenberg[2,*], Rachel Patton McC
Anton Goloborodko[1], Bryan R. Lajoie[3], Job Dekker[3,#], and Le
[1]Department of Physics, MIT, Cambridge, MA
[2]Graduate Program in Biophysics, Harvard University, Cambridge

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 28 no. 23 2012, pages 3131–3133*
doi:10.1093/bioinformatics/bts570

Genome analysis
Advance Access publication September 27, 2012

HiCNorm: removing biases in Hi-C data via Poisson regression

Ming Hu[1], Ke Deng[1], Siddarth Selvaraj[2,3], Zhaohui Qin[4], Bing Ren[2] and Jun S. Liu[1,*]
[1]Department of Statistics, Harvard University, Cambridge, MA 02138, USA, [2]Department of Cellular and Molecular Medicine, UCSD School of Medicine, La Jolla, CA 92093, USA, [3]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA and [4]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322 USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: We propose a parametric model, HiCNorm, to remove systematic biases in the raw Hi-C contact maps, resulting in a

biases, a non-parametric probabilistic model (referred to hereafter as the YT approach) was proposed that explicitly models the probability of observing a paired-end read spanning two

Category

HiCorrector: A fast, scalable and memory-efficient package normalizing large-scale Hi-C data

Wenyuan Li[1], Ke Gong[1], Qingjiao Li[1], Frank Alber[1*] and Xianghong Jasmine Zl
[1] Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, US

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Summary: Genome-wide proximity ligation assays, e.g. Hi-C and its variant TCC, have recently become important tools to study spatial genome organization. Removing biases from chromatin contact

and column sums of the matrix are equal to one. How C chromatin interaction matrix is of the massive size $O(N^2)$, where N is the number of genomic regions. Thus, it requires expensive computing resources such as large memory and long computation

Hi-C data analysis: matrix normalization

A FAST ALGORITHM FOR MATRIX BALANCING

PHILIP A. KNIGHT* AND DANIEL RUIZ†

Abstract. As long as a square nonnegative matrix A contains sufficient nonzero elements, then the matrix can be balanced, that is we can find a diagonal scaling of A that is doubly stochastic. A number of algorithms have been proposed to achieve the balancing, the most well known of these being Sinkhorn-Knopp. In this paper we derive new algorithms based on inner-outer iteration schemes. We show that Sinkhorn-Knopp belongs to this family, but other members can converge much more quickly. In particular, we show that while stationary iterative methods offer little or no improvement in many cases, a scheme using a preconditioned conjugate gradient method as the inner iteration can give quadratic convergence at low cost.

Key words. Matrix balancing, Sinkhorn-Knopp algorithm, doubly stochastic matrix, conjugate gradient iteration.

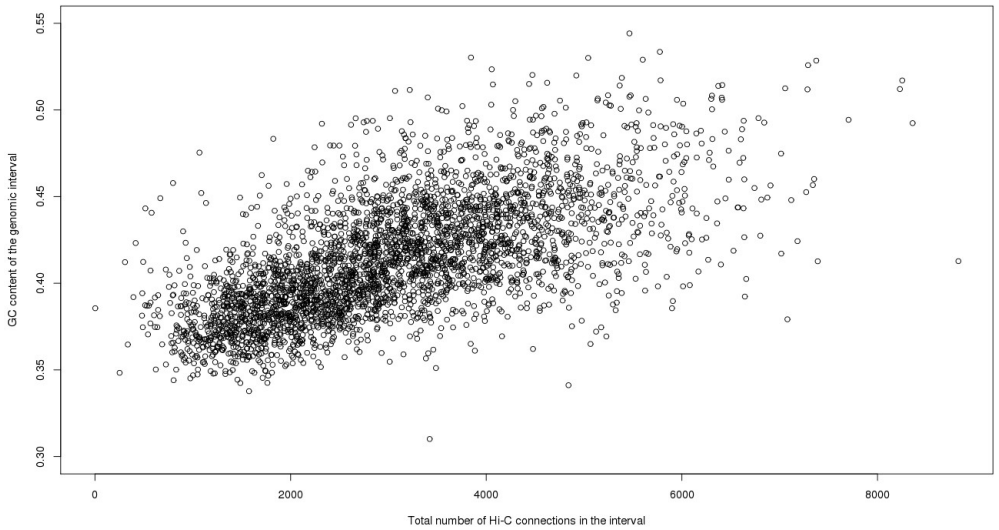
AMS subject classifications. 15A48, 15A51, 65F10, 65H10.

1. Introduction. For at least 70 years, scientists in a wide variety of disciplines have attempted to transform square nonnegative matrices into doubly stochastic form by applying diagonal scalings. That is, given $A \in \mathbb{R}^{n\times n}$, $A \geq 0$, find diagonal matrices D_1 and D_2 so that $P = D_1 A D_2$ is doubly stochastic. Motivations for achieving this balance include interpreting economic data [1], preconditioning sparse matrices [16], understanding traffic circulation [14], assigning seats fairly after elections [3], matching protein samples [4] and ordering nodes in a graph [12]. In all of these applications, one of the main methods considered is SK[1]. This is an iterative process that attempts to find D_1 and D_2 by alternately normalising columns

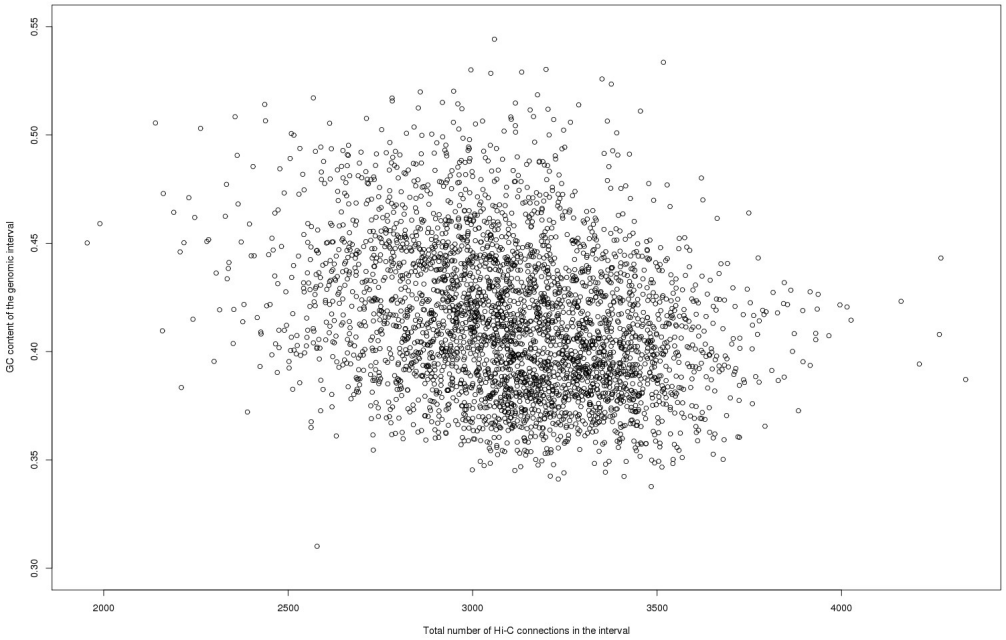
Knight & Ruiz, IMA J. Numer. Anal., 2013

Hi-C data analysis: matrix normalization

example with number of reads vs. GC%

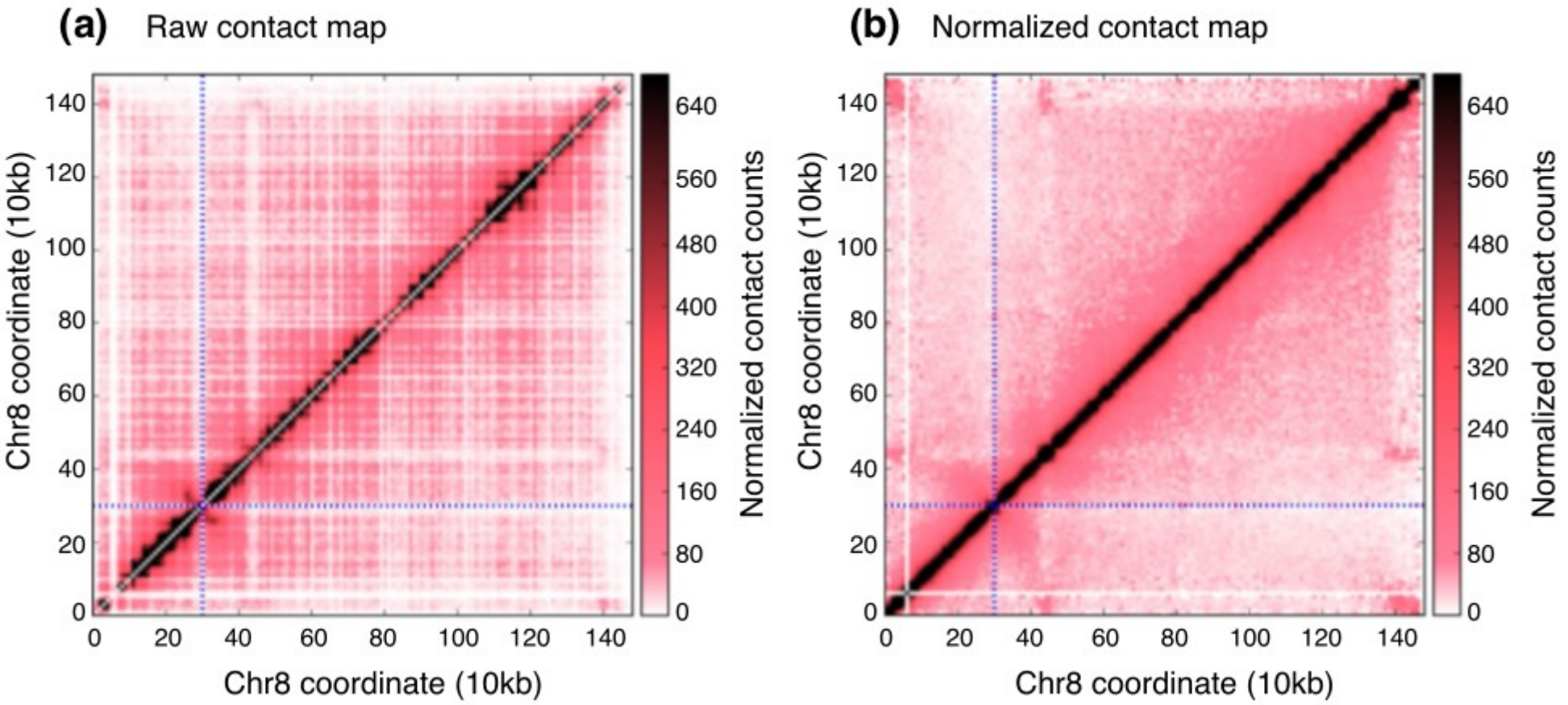


before normalization



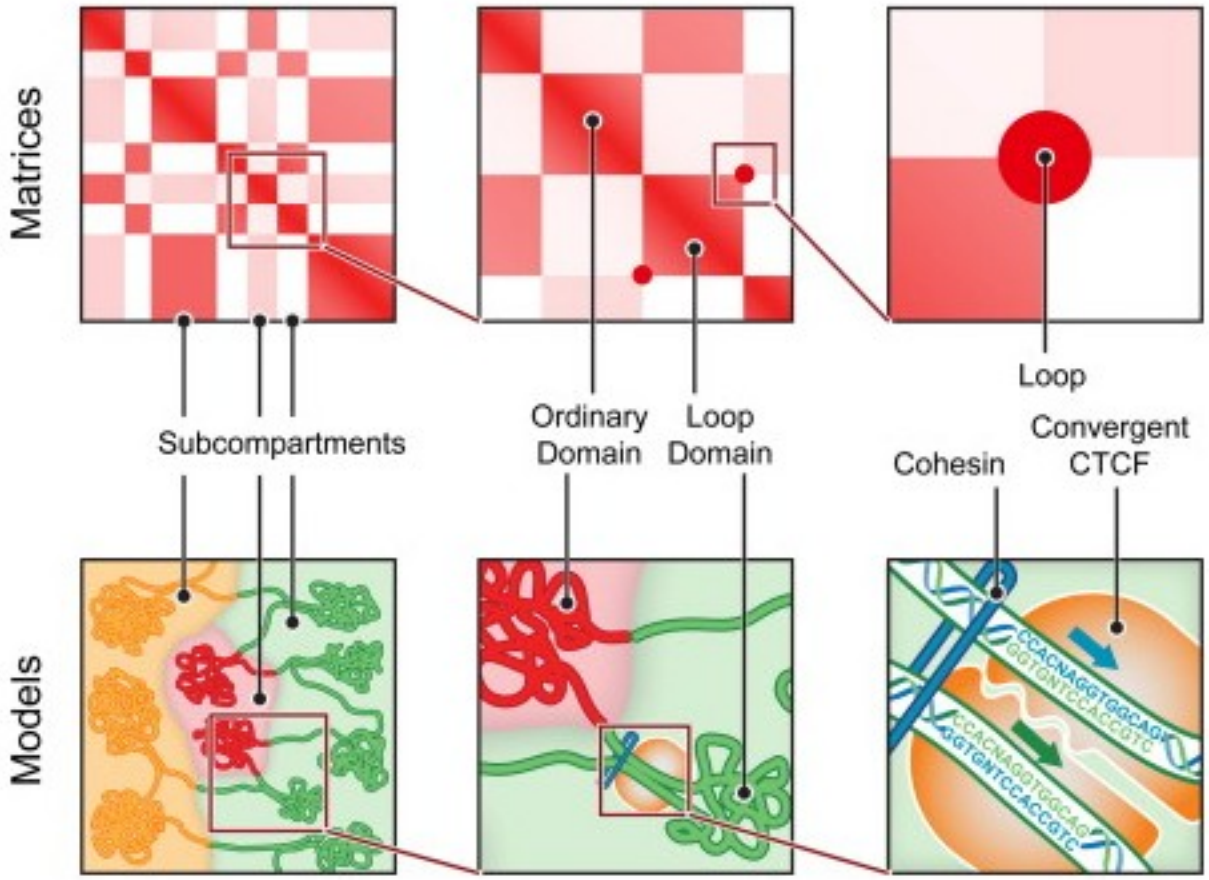
after normalization

Hi-C data analysis: matrix normalization



Ay & Noble, Genome Biology, 2015

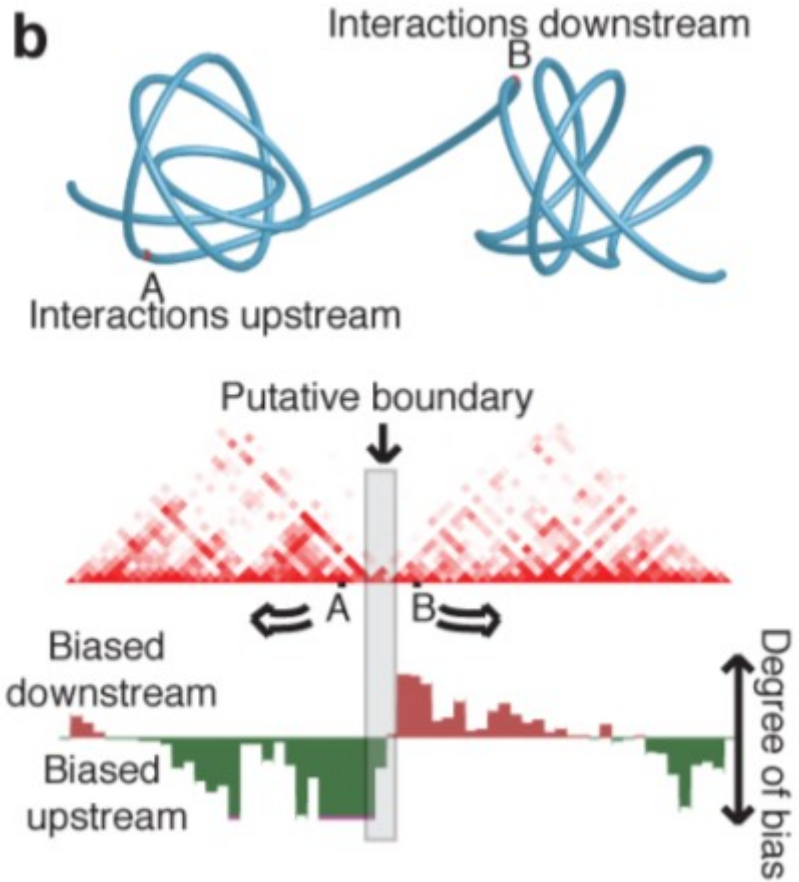
Hi-C data analysis: finding topologically associated domains (TADs)



Rao et al, Cell, 2014

- methods: clustering, 2D-segmentation, etc

Hi-C data analysis: finding topologically associated domains (TADs)



Dixon et al., Nature, 2012

Published in final edited form as:
Nature. ; 485(7398): 376–380. doi:10.1038/nature11082.

Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions

Jesse R. Dixon[1,3,4], **Siddarth Selvaraj**[1,5], **Feng Yue**[1], **Audrey Kim**[1], **Yan Li**[1], **Yin Shen**[1], **Ming Hu**[6], **Jun S. Liu**[6], and **Bing Ren**[1,2,*]

[1]Ludwig Institute for Cancer Research

[2]University of California, San Diego School of Medicine, Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, 9500 Gilman Drive, La Jolla, CA 92093

[3]Medical Scientist Training Program, University of California, San Diego, La Jolla CA 92093

[4]Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla CA 92093

[5]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla CA 92093

[6]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138

Abstract

The spatial organization of the genome is intimately linked to its biological function, yet our understanding of higher order genomic structure is coarse, fragmented and incomplete. In the nucleus of eukaryotic cells, interphase chromosomes occupy distinct chromosome territories (CT),

Directionality Index

$$DI = \left(\frac{B - A}{|B - A|}\right)\left(\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}\right)$$

DI HMM => TADs

Hi-C data analysis: finding topologically associated domains (TADs)

Identification of hierarchical chromatin domains

Caleb Weinreb [1], and Benjamin J. Raphael [1,2*]

[1] Center for Computational Molecular Biology, Brown University, Providence, RI
[2] Department of Computer Science, Brown University, Providence, RI

Associate Editor: Prof. Gunnar Ratsch

ABSTRACT

Motivation: The 3D structure of the genome is an important regulator of many cellular processes including differentiation and gene regulation. Recently, technologies such as Hi-C that combine proximity ligation with high-throughput sequencing have revealed domains of self-interacting chromatin, called topologically associating domains (TADs), in many organisms. Current methods for identifying TADs using Hi-C data assume that TADs are non-overlapping, despite evidence for a nested structure in which TADs and sub-TADs form a complex hierarchy.

Results: We introduce a model for hierarchical decomposition of contact frequencies into a hierarchy of nested TADs. This model is based empirical distributions of contact frequencies within TADs, where positions that are far apart have a greater enrichment of contacts than

resulting in a contact matrix A, where A_{ij} is the number of contacts between bins i and j, normalized for experimental bias. Several methods have been developed for the identification of TADs from Hi-C data. These methods may be roughly classified into two categories: (1) methods that define a 1D test statistic from the contact matrix A_{ij}; (2) methods that exploit the 2D structure of the contact matrix.

Dixon *et al.* (2012) compute a 1D "directionality index" (DI) from the contact matrix. This index defines whether contacts have an upstream bias, downstream bias or no bias. Next, they use a hidden Markov model (HMM) to partition the genome into regions defined by changes in the directionality index. Each transition into downstream bias marks the start of a domain and the next transition out of upstream bias marks its end. Sauria *et al.* (2014) introduce a 1D

defined below.

DEFINITION 2. *Consider a TAD D and interval $[i,j] \subseteq [D_L, D_R]$. Let the* error compensation $\mathcal{E}_C(i,j,D)$ *be*

$$\mathcal{E}_C(i,j,D) = \sum_{l=i}^{j}\sum_{k=l}^{j}\left(\tilde{A}_D(l,k) - A_{lk}\right)^2. \quad (8)$$

Using the error compensation, we derive an expression for the score of a TAD tree in terms of its root TAD and sub-trees.

PROPOSITION 1. *Let T be a TAD tree consisting of a root TAD D and a collection of non-overlapping sub-trees $T_1, \ldots, T_m$, spanning the intervals $[i_1, j_1], \ldots, [i_m, j_m]$. The score $\mathcal{O}_\gamma(T)$ can be decomposed as*

$$\mathcal{O}_\gamma(T) = \mathcal{O}_\gamma(D) + \sum_{x=1}^{m}\left(\mathcal{O}_\gamma(T_x) + \mathcal{E}_C(i_x, j_x, D)\right). \quad (9)$$

We now describe steps (1-3) above in greater detail. To perform step (1), recall that a TAD is defined by four parameters, $(L_D, R_D, \delta_D, \beta_D)$. Thus, in choosing the root TAD D, two parameters are given ahead of time ($[L_D, R_D] = [i,j]$), meaning we only need to select optimal values for δ_D and β_D. Next, for a given choice of δ_D and β_D, we must choose a set of non-overlapping sub-trees, defined by sub-intervals $[i_x, j_x]$ and multiplicites n_x (steps 2-3). To that end, let $\mathcal{I}(i,j,N)$ be the collection of sets $\{(i_x, j_x, n_x)\}$ that satisfy the following properties: (i) $[i_x, j_x]$ are non-overlapping sub-intervals of $[i,j]$; (ii) $\sum n_x = N - 1$; (iii) i_x and j_x are valid boundaries. Using $\mathcal{I}(i,j,N)$ as a search space, we evaluate $\Phi(i,j,N,\delta)$ as follows.

PROPOSITION 2. *For each interval $[i,j]$ and positive integer N,*

$$\Phi(i,j,N,\delta) = \max \left(\mathcal{O}_\gamma(D) + \max \left(\sum \mathcal{W}_x\right)\right)$$

PROBLEM 2. *Given $N \in \mathbb{N}$ and $\gamma \in \mathbb{R}^+$, find the TAD forest F that maximizes the objective $\mathcal{O}_\gamma(F) = \gamma\bar{\mathcal{B}}_{p,q}(F) - \mathcal{E}(F)$ such that $|F| = N$, and each $D \in F$ is locally fitted and has valid boundaries.*

Once again, our first step in solving Problem 2 will be to find optimal TAD trees over every interval.

DEFINITION 4. *Given $N \in \mathbb{N}$, and the interval $[i,j]$, define $\widehat{\Phi}(i,j,N) := \max \mathcal{O}_\gamma(T)$ over all TAD trees T such that (i) T is rooted at the interval $[i,j]$, (ii) T contains N TADs ($|T| = N$), and (iii) each $D \in T$ is locally fitted has valid boundaries.*

In contrast to $\Phi(i,j,N,\delta)$, $\widehat{\Phi}(i,j,N)$ does not take δ as an argument, since it maximizes over TAD trees whose δ values are fixed by the requirement that they be locally fitted. This leads to the following proposition, which shows how to evaluate $\widehat{\Phi}(i,j,N)$.

PROPOSITION 3. *For each interval $[i,j]$ and positive integer N,*

$$\widehat{\Phi}(i,j,N) = \mathcal{O}_\gamma(\hat{D}_{ij}) + \max_{\{(i_x,j_x,n_x)\}\in\mathcal{I}(i,j,N)}\left(\sum_x \mathcal{W}_x\right) \quad (12)$$

where

$$\mathcal{W}_x = \begin{cases} \widehat{\Phi}(i_x, j_x, n_x) + \mathcal{E}_C(i_x, j_x, \hat{D}_{ij}) & \text{if } \hat{\delta}(i_x, j_x) \geq \hat{\delta}(i,j) \\ -\infty & \text{otherwise.} \end{cases}$$

2.3 Algorithm

To evaluate equation (12), we must choose a set of non-overlapping intervals $[i_x, j_x]$ and multiplicities n_x that maximize $\sum_x \mathcal{W}_x$ and satisfy $\sum n_x = N - 1$. Similarly, to assemble a TAD forest from TAD trees, we will likewise be choosing a non-overlapping set of intervals (leaves of TAD trees) with multiplicities (number of TADs in each tree) such that the sum of their scores is maximized and the multiplicites sum to a predefined N. These tasks are both similar to the weighed interval scheduling problem (Kleinberg

Weireb & Raphael, Bioinformatics, 2015

Hi-C data analysis: finding topologically associated domains (TADs)

Identification of hierarchical chromatin domains

Caleb Weinreb [1], and Benjamin J. Raphael [1,2*]

[1]Center for Computational Molecular Biology, Brown University, Providence, RI
[2]Department of Computer Science, Brown University, Providence, RI

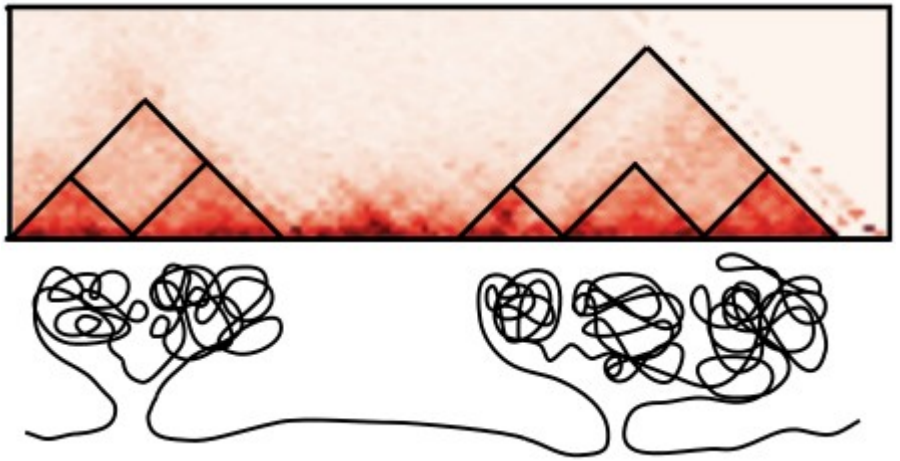
Associate Editor: Prof. Gunnar Ratsch

ABSTRACT

Motivation: The 3D structure of the genome is an important regulator of many cellular processes including differentiation and gene regulation. Recently, technologies such as Hi-C that combine proximity ligation with high-throughput sequencing have revealed domains of self-interacting chromatin, called topologically associating domains (TADs), in many organisms. Current methods for identifying TADs using Hi-C data assume that TADs are non-overlapping, despite evidence for a nested structure in which TADs and sub-TADs form a complex hierarchy.

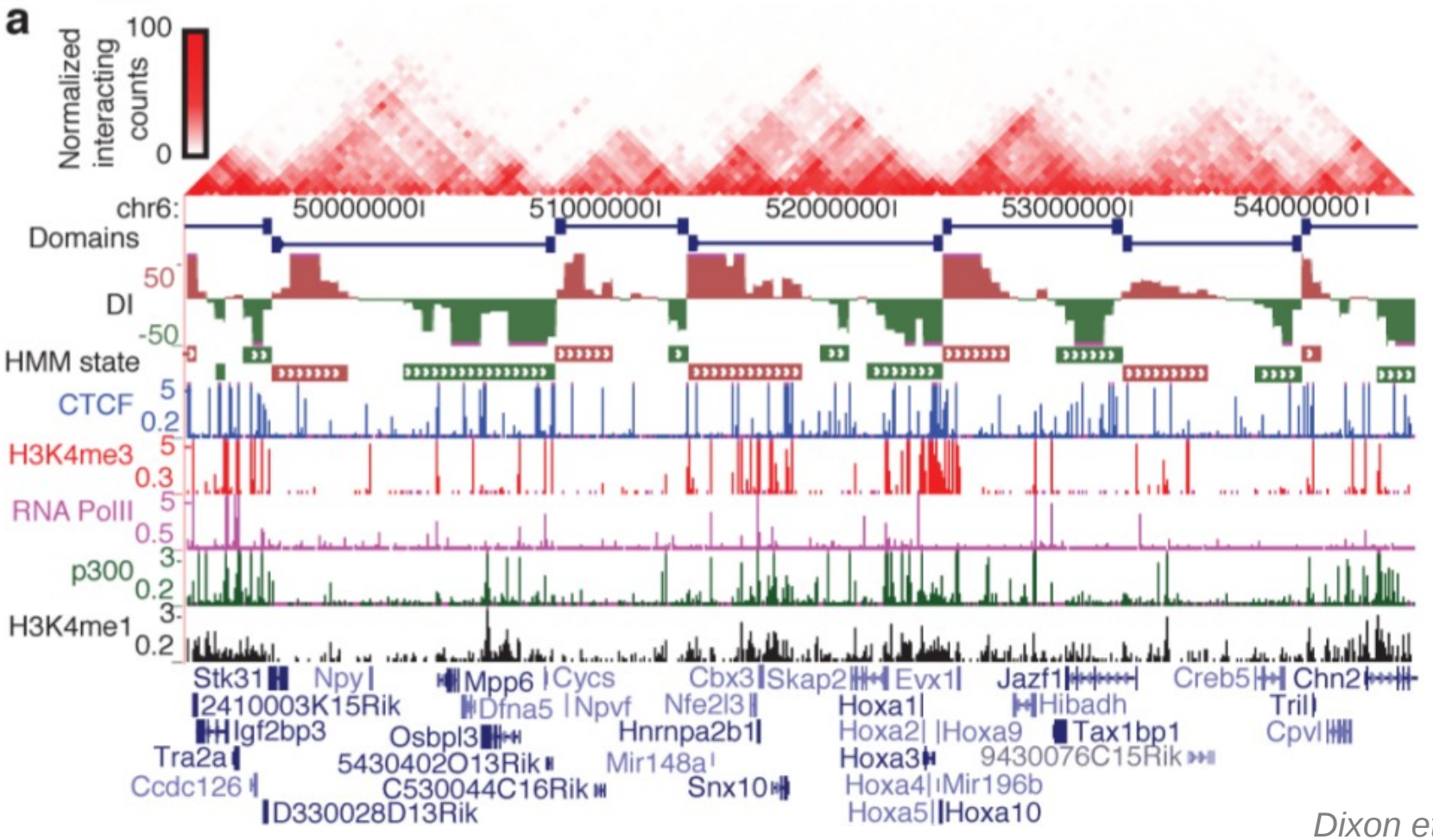
Results: We introduce a model for hierarchical decomposition of contact frequencies into a hierarchy of nested TADs. This model is based empirical distributions of contact frequencies within TADs, where positions that are far apart have a greater enrichment of contacts than resulting in a contact matrix A, where A_{ij} is the number of contacts between bins i and j, normalized for experimental bias. Several methods have been developed for the identification of TADs from Hi-C data. These methods may be roughly classified into two categories: (1) methods that define a 1D test statistic from the contact matrix A_{ij}; (2) methods that exploit the 2D structure of the contact matrix.

Dixon *et al.* (2012) compute a 1D "directionality index" (DI) from the contact matrix. This index defines whether contacts have an upstream bias, downstream bias or no bias. Next, they use a hidden Markov model (HMM) to partition the genome into regions defined by changes in the directionality index. Each transition into downstream bias marks the start of a domain and the next transition out of upstream bias marks its end. Sauria *et al.* (2014) introduce a 1D



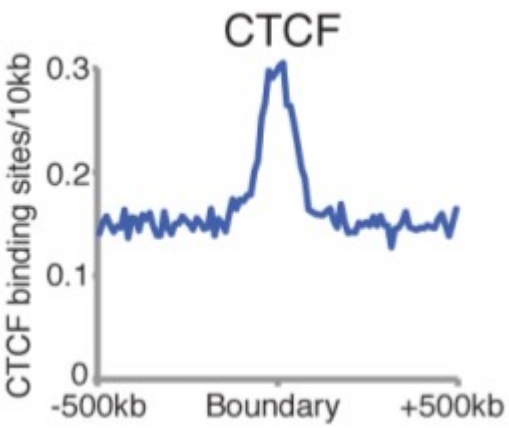
Weireb & Raphael, Bioinformatics, 2015

Hi-C data analysis: integrative analysis

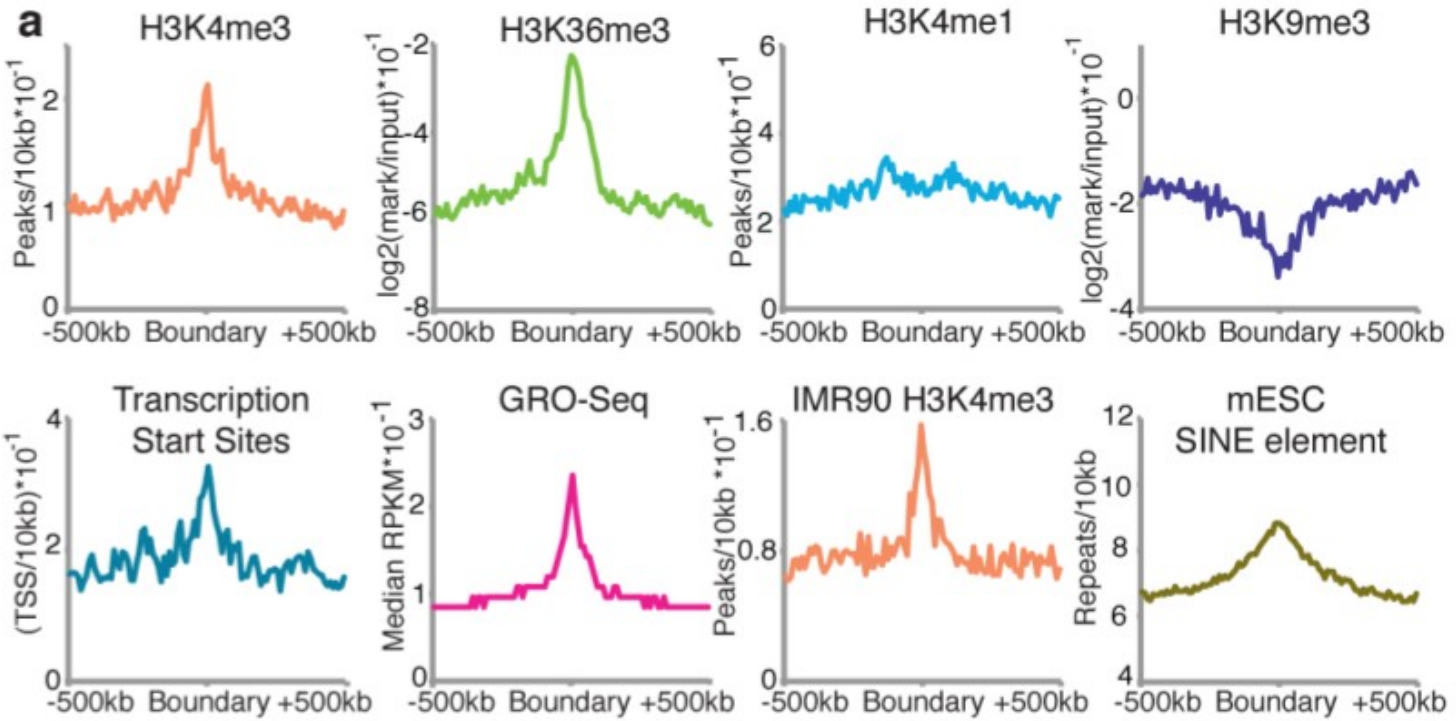


Dixon et al 2012

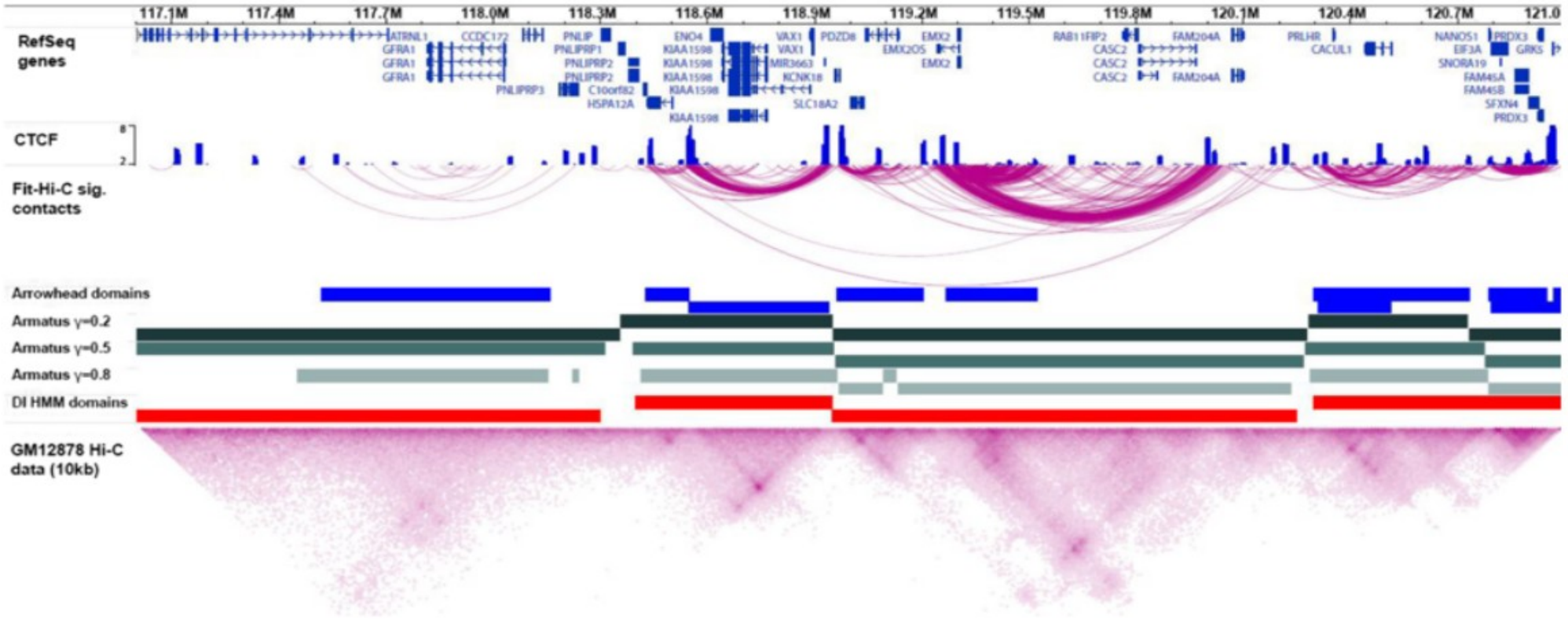
Hi-C data analysis: comparisons



Dixon et al 2012

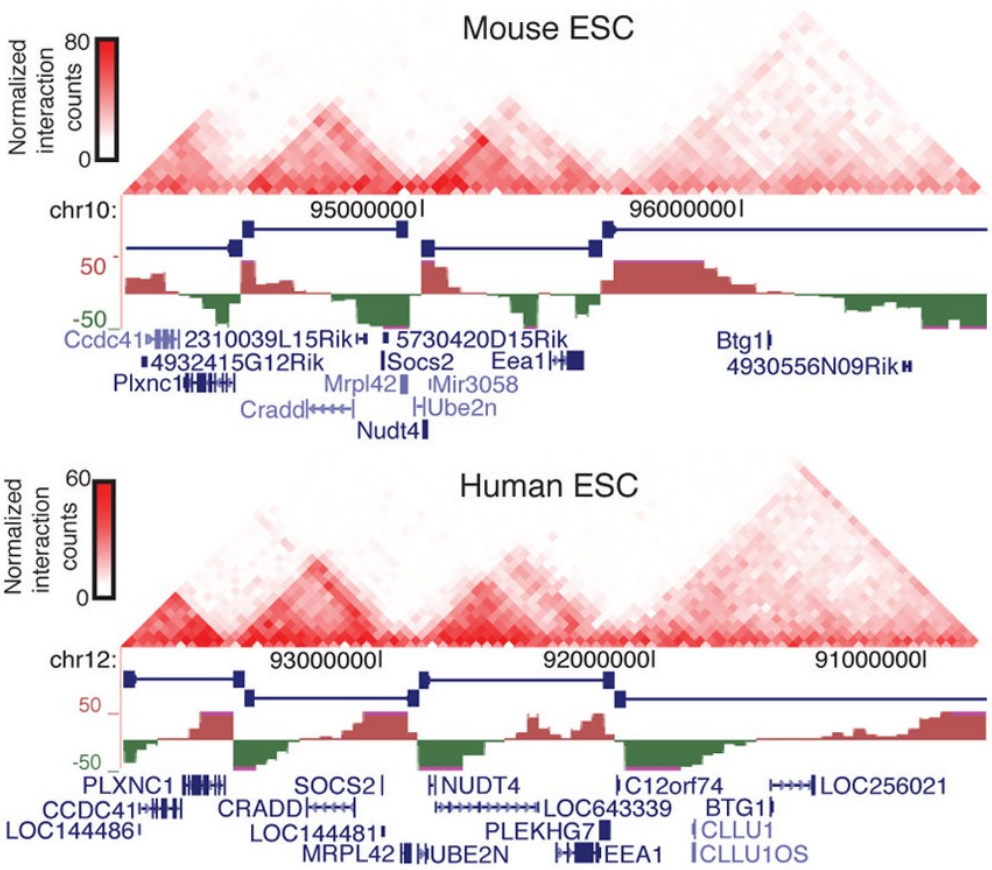


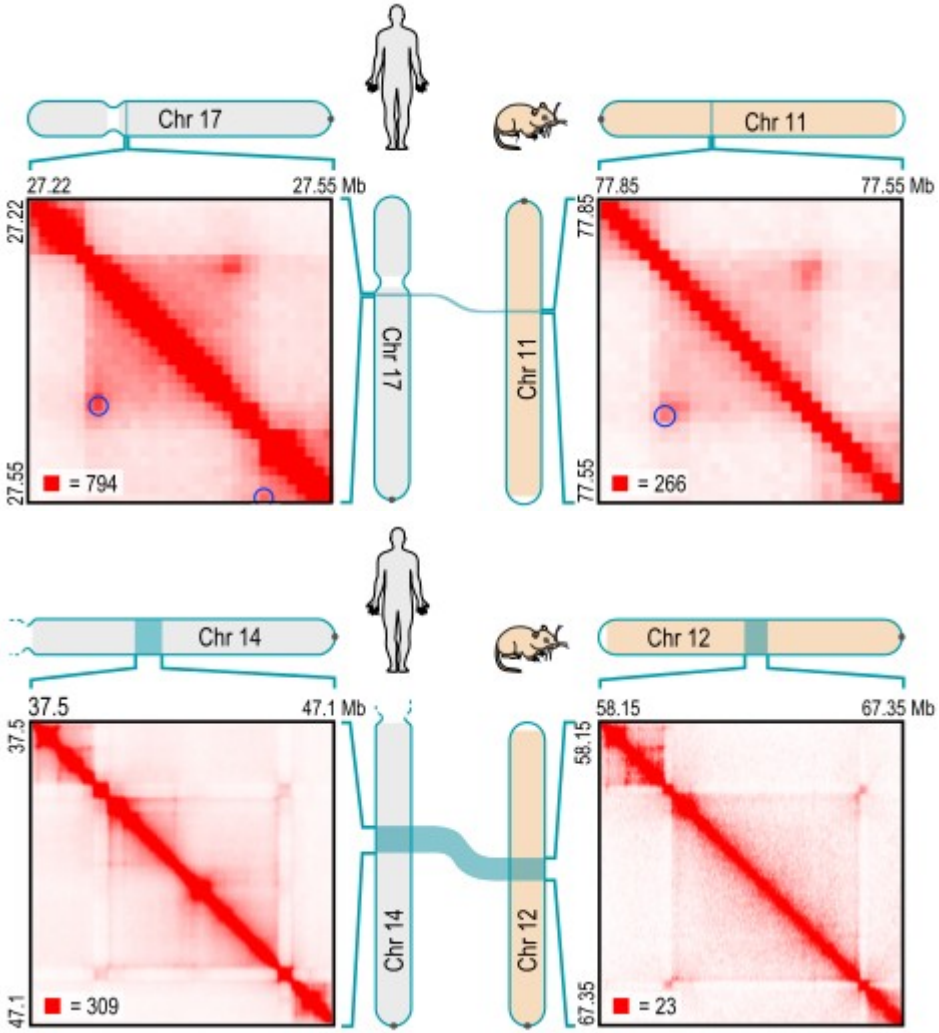
Hi-C data analysis: integrative analysis



Ay & Noble, Genome Biology, 2015

Hi-C data analysis: integrative analysis





FR-AgENCODE: livestock genome annotation

2x♂
2x♀

Sus scrofa
(Large White)

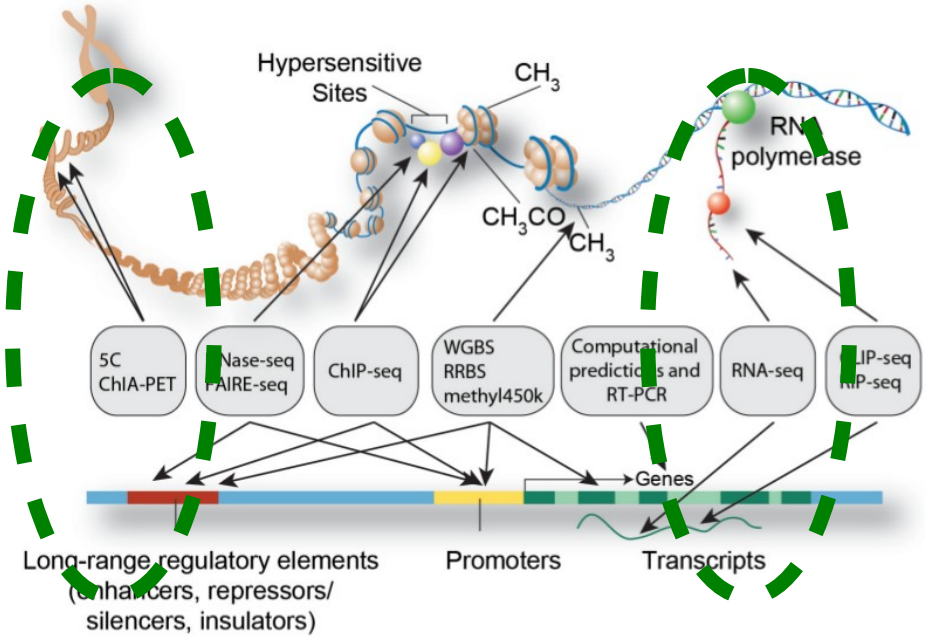
Gallus gallus
(White Leghorn)

Bos Taurus
(Holstein)

Capra hircus
(Alpine)

- Sampling: 34 somatic tissues + 13 reproductive tissues
  => INRA CRB-Anim biorepository

- Molecular assays
  - RNA-seq
  - Hi-C
- Data analysis

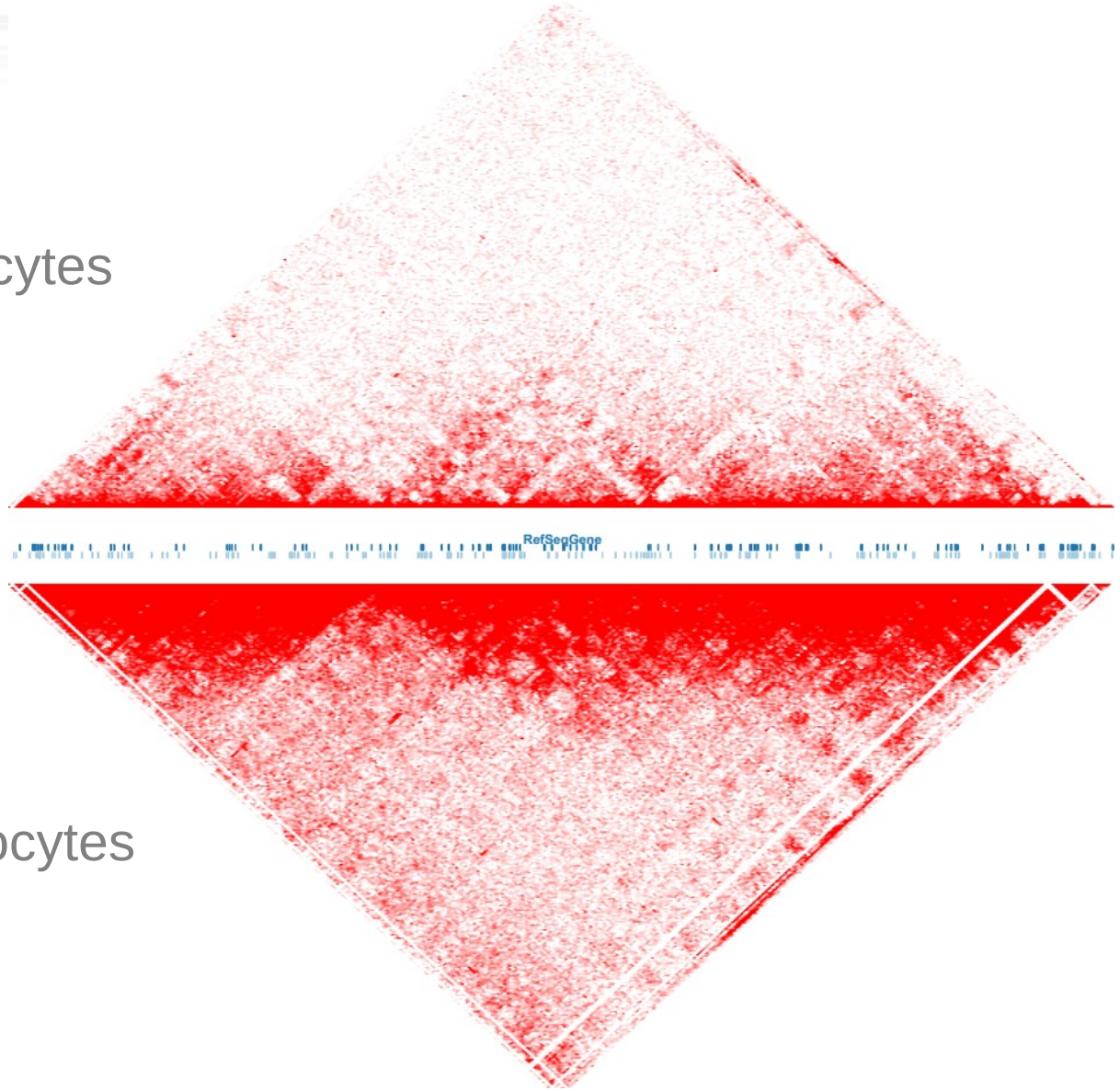


Modified from PLoS Biol 9-e1001046, 2011 & Science 306:636, 2004
Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Check out:
www.faang.org

FR-AgENCODE: livestock genome annotation

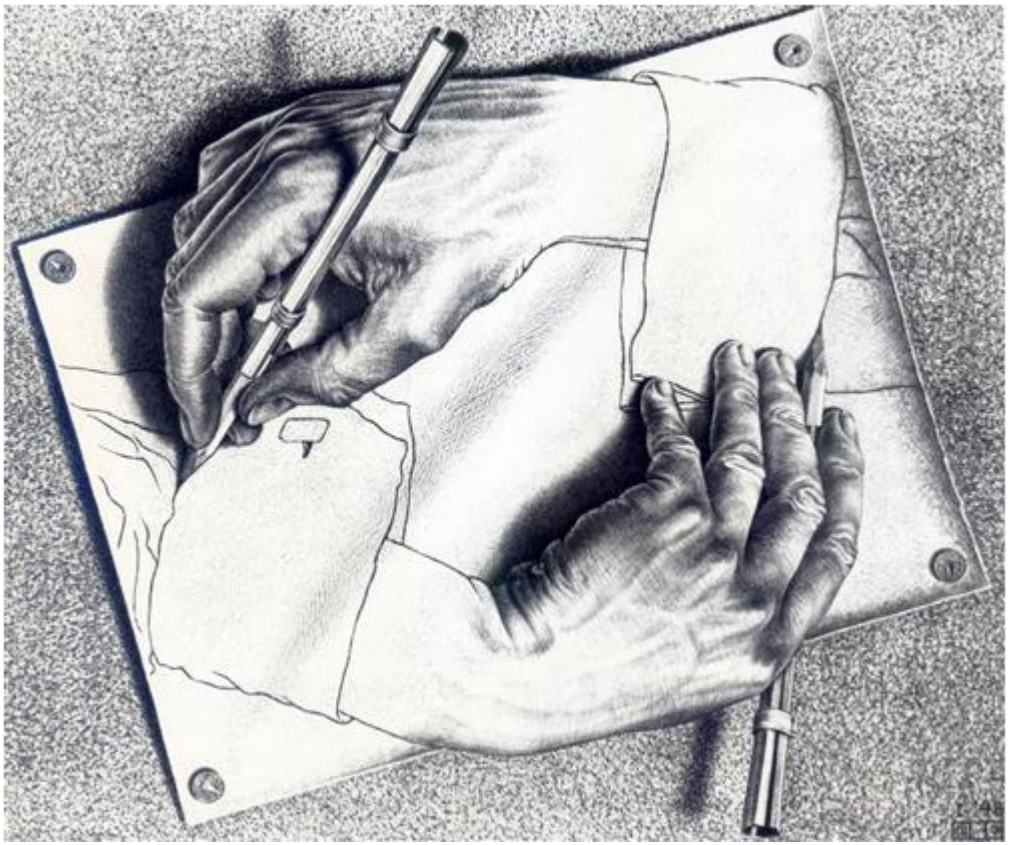
Pig hepatocytes



Pig lymphocytes

Outline

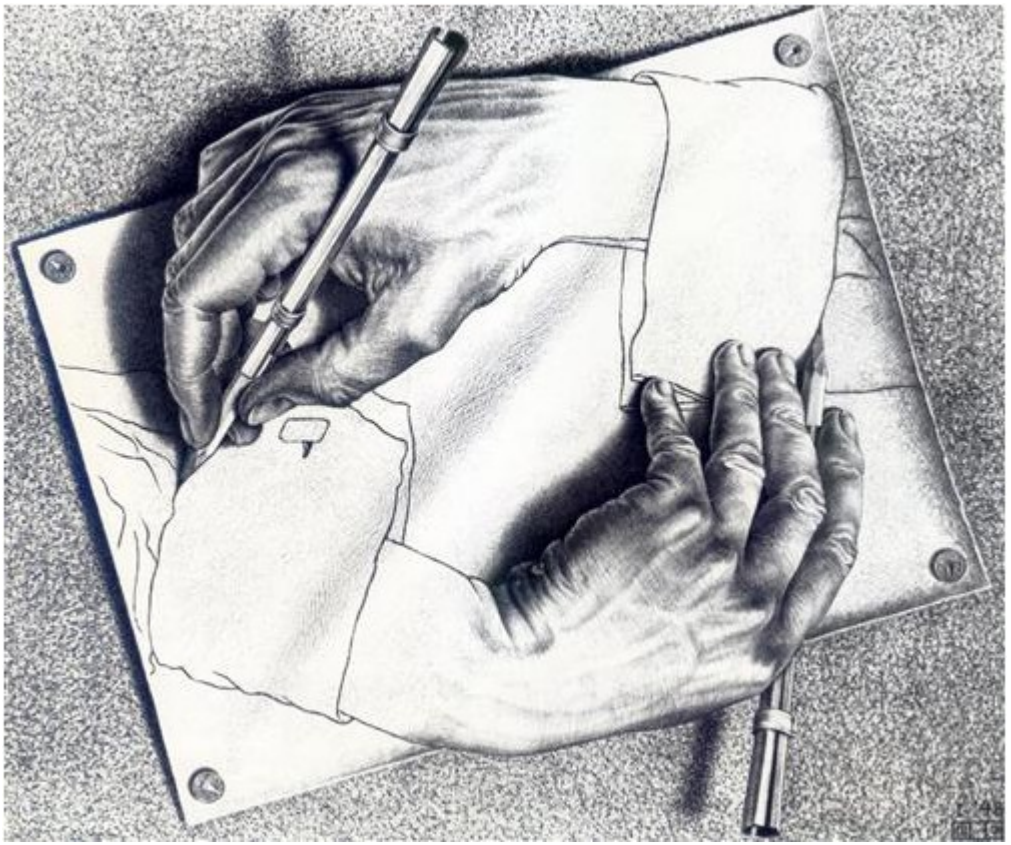
- Biological context
- More biological context
- Hi-C data processing
  - map
  - filter
  - count
  - normalize
  - segment
  - compare
- Conclusion, discussion, NETBIO lunch



M.C. Escher, 1948

Outline

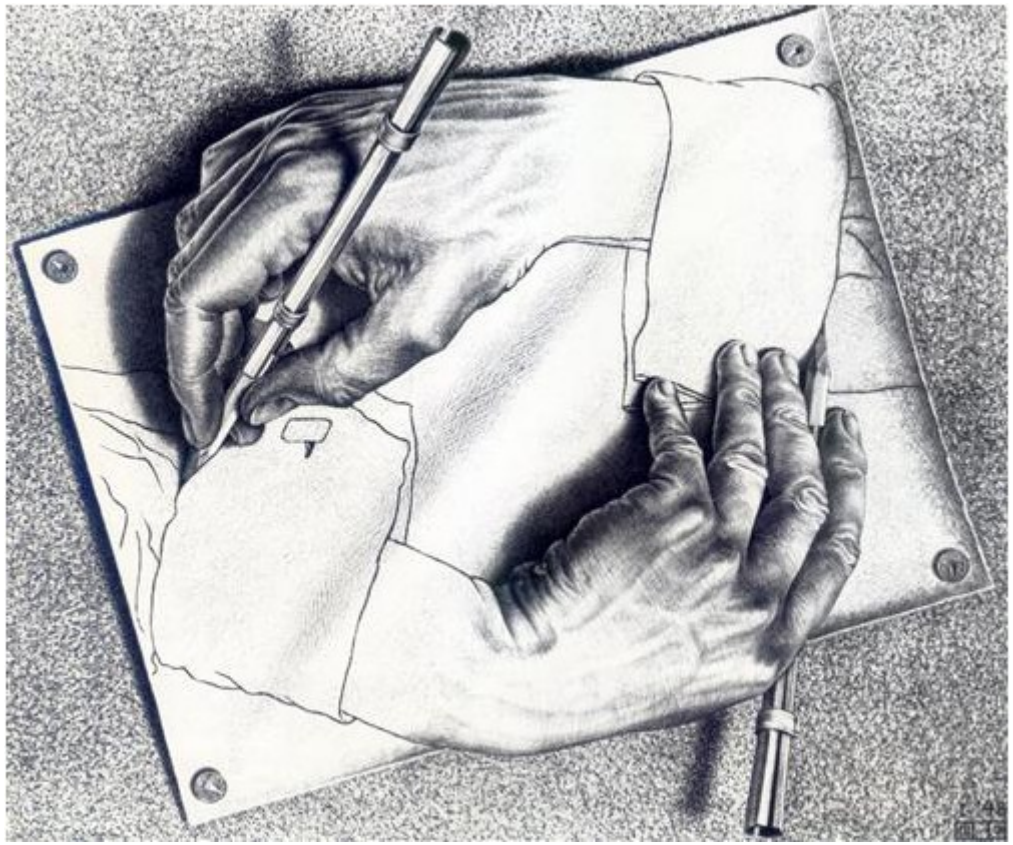
- Biological context
- More biological context
- Hi-C data processing
  - map
  - filter
  - count
  - normalize
  - segment
  - compare
- Conclusion, discussion, NETBIO lunch



M.C. Escher, 1948

Acknowledgements

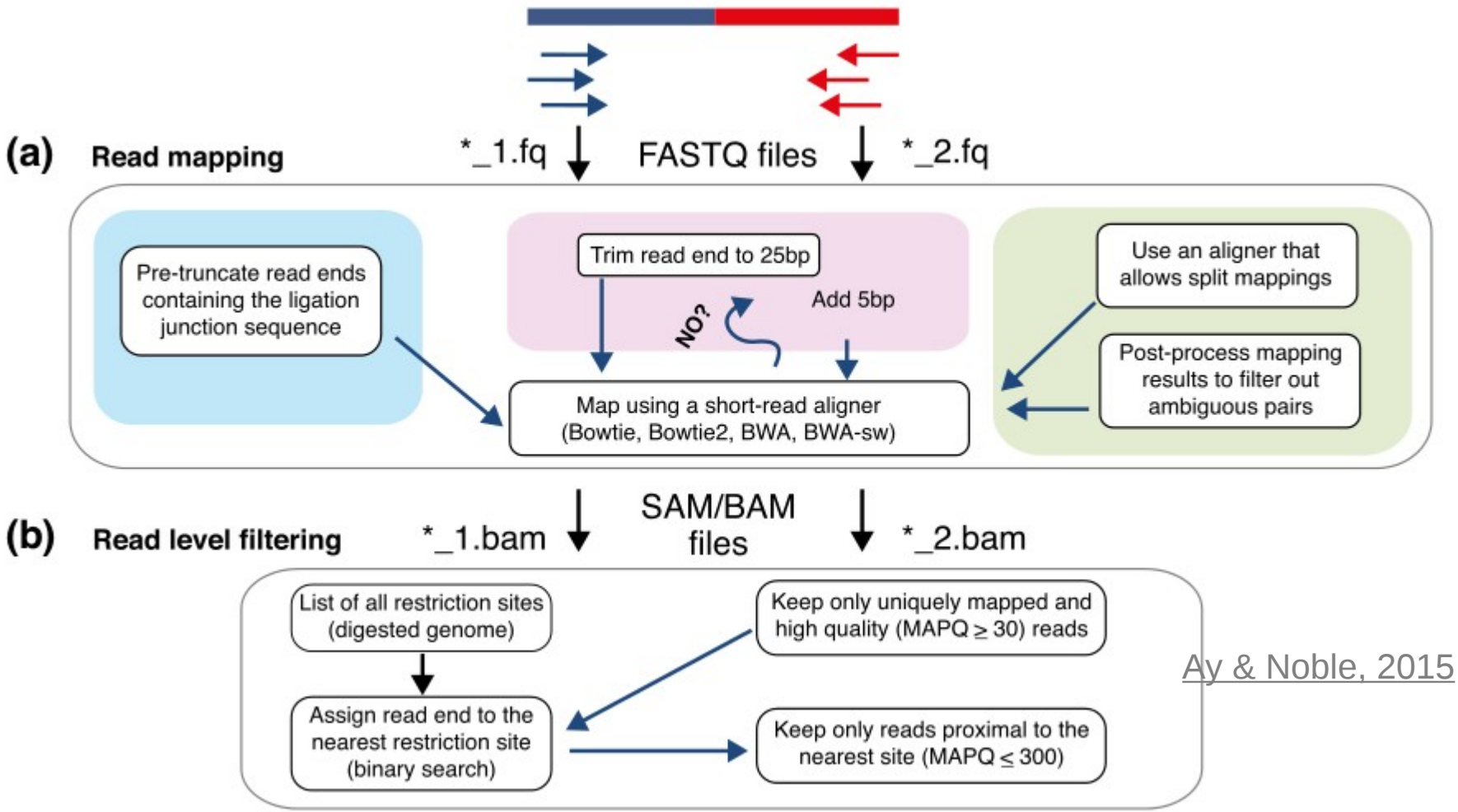
- NETBIO
- GenPhySE
  - Hervé Acloque
  - Diane Esquerré
  - Magali San Cristobal
  - Matthias Zytnicki
  - David Robelin
  - Martine Yerle
  - Yvette Lahbib
- Nicolas Servant
- FR-AgENCODE
  - Elisabetta Giuffra
- FAANG consortium
- Etc (sorry!)



M.C. Escher, 1948

Analysis: example on Hi-C data

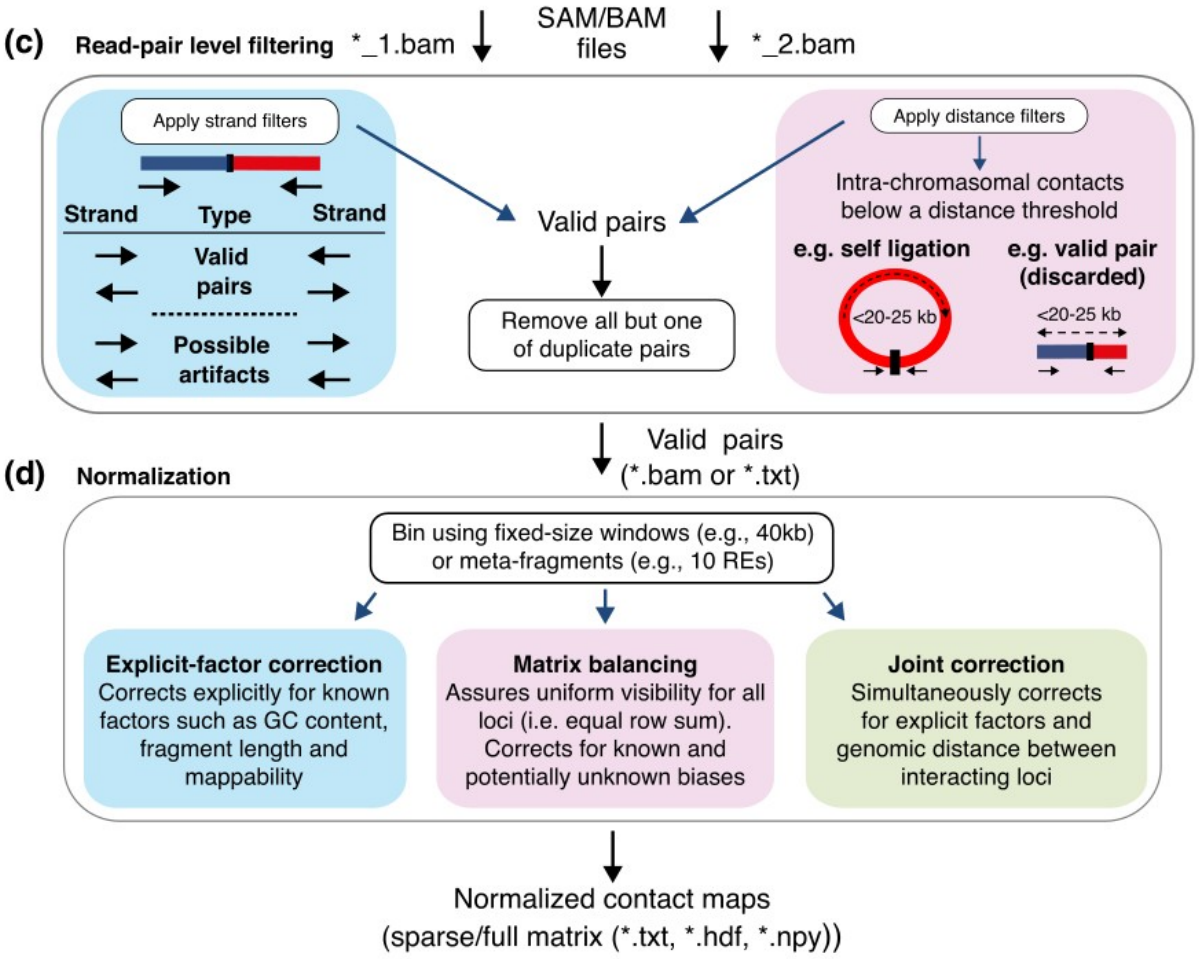
- Analysis pipeline (first part)



Ay & Noble, 2015

Analysis: example on Hi-C data

- Analysis pipeline (first part)



Ay & Noble, 2015

- + TAD calling, differential analysis, integrative analysis, ...