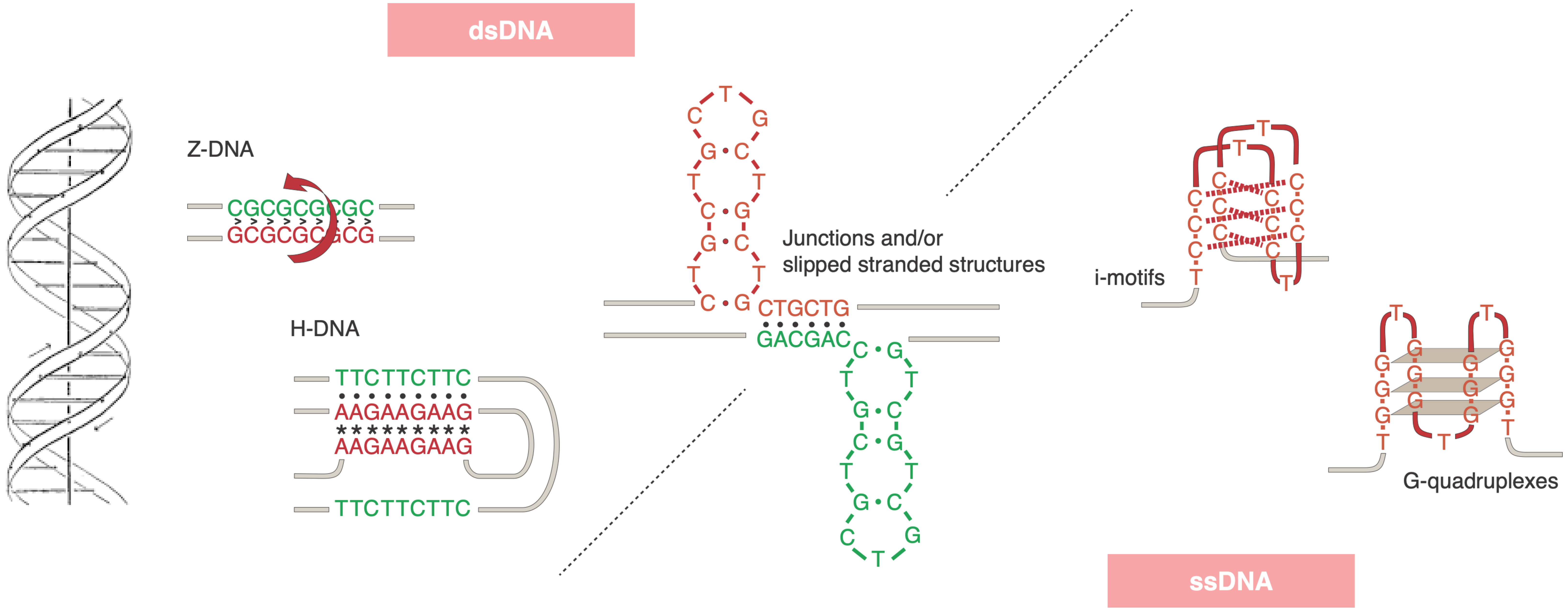
# Sequences that stall replication and shape the genome

Julian E. Sale
MRC Laboratory of Molecular Biology
Cambridge

NIH DNA Repair Interest Group
January 10th 2023

# The challenge to replication posed by alternative DNA structures



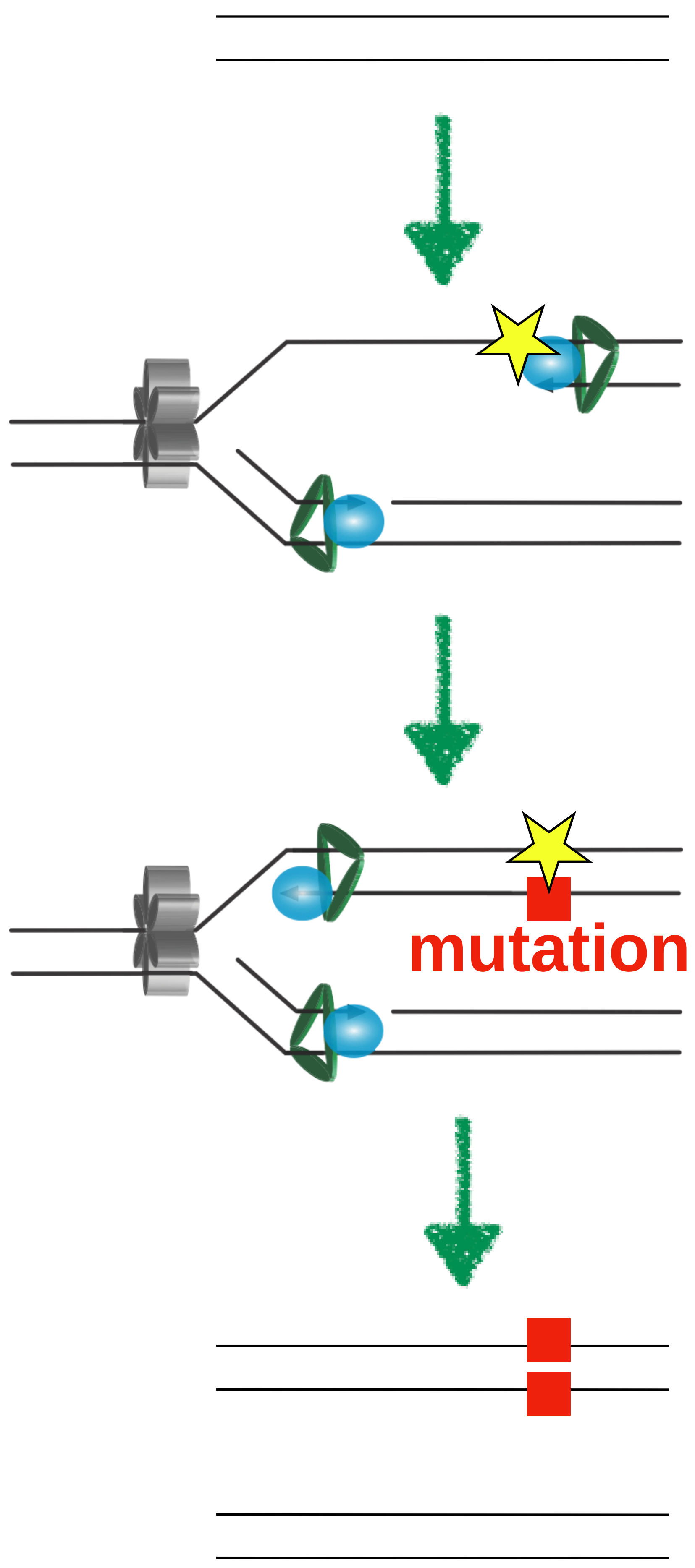
- millions of loci in the human genome - often found in repetitive / low complexity sequence
- Hotspots for chromosomal rearrangements, copy number variations, mutagenesis and epigenetic instability
- Trinucleotide repeat expansion disorders

| | |
|---|---|
| Huntington’s disease | (CAG)n |
| Friedreich’s ataxia | (GAA)n |
| Fragile X syndrome | (CGG)n |
| etc. etc. | |

**How does the replisome deal with secondary structures in the template?**

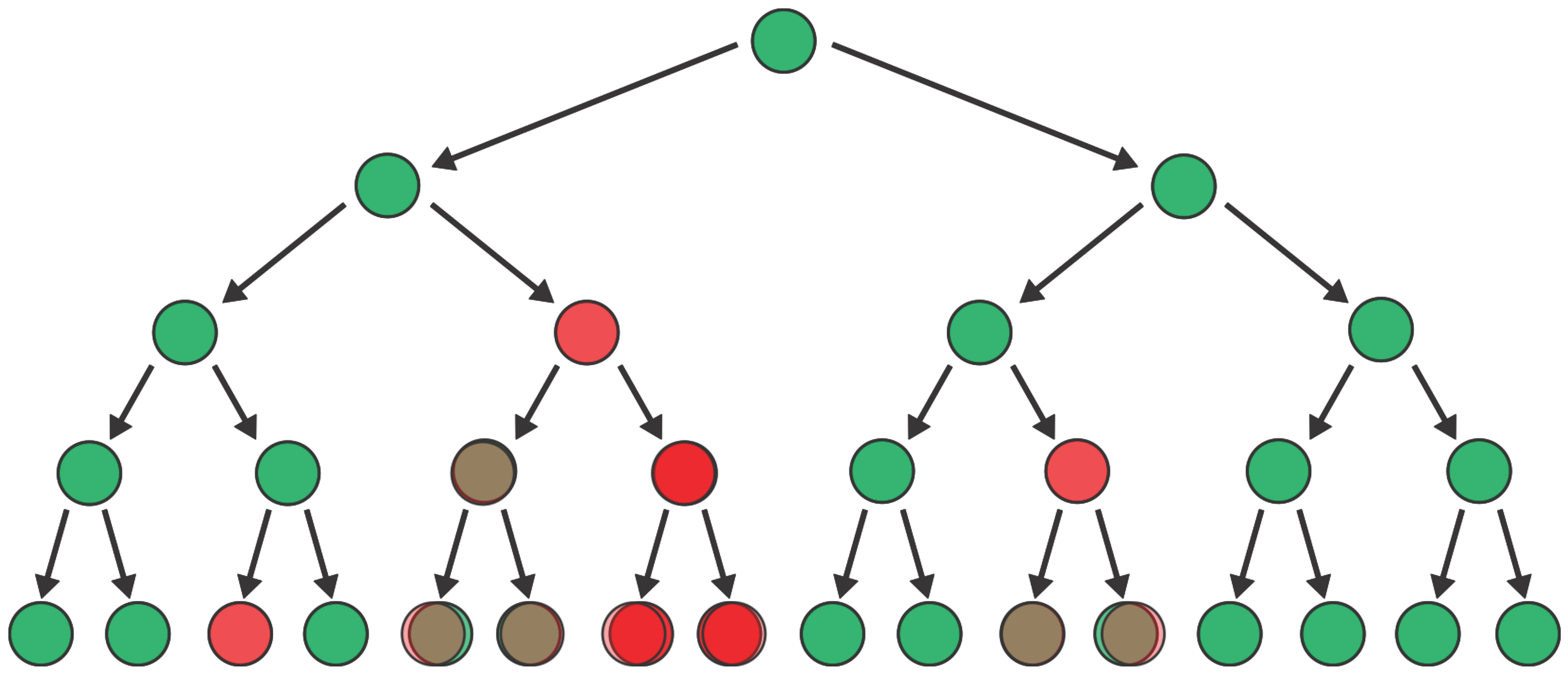
**What drives the evolution of genomic sequences with structure-forming potential?**

# The problem of monitoring replication fork stalling *in vivo*



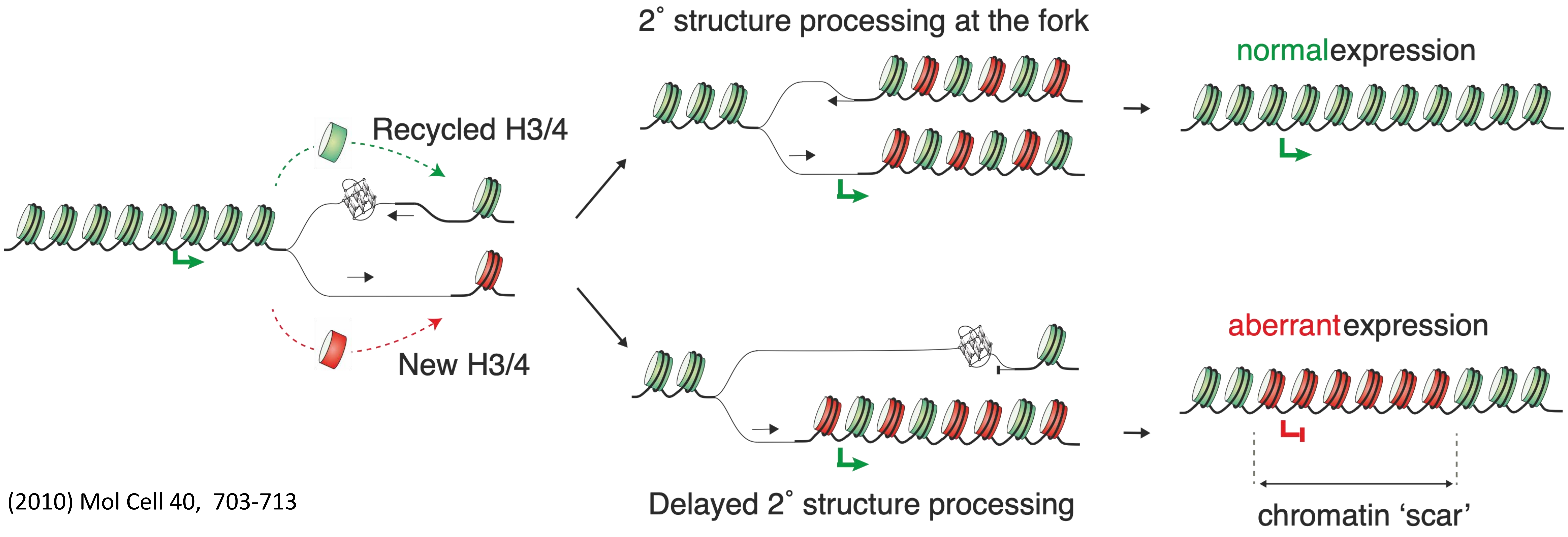
polymerase stalling is transient
and therefore hard to detect …

… mutagenic outcomes are traceable but relatively rare

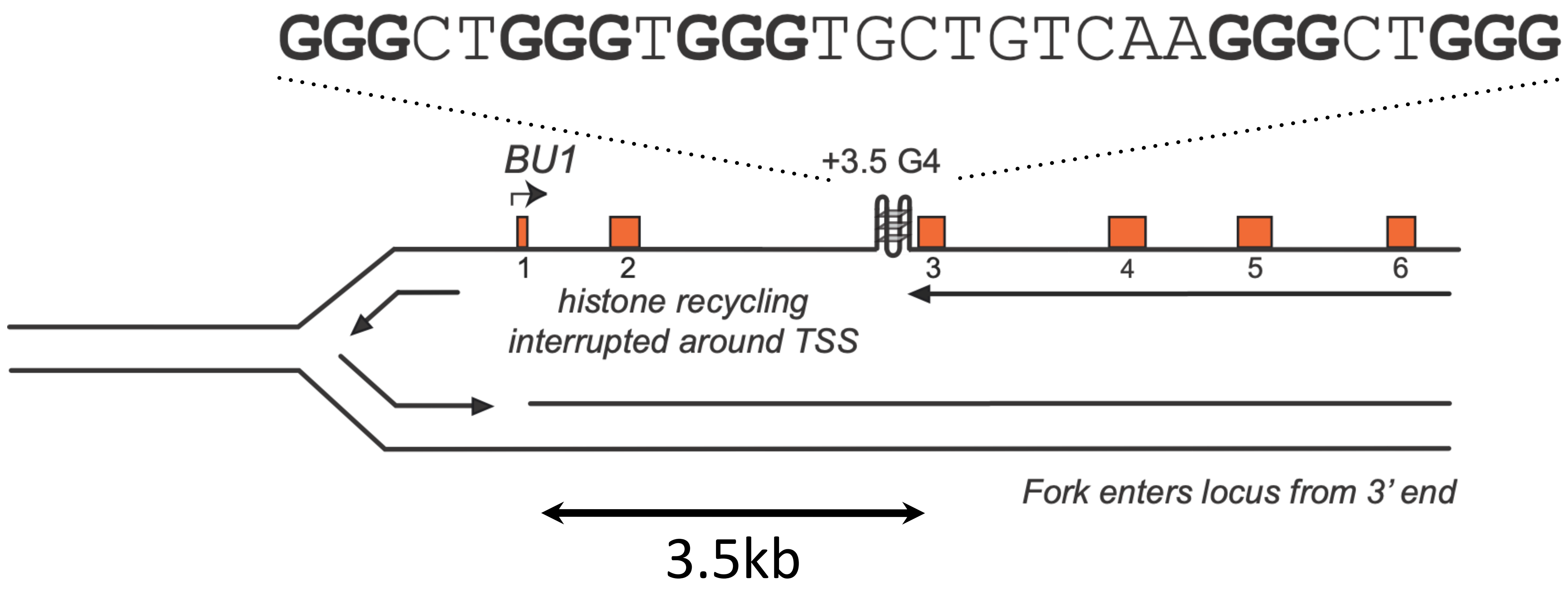


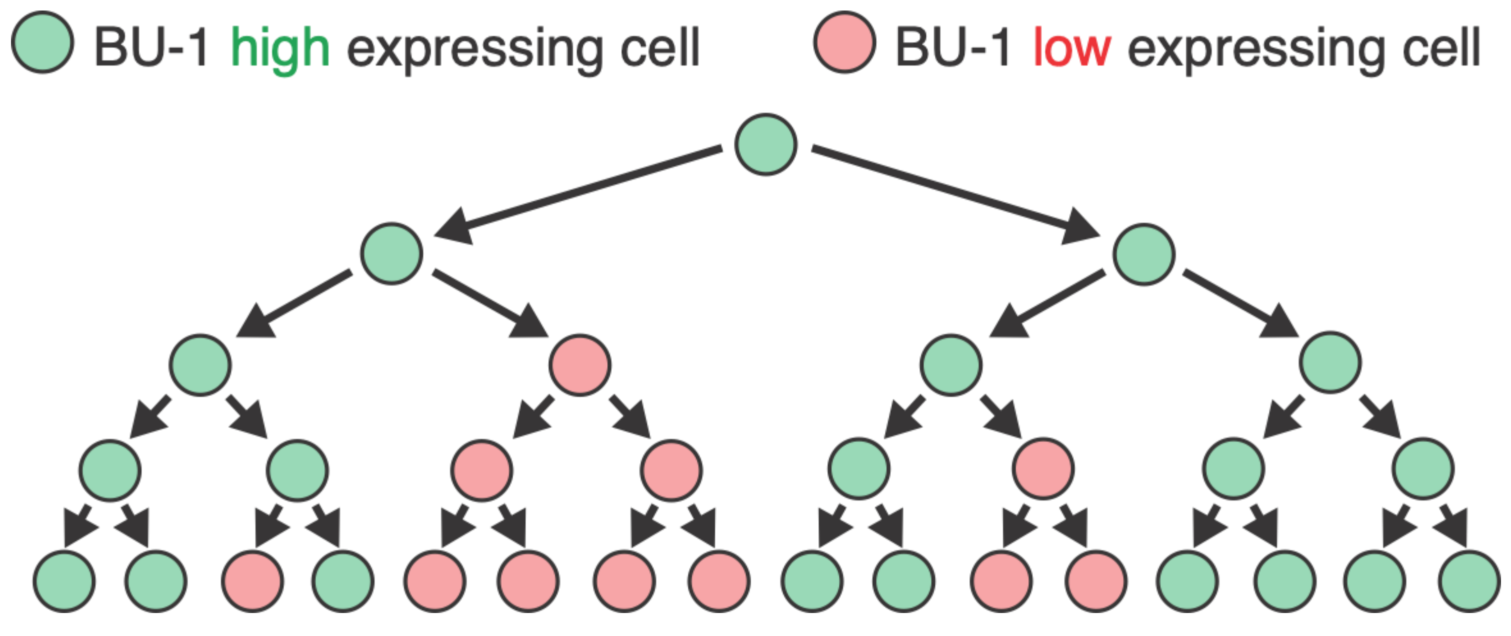
How can non-mutagenic episodes of fork stalling
be reported in an expanding cell population?

# Using local loss of epigenetic memory to monitor delayed replication of DNA secondary structures



Sarkies et al. (2010) Mol Cell 40, 703-713

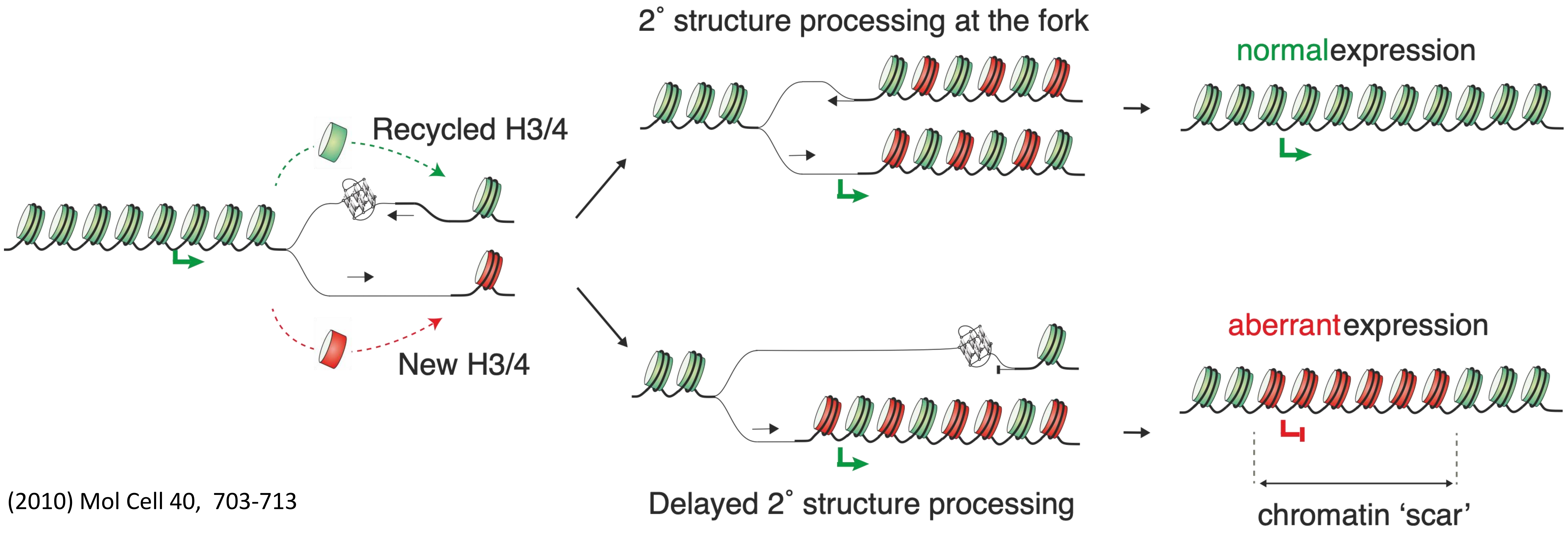

GGGCTGGGTGGGTGCTGTCAAGGGCTGGG
BU1
+3.5 G4
1
2
3
4
5
6
histone recycling interrupted around TSS
Fork enters locus from 3’ end
3.5kb



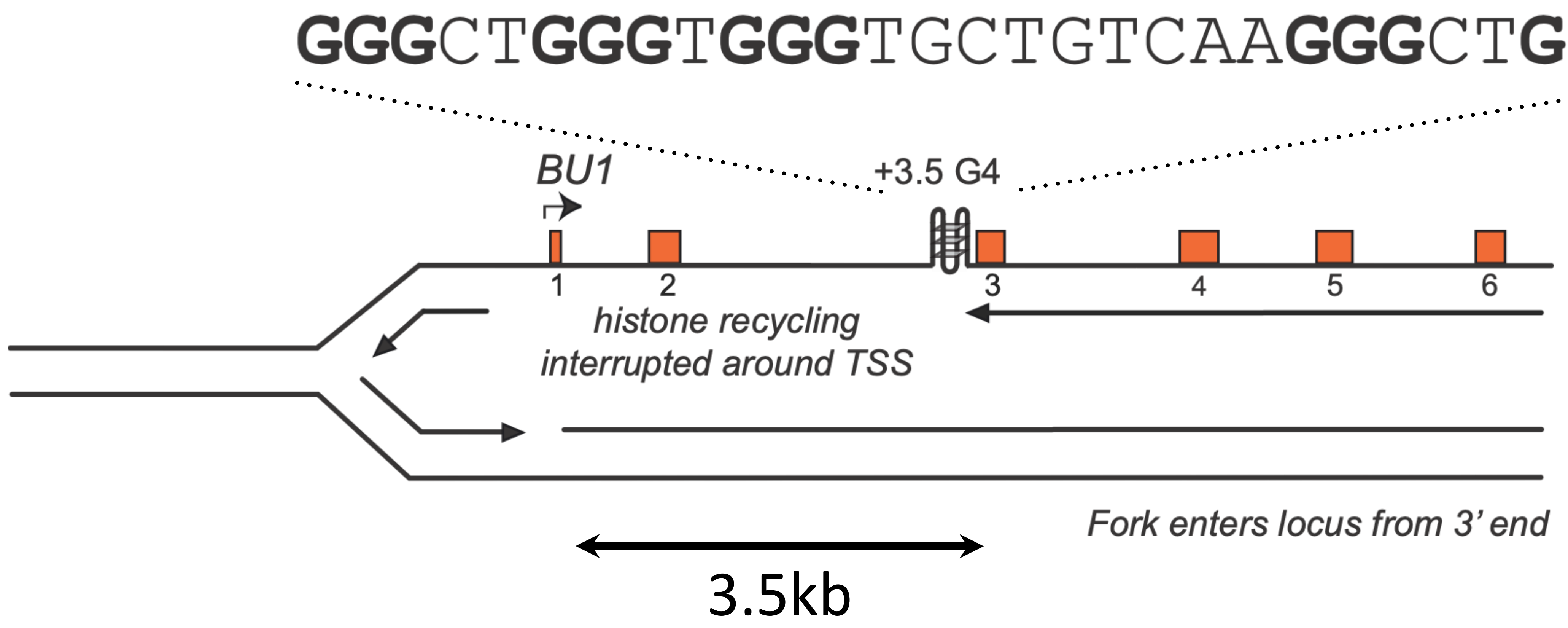
$$p_{\text{Bu-1 loss}} = p_{\text{stall}} + p_{\text{gap}}$$

Sarkies et al. (2012) NAR 40, 1485-1498
Schiavone, Guilbaud et al. (2014) EMBO J 33, 2507-20

# Using local loss of epigenetic memory to monitor delayed replication of DNA secondary structures



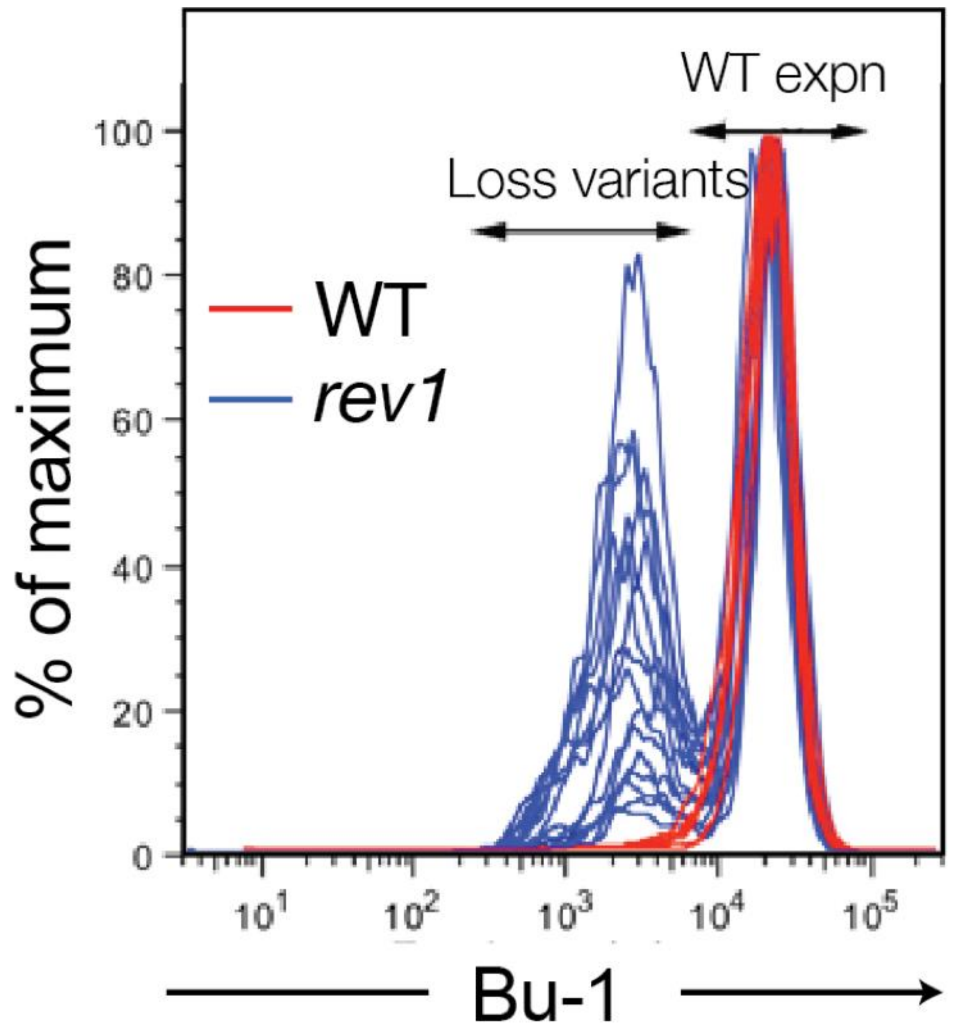
Sarkies et al. (2010) Mol Cell 40, 703-713



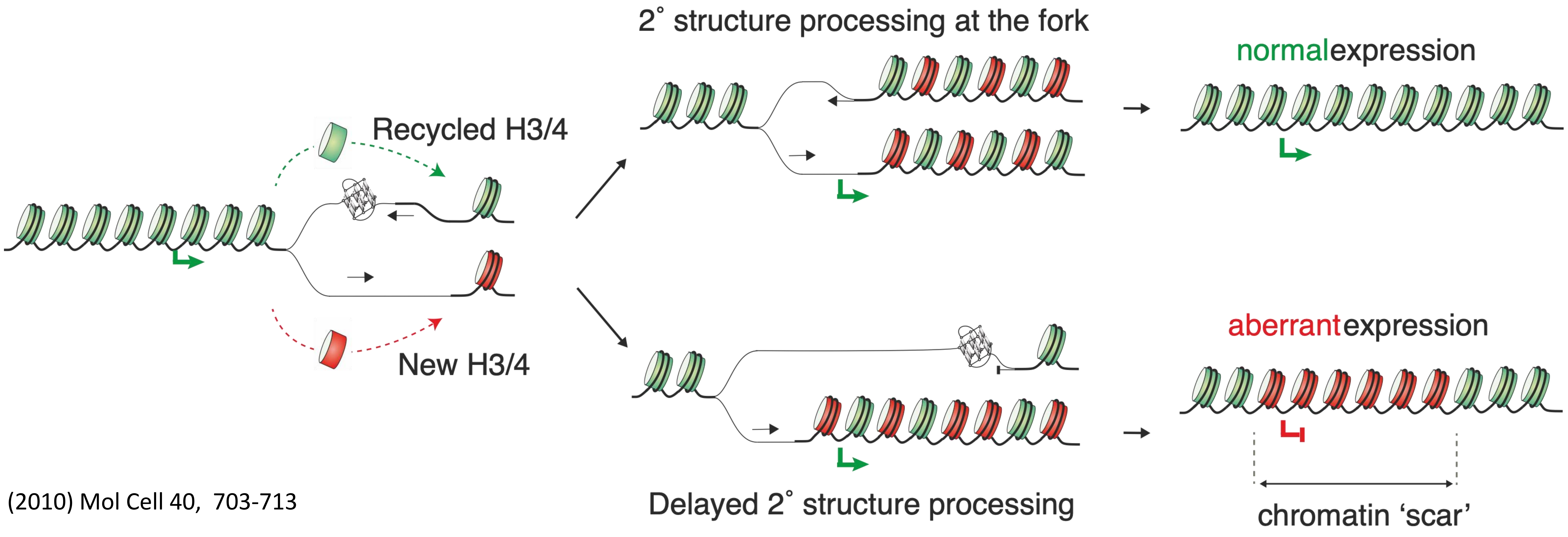
GGGCTGGGTGGGTGCTGTCAAGGGCTGGG
BU1
+3.5 G4
1
2
3
4
5
6
histone recycling interrupted around TSS
Fork enters locus from 3' end
3.5kb

Sarkies et al. (2012) NAR 40, 1485-1498

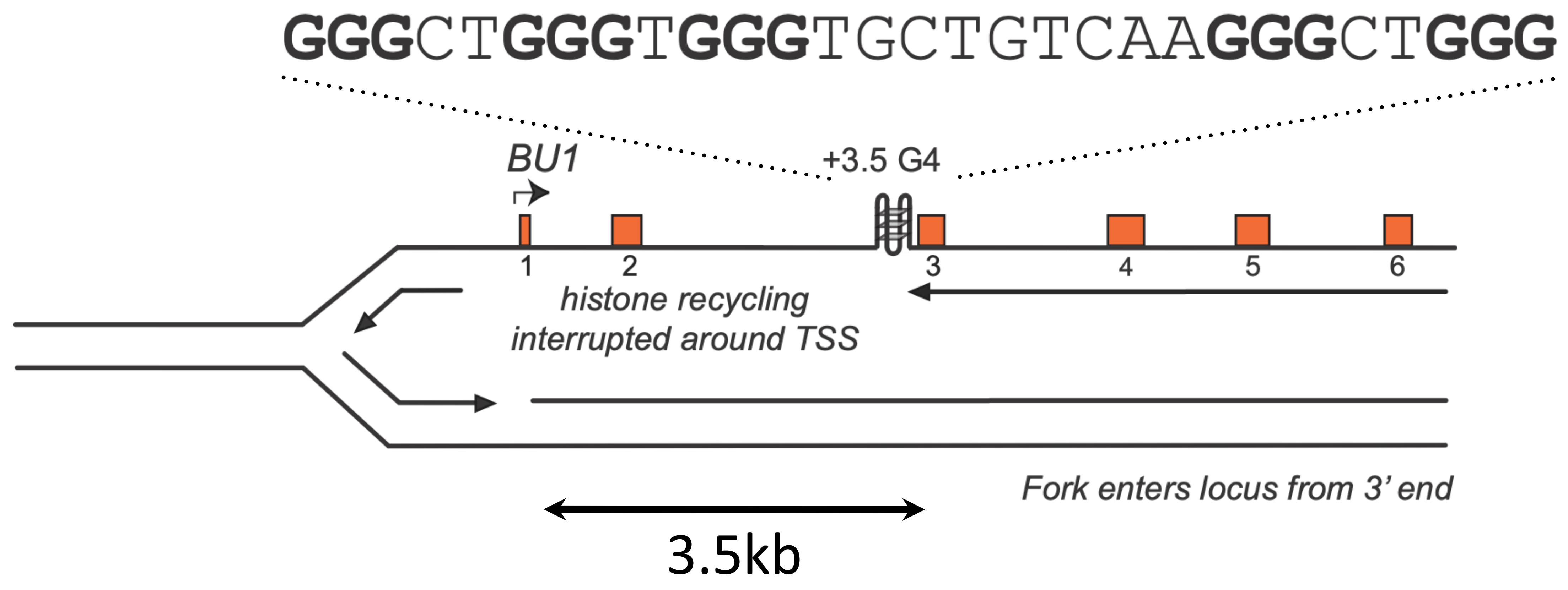
Schiavone, Guilbaud et al. (2014) EMBO J 33, 2507-20



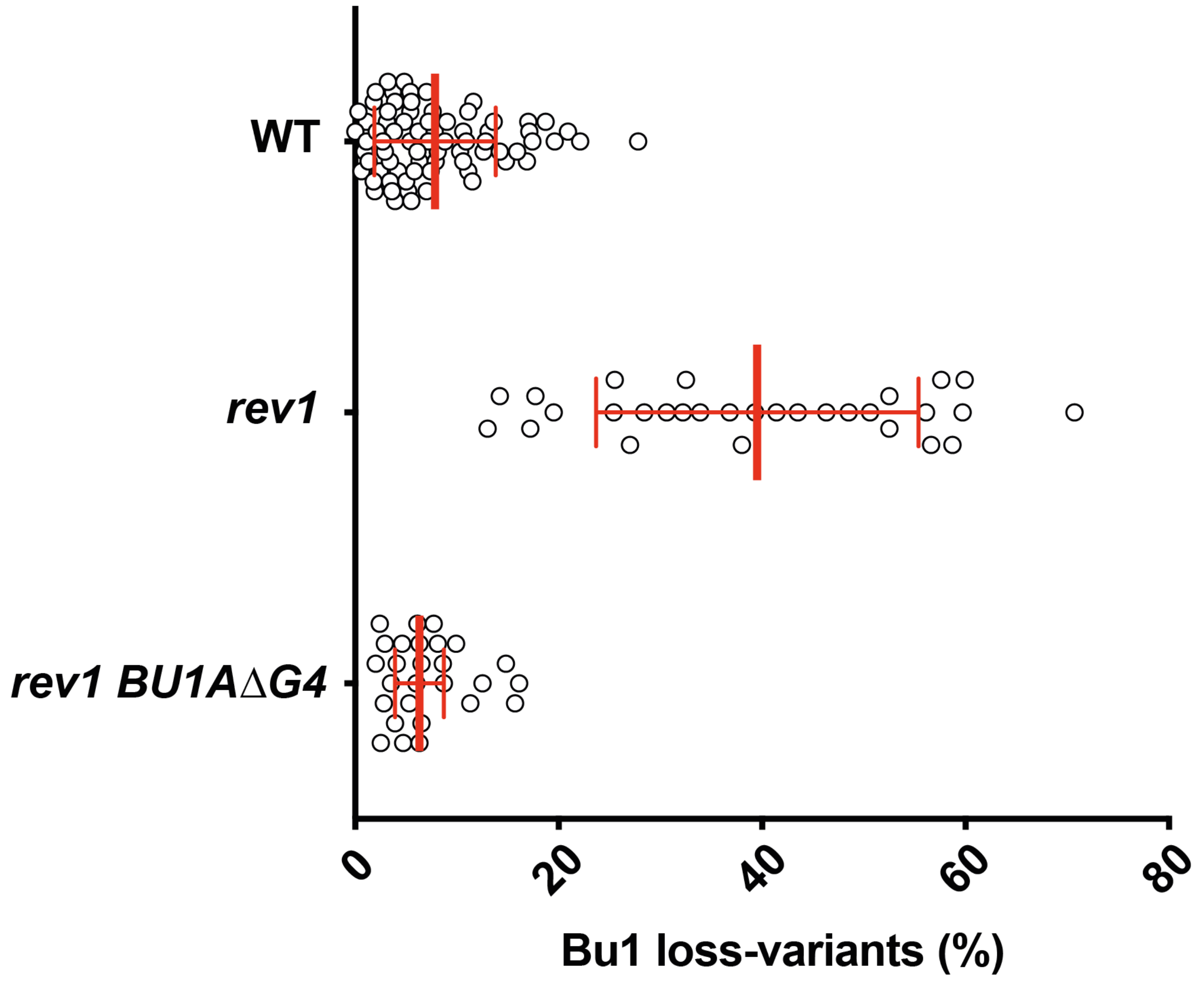
# Using local loss of epigenetic memory to monitor delayed replication of DNA secondary structures



Sarkies et al. (2010) Mol Cell 40, 703-713


GGGCTGGGTGGGTGCTGTCAAGGGCTGGG
BU1
+3.5 G4
1
2
3
4
5
6
histone recycling interrupted around TSS
Fork enters locus from 3' end
3.5kb

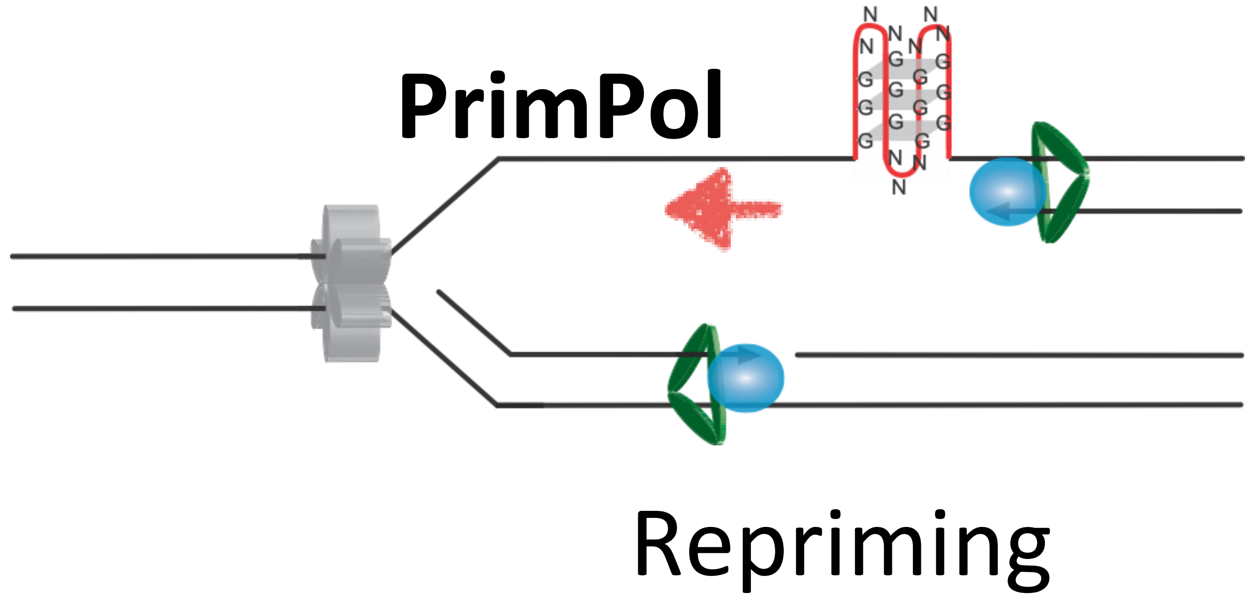
Sarkies et al. (2012) NAR 40, 1485-1498
Schiavone, Guilbaud et al. (2014) EMBO J 33, 2507-20

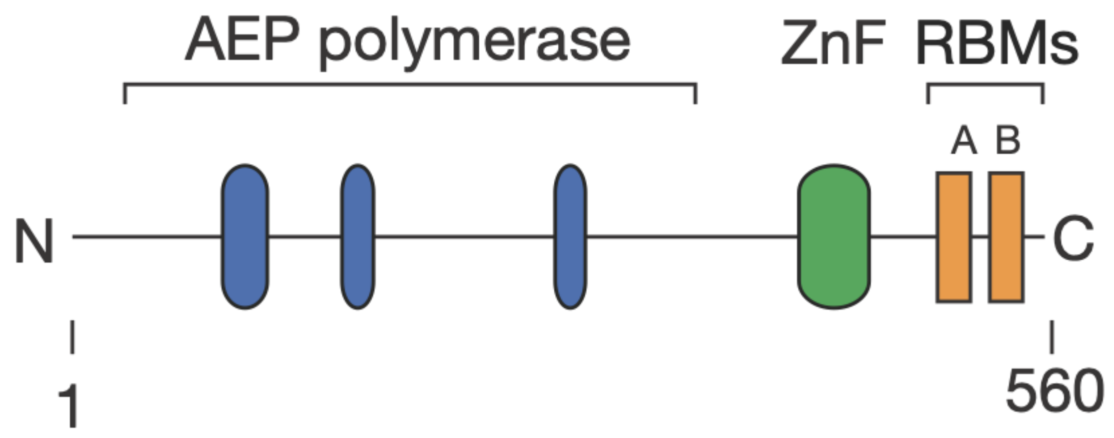


also:

*fancj*
*wrn*
*blm*
*pif1*

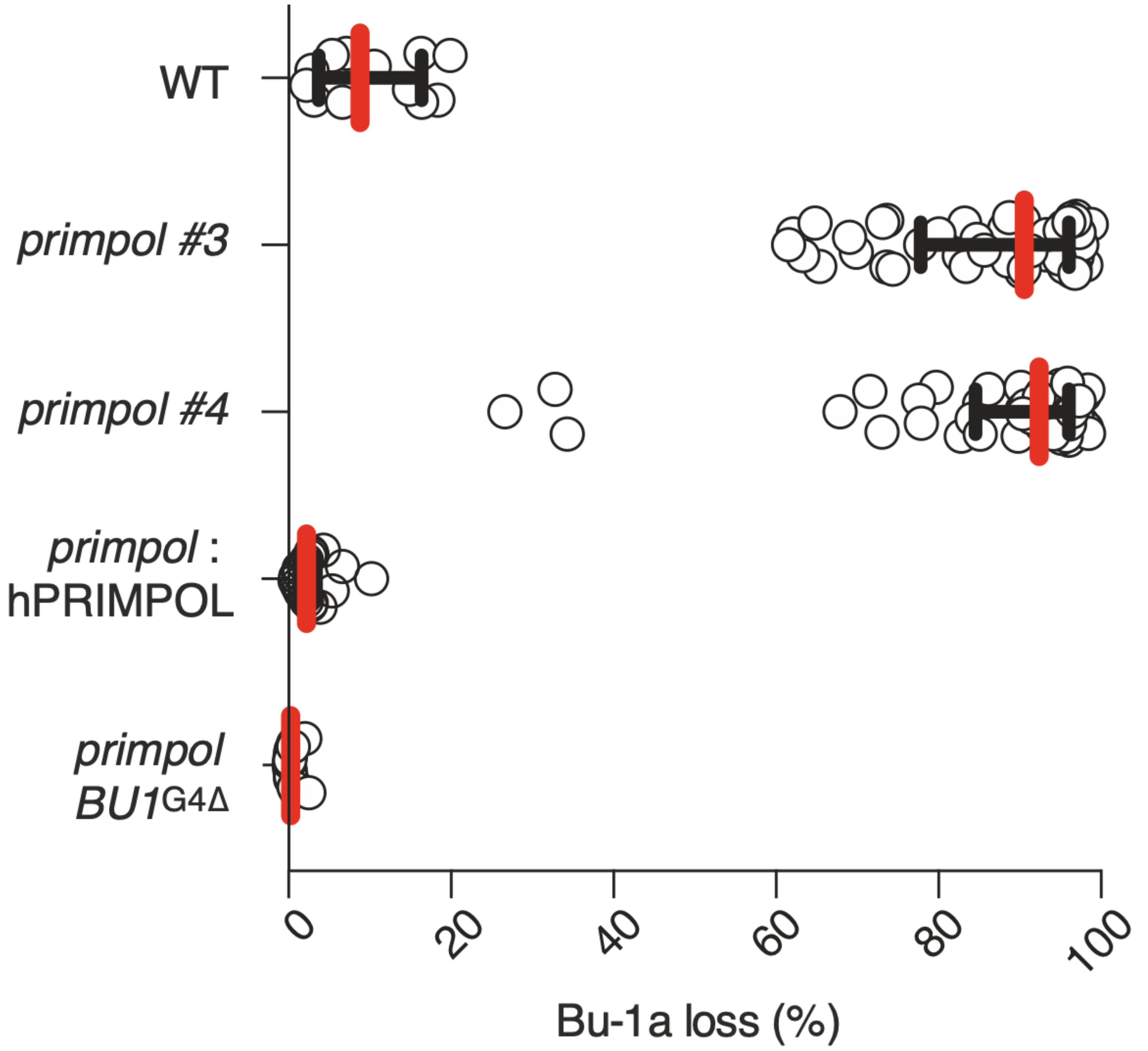
# Structure formation is likely a frequent event during replication





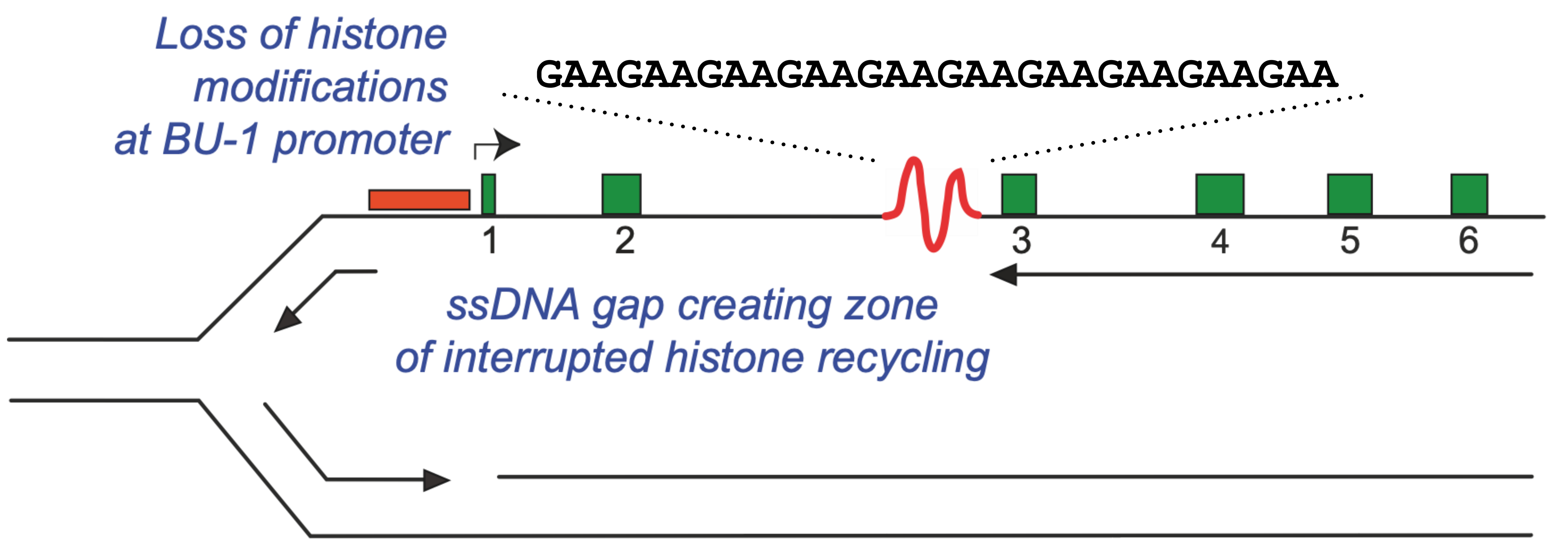
- **PrimPol** is the second identified primase in vertebrates
- RNA / DNA primase
- DNA polymerase with some capacity for lesion bypass

- Unable to replicate G4s
- Binds to G4s and efficiently reprimes close by the structure



Schiavone et al. Mol Cell (2016) 61, 161-169
(with Aidan Doherty & Stan Jozwiakowski, University of Sussex)

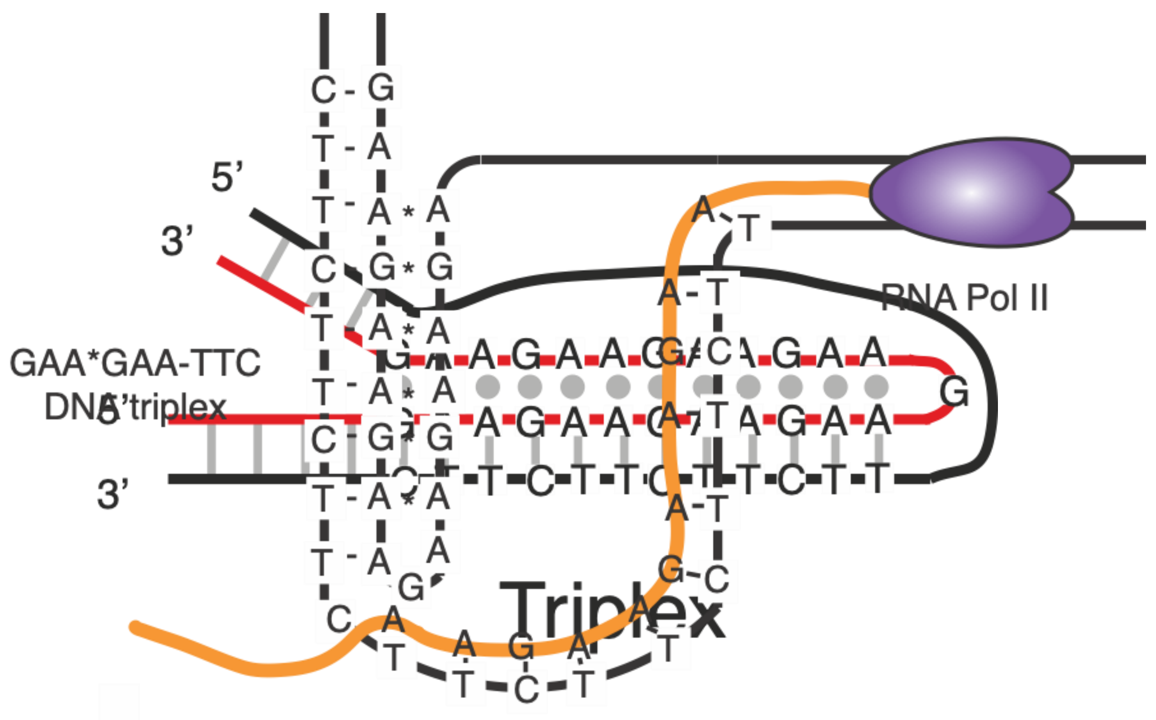
# PrimPol loss reveals that even short repeats can be replication impediments



*Loss of histone modifications at BU-1 promoter*

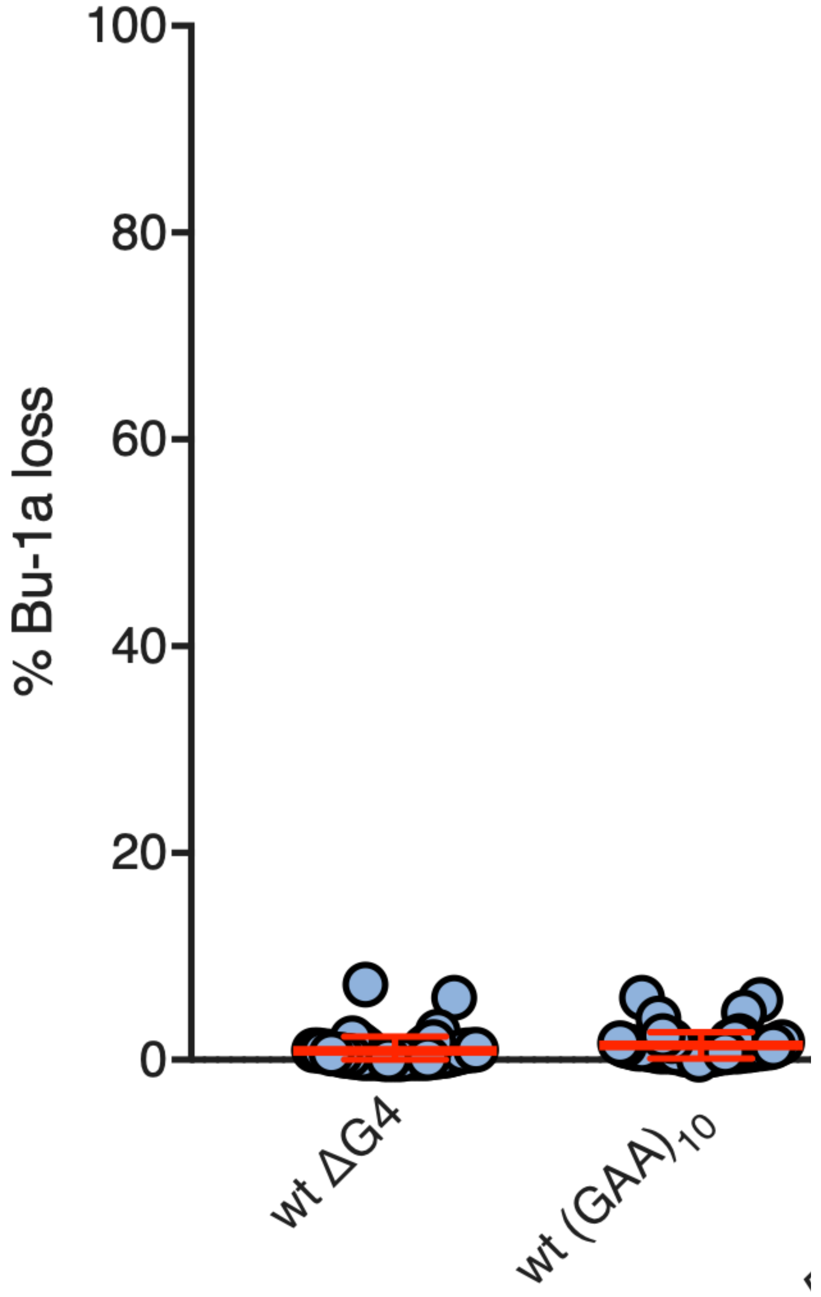
*ssDNA gap creating zone of interrupted histone recycling*

*Fork enters locus from 3' end and pauses at +3.5 G4*



The 'transcriptional diode' Grabczyk & Fishman JBC 1995

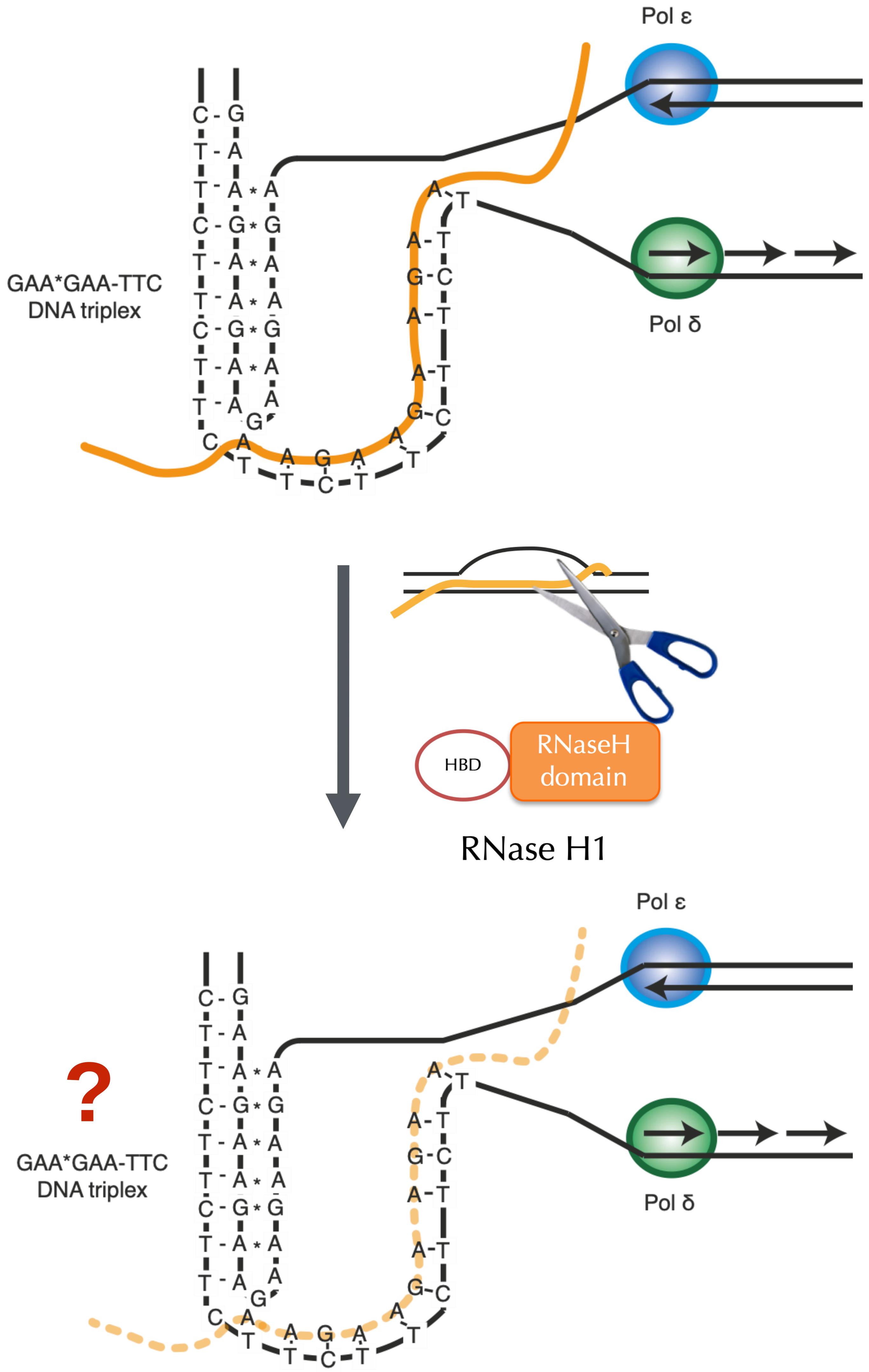
(GAA)n repeats will form triplexes when transcribed as the coding strand

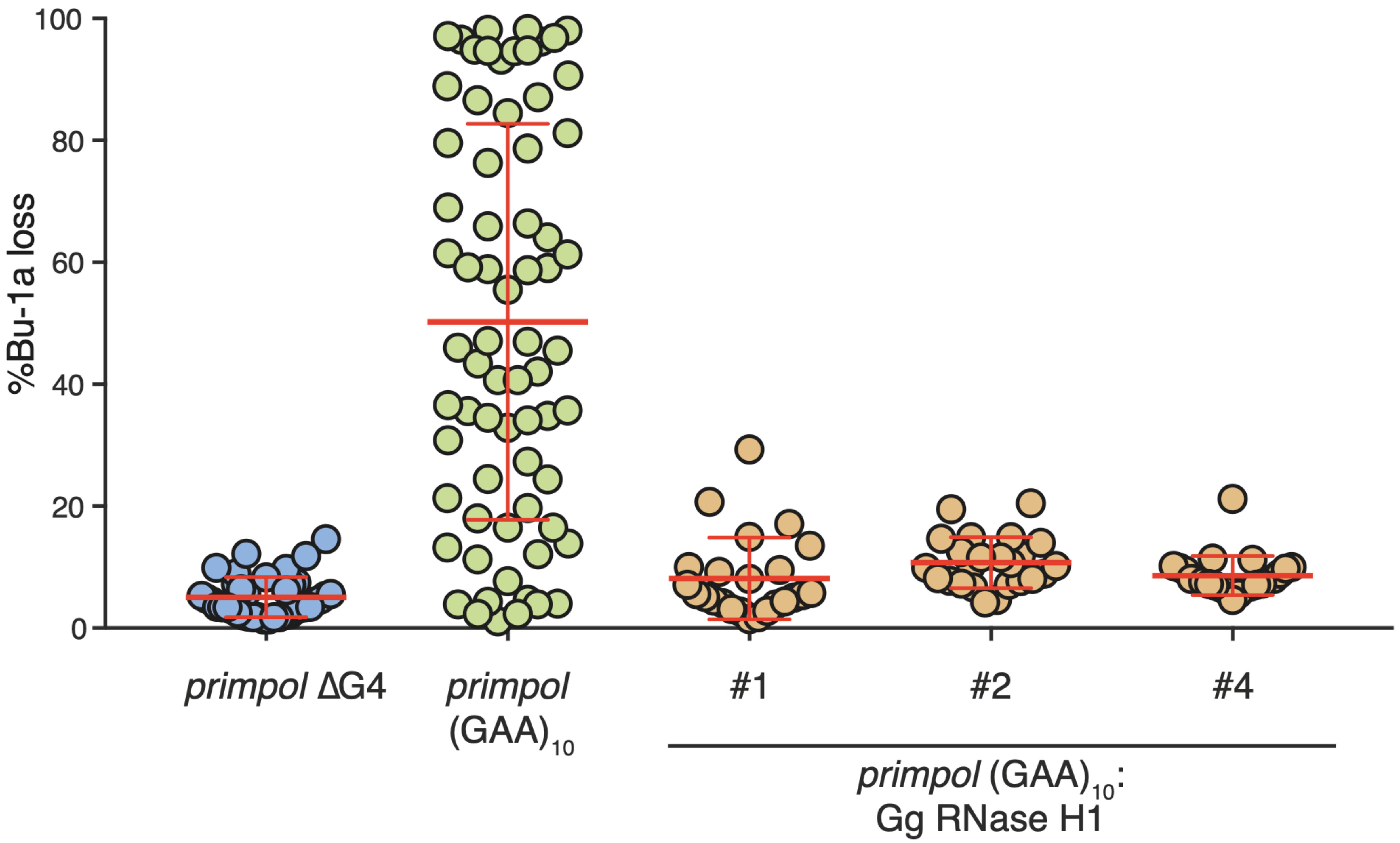


- **(GAA)$_{10}$ on the leading strand template is able to block DNA synthesis**

Šviković et al. (2019) *EMBO J* 38(3):e99793

# Is R-loop formation necessary for $(GAA)_{10}$ to trigger instability of BU-1 expression

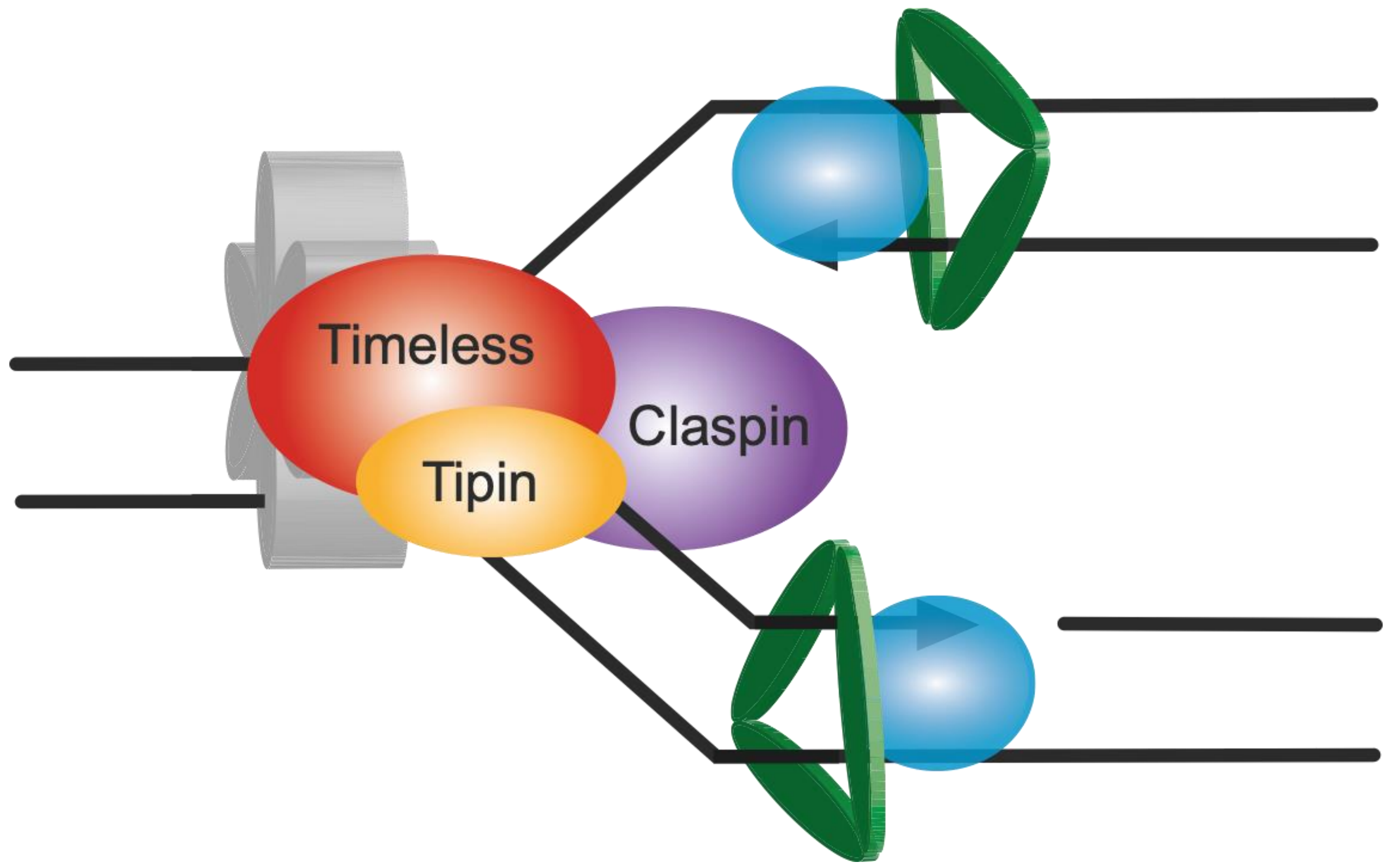


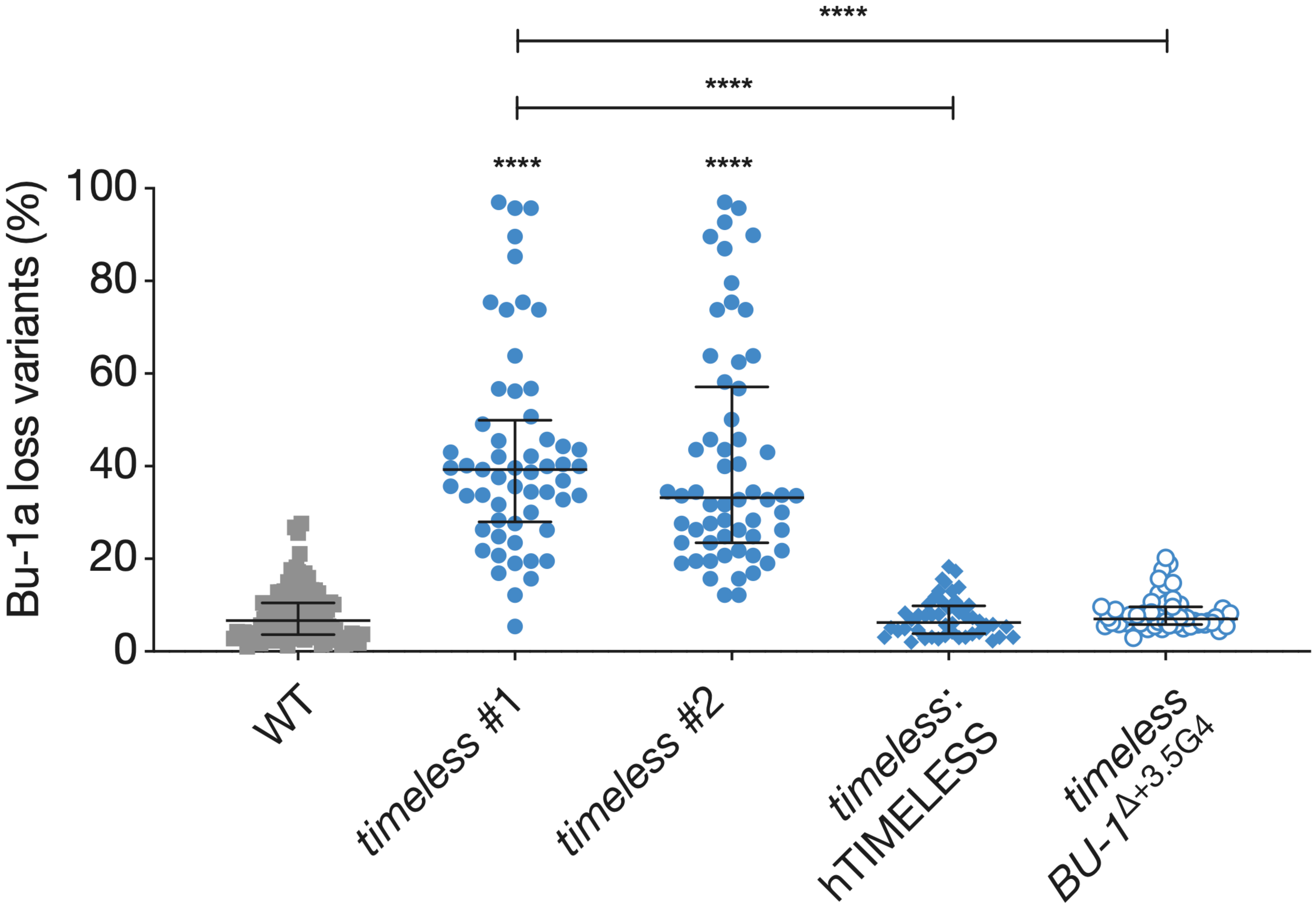


- **RNaseH1 removes the need for PrimPol-mediated repriming suggesting that (GAA)10 requires RNA:DNA hybrid formation for it to become a replication impediment**

Šviković et al. (2019) *EMBO J* 38, pii: 399793

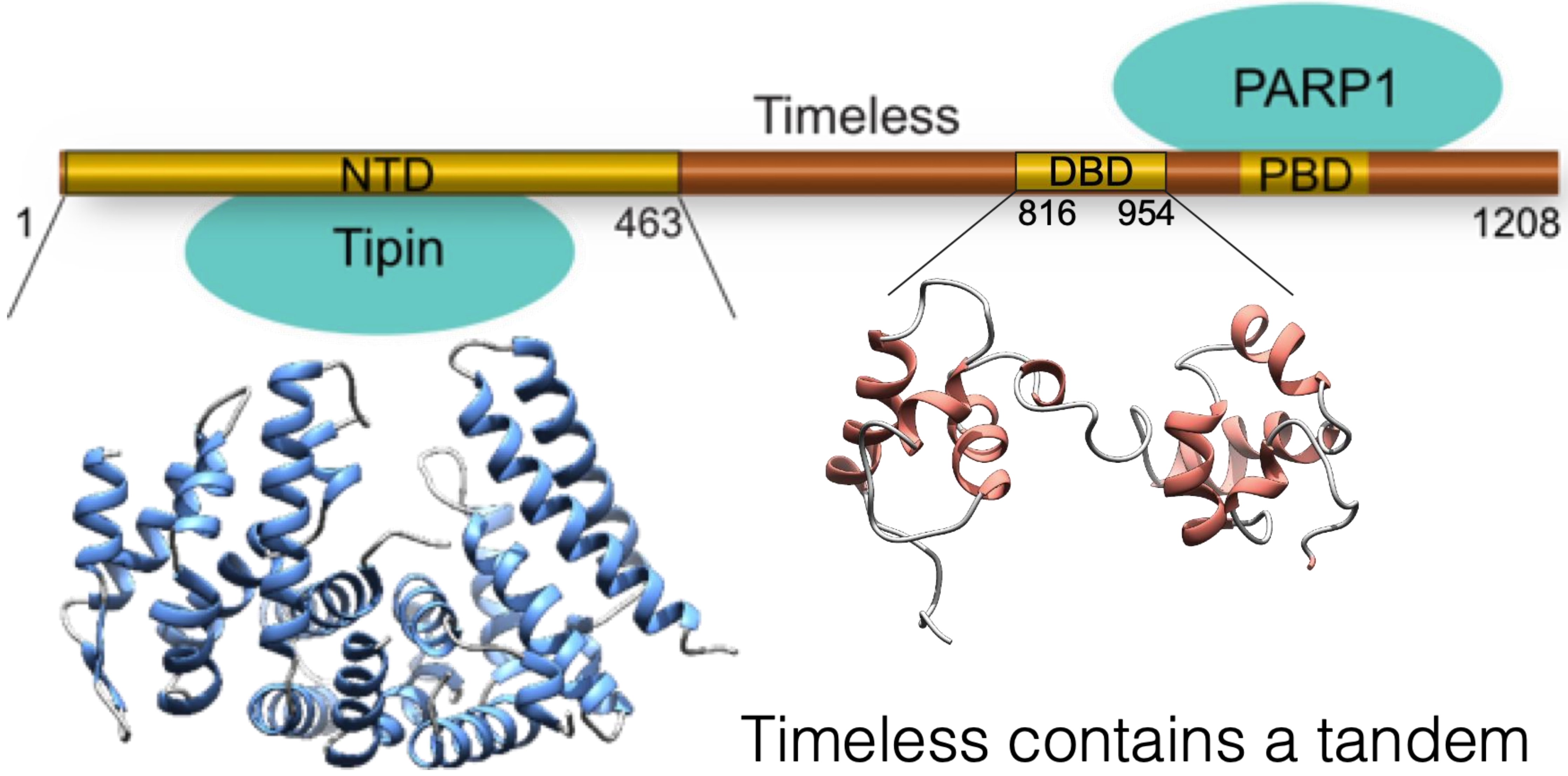
# Does the replisome have specific mechanisms for surveillance of structure formation: the fork protection complex



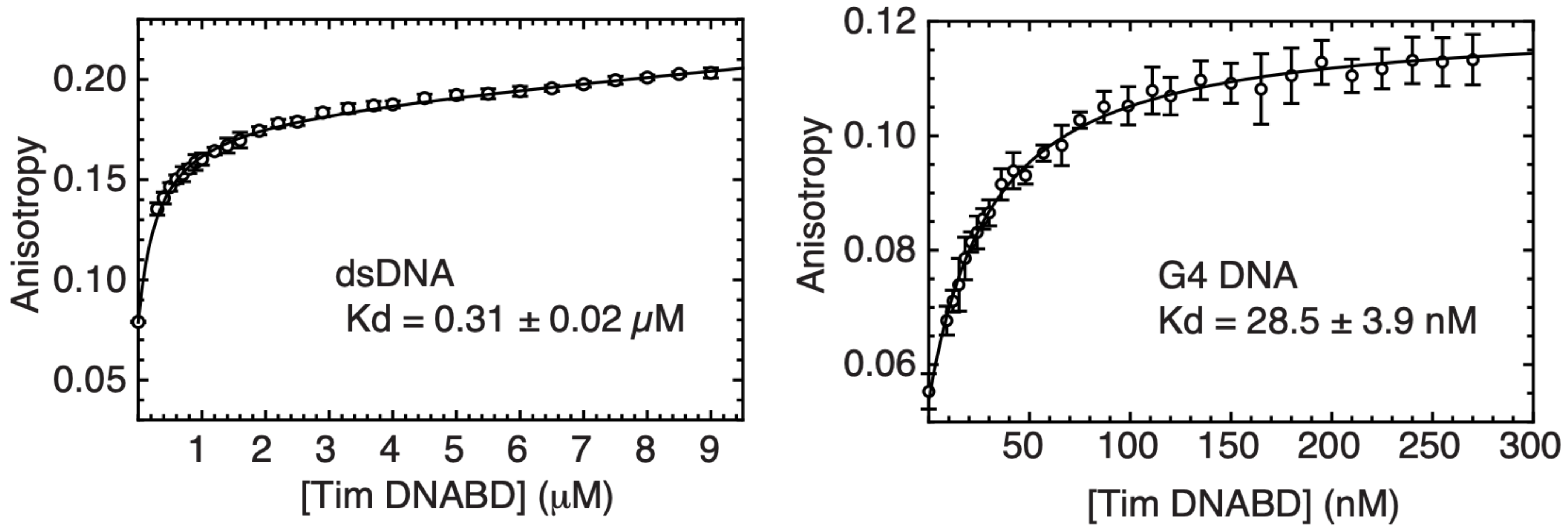


Koch Lerner, Holzer et al (2020) *EMBO J* 39(18):e104185

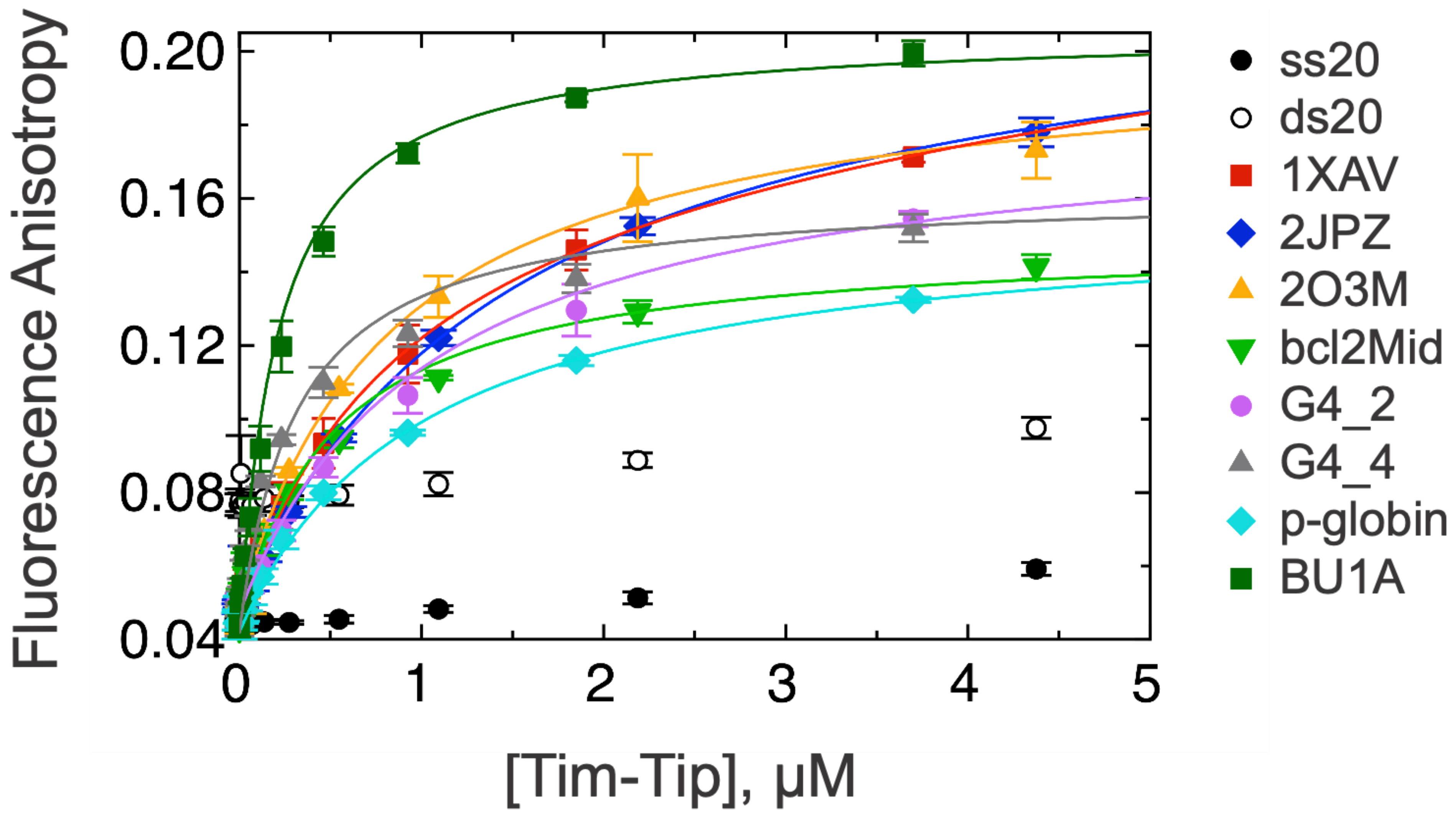
# A newly identified DNA binding domain in Timeless



Timeless contains a tandem repeat of helical domains (TimDBD)



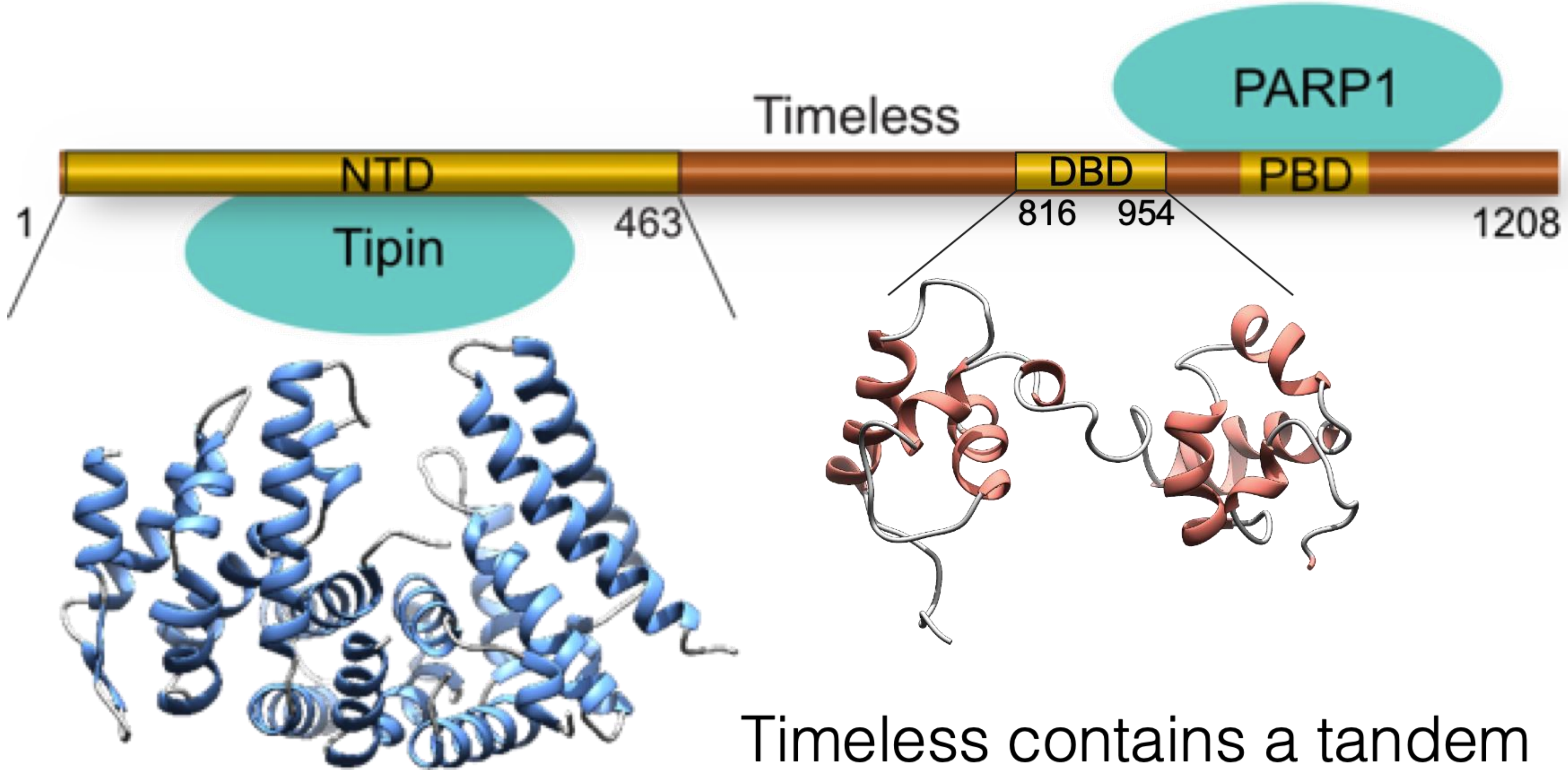
G4 DNA : **6FAM-TGAGGGTGGGTAGGGTGGGTAA**



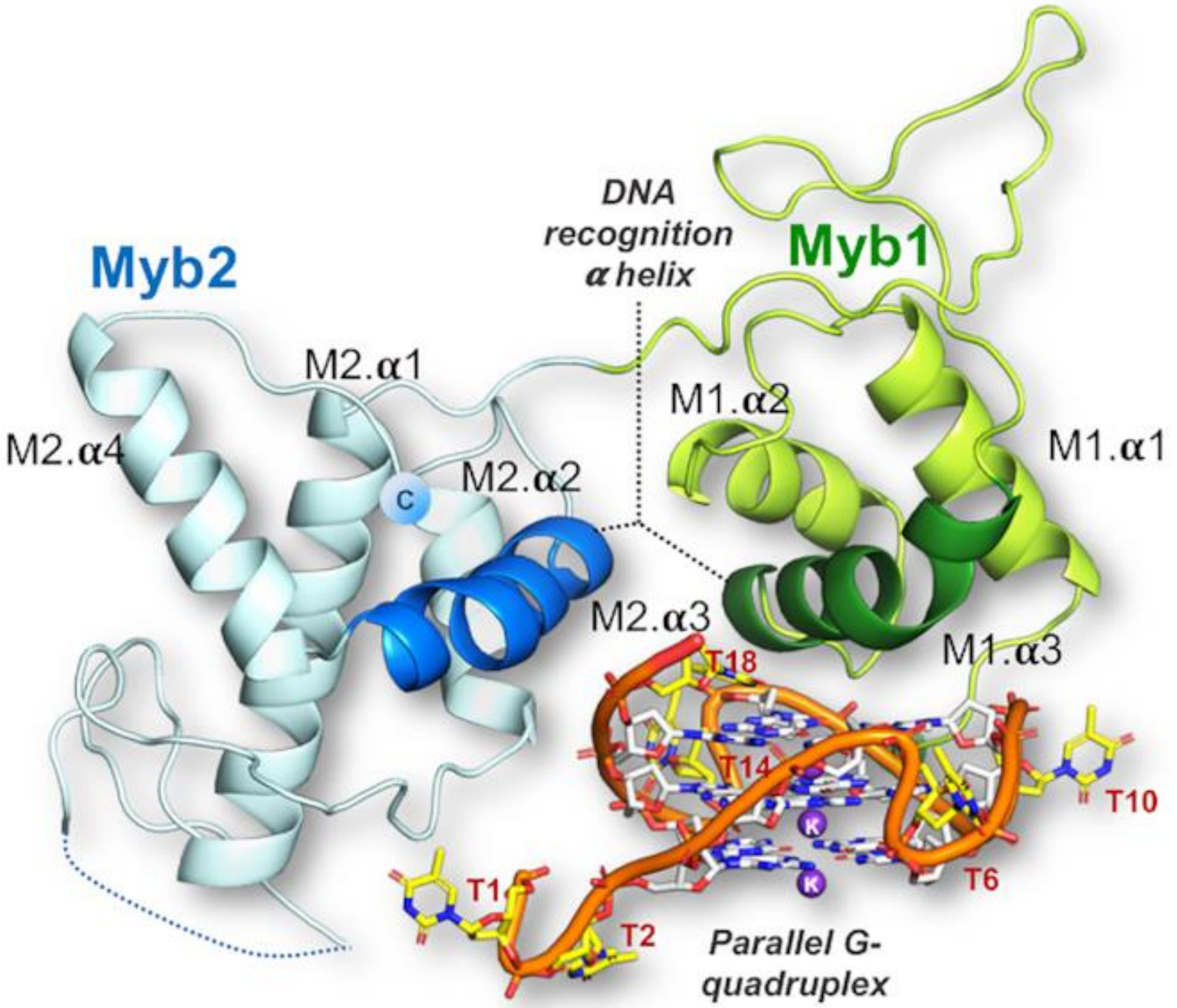
Luca Pellegrini lab

Koch Lerner, Holzer et al (2020) *EMBO J* 39(18):e104185

# A newly identified DNA binding domain in Timeless



Timeless contains a tandem repeat of helical domains (TimDBD)



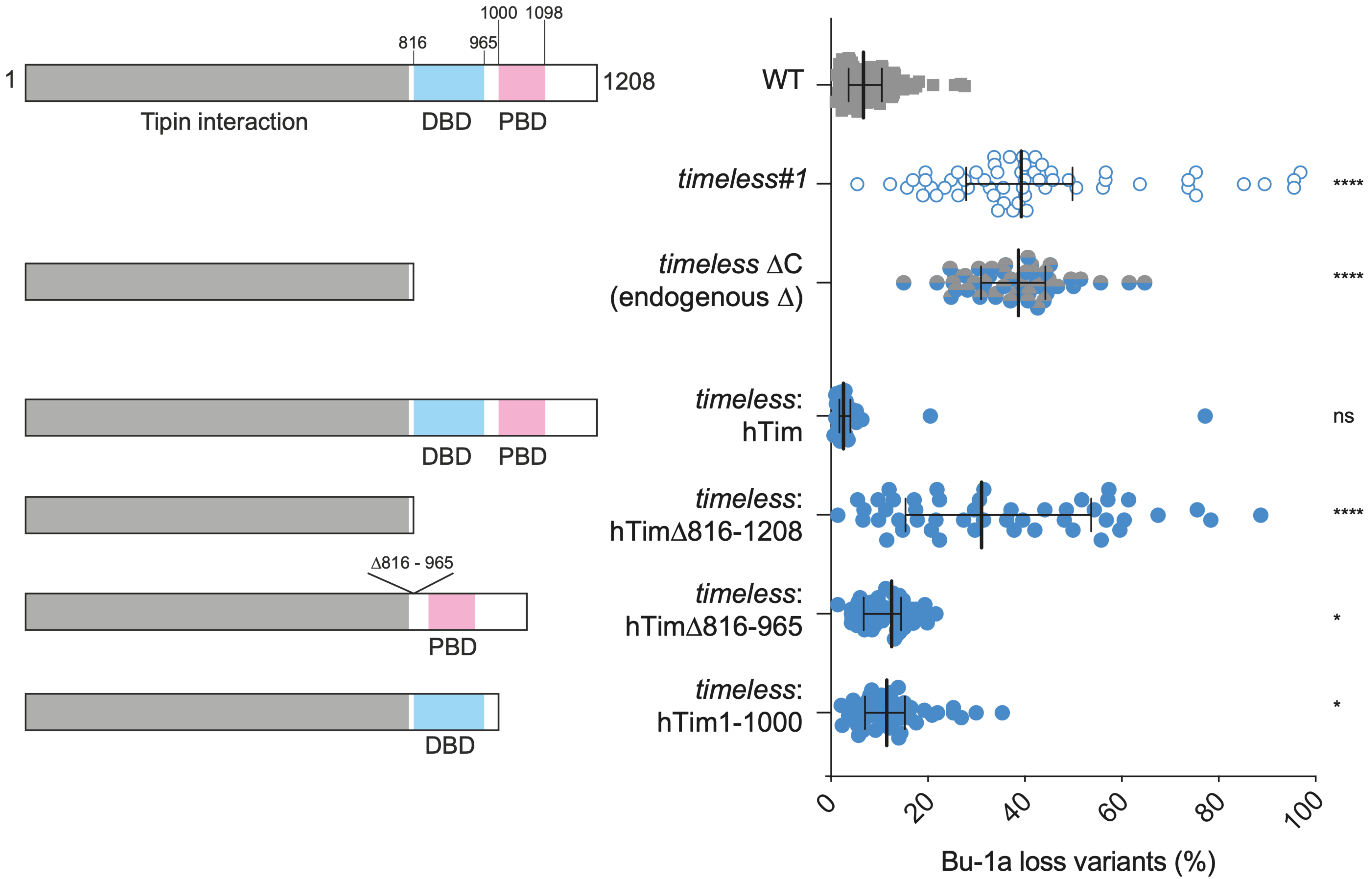
G4 recognition by the tandem Myb domains of RAP1

Traczyk … Rhodes *NAR 2020*

Luca Pellegrini lab

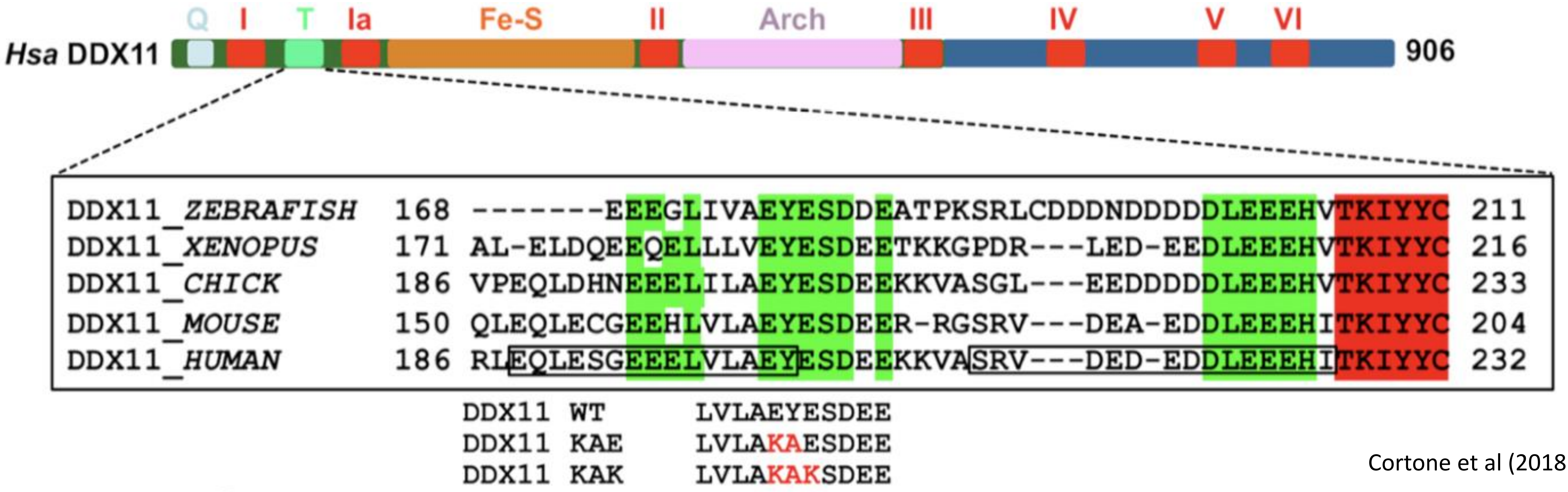
Koch Lerner, Holzer et al (2020) *EMBO J* 39(18):e104185

# The C-terminus of Timeless is required to prevent G4-induced instability of *BU-1*



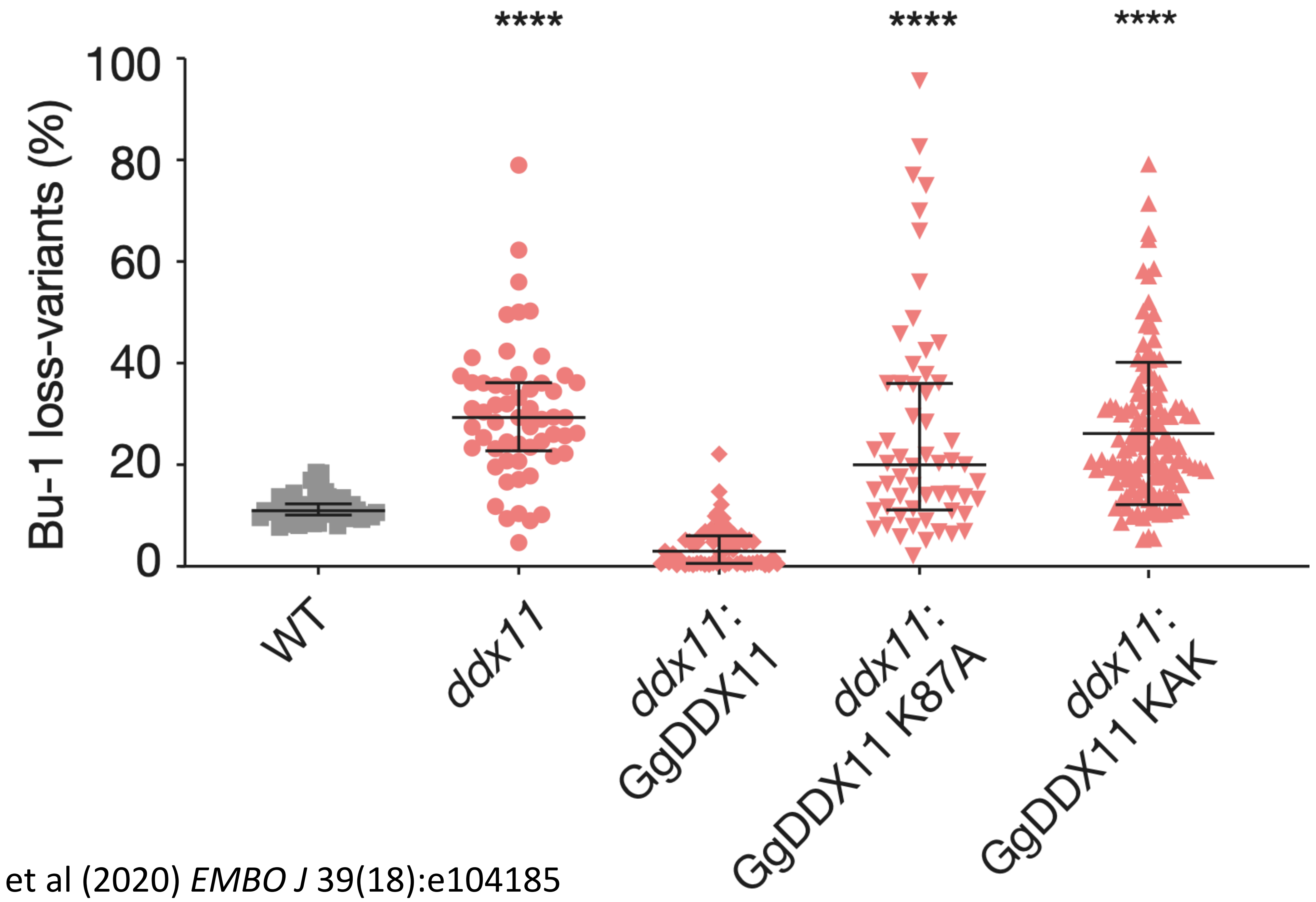
Koch Lerner, Holzer et al (2020) *EMBO J* 39(18):e104185

# The interaction of Timeless with the helicase DDX11 is required for processing fork-stalling G4s



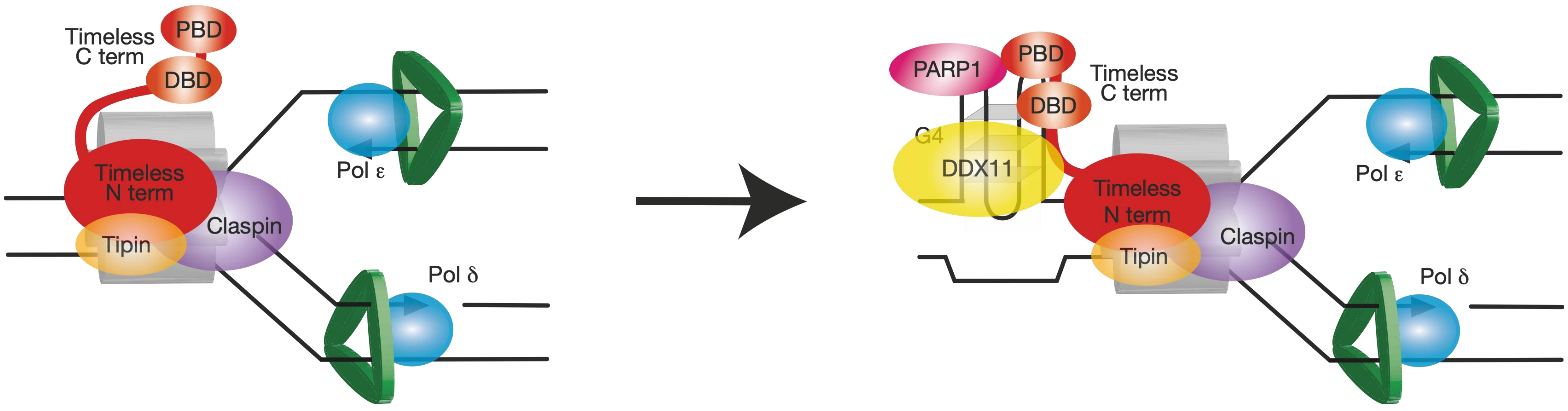
```
DDX11_ZEBRAFISH   168 -------EEEGLIVAEYESDDEATPKSRLCDDDNDDDDLEEEHVTKIYYC 211
DDX11_XENOPUS     171 AL-ELDQEEQELLLVEYESDEETKKGPDR---LED-EEDLEEEHVTKIYYC 216
DDX11_CHICK       186 VPEQLDHNEEELILAEYESDEEKKVASGL---EEDDDDDLEEEHVTKIYYC 233
DDX11_MOUSE       150 QLEQLECGEEHLVLAEYESDEER-RGSRV---DEA-EDDLEEEHITKIYYC 204
DDX11_HUMAN       186 RLEQLESGEEELVLAEYESDEEKKVASRV---DED-EDDLEEEHITKIYYC 232

              DDX11 WT     LVLAEYESDEE
              DDX11 KAE    LVLAKAESDEE
              DDX11 KAK    LVLAKAKSDEE
```

Cortone et al (2018) *PLOS Genetics* 14(10):e1007622



Koch Lerner, Holzer et al (2020) *EMBO J* 39(18):e104185

# Timeless detects G4s at the fork and coordinates their resolution by DDX11

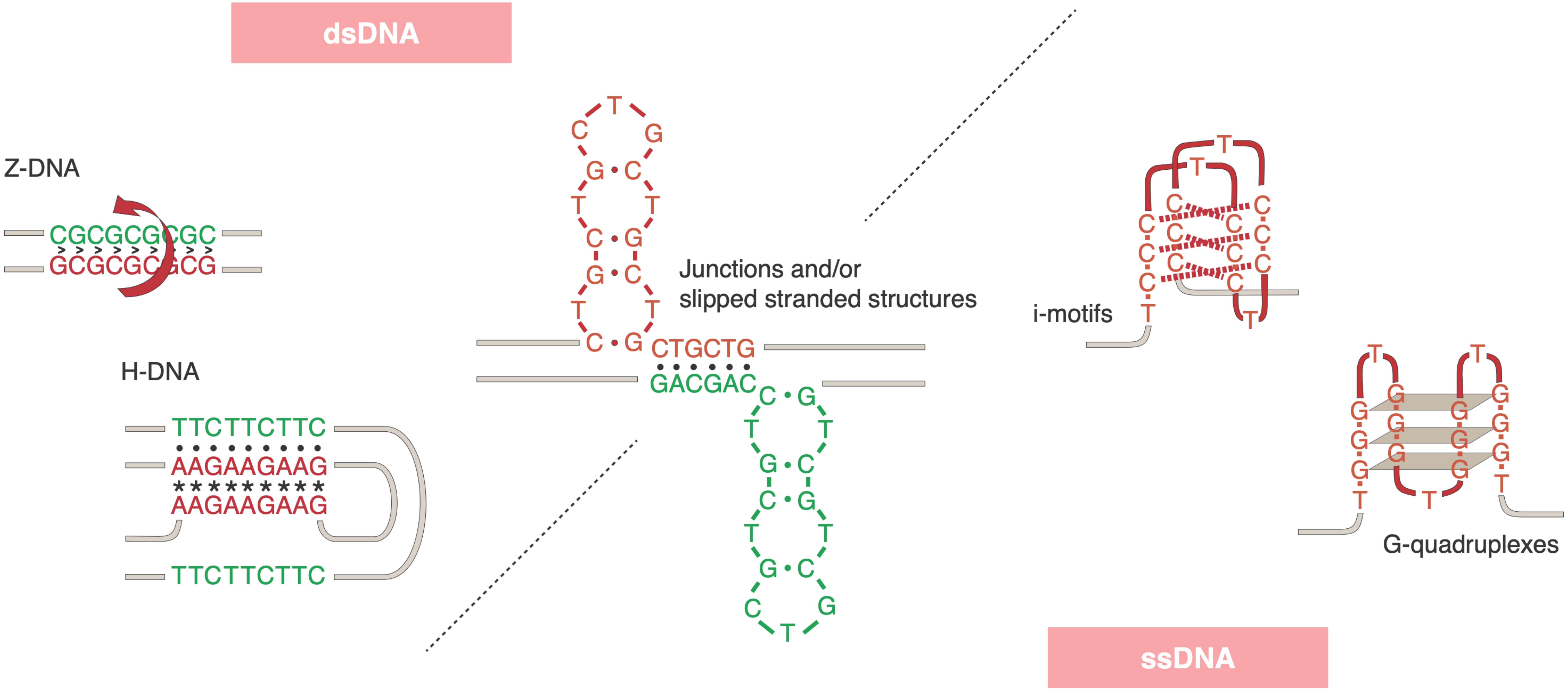


Unperturbed

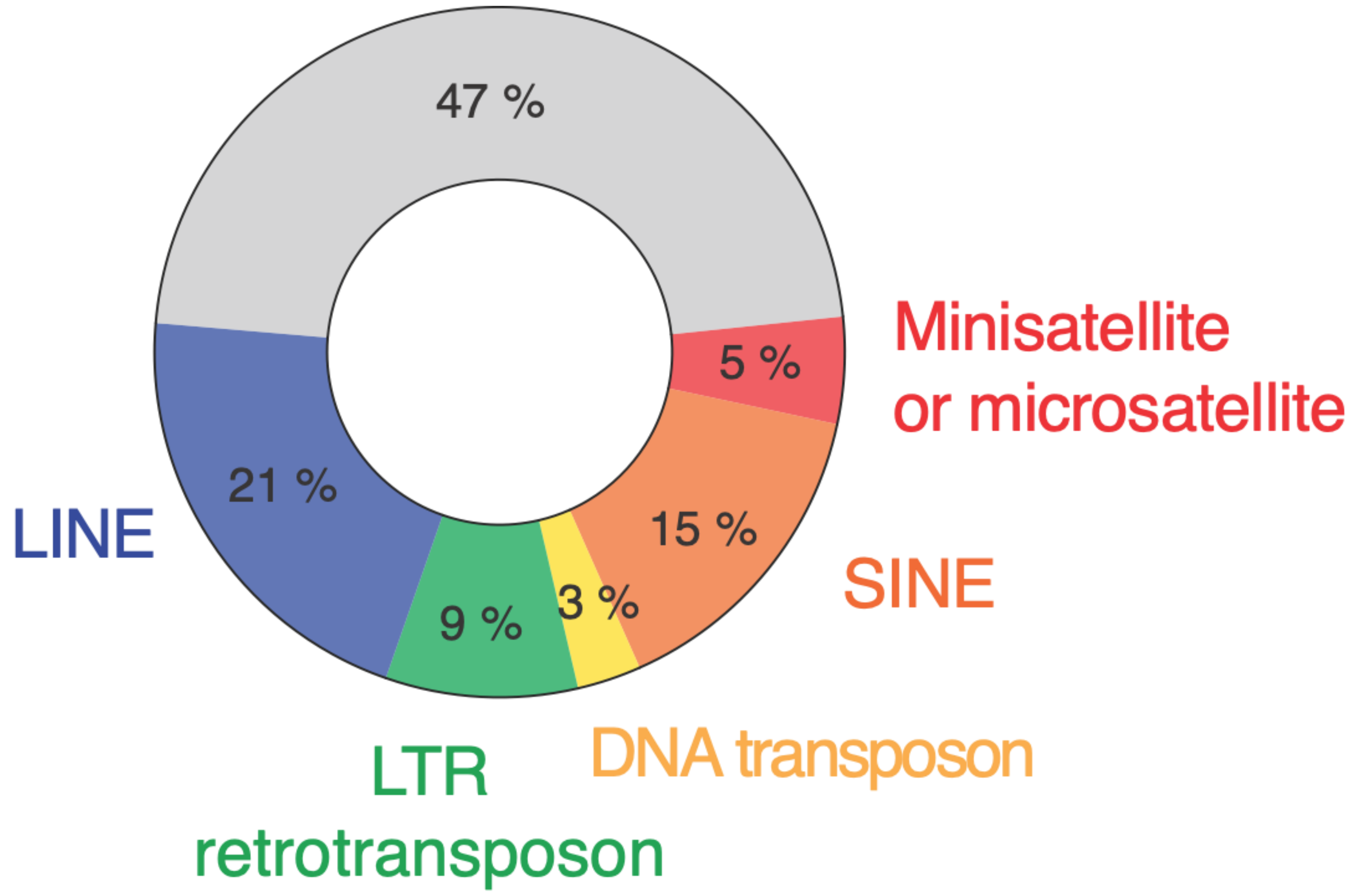
G4-detected ahead of fork

Koch Lerner, Holzer et al (2020) *EMBO J* 39(18):e104185

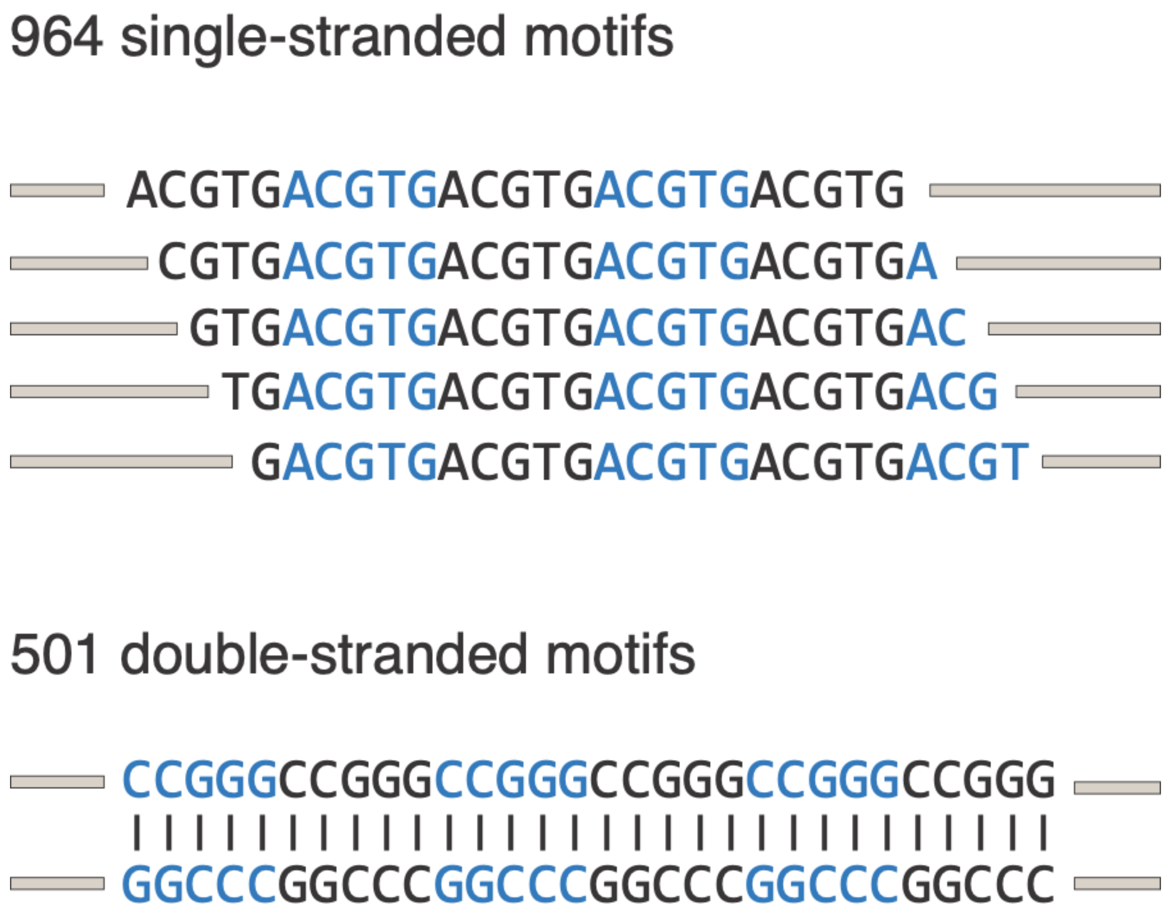
# Which sequences are intrincially capable of forming secondary structures that impede DNA synthesis?



# DNA repeats contribute to gene function, genome structure and evolution, but repeat distribution is highly dependent on sequence



STRs : Tracts of repetitive 2-6 bp motifs
4,500,000 loci covering 2.5% of the human genome

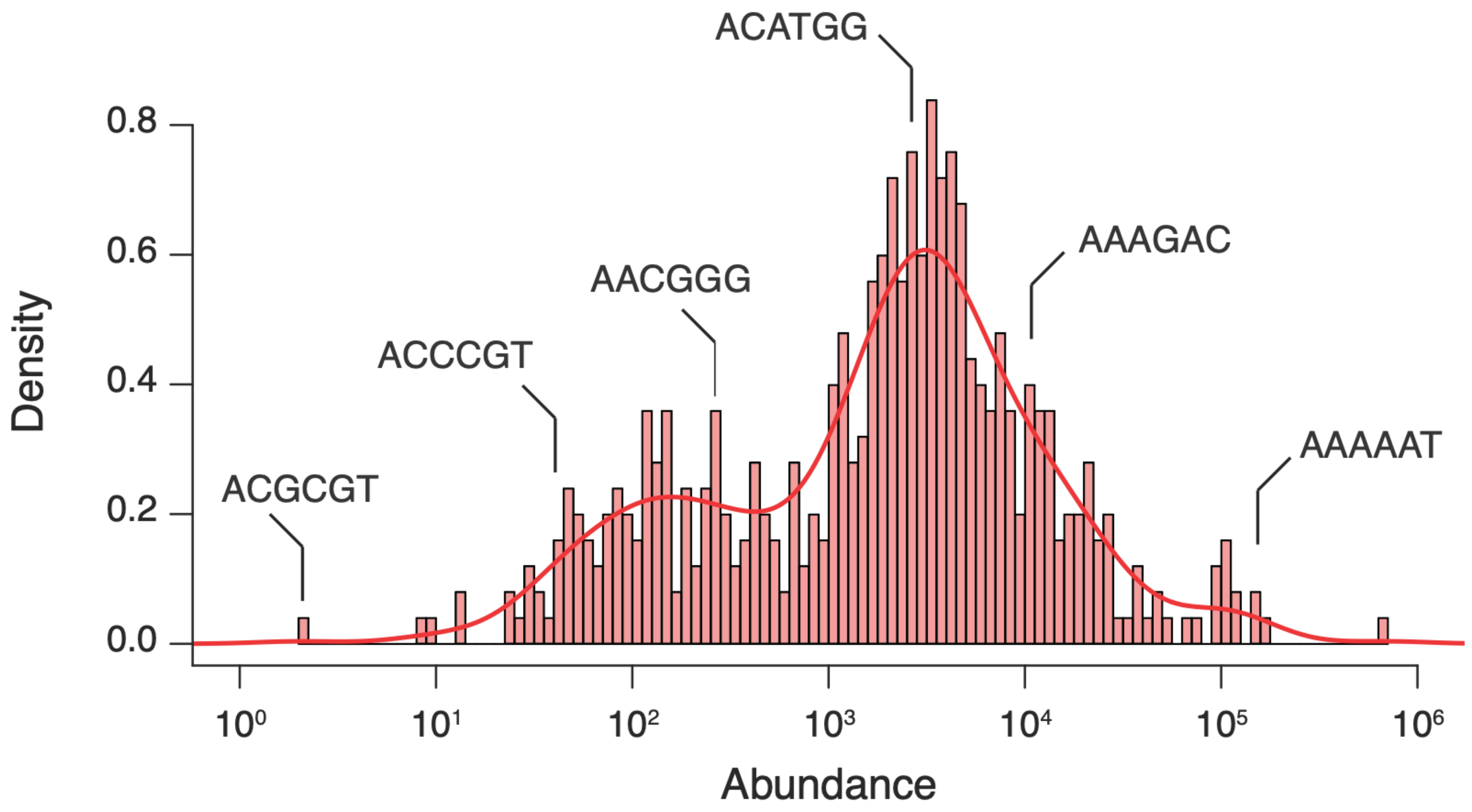


964 single-stranded motifs

ACGTGACGTGACGTGACGTGACGTG
CGTGACGTGACGTGACGTGACGTGA
GTGACGTGACGTGACGTGACGTGAC
TGACGTGACGTGACGTGACGTGACG
GACGTGACGTGACGTGACGTGACGT

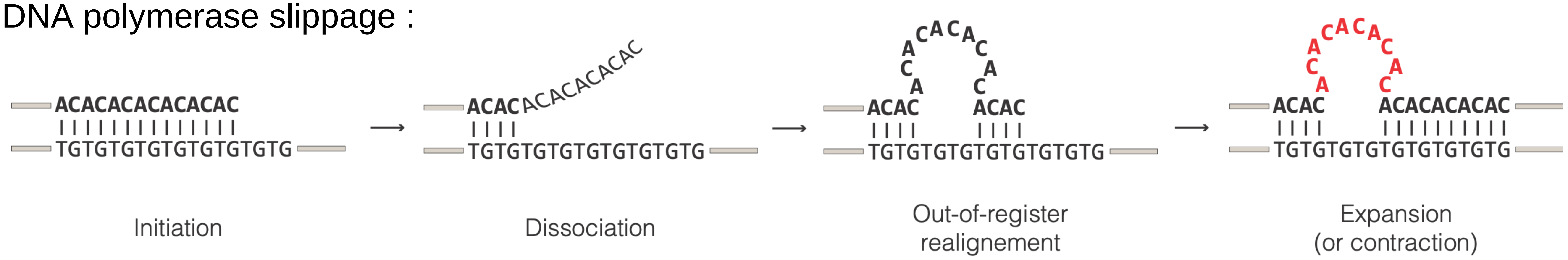
501 double-stranded motifs

CCGGGCCGGGCCGGGCCGGGCCGGGCCGGG
GGCCCGGCCCGGCCCGGCCCGGCCCGGCCC



# Repeat length is also highly polymorphic and related to repeat sequence

DNA polymerase slippage :



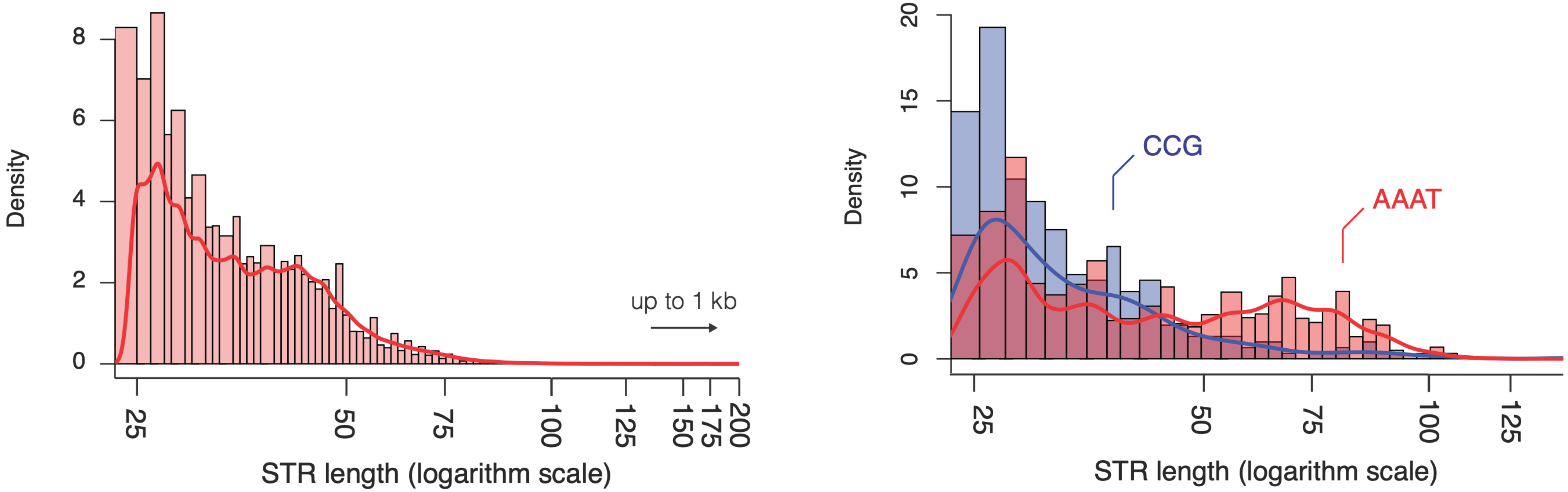
Initiation

Dissociation

Out-of-register realignement

Expansion (or contraction)

But wide distribution of STR length at equilibrium :

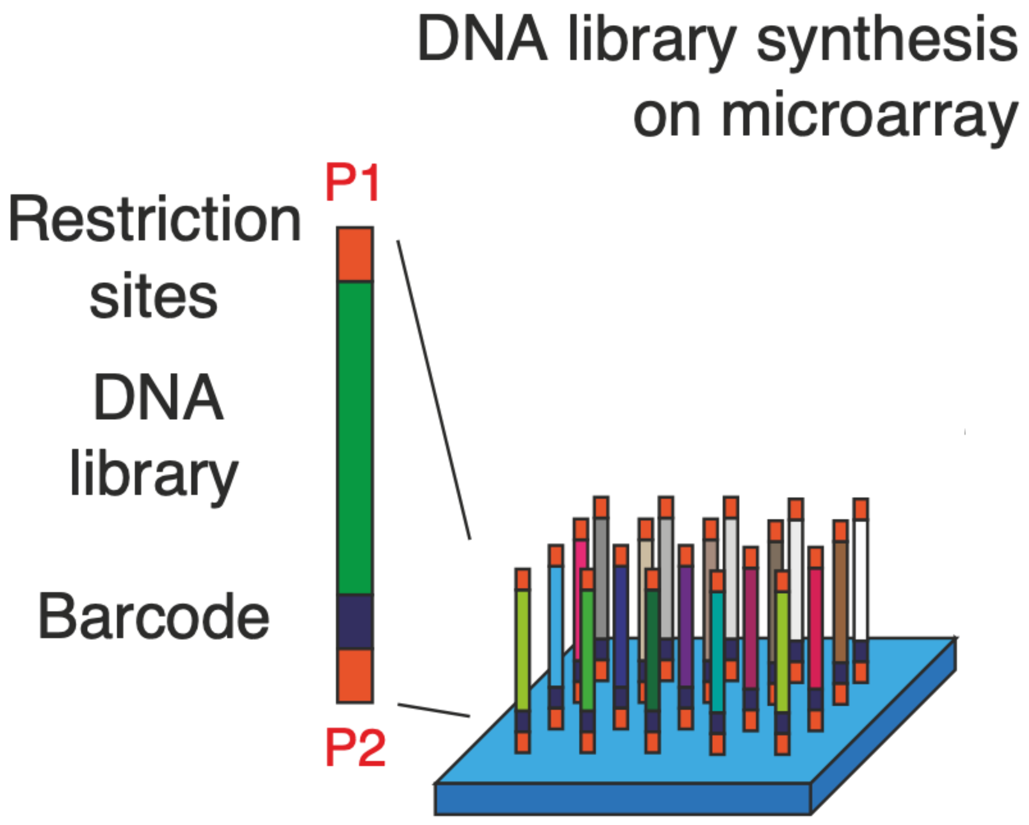


What determines the representation of STRs in the human genome?

Is STR abundance and length determined by DNA polymerase behaviour?

Murat, P, Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

# Which sequences are intrinsically able to stall DNA synthesis?

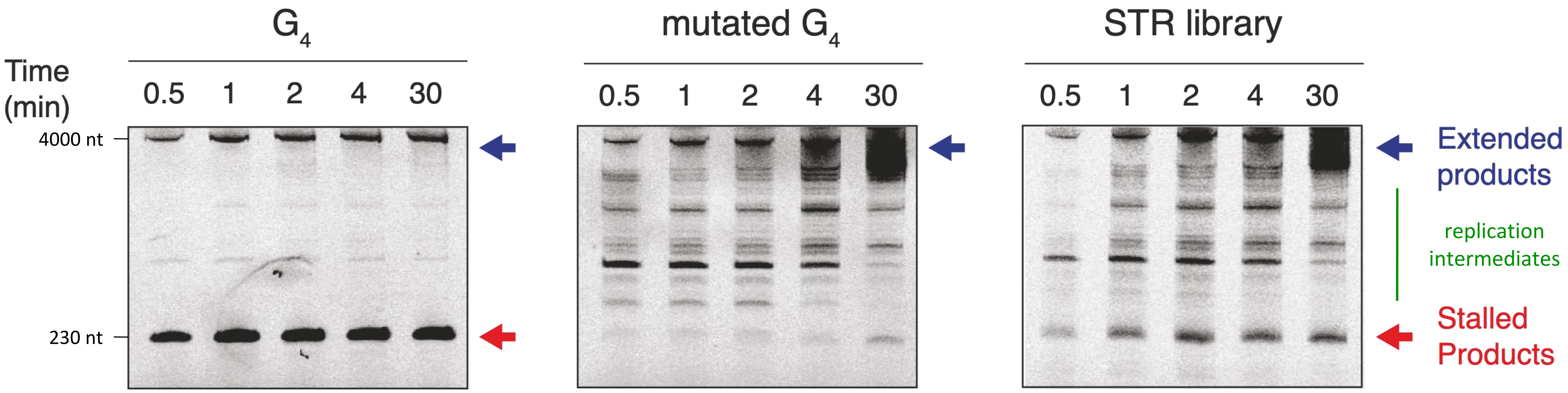


*DNA library composition :*

| | |
|---|---|
| All STR permutations (2-6 bp sequences) : | 5,356 sequences |
| in three different lengths (24, 48, 72 nt) | 16,068 sequences |
| Positive controls<br>Hairpin, G-quadruplex and I-motif forming sequences | 2,932 sequences |
| Negative controls<br>Random sequences of varying GC content | 1,000 sequences |
| Total : | 20,000 sequences |

Murat, P, Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

# Which sequences are intrinsically able to stall DNA synthesis?



Use of high-throughput sequencing to quantify:

- DNA synthesis efficiency / stalling

$$\textit{Stall score}_{(t)} = \sigma_{(t)} = \frac{\#\ \textit{Reads in the stalled fraction}}{\#\ \textit{Reads in the stalled and extended fraction}}$$

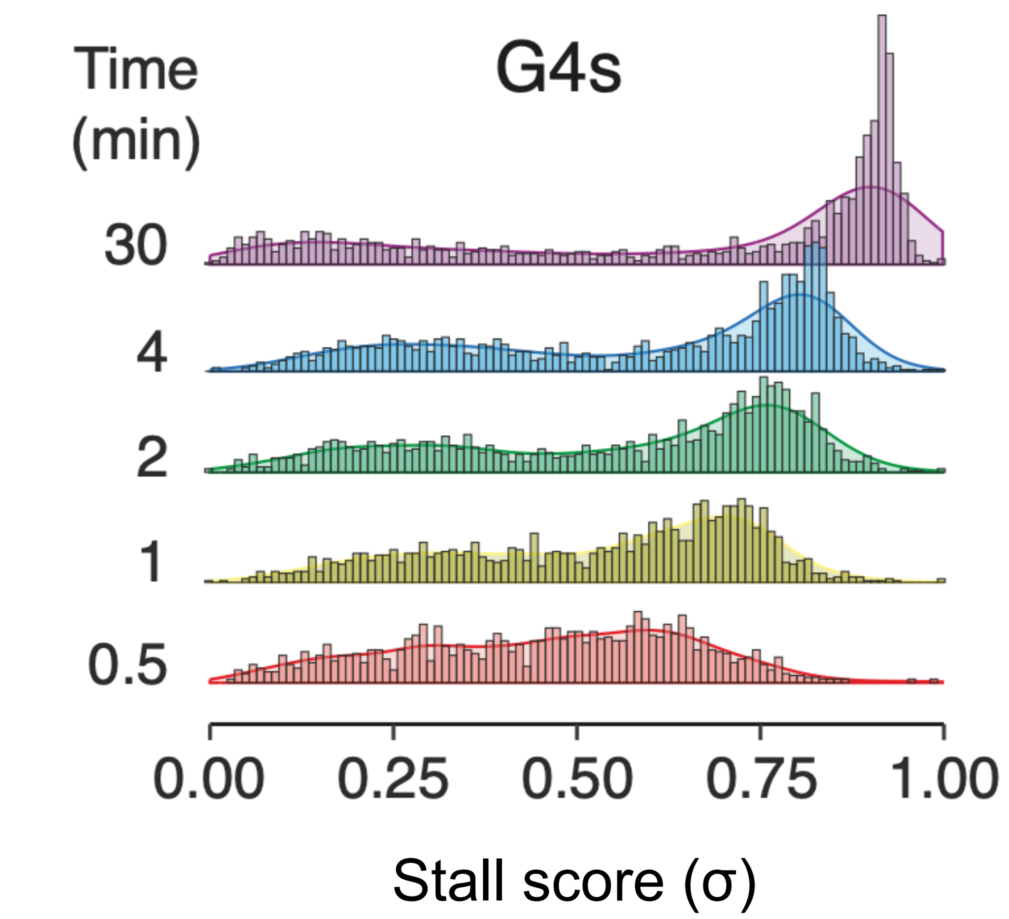
- DNA synthesis fidelity

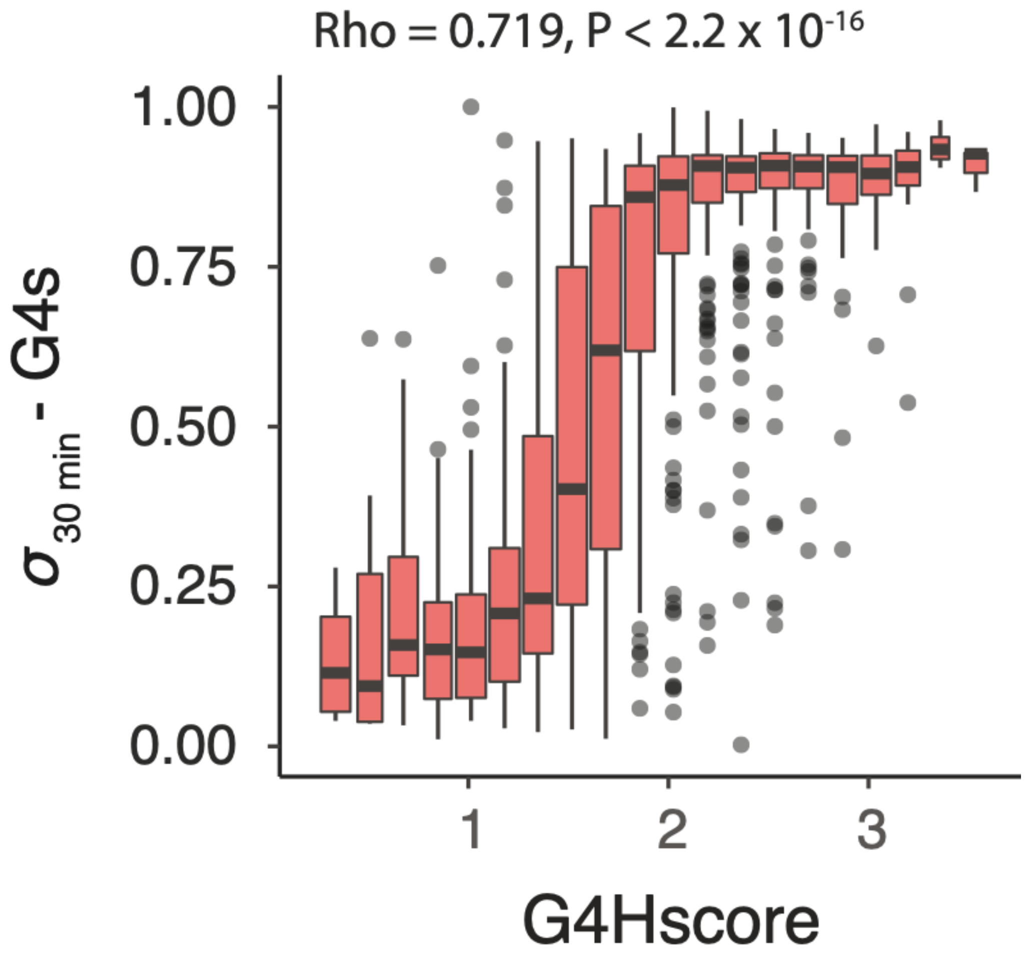
$$\textit{Errors}_{(t)} = \#\ \textit{Mutations in extended fraction}$$

Murat, P,  Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

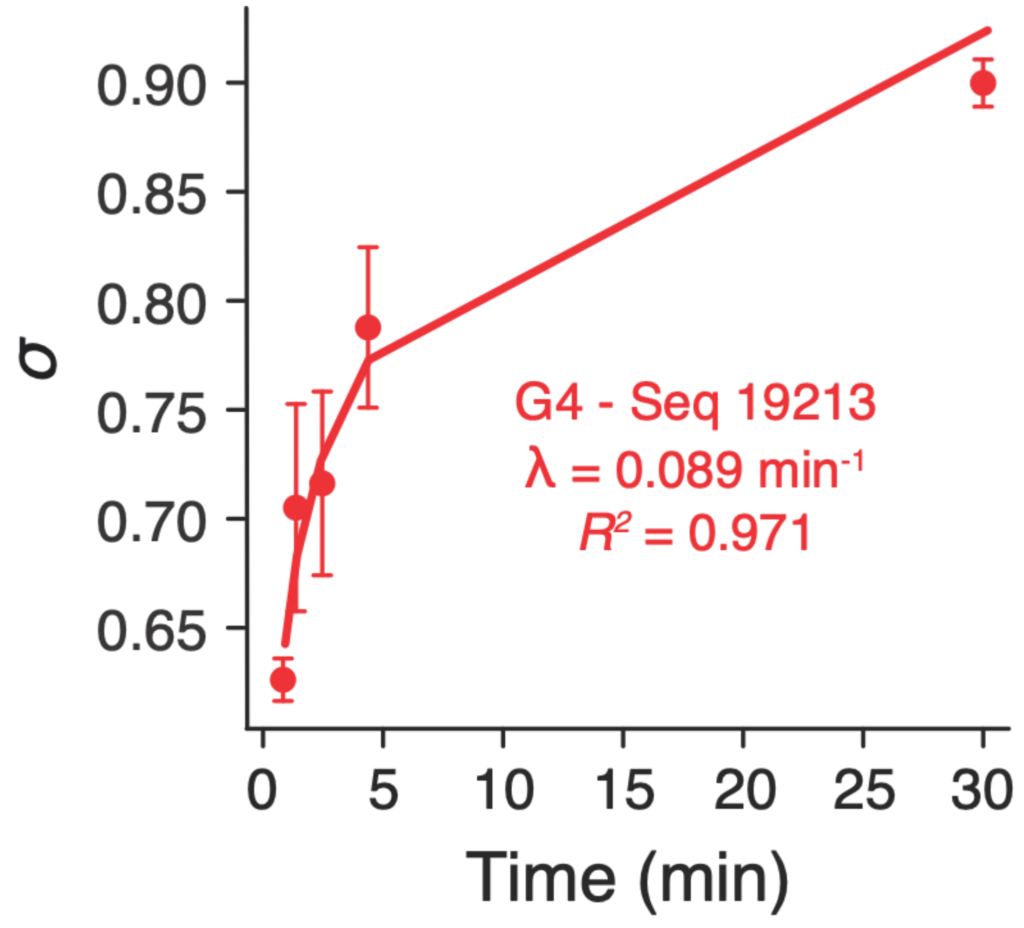
# Structure-dependent transient and persistent stalling events

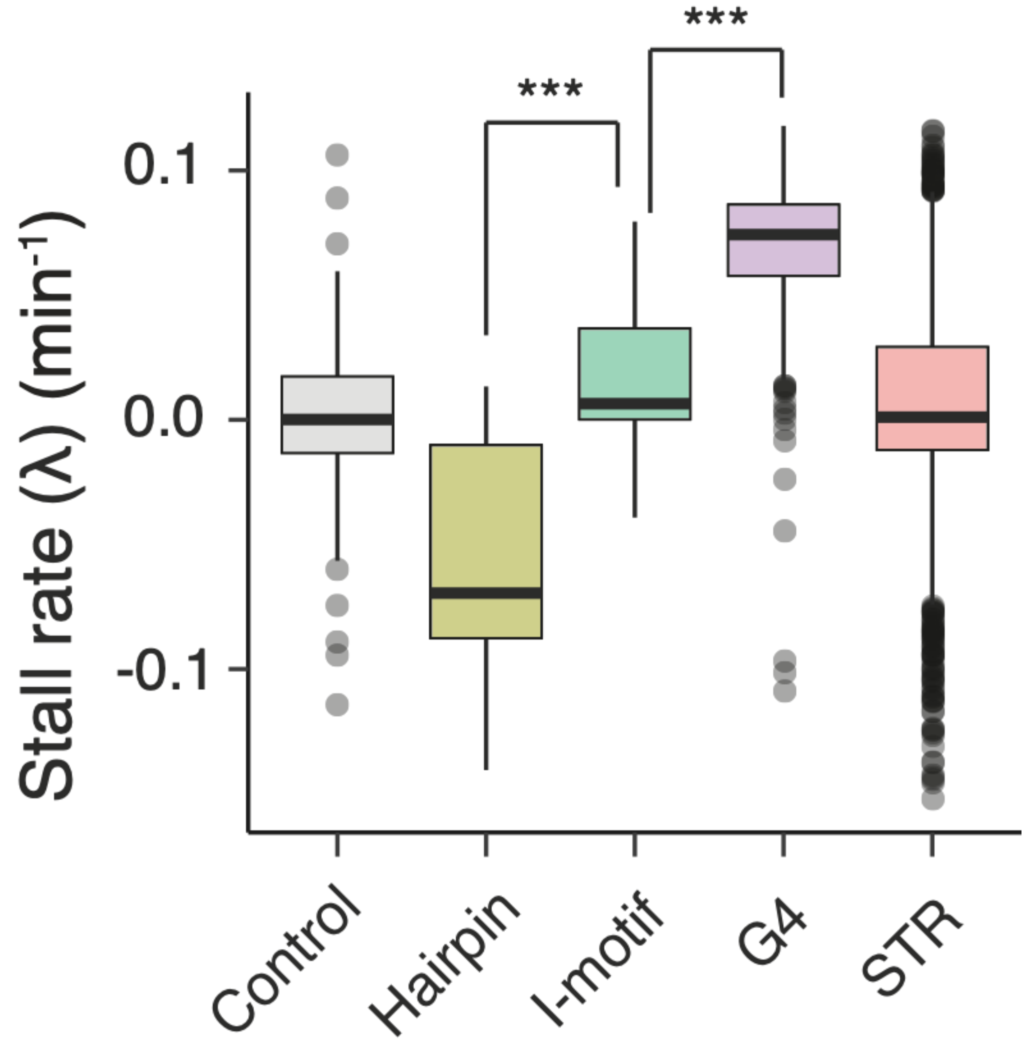
1,500 sequences



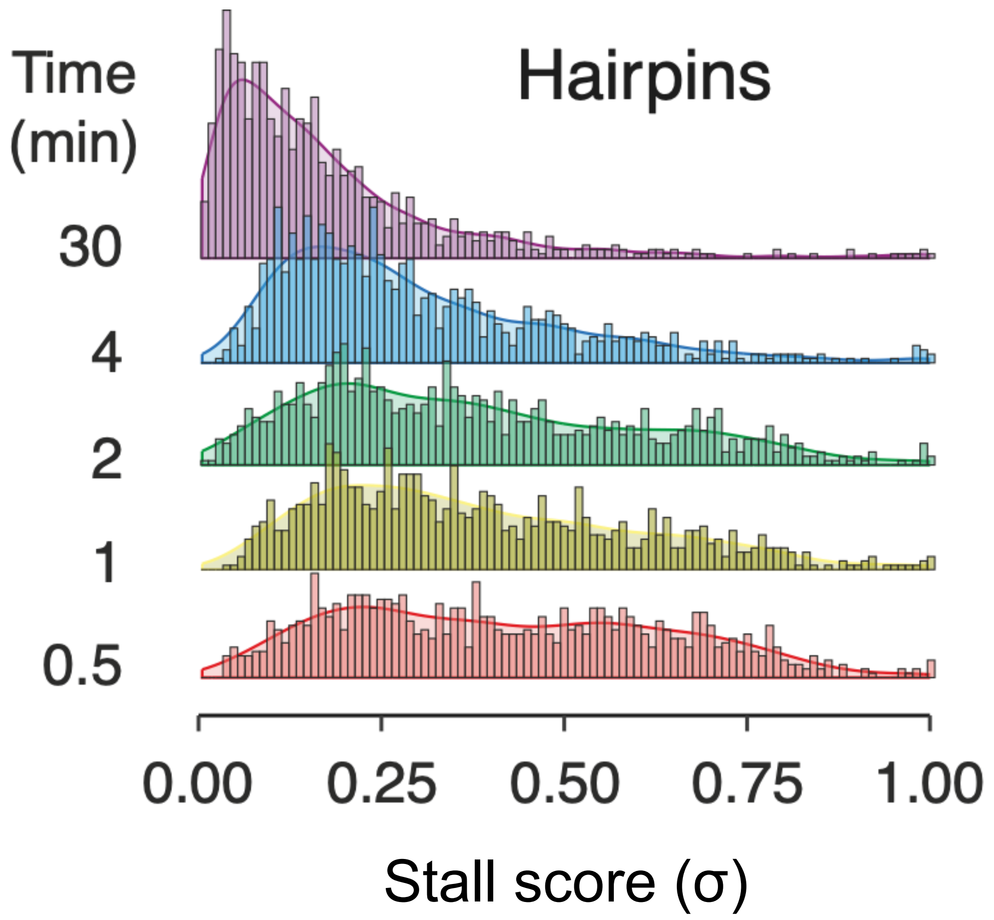


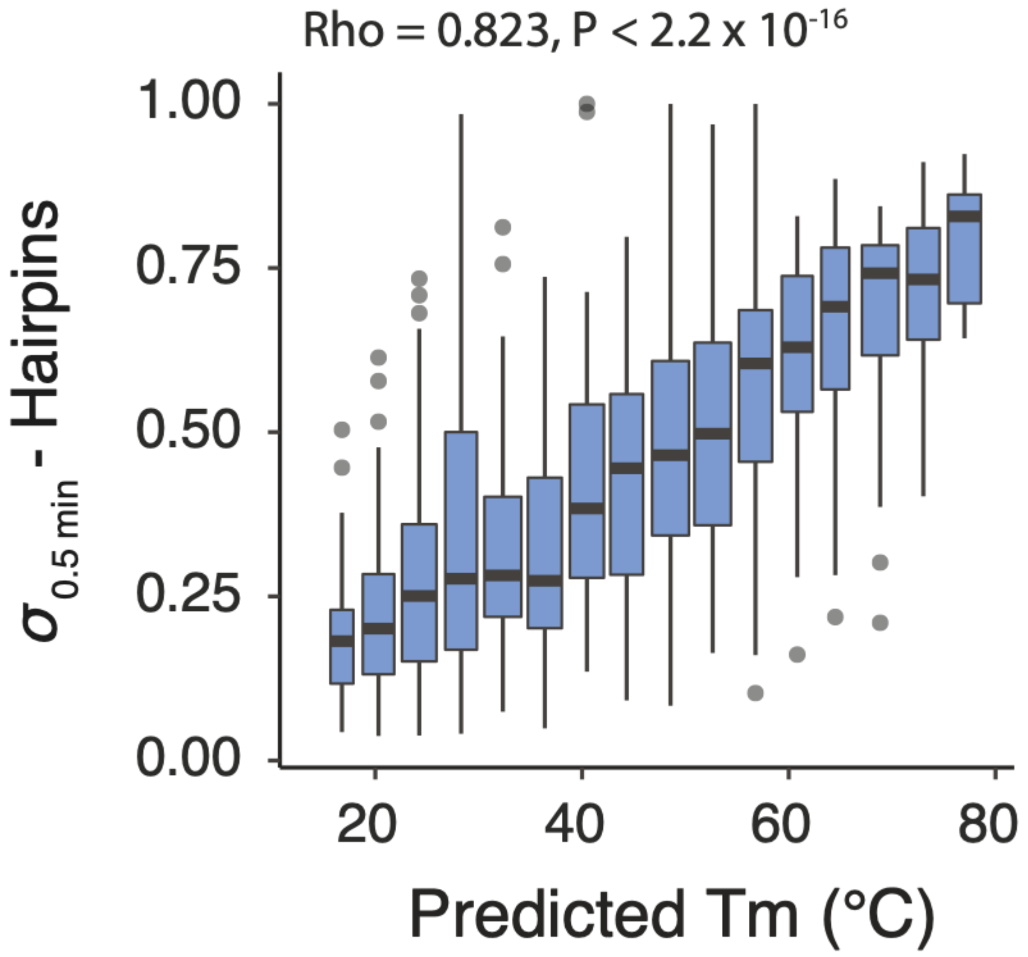
G4 = persistent



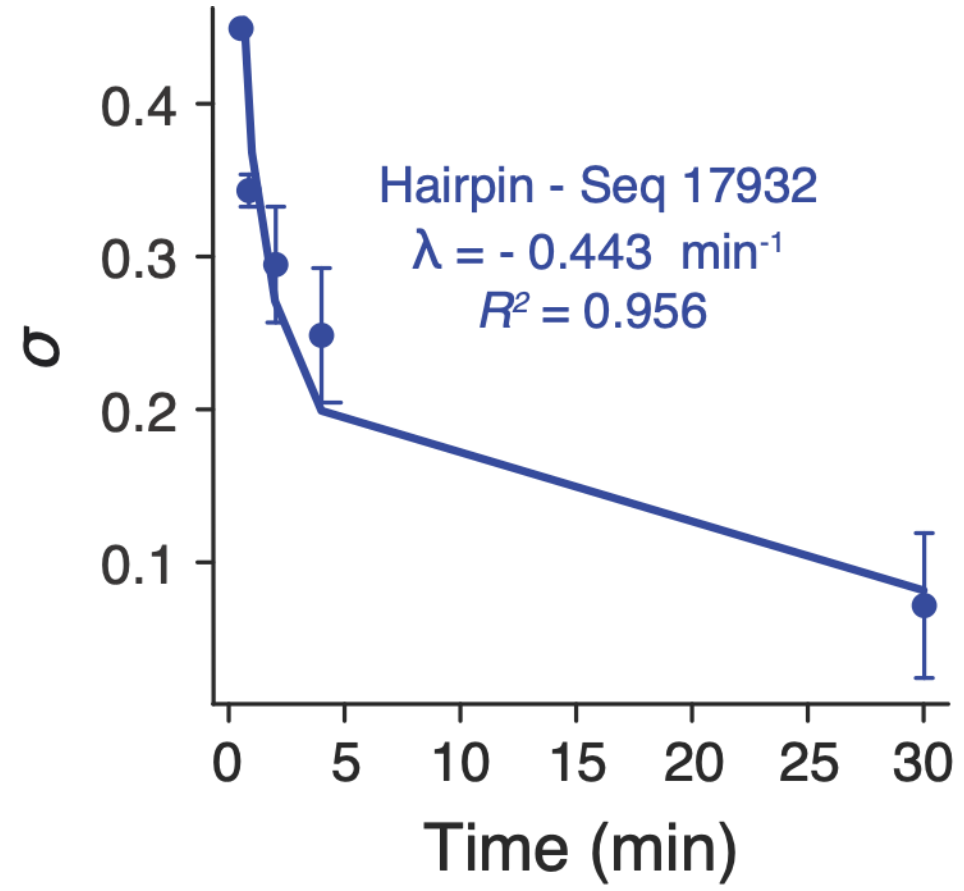


960 sequences





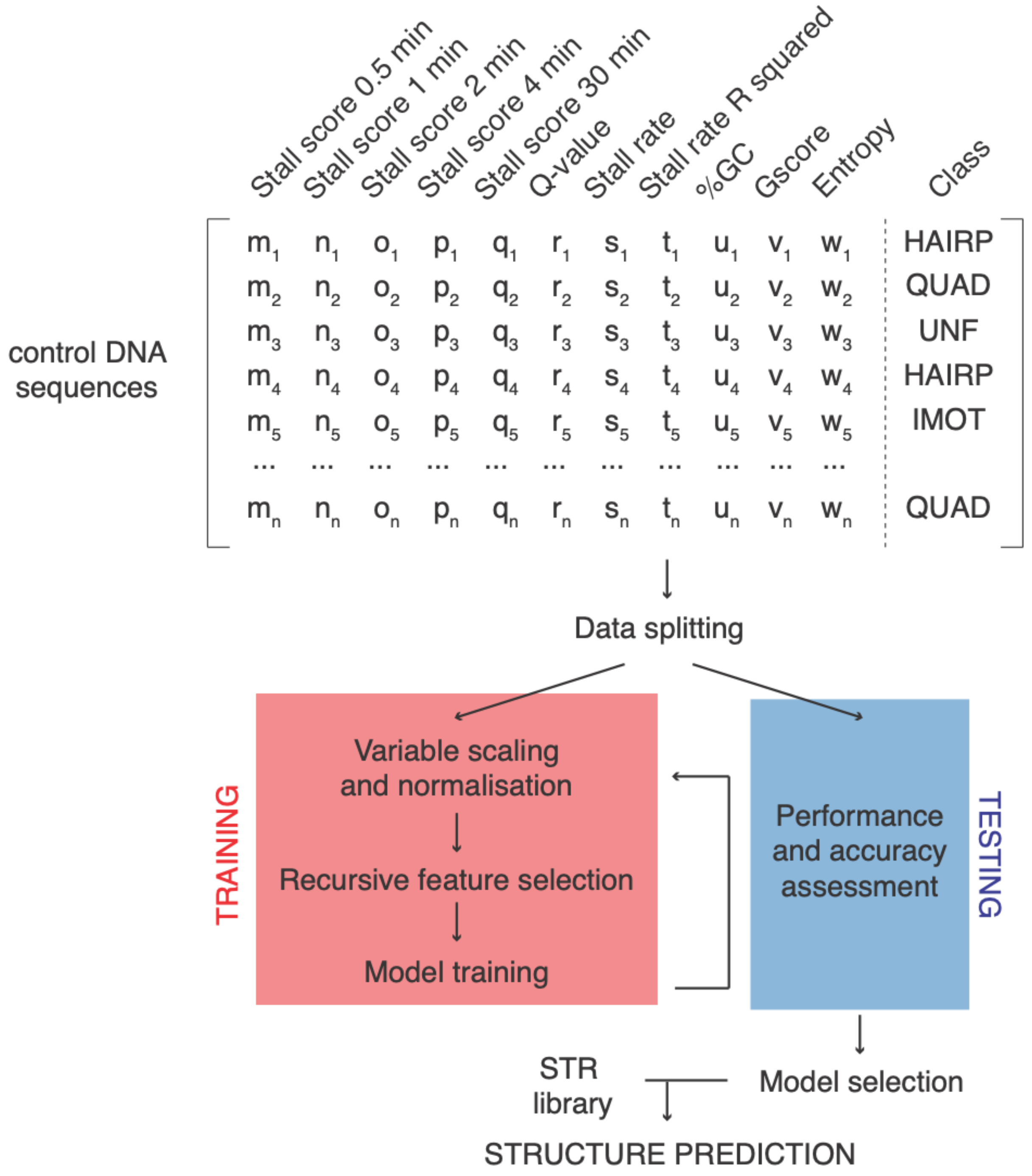
Hairpin = transient



can polymerase response be used to to categorise STRs by structure

Murat, P, Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

# A machine learning approach allows structural categorisation of STRs based on polymerase response



|  | Stall score 0.5 min | Stall score 1 min | Stall score 2 min | Stall score 4 min | Stall score 30 min | Q-value | Stall rate | Stall rate R squared | %GC | Gscore | Entropy | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| control DNA sequences | $m_1$ | $n_1$ | $o_1$ | $p_1$ | $q_1$ | $r_1$ | $s_1$ | $t_1$ | $u_1$ | $v_1$ | $w_1$ | HAIRP |
|  | $m_2$ | $n_2$ | $o_2$ | $p_2$ | $q_2$ | $r_2$ | $s_2$ | $t_2$ | $u_2$ | $v_2$ | $w_2$ | QUAD |
|  | $m_3$ | $n_3$ | $o_3$ | $p_3$ | $q_3$ | $r_3$ | $s_3$ | $t_3$ | $u_3$ | $v_3$ | $w_3$ | UNF |
|  | $m_4$ | $n_4$ | $o_4$ | $p_4$ | $q_4$ | $r_4$ | $s_4$ | $t_4$ | $u_4$ | $v_4$ | $w_4$ | HAIRP |
|  | $m_5$ | $n_5$ | $o_5$ | $p_5$ | $q_5$ | $r_5$ | $s_5$ | $t_5$ | $u_5$ | $v_5$ | $w_5$ | IMOT |
|  | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |  |
|  | $m_n$ | $n_n$ | $o_n$ | $p_n$ | $q_n$ | $r_n$ | $s_n$ | $t_n$ | $u_n$ | $v_n$ | $w_n$ | QUAD |

Data splitting

TRAINING

Variable scaling and normalisation

Recursive feature selection

Model training

TESTING

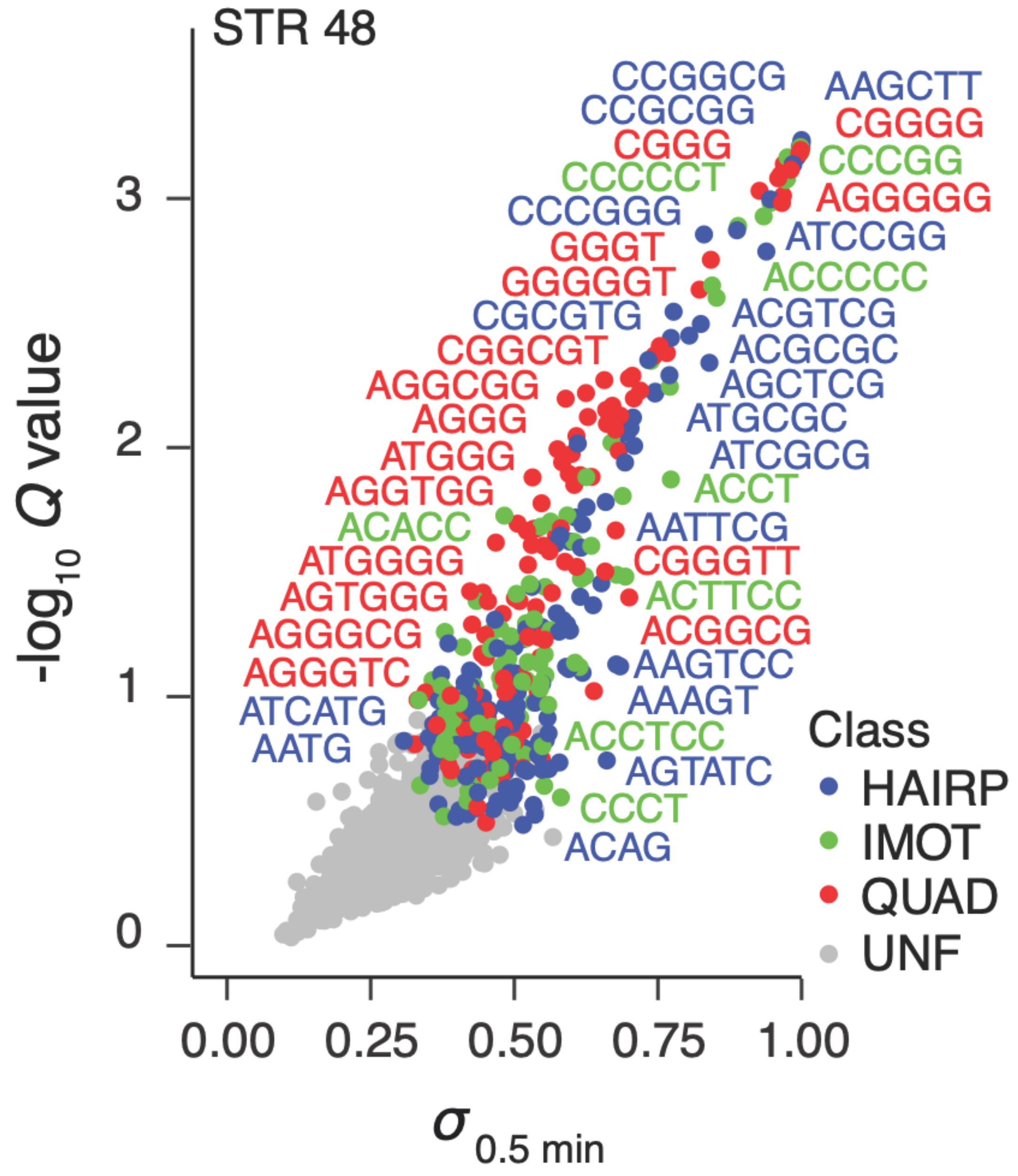
Performance and accuracy assessment

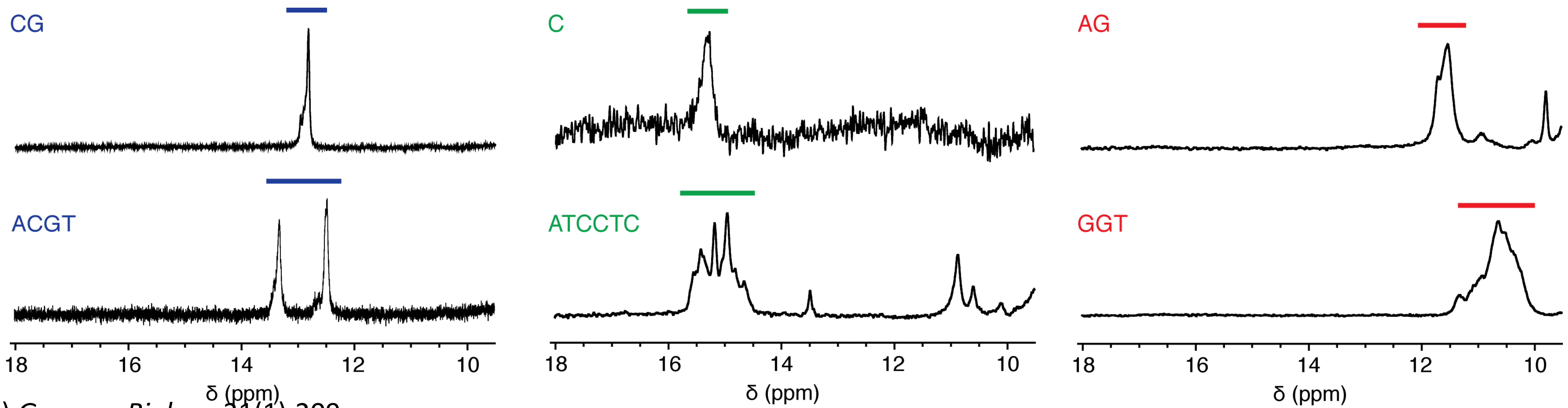
Model selection

STR library

STRUCTURE PREDICTION

accuracy 0.96 +/- 0.03
on test dataset





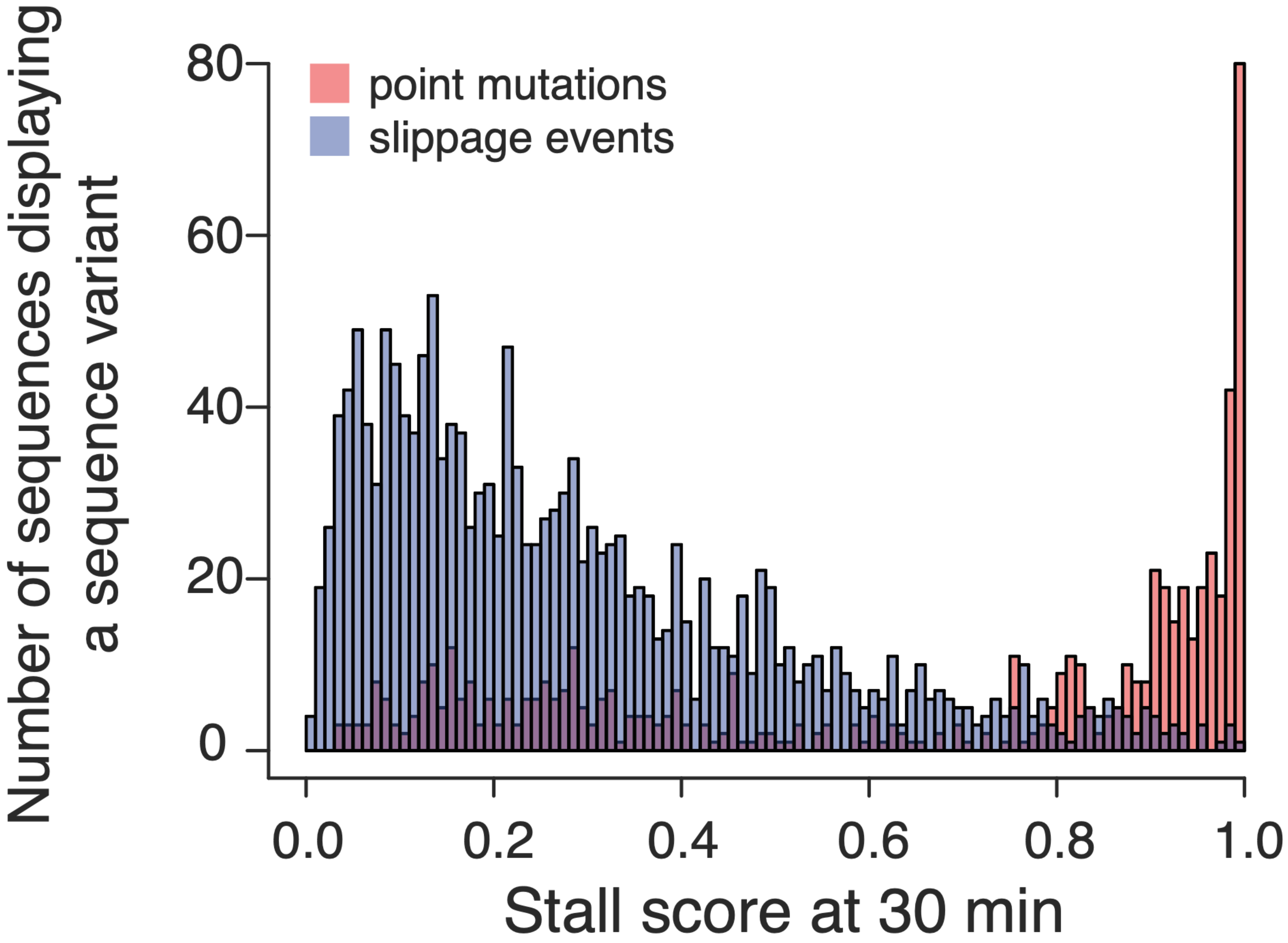
**Hairpin** - WC base pairing

**I-motif** - Hemiprotonated Cs

**G4** - Hoogsteen base pairing

Murat, P, Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

# Polymerase stalling promotes sequence instability

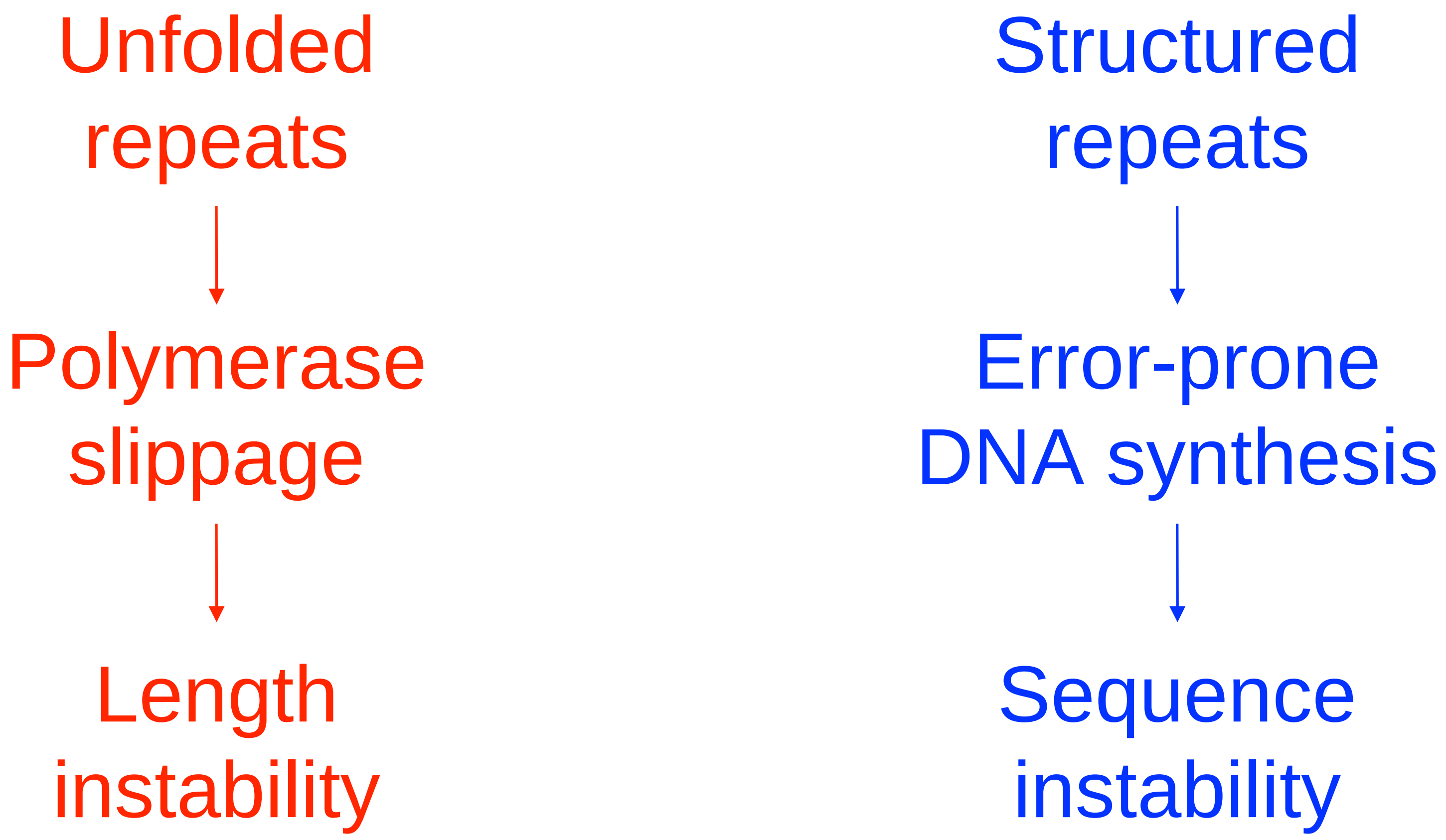


Murat, P,  Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

# How does structure formation by STRs affect their evolutionary behaviour?

*A high throughput replication assay that:*

- Quantifies the efficiency and fidelity of DNA synthesis at all STR permutations
- Infers STR structure from polymerase stalling events
- Establishes general principles for synthesis-dependent STR instability:

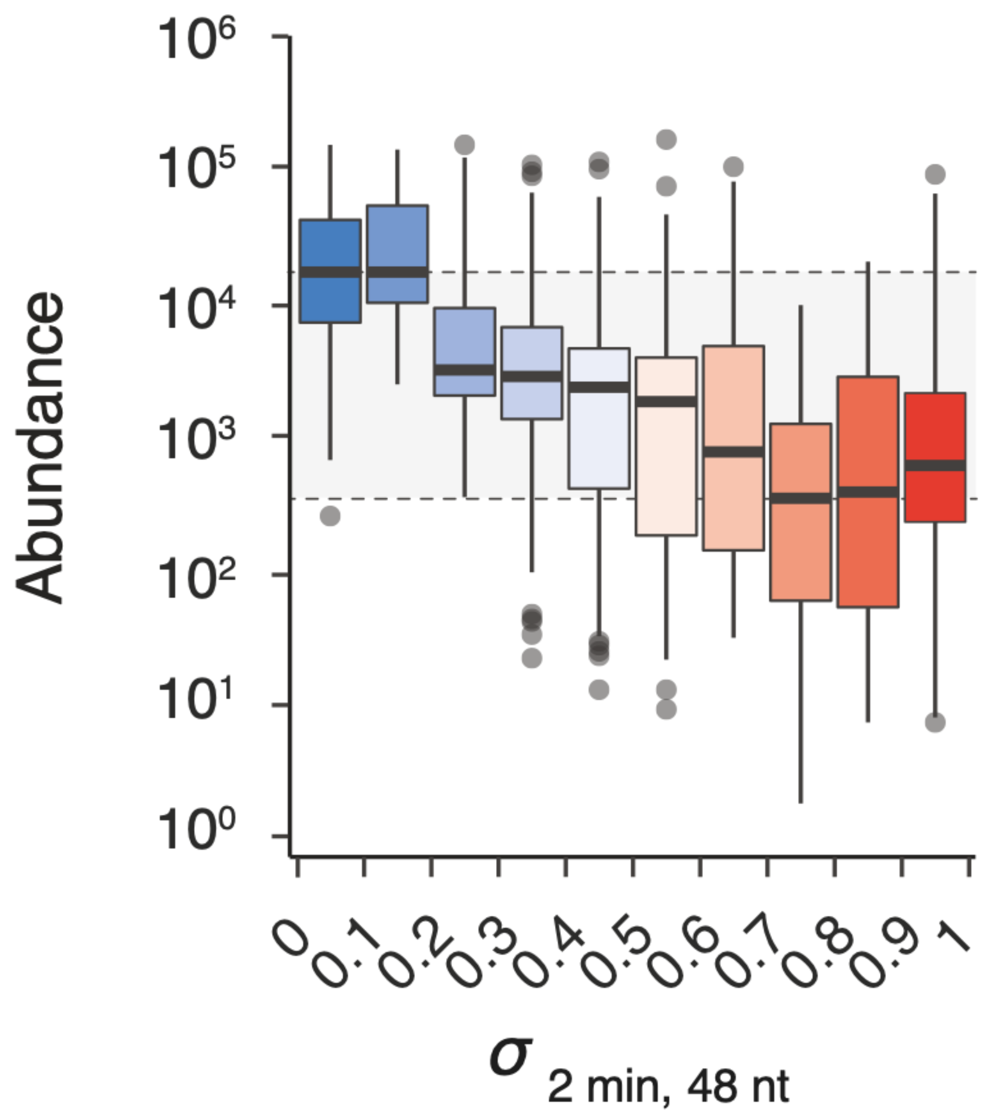


Do these observations have any *in vivo* correlate?

Murat, P, Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

# DNA polymerase stalling at DNA structures predicts STR abundance and length in eukaryotic genomes

*Human* genome (4,500,000 STR loci)

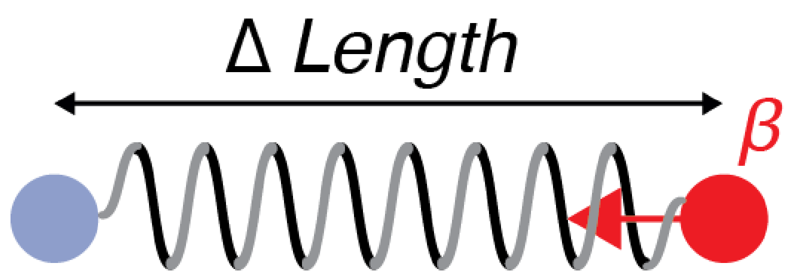


Same trends observed for the *Mouse*, *Chicken*, *Zebrafish*, *Fly* and *Yeast* genomes

Murat, P, Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

# Expansion is favoured and less constrained in weakly stalling repeats

Multi-step
Ornstein-Uhlenbeck
process



Length at equilibrium

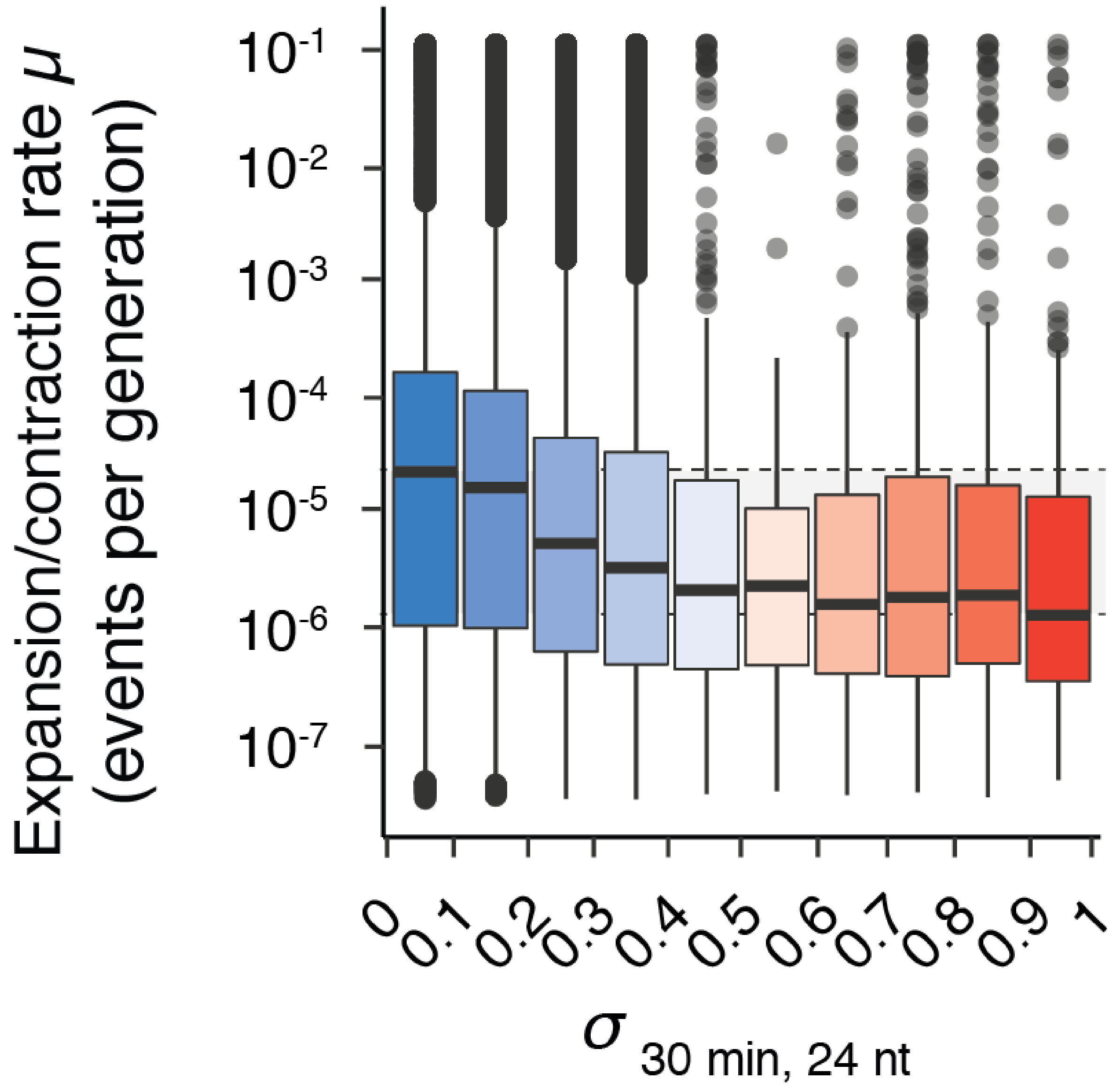
Length after expansion

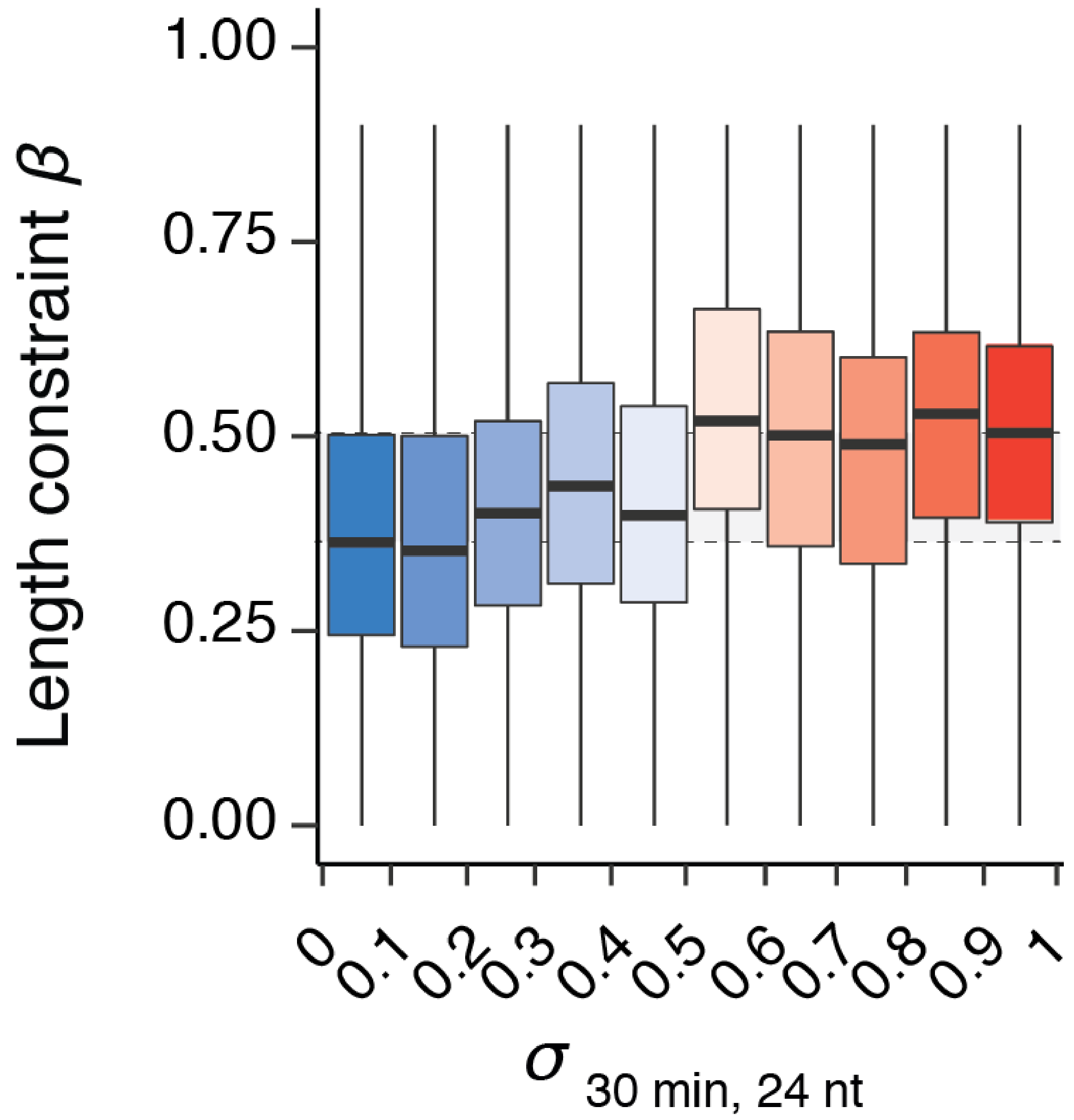
$$\mu = \frac{\mathrm{d}Length}{\mathrm{d}t} \text{ : mutation rate}$$

$\beta$ : Length constraint

Mutation rate

Evolutionary constraint





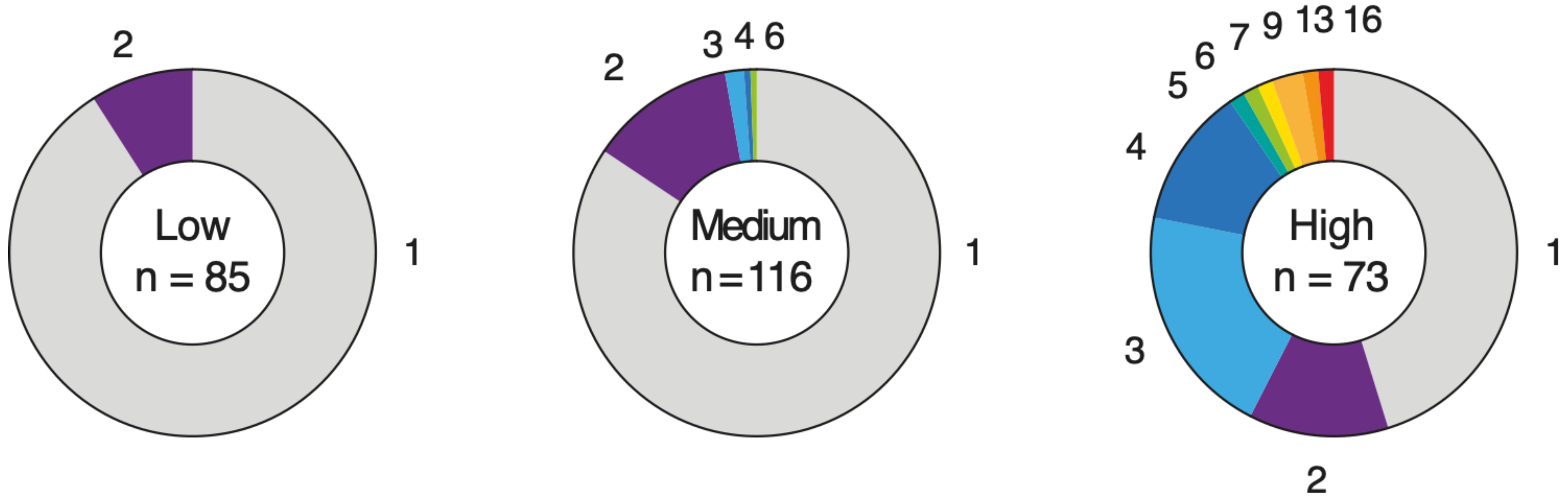
What is the basis for the increased length constraint on structured STRs?

Murat, P, Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

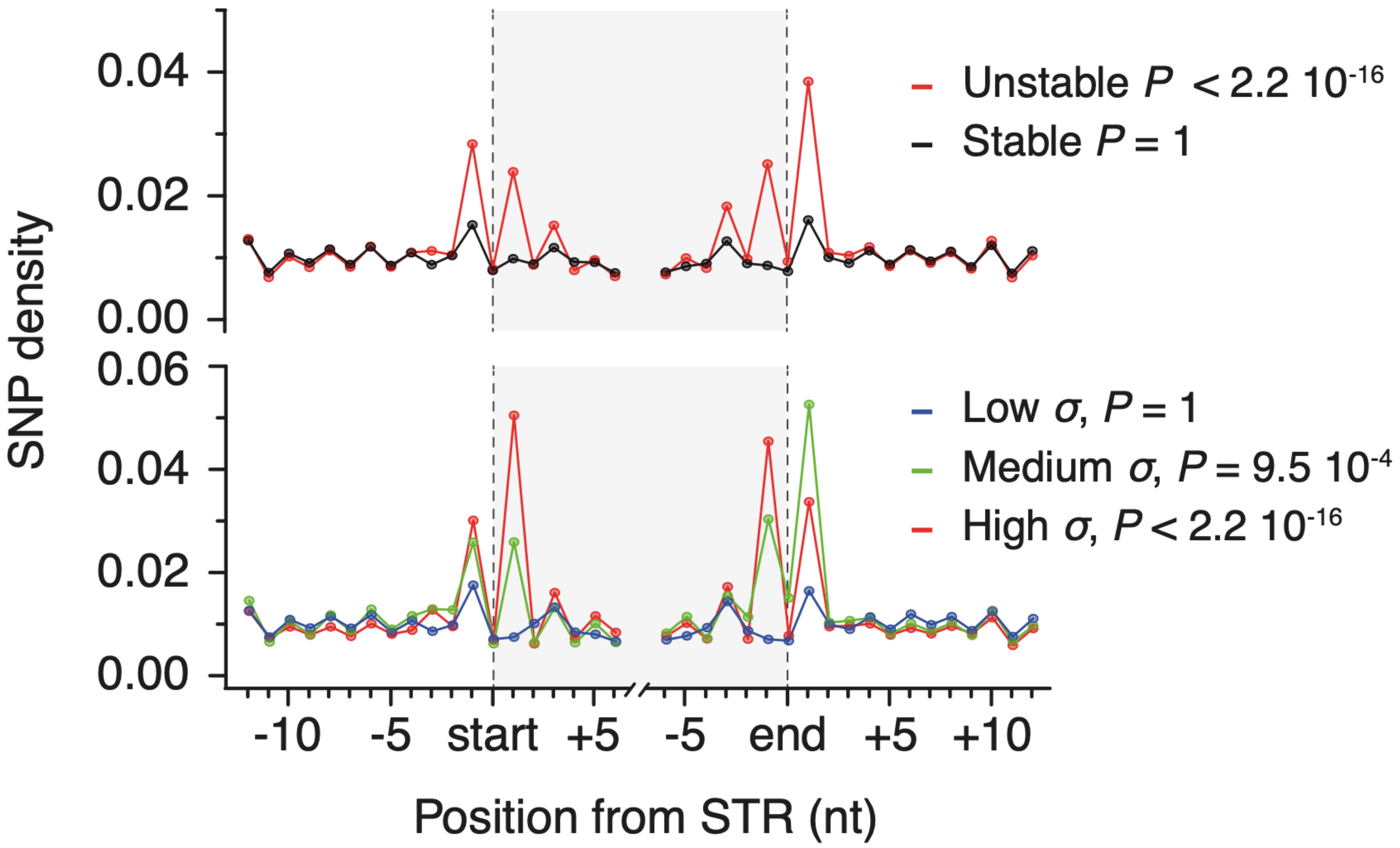
Gymrek et al. Nature Gen. 2017, 49, 1495-1501

# Structured STRs are prone to point mutation in the human genome: a mechanism for length constraint?

Highly stalling (high σ) sequences are more prone to point mutation

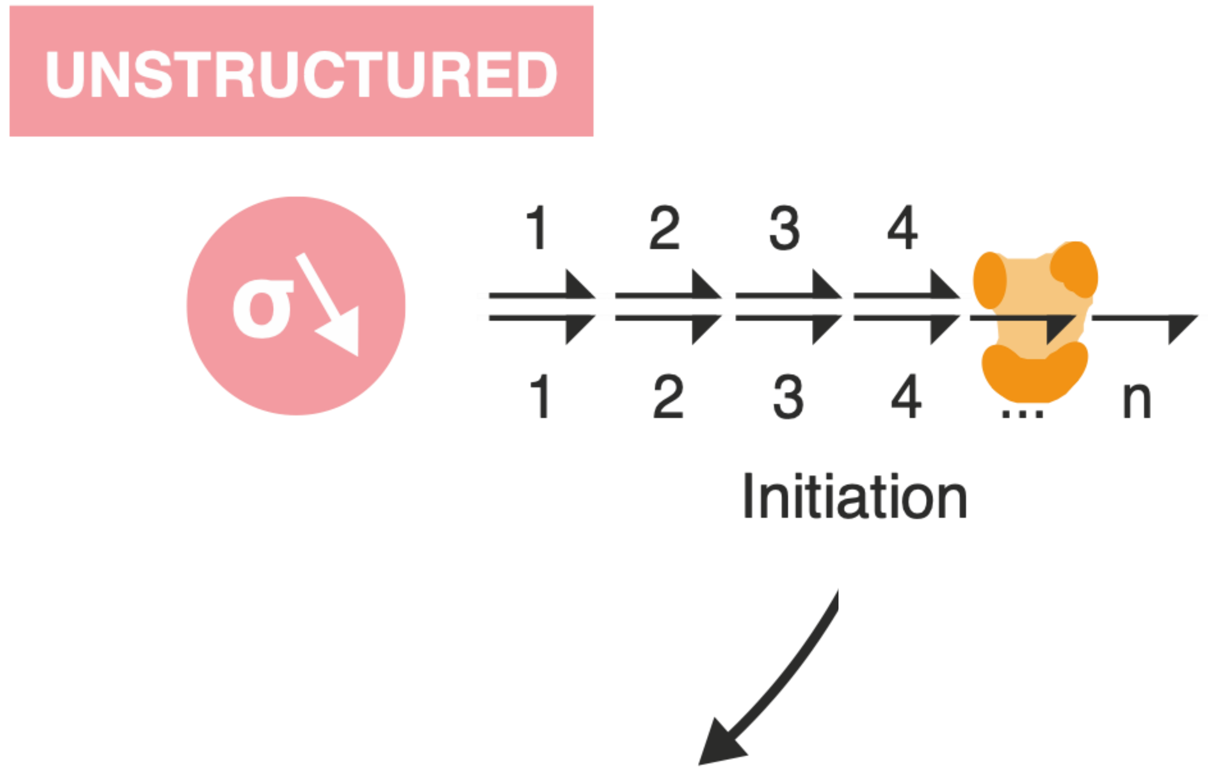


Density of germline mutations in the vicinity of STRs :



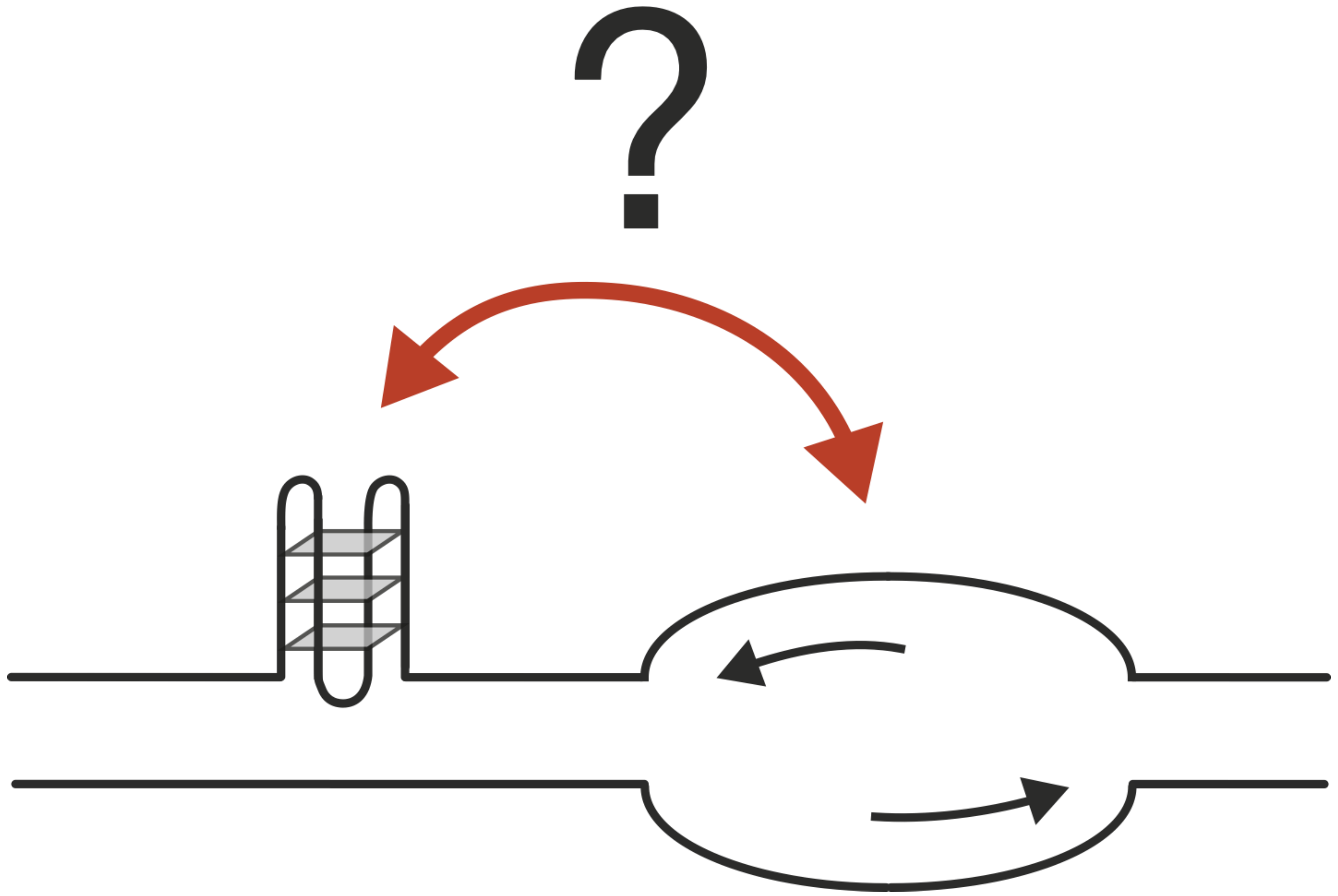
Murat, P, Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

# DNA polymerase stalling at structured DNA constrains the expansion of Short Tandem Repeats



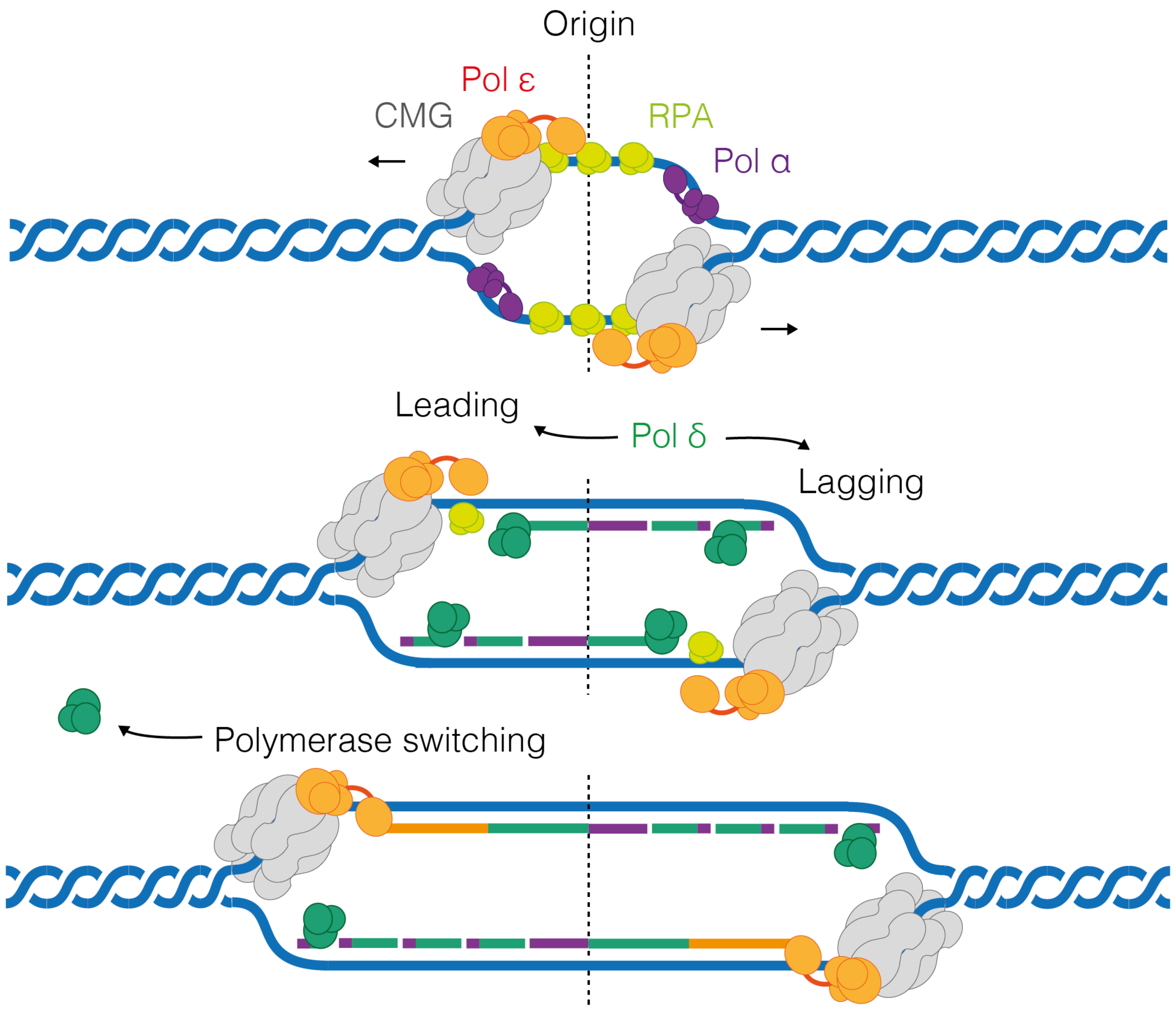
Murat, P,  Guilbaud, G, & Sale, JE (2020) *Genome Biology* 21(1):209

# The origin of origins



Why are G4s associated with replication origins?

# Hypothesis: replication origins are hotspots for mutagenesis



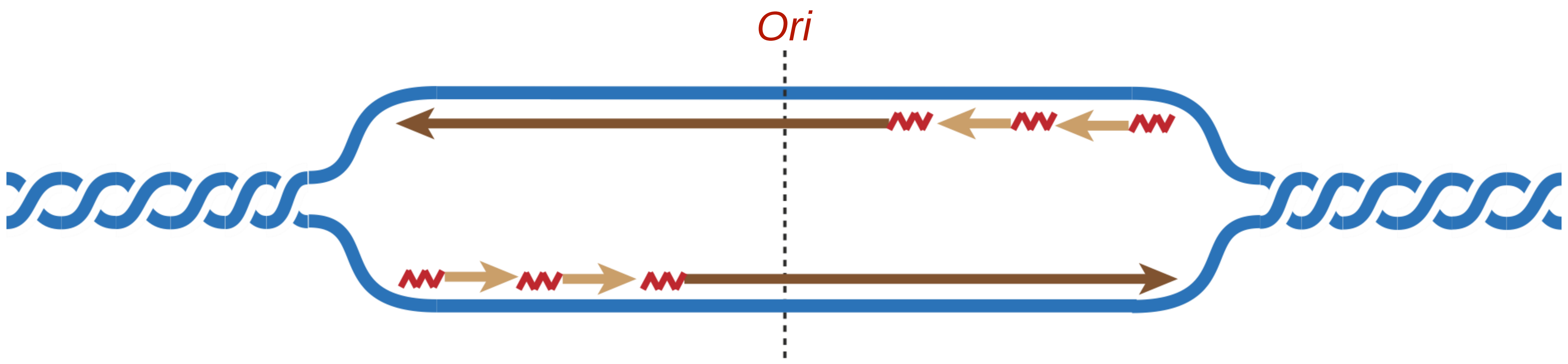
Decreased accessibility

Torsional constraints

Ribonucleotide persistance

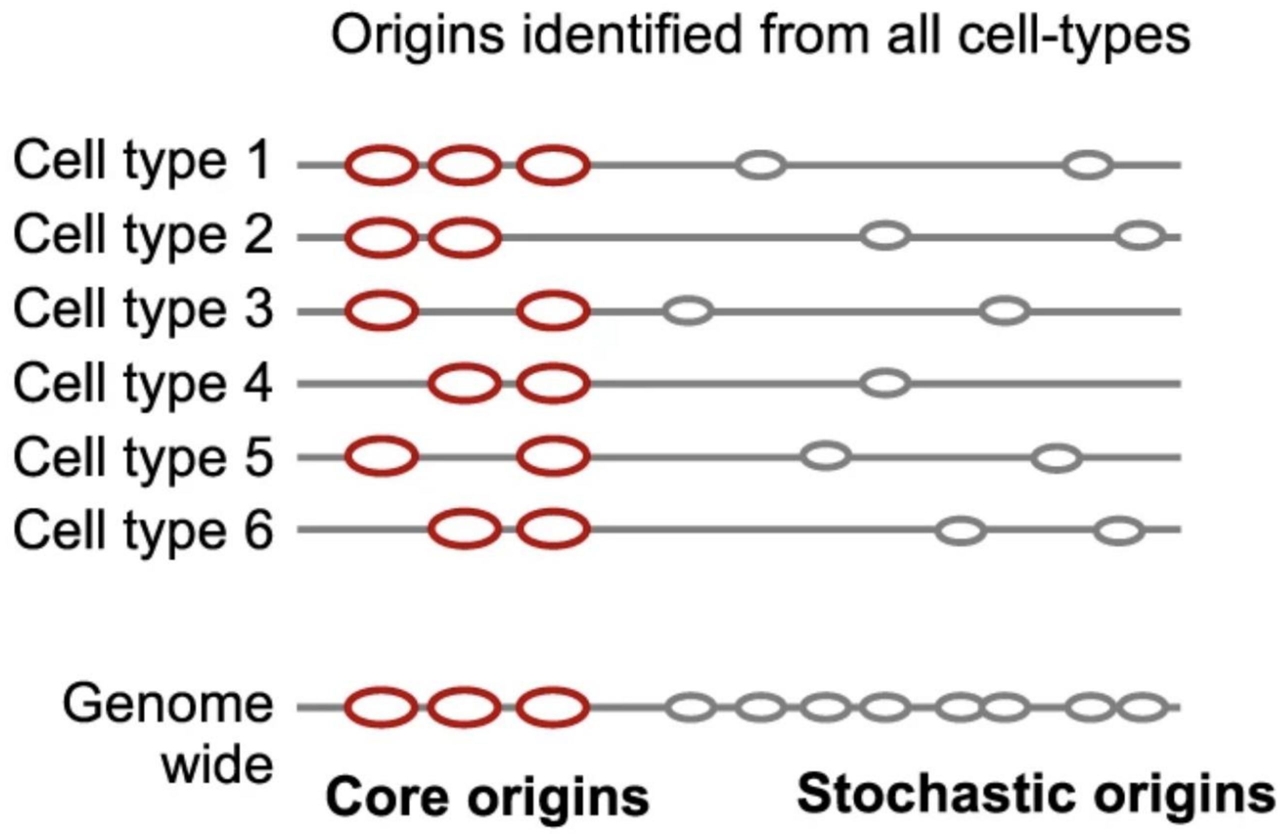
Uncoupled DNA synthesis

# Identifying origins of replication in human cells



## Short Nascent Strand sequencing (SNS-seq)

> 300,000 origins
1-4kb resolution

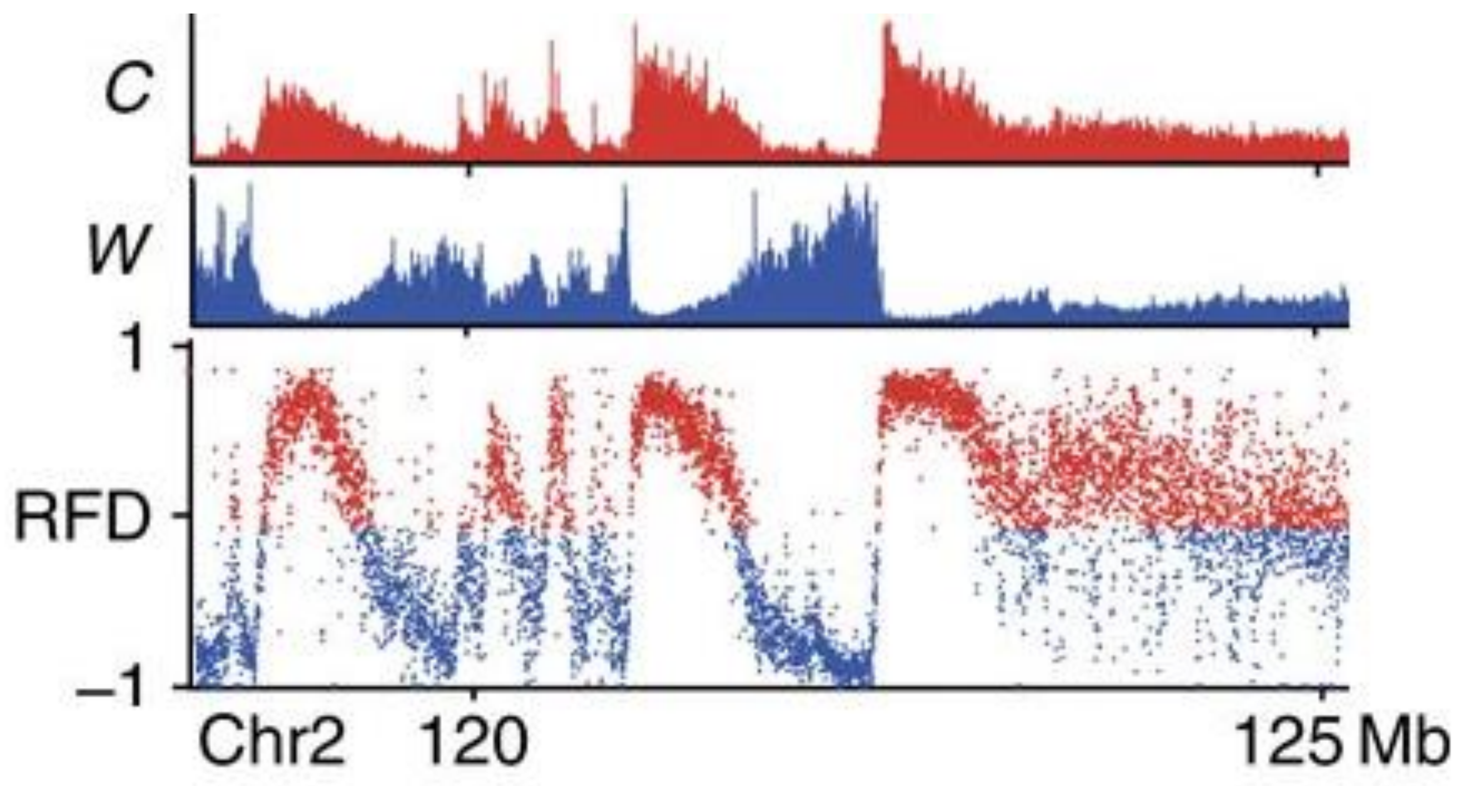


**Core origins**
Highly ubiquitous
High activity
~64,000 regions

**Stochastic origins**
Stochastic
Low activity
~250,000 regions

Akerman et al. 2020

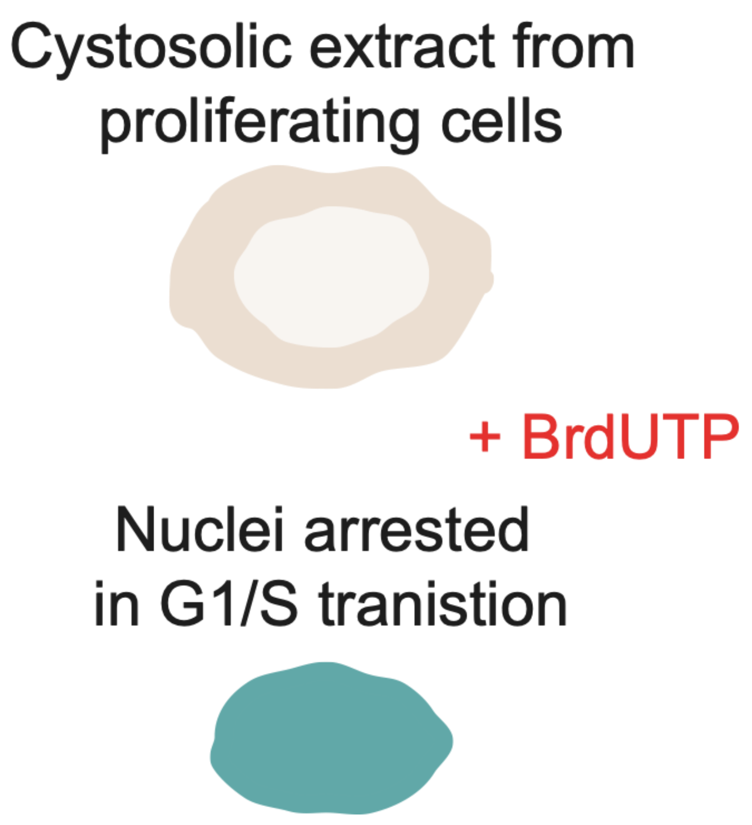
## Okazaki fragment sequencing (Ok-seq)



~ 22,000 initiation zones
~50kb resolution

Petryk et al. 2015

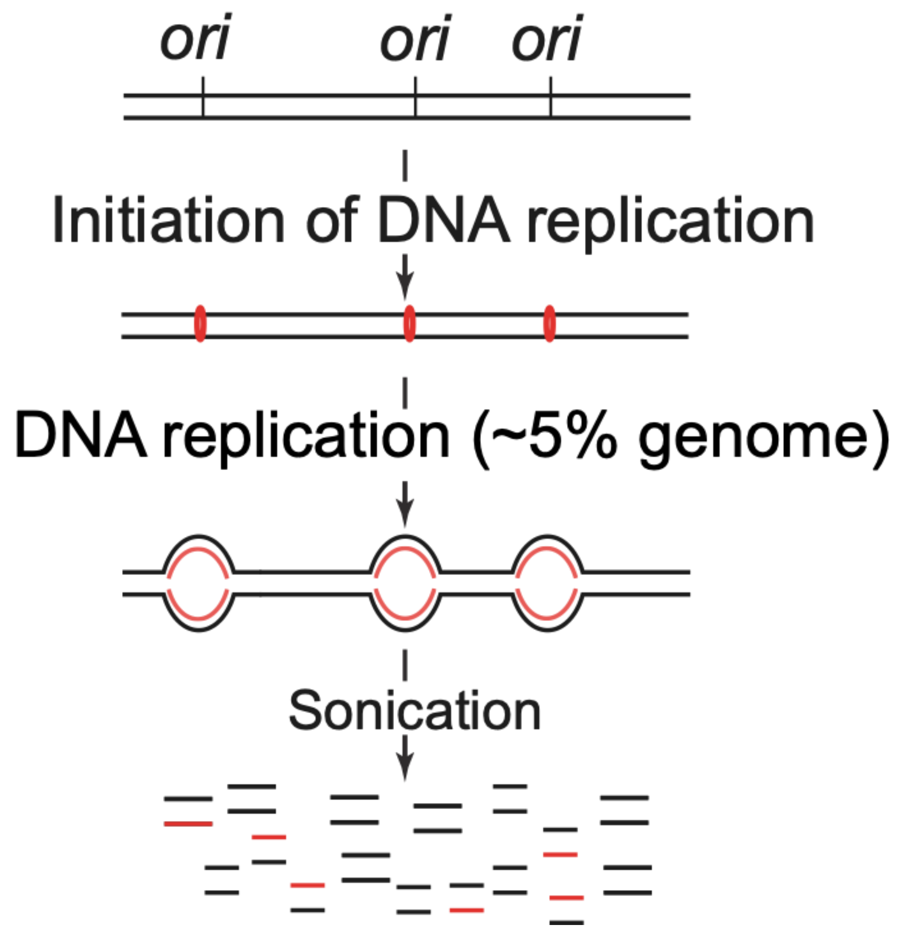
# Ini-seq 2: a method to map both replication origin location and efficiency



Cystosolic extract from proliferating cells

+ BrdUTP

Nuclei arrested in G1/S tranistion

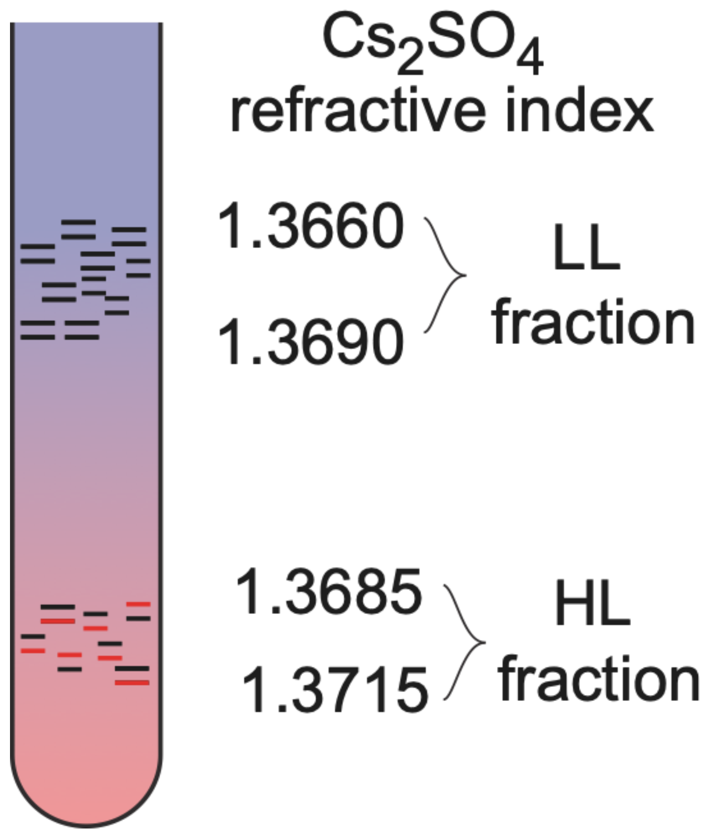


ori ori ori

Initiation of DNA replication

DNA replication (~5% genome)

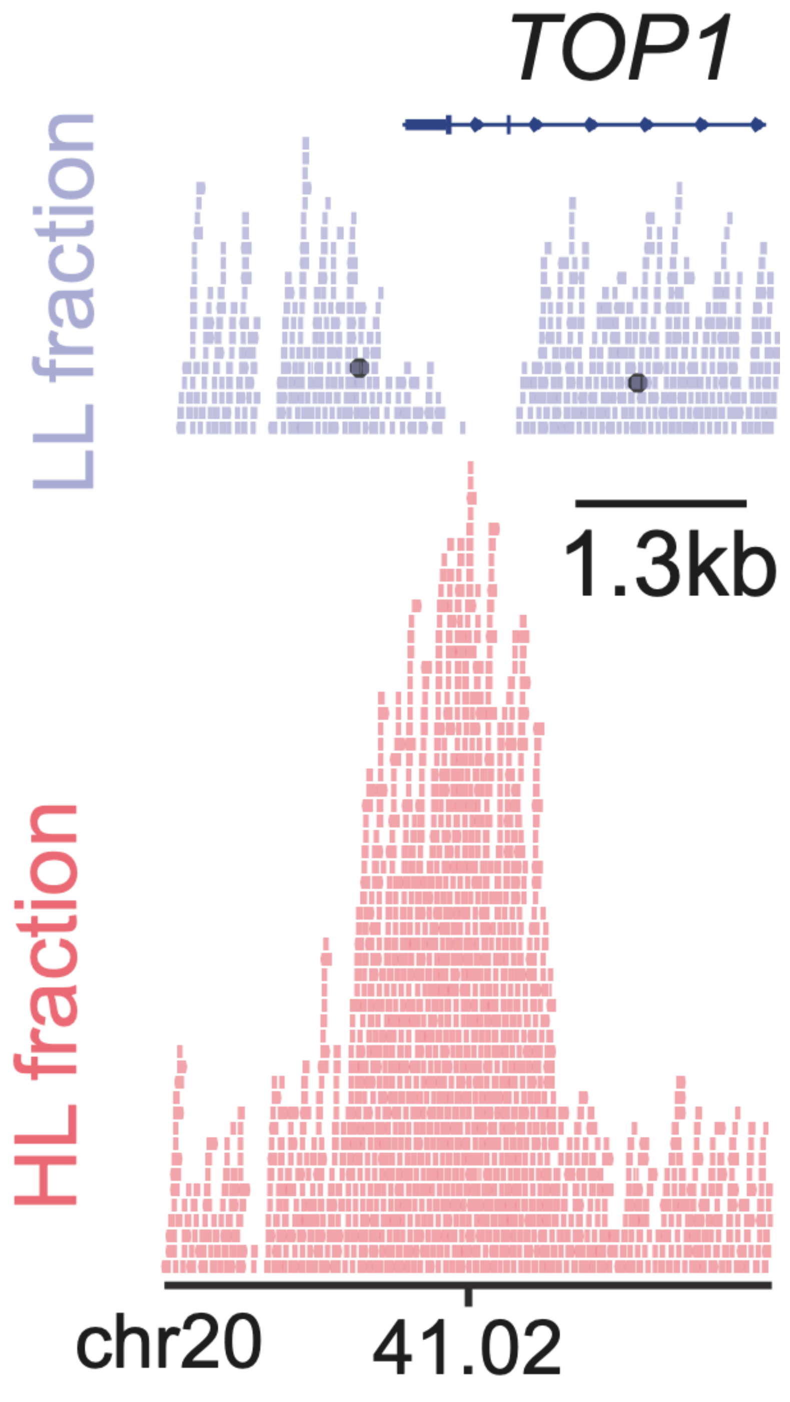
Sonication



$Cs_2SO_4$ refractive index

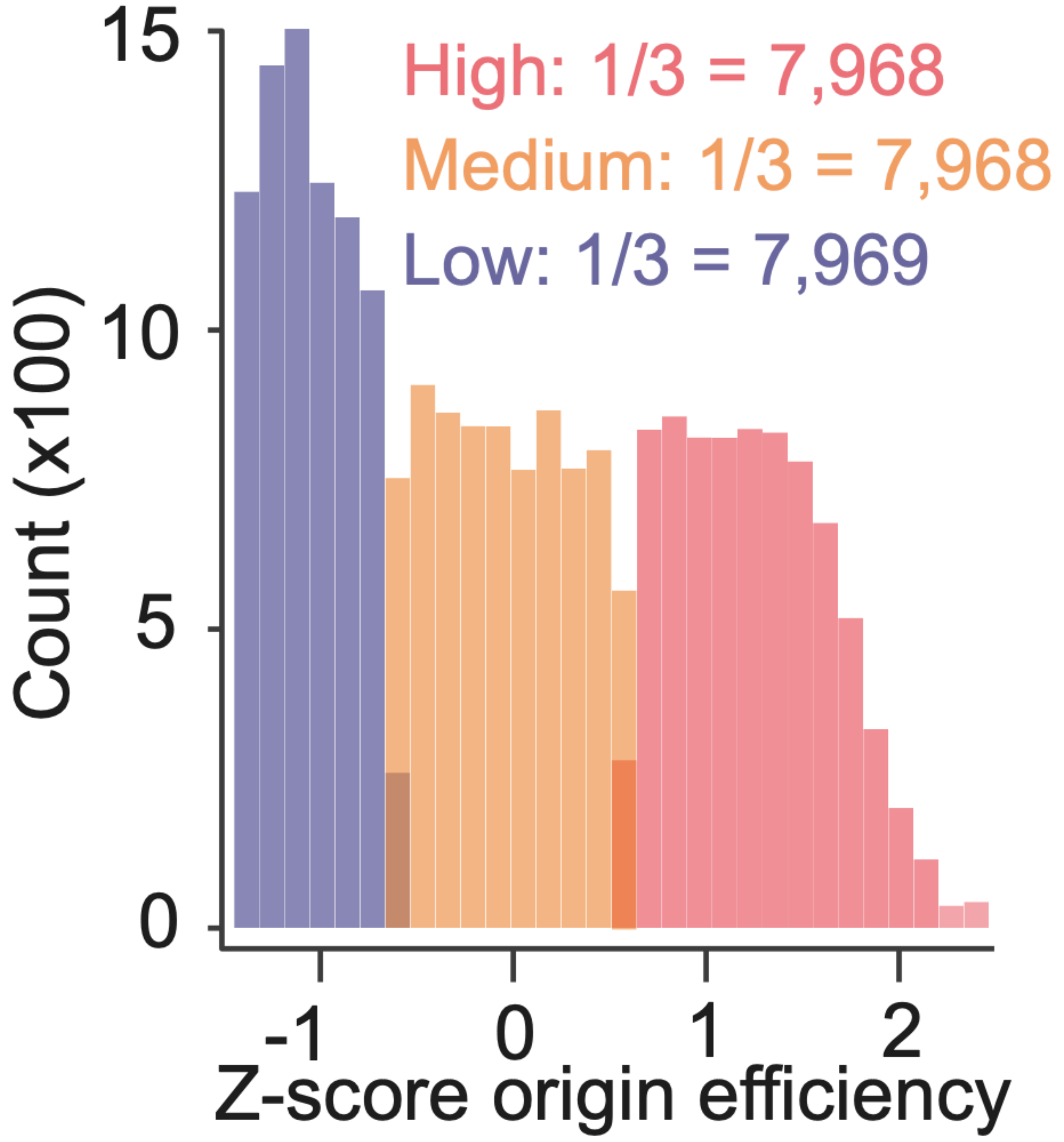
1.3660
1.3690
LL fraction

1.3685
1.3715
HL fraction



Guilbaud et al. (2022) Nucleic Acids Res. 50, 7436 - 7450

**23,905 called origins**

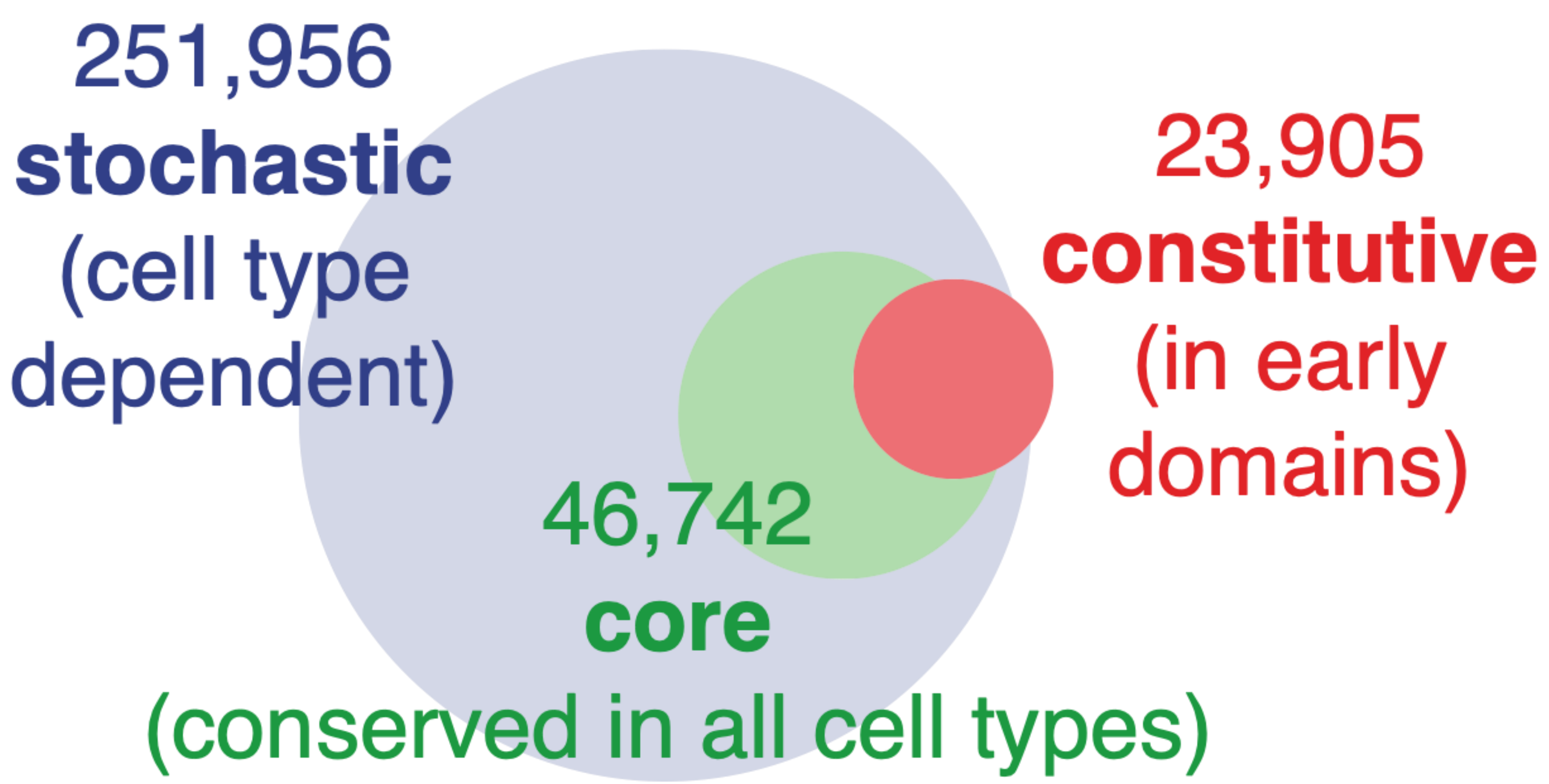


High: 1/3 = 7,968

Medium: 1/3 = 7,968

Low: 1/3 = 7,969

322,603 human origins



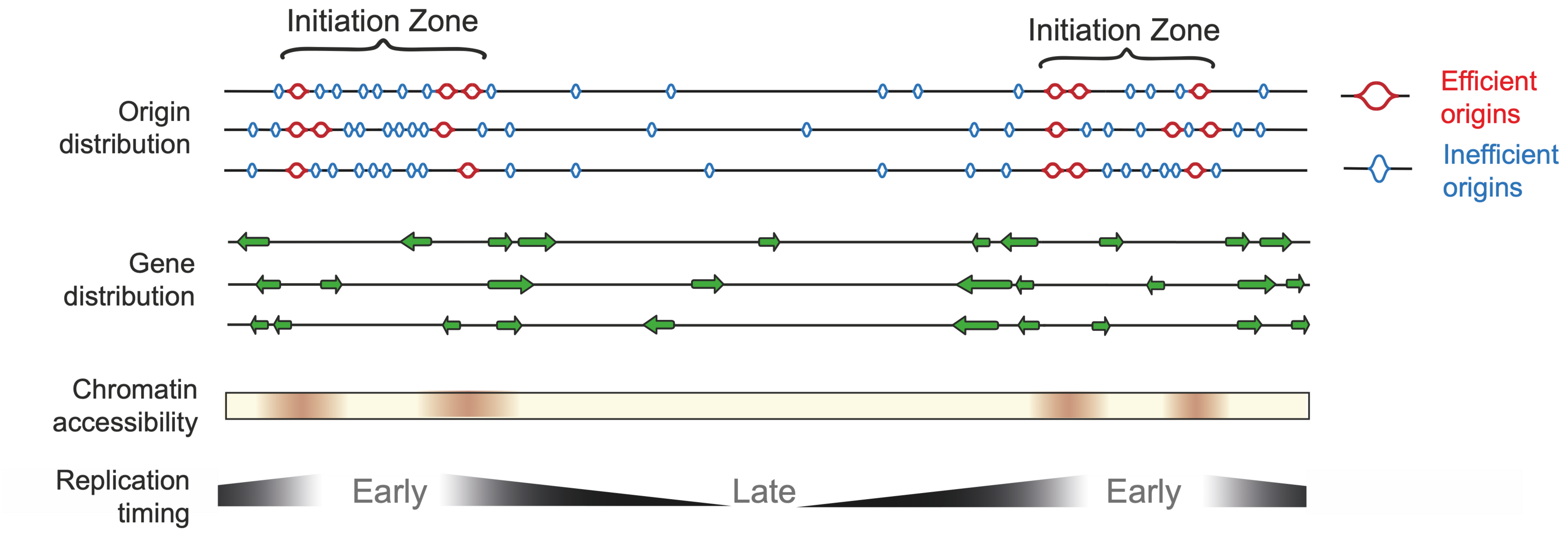
251,956 **stochastic** (cell type dependent)

23,905 **constitutive** (in early domains)

46,742 **core** (conserved in all cell types)

Datasets:
Akerman et al. Nature Commun. 2020, 11, 1-15
Guilbaud et al. (2022) Nucleic Acids Res. 50, 7436 - 7450

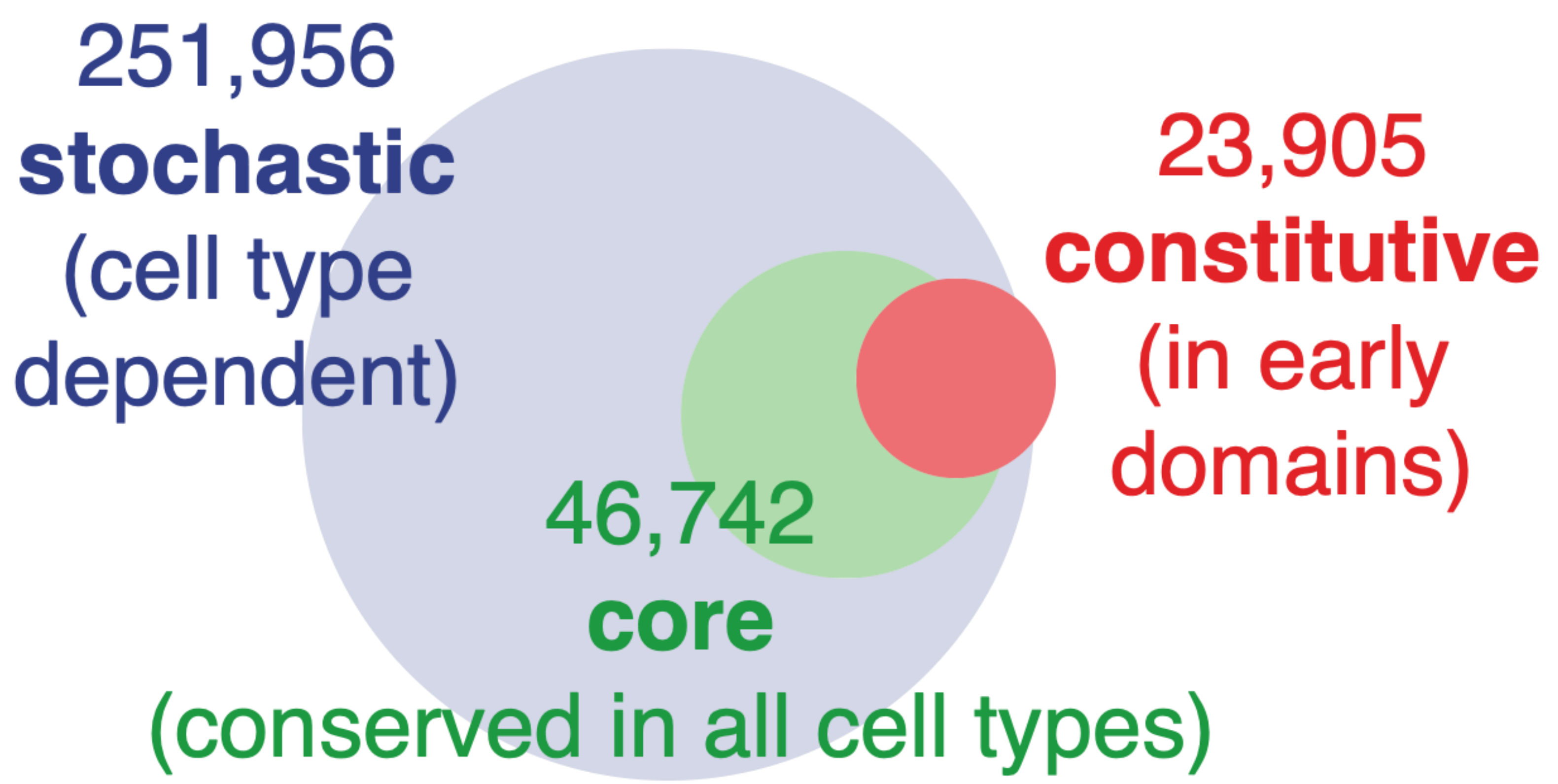
# Organisation of replication origins by efficiency in the human genome



Guilbaud et al. (2022) Nucleic Acids Res. 50, 7436 - 7450

# The mutational footprint of replication initiation is revealed at ‘constitutive’ ini-seq origins

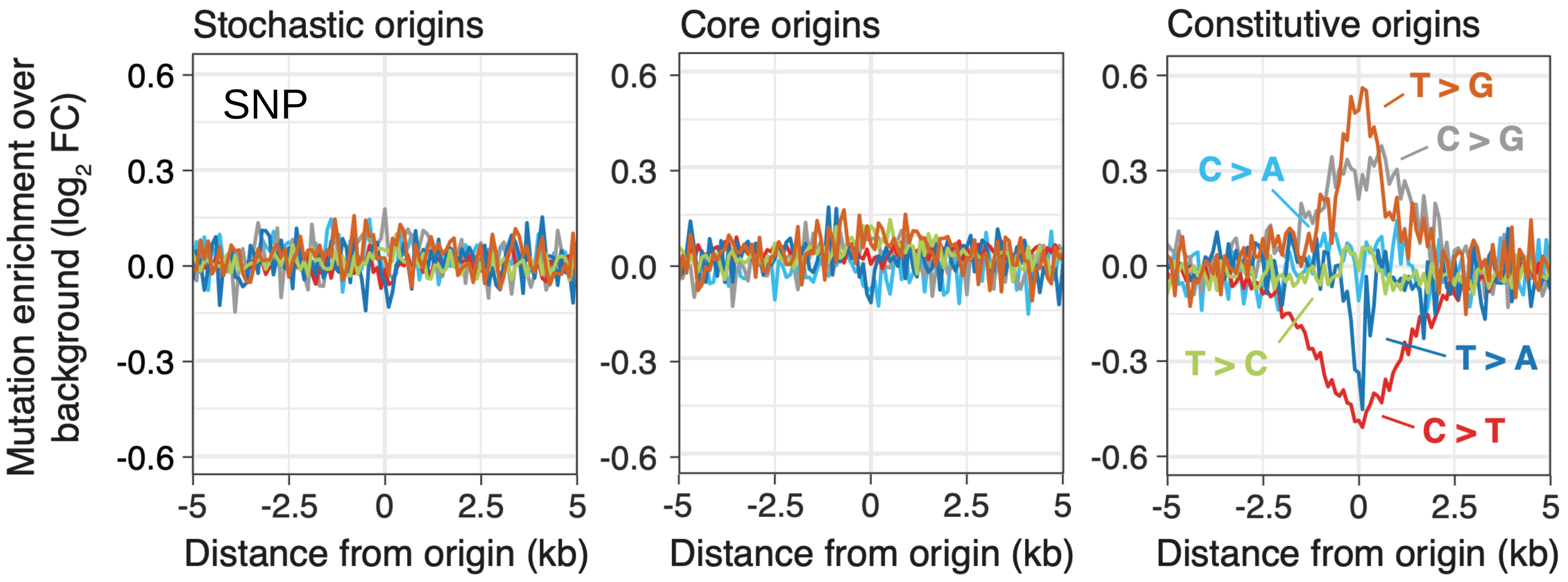
322,603 human origins

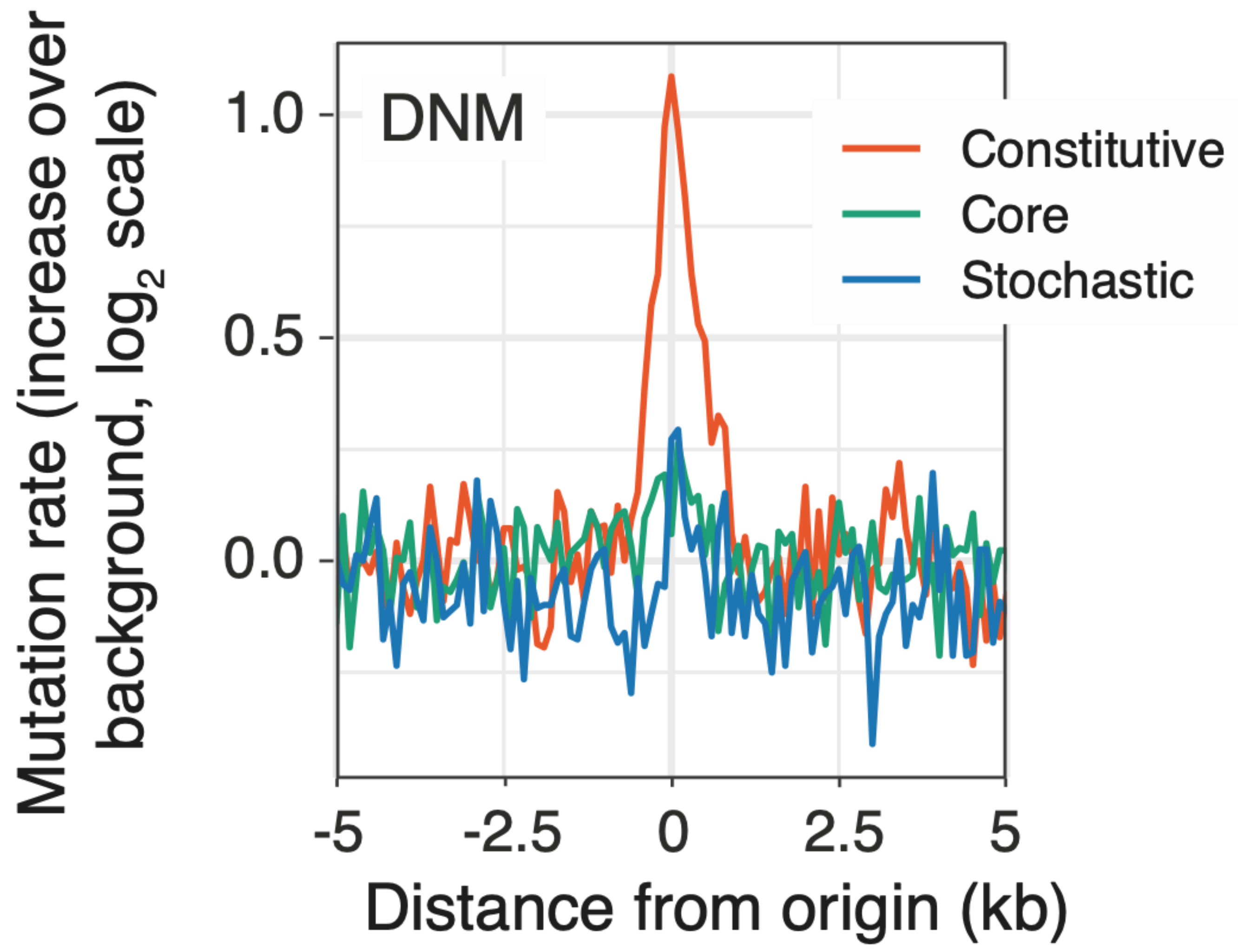


251,956 **stochastic** (cell type dependent)

23,905 **constitutive** (in early domains)

46,742 **core** (conserved in all cell types)

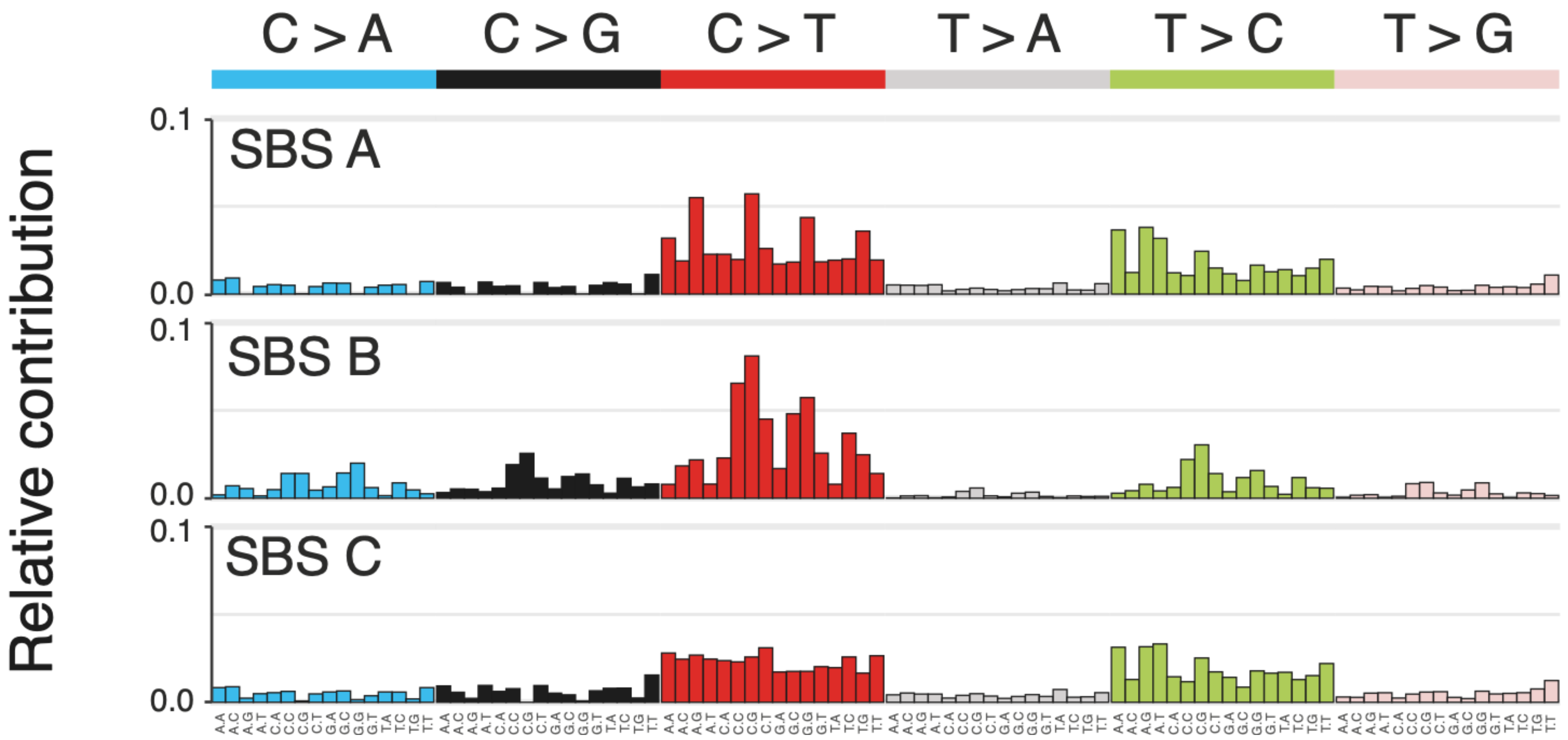


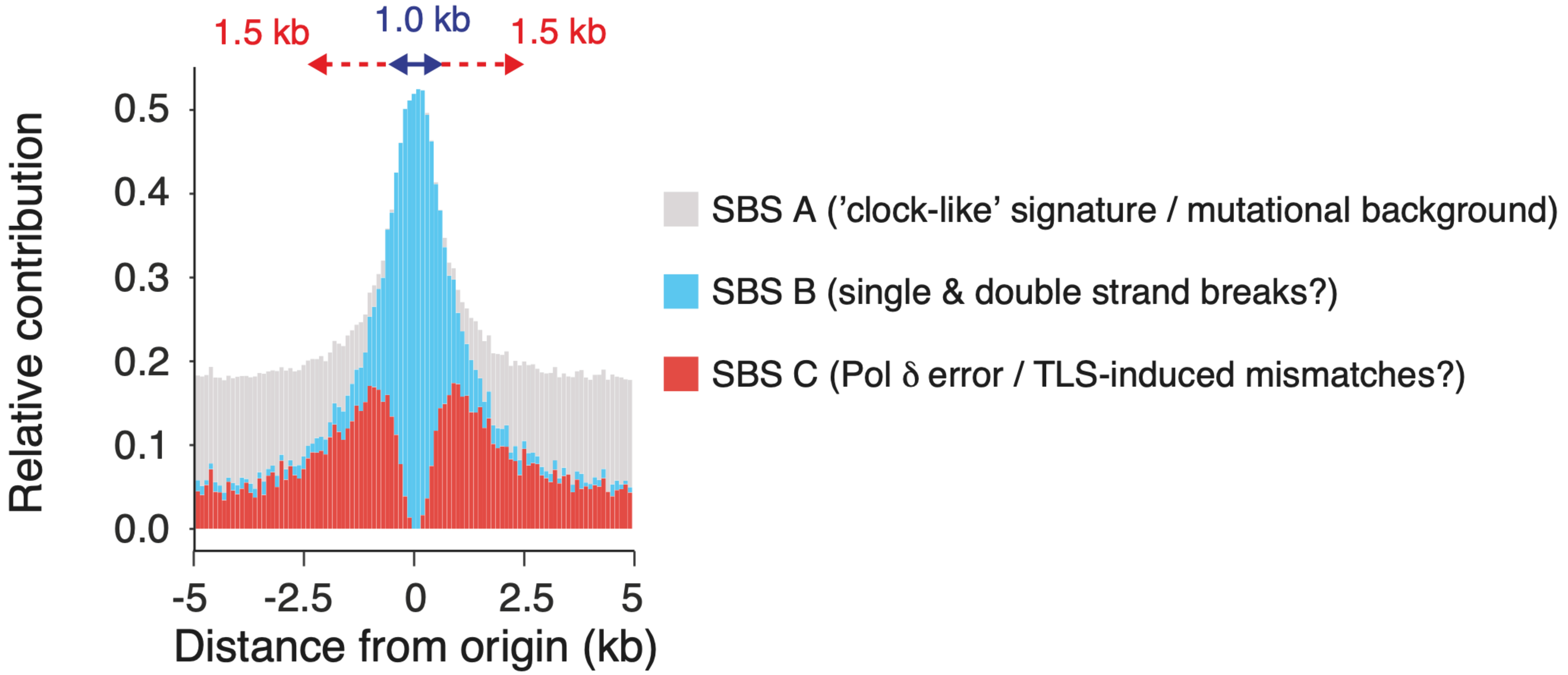


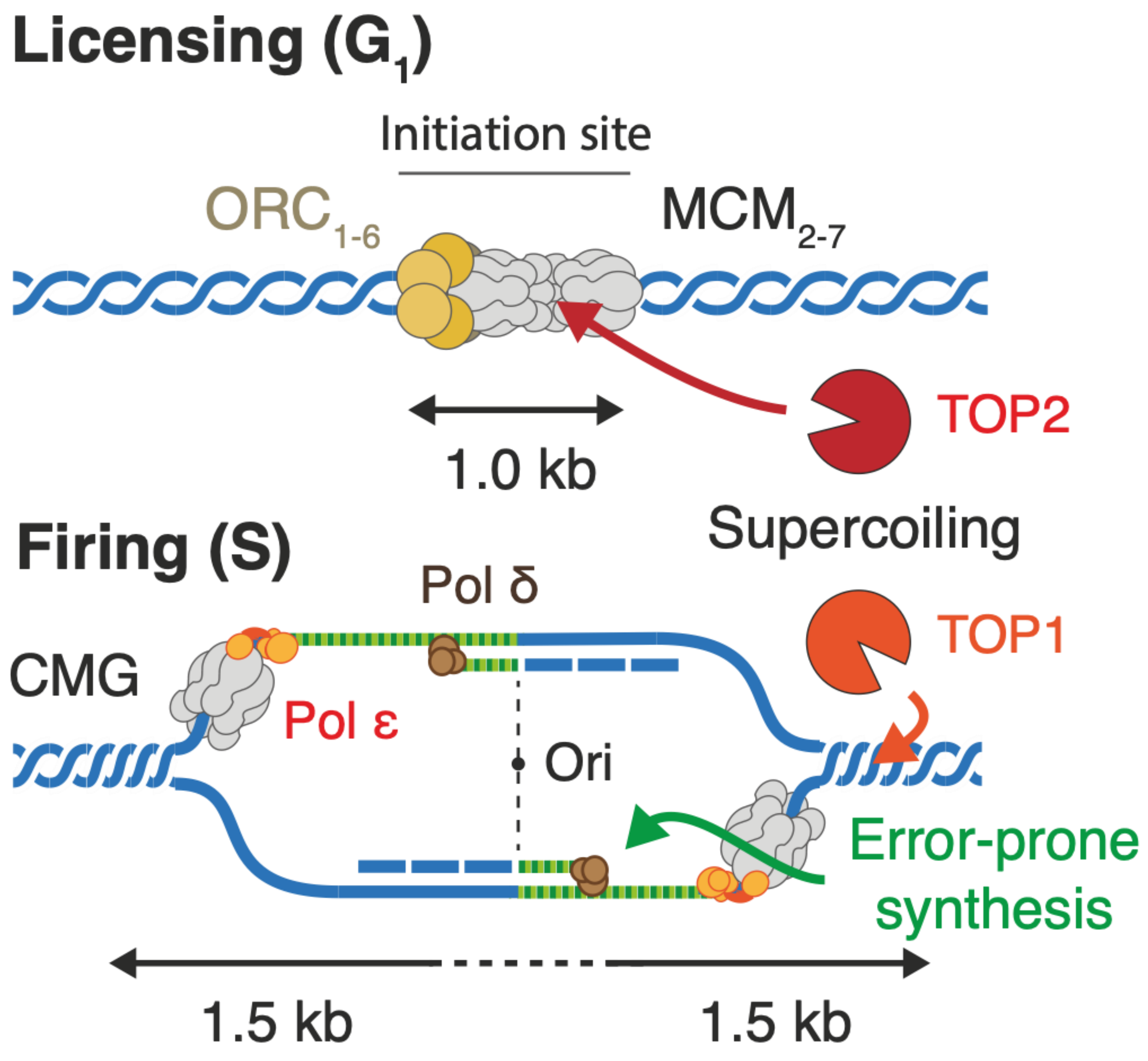
Akerman et al. Nature Commun. 2020, 11, 1-15
Guilbaud et al. (2022) Nucleic Acids Res. 50, 7436 - 7450

Murat et al., (2022) *Sci. Adv.* 8(45):eadd3686

# The mutational signature of replication origins

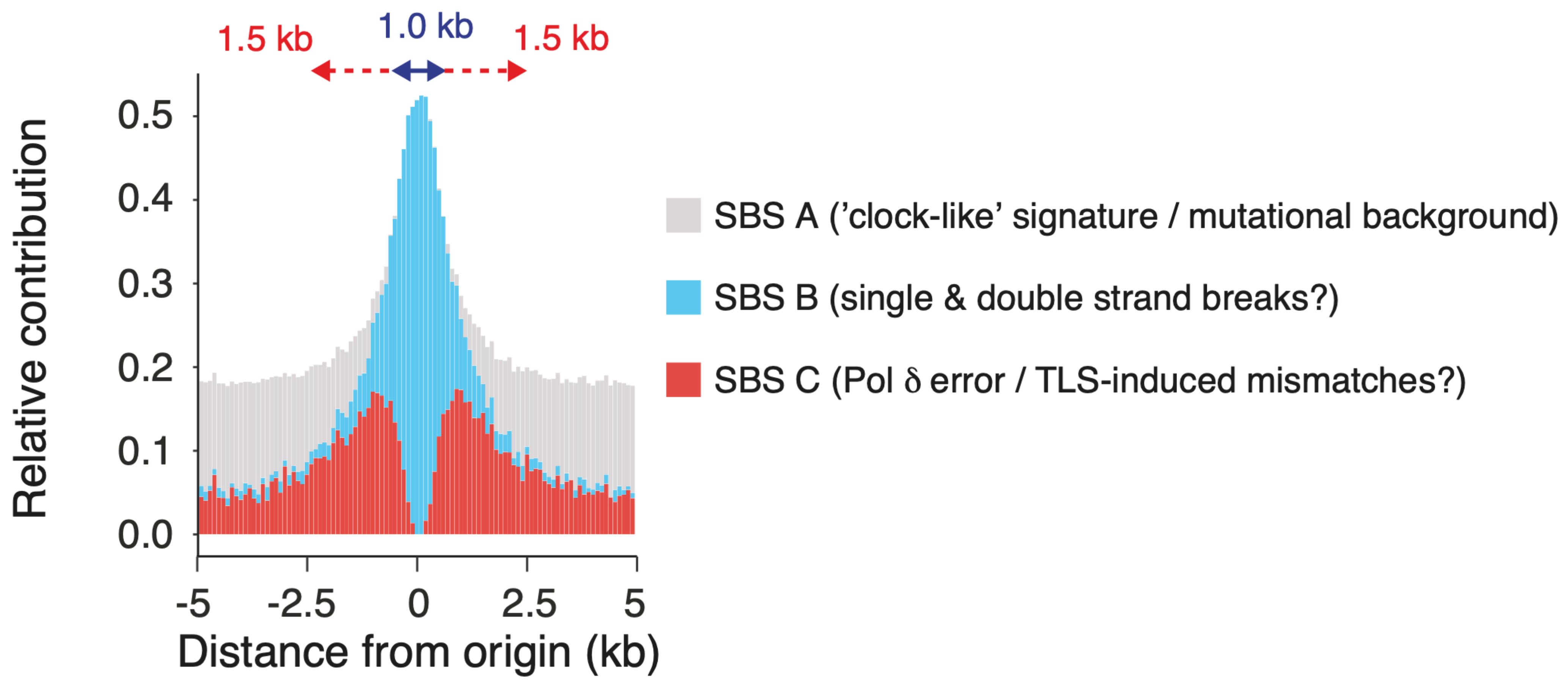


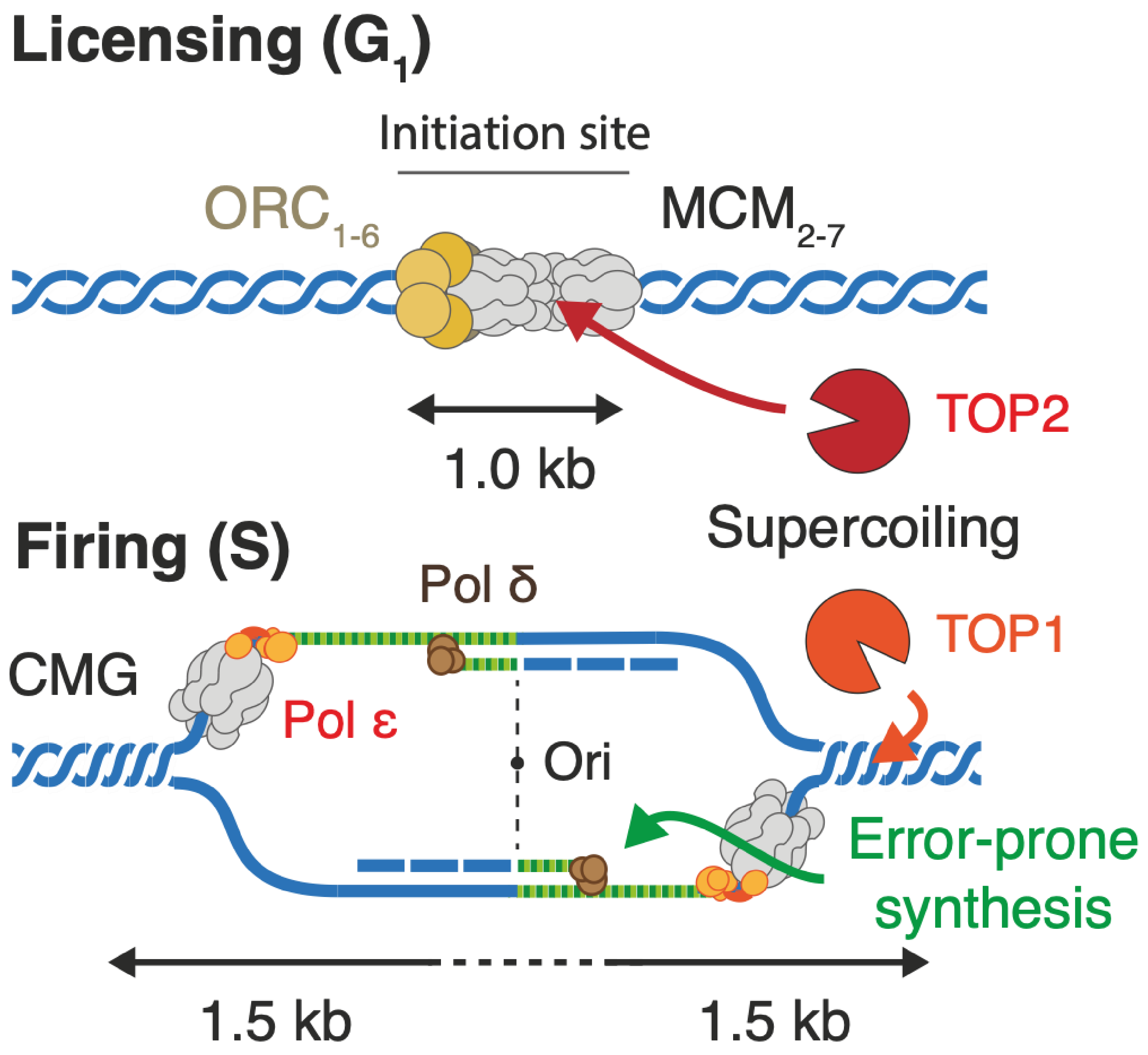


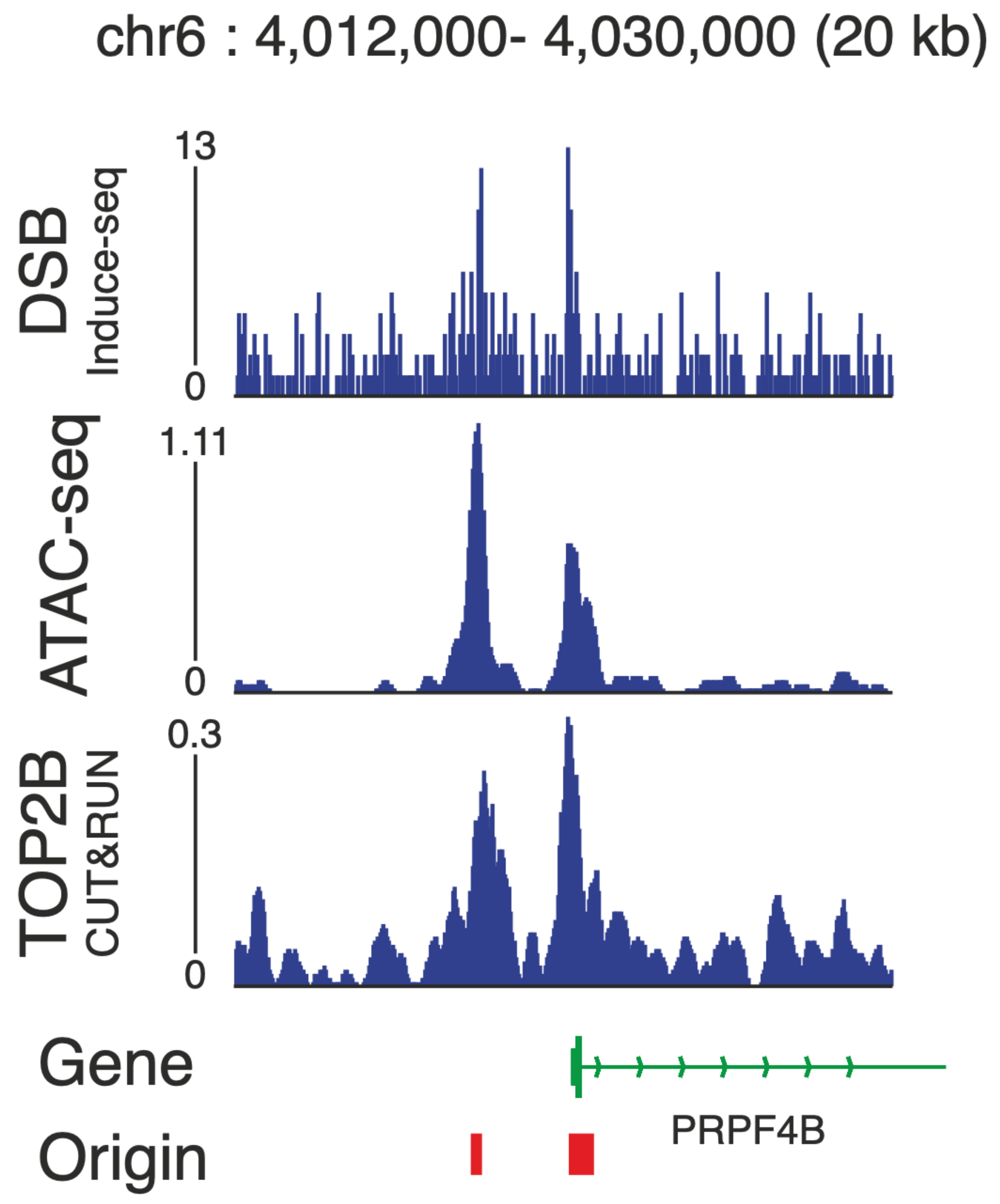


Murat et al., (2022) *Sci. Adv.* 8(45):eadd3686

# Ori SBS B reflects double strand breaks at replication origins

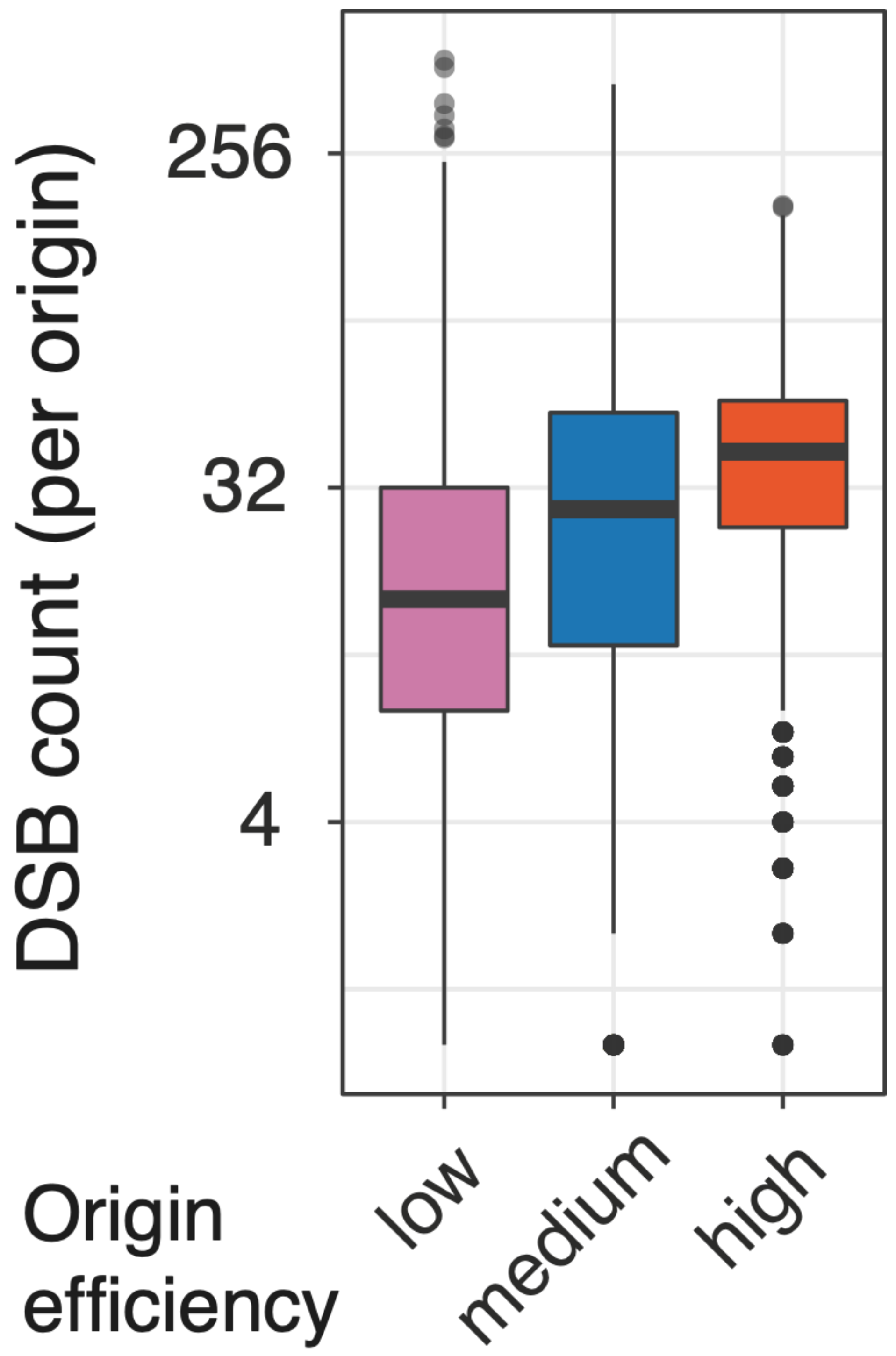






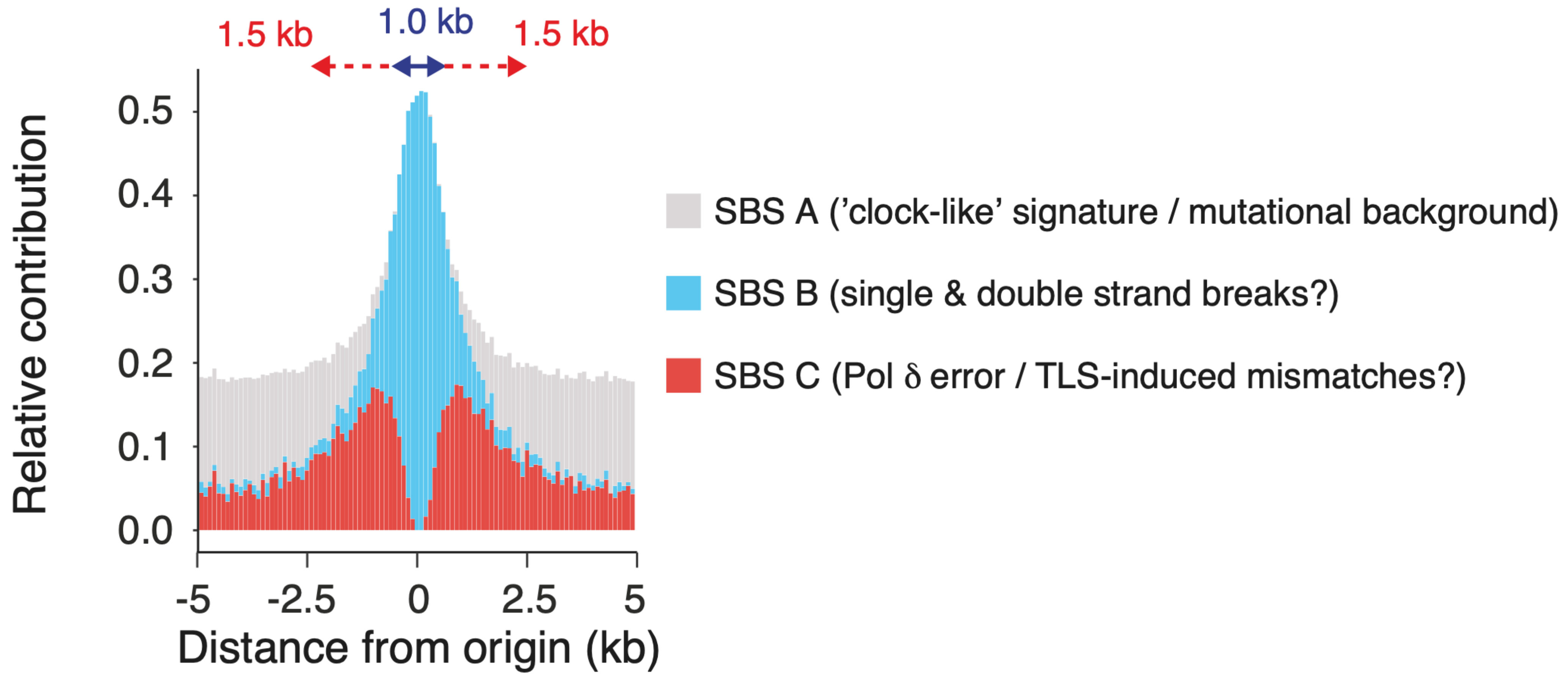
H9 hES cells

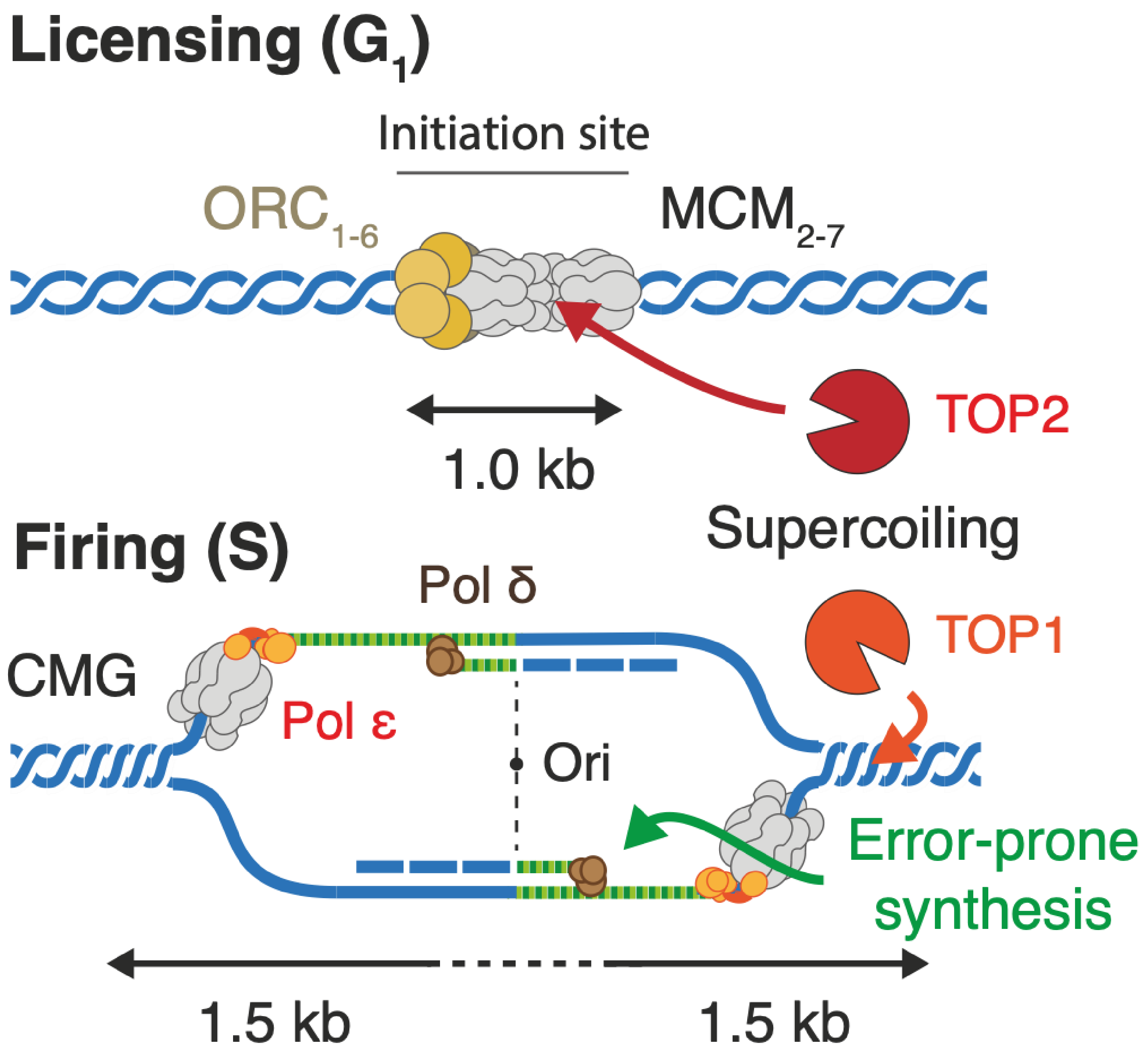
with Simon Reed & Pat van Eijk
INDUCE-seq: Dobbs et al. (2022), Nat Comms13:3989

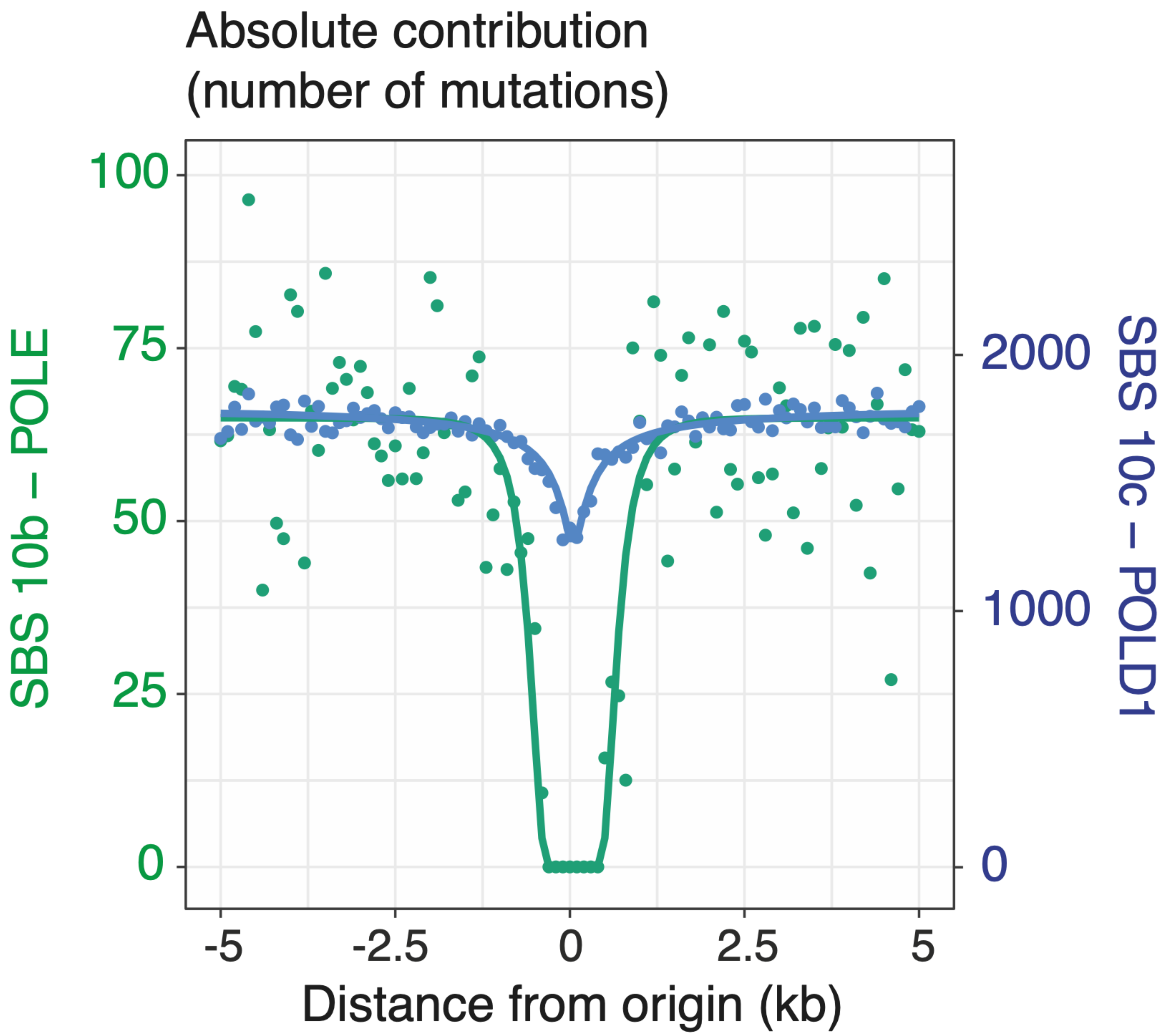


Murat et al., (2022) *Sci. Adv.* 8(45):eadd3686

# Ori SBS C: Error-prone DNA synthesis in the vicinity of origins: exclusion of Pol ε?

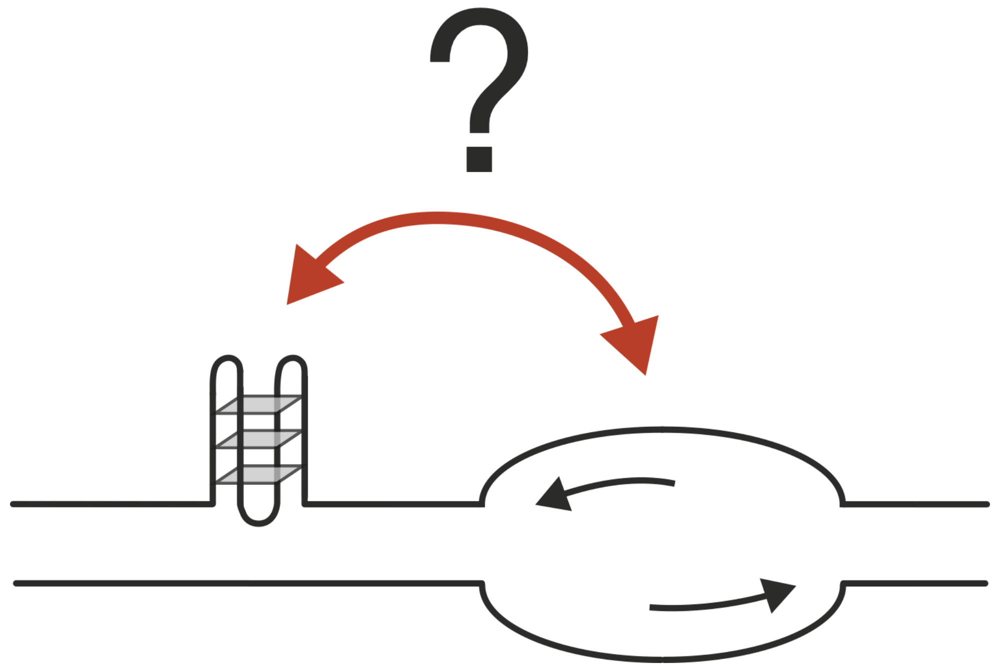




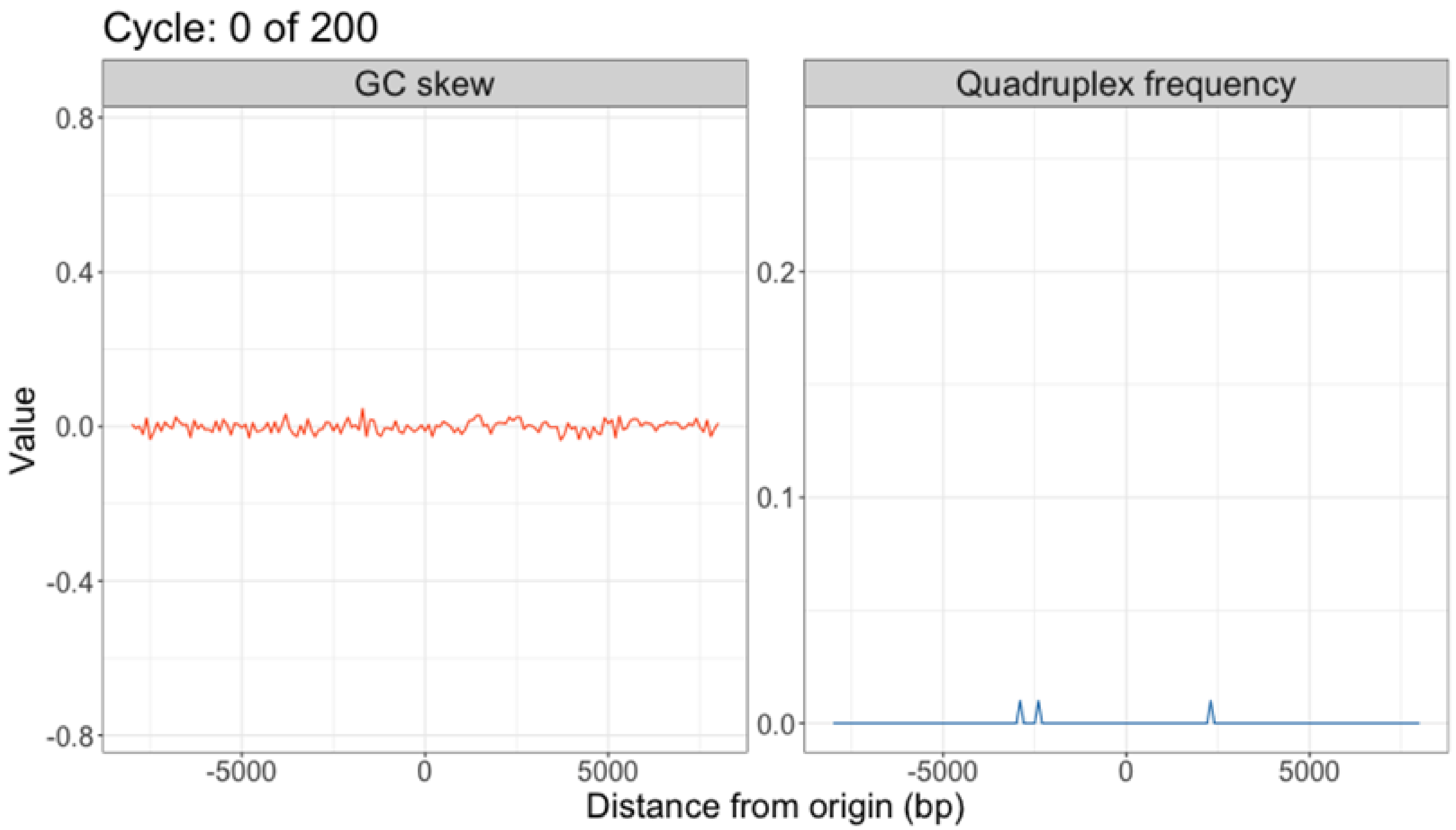


Murat et al., (2022) *Sci. Adv.* 8(45):eadd3686

# The mutagenic signatures of replication origins are sufficient to create the sequence environment observed at efficient origins

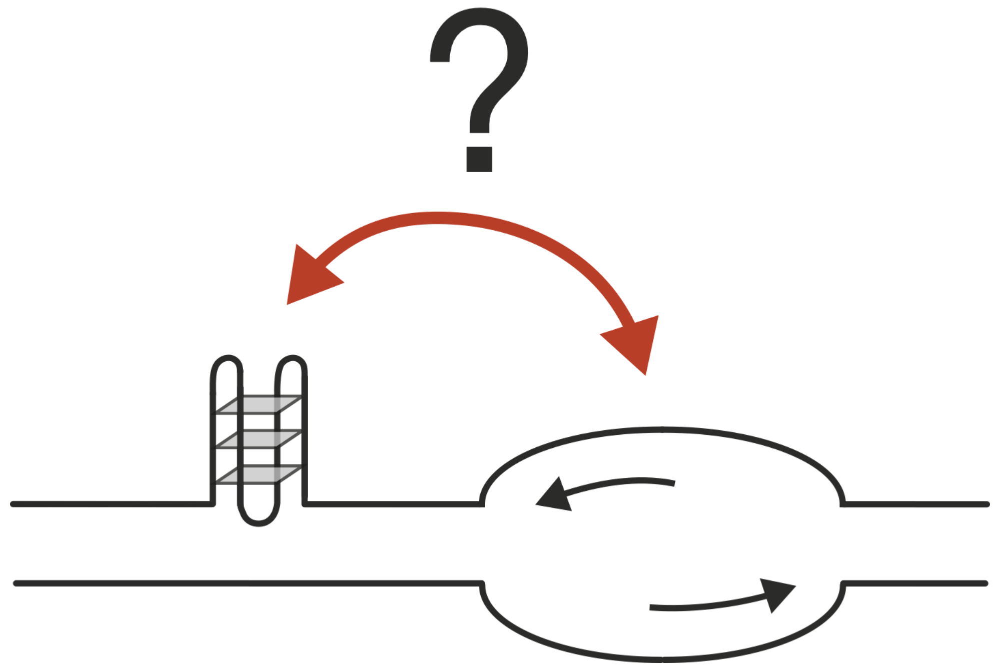


The origin of origins…
Why are G4s associated with replication origins?



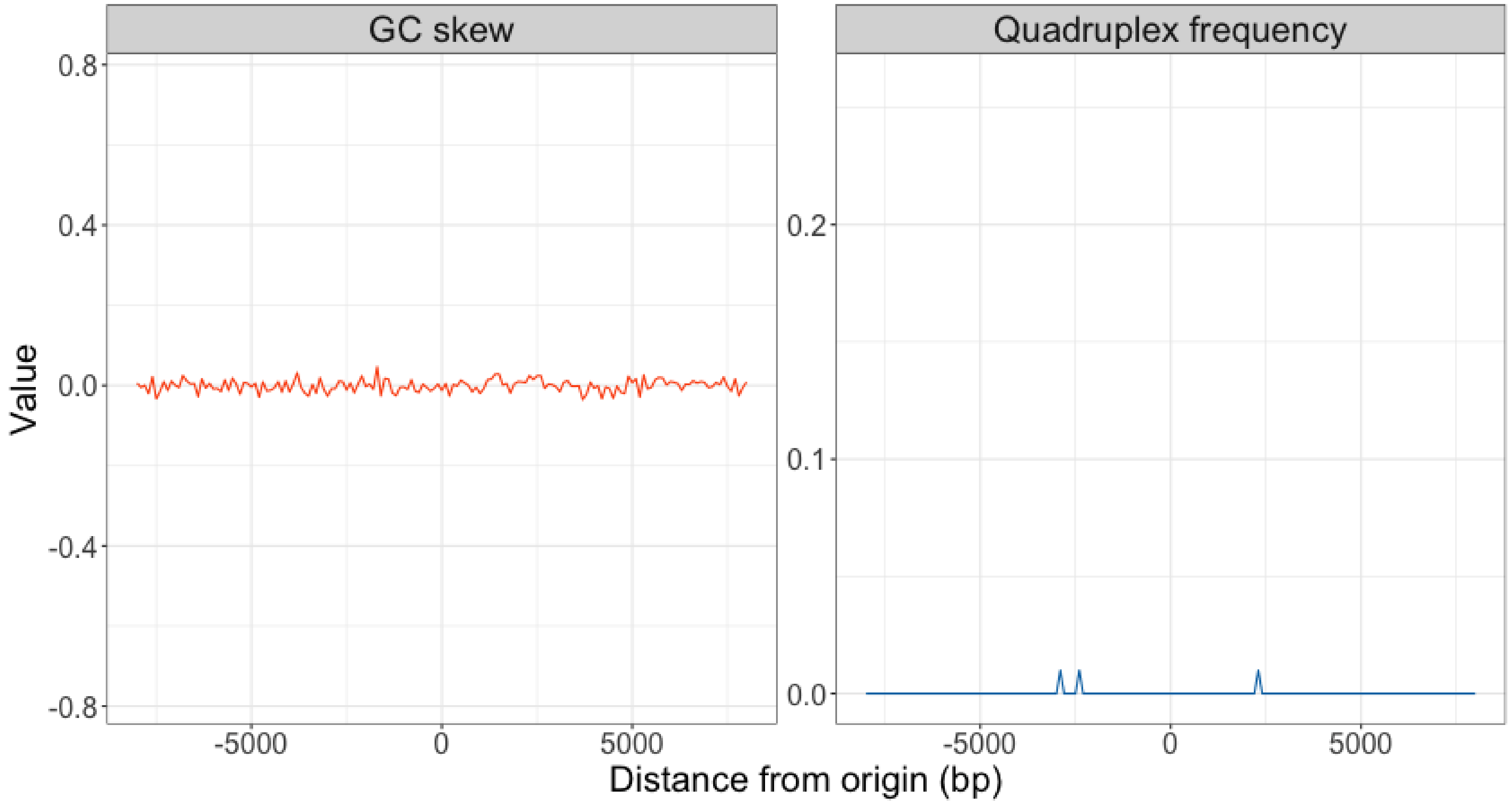
Murat et al., (2022) *Sci. Adv.* 8(45):eadd3686

# The mutagenic signatures of replication origins are sufficient to create the sequence environment observed at efficient origins



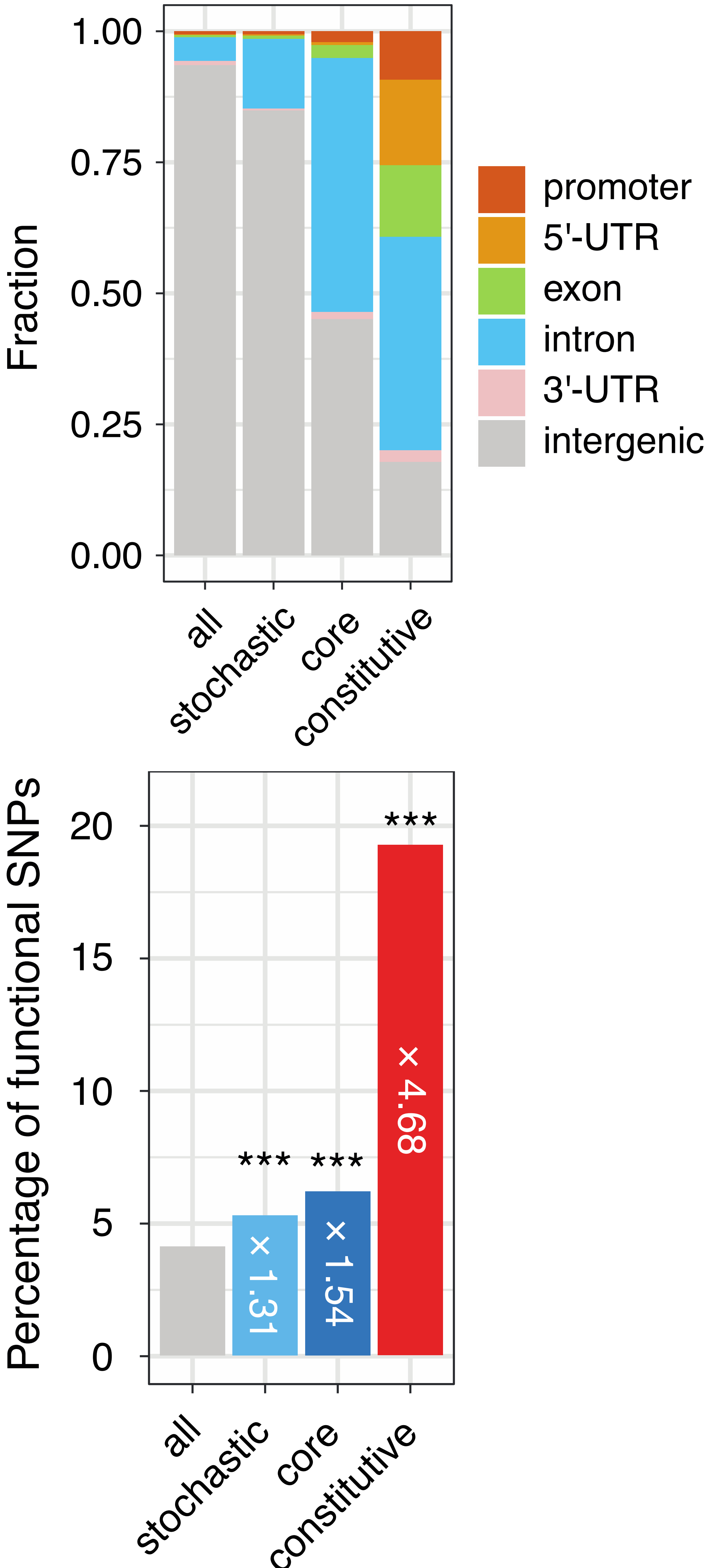
The origin of origins…
Why are G4s associated with replication origins?

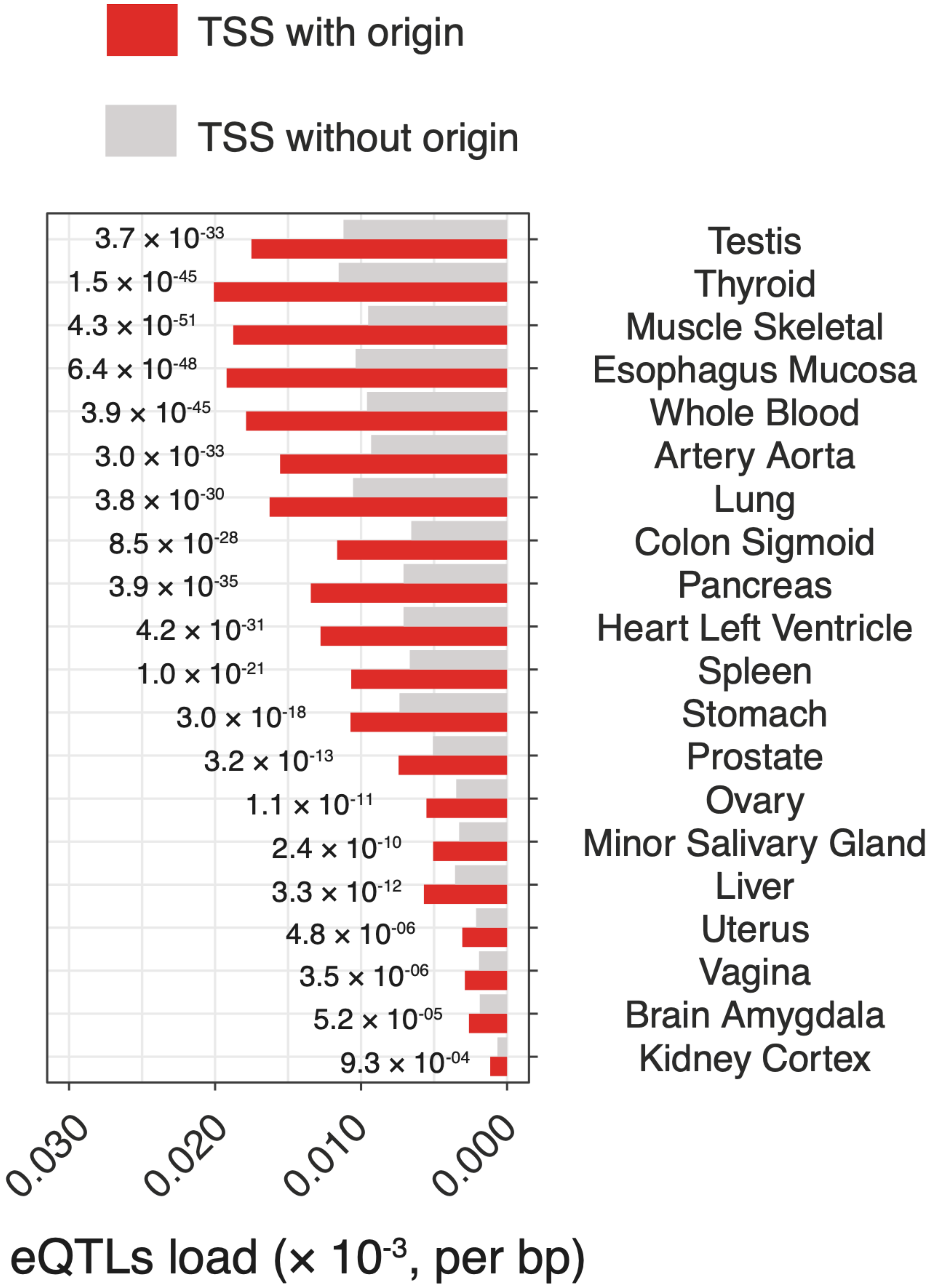
Cycle: 0 of 200

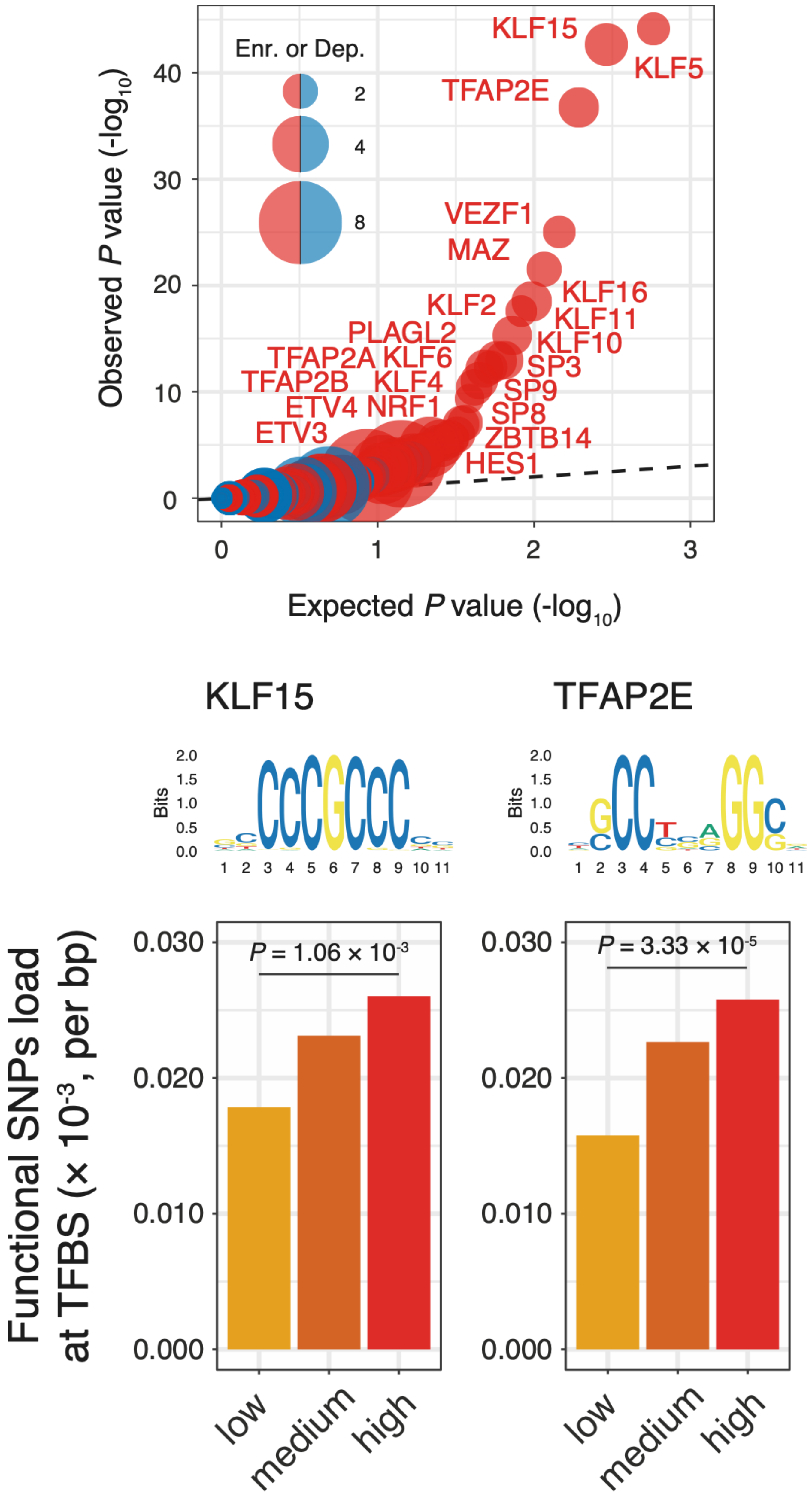


Murat et al., (2022) *Sci. Adv.* 8(45):eadd3686

# Constitutive origins are more likely to generate functionally important mutations







Murat et al., (2022) *Sci. Adv.* 8(45):eadd3686

# Summary

- Short repeat sequences with structure-forming potential are common in vertebrate genomes and have the capacity to trip up DNA replication

- Repriming is deployed frequently during replication of these sequences suggesting that they readily form replication impediments (and this can be promoted by RNA:DNA hybrid formation)

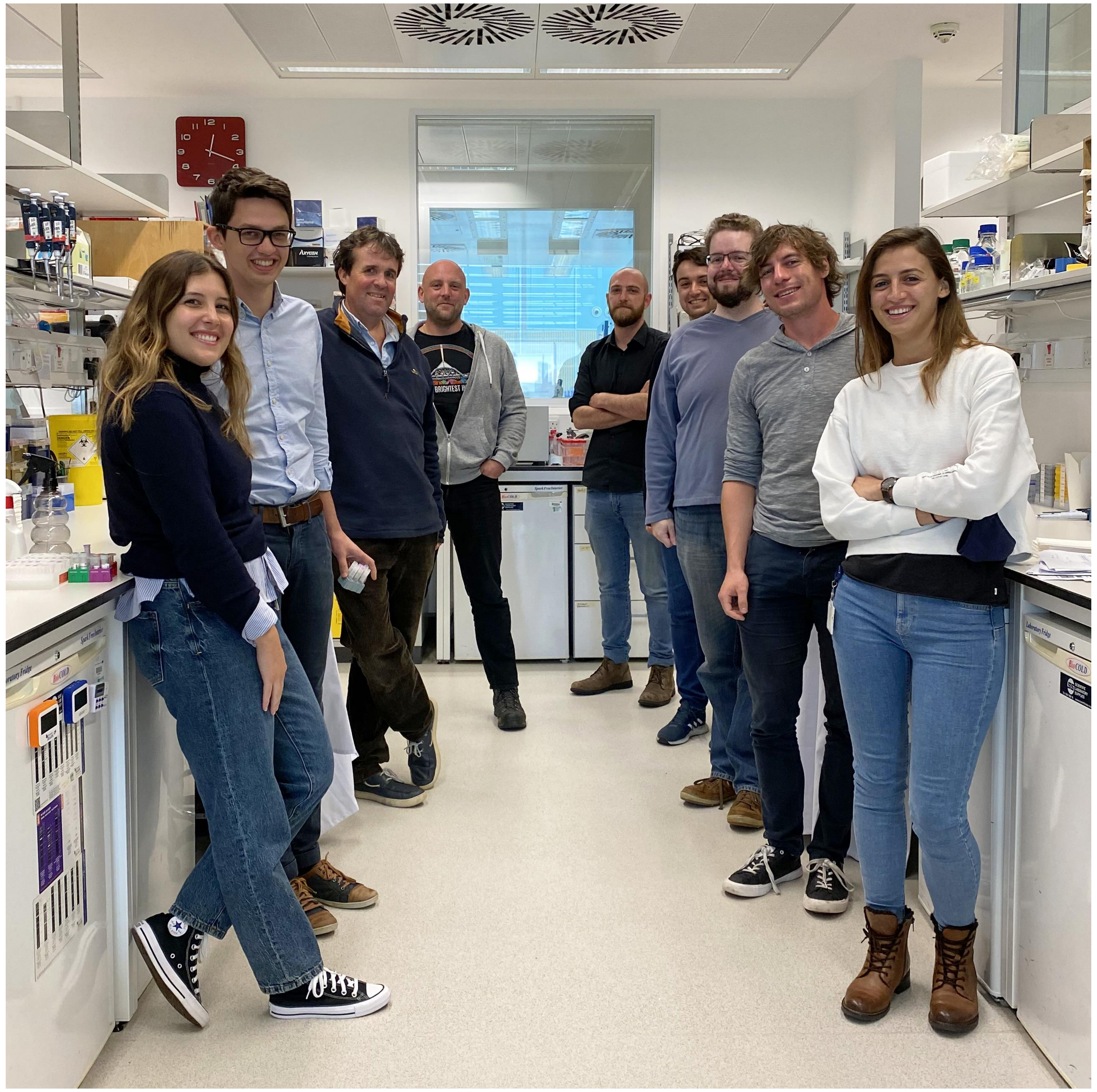
- The fork protection component Timeless links recognition of G4s via a novel DNA binding domain with recruitment of the DDX11 helicase

- The response of a model replicative polymerase to structure forming sequences *in vitro* makes powerful predictions about STR behaviour in genomes

- The identification of highly efficient sites of replication initiation has allowed the detection of replication origin dependent mutagenesis

- Highly efficient replication origins create their own sequence environment, including G4s, and are positioned to exert a significant impact on genome evolution

# Acknowledgements



**Current Group**
Consuelo Perez, Jedrzej Jaworski, Guillaume Guilbaud, Pierre Murat, Chris Mellor, Alastair Crisp, Andrew Zeller, Joelle Nassar

**Contributing former group members**
Peter Sarkies, Harris Papadopoulou,
Davide Schiavone,
Leticia Koch Lerner, Saša Šviković

**Thanks also to…**

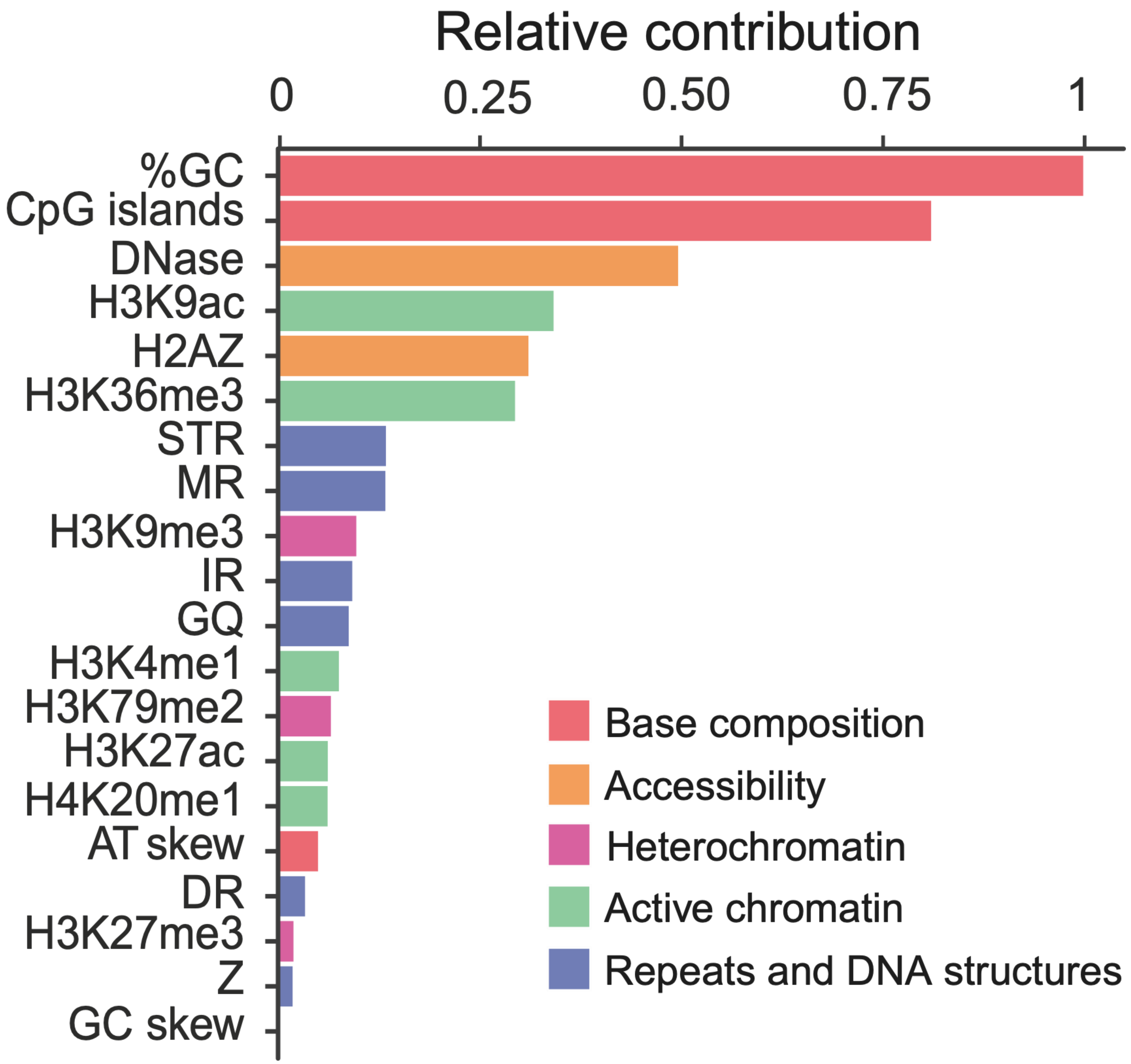
Luca Pellegrini & Sandro Holzer
(Department of Biochemistry, University of Cambridge)

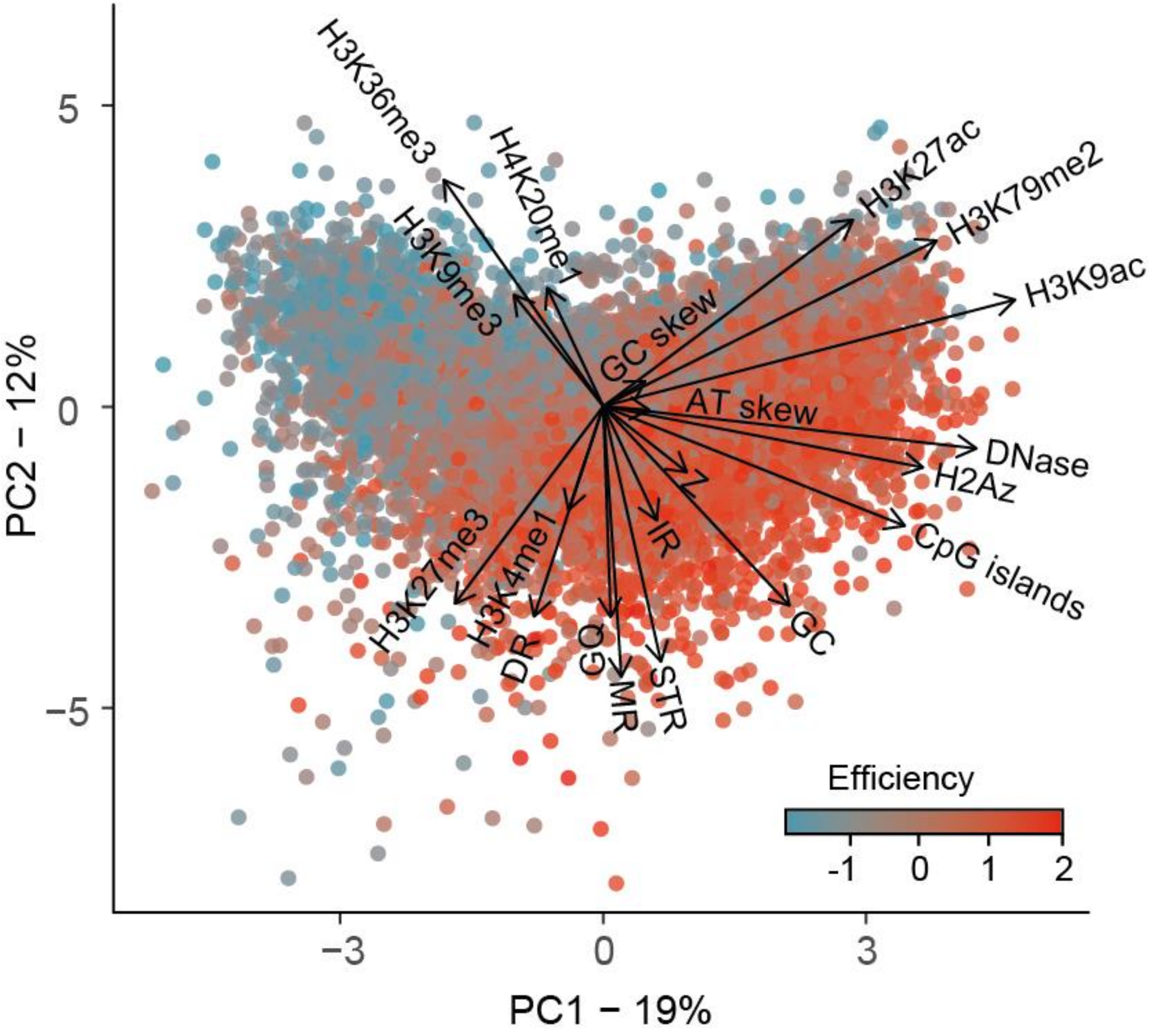
Torsten Krude
(Department of Zoology, University of Cambridge)

Simon Reed & Pat van Eijk
(University of Cardiff & Broken String Biosciences)

MRC Laboratory of Molecular Biology

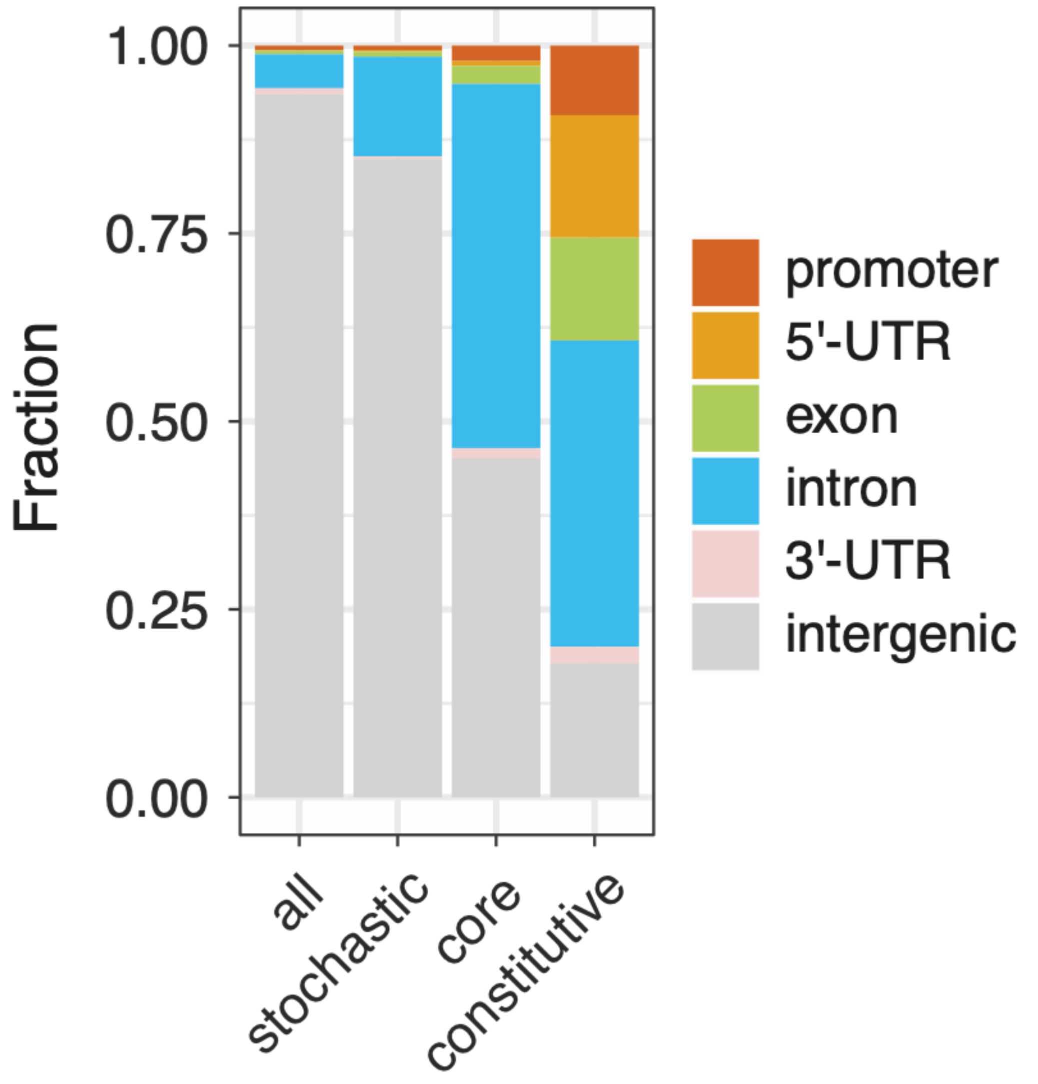
# Features of highly efficient human replication origins

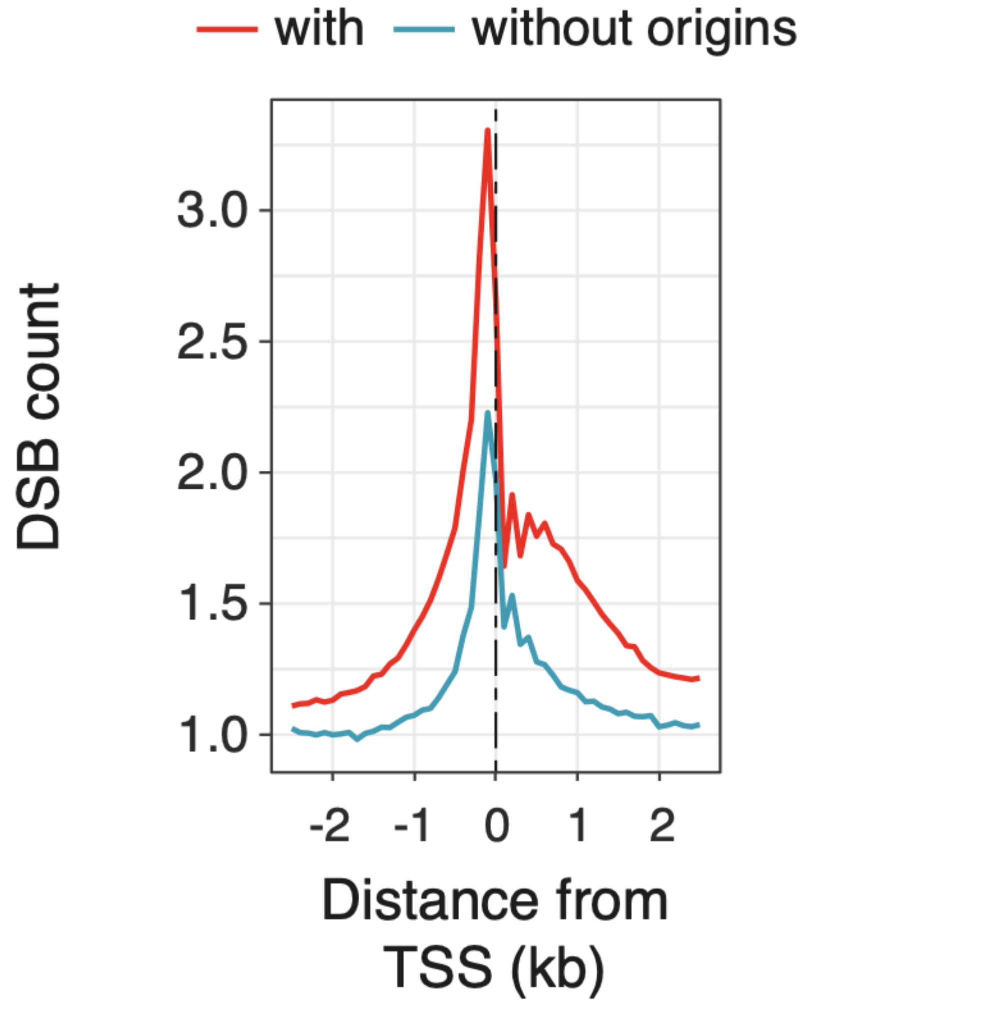


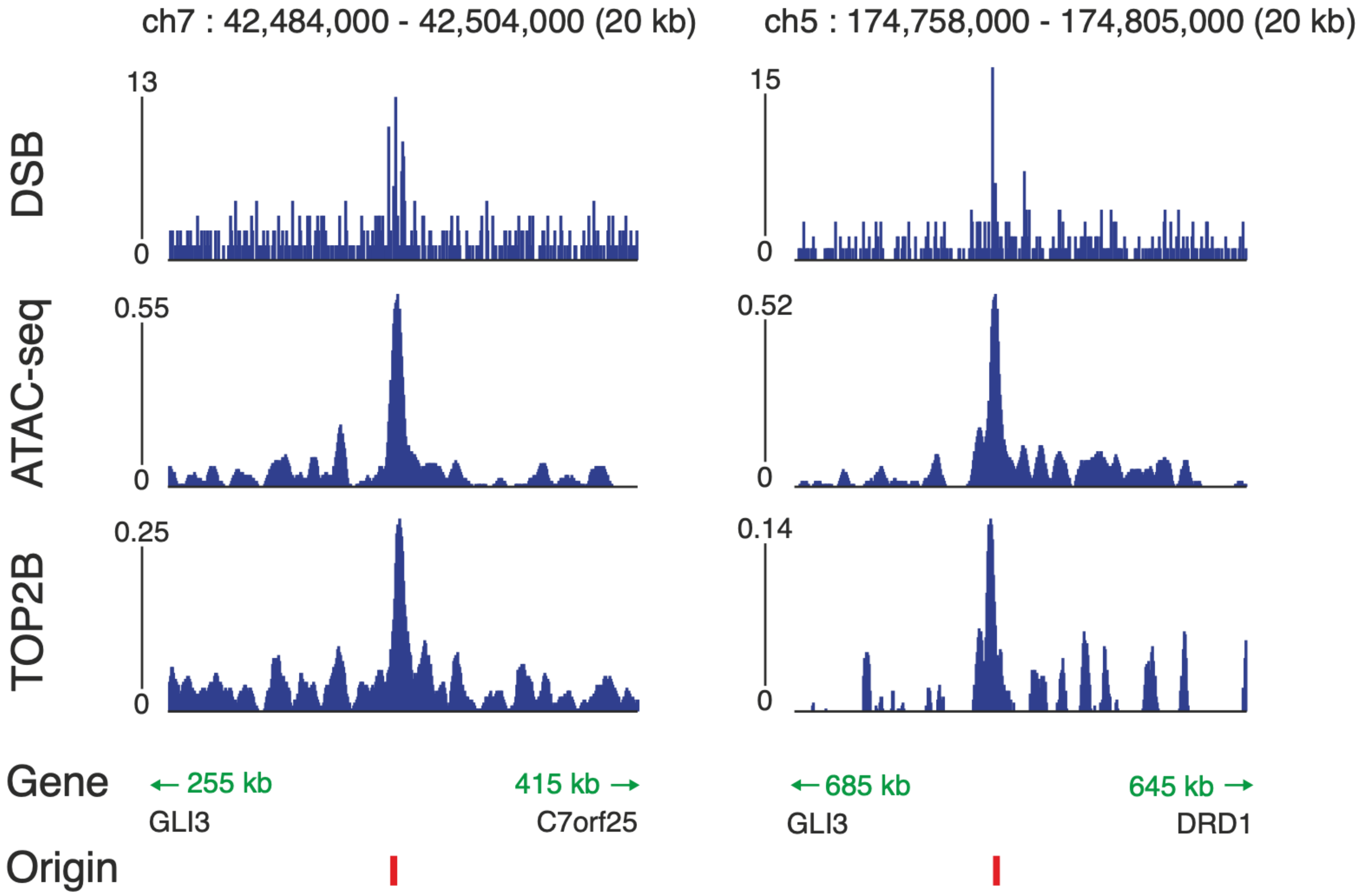


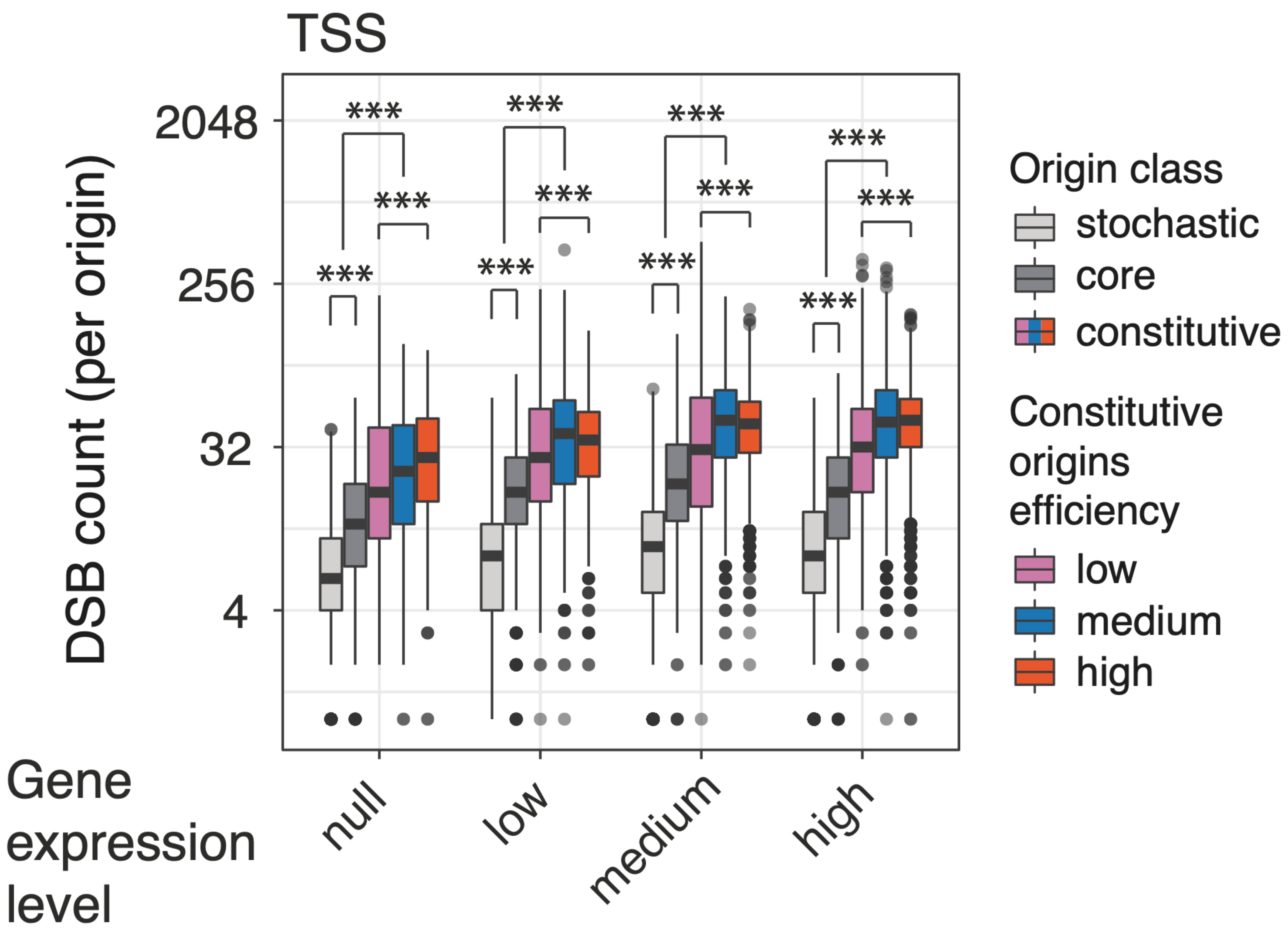
Guilbaud et al. (2022) *Nucleic Acids Res.* 50, 7436 - 7450

# Double strand breaks at origins occur independently of transcription









Increase in DSBs:
Ori: 5.1x
TSS: 1.9x

Murat et al., (2022) *Sci. Adv.* 8(45):eadd3686