Origins of Replication and Cancer

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

Laboratoire de Physique de l'ENS-Lyon

ENS de Lyon

Benjamin Audit
Antoine Baker
Rasha Boulos
Guenola Drillon
Hanna Julienne
Lamia Zaghloul

ENS de Lyon / CRCL

Françoise Argoul
Bastien Laperrousaz
Véronique Maguer-Satta
Franck Nicolini

ENS, Paris

Guillaume Guilbaud
Olivier Hyrien
Malik Kahli
Nataliya Petryk
Aurélien Rappailles

CGM, Gif-sur-Yvette

Yves d'Aubenton-Carafa
Chun-Long Chen
Claude Thermes

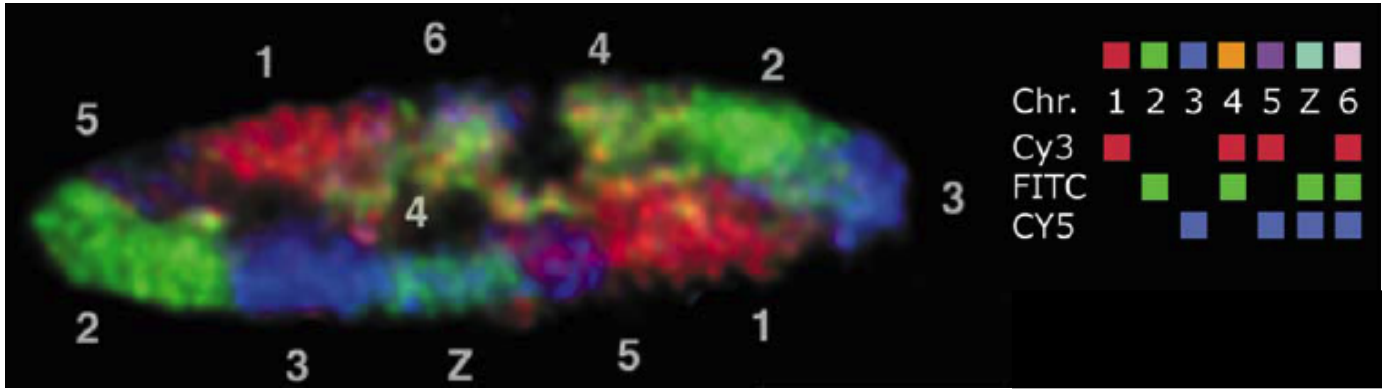
CEA, Saclay

Arach Goldar

Organisation of eukaryotic nucleus

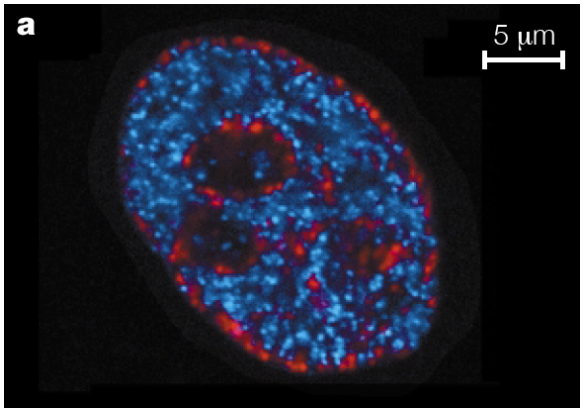
Cremer, Nat Rev Genet 2, 2001 – Cook, Science 284, 1999 - Pombo, EMBO J 18, 1999

Chromosome territories



Chicken fibroblast nucleus where the 7 macro-chromosome are revealed by FISH

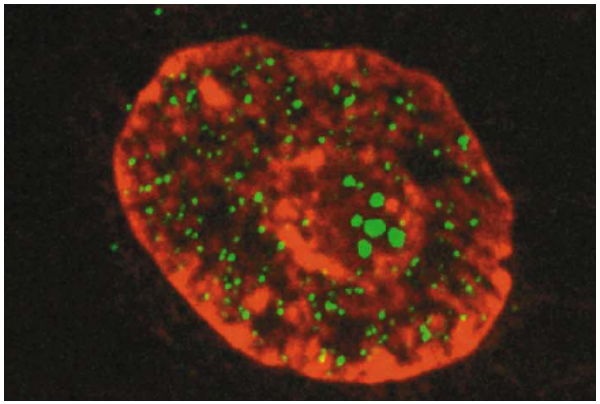
Regionalisation of nuclear functions



Human neuroblastoma

Early replication *(Cy3)*

Mid- to late- replication *(Cy5)*

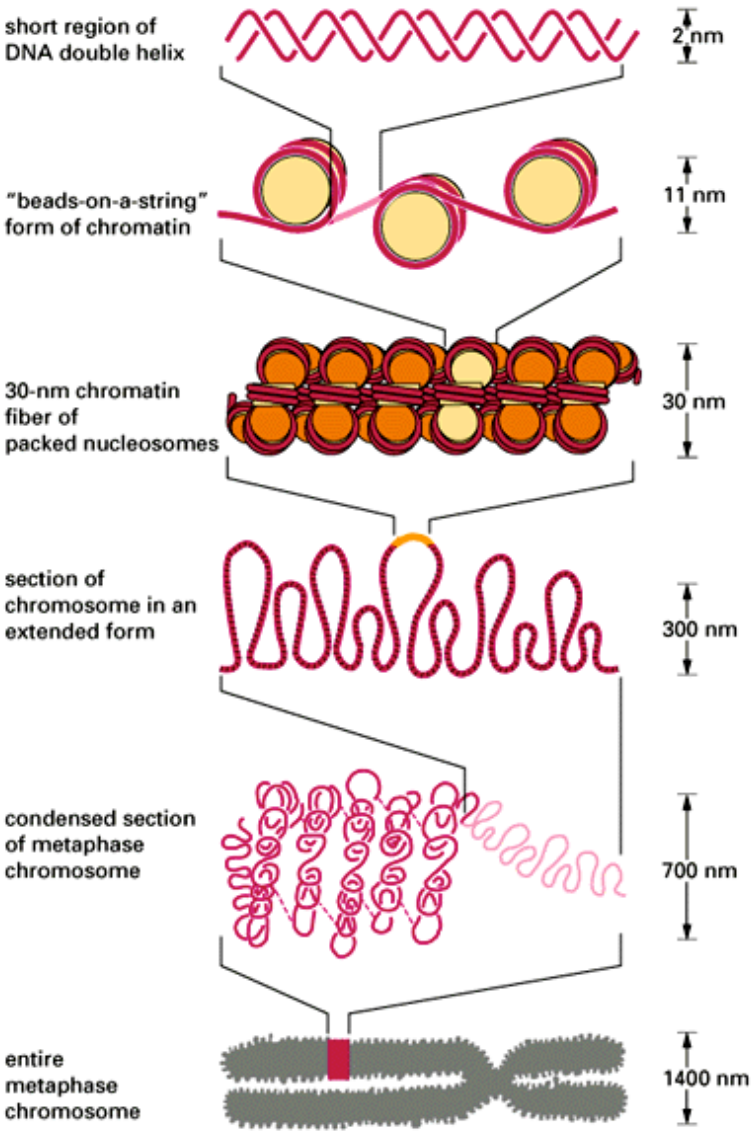


HeLa Cell

RNA (5-bromo-UTP / FITC)

DNA (TOTO-3)

To which extent does the DNA sequence “code” for the structure of chromatin?



Sequencing projects result in 4 letter texts :

```
gtcagttcctgaggcgggtcgggacccaggcgtgagactggagtctgcc
caggggcccagctgagccagcctcctcgtcagctgcttgggccgccagga
cgccgccgggggtgcgccgcgcttccctggatggggtgcccccactcccc
tcggagccccagggagaccccccgaactcagctcctctcaggggtgccag
ggggaccctcaaactccactccccgcaggttcctggggagacgccccct
gctcgattcccctcagggtcccagggagacccccctaattcagctcctctc
aggggtactgggggacctctcgagctccactcccatcagggtcccaggga
gacccccccaactatgctcaggggtcccagggagatgccagcaccccaact
ccgcttccctggggcccccctcccccttacagctcaacttccctcgagagt
ctggggctggggctccgttcagttcttgagtcccccttccctcggggtgtc
ccggggccgcccacccccacactgtctgtgattccccaaggcgcgggtct
cgggccgcagcctgttccacgttctgctgctcgttcttttctggctcctt
gctttcgaaggagagaaggaggccttcgtttccagtctttttgccttttc
taatggagccctgcttttccttccgtgtcccttcaggctacttctgccag
gtttctatttttcattctttattatgacttcgcccaaaatattcttgact
tctattgagaaggattcgggggtctatttcttattcggaggcgtgtgctt
aagttccaaacagatgaggattttccagttaatccttctggggtgactta
ttgcttaatgccaccatagccagaaaatggactctcagtgtccgaaactg
cattcggctctgaagtgtctgtccttgtcacctcttgcaatgtttcgcgg
cgggaagcctgcactcgccgacgctgacgtaactgtttctgtctttcagg
tctacagcctcctgtgggtgggcgatattgacatatactttatttctata
tatgttatgaactcaatatttcttgcagcgggtctgctgataataagata
tgcctactctgcgagtctggaagccatcttaagcttaccctgtatgtgcc
ccatgcatctcttccgttacacggctcctgagttgacacctgtgtgataa
actggtaatagcaagtaaactgttttcttgtgctctgtaagctgctctag
caaattatctaggaggaggtggtcttggaaacccctgatttataagcggg
cagtcagcagtacacgtggcccagaatcgtgattggcatttgaagtgggg
gcagtagggtgggactgagcccttcacctgtggggtctgccctgctcaag
gcagtgtcagaattgaagtgaaatgttggacggtcggtgtccagagagtt
ggagaactggtttgtgtgtaaaaactnacatatttagggtcagaagtatg
...
```

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

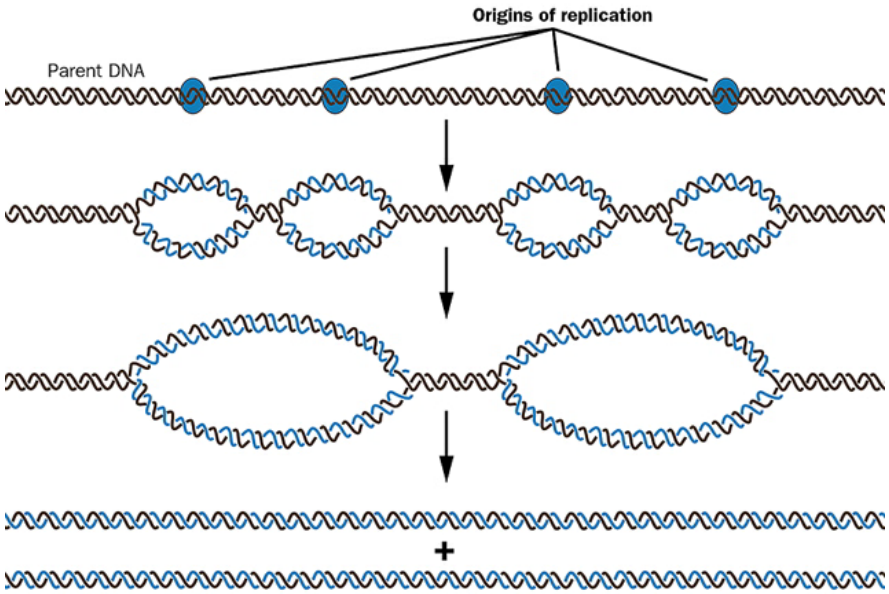
Article history:
Accepted 24 September 2010
Available online 8 October 2010
editor: H. Orland

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.

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

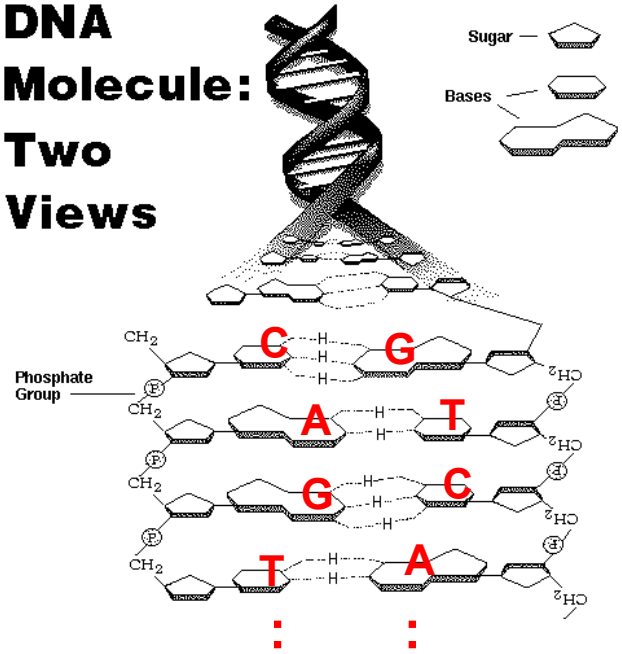
➢ **multicellular eukaryotes :** *replication origins are « terra incognita » !*

- 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**

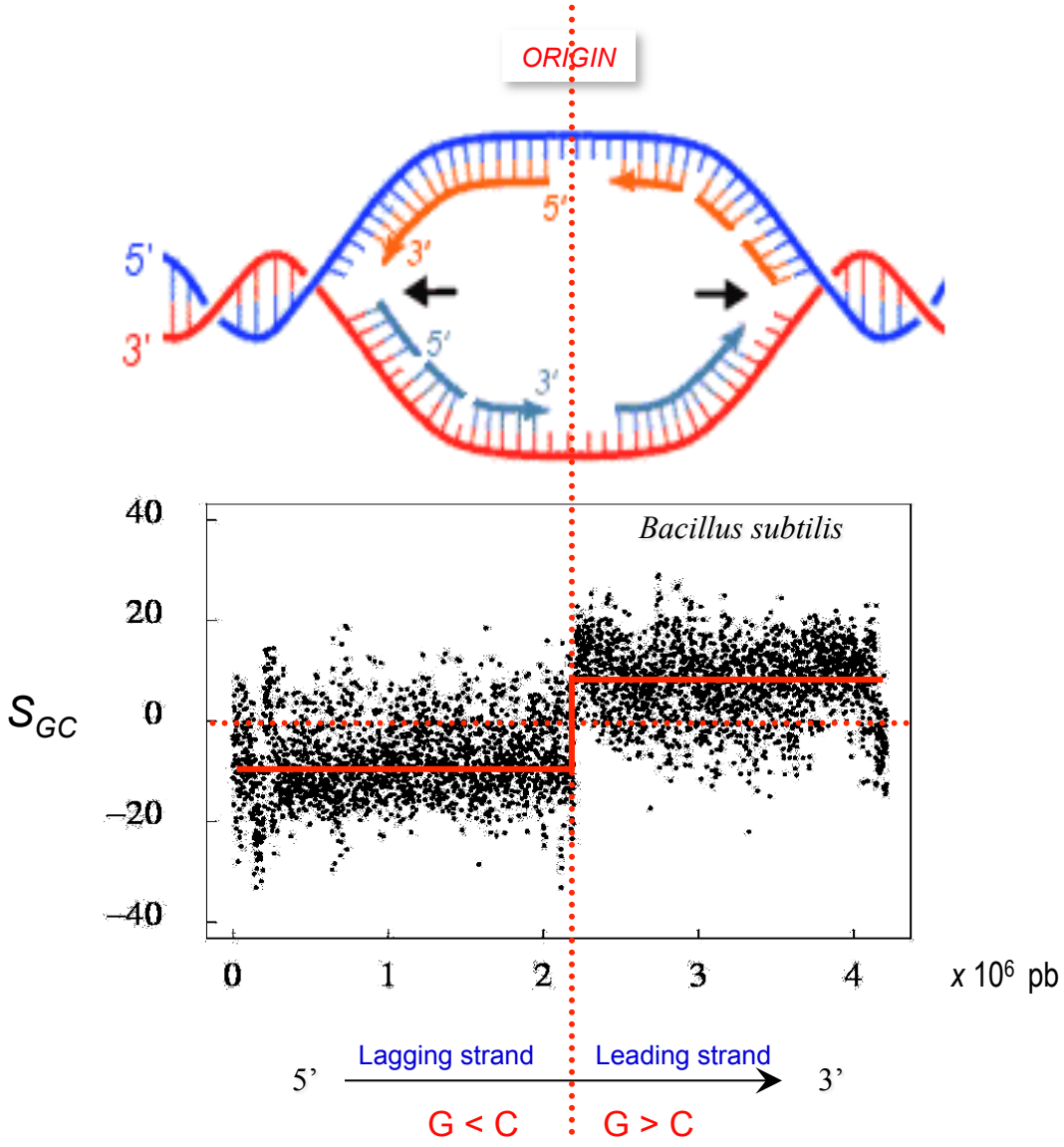
Different nucleotide substitution rates in the leading and lagging strands

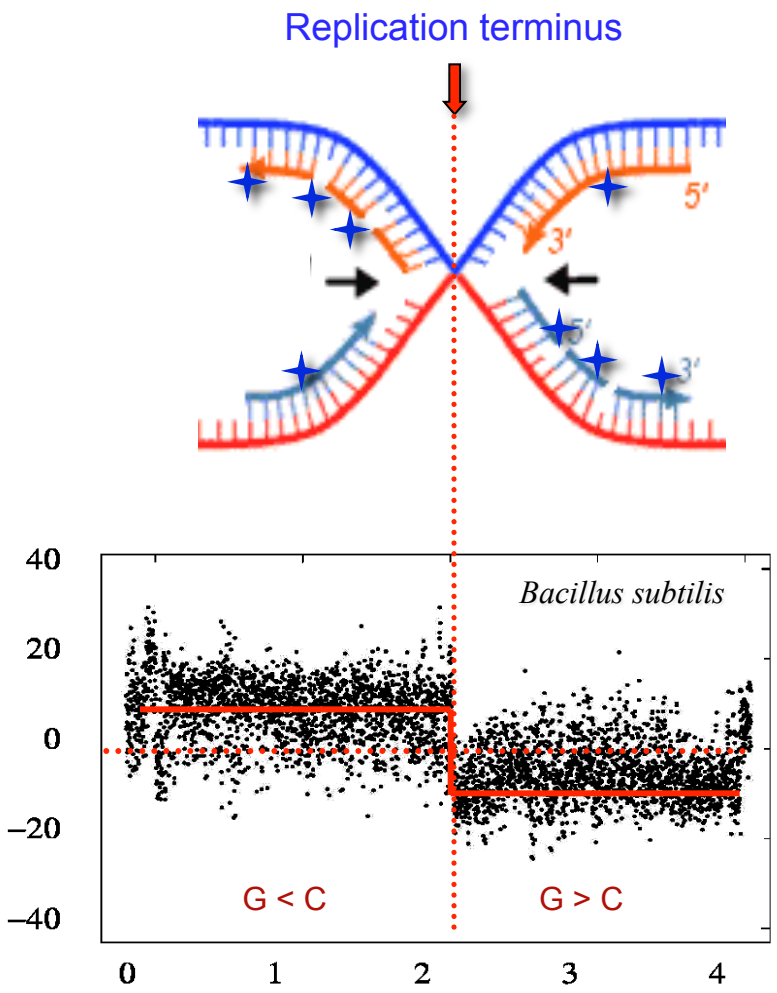


Symmetry breaking

[G] ≠ [C] ; [A] ≠ [T]

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

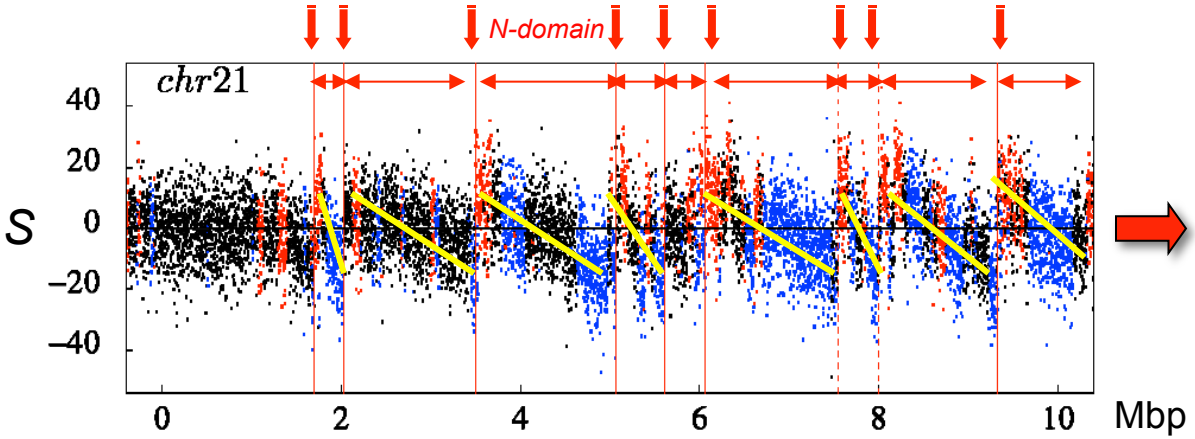




Used to detect replication origins and terminus in bacterial genomes

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

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

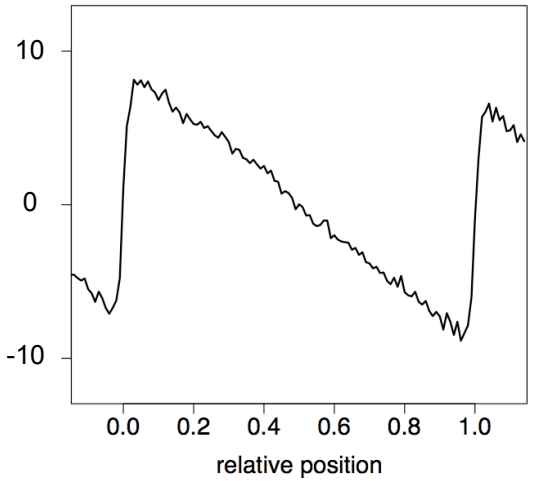


~ 700 N-domains

> 1/3 of the genome

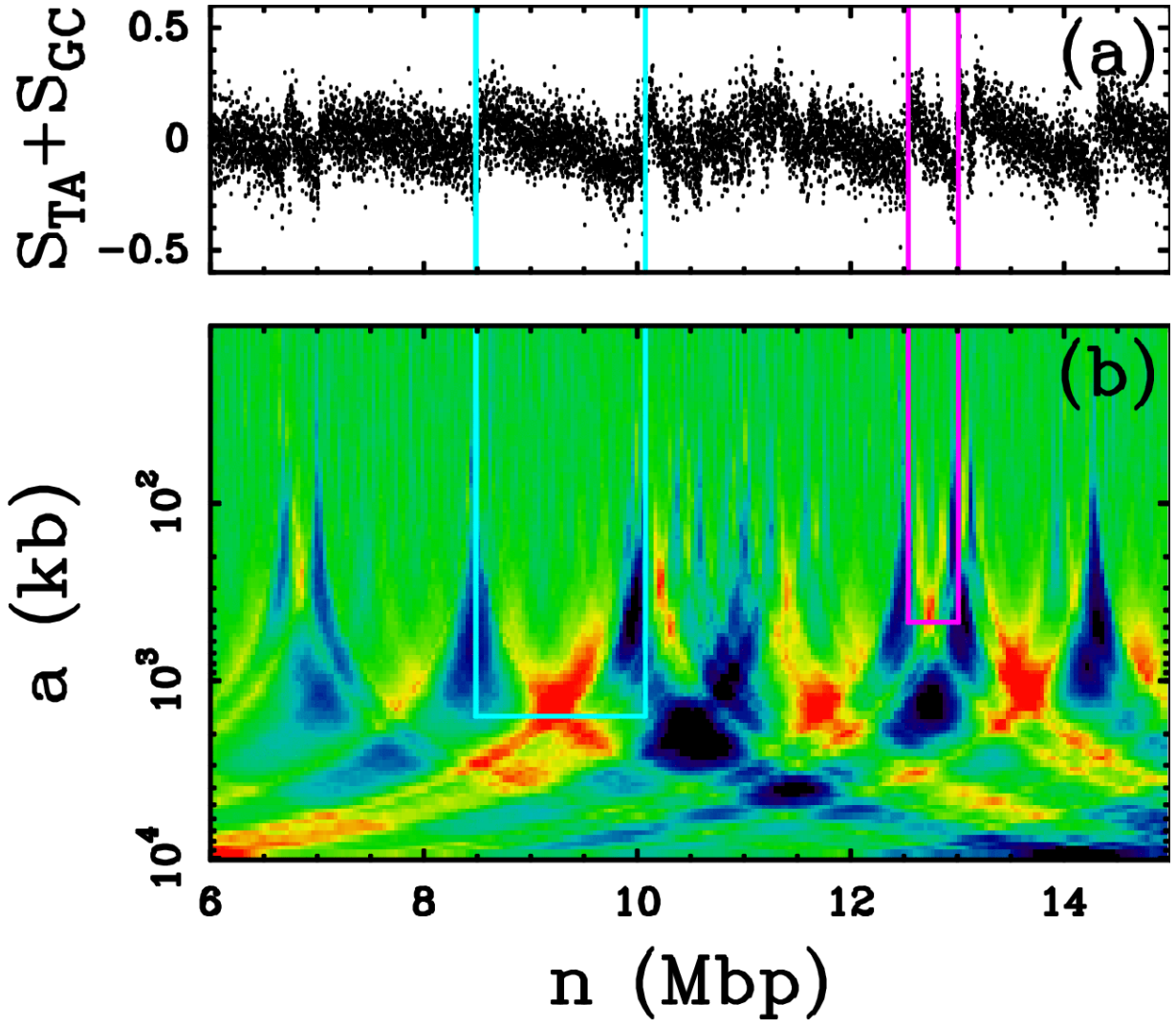
> 1500 putative replication origins

Mean skew *S*

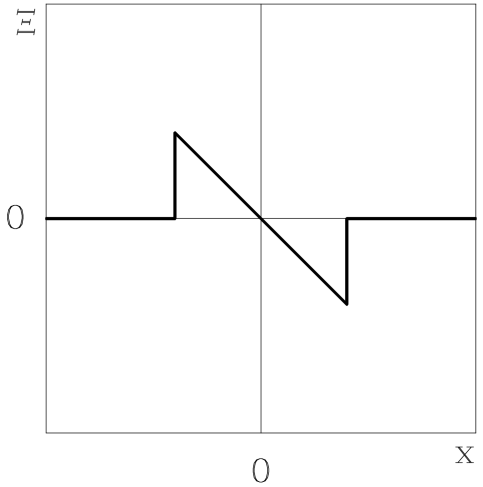


Touchon, Proc. Natl. Acad. Sci. USA (2005)
Brodie and Brodie, Phys. Rev. Lett. (2005)
Huvet, Genome Research (2007)

Wavelet-based multi-scale detection of replication N-domains



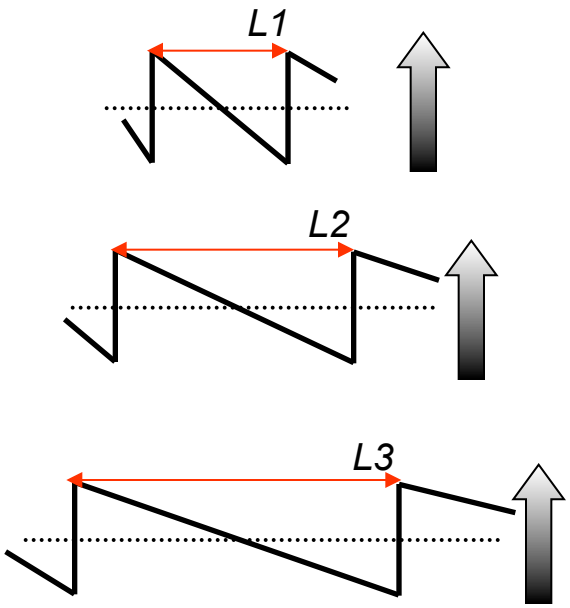
adapted analyzing wavelet

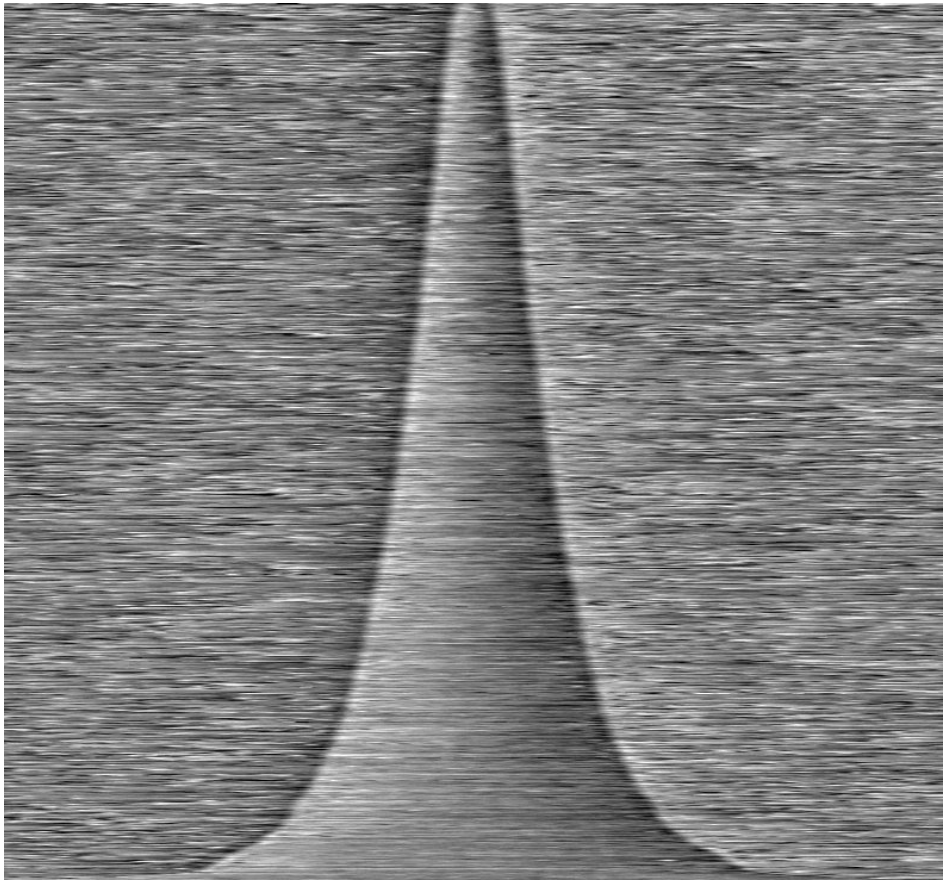


Audit, Phys. Rev. Lett. (2007)
Huvet, Genome Res. (2007)
Baker, ACHA (2010)

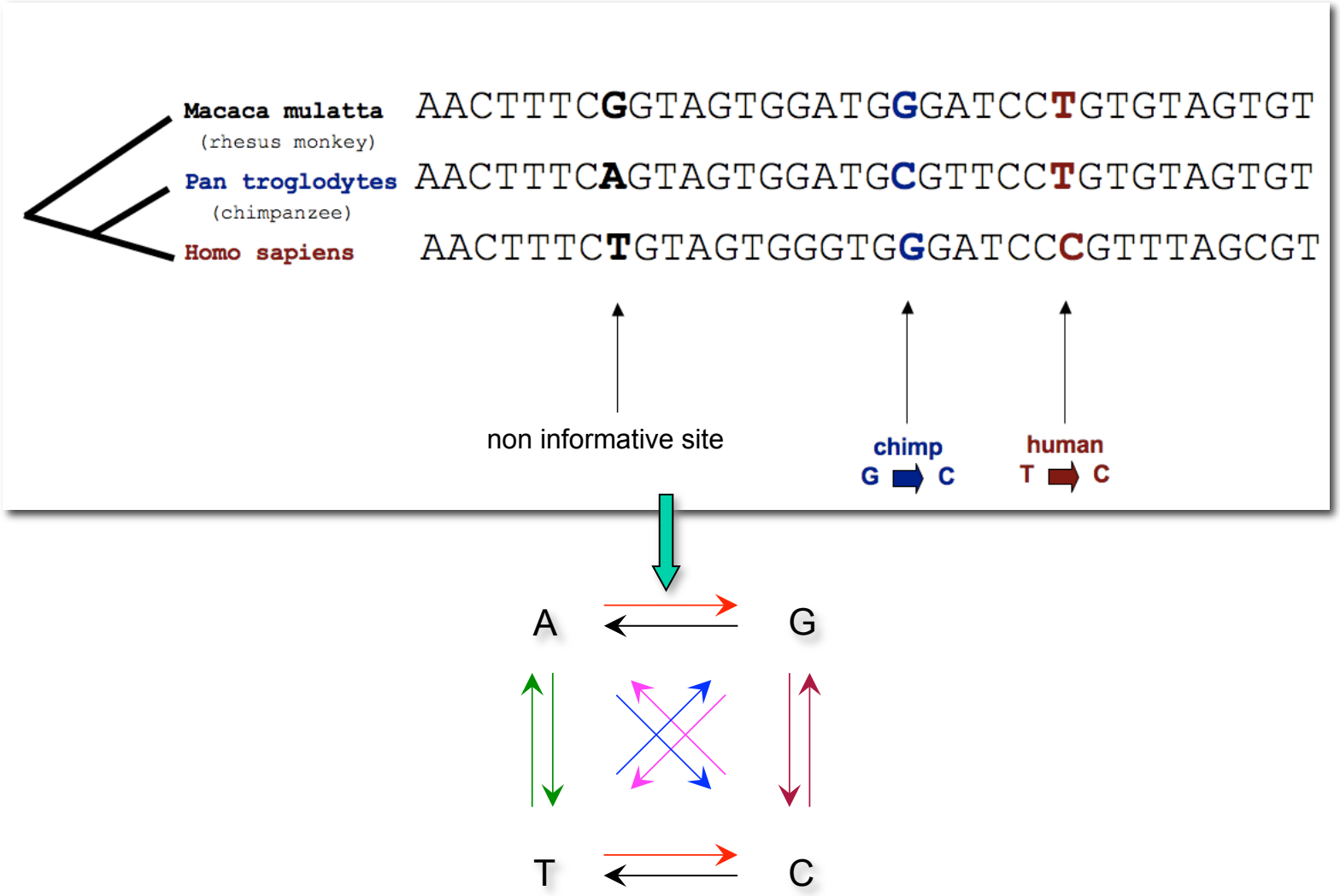
Skew profile of the N-domains

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



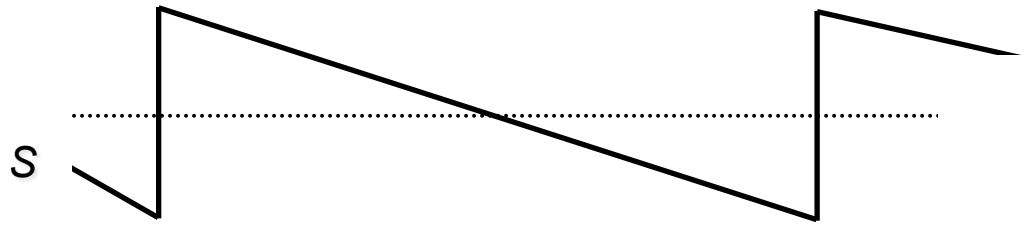


Determination of substitution rates

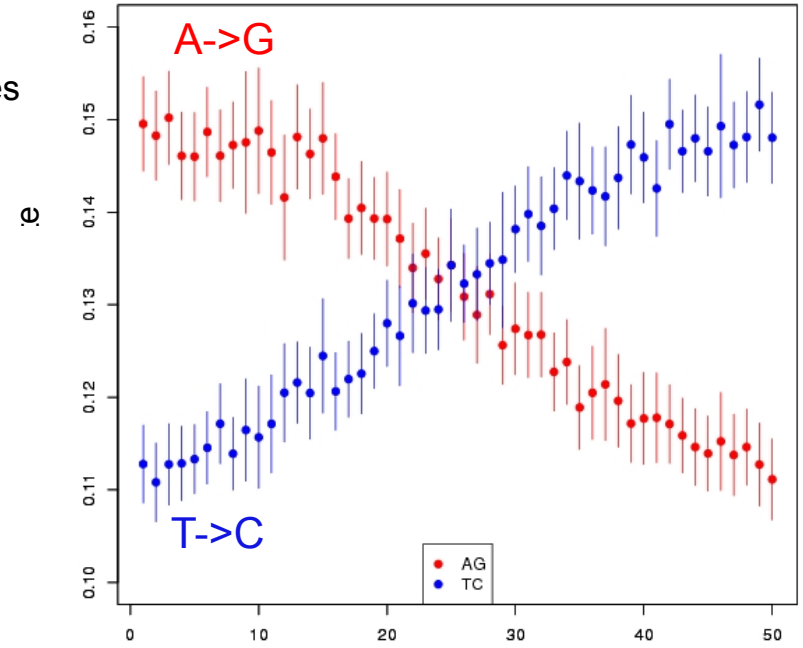


Chen, Genome Res. (2010)
Chen, Mol. Biol. Evol. (2011)

Substitution rates along the N-domains



Normalized substitution rates

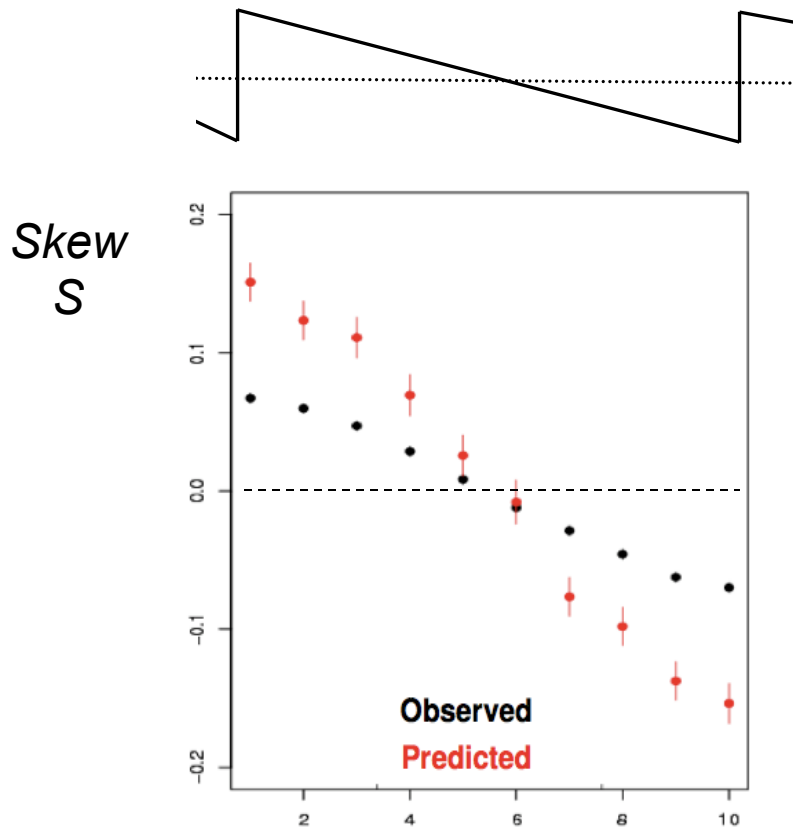


Relative position in the domain

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

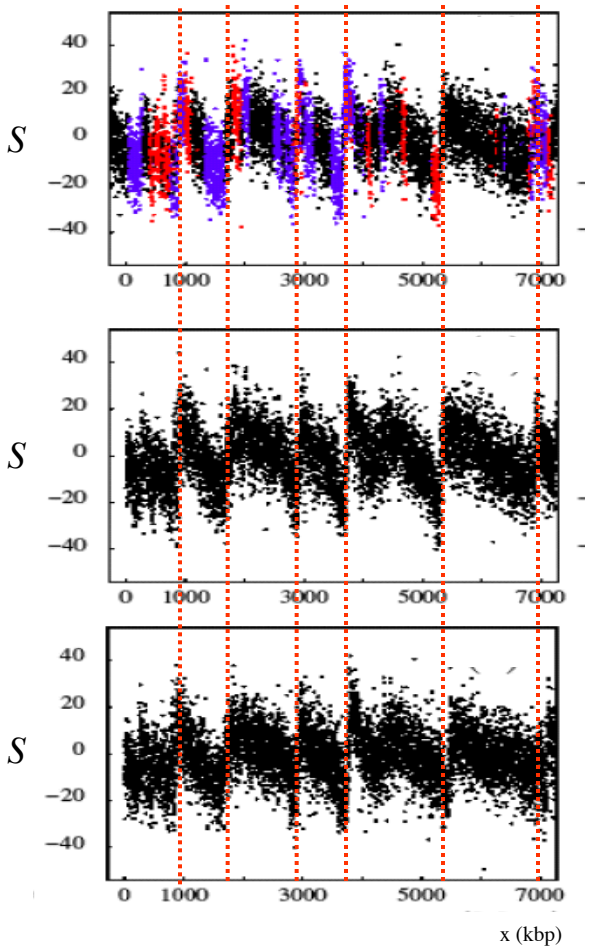
N-domains result from mutation asymmetry in germline cells

Conservation of N-domains in mammalian genomes

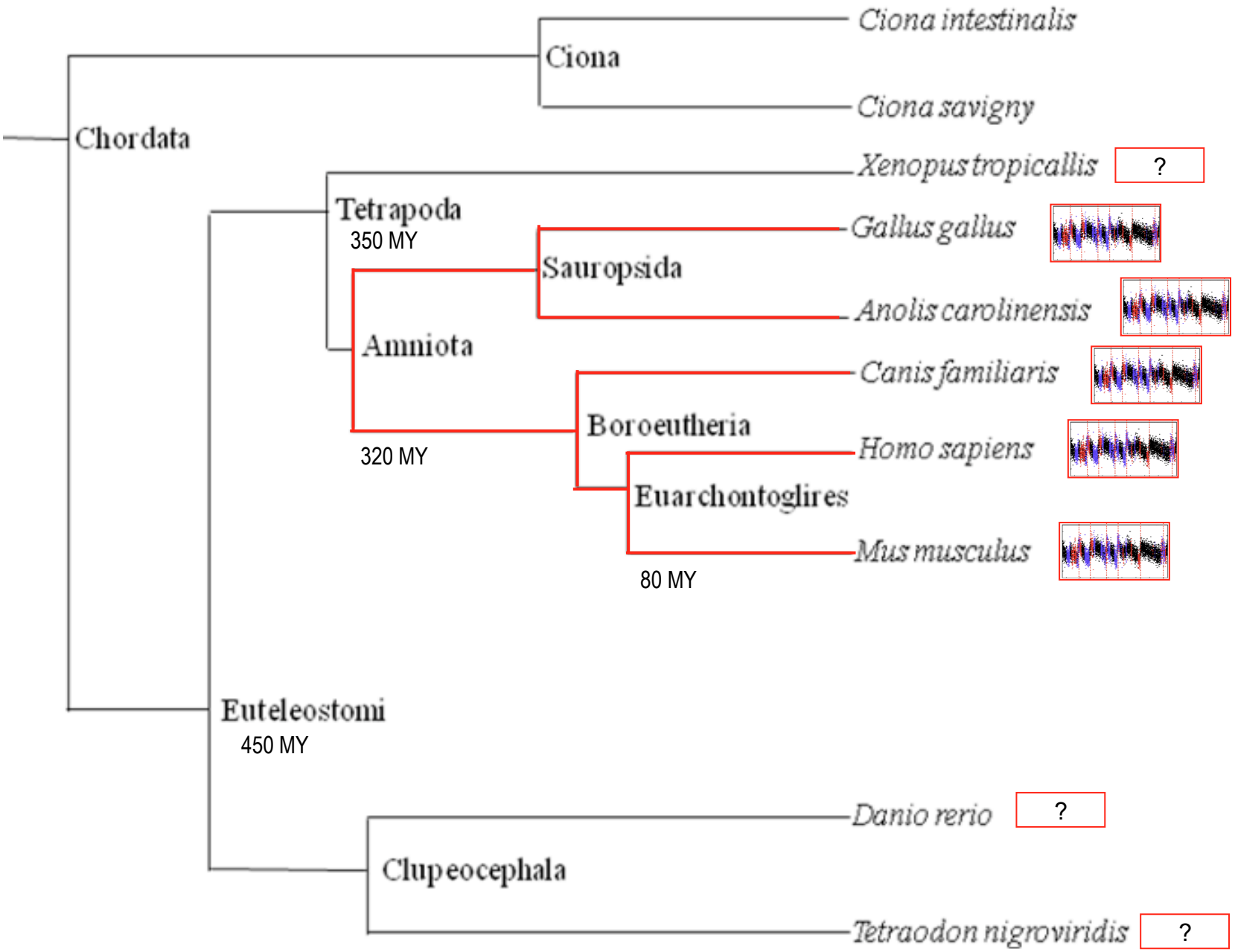
human

mouse

dog



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

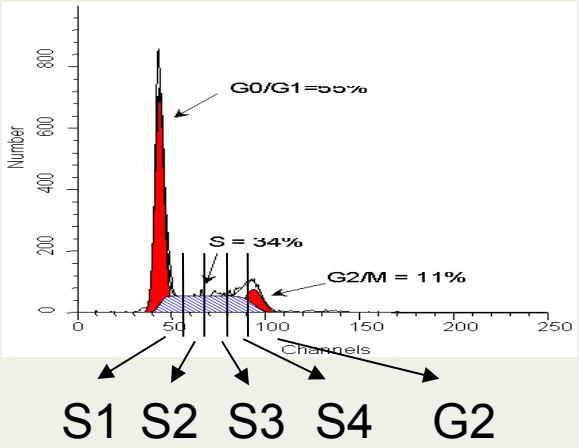
S1, S2, S3, S4 | S | G2

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

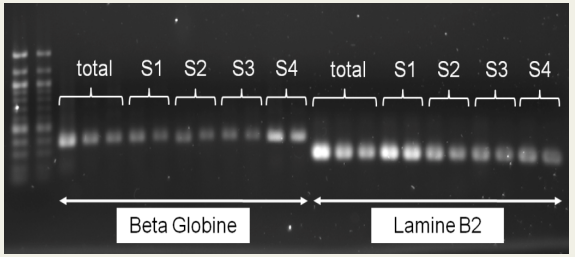
Immuno-precipitation

Random priming 10 minutes (generate double strand)

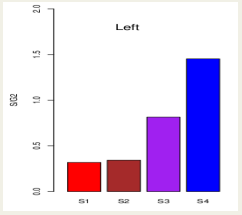
Ilumina Solexa sequencing
(Sequence reads: 43,974,937)



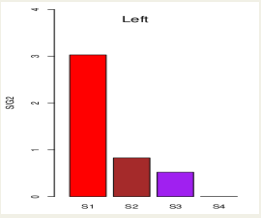
qPCR test



Beta Globin
(late replication)



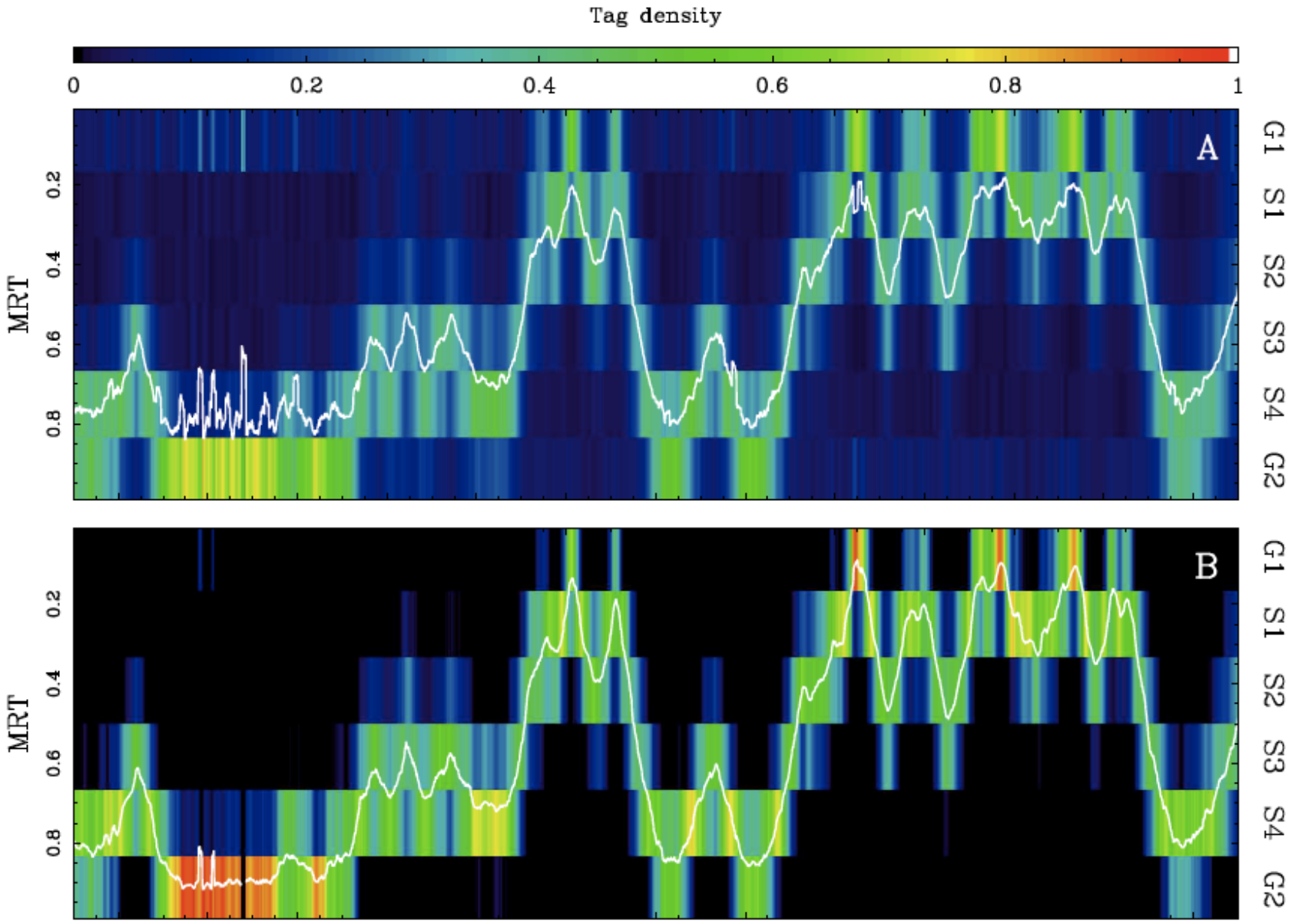
Lamin B2
(early replication)



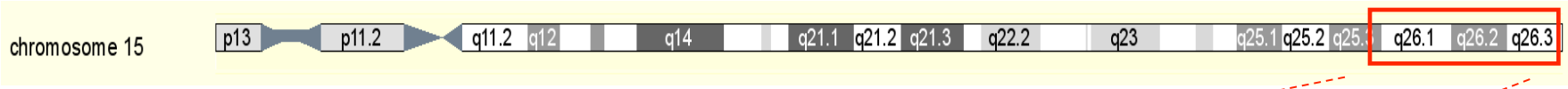
Chen, Genome Research (2010); Rappailles, PLoS Genetics (submitted)

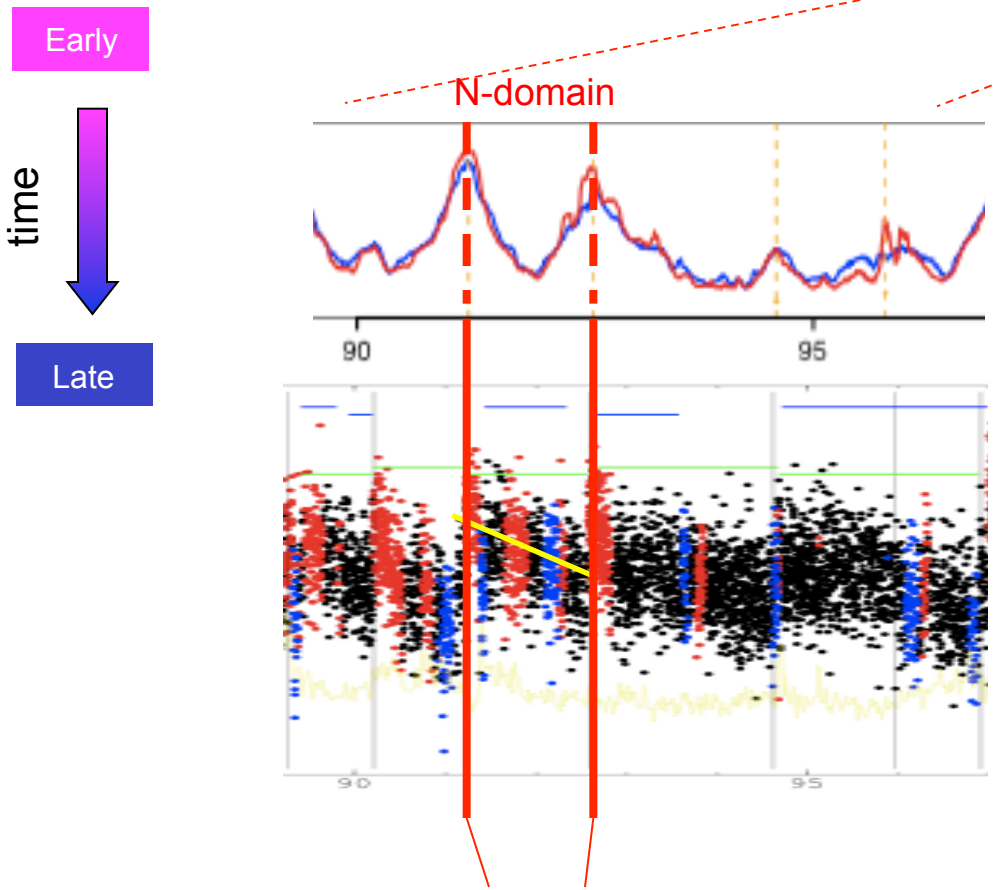
Computing Mean Replication Timing profiles from RepliSeq data

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



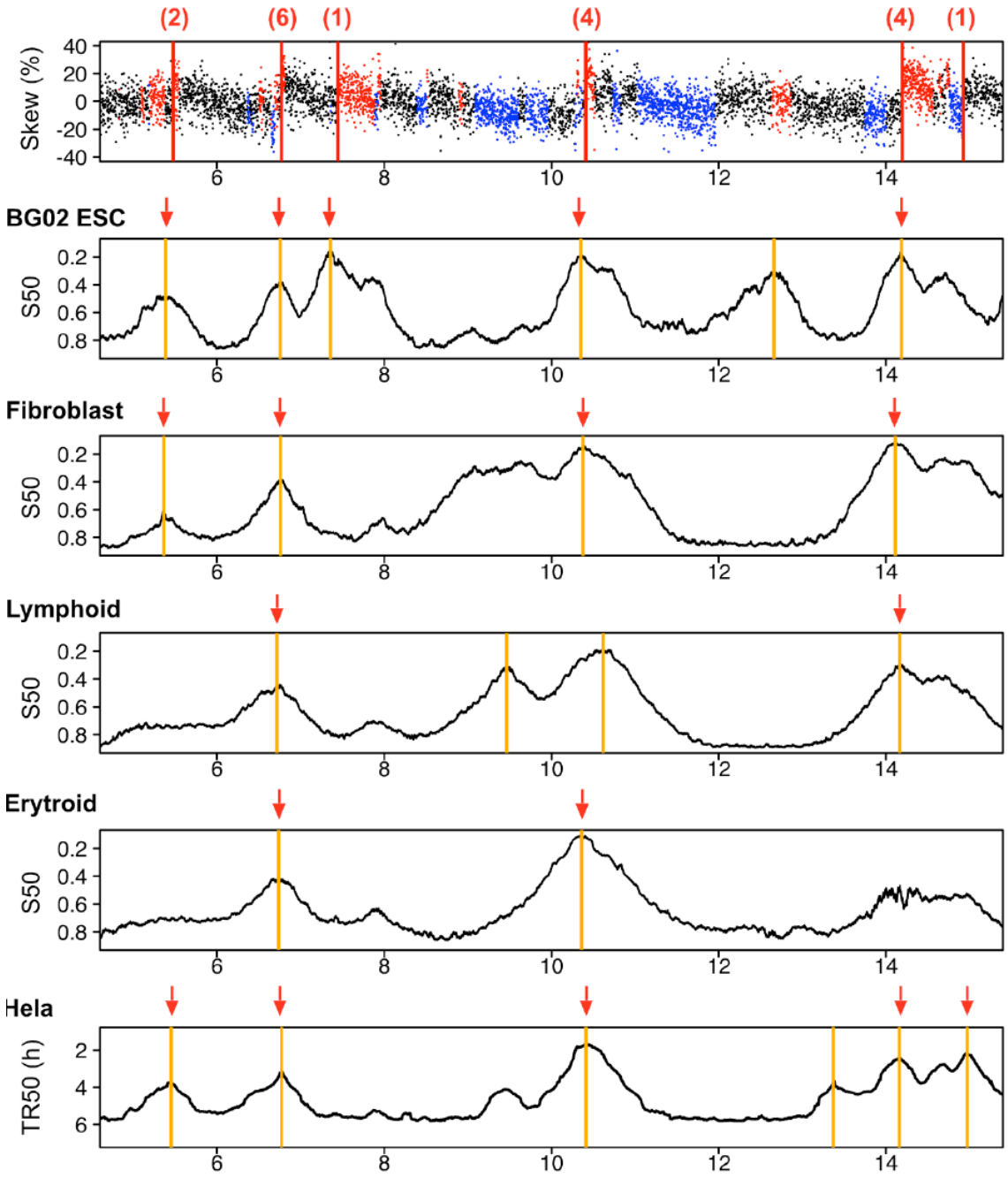
Comparison of upward jumps with initiation zones





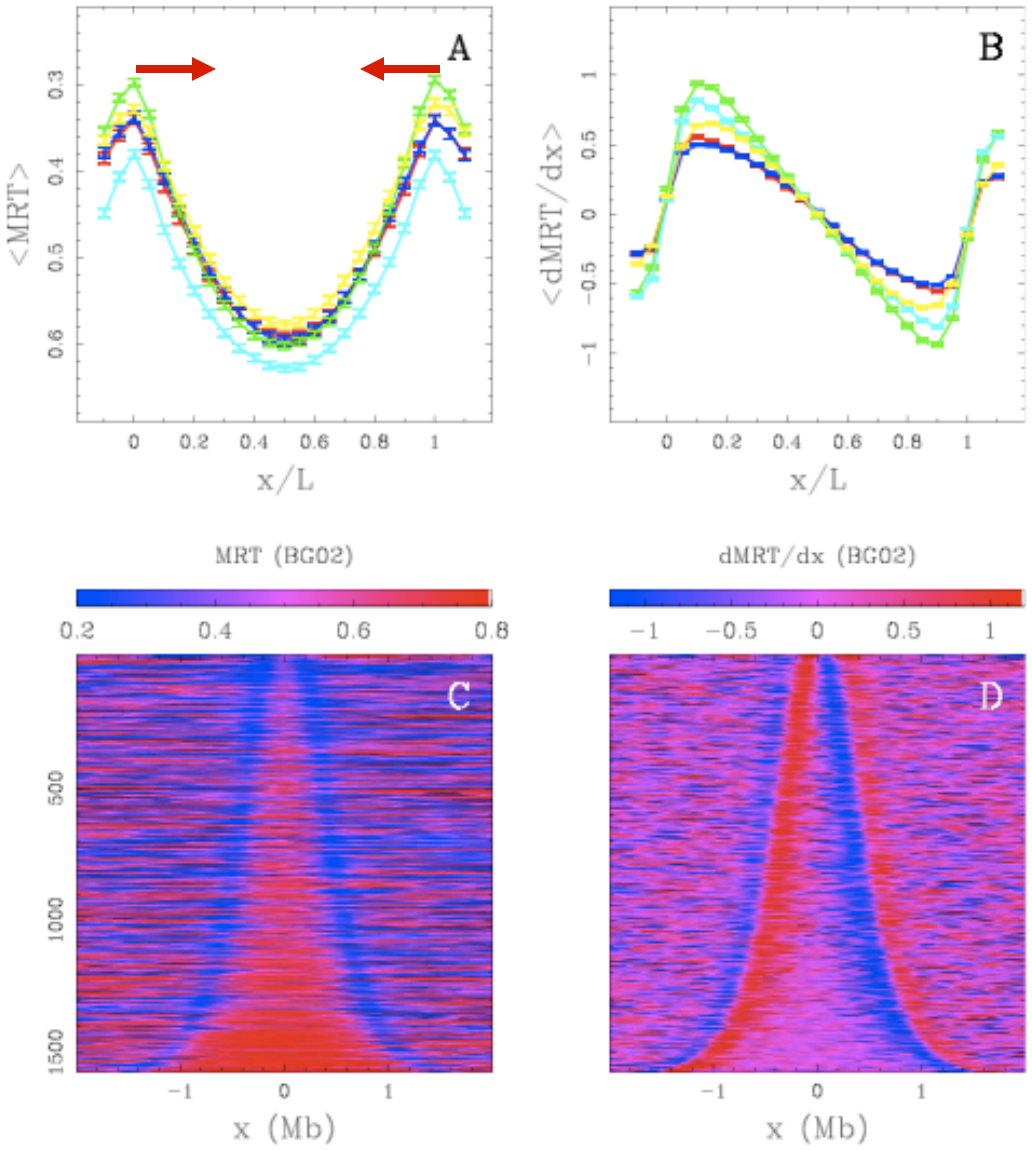
Audit, Phys. Rev. Lett. (2007)
Huvet, Genome Res. (2007)

Replication timing across cell differentiation



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

Replication timing | **Fork polarity**

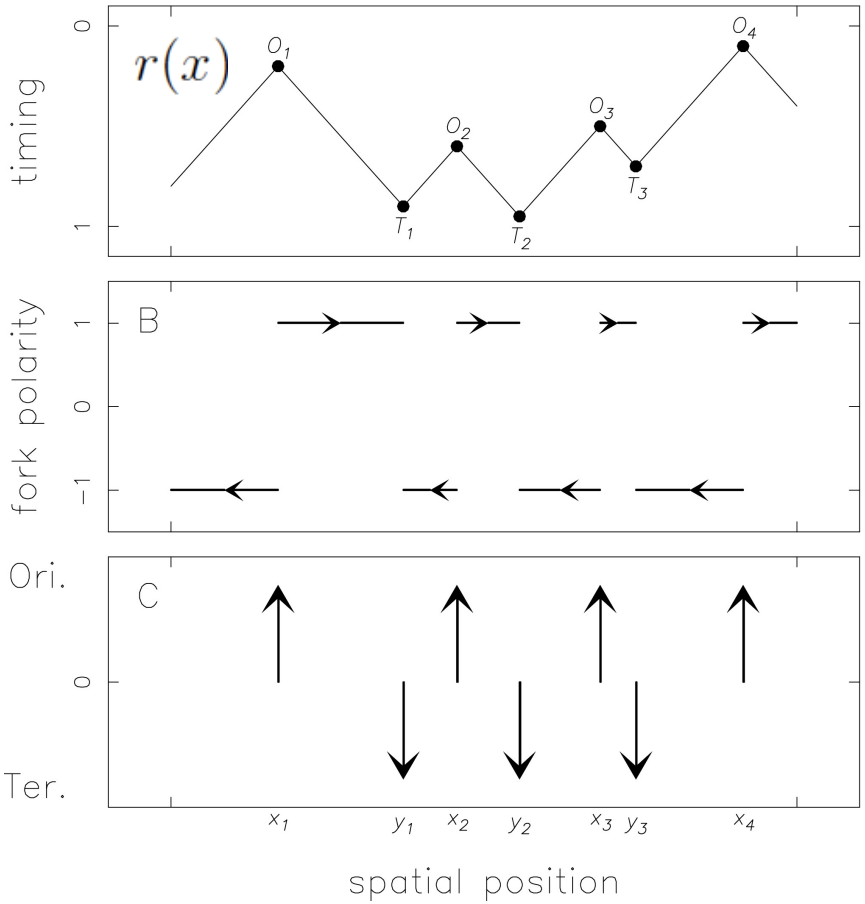


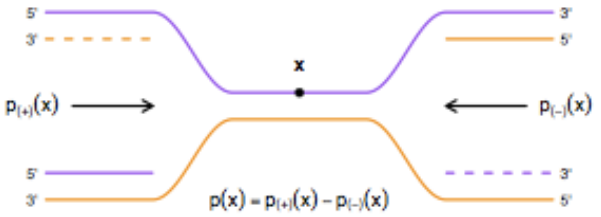
Baker, PLoS Comput Biol (2012).

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

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 :

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

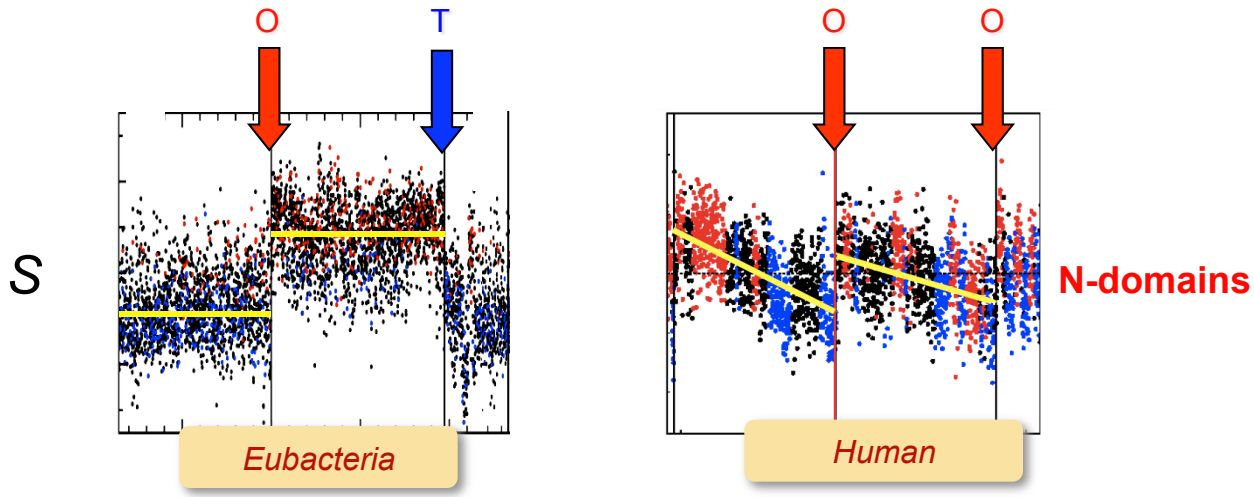
$$\frac{d^2}{dx^2}\langle r(x)\rangle_{\mathrm{Cells},\Delta x} = \frac{1}{v}\left(N^{\mathrm{Ori}}_{\mathrm{Cells},\Delta x}(x) - N^{\mathrm{Ter}}_{\mathrm{Cells},\Delta x}(x)\right)$$

Apparent replication speed :

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

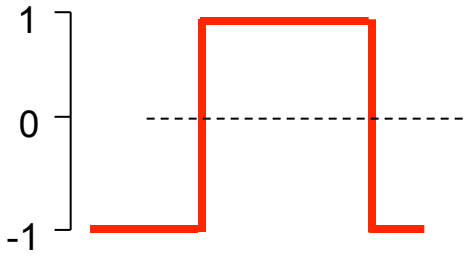
Guilbaud, PLoS Comput Biol (2011)
Baker, PLoS Comput Biol (2012)

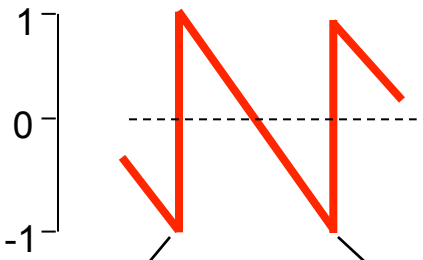
N-shape results from gradient of replication fork polarity



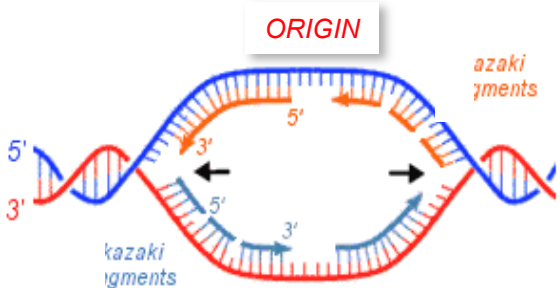
Replication fork polarity

Leading - lagging

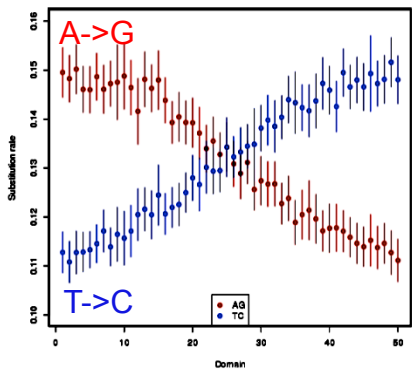




Replication bias
⇒ S ∝ fork polarity

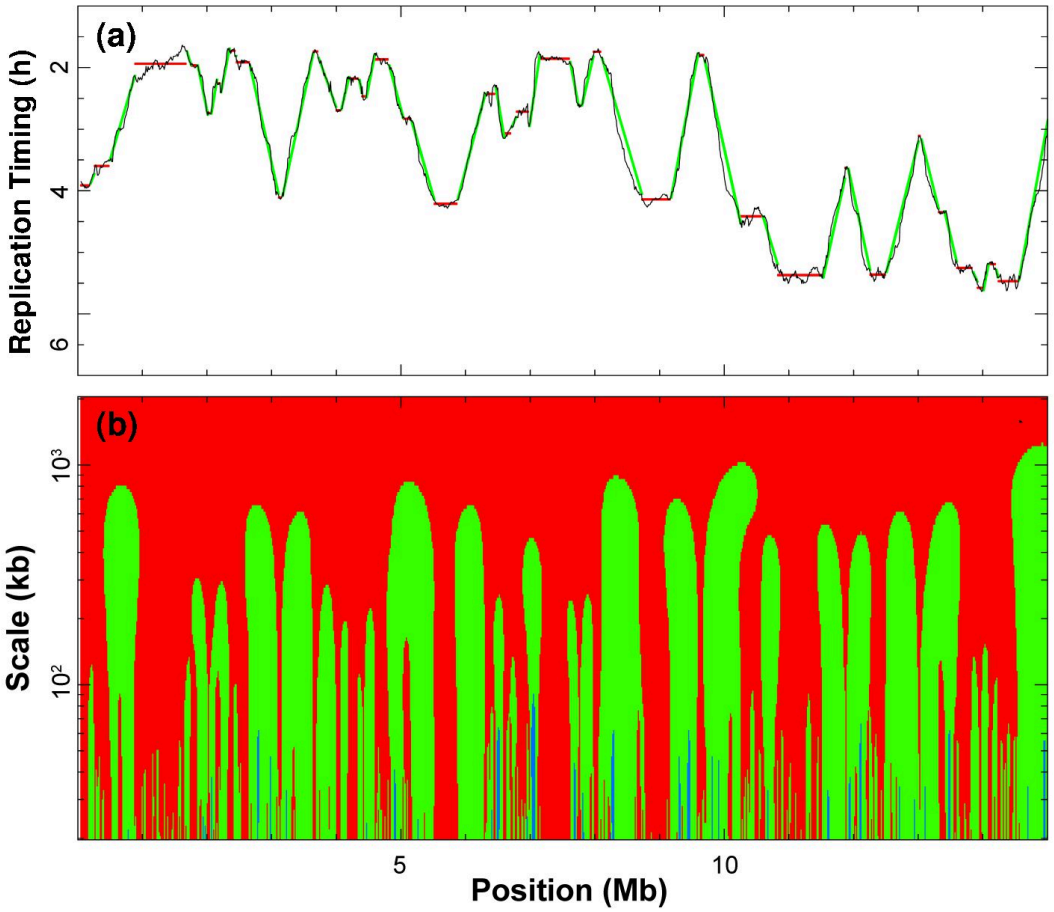


Different mutation rates
on leading and lagging strands

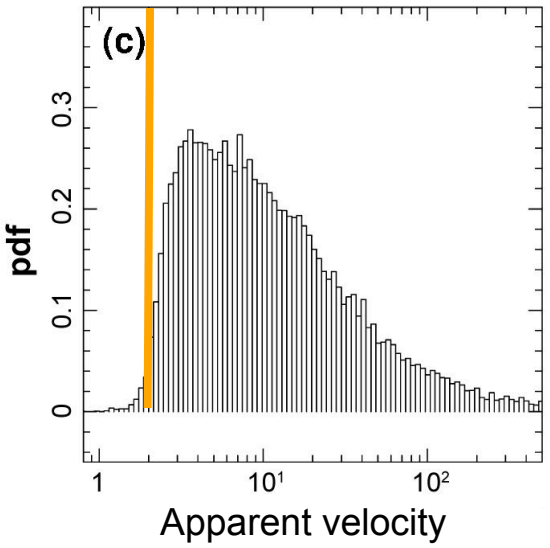


Exploring the space-scale map of apparent speed of replication

Guilbaud, PLoS Comput Biol (2011)
Audit, Nat Protoc (2013)



Vmax = 2 kb/min

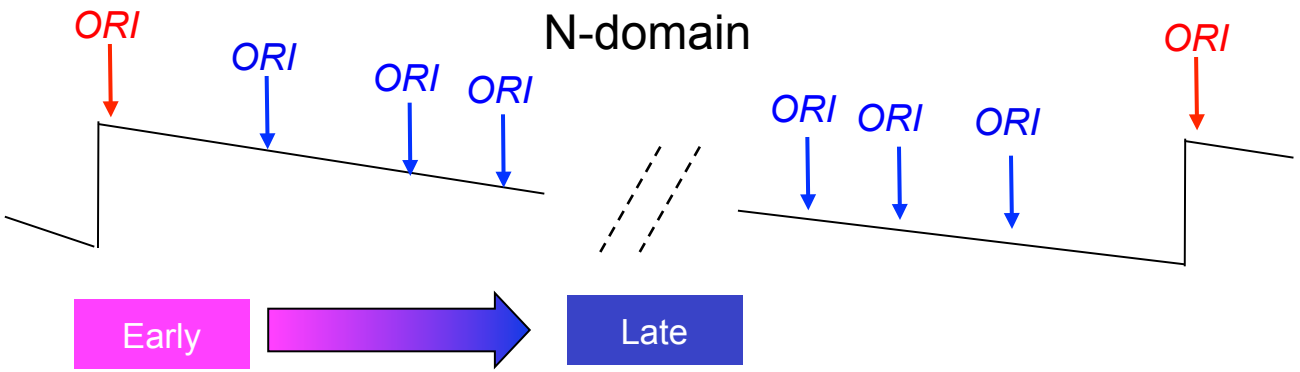


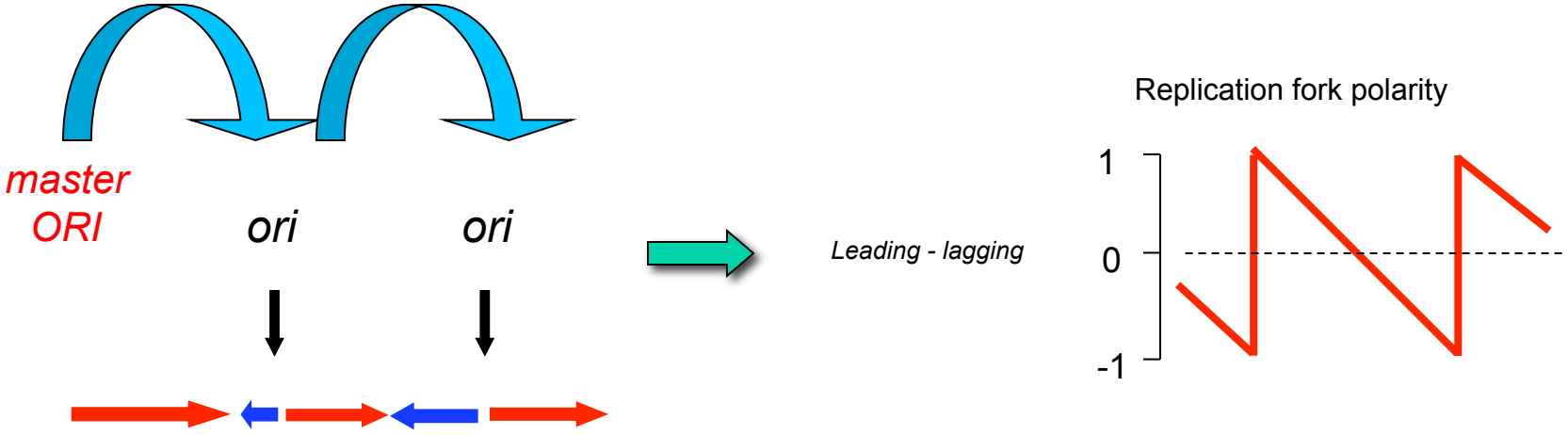
V > 10 kb/min
10 kb/min > V > 2 kb/min
V < 2 kb/min

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

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

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



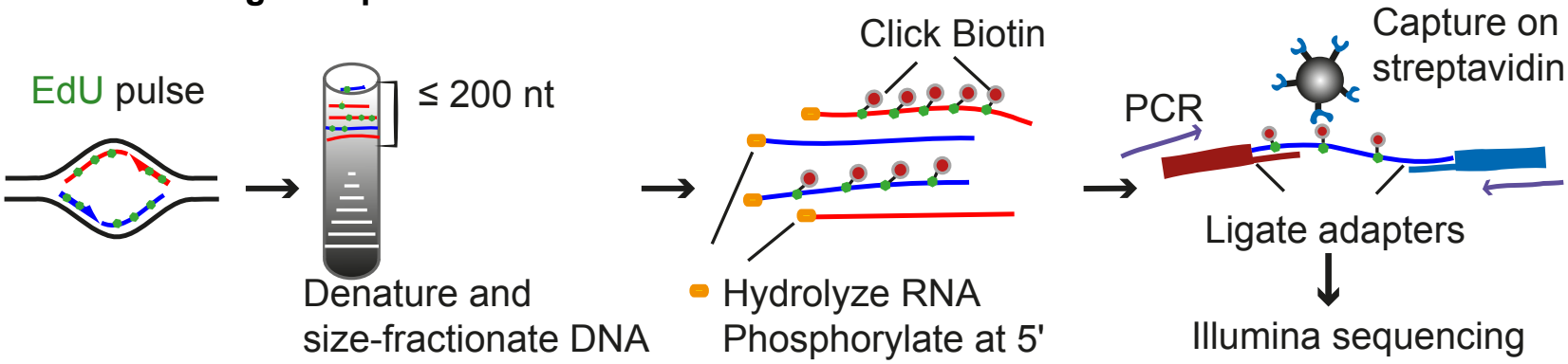


Activation of replication origins propagate along N-domains

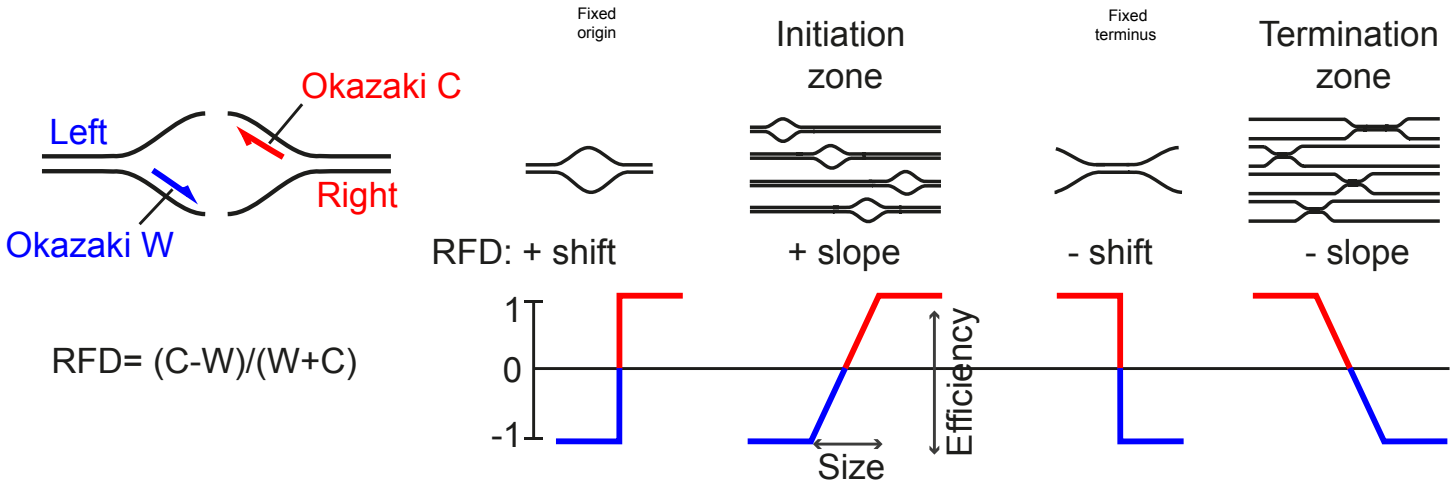
Massive sequencing of Okazaki fragments

N. Petryk, M. Kahli, O. Hyrien (IBENS, Paris)

Okasaki fragment purification

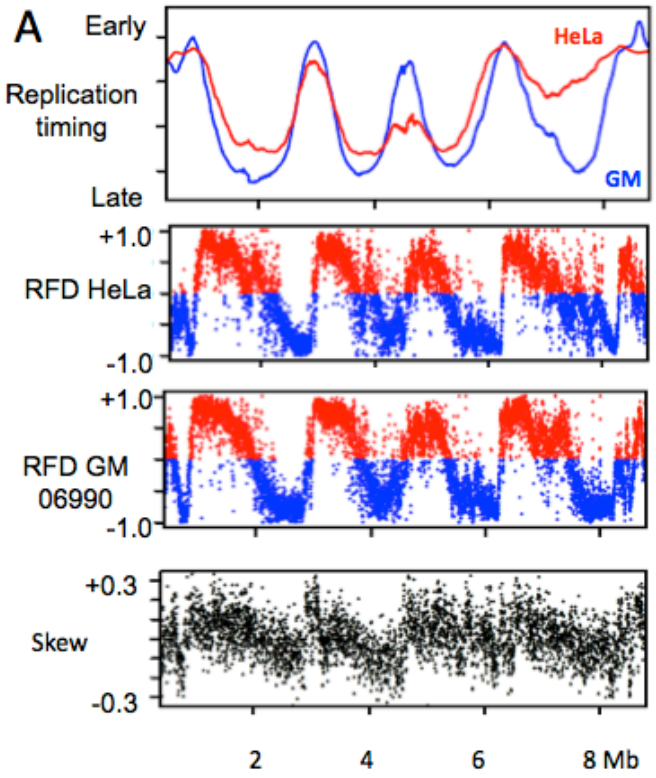


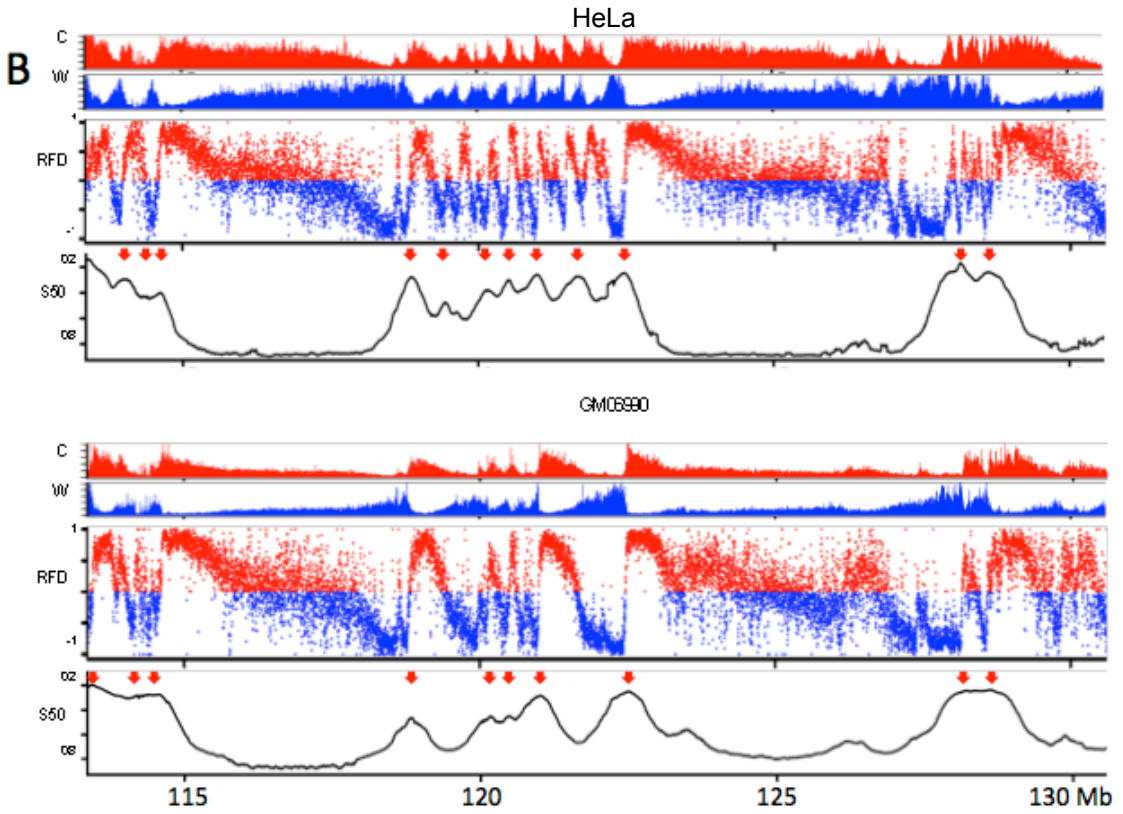
Fork polarity measurement from Okasaki fragment abundance

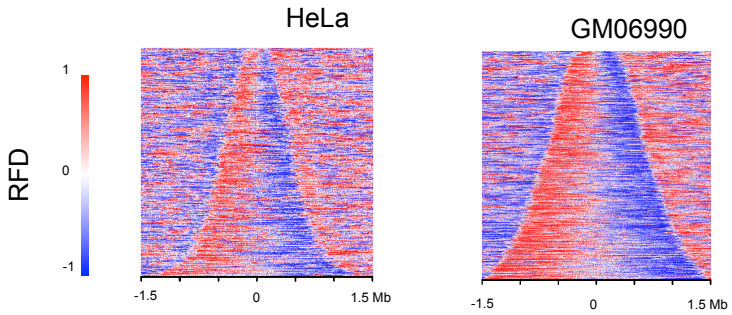


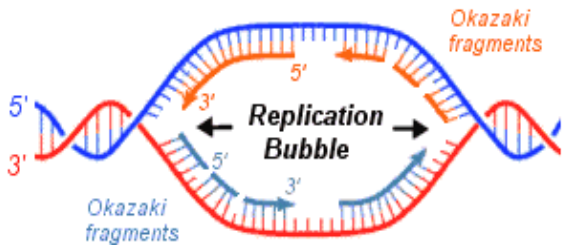
Massive sequencing of Okazaki fragments

N. Petryk, M. Kahli, O. Hyrien (IBENS, Paris)



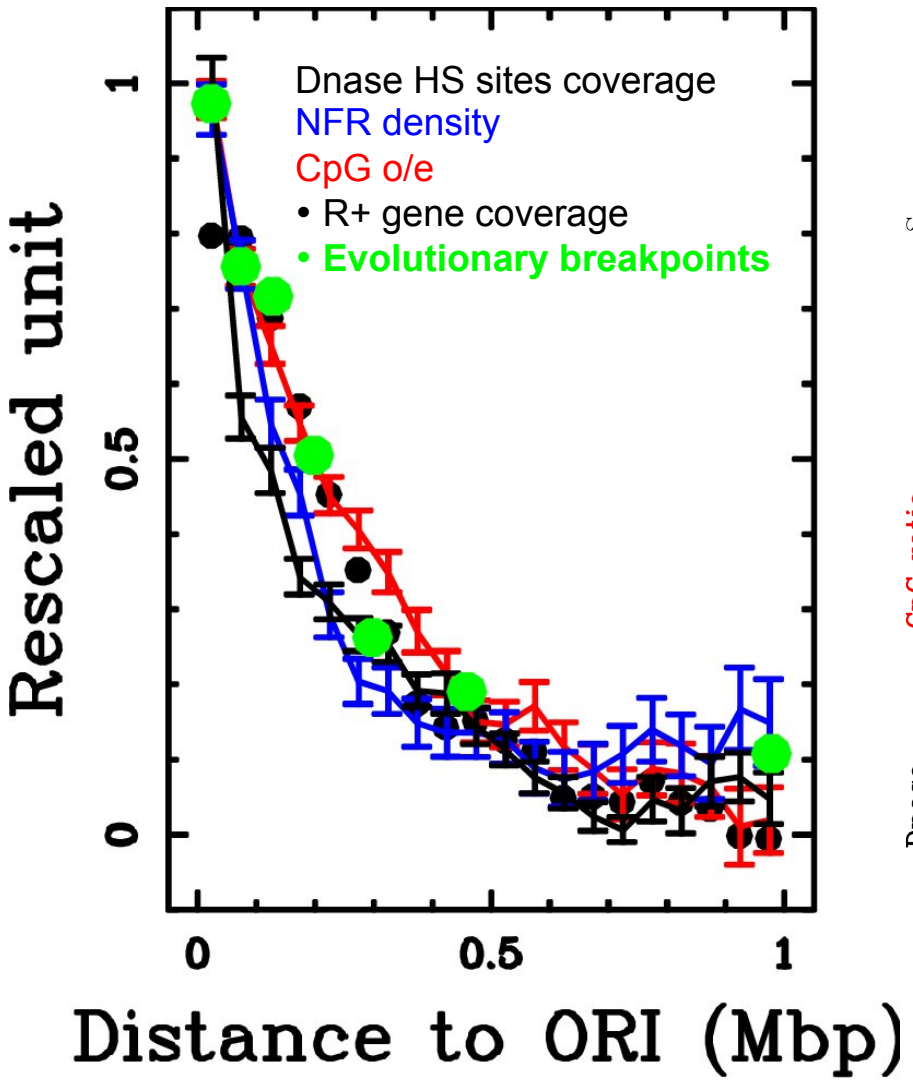


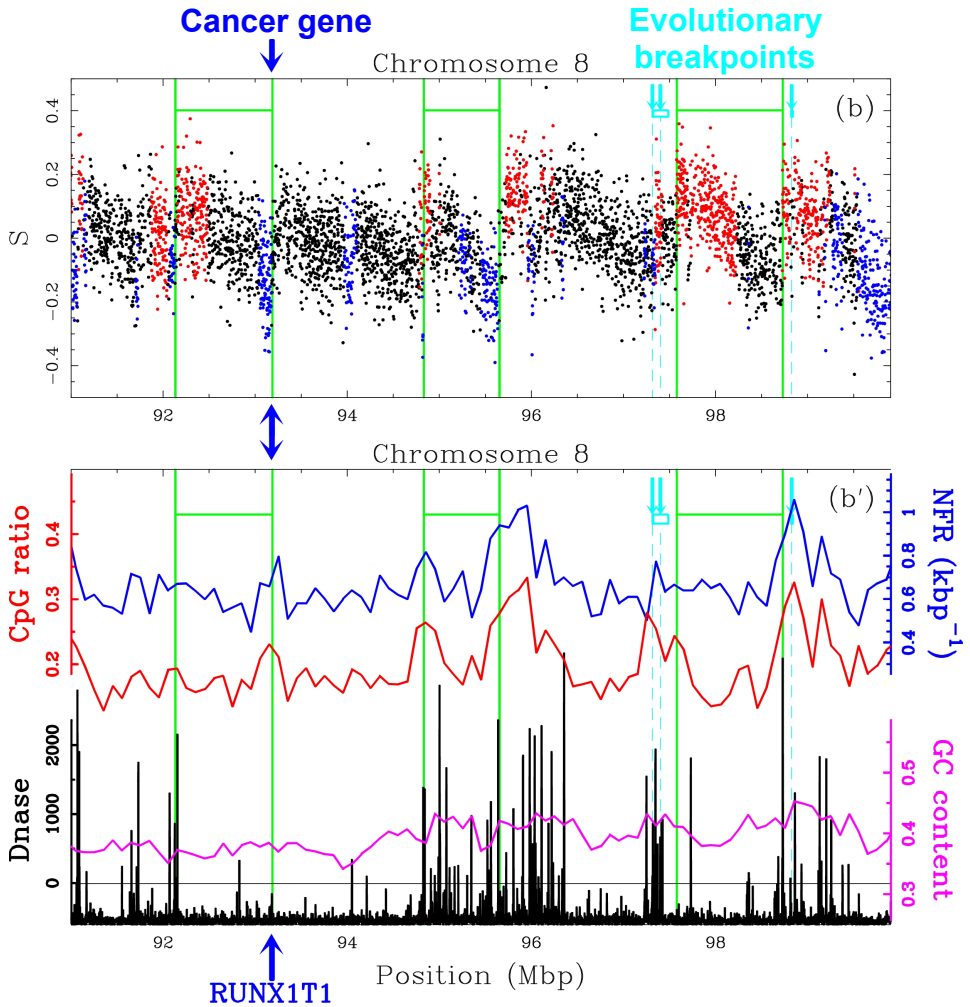




Fragility of the human genome at N-domain borders

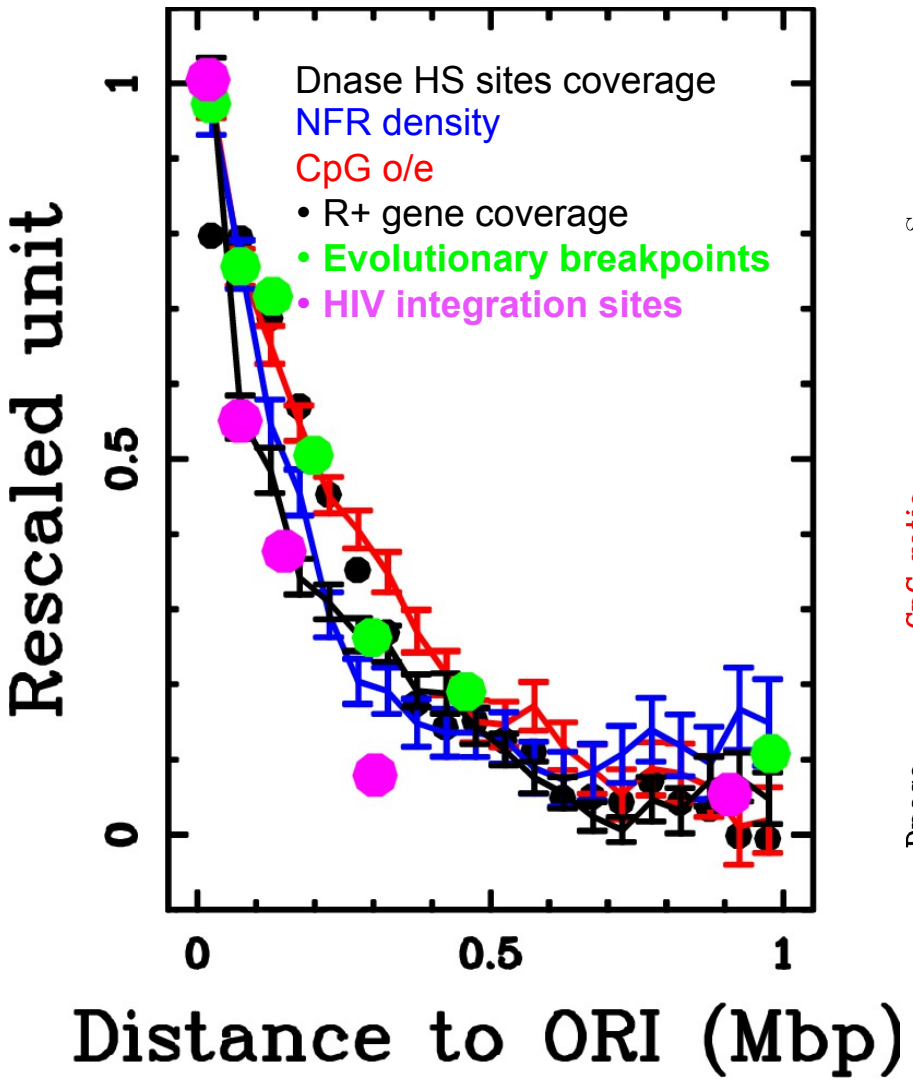
Huvet, Genome Research (2007)
Lemaitre, BMC Genomics (2009)

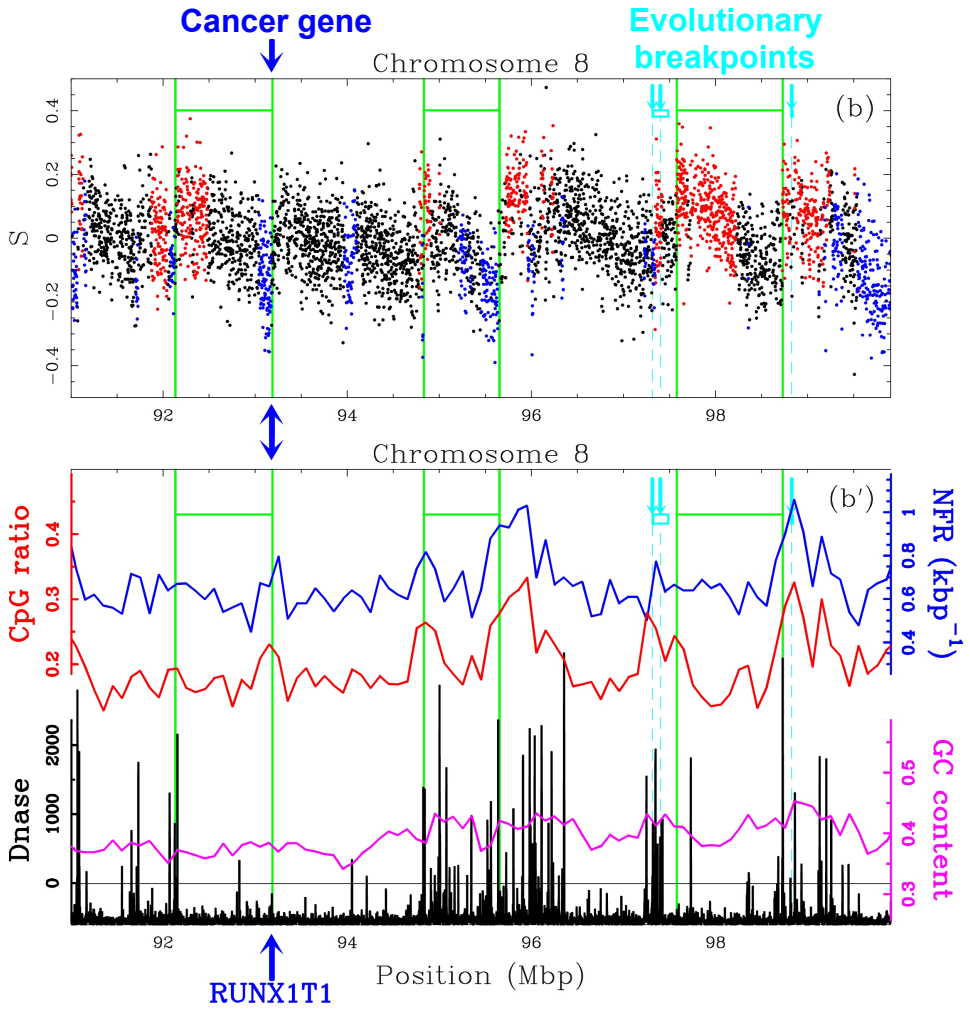




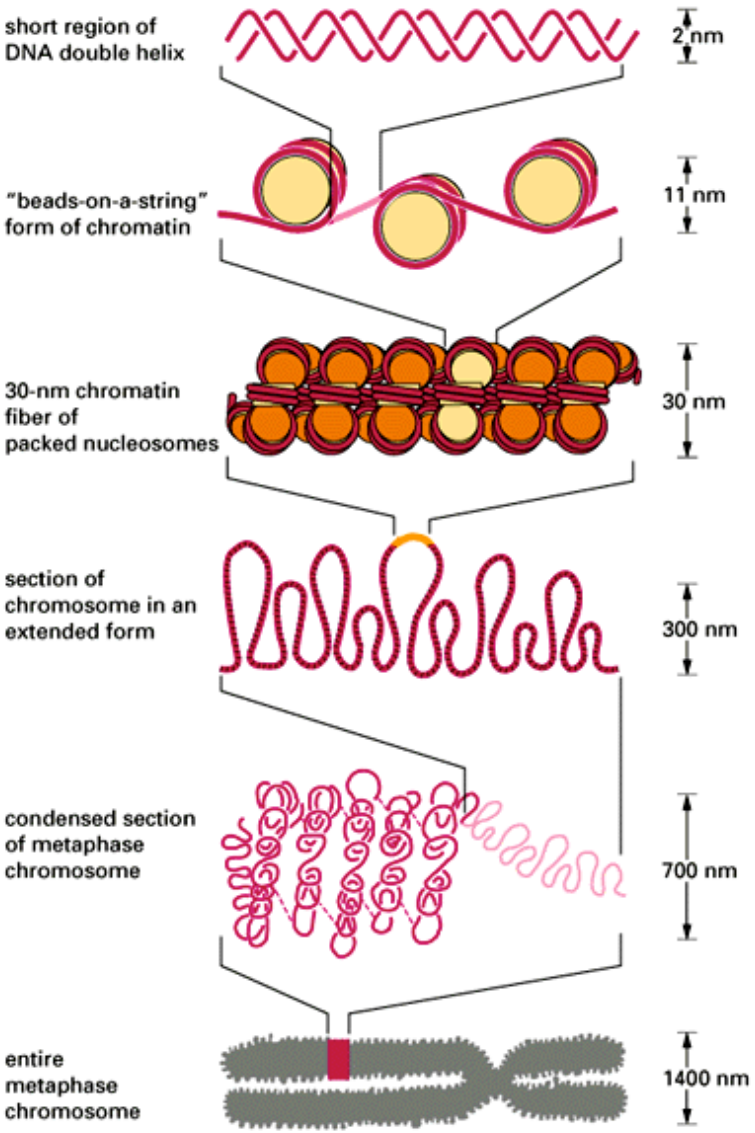
Fragility of the human genome at N-domain borders

Huvet, Genome Research (2007)
Lemaitre, BMC Genomics (2009)





To which extent does the DNA sequence “code” for the structure of chromatin?



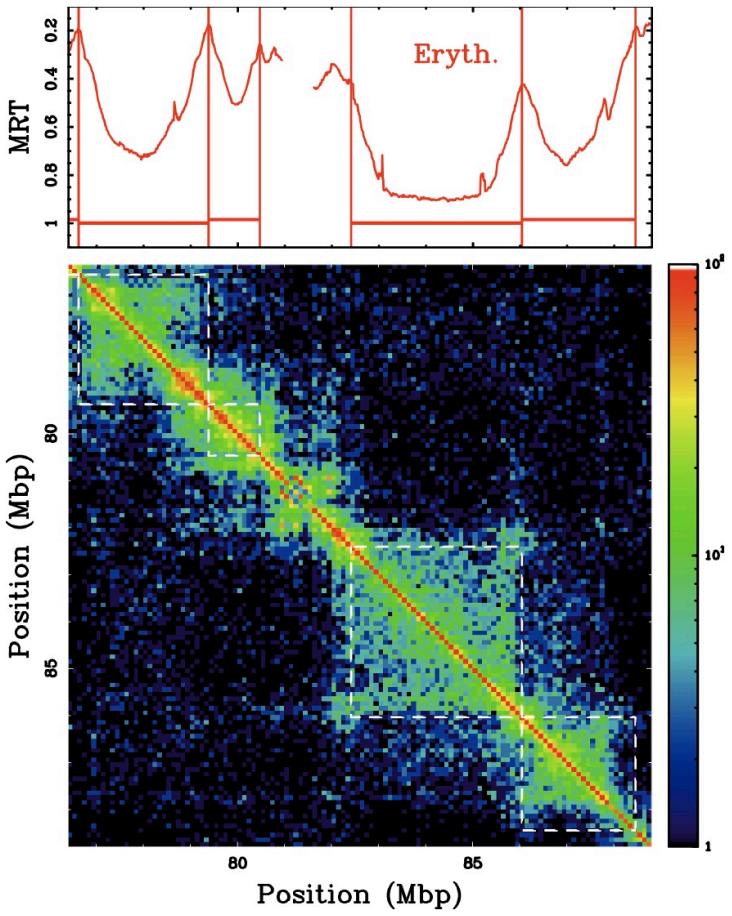
Sequencing projects result in 4 letter texts :

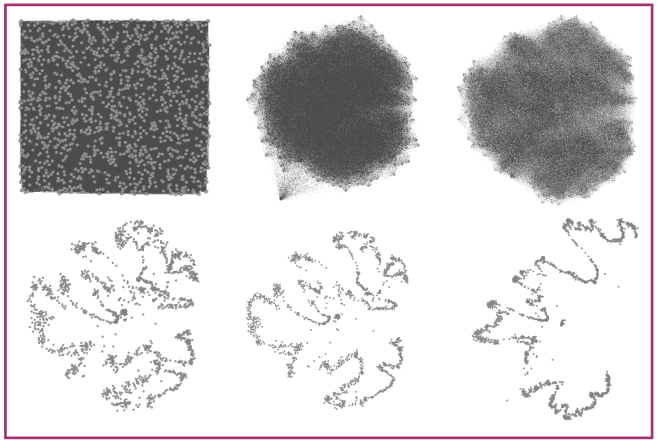
```
gtcagttcctgaggcgggtcgggacccaggcgtgagactggagtctgcc
caggggcccagctgagccagcctcctcgtcagctgcttgggccgccagga
cgccgccgggggtgcgccgcgcttccctggatggggtgcccccactcccc
tcggagccccagggagaccccccgaactcagctcctctcaggggtgccag
ggggaccctcaaactccactccccgcaggttcctggggagacgccccct
gctcgattcccctcagggtcccagggagacccccctaattcagctcctctc
aggggtactgggggacctctcgagctccactcccatcagggtcccaggga
gacccccccaactatgctcaggggtcccagggagatgccagcaccccaact
ccgcttccctggggcccccctcccccttacagctcaacttccctcgagagt
ctggggctggggctccgttcagttcttgagtcccccttccctcggggtgtc
ccggggccgcccacccccacactgtctgtgattccccaaggcgcgggtct
cgggccgcagcctgttccacgttctgctgctcgttcttttctggctcctt
gctttcgaaggagagaaggaggccttcgtttccagtctttttgccttttc
taatggagccctgcttttccttccgtgtcccttcaggctacttctgccag
gtttctatttttcattctttattatgacttcgcccaaaatattcttgact
tctattgagaaggattcgggggtctatttcttattcggaggcgtgtgctt
aagttccaaacagatgaggattttccagttaatccttctggggtgactta
ttgcttaatgccaccatagccagaaaatggactctcagtgtccgaaactg
cattcggctctgaagtgtctgtccttgtcacctcttgcaatgtttcgcgg
cgggaagcctgcactcgccgacgctgacgtaactgtttctgtctttcagg
tctacagcctcctgtgggtgggcgatattgacatatactttatttctata
tatgttatgaactcaatatttcttgcagcgggtctgctgataataagata
tgcctactctgcgagtctggaagccatcttaagcttaccctgtatgtgcc
ccatgcatctcttccgttacacggctcctgagttgacacctgtgtgataa
actggtaatagcaagtaaactgttttcttgtgctctgtaagctgctctag
caaattatctaggaggaggtggtcttggaaacccctgatttataagcggg
cagtcagcagtacacgtggcccagaatcgtgattggcatttgaagtgggg
gcagtagggtgggactgagcccttcacctgtggggtctgccctgctcaag
gcagtgtcagaattgaagtgaaatgttggacggtcggtgtccagagagtt
ggagaactggtttgtgtgtaaaaactnacatatttagggtcagaagtatg
...
```

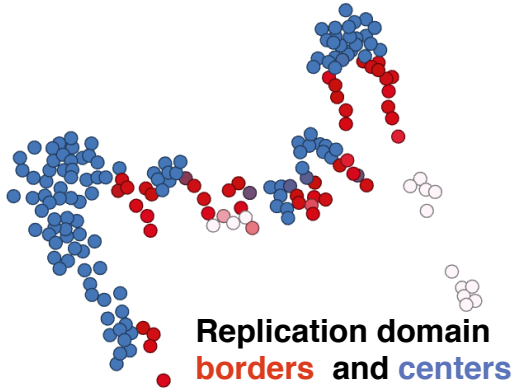
Replication domain borders are giant hubs in the chromatin conformation graph in K562

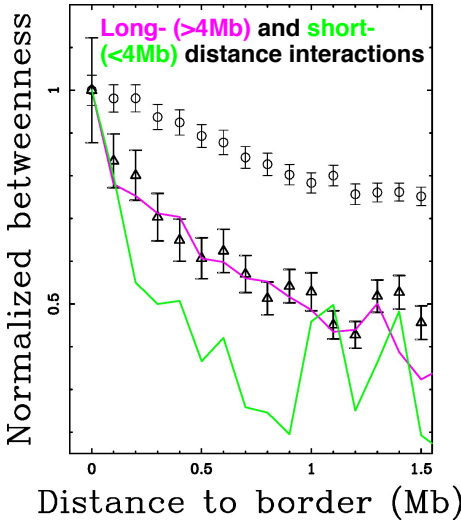
Boulos, Physical Review Letters (2013)

Co-localisation matrix as the adjacency matrix of the **chromatin interaction graph**



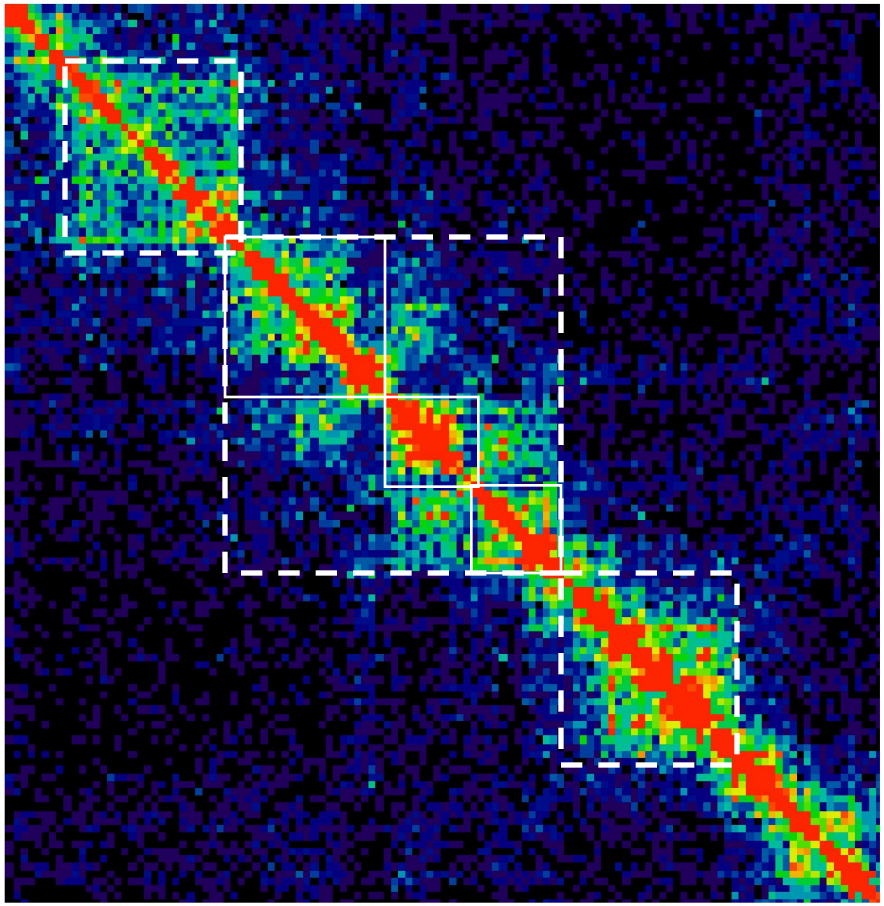


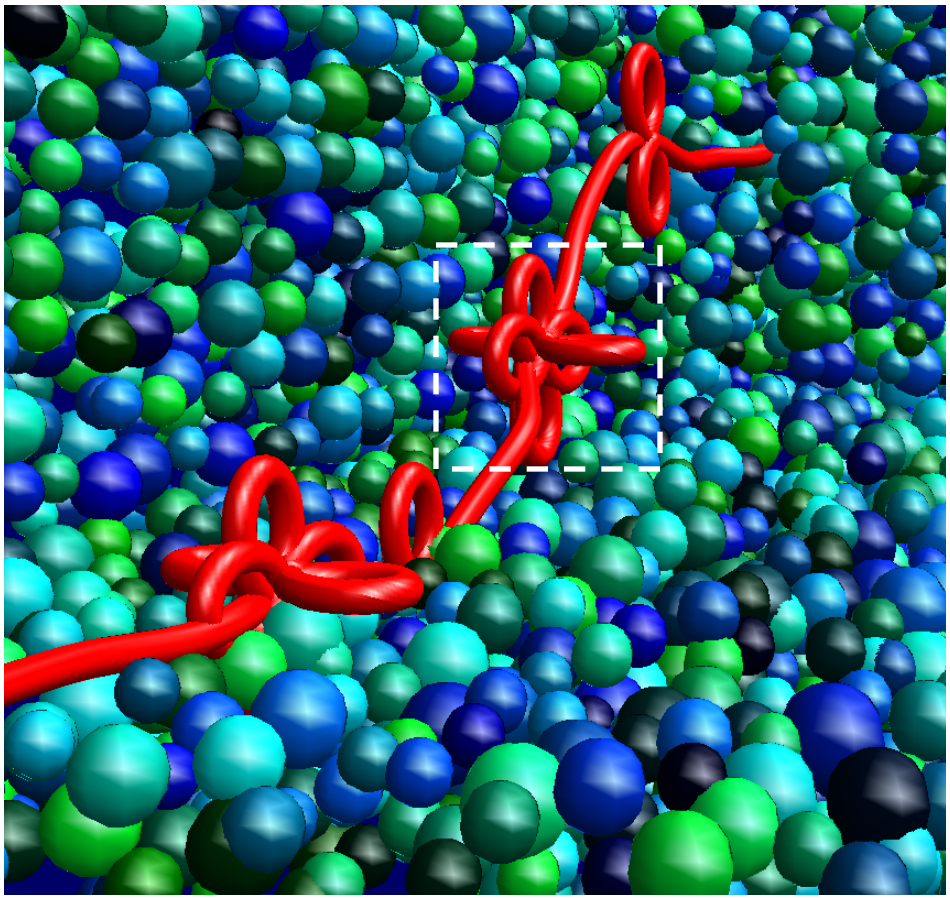




Intra-chromosomal betweenness centrality along replication domains

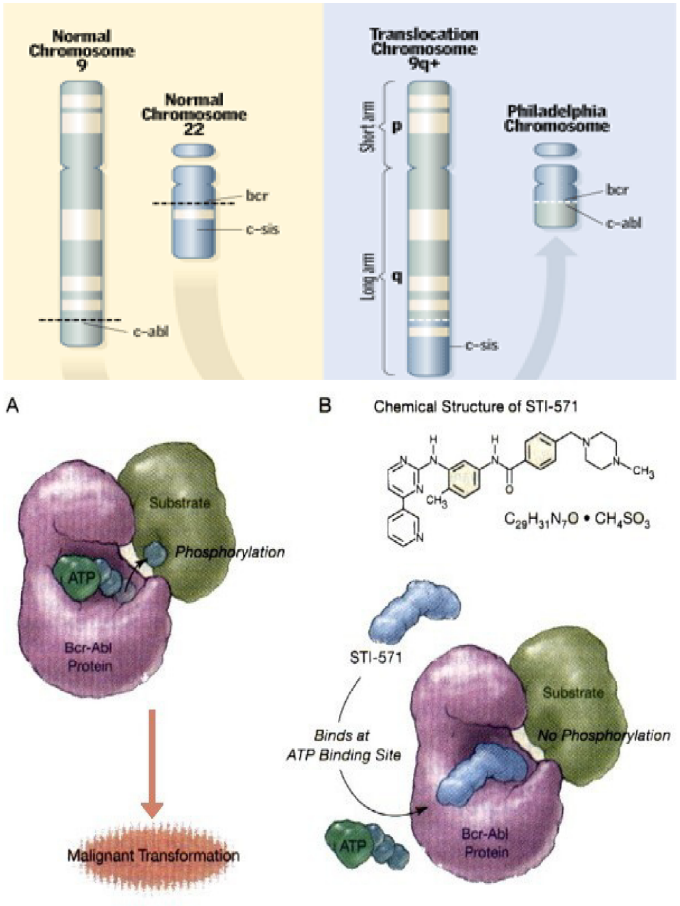
Master replication origins at the heart of parallel genome functioning ?

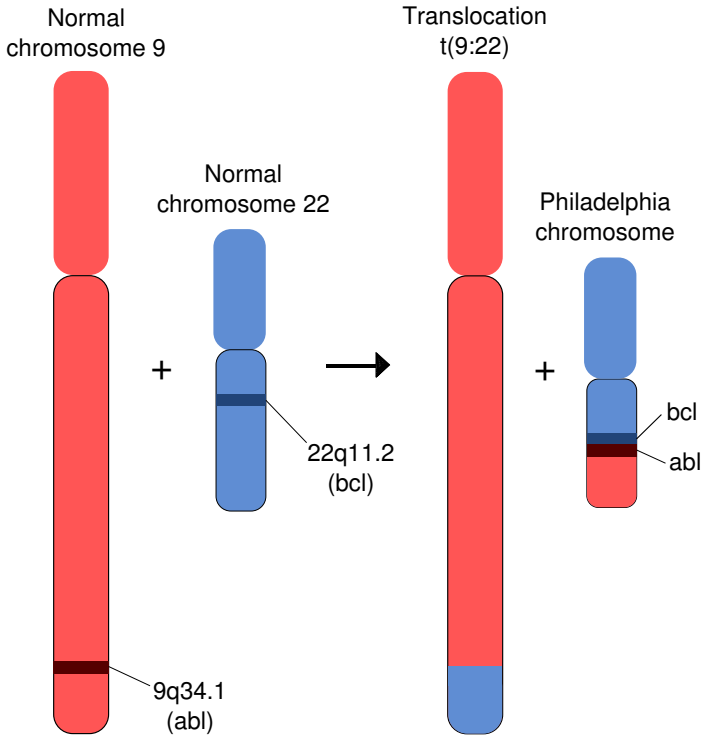




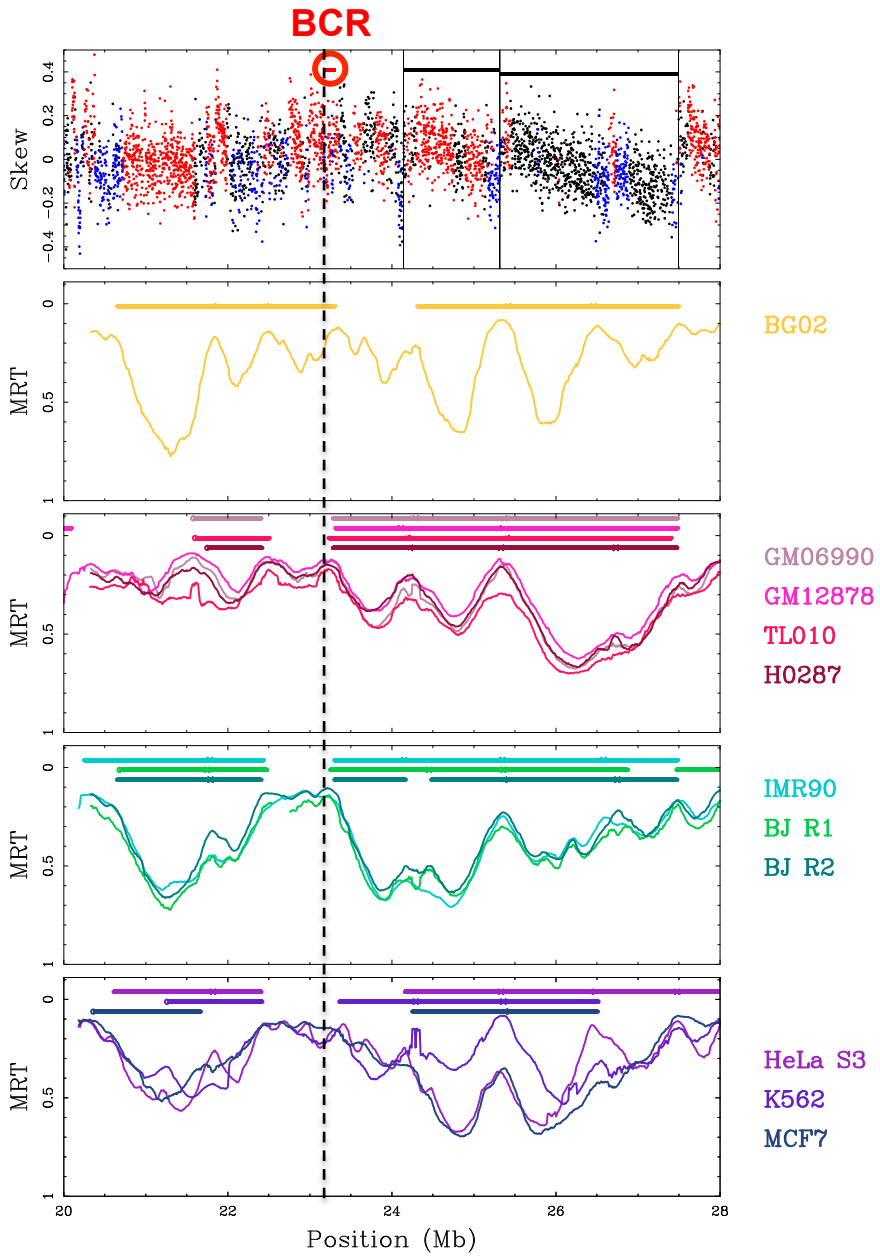
The Chronic Myelogenous Leukemia
a disease of hematopoietic cells

Chronic Myeloid Leukemia (CML) : targeted therapy model

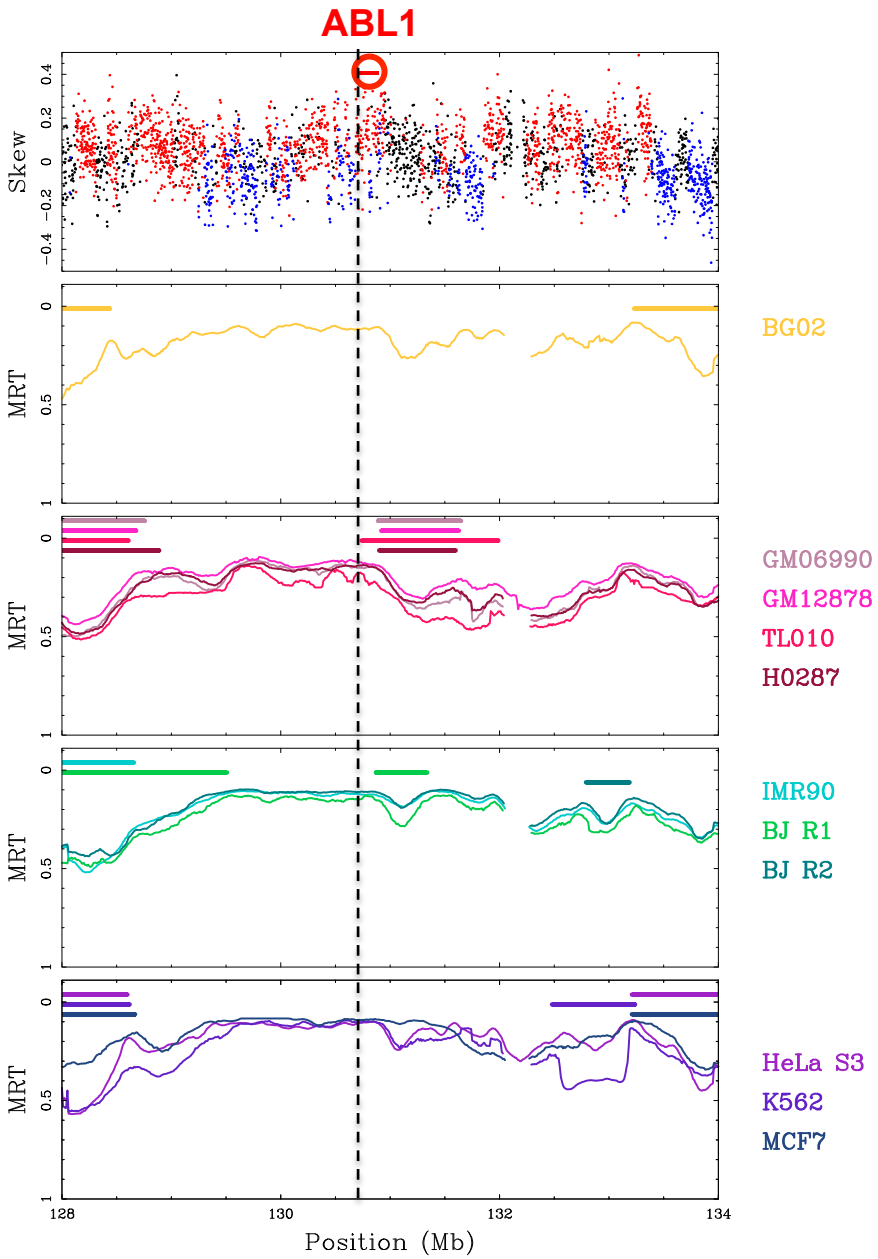




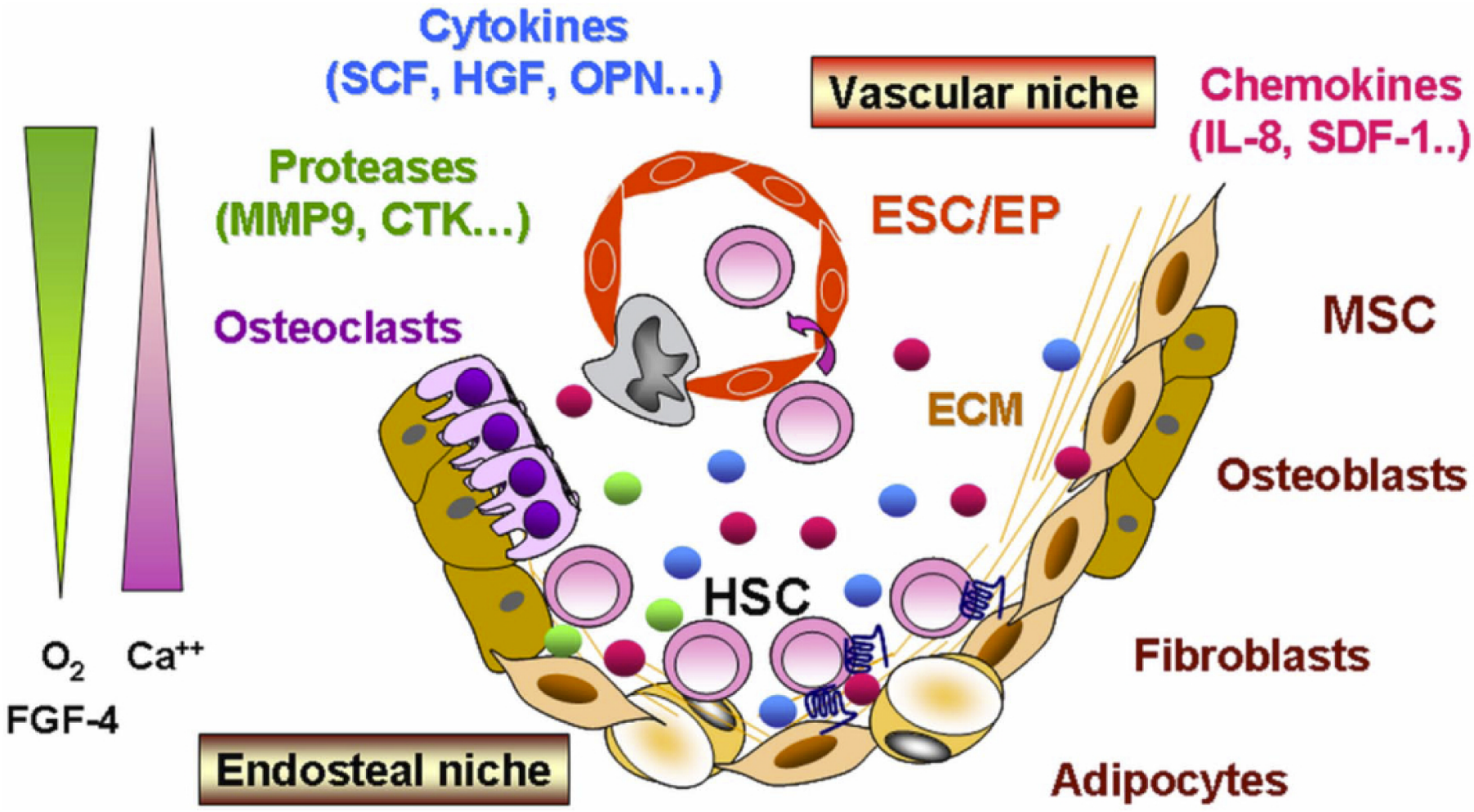
Replication domain organization at the BCR locus in chromosome 22



Replication domain organization at the ABL1 locus in chromosome 9

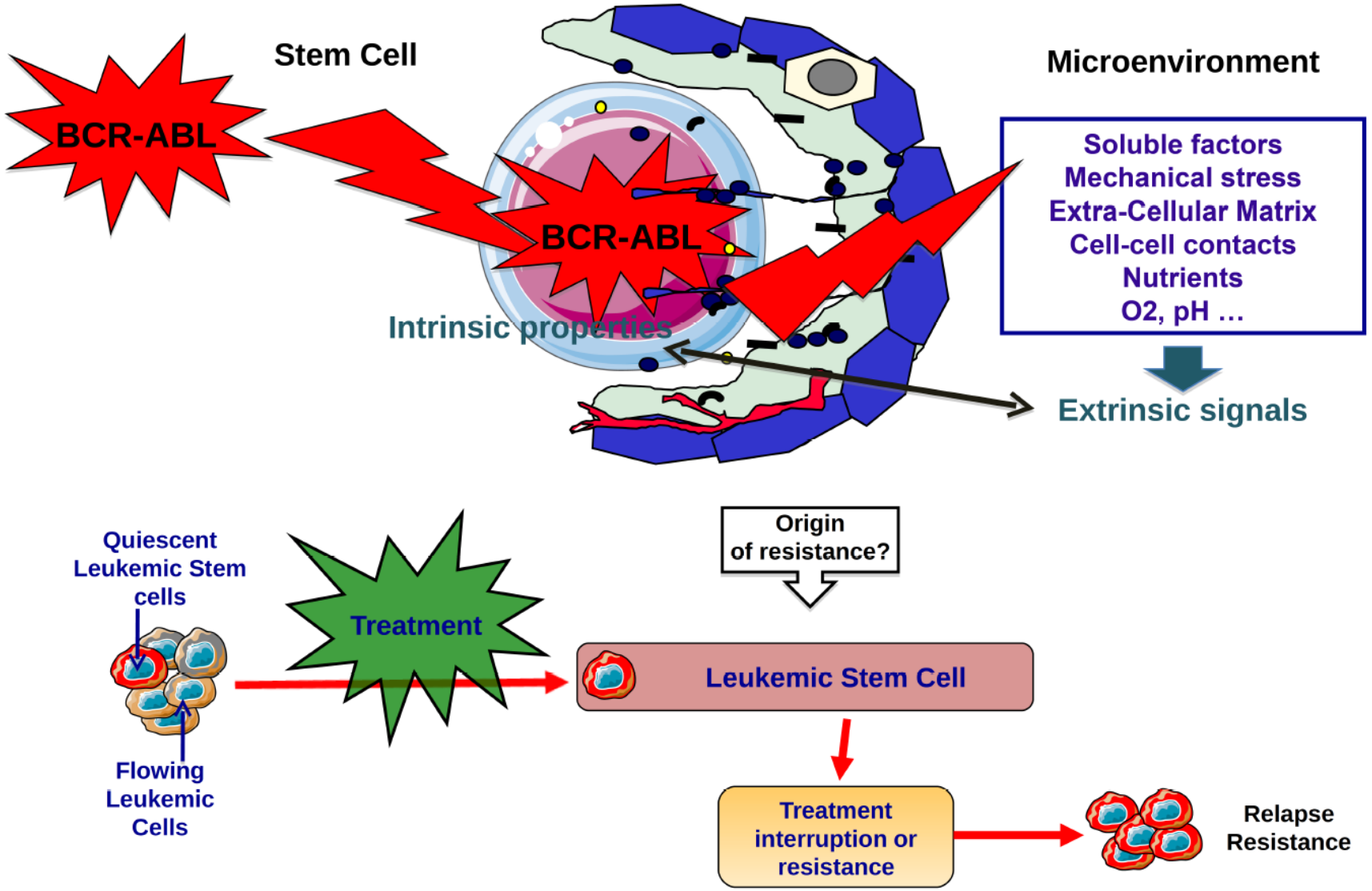


Mesenchymal Stem Cells and their Niche



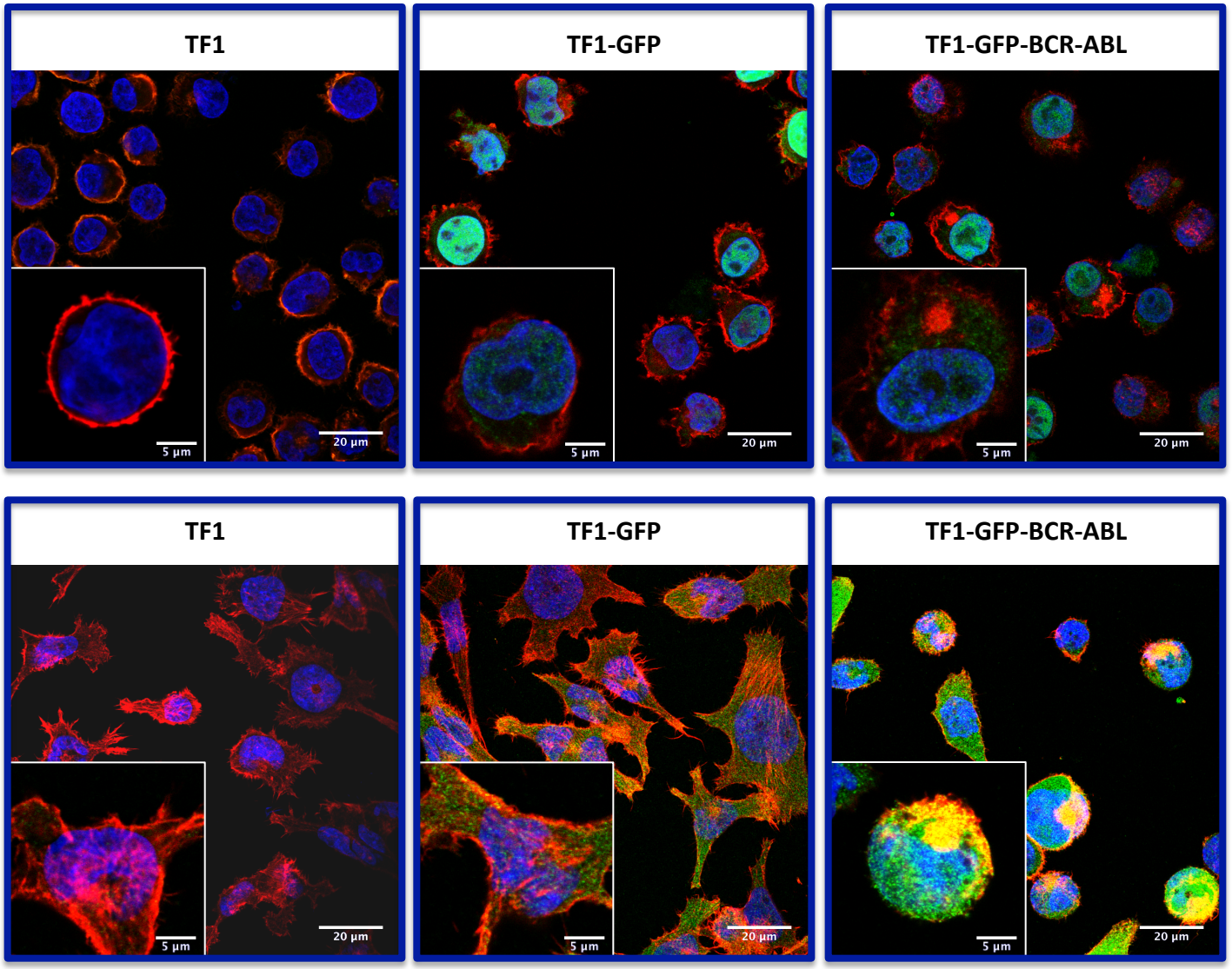
MESENCHYMAL STEM CELLS and their NICHE

Dysregulation of the Niche of Hematopoietic Stem Cells in Chronic Myelogenous Leukemia

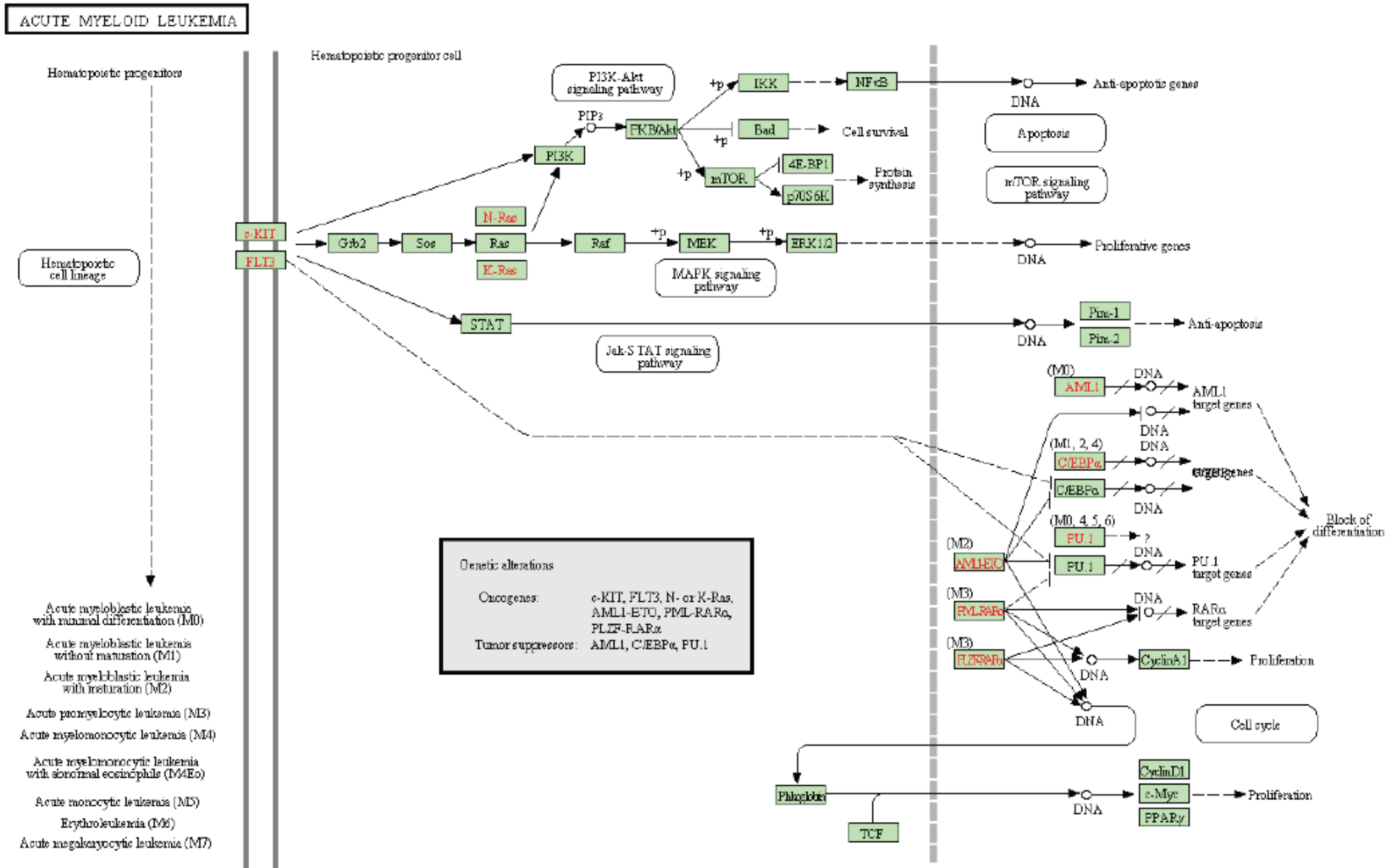


BCR-ABL alters F-actin distribution in Chronic Myelogenous Leukemia cells

J. R. McWhirter and J. Y. Wang
The EMBO Journal 1993.



Chronic Myelogenous Leukemia signaling pathway



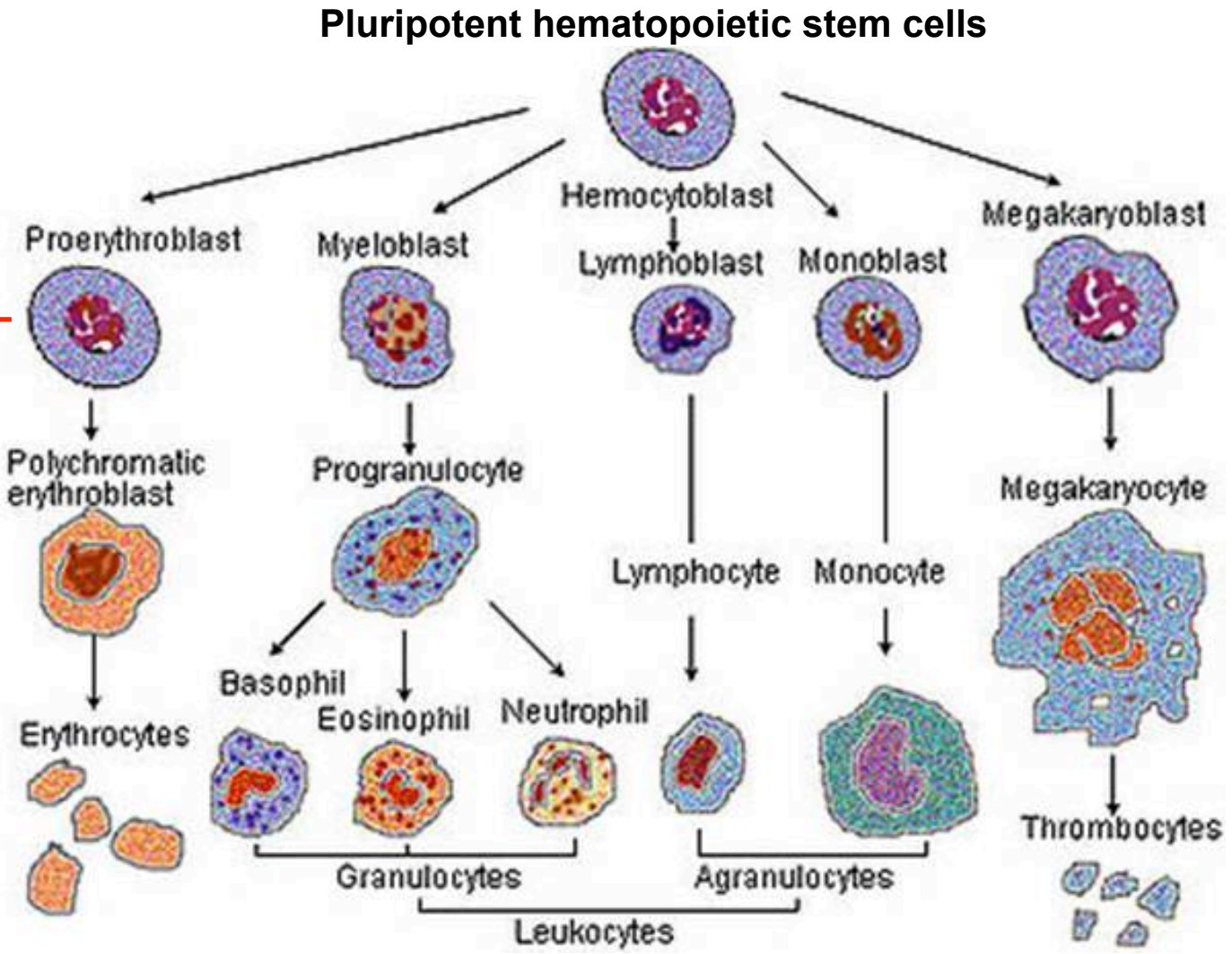
Hematopoiesis from pluripotent stem cells

TF1

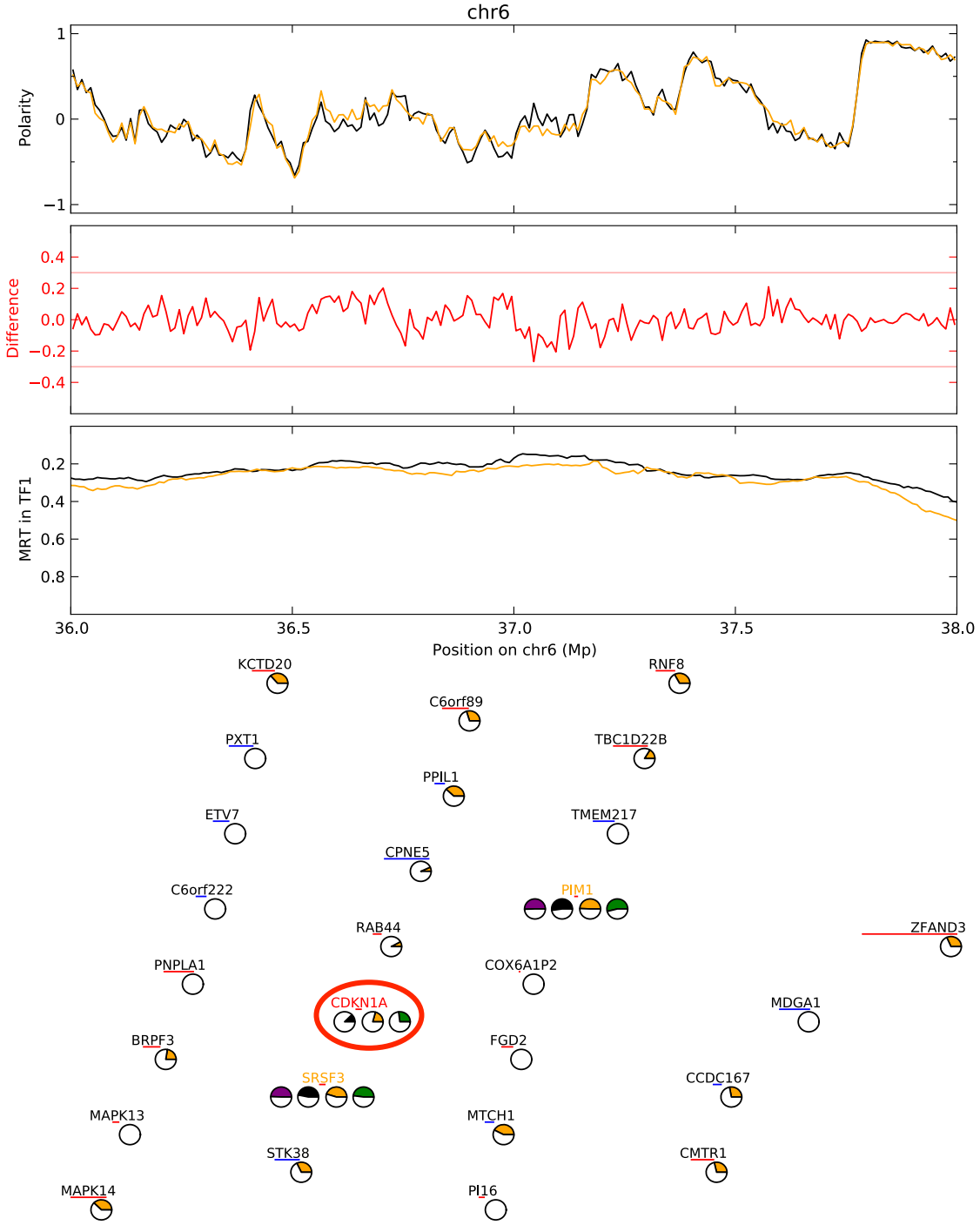
Erythroleukemic cell line

TF1-GFP
TF1-BCR-ABL (1 month)
TF1-BCR-ABL (6 months)

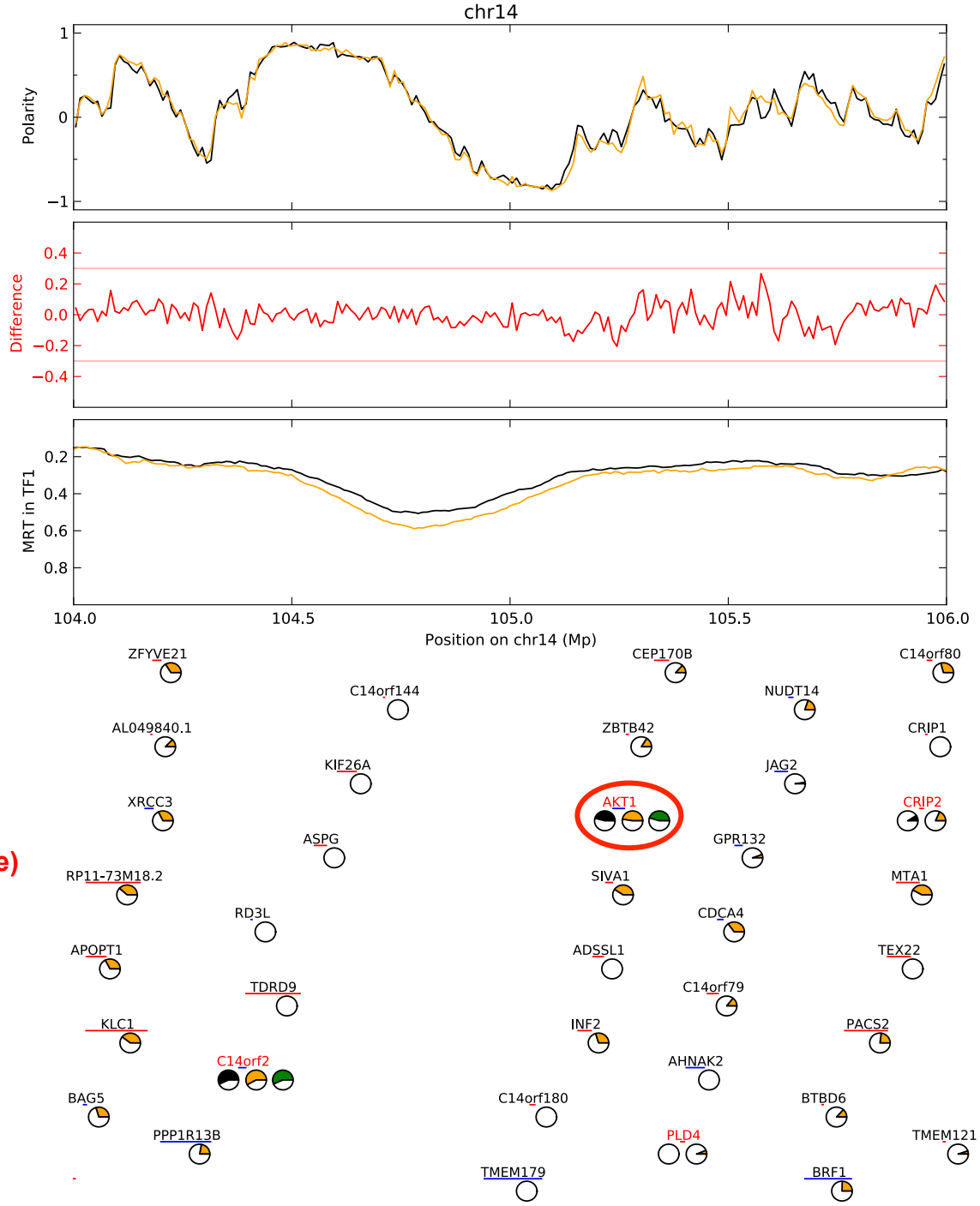
Laperrousaz, Blood (2013)

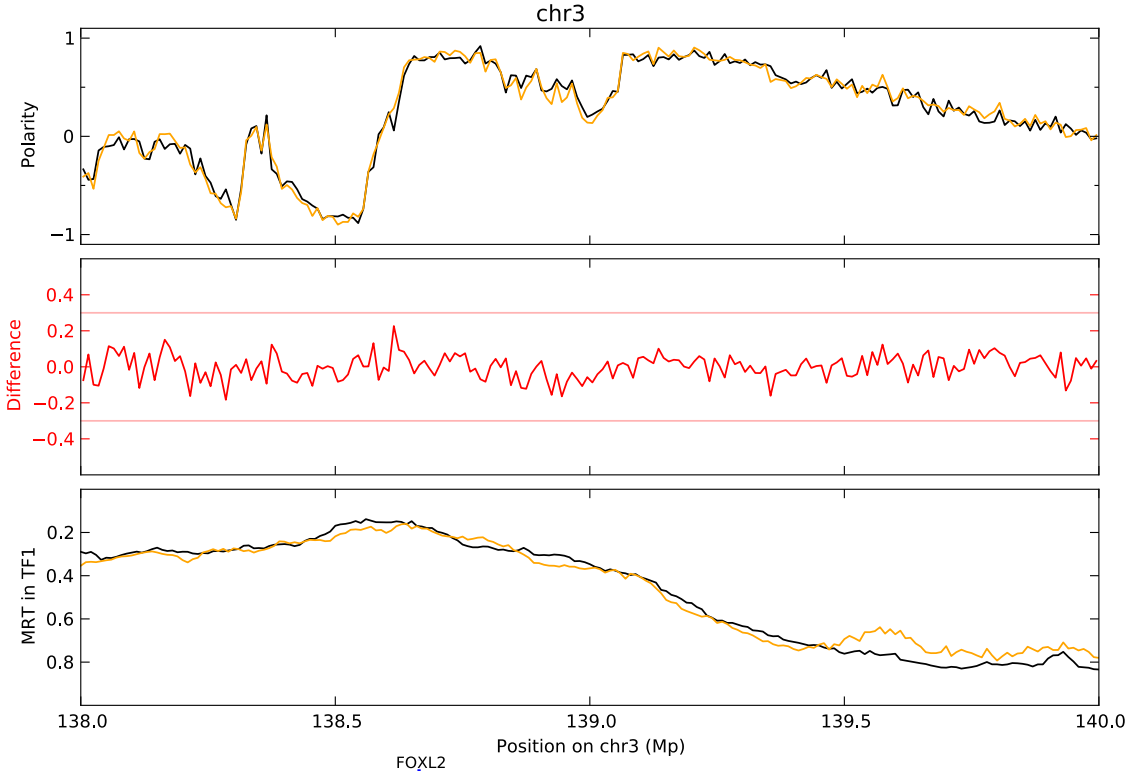


Cyclin-dependent Kinase Inhibitor 1A

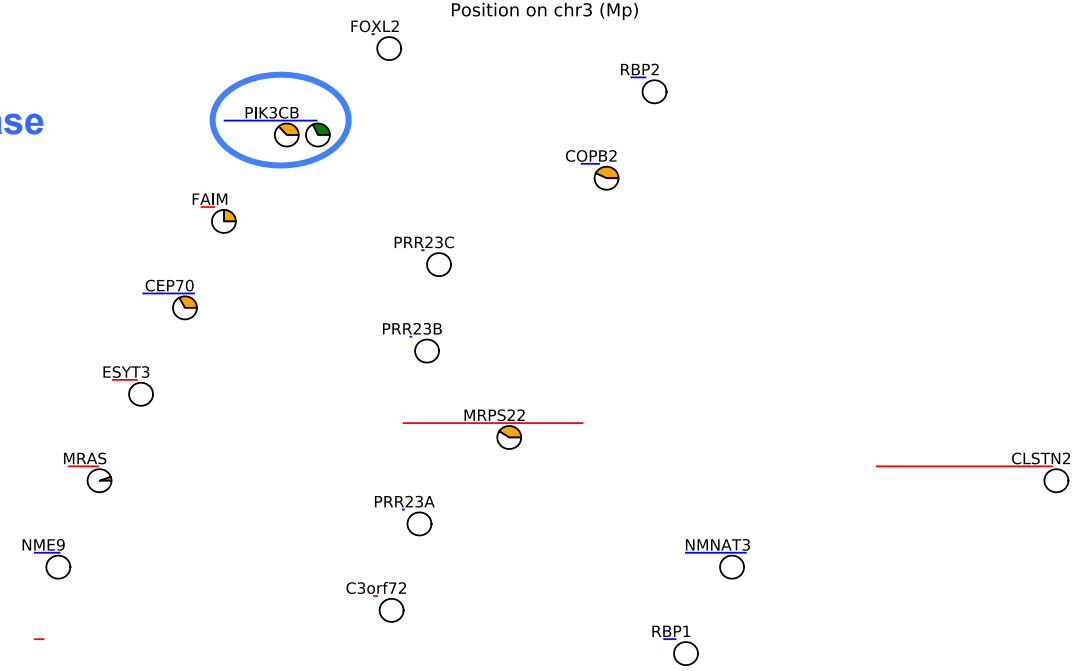


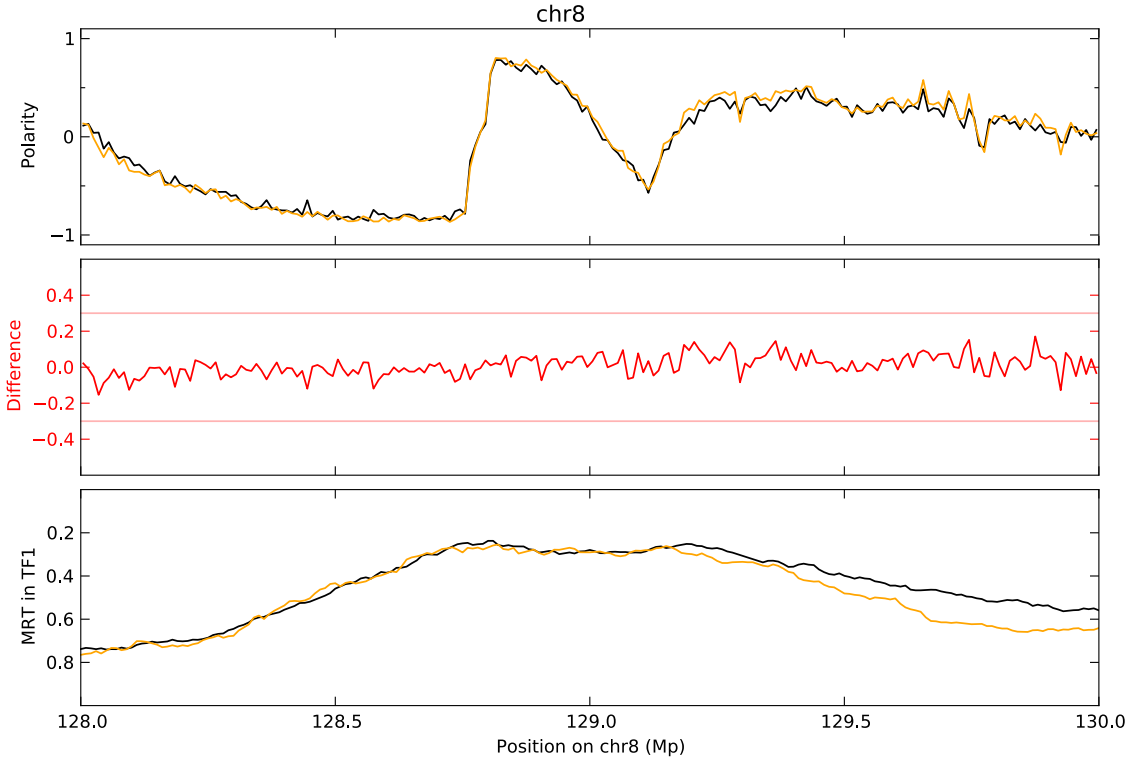
V-Akt Murine Thymoma Viral Oncogene Homolog 1
(serine/threonine-protein kinase)



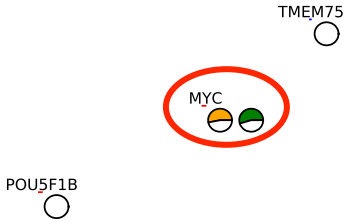


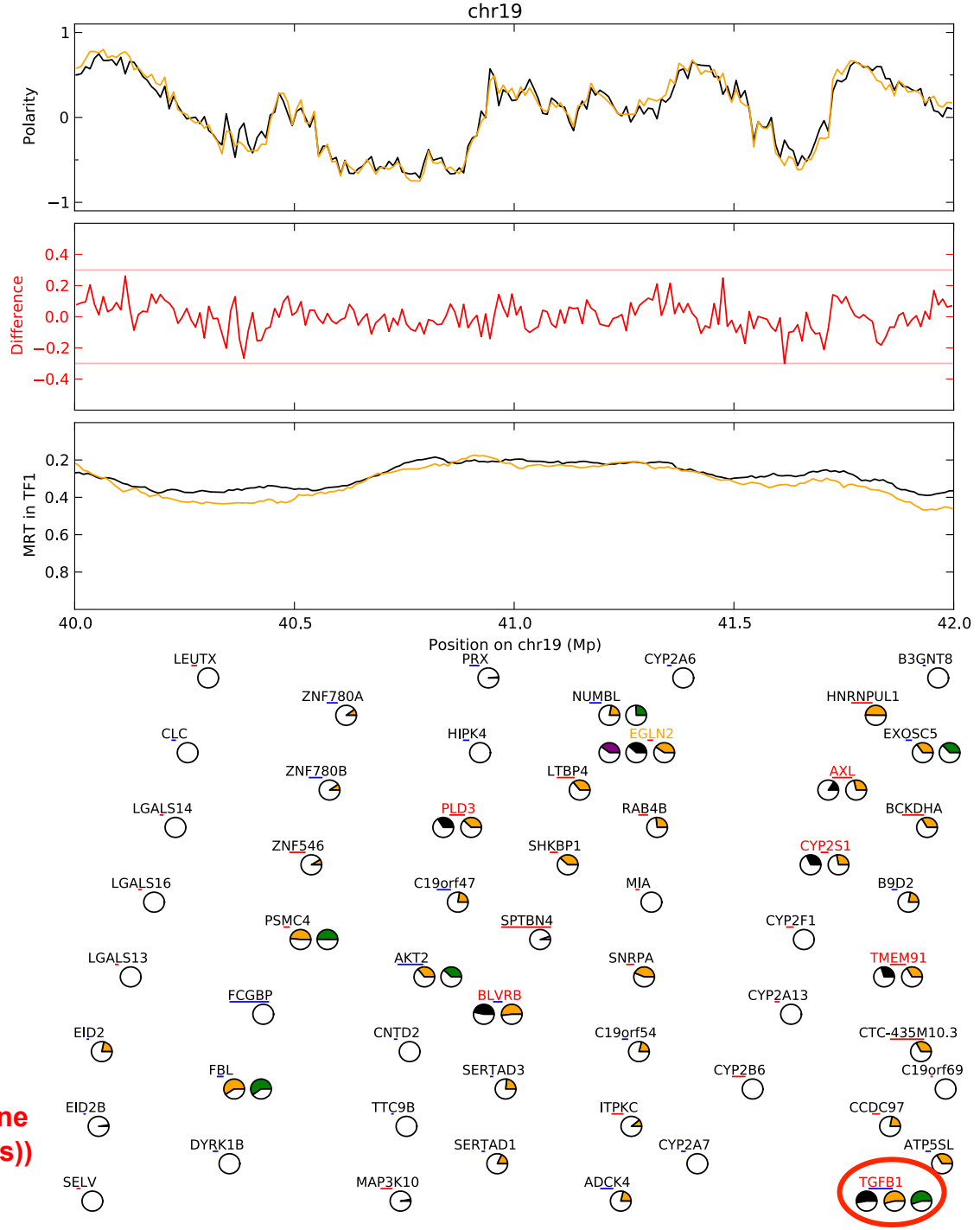
catalytic subunit of phosphoinositide 3-kinase (PI3K)



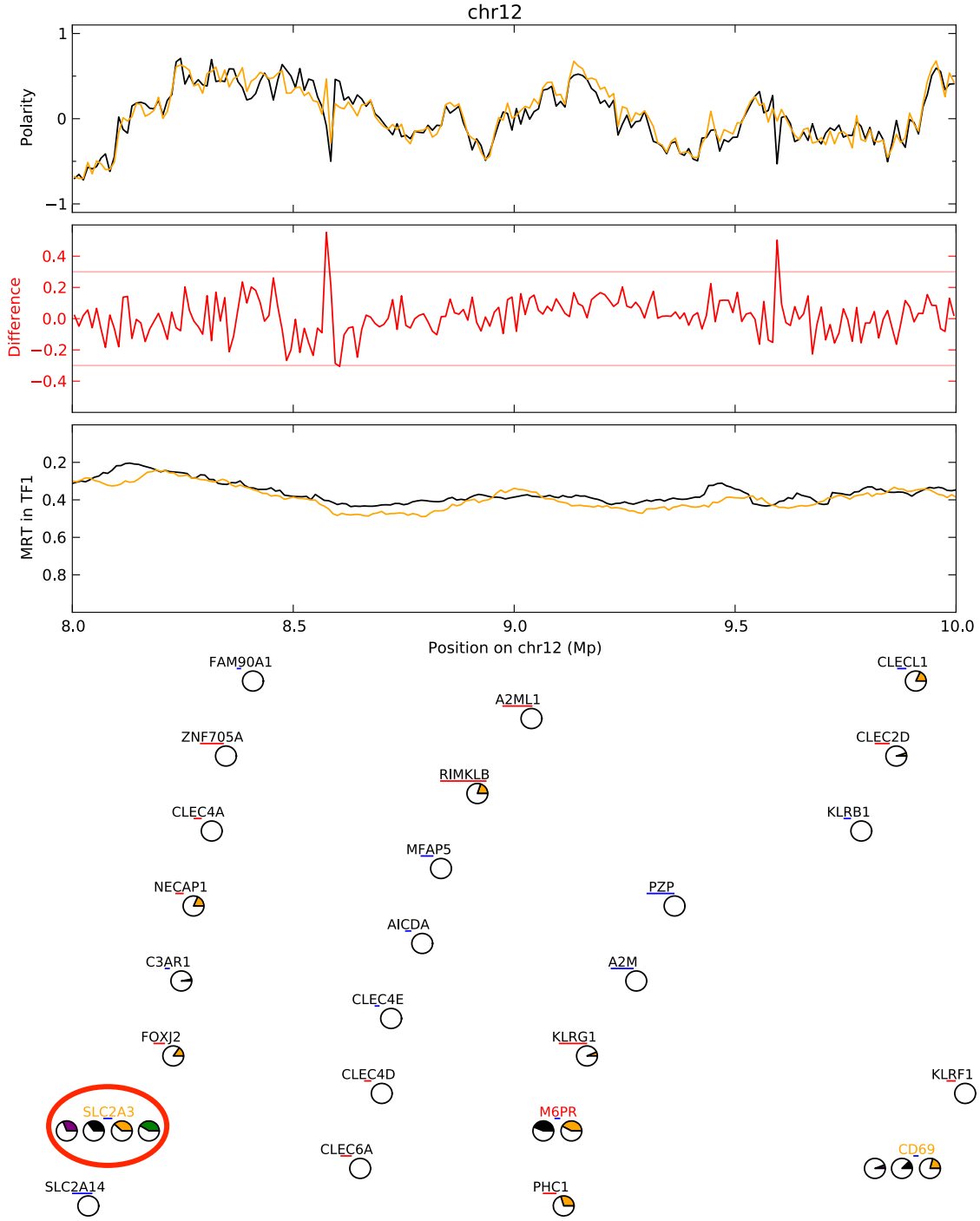


V-Myc Avian Myelocytomatosis Viral Oncogene Homolog
(DNA binding protein)



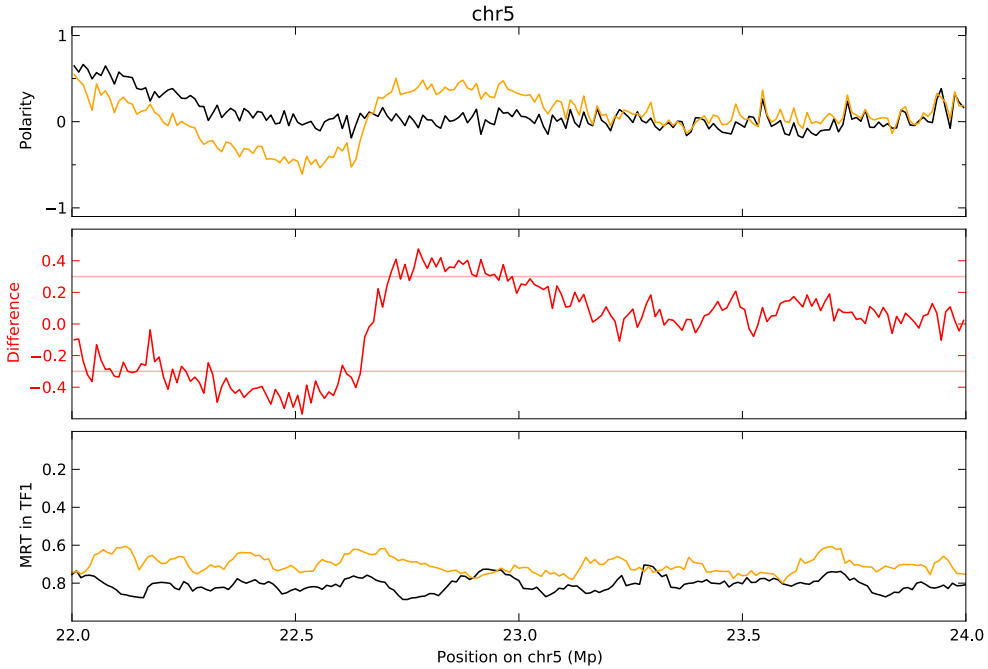


Transforming Growth Factor, Beta 1
(cytokine, ligands include: Bone morphogenetic proteins (BMPs))



Solute Carrier Family 2 (Facilitated Glucose Transporter), Member 3
(glucose transmembrane transporter activity)

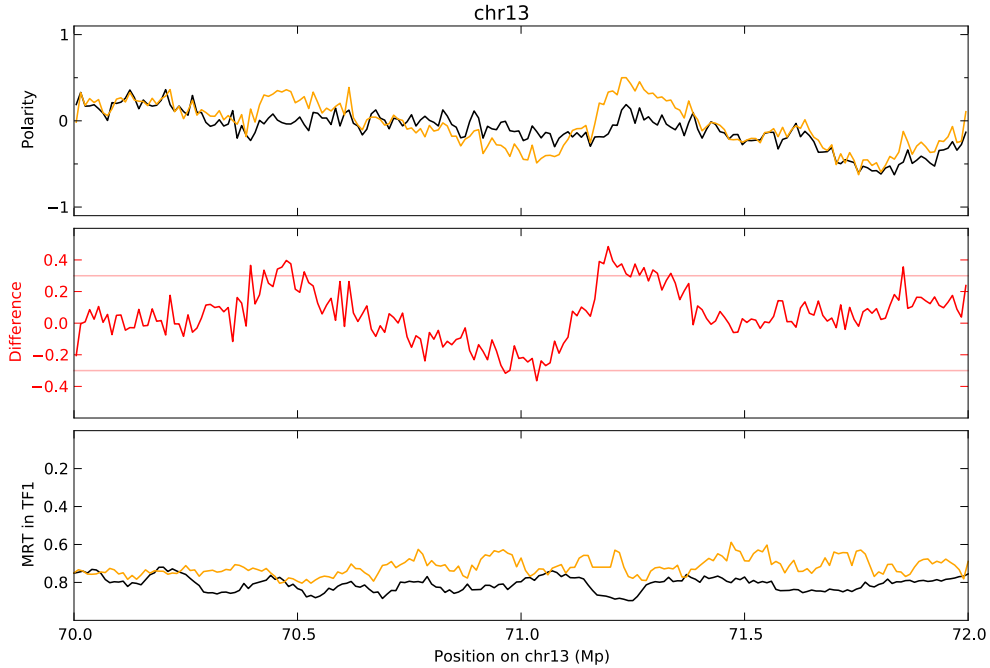
Activated replication origins are in late timing regions

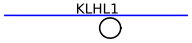


PRDM9

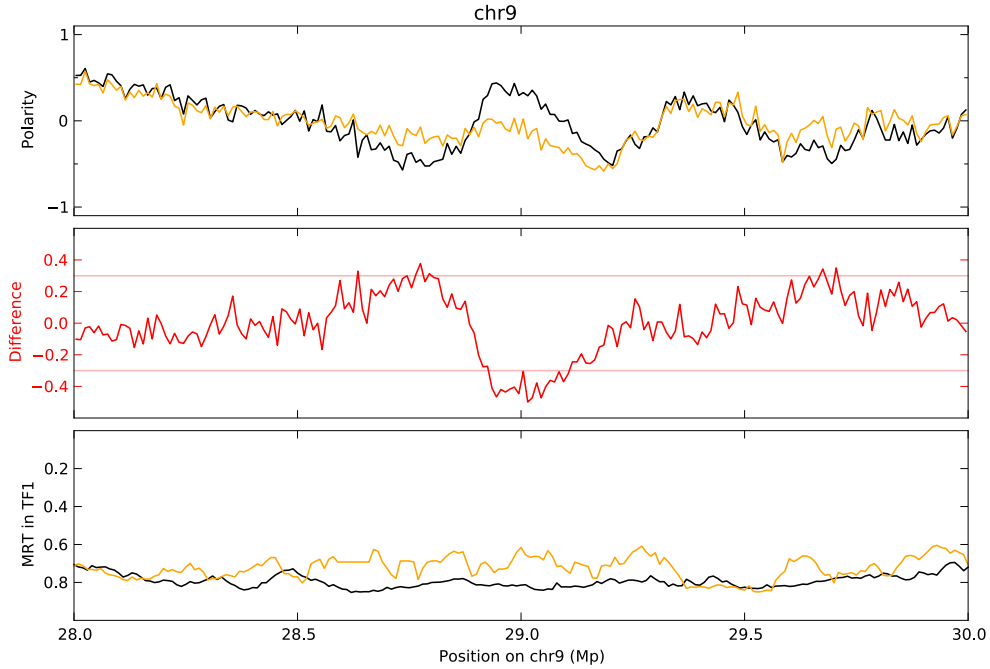
CDH12

Activated replication origins are in late timing regions

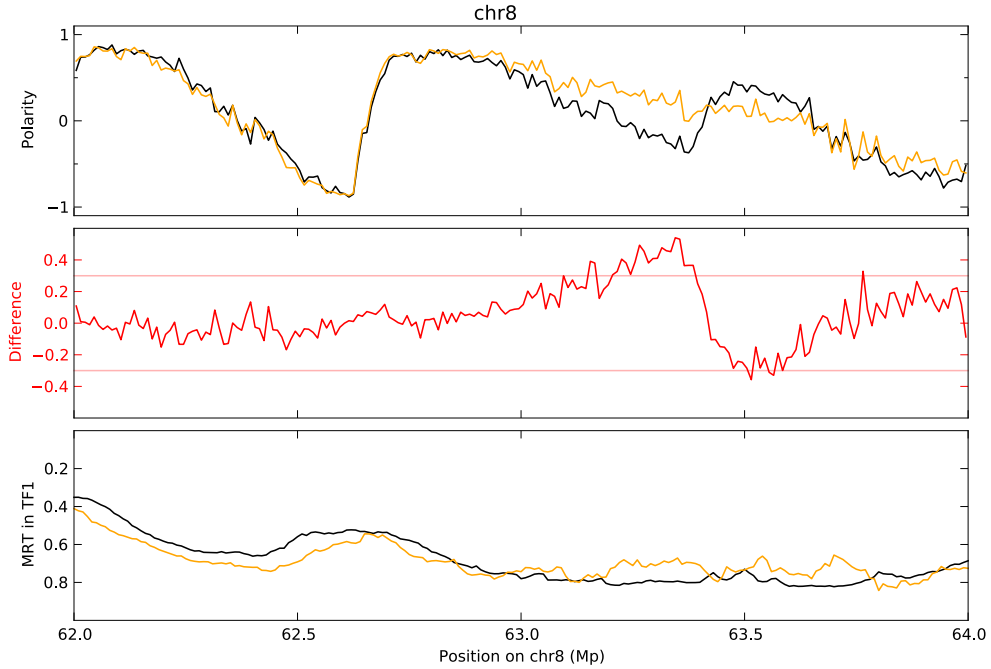




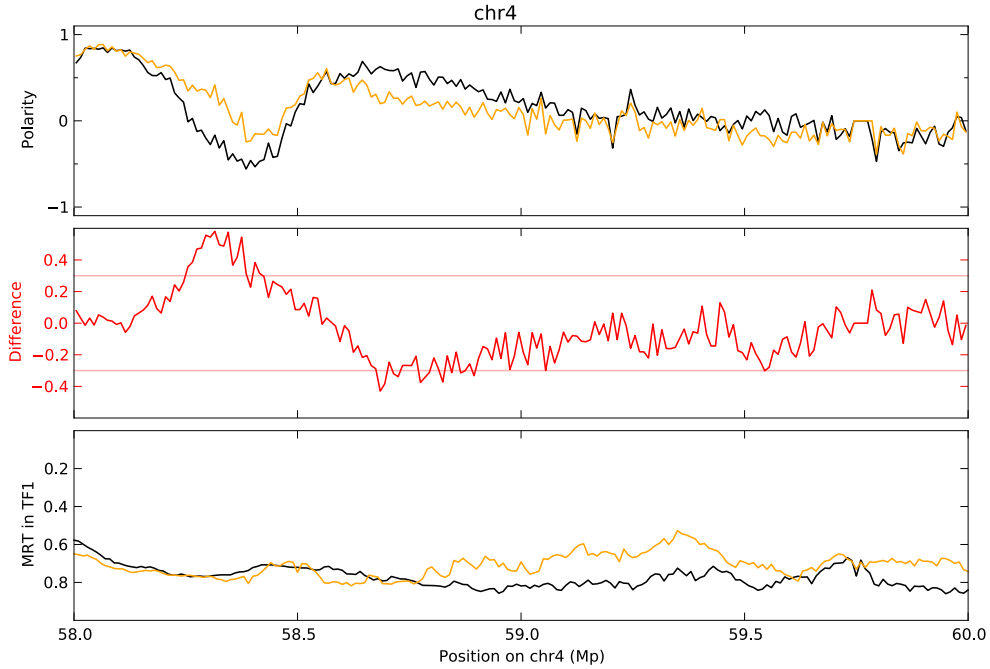
Silenced replication origins are in late timing regions



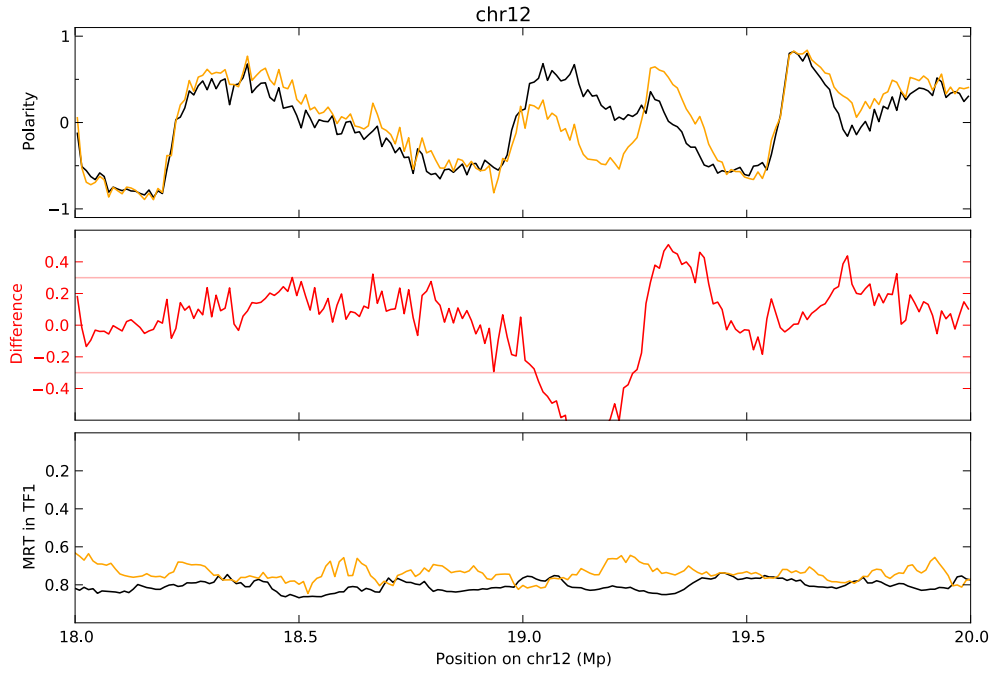
Silenced replication origins are in late timing regions



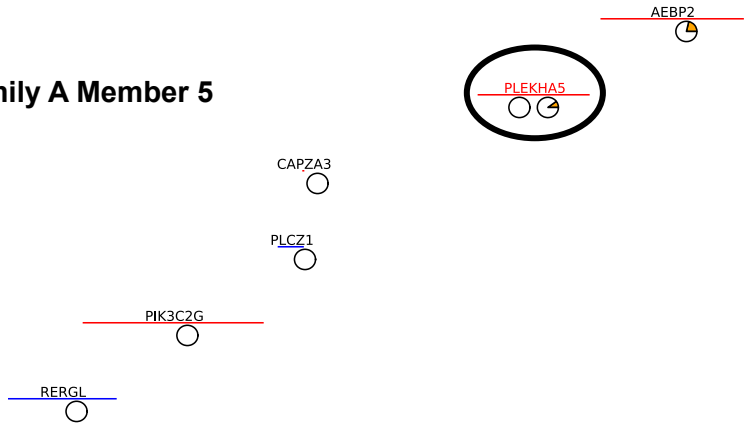
Weakened replication origins are in late timing regions



Changes in replication origin activity are in late timing regions



Pleckstrin Homology Domain Containing, Family A Member 5



Changes in origin activity are in late replicating gene deserts

New origins: 17

12 late - 4 mid S - 1 midS to late

Enhanced activity: 20

16 late - 2 mid S – 2 early

Weakened activity: 32

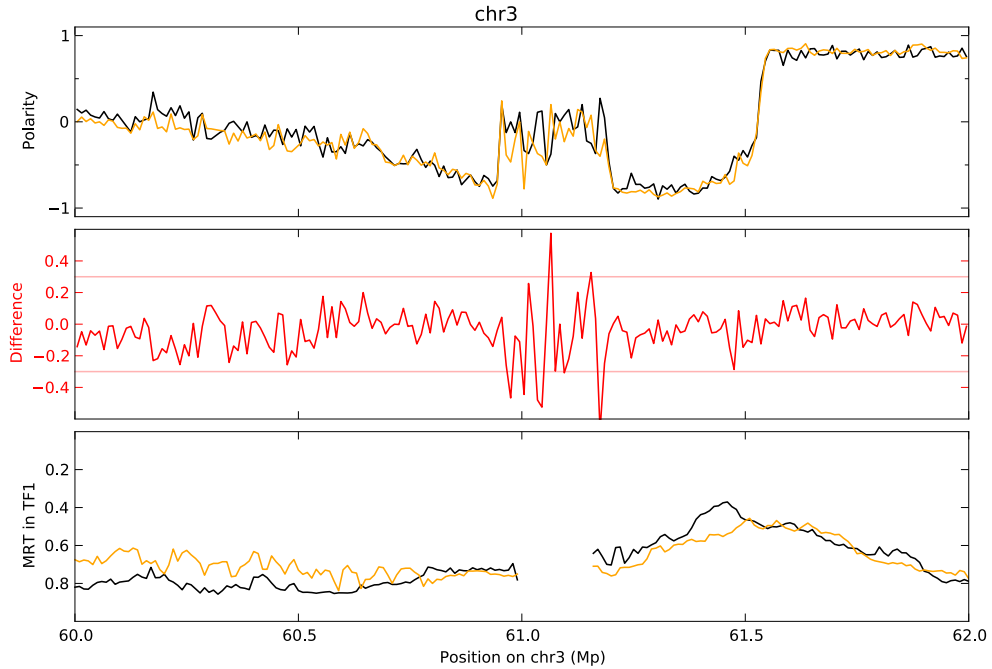
28 late - 1 mid S – 3 midS to late

Silenced origins: 23

23 late

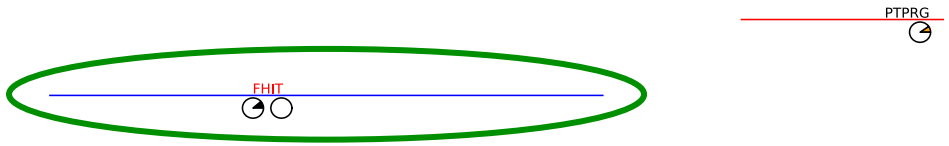
Active fragile sites are in late timing regions

Letessier, Nature (2011)



FHIT (fragile histidine triad): Tumor supressor gene

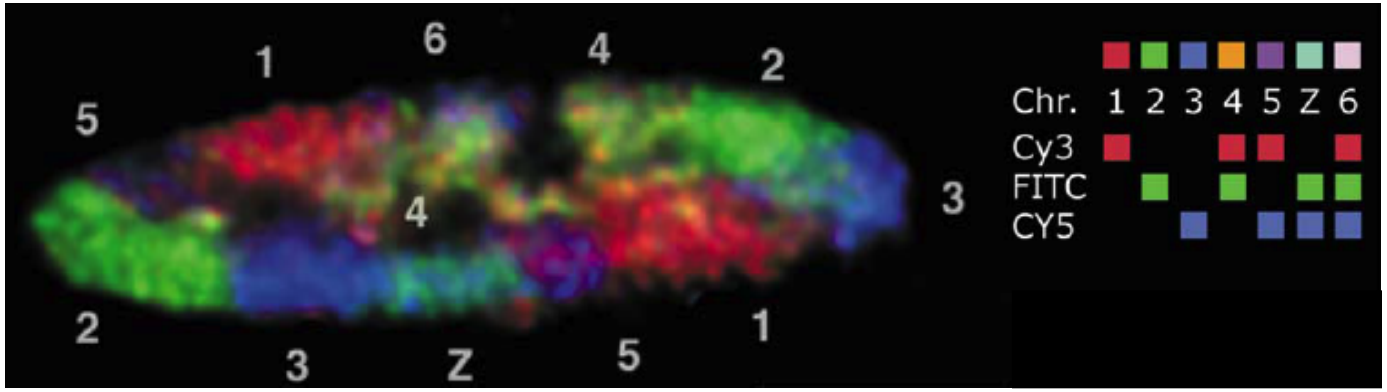
Part of the FRA3B most active common fragile site in lymphocytes



Organisation of eukaryotic nucleus

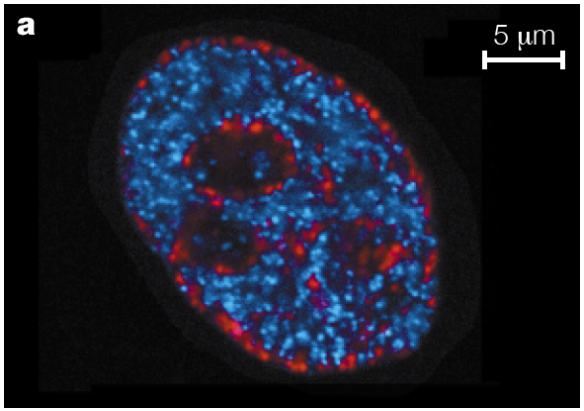
Cremer, Nat Rev Genet 2, 2001 – Cook, Science 284, 1999 - Pombo, EMBO J 18, 1999

Chromosome territories



Chicken fibroblast nucleus where the 7 macro-chromosome are revealed by FISH

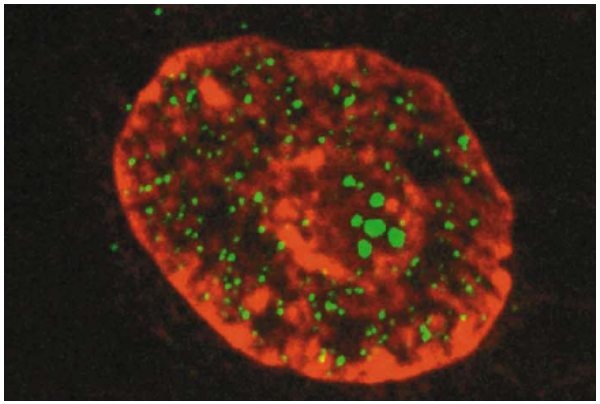
Regionalisation of nuclear functions



Human neuroblastoma

Early replication *(Cy3)*

Mid- to late- replication *(Cy5)*



HeLa Cell

RNA (5-bromo-UTP / FITC)

DNA (TOTO-3)

Origins of Replication and Cancer

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

Laboratoire de Physique de l'ENS-Lyon

ENS de Lyon

Benjamin Audit
Antoine Baker
Rasha Boulos
Guenola Drillon
Hanna Julienne
Lamia Zaghloul

ENS de Lyon / CRCL

Françoise Argoul
Bastien Laperrousaz
Véronique Maguer-Satta
Franck Nicolini

ENS, Paris

Guillaume Guilbaud
Olivier Hyrien
Malik Kahli
Nataliya Petryk
Aurélien Rappailles

CGM, Gif-sur-Yvette

Yves d'Aubenton-Carafa
Chun-Long Chen
Claude Thermes

CEA, Saclay

Arach Goldar

Disentangling replication and transcription contributions to skew profiles