The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions by

Camille Berthelot, Matthieu Muffato,
Judith Abecassis and Hugues Roest Crollius

Speaker: Ilia Minkin

24th April 2015

Two types of genome alterations

1. Small point mutations:

ACTTG
AGT–G

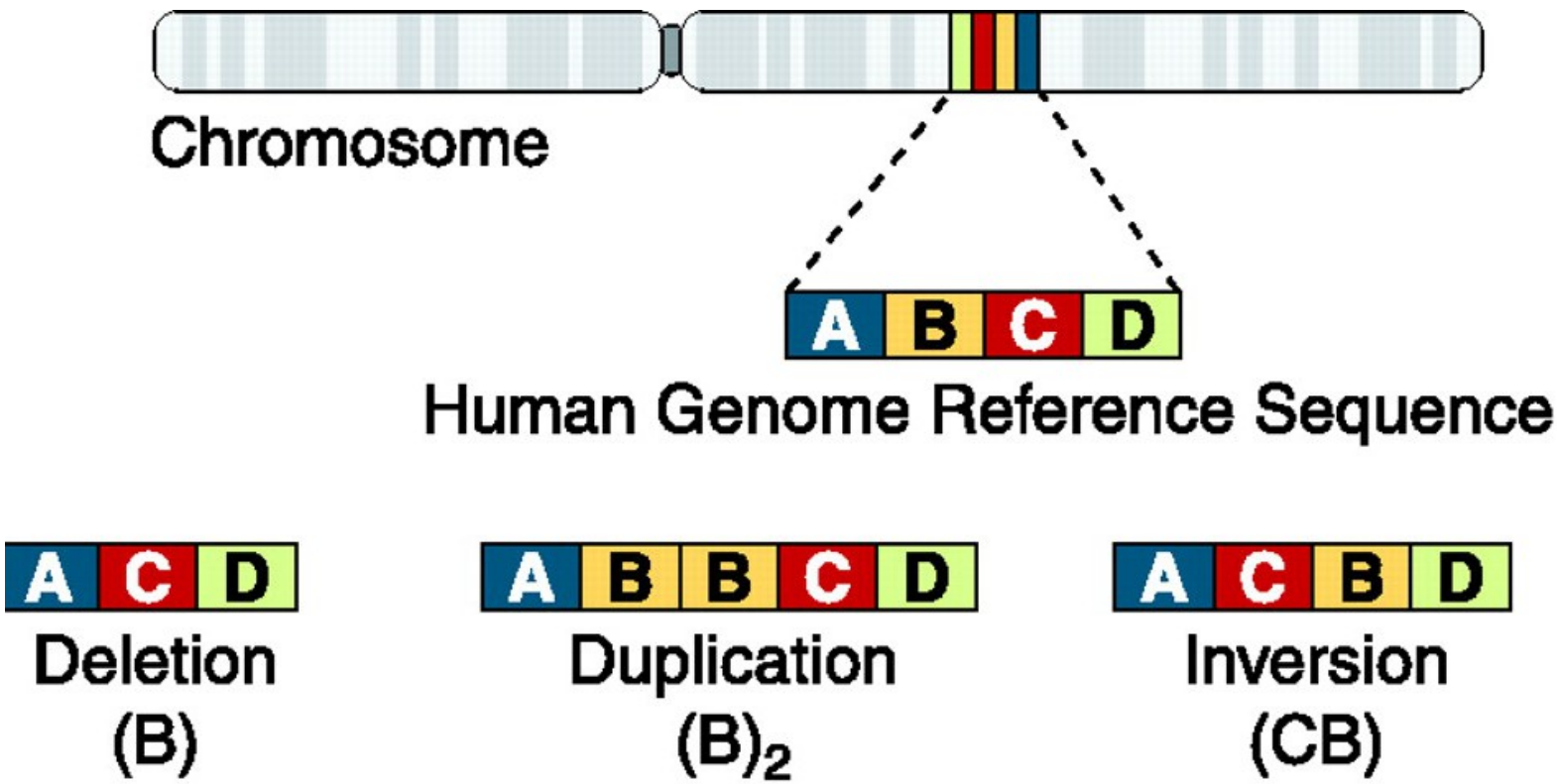
Two types of genome alterations

1. Small point mutations:

ACTTG
AGT–G

2. Large rearrangements:
   - Inversions
   - Transpositions
   - Fusions
   - ...

Genome Rearrangements



Source: *Dierssen et al, 2009*

Motivation

Rearrangements:

- Are a major driving force in evolution
- Play large role in diseases (e.g. cancer)

Known mechanisms:

- Non-homologous end joining
- Non-allelic homologous recombination
- Replication fork stalling
- ...

The Big Question

Are rearrangements more likely to happen in one parts of a genome than the others?

The Big Question

Are rearrangements more likely to happen in one parts of a genome than the others?

Two hypotheses:

1. Rearrangement rates are the same for all parts of a genome
2. Some regions are more likely to be disrupted than the others

A Short Survey

In 1984, Nadeau and Taylor presented first arguments in favor of random breakage

A Short Survey

In 1984, Nadeau and Taylor presented first arguments in favor of random breakage

Pevzner and Tesler in 2003 showed evidence of fragile regions

A Short Survey

In 1984, Nadeau and Taylor presented first arguments in favor of random breakage

Pevzner and Tesler in 2003 showed evidence of fragile regions

Ma et al., 2006 argued for random model with higher resolution analysis of rearrangements

The Study

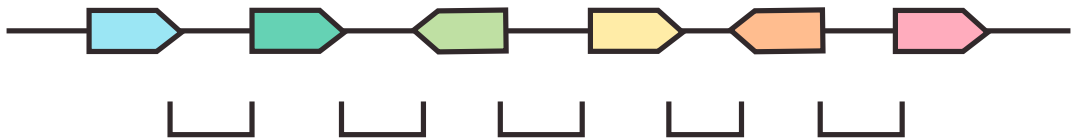
Two questions:

- Do fragile regions exist?
- If they do, what is the cause of fragility?

A note: fragility is not "physical", it only means higher possibility of rearrangements

Genomes Representations

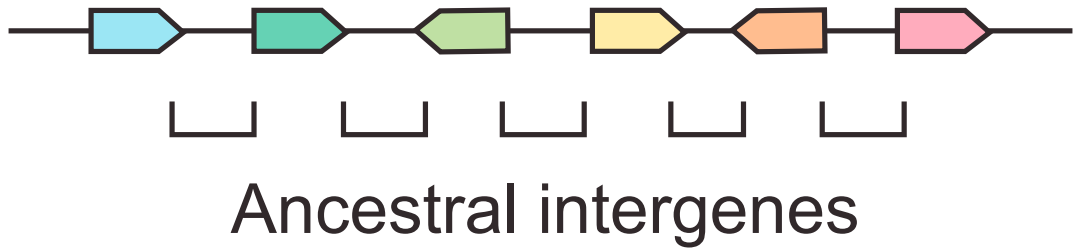
Genomes are sequences of gene markers that are unbreakable:



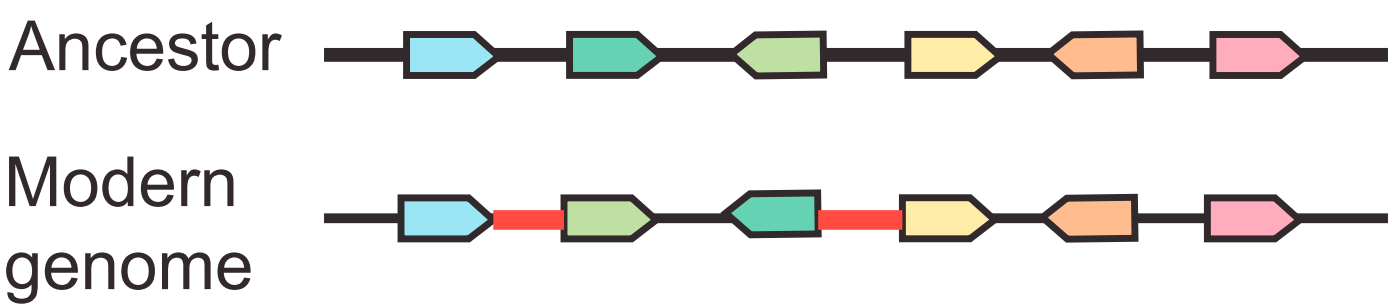
Ancestral intergenes

Genomes Representations

Genomes are sequences of gene markers that are unbreakable:



Here red dashes are breakpoints:



Methodology

Boreoeutheria: the last common ancestor of primates, rodents, and laurasiatherians

There are representations of human, mouse, dog, cow and horse genomes as gene sequences

We can reconstruct the gene order of Boreoeutheria

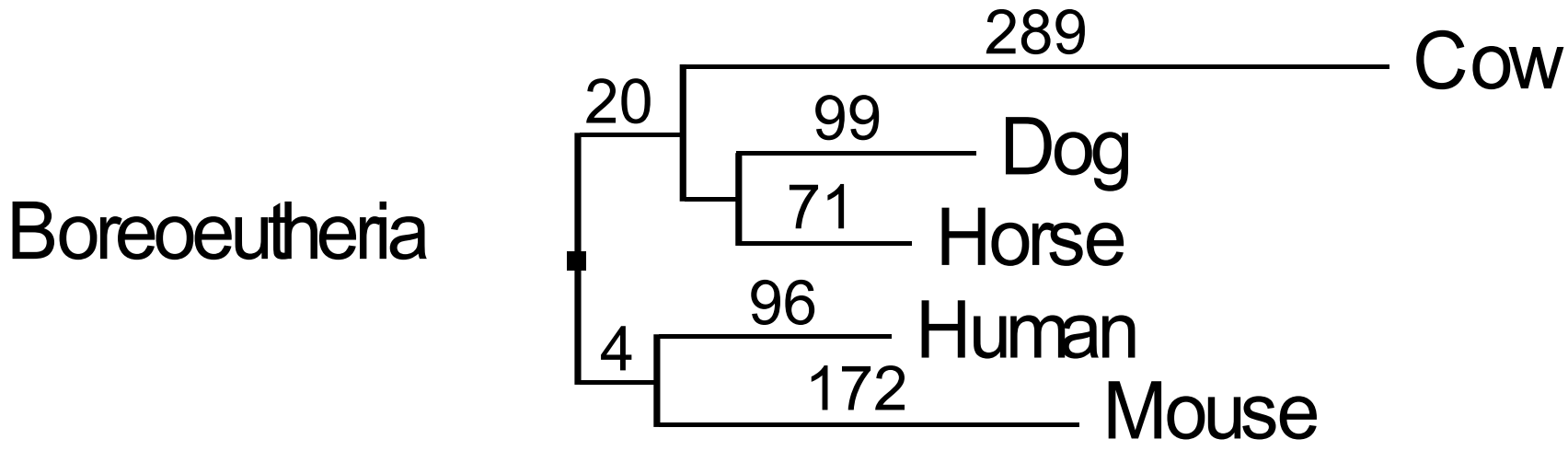
Idea: if two genes are adjacent modern genomes, the adjacency is likely to be in the ancestor

Methodology

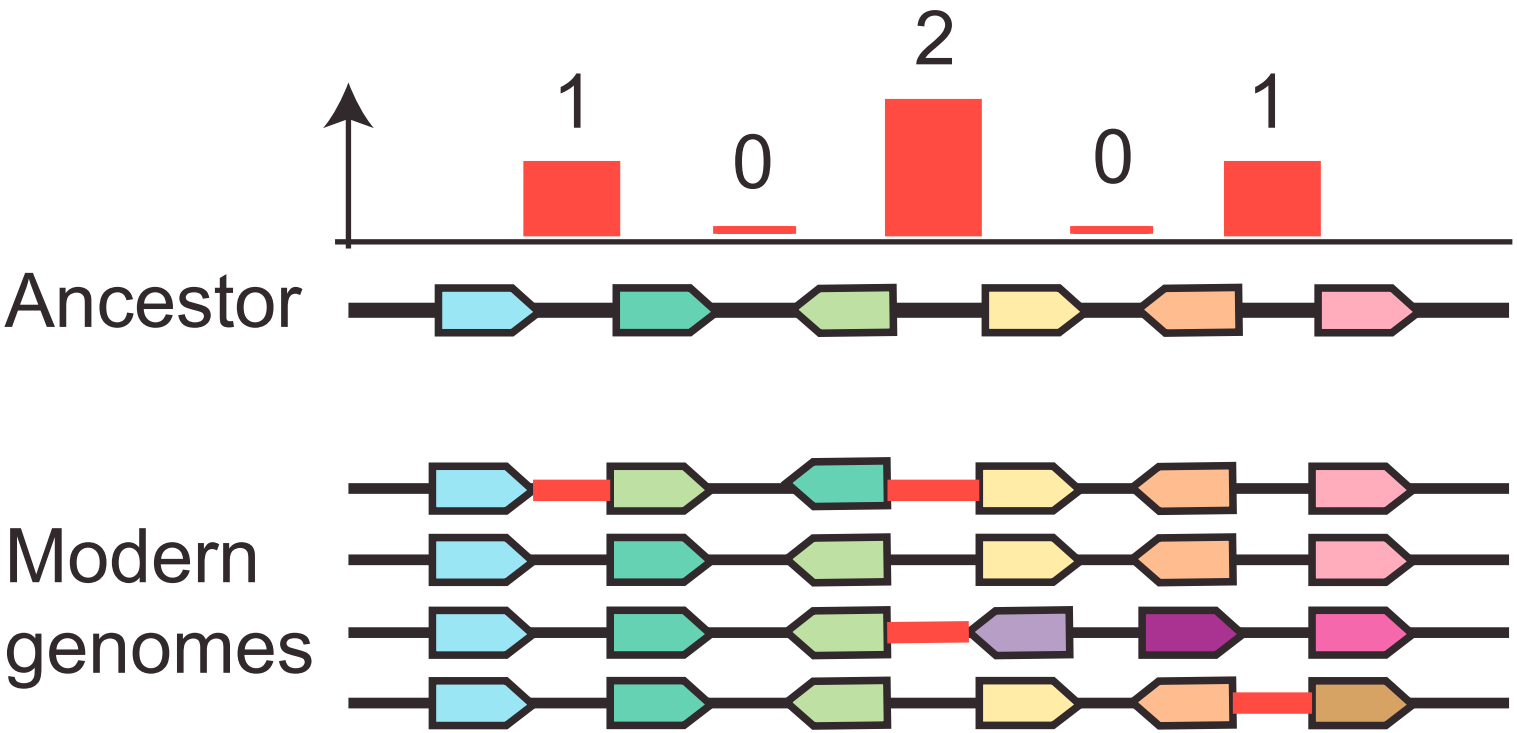
Stages of the study:

- Reconstruct gene order of Boreoeutheria
- Annotate ancestral intergenes
- Identify breakpoints w.r.t. human, mouse, dog, cow and horse
- Count breakpoints per intergene
- Do regression of breakpoint counts

The Phylogenetic Tree



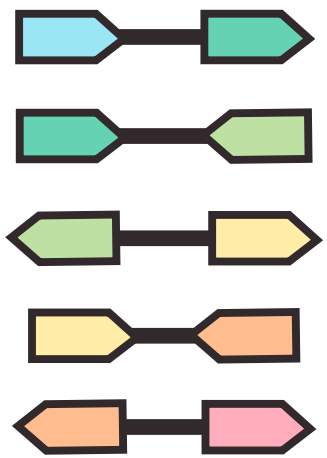
Breakpoint Counting



Intergene Annotation

CNEs – conservative non-coding elements

Characteristics of ancestral intergenes



	Length	%GC	%CNE
	X_1	Y_1	Z_1
	X_2	Y_2	Z_2
	X_3	Y_3	Z_3
	X_4	Y_4	Z_4
	X_5	Y_5	Z_5

Poisson Regression

Put intergenes into bins according to predictors:

1. Intergene length, L
2. GC content, %
3. Conservative non-coding elements, %

Response: breakpoints count / intergene count $=$ or breakage rate r

Regression

Null Hypothesis: pure Poission process

Breakpoints are distributed randomly, in direct proportion to the size of intergenes:

$$r = a \cdot L \Leftrightarrow log(r) = log(L) + \alpha$$

a = average number of breakpoints per intergenic base pair

Regression

Null Hypothesis: pure Poission process

Breakpoints are distributed randomly, in direct proportion to the size of intergenes:

$$r = a \cdot L \Leftrightarrow log(r) = log(L) + \alpha$$

a = average number of breakpoints per intergenic base pair

Regression equation:

$$log(r) = \alpha \cdot log(L) + \beta \cdot \%GC + \gamma \cdot CNE + \delta$$

or

$$r = d \cdot L^{\alpha} \cdot e^{\beta \cdot \%GC} \cdot e^{\gamma \cdot CNE}$$

The Result

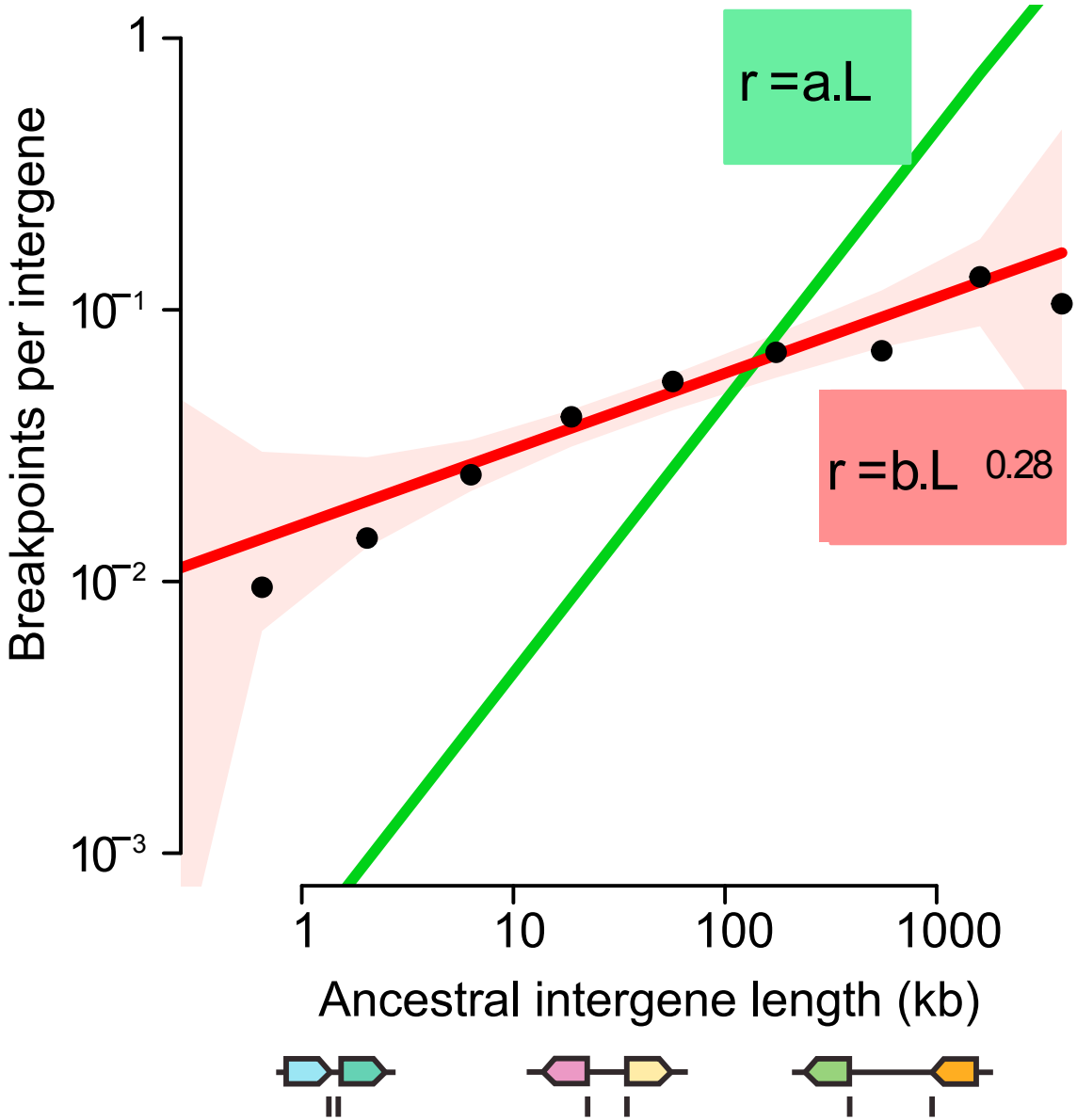

r =a.L
r =b.L 0.28
Breakpoints per intergene
1
10^{-1}
10^{-2}
10^{-3}
1
10
100
1000
Ancestral intergene length (kb)

Table 1. Coefficients and Statistics of Poisson Regression Models Describing the Average Number of Breakpoints per Intergene as a Function of Intergene Length, %GC, and %CNE

	Coefficients					Goodness of Fit		
	Simple Regression	Stepwise Regression	P(> |z|)	Null Deviance (df)	Residual Deviance (df)	c^2 p value	Stepwise c^2 p value	Pseudo R^2
Model 1: length only								
Intergene length	0.28	—	< 2.10^{À16}[a]	167.3 (10)	12.4 (9)	0.19[a]	—	0.93[a]
Model 2: length + %GC								
Intergene length	0.26	0.27	< 2.10^{À16}[a]	137.8 (28)	25.7 (27)	0.53[a]	—	0.81[a]
%GC	—	0.003	0.44	137.8 (28)	25.1 (26)	0.52	0.42	0.82
Model 3: length + %CNE								
Intergene length	0.28	0.30	< 2.10^{À16}[a]	179.2 (19)	26.3 (18)	0.09[a]	—	0.85[a]
%CNE	—	À4.55	0.01[a]	179.2 (19)	20.7 (17)	0.24[a]	0.02[a]	0.88[a]
Simulation: 3D contacts in open chromatin								
Intergene length	0.28	—	< 2.10^{À16}	253.8 (14)	29.6 (13)	0.005	—	0.88

A parameter significantly affecting the breakage rate has a regression coefficient statistically different from 0 (P(> jzj) < 0.05). The goodness of fit of each model is assessed by a c^2 test on the residual deviance and degrees of freedom (i.e., likelihood ratio test): a non-significant p value means that the residual deviance may be attributed to statistical noise. The effect of an additional parameter on the fit is assessed by a c^2 test on the difference in residual deviances and degrees of freedom with and without the parameter: a significant p value means that the fit is significantly better with the additional parameter. The pseudo R^2 corresponds to McFadden's pseudo R^2 (proportion of null deviance explained by the model).
For methods, see the Supplemental Information.
[a]Values indicative of an improvement in the model.

How to Explain the Equation?

$$r = 2.410^{-3} \times L^{0.28}$$

Intergene length explains 93% of variation in breakpoint occurrence

Short intergenes are more breakable than under pure random model, while longer ones are less

Why?

Is GC Content has any Influence?

GC content strongly correlates with gene density

Maybe GC content can be a physical explanation?

Is GC Content has any Influence?

GC content strongly correlates with gene density

Maybe GC content can be a physical explanation?

Added GC content in regression — got a non-significant coefficient

Table 1. Coefficients and Statistics of Poisson Regression Models Describing the Average Number of Breakpoints per Intergene as a Function of Intergene Length, %GC, and %CNE

	Coefficients					Goodness of Fit		
	Simple Regression	Stepwise Regression	P(> |z|)	Null Deviance (df)	Residual Deviance (df)	c^2 p value	Stepwise c^2 p value	Pseudo R^2
Model 1: length only								
Intergene length	0.28	—	< 2.10À16[a]	167.3 (10)	12.4 (9)	0.19[a]	—	0.93[a]
Model 2: length + %GC								
Intergene length	0.26	0.27	< 2.10À16[a]	137.8 (28)	25.7 (27)	0.53[a]	—	0.81[a]
%GC	—	0.003	0.44	137.8 (28)	25.1 (26)	0.52	0.42	0.82
Model 3: length + %CNE								
Intergene length	0.28	0.30	< 2.10À16[a]	179.2 (19)	26.3 (18)	0.09[a]	—	0.85[a]
%CNE	—	À4.55	0.01[a]	179.2 (19)	20.7 (17)	0.24[a]	0.02[a]	0.88[a]
Simulation: 3D contacts in open chromatin								
Intergene length	0.28	—	< 2.10À16	253.8 (14)	29.6 (13)	0.005	—	0.88

A parameter significantly affecting the breakage rate has a regression coefficient statistically different from 0 (P(> jzj) < 0.05). The goodness of fit of each model is assessed by a c^2 test on the residual deviance and degrees of freedom (i.e., likelihood ratio test): a non-significant p value means that the residual deviance may be attributed to statistical noise. The effect of an additional parameter on the fit is assessed by a c^2 test on the difference in residual deviances and degrees of freedom with and without the parameter: a significant p value means that the fit is significantly better with the additional parameter. The pseudo R^2 corresponds to McFadden's pseudo R^2 (proportion of null deviance explained by the model).
For methods, see the Supplemental Information.
[a]Values indicative of an improvement in the model.

Are CNEs Affect Fragility?

CNEs – conservative non-coding elements

Logic: regulative elements may affect genes that are nearby

Disrupting synteny between CNEs and genes may have impact on rearrangements that we observe

Do they?

Are CNEs Affect Fragility?

CNEs – conservative non-coding elements

Logic: regulative elements may affect genes that are nearby

Disrupting synteny between CNEs and genes may have impact on rearrangements that we observe

Do they?

Not that much

Added CNE rate in regression — got a significant coefficient, improved explanation rate only by 3%

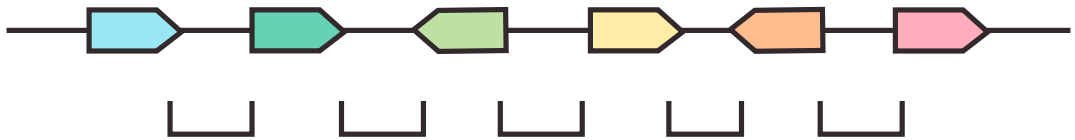
Table 1. Coefficients and Statistics of Poisson Regression Models Describing the Average Number of Breakpoints per Intergene as a Function of Intergene Length, %GC, and %CNE

	Coefficients					Goodness of Fit		
	Simple Regression	Stepwise Regression	P(> |z|)	Null Deviance (df)	Residual Deviance (df)	c^2 p value	Stepwise c^2 p value	Pseudo R^2
Model 1: length only								
Intergene length	0.28	—	< 2.10^{À16}[a]	167.3 (10)	12.4 (9)	0.19[a]	—	0.93[a]
Model 2: length + %GC								
Intergene length	0.26	0.27	< 2.10^{À16}[a]	137.8 (28)	25.7 (27)	0.53[a]	—	0.81[a]
%GC	—	0.003	0.44	137.8 (28)	25.1 (26)	0.52	0.42	0.82
Model 3: length + %CNE								
Intergene length	0.28	0.30	< 2.10^{À16}[a]	179.2 (19)	26.3 (18)	0.09[a]	—	0.85[a]
%CNE	—	À4.55	0.01[a]	179.2 (19)	20.7 (17)	0.24[a]	0.02[a]	0.88[a]
Simulation: 3D contacts in open chromatin								
Intergene length	0.28	—	< 2.10^{À16}	253.8 (14)	29.6 (13)	0.005	—	0.88

A parameter significantly affecting the breakage rate has a regression coefficient statistically different from 0 (P(> jzj) < 0.05). The goodness of fit of each model is assessed by a c^2 test on the residual deviance and degrees of freedom (i.e., likelihood ratio test): a non-significant p value means that the residual deviance may be attributed to statistical noise. The effect of an additional parameter on the fit is assessed by a c^2 test on the difference in residual deviances and degrees of freedom with and without the parameter: a significant p value means that the fit is significantly better with the additional parameter. The pseudo R^2 corresponds to McFadden's pseudo R^2 (proportion of null deviance explained by the model).
For methods, see the Supplemental Information.
[a]Values indicative of an improvement in the model.

Inversions within Intergenes

A reminder: we work with gene markers $\Rightarrow$ see only rearrangements disrupting their order



Ancestral intergenes

Regression showed that longer intergenes have smaller breaks than expected

What if really missing breakpoints within long intergenes?

Inversions within Intergenes

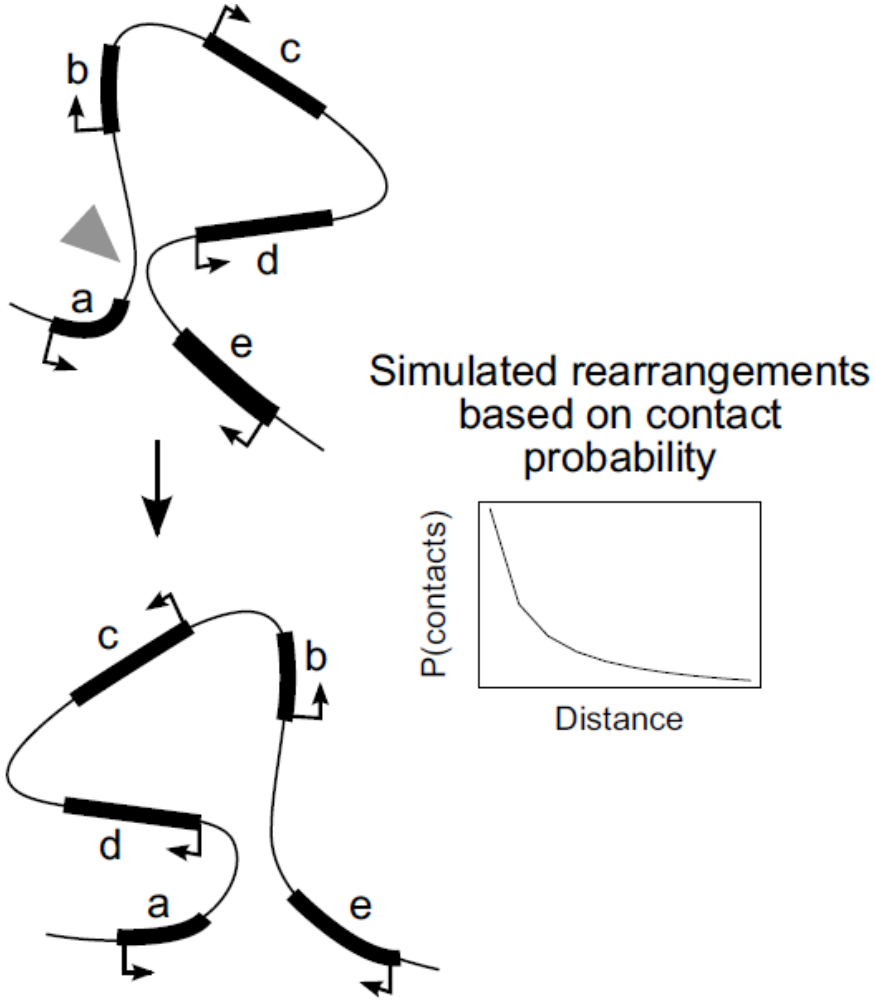
Solution: simulate rearrangements, select detectable ones and compare with the real data

If bias exists, the results should be very close

Rearrangements have been shown to occur between regions in close 3D proximity in the nucleus

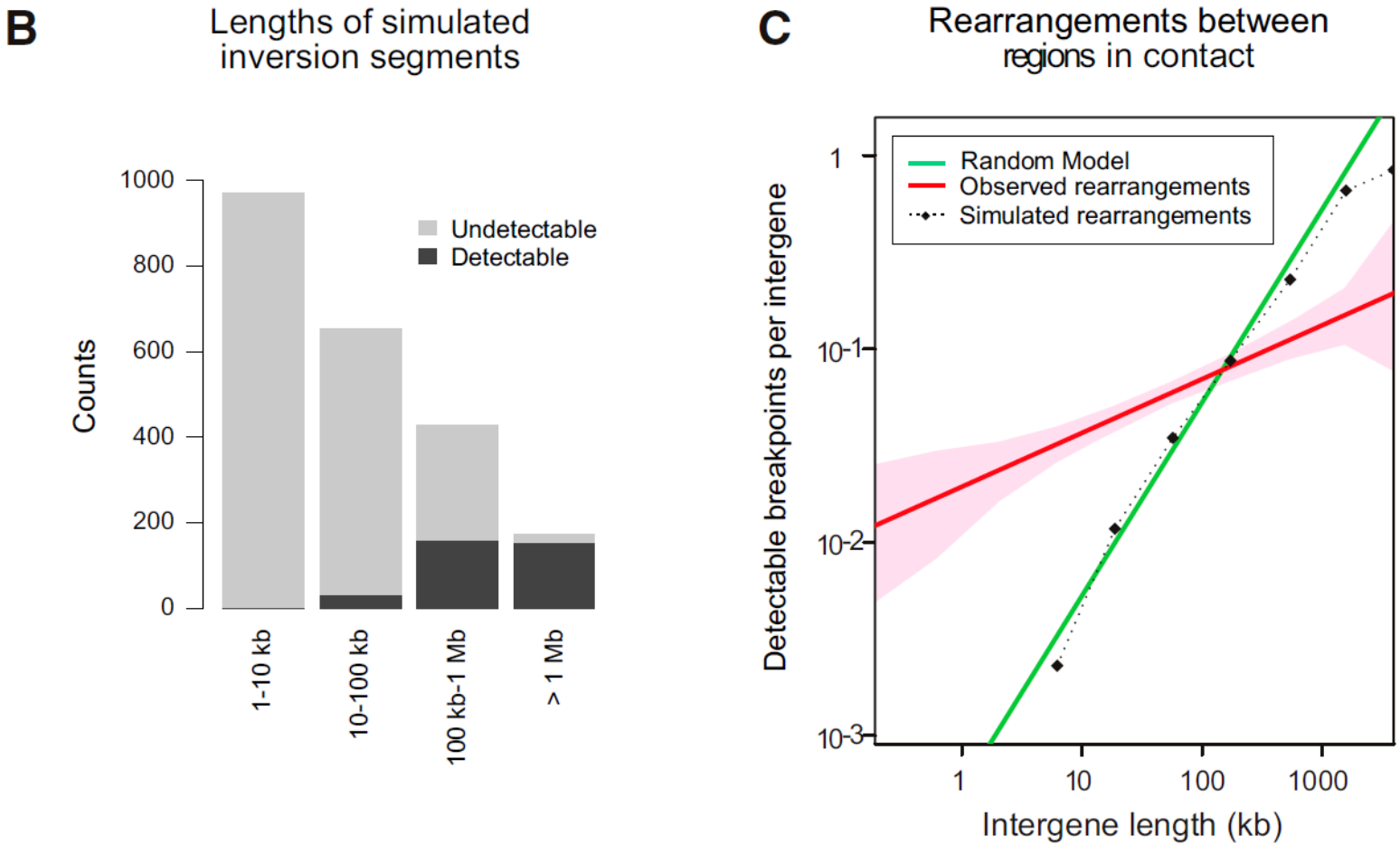
Contact probability is a good proxy for rearrangement probability

Simulation



Inversions are simulated in the human genome (gray arrow) based on the probability of 3D DNA contacts experimentally derived from Hi-C studies (right inset).

Simulation Result



Conclusion: non-detected rearrangements do not introduce enough bias

Open Chromatin is the Culprit

Stick with the simulation – restrict rearrangements to only **open chromatin** regions

ENCODE published chromatin state profiles

The study used four different cell types

Open Chromatin is the Culprit

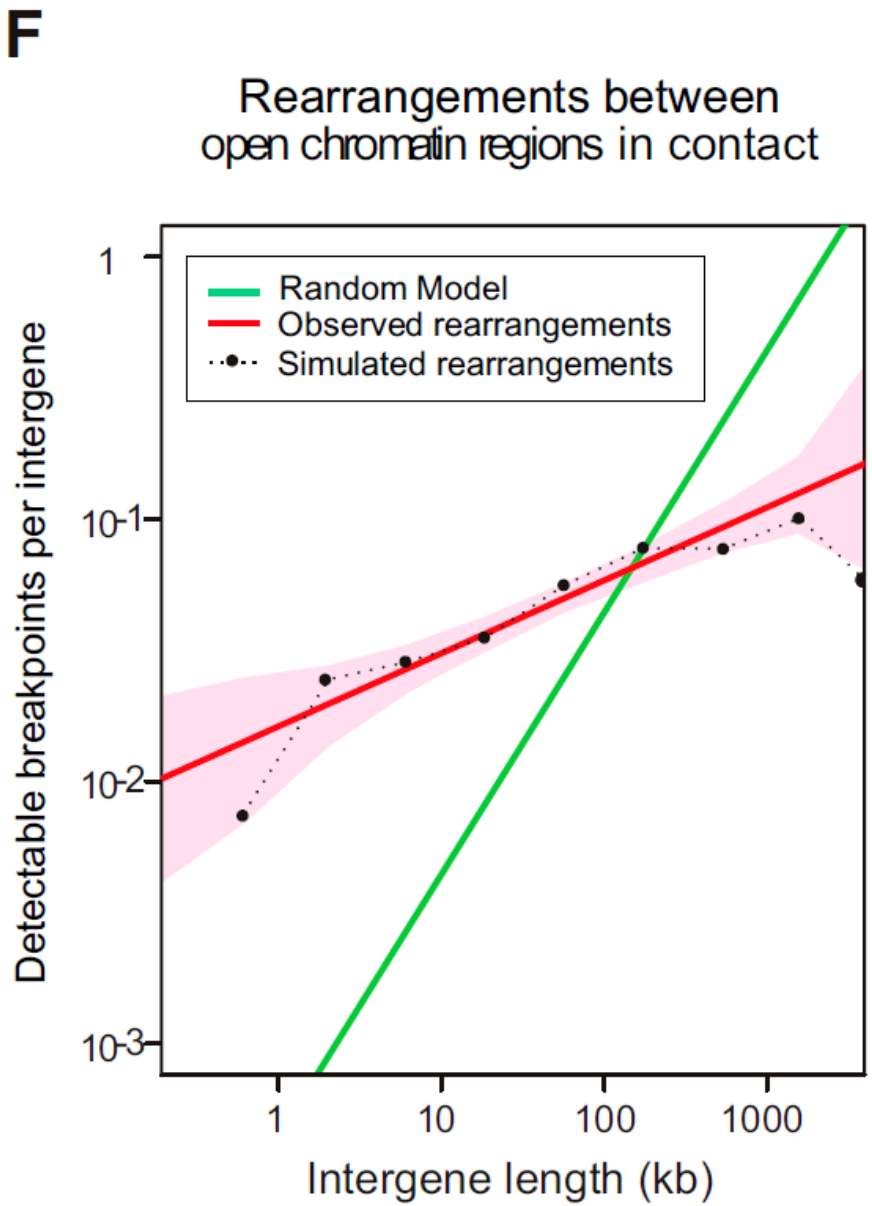
Stick with the simulation – restrict rearrangements to only **open chromatin** regions

ENCODE published chromatin state profiles

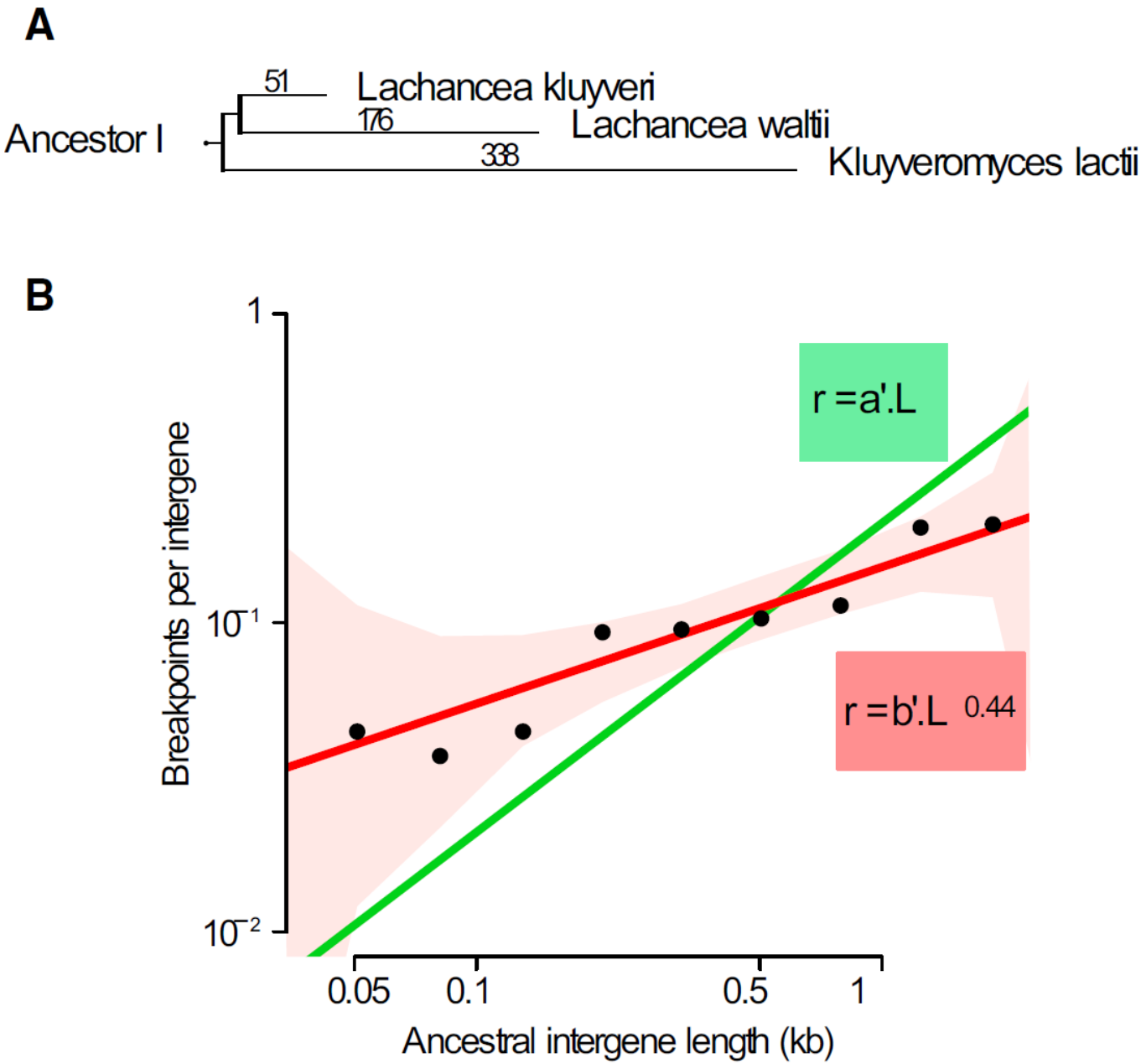
The study used four different cell types

Voilà – simulation coincides with the model!

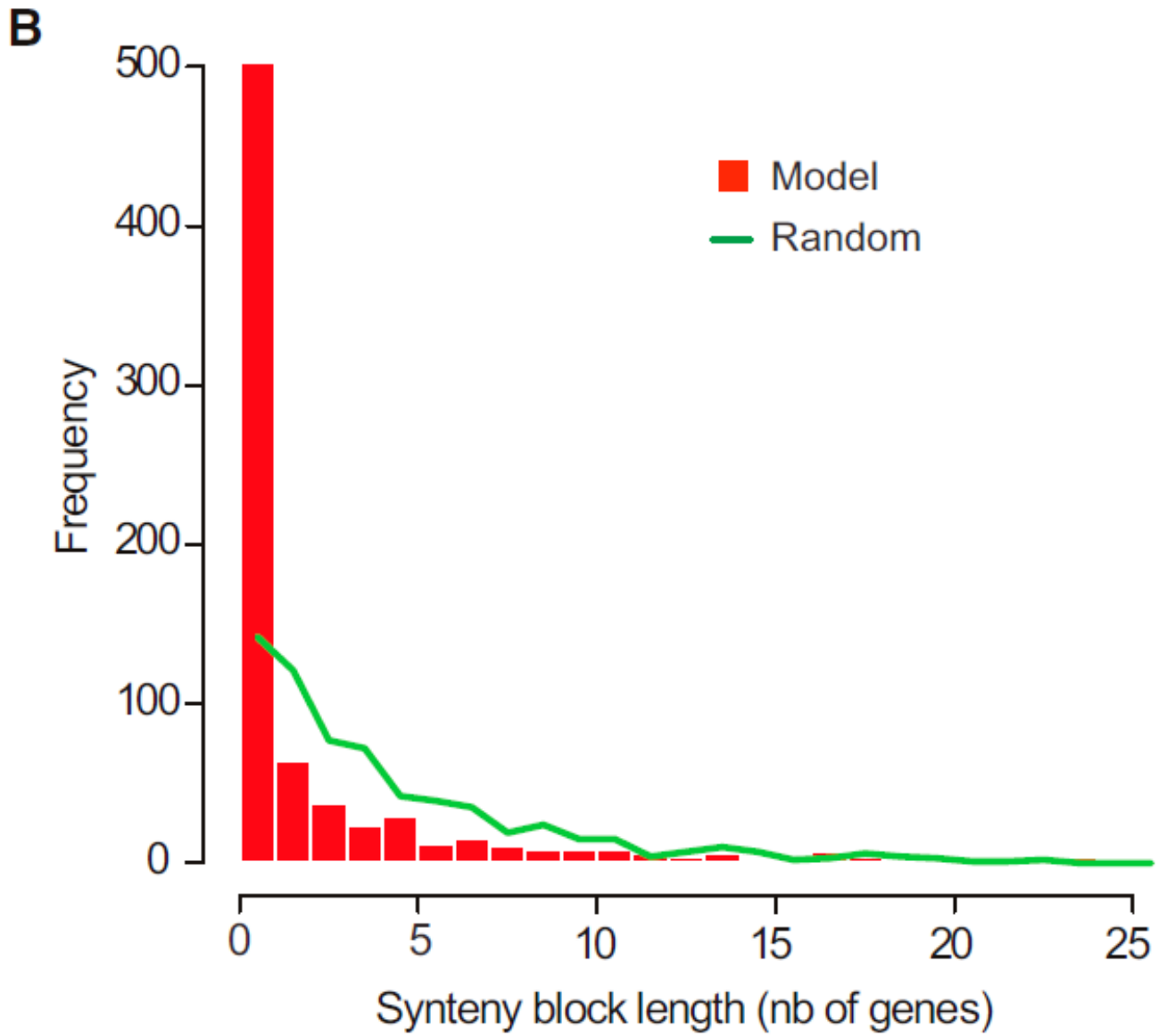
Simulation Result



Simulation Result



Synteny blocks



Conclusion

A number of interesting insights:

- Rearrangements are likely to happen in gene dense regions
- Rearrangement rates depend on location, not content
- Chromatin state and 3D proximity can be used to predict rearrangements
- The patterns are consistent in mammals and yeasts

Thank you!