Three-dimensional organization of genomes: interpreting chromatin interaction data

3C to hi-C

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

Genomics: Lecture #15

Outline

1. Gene Regulation in Eukaryotes
2. 3D Organization of Genomes
3. hi-C
4. Normalizing Hi-C Data
5. Poisson regression: GLMs
6. HiCNorm

The Purpose of Gene Regulation

Housekeeping genes are typically constitutive genes that are required for the maintenance of basic cellular function, and are expressed in all cells of an organism under normal and patho-physiological condition

Examples: Genes involved in . . .

Carbohydrate metabolism

Citric acid cycle

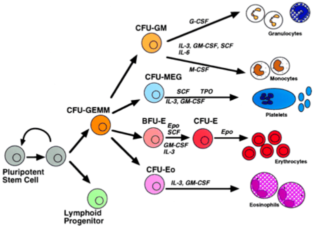
Cytoskeleton

. . .

Gene Regulation

Some genes are expressed only when a cell **enters a particular pathway of differentiation**.

Bone marrow pluripotent stem cells in the hematopoietic compartment may either self-renew or give rise to eight different hematopoietic lineages through a gradual process of commitment and differentiation.

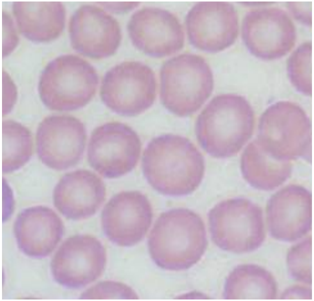


Socolovsky M et al. (1998) *PNAS* **95**:6573-5.

Gene Regulation

Some genes are expressed are expressed constitutively in only those cells that have differentiated into a particular type of cell

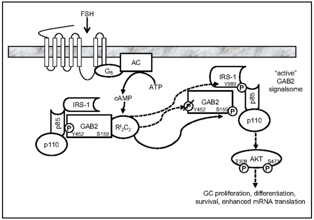
Beta-globin and alpha-globin make up HbA, the most common form of hemoglobin in adults. Expression of this gene is highly specific for the erythrocyte lineage



Gene Regulation

Some genes are expressed are expressed in response to stimuli.

Ovarian follicles are restrained at an immature stage until stimulated by Follicle-stimulating hormone (FSH) secreted by pituitary gonadotropes. FSH regulates the development, growth, pubertal maturation and, reproductive processes of the body. AKT is an essential downstream effector of the FSH signal.



Hunzicker-Dunn ME et al. (2012) *PNAS* **109**:E2979-88.

Gene Regulation

There are many methods by which the cell can regulate gene expression

- Alter rate of transcription of a gene (probably most important)
- Alter rate of processing of the RNA transcripts, including splicing
- Alter the stability of the mature mRNA
- Alter the rate of translation of the mRNA into polypeptide
- (Alter the stability and activity of the protein)

Gene Regulation

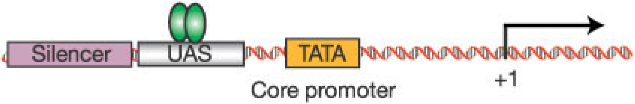
Today we are going to talk about regulation of the rate of transcription of genes by means of long-range three dimensional interactions between promoters and enhancers.

The following terminology is not used in a consistent way in the literature, but we will give reasonable definitions of the following elements that are the main "actors" in our story.

- Transcription start site
- The core/upstream promoter
- enhancers
- silencers

Simple eukaryotic transcriptional unit

A simple core promoter (TATA), upstream activator sequence (UAS) and silencer element spaced within 100–200 bp of the TATA box that is typically found in unicellular eukaryotes.

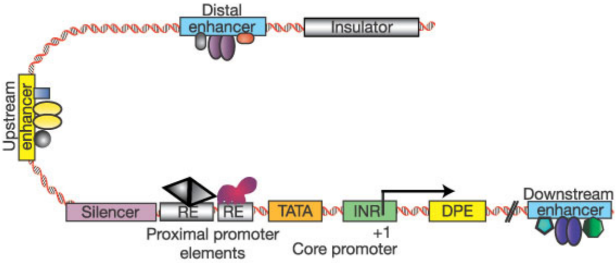


Organismal complexity correlates with an increase in both the ratio and absolute number of transcription factors per genome.

Organism	Genes	Transcription Factors	Ratio TF:Gene
Yeast	$\sim$ 6275	$\sim$ 300	1:20
Human	$\sim$ 20,000	$\sim$ 3000	1:7

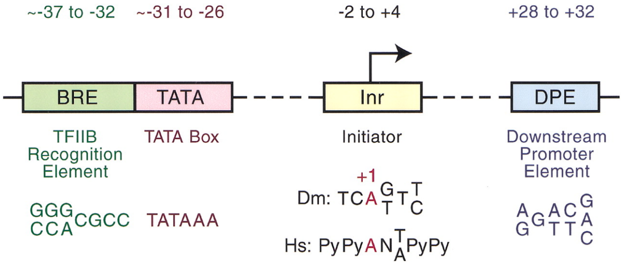
Simple eukaryotic transcriptional unit

A complex arrangement of multiple clustered enhancer modules interspersed with silencer and insulator elements which can be located 1050 kb either upstream or downstream of a composite core promoter containing TATA box (TATA), Initiator sequences (INR), and downstream promoter elements (DPE).



Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* **424**:147-51.

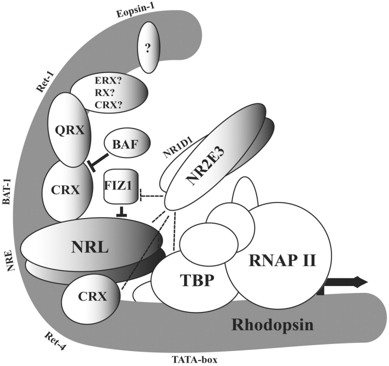
The core promoter



Butler JE, Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16**:2583-92.

- The core promoter (aka: "basal" promoter) is located within about 40 base pairs (bp) of the start site
- Found in all protein-coding genes, relatively similar, bound by a large complex of about 50 proteins including Transcription Factor IID (TFIID). itself a complex of TATA-binding protein and 13 other proteins,
- Transcription Factor IIB (TFIIB) which binds both the DNA and Pol II.

The upstream (regulatory) promoter

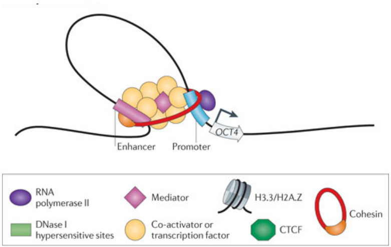


Cheng H,et al (2004) *Hum Mol Genet* **13**:1563-75.

- The "upstream" promoter, which may extend over as many as 200 bp farther upstream and that greatly differ from gene to gene.
- Cis-regulatory elements upstream of the transcription start site in the rhodopsin promoter include Eopsin-1, Ret-1, BAT-1, NRE, Ret-4 and TATA-box.

Enhancers

Enhancers are short DNA sequences with binding sites for several transcription factors that are largely responsible for specificity of gene expression patterns in a given cell. Enhancers may be located thousands (or tens/hundreds of thousands) of base pairs away from the gene they control.

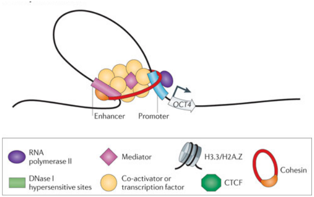


Nature Reviews | Genetics

Ong CT, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**:283-93.

Enhancer characteristics

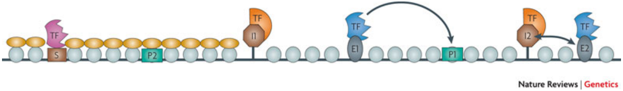
- Cohesin may facilitate enhancer–promoter interaction by bringing them into close proximity
- Mediator (a $\sim$ 30 subunit complex) coordinates signals between enhancers and general transcription factors by interaction between RNAP2 and site-specific factors
- Histone modifications mark active enhancers is specific cell types
- Enhancers themselves may be transcribed (eRNA), the function of eRNAs is still largely unknown
- Enhancers/Promoters interact with other non-coding RNAs to implement gene regulatory programs



Nature Reviews | Genetics

Silencers and Insulators

Silencer elements (S) are sites of initiation of heterochromatin, which spreads and encompasses promoters (P2 in the diagram), silencing transcription.

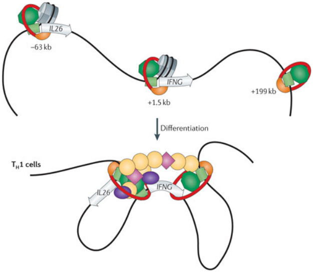


Raab JR, Kamakaka RT (2010) Insulators and promoters: closer than we think. *Nat Rev Genet* **11**:439-46.

- The I1 **insulator** functions to restrict the spread of heterochromatin.
- An enhancer (E1) that is present in an active chromatin domain flanked by insulators (I1 and I2) and that is bound by a transcription factor (TF) is able to communicate with a promoter (P1) in the same domain
- Another enhancer (E2) is unable to communicate with promoter P1 because of an intervening insulator (I2).

Enhancers and Looping

Chromatin looping facilitates interactions between enhancers and promoters. Looping is tissue and development specific, dynamically regulating gene expression. In the rest of this lecture, we will concentrate on the use of genomic chromatin interaction data to investigate looping.



Ong CT, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**:283-93.

Outline

1. Gene Regulation in Eukaryotes
2. 3D Organization of Genomes
3. hi-C
4. Normalizing Hi-C Data
5. Poisson regression: GLMs
6. HiCNorm

Chromosome conformation capture

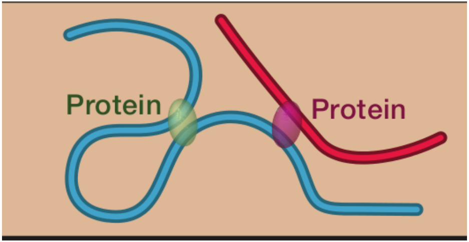
The **family** of chromosome conformation capture (CCC) techniques is a set of biochemical approaches to determine the physical interaction of genome regions.

CCC-technology approaches:

- invariably involve five wetlab steps
- require computational analysis to determine interaction frequencies captured in the ligation of the crosslinked chromatin.

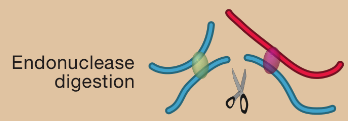
CCC: 1- Fixation

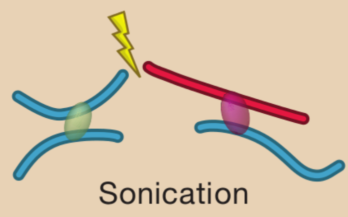
(1) formaldehyde fixation to crosslink chromatin at sites of physical interaction



CCC: 2- Cleavage

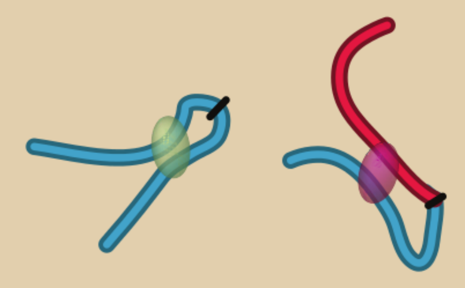
(2) cleavage of chromatin by restriction enzyme or sonication





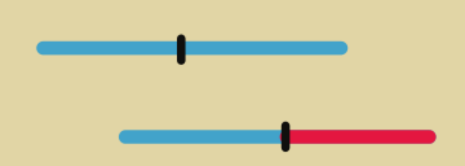
CCC: 3- Ligation

(3) ligation under dilute conditions favoring ligation between DNA ends captured on the same complex (**intramolecular** ligation) over ligations from random collisions



CCC: 4- Reverse crosslinks

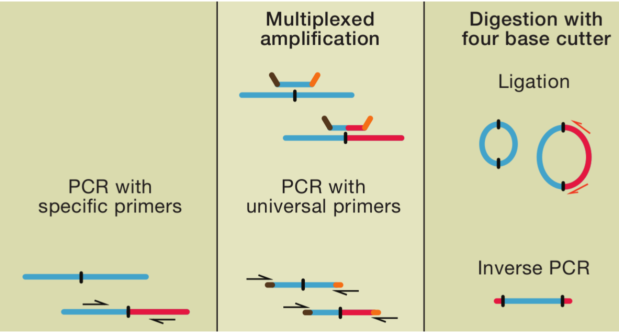
(4) Reverse the formaldehyde-induced crosslinks



Cross-links are reversed by heating, Proteinase K digestion, and phenol-chloroform extraction.

CCC: 5- Detection

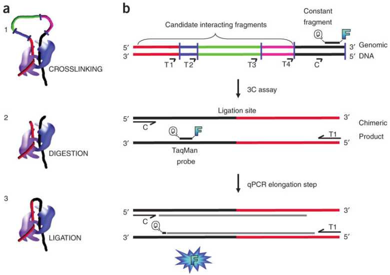
(5) detection of ligation junctions using variable molecular biology steps. Methodology depends on whether we are doing 3C, 4C, 5C, hi-C, etc.


Multiplexed amplification
Digestion with four base cutter
Ligation
PCR with specific primers
PCR with universal primers
Inverse PCR

Images from Hakim O, Misteli T (2012) Cell 148:e1068

CCC: 3C

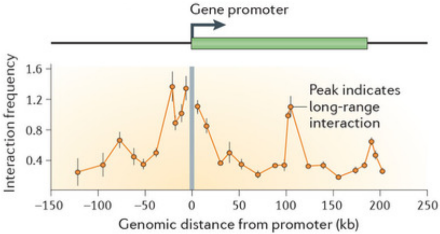
3C: The simplest CCC technology. We use one targeted quantitative PCR (Q-PCR) for each interaction we want to measure. Shown here: constant fragment (black segment) and candidate interacting fragments (red, blue, green and pink segments). Restriction sites that will be used in the 3C assay are depicted as small vertical bars in blue.



Hagege H et al (2007) *Nat Protoc* **2**:1722-33.

CCC: 3C

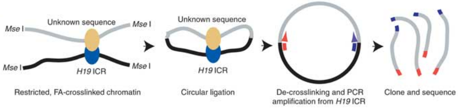
3C data for the CFTR gene. Interaction frequencies between the promoter (grey bar) and various other regions are measured with specific Q-PCR primer combinations. Peaks suggest increased frequency, ergo, substantial interaction between the promoter and some distant element.



Dekker J et al (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**:390-403.

CCC: 4C

A particular derivative of 3C method named **circular chromosome conformation capture (4C)** enables de novo detection of all interacting partners of a known genomic region,

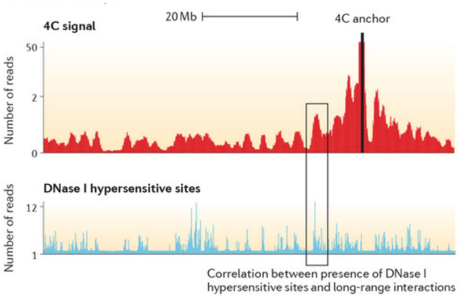


Zhao Z et al. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* **38**:1341-7.

- Red and blue arrows and rectangles ⇒ the nested primers within the H19 imprinting control region, which is common to all 4C products.
- Gray line ⇒ any sequence interacting with the H19 ICR and captured by the 4C approach.

CCC: 4C

3C and 4C generate single interaction profiles for individual loci (anchor). In contrast to 3C, 4C generates a genome-wide interaction profile for the single locus.

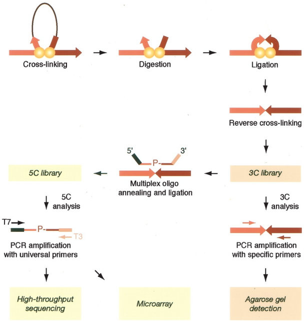


Correlation between presence of DNase I hypersensitive sites and long-range interactions

Dekker J et al (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**:390-403.

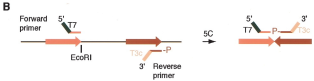
CCC: 3C-Carbon Copy – 5C

5C uses highly multiplexed ligation-mediated amplification (LMA) to first copy and then amplify parts of the 3C library followed by detection on microarrays or by quantitative DNA sequencing.


Cross-linking
Digestion
Ligation
Reverse cross-linking
3C library
5'
3'
P-
Multiplex oligo annealing and ligation
5C library
5C analysis
3C analysis
T7
P-
T3
PCR amplification with universal primers
PCR amplification with specific primers
High-throughput sequencing
Microarray
Agarose gel detection

Dostie J et al (2006) *Genome Res* **16**:1299-309.

CCC: 5C primer design

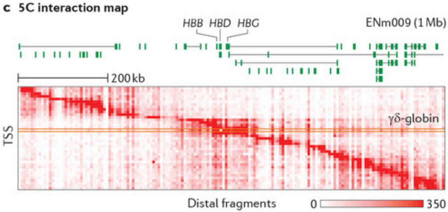


Dostie J et al (2006) *Genome Res* **16**:1299-309.

- Forward 5C primers anneal to the sense strand of the 3-end of restriction fragments and include half of the selected restriction site.
- All forward primers feature a common 5-end tail containing the T7 promoter sequence.
- Reverse 5C primers anneal to the antisense strand of the 3-end of restriction fragments, including half of the restriction site.
- All reverse primers contain a common 3-end tail featuring the complementary T3 sequence (T3c) and are phosphorylated at the 5-end.
- **5C forward and reverse primers anneal to the same strand of head-to-head ligation products present in the 3C library.**

CCC: 5C

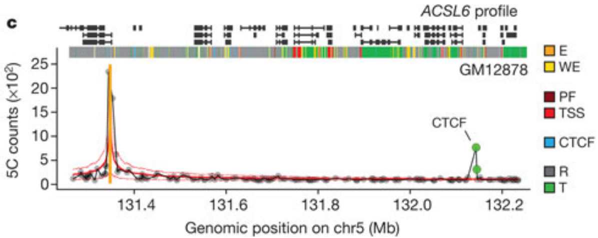
5C interaction map. Each row represents an interaction profile of a transcription start site (TSS) across the 1 Mb region on human chromosome 11 that contains the β-globin locus. 5C is many-vs-many and requires a large set of specific Q-PCR primers. **5C essentially measure many anchored interaction profiles in parallel.**



Dekker J et al (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**:390-403.

CCC: 5C

5C analysis strategy (for one particular reference point)



Sanyal A et al (2012) *Nature* **489**:109-13.

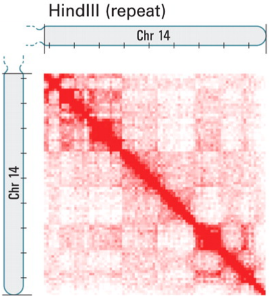
- W5C interaction profiles for promoter of *ACSL6* gene
- solid red lines ⇒ the expected interaction level
- The dashed red lines ⇒ ±1 standard deviation.
- 5C signals that are significantly higher than expected are considered **looping interactions**.

Outline

1. Gene Regulation in Eukaryotes
2. 3D Organization of Genomes
3. hi-C
4. Normalizing Hi-C Data
5. Poisson regression: GLMs
6. HiCNorm

hi-C

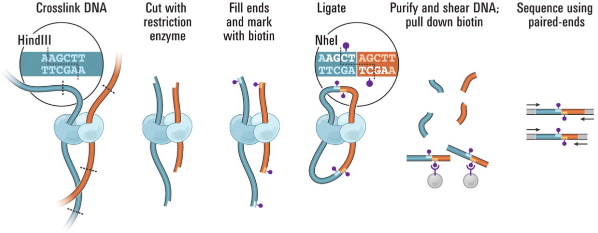
HiC enables an all-versus-all interaction profiling.



Lieberman-Aiden E et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**:289-93.

- unbiased genomewide analysis
- Heat map shows contact matrix for chromosome 14

hi-C: Methodology


Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Lieberman-Aiden E et al (2009) *Science* **326**:289-93.

- Here, two spatially adjacent chromatin segments (blue/orange) are connected by protein-protein interactions
- They are crosslinked by formaldehyde and digested by *Hind*III (aaactt)

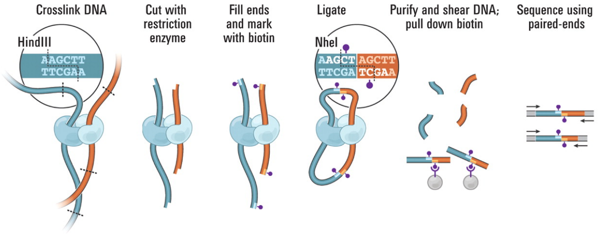
hi-C: Methodology

- *Hind*III digestion leaves a four-nucleotide overhang on the restriction fragments

```
5’-AGCTT-3’
    3’-A-5’
```

- This cleavage provides a template for labeling the restriction fragments with biotin-14-dCTP
- The overhang is filled in by the Klenow fragment of DNA polymerase I using equimolar amounts of all deoxyribonucleotides with the substitution of biotin-14-dCTP for dCTP.

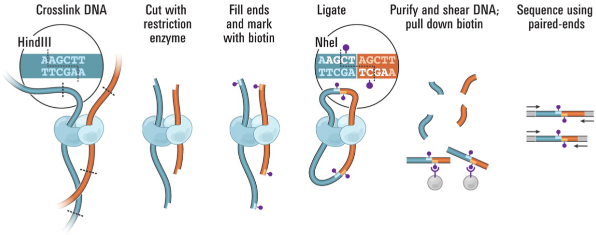
hi-C: Methodology


Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
HindIII
AAGCTT
TTCGAA
NheI
AAGCT AGCTT
TTCGA TCGAA

Lieberman-Aiden E et al (2009) *Science* **326**:289-93.

- Processing by the Klenow fragment results in two fragments with blunt ends that are still cross-linked to one another
- They are ligated to one another in a dilute solution that favors intra-molecular ligation
- The ligation of two filled in HindIII sites results in a new NheI site (5'-GCTAGC-3')

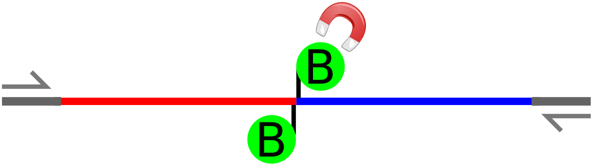
hi-C: Methodology


Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
HindIII
AAGCTT
TTCGAA
NheI
AAGCT AGCTT
TTCGA TCGAA

Lieberman-Aiden E et al (2009) *Science* **326**:289-93.

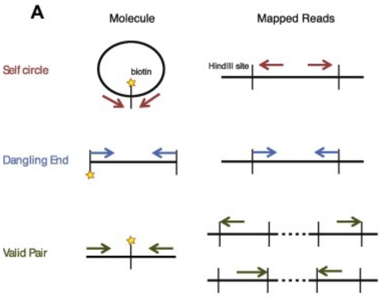
- Process fragments with proteinase K at 65°C
- Purify/shear DNA, pull down with streptavidin beads
- Illumina paired-end sequencing (50bp)

hi-C: Sequence Read Mapping/Filtering



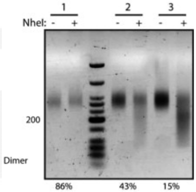
- Paired end sequencing of the two fragments
- Optimally, we pull down a fragment that is biotin-marked at the ligation point of two distinct fragments
- We can now count the number of times distinct pairs of genomic regions are connected

hi-C: Artefacts



- If two reads come from the same HindIII fragment, it represents an artefact
- Self circle: self-circularized ligation product
- Dangling end: An unligated product

hi-C: Quality control

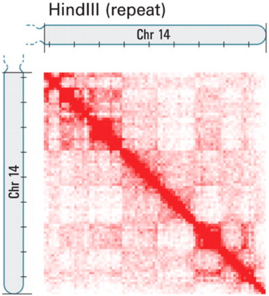


Belton JM et al (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**:268-76.

- Recall that HindIII digestion leaves us with the overhang AGCT. If two such overhangs are filled in and then blunt-end ligated, we obtain a new NheI site: 5'-GCTACG-3'.
- The shift of the size distribution of the library following digestion with NheI estimates the proportion of the library that consists of real HiC ligation products.
- Here: library 1 ⇒ poor performance, library 2 ⇒ medium performance and library 3 ⇒ good performance.

hi-C: Computational analysis

Basic analysis strategy: divide the genome into 1-Mb regions, and define the matrix entry m_{ij} to be the number of ligation products between region i and region j.

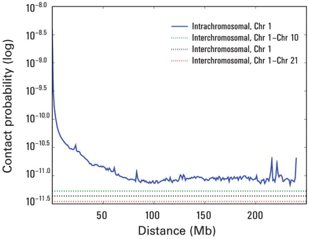


Lieberman-Aiden E et al (2009) *Science* **326**:289-93.

- The figure represents the contact matrix M for chromosome 14 as a heatmap

hi-C: Computational analysis

Hi-C can be used to investigate the **organization of chromosome territories**



Lieberman-Aiden E et al (2009) *Science* **326**:289-93.

- We will call the 1-Mb regions "loci"
- Let $I_n(s)$ be the average intrachromosomal contact probability on chromosome n for pairs of loci separated by a genomic distance s
- Probability of contact decreases with genomic distance
- Intrachromosome contacts more common than interchromosomal

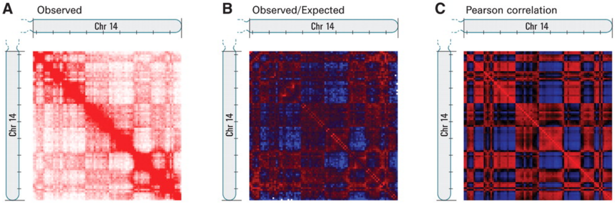
hi-C: Computational analysis

- We define a normalized contact matrix M^* by dividing each entry of the contact matrix M by the genome wide-average contact probability for loci at that genomic distance
- For loci i and j, which are at a distance of $s(i,j)$ to one another, the element of M^* is thus

$$\frac{m_{ij}}{I(s(i,j))}$$

- The normalized matrix shows many patterns that were less apparent in the original matrix

hi-C: Computational analysis

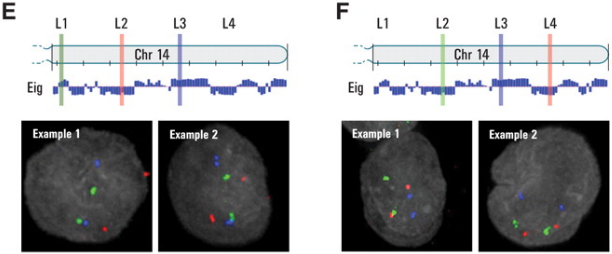


Lieberman-Aiden E et al (2009) *Science* **326**:289-93.

- O/E matrix: loci with either more (red) or less (blue) interactions than expected given their genomic distance
- Pearson matrix: Correlation between the intrachromosomal interaction profiles for every pair of 1-Mb loci
- The plaid pattern indicates two compartments within the chromosome

hi-C: Computational analysis

- plaid patterns were consistently seen for all chromosomes. The labels A and B could be assigned so that sets on each chromosome with the same label had correlated contact profiles.
- These results imply that the genome can be divided into two spatial compartments.



Lieberman-Aiden E et al (2009) *Science* **326**:289-93.

- **(E)** probes L1 and L3 (compartment A) are close to one another than to L2 (compartment B)
- **(F)** probes L2 and L4 (compartment B) are close to one another than to L3 (compartment A)

hi-C: Computational analysis

Compartment A correlates strongly with the presence of

- Genes
- higher expression
- accessible chromatin activating (H3K36 trimethylation) and repressive (H3K27 trimethylation) chromatin marks .

Thus, compartment A is more closely associated with open, accessible, actively transcribed chromatin.

Lieberman-Aiden E et al (2009) *Science* **326**:289-93.

Outline

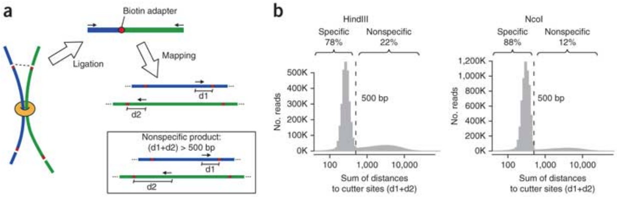
1. Gene Regulation in Eukaryotes
2. 3D Organization of Genomes
3. hi-C
4. Normalizing Hi-C Data
5. Poisson regression: GLMs
6. HiCNorm

Systematic biases in Hi-C data

In Hi-C, the frequency of chromatin interaction is represented by the number of paired-end reads linking two genomic sequences. In principle, higher counts indicate increased frequency of chromatin interaction and closer spatial distance between the two sequences, but **systematic biases** in the data can greatly affect interpretation.

- In the remaining time, we will examine the reasons for these biases and then present a methodology that uses a generalized linear model to correct for bias.

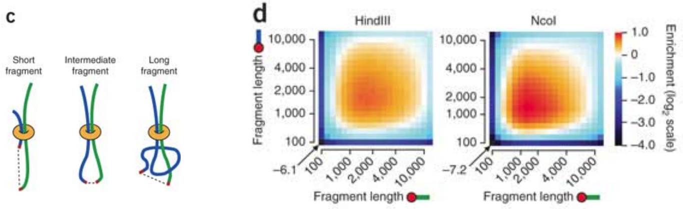
Systematic biases in Hi-C data



Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**:1059-65.

- Note here that most of the reads are located with 500 bp of the HindIII (or NcoI) restriction sites
- This is as expected given that the size selection parameter in the experiment was 500bp
- However, there are additional reads distributed nearly uniformly across the entire fragment – likely to represent artifacts

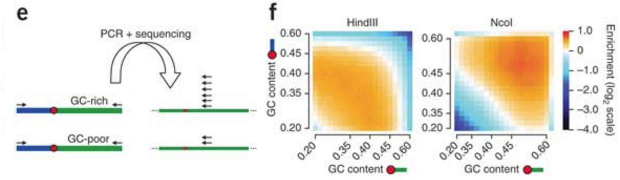
Systematic biases in Hi-C data



Yaffe E, Tanay A (2011) *Nat Genet* **43**:1059-65.

- **(C)** Length of restriction fragments (in other words, the distance between adjacent cutter sites) represents another source of bias
- For example, long and short fragments may have variable ligation efficiencies or compete differently on ligations with cis and trans fragment ends
- **(D)** Restriction fragment lengths are indeed correlated with trans-contact probabilities.

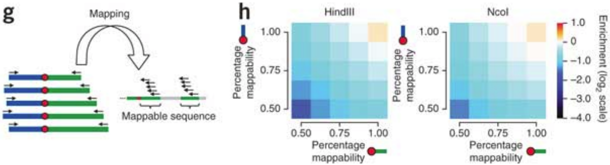
Systematic biases in Hi-C data: G/C content



Yaffe E, Tanay A (2011) *Nat Genet* **43**:1059-65.

- A known major source of bias in sequencing experiments is the nucleotide composition of the DNA under study
- G/C content had a major effect on the hi-C data, but in different ways for HindIII and NcoI data

Systematic biases in Hi-C data: Mapability



Yaffe E, Tanay A (2011) *Nat Genet* **43**:1059-65.

- Mappability is predicted and confirmed to have a linear effect on the estimated trans-contact probabilities.

Outline

1. Gene Regulation in Eukaryotes
2. 3D Organization of Genomes
3. hi-C
4. Normalizing Hi-C Data
5. **Poisson regression: GLMs**
6. HiCNorm

Poisson regression

In the final section of this lecture, we will examine a method called HiCNorm, which is a normalization approach for removing biases in Hi-C data via Poisson regression. We will first introduce the topic of generalized linear model, Poisson regression, and then show how it was used to allow an efficient and effective normalization of Hi-C data

Generalized linear models (GLM)

The simple straight-line regression model has the form:

$$y_i = \alpha + \beta x_i + \epsilon \tag{1}$$

The regression model can be written, equivalently

$$\mathbb{E}[y] = \alpha + \beta x \tag{2}$$

Generalized linear models (GLM)

GLM is a generalization of equation (2), which allows a transformation of the left-hand side of the equation. In other words:

$$f\left(\mathbb{E}[y]\right) = \alpha + \beta x \tag{3}$$

The result specifies a linear relation with x. The function $f(\cdot)$ is called the link function; common examples of link functions are $f(x) = x$, $f(x) = \frac{1}{x}$, $f(x) = \log(x)$ and $f(x) = \log(x/(1-x))$ (the latter function is the logit link). Note all of these functions are monotonic.

Generalized linear models (GLM)

Let us begin with a simple example. Categorical (yes/no) responses do not fit naturally into the framework linear regression, which assumes normally distributed responses.

Suppose we are modeling some response (Y) that is categorical yes/no according to a Bernoulli distribution

$$Y_i \sim \mathrm{Bernoulli}(p_i) \tag{4}$$

This is in contrast to the assumption of linear regression: $Y_i \sim \mathrm{Normal}(\mu_i, \sigma)$, which corresponds to the least squares equation

$$Y = X\beta + \epsilon \tag{5}$$

where β is a $k \times 1$ vector of unknown parameters and ϵ is an $n \times 1$ vector of unobserved disturbances.

Generalized linear models (GLM)

- Lets imagine we have data with $y \in \{0,1\}$ and $x \in \mathbb{Z}^+$.

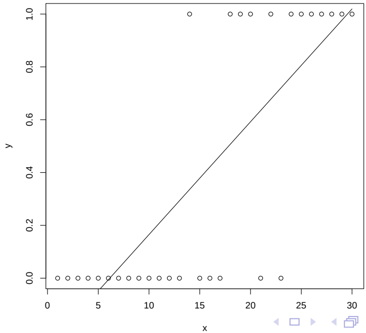
```
 x y                       x y
 1 0                      16 0
 2 0                      17 0
 3 0                      18 1
 4 0                      19 1
 5 0                      20 1
 6 0                      21 0
 7 0                      22 1
 8 0                      23 0
 9 0                      24 1
10 0                      25 1
11 0                      26 1
12 0                      27 1
13 0                      28 1
14 1                      29 1
15 0                      30 1
```

- Let's say we're feeling really stupid today and decide to do standard linear regression (which is equivalent to GLM with an identity link function)

Generalized linear models (GLM)

In R, we could code this as follows:

```
dat <- read.table("example.txt",header=T)
attach(dat)
new <- data.frame(x = seq(1, 30, 0.5))
pr <- predict(lm(y ~ x), new)
plot(x, y)
matplot(new$x, pr, type = "l", ylab = "predicted y",add=TRUE)
```



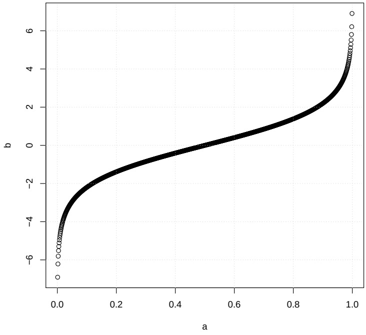
Dumb Regression

- The scatter plot on the previous slide is almost impossible to interpret (There is some relationship between $y = 1$ and high x values compared to low x values, but it's impossible to see much else, much less to visualize the correct regression line).
- The predicted values for y are not even constrained to be within $[0, 1]$.
- The only reasonable solution would seem to be giving up our assumption of normality.
- Briefly, the GLM used for two-category response variables involves the logit link

Bernoulli Regression

The logit link is defined as

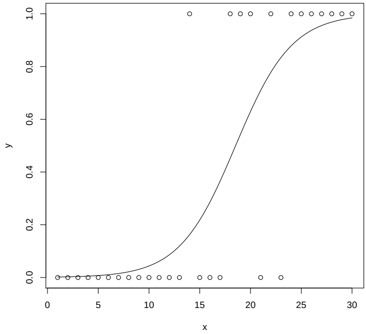
$$f(y) = \log \frac{x}{1-x} \qquad \text{with } x \in (0,1) \tag{6}$$



Bernoulli Regression

In R, we can code Bernoulli GLM as

```
out <- glm(y ~ x, family = binomial)
#summary(out)
plot(x, y)
curve(predict(out, newdata = data.frame(x = x),
    type = "response"), add = TRUE)
```



Poisson Regression

Poisson regression is also a type of GLM where the response variable is a count

$$Y_i \sim \mathrm{Poisson}(\lambda_i) \tag{7}$$

Imagine we are naturalists in Kenia and have data on the number of matings for Bull elephants arranged according to age. The age of the elephant at the beginning of the study and the number of successful matings during the 8 years were recorded. We assume the number of matings follows a Poisson distribution, where the mean depends on the age of the elephant in question.

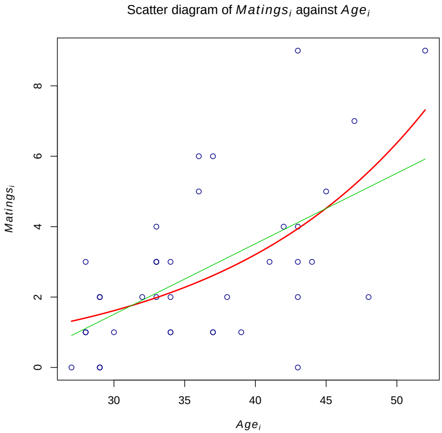
Age	Matings
27	0
28	1
28	1
28	1
28	3
29	0
29	0
29	0
29	2
29	2
29	2
30	1
32	2

Age	Matings
33	4
33	3
33	3
33	3
33	2
34	1
34	1
34	2
34	3
36	5
36	6
37	1
37	1
37	6

Age	Matings
38	2
39	1
41	3
42	4
43	0
43	2
43	3
43	4
43	9
44	3
45	5
47	7
48	2
52	9

Poisson regression

Here we see Poisson regression (GLM) in red vs. standard linear regression in green.



Poisson regression

– R code

```
dat <- read.table("elephants.txt",header=TRUE)
attach(dat)
x<-as.numeric(Age)
y<-as.numeric(Matings)
plot(x,y,col = "darkblue",
      main = expression(paste("Scatter diagram of ",
      italic(Matings[i]), " against ",italic(Age[i]))),
      xlab = expression(italic(Age[i])),
      ylab = expression(italic(Matings[i])))
model<-glm(y ~ x, family = poisson)
predProbs<-predict(model,data.frame(x=seq(min(x), max(x), length.out=100)), type="response")
lines(seq(min(x), max(x), length.out=100), predProbs, col=2, lwd=2)

## Now linear model
ep.lm <- predict(lm(y ~ x), new)
matplot(new$x, ep.lm, type = "l", col=3,add=TRUE)
```

Poisson regression

Recall that in linear regression, if the response variable has a normal distribution, its mean can be linked to a set of explanatory variables using a linear function:

$$Y = \beta_0 + \sum_i \beta_i X_i \tag{8}$$

If the response variable is a count (all positive integers), the Poisson is more appropriate. The logarithm of the response variable is linked to a linear function of response variables

$$\log(Y) = \beta_0 + \sum_i \beta_i X_i \tag{9}$$

equivalently

$$Y = e^{\beta_0} \times \prod_i e^{\beta_i X_i} \tag{10}$$

Poisson regression

In other words, Poisson regression expresses the log outcome as a linear function of a set of predictors

Assumptions:

- Logarithm of the response rate changes linearly with equal increment increases in the exposure variable.
- Changes in the rate from combined effects of different exposures or risk factors are multiplicative.
- At each level of the covariates the number of cases has variance equal to the mean.
- Observations are independent

Source: J Tropical Ped, Research Methods II: Multivariant analysis

Outline

1. Gene Regulation in Eukaryotes
2. 3D Organization of Genomes
3. hi-C
4. Normalizing Hi-C Data
5. Poisson regression: GLMs
6. **HiCNorm**

HiCNorm

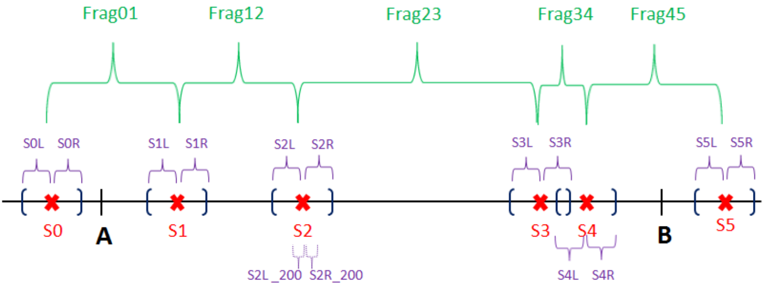
HiCNorm uses three local genomic features to predict the hi-C read count via Poisson regression, using this to correct the raw counts before downstream analysis.

The genome is divided into 1-Mb bins, where L_i^j is the jth bin on chromosome i. For each such bin, three attributes are measures

- Effective length
- GC content
- Mapability

Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**:3131-3.

HiCNorm



- A–B represents one genomic locus (i.e., 1-Mb bin)
- 6 restriction sites (red cross) partition the genomic region into consecutive, disjoint fragments.
- The fragment end is the 500bp genomic region next to the restriction site (Frag01,...). **Effective length** is sum of fragment ends minus overlap
- **GC content** is calculated within a 200 bp region upstream (e.g., S2L_200 and S2R_200)
- **Mappability**: create 55 subsequences (36bp each) in 500 bp around restriction site. Calculate percentage of uniquely mappable reads

HiCNorm

The Poisson regression in HiCNorm works as follows.

- Let $\mathbf{U}^i = \left\{u^i_{jk}\right\}_{1 \le j,k \le n_i}$ represent the $n_i \times n_i$ Hi-C contact map for chromosome i (n_i is the number of 1-Mb bins)
- Thus, u^i_{jk} represents the number of reads spanning bins L^i_j and L^i_k.
- Let x^i_j, y^i_j, and z^i_j, represent the effective length, the GC content, and the mappability at bin L^i_j.

We then assume that u^i_{jk} follows a Poisson distribution with rate parameter λ^i_{jk}

$$\log \lambda^i_{jk} = \beta^i_0 + \beta^i_{\mathrm{len}} \log(x^i_j x^i_k) + \beta_{\mathrm{g/c-c}} \log(y^i_j y^i_k) + \log(z^i_j z^i_k) \quad (11)$$

HiCNorm

The value for λ^i_{jk} is estimated from the Poisson regression[1]
The residual:

$$\hat{e}^i_{jk}$$

is the normalized cis interaction between two bins L^i_j and L^i_k.
Thus, we obtain a normalized matrix

$$\mathbf{U}^* = \left\{\hat{e}^i_{jk}\right\}_{1\leq j,k\leq n_i}$$

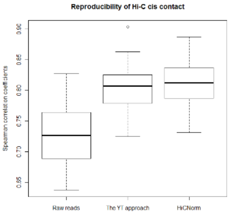
This matrix is used for all downstream applications such as that discussed previously in this lecture

[1] This is performed by convex optimization, we have not covered this in this lecture.

HiCNorm

The authors validated their program in various ways in general the and the computing time for a typical dataset was about .

- reproducibility between experiments was better than that of competing approaches
- Compute speed: 2 seconds compared to about 4 hours for the major competing approach



Finally

- Email: peter.robinson@charite.de
- Office hours by appointment

Further reading

- Belton JM et al (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**:268-76.
- Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**:3131-3
- Yaffe E,Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**:1059-1065.