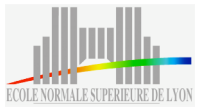
Replication-associated mutational strand asymmetry in the human genome

Chun-Long Chen[1], Benjamin Audit[2], Lauranne Duquenne[1], Guillaume Guilbaud[3], Aurélien Rappailles[3], Yves d'Aubenton-Carafa[1], Olivier Hyrien[3], Alain Arneodo[2], & Claude Thermes[1]

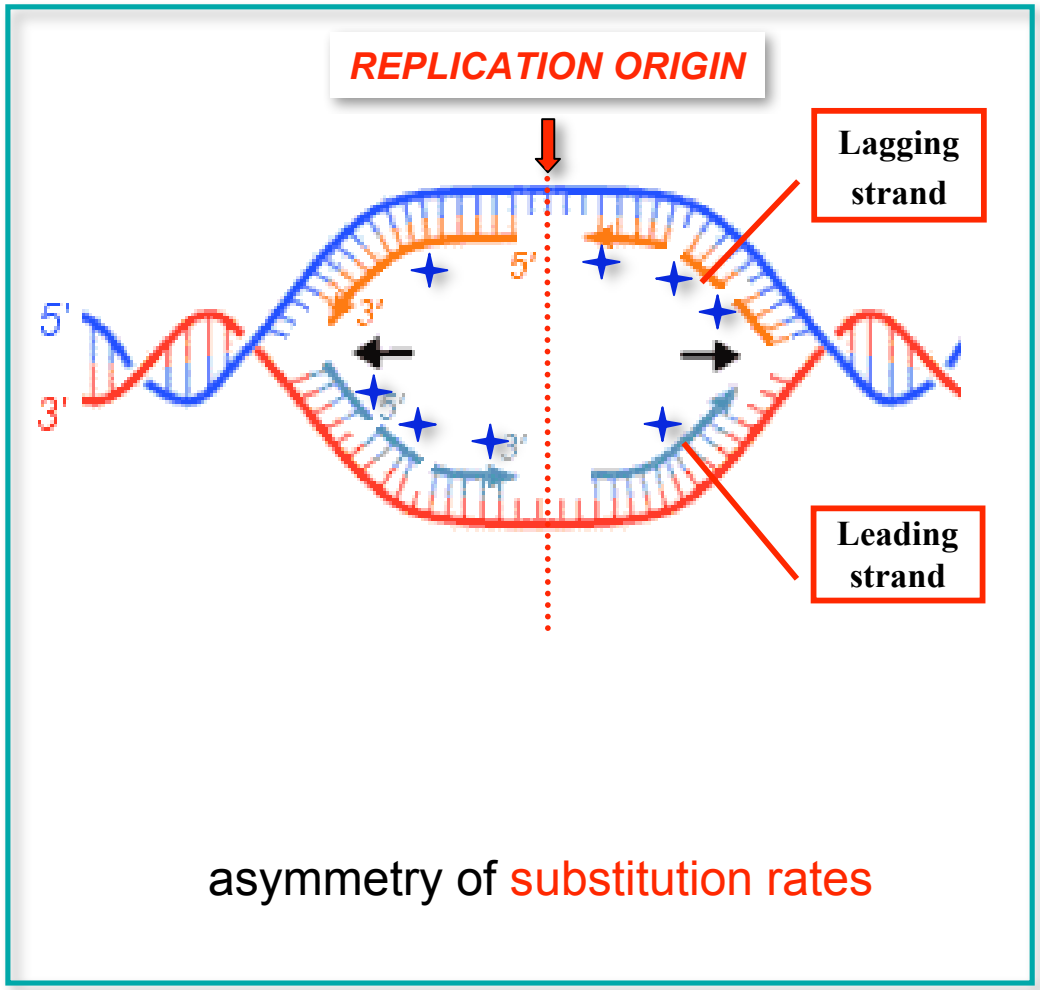
[1] Centre de Génétique Moléculaire, CNRS, Gif-sur-Yvette
[2] Laboratoire Joliot Curie et Laboratoire de Physique, Ecole Normale Supérieure de Lyon
[3] Ecole Normale Supérieure de Paris



07/09/10, Jobim 2010, Montpellier

INTRODUCTION

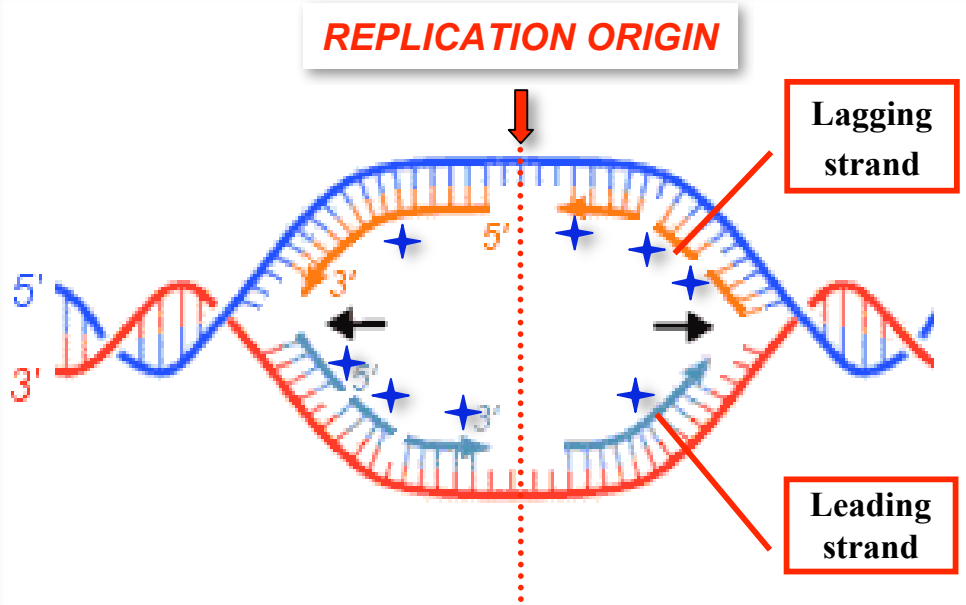
Prokaryotes: mutation rates differ between leading and lagging strands



asymmetry of substitution rates

Lobry JR. *MBE*. (1996) 13:660

Prokaryotes:



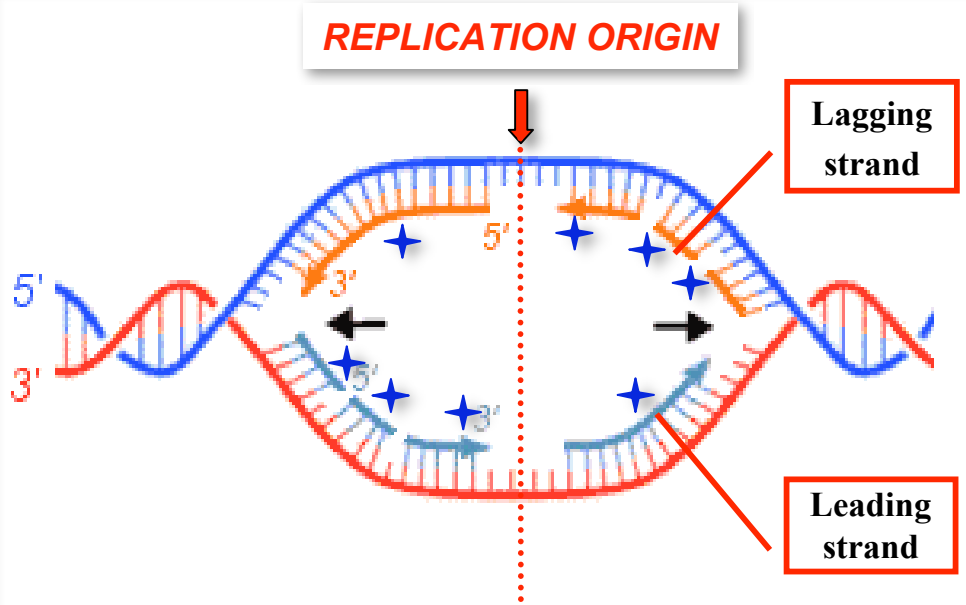
asymmetry of substitution rates

G > C and T > A on the leading strand
G < C and T < A on the lagging strand

asymmetry of nucleotide composition

Lobry JR. *MBE*. (1996) 13:660

Prokaryotes:



asymmetry of substitution rates

Compositional skew

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

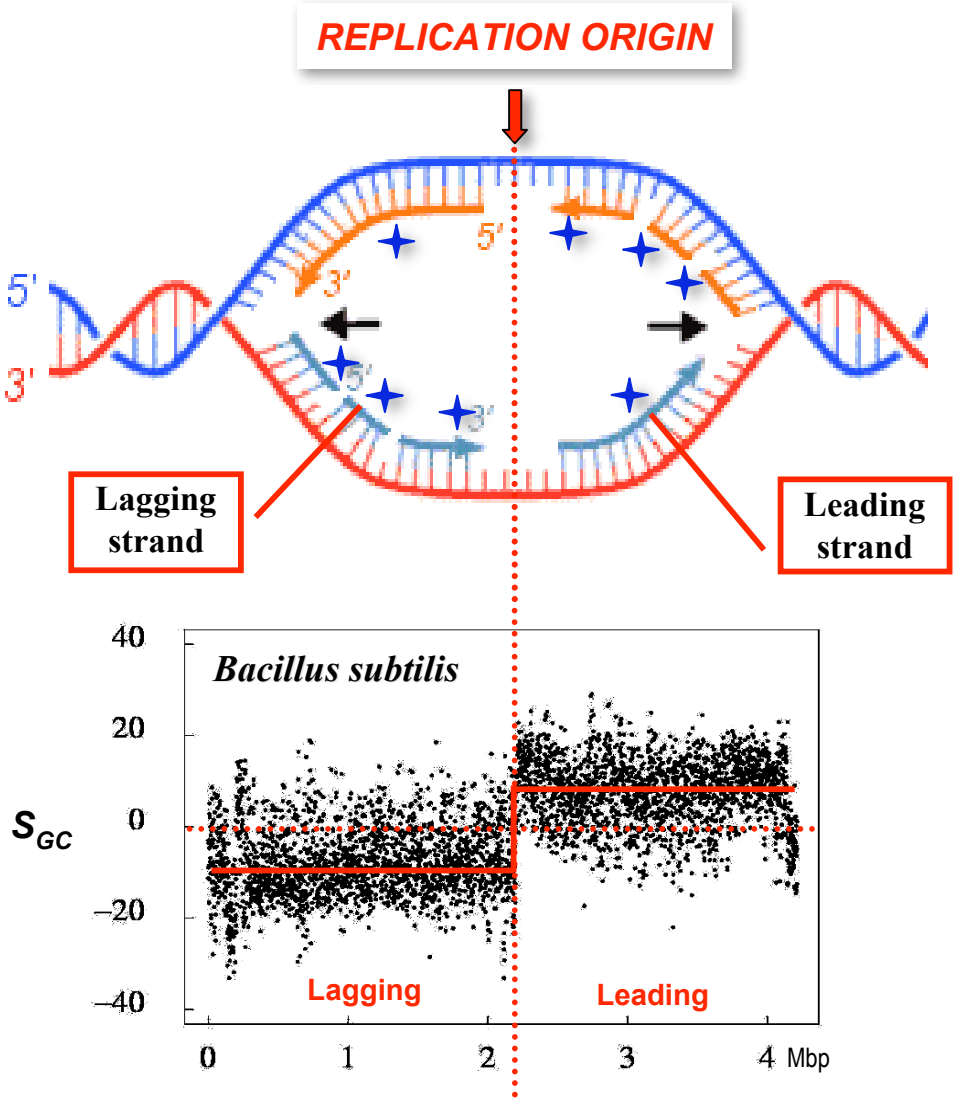
$$S_{TA} = \frac{n_T - n_A}{n_T + n_A}$$

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

asymmetry of nucleotide composition

Lobry JR. *MBE*. (1996) 13:660

Prokaryotes:

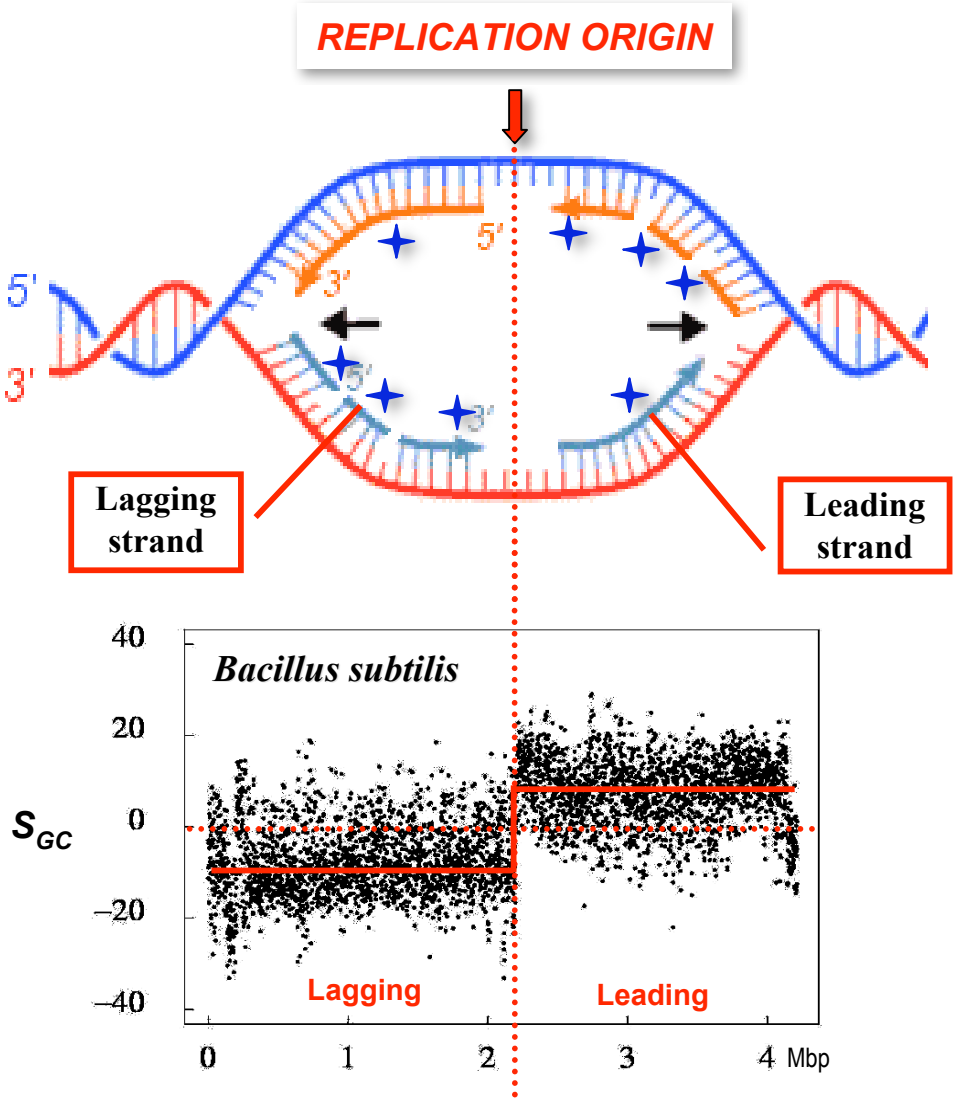


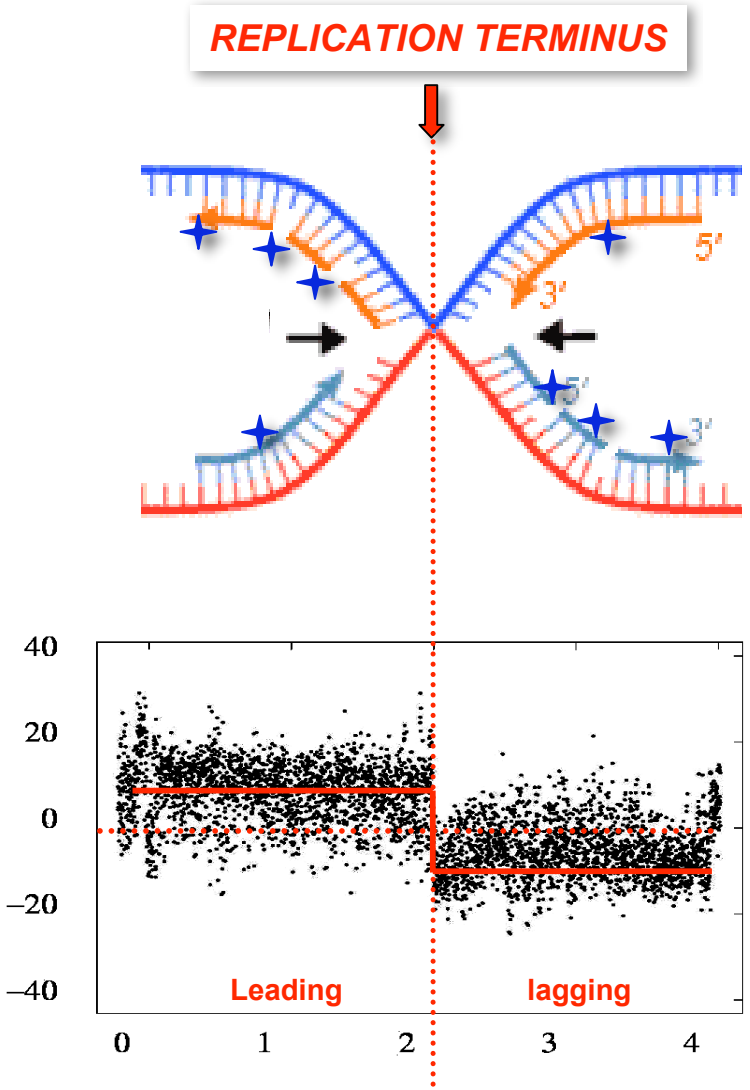
Replication origin

⇨ **upward jump in the skew profile**

Lobry JR. *MBE*. (1996) 13:660

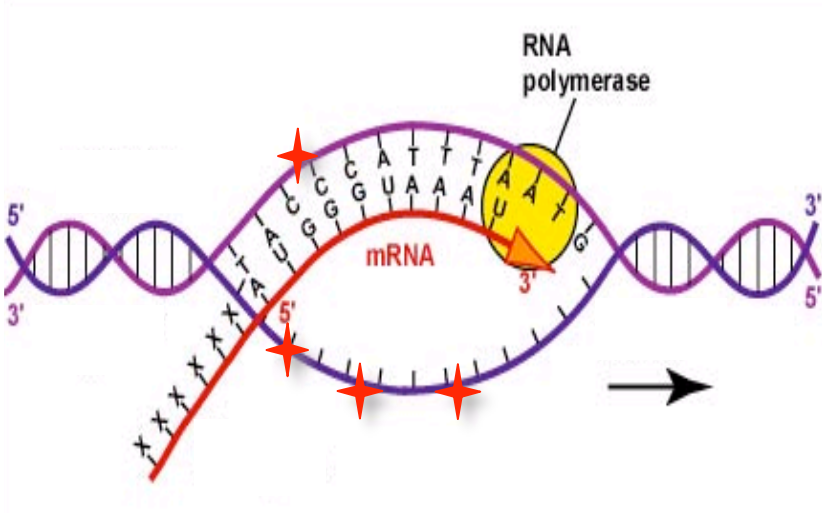
Prokaryotes:





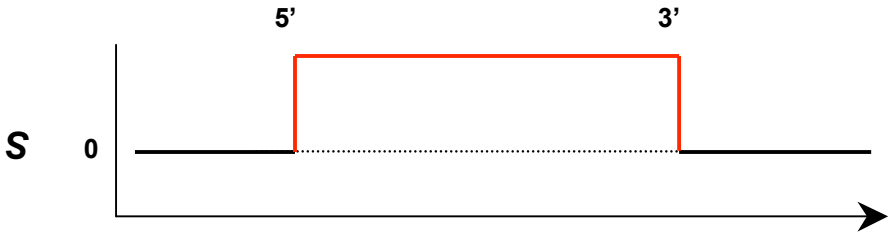
Lobry JR. *MBE*. (1996) 13:660

Prokaryotes:



asymmetry of substitution rates

between transcribed and non-transcribed strands



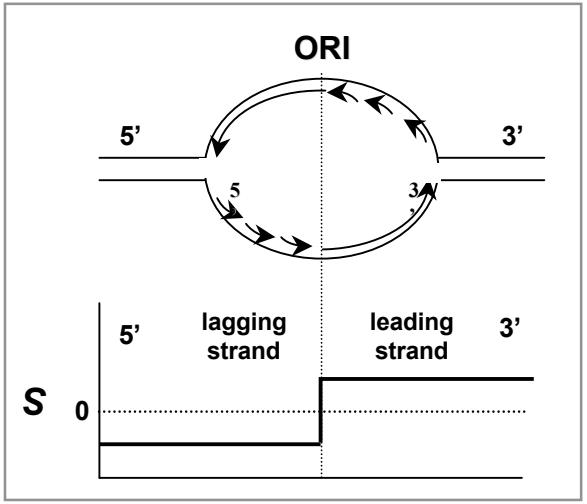
asymmetry of nucleotide composition

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

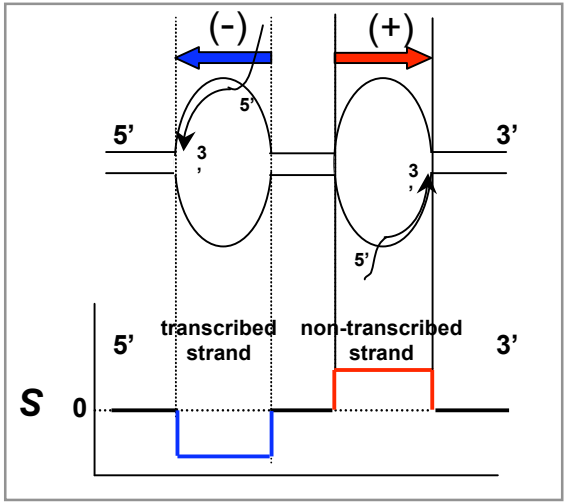
Beletskii A. *Biol.Chem*, (1998) 379:549

Total skew = superposition of skews due to replication and transcription

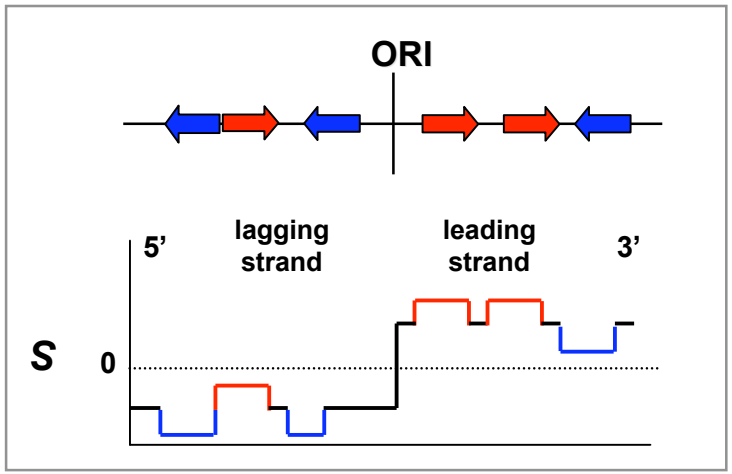
Replicative skew profile



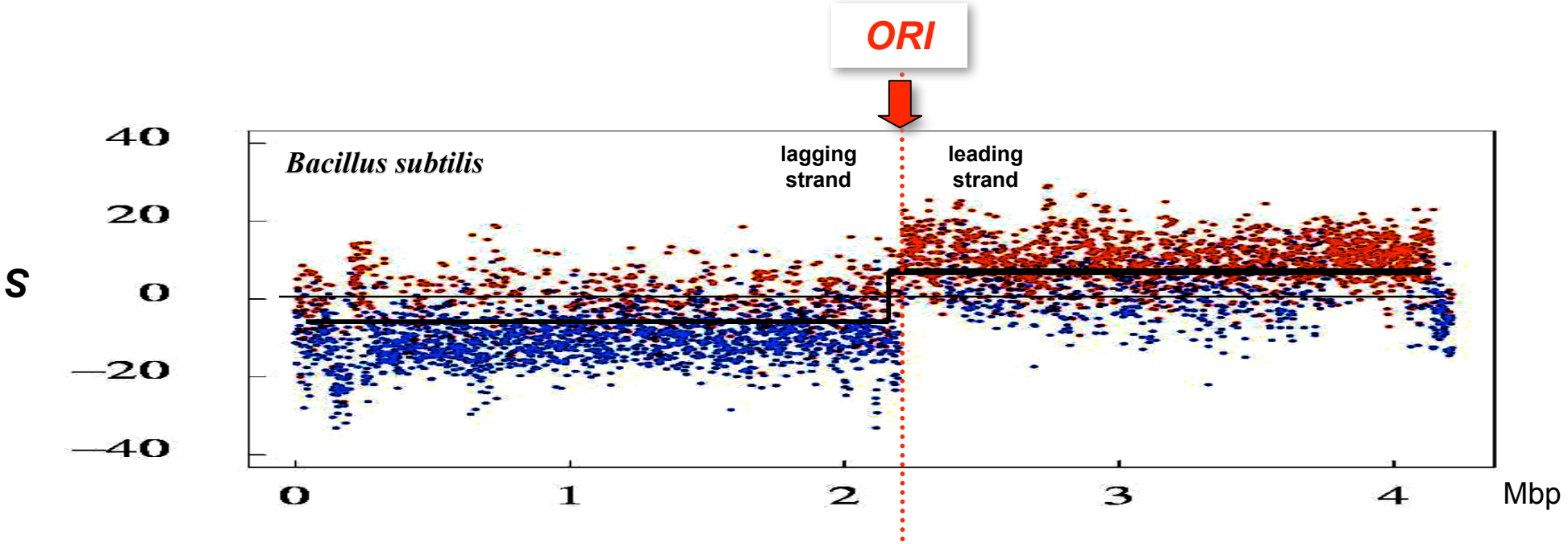
Transcriptional skew profile

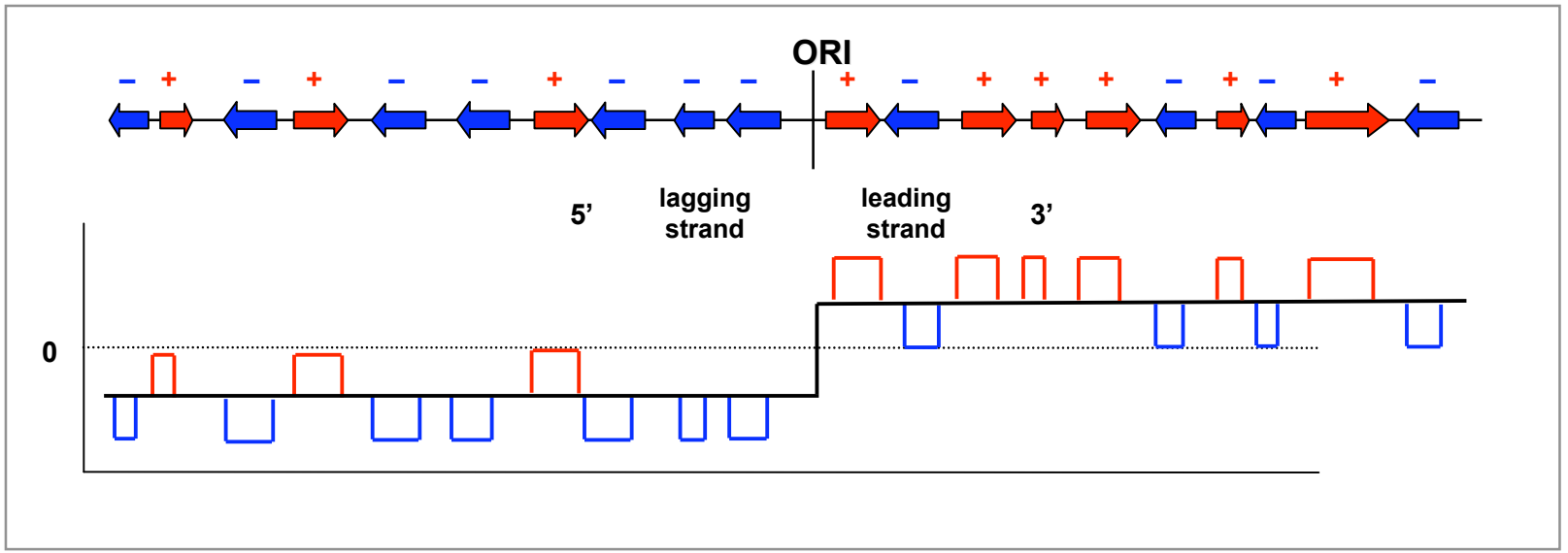


Superposition of replication and transcription



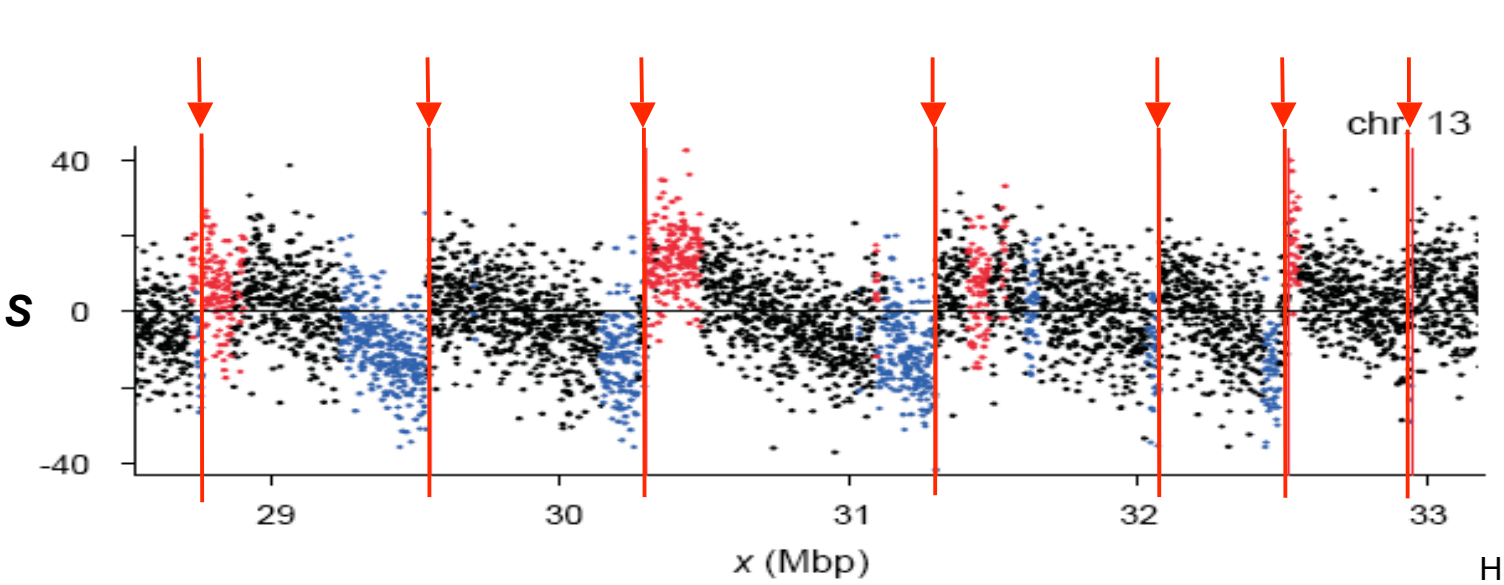
Total skew = superposition of skews due to replication and transcription

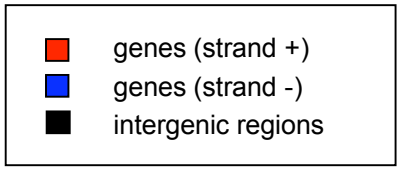




COMPOSITIONAL SKEWS IN THE HUMAN GENOME

COMPOSITIONAL SKEWS IN THE HUMAN GENOME

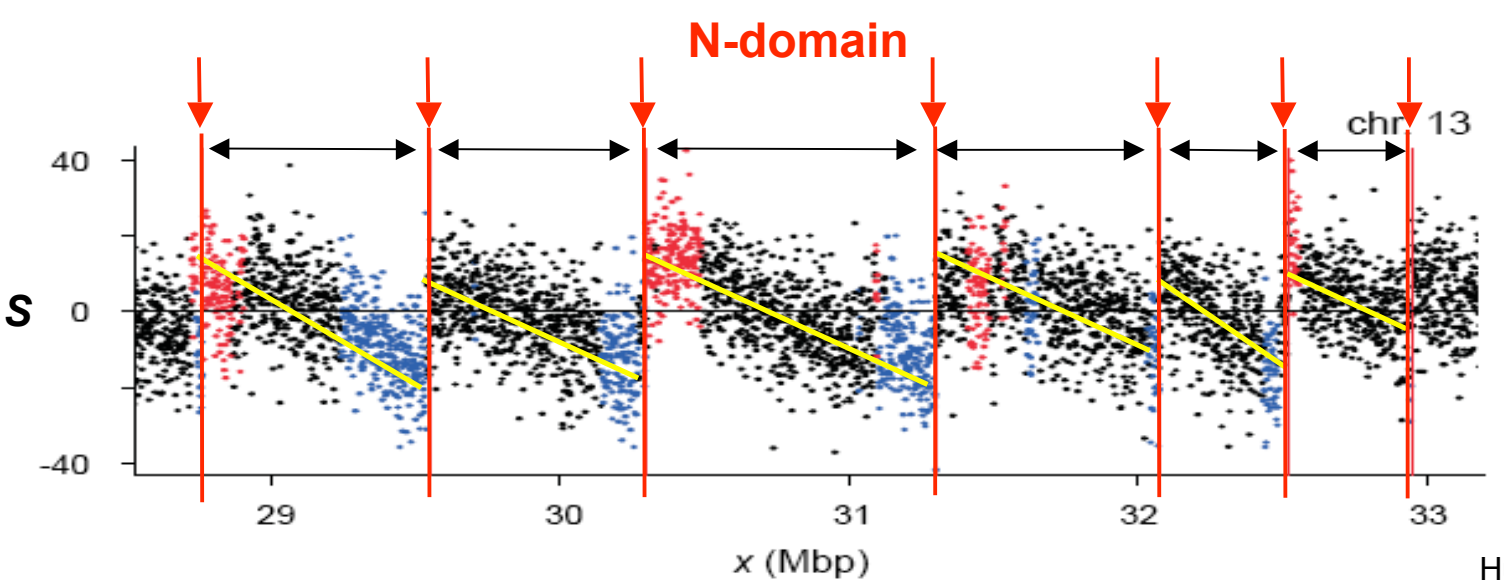




1546 upward jumps were detected

Touchon M. et al. *PNAS*. (2005) 102:9836
Huvet M. et al. *Genome Res*. (2008) 17:1278

COMPOSITIONAL SKEWS IN THE HUMAN GENOME



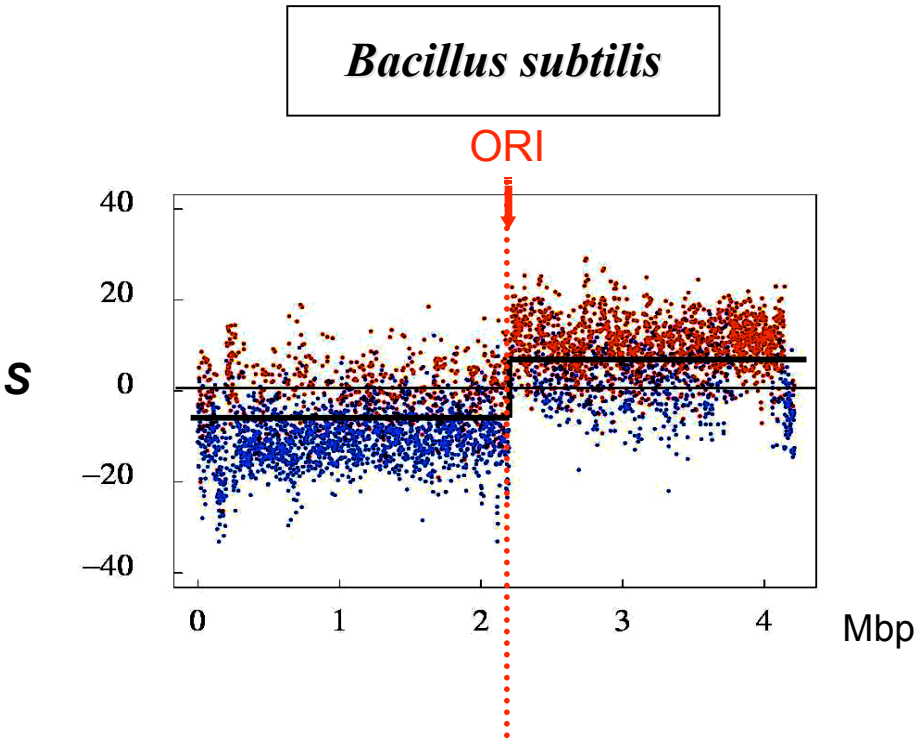
genes (strand +)
genes (strand -)
intergenic regions

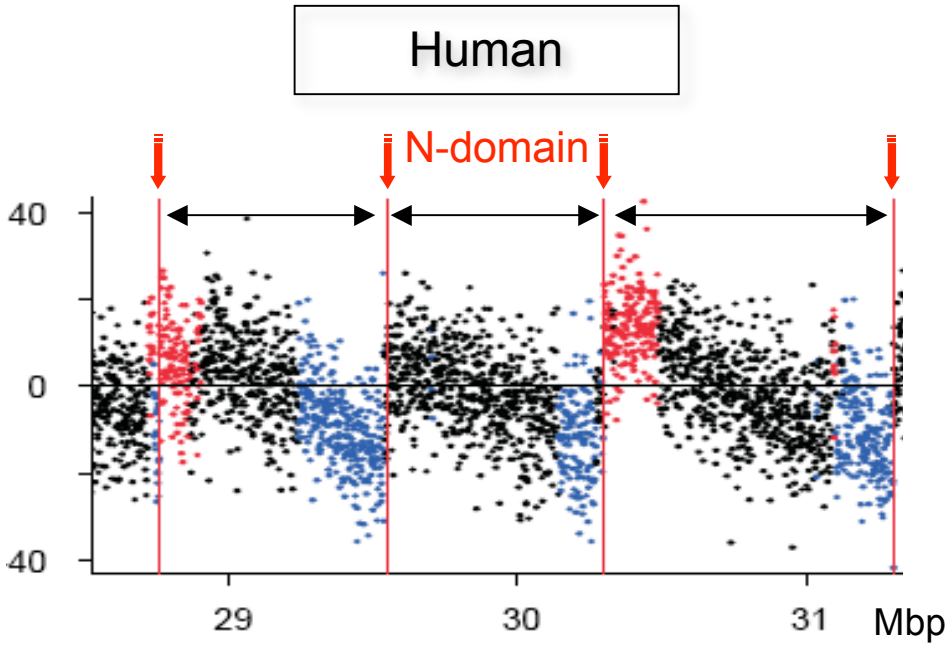
1546 Upward jumps were detected

Touchon M. et al. *PNAS*. (2005) 102:9836
Huvet M. et al. *Genome Res*. (2008) 17:1278

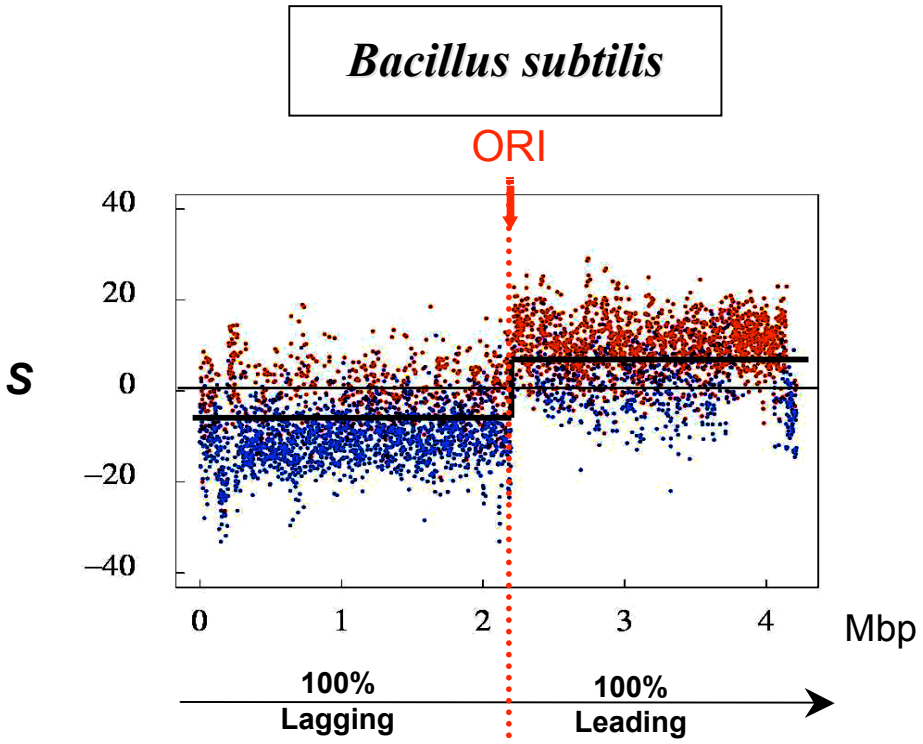
length between 0.5-2 Mbp covering >1/3 of the genome

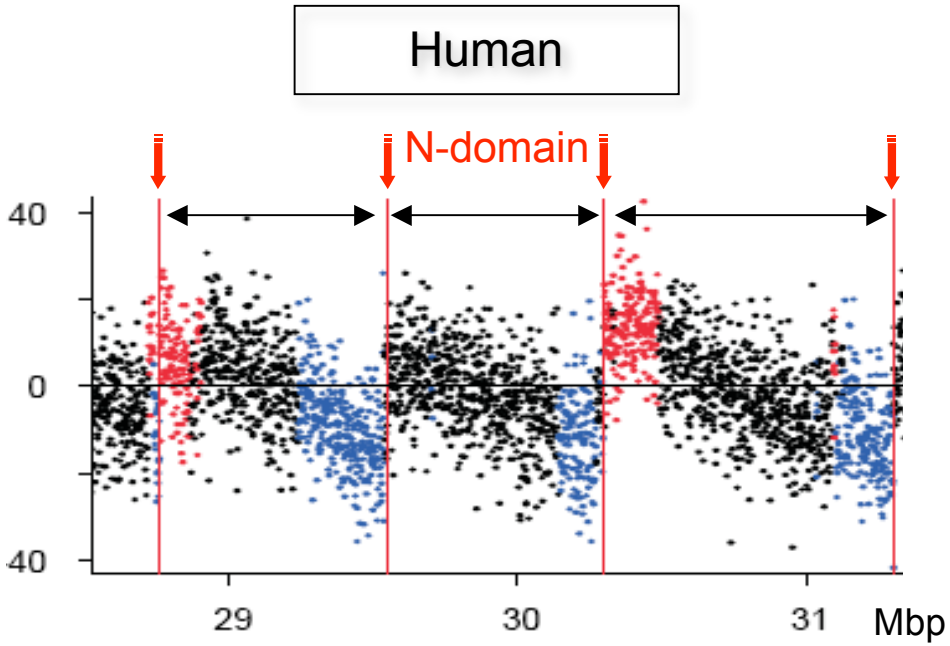
COMPOSITIONAL SKEWS IN THE HUMAN GENOME

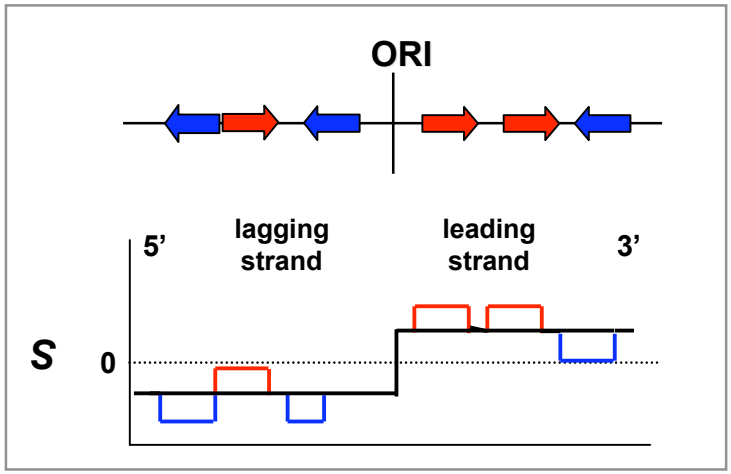




COMPOSITIONAL SKEWS IN THE HUMAN GENOME

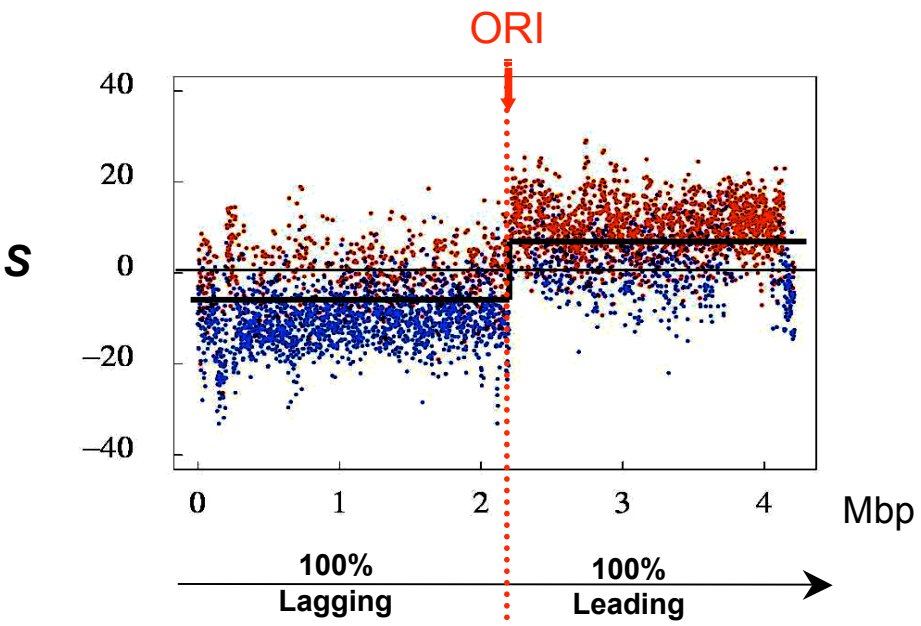




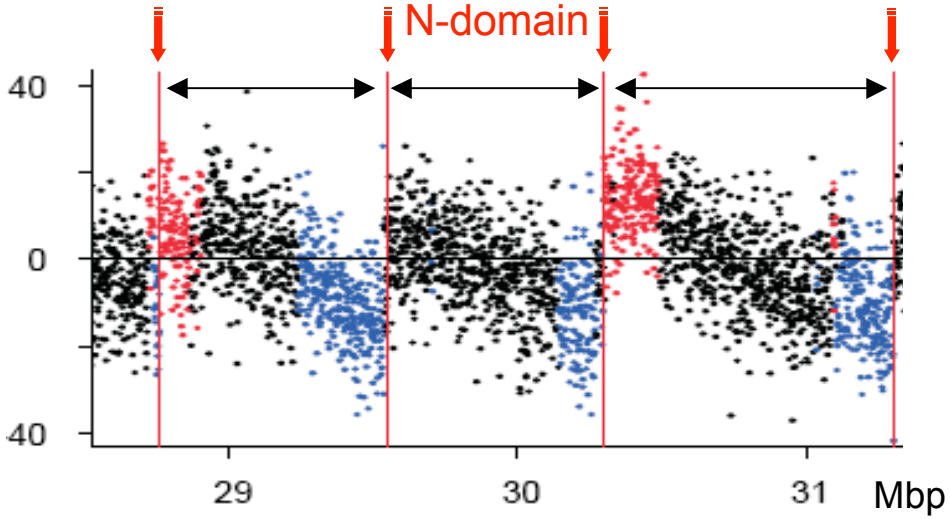


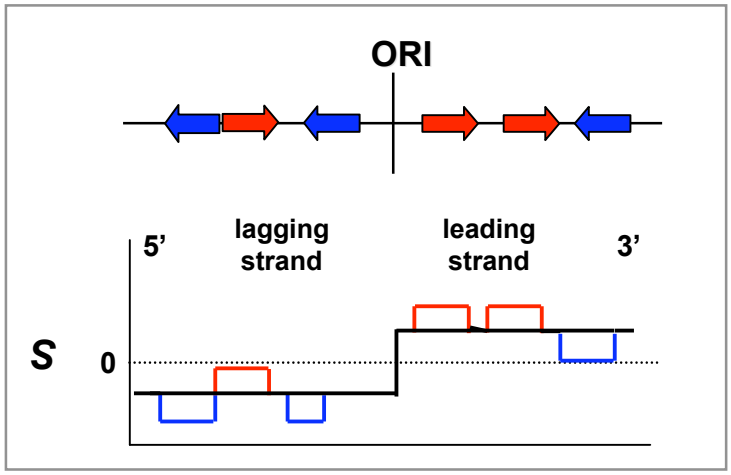
COMPOSITIONAL SKEWS IN THE HUMAN GENOME

Bacillus subtilis



Human



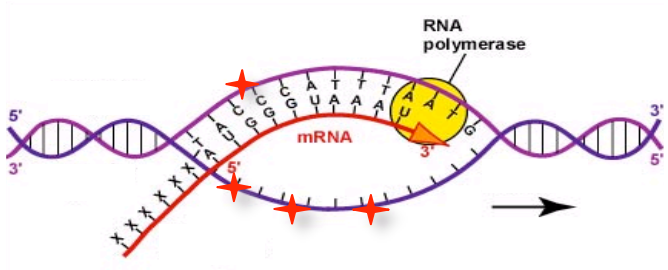


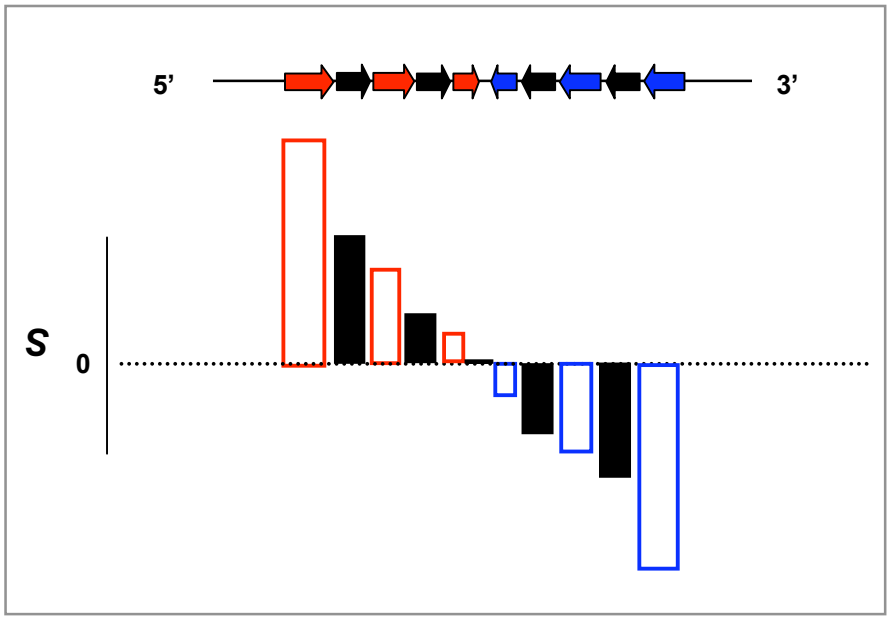
?

COMPOSITIONAL SKEWS IN THE HUMAN GENOME

MODEL 1

N-PATTERN DUE TO TRANSCRIPTION



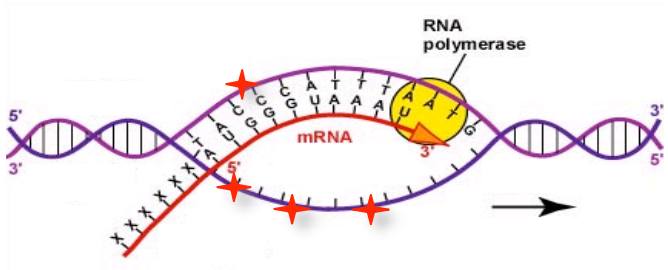


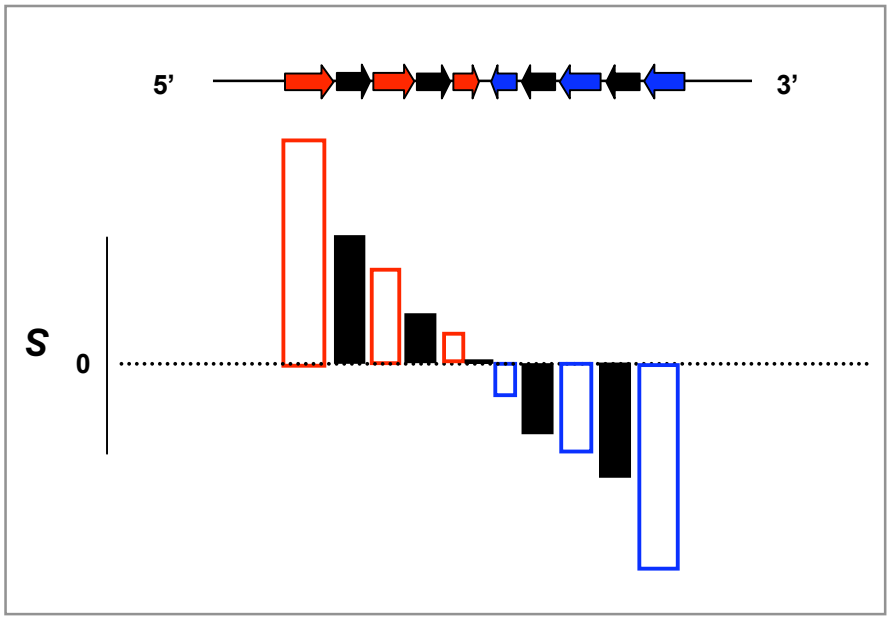
Transcription specifically organized

COMPOSITIONAL SKEWS IN THE HUMAN GENOME

MODEL 1

N-PATTERN DUE TO TRANSCRIPTION

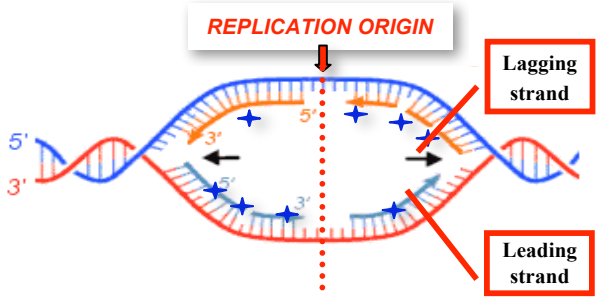


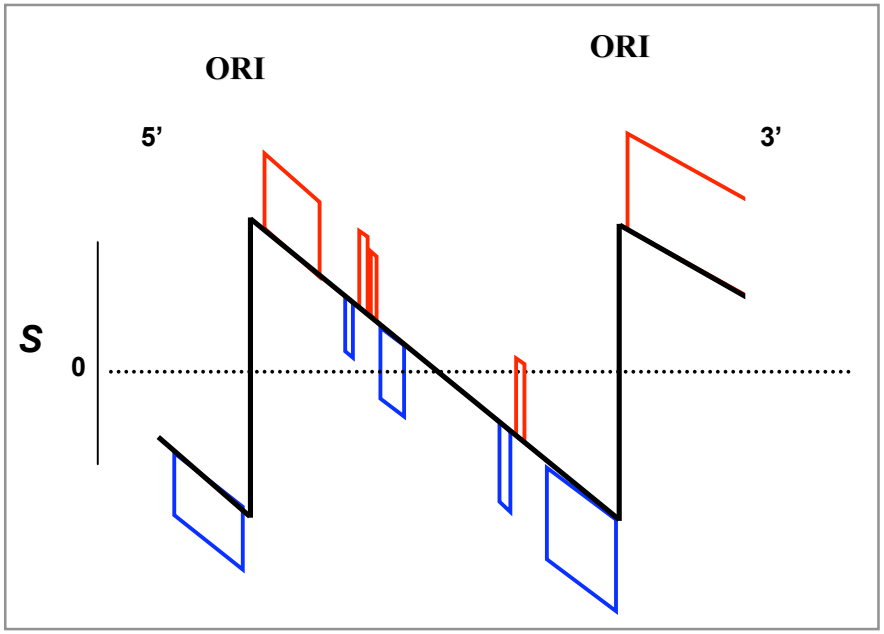


Transcription specifically organized

MODEL 2

N-PATTERN DUE TO REPLICATION



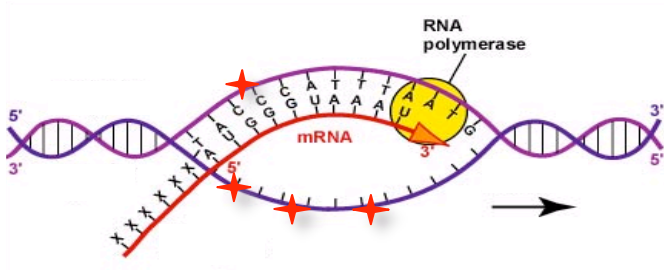


Replication specifically organized

COMPOSITIONAL SKEWS IN THE HUMAN GENOME

MODEL 1

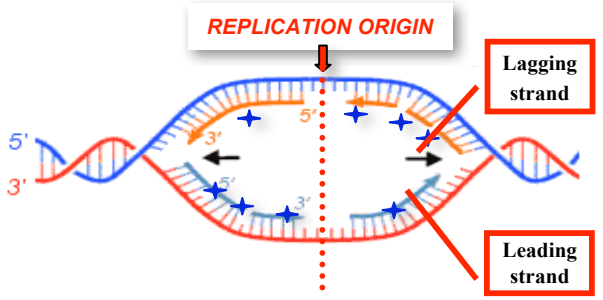
N-PATTERN DUE TO TRANSCRIPTION

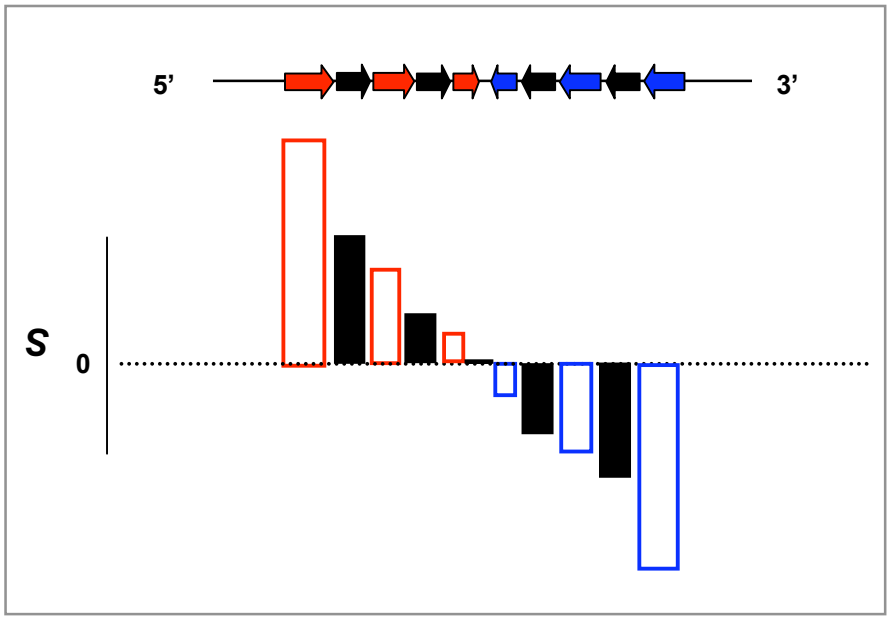


?

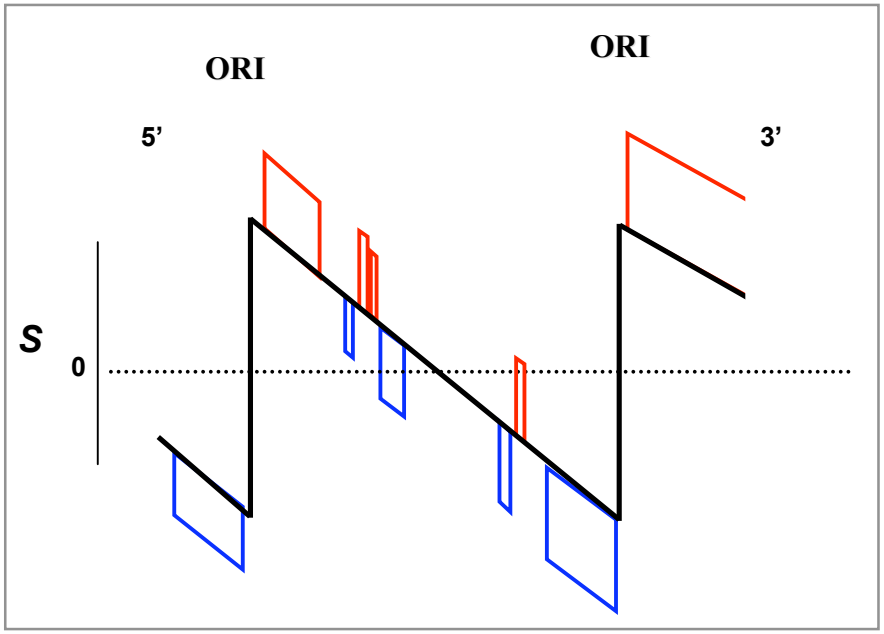
MODEL 2

N-PATTERN DUE TO REPLICATION





Transcription specifically organized

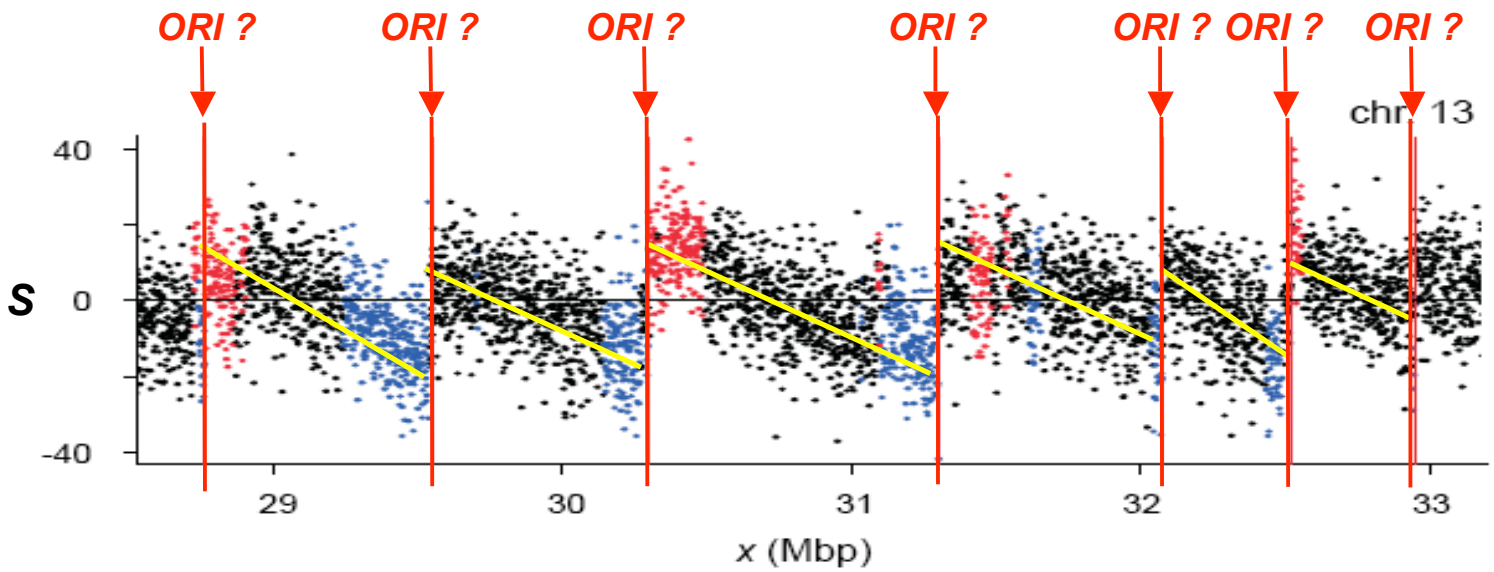


Replication specifically organized

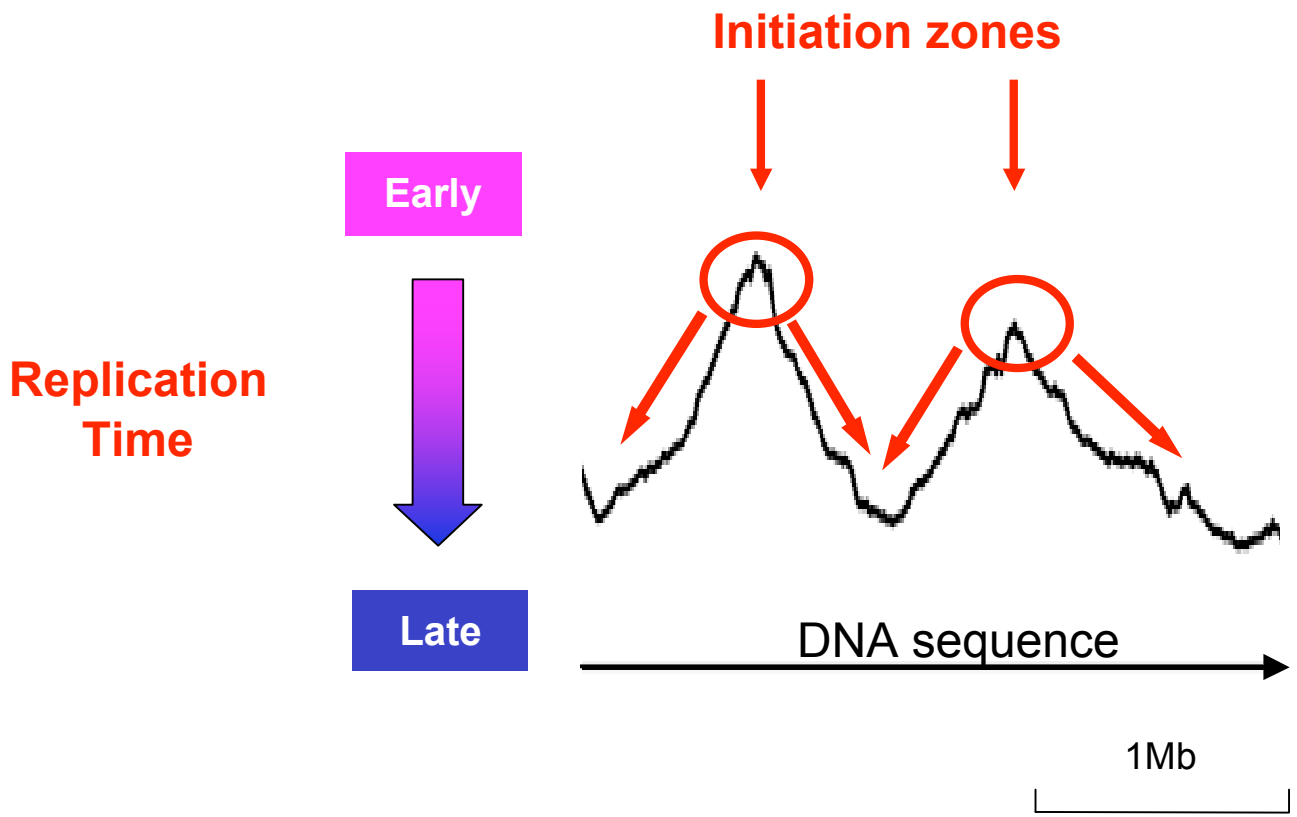
RESULTS

QUESTION 1:

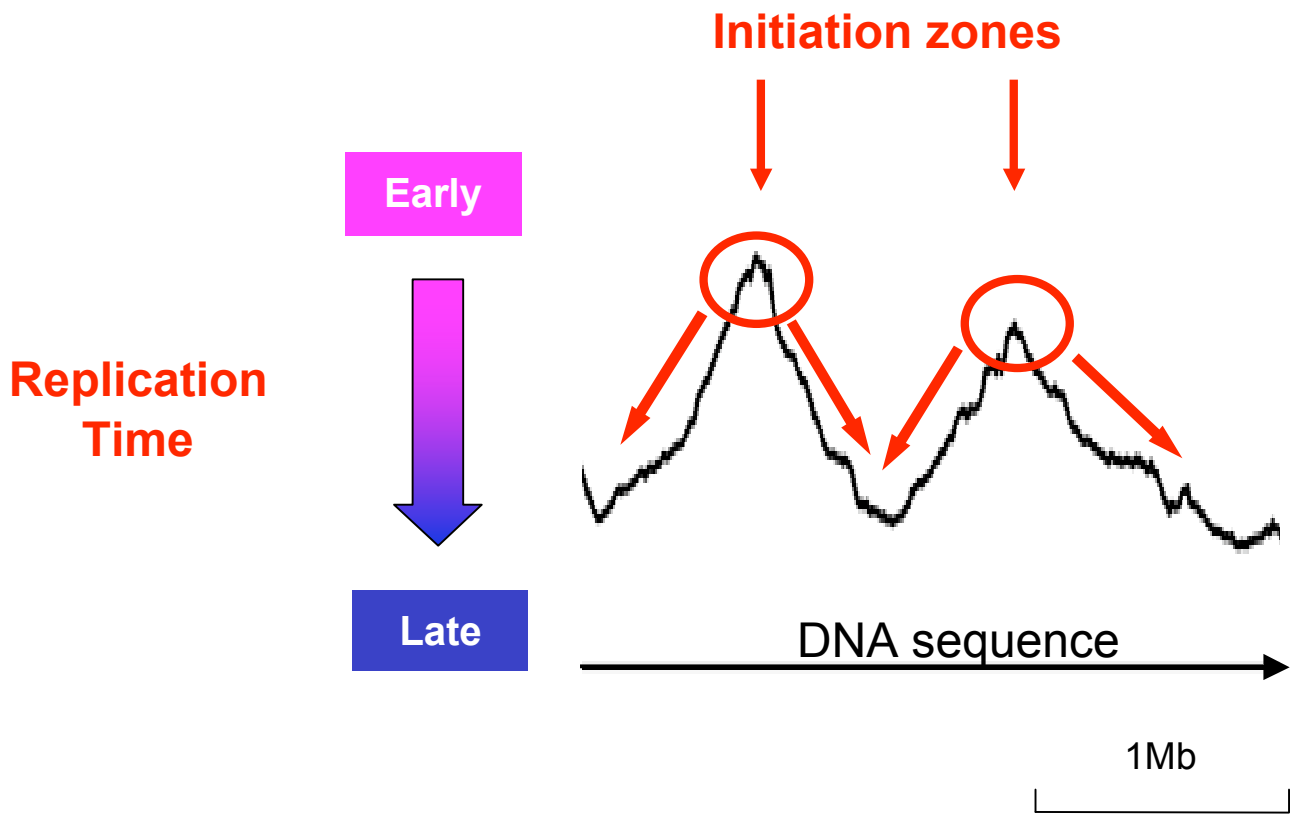
Are the upward jumps associated with replication origins?



REPLICATION TIMING PROFILE



GERMLINE CELLS



Determination of a human replication timing profile

A. Rappailles, G. Guilbaud, O. Hyrien

Asynchronous HeLa cells

↓

Incorporation BrdU+Fixation+Cell sorting

↙ ↘

S1, S2, S3, S4 | S (Control)

↓ ↓

DNA Extraction+Sonication+Denature

↓ ↓

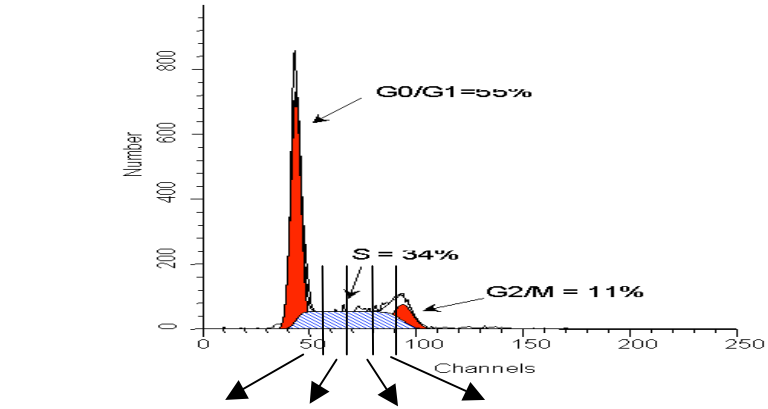
Immuno-precipitation

↓ ↓

Random priming 10 minutes
(generate double strand)

↓ ↓

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

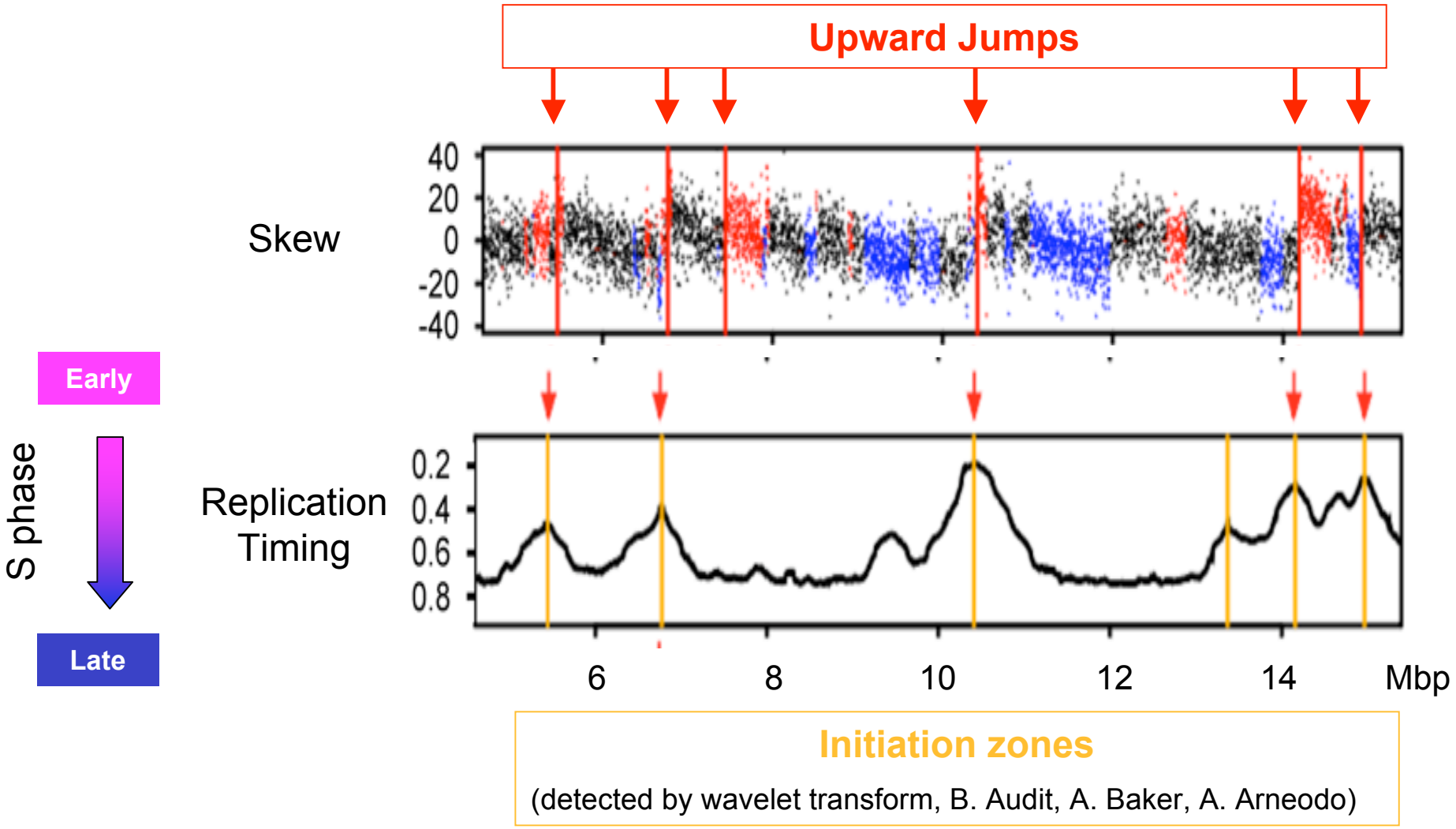


Early → Late

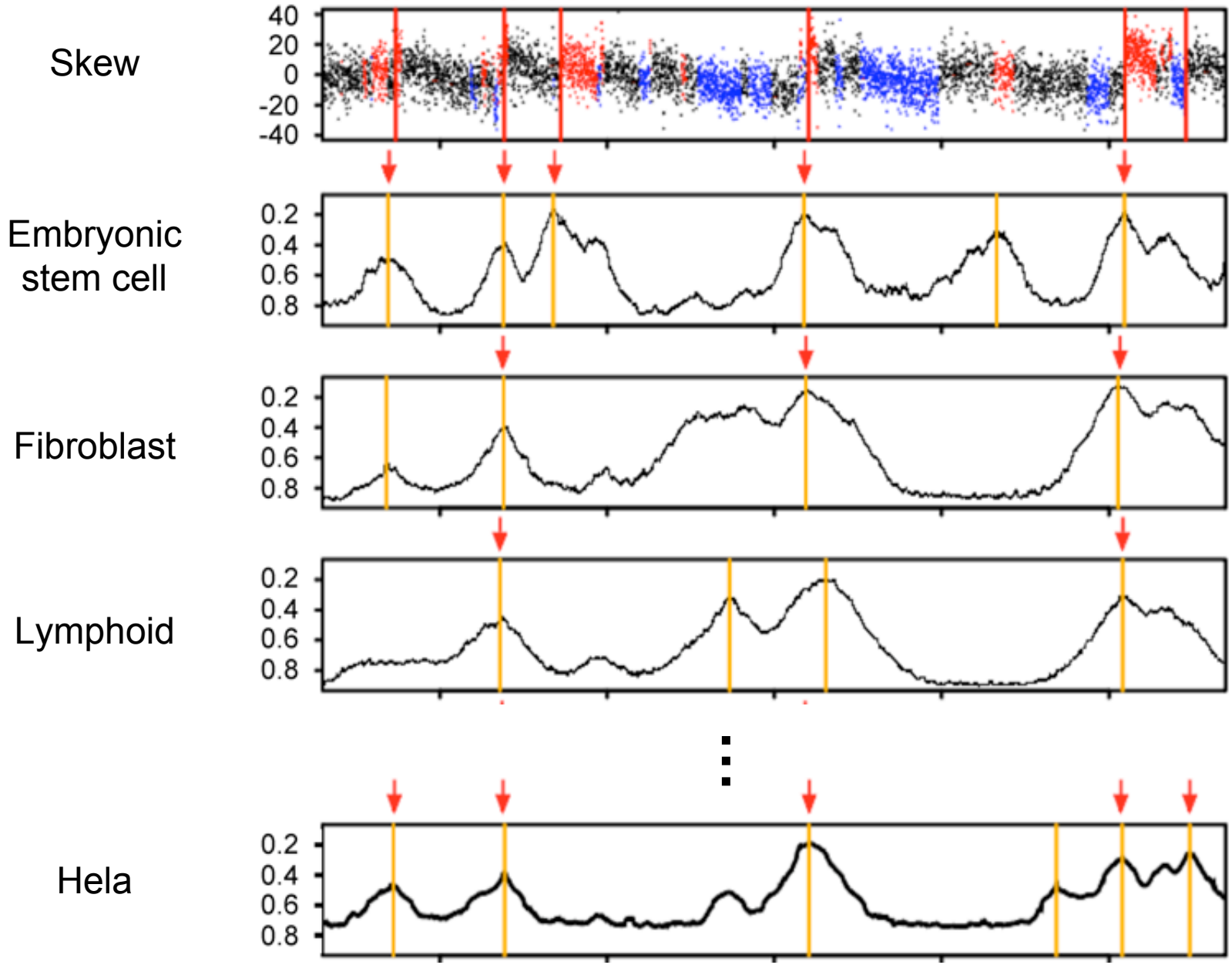
S phase

Chen CL. et al. *Genome Res*. (2010) 20:447

Comparison of upward jumps with initiation zones

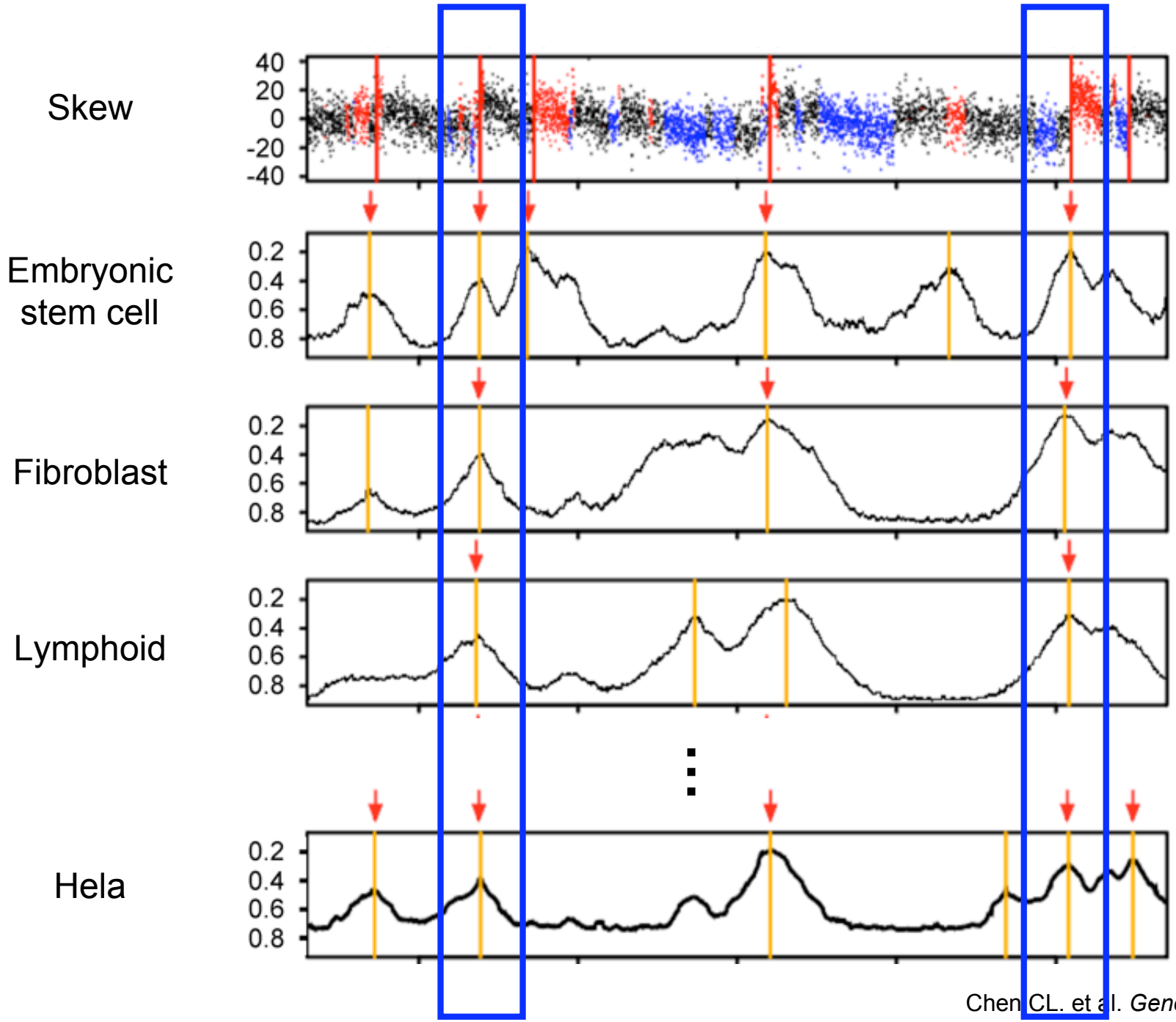


Upward jump positions correspond to replication initiation zones



Chen CL. et al. *Genome Res*. (2010) 20:447

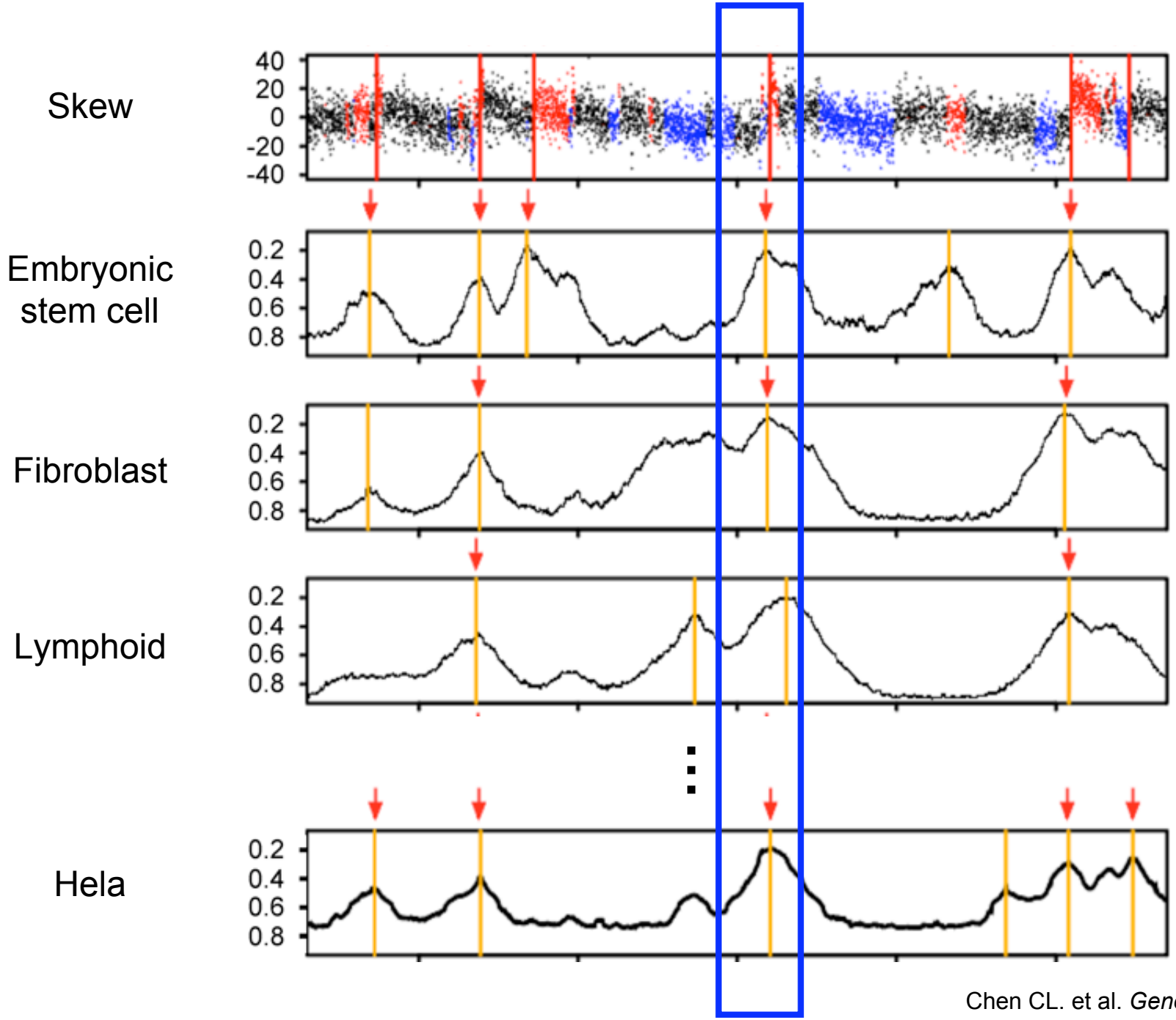
Hansen RS. et al. *PNAS*. (2010) 107:139



Chen CL. et al. *Genome Res*. (2010) 20:447

Hansen RS. et al. *PNAS*. (2010) 107:139

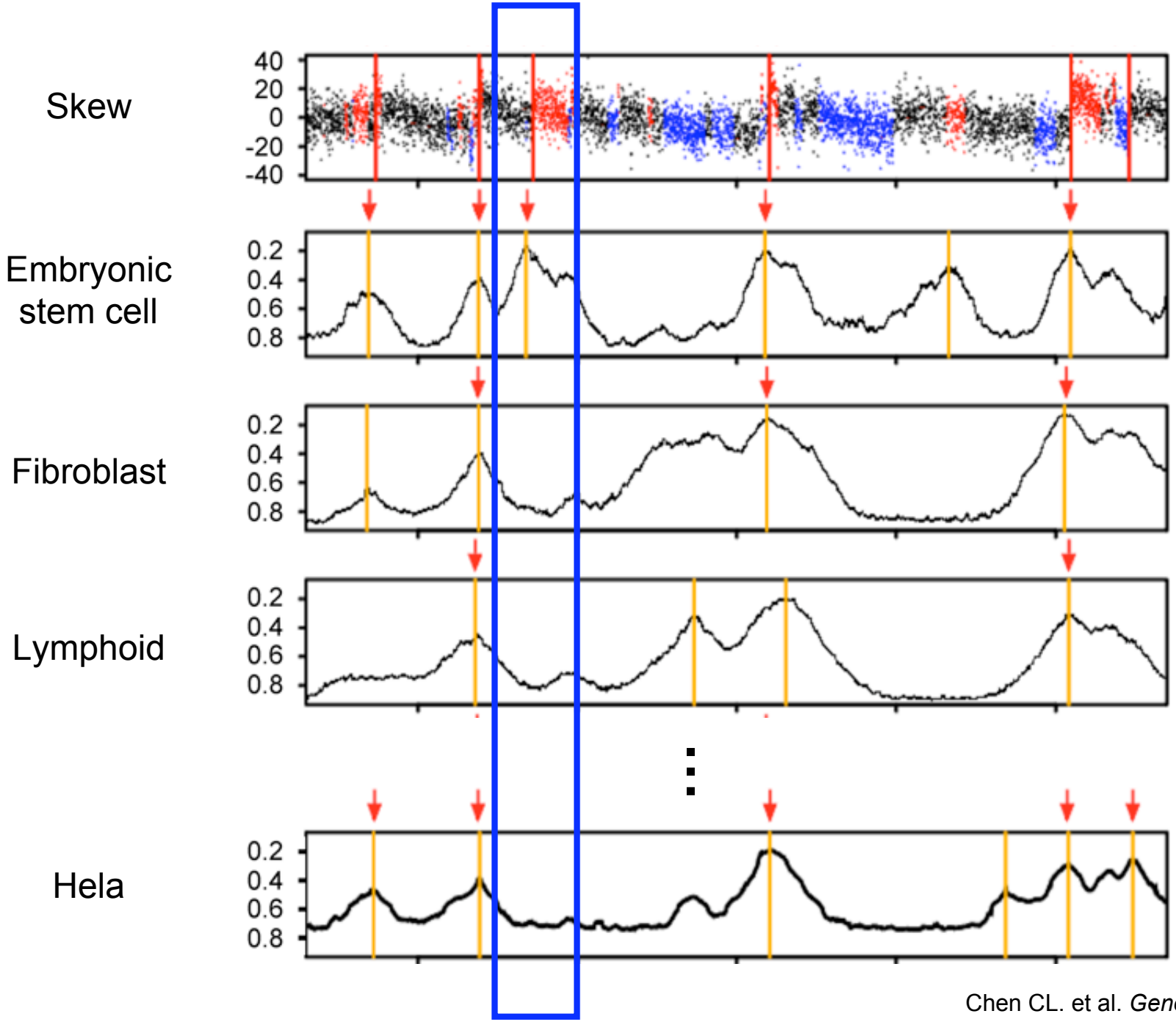
Desprat R. et al. *Genome Res*. (2009) 19:2288



Chen CL. et al. *Genome Res*. (2010) 20:447

Hansen RS. et al. *PNAS*. (2010) 107:139

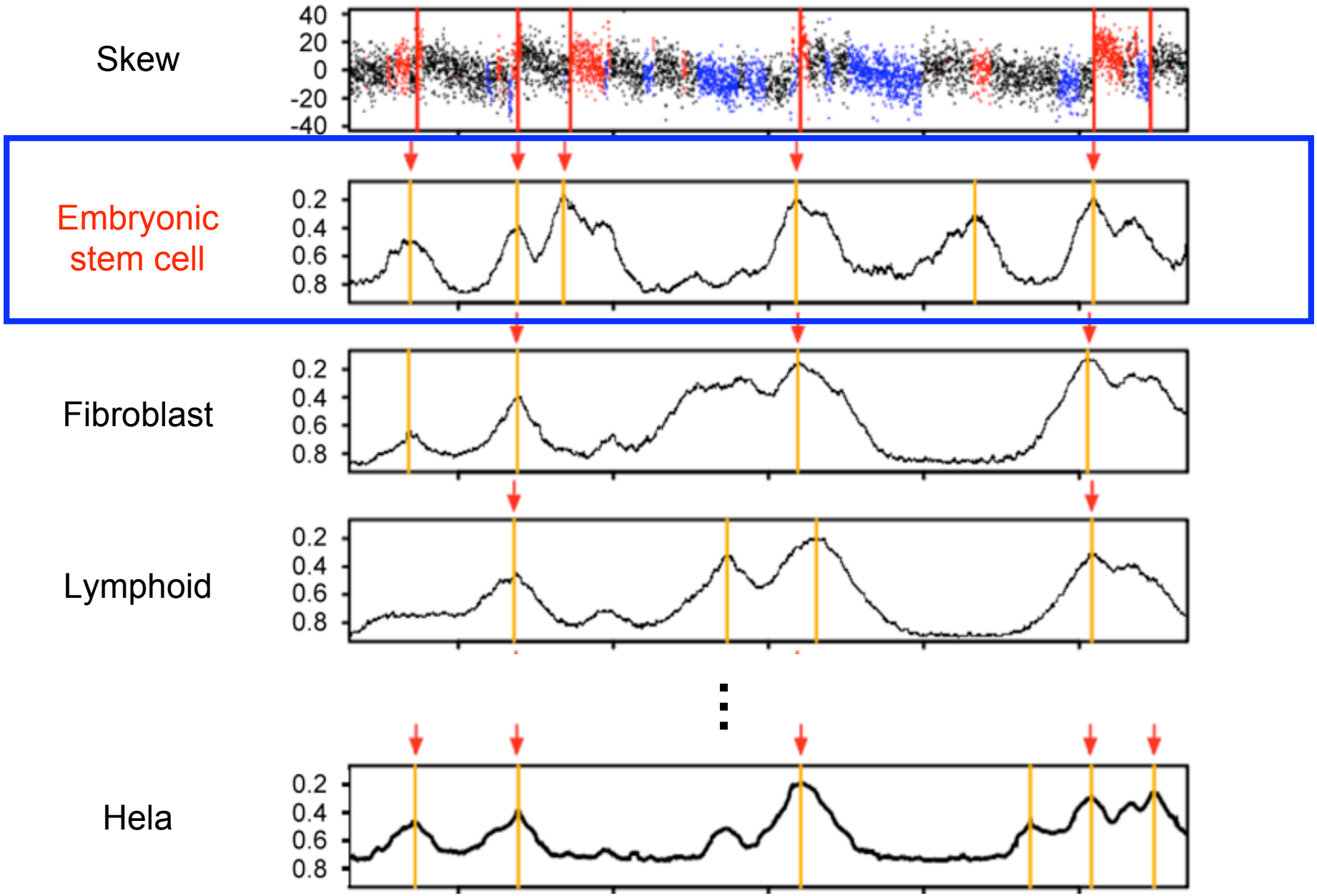
Desprat R. et al. *Genome Res*. (2009) 19:2288



Chen CL. et al. *Genome Res*. (2010) 20:447

Hansen RS. et al. *PNAS*. (2010) 107:139

Desprat R. et al. *Genome Res*. (2009) 19:2288

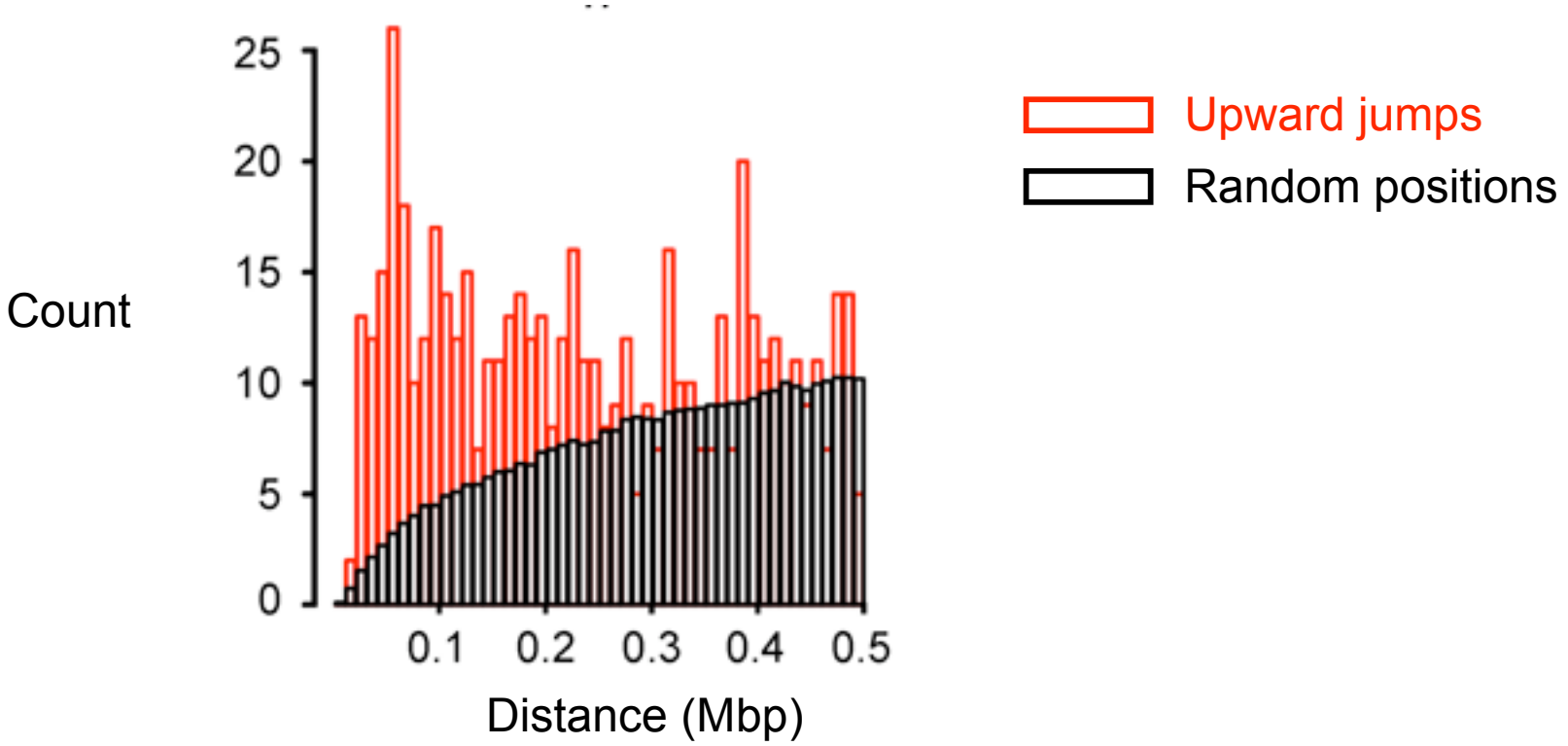


Chen CL. et al. *Genome Res*. (2010) 20:447

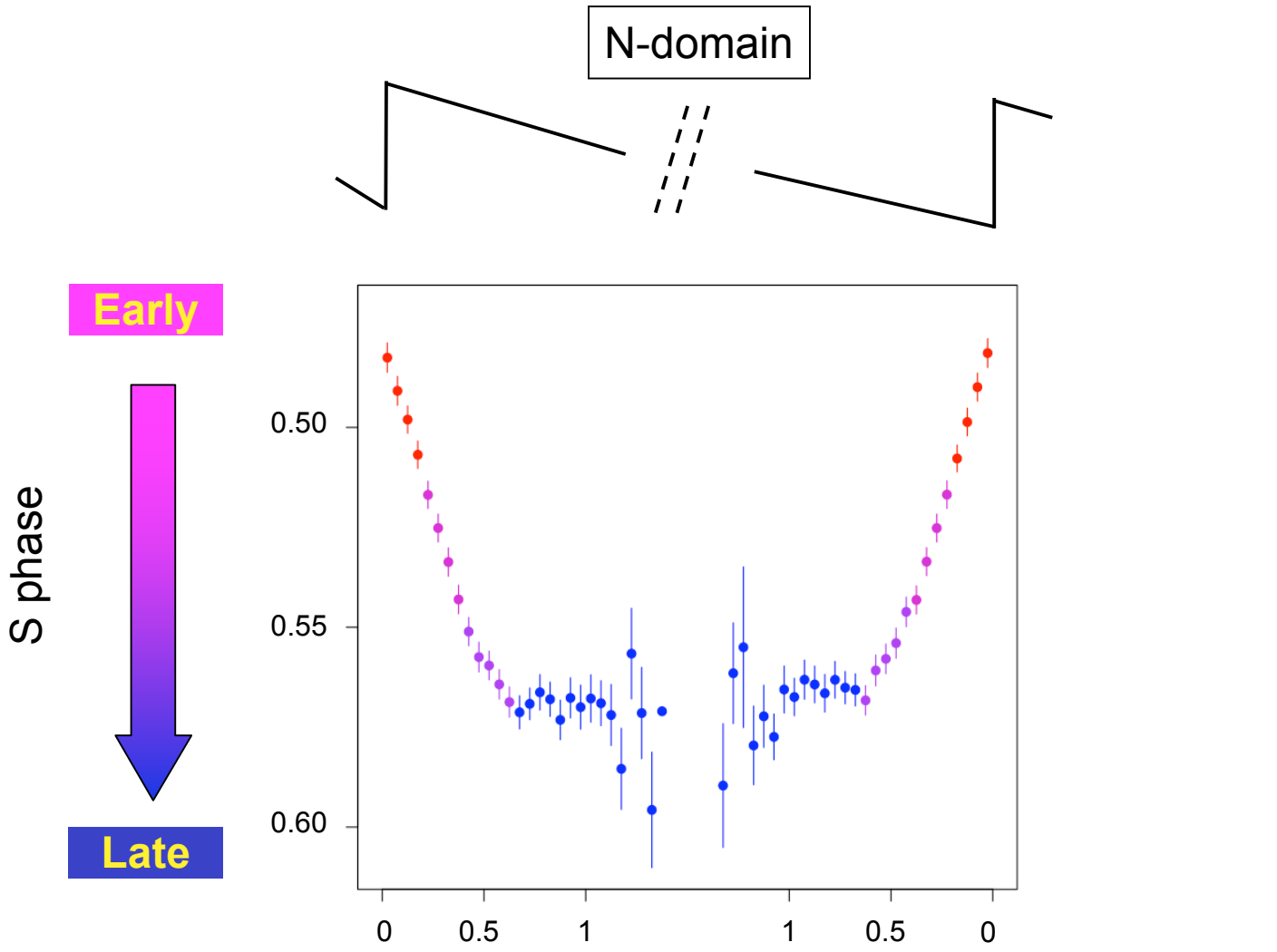
Hansen RS. et al. *PNAS*. (2010) 107:139

Desprat R. et al. *Genome Res*. (2009) 19:2288

Statistical evaluation : comparison with null distribution of simulation



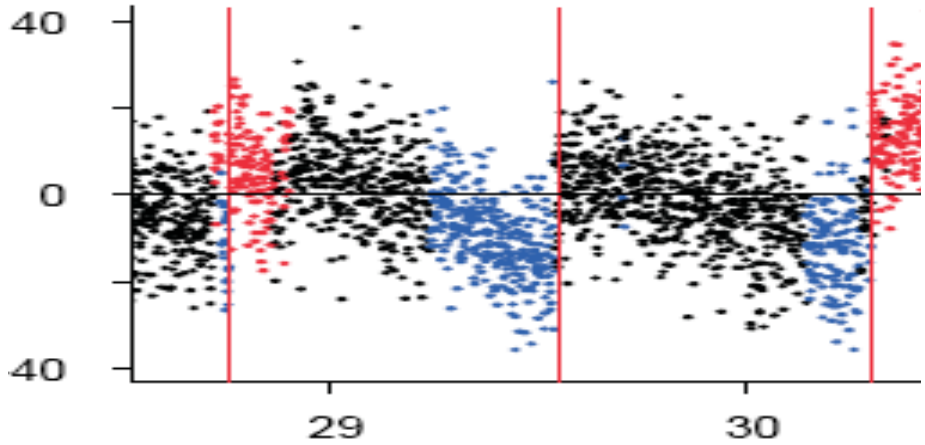
Upward jumps are significantly associated with replication initiation zones



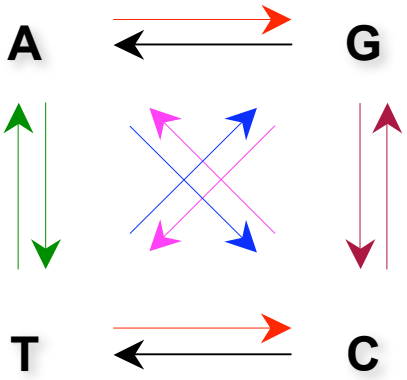
Replication starts from N domain borders and propagate to center in later S phase

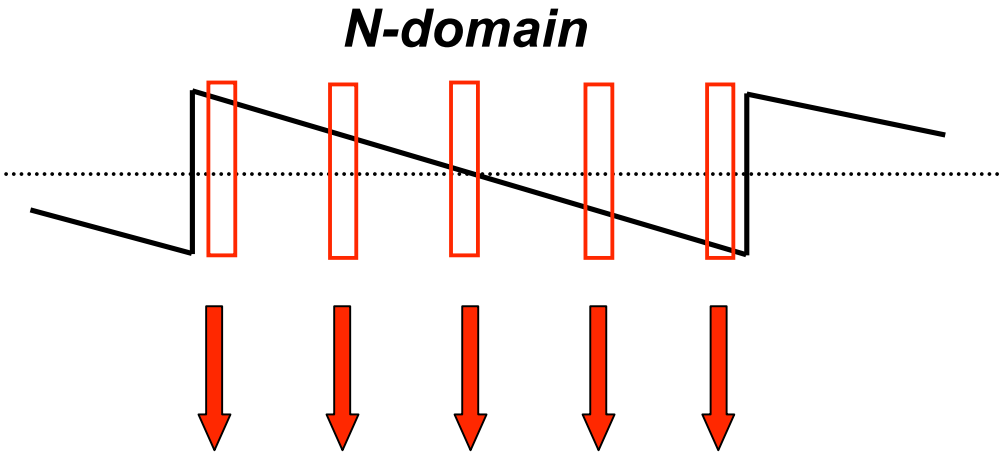
QUESTION 2:

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



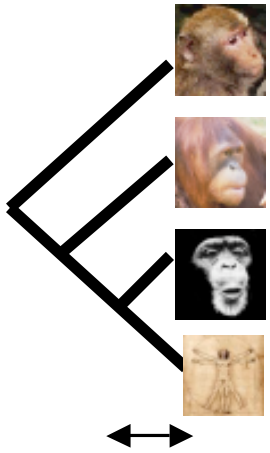
Computation of nucleotide substitution rates





A -> C	?	?	?	?	?
A -> G	?	?	?	?	?
A -> T	?	?	?	?	?
C -> A	?	?	?	?	?
C -> G	?	?	?	?	?
C -> T	?	?	?	?	?
G -> A	?	?	?	?	?
G -> C	?	?	?	?	?
G -> T	?	?	?	?	?
T -> A	?	?	?	?	?
T -> C	?	?	?	?	?
T -> G	?	?	?	?	?

Computation of nucleotide substitution rates



Macaca mulatta (rhesus monkey) AACTTTCG**G**TAG**T**GGATGA**C**ATCC**C**ATGTCGTGT

Pongo pygmaeus abelii (orangutan) AACTTTCG**G**TAG**C**GGATGA**C**ATCC**C**ATGTCGTGT

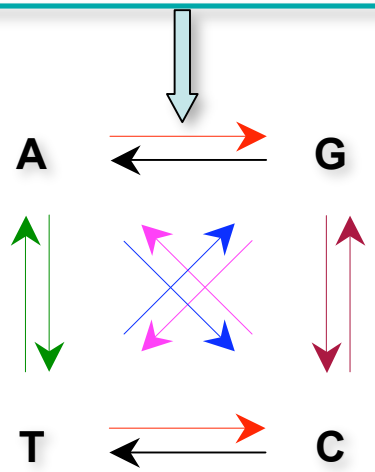
Pan troglodytes (chimpanzee) AACTTTCG**A**TAG**T**GGATGA**G**ATCC**C**ATGTCGTGT

Homo sapiens (human) AACTTTCG**T**TAG**T**GGATGA**C**ATCC**T**ATGTCGTGT

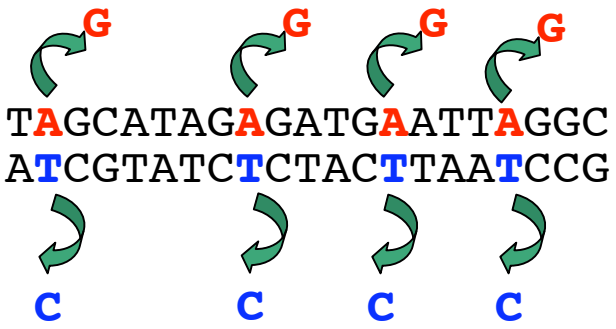
non-informative site

chimp C ➡ G

human C ➡ T



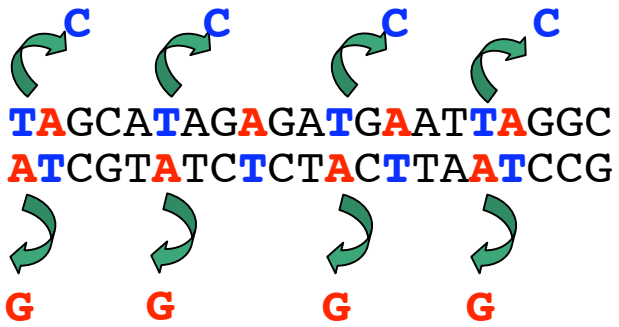
Alignment data retrieved from *Ensembl*



upper strand: n(A→G)

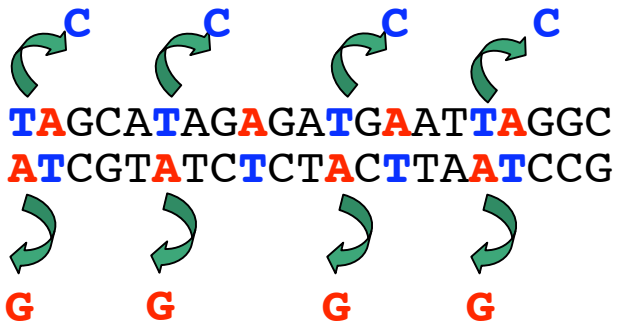
=

lower strand: n(T→C)



upper strand: n(A→G) = n(T→C)

lower strand: n(T→C) = n(A→G)

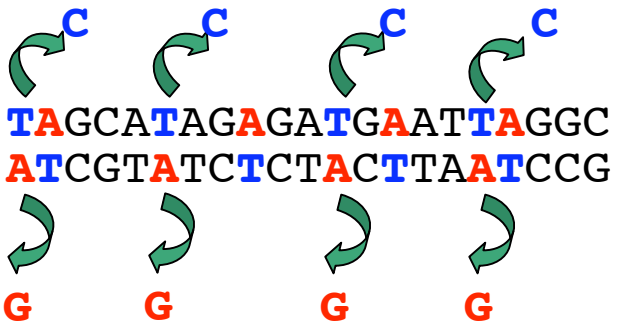


upper strand: **n(A→G)** n(T→C)

= =

lower strand: n(T→C) **n(A→G)**

Compare the same substitution rate on two strands

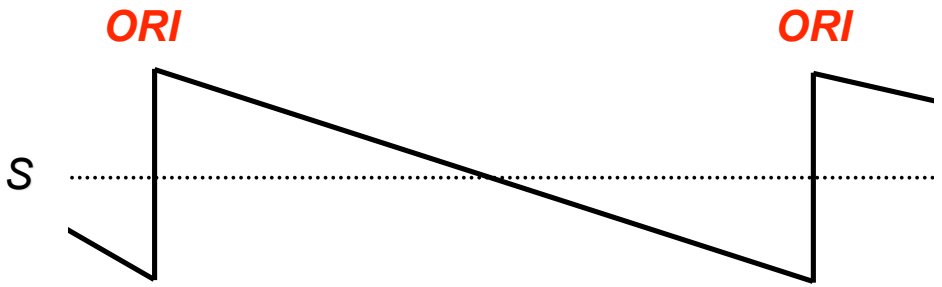


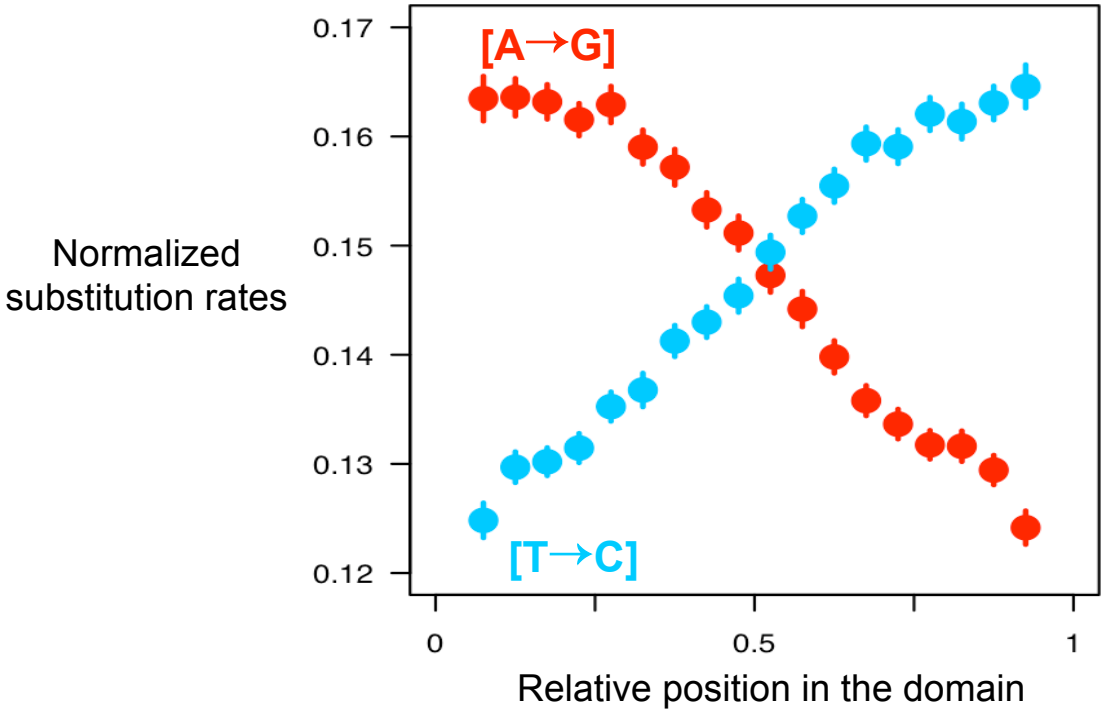
upper strand: n(A→G) ⟷ n(T→C)

Compare the same substitution rate on two strands

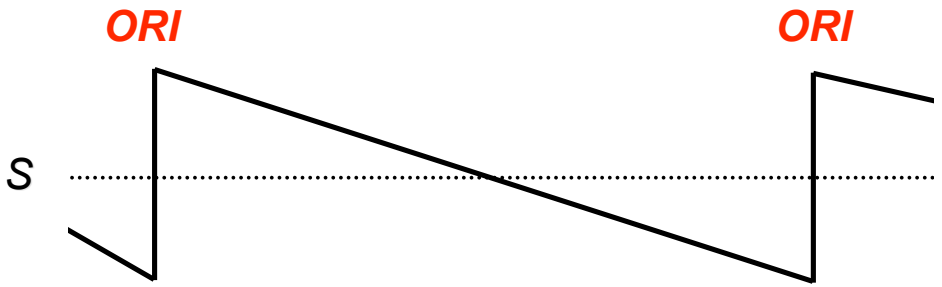
→ **Compare the complementary substitution rates on the same strand**

Complementary substitution rates along N-domains



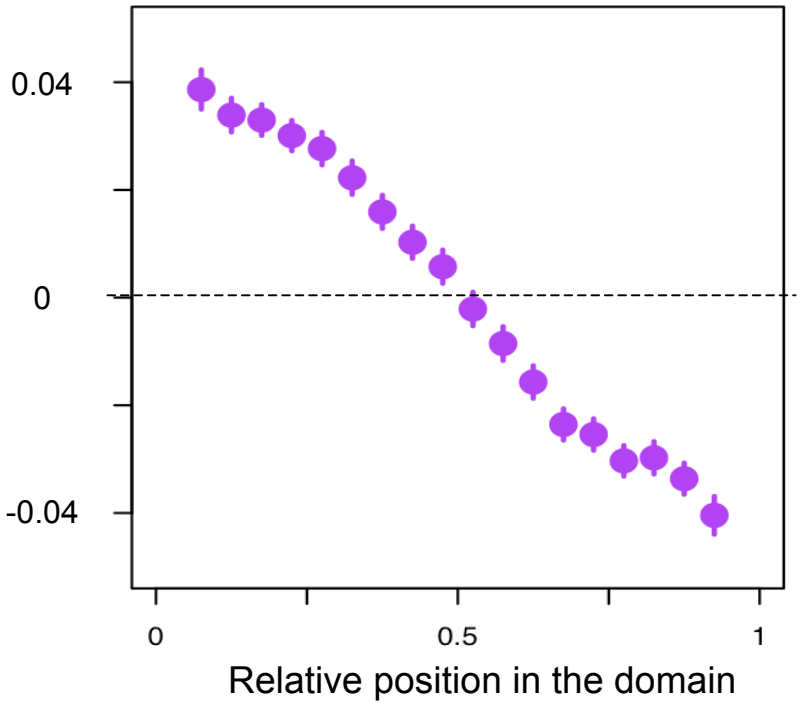


Complementary substitution rates along N-domains



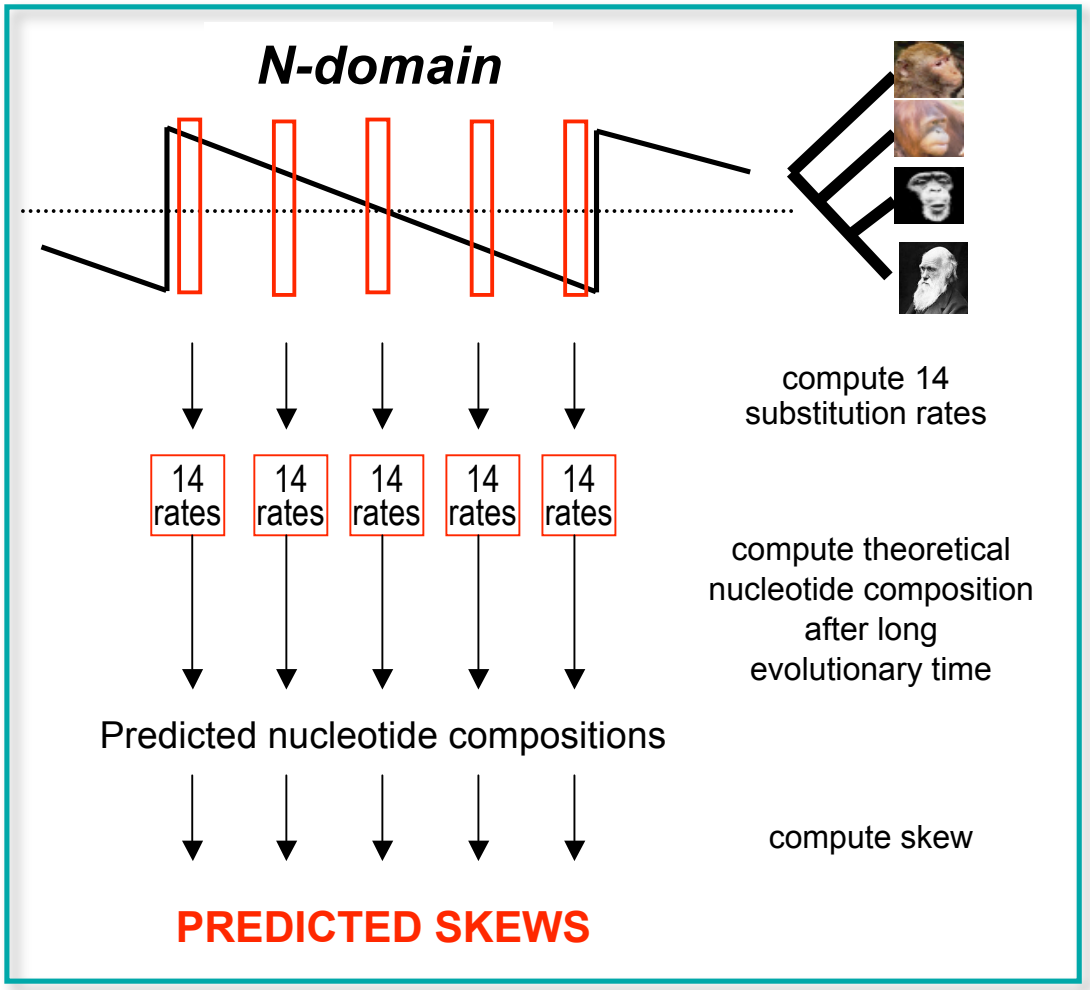
Asymmetry

△ = [A→G] – [T→C]



Reproduces perfectly the “N” pattern of skew profile

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



$$\frac{\partial f_{aa}}{\partial t} = caf_{ca} + gaf_{ga} + taf_{ta} - acf_{aa} - agf_{aa} - atf_{aa} + caf_{ac} + gaf_{ag} + taf_{at} - acf_{aa} - agf_{aa} - atf_{aa} + \frac{pgpaf_{cg}f_{ga}}{f_{ga}+f_{gc}+f_{gg}+f_{gt}}$$

$$\frac{\partial f_{ac}}{\partial t} = caf_{cc} + gaf_{gc} + taf_{tc} - acf_{ac} - agf_{ac} - atf_{ac} + acf_{aa} + gcf_{ag} + tcf_{at} - caf_{ac} - cgf_{ac} - ctf_{ac} + \frac{pgpaf_{cg}f_{gc}}{f_{ga}+f_{gc}+f_{gg}+f_{gt}} - \frac{cptpf_{ac}f_{cg}}{f_{ca}+f_{cc}+f_{cg}+f_{ct}}$$

$$\frac{\partial f_{ag}}{\partial t} = caf_{cg} + gaf_{gg} + taf_{tg} - acf_{ag} - agf_{ag} - atf_{ag} + agf_{aa} + cgf_{ac} + tgf_{at} - gaf_{ag} - gcf_{ag} - gtf_{ag} + \frac{pgpaf_{cg}f_{gg}}{f_{ga}+f_{gc}+f_{gg}+f_{gt}}$$

$$\frac{\partial f_{at}}{\partial t} = caf_{ct} + gaf_{gt} + taf_{tt} - acf_{at} - agf_{at} - atf_{at} + atf_{aa} + ctf_{ac} + gtf_{ag} - taf_{at} - tcf_{at} - tgf_{at} + \frac{pgpaf_{cg}f_{gt}}{f_{ga}+f_{gc}+f_{gg}+f_{gt}} + \frac{cptpf_{ac}f_{cg}}{f_{ca}+f_{cc}+f_{cg}+f_{ct}}$$

$$\frac{\partial f_{ca}}{\partial t} = acf_{aa} + gcf_{ga} + tcf_{ta} - caf_{ca} - cgf_{ca} - ctf_{ca} + caf_{cc} + taf_{ct} - acf_{ca} - agf_{ca} - atf_{ca} + pgpaf_{cg}$$

$$\frac{\partial f_{cc}}{\partial t} = acf_{ac} + gcf_{gc} + tcf_{tc} - caf_{cc} - cgf_{cc} - ctf_{cc} + acf_{ca} + gcf_{cg} + tcf_{ct} - caf_{cc} - cgf_{cc} - ctf_{cc} - \frac{cptpf_{cc}f_{cg}}{f_{ca}+f_{cc}+f_{cg}+f_{ct}}$$

$$\frac{\partial f_{cg}}{\partial t} = acf_{ag} + gcf_{gg} + tcf_{tg} - caf_{cg} - cgf_{cg} + agf_{ca} + cgf_{cc} + tgf_{ct} - gcf_{cg} - gtf_{cg} - cptpf_{cg} - pgpaf_{cg}$$

$$\frac{\partial f_{ct}}{\partial t} = acf_{at} + gcf_{gt} + tcf_{tt} - caf_{ct} - cgf_{ct} - ctf_{ct} + atf_{ca} + ctf_{cc} + gtf_{cg} - taf_{ct} - tcf_{ct} - tgf_{ct} + \frac{cptpf_{cc}f_{cg}}{f_{ca}+f_{cc}+f_{cg}+f_{ct}}$$

$$\frac{\partial f_{ga}}{\partial t} = agf_{aa} + cgf_{ca} + tgf_{ta} - gaf_{ga} - gcf_{ga} - gtf_{ga} + caf_{gc} + gaf_{gg} + taf_{gt} - acf_{ga} - agf_{ga} - atf_{ga} - \frac{pgpaf_{cg}f_{ga}}{f_{ga}+f_{gc}+f_{gg}+f_{gt}}$$

$$\frac{\partial f_{gc}}{\partial t} = agf_{ac} + cgf_{cc} + tgf_{tc} - gaf_{gc} - gcf_{gc} - gtf_{gc} + acf_{ga} + gcf_{gg} + tcf_{gt} - caf_{gc} - cgf_{gc} - ctf_{gc} - \frac{pgpaf_{cg}f_{gc}}{f_{ga}+f_{gc}+f_{gg}+f_{gt}} - \frac{cptpf_{gc}f_{cg}}{f_{ca}+f_{cc}+f_{cg}+f_{ct}}$$

$$\frac{\partial f_{gg}}{\partial t} = agf_{ag} + cgf_{cg} + tgf_{tg} - gaf_{gg} - gcf_{gg} - gtf_{gg} + agf_{ga} + cgf_{gc} + tgf_{gt} - gaf_{gg} - gcf_{gg} - gtf_{gg} - \frac{pgpaf_{cg}f_{gg}}{f_{ga}+f_{gc}+f_{gg}+f_{gt}}$$

$$\frac{\partial f_{gt}}{\partial t} = agf_{at} + cgf_{ct} + tgf_{tt} - gaf_{gt} - gcf_{gt} - gtf_{gt} + atf_{ga} + ctf_{gc} + gtf_{gt} - taf_{gt} - tcf_{gt} - tgf_{gt} - \frac{pgpaf_{cg}f_{gt}}{f_{ga}+f_{gc}+f_{gg}+f_{gt}} + \frac{cptpf_{gc}f_{cg}}{f_{ca}+f_{cc}+f_{cg}+f_{ct}}$$

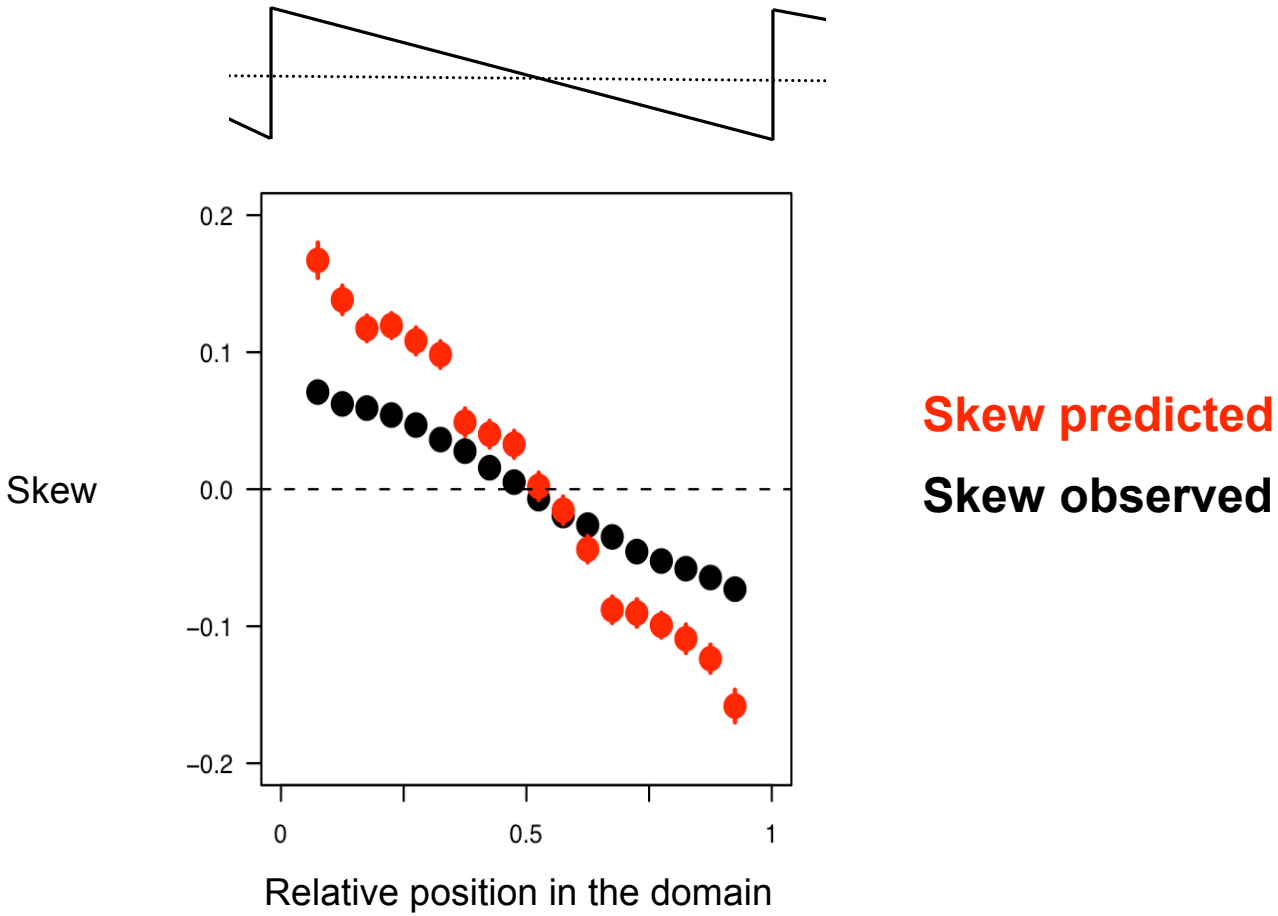
$$\frac{\partial f_{ta}}{\partial t} = atf_{aa} + ctf_{ca} + gtf_{ga} - taf_{ta} - tcf_{ta} - tgf_{ta} + caf_{tc} + gaf_{tg} + taf_{tt} - acf_{ta} - agf_{ta} - atf_{ta}$$

$$\frac{\partial f_{tc}}{\partial t} = atf_{ac} + ctf_{cc} + gtf_{gc} - taf_{tc} - tcf_{tc} - tgf_{tc} + acf_{ta} + gcf_{tg} + tcf_{tt} - caf_{tc} - cgf_{tc} - ctf_{tc} - \frac{cptpf_{tc}f_{cg}}{f_{ca}+f_{cc}+f_{cg}+f_{ct}}$$

$$\frac{\partial f_{tg}}{\partial t} = atf_{ag} + gtf_{gg} - taf_{tg} - tcf_{tg} - tgf_{tg} + agf_{ta} + cgf_{tc} + tgf_{tt} - gaf_{tg} - gcf_{tg} - gtf_{tg} + cptpf_{cg}$$

$$\frac{\partial f_{tt}}{\partial t} = atf_{at} + ctf_{ct} + gtf_{gt} - taf_{tt} - tcf_{tt} - tgf_{tt} + atf_{ta} + ctf_{tc} + gtf_{tg} - taf_{tt} - tcf_{tt} - tgf_{tt} + \frac{cptpf_{tc}f_{cg}}{f_{ca}+f_{cc}+f_{cg}+f_{ct}}$$

Skew at equilibrium reproduces perfectly the “N” pattern of skew profile

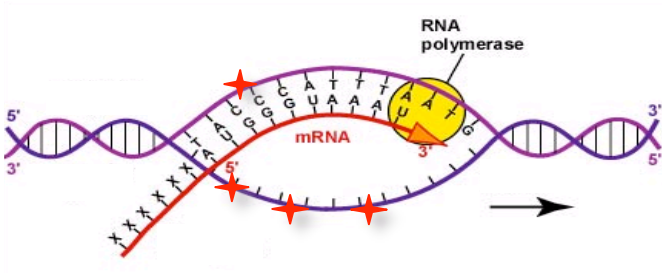


The skew is not at equilibrium

N-domains result from mutation asymmetry in germline cells

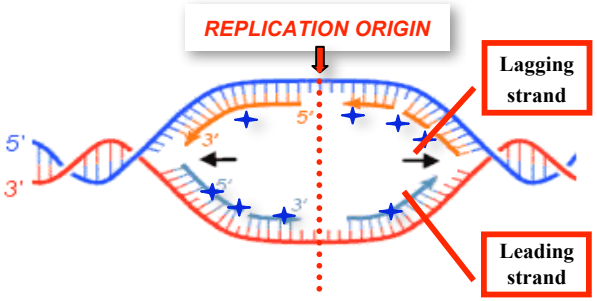
➢ **Does the "N" result from:**

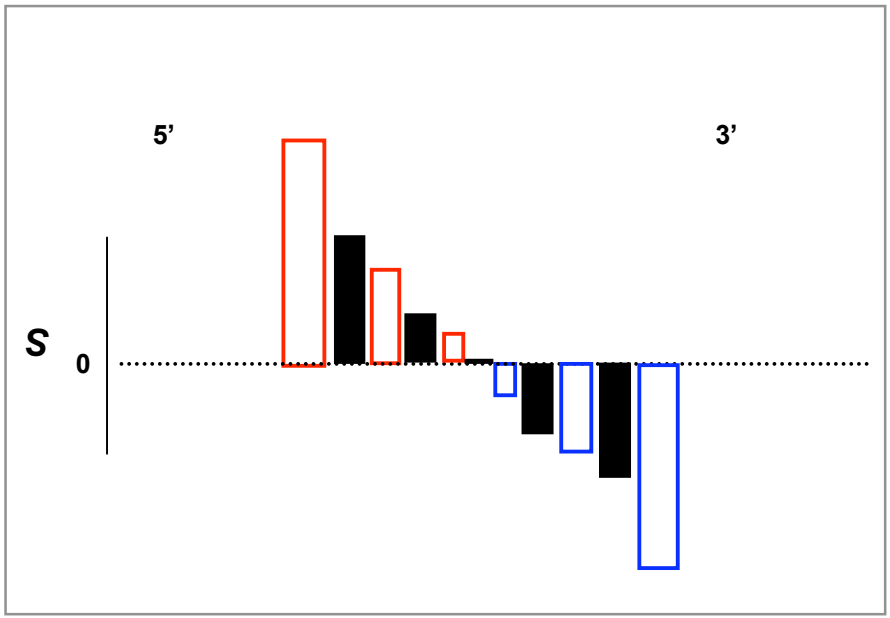
Transcription



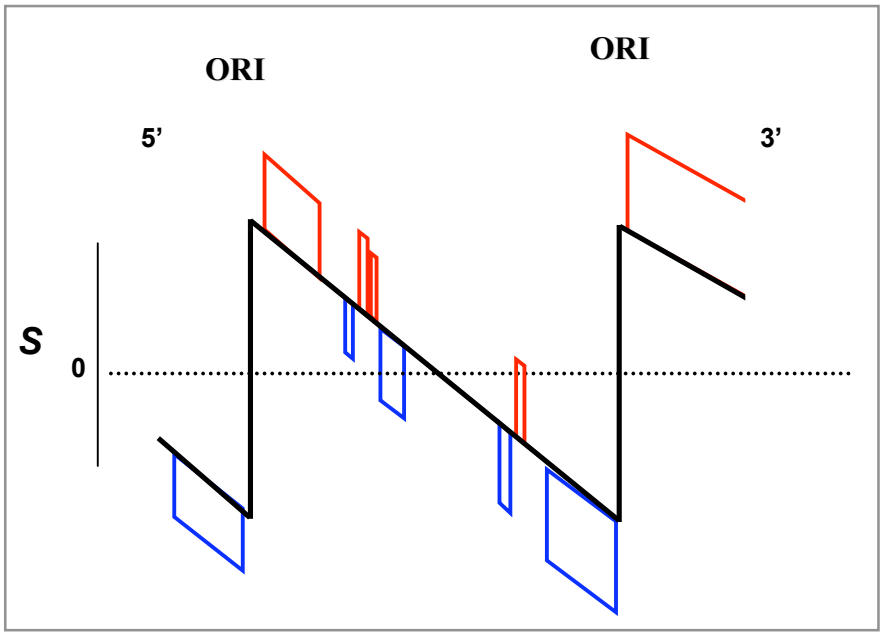
?

Replication





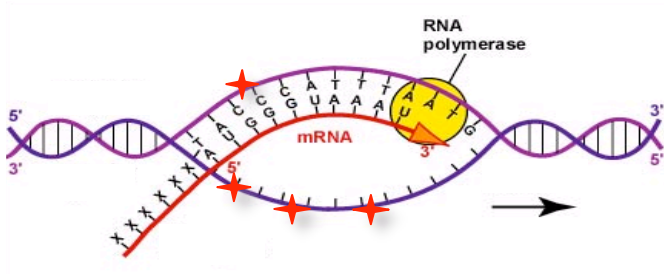
Transcription specifically organized

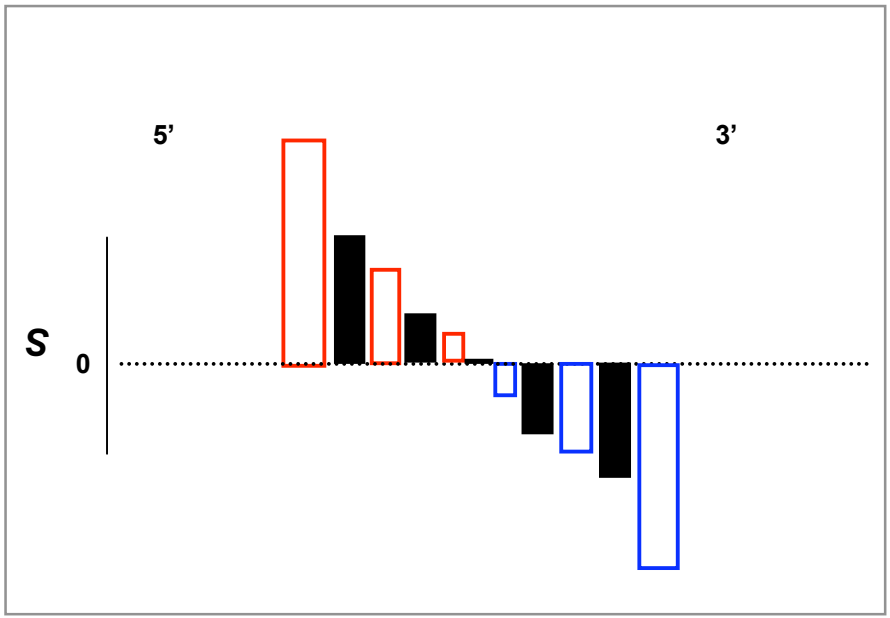


Replication specifically organized

➢ Does the “N” result from:

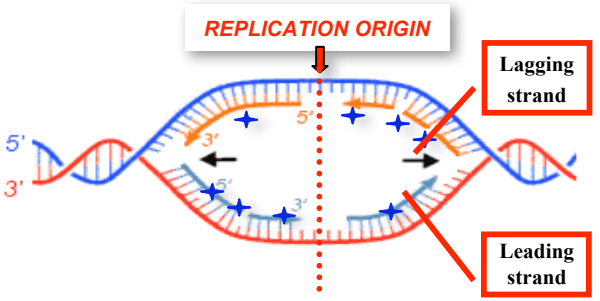
Transcription

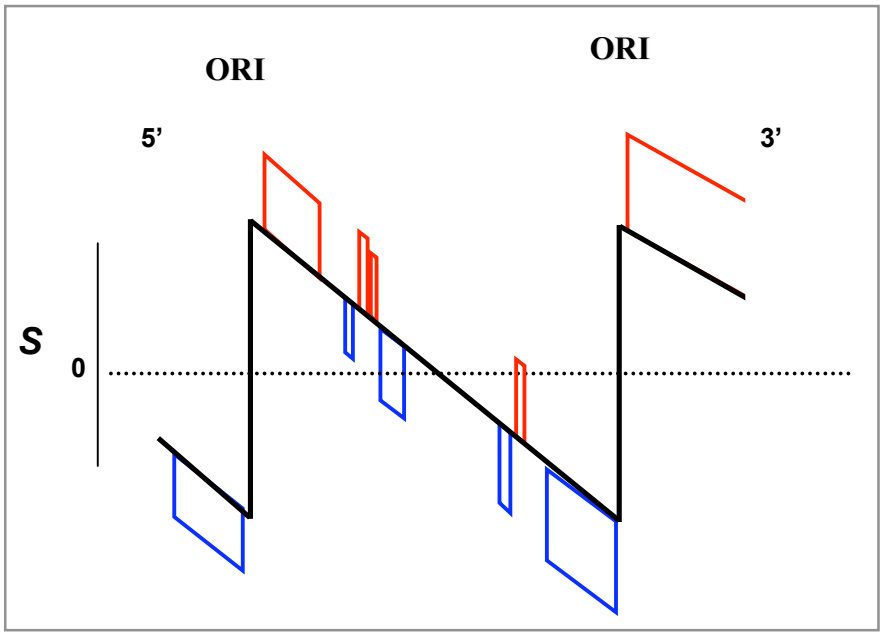




Transcription specifically organized

Replication

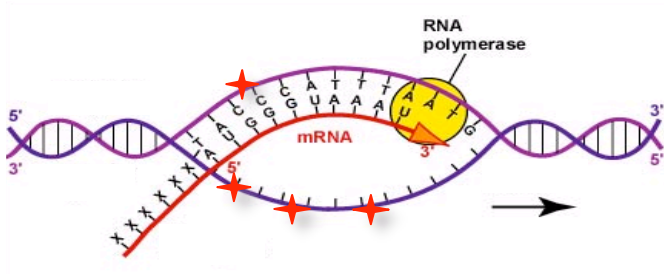


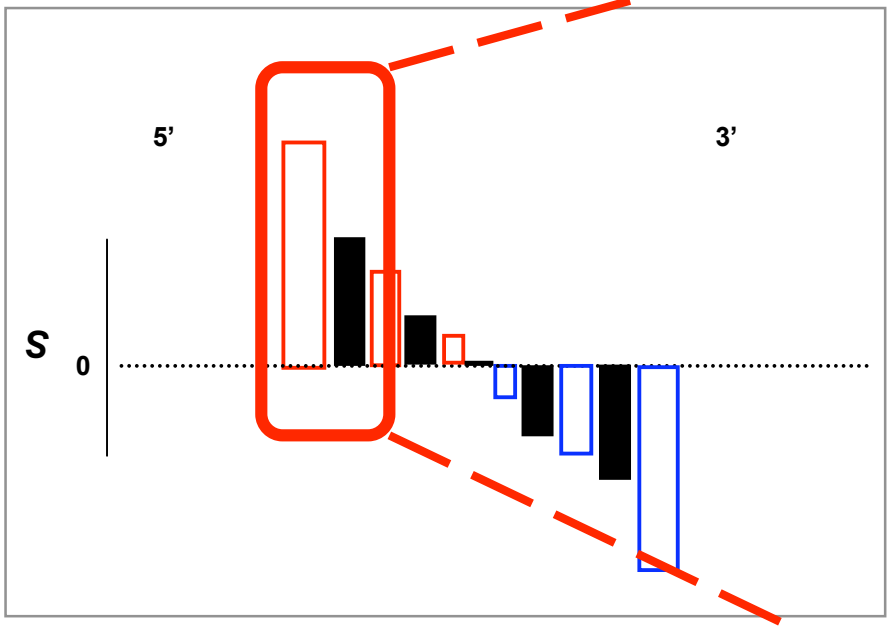


Replication specifically organized

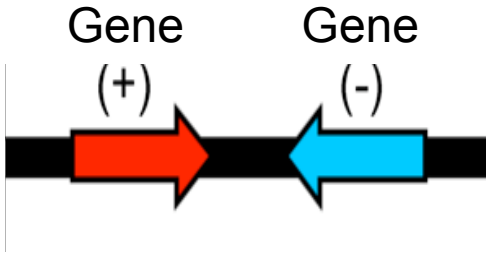
➢ **Does the "N" result from:**

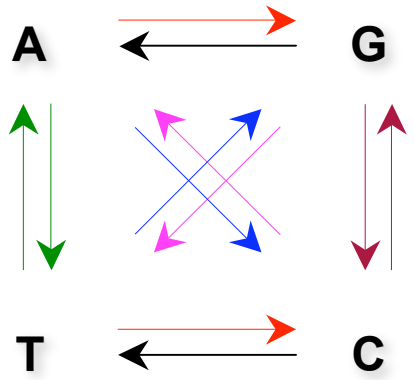
Transcription





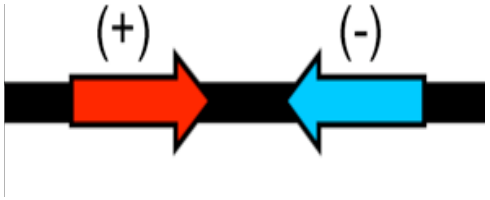
Transcription specifically organized

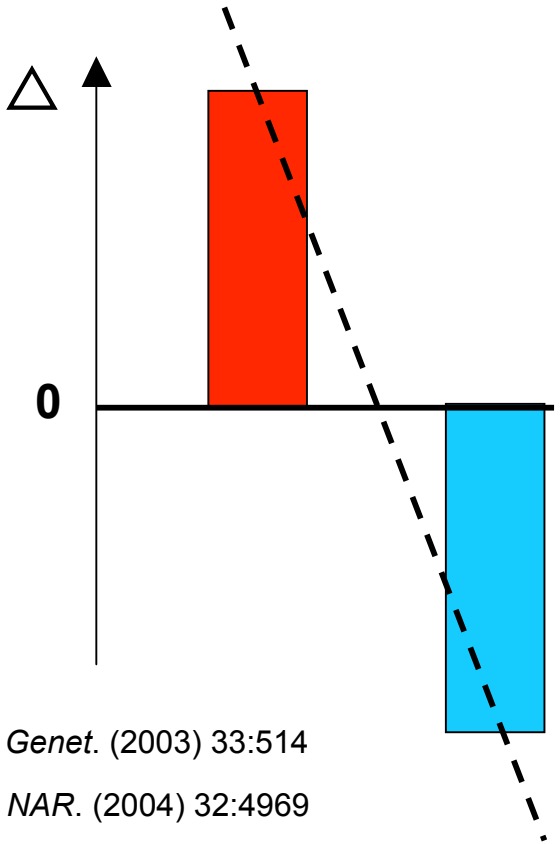




Group 1: △ = [A→G] – [T→C]

Previous study on transcriptional asymmetry





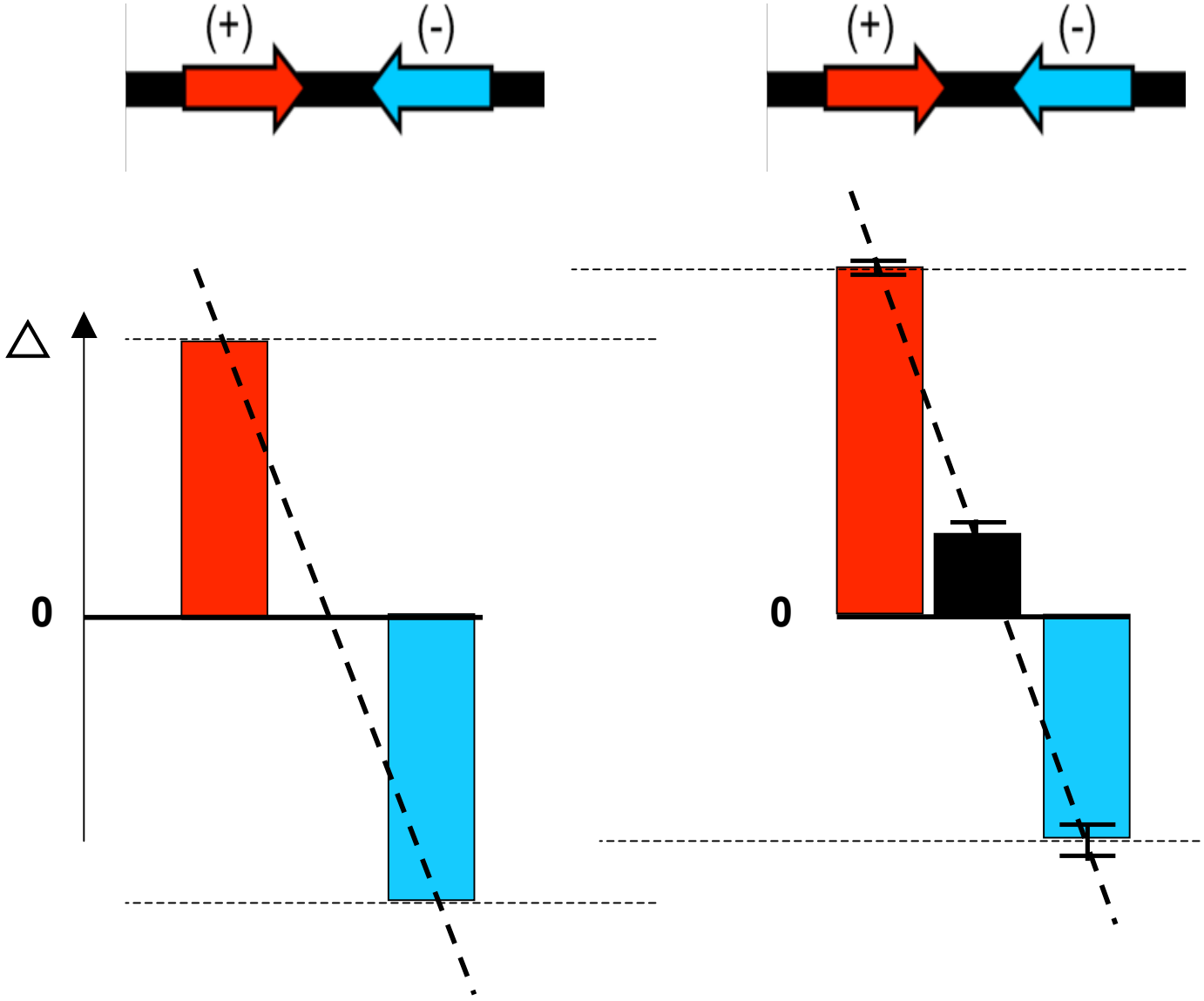
Green P. et al. *Nat. Genet*. (2003) 33:514

Touchon M. et al. *NAR*. (2004) 32:4969

Group 1: △ = [A→G] – [T→C]

Previous study on transcriptional asymmetry

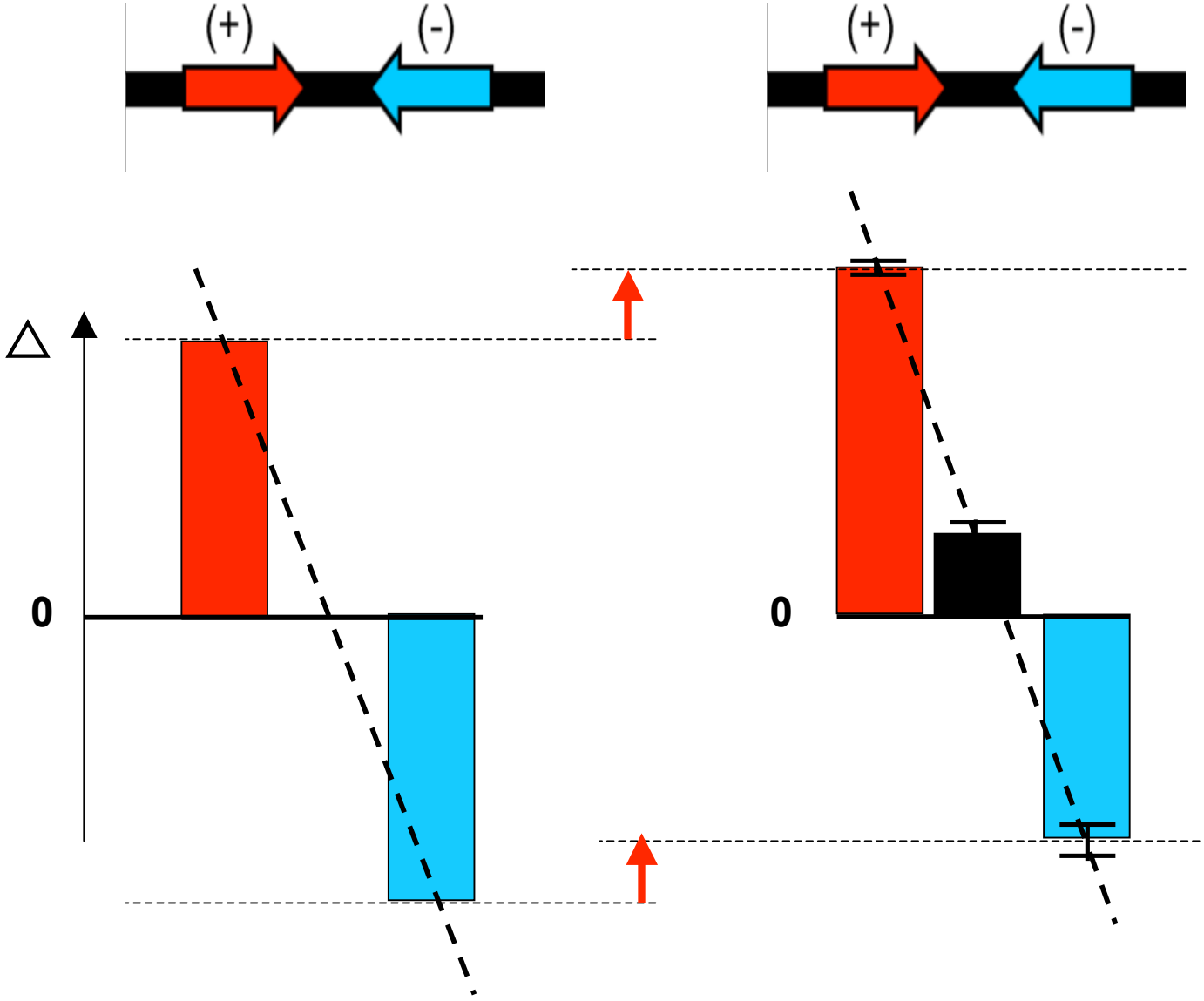
Observed in this study



Group 1: △ = [A→G] – [T→C]

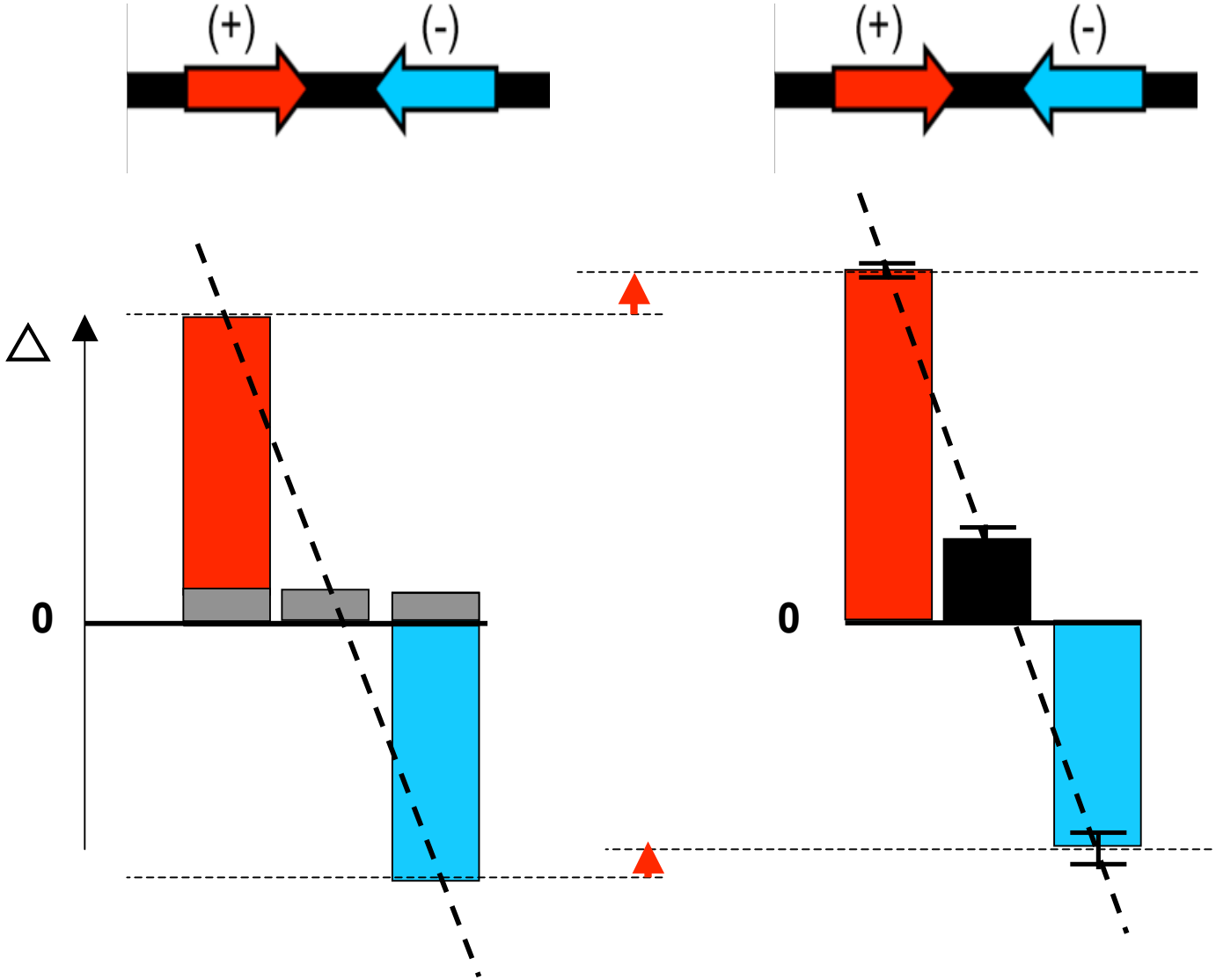
Previous study on transcriptional asymmetry

Observed in this study



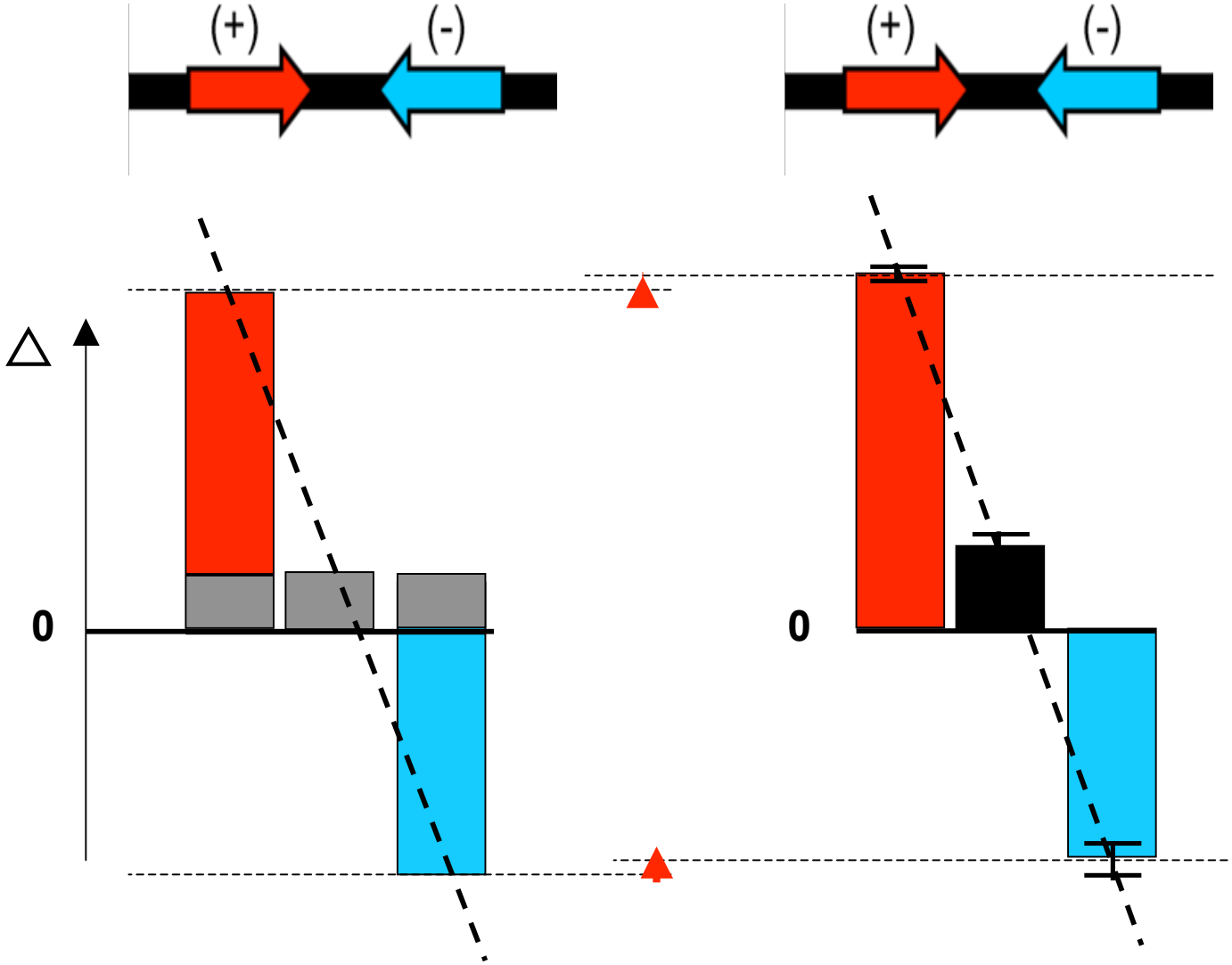
Group 1: △ = [A→G] – [T→C]

Observed in this study



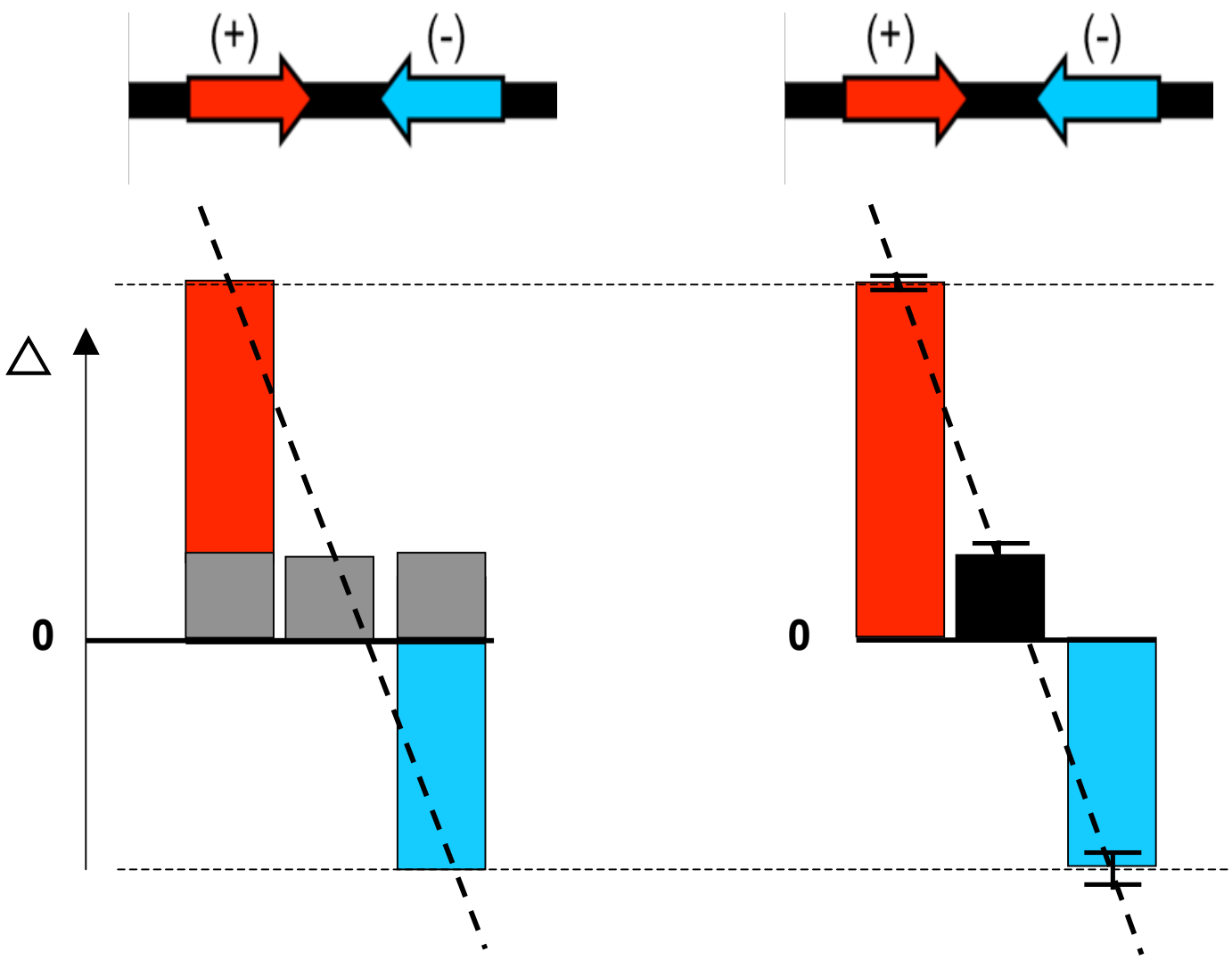
Group 1: △ = [A→G] – [T→C]

Observed in this study

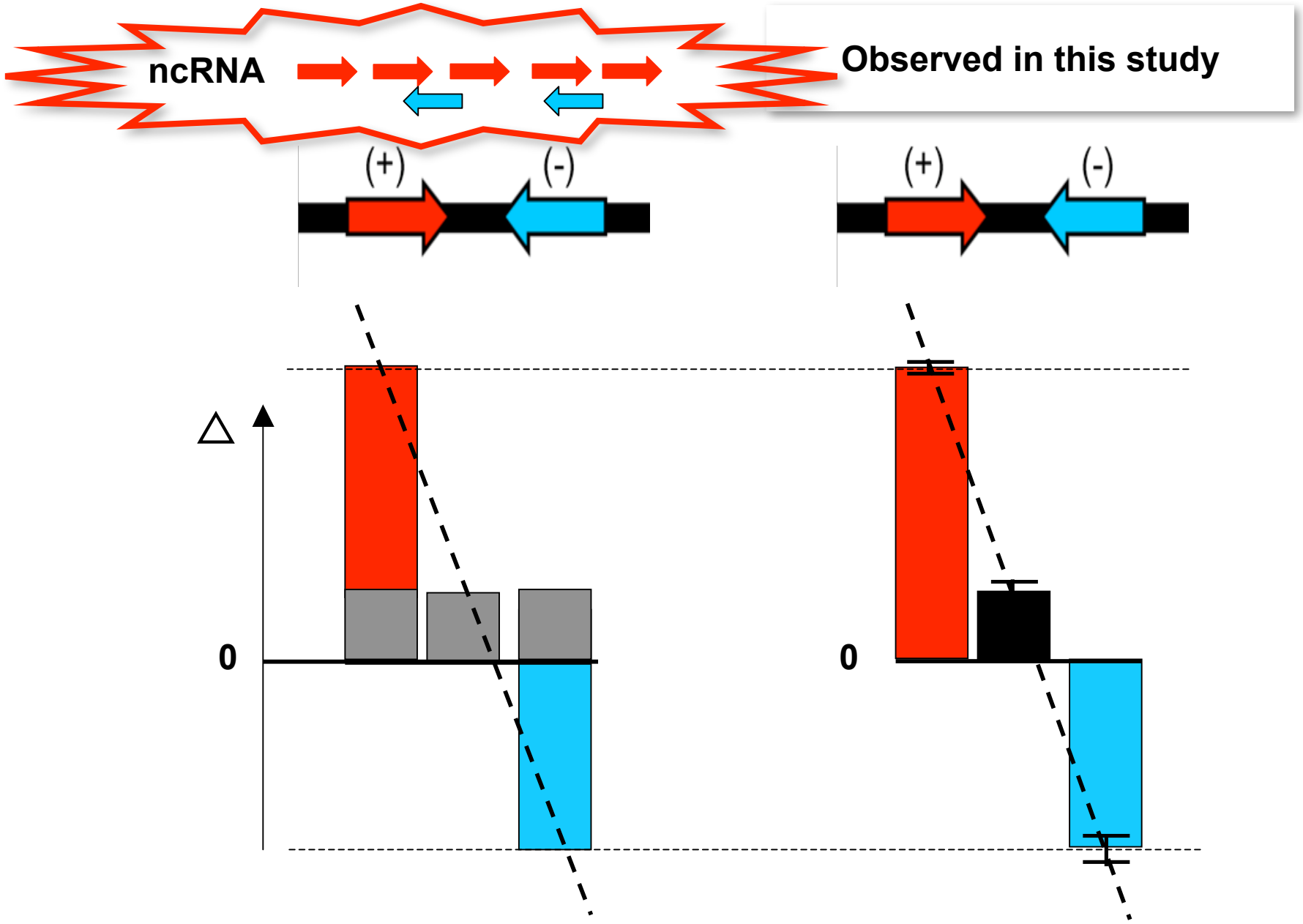


Group 1: △ = [A→G] – [T→C]

Observed in this study

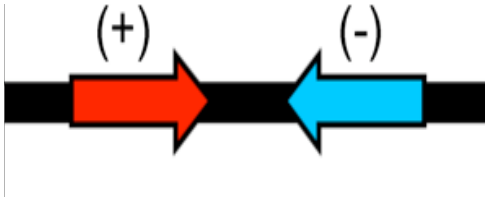


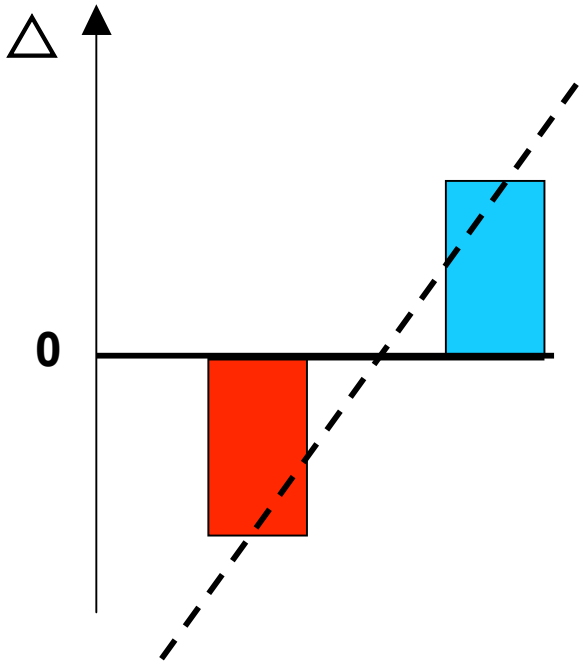
Group 1: △ = [A→G] – [T→C]



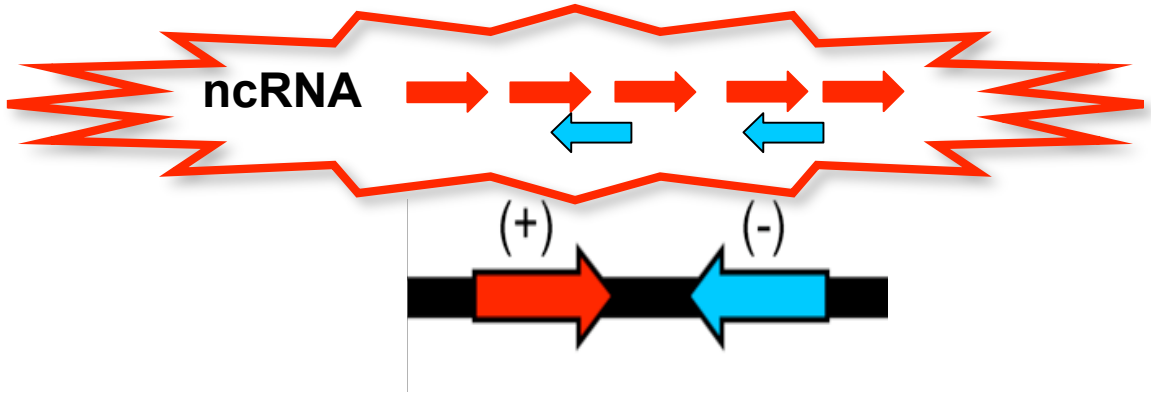
Group 2: △ = [C→T] – [G→A]

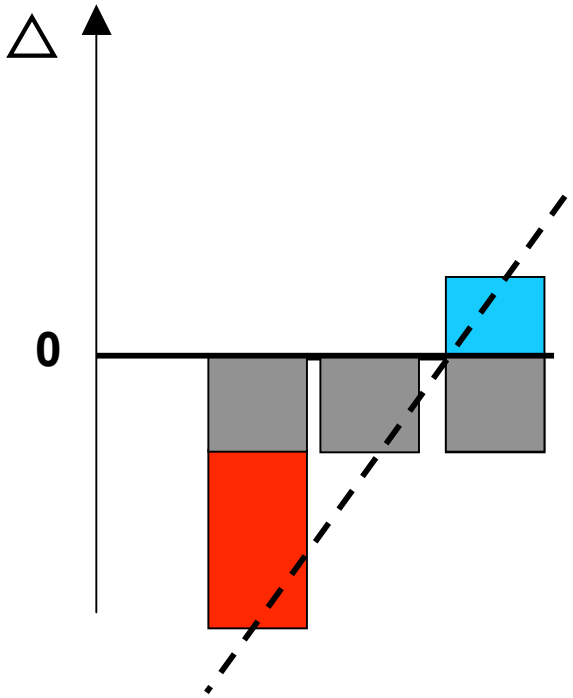
Previous study on transcriptional asymmetry



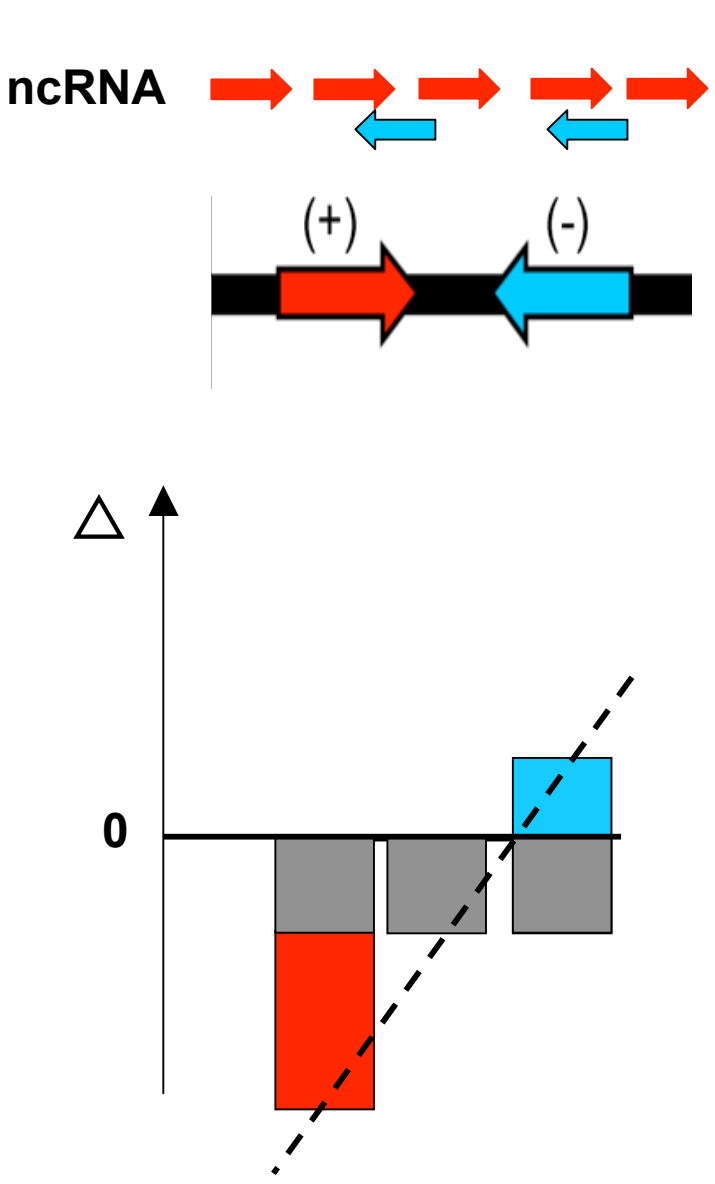


Group 2: △ = [C→T] – [G→A]

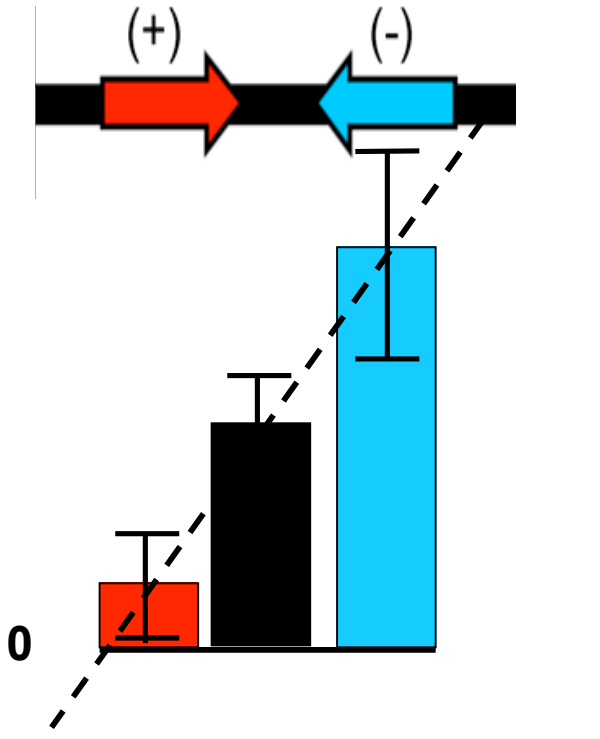




Group 2: △ = [C→T] – [G→A]



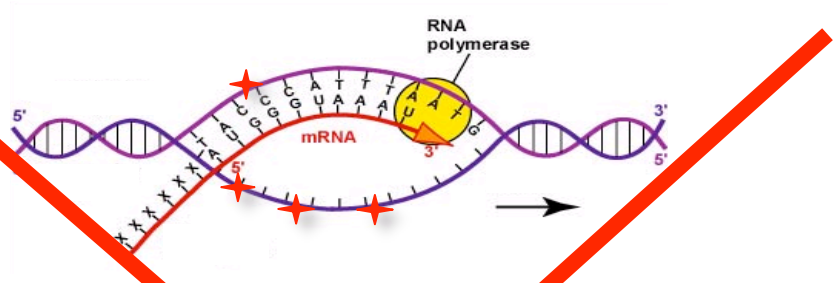
Observed in this study

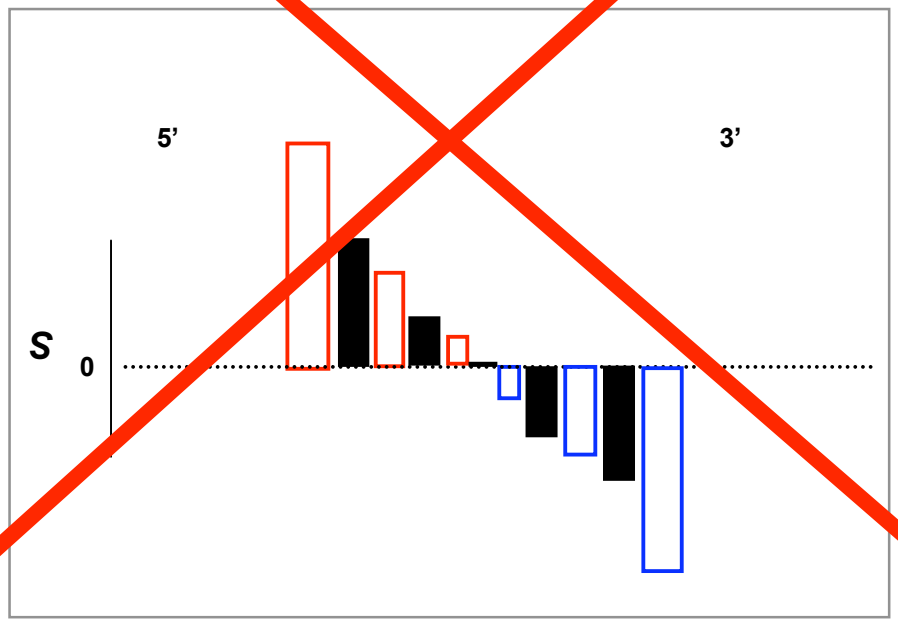


INCONSISTENCY !

Transcription only cannot explain the observed substitution patterns.

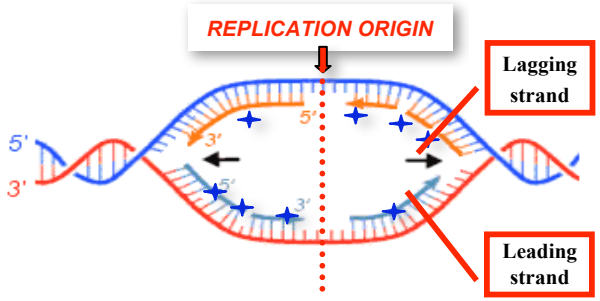
Transcription

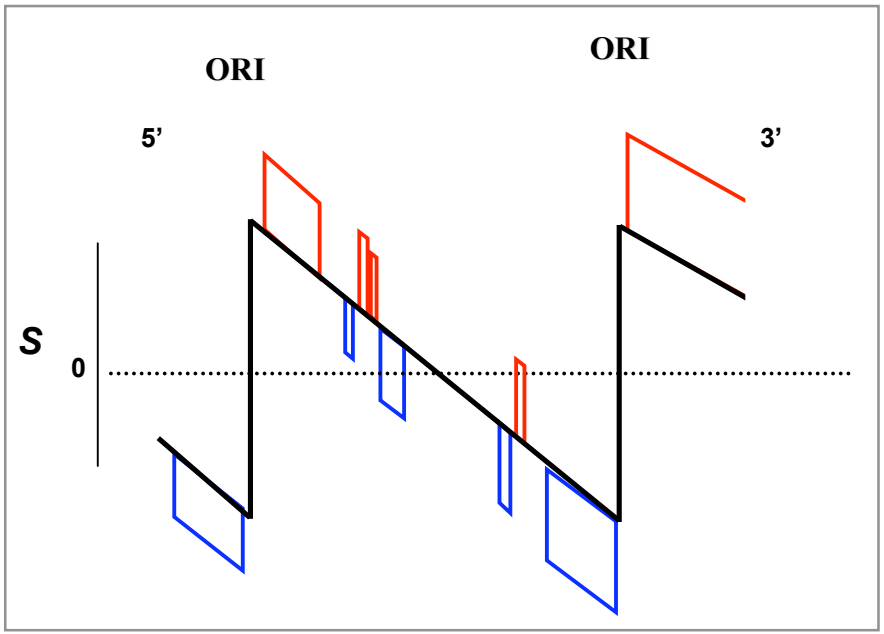




Transcription specifically organized

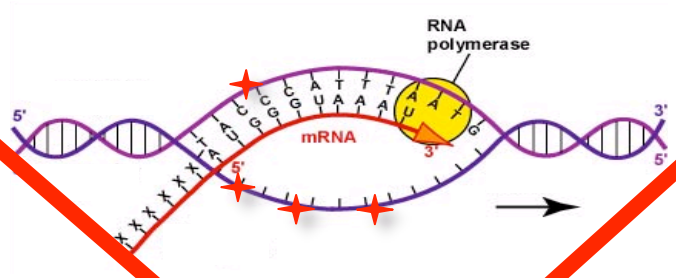
Replication

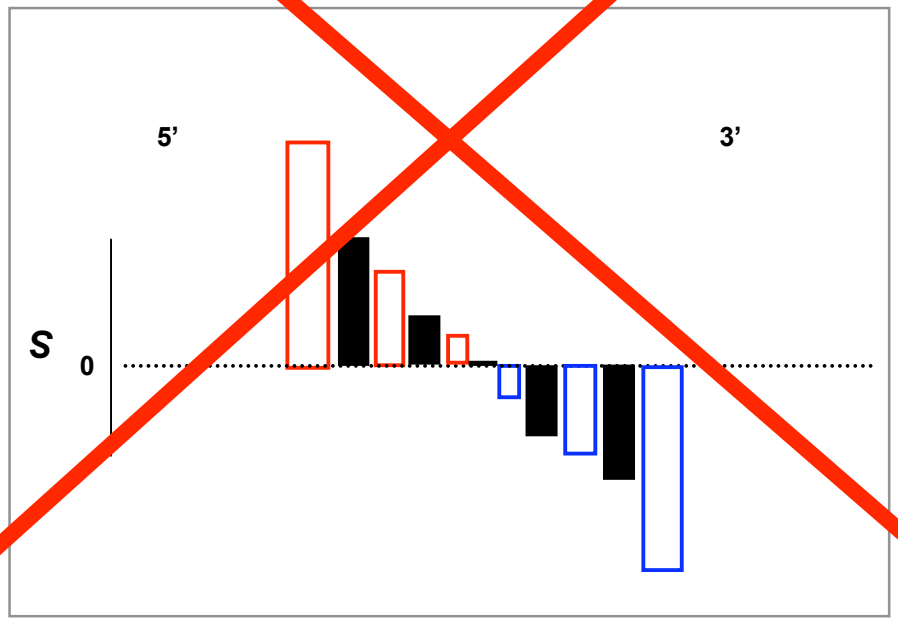




Replication specifically organized

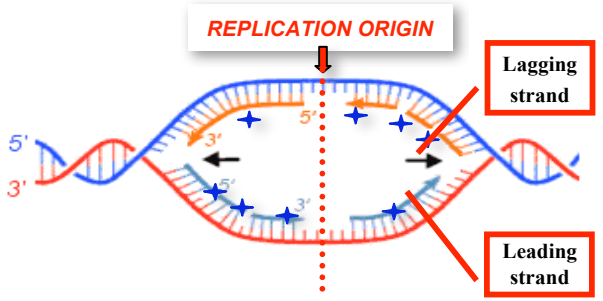
Transcription

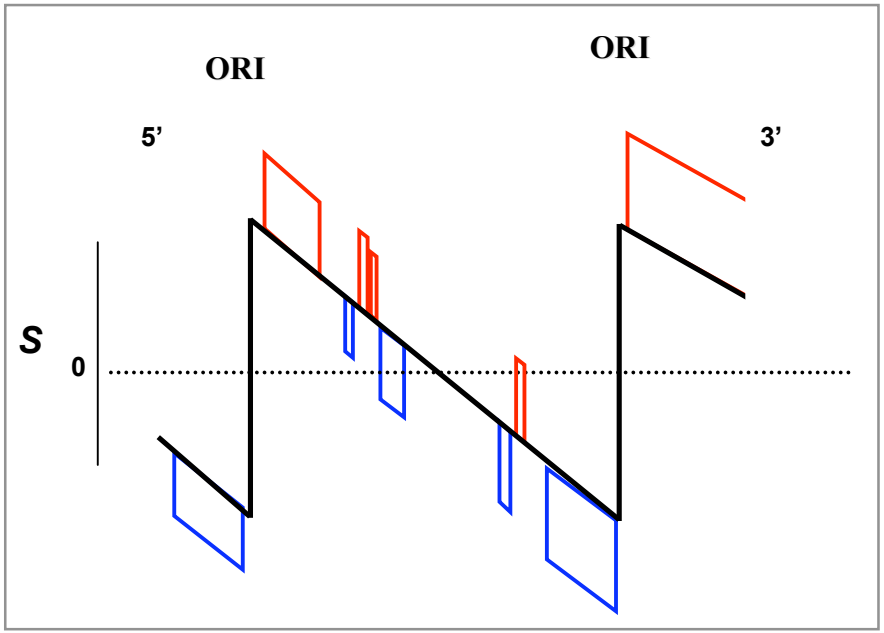




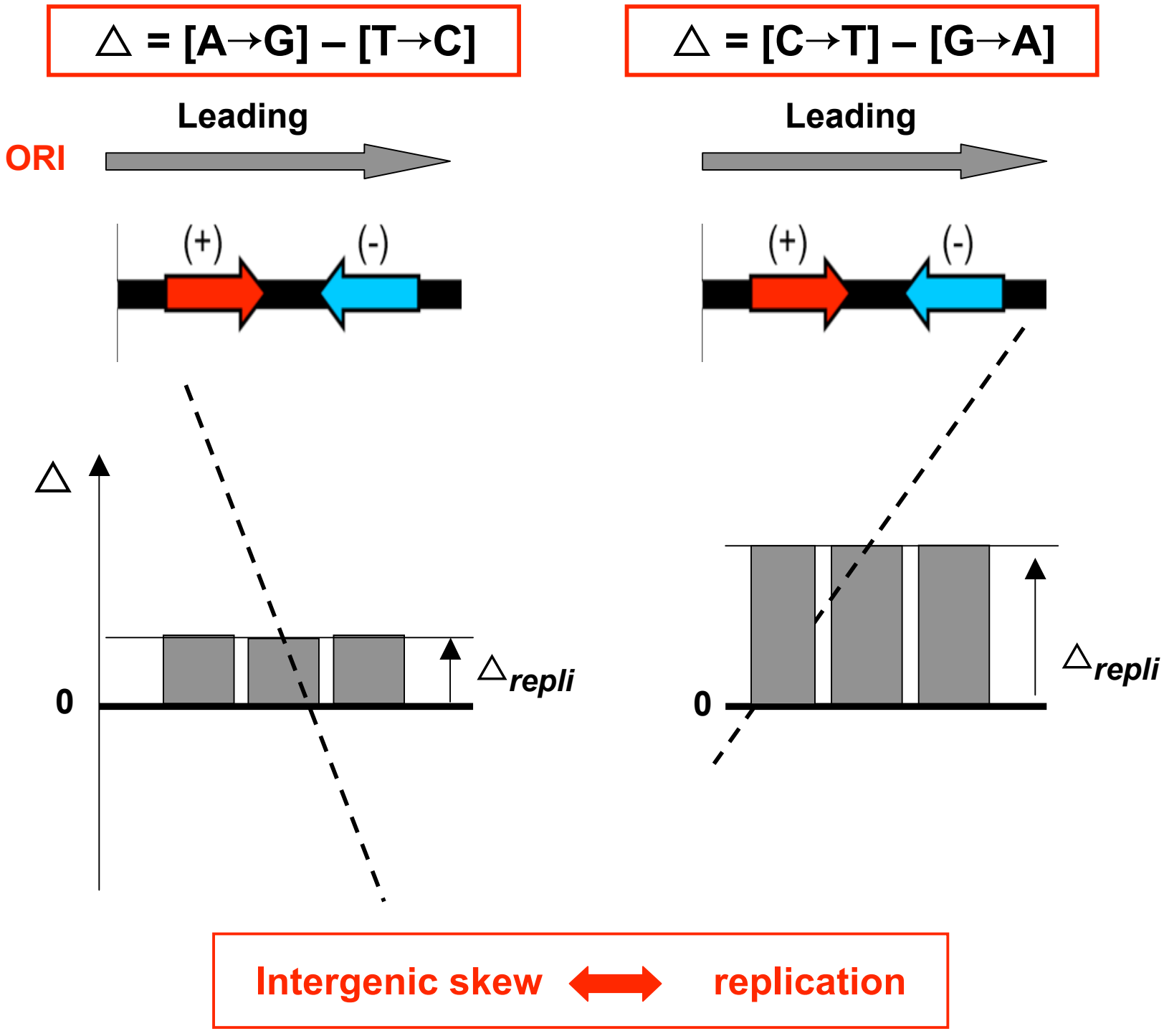
Transcription specifically organized

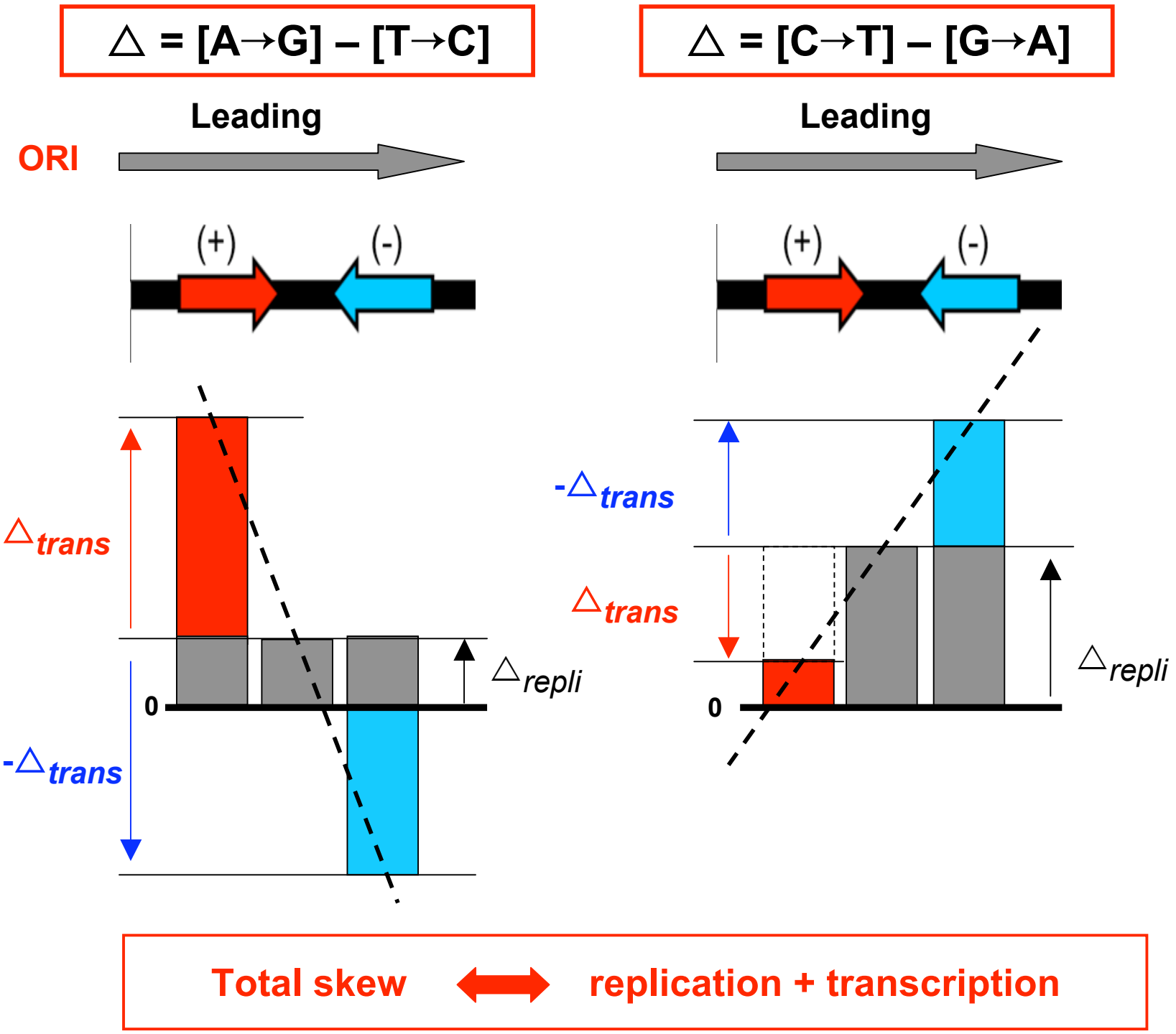
Replication





Replication specifically organized


△ = [A→G] – [T→C]
△ = [C→T] – [G→A]
Leading
Leading
ORI
(+)
(-)
(+)
(-)
△
0
△repli
0
△repli
Intergenic skew
replication


△ = [A→G] – [T→C]
△ = [C→T] – [G→A]
Leading
Leading
ORI
(+)
(-)
(+)
(-)
△trans
-△trans
△repli
0
-△trans
△trans
△repli
0
Total skew
replication + transcription

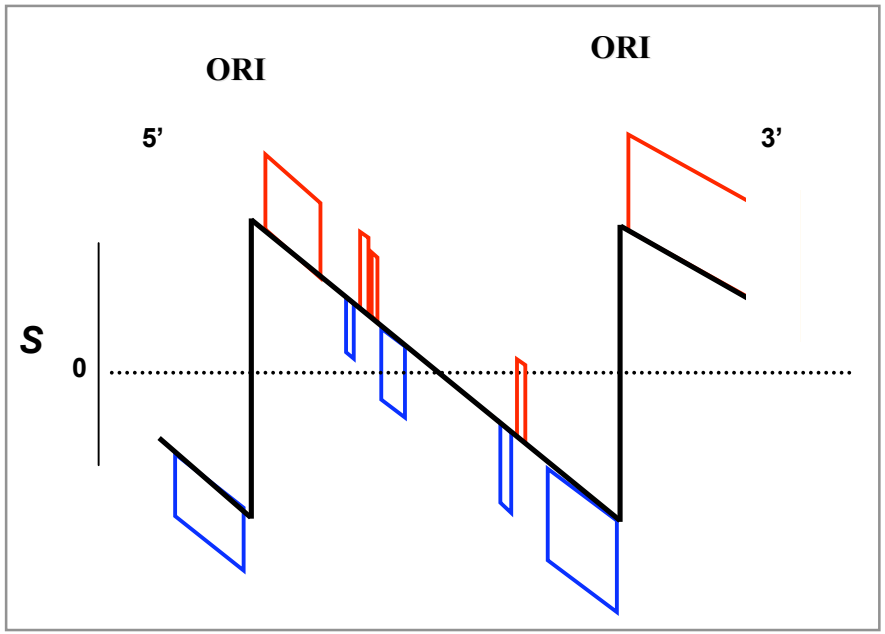
“N” pattern of skew profile results from replication-associated mutational strand asymmetry.

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

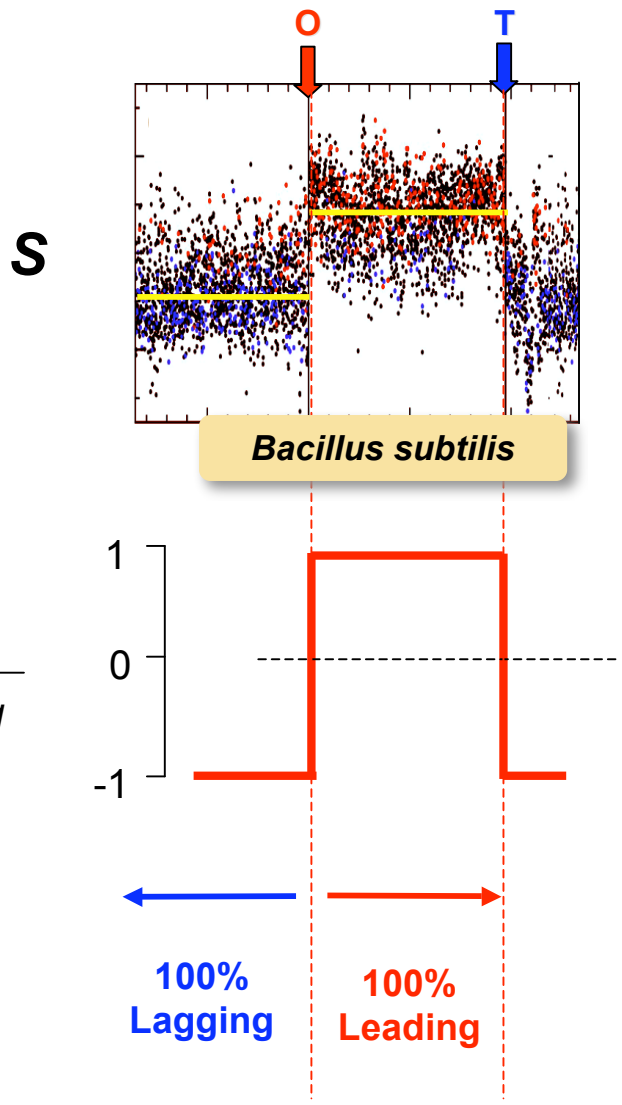
For the first time, the existence of replication-associated mutational asymmetry in a eukaryotic organism is demonstrated.

QUESTION 3:

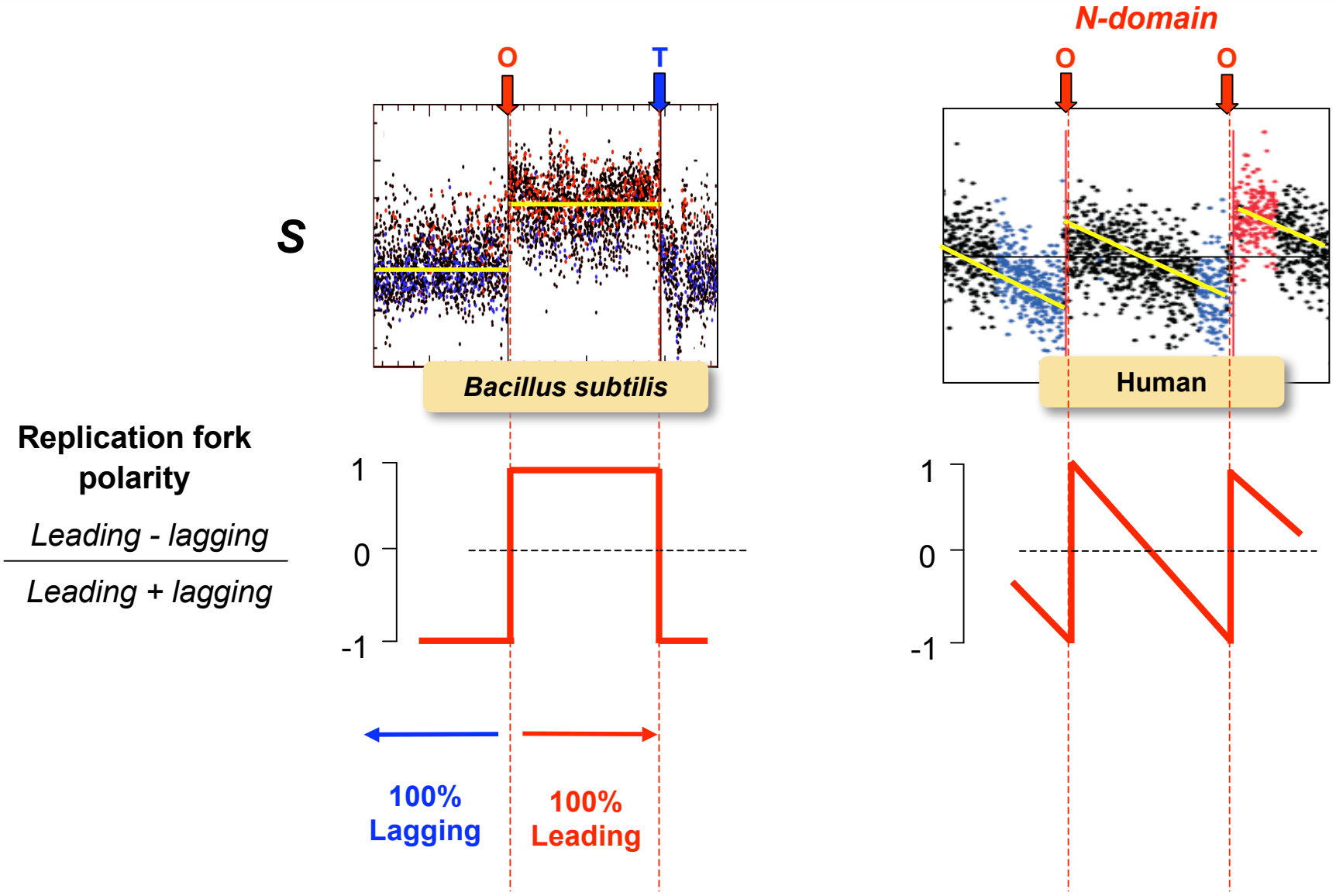
What model of replication can explain this N-pattern ?



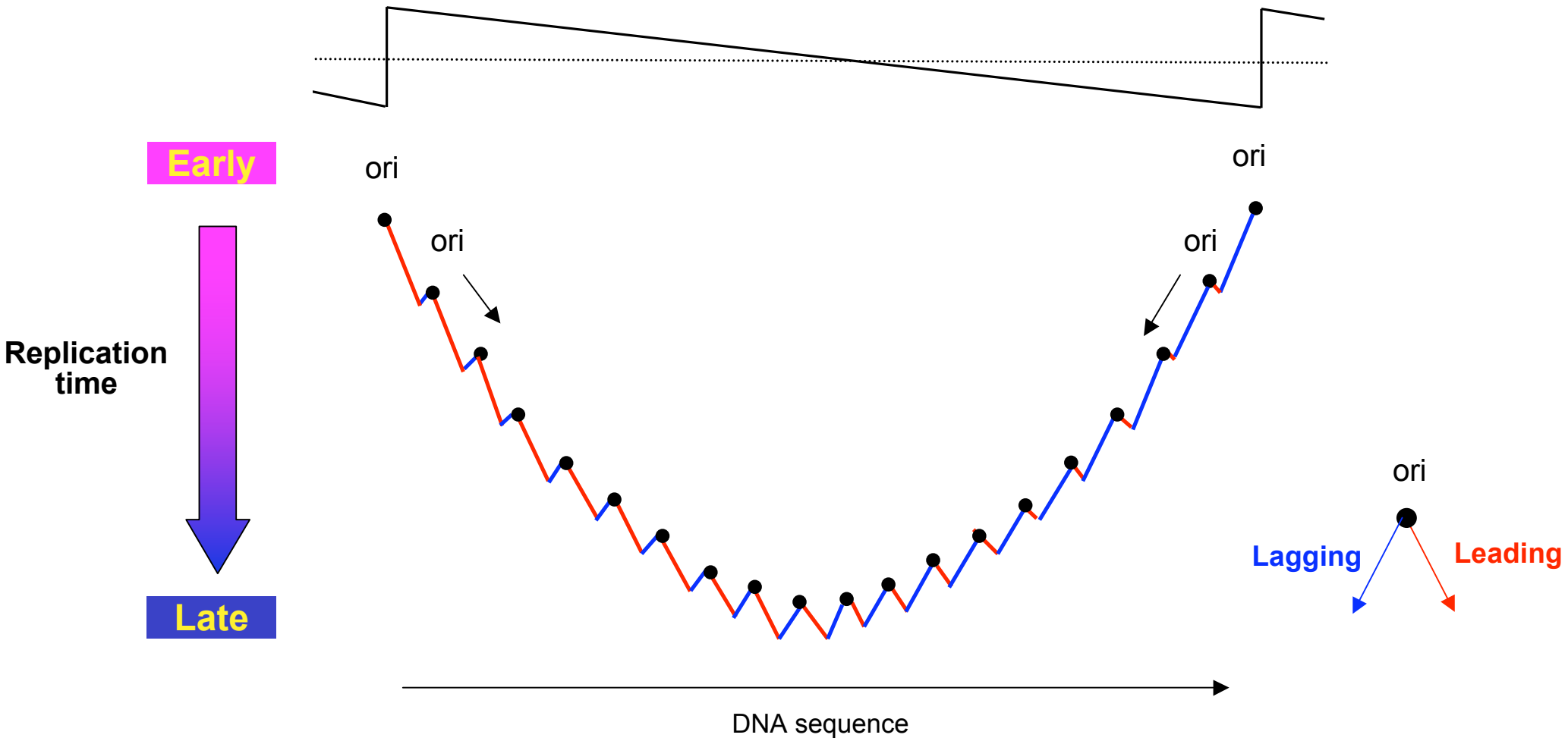
MODEL: N-pattern results from gradient of replication fork polarity



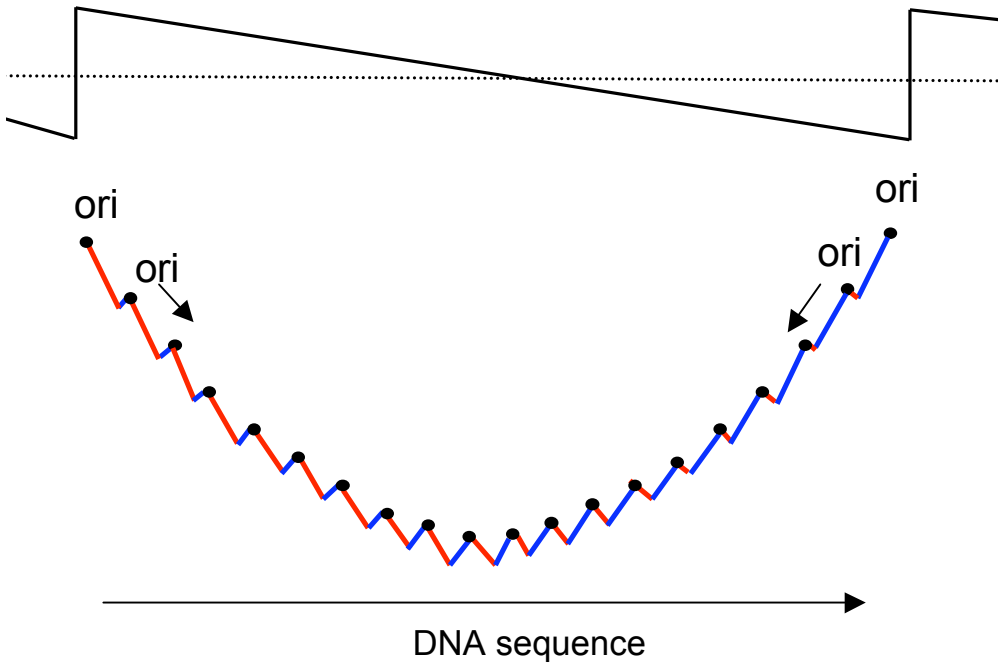
MODEL: N-pattern results from gradient of replication fork polarity



MODEL: N-pattern results from gradient of replication fork polarity



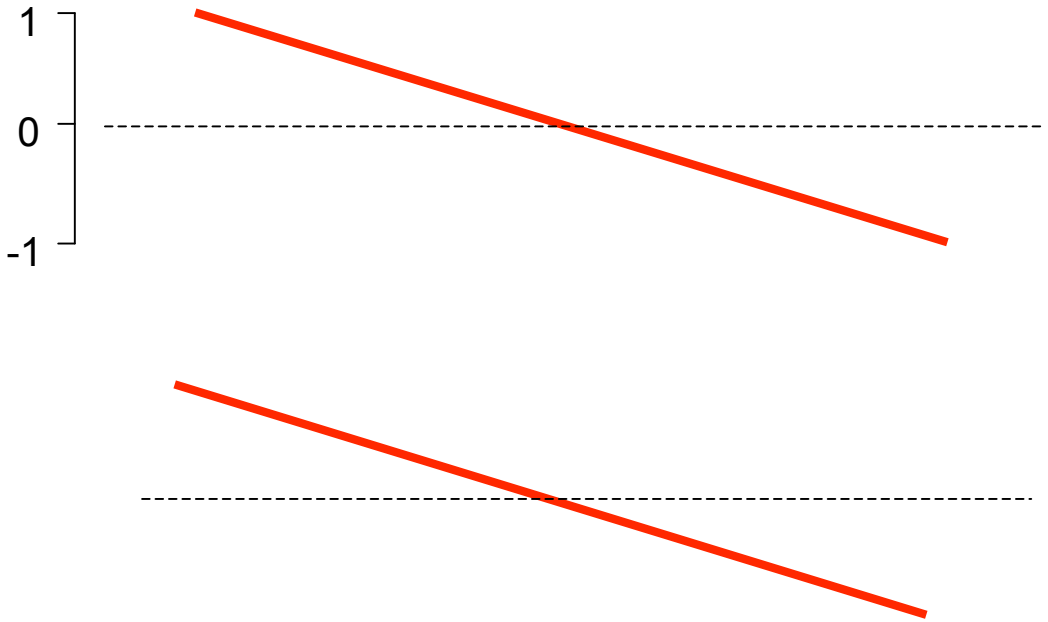
Derivative of replication timing ≈ replication fork polarity



Replication time

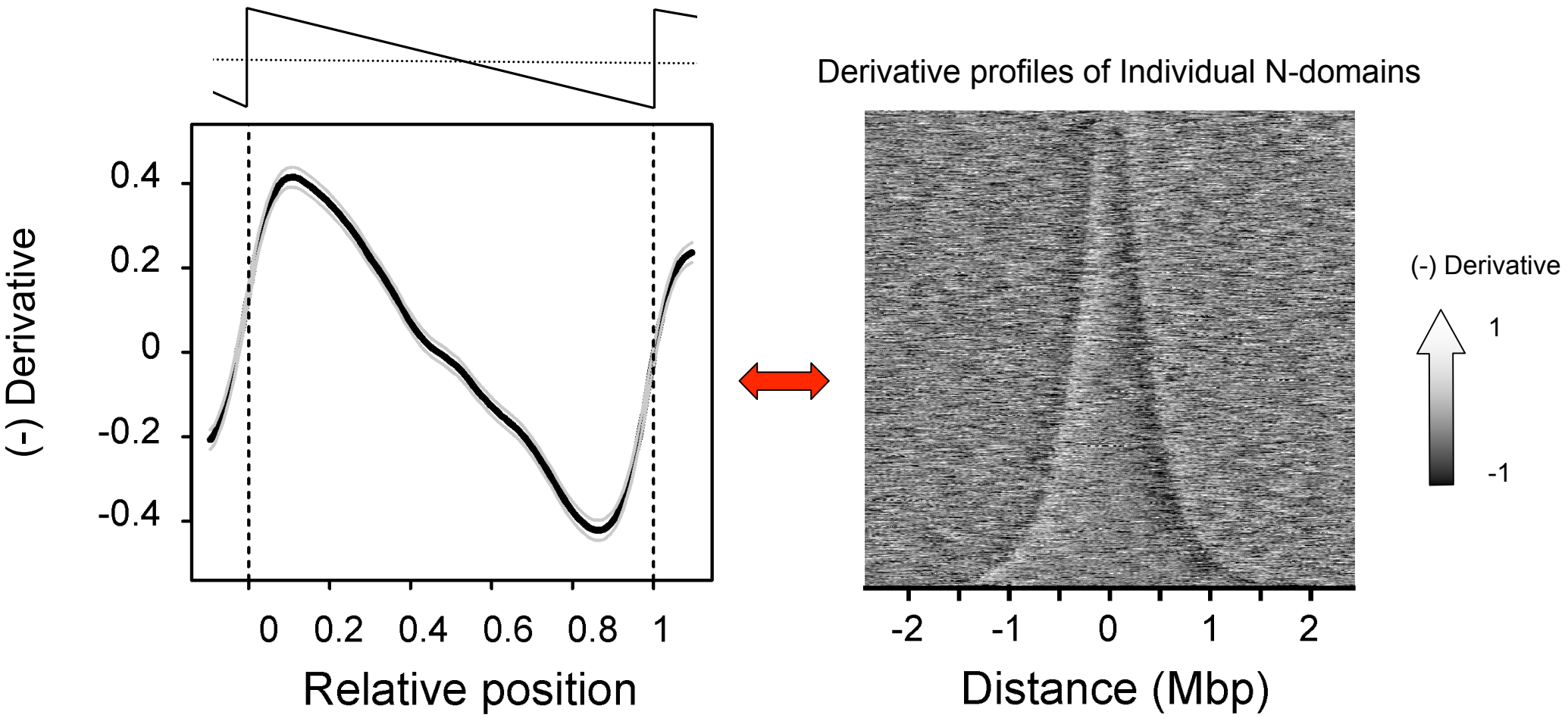
Replication fork polarity

$$\frac{\textit{Leading - lagging}}{\textit{Leading + lagging}}$$



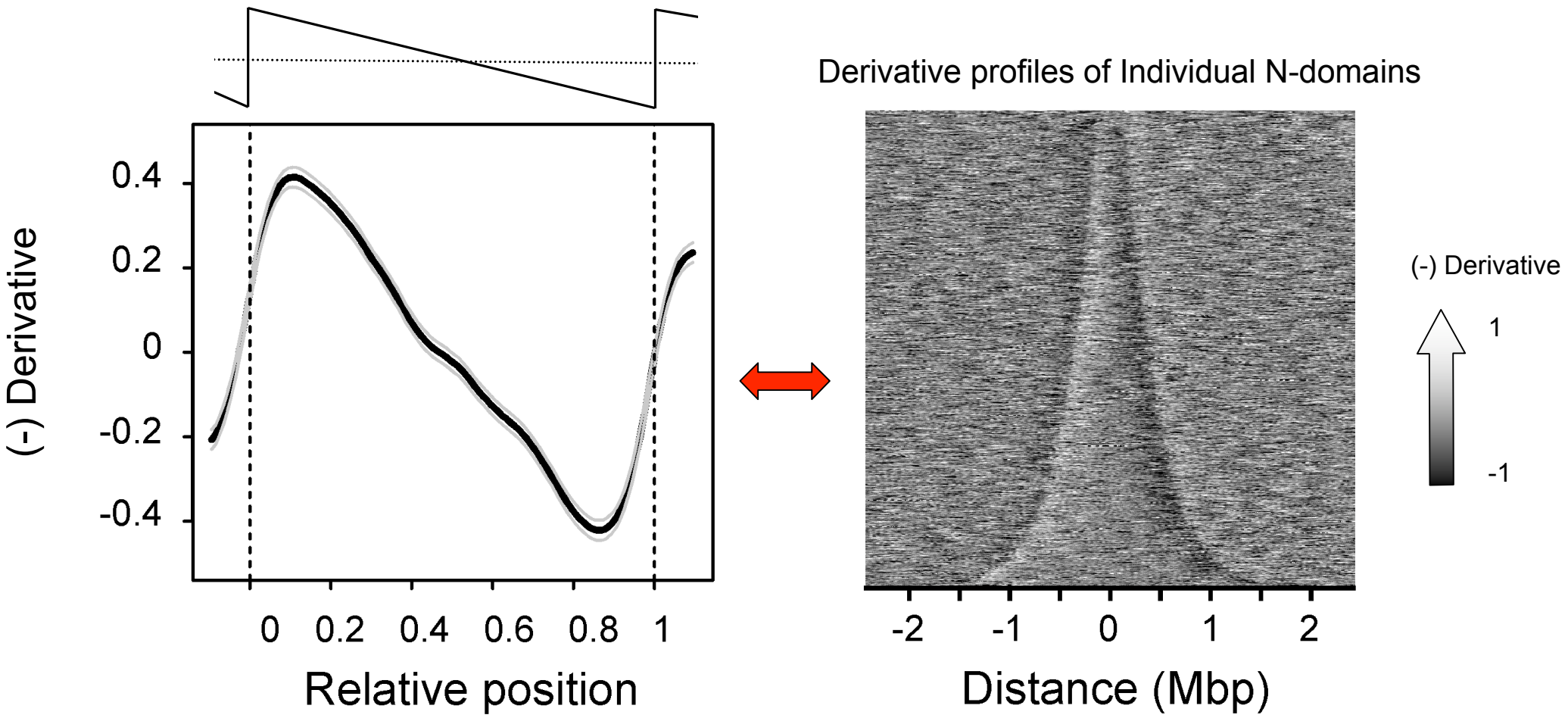
(-) derivative of replication timing profile

EXPERIMENTAL VERIFICATION



Derivative of replication timing profile reproduces perfectly the “N” pattern

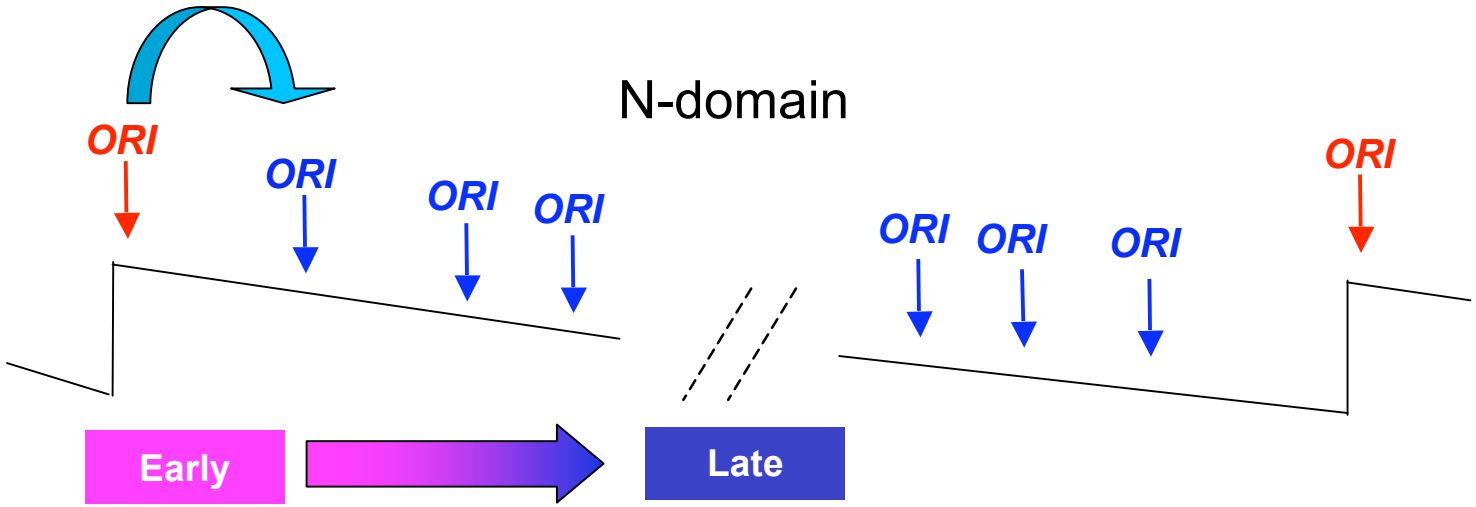
EXPERIMENTAL VERIFICATION

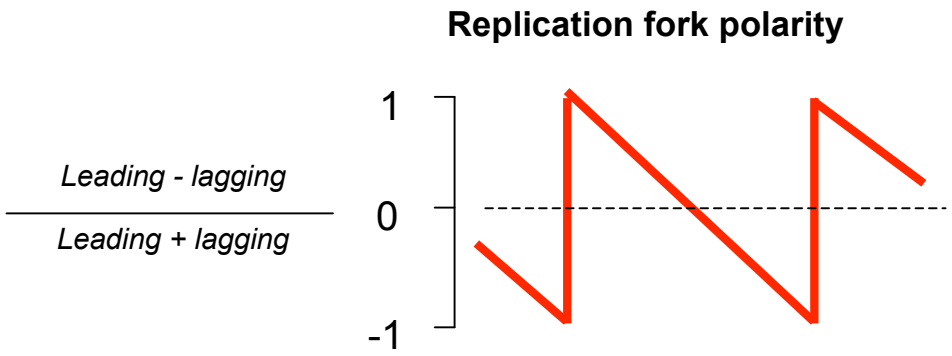


Derivative of replication timing profile reproduces perfectly the “N” pattern

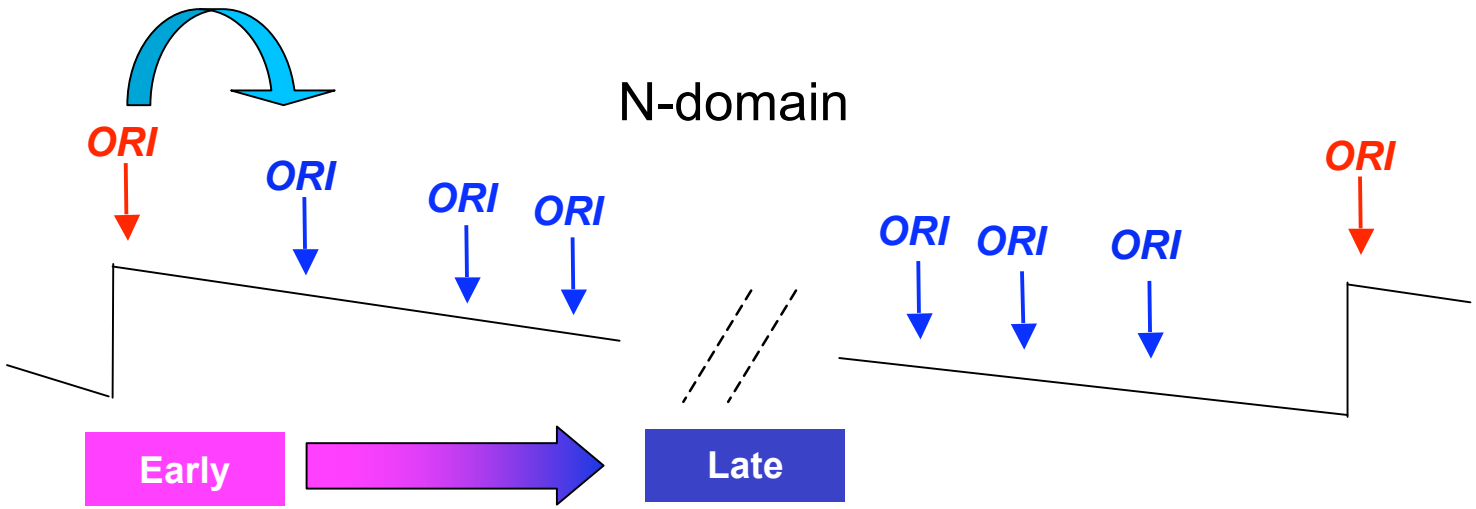
N-pattern results from gradient of replication fork polarity

DOMINO MODEL FOR N-DOMAIN REPLICATION





DOMINO MODEL FOR N-DOMAIN REPLICATION



PERSPECTIVES

Does this kind of replication program reflect a higher order chromatin structure ?

Study by using the data of :

- Nucleosome position
- Epigenetic markers
- 3C (Chromosome Conformation Capture)
- Hi-C chromosomal interaction

Acknowledgments and collaborations

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

(CGM, Gif sur Yvette)

Arach Goldar

(CEA, iBiTec-S, Gif-sur-Yvette)

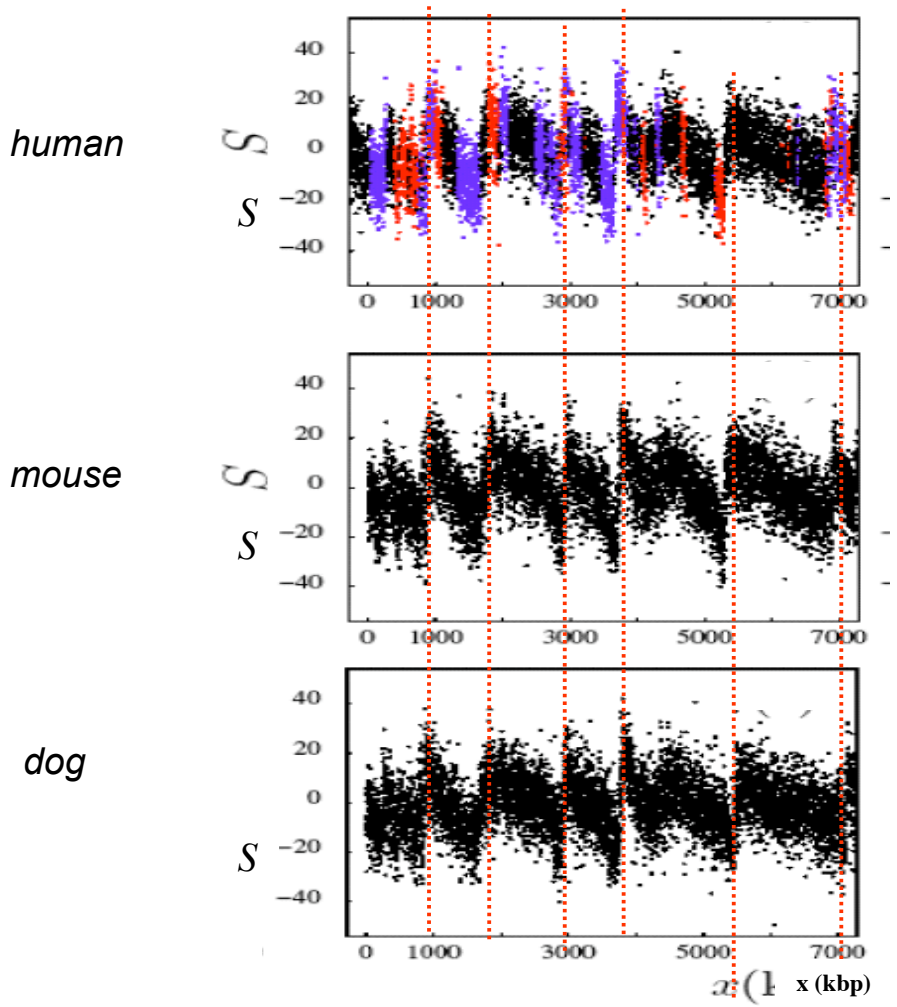
Aurélien Rappailles
Guillaume Guilbaud
Olivier Hyrien
(ENS-Paris)

Benjamin Audit
Antoin Baker
Lamia Zaghloul
Alain Arneodo

(ENS-Lyon)

Merci!

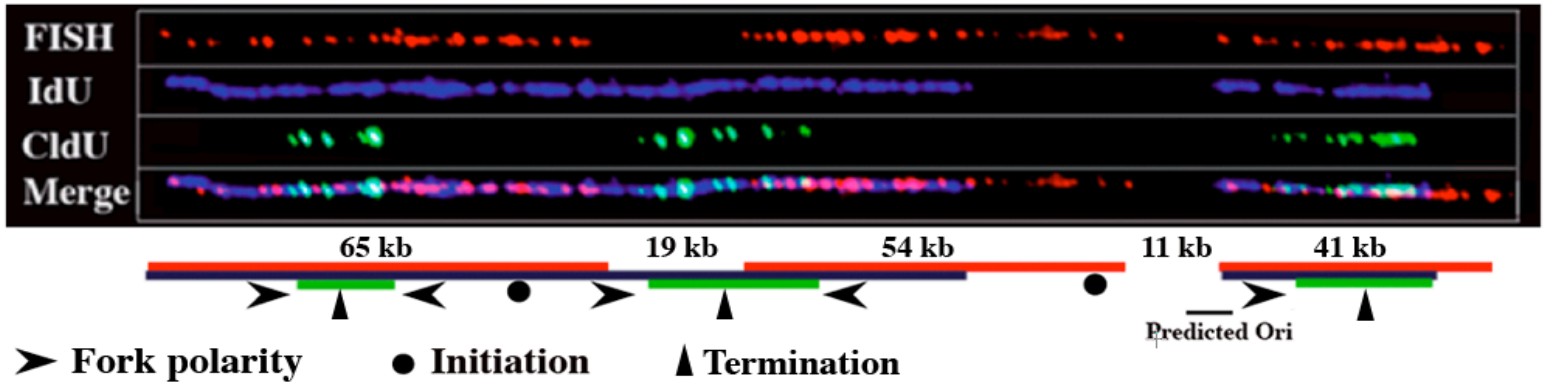
CONSERVATION OF N-DOMAINS IN MAMMALIAN GENOMES

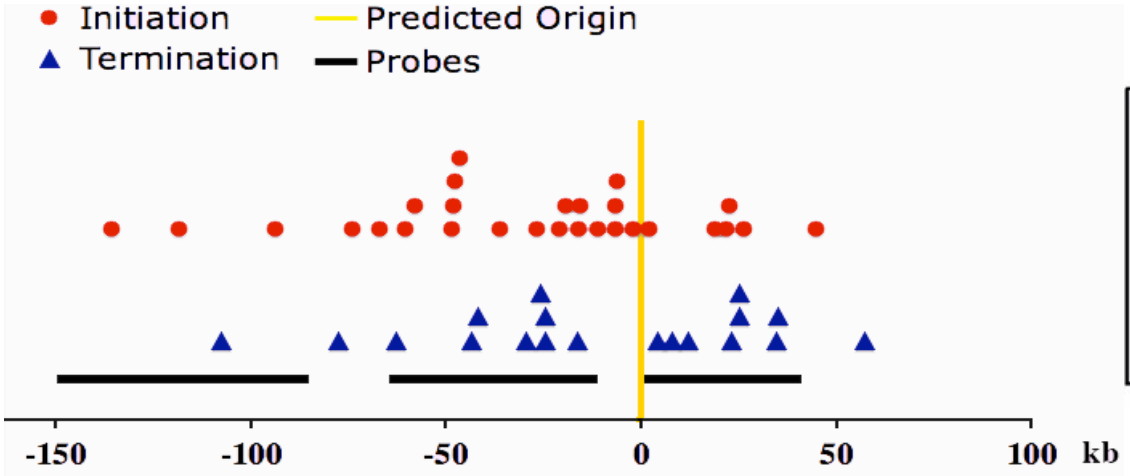


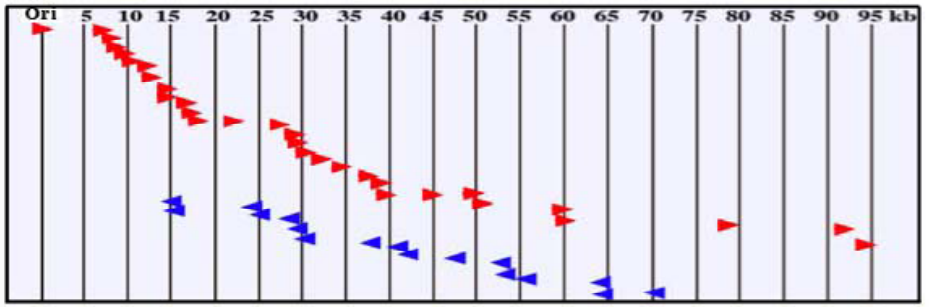
Genes have changed but the replication program has remained identical

Mapping of initiations/terminations along a predicted origin

G. Guilbaud, O. Hyrien (ENS Paris)



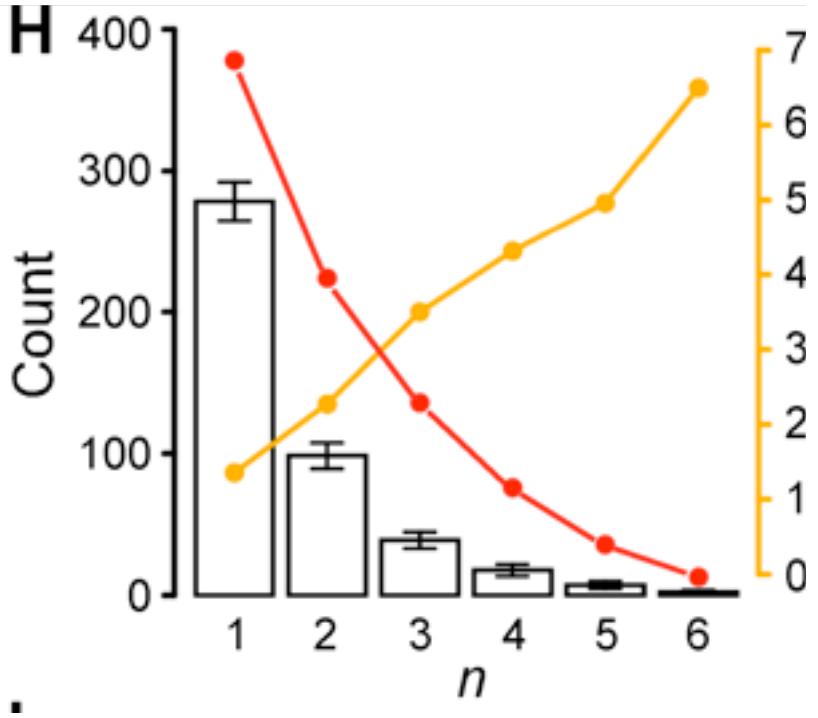


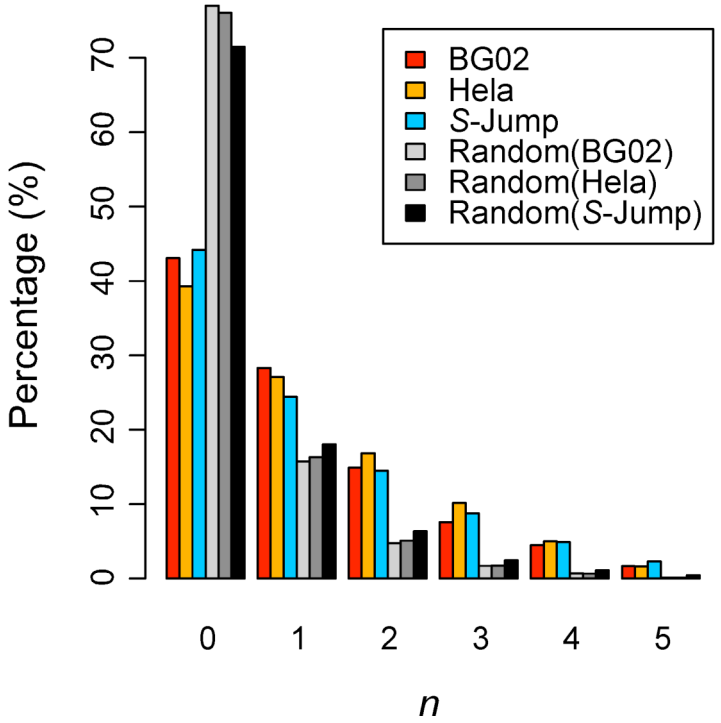


Initiations occur both at the predicted origin and elsewhere in the N-domain

A majority of forks move away from the predicted origin

Distance between each upward jump and the closest initiation zone





Upward jumps are significantly associated with replication initiation zones