

# Impact of replication on the evolution of human genome nucleotide composition

**Chun-Long CHEN**

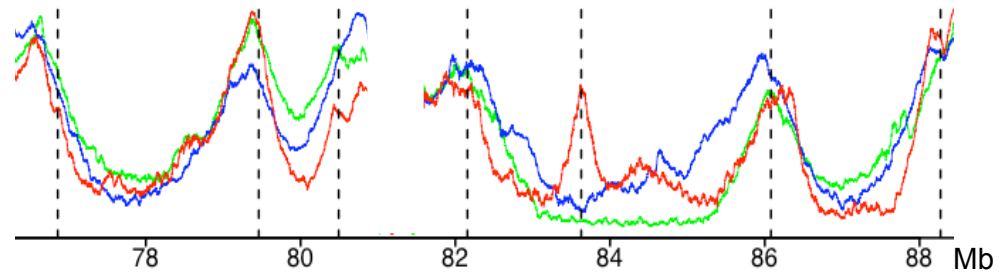
Genome Analysis lab, Centre de Génétique Moléculaire



01/03/13, BioInfo Club, IJM, Paris

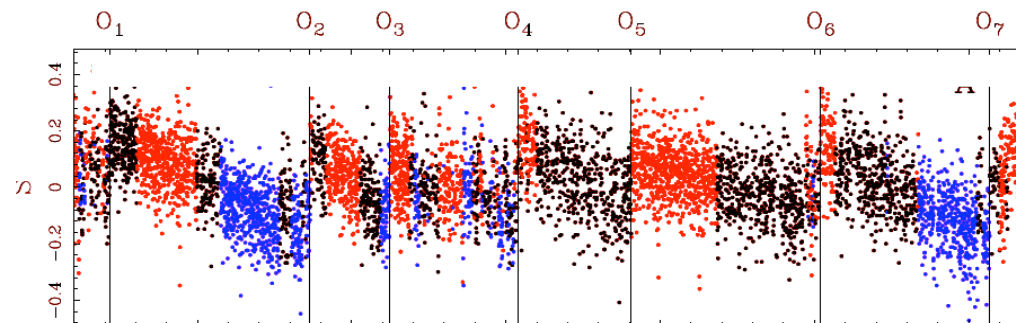
- **Part I: Spatio-temporal replication program of the human genome**

Replication timing



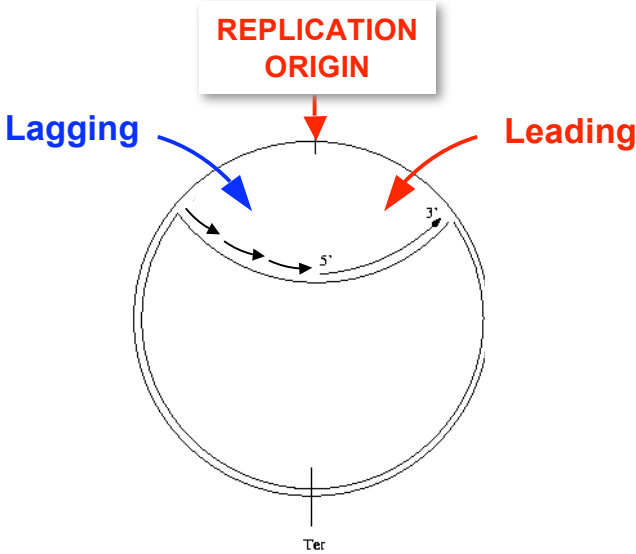
- **Part II: Impact of replication on the evolution and organization of the genome**

Nucleotide compositional skew



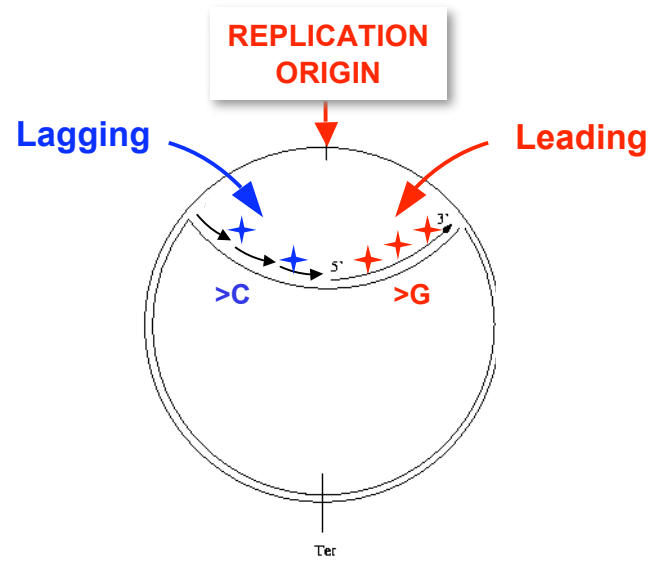
# INTRODUCTION

Eubacteria:



Eubacteria:

Substitution rates  
differ between  
leading and lagging strands



Francino & Ochman *Trends Genet.* 1997  
Frank & Lobry *Gene.* 1999  
Mrazek & Karlin *PNAS.* 1998

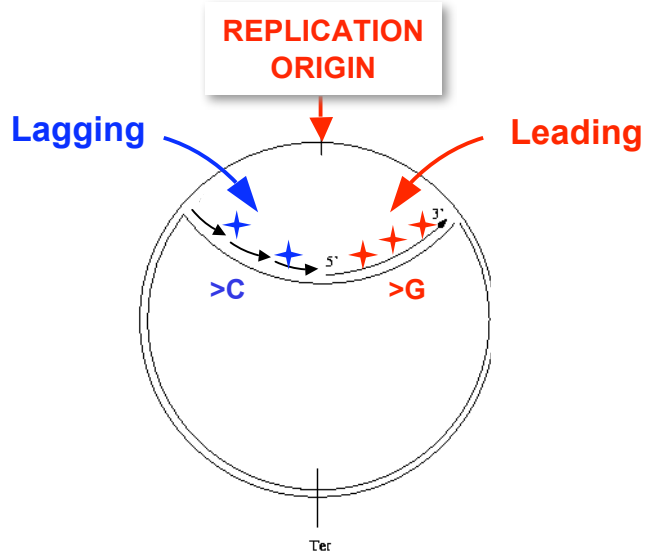
Eubacteria:

Substitution rates  
differ between  
leading and lagging strands

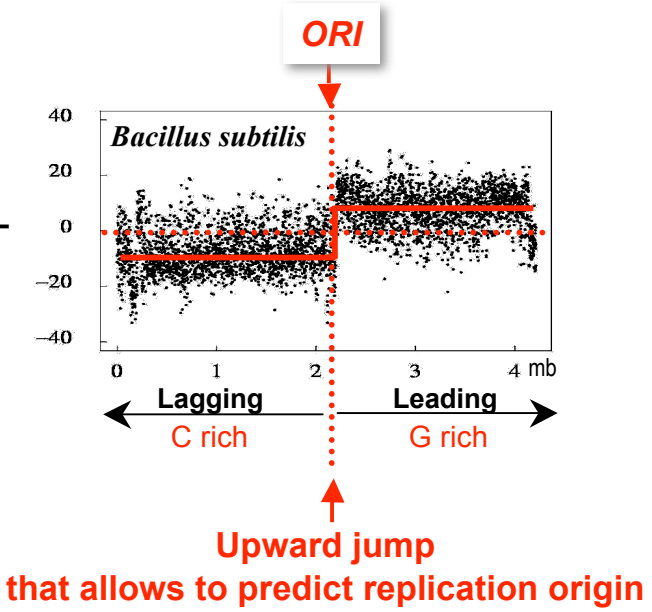


Compositional skew

$$S_{GC} = \frac{n_G - n_C}{n_G + n_C}$$

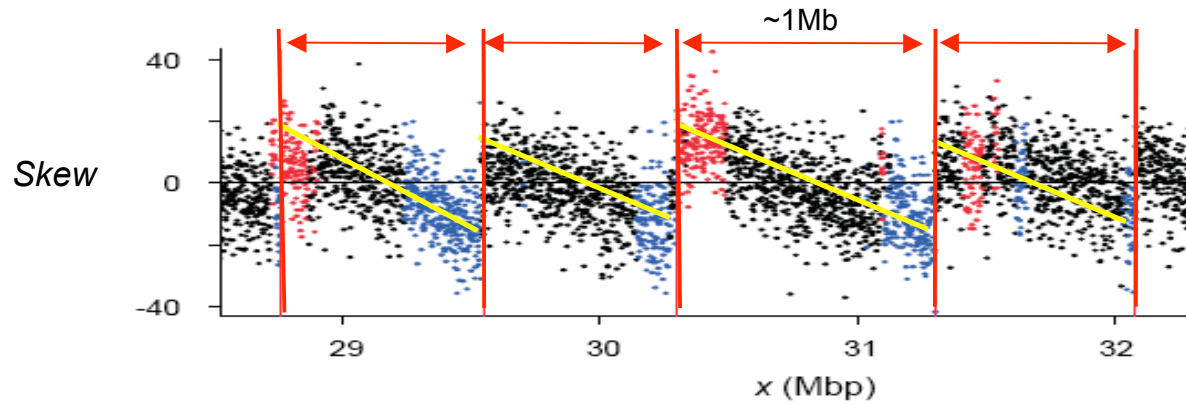


Francino & Ochman *Trends Genet.* 1997  
Frank & Lobry *Gene.* 1999  
Mrazek & Karlin *PNAS.* 1998



Human:

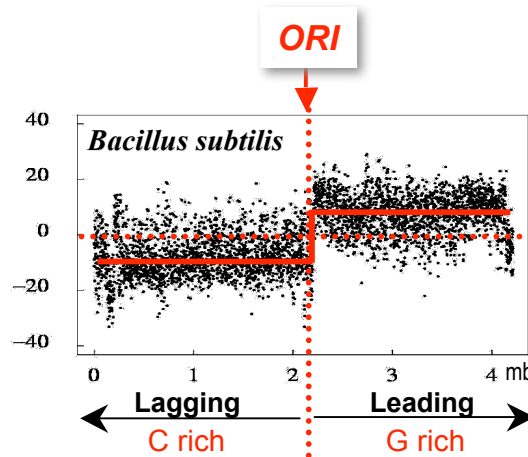
N-domains : > 1/3 of the genome



Touchon et al. *PNAS*. 2005  
Huvel et al. *Genome Res.* 2008

Compositional skew

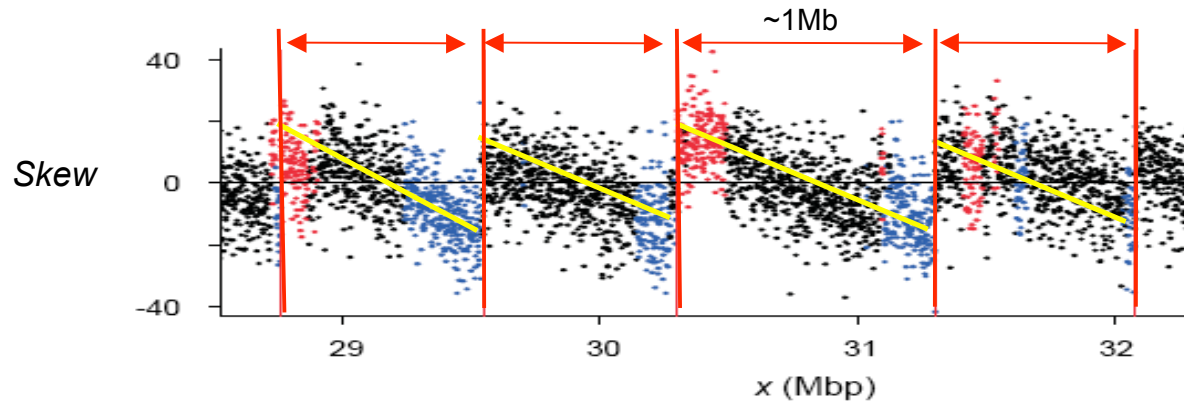
$$S_{GC} = \frac{n_G - n_C}{n_G + n_C}$$



Upward jump  
that allows to predict replication origin

Human:

N-domains : > 1/3 of the genome



Touchon et al. *PNAS*. 2005  
Huvet et al. *Genome Res.* 2008

**Do N-domains result from replication?**

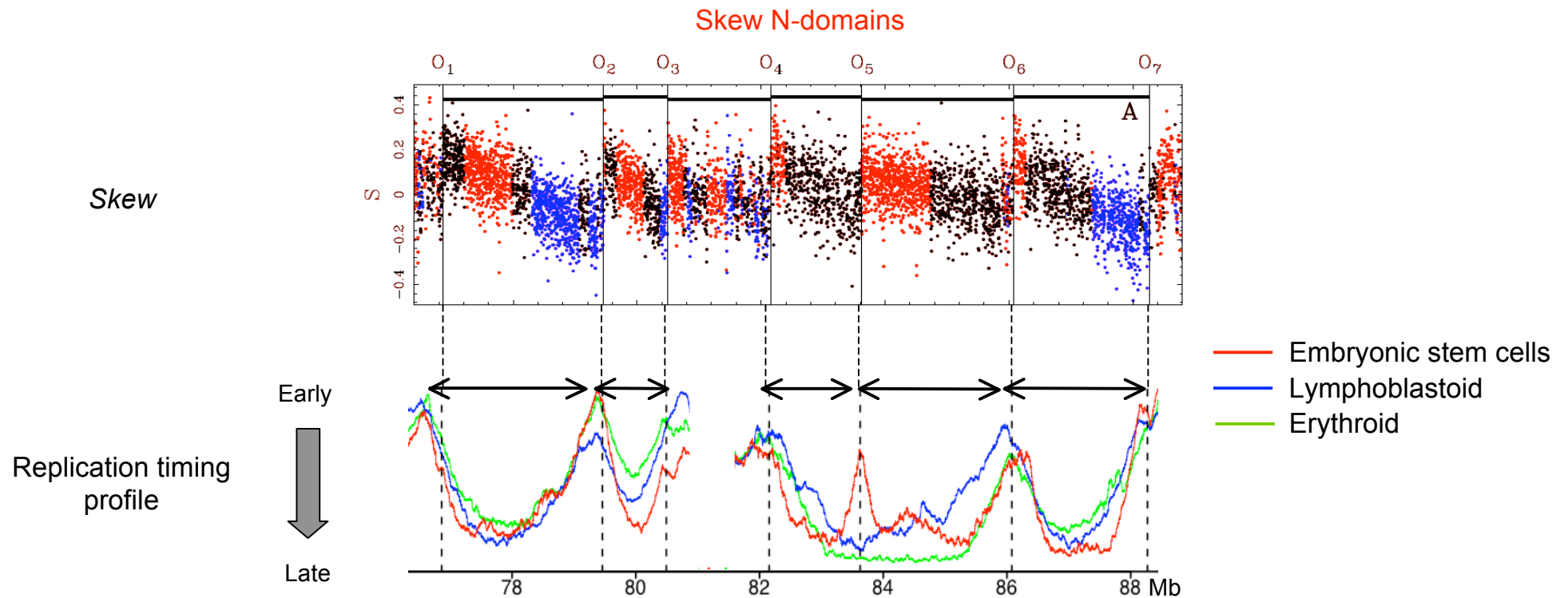
**If yes, what kind of replication program generates the N-shape?**



## **QUESTION 1**

**N-DOMAINS BORDERS HARBOR  
EARLY REPLICATION INITIATION ZONES ?**

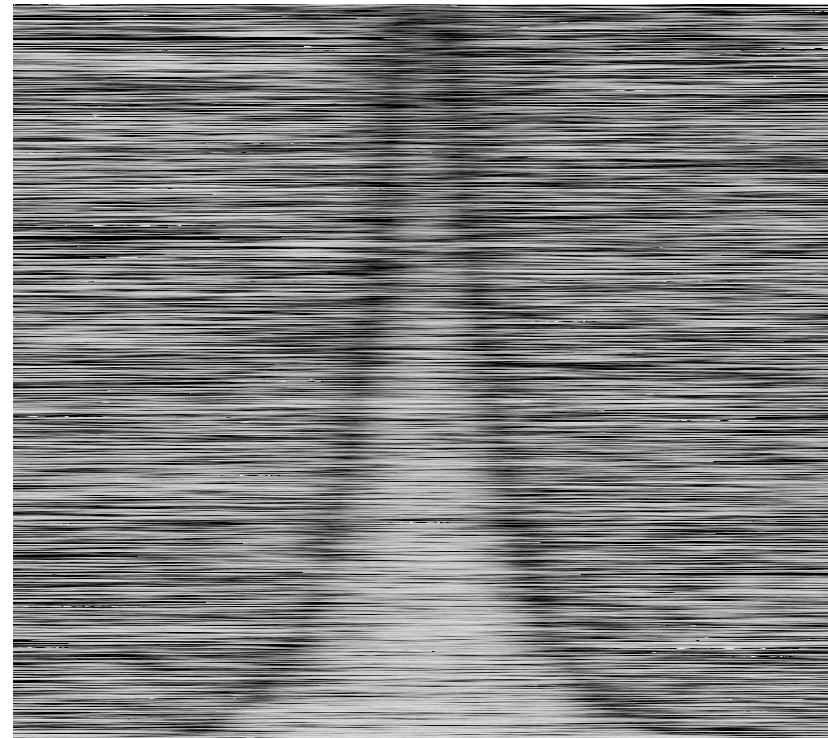
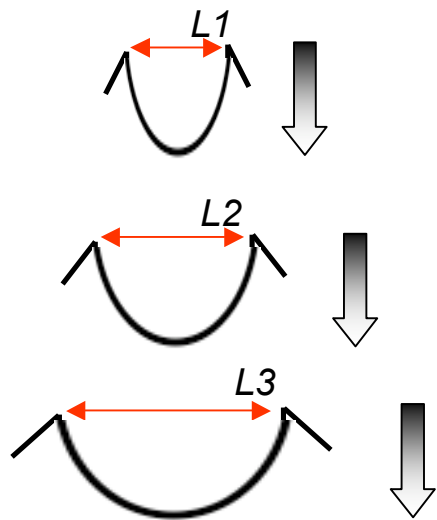
## COMPARISON OF SKEW PROFILE AND REPLICATION TIMING PROFILES



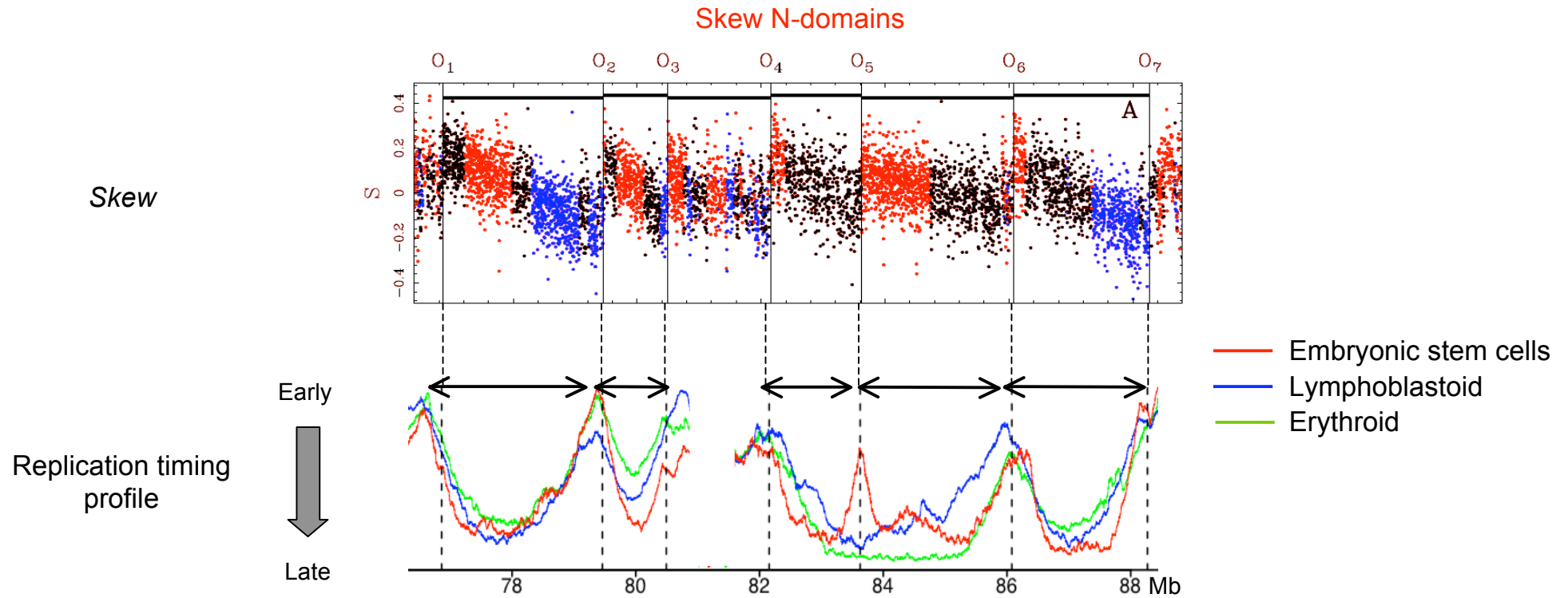
- N-domain extremities are significantly associated with replication initiation zones
- Replication starts from N domain borders and propagates to center in late S phase

## Replication timing profiles within N-domains

Embryonic stem cells

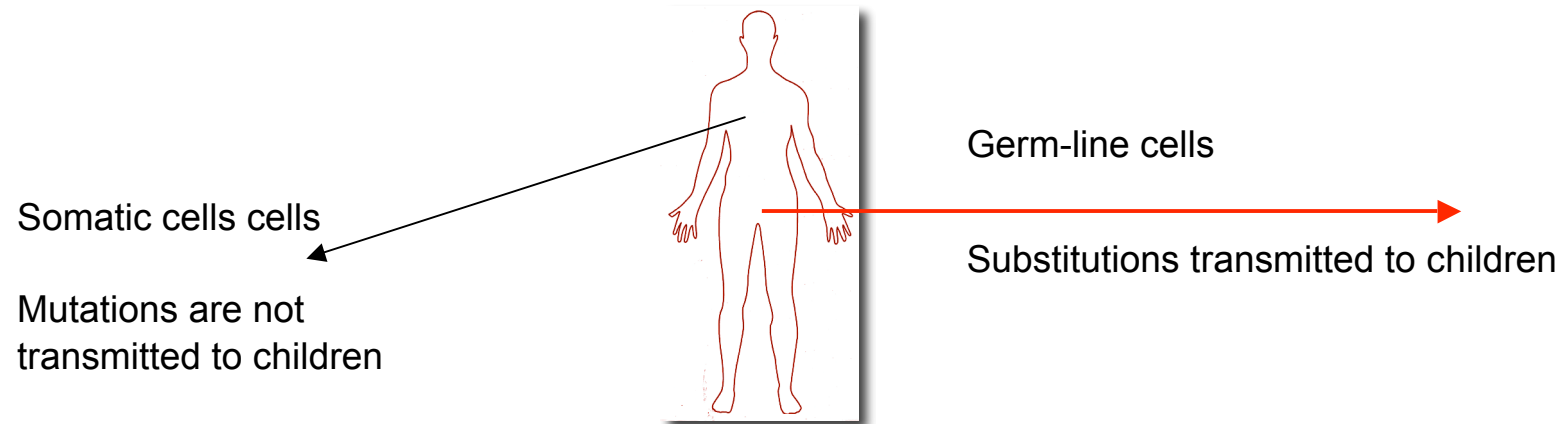


# COMPARISON OF SKEW PROFILE AND REPLICATION TIMING PROFILES



U-domains  $\longleftrightarrow$  replication domains in somatic cells  
N-domains  $\longleftrightarrow$  replication domains in germline cells

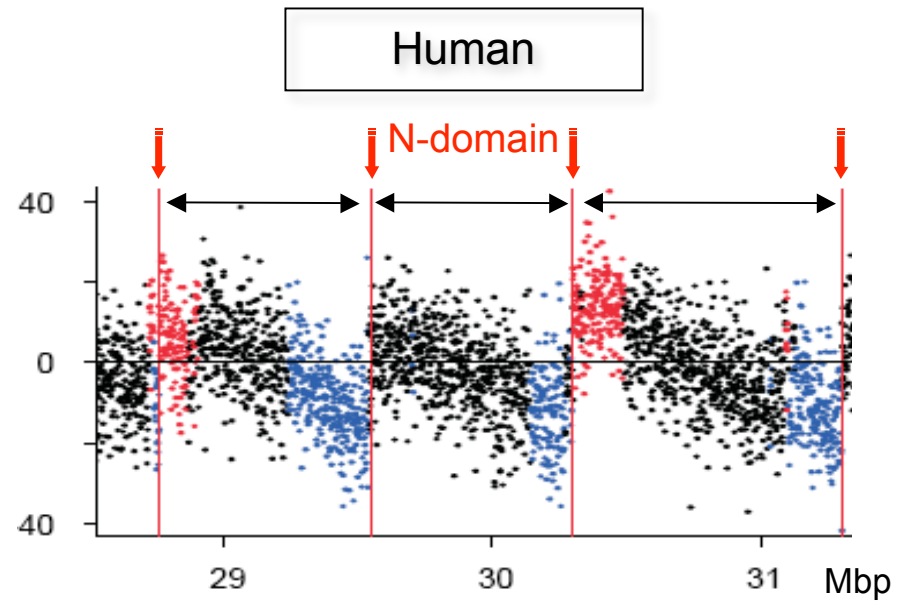
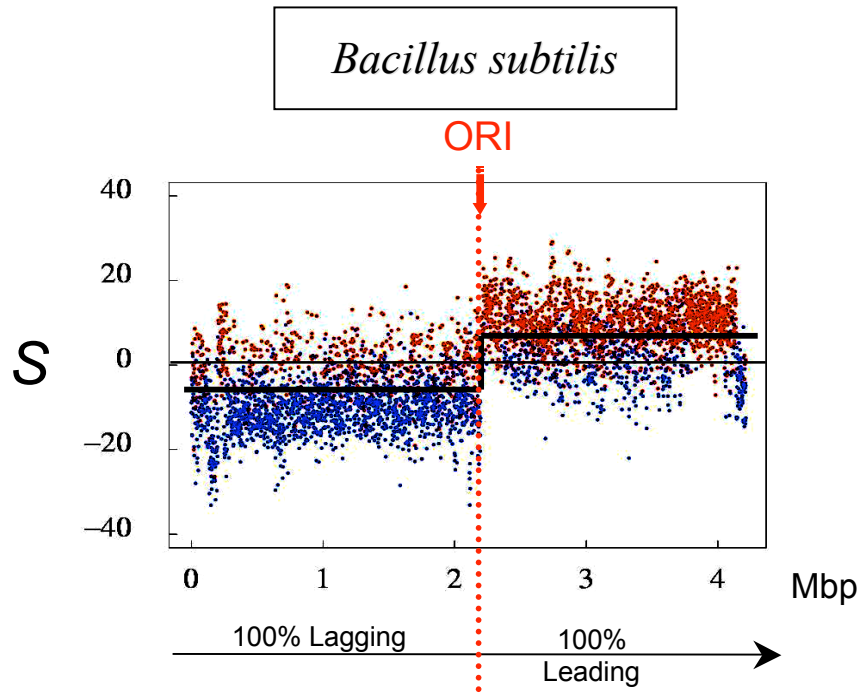
## SUBSTITUTIONS OBSERVED IN GENOME SEQUENCE OCCUR IN GERM-LINE CELLS



## N-DOMAIN PROFILE RESULTS FROM REPLICATION IN GERM-LINE CELLS

## **QUESTION 2**

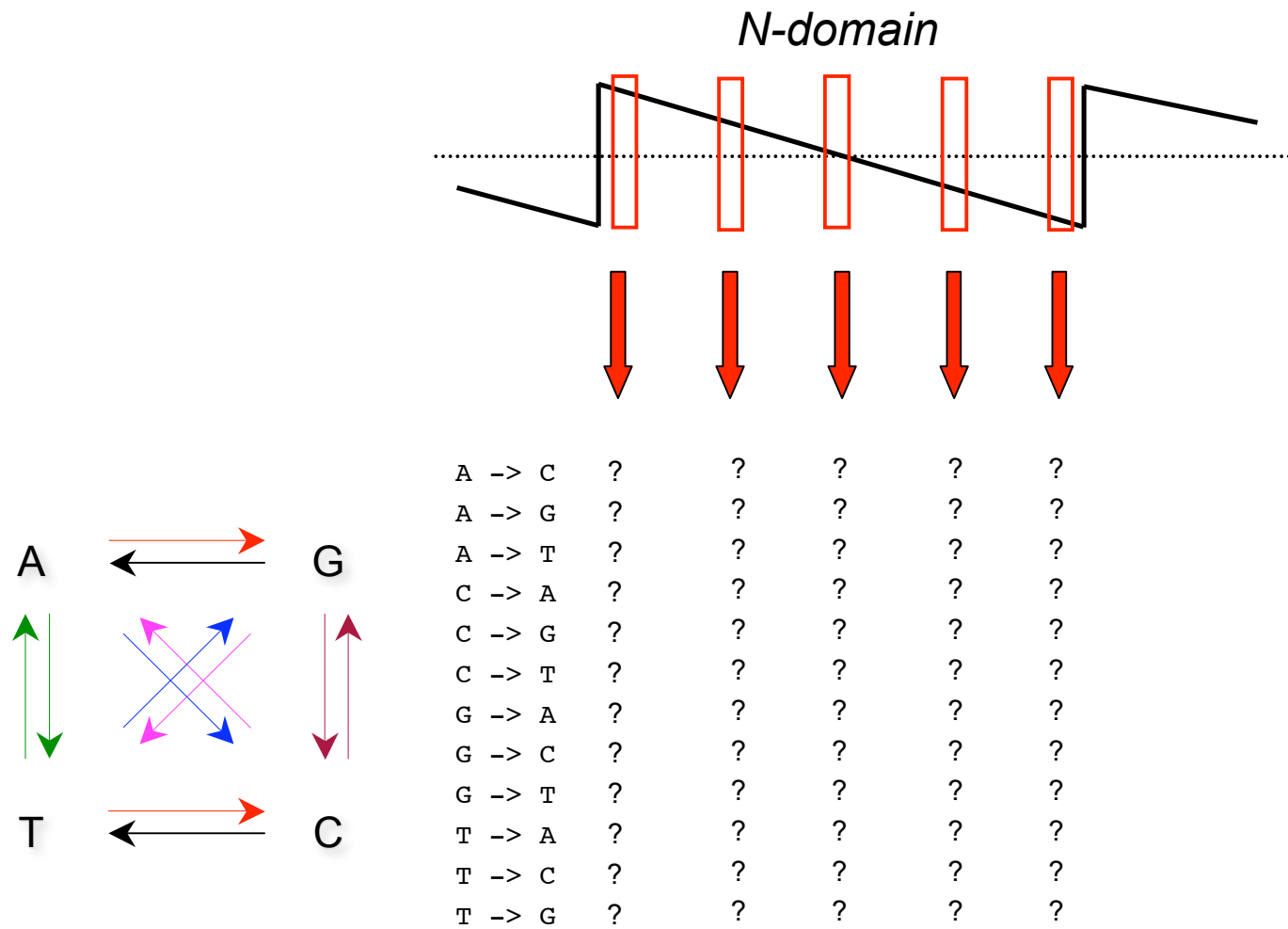
**N-DOMAINS RESULT FROM  
ASYMMETRIC MUTATION PATTERNS ?**



Is the “N” skew pattern generated by asymmetric nucleotide substitution rates ?

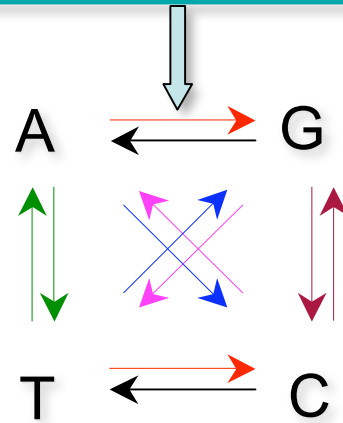
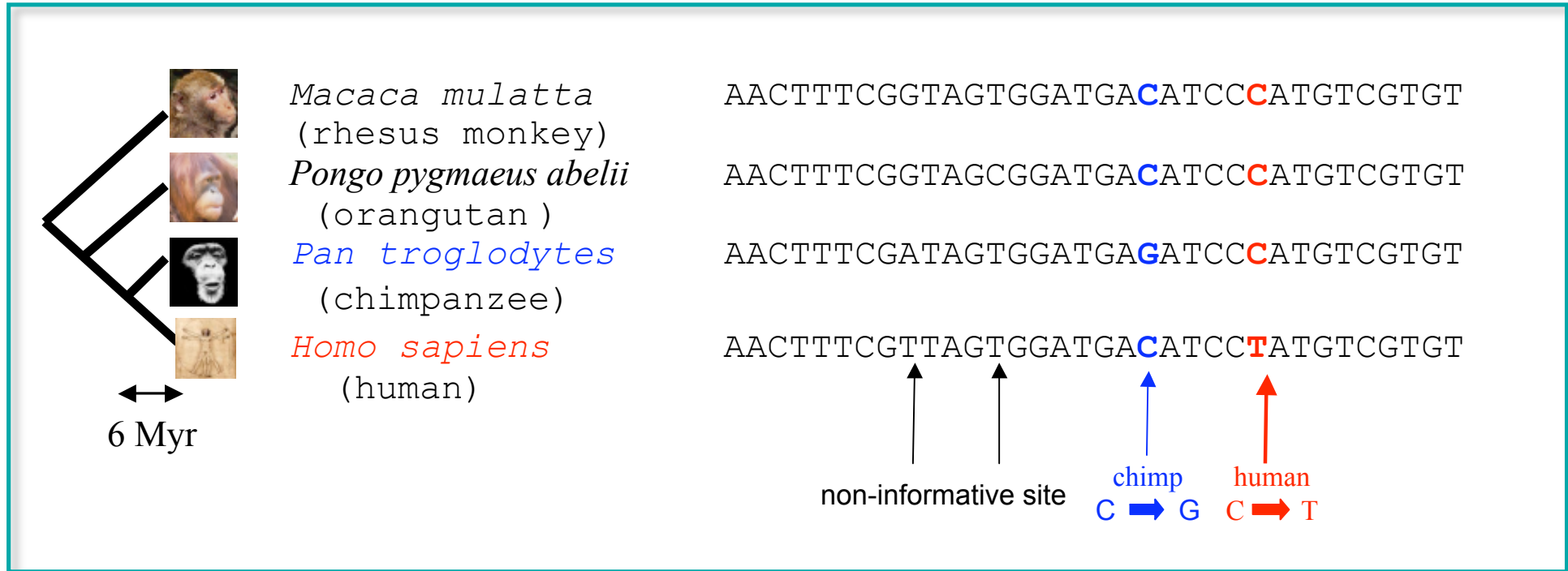


# Computation of nucleotide substitution rates





# Computation of nucleotide substitution rates



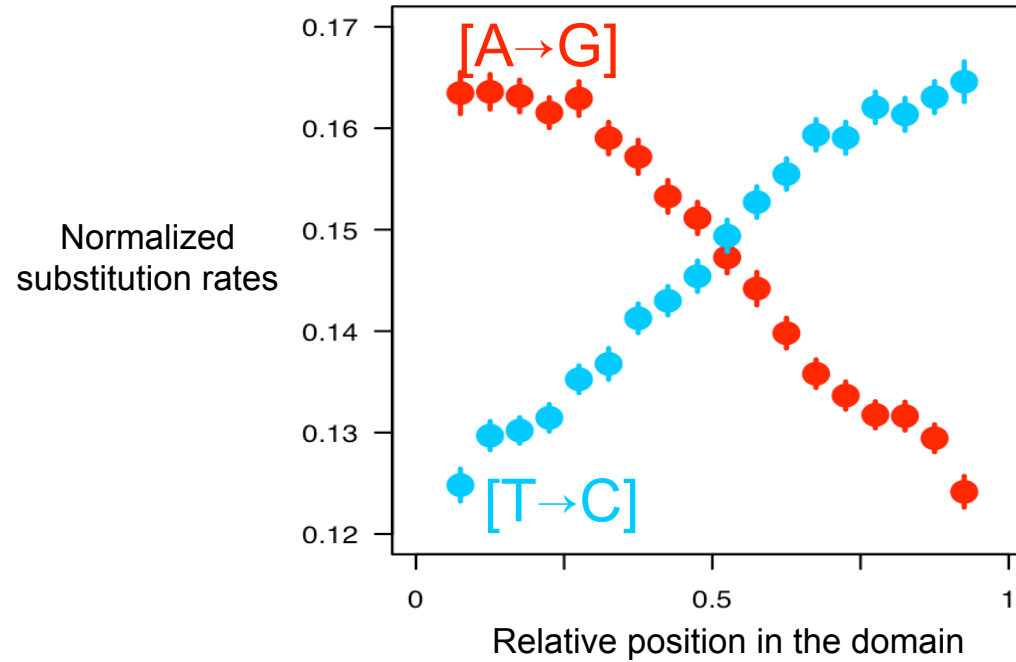
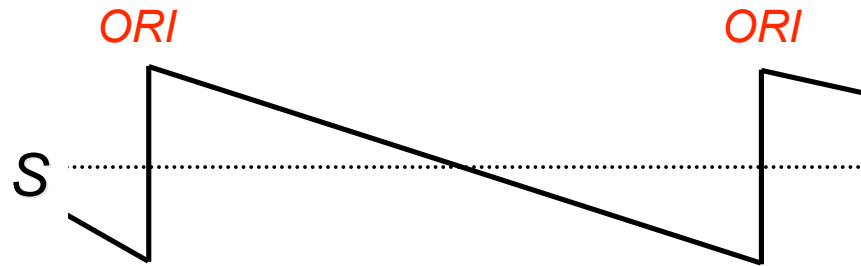
Compare A→G values on the two strands



Compare A→G to T→C on the same strand

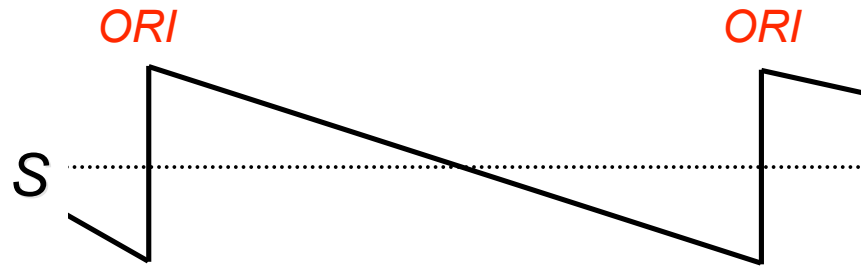
Compare complementary substitution rates on the same strand

# Complementary substitution rate along N-domains



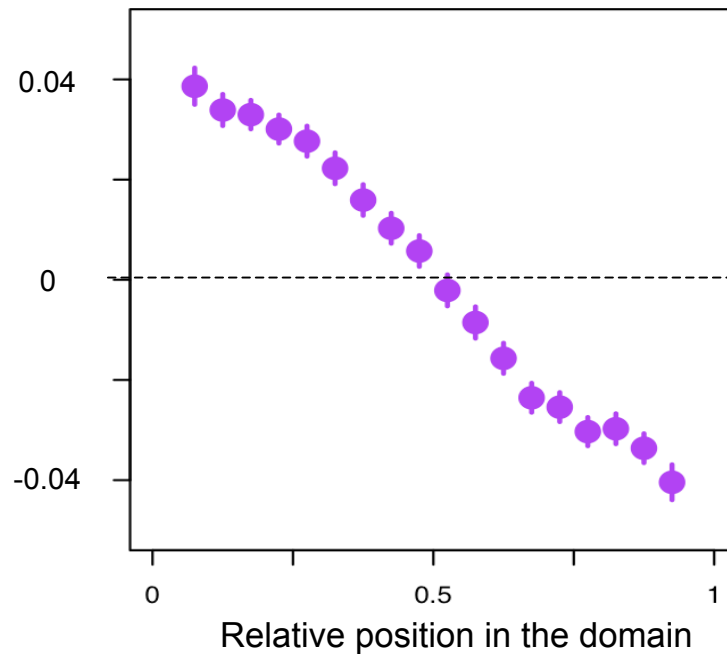
Unpublished

## Complementary substitution rate along N-domains



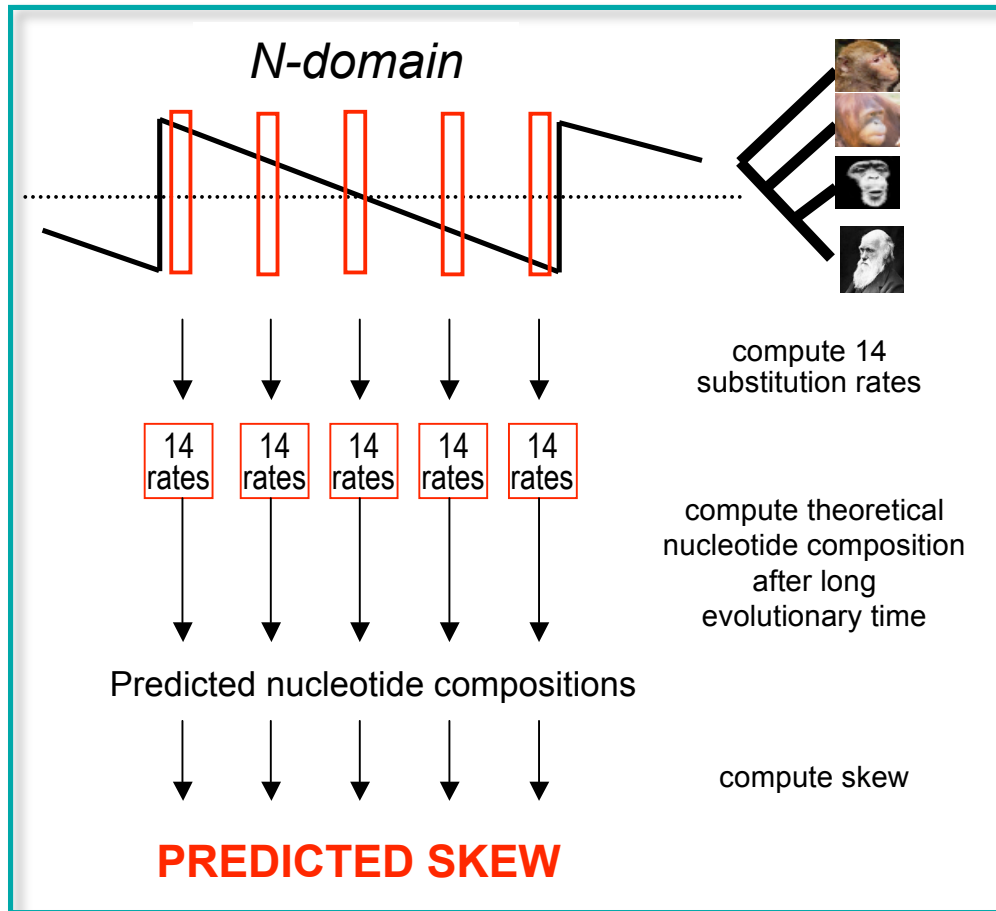
Asymmetry

$$\Delta = [A \rightarrow G] - [T \rightarrow C]$$

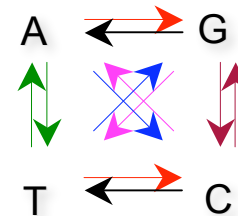


Reproduces perfectly the "N" skew profile

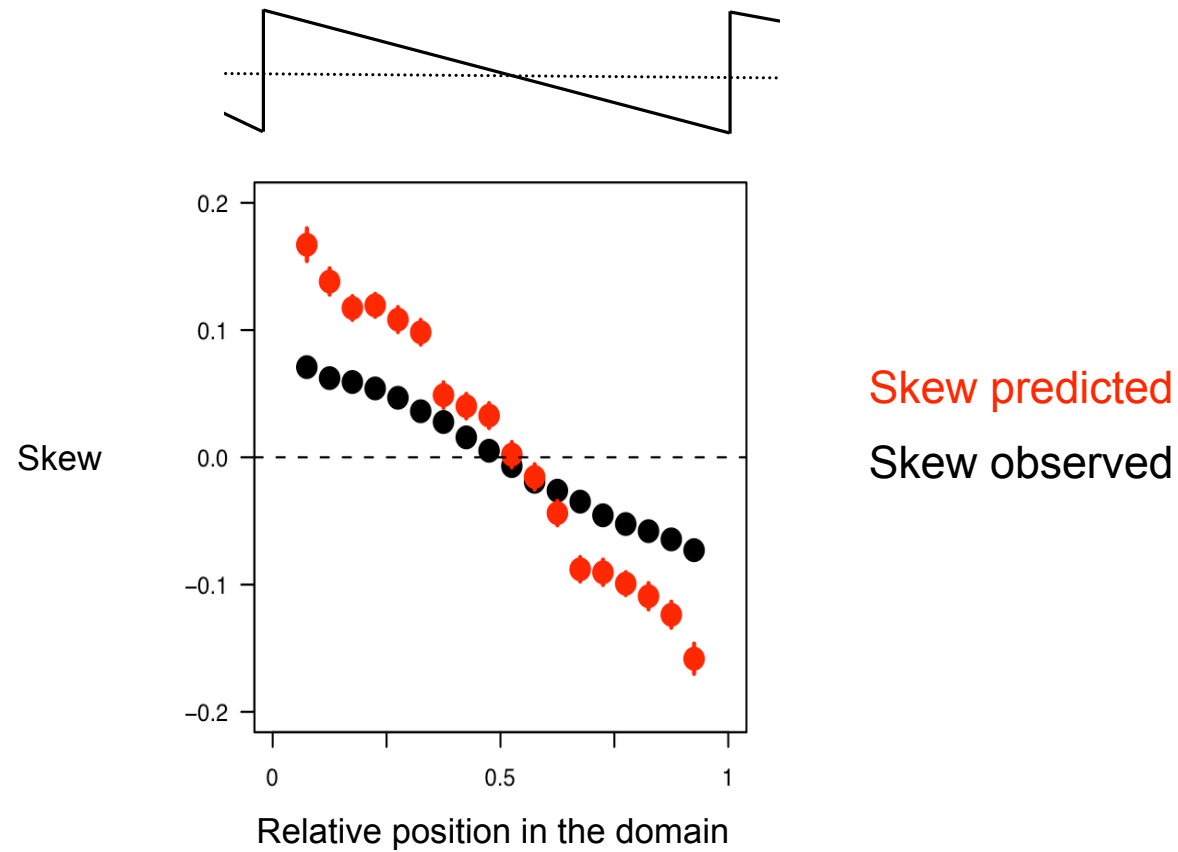
# Compute the predicted skew (S at equilibrium) along N-domain



$$\begin{aligned} \frac{\partial f_{aa}}{\partial t} &= ca f_{ca} + ga f_{ga} + ta f_{ta} - ac f_{aa} - ag f_{aa} - at f_{aa} + ca f_{ac} + ga f_{ag} + ta f_{at} - ac f_{aa} - ag f_{aa} - at f_{aa} + \frac{pgpa f_{cg} f_{ga}}{f_{ga} + f_{gc} + f_{gg} + f_{gt}} \\ \frac{\partial f_{ac}}{\partial t} &= ca f_{cc} + ga f_{gc} + ta f_{tc} - ac f_{ac} - ag f_{ac} - at f_{ac} + ac f_{aa} + gc f_{ag} + tc f_{at} - ca f_{ac} - cg f_{ac} - ct f_{ac} + \frac{pgpa f_{cg} f_{gc}}{f_{ga} + f_{gc} + f_{gg} + f_{gt}} - \frac{cptp f_{ac} f_{cg}}{f_{ca} + f_{cc} + f_{cg} + f_{ct}} \\ \frac{\partial f_{ag}}{\partial t} &= ca f_{cg} + ga f_{gg} + ta f_{tg} - ac f_{ag} - ag f_{ag} - at f_{ag} + ag f_{aa} + cg f_{ag} + tg f_{at} - ga f_{ag} - gc f_{ag} - gt f_{ag} + \frac{pgpa f_{cg} f_{gg}}{f_{ga} + f_{gc} + f_{gg} + f_{gt}} \\ \frac{\partial f_{at}}{\partial t} &= ca f_{ct} + ga f_{gt} + ta f_{tt} - ac f_{at} - ag f_{at} - at f_{at} + at f_{aa} + ct f_{ac} + gt f_{ag} - ta f_{at} - tc f_{at} - tg f_{at} + \frac{pgpa f_{cg} f_{gt}}{f_{ga} + f_{gc} + f_{gg} + f_{gt}} + \frac{cptp f_{ac} f_{cg}}{f_{ca} + f_{cc} + f_{cg} + f_{ct}} \\ \frac{\partial f_{ca}}{\partial t} &= ac f_{aa} + gc f_{ga} + tc f_{ta} - ca f_{ca} - cg f_{ca} - ct f_{ca} + ca f_{ca} + ta f_{ct} - ac f_{ca} - ag f_{ca} - at f_{ca} + pgpa f_{cg} \\ \frac{\partial f_{cc}}{\partial t} &= ac f_{ac} + gc f_{gc} + tc f_{tc} - ca f_{cc} - cg f_{cc} - ct f_{cc} + ac f_{ca} + gc f_{cg} + tc f_{ct} - ca f_{cc} - cg f_{cc} - ct f_{cc} - \frac{cptp f_{cc} f_{cg}}{f_{ca} + f_{cc} + f_{cg} + f_{ct}} \\ \frac{\partial f_{cg}}{\partial t} &= ac f_{ag} + gc f_{gg} + tc f_{tg} - ca f_{cg} - cg f_{cg} + ag f_{ca} + cg f_{cc} + tg f_{ct} - gc f_{cg} - gt f_{cg} - cptp f_{cg} - pgpa f_{cg} \\ \frac{\partial f_{ct}}{\partial t} &= ac f_{at} + gc f_{gt} + tc f_{tt} - ca f_{ct} - cg f_{ct} - ct f_{ct} + at f_{ca} + ct f_{cc} + gt f_{cg} - ta f_{ct} - tc f_{ct} - tg f_{ct} + \frac{cptp f_{cc} f_{cg}}{f_{ca} + f_{cc} + f_{cg} + f_{ct}} \\ \frac{\partial f_{ga}}{\partial t} &= ag f_{aa} + cg f_{ca} + tg f_{ta} - ga f_{ga} - gc f_{ga} - gt f_{ga} + ca f_{gc} + ga f_{gg} + ta f_{gt} - ac f_{ga} - ag f_{ga} - at f_{ga} - \frac{pgpa f_{cg} f_{ga}}{f_{ga} + f_{gc} + f_{gg} + f_{gt}} \\ \frac{\partial f_{gc}}{\partial t} &= ag f_{ac} + cg f_{cc} + tg f_{tc} - ga f_{gc} - gc f_{gc} - gt f_{gc} + ac f_{ga} + gc f_{gg} + tc f_{ct} - ca f_{gc} - cg f_{gc} - ct f_{gc} - \frac{pgpa f_{cg} f_{gc}}{f_{ga} + f_{gc} + f_{gg} + f_{gt}} - \frac{cptp f_{gc} f_{cg}}{f_{ca} + f_{cc} + f_{cg} + f_{ct}} \\ \frac{\partial f_{gg}}{\partial t} &= ag f_{ag} + cg f_{cg} + tg f_{tg} - ga f_{gg} - gc f_{gg} - gt f_{gg} + ag f_{ga} + cg f_{gc} + tg f_{gt} - ga f_{gg} - gc f_{gg} - gt f_{gg} - \frac{pgpa f_{cg} f_{gg}}{f_{ga} + f_{gc} + f_{gg} + f_{gt}} \\ \frac{\partial f_{gt}}{\partial t} &= ag f_{at} + cg f_{ct} + tg f_{tt} - ga f_{gt} - gc f_{gt} - gt f_{gt} + at f_{ga} + ct f_{gc} + gt f_{gt} - ta f_{gt} - tc f_{gt} - tg f_{gt} - \frac{pgpa f_{cg} f_{gt}}{f_{ga} + f_{gc} + f_{gg} + f_{gt}} + \frac{cptp f_{gc} f_{cg}}{f_{ca} + f_{cc} + f_{cg} + f_{ct}} \\ \frac{\partial f_{ta}}{\partial t} &= at f_{aa} + ct f_{ca} + gt f_{ga} - ta f_{ta} - tc f_{ta} - tg f_{ta} + ca f_{tc} + ga f_{tg} + ta f_{tt} - ac f_{ta} - ag f_{ta} - at f_{ta} \\ \frac{\partial f_{tc}}{\partial t} &= at f_{ac} + ct f_{cc} + gt f_{gc} - ta f_{tc} - tc f_{tc} - tg f_{tc} + ac f_{ta} + gc f_{tg} + tc f_{tt} - ca f_{tc} - cg f_{tc} - ct f_{tc} - \frac{cptp f_{tc} f_{cg}}{f_{ca} + f_{cc} + f_{cg} + f_{ct}} \\ \frac{\partial f_{tg}}{\partial t} &= at f_{ag} + gt f_{gg} - ta f_{tg} - tc f_{tg} - tg f_{tg} + ag f_{ta} + cg f_{tc} + tg f_{tt} - ga f_{tg} - gc f_{tg} - gt f_{tg} + cptp f_{cg} \\ \frac{\partial f_{tu}}{\partial t} &= at f_{at} + ct f_{ct} + gt f_{gt} - ta f_{tu} - tc f_{tu} - tg f_{tu} + at f_{ta} + ct f_{tc} + gt f_{tg} - ta f_{tu} - tc f_{tu} - tg f_{tu} + \frac{cptp f_{tc} f_{cg}}{f_{ca} + f_{cc} + f_{cg} + f_{ct}} \end{aligned}$$



# COMPOSITION AT EQUILIBRIUM REPRODUCES PERFECTLY THE “N” SKEW PROFILE

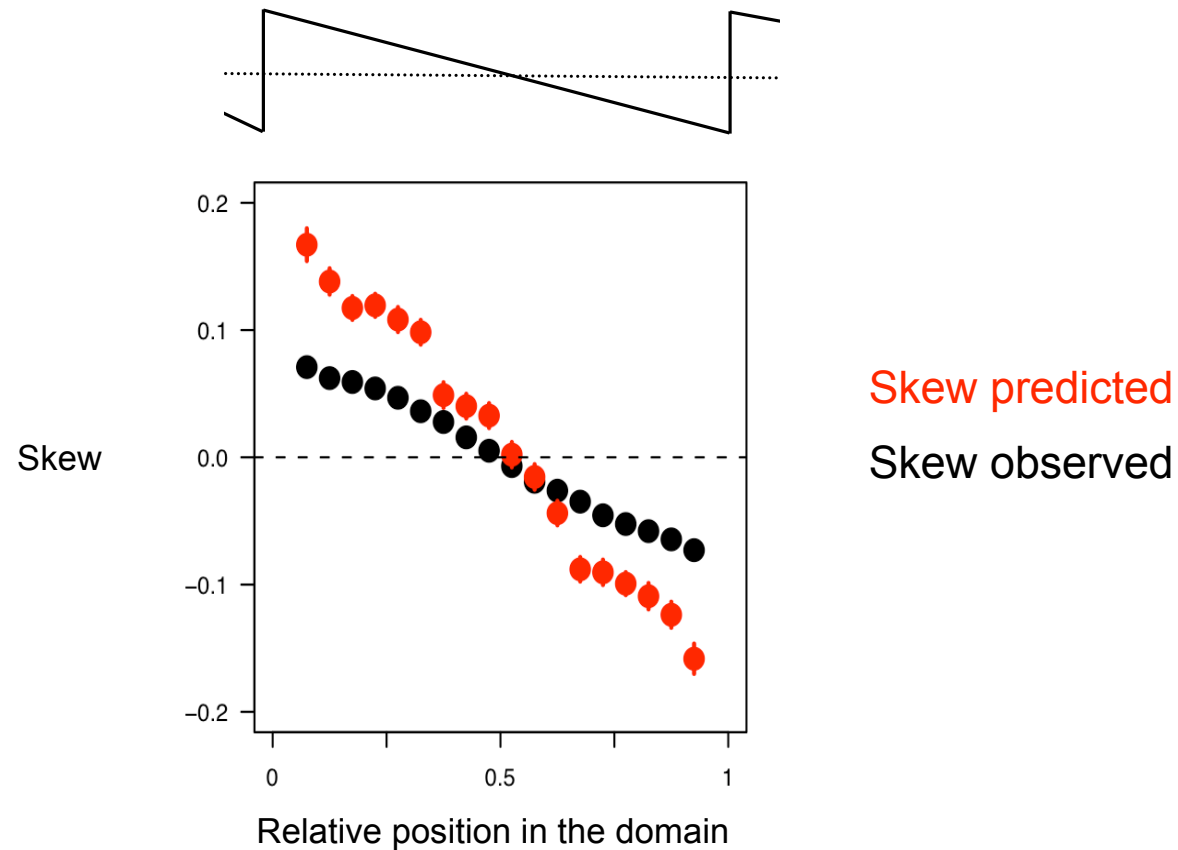


N-domains result from mutation asymmetry in germline cells



Unpublished

# COMPOSITION AT EQUILIBRIUM REPRODUCES PERFECTLY THE “N” SKEW PROFILE



N-domains result from mutation asymmetry in germline cells

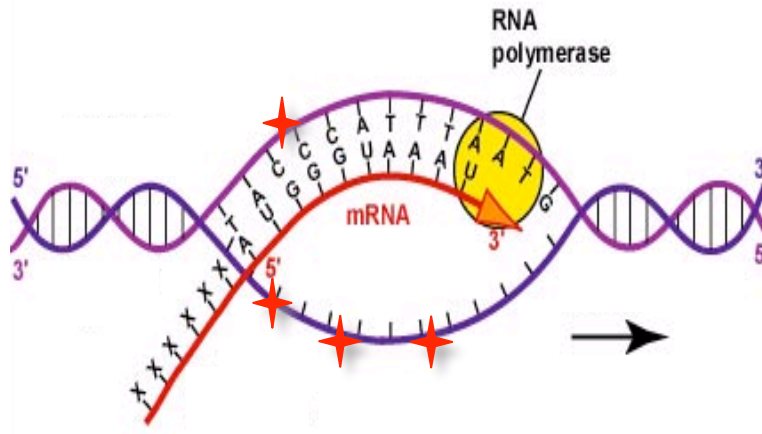
The skew is not at equilibrium. Time to reach the observed skew : 300 – 400 million years

## **QUESTION 3**

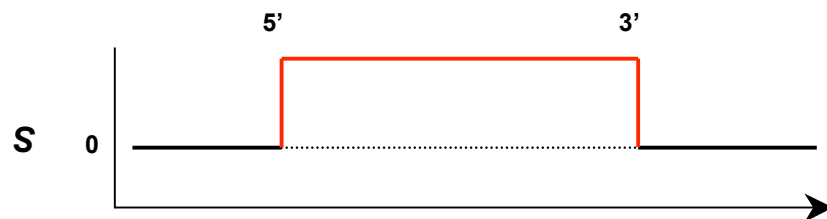
**N-DOMAINS RESULT FROM ASYMMETRIC MUTATIONS  
ASSOCIATED WITH TRANSCRIPTION OR REPLICATION ?**



# TRANSCRIPTIONAL MUTATIONAL ASYMETRIES



asymmetry of substitution rates  
between **transcribed** and **non-transcribed** strands



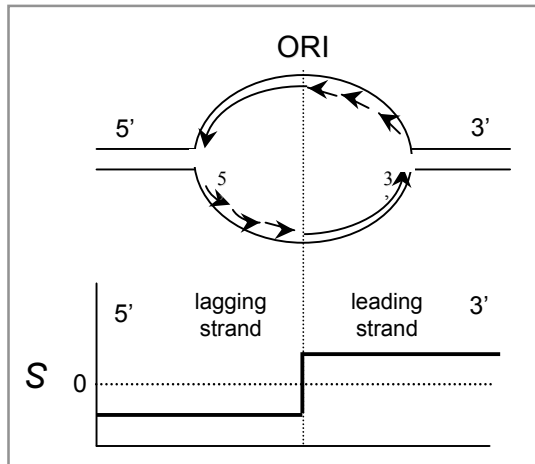
asymmetry of nucleotide composition

⇒ Transcription: **G > C** and **T > A** on the non-transcribed strand

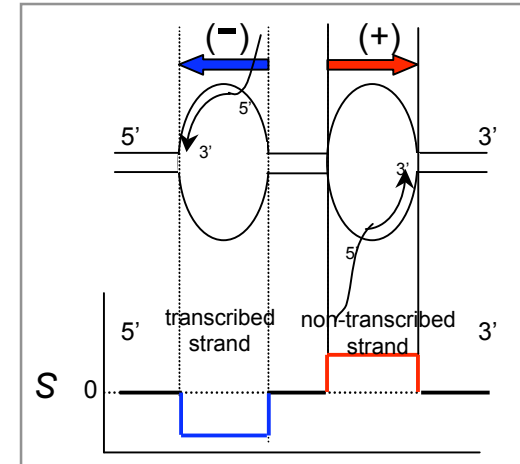
Beletskii A. *Biol.Chem.*, (1998) 379:549  
Green P. et al. *Nat. Genet.* (2003) 33:514  
Touchon M. et al. *NAR.* (2004) 32:4969

Total skew = superposition of skews due to replication and transcription

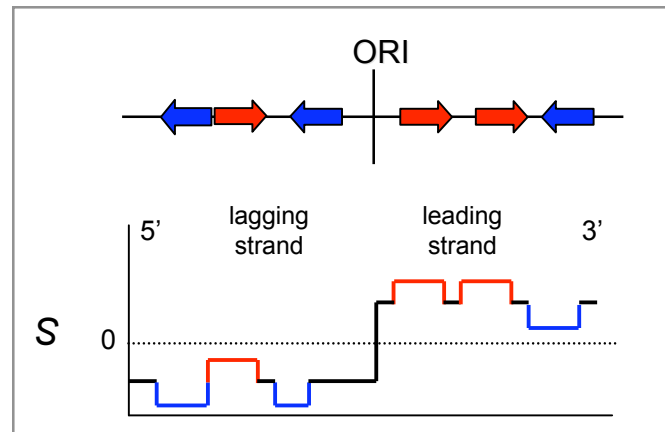
Replicative skew profile



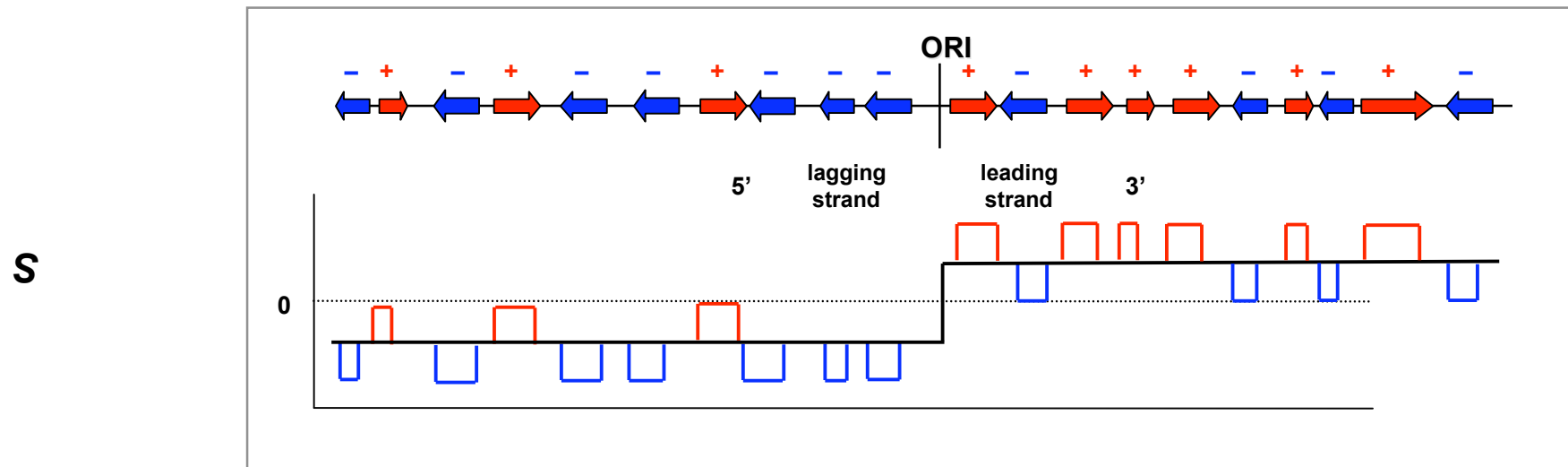
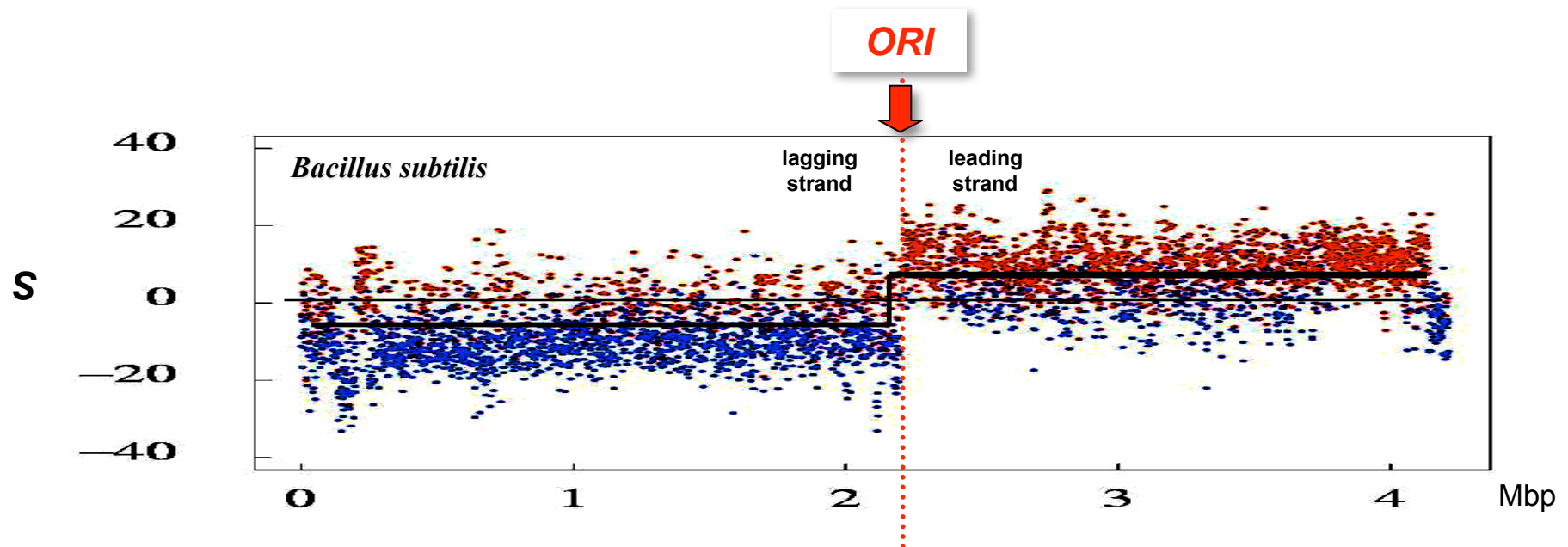
Transcriptional skew



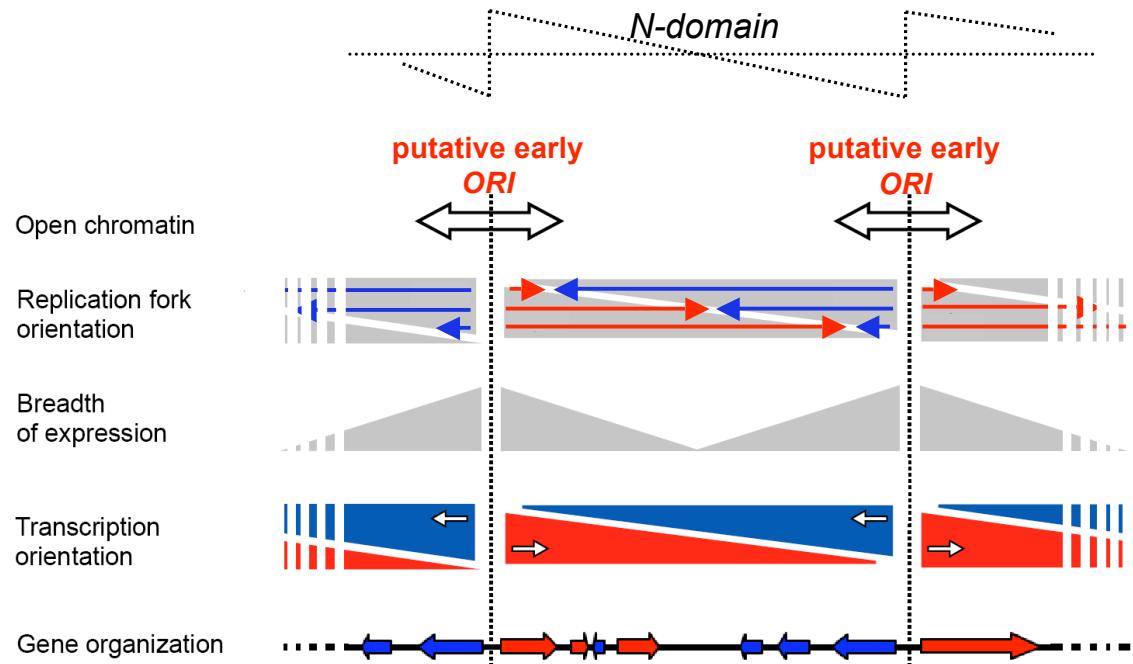
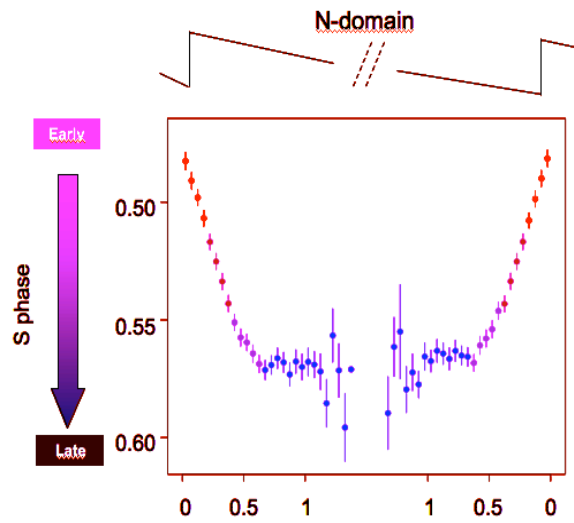
Superimposition of replication and transcription



Total skew = superposition of skews due to replication and transcription

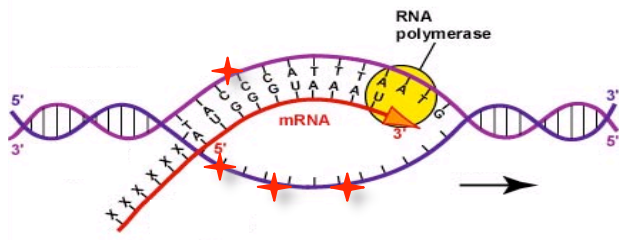


# COORDINATION OF REPLICATION AND TRANSCRIPTION

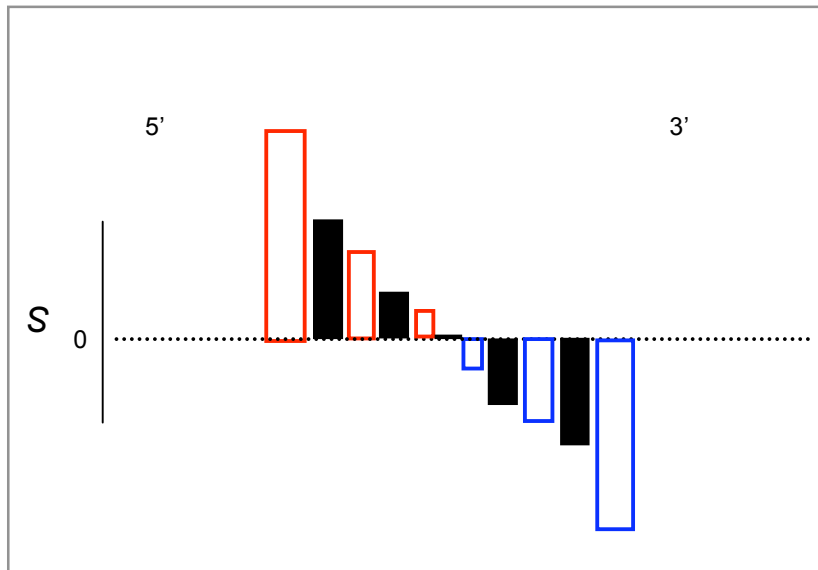
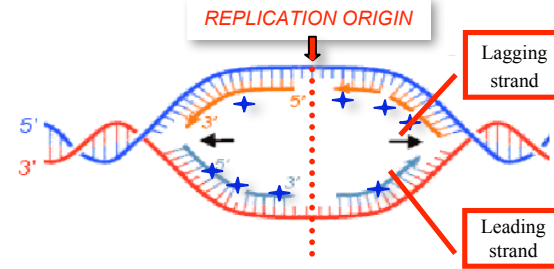


➤ Does the “N” result from:

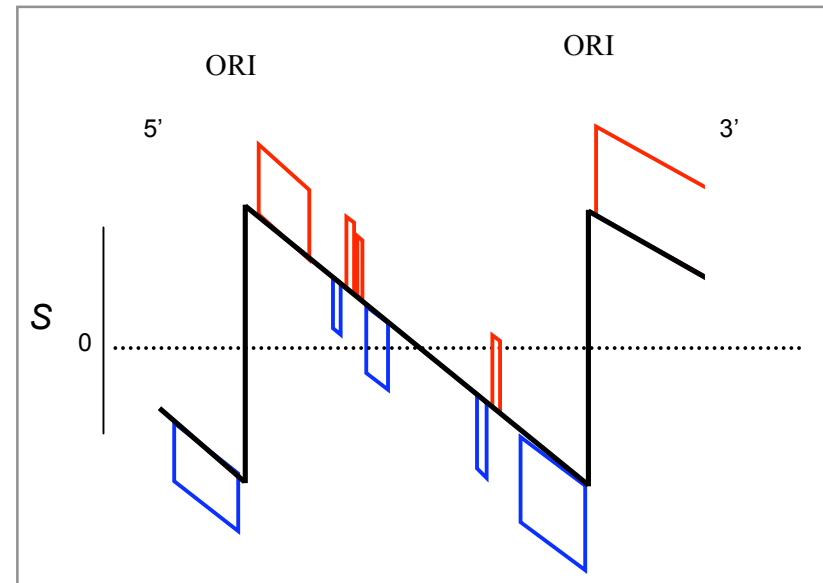
Transcription



Replication

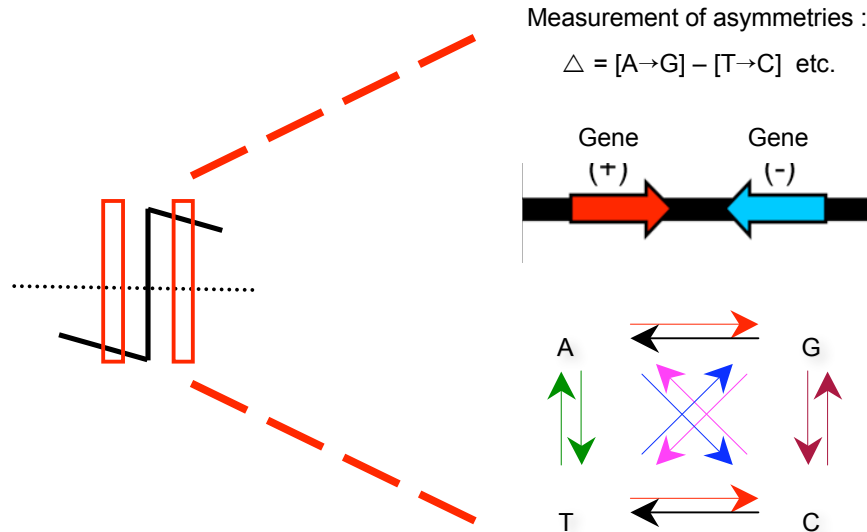


Transcription specifically organized



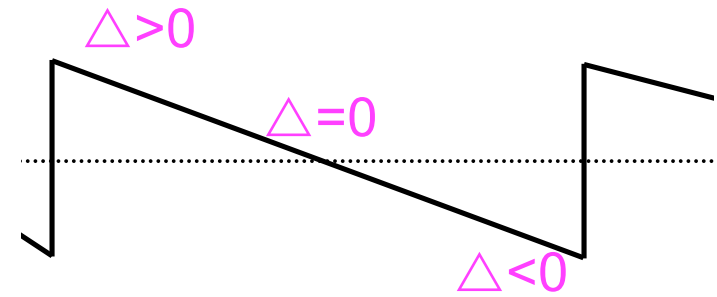
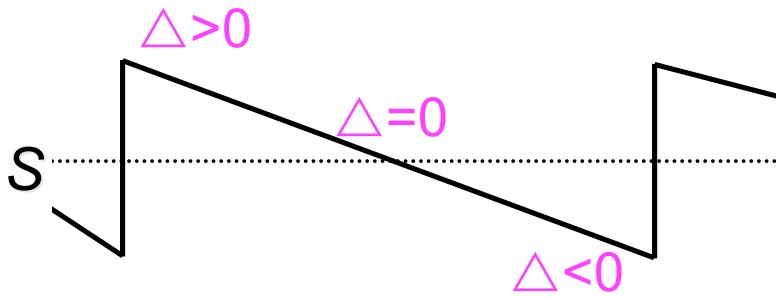
Replication specifically organized

# ASYMETRIES ASSOCIATED WITH REPLICATION AND TRANSCRIPTION

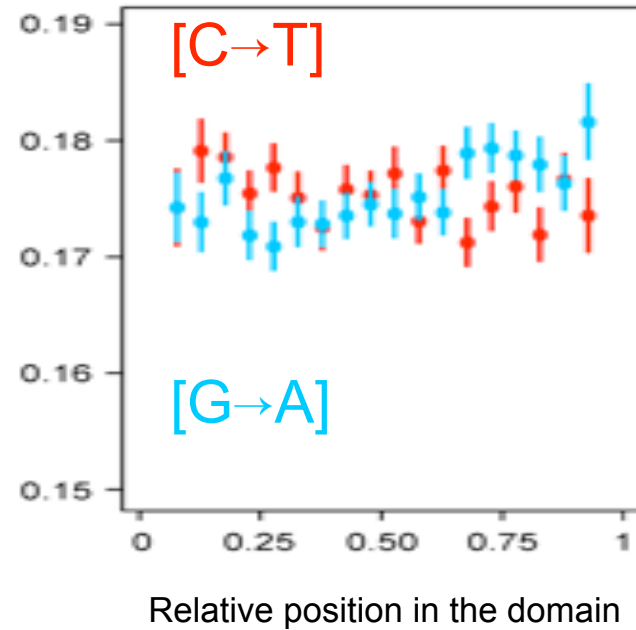
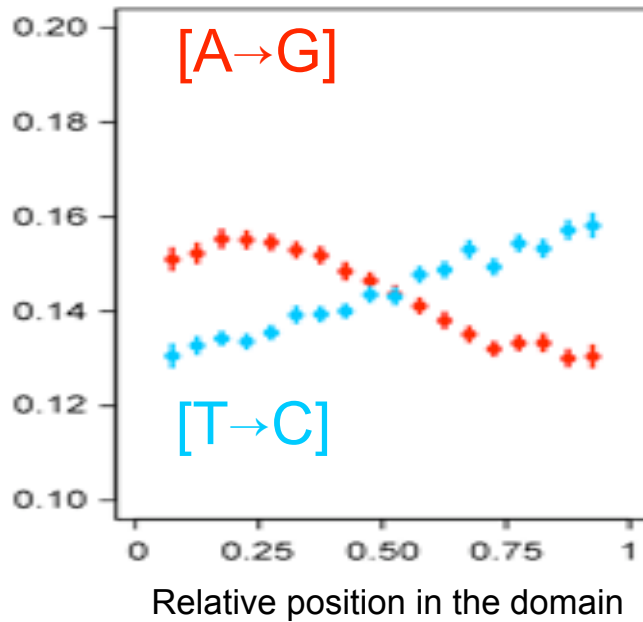


	Replication	Transcription
$r(A \rightarrow G) - r(T \rightarrow C)$	$0.0329 \pm 0.0022$	$0.0939 \pm 0.0029$
<i>P</i> value	$3 \times 10^{-46}$	$< 1 \times 10^{-100}$
$r(C \rightarrow T) - r(G \rightarrow A)$	$0.0130 \pm 0.0021$	$-0.0098 \pm 0.0028$
<i>P</i> value	$6 \times 10^{-10}$	$6 \times 10^{-4}$
$r(A \rightarrow T) - r(T \rightarrow A)$	$0.0029 \pm 0.0007$	$0.0077 \pm 0.0009$
<i>P</i> value	$9 \times 10^{-5}$	$8 \times 10^{-17}$
$r(C \rightarrow G) - r(G \rightarrow C)$	$0.0077 \pm 0.0013$	$0.0227 \pm 0.0017$
<i>P</i> value	$3 \times 10^{-9}$	$4 \times 10^{-39}$
$r(A \rightarrow C) - r(T \rightarrow G)$	$0.0014 \pm 0.0008$	$-0.0005 \pm 0.0011$
<i>P</i> value	0.09	0.64
$r(G \rightarrow T) - r(C \rightarrow A)$	$0.0039 \pm 0.0012$	$0.0130 \pm 0.0015$
<i>P</i> value	$1 \times 10^{-3}$	$3 \times 10^{-17}$

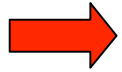
# Complementary substitution rates along N-domains



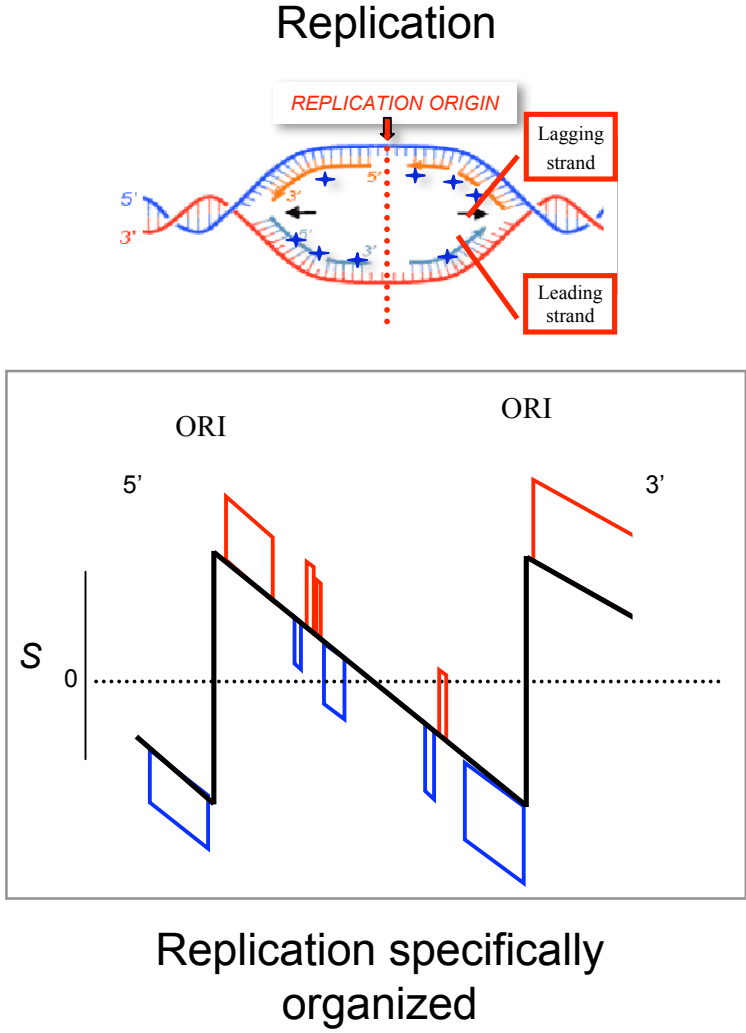
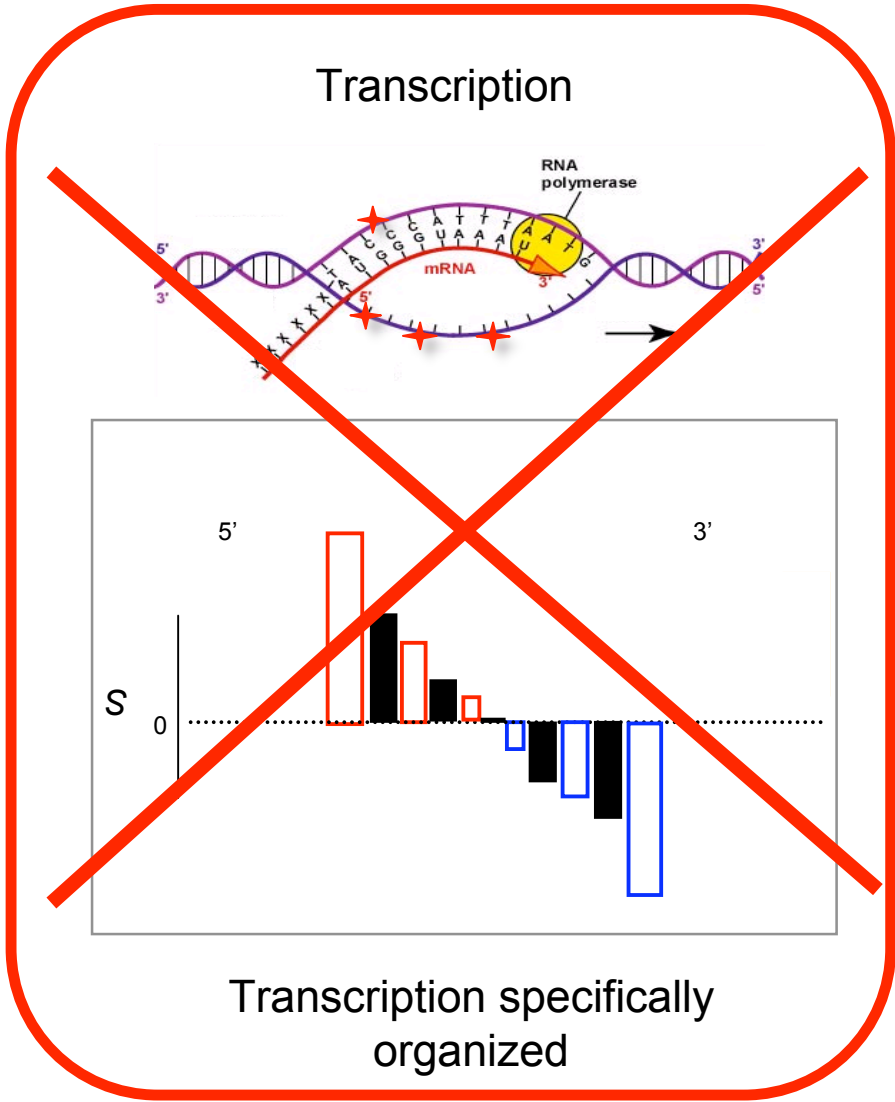
Normalized substitution rates



Unpublished

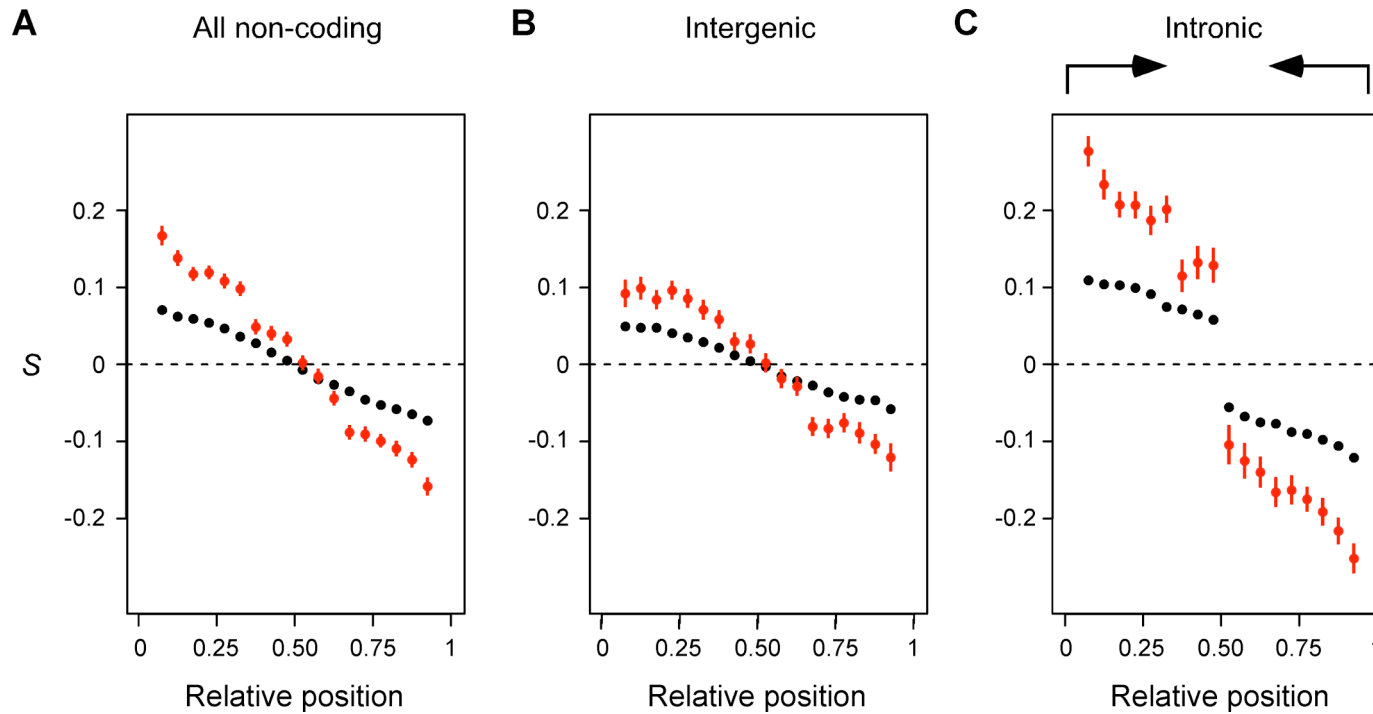


“N” pattern of skew profile results from replication-associated mutational strand asymmetries





# SUPERPOSITION OF SKEWS DUE TO REPLICATION AND TRANSCRIPTION



Unpublished

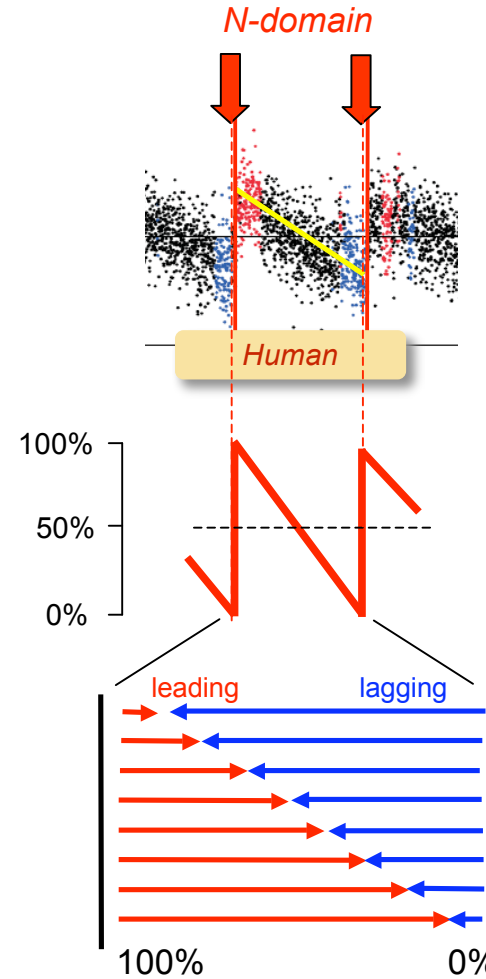
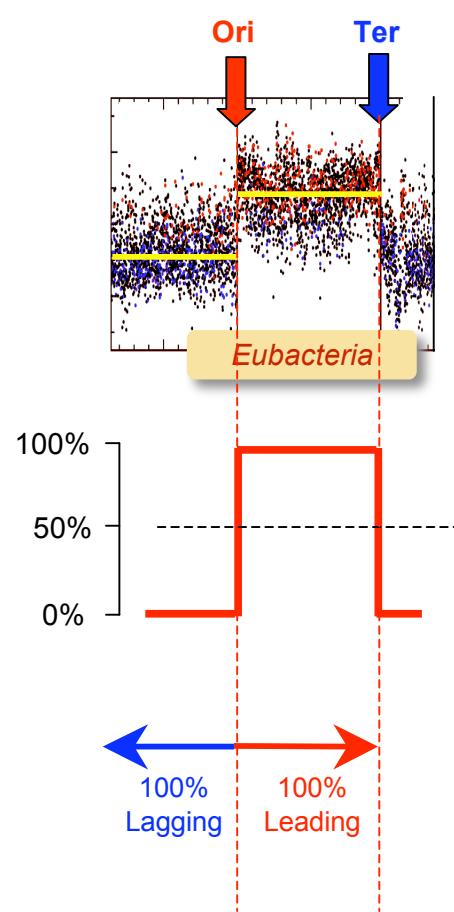
## **QUESTION 4**

**WHAT MODEL OF REPLICATION  
CAN EXPLAIN THE N-PATTERN ?**

# MODEL: N-shape results from gradient of replication fork polarity

## Replication fork polarity

Leading  
Leading + lagging

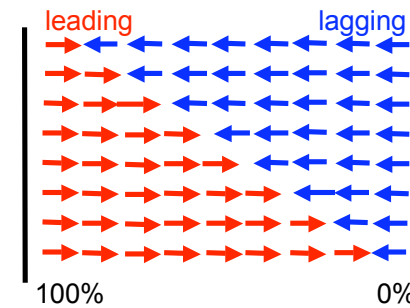
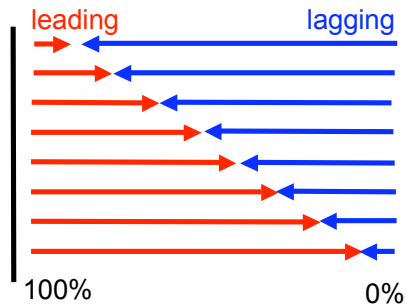


# Replication models that can generate N-shape of replication fork polarity

Single replication fork  
with random termination

Multiple internal replication origins  
with increase of fork speed and/or of initiation rate

gradient of  
replication fork  
polarity

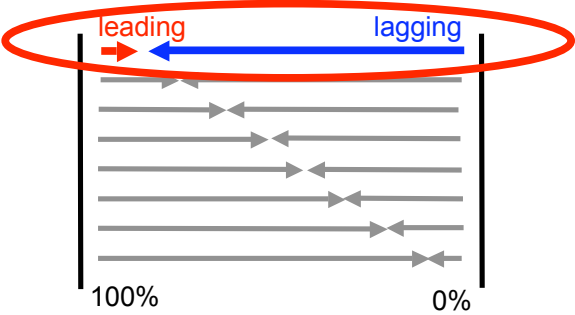


# Replication models that can generate N-shape of replication fork polarity

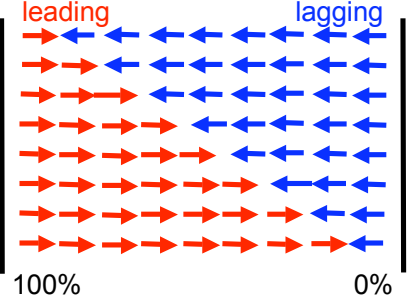
Single replication fork with random termination

Multiple internal replication origins with increase of fork speed and/or of initiation rate

gradient of replication fork polarity



Cell 1

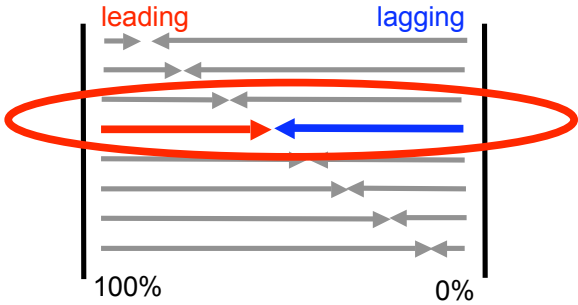


# Replication models that can generate N-shape of replication fork polarity

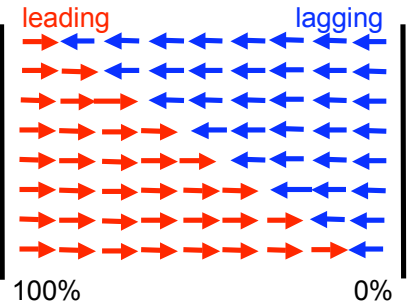
Single replication fork with random termination

Multiple internal replication origins with increase of fork speed and/or of initiation rate

gradient of replication fork polarity



Cell 2

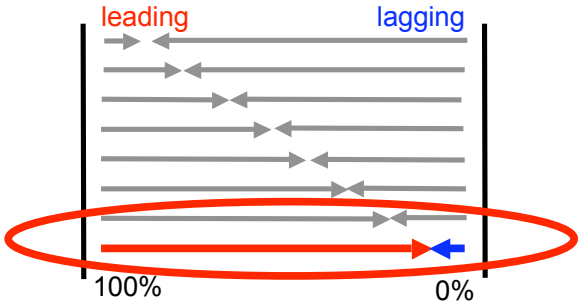


# Replication models that can generate N-shape of replication fork polarity

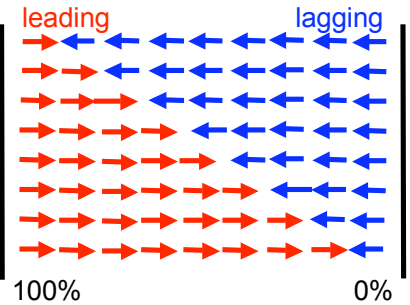
Single replication fork with random termination

Multiple internal replication origins with increase of fork speed and/or of initiation rate

gradient of replication fork polarity



⋮  
Cell n

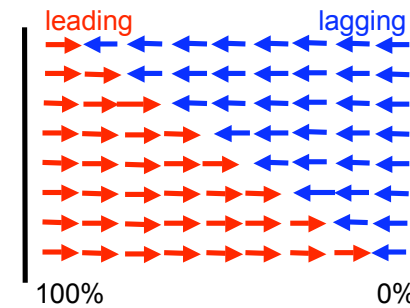
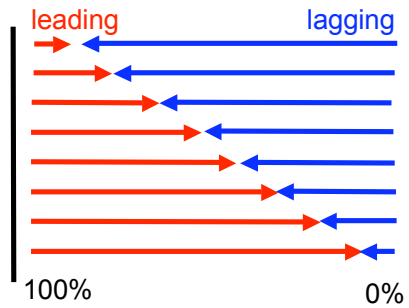


# Replication models that can generate N-shape of replication fork polarity

Single replication fork with random termination

Multiple internal replication origins with increase of fork speed and/or of initiation rate

gradient of replication fork polarity



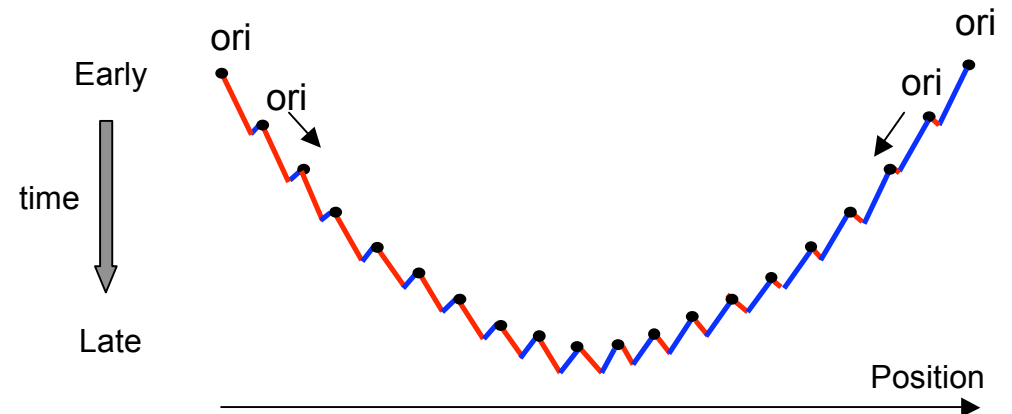
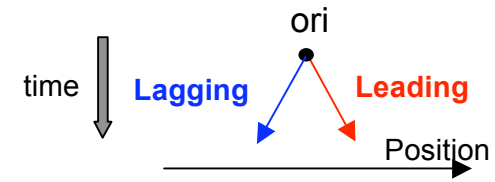
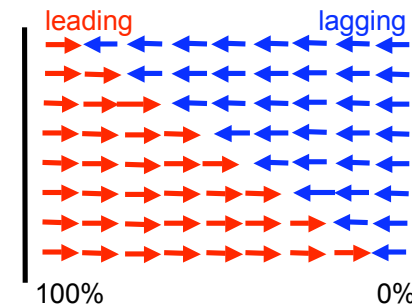
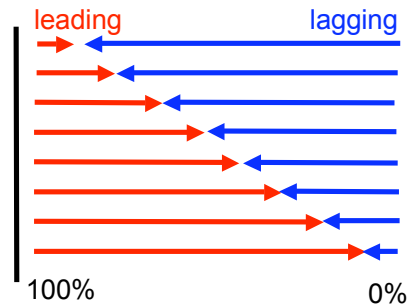


# Replication models that can generate N-shape of replication fork polarity

Single replication fork with random termination

Multiple internal replication origins with increase of fork speed and/or of initiation rate

gradient of replication fork polarity

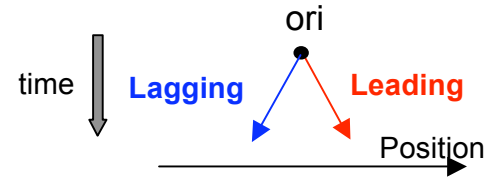
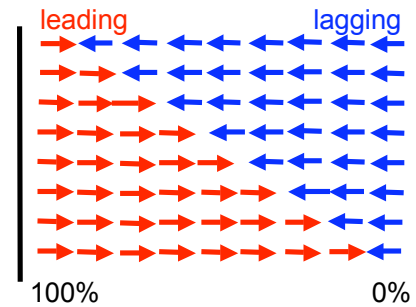
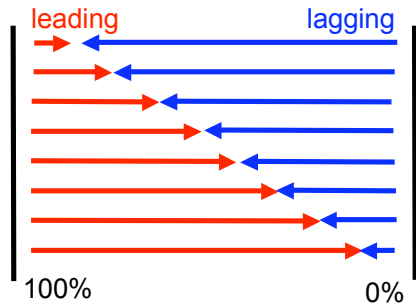


# Replication models that can generate N-shape of replication fork polarity

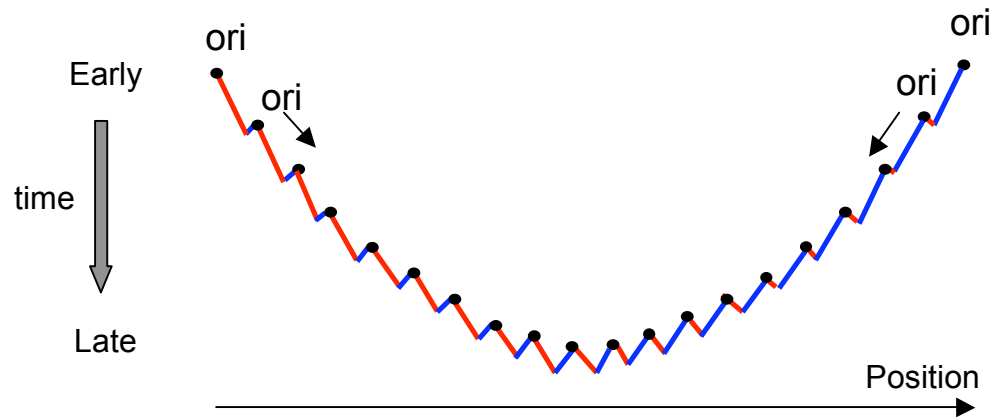
Single replication fork with random termination

Multiple internal replication origins with increase of fork speed and/or of initiation rate

gradient of replication fork polarity



To generate the "N" replication speed needs to be increase



**Genome-wide quantitative analysis of  
replicating DNA molecules stretched by DNA combing  
at different stages of the S phase.**

# DNA combing in HeLa cells

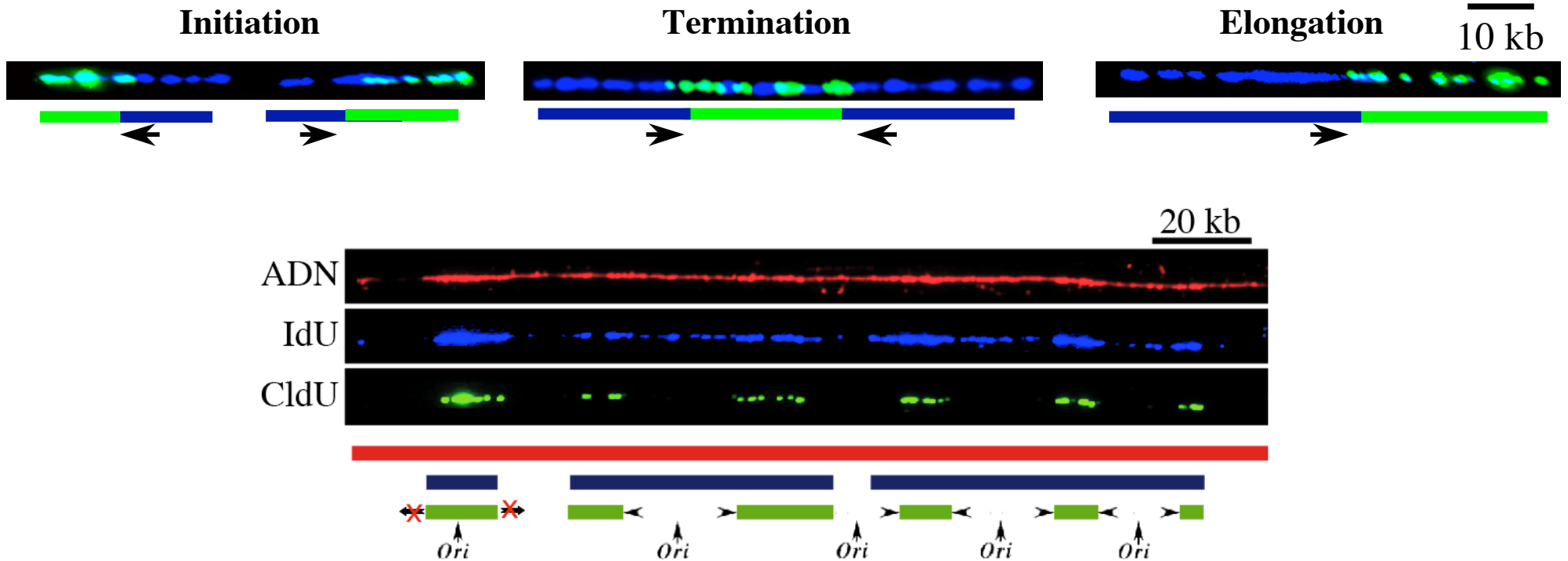
A. Rappailles, G. Guilbaud, O. Hyrien



- Replicative labeling (thymidine analog)



- Some observed replication patterns



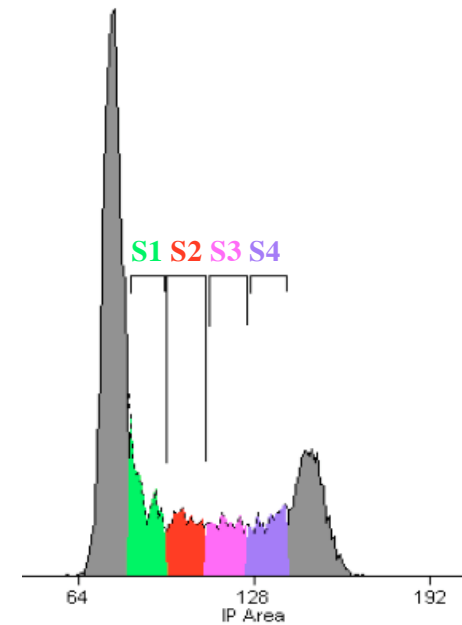
# DNA combing in HeLa cells

A. Rappailles, G. Guilbaud, O. Hyrien

## A quantitative genome-wide analysis by DNA combing in cells at different stages of the S phase

number of fork-containing fibers:

S1 ( $n=202$ ), S2 ( $n=202$ ), S3 ( $n=225$ ), S4 ( $n=203$ )



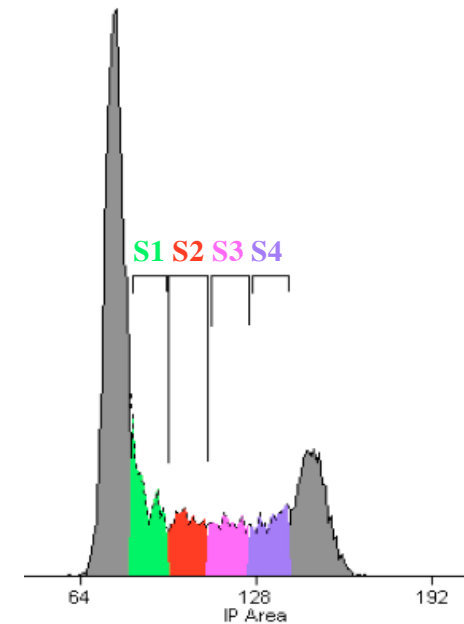
# DNA combing in HeLa cells

A. Rappailles, G. Guilbaud, O. Hyrien

## A quantitative genome-wide analysis by DNA combing in cells at different stages of the S phase

number of fork-containing fibers:

S1 ( $n=202$ ), S2 ( $n=202$ ), S3 ( $n=225$ ), S4 ( $n=203$ )



- Origins fire as clusters
- Inter-origin distances are  $\approx 40$  kb and constant through S phase

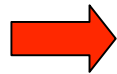
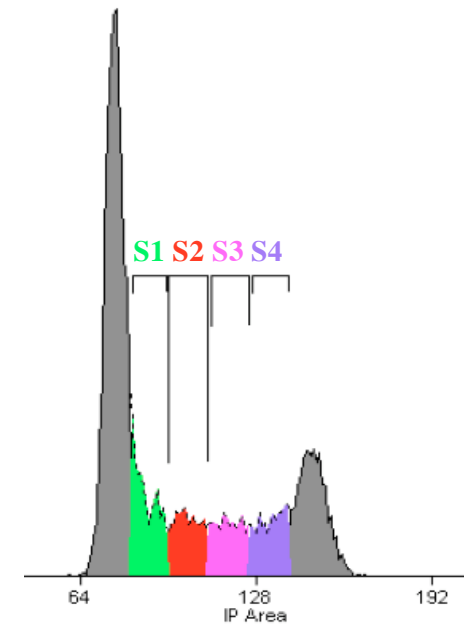
# DNA combing in HeLa cells

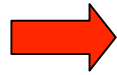
A. Rappailles, G. Guilbaud, O. Hyrien

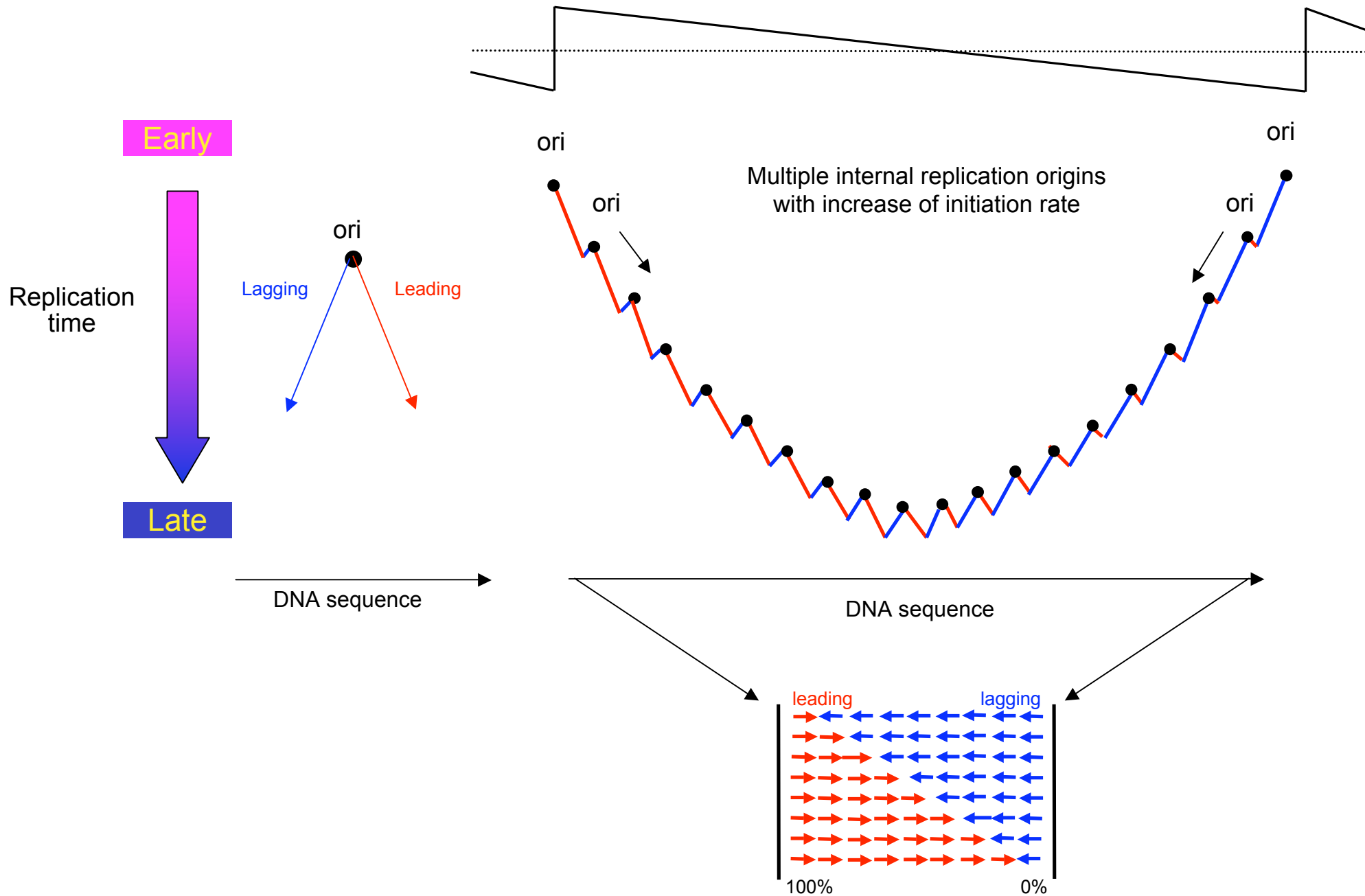
## A quantitative genome-wide analysis by DNA combing in cells at different stages of the S phase

number of fork-containing fibers:

S1 ( $n=202$ ), S2 ( $n=202$ ), S3 ( $n=225$ ), S4 ( $n=203$ )



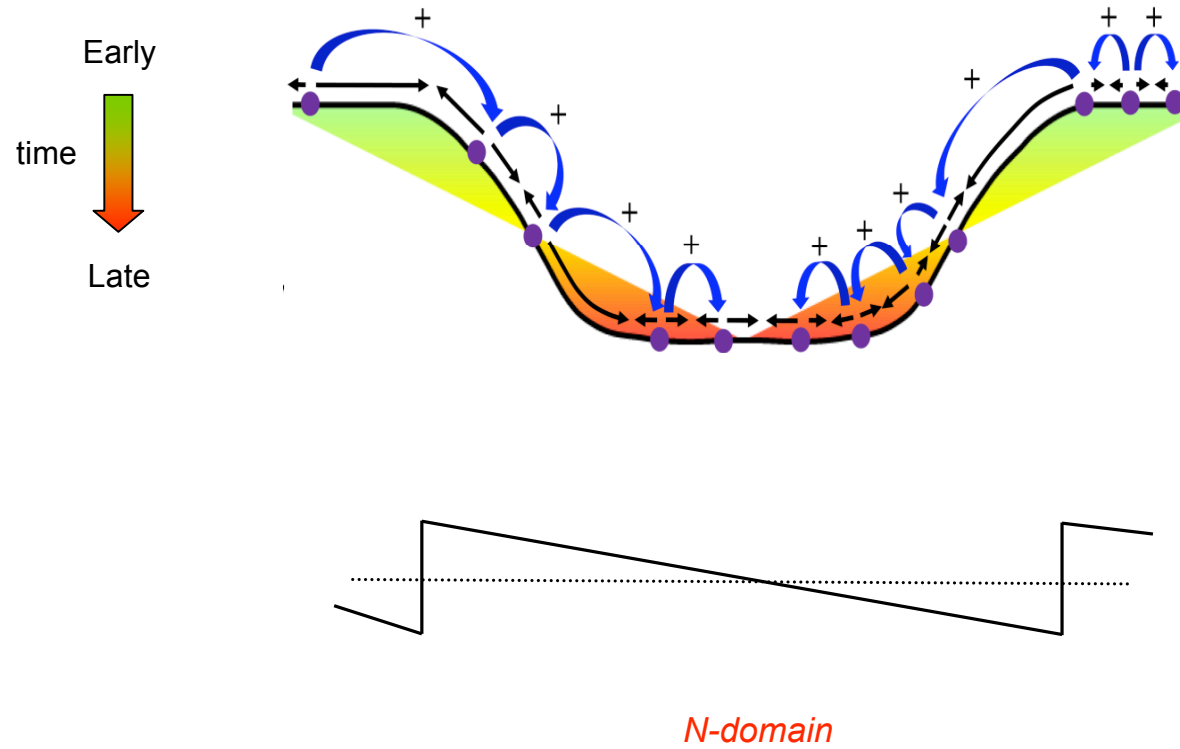
- Origins fire as clusters
  - Inter-origin distances are  $\approx 40$  kb and constant through S phase
- 
- Single replication fork speed is constant during S phase
  - The rate of initiation increases with more and more origins fire during S phase





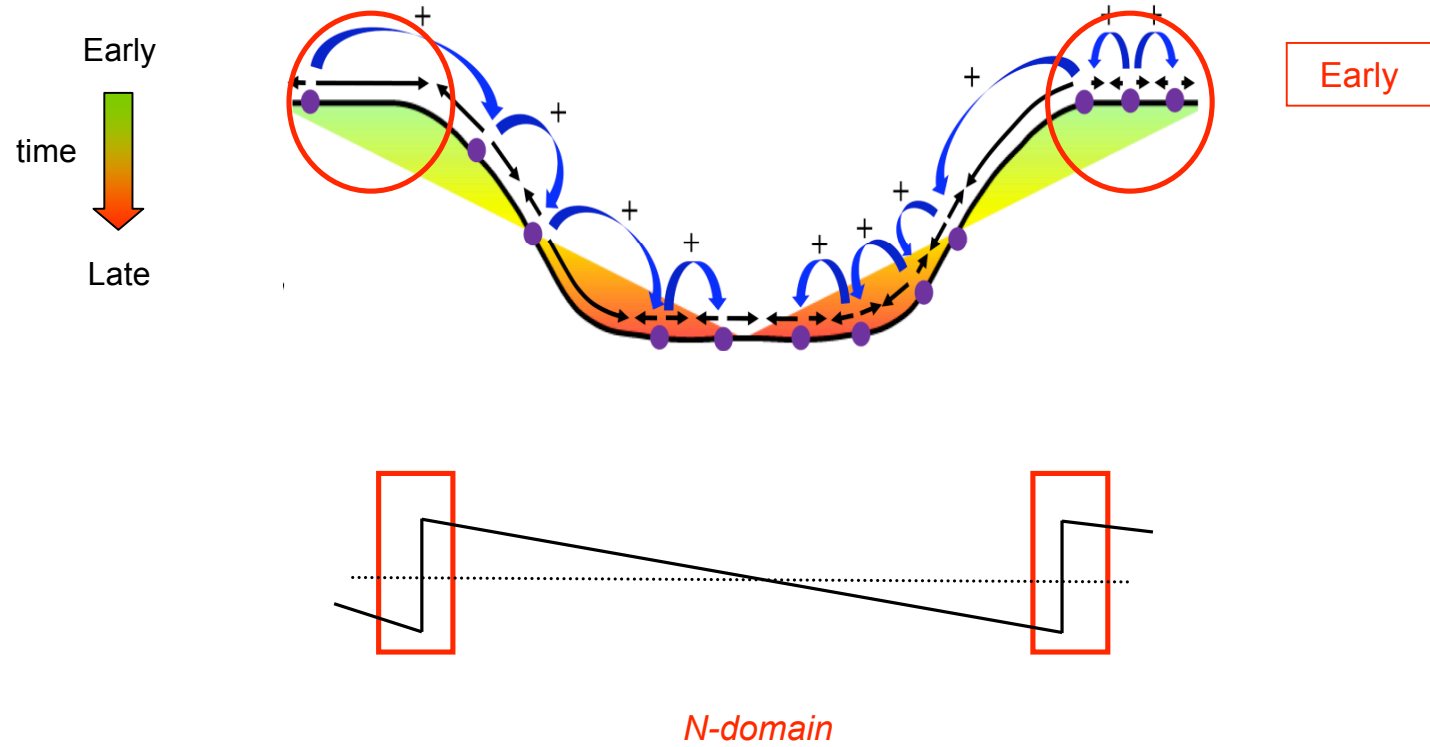
## Domino replication model

(based on our timing and DNA combing data)



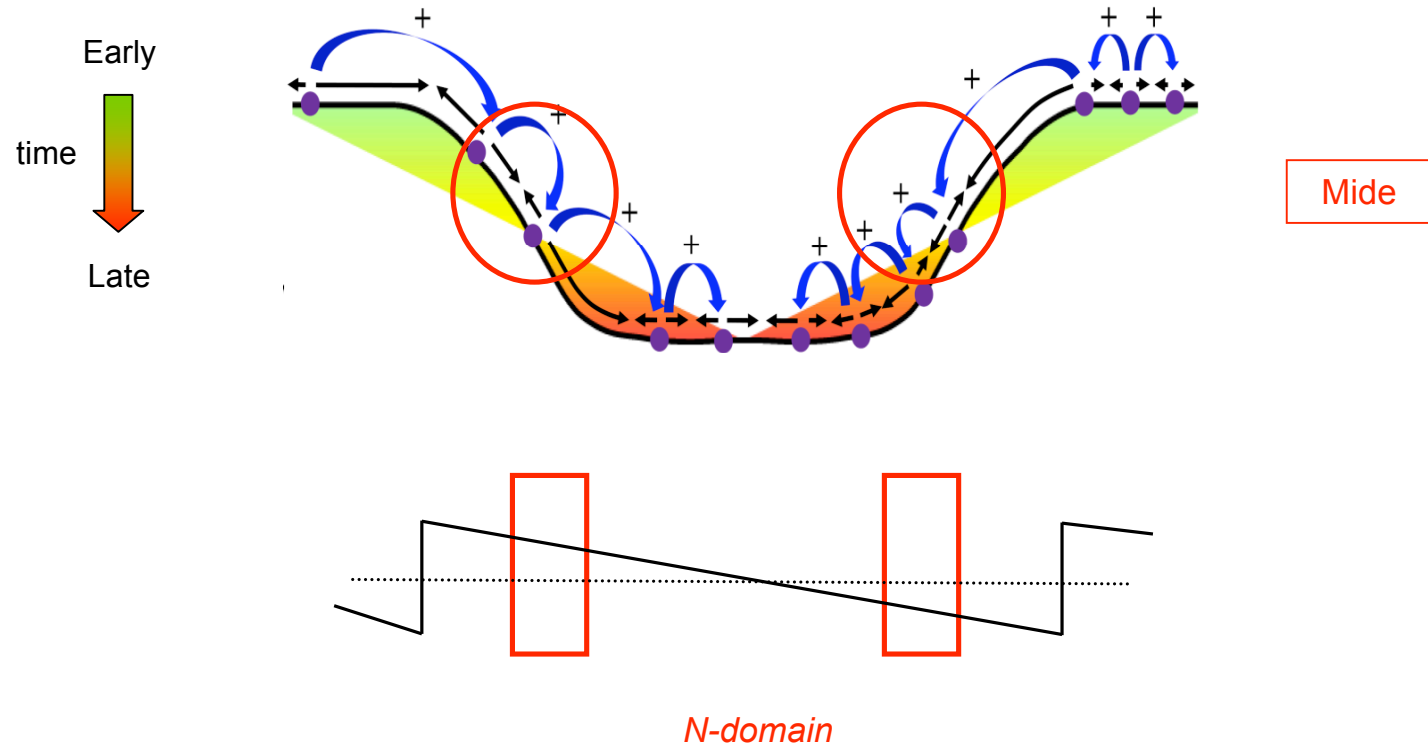
## Domino replication model

(based on our timing and DNA combing data)



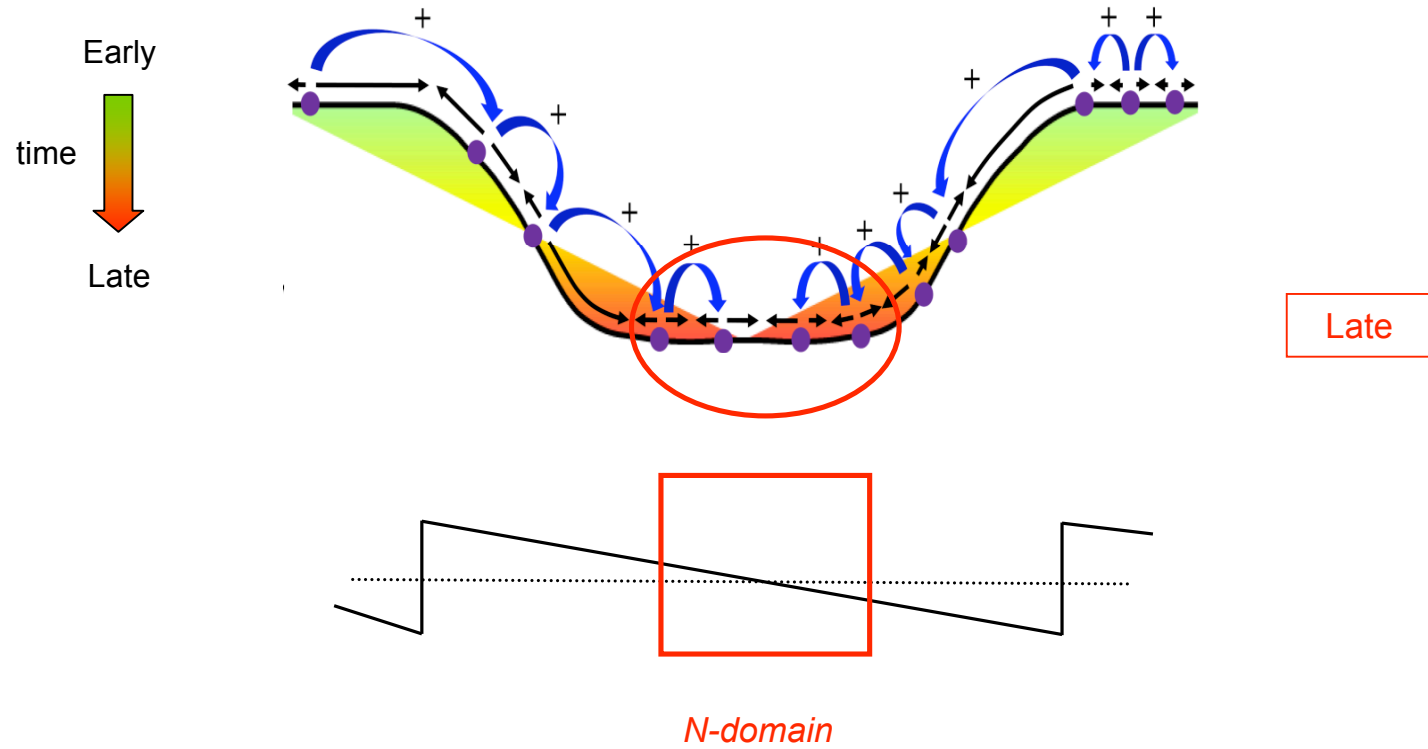
## Domino replication model

(based on our timing and DNA combing data)



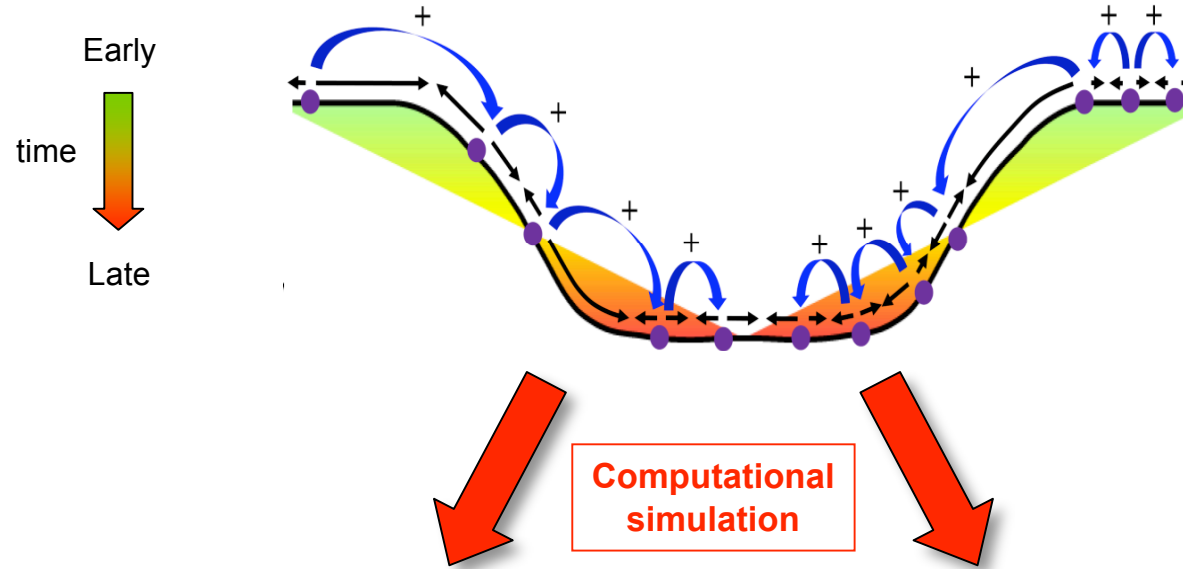
## Domino replication model

(based on our timing and DNA combing data)



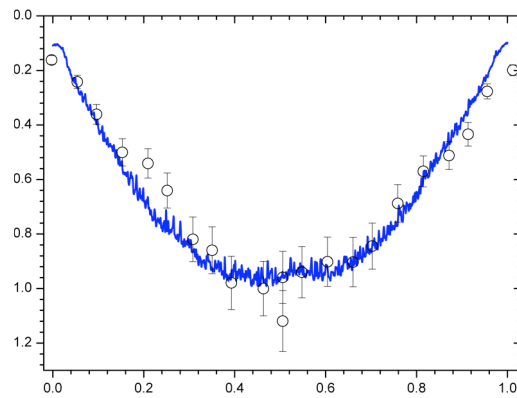
# Domino replication model

(based on our timing and DNA combing data)

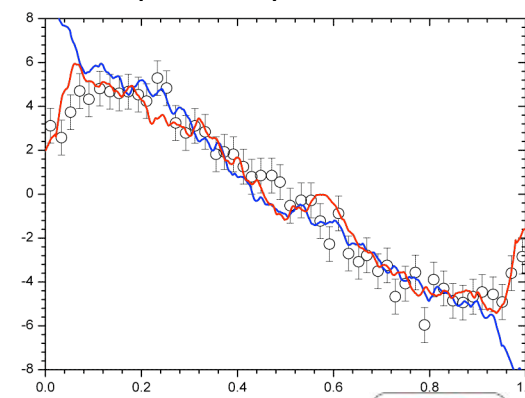


A. Goldar, O. Hyrien

U-shaped replication timing

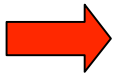
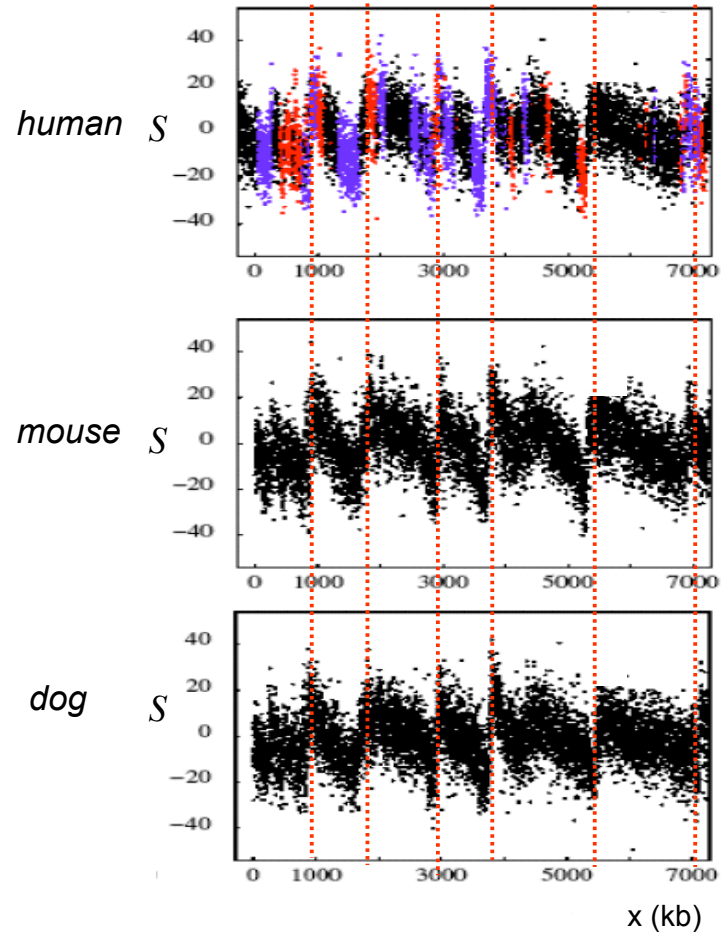


N-shaped compositional skew



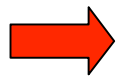
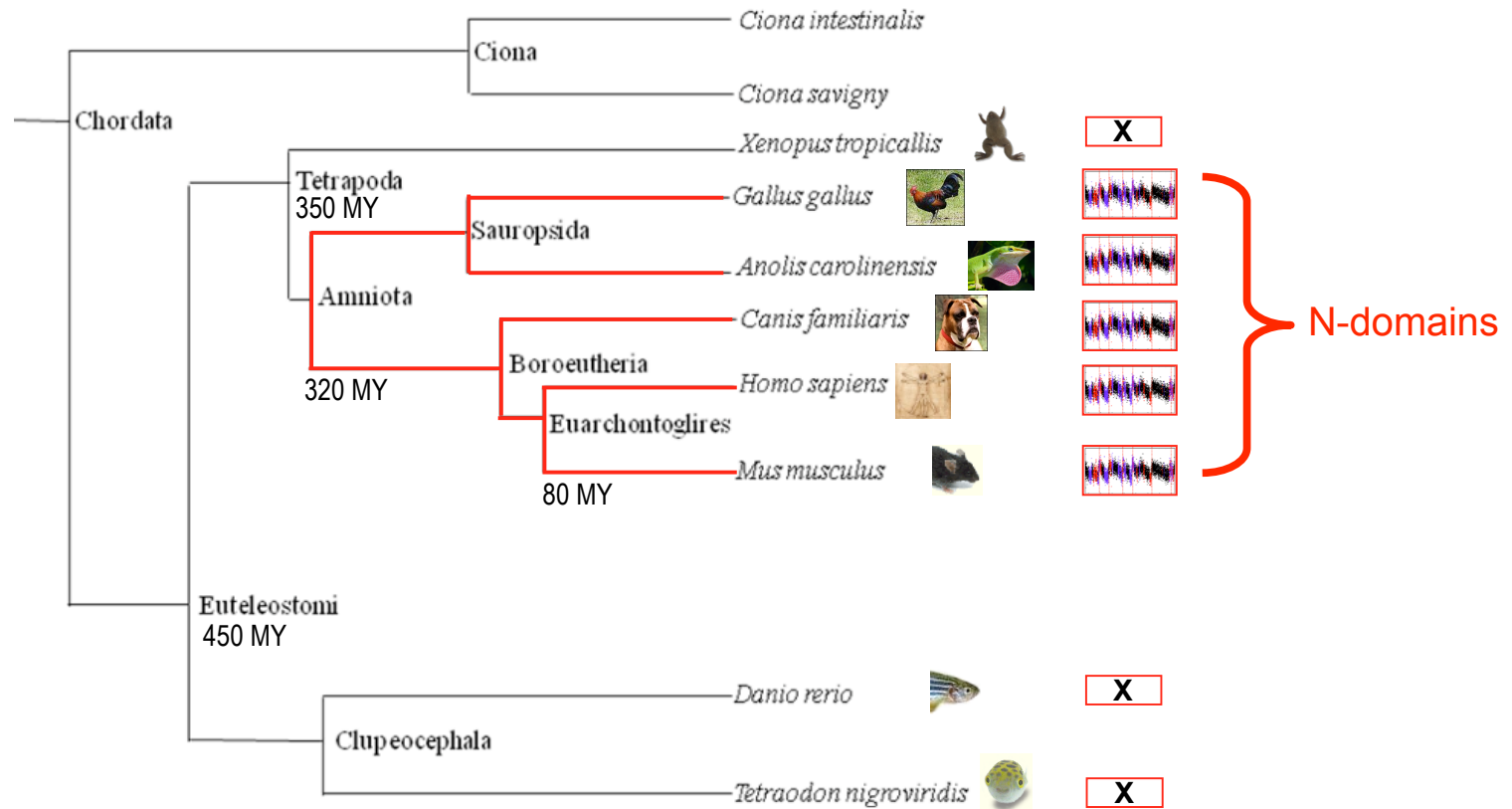
Unpublished

# Skew N-domains in mammalian genomes



The N-domains are conserved during mammalian evolution.

# N-domains are 320 million years old



**This replication program has been conserved since amniota divergence**



Unpublished

# Conclusions

- Skew N-domains correspond to **U-shaped timing domains of germline cells**
- N-shape of skew profile is generated by **gradient of replication fork polarity**
- We construct a **domino-model** of replication : replication initiates at master origins and propagates by cascade of secondary initiations associated with **a gradient of open chromatin structure**
- This replication program has been **conserved during mammalian evolution**

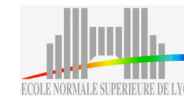


# Acknowledgments and collaborations

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

Arach Goldar  
(CEA, iBiTec-S, Gif-sur-Yvette)  
Guillaume Guilbaud  
Aurélien Rappailles  
Olivier Hyrien  
(ENS-Paris)

Benjamin Audit  
Antoine Baker  
Alain Arneodo  
(ENS-Lyon)



Merci!

謝謝



