3D genome sequencing

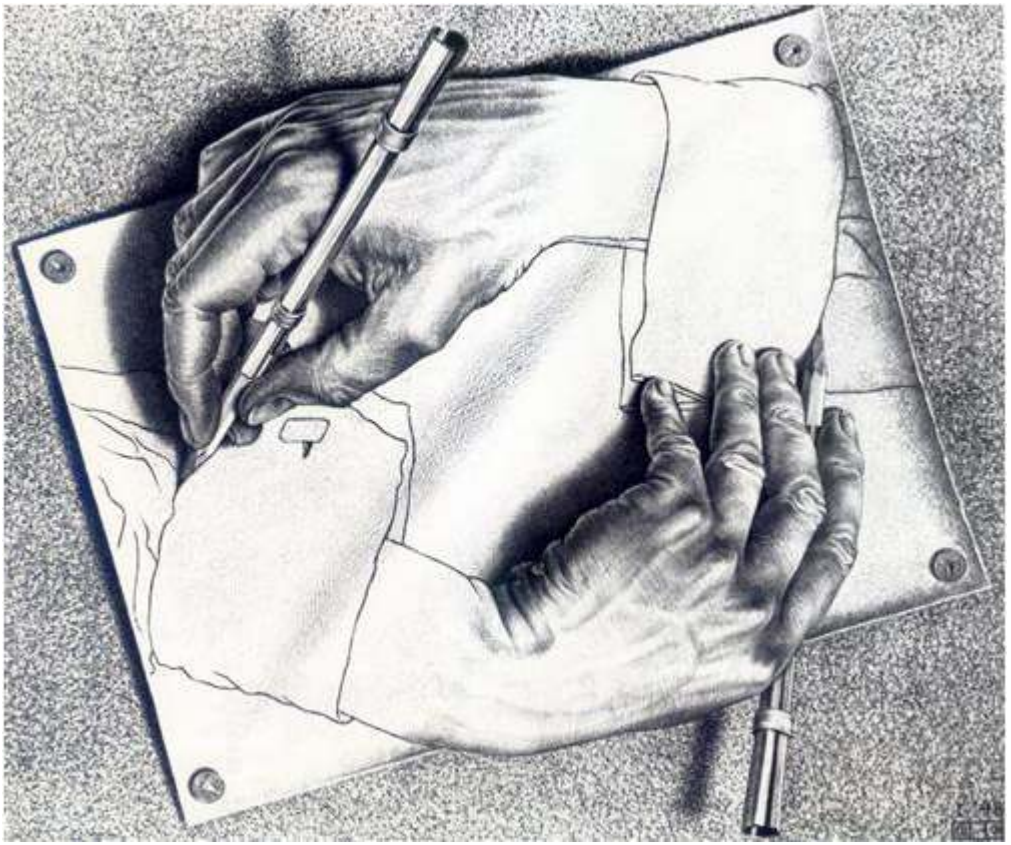
Hi-C data analysis

Sylvain Foissac, INRA Toulouse

GenoToul GetPlage - Mai 2016

Outline

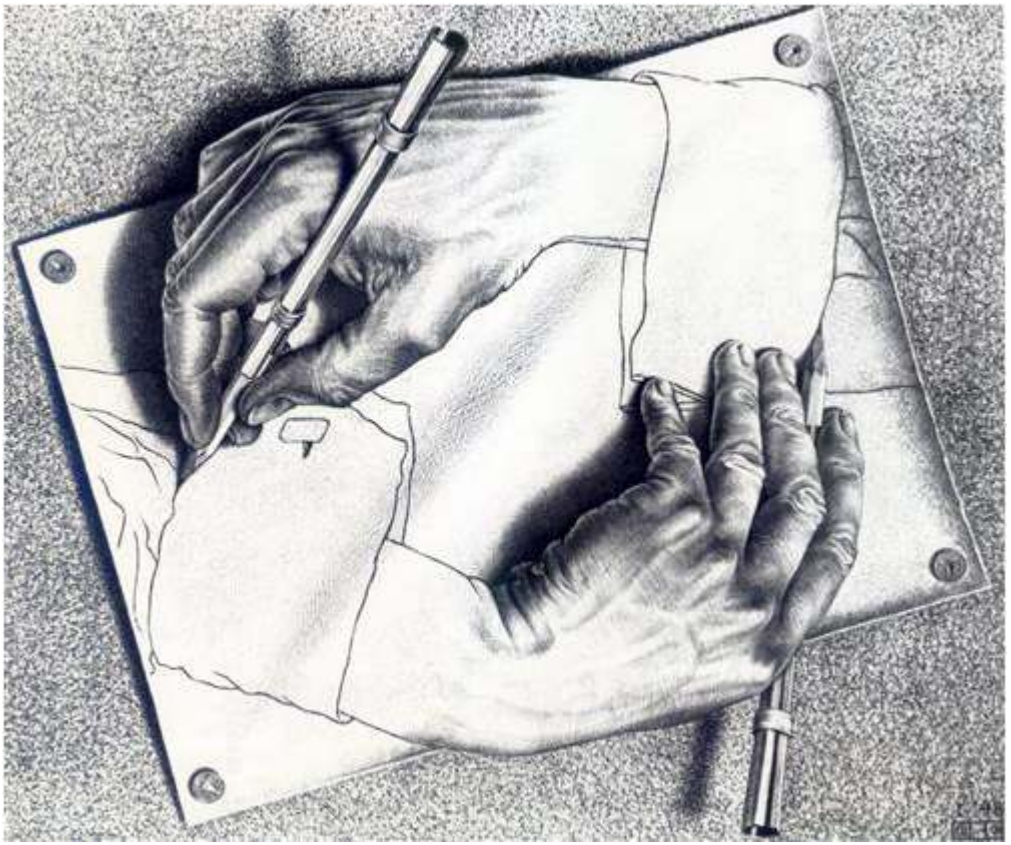
- Why
- What
- How
- Where
- Who



M.C. Escher, 1948

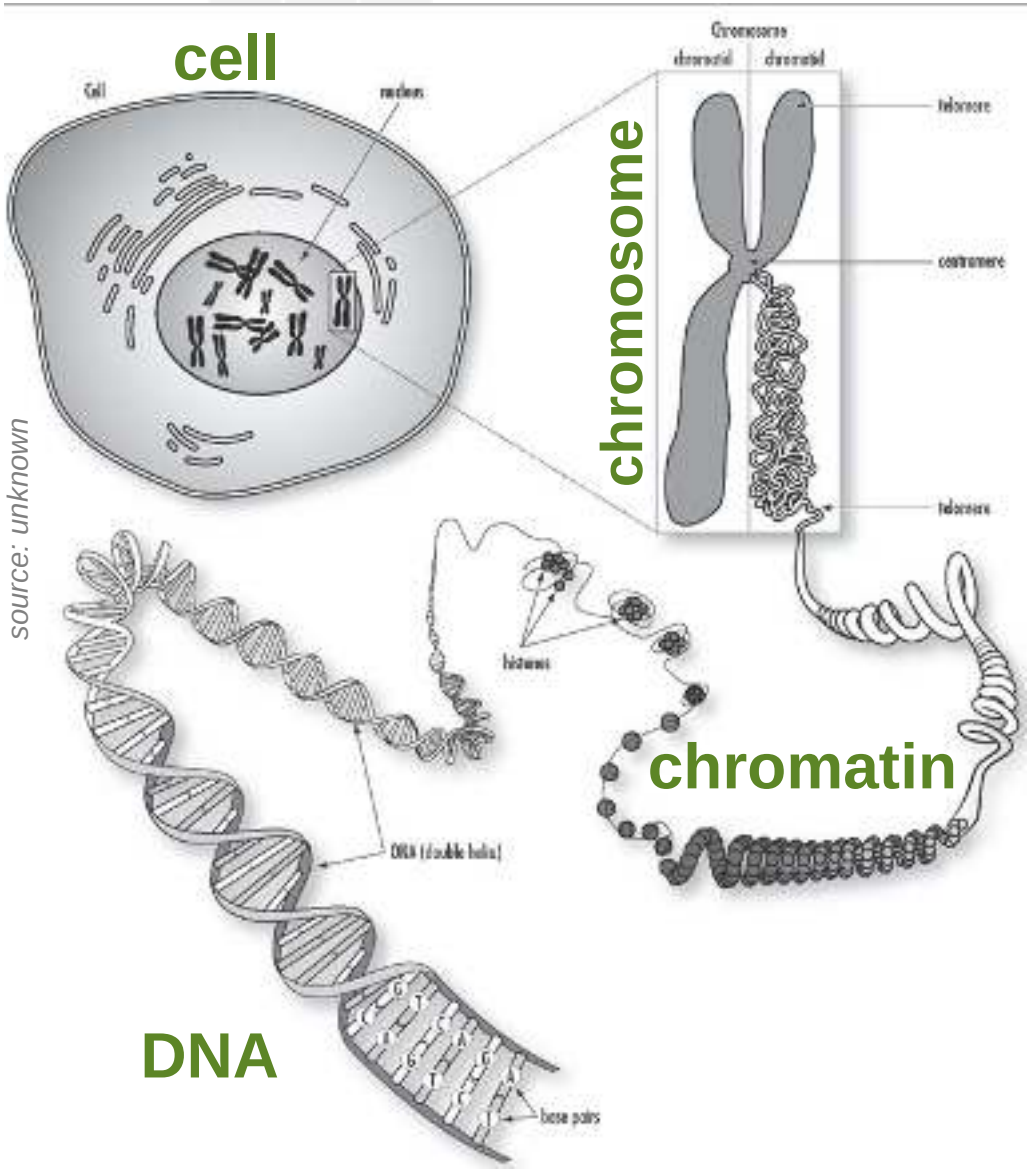
Outline

- Why
- What
- How
- Where
- Who

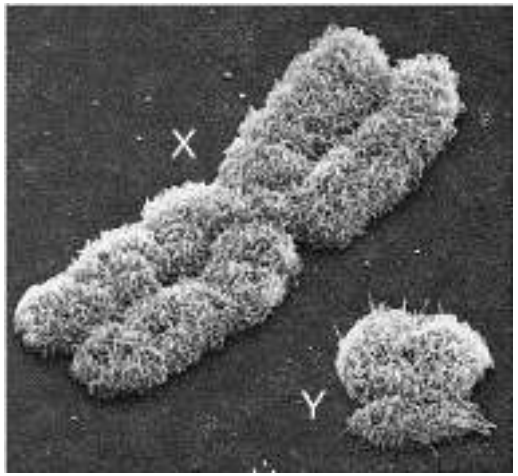


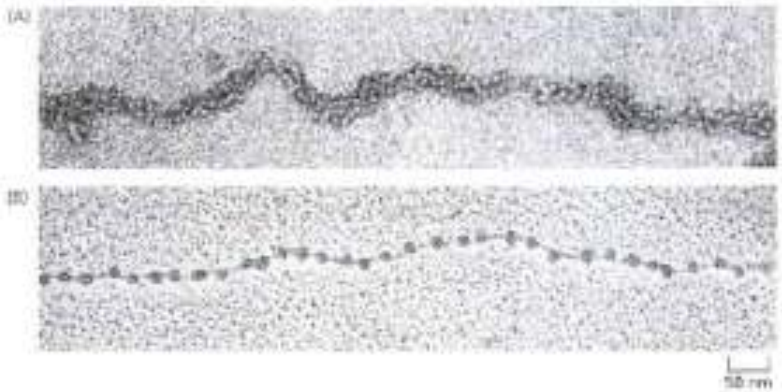
M.C. Escher, 1948

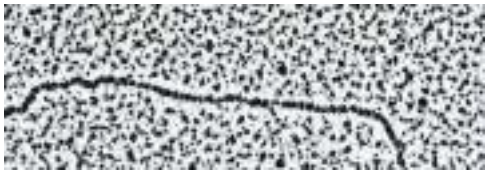
Life, cell, chromosome & DNA



source: unknown





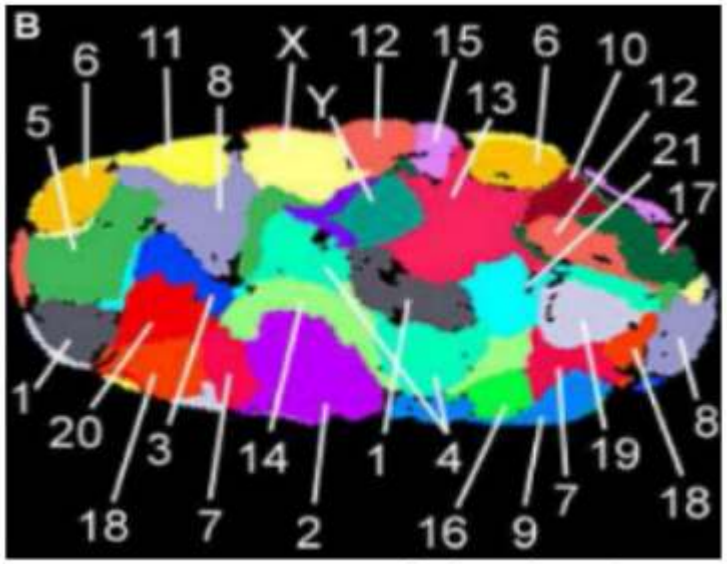


Life, cell, chromosome & DNA

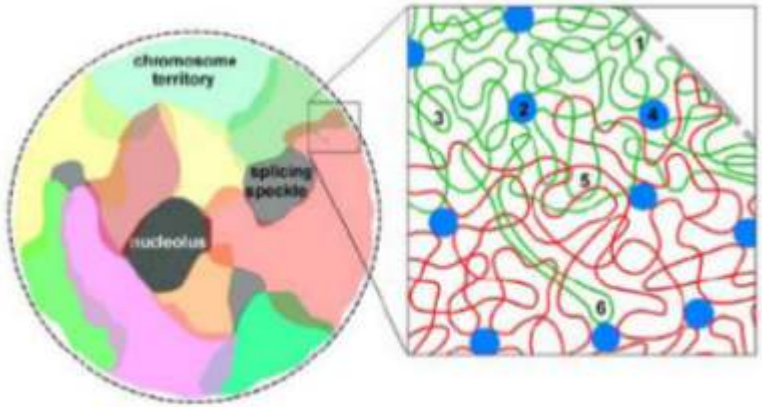
?

Life, cell, chromosome & DNA

≠



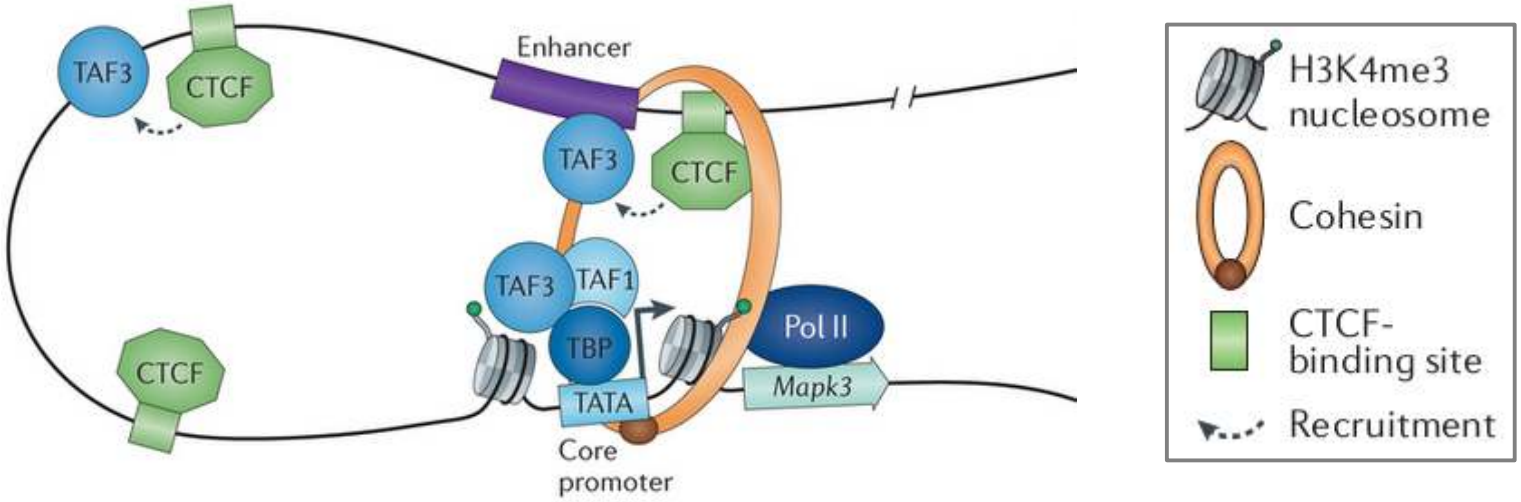
Bolzer A et al. 2005



Branco MR & Pombo A, 2006

Genome 3D structure is organized

From structure to function



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

3D structure regulates 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: ycde@mhlangalab.org
http://dx.doi.org/10.1016/j.cell.2013.09.051

SUMMARY

Transcription of coregulated genes occurs in the

GENOME RESEARCH

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

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

Dario G. Lupiáñez,[1,2] Katerina Kraft,[1,2] Verena Heinrich,[2] Peter Krawitz,[1,2] Francesco Brancati,[3] Eva K
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,[1
Bernd Timmermann,[13] Jochen Hecht,[1,14] Malte Spielmann,[1,2,14] Axel Visel,[12,15,16] and Stefan Mundlos
[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
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

TRANSCRIPTION

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

Varun Narendra,[1,5] Pedro P. Rocha,[2] Dili An,[4] Ramya Ravi
Esteban O. Mazzoni,[4,*] Danny Reinberg[1,3,*]

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

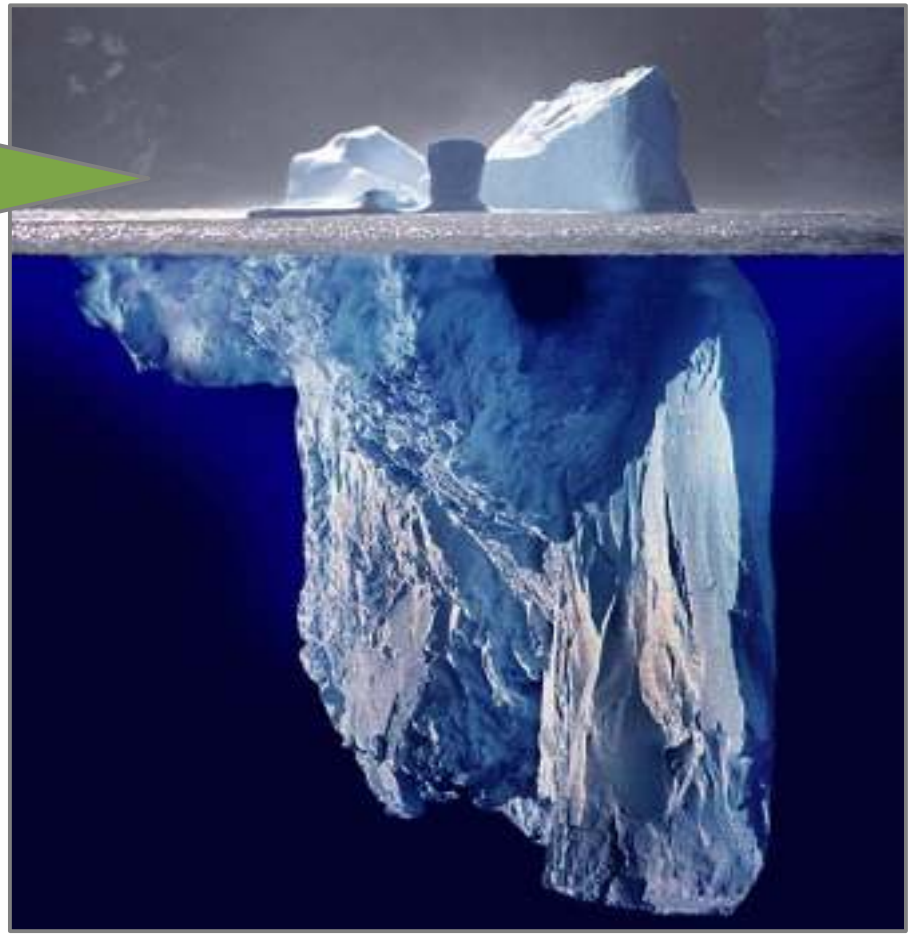
Cell

A CRISPR Connection between Chromatin Topology and Genetic Disorders

3D structure regulates gene expression

From structure to function

?



3D structure regulates gene expression

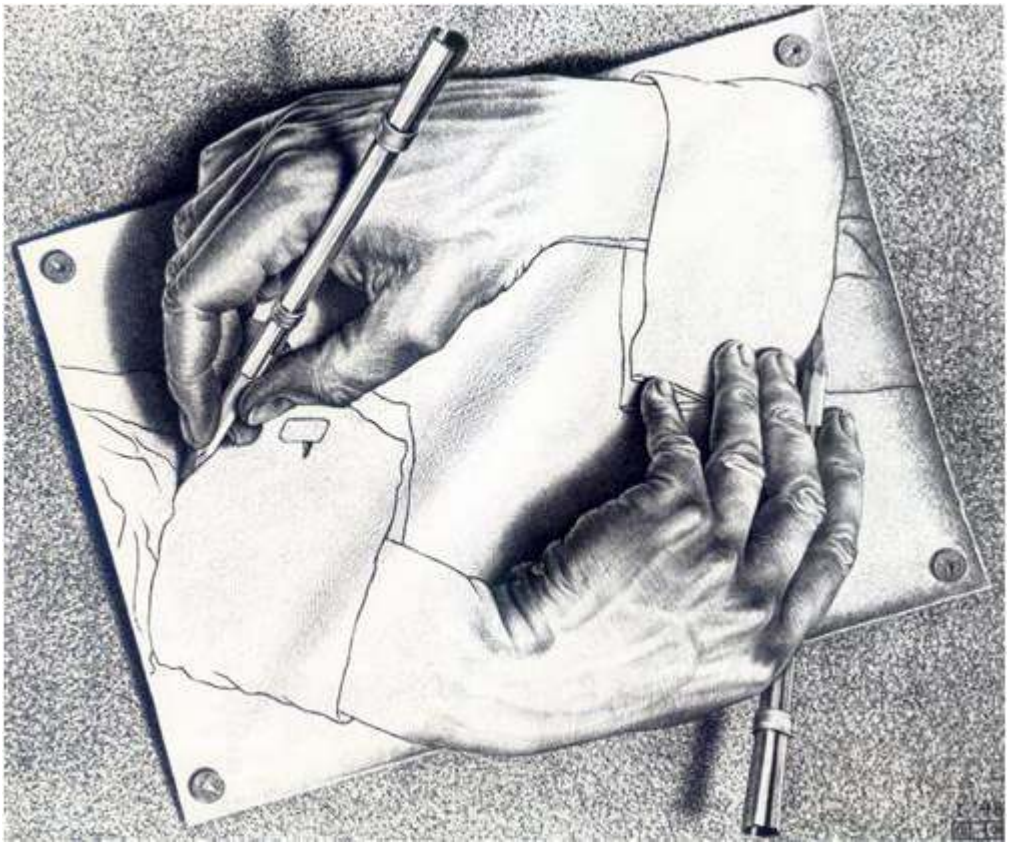
From structure to function

?

3D structure regulates gene expression

Outline

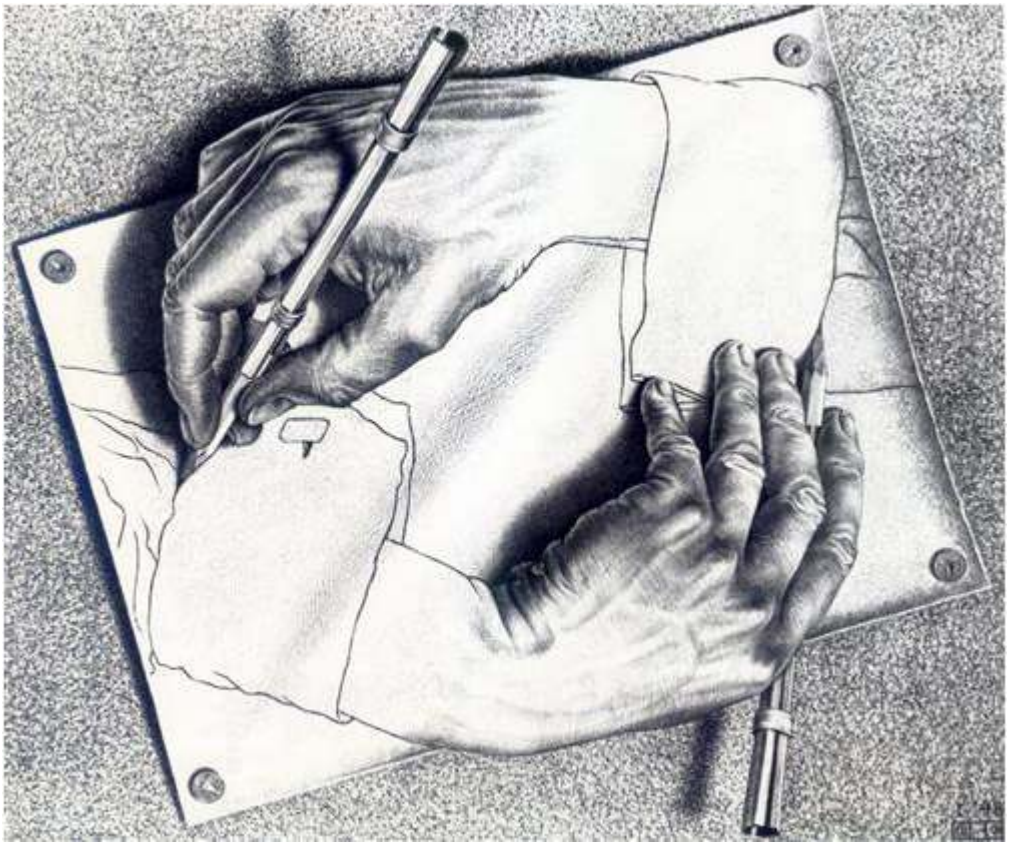
- Why
- What
- How
- Where
- Who



M.C. Escher, 1948

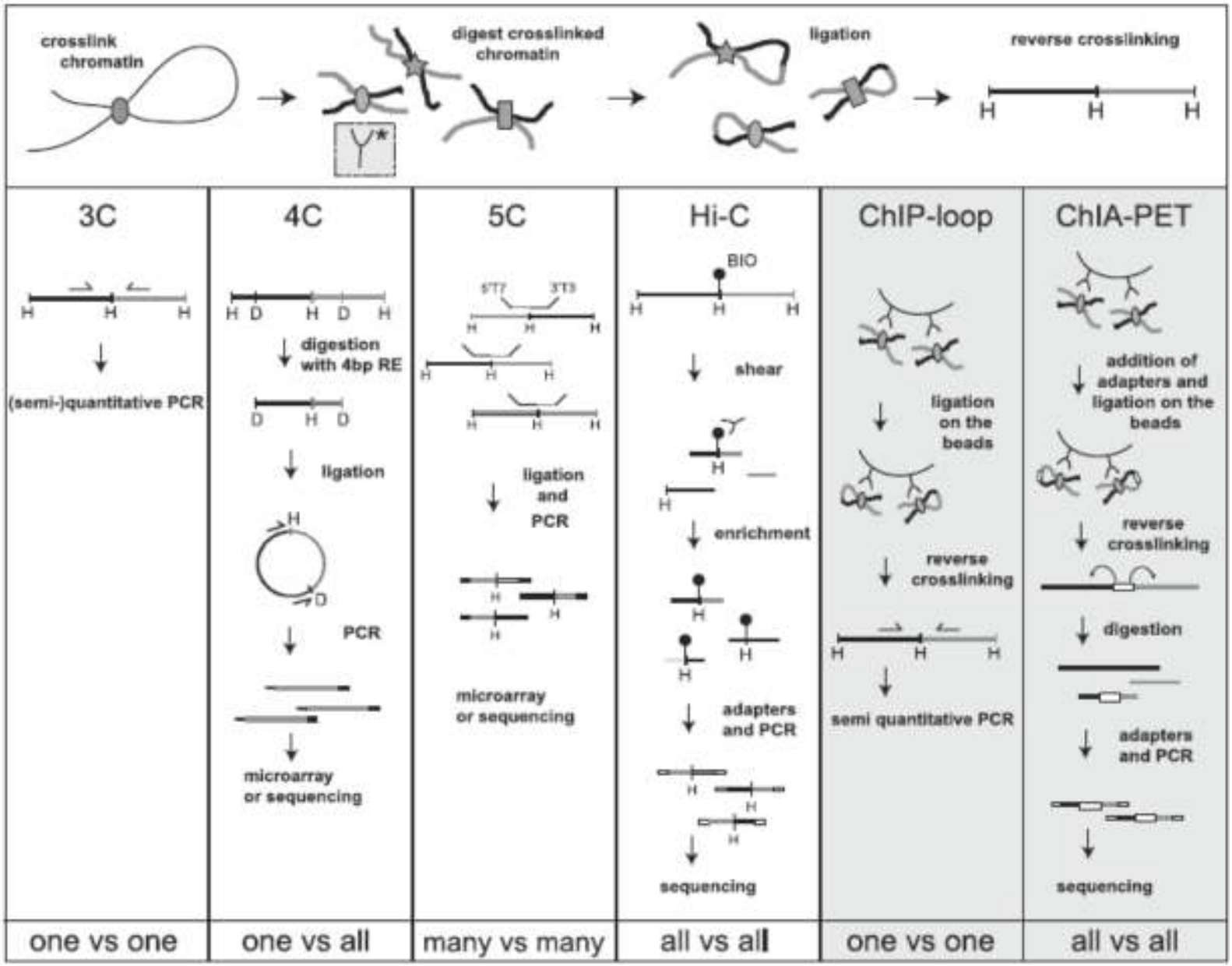
Outline

- Why
- What
- How
- Where
- Who



M.C. Escher, 1948

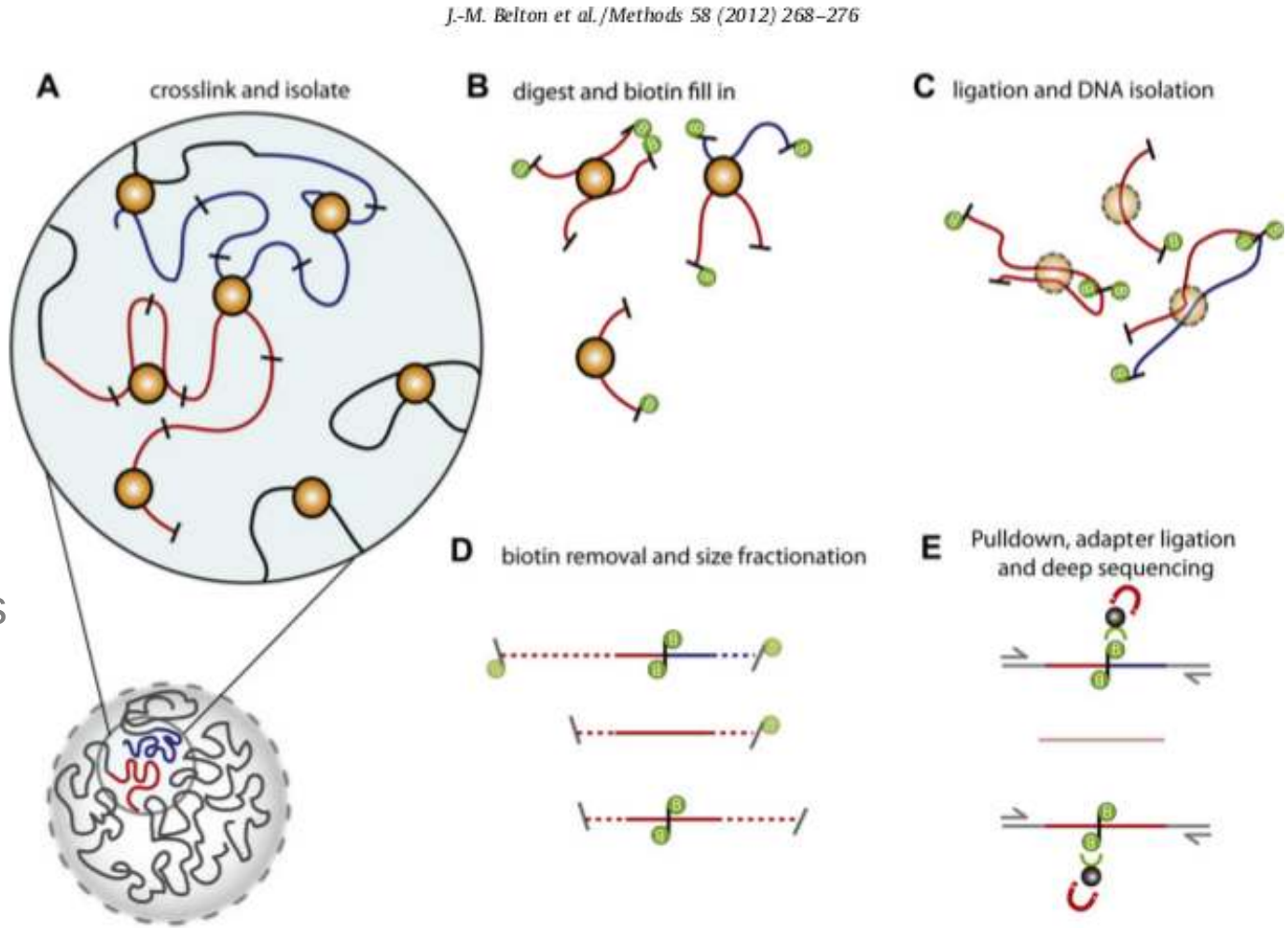
Chromosome Conformation Capture assays



Hi-C: the experiment

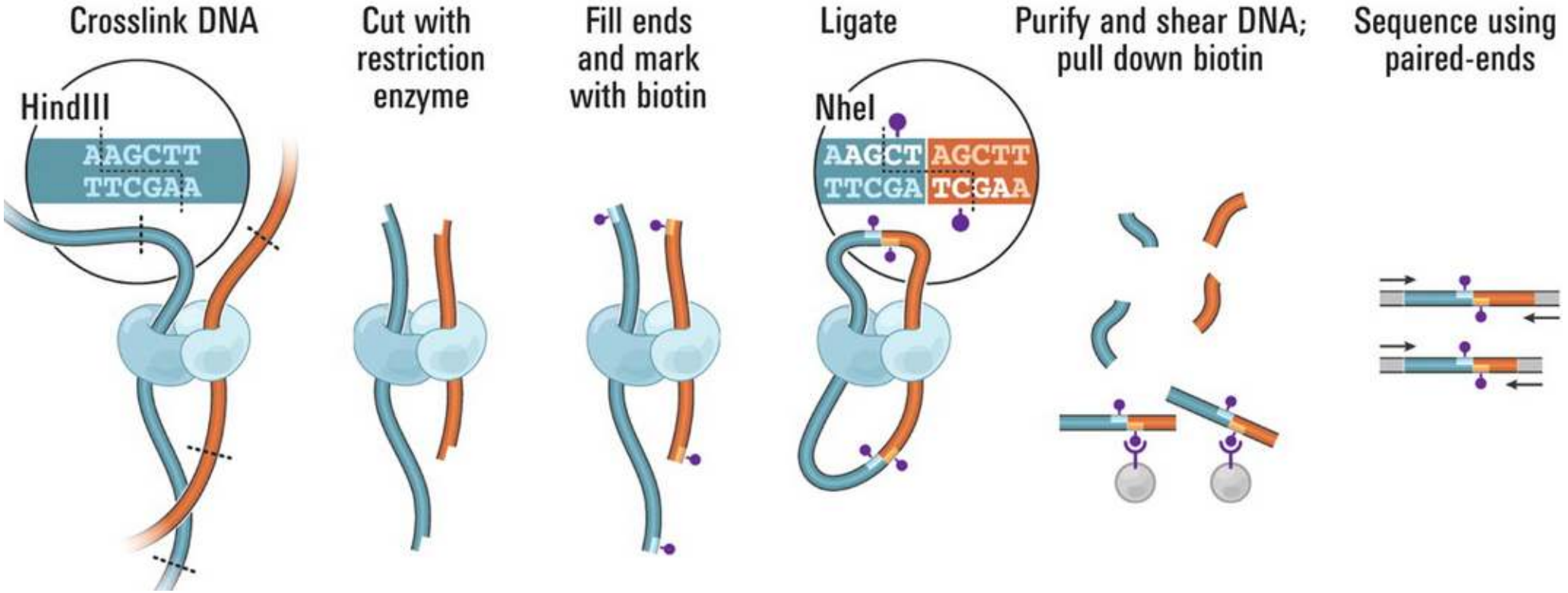
Hi-C: high-throughput chromatin conformation capture
(Lieberman-Aiden et al, Science, 2009, Rao et al, Cell, 2014)

- 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

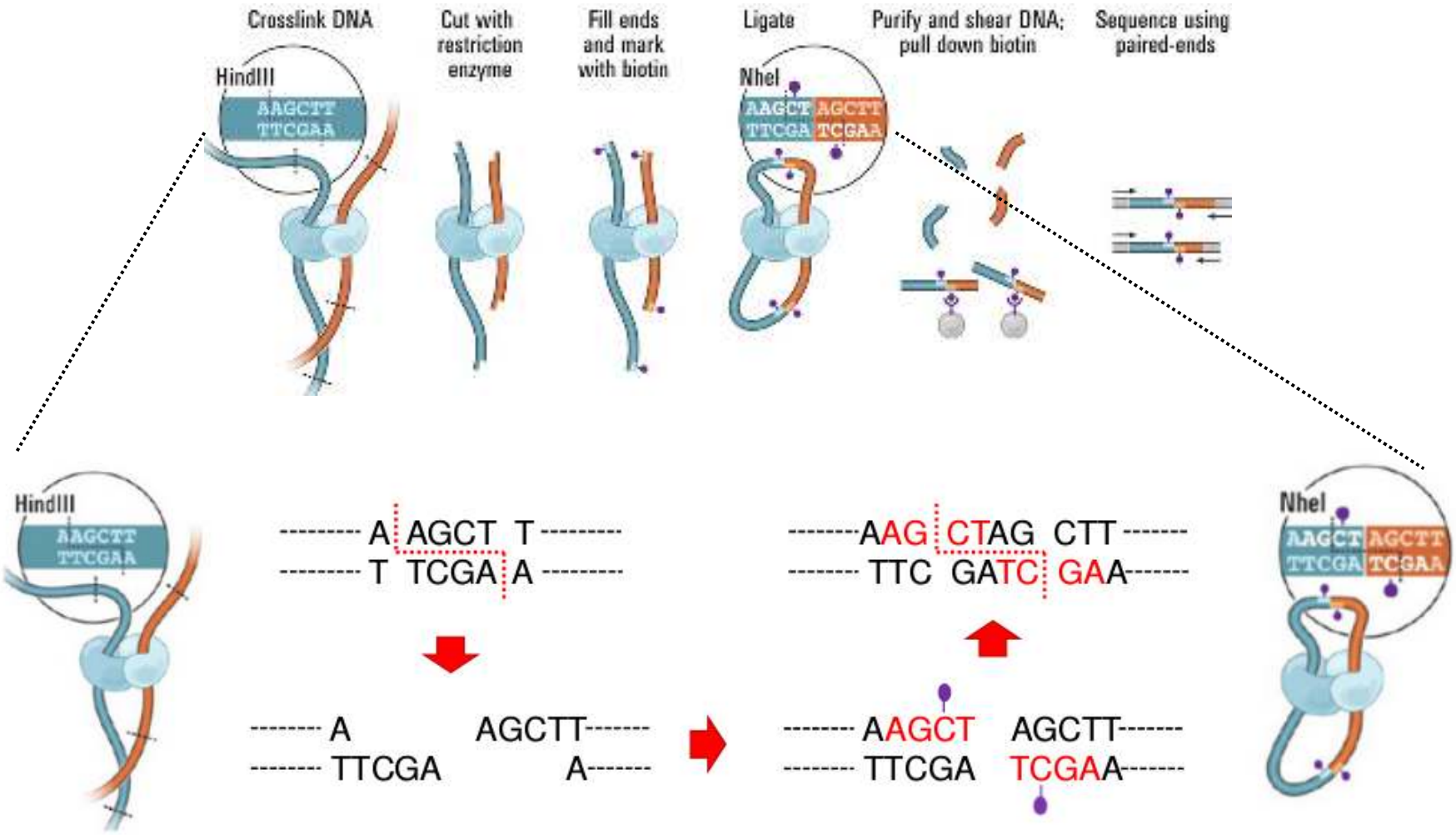
Hi-C: high-throughput chromatin conformation capture
(Lieberman-Aiden et al, Science, 2009, Rao et al, Cell, 2014)



Rao et al, Cell, 2014

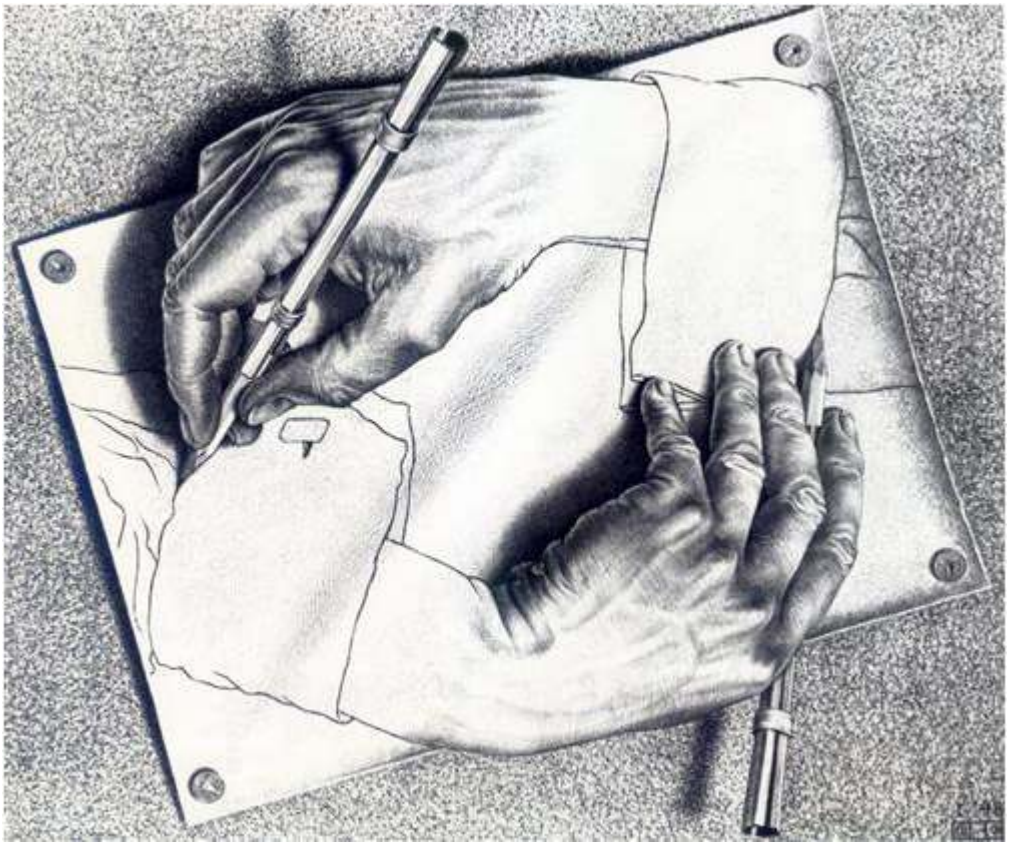
Hi-C: the experiment

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



Outline

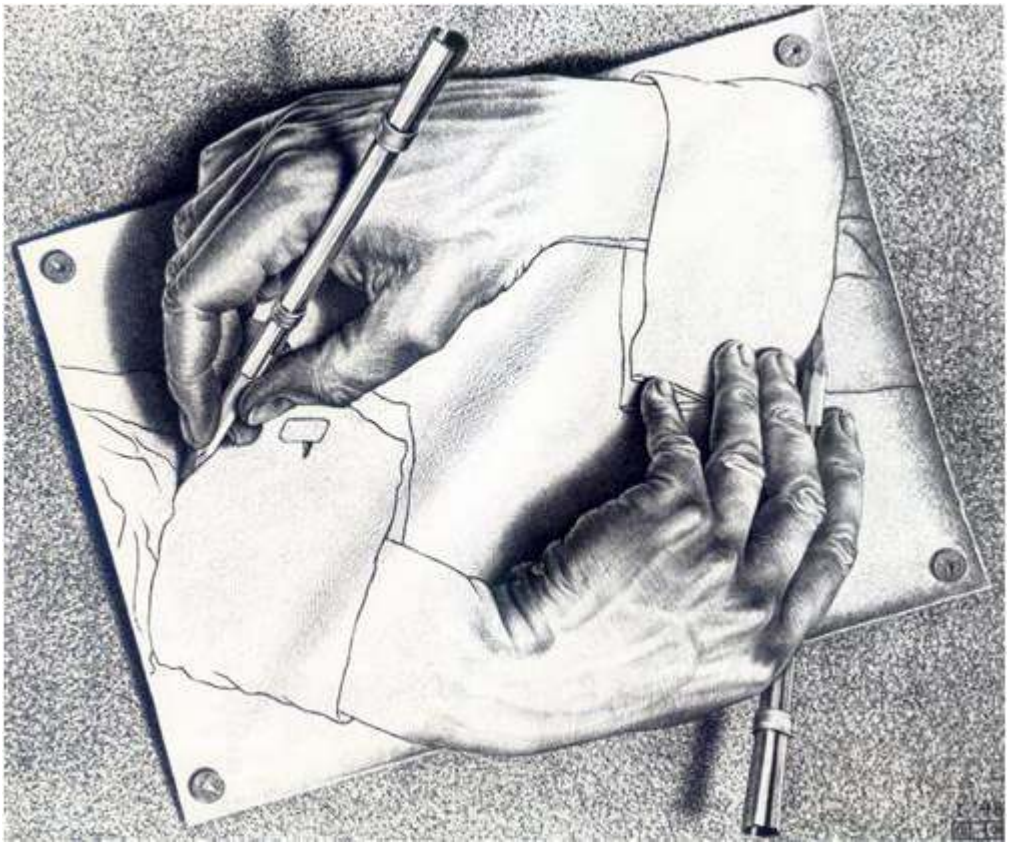
- Why
- What
- How
- Where
- Who



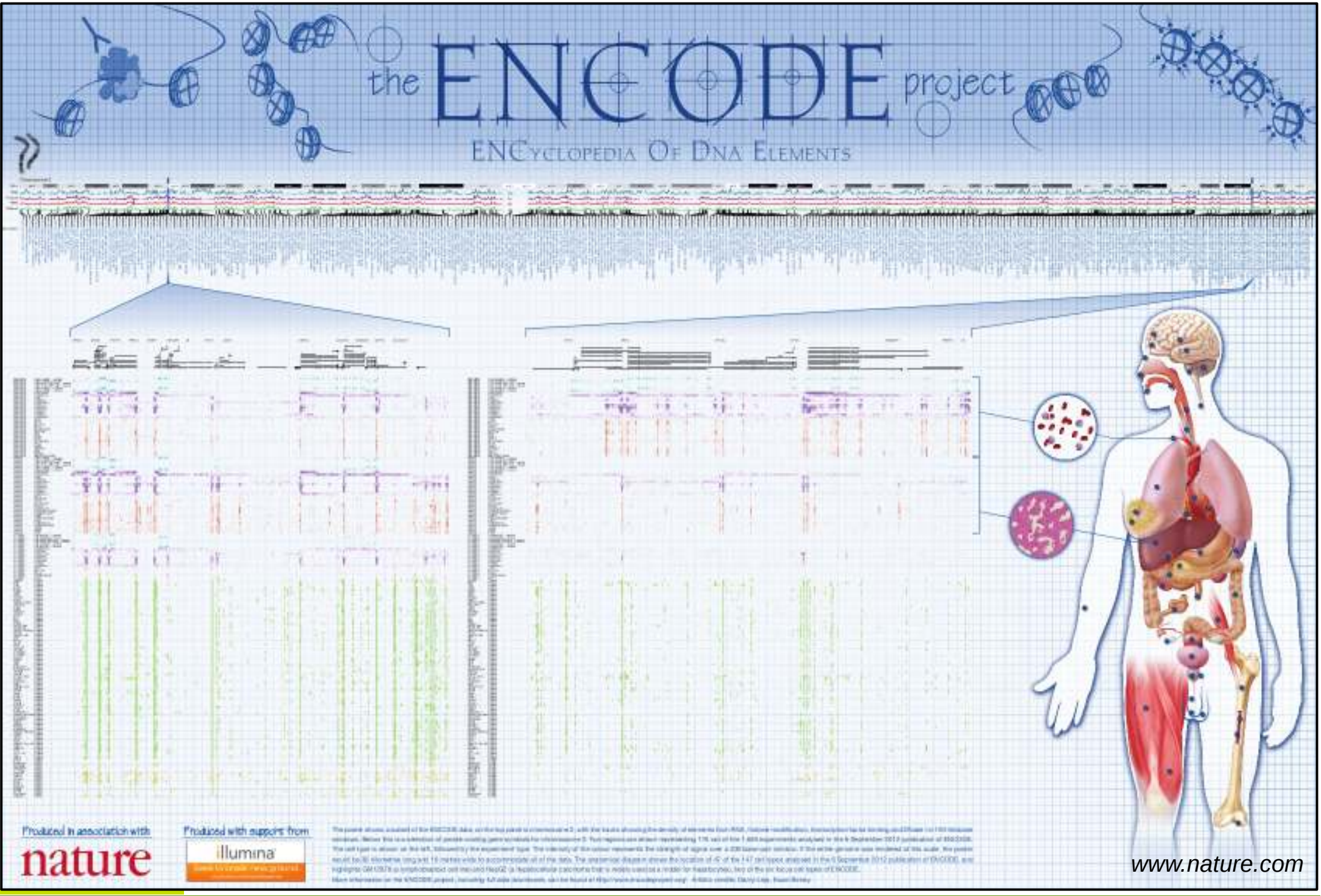
M.C. Escher, 1948

Outline

- Why
- What
- How
- Where
- Who



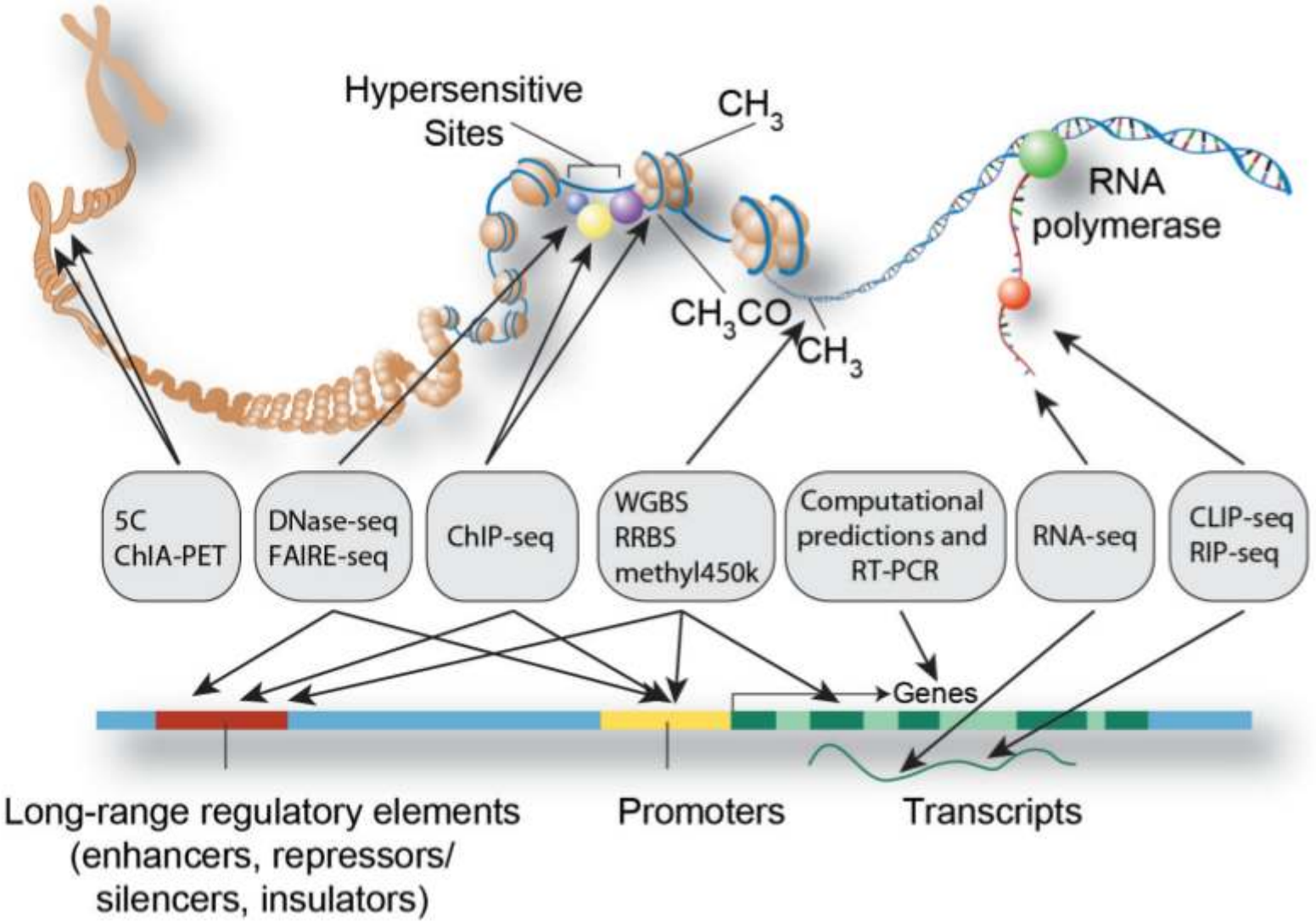
M.C. Escher, 1948


the ENCODE project
ENCyclopedia Of DNA Elements
Produced in association with
nature
Produced with support from
illumina
www.nature.com

The ENCODE project

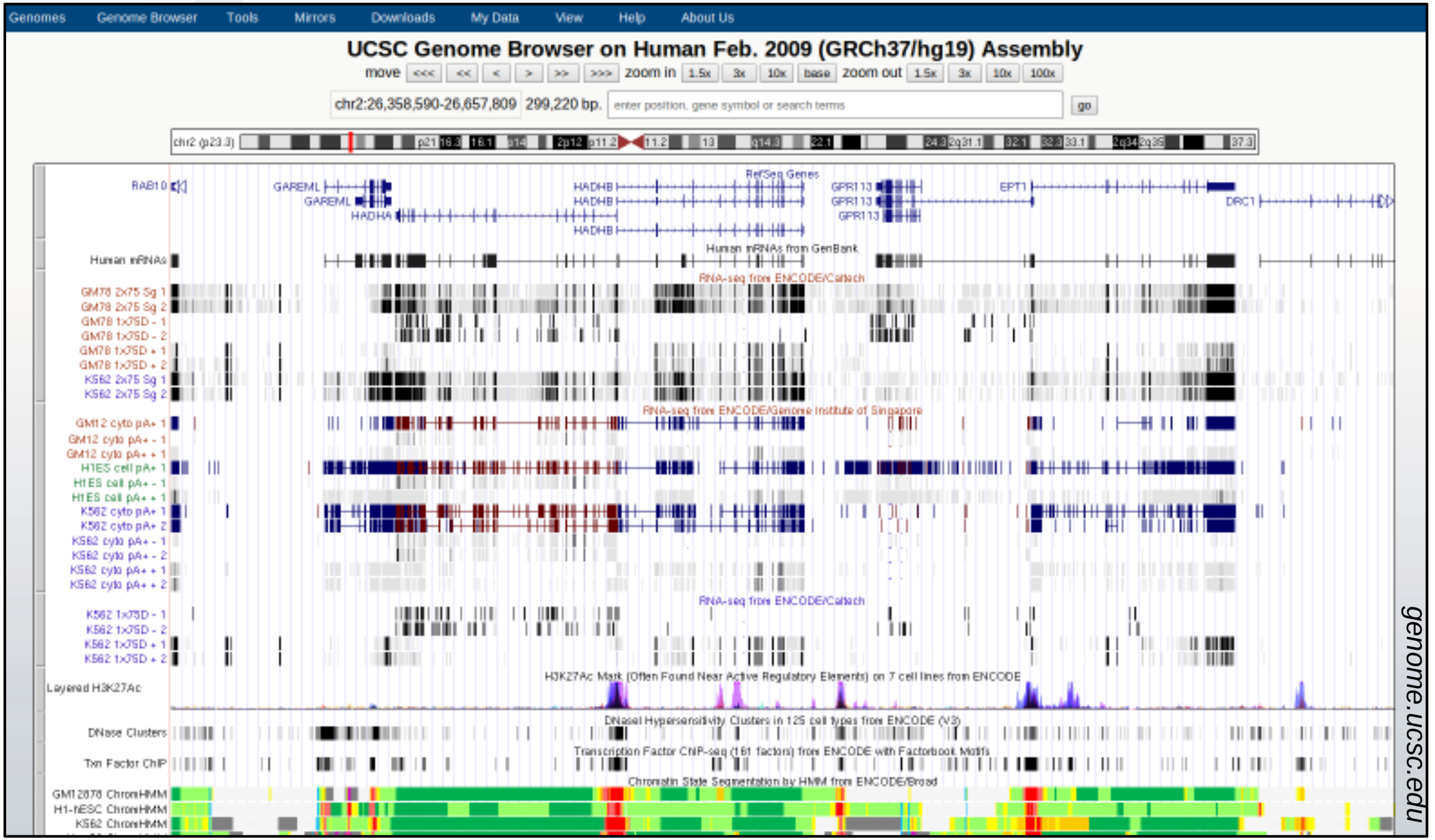
- Encyclopedia of DNA elements
- goal: characterize (annotate) all functional elements of the human genome
- mainly USA, UK, Spain, Singapore, Japan
- 32 institutes
- 440 scientists
- $300M budget from NHGRI (pilot phase + production, 2003-2012)

ENCODE experiments

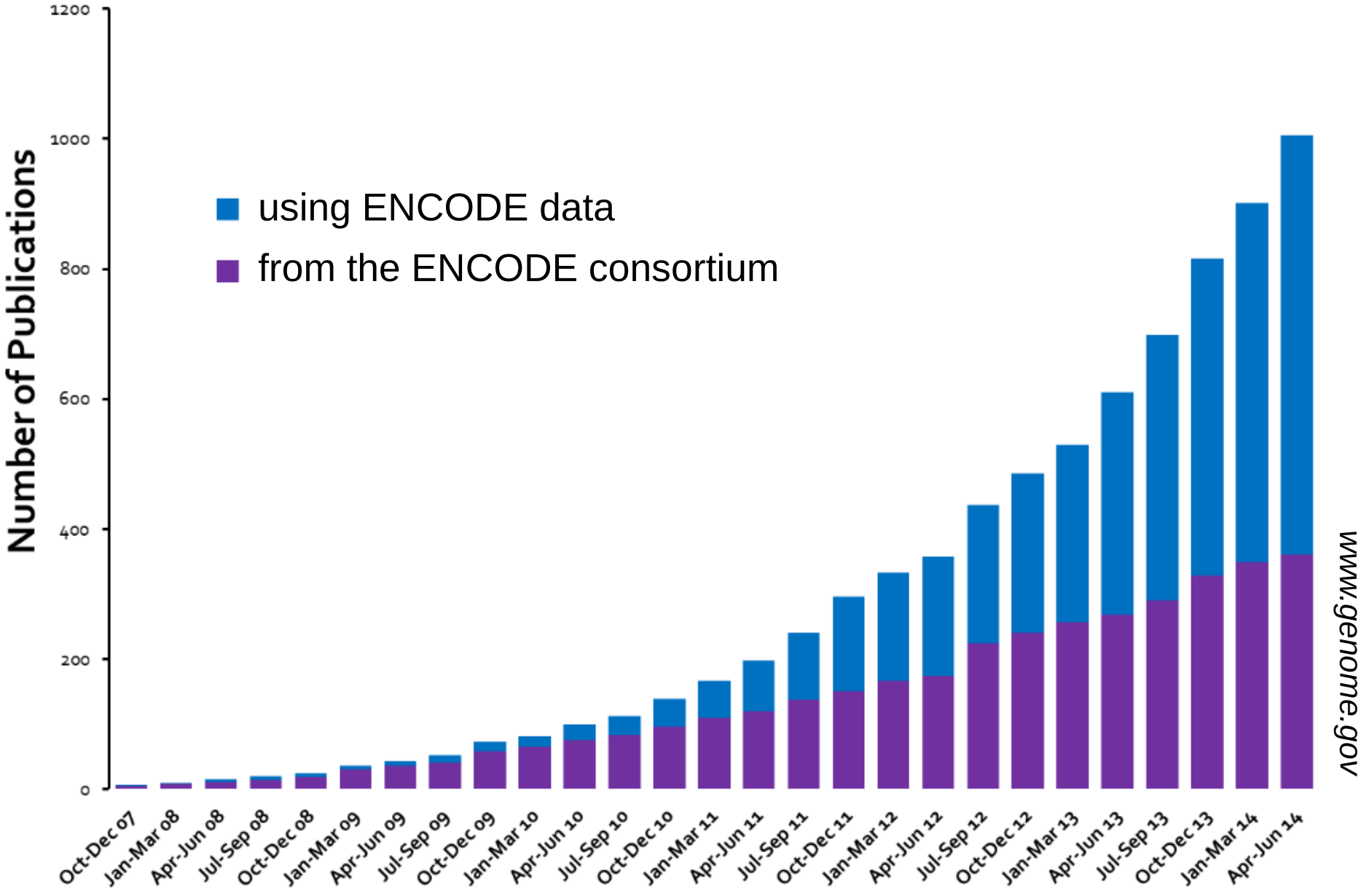


Modified from PLoS Biol 9-e1001046, 2011 & Science 306:636, 2004
Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

ENCODE data

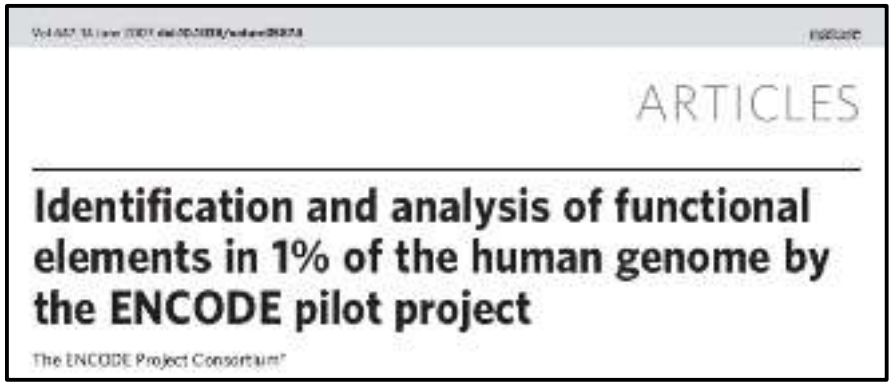

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move
zoom in
zoom out
chr2:26,358,590-26,657,809
299,220 bp.
enter position, gene symbol or search terms
go
RefSeq Genes
Human mRNAs from GenBank
RNA-seq from ENCODE/Caltech
RNA-seq from ENCODE/Genome Institute of Singapore
RNA-seq from ENCODE/Caltech
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Chromatin State Segmentation by HMM from ENCODE/Broad
genome.ucsc.edu

ENCODE publications



ENCODE publications

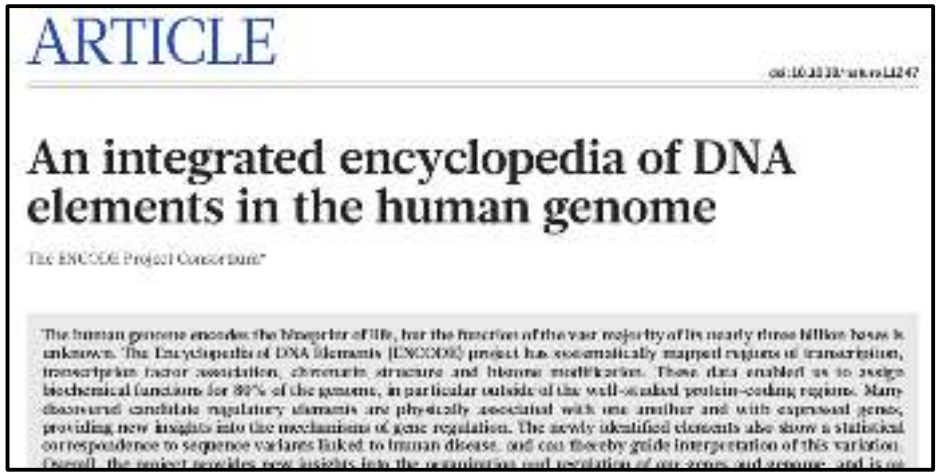
2007: pilot project


ARTICLES

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

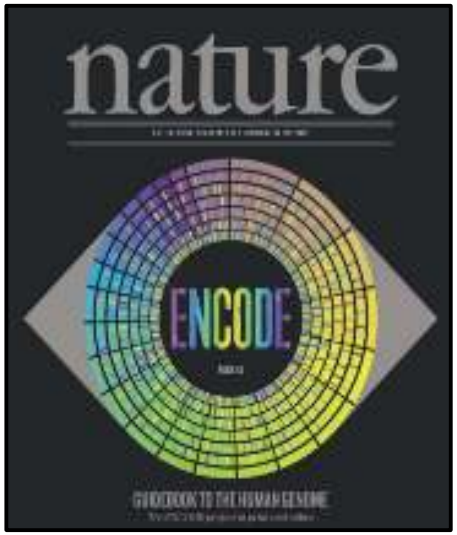
The ENCODE Project Consortium*

2012: 30 articles


ARTICLE

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*



THE WALL STREET JOURNAL.

HEALTH

'Junk DNA' Debunked

Studies Find Human Genomic Makeup Is Vastly Messier; New

By GAUTAM NAIK and ROBERT LEE HOTZ

Updated Sept. 5, 2012 2:01 p.m. ET

BBC

News | Sport | Weather | Earth | Future | Shop

NEWS HEALTH

Home | UK | Africa | Asia | Australia | Europe | Latin America | Mid-East | US & Canada | Business | Health

Human genome 'more active than

The Economist

World politics | Business & finance | Economics | Science & techno

Human genomics

The new world of DNA

A long-term effort to catalogue all the bits of the human genome that do

has

The New York Times

Science

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH

ENVIRONMENT

Bits of Mystery DNA, Far From 'Junk,' Play

By G

theguardian

Winner of the Pulitzer prize

home › science | UK | world | sport | football | comment | culture | eco | all sections

Genetics

Breakthrough study overturns theory of 'junk DNA' in genome

The international Encode project has found that about a fifth of the human genome regulates the 2% that makes proteins

Los Angeles Times | ARTICLE

← Back to Original Article

ENCODE project sheds light

New findings from the ENCODE project provid

personalized medicine.

September 05, 2012 | By Rosie Mestel and Eryn Brow

The Tele

Home | Video | News | Wo

Politics | Investigations

Science News | Space | Night Sky | Roger Highfield | Dinosaurs | Evolution | Steve Jones | Science

HOME » SCIENCE » SCIENCE NEWS

Worldwide army of scientists cracks the 'junk DNA' code

One of the biggest mysteries in genetics has been solved after an international team of hundreds of scientists uncovered the secrets of "junk DNA".

DNA project interprets 'book of life'

By Elizabeth Landau , CNN

September 5, 2012 -- Updated 1841 GMT (0241 HKT)

CNN

(CNN) -- Our genes play a majo

but a lot of information about th

mysterious.

That's why an international tea

out what the working parts of th

they mean for the human body as we know it.

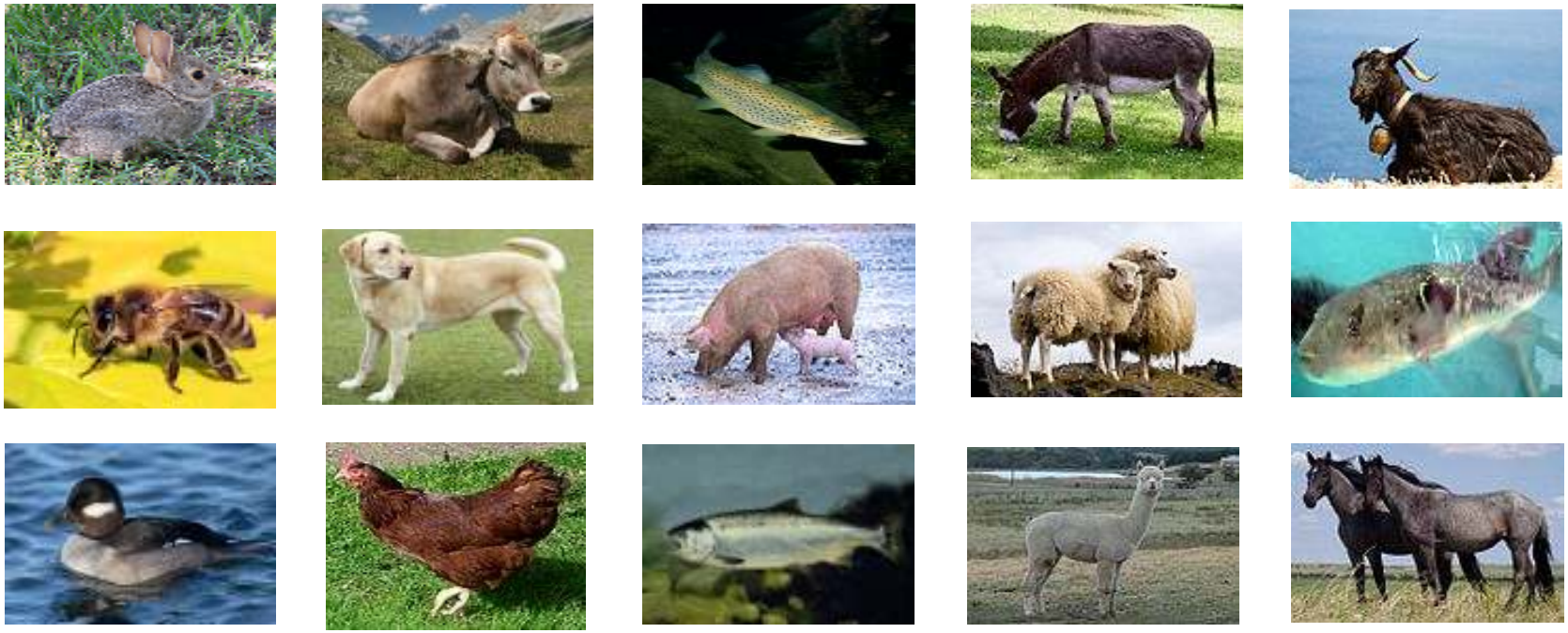
The project is called the Encyclopedia of DNA Elements (E

The FAANG project

FAANG: Functional Analysis of ANimal Genomes

Andersson et al, Genome Biology, 2015

Coordinated international action to accelerate Genome to Phenome - the Functional Annotation of Animal Genome (FAANG) project

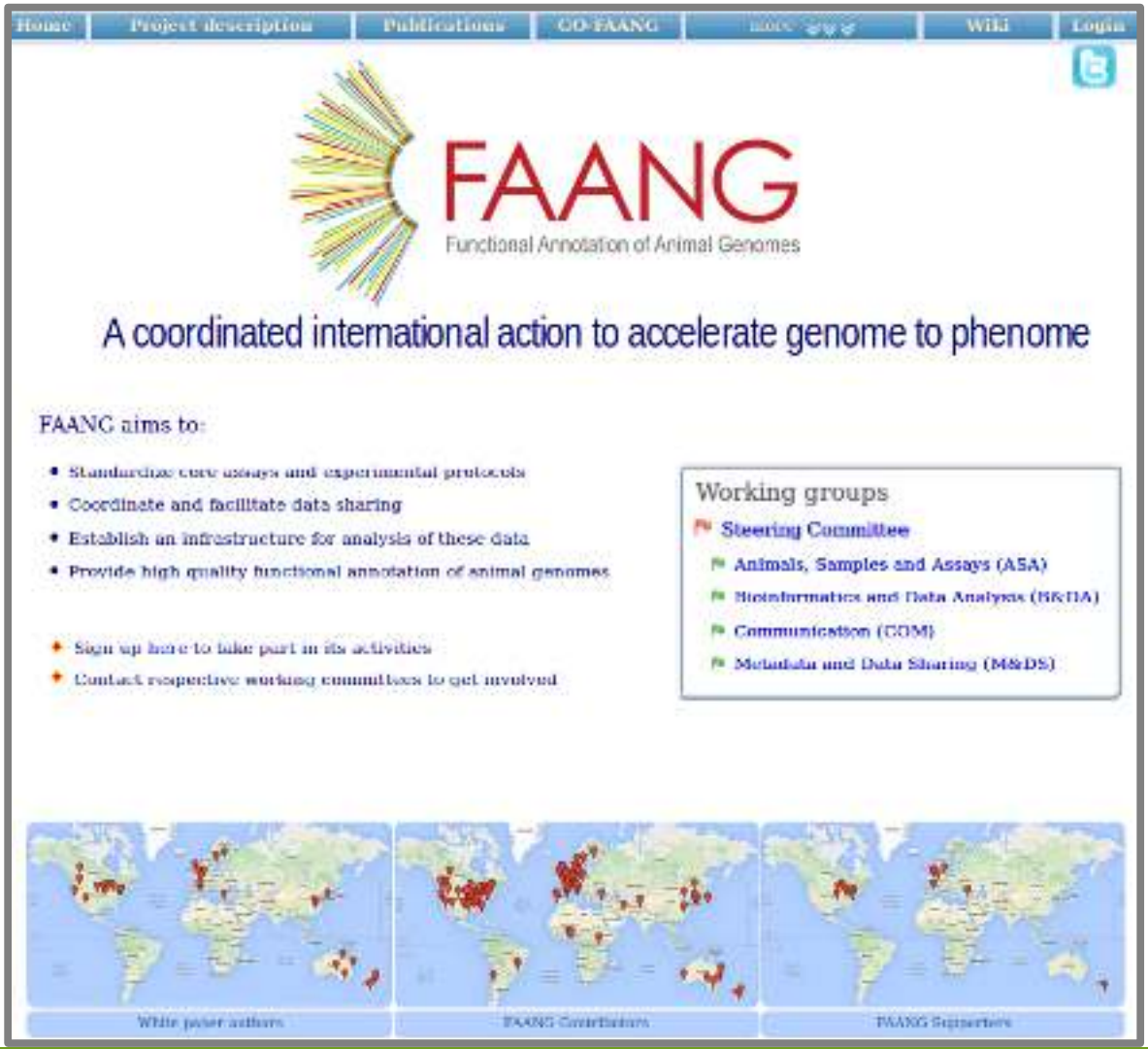


wikipedia.org

The FAANG project

FAANG: Functional Analysis of ANimal Genomes

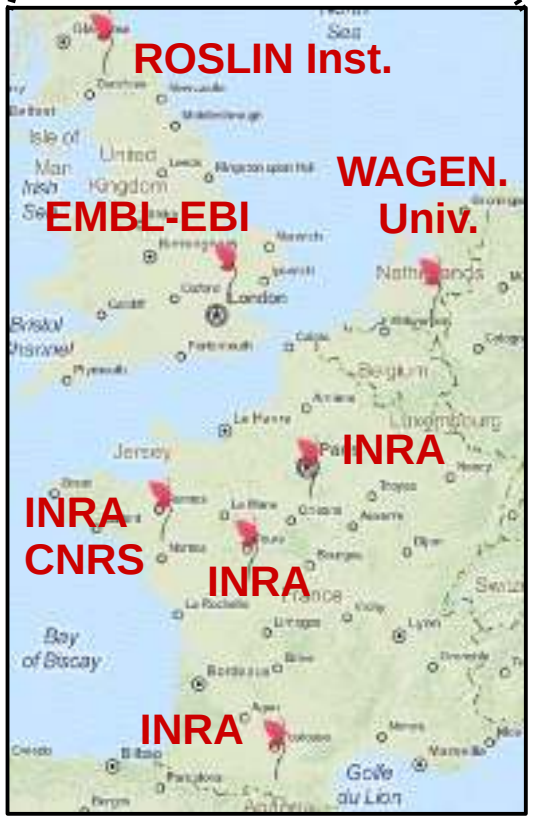
www.faang.org



=> 3 pilot projects

FR-AgENCODE: overview

- a French pilot project of FAANG
- goal: improve functional annotation of livestock genomes
- founding: INRA, France (300KE)
- 4 INRA sites, 9 labs, 58 scientists
- 4 species: pig, chicken, cattle, goat
- primary targets: liver & blood cells (CD4+ & CD8+)
- molecular assays: RNA-seq, Hi-C & ATAC-seq
- duration: 2015-2017



E. Giuffra, INRA GABI

FR-AgENCODE: overview

Sampling: 40+ tissues

(**liver, CD4+, CD8+**, sperm, plasma, heart, lung, skin, fat, duodenum, ileum, jejunum, cerebellum, frontal lobe, olfactory bulb, trigeminal ganglia, hypotalamus, pancreas, andrenals, kidney, muscle, bone, joints, spleen, lymphatic nodes, peyer's patches, ovary, oocytes, oviduct, uterus, mammary gland, acini, testis, seminal vesicle, etc)

2x♂
2x♀

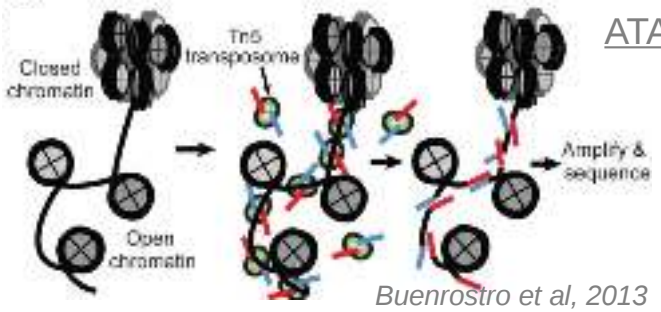
Sus scrofa (Large White)

Gallus gallus (White Leghorn)

Bos Taurus (Holstein)

Capra hircus (Alpine)

=> INRA CRB-Anim biorepository



Buenrostro et al, 2013

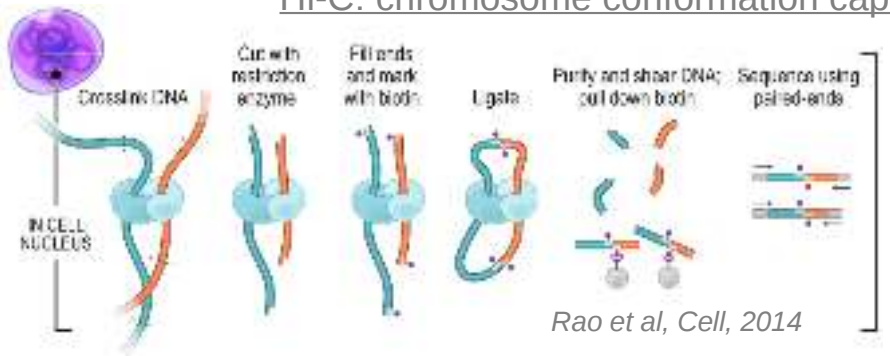
ATAC-seq: chromatin accessibility

RNA-seq: long & short RNAs

Molecular assays: 3 target tissues

transcriptome & chromatin structure profiling
polyA+ RNA-seq (mRNAs & lncRNAs, 130M RP/lib)
small RNA-seq (miRNAs & <200nt RNAs, 40MR/lib)
Hi-C (130M RP/lib) & ATAC-seq (40M RP/lib)

Hi-C: chromosome conformation capture



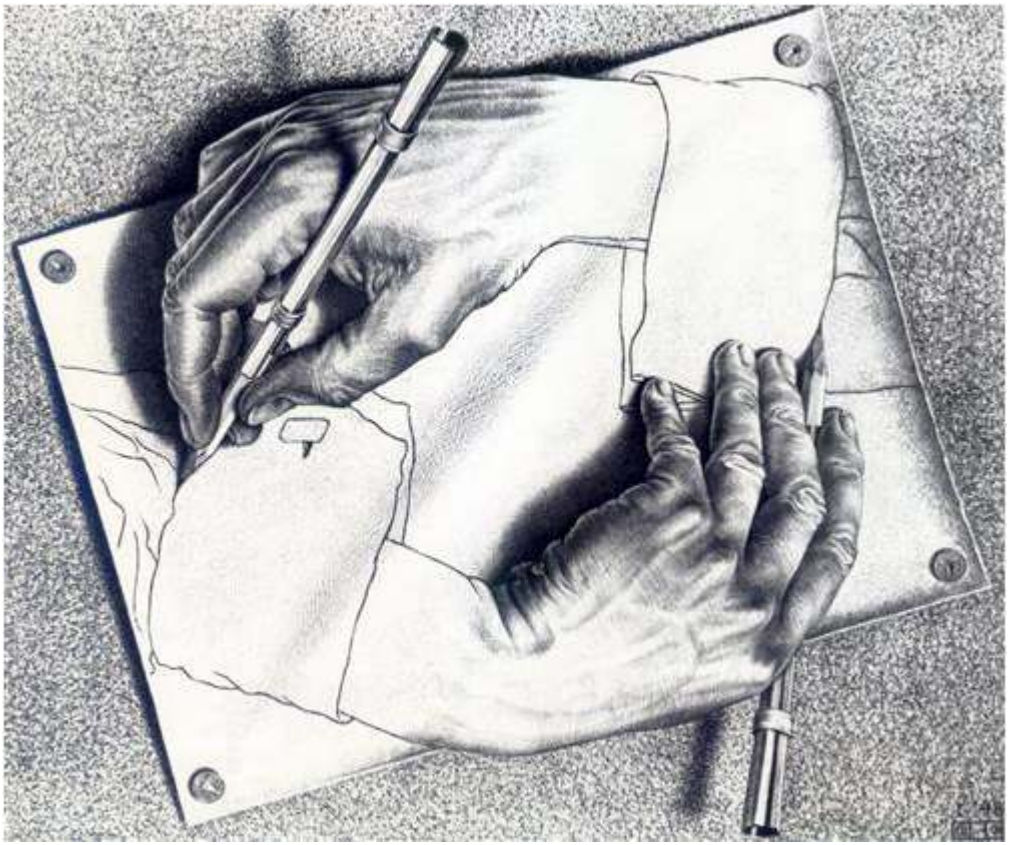
Rao et al, Cell, 2014

Bioinformatics data analysis

genome annotation, gene expression, lncRNAs & sRNAs annotation/prediction, chromosome interaction matrices & contact heatmaps, allele-specific expression, chimeric transcripts detection, comparative genomics, etc

Outline

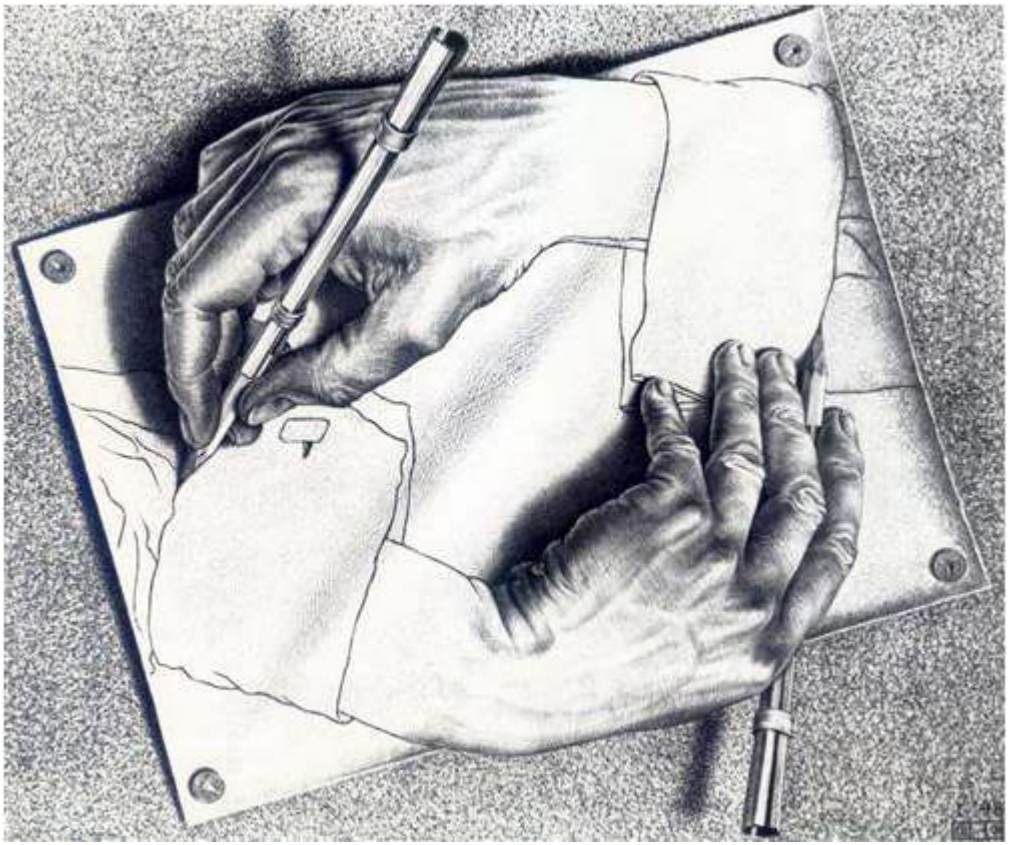
- Why
- What
- How
- Where
- Who



M.C. Escher, 1948

Outline

- Why
- What
- How
- Where
- Who



M.C. Escher, 1948

FR-AgENCODE: sampling

2x ♂

2x ♀

Sus scrofa
(Large White)

Gallus gallus
(White Leghorn)

Bos Taurus
(Holstein)

Capra hircus
(Alpine)

- 34 somatic tissues + 13 reproductive tissues (8 female + 5 male):

  liver, sperm, CD4+, CD8+, plasma, heart, lung, skin, fat, duodenum, ileum, jejunum, cerebellum, frontal lobe, olfactory bulb, trigeminal ganglia, hypotalamus, pancreas, andrenals, kidney, muscle, bone, joints, spleen, lymphatic nodes, peyer's patches, ovary, oocytes, oviduct, uterus, mammary gland, acini, testis, seminal vesicle, etc...

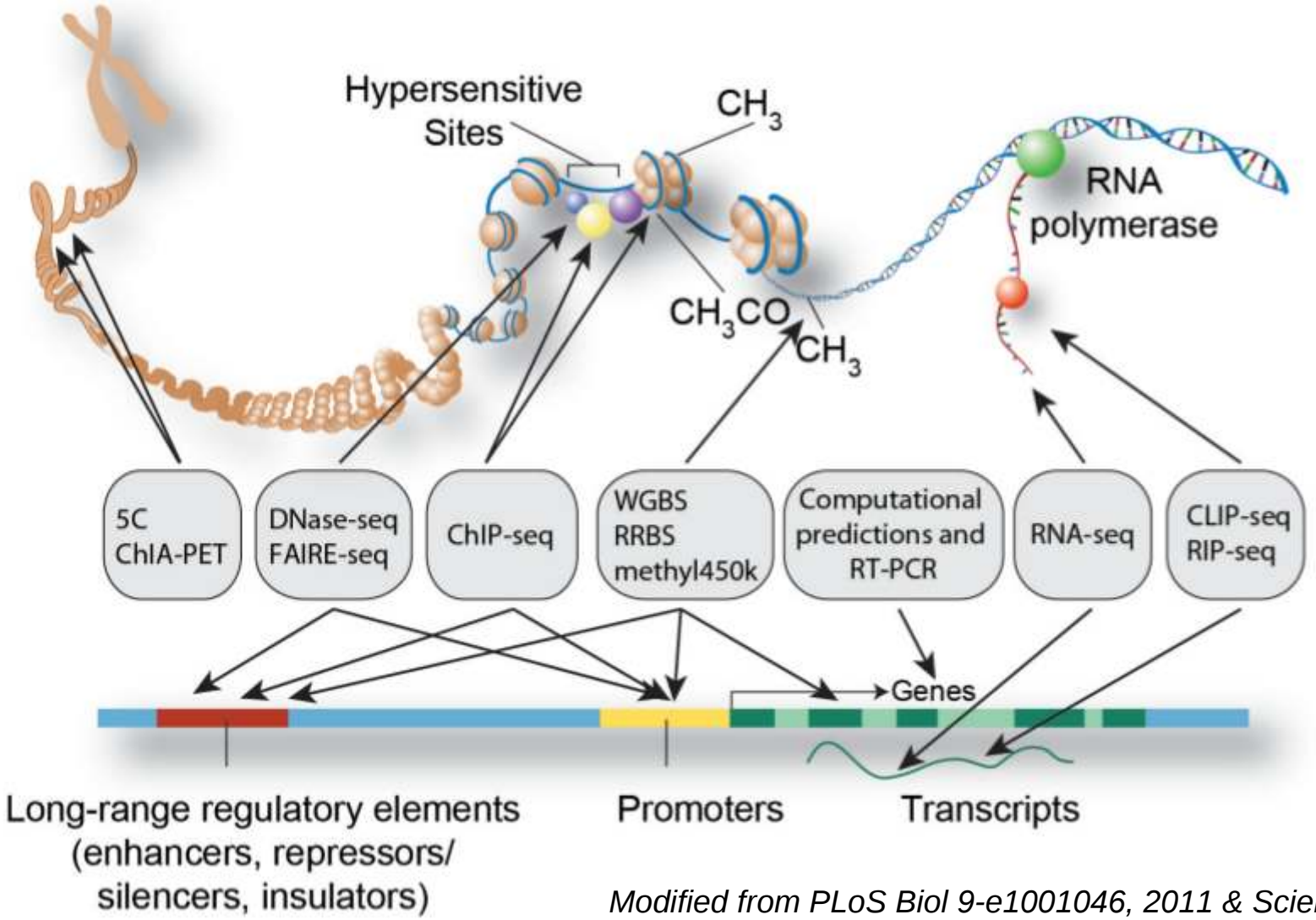
- total: 2,000 to 6,000 samples

M. Tixier-Boichard, INRA GABI

S. Fabre INRA GenPhySE

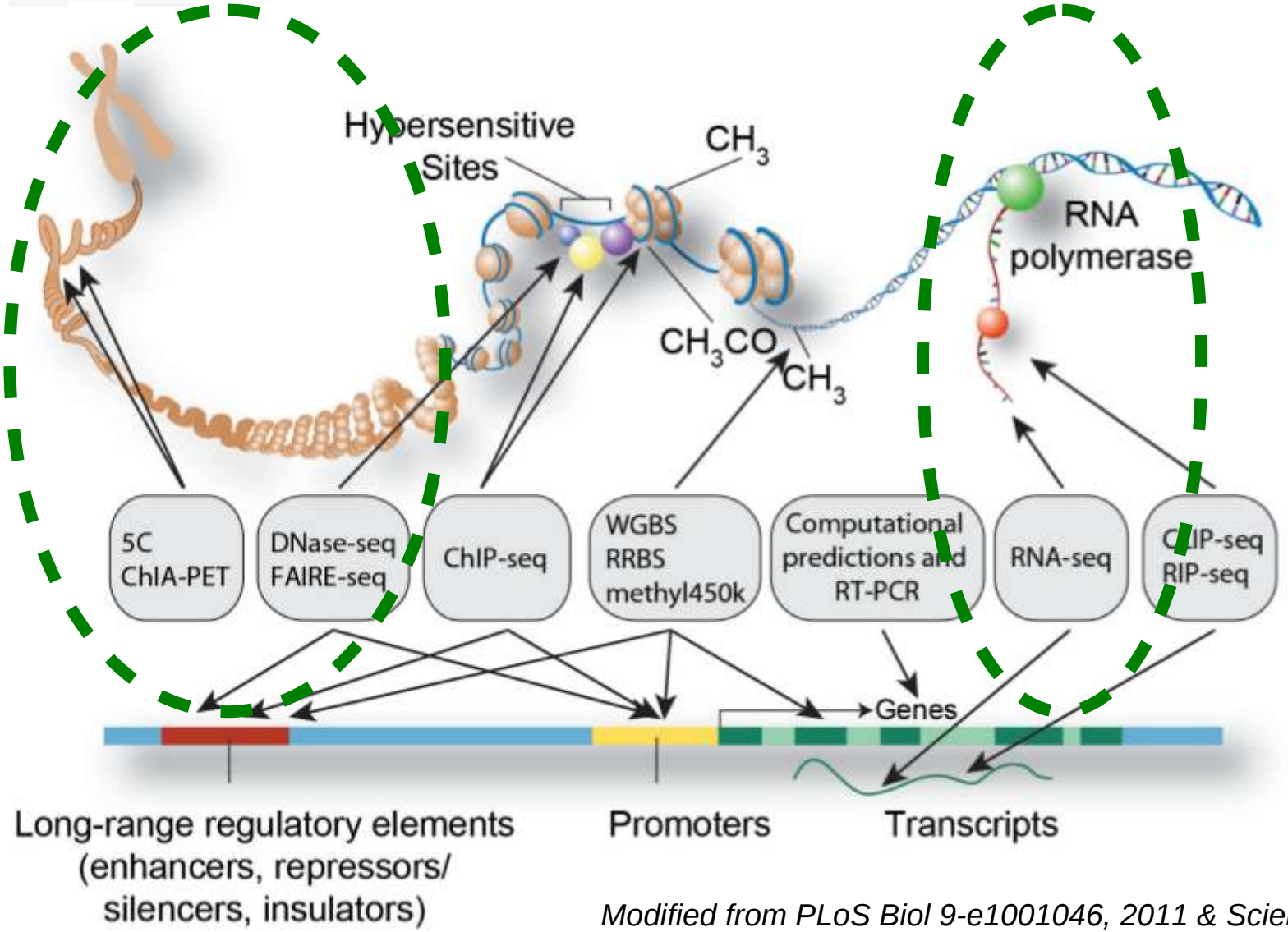
→ **INRA CRB-Anim biorepository**

FR-AgENCODE: molecular assays



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

FR-AgENCODE: molecular assays



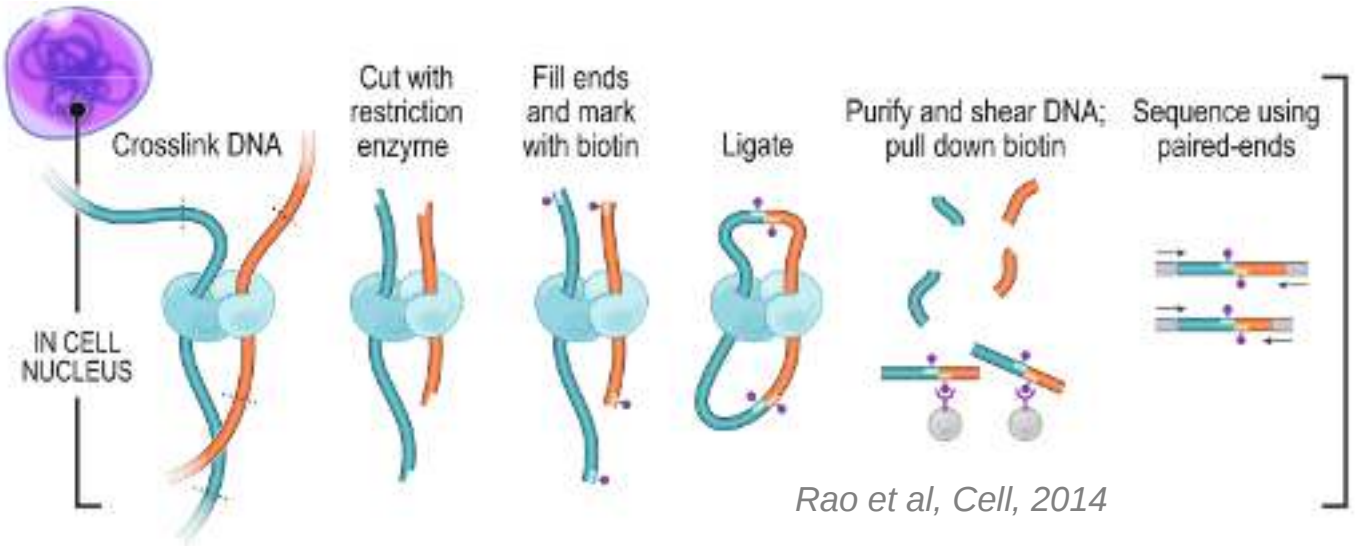
Modified from PLoS Biol 9-e1001046, 2011 & Science 306:636, 2004
Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

FR-AgENCODE: molecular assays

- RNA-seq
  directional protocol, Illumina Hi-Seq3000
  - polyA+ RNAs (mRNAs + lncRNAs)
    100M+ pairs (2x150bp) per sample (2/lane)
  - sRNAs (miRNAs and others)
    35M reads (1x50bp) per sample (3/lane)

FR-AgENCODE: molecular assays

- RNA-seq
  directional protocol, Illumina Hi-Seq3000
  - polyA+ RNAs (mRNAs + lncRNAs)
    100M+ pairs (2x150bp) per sample (2/lane)
  - sRNAs (miRNAs and others)
    35M reads (1x50bp) per sample (3/lane)
- Hi-C: chromosome conformation capture
  (Lieberman-Aiden et al, Science, 2009, Rao et al, Cell, 2014)



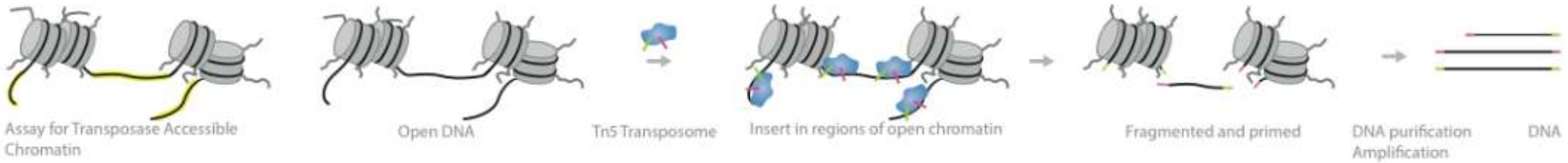
Rao et al, Cell, 2014

FR-AgENCODE: molecular assays

- RNA-seq
  directional protocol, Illumina Hi-Seq3000
  - polyA+ RNAs (mRNAs + lncRNAs)
    100M+ pairs (2x150bp) per sample (2/lane)
  - sRNAs (miRNAs and others)
    35M reads (1x50bp) per sample (3/lane)
- Hi-C: chromosome conformation capture
  - 16 samples (4 replicates, 4 species)
  - 130M read pairs per sample (2 lib/lane)

FR-AgENCODE: molecular assays

- RNA-seq
  directional protocol, Illumina Hi-Seq3000
  - polyA+ RNAs (mRNAs + lncRNAs)
    100M+ pairs (2x150bp) per sample (2/lane)
  - sRNAs (miRNAs and others)
    35M reads (1x50bp) per sample (3/lane)
- Hi-C: chromosome conformation capture
  - 16 samples (4 replicates, 4 species)
  - 130M read pairs per sample (2 lib/lane)
- ATAC-seq: chromatin accessibility



Adapted from: www.illumina.com/techniques

FR-AgENCODE: molecular assays

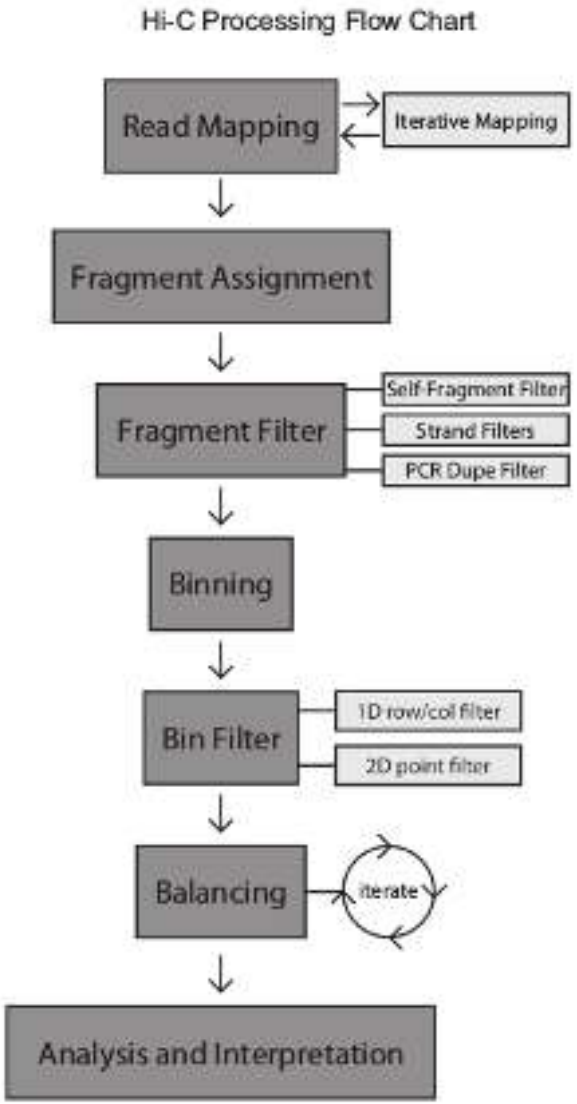
- RNA-seq
  directional protocol, Illumina Hi-Seq3000
  - polyA+ RNAs (mRNAs + lncRNAs)
    100M+ pairs (2x150bp) per sample (2/lane)
  - sRNAs (miRNAs and others)
    35M reads (1x50bp) per sample (3/lane)
- Hi-C: chromosome conformation capture
  - 16 samples (4 replicates, 4 species)
  - 130M read pairs per sample (2 lib/lane)
- ATAC-seq: chromatin accessibility

D. Esquerré
INRA GetPlage

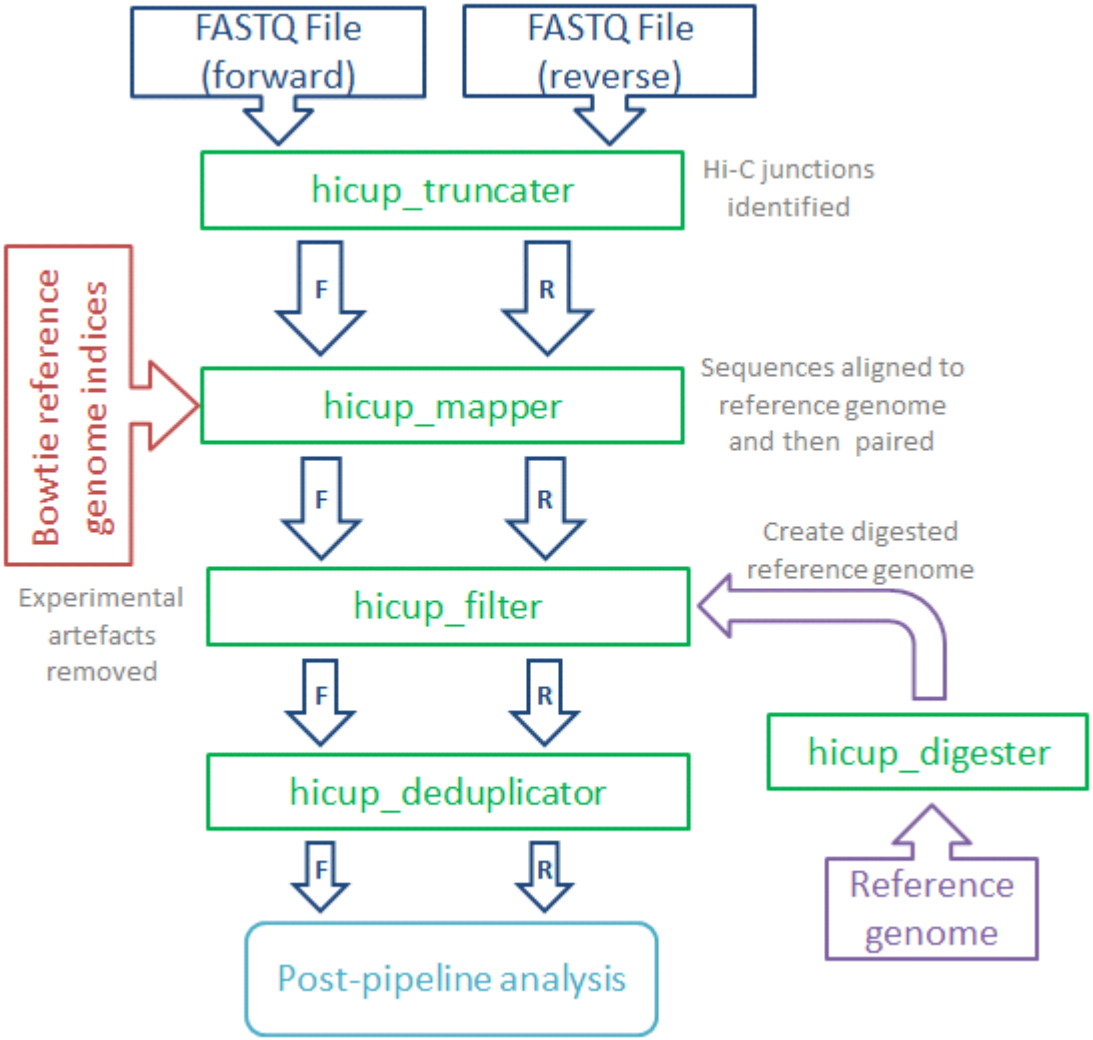
H. Acloque,
INRA GenPhySE

Comparative analysis of genome topology and expression

Hi-C data analysis: overview



Lajoie et al, 2015

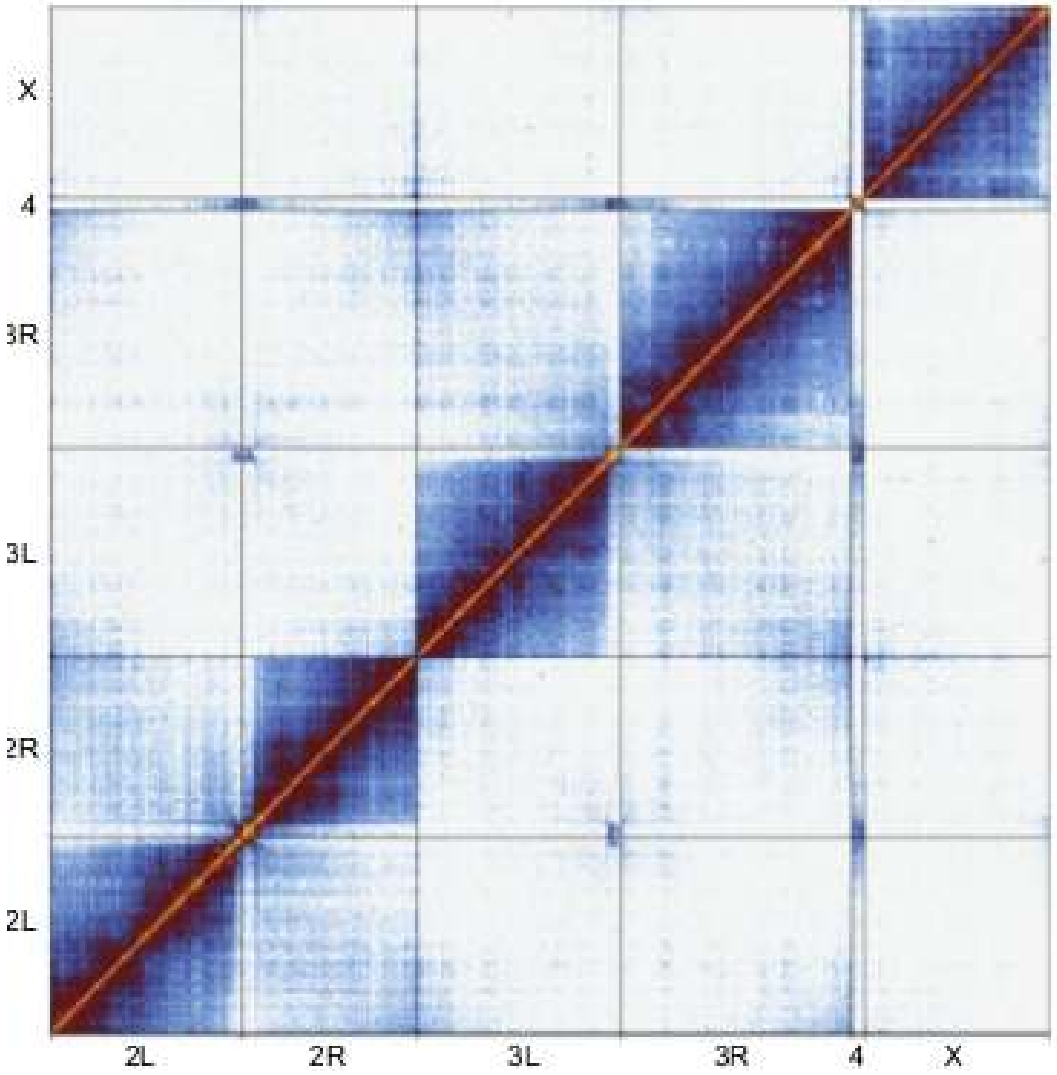


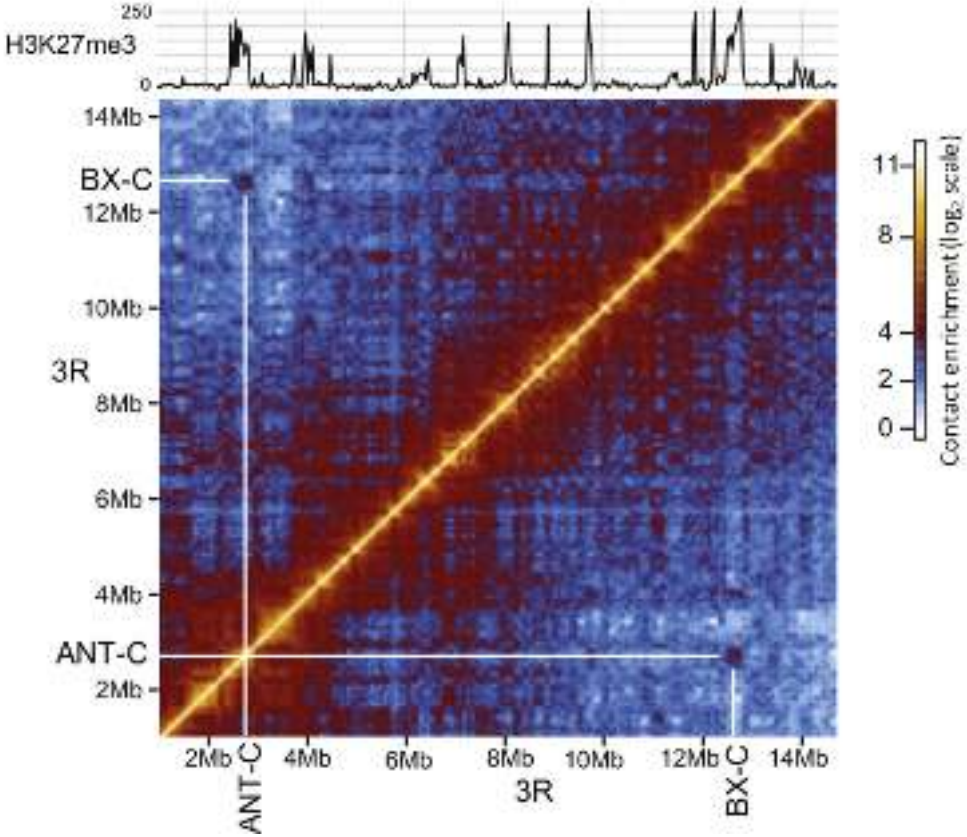
HiCUP, *www.bioinformatics.babraham.ac.uk/*

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: the contact matrix

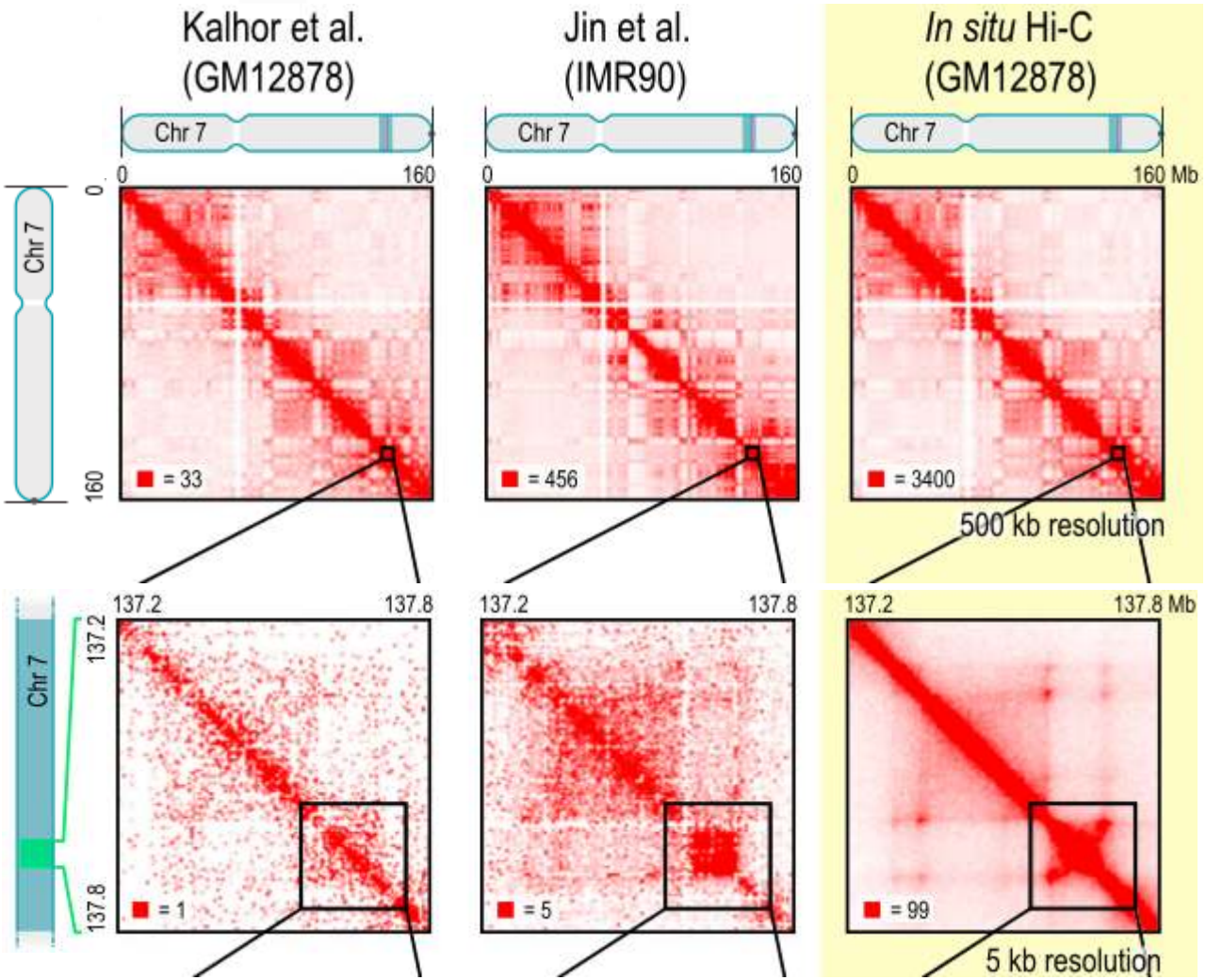


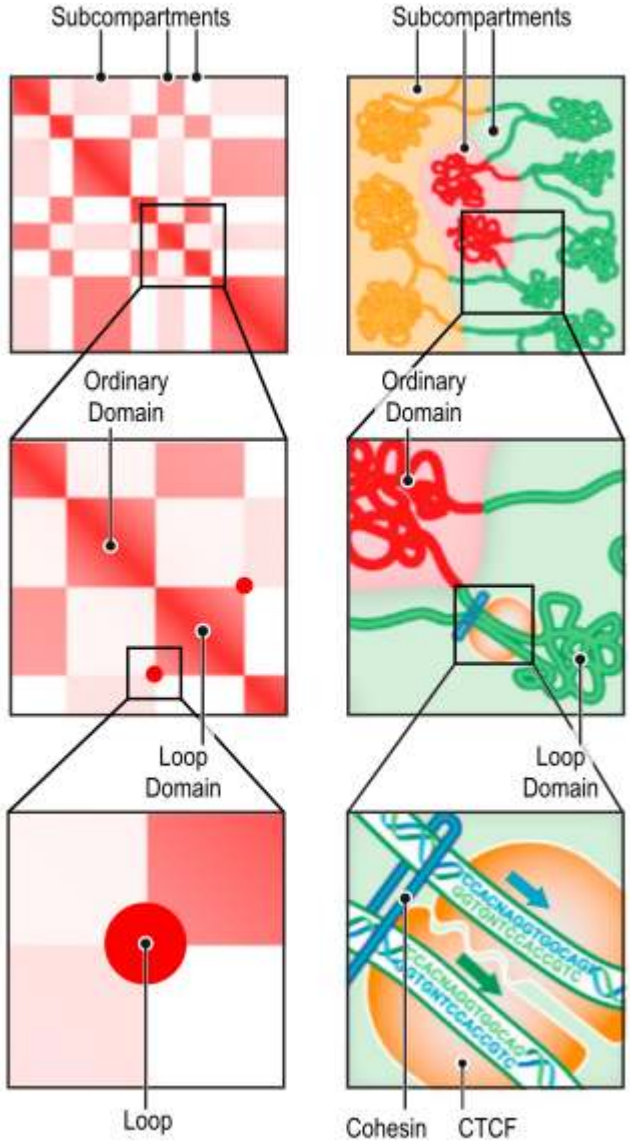


Sexton et al 2012

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

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.

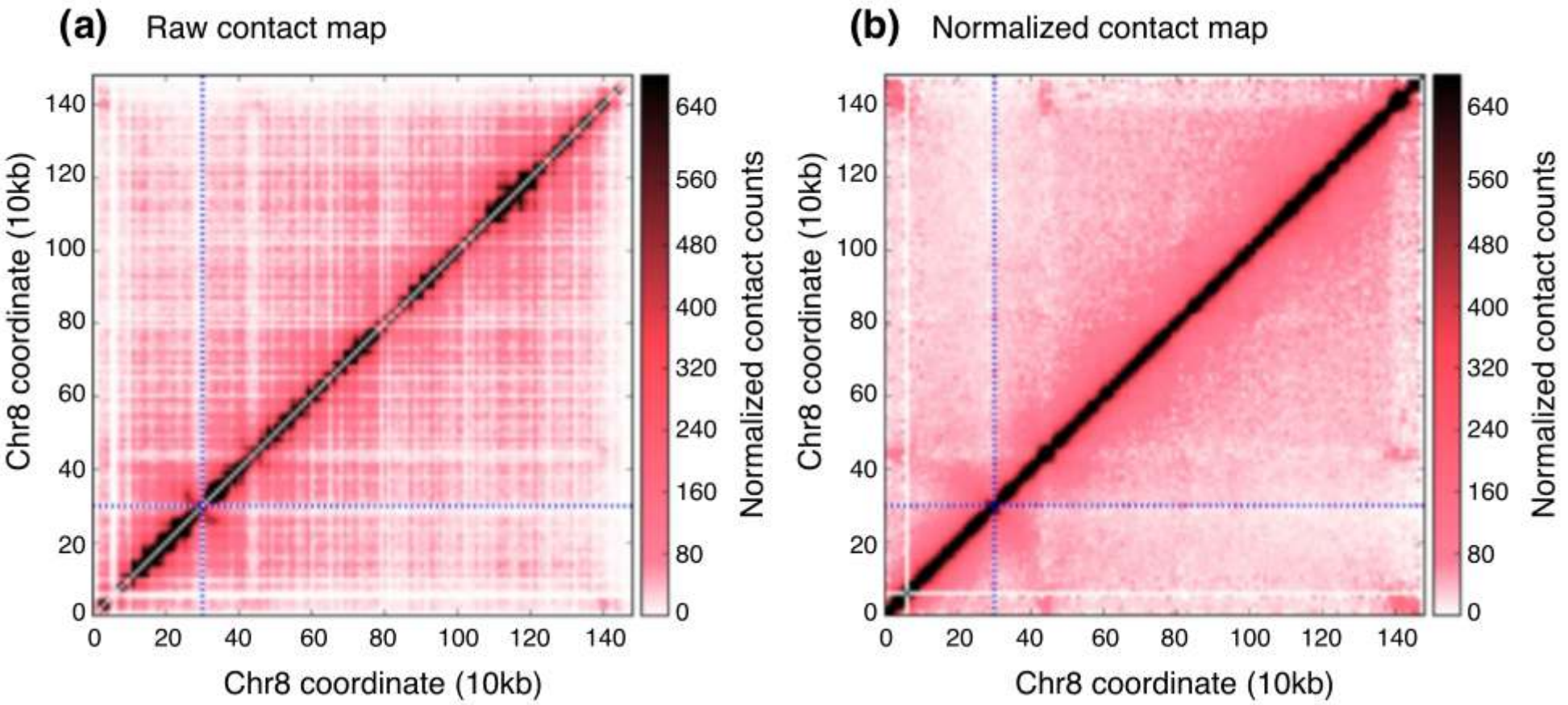
Key words. Matrix balancing, Sinkhorn-Knopp algorithm, doubly stochastic matrix, conjugate gradient iteration.

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

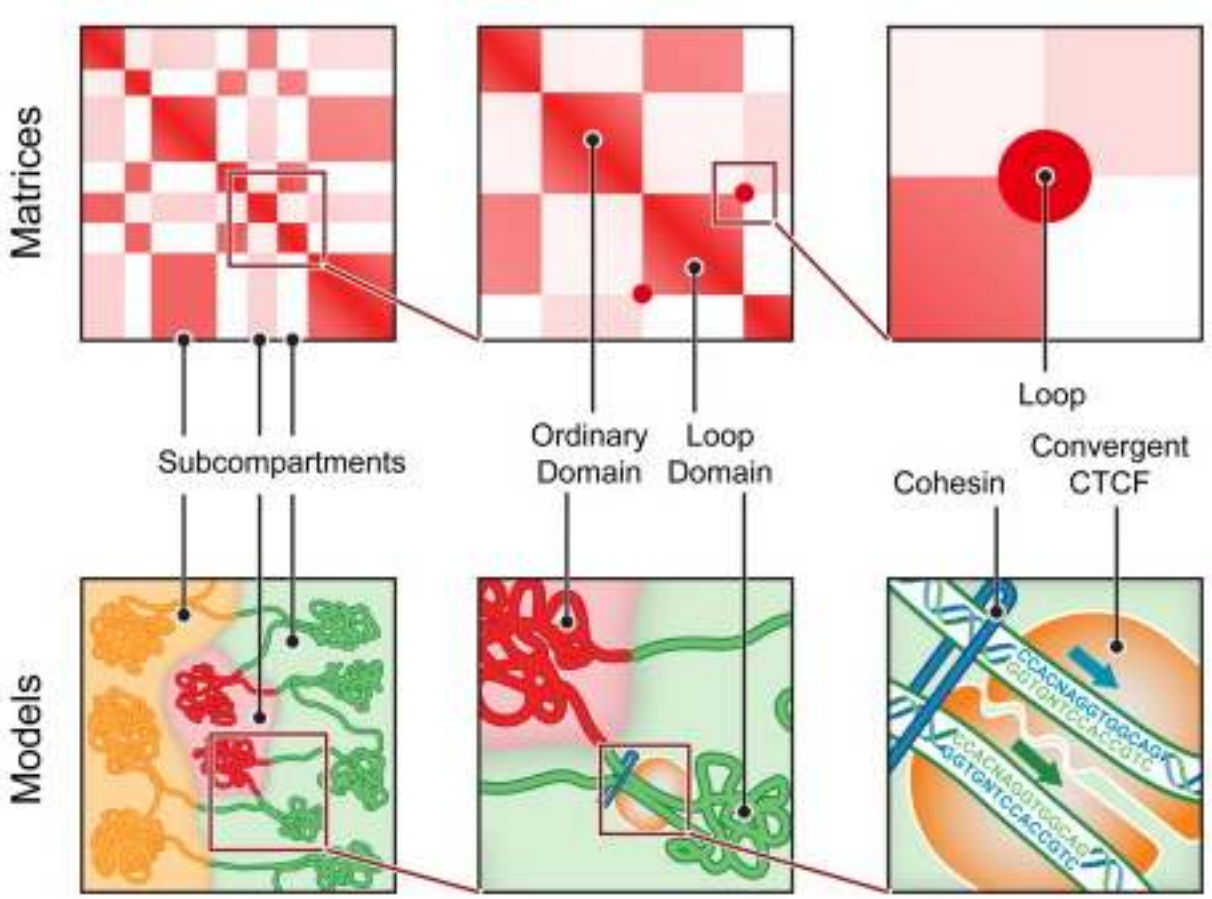
Knight & Ruiz, IMA J. Numer. Anal., 2013

Hi-C data analysis: matrix normalization



Ay & Noble, Genome Biology, 2015

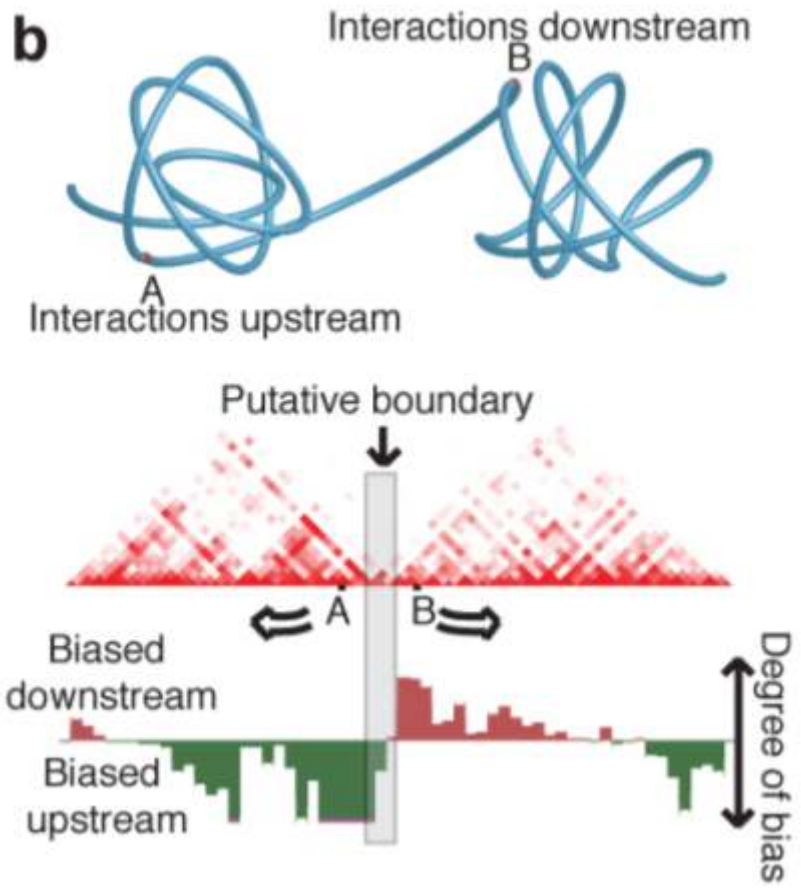
Hi-C data analysis: finding topologically associated domains (TADs)



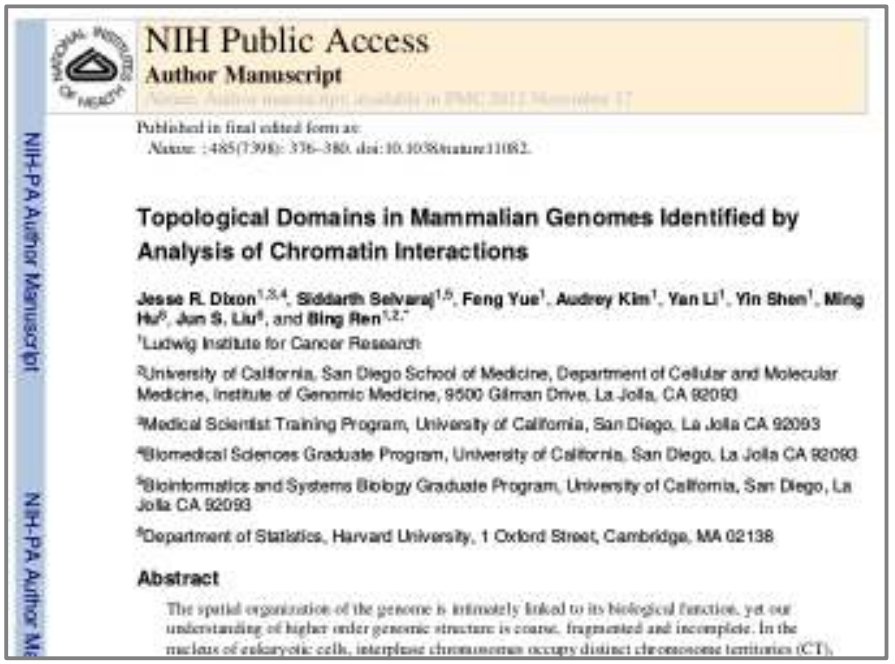
Rao et al, Cell, 2014

- methods: clustering, 2D-segmentation, etc

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



Dixon et al., Nature, 2012



NIH Public Access
Author Manuscript

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[1,2,3,4], 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

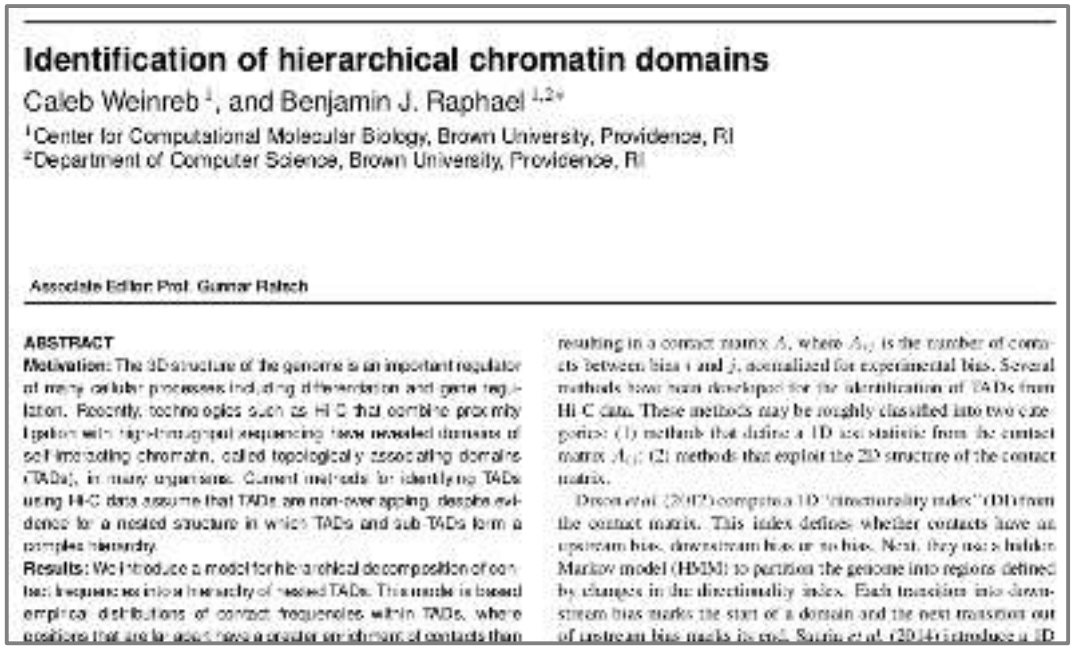
The spatial organization of the genome is intimately linked to its biological function, yet our understanding of higher order genomic structure is coarse, fragmented and incomplete. In the nucleus of eukaryotic cells, interphase chromosomes occupy distinct chromosome territories (CT),

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 Rätsch

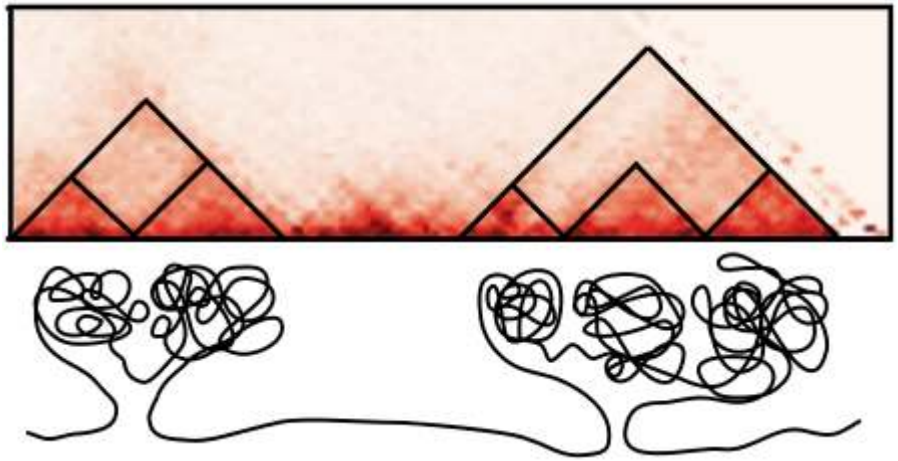
ABSTRACT

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 on empirical distributions of contact frequencies within TADs, where positions that are far apart have a smaller enrichment of contacts than

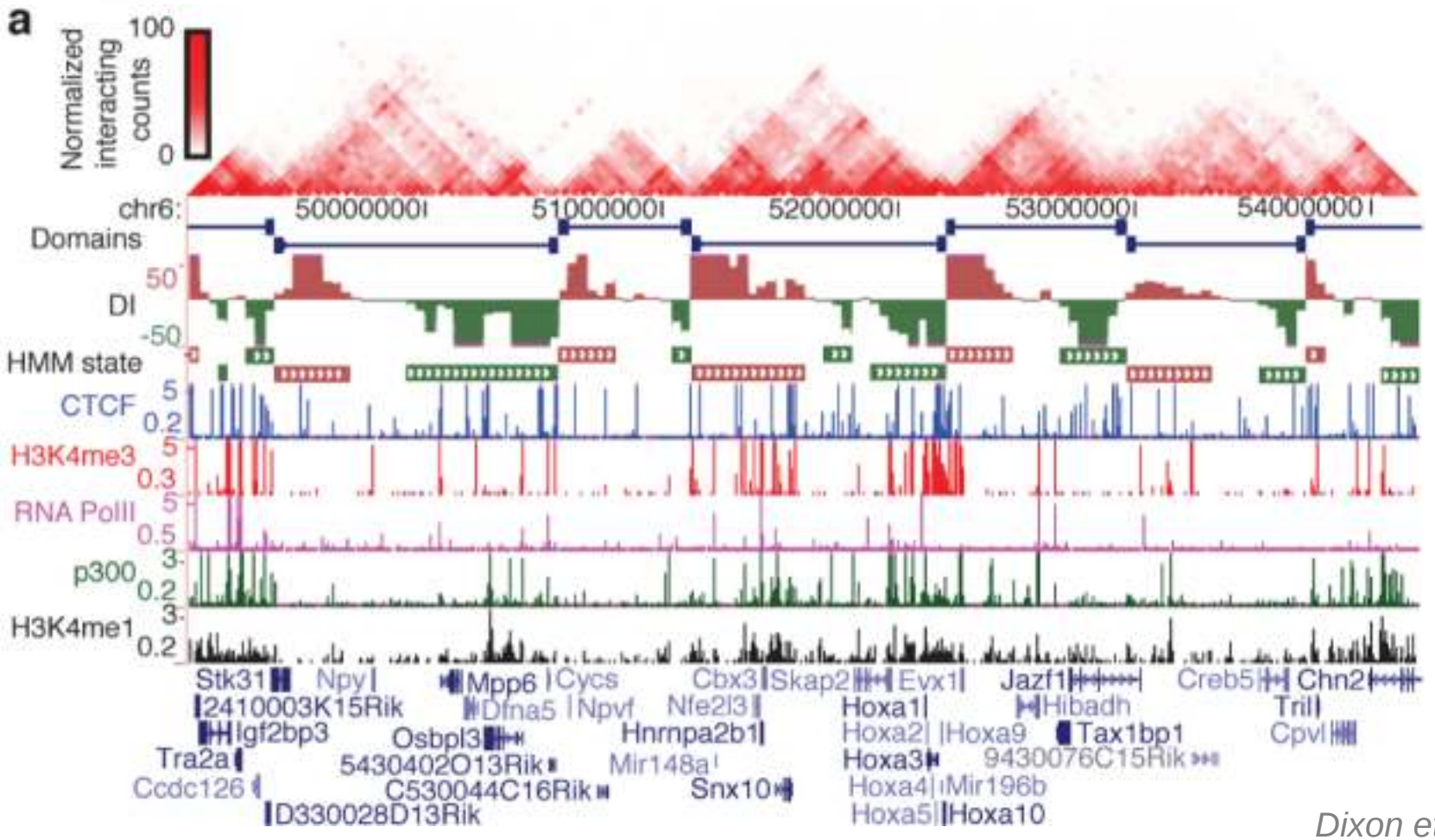
resulting in a contact matrix A, where A_{ij} is the number of contacts between bins i and j, normalized for experimental bias. Several methods have been developed for the identification of TADs from Hi-C data. These methods may be roughly classified into two categories: (1) methods that define a 1D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

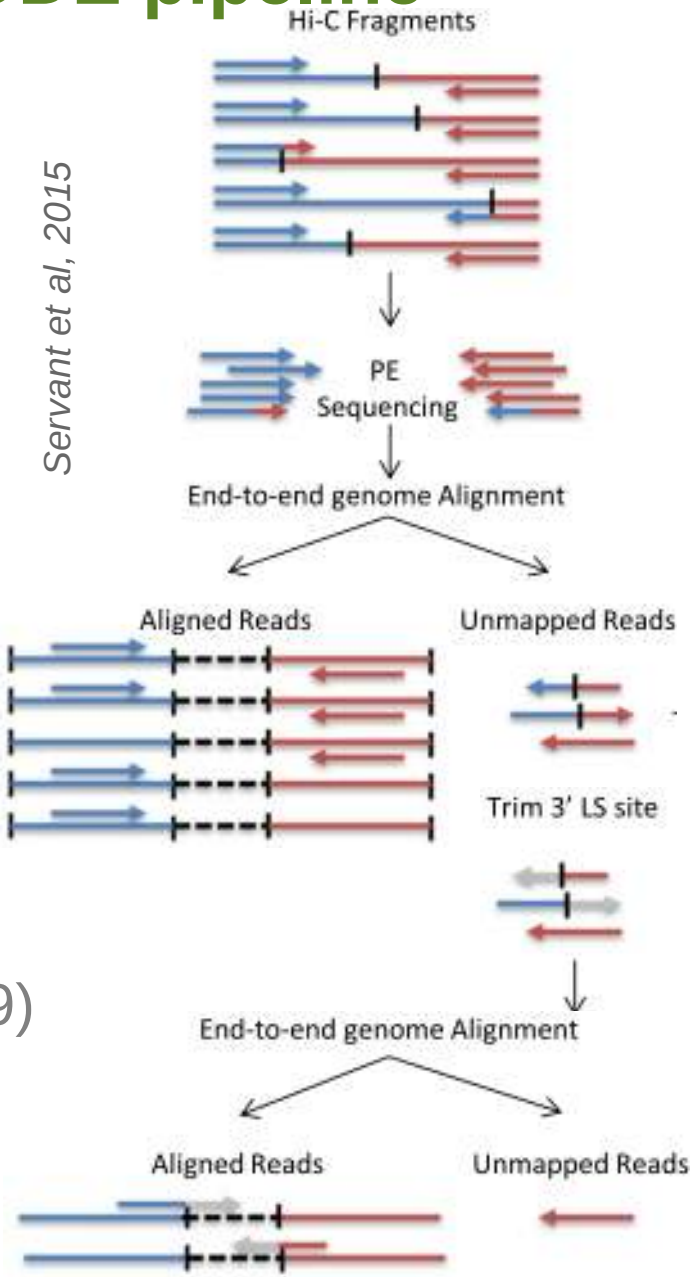
Data analysis

Pipeline

- Trim reads (ligation site)
- Map on reference genome
- Discard inconsistent pairs
- Count reads in pairs of genomic bins & generate contact matrix
- Normalize contact matrix (non parametric, matrix balancing)
- Identify Topologically Associated Domains, *cis* and *trans* interactions

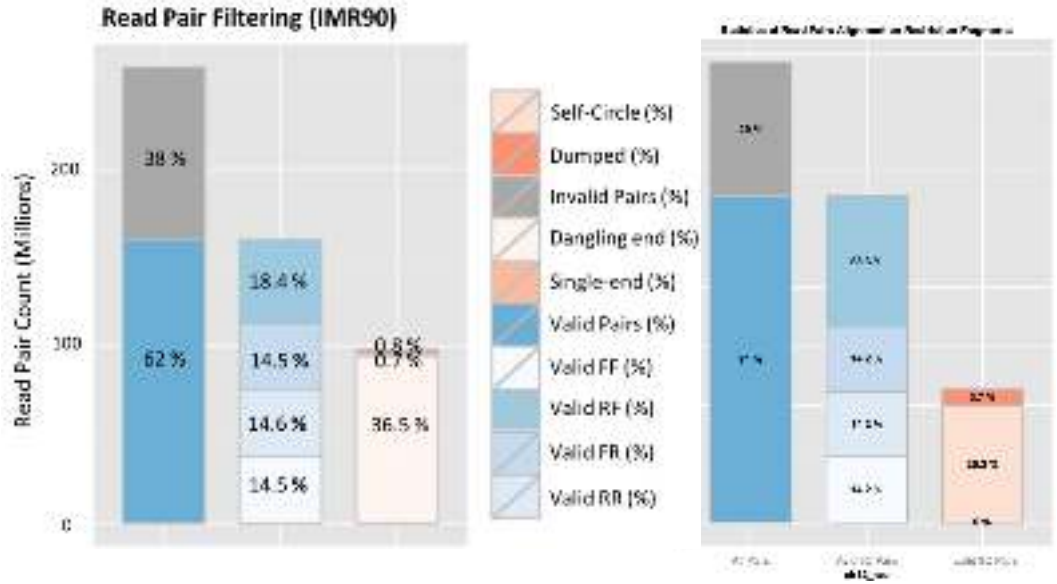
Software

- HiC-Pro pipeline (Servant et al 2015)
- Bowtie2 mapping (Langmead et al, 2009)
- ICE normalization (Imakaev et al, 2012)
- HiTC display (Servant et al, 2012)
- HiFive pipeline (Sauria et al, 2015)



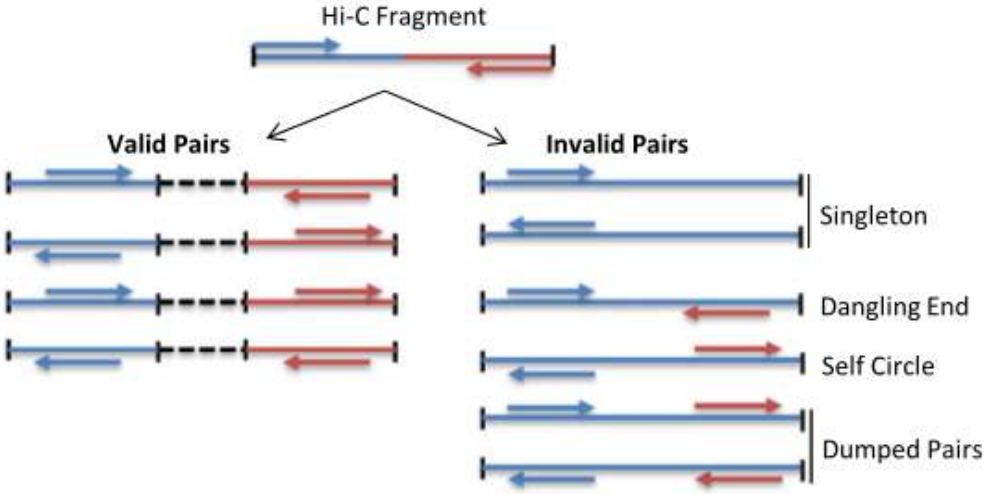
FR-AgENCODE Hi-C preliminary results

Read pairs status after mapping



Dixon et al data
(human, from Servant et al 2015)

Rao et al data
(mouse, CH12 cells)



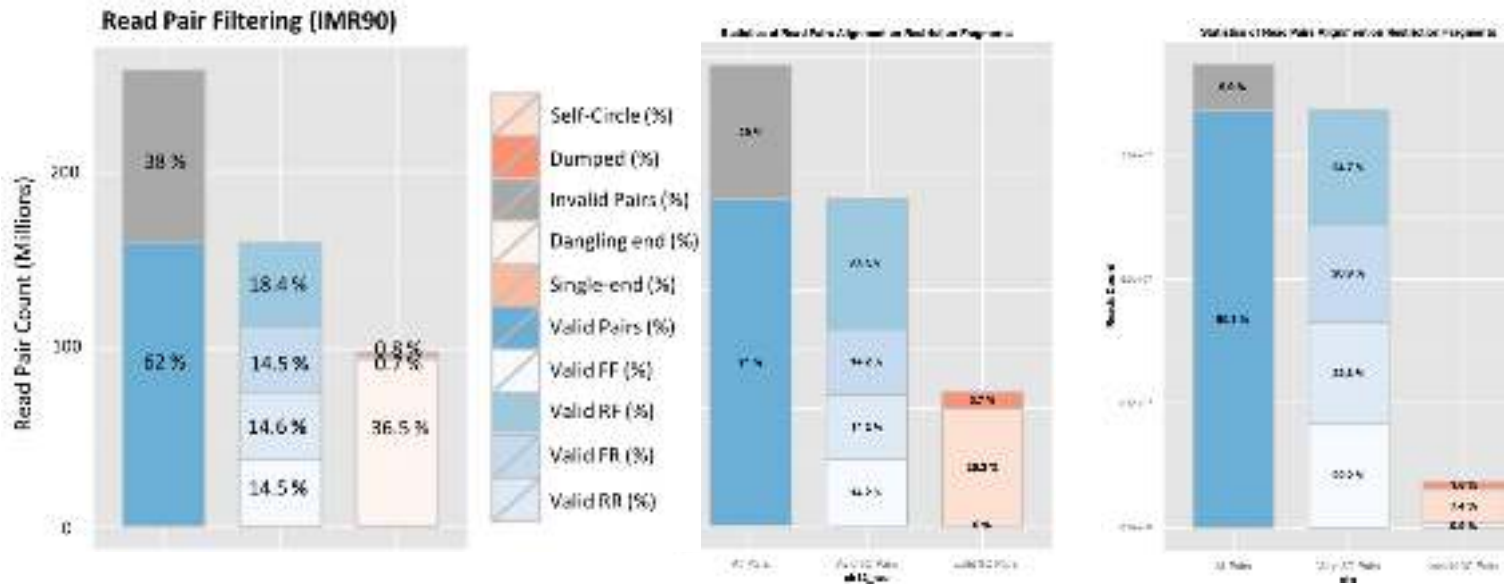
Servant et al 2015

62%
valid pairs

71%
valid pairs

FR-AgENCODE Hi-C preliminary results

Read pairs status after mapping



Dixon et al data
(human, from Servant et al 2015)

62%
valid pairs

Rao et al data
(mouse, CH12 cells)

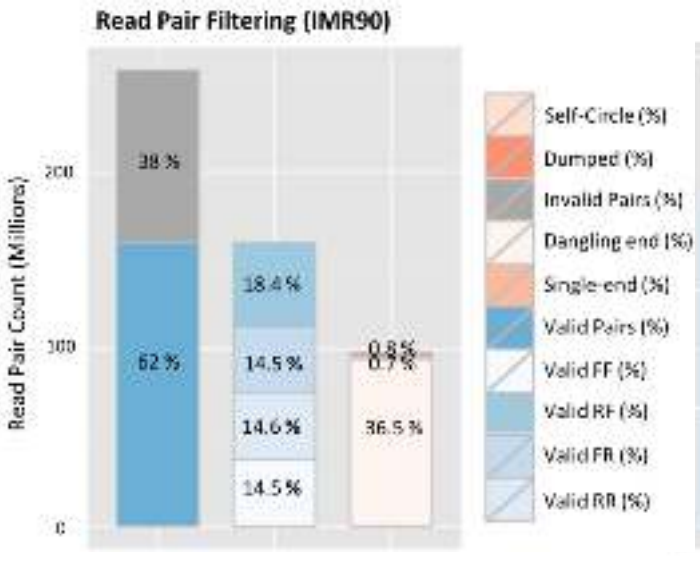
71%
valid pairs

FR-AgENCODE data
(mouse, STO cells)

90%
valid pairs

FR-AgENCODE Hi-C preliminary results

Read pairs status after mapping

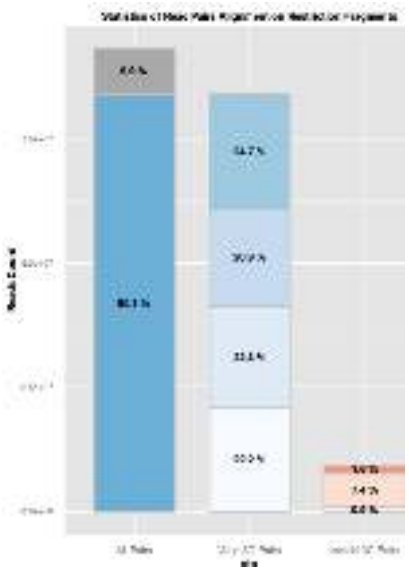


Dixon et al data
(human, from Servant et al 2015)

62%
valid pairs

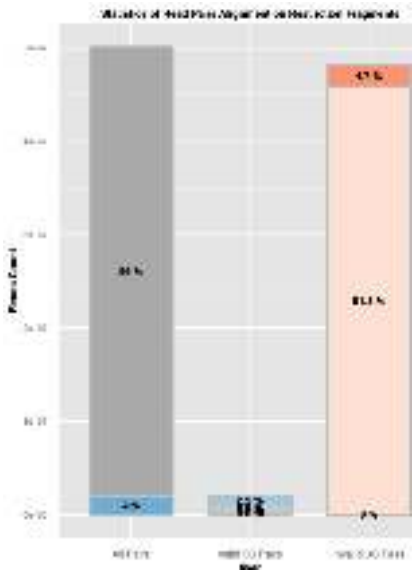
Rao et al data
(mouse, CH12 cells)

71%
valid pairs



FR-AgENCODE data
(mouse, STO cells)

90%
valid pairs

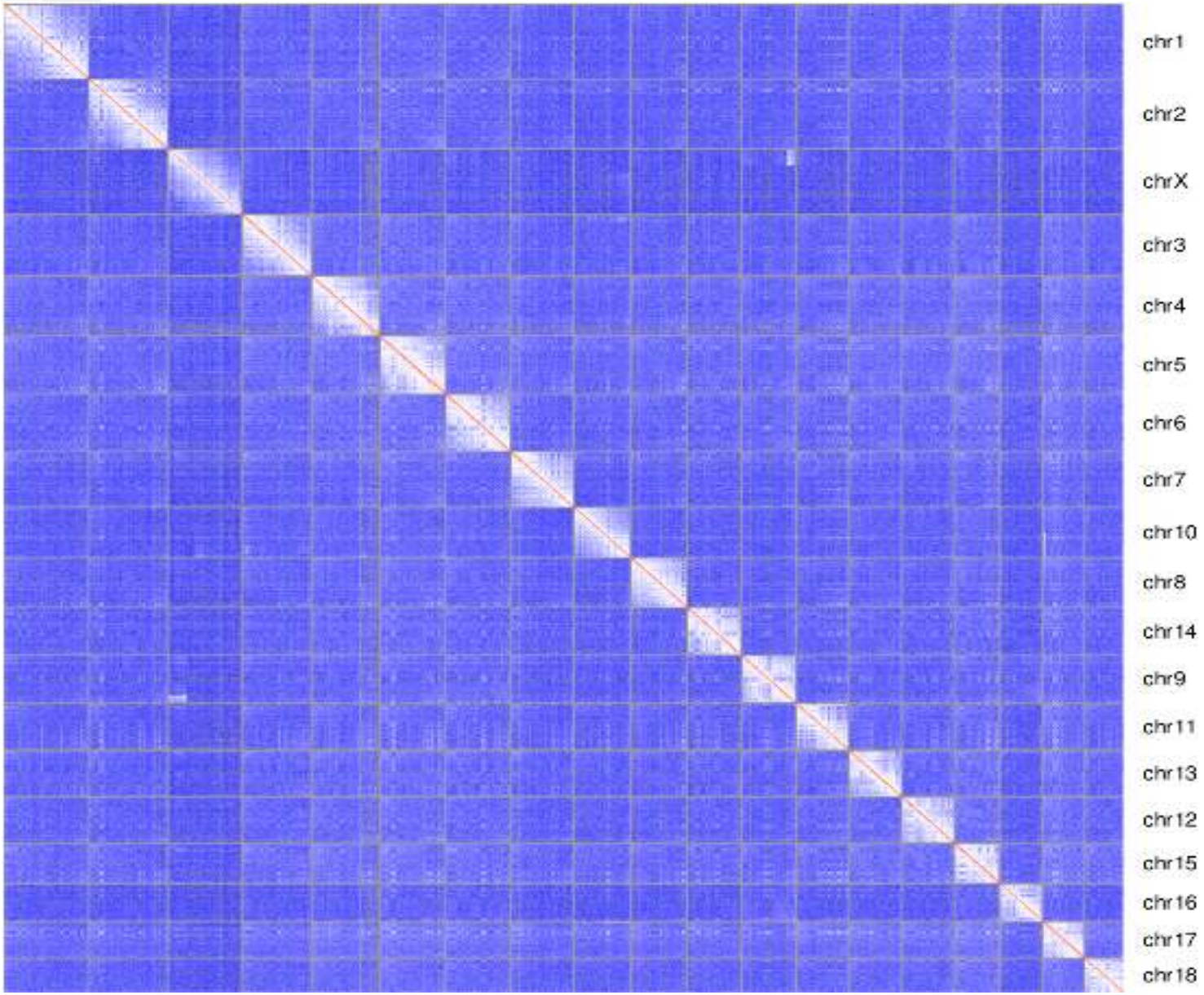


FR-AgENCODE data
(pig, hepatocytes)

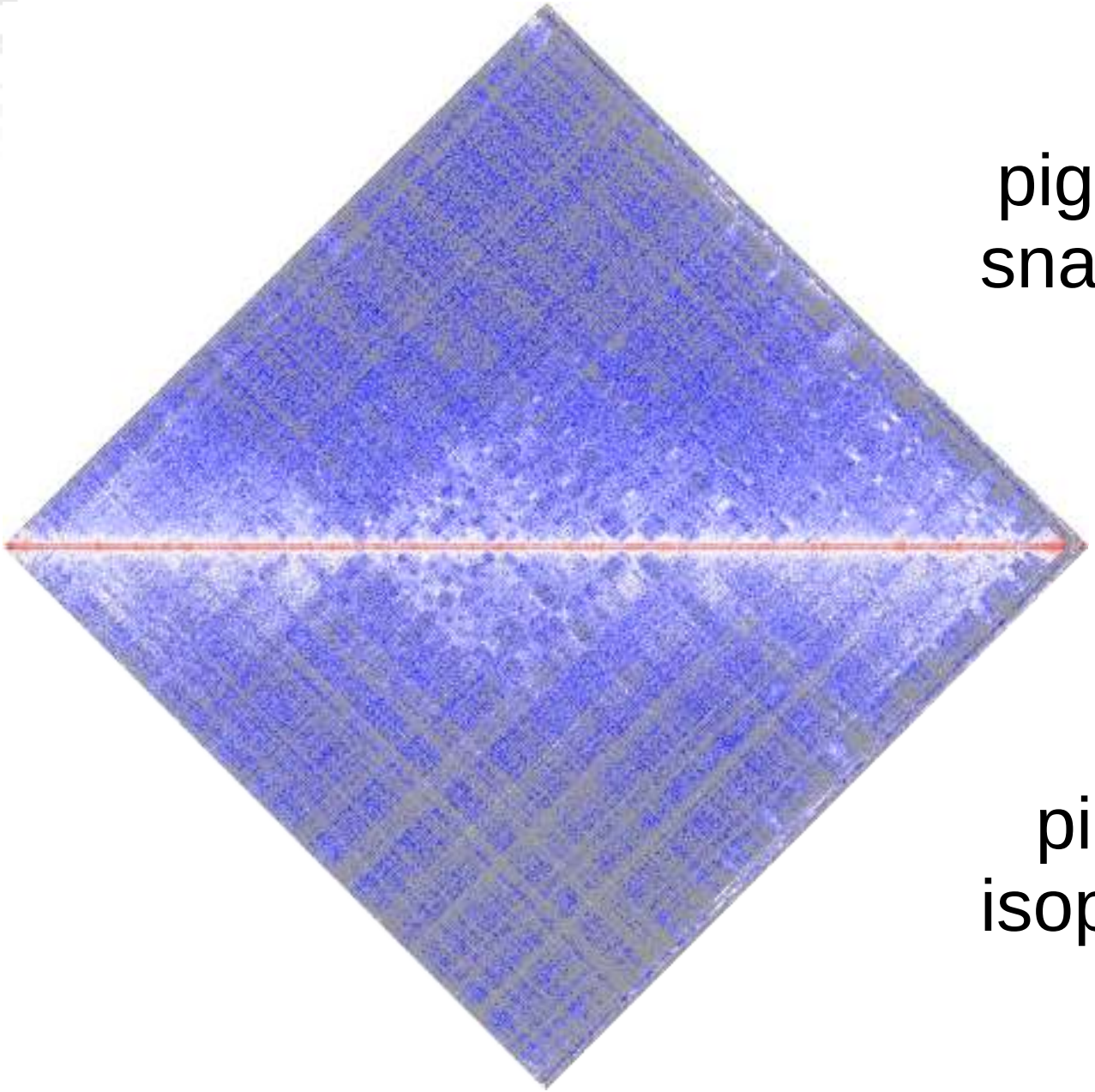
4%
valid pairs

FR-AgENCODE: Hi-C preliminary set-up

Mouse STO cells, whole genome



FR-AgENCODE: Hi-C preliminary set-up

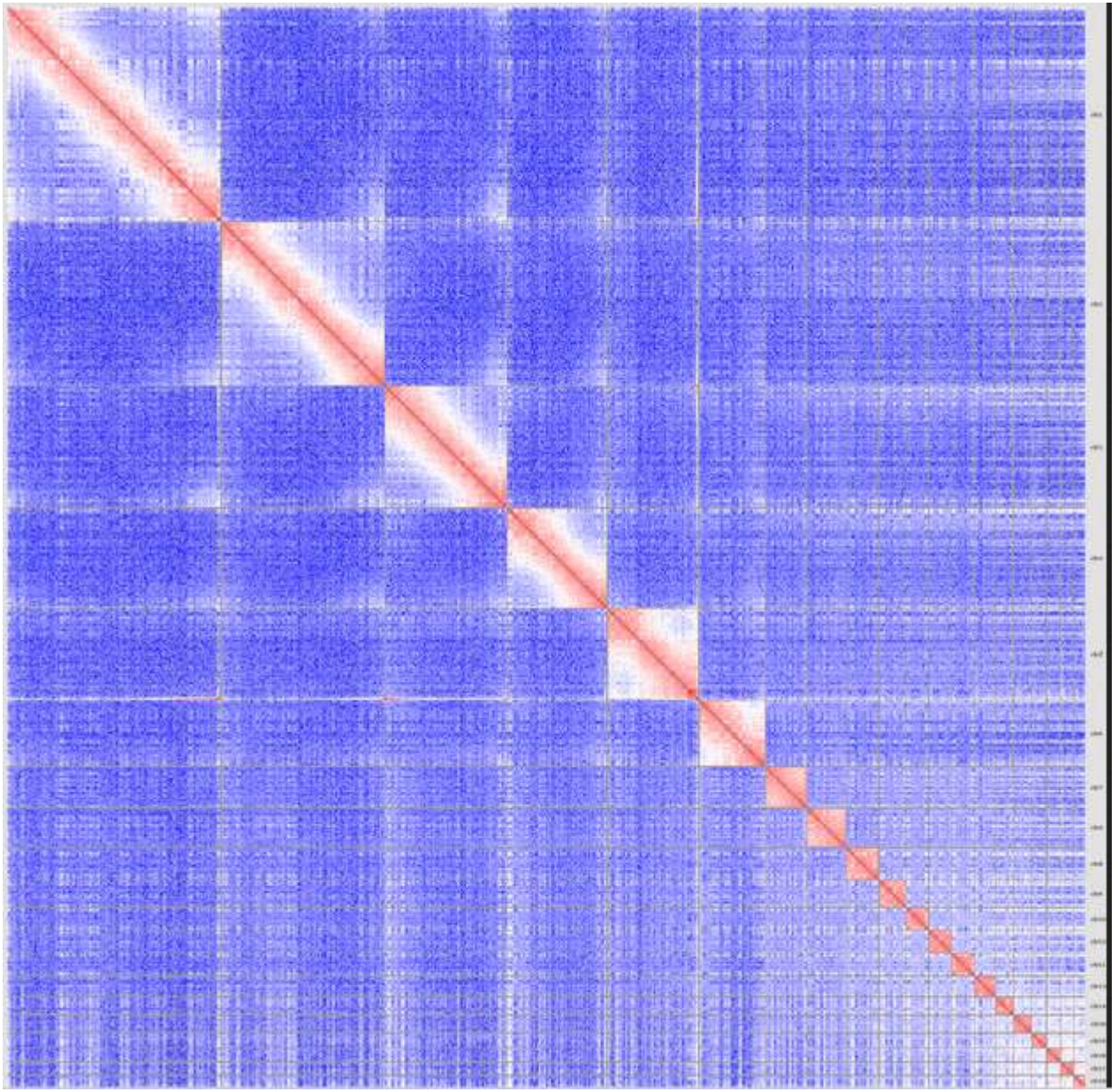


pig muscle,
snap frozen,
chr1

pig liver,
isopentane,
chr1

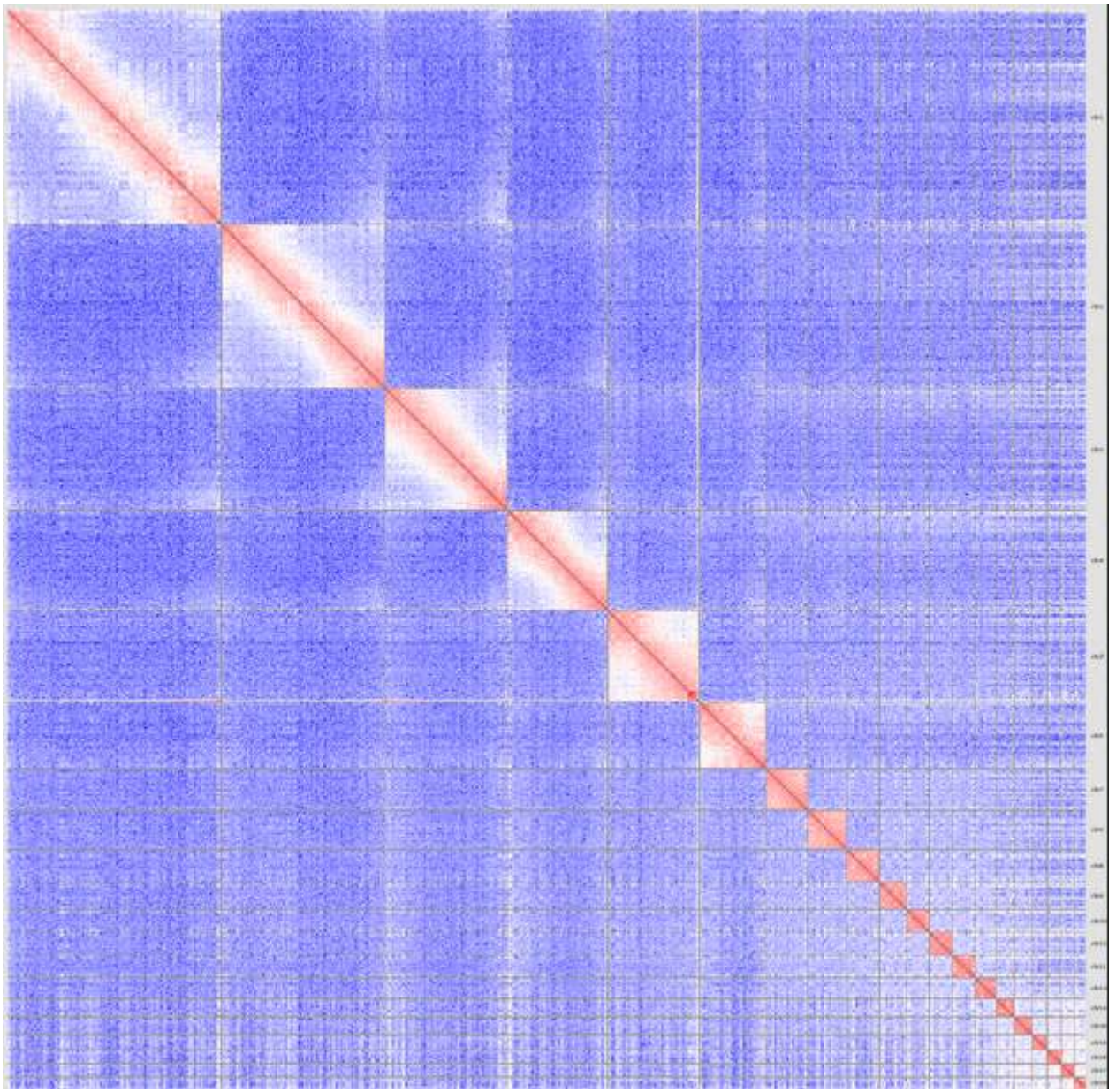
FR-AgENCODE: Hi-C preliminary results

chicken liver, animal #2



FR-AgENCODE: Hi-C preliminary results

chicken liver, animal #4



FR-AgENCODE: data management and analysis

- Data production
  - RNA-seq: 2.2T
  - Hi-C: 0.8T
  - ATAC-seq: 1.1T

  => Expected total: > 3T of raw sequence data
- Data storage
  - GenoToul ng6
  - EMBL-EBI (FAANG rapid data release policy)
- Data analysis
  - INRA units
  - EMBL-EBI (FAANG analysis pipelines)

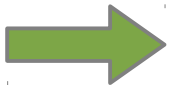
FR-AgENCODE in FAANG

Bioinformatics and Data Analysis Committee

- aim: define standard pipelines
- Working Groups
  - transcriptome: RNA-seq, lncRNA-seq, sRNA-seq
  - regulation: ChIP-seq
  - methylation: WGBS, RRBS
  - chromatin structure: Hi-C, DNAse-seq, ATAC-seq
- activities: teleconferences, seminars, hackatons
  - identification of reference datasets
  - list and benchmark tools
  - publicly report

Summary / Conclusion

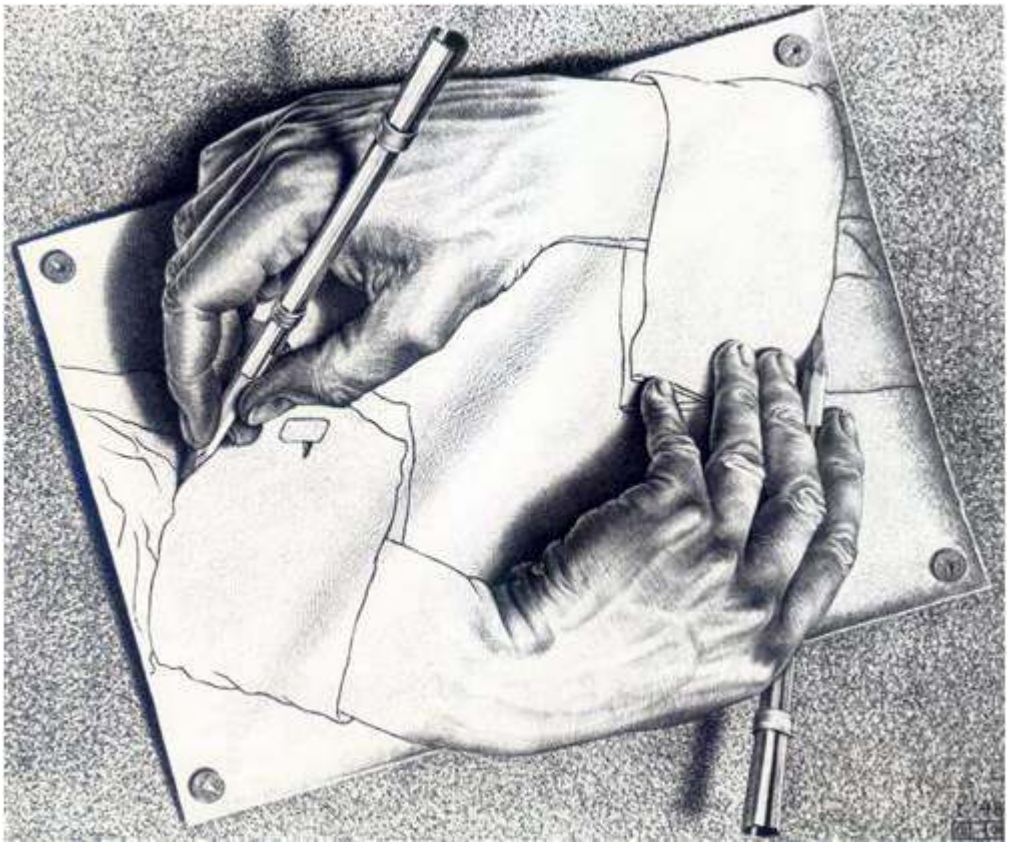
- Hi-C allows to capture 3D conformation of the chromatin
- INRA protocol for tissue samples from livestock species
- INRA contribution to the FAANG action and the Genome to Phenome challenge
  - FR-AgENCODE pilot project
  - Samples and Assays committee
  - Bioinformatics and Data Analysis committee
- Success story incoming?



To Be Continued...

Outline

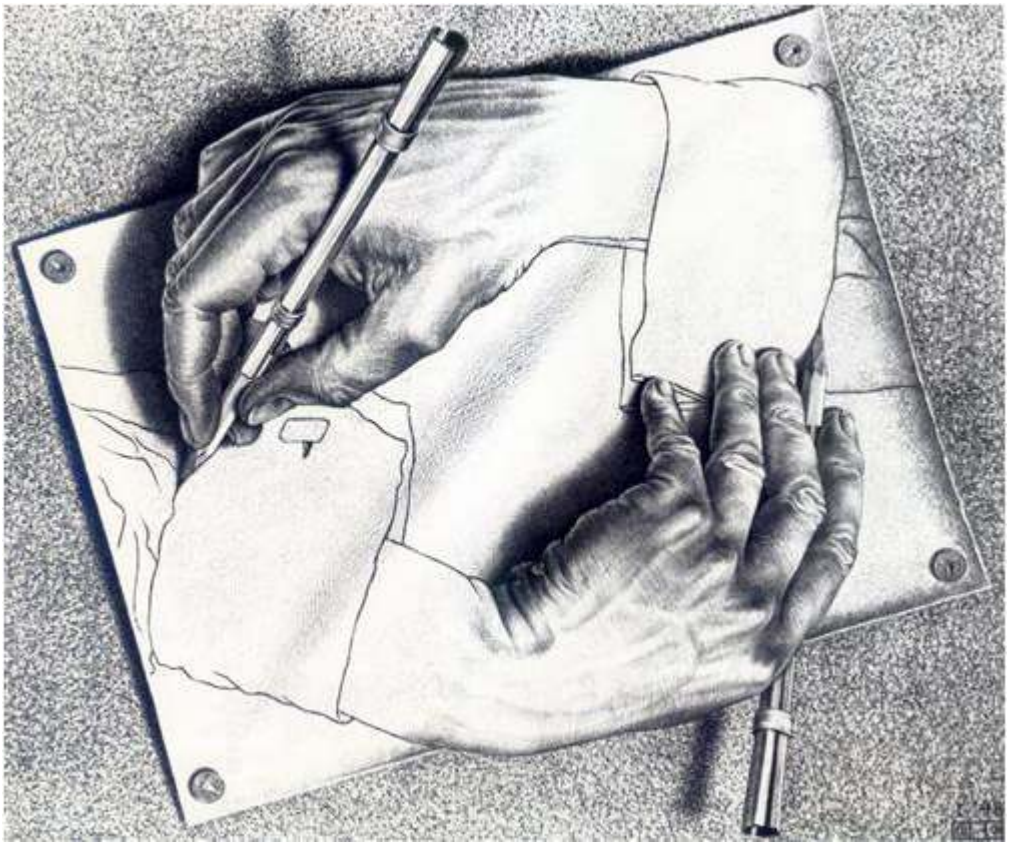
- Why
- What
- How
- Where
- Who



M.C. Escher, 1948

Outline

- Why
- What
- How
- Where
- Who



M.C. Escher, 1948

Acknowledgments

FR-AgENCODE members

- Management: Elisabetta Giuffra (coordination), Sandrine Lagarrigue, Marie Hélène Pinard
- Sampling: Michèle Tixier-Boichard, Stéphane Fabre et al.
- Assays: **Diane Esquerré, Hervé Acloque** et al.
- Analysis: Christophe Klopp, Christine Gaspin, **David Robelin, Matthias Zytnicki, Sarah Djebali, Magali San Cristobal, Ignacio Gonzalez,** Kylie Munyard, Céline Noirot, Nathalie Villa Vialaneix, Gaelle Lefort, Marjorie Mersch, Frédérique Pitel et al.

Hi-C team @ GenPhySE, INRA

- H. Acloque, M. Yerle, Y. Lahbib, F. Mompart, M. Marti et al.

FAANG B&DA committee

- L. Clarke, D. Zerbino, J. Reecy, P. Ross, L. Eory, M. Watson et al.