The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

2585-2

Joint ICTP-TWAS School on Coherent State Transforms, Time-Frequency and Time-Scale Analysis, Applications

2 - 20 June 2014

What can we learn on genome structure and function from a wavelet-based multi-scale analysis of genomic and epigenetic data contd

A. Arneodo
ENS, Lyon
France

Large-scale analysis of genomic sequences

From the detection of replication origins to the modeling of replication in higher eukaryotes

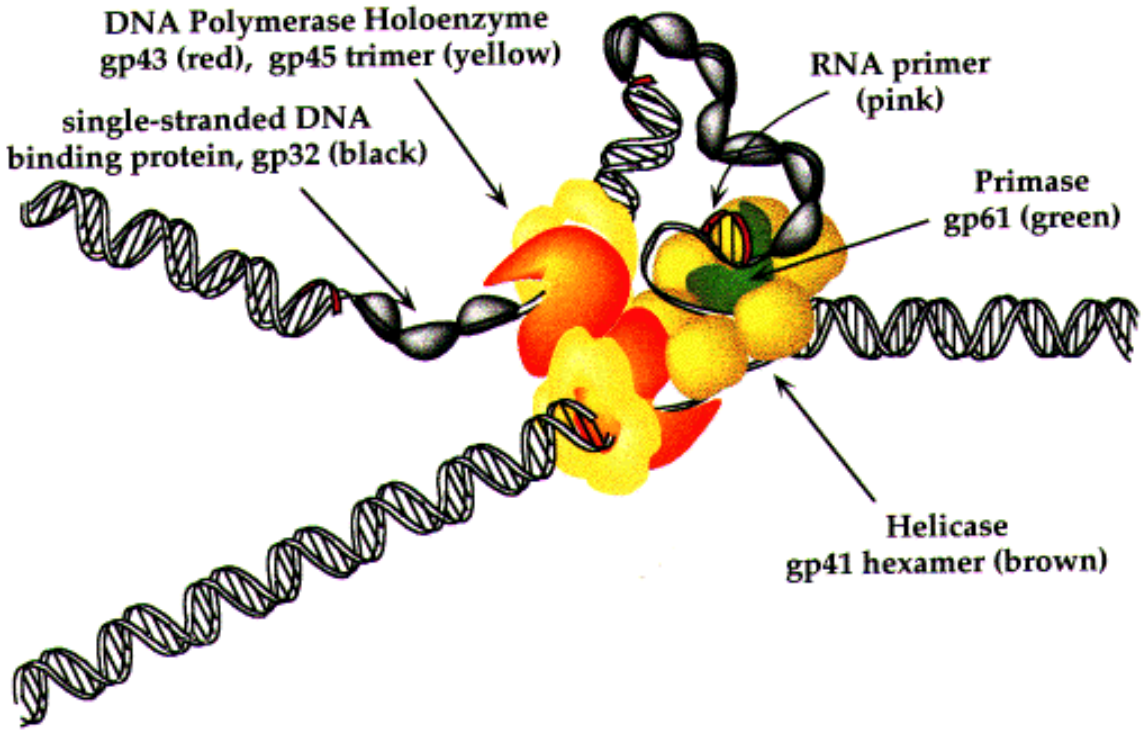
Alain Arneodo

Laboratoire de Physique
Ecole Normale Supérieure de Lyon
46 allée d'Italie, 69364 Lyon Cedex 07, FRANCE

Benjamin Audit
Antoine Baker
Edward-Benedict Brodie of Brodie
Samuel Nicolay — ENS de Lyon, France
Cédric Vaillant

Yves d'Aubenton-Carafa — CGM, Gif-sur-Yvette, France
Chun-Long Chen
Maxime Huvet
Claude Thermes
Marie Touchon

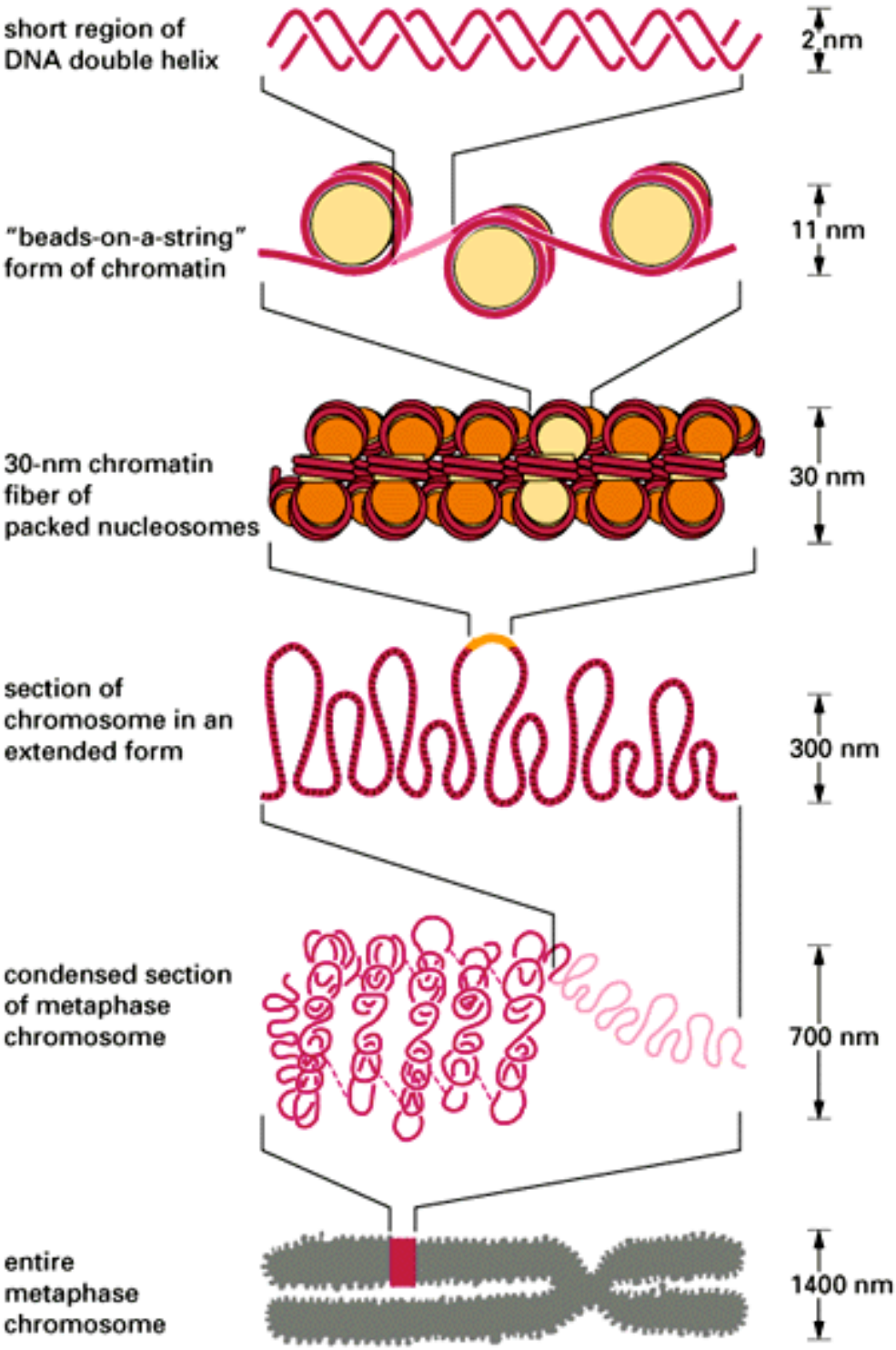
Large-scale analysis of the Human genome: From DNA sequence analysis to the modelling of replication in higher eukaryotes



Sequencing projects result in 4 letter texts :

gtcagttcctgaggcgggtcgggacccaggcgtgagactggagtctgcc
caggggcccagctgagccagcctcctcgtcagctgcttgggccgccagga
cgccgccgggggtgcgccgcgcttccctggatggggtgcccccactcccc
tcggagccccagggagacccccgaactcagctcctctcaggggtgccag
ggggacccctcaaactccactccccgcaggttcctggggagacgccccct
gctcgattcccctcagggtcccagggagacccccctaattcagctcctctc
aggggtactgggggacctctcgagctccactcccatcagggtcccaggga
gacccccccaactatgctcaggggtcccagggagatgccagcaccccaact
ccgcttccctggggcccccctcccccttacagctcaacttccctcgagagt
ctggggctggggctccgttcagttcttgagtcccccttccctcggggtgtc
ccggggccgcccacccccacactgtctgtgattccccaaggcgcgggtct
cgggccgcagcctgttccacgttctgctgctcgttctttttctggctcctt
gctttcgaaggagagaaggaggccttcgtttccagtctttttgccttttc
taatggagccctgcttttccttccgtgtcccttcaggctacttctgccag
gtttctatttttcattctttattatgacttcgcccaaaatattcttgact
tctattgagaaggattcgggggtctatttcttattcggaggcgtgtgctt
aagttccaaacagatgaggattttccagttaatccttctggggtgactta
ttgcttaatgccaccatagccagaaaatggactctcagtgtccgaaactg
cattcggctctgaagtgtctgtccttgtcacctcttgcaatgtttcgcgg
cgggaagcctgcactcgccgacgctgacgtaactgtttctgtctttcagg
tctacagcctcctgtgggtgggcgatattgacatatactttatttctata
tatgttatgaactcaatatttcttgcagcgggtctgctgataataagata
tgcctactctgcgagtctggaagccatcttaagcttaccctgtatgtgcc
ccatgcatctcttccgttacacggctcctgagttgacacctgtgtgataa
actggtaatagcaagtaaactgttttcttgtgctctgtaagctgctctag
caaattatctaggaggaggtggtcttggaaaccccctgatttataagcggg
cagtcagcagtacacgtggcccagaatcgtgattggcatttgaagtgggg
gcagtagggtgggactgagcccttcacctgtggggtctgccctgctcaag
gcagtgtcagaattgaagtgaaatgttggacggtcggtgtccagagagtt
ggagaactggtttgtgtgtaaaaactnacatatttagggtcagaagtatg
...

HIERARCHICAL STRUCTURE OF EUCARYOTIC DNA

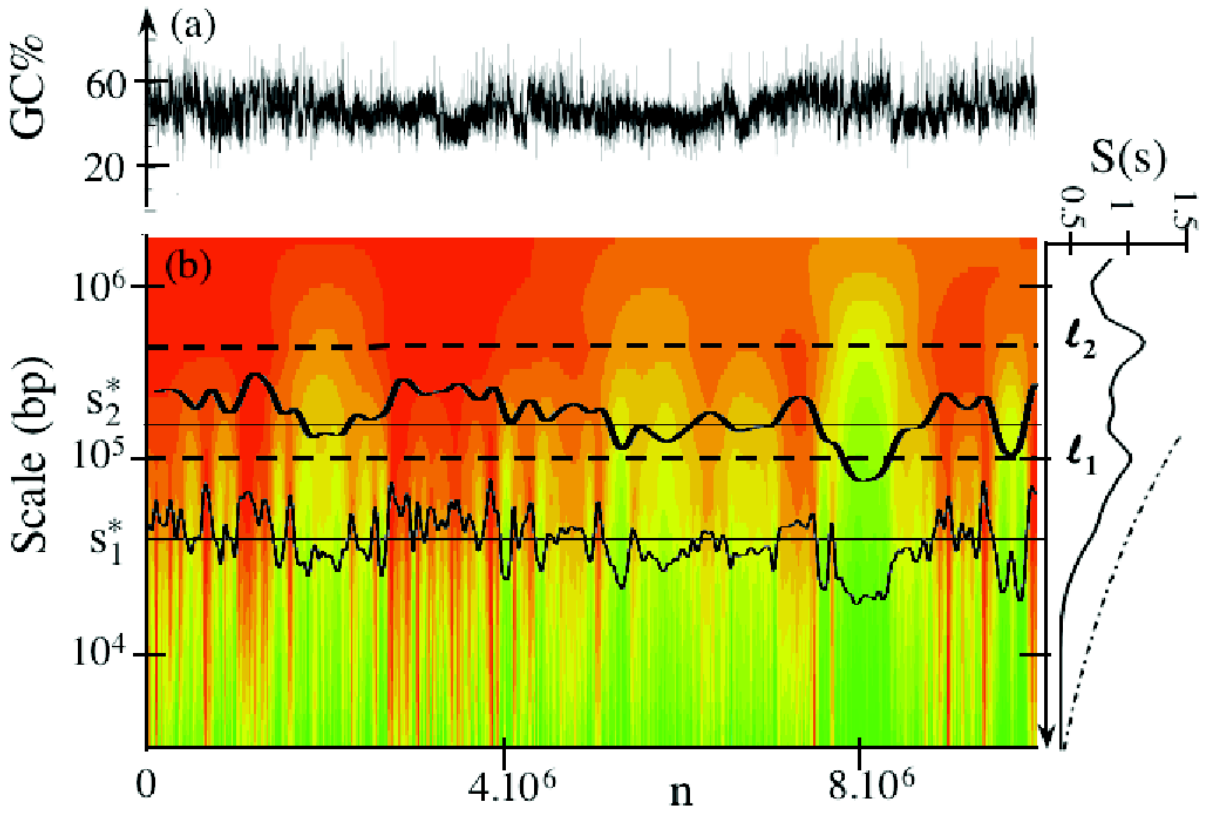


NET RESULT : EACH DNA MOLECULE HAS BEEN PACKAGED INTO A MITOTIC CHROMOSOME THAT IS 50.000x SHORTER THAN ITS EXTENDED LENGTH

LARGE SCALE REPRESENTATION OF GENOMIC SEQUENCES

Space-Scale Representation of the GC Content with a Smoothing Gaussian Filter

Chromosome 22 (Human)



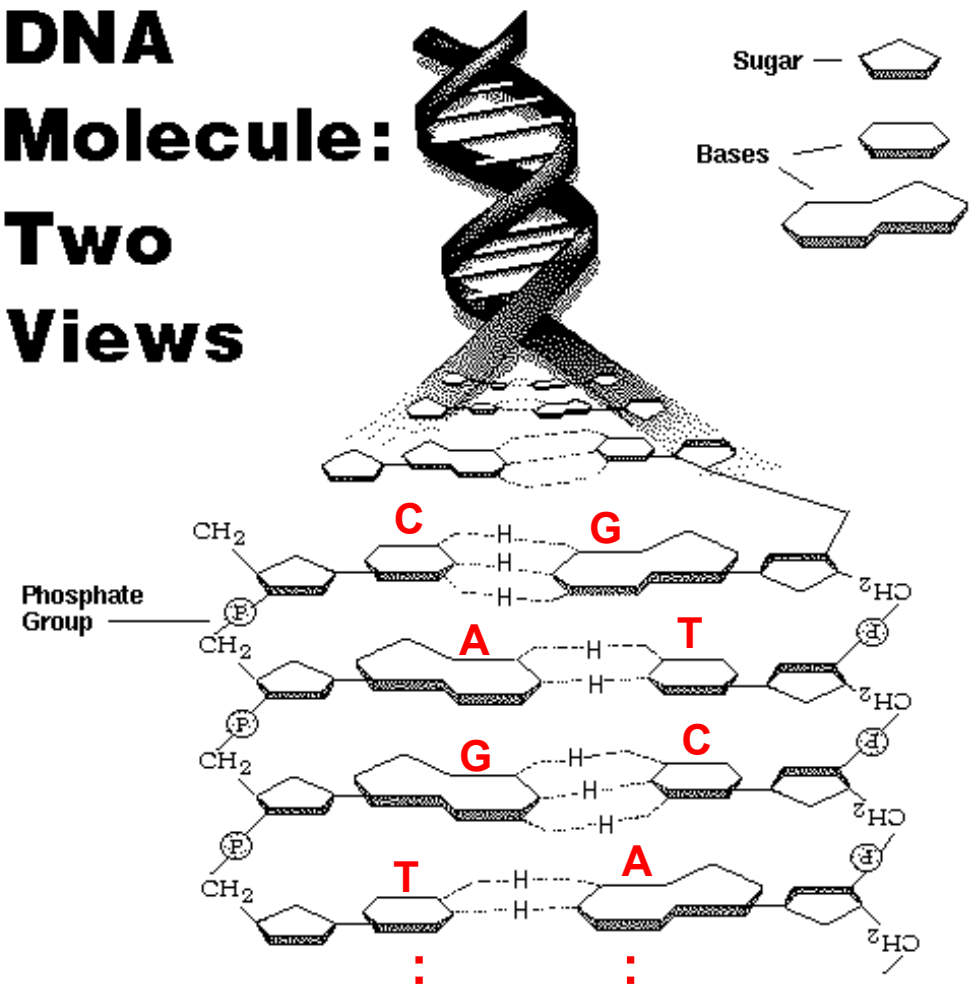
Filtering scales: $a_1^* = 40\text{kb}$, $a_2^* = 160\text{kb}$

Space-scale content: $S(a) = \sum_n |T_{\psi_M}(n, a)|$,
where ψ_M is the Morlet wavelet

Symmetrical properties of the strands: “Parity Rule type 2”

$$[A] = [T] \quad \& \quad [G] = [C]$$

in each strand



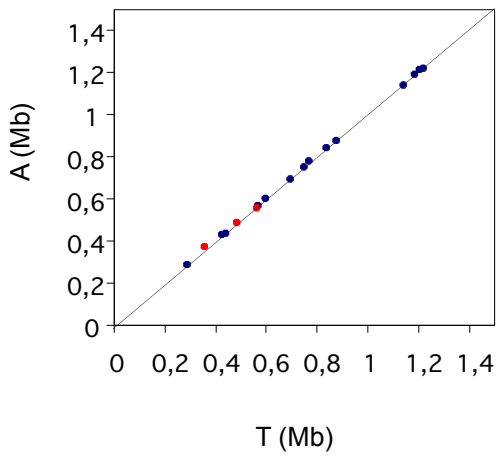
Symmetrical properties of the strands: “Parity Rule type 2”

$$[A] = [T] \quad \& \quad [G] = [C]$$

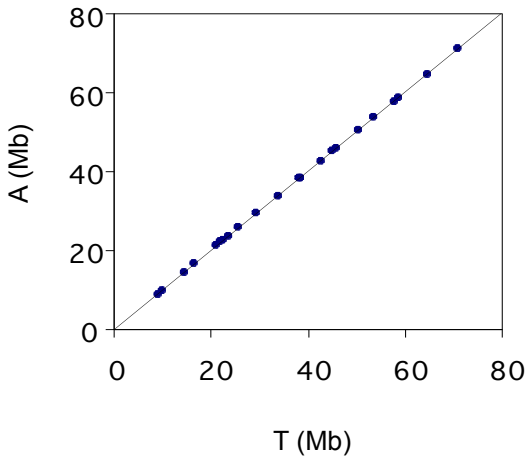
in each strand

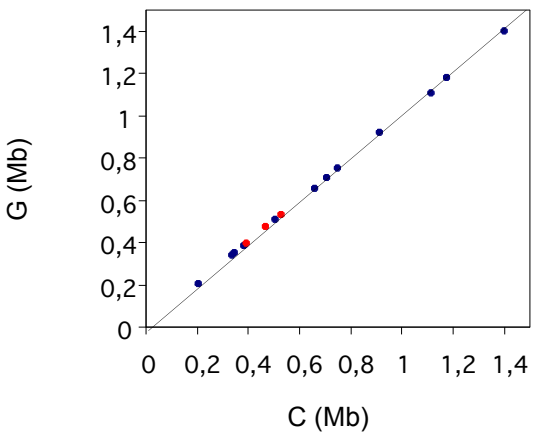
Second parity rule is verified for complete chromosomes

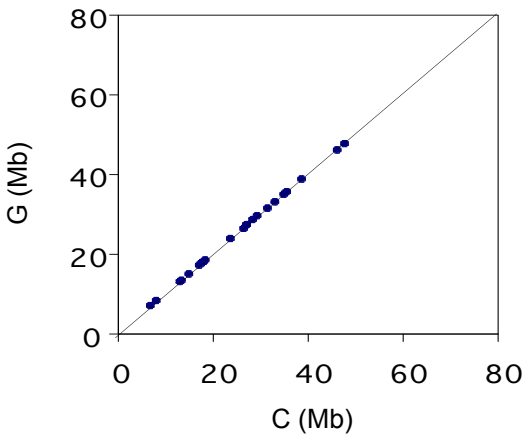
Bacteria/Archaebacteria



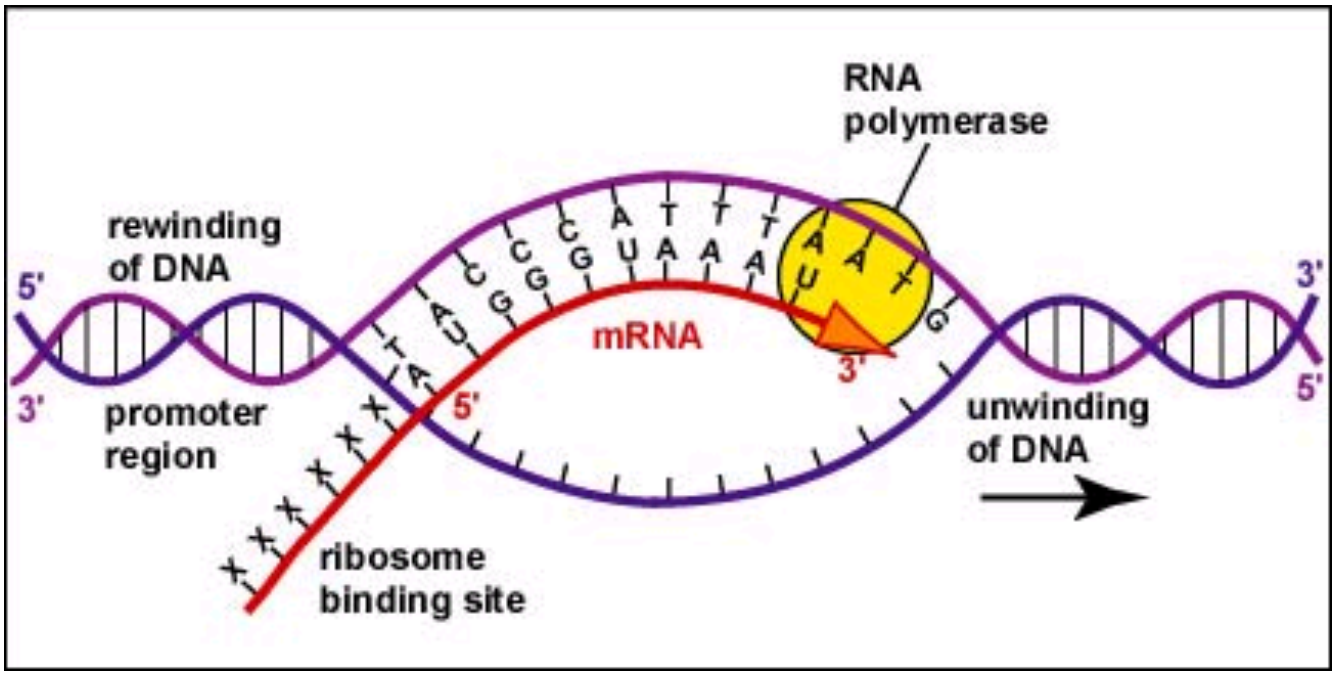
Human chromosomes



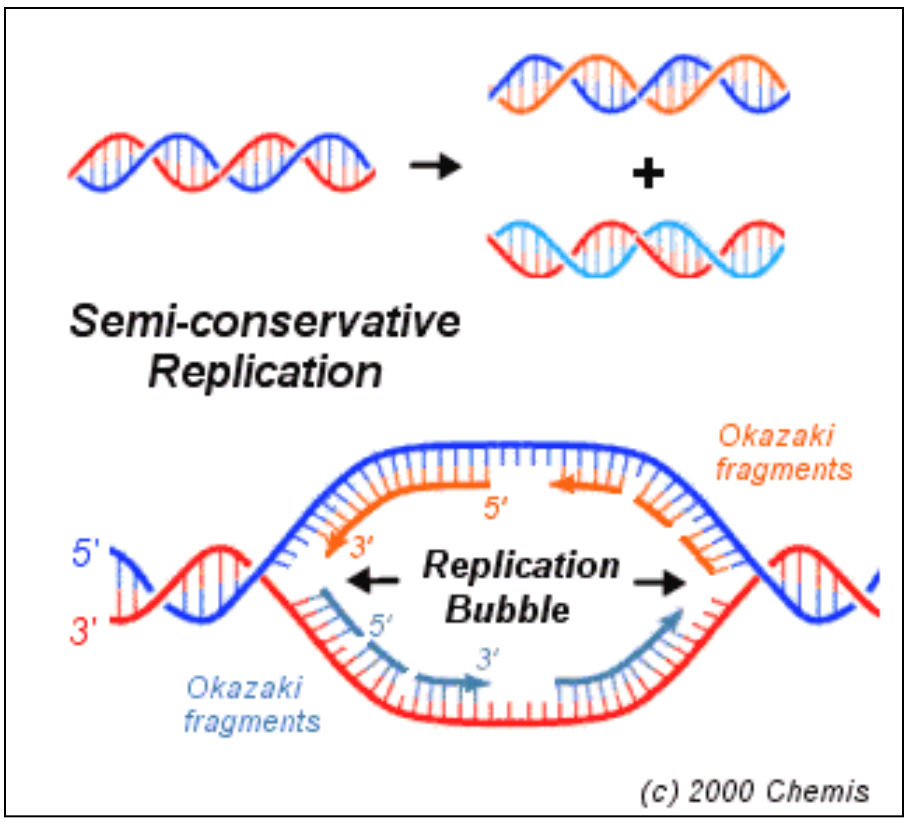




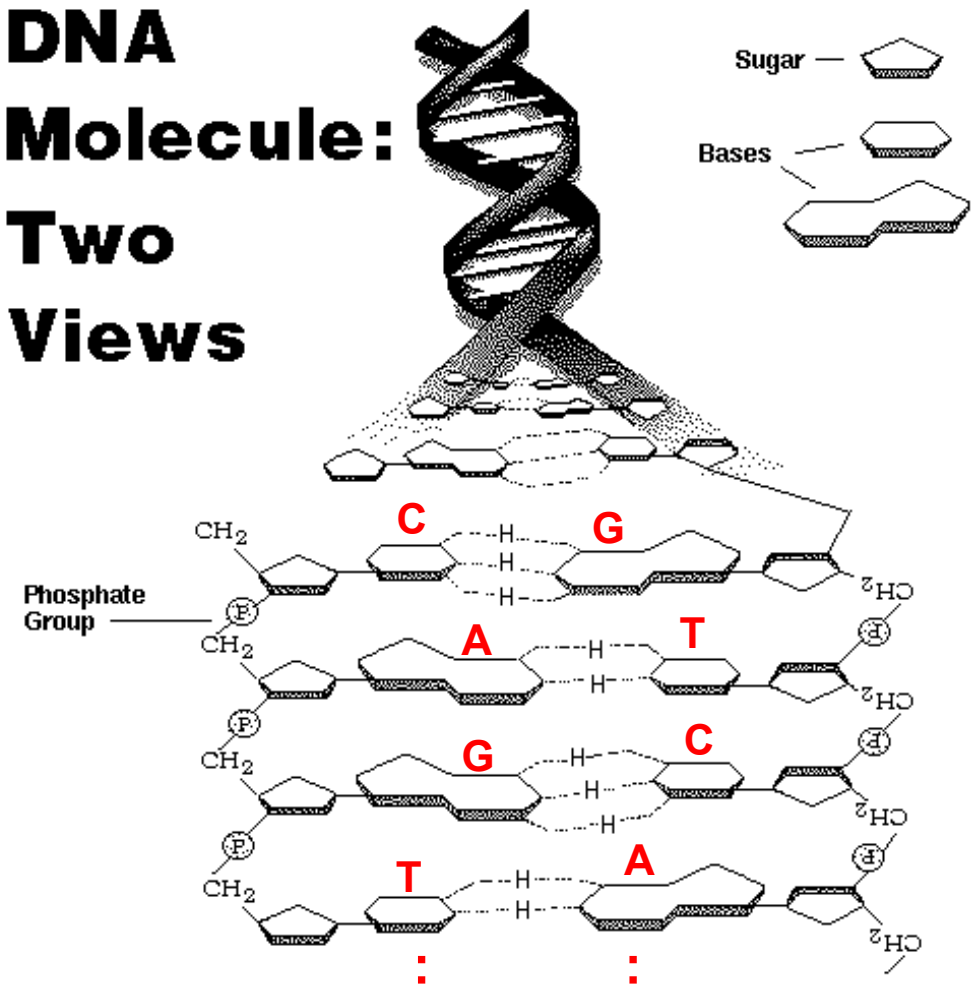
Transcription



Replication



Opening of the double helix with a different environment for each strand => asymmetrical process

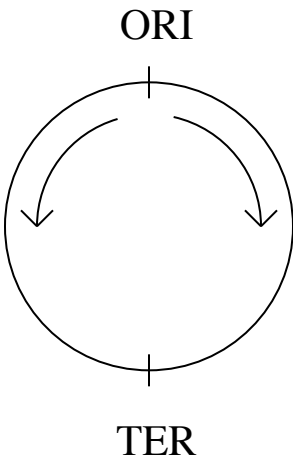


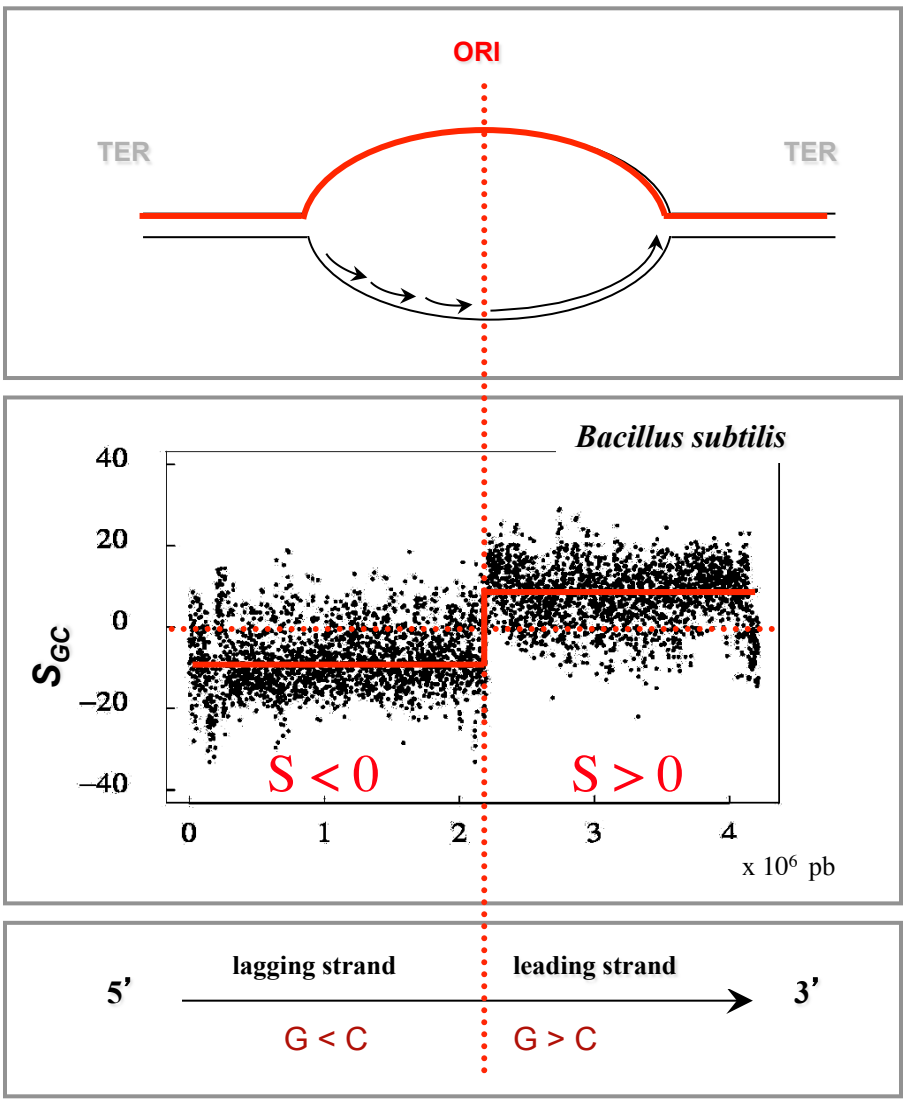
Local symmetry breaking

[G] ≠ [C] ; [A] ≠ [T]

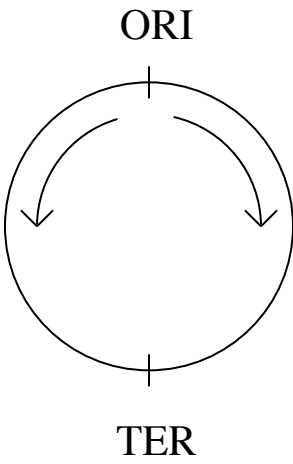
$$S_{GC} = \frac{[G] - [C]}{[G] + [C]} \qquad S_{TA} = \frac{[T] - [A]}{[T] + [A]}$$

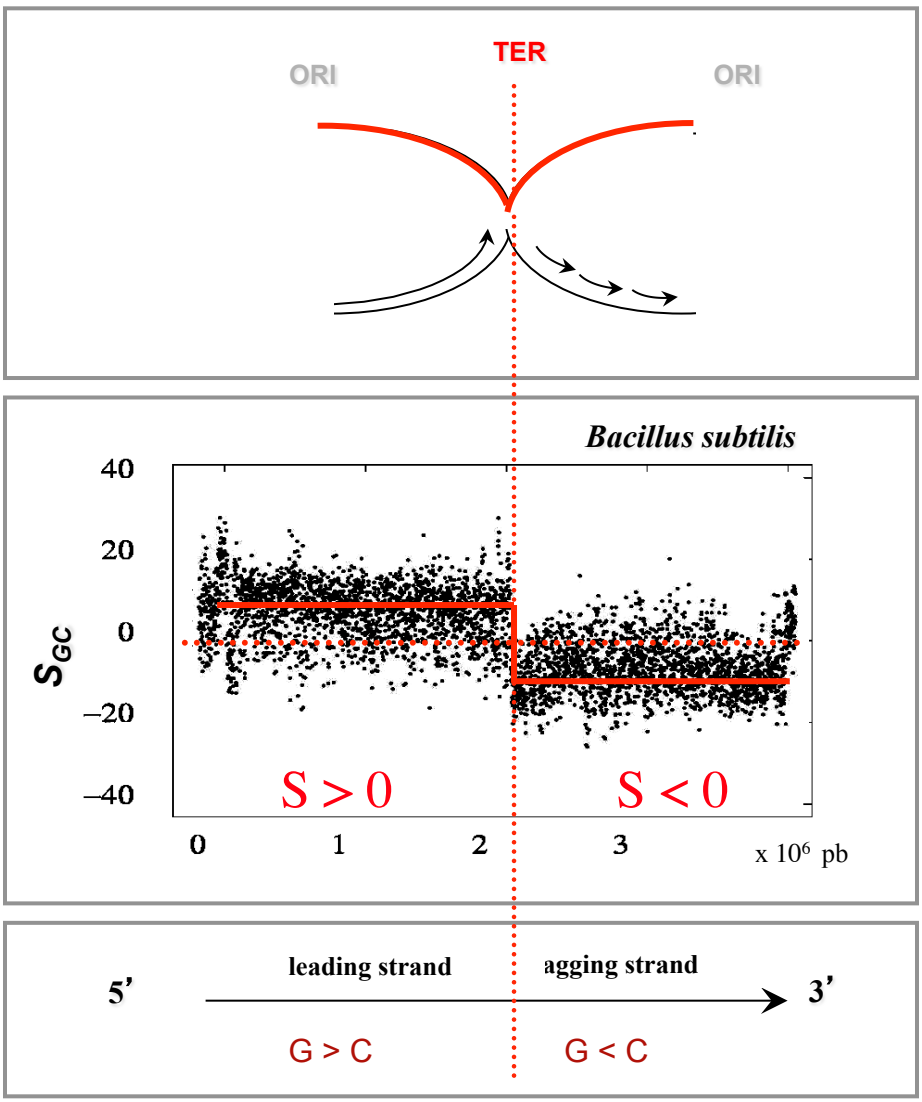
Replication bias in bacteria: An upward jump at the origin



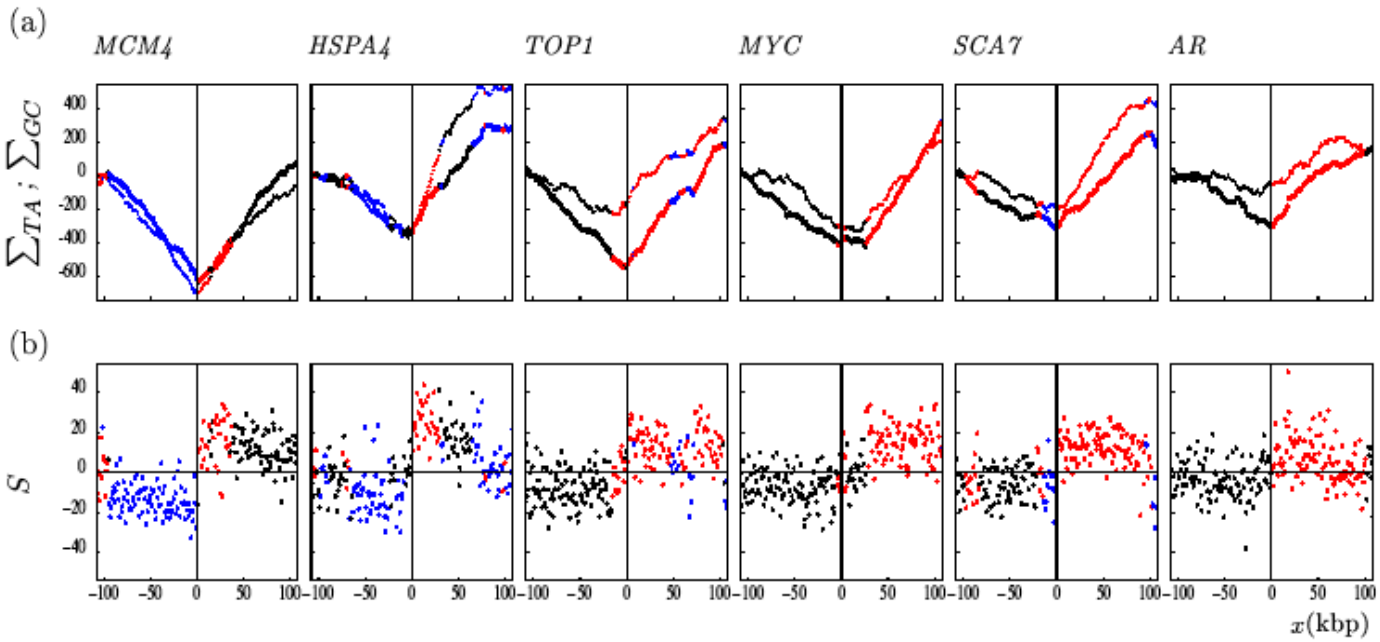


Replication bias in bacteria: A downward jump at the terminus





Skew profiles around human replication origins

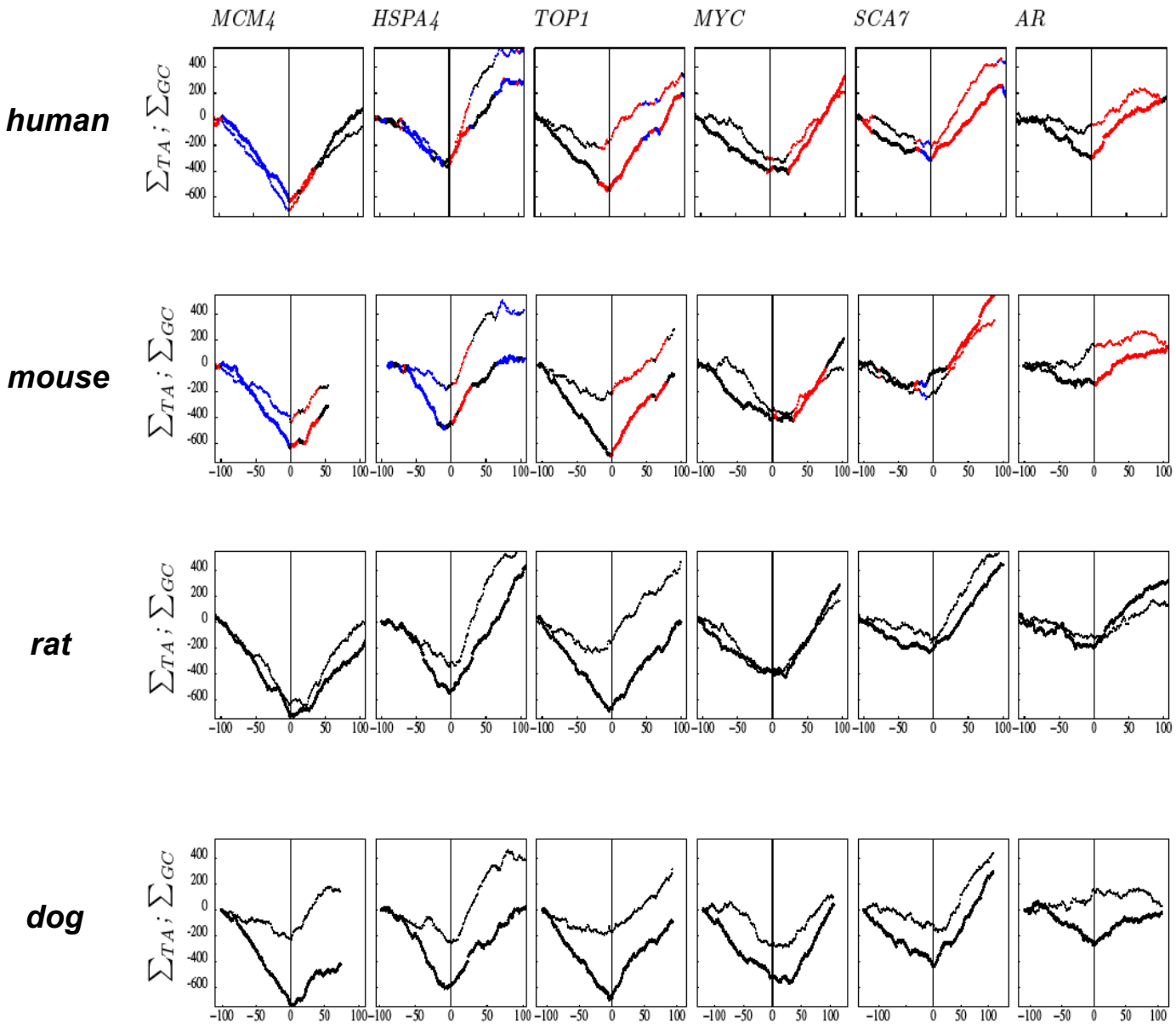


upward jumps : ΔS ~ 24%

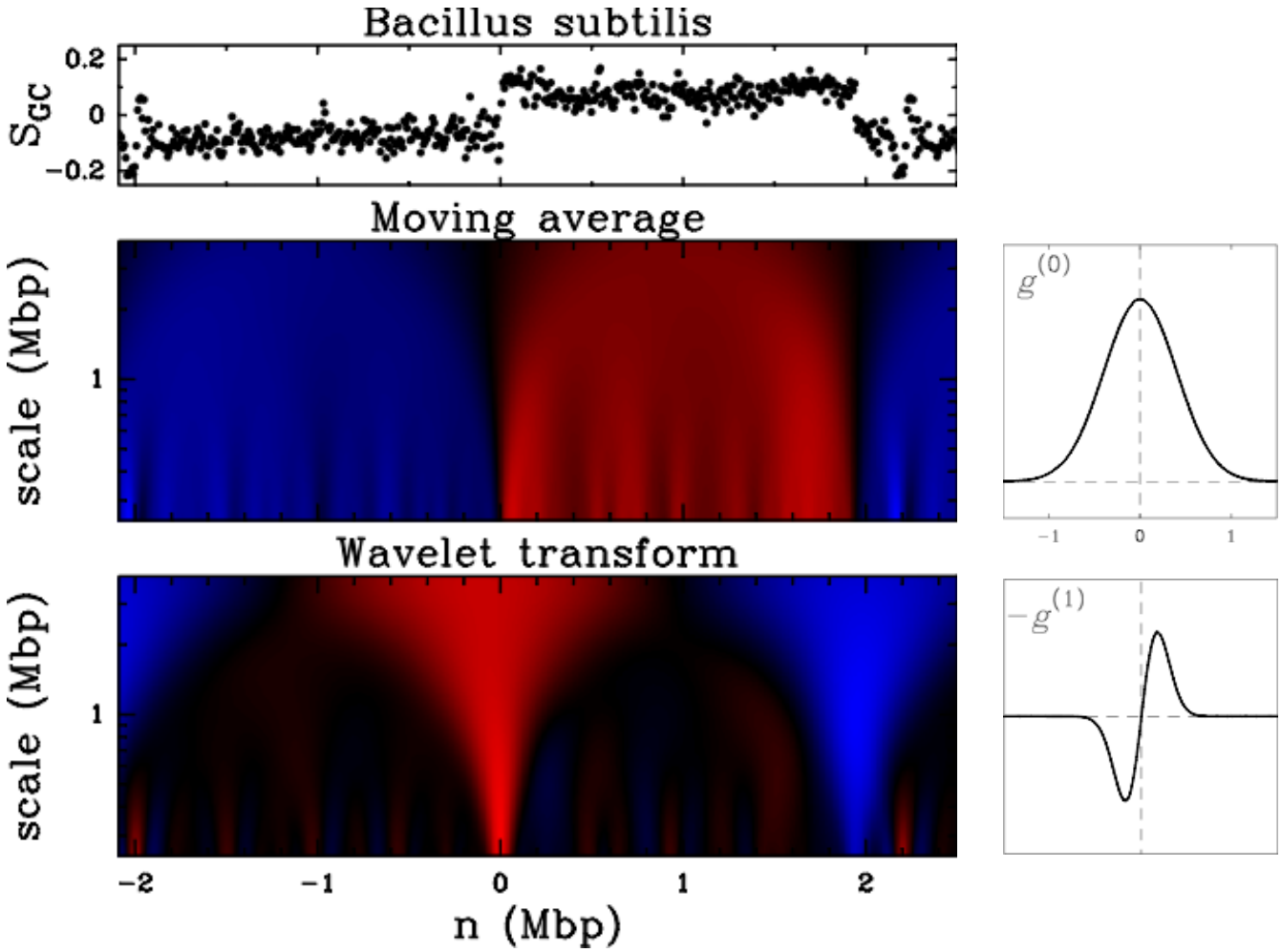
Can these profiles be explained by transcription only ?

Do intergenic regions show upward transitions of the skew *S* ?

Conservation of profiles in mammalian genomes

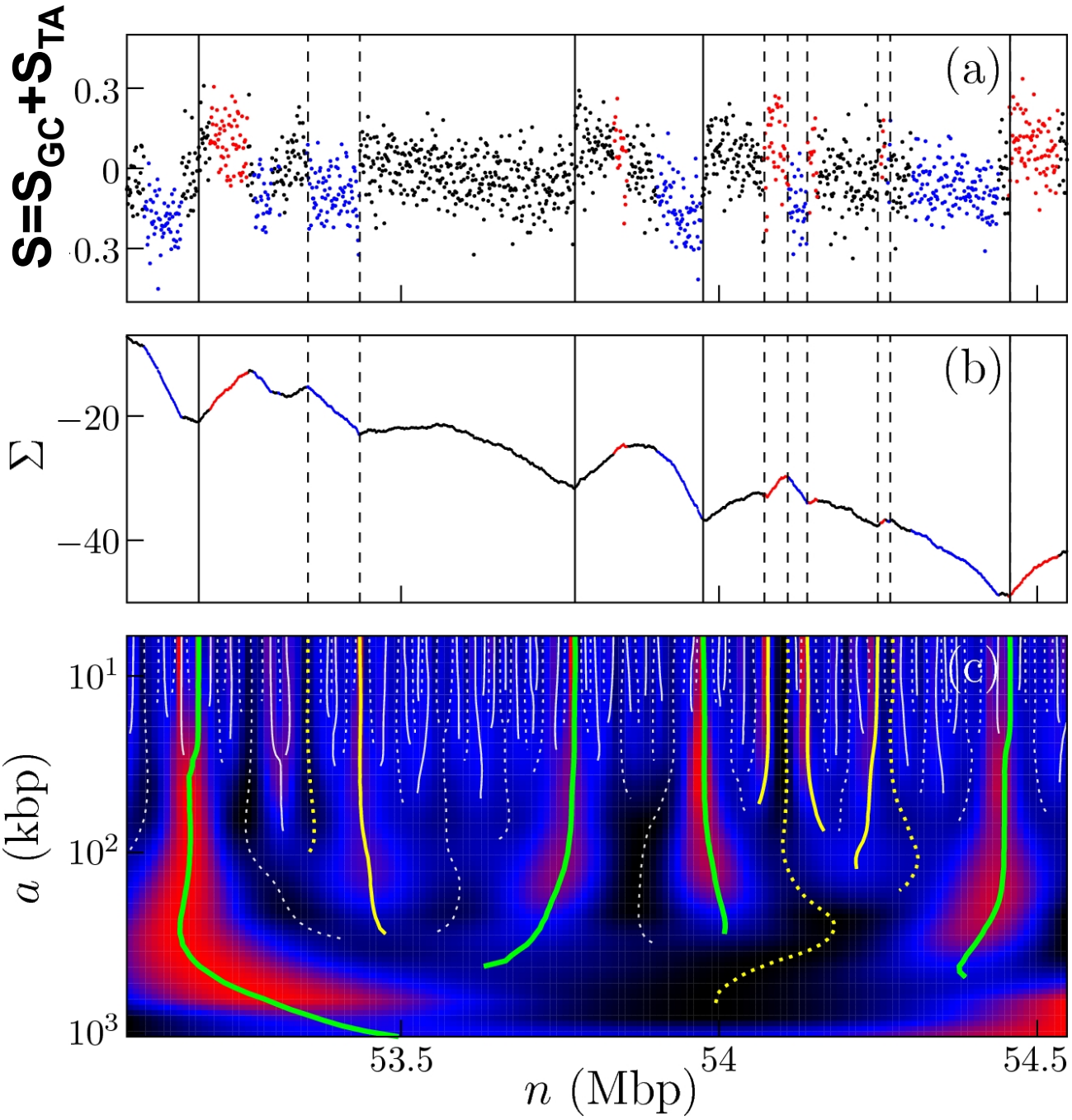


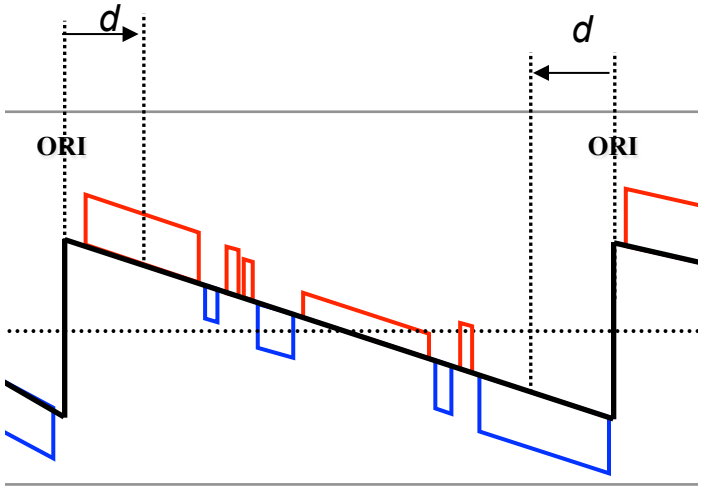
Multi-scale detection of jumps in skew profiles using the wavelet transform



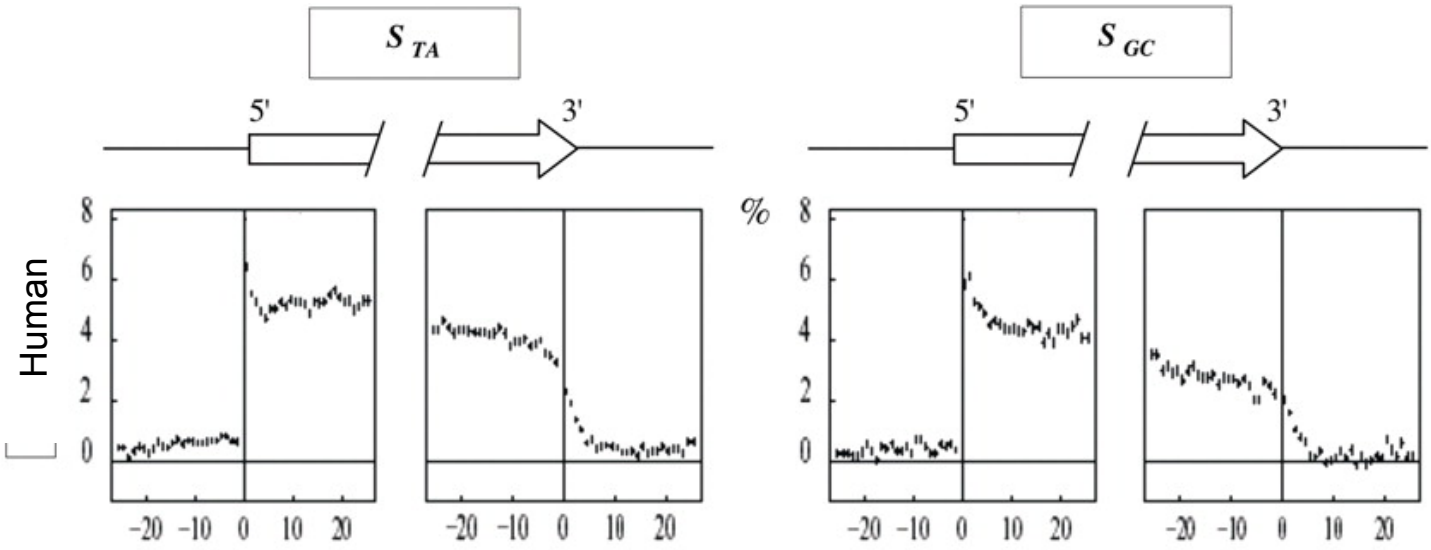
Wavelet transform of the skew profile

Human chromosome 6

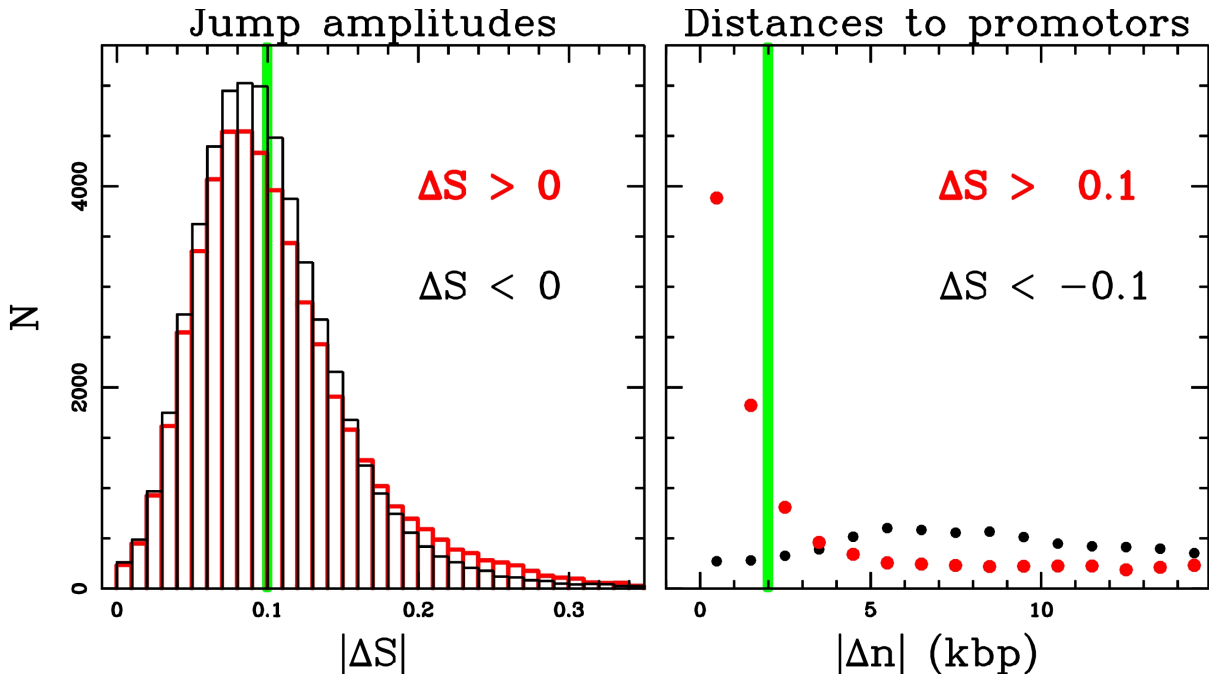




Transcription bias



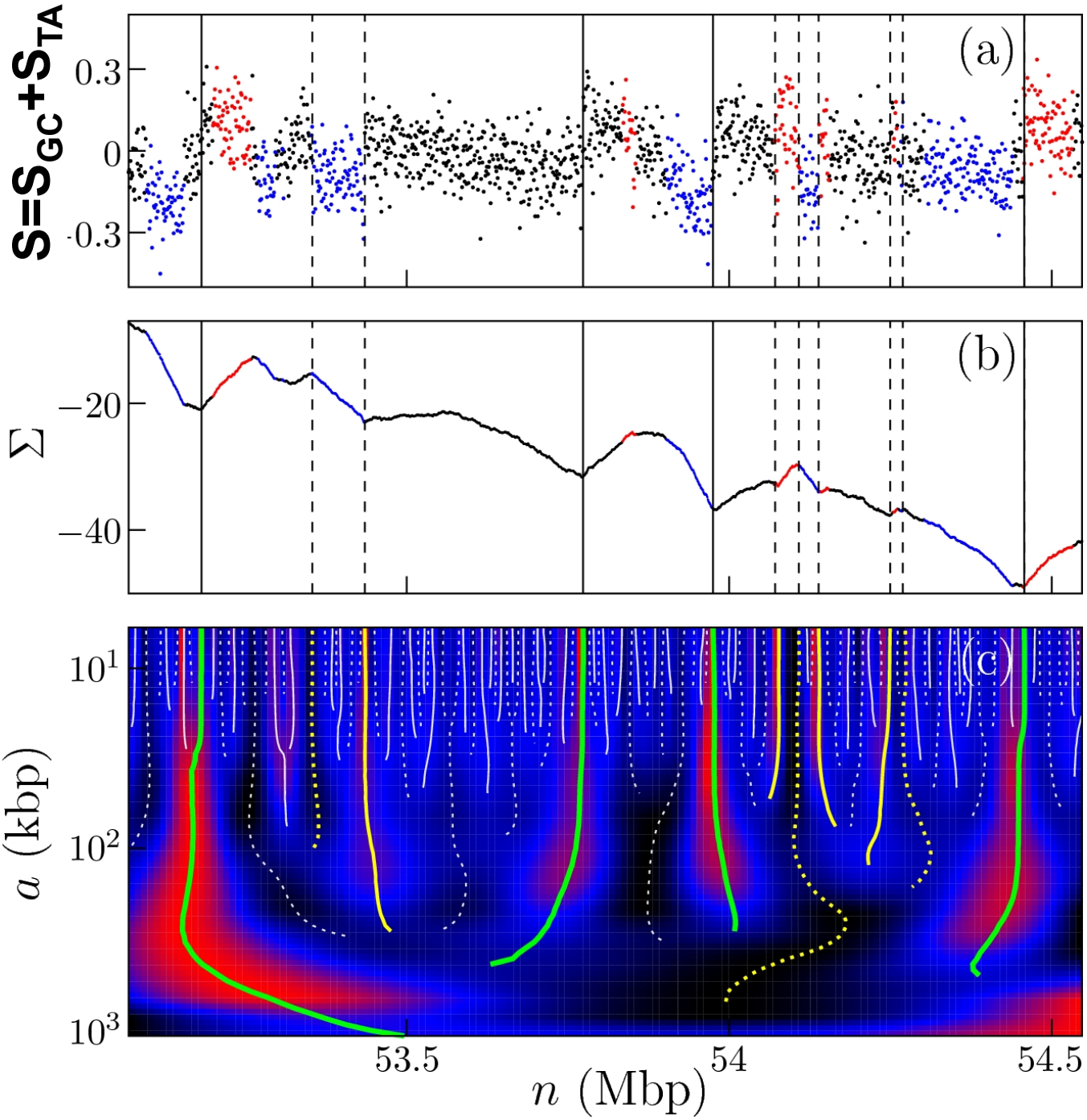
Statistics of upward and downward jumps at scale 10 kbp



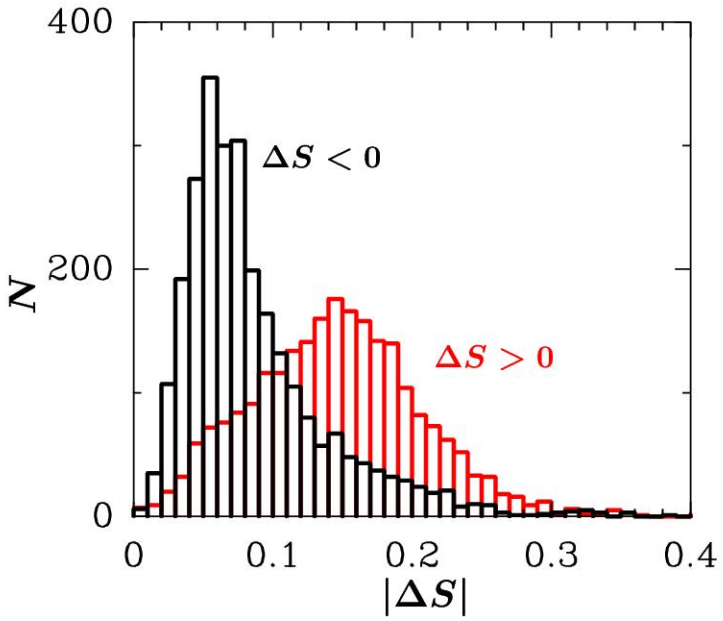
ΔS>0.1, Δn<2kbp: **36%** (7228/20023) significantly biased human genes likely expressed in germ line cells.

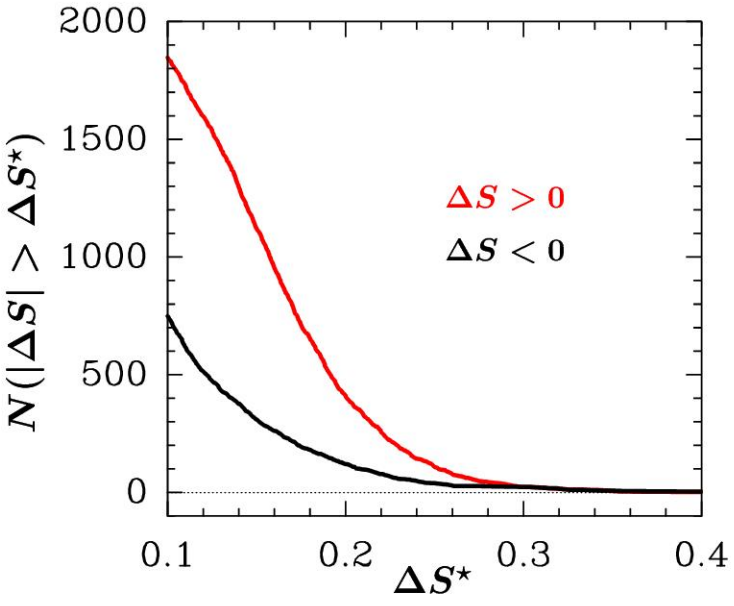
T.I. Lee *et al.*, Cell 125 (2006): **32%** of human ES cell genes bound to Pol II

Jumps in the skew profiles at scales a > 200kbp



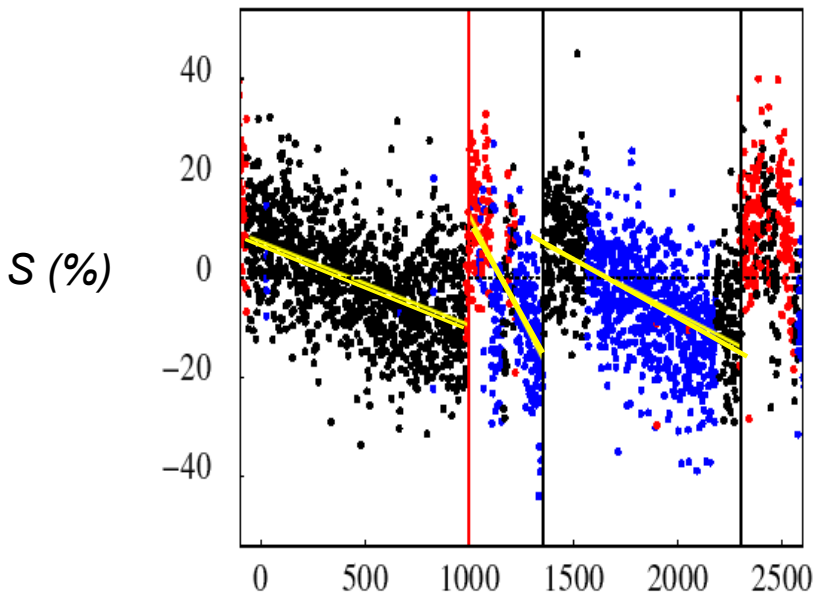
Statistics of upward and downward jumps in Human skew profiles



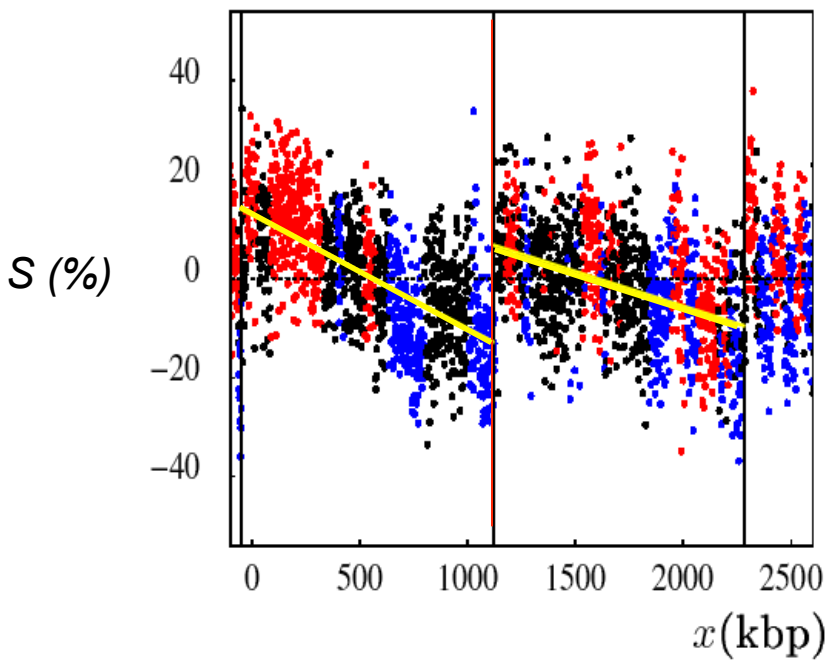


« Factory roof » profiles

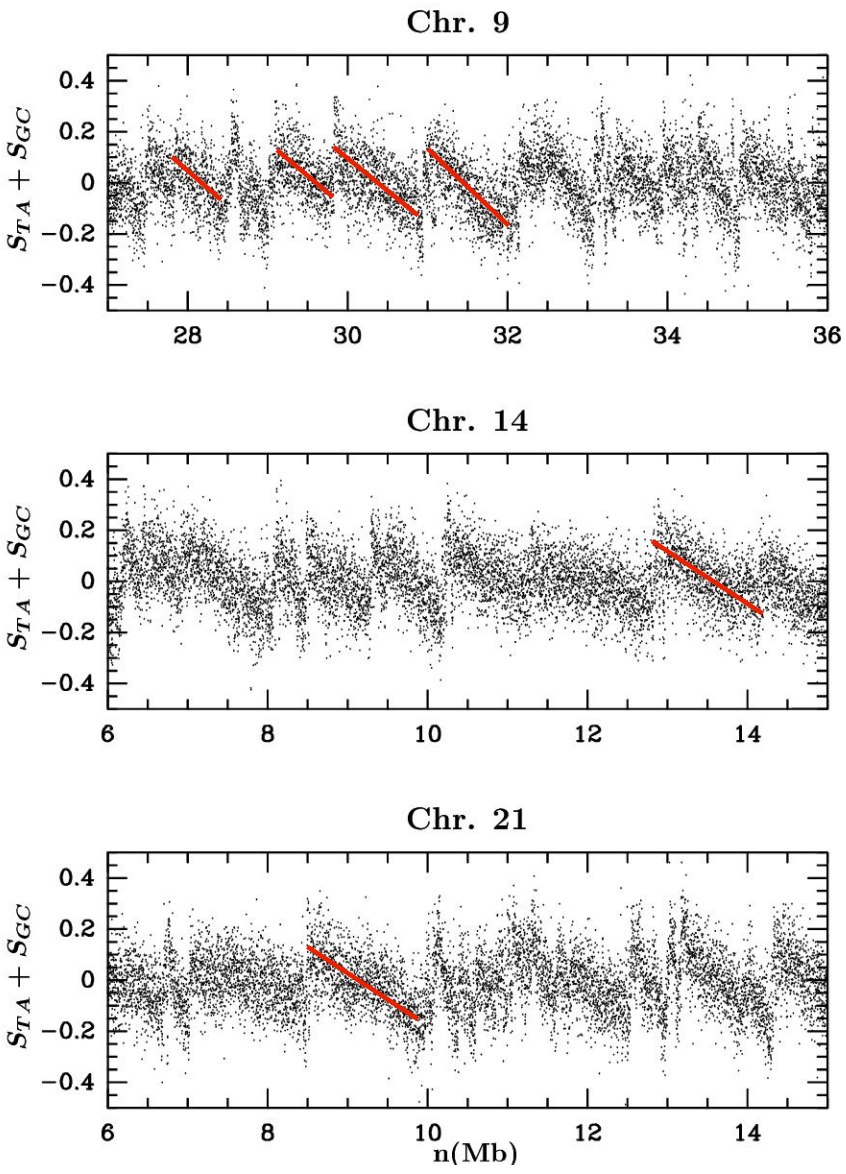
TOP1



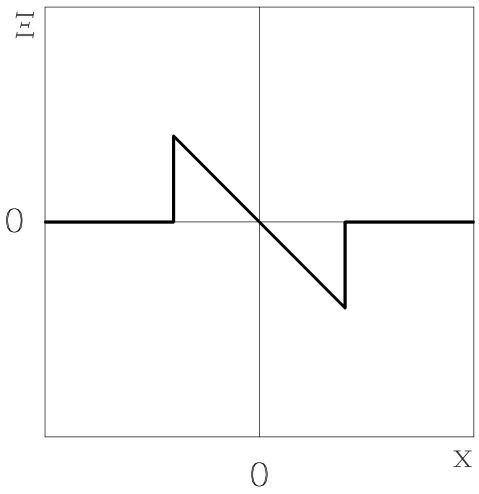
MCM4

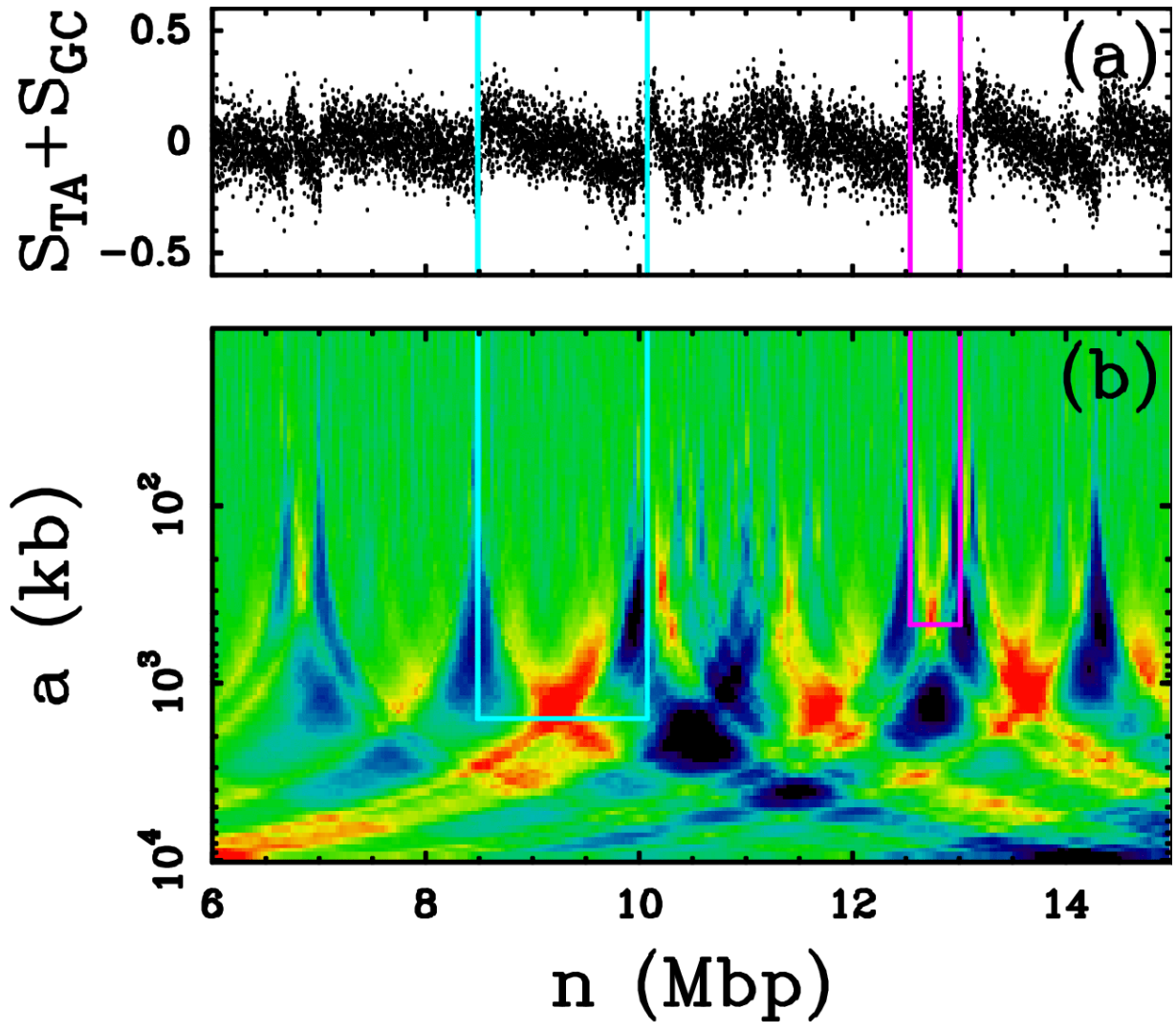


Skew profiles along 9Mb Human contigs

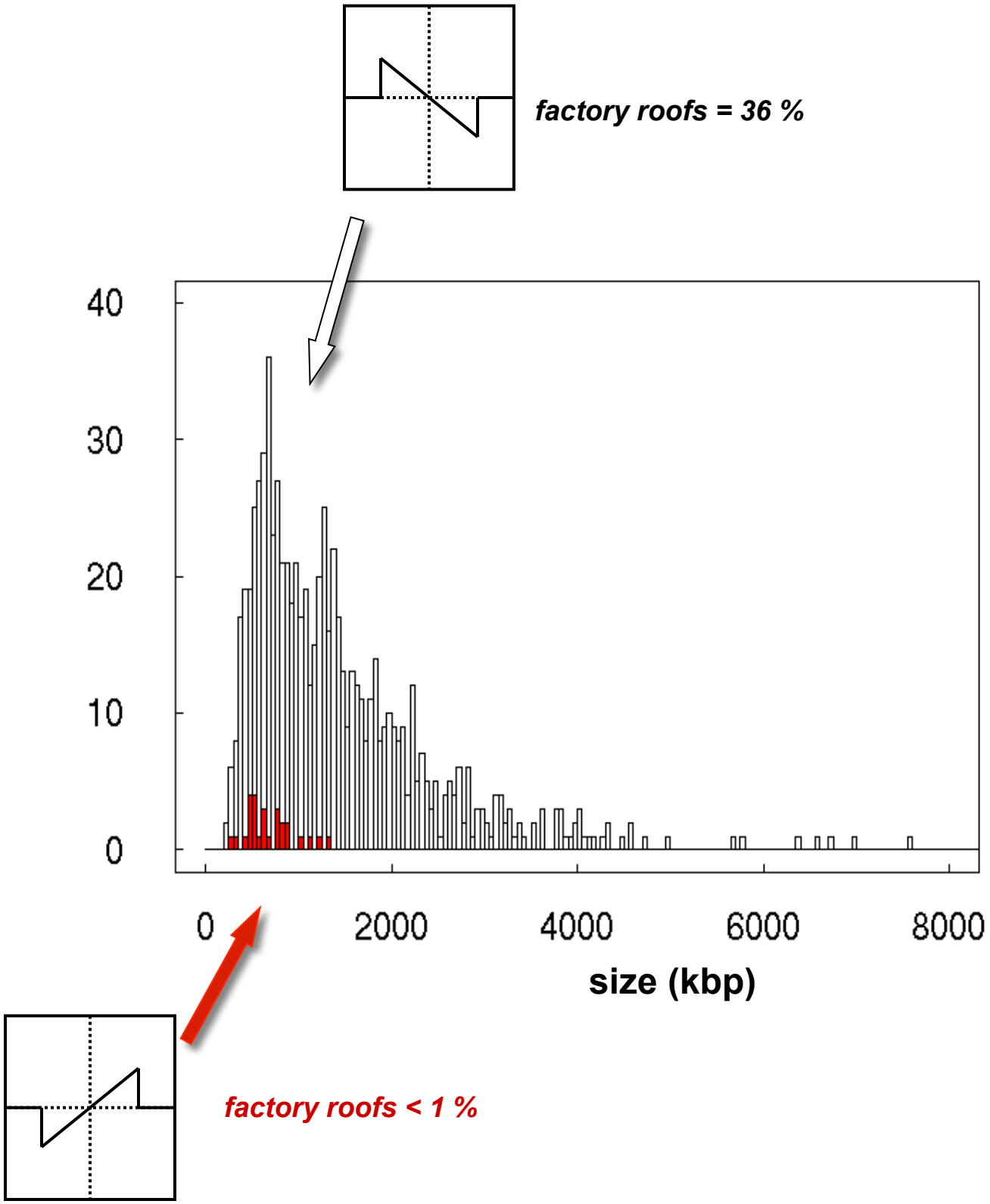


Replication domain detection using an adapted analyzing wavelet

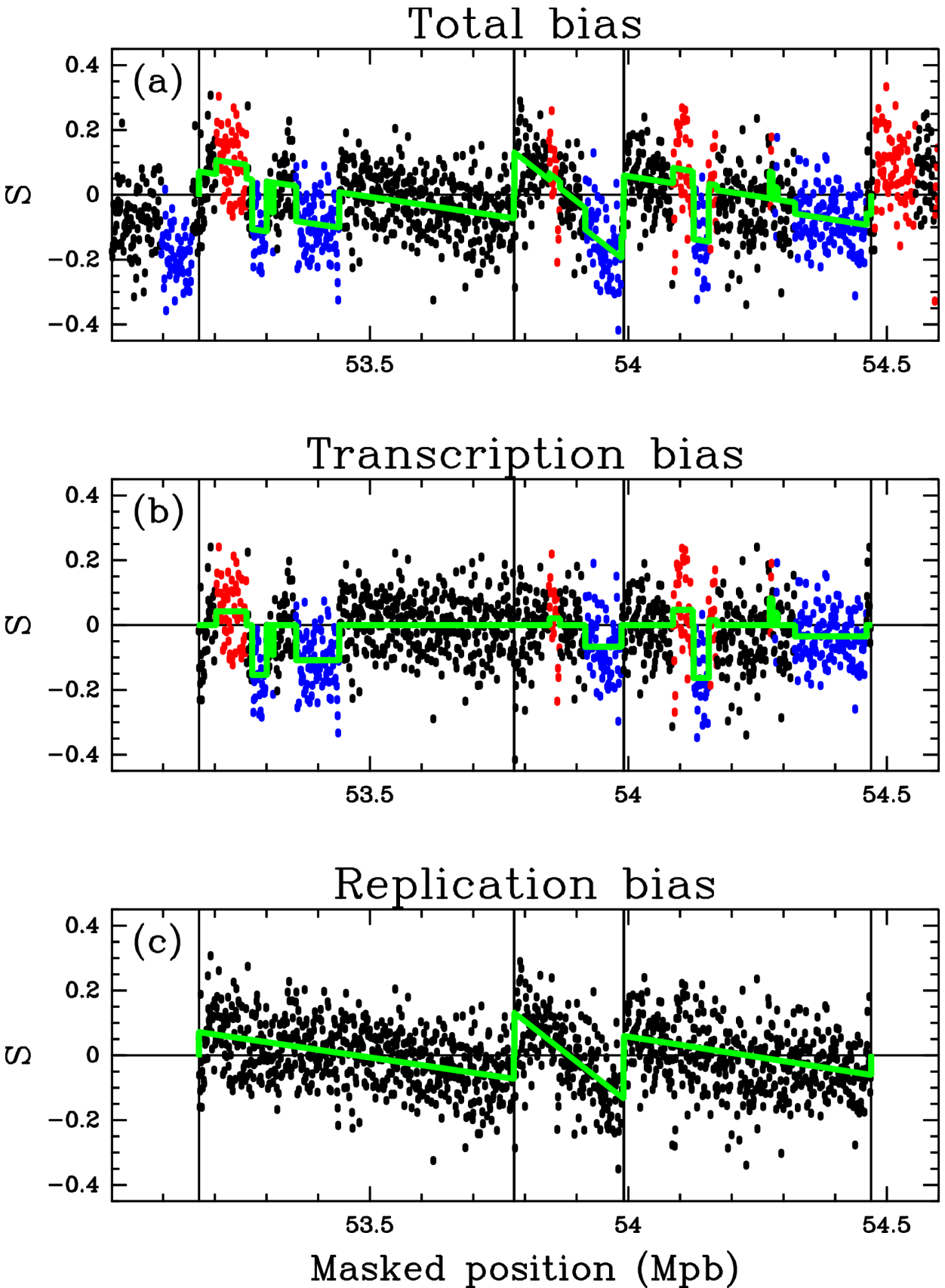




ASYMMETRY OF HUMAN GENOME



Disentangling replication and transcription contributions to skew profiles



Master replication origins at the heart of the organization and fragility of the human genome

Alain Arneodo – alain.arneodo@ens-lyon.fr

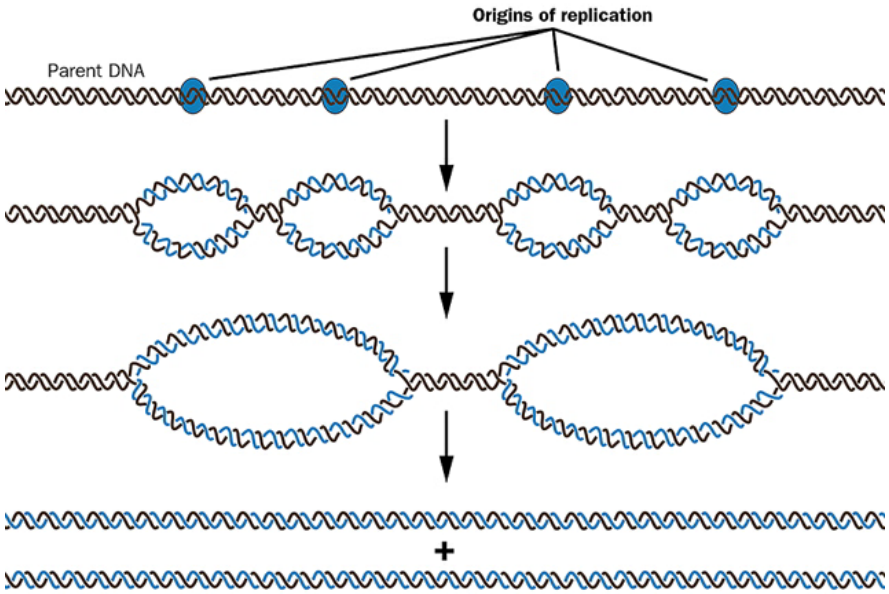
ENS de Lyon
Benjamin Audit
Antoine Baker
Rasha Boulos
Guillaume Chevereau
Pablo Jensen
Hanna Julienne
Antoine Leleu
Cédric Vaillant
Lamia Zaghloul

CGM, Gif-sur-Yvette
Yves d'Aubenton-Carafa
Chun-Long Chen
Claude Thermes

ENS, Paris
Guillaume Guilbaud
Olivier Hyrien
Aurélien Rappailles

CEA, Saclay
Arach Goldar

Identification of replication origins



➢ **Prokaryotes :** computer detection easy and efficient in many eubacteria ; confirmed by experiments

➢ ***S. cerevisiae :*** ARS regions (~ 125 bp ; 11 bp ACS consensus) ; all origins experimentally determined

➢ ***S. pombe :*** ARS (~ 750 bp; no consensus, but AT-stretch) a number of origins experimentally determined

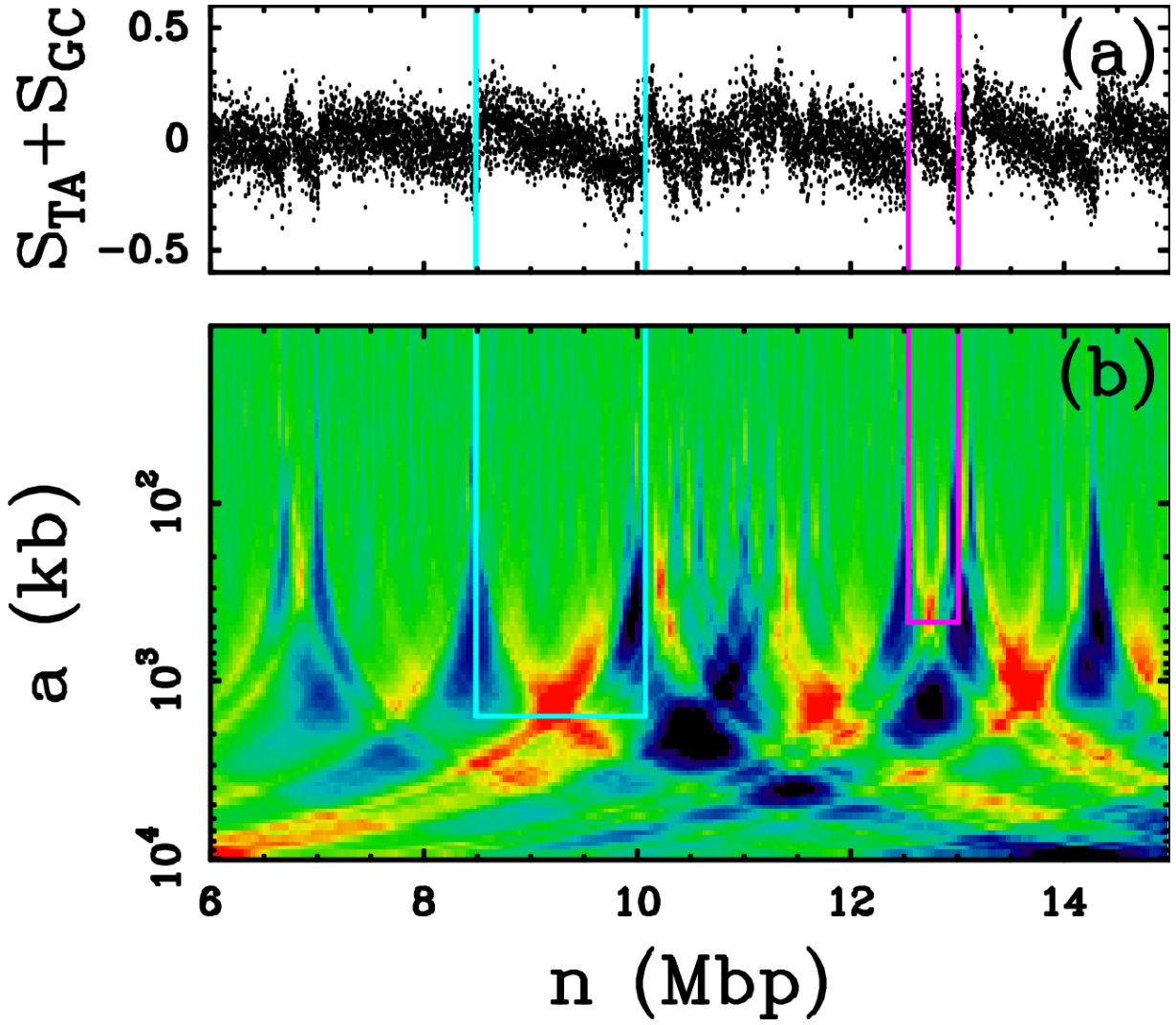
➢ **multicellular eukaryotes :** *replication origins are « terra incognita » !*

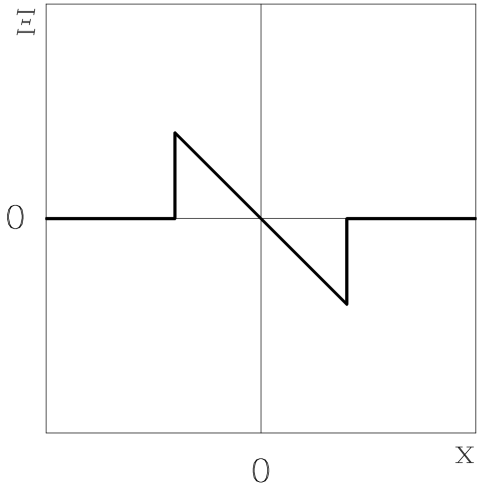
- very few origins experimentally determined
- no consensus sequence (epigenetic elements)

➢ **Human :**

- **10 000 - 30 000 replication origins expected**
- **~ 10 precisely determined**
- **High-throuput methods are now emerging**

Replication domain detection using an adapted analyzing wavelet

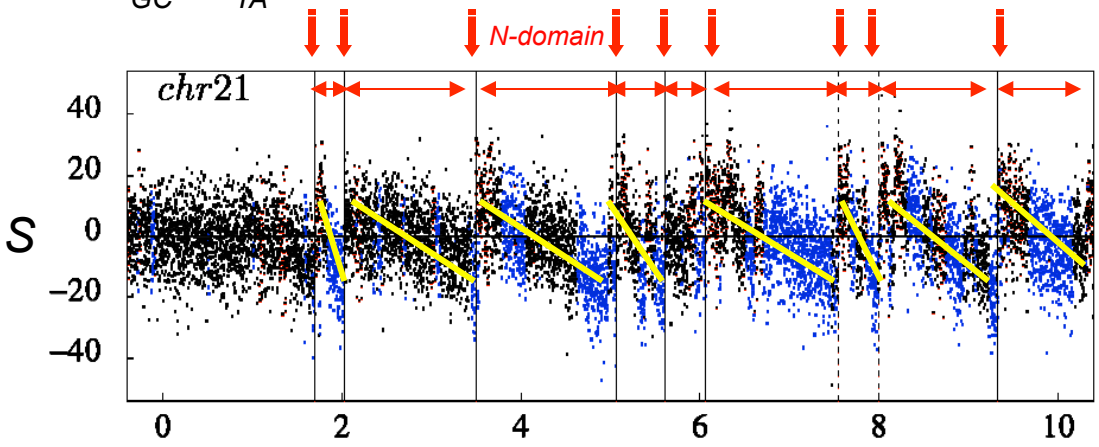




Audit, Phys. Rev. Lett. (2007)
Huvet, Genome Res. (2007)
Baker, ACHA (2010)

Detection of upward jumps of the skew profile in the human genome

Total skew $S = S_{GC} + S_{TA}$

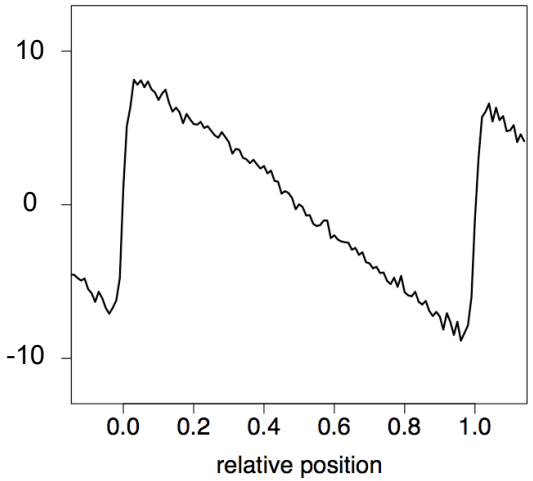


~ 700 N-domains

> 1/3 of the genome

> 1500 putative replication origins

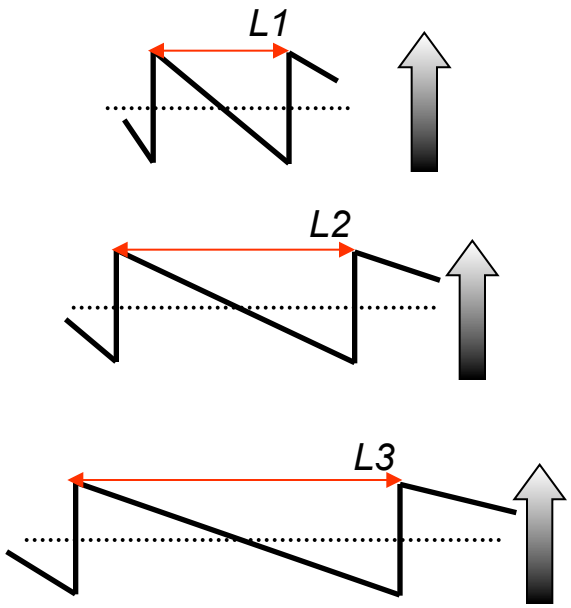
Mean skew *S*

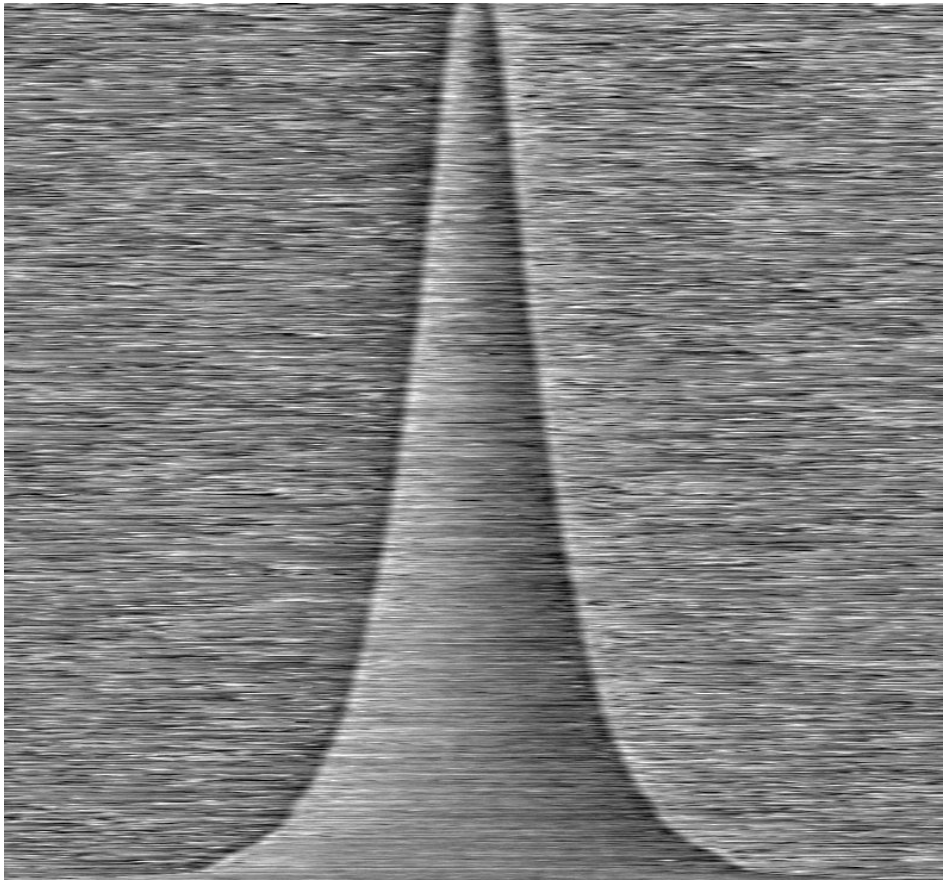


Touchon, Proc. Natl. Acad. Sci. USA (2005)
Brodie and Brodie, Phys. Rev. Lett. (2005)
Huvet, Genome Research (2007)

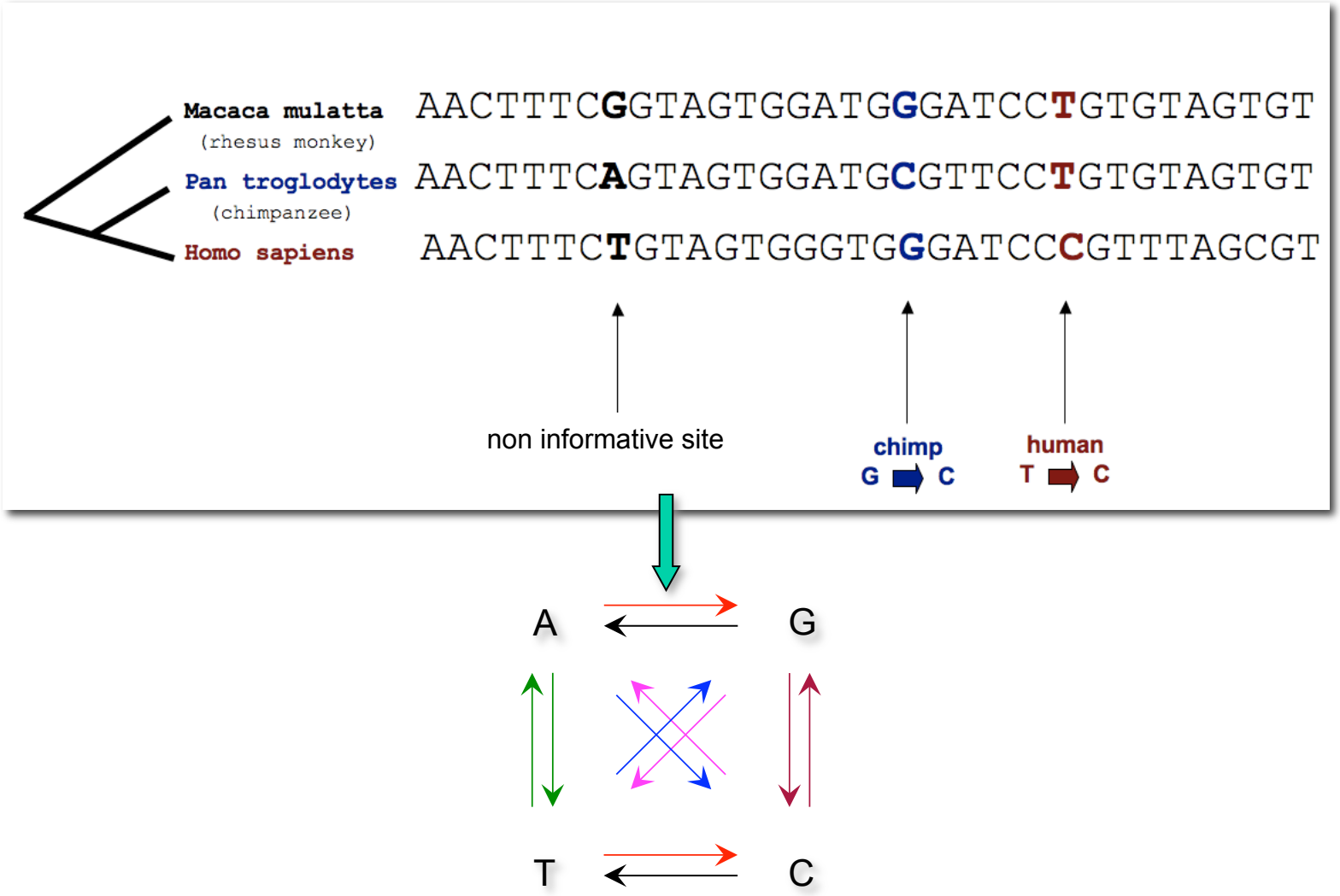
Skew profile of the N-domains

$S = S_{GC} + S_{TA}$



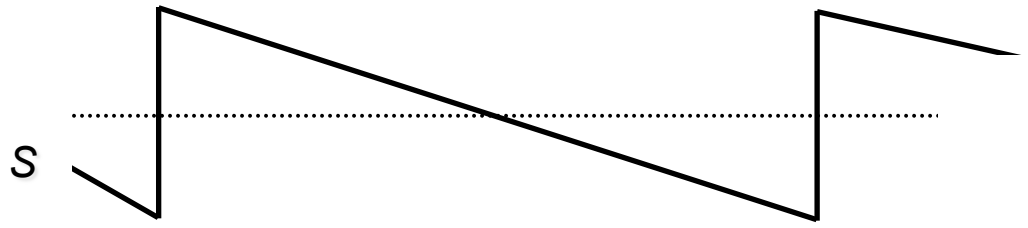


Determination of substitution rates

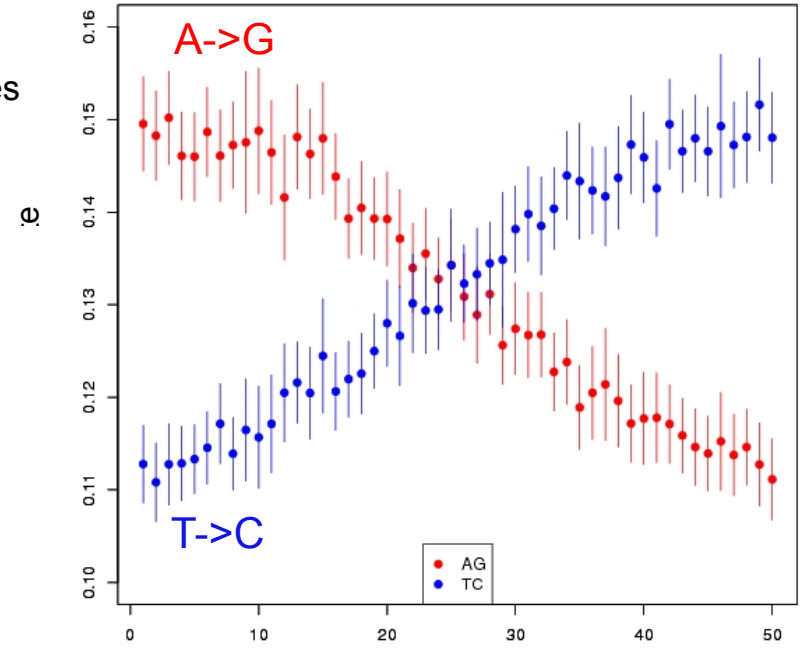


Chen, Genome Res. (2010)
Chen, Mol. Biol. Evol. (2011)

Substitution rates along the N-domains



Normalized substitution rates

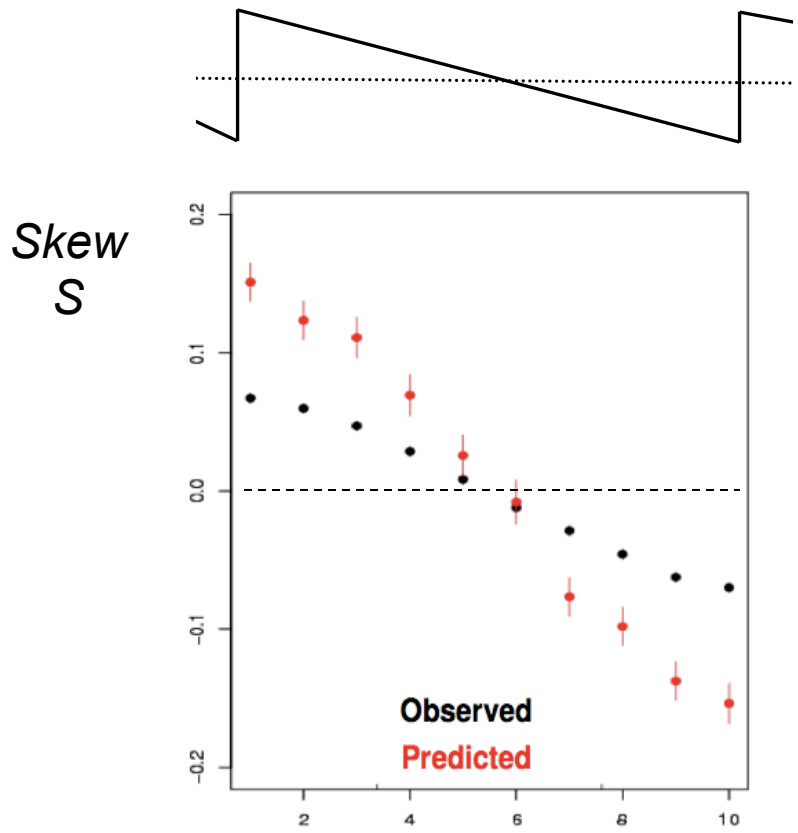


Relative position in the domain

Replication induces more A->G than T->C on the leading strand

Chen, Genome Res. (2010)
Chen, Mol. Biol. Evol. (2011)

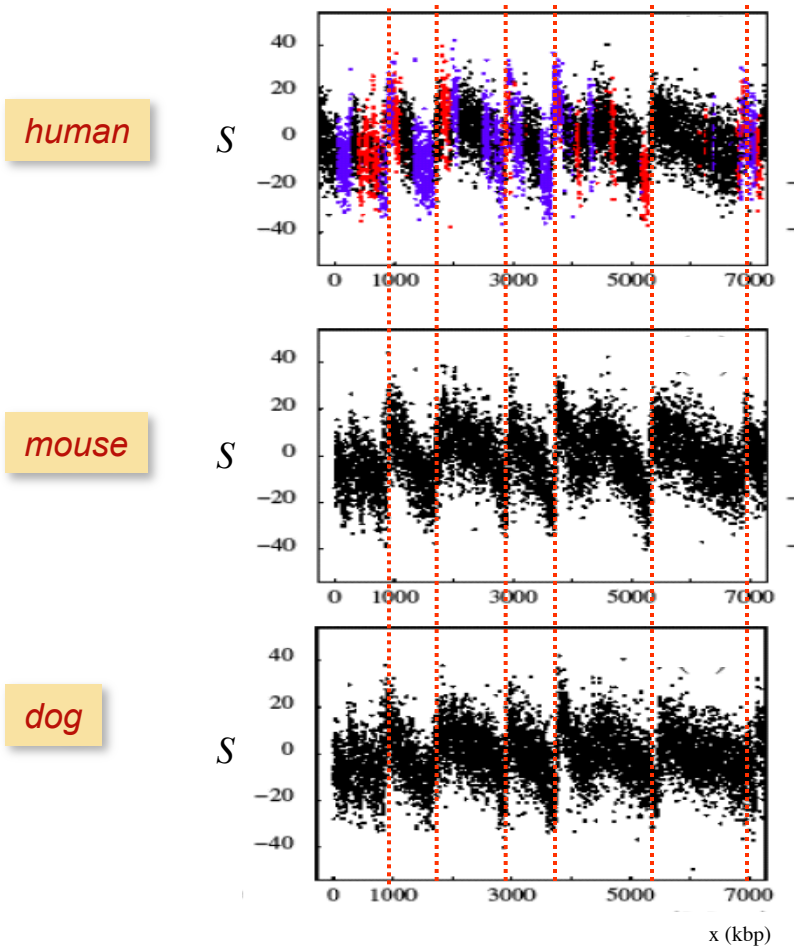
Composition at equilibrium reproduces perfectly the N skew profile



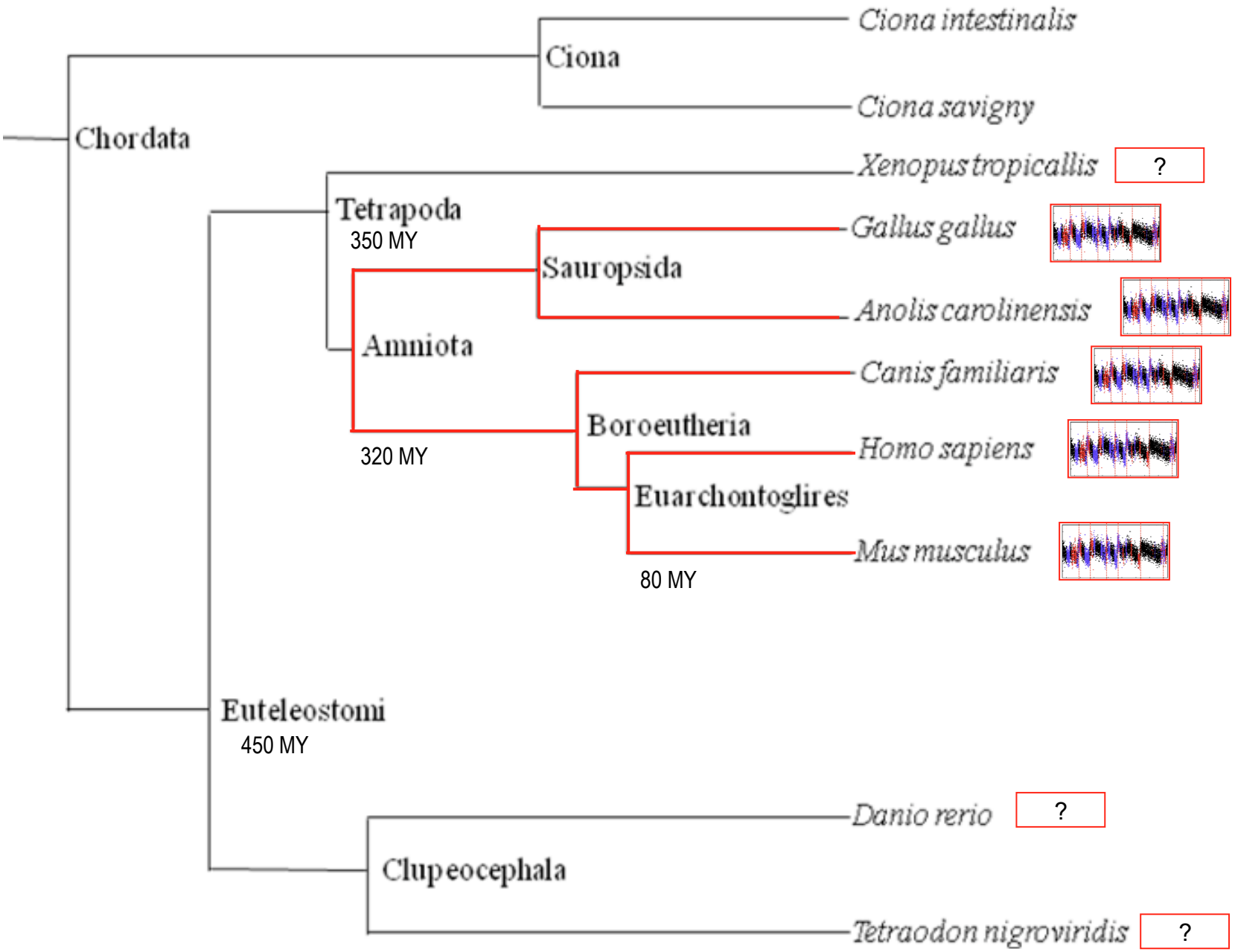
The skew is not at equilibrium

N-domains result from mutation asymmetry in germline cells

Conservation of N-domains in mammalian genomes



N-domains are at least 320.10^6 years old



Determination of replication timing profile by massive sequencing of new born replicated DNA

A. Rappailles, G. Guilbaud, O. Hyrien

Asynchronous HeLa cells

Incorporation BrdU + Fixation + Cell sorting

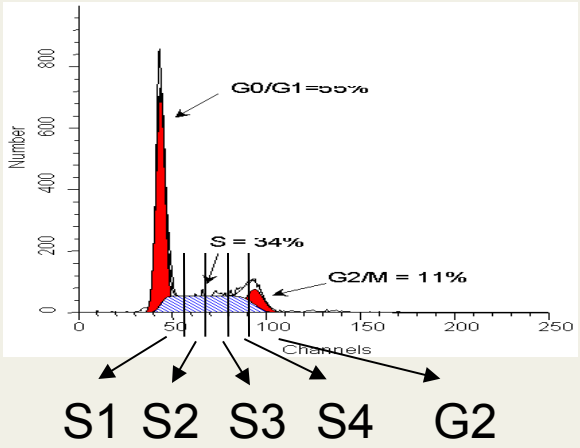
S1, S2, S3, S4 | S | G2

DNA Extraction+Sonication (250 à 4000 pb)+Denature

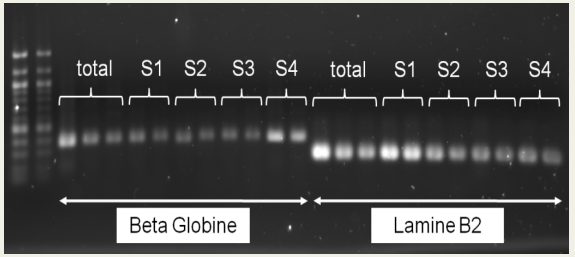
Immuno-precipitation

Random priming 10 minutes (generate double strand)

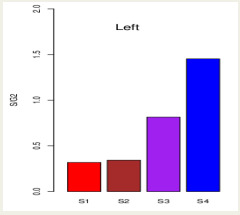
Ilumina Solexa sequencing
(Sequence reads: 43,974,937)



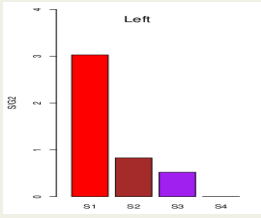
qPCR test



Beta Globin
(late replication)



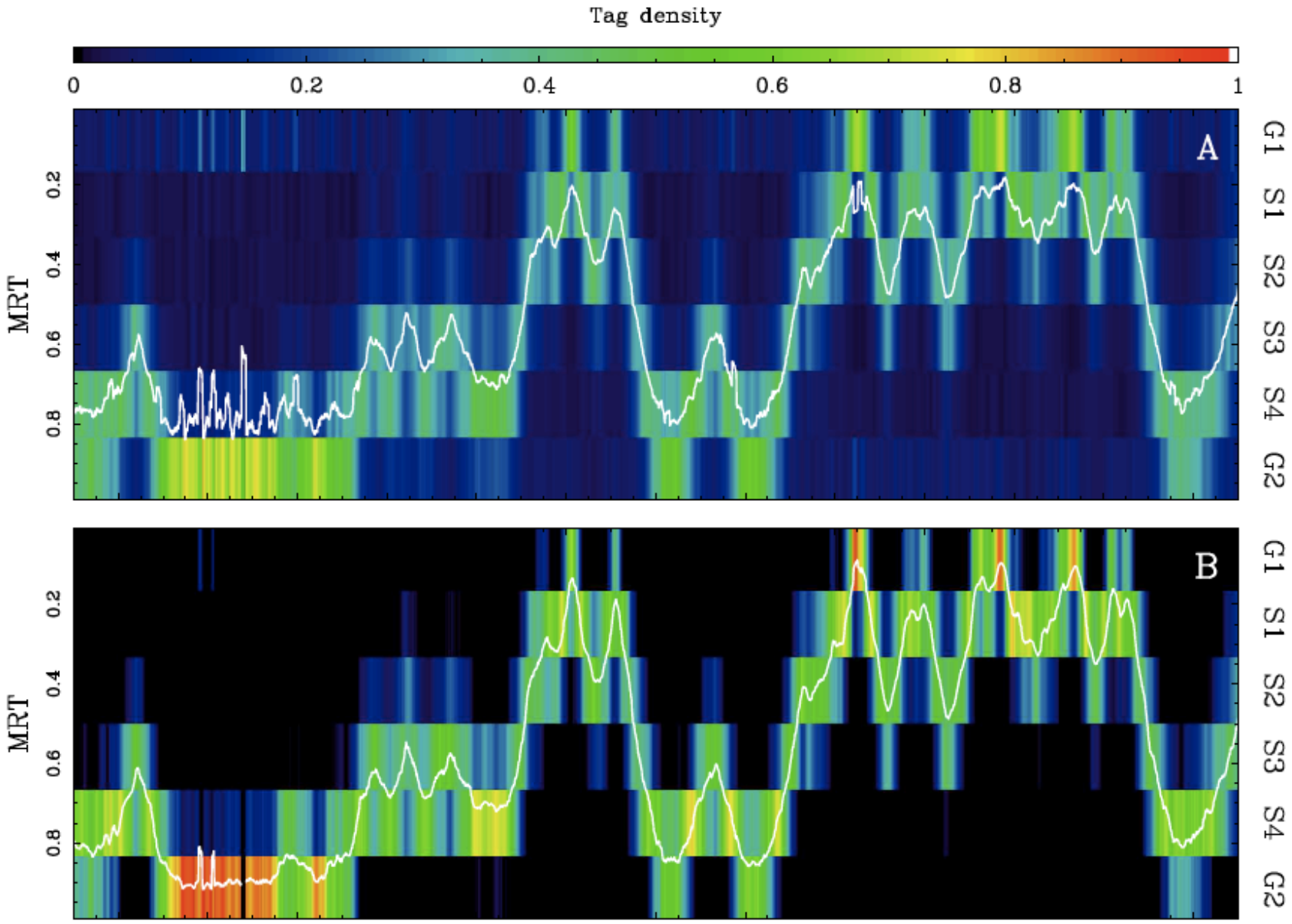
Lamin B2
(early replication)



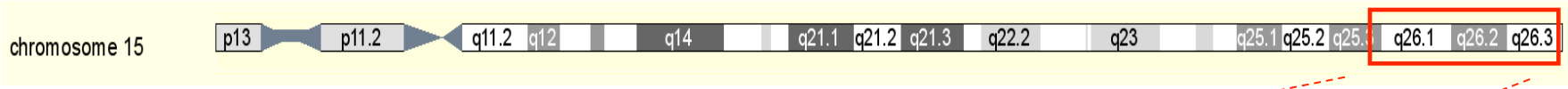
Chen, Genome Research (2010); Rappailles, PLoS Genetics (submitted)

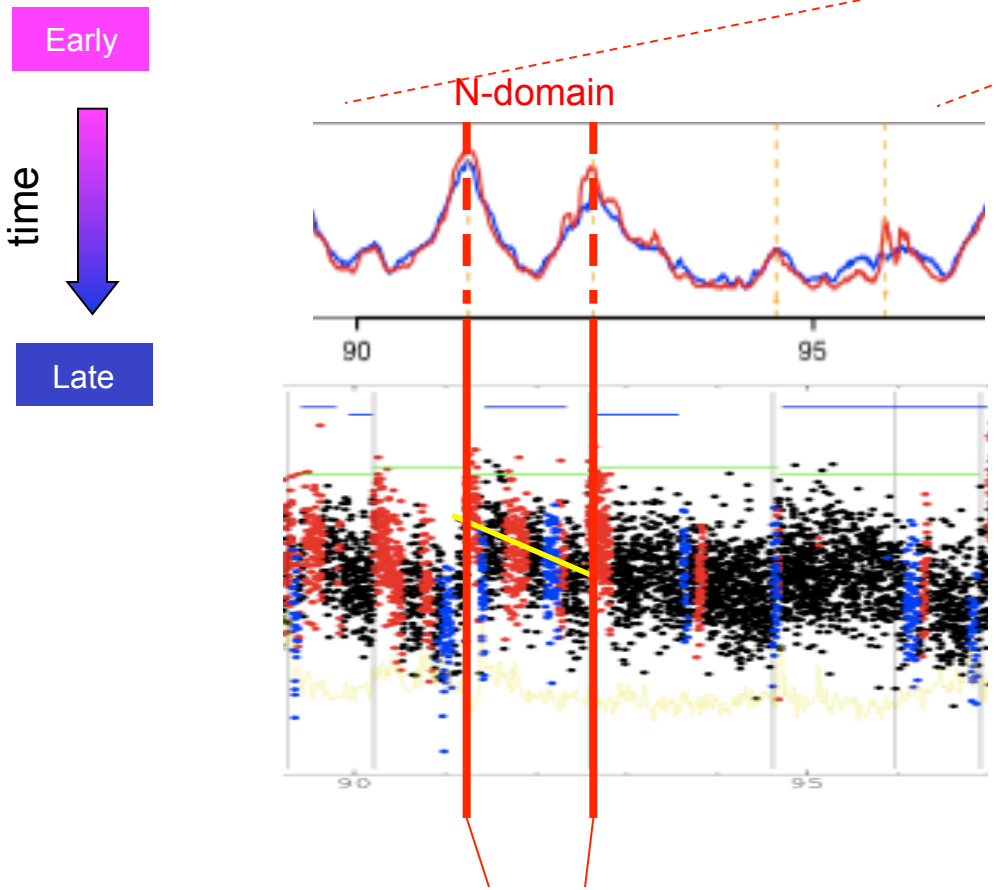
Computing Mean Replication Timing profiles from RepliSeq data

(data from Chen, Genome Research (2010) and Hansen, PNAS (2010))



Comparison of upward jumps with initiation zones

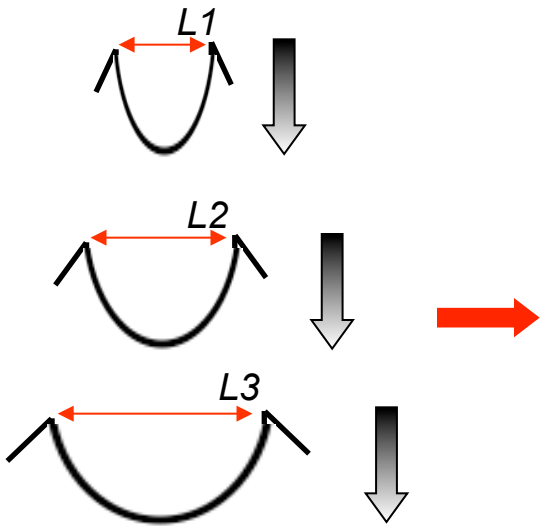




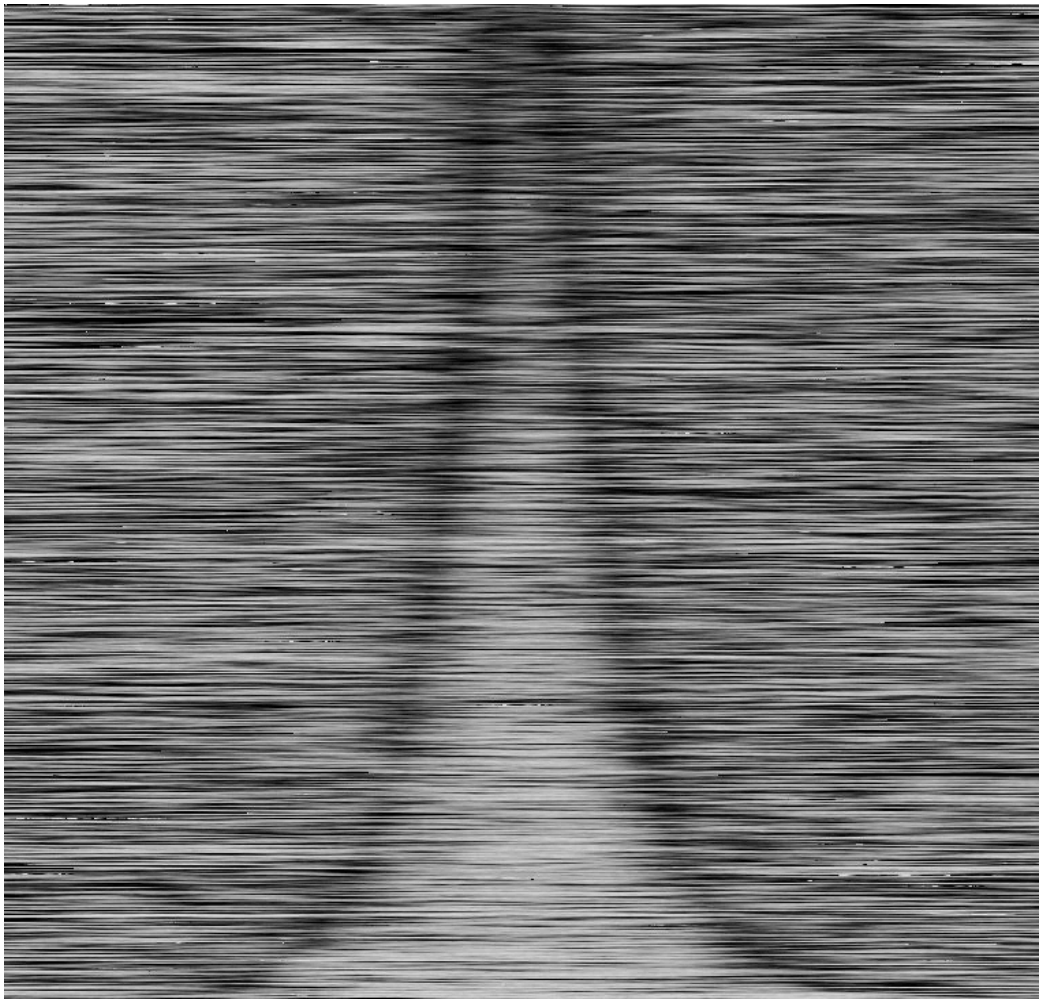
Audit, Phys. Rev. Lett. (2007)
Huvet, Genome Res. (2007)

Replication timing profile within N-domains

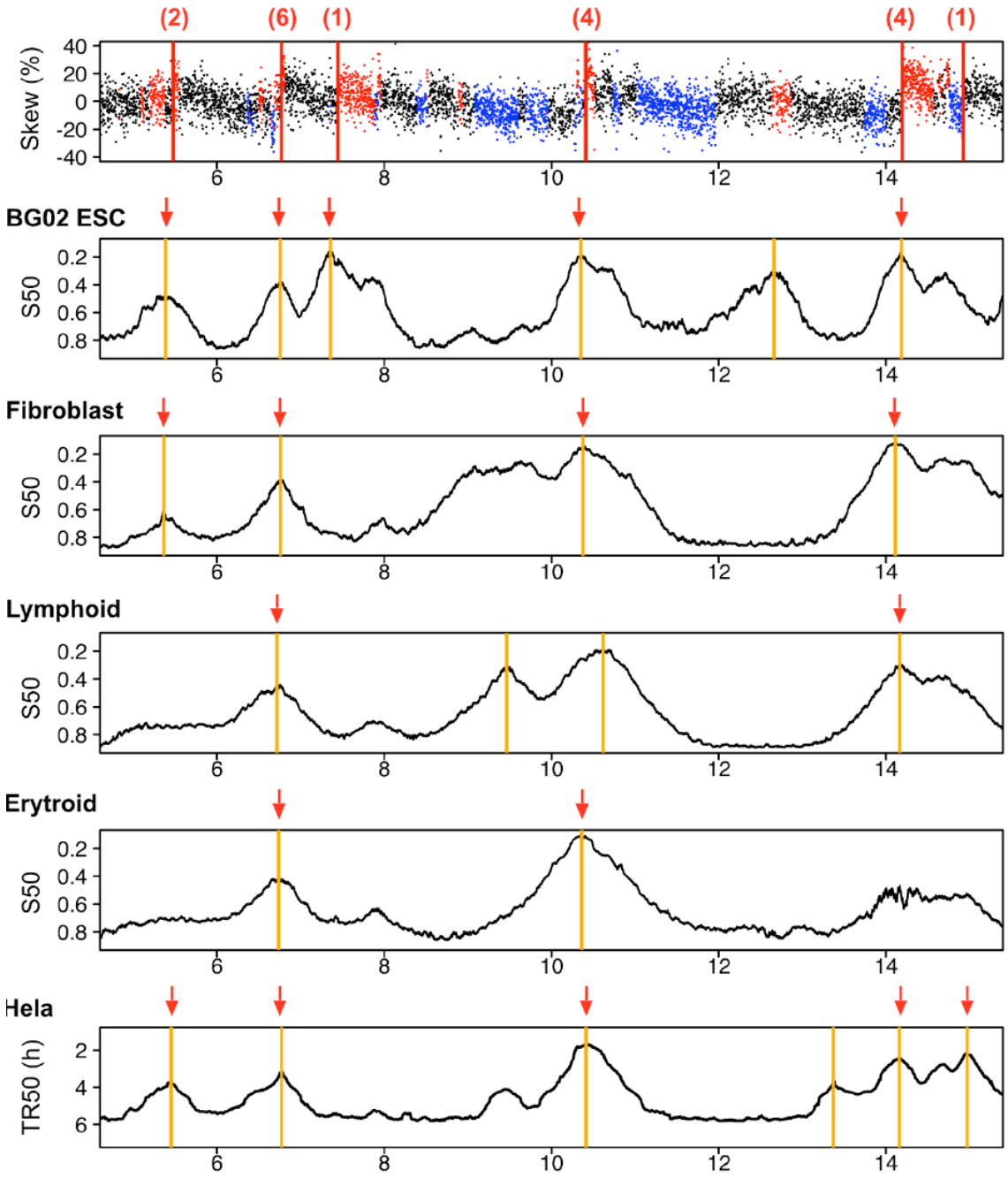
Replication timing



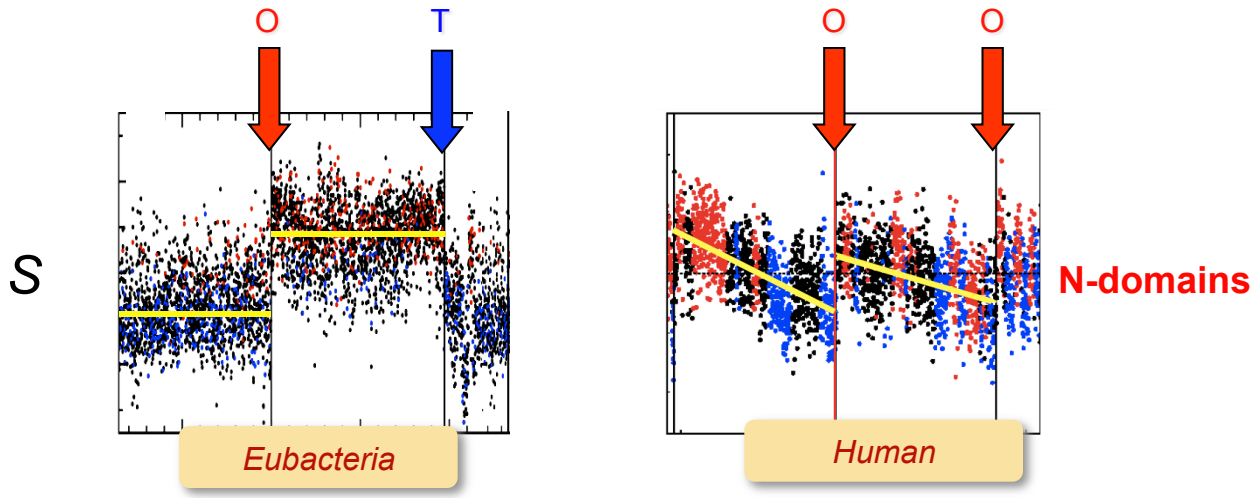
Embryonic stem cells
data from Hansen, PNAS (2010)



Replication timing across cell differentiation

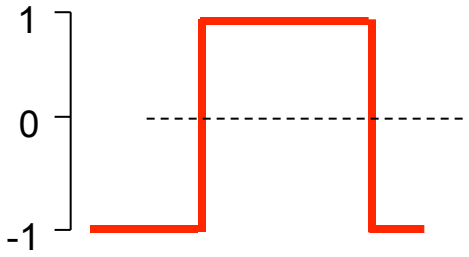


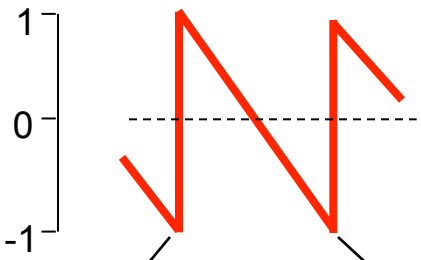
N-shape results from gradient of replication fork polarity



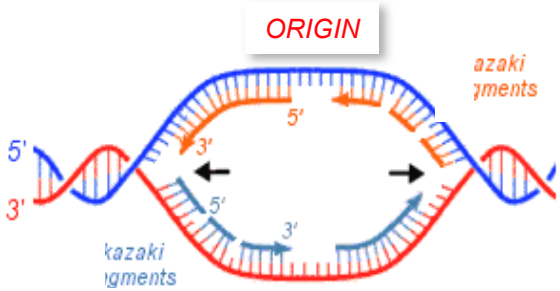
Replication fork polarity

Leading - lagging

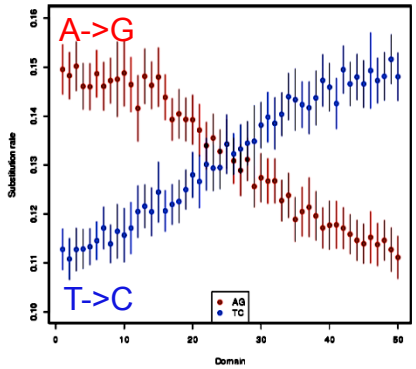




Replication bias
⇒ S ∝ fork polarity

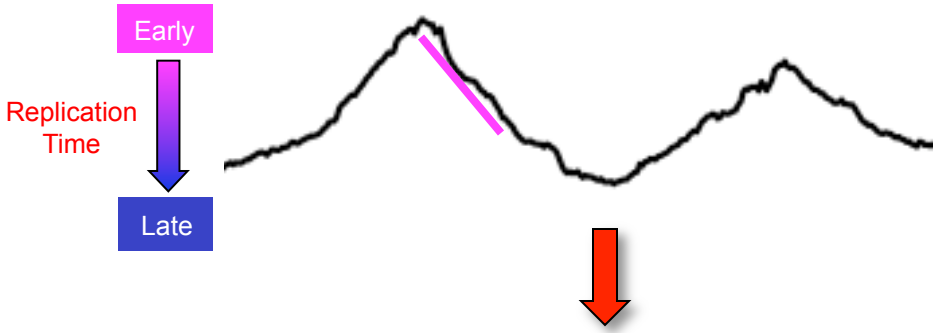


Different mutation rates
on leading and lagging strands



Derivative of replication timing profile = replication fork polarity

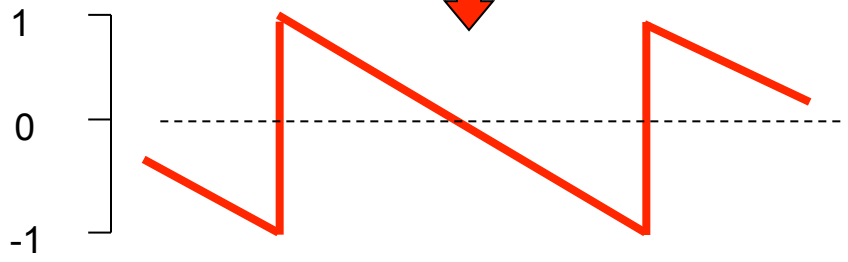
Replication timing profile
≈ parabolic shape



Derivative of
replication timing profile
≈ N shape

Replication fork polarity

Leading - lagging

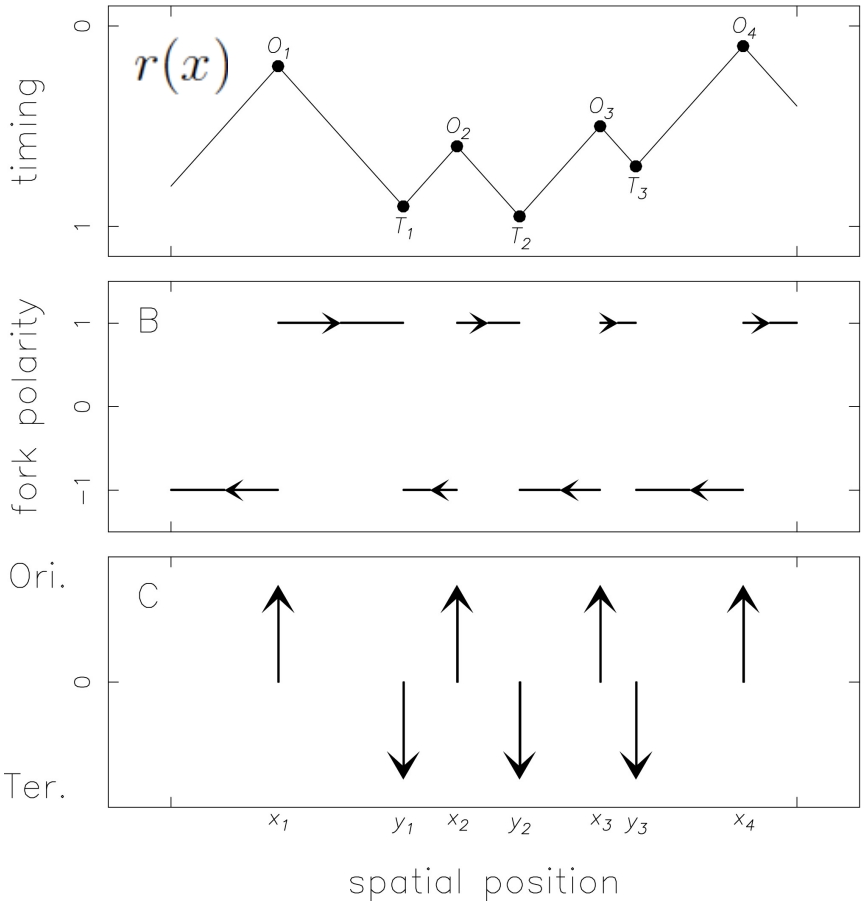


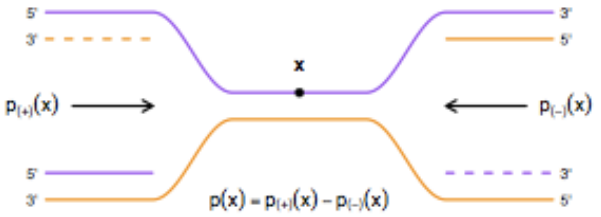
Parabolic shape of replication timing profile

→ gradient of replication fork polarity

Mathematical modelling of replication timing profile

For constant replication fork velocity v :





$$v\,\frac{d}{dx}r(x) = p(x),$$

$$v\,\frac{d^2}{dx^2}r(x) = \frac{d}{dx}p(x) = \sum_i \delta(x - x_i) - \sum_i \delta(y - y_i).$$

Averaging over many cell cycles, it results :

$$\frac{d}{dx}\langle r(x)\rangle_{\mathrm{Cells},\Delta x} = \frac{1}{v}\langle p(x)\rangle_{\mathrm{Cells},\Delta x}\ ,$$

$$\frac{d^2}{dx^2}\langle r(x)\rangle_{\mathrm{Cells},\Delta x} = \frac{1}{v}\left(N^{\mathrm{Ori}}_{\mathrm{Cells},\Delta x}(x) - N^{\mathrm{Ter}}_{\mathrm{Cells},\Delta x}(x)\right)$$

Apparent replication speed :

$$v_{app,\Delta x} = \frac{1}{\frac{d}{dx}\langle r(x)\rangle_{\mathrm{Cells},\Delta x}}$$

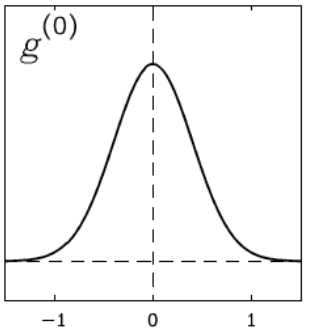
Guilbaud, PLoS Comput Biol (2011)
Baker, PLoS Comput Biol (2012)

Defining scale-derivatives using the wavelet transform

Moving average

$$M_\phi(b,a) = \int s(t)\ \phi(\tfrac{t-b}{a})\mathrm{d}t$$

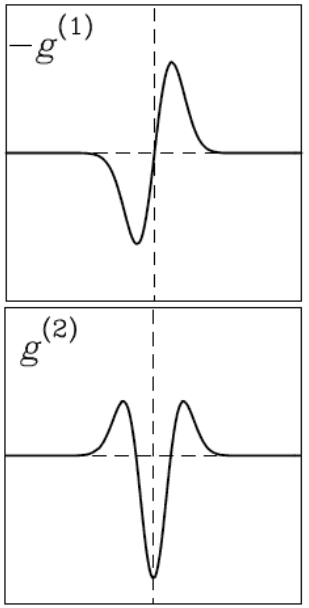
$\Rightarrow$ Smoothing signal $s(t)$ at scale a



Wavelet transform

$$\frac{\partial}{\partial b} M_\phi(b,a) = T_{\psi=-\phi'}(b,a)$$

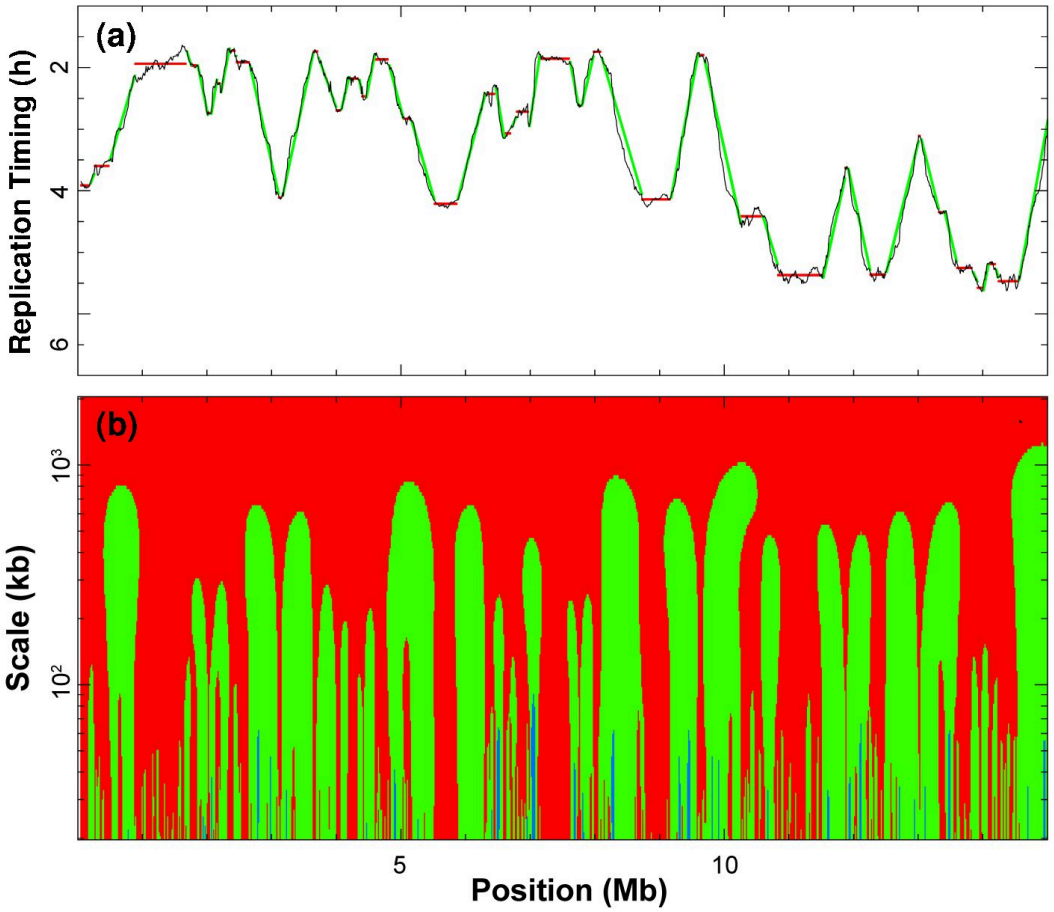
$$\frac{\partial^2}{\partial b^2} M_\phi(b,a) = T_{\psi=\phi''}(b,a)$$



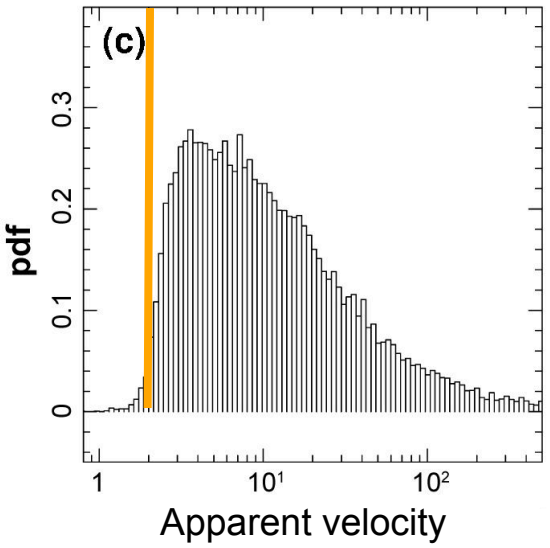
$T_{\phi^{(n)}}(b,a)$ defines the n^{th} derivative at scale a of $s(t)$ at $t=b$

Exploring the space-scale map of apparent speed of replication

Guilbaud, PLoS Comput Biol (2011)
Audit, Nat Protoc (in press)



Vmax = 2 kb/min

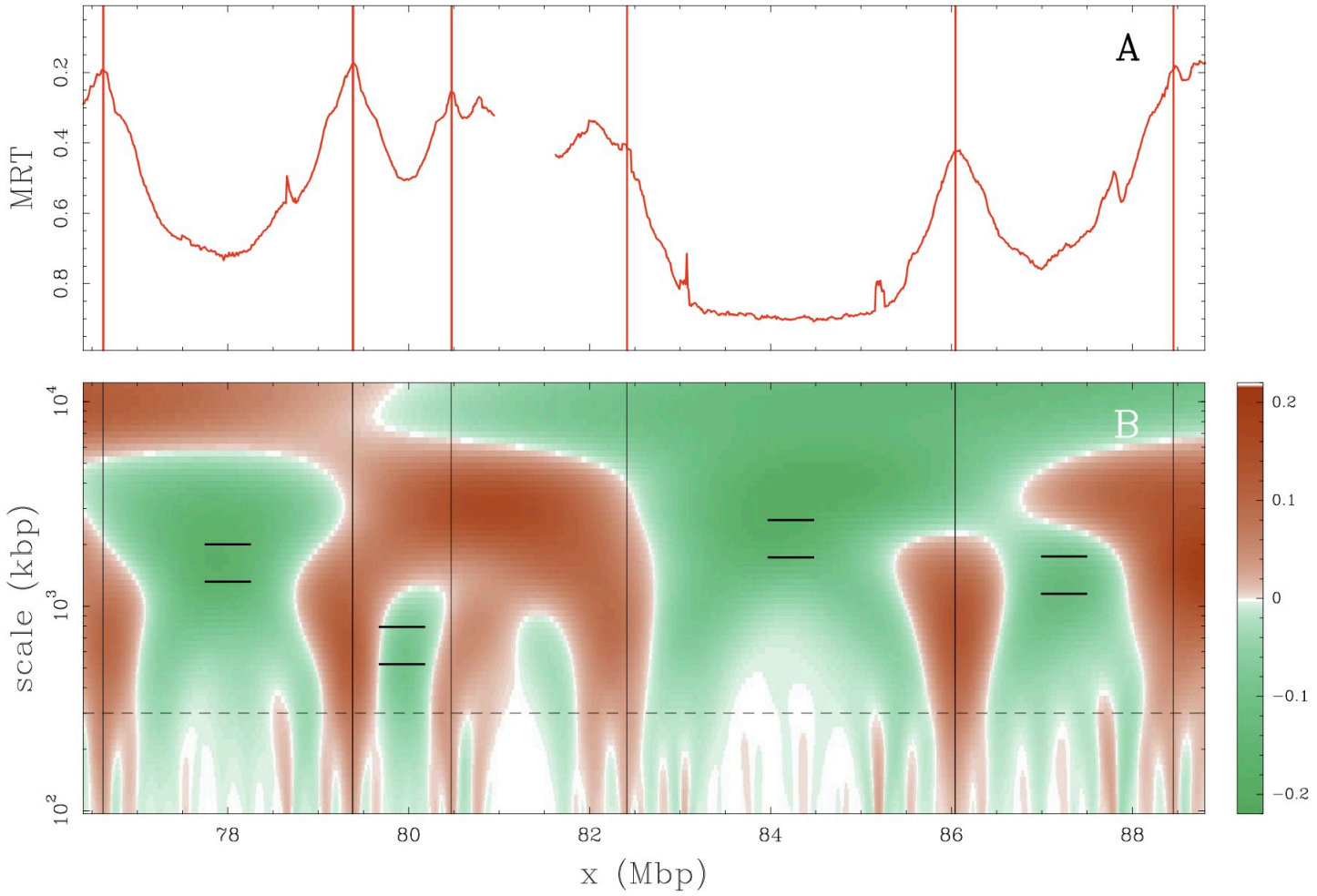


V > 10 kb/min
10 kb/min > V > 2 kb/min
V < 2 kb/min

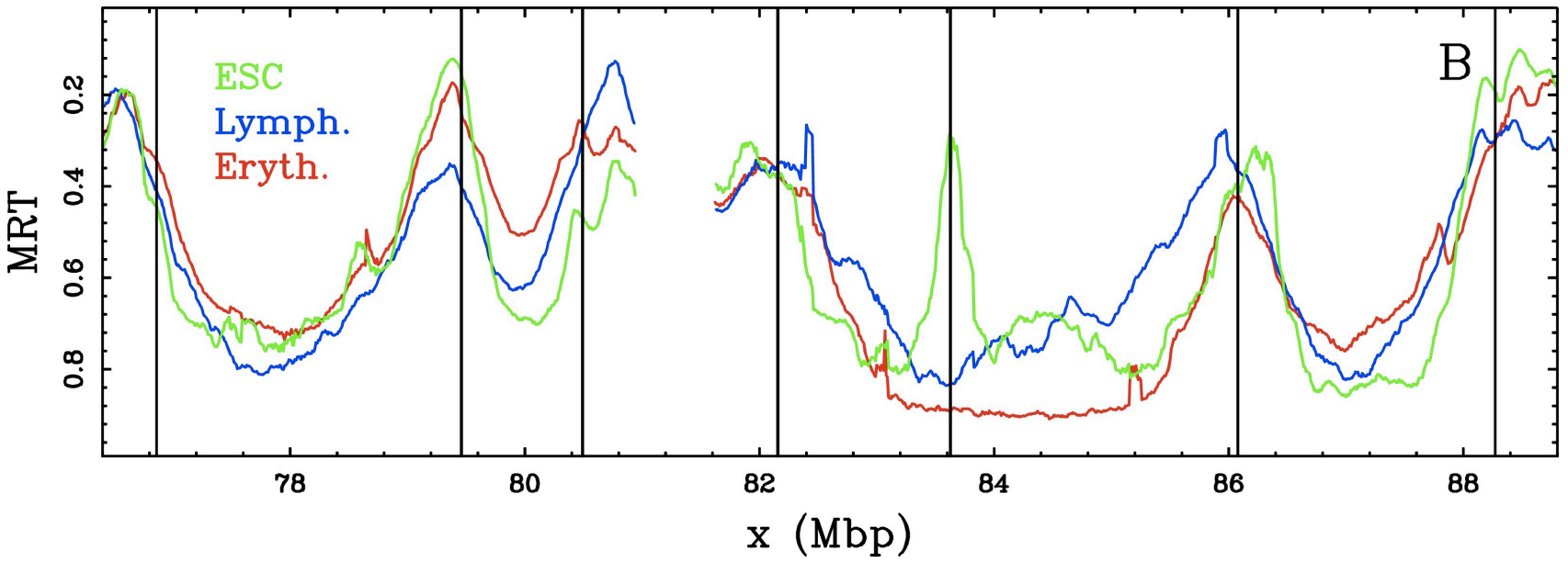
Practically no regions have an apparent speed of replication compatible with unidirectional progression of a single fork

Multiscale detection of replication U-domains genome-wide

Baker, PLoS Comput Biol (2012)



U-shaped replication timing domains along the human genome

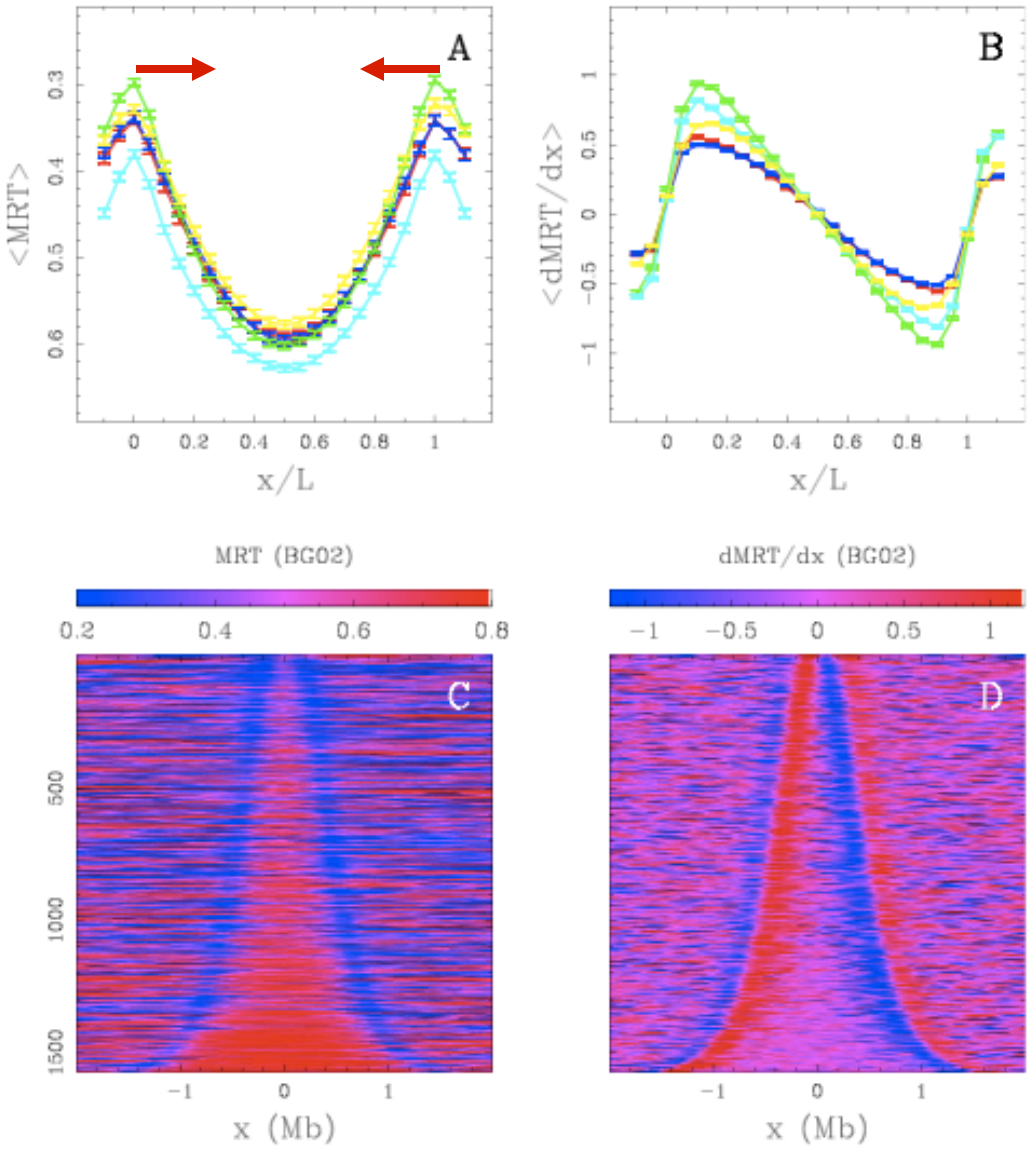


Replication U-domains are robustly found covering about half of the genome in seven cell lines. About half U-domains are common to several cell lines.

	Ndom	ESC	Erythroid	Lymphoblastoid			Fibroblast		HeLa	
N	663	1534	876	882	830	664	1150	1247	1422	1498
L	1.19	1.09	1.42	1.52	1.57	1.62	1.19	1.15	1.06	0.966
G	29.2	61.9	46.1	49.5	48.1	39.6	50.5	53.2	55.7	53.5
GC	40.30	40.25	40.84	40.85	40.94	41.13	40.84	40.60	40.72	40.99

U-domains correspond to large-scale gradients of the replication fork polarity

Replication timing | **Fork polarity**

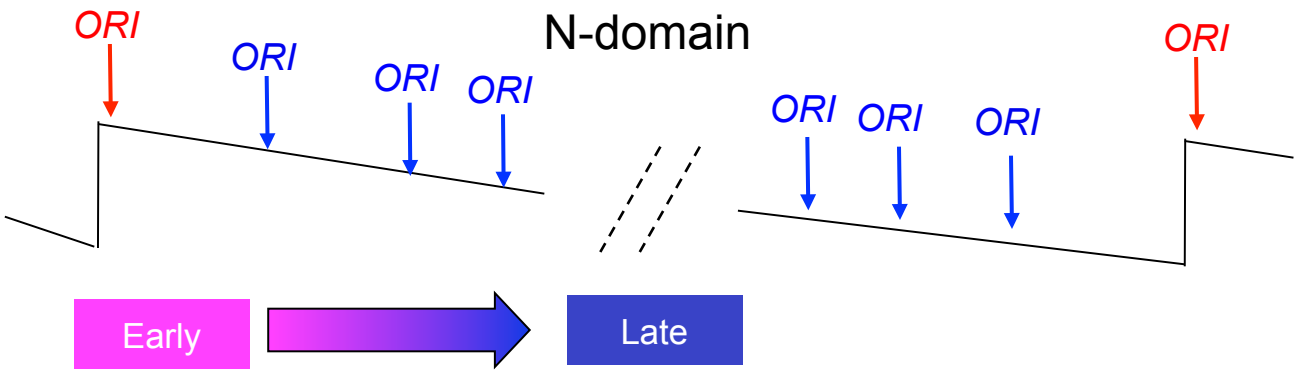


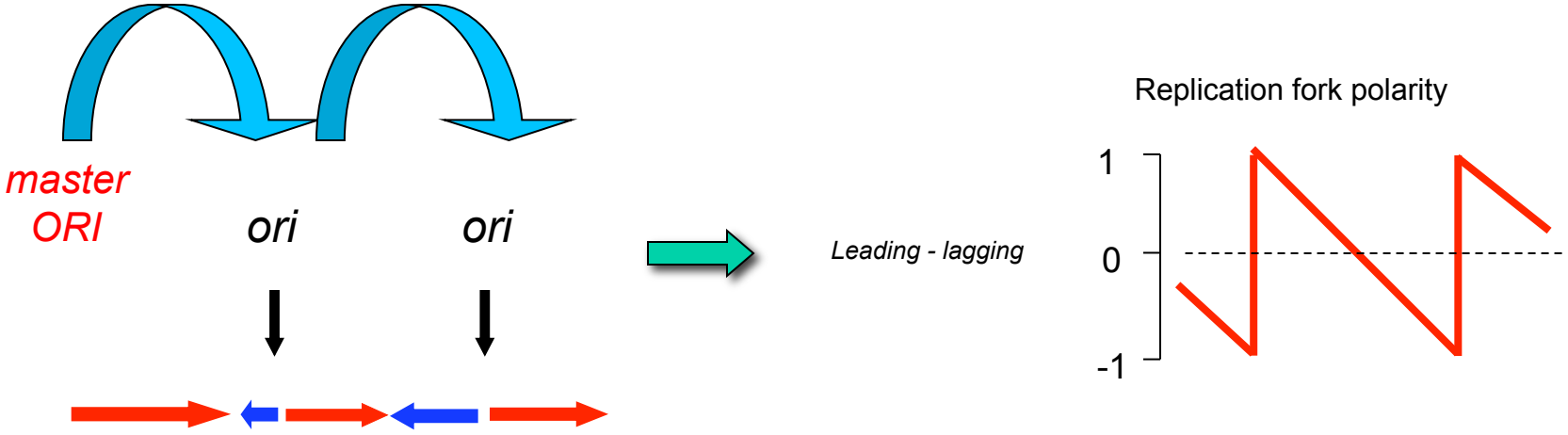
Baker, PLoS Comput Biol (2012).

The derivative of replication timing profiles displays a N-shape in U-domains sustaining the existence of large-scale gradients of replication fork polarity.

A model for the spatio-temporal replication program in mammalian cells

"Master Origins" specified by a particular chromatin environment coded in the sequence

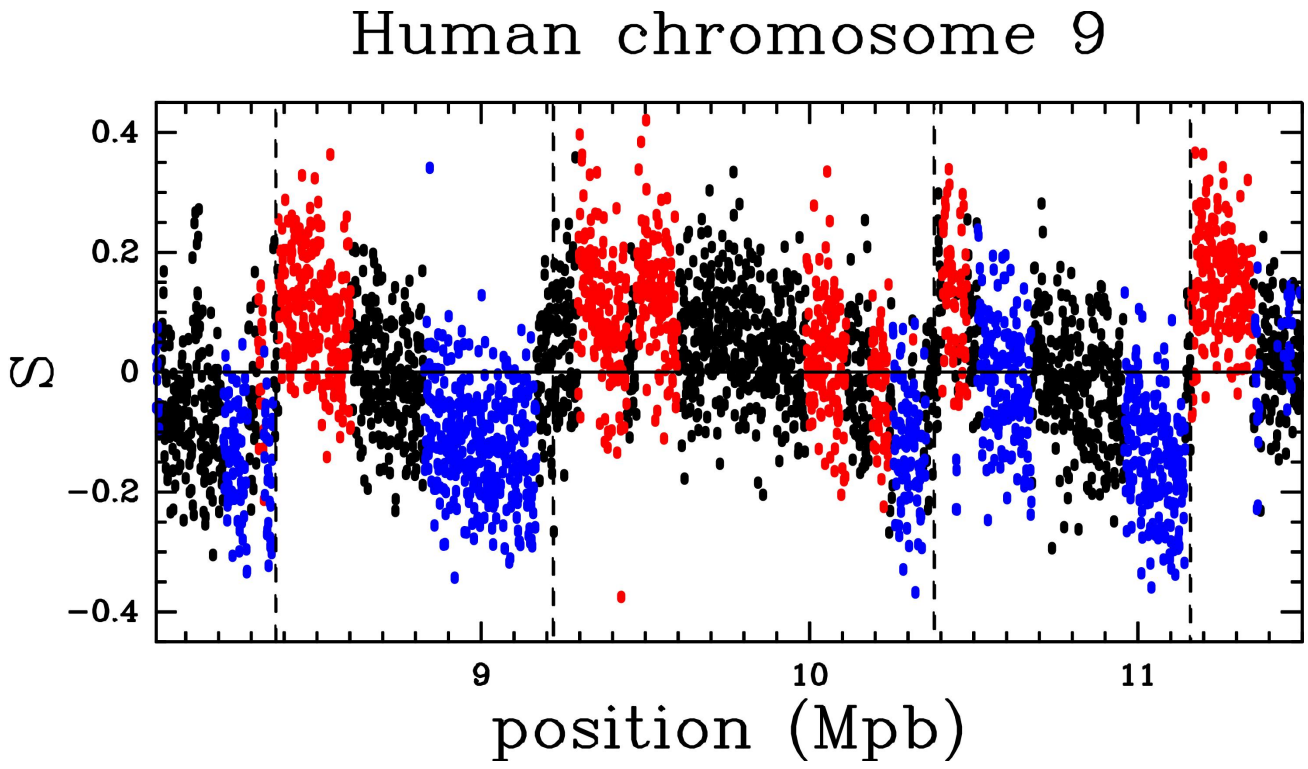




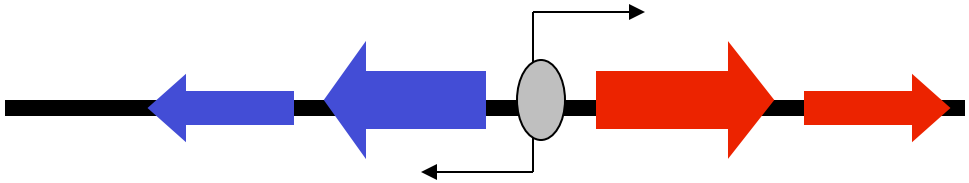
Activation of replication origins propagate along N-domains

Genome organisation inside N-domains

Huvet, Genome Res. (2007)

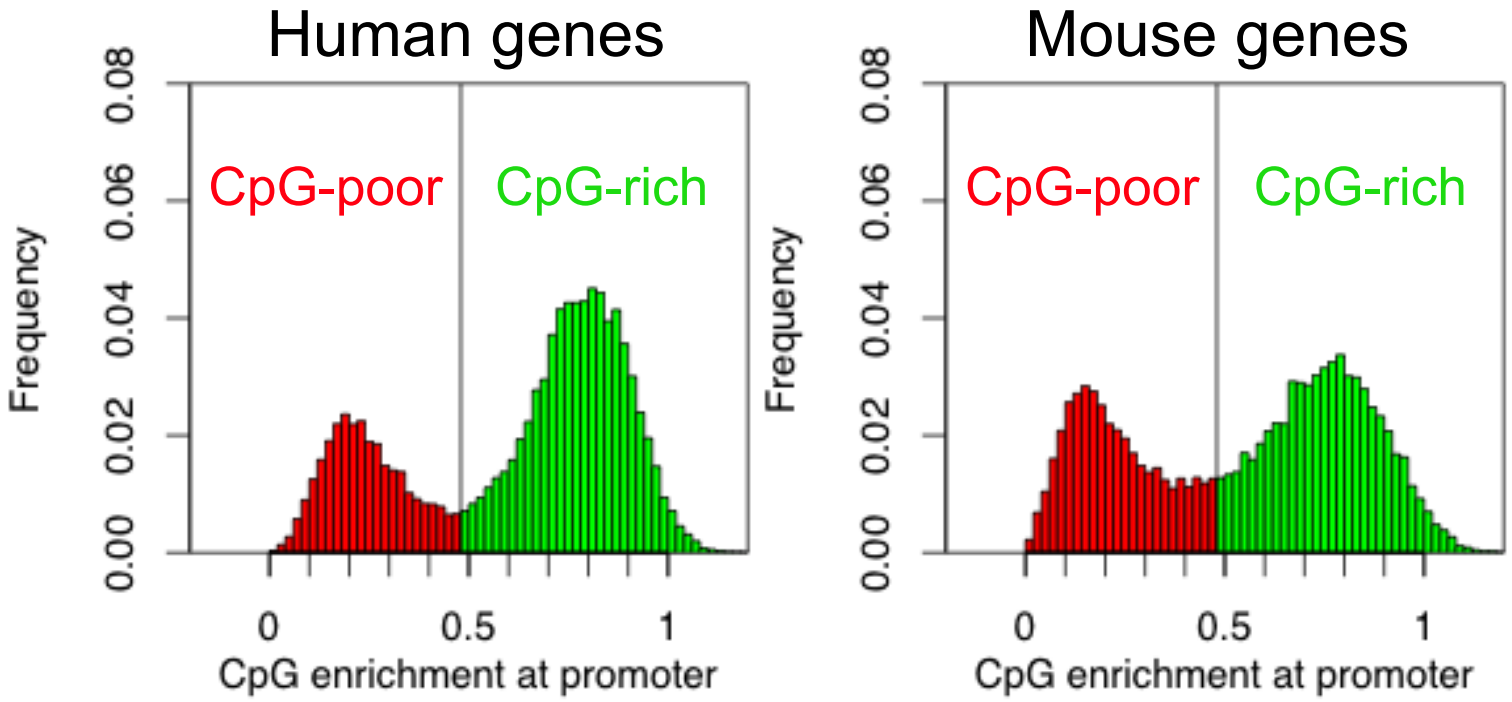


sense gene
anti-sense gene
intergene



CpG enrichment at human gene promoters

CpG-rich vs CpG-poor

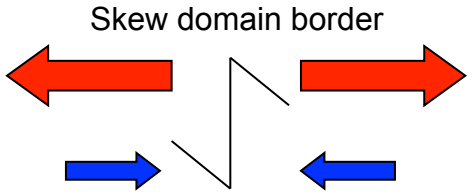


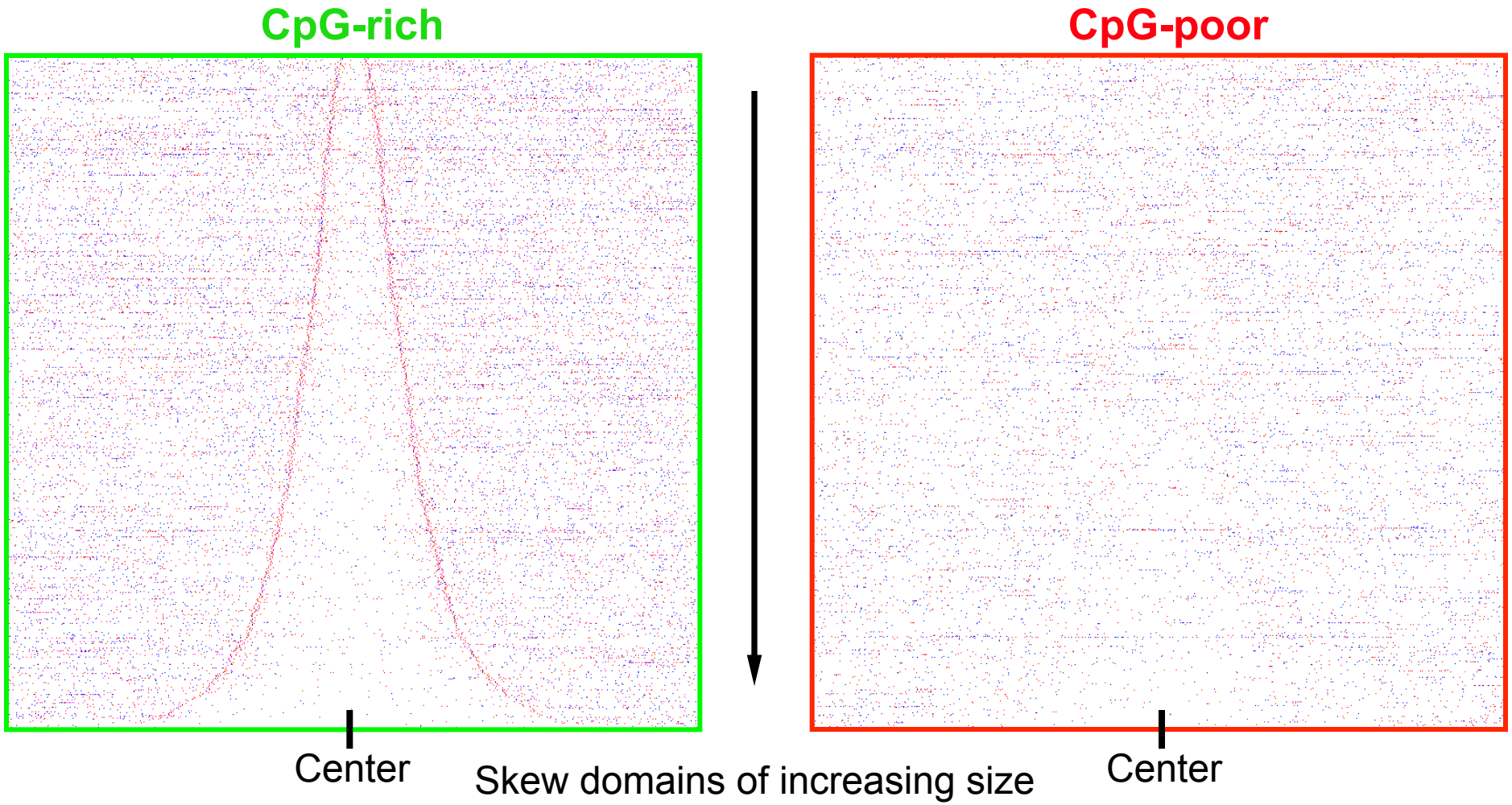
Antequera, PNAS, 1993
Antequera, Cell Mol Life Sci, 2003

CpG enrichment = [CpG] / [C].[G]

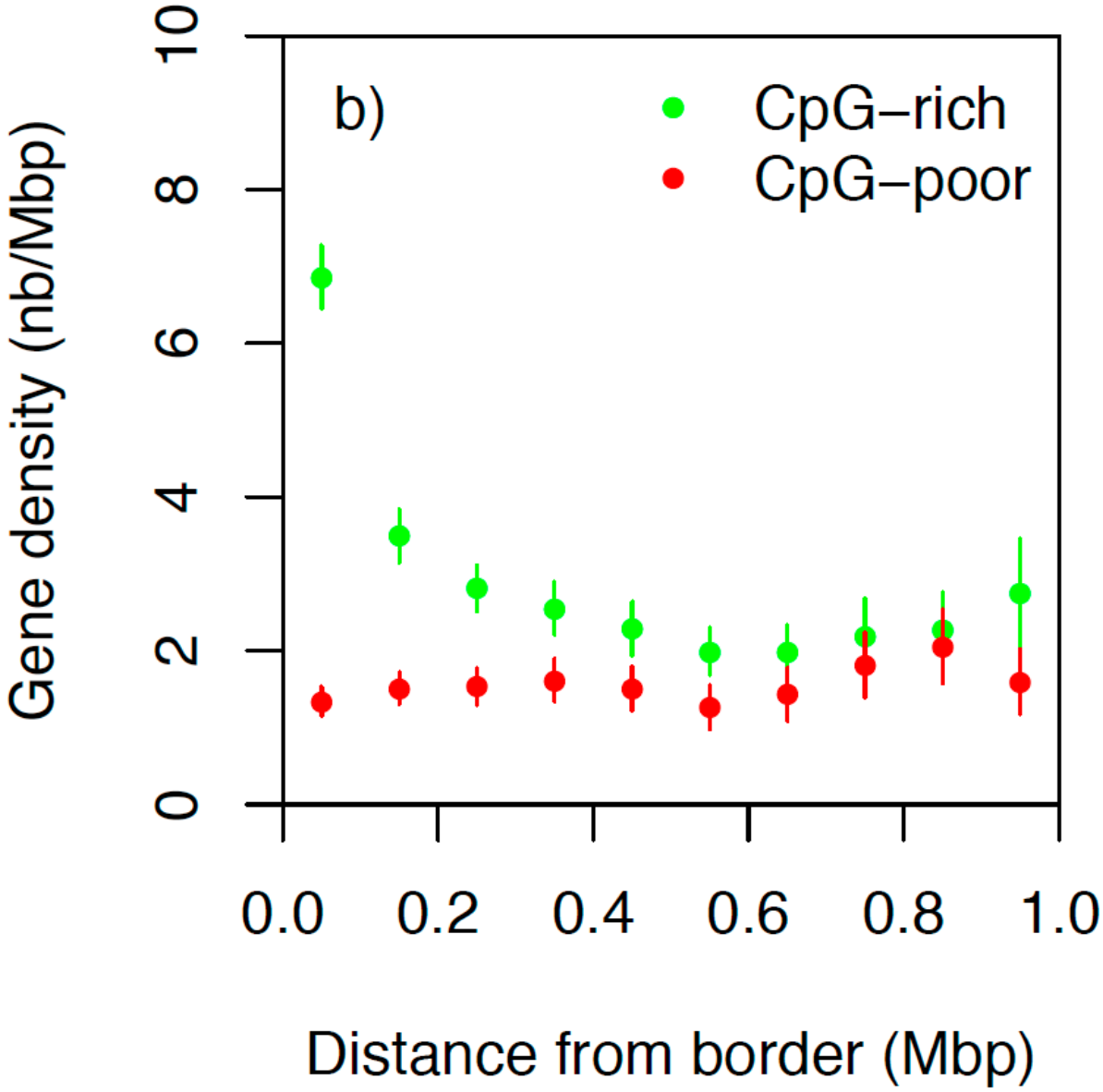
Genome average CpG enrichment ~ 0.2

CpG-rich genes are over-represented around skew domains borders

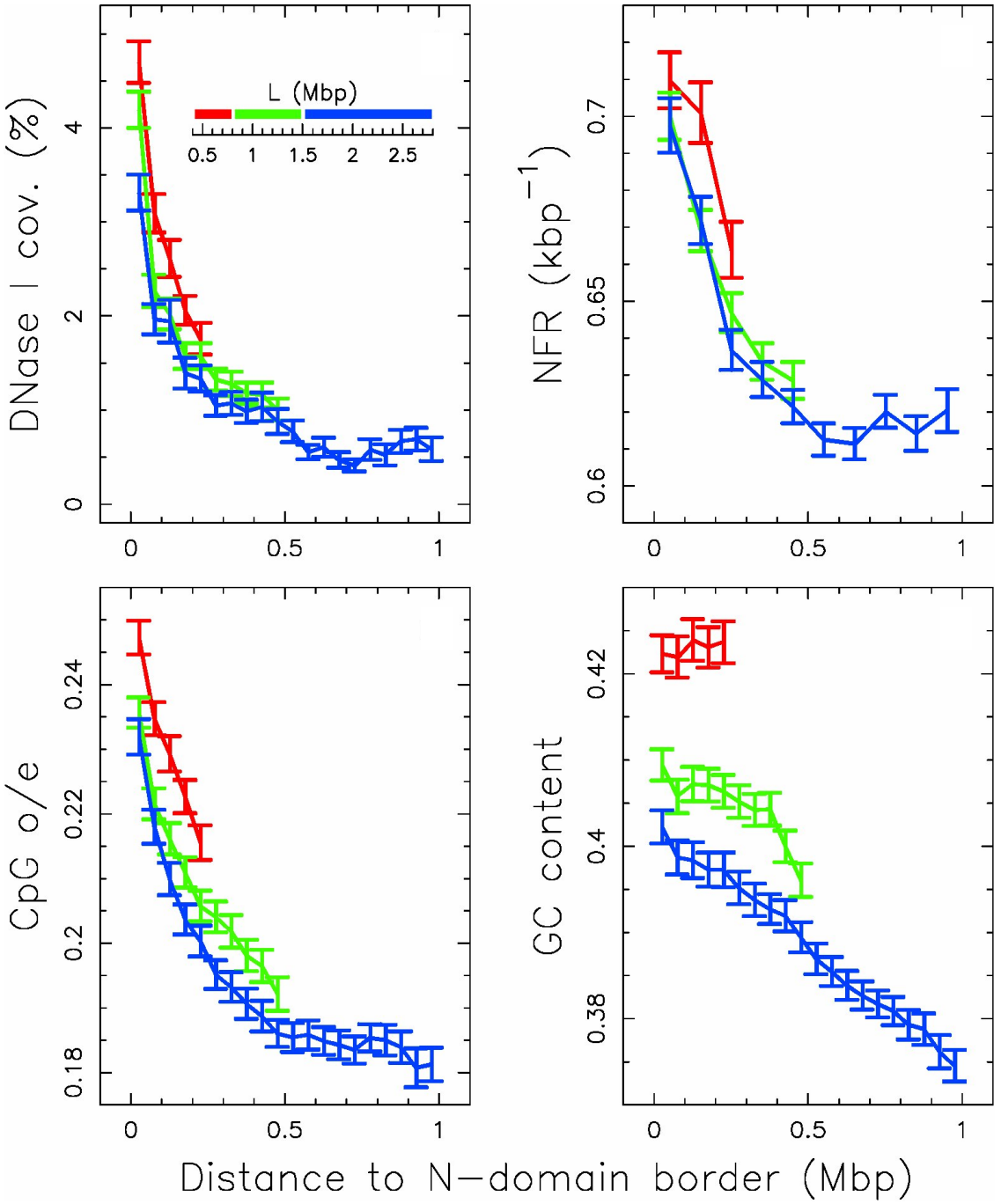




CpG-rich genes are over-represented around skew domains borders

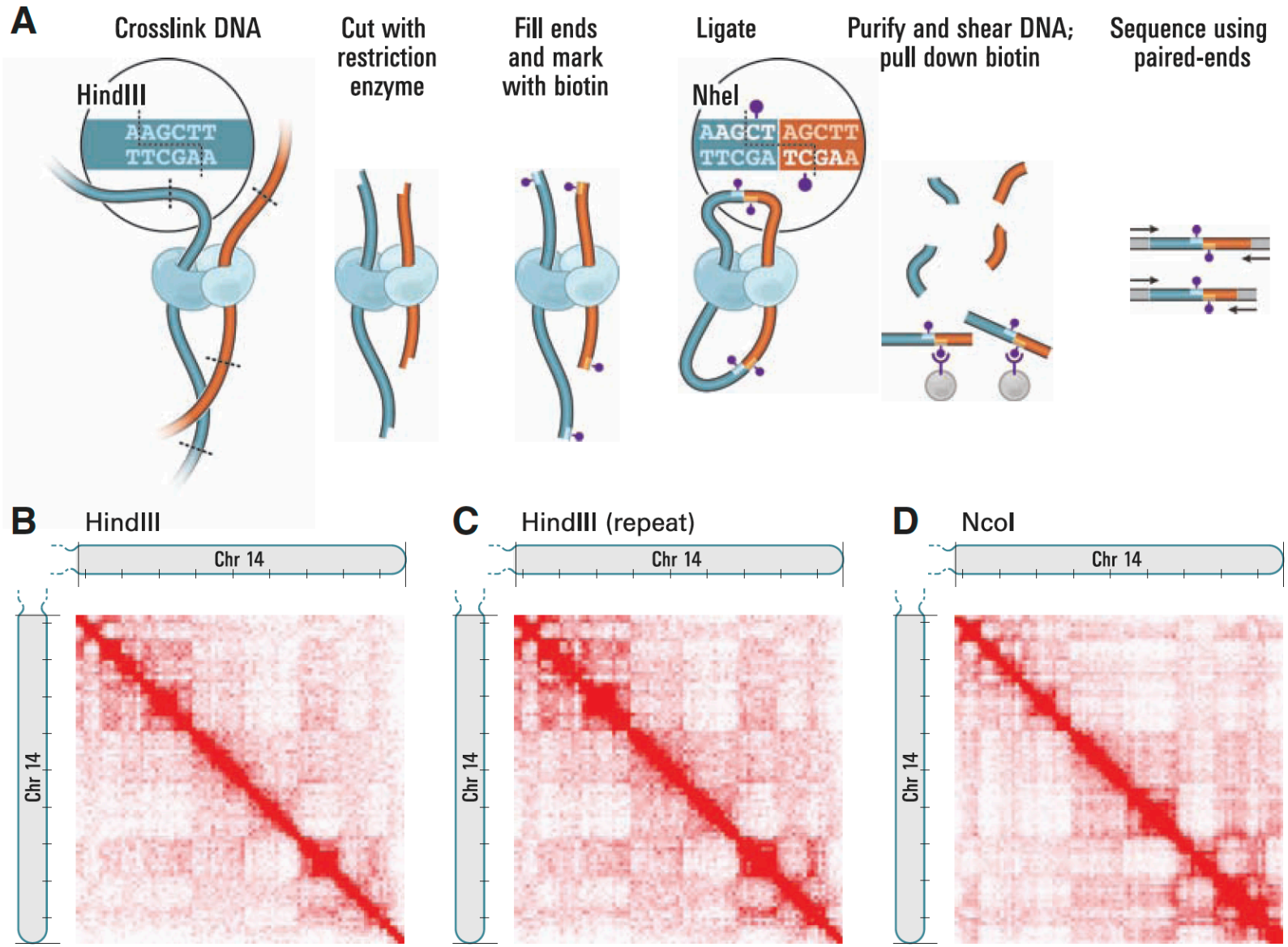


Open chromatin regions around N-domain borders have a characteristic size ~300kbp



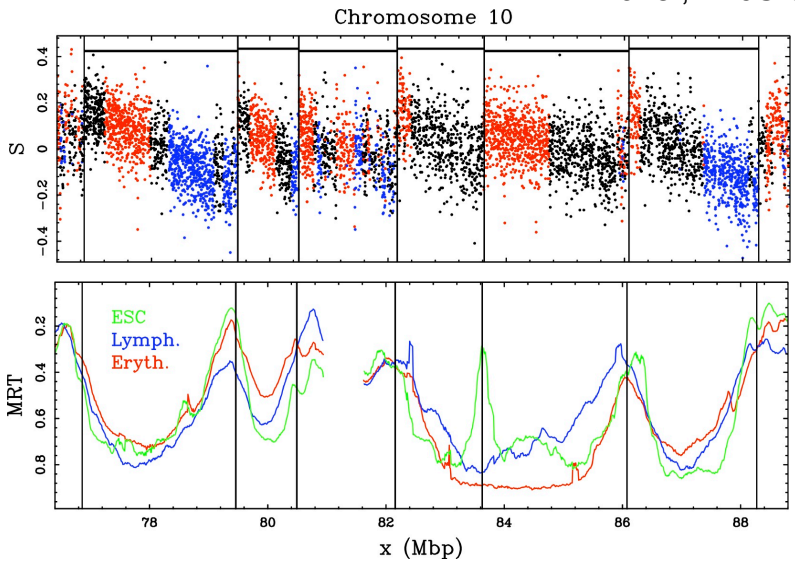
Chromatin conformation capture for the human genome

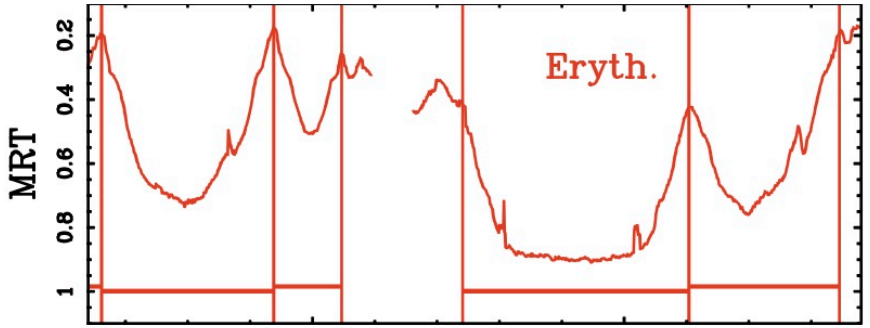
Hi-C data from Lieberman-Aiden, Science 326 (2009)

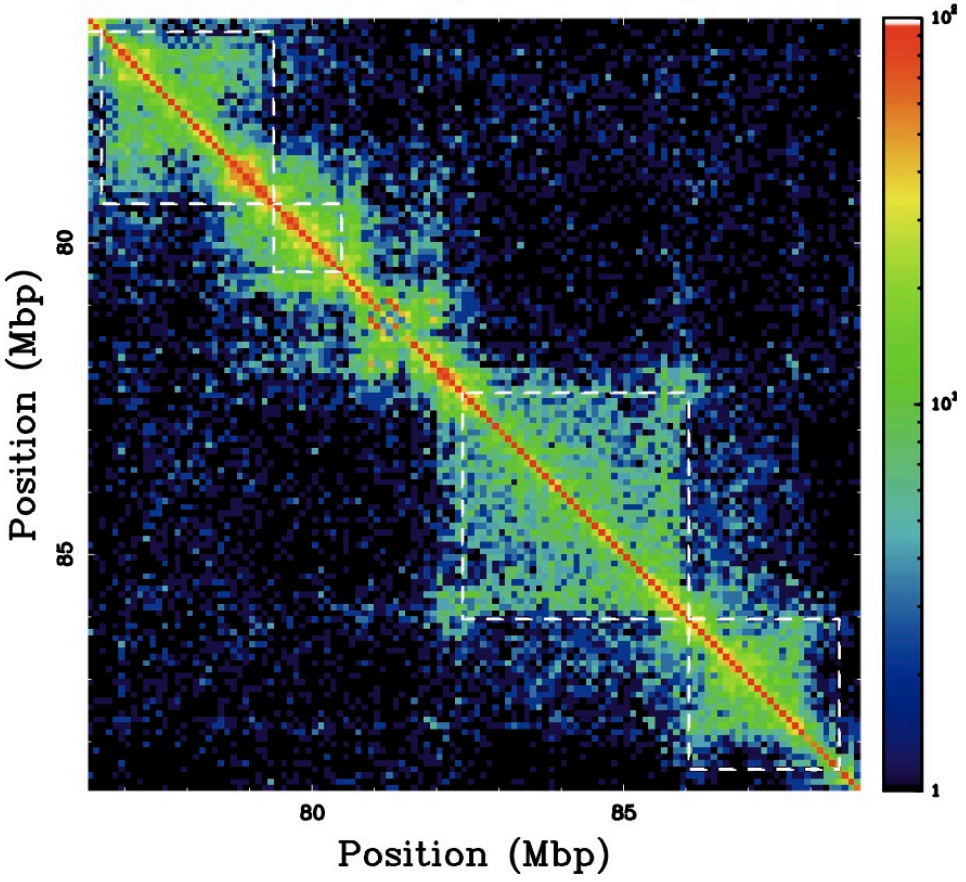


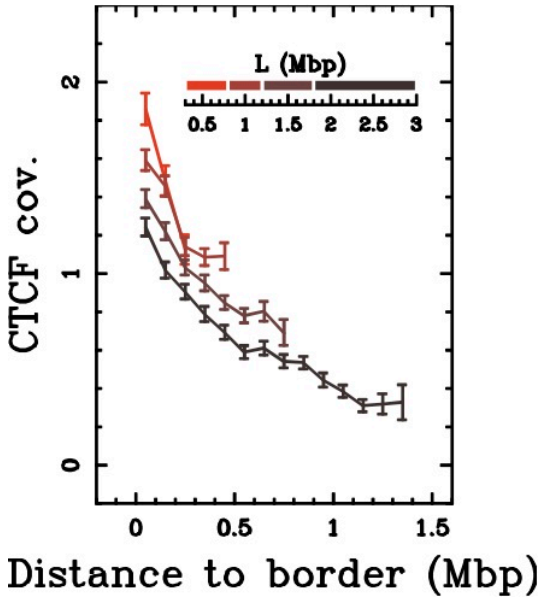
Large-scale chromatin conformation along replication domains

Baker, PLoS Comput Biol (2012). Hi-C data from Lieberman-Aiden, Science (2009)





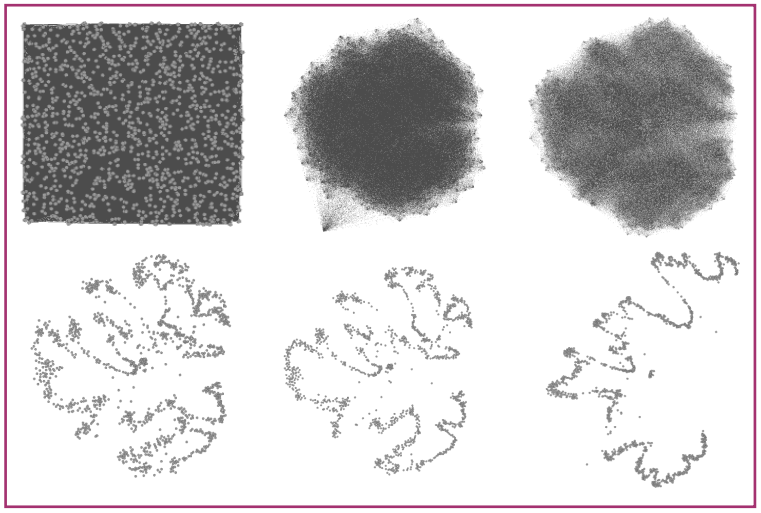




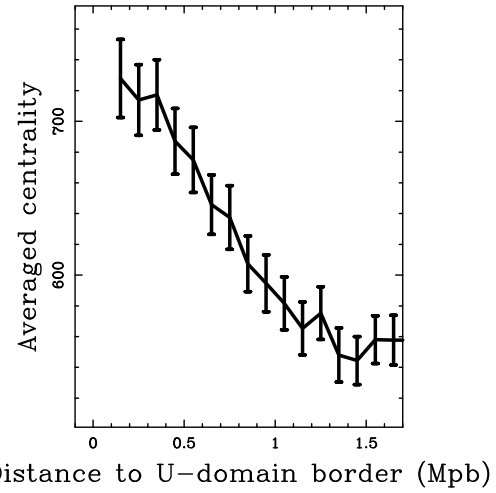
Replication domains are self-interacting structural units

Replication domain borders are long-range interconnected hubs in human chromatin interaction graph

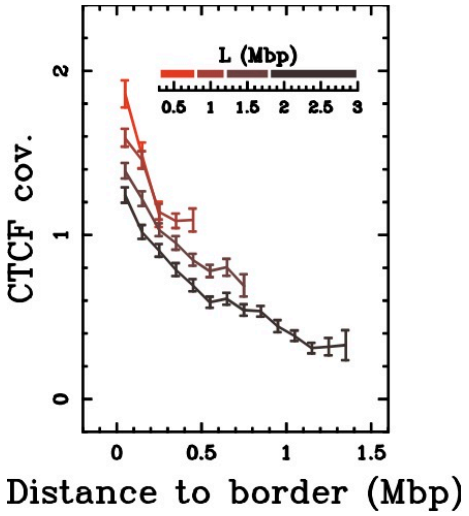
Developing a graph theoretical approach to chromatin conformation data

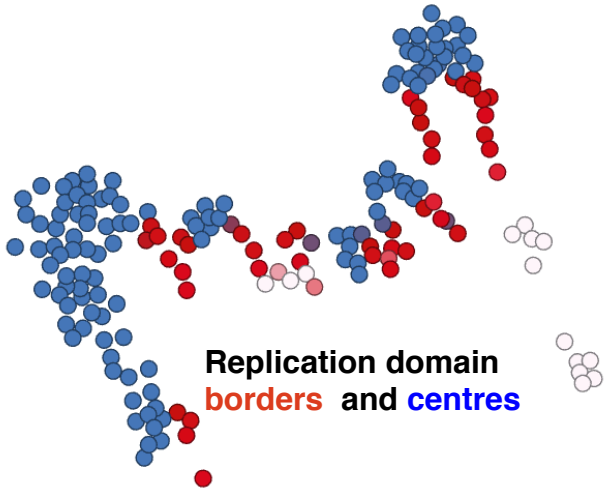


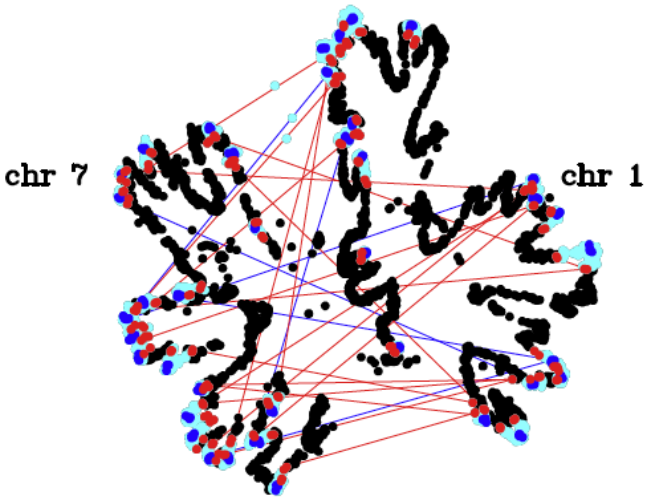
Betweenness centrality



Insulator DNA binding protein

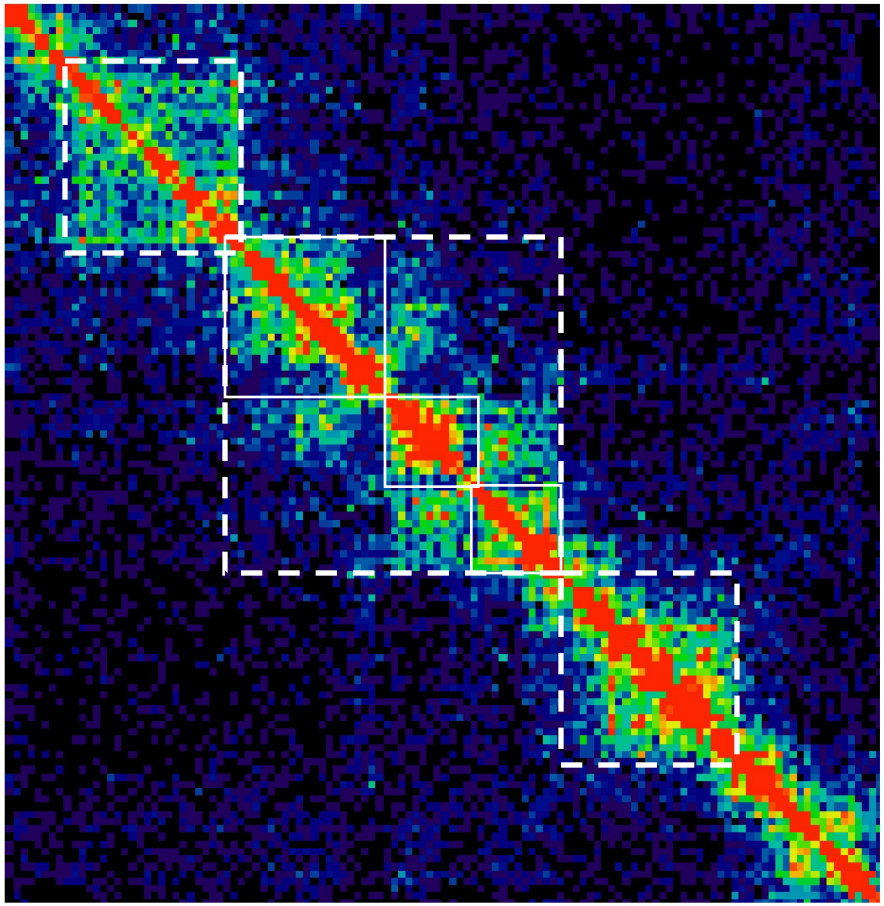


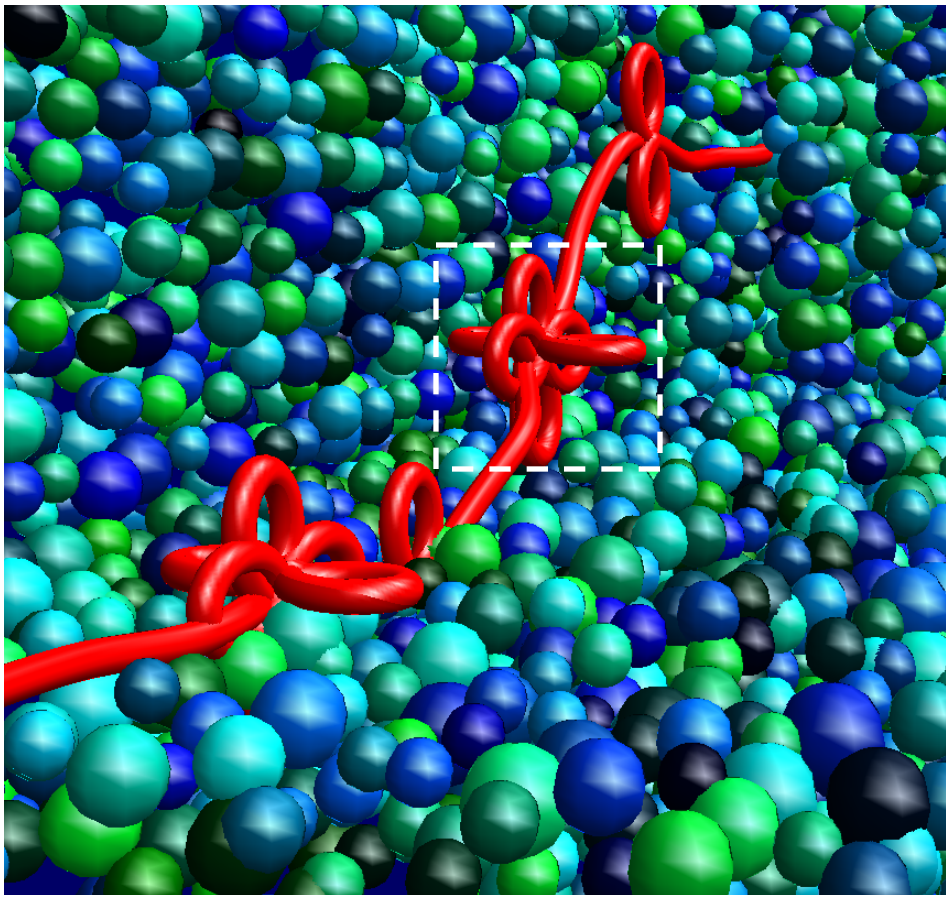




Boulos, preprint (2013)

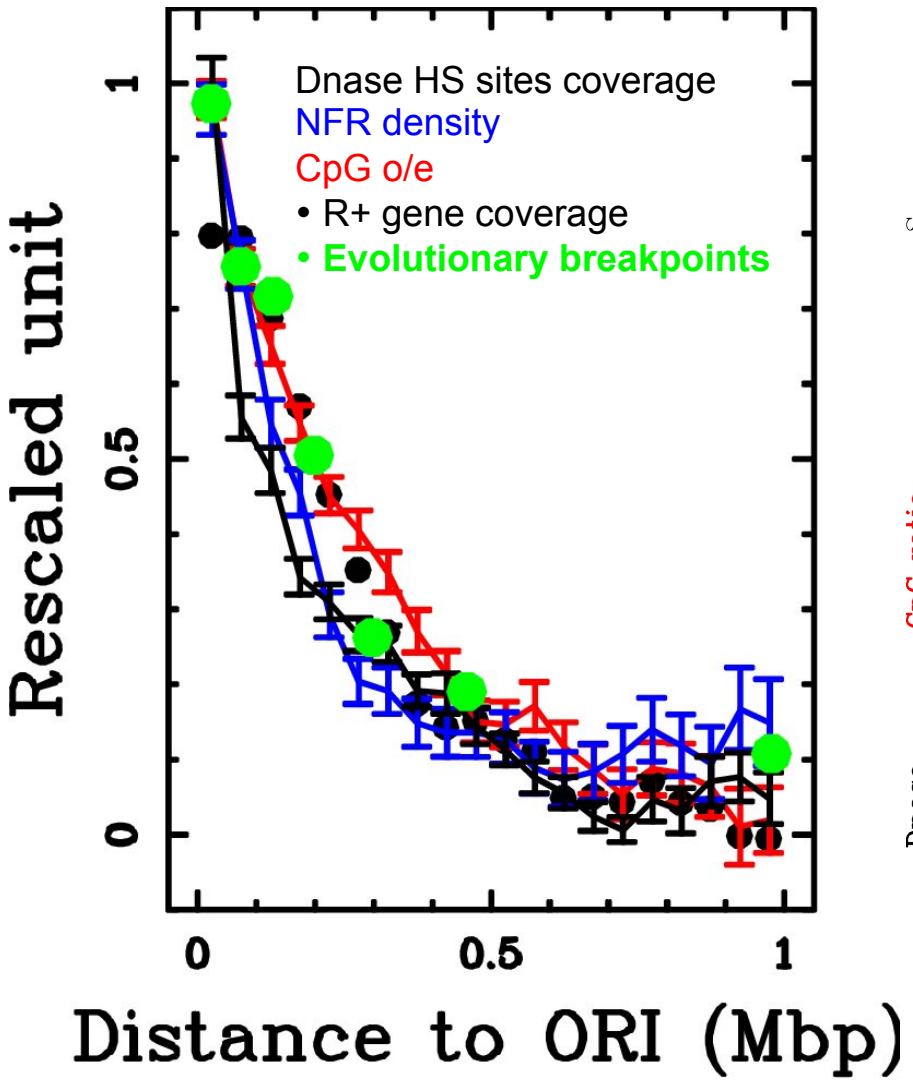
Master replication origins at the heart of parallel genome functioning ?

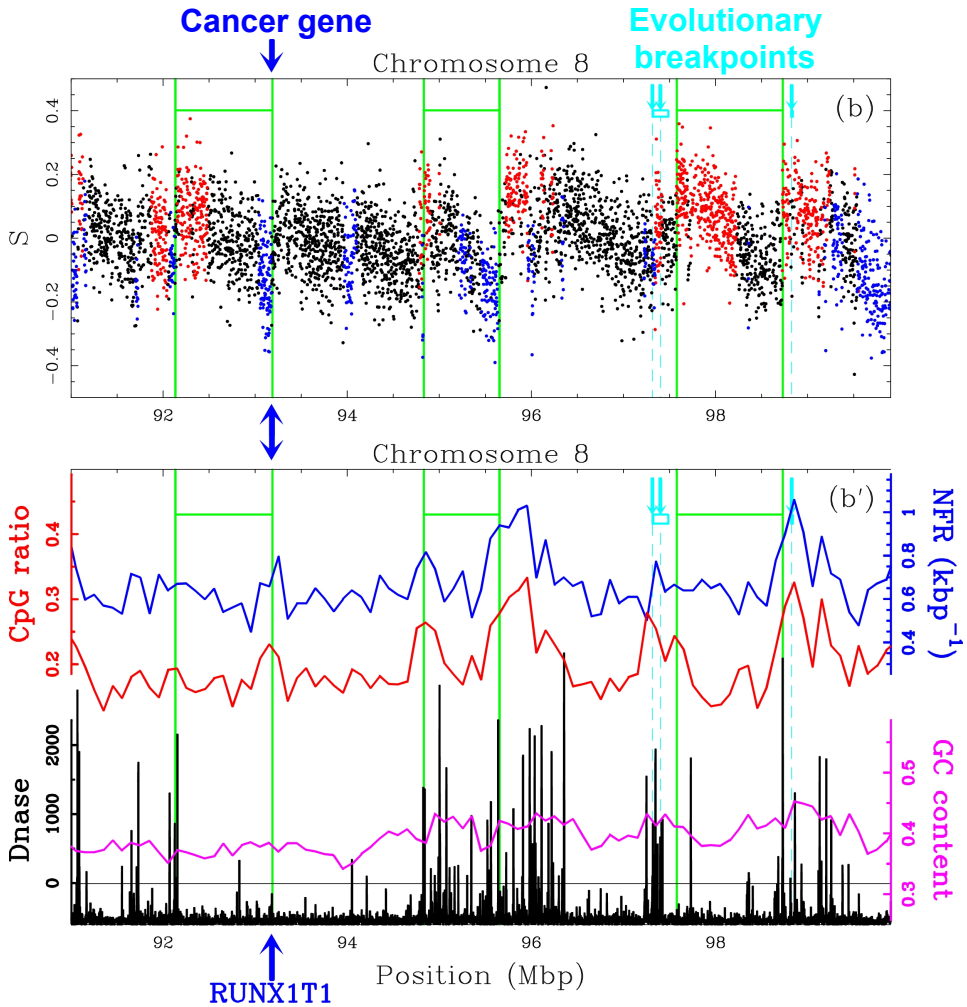




Fragility of the human genome at N-domain borders

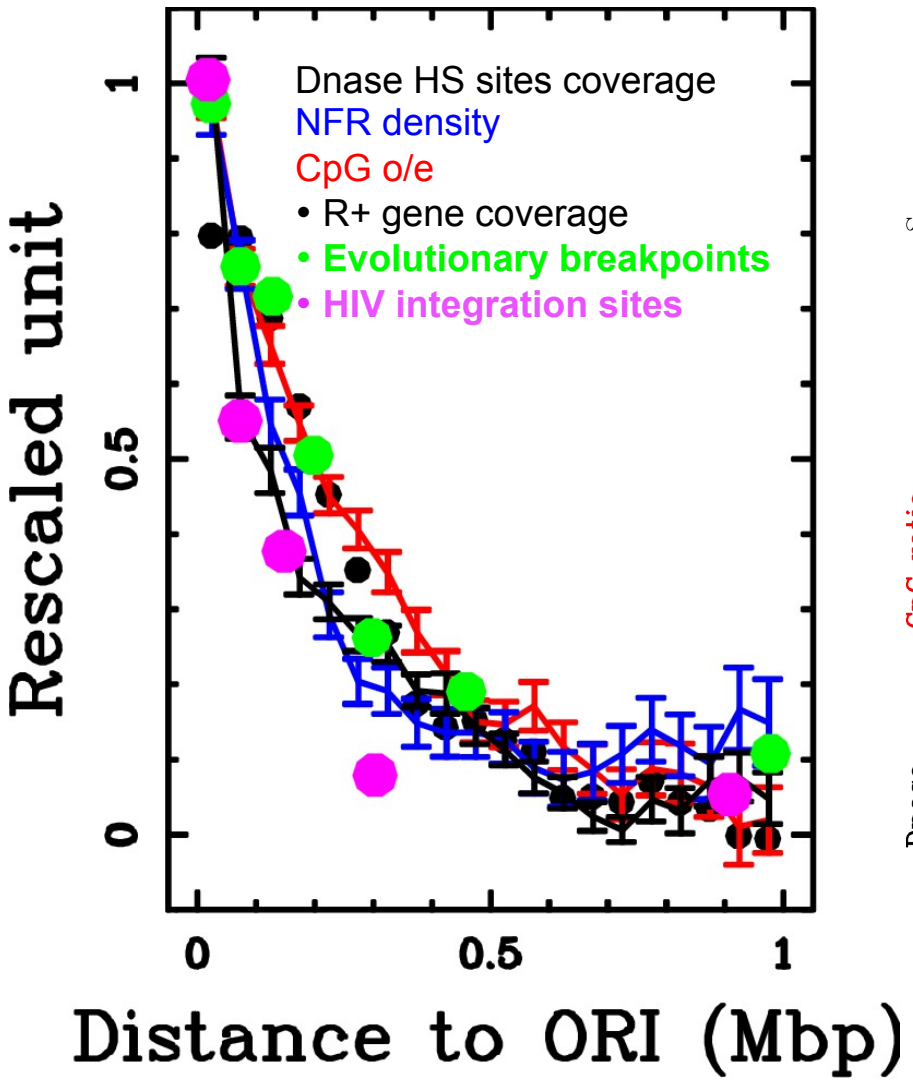
Lemaitre, BMC Genomics (2009)

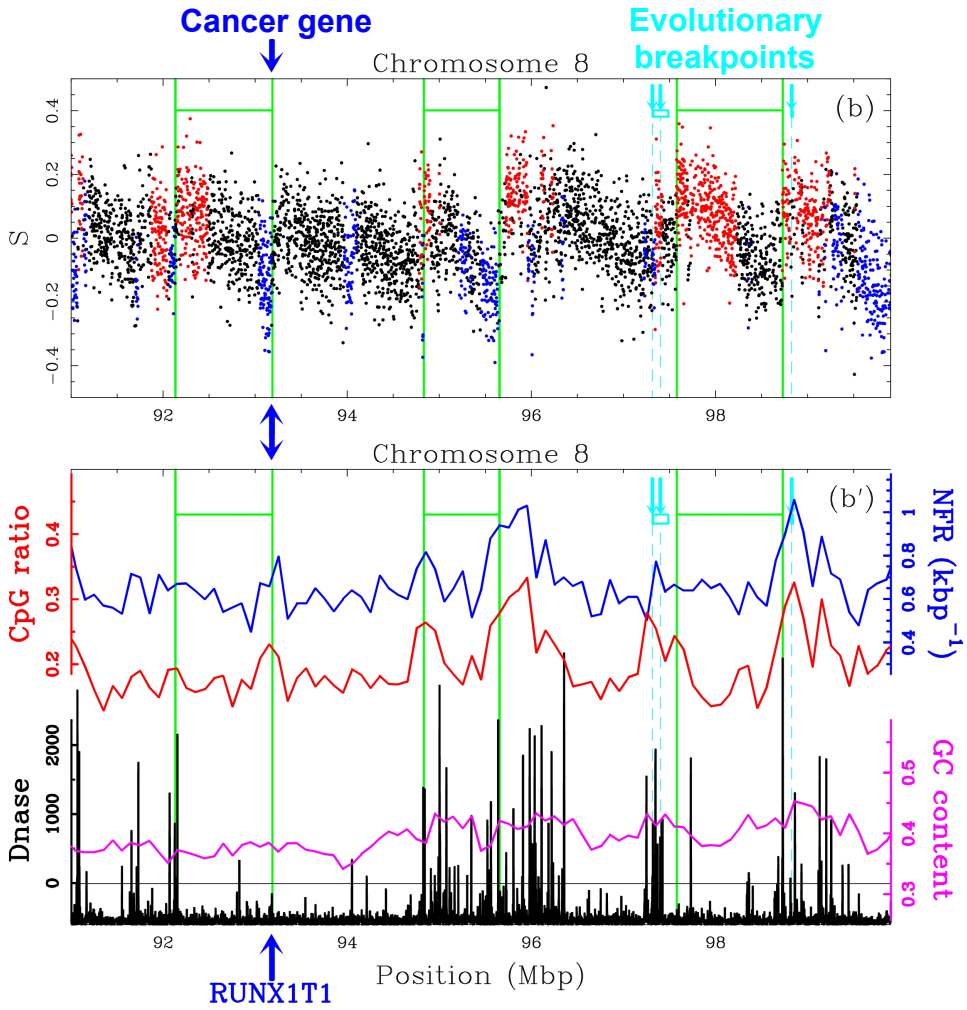




Fragility of the human genome at N-domain borders

Lemaitre, BMC Genomics (2009)





Physics Reports 498 (2011) 45–188

Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

Multi-scale coding of genomic information: From DNA sequence to genome structure and function

Alain Arneodo [a,b,*], Cédric Vaillant [a,b], Benjamin Audit [a,b], Françoise Argoul [a,b], Yves d'Aubenton-Carafa [c], Claude Thermes [c]

[a] Université de Lyon, F-69000 Lyon, France

[b] Laboratoire Joliot-Curie and Laboratoire de Physique, CNRS, Ecole Normale Supérieure de Lyon, F-69007 Lyon, France

[c] Centre de Génétique Moléculaire, CNRS, Allée de la Terrasse, 91198 Gif-sur-Yvette, France

ARTICLE INFO

Article history:
Accepted 24 September 2010
Available online 8 October 2010
editor: H. Orland

Keywords:
DNA sequence
Chromatin
Nucleosome
Genome organization
Epigenetics
Transcription
Replication
Compositional strand asymmetry
Statistical physics
Hetero-polymer
Generalized worm-like-chain model
Scale-invariance
Multi-fractal
Multi-scale analysis
Wavelet transform
Long-range correlations
Atomic force microscopy

ABSTRACT

Understanding how chromatin is spatially and dynamically organized in the nucleus of eukaryotic cells and how this affects genome functions is one of the main challenges of cell biology. Since the different orders of packaging in the hierarchical organization of DNA condition the accessibility of DNA sequence elements to trans-acting factors that control the transcription and replication processes, there is actually a wealth of structural and dynamical information to learn in the primary DNA sequence. In this review, we show that when using concepts, methodologies, numerical and experimental techniques coming from statistical mechanics and nonlinear physics combined with wavelet-based multi-scale signal processing, we are able to decipher the multi-scale sequence encoding of chromatin condensation–decondensation mechanisms that play a fundamental role in regulating many molecular processes involved in nuclear functions.