Lecture 2

OriC finding

Aaron McKenna - 2023/4/5

Survey results

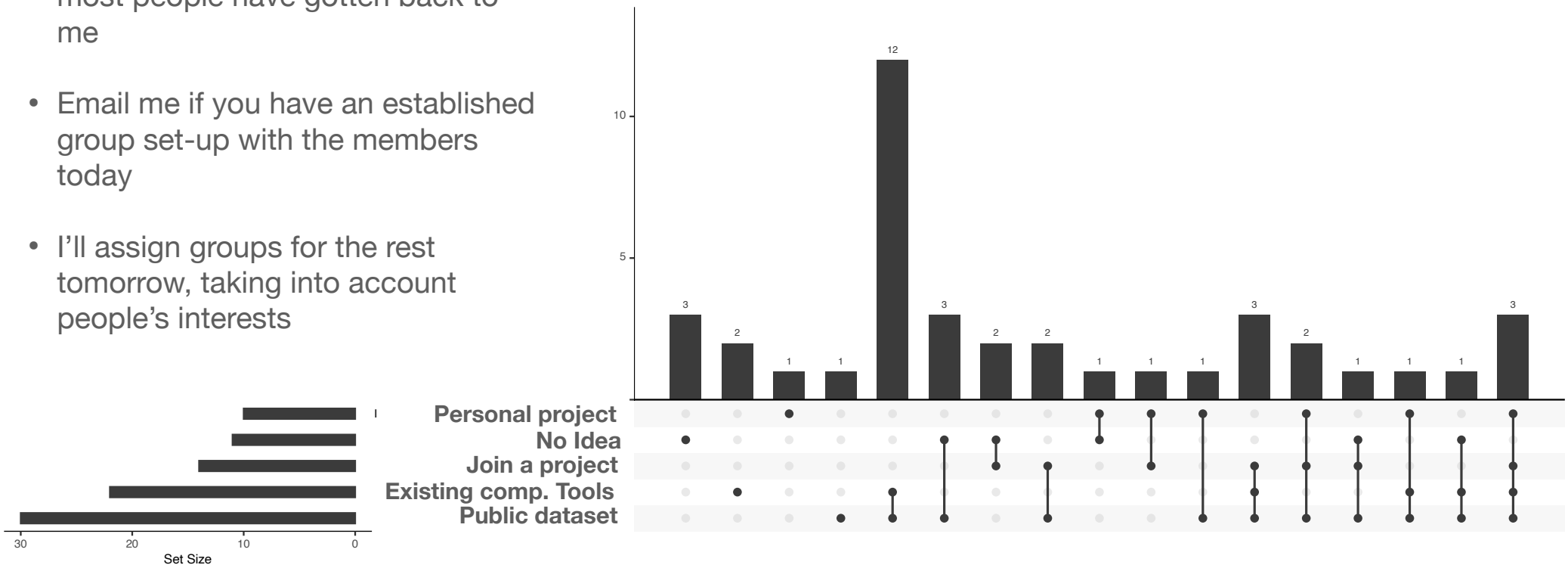
- We talked about this a little bit last time, but I got a chance to read through all the responses
- Lots of interest in RNA sequencing and single-cell sequencing, the needed tools / applications
- Good feedback on the grades, lots of thoughtful comments, everyone has a personal but generally well thought-out idea on passing
- Some interest in pipelines, specific tools, things we'd like to cover but are going to be stretched already. Projects can be a place to fit this in

Schedule (draft)

Month	Day	Room	Class #	Content
March	28	-	1	I'm out of town, no class
	30	Kellogg 100	2	Overview of the class, 'the semi-flipped experiment', final project discussions. python notebooks, genomic structure and information content, genome assembly
April	4	Kellogg 100	3	chapter 1: finding enriched sequences, motif finding, kmers, regulatory sequences
	6	Kellogg 100	4	chapter 2: In-class entropy and hidden messages
	11	Kellogg 100	5	chapter 2: probabilistic motif finding, gibbs sampling
	13	**No class**	6	~~chapter 2: In-class motif 'thought experiment'~~
	18	Kellogg 100	7	**chapter 3: How do we assemble genomes? -**
	20	Kellogg 200	8	**chapter 3: In class assembly exercise - maybe move down?**
	25	Kellogg 100	9	chapter 5: Aligning two sequences: Dynamic programming
	27	Kellogg 200	10	chapter 5: In-class walking around NYC exercises
May	2	Kellogg 100	11	**chapter 8/9: RNA sequencing, read mapping, counting, and enrichment - - maybe move up?**
	4	Kellogg 200	12	**chapter 8/9: In-class RNA sequencing experimental design exercise**
	9	Chilcott	13	chapter 8: clustering: RNA to identity
	11	Chilcott	14	chapter 8: In-class exploring k-means
	16	Kellogg 200	15	chapter 10: probabilistic modeling of hidden states using HMMs
	18	Vail 120 Auditorium	16	chapter 10: In-class HMM and CpG island exercise
	23	Vail 120 Auditorium	17	chapter 10: HMMs wrap-up, extensions to more complex models
	25	Chilcott	18	final project presentations
	30	Kellogg 200	19	final project presentations
June	1	Kellogg 100	20	final project presentations (if needed)

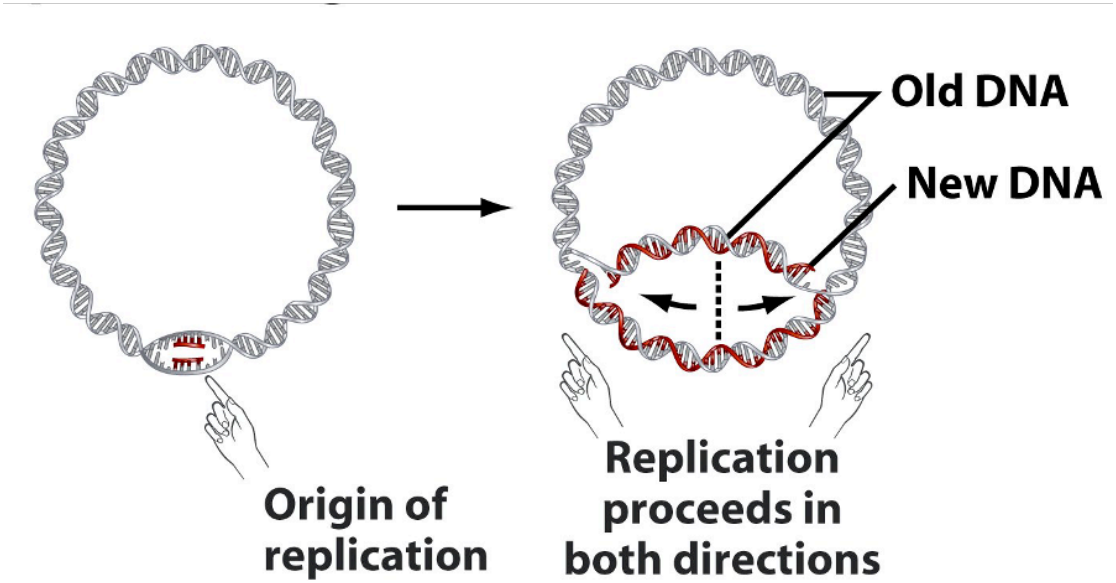
Group choices

- Contacted personal project people, most people have gotten back to me
- Email me if you have an established group set-up with the members today
- I'll assign groups for the rest tomorrow, taking into account people's interests



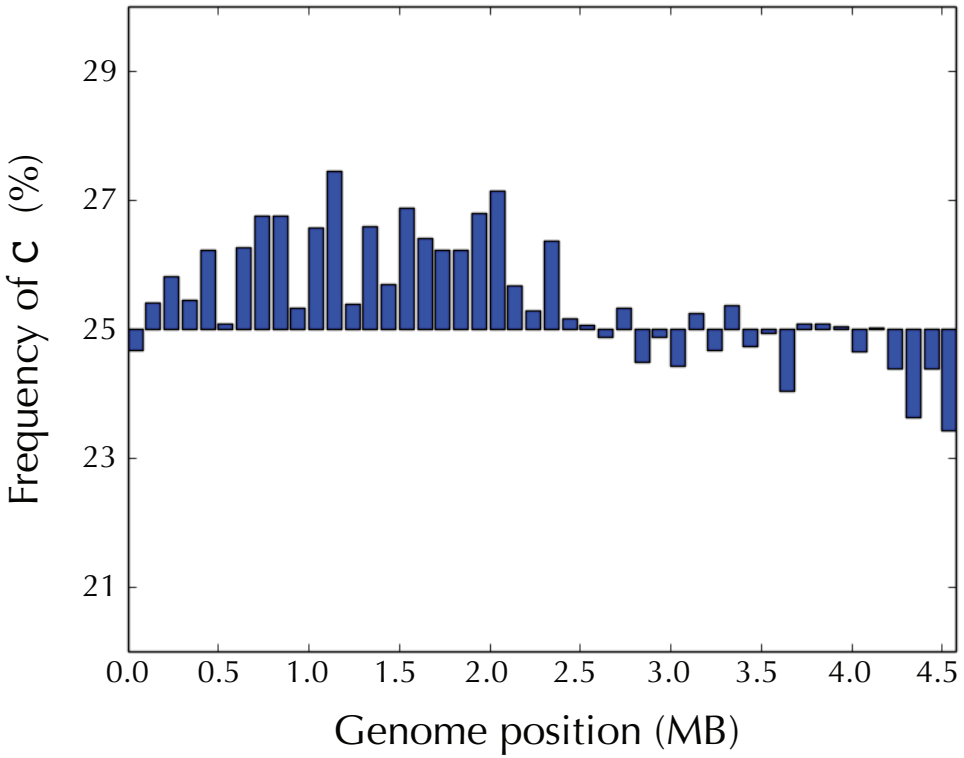
Where we left off

- We’re interested in solving a central problem for cells: where do you start replicating a genome



Where we left off

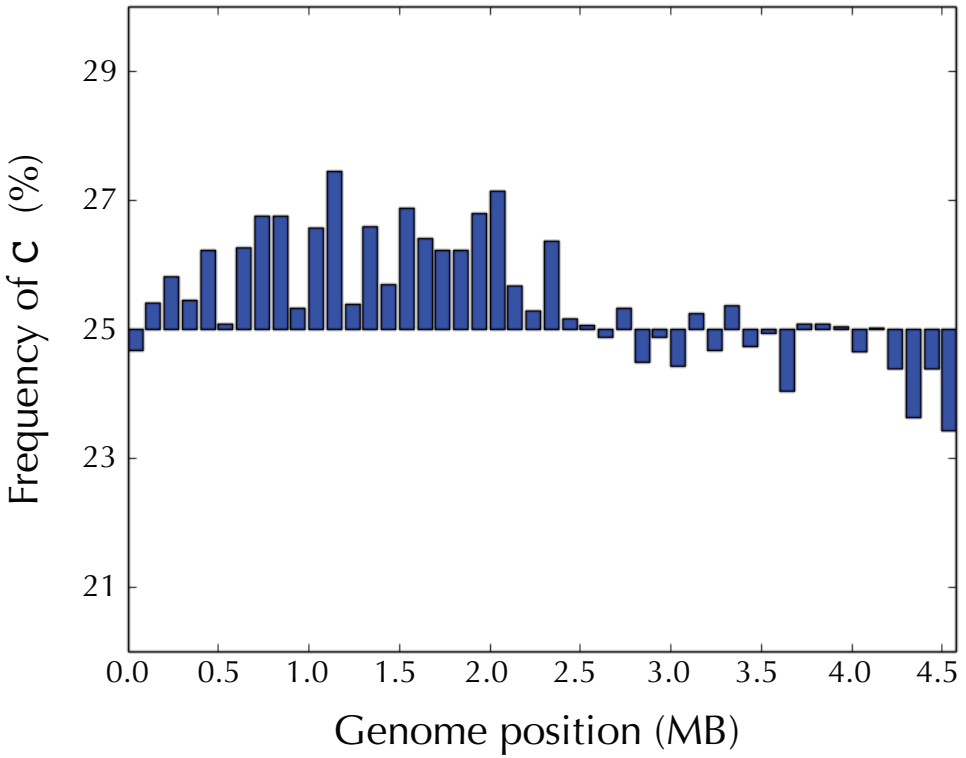
Let's run a very simple computational analysis: take frequency of each nucleotide in 100,000 nucleotide windows of *E. coli*



A Surprising Pattern in Nucleotide Counts

Let's run a very simple computational analysis: take frequency of each nucleotide in 100,000 nucleotide windows of *E. coli*

Why would there be more C on half the genome?

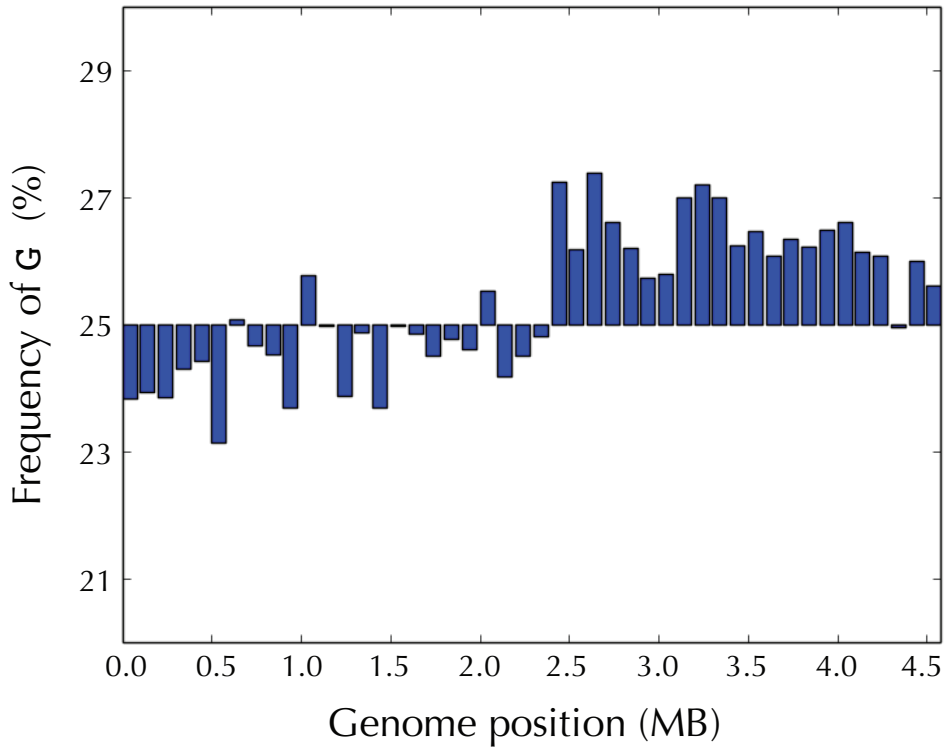


A Surprising Pattern in Nucleotide Counts

Let's run a very simple computational analysis: take frequency of each nucleotide in 100,000 nucleotide windows of *E. coli*

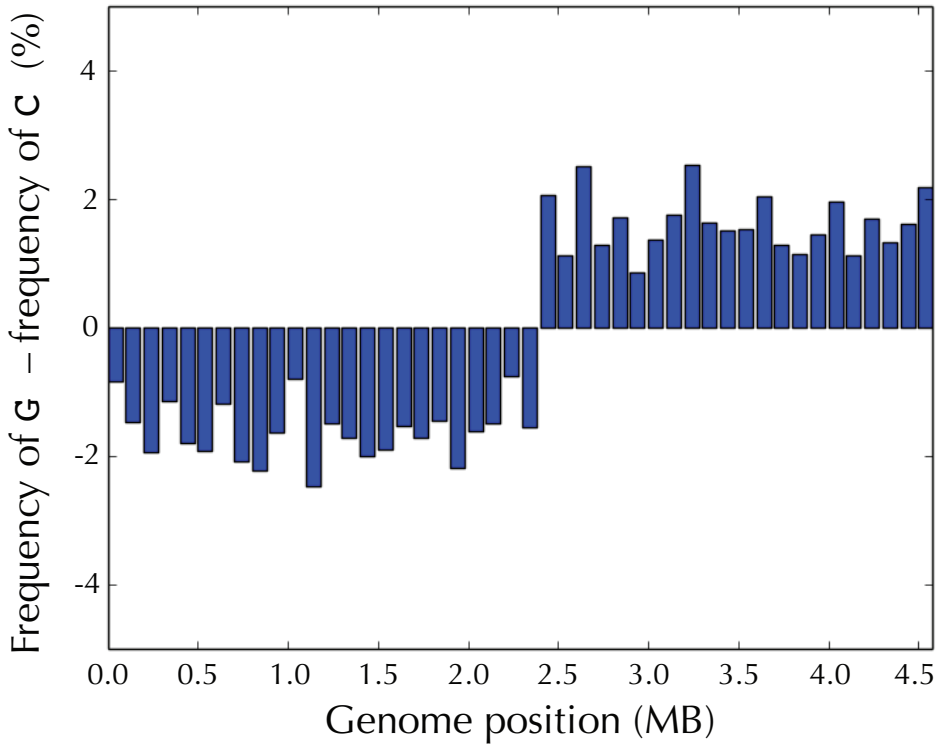
Why would there be more C on half the genome?

And why would the story be opposite when we count G's?



A Surprising Pattern in Nucleotide Counts

The pattern is even more stark if we take the difference between the frequency of G and the frequency of C

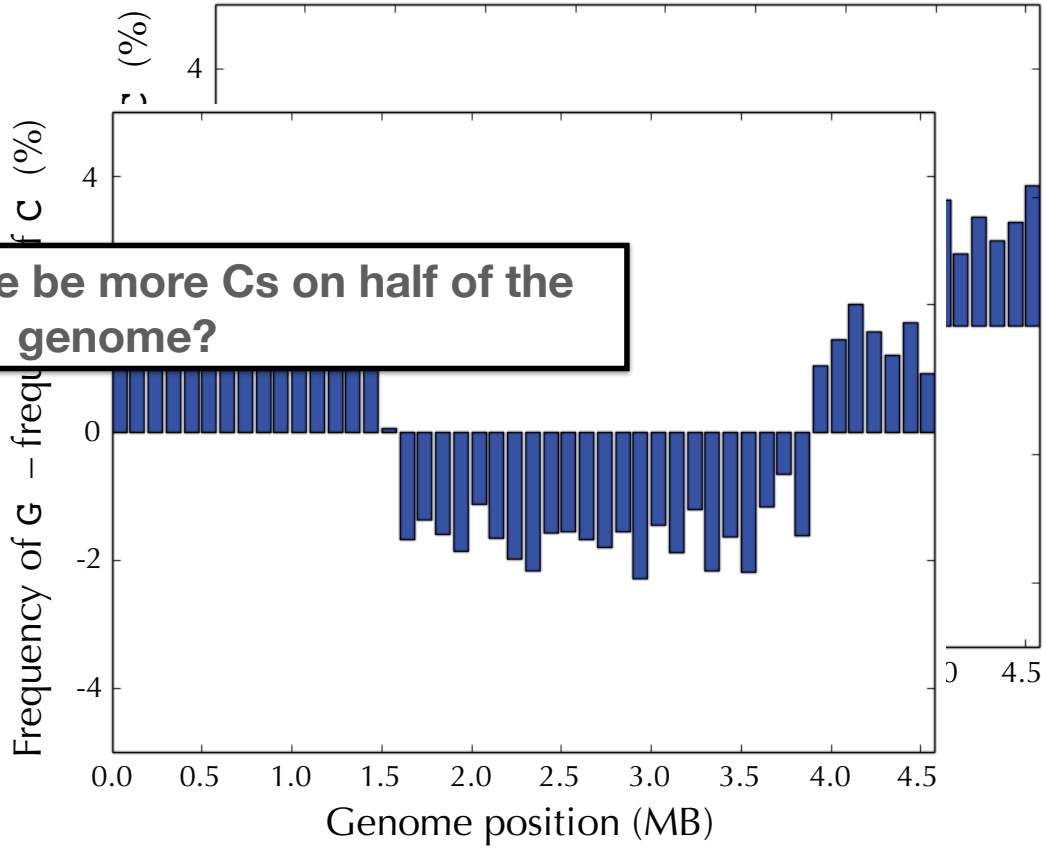


A Surprising Pattern in Nucleotide Counts

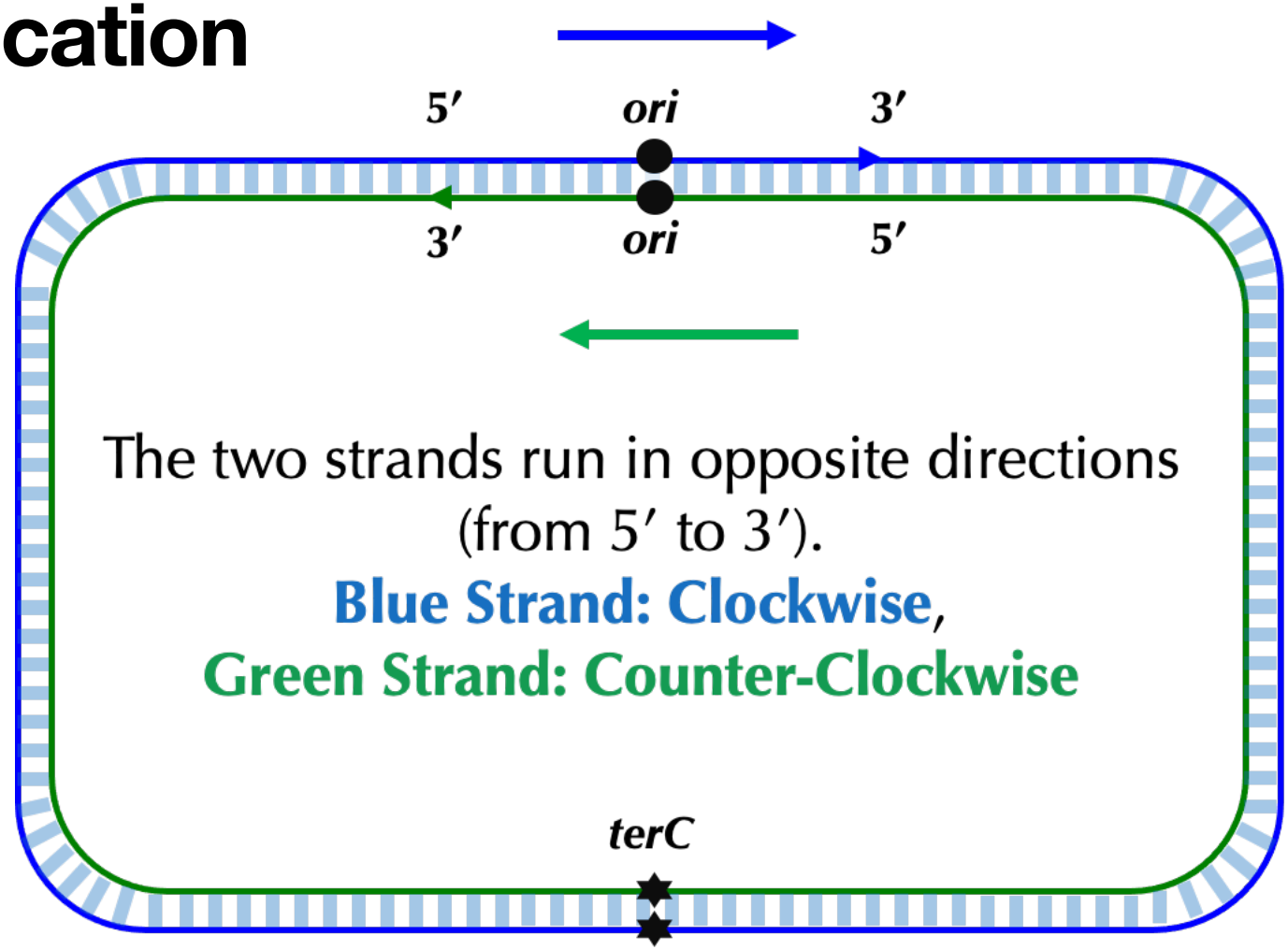
The pattern is even
take the difference b
frequency of G and

It's not an effect of where we start counting either or where the OriC is

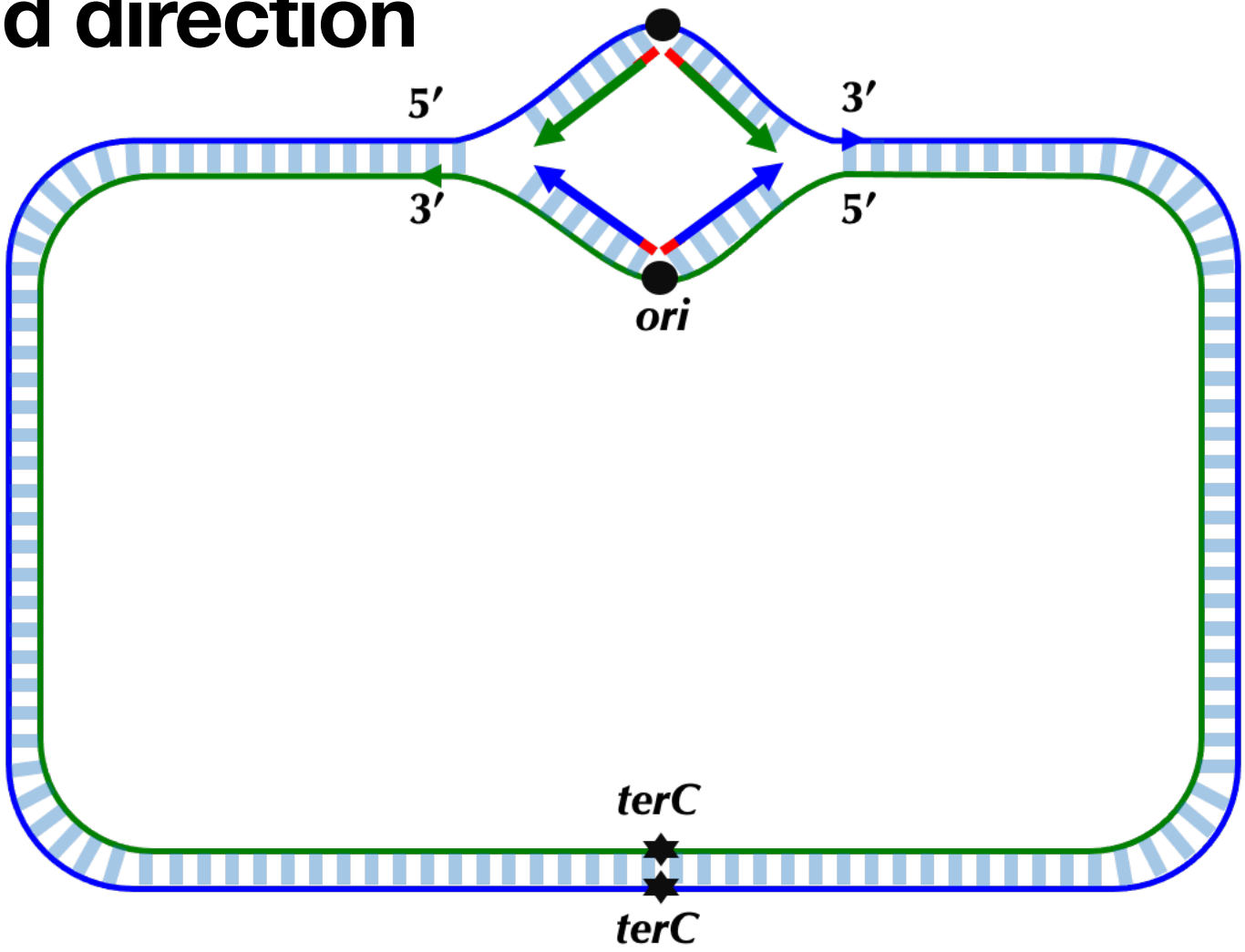
Why would there be more Cs on half of the genome?



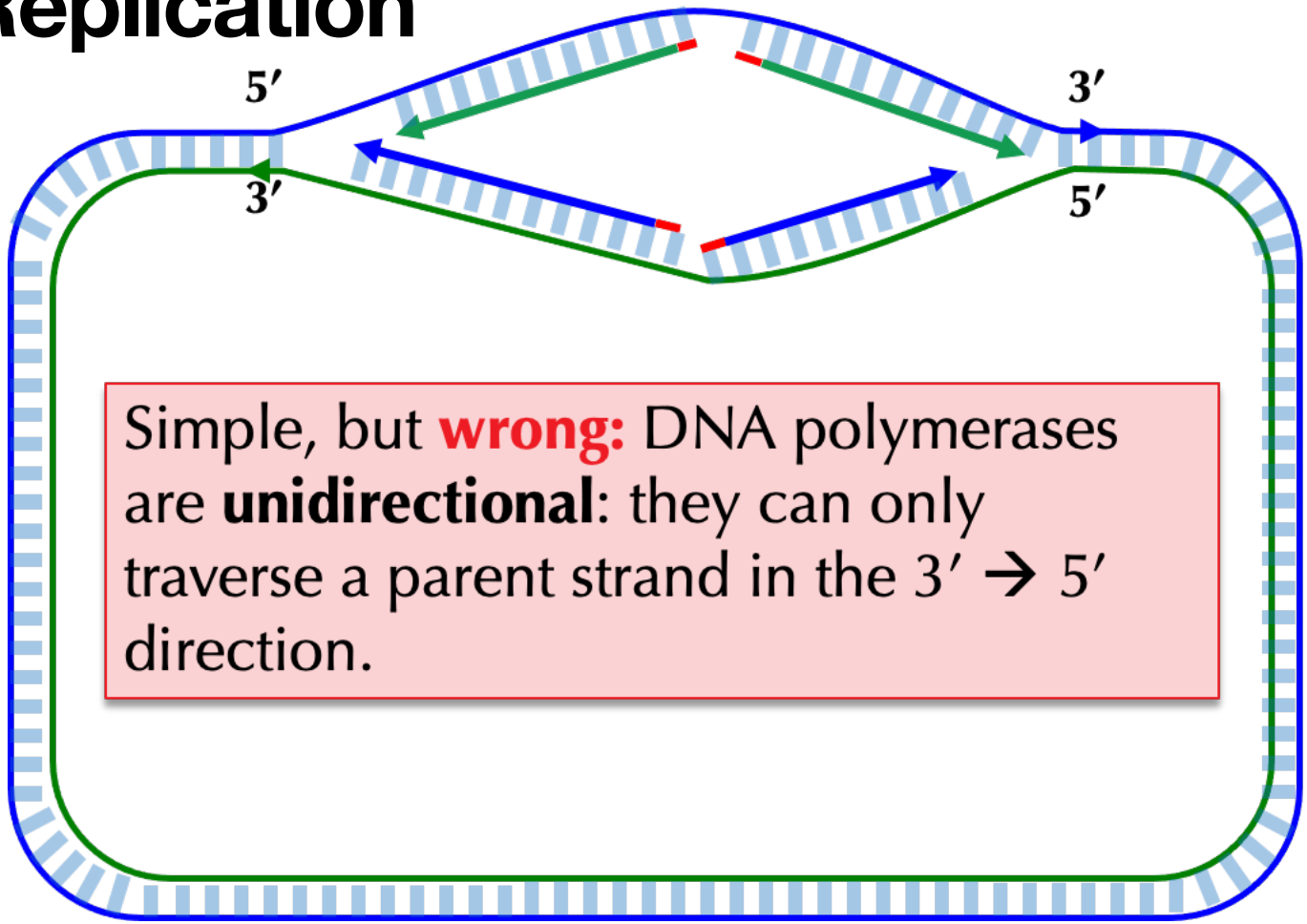
DNA Replication



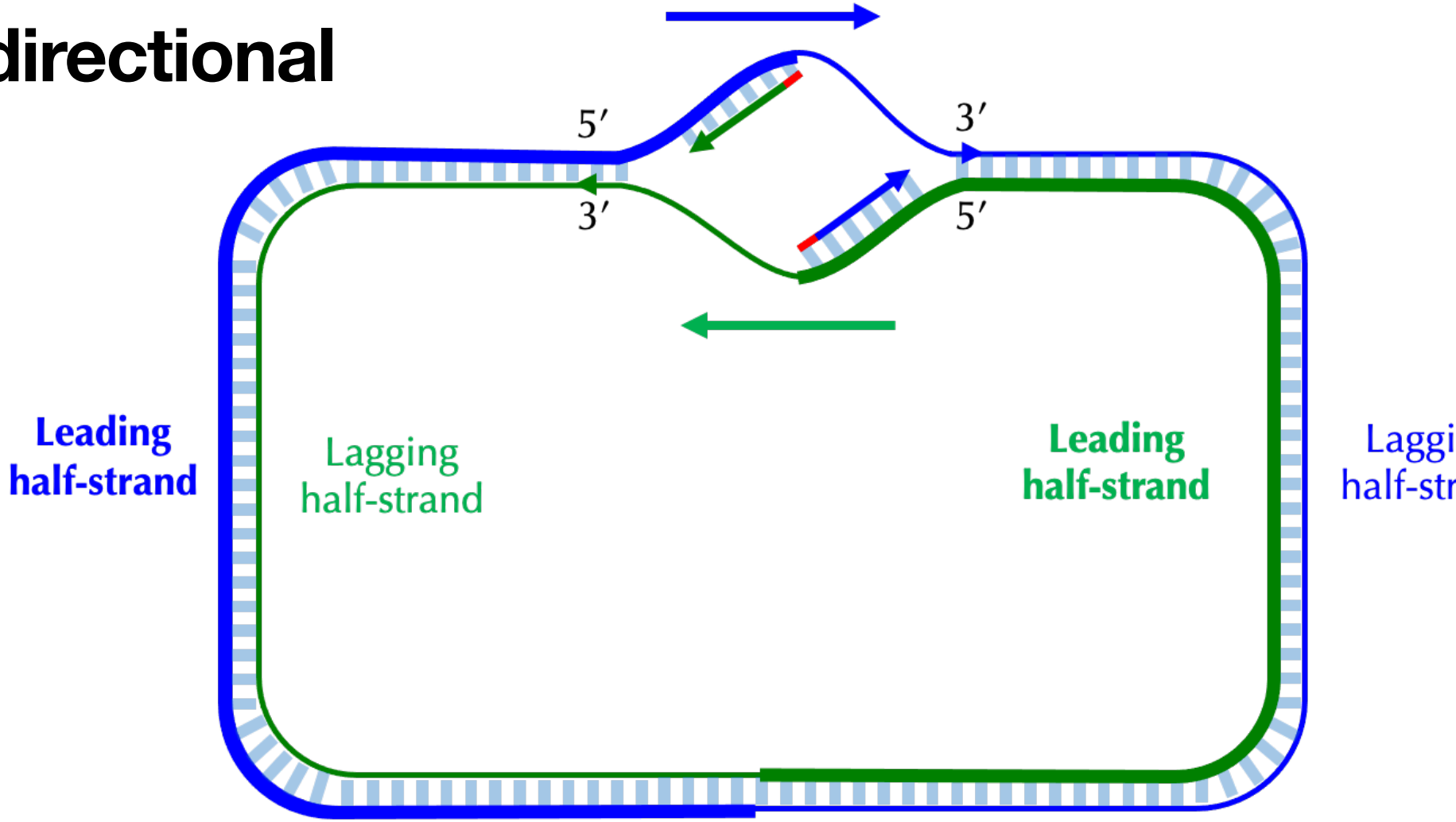
DNA Strand direction



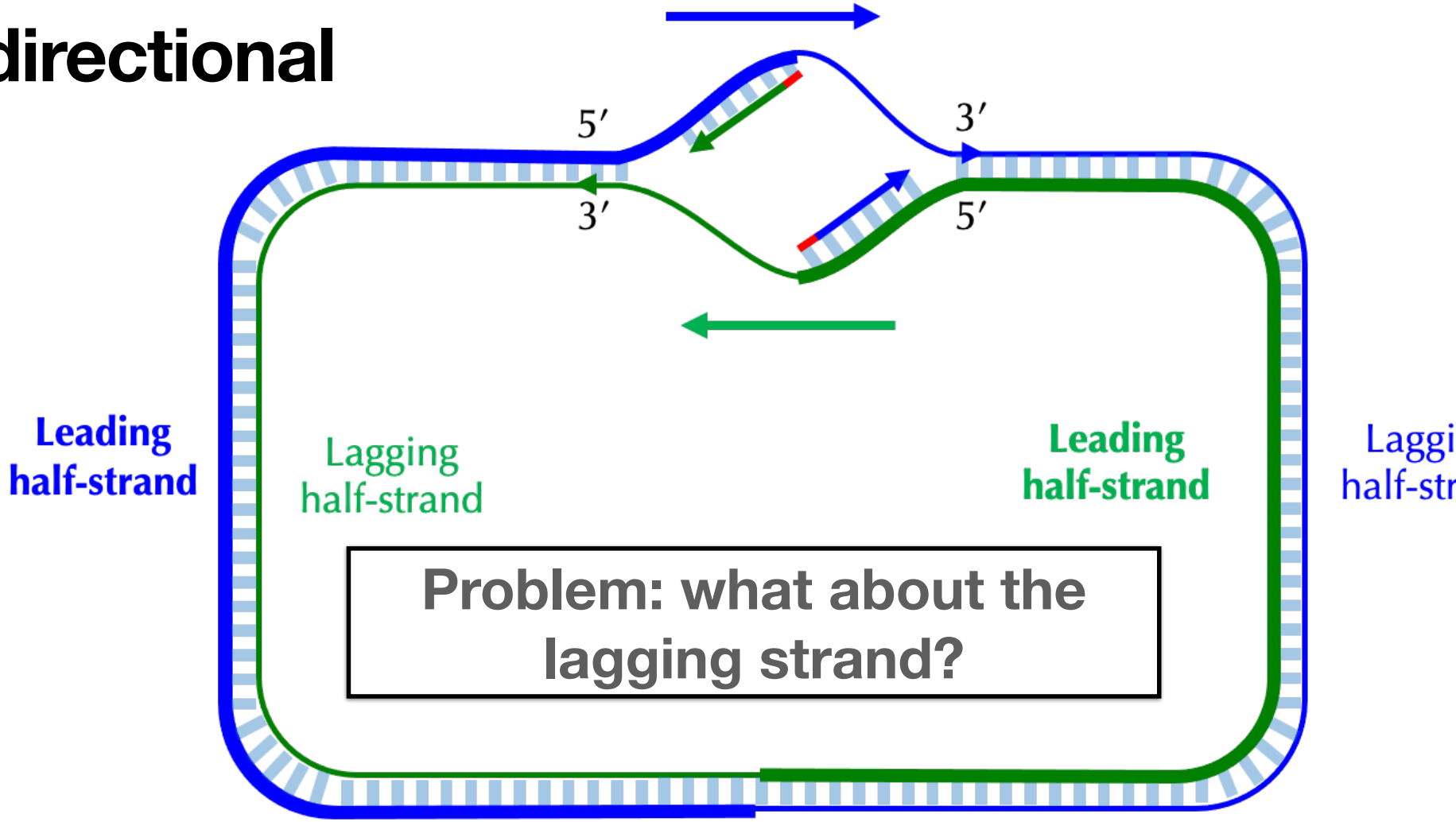
Not DNA Replication



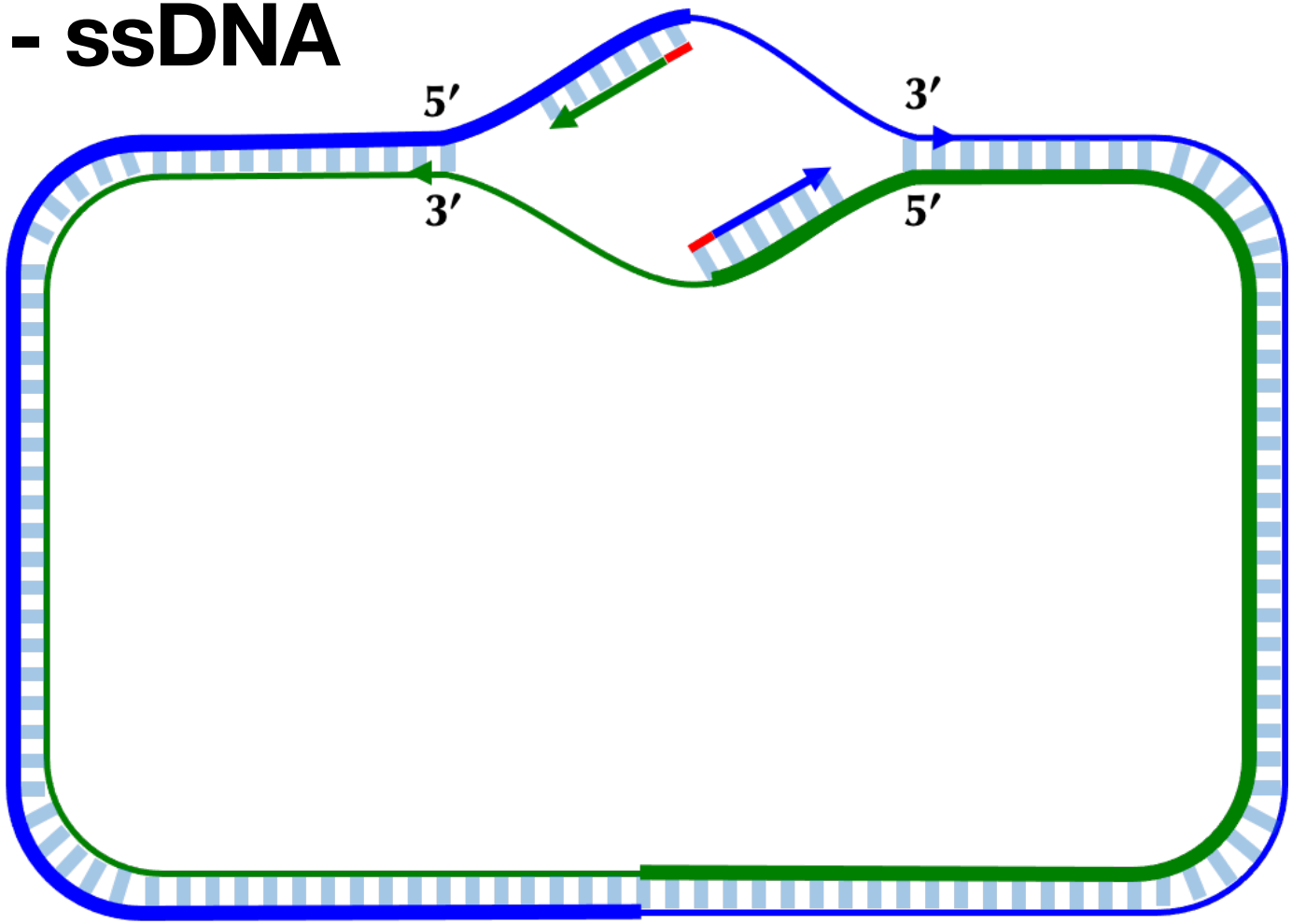
Unidirectional



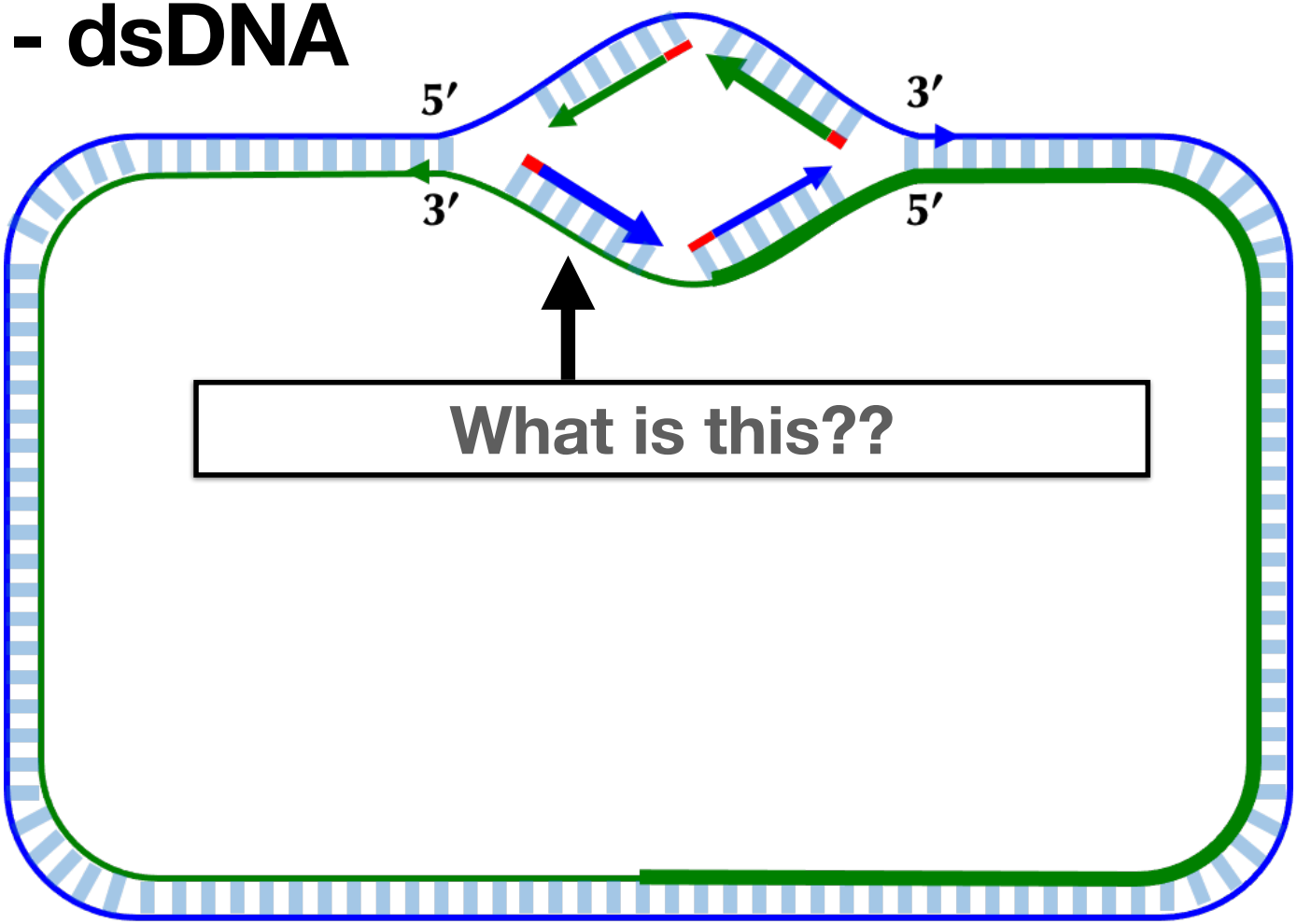
Unidirectional



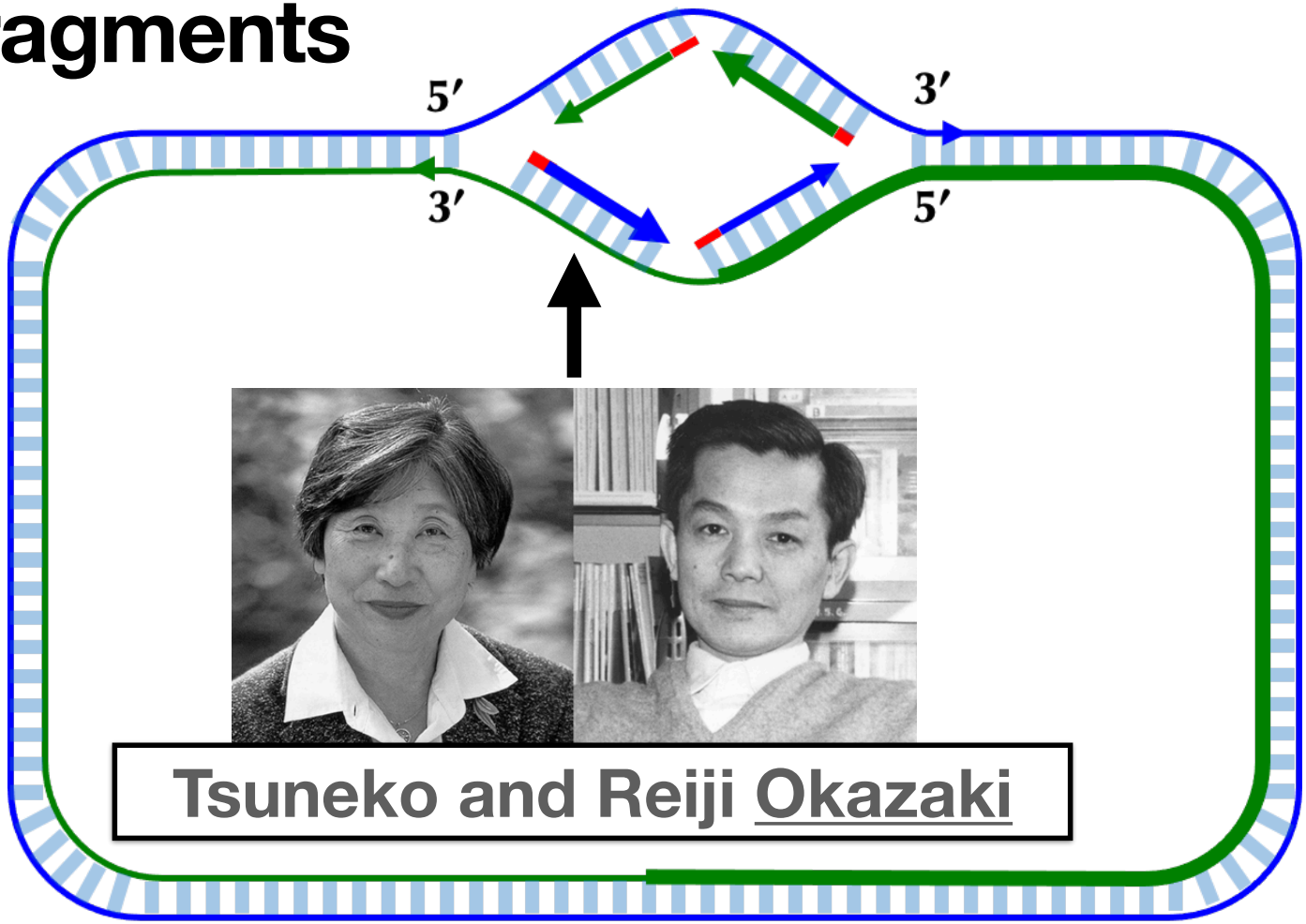
Open fork - ssDNA



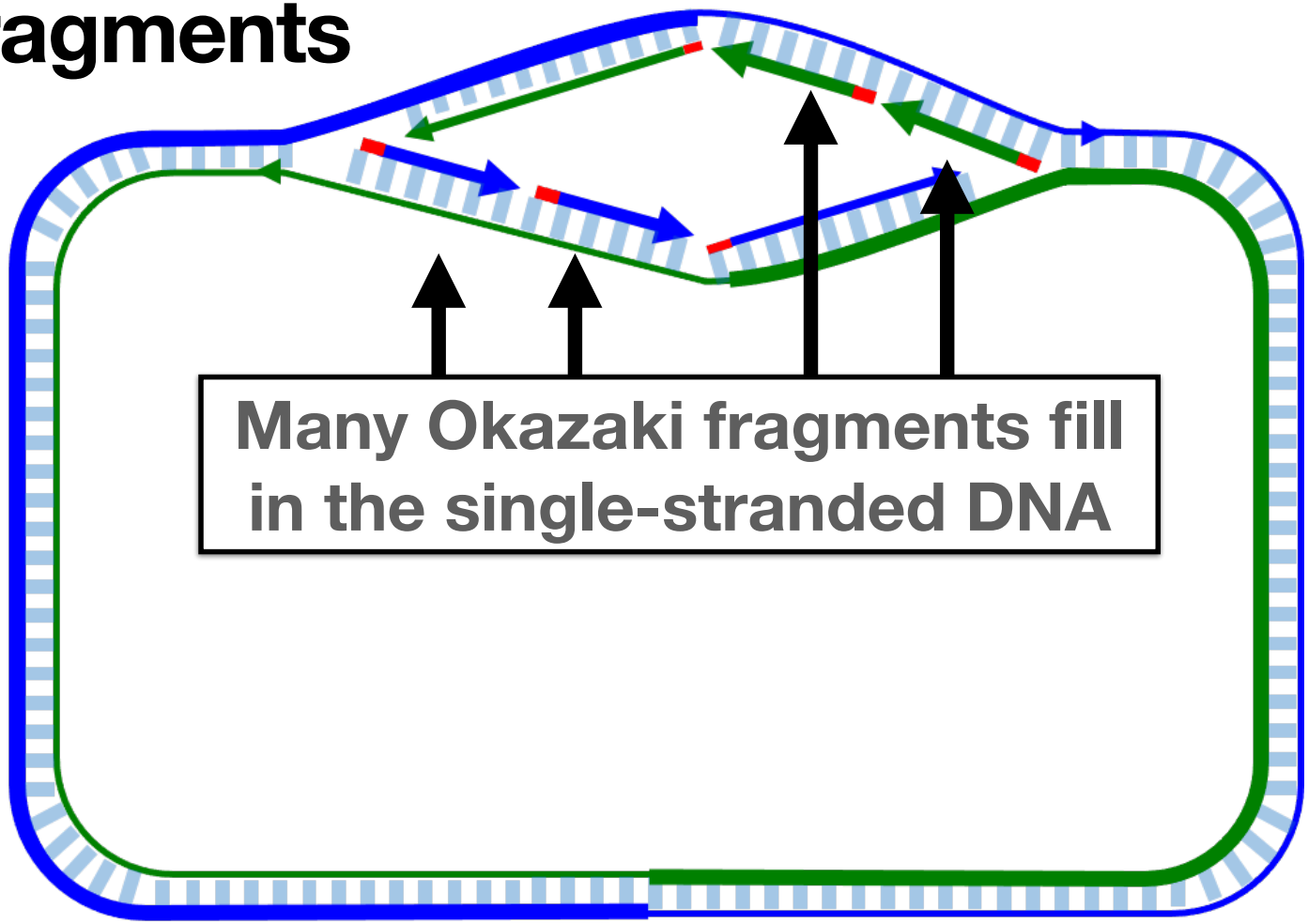
Open fork - dsDNA



Okazaki fragments

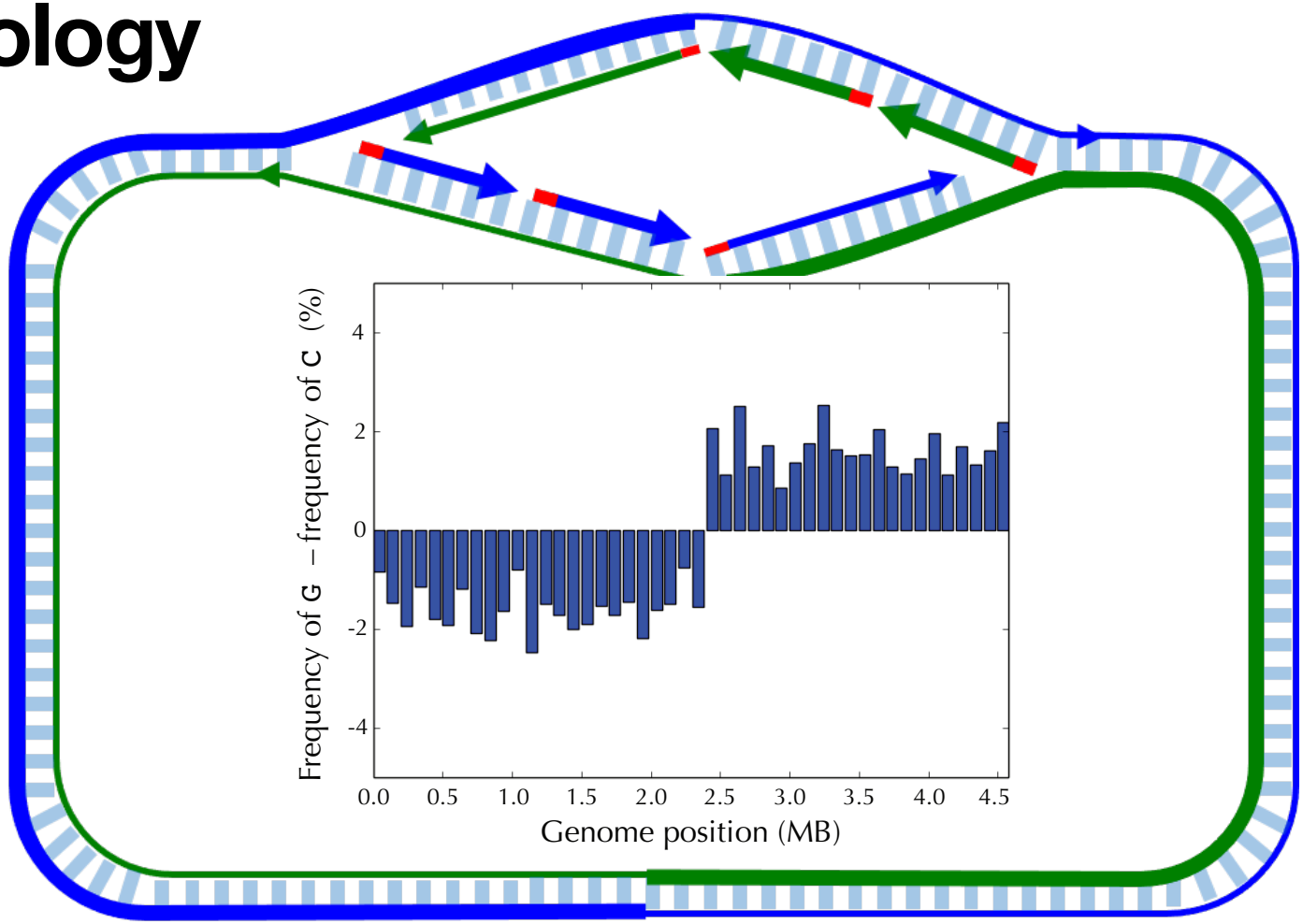


Okazaki fragments



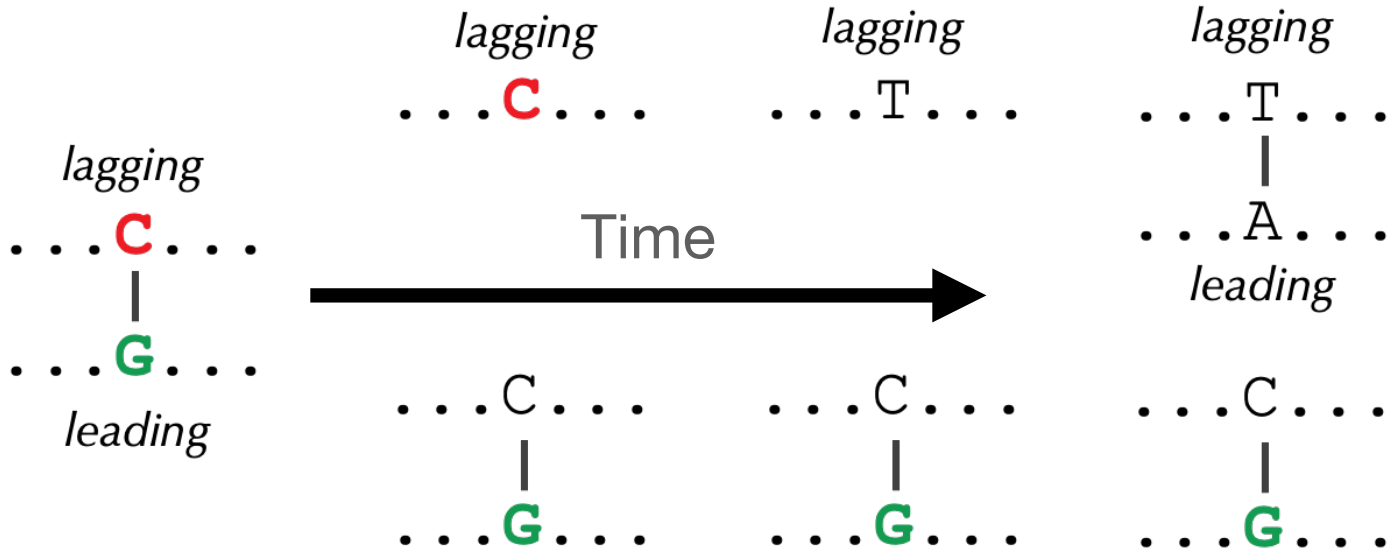
DNA Replication is done, with implications

Data to biology



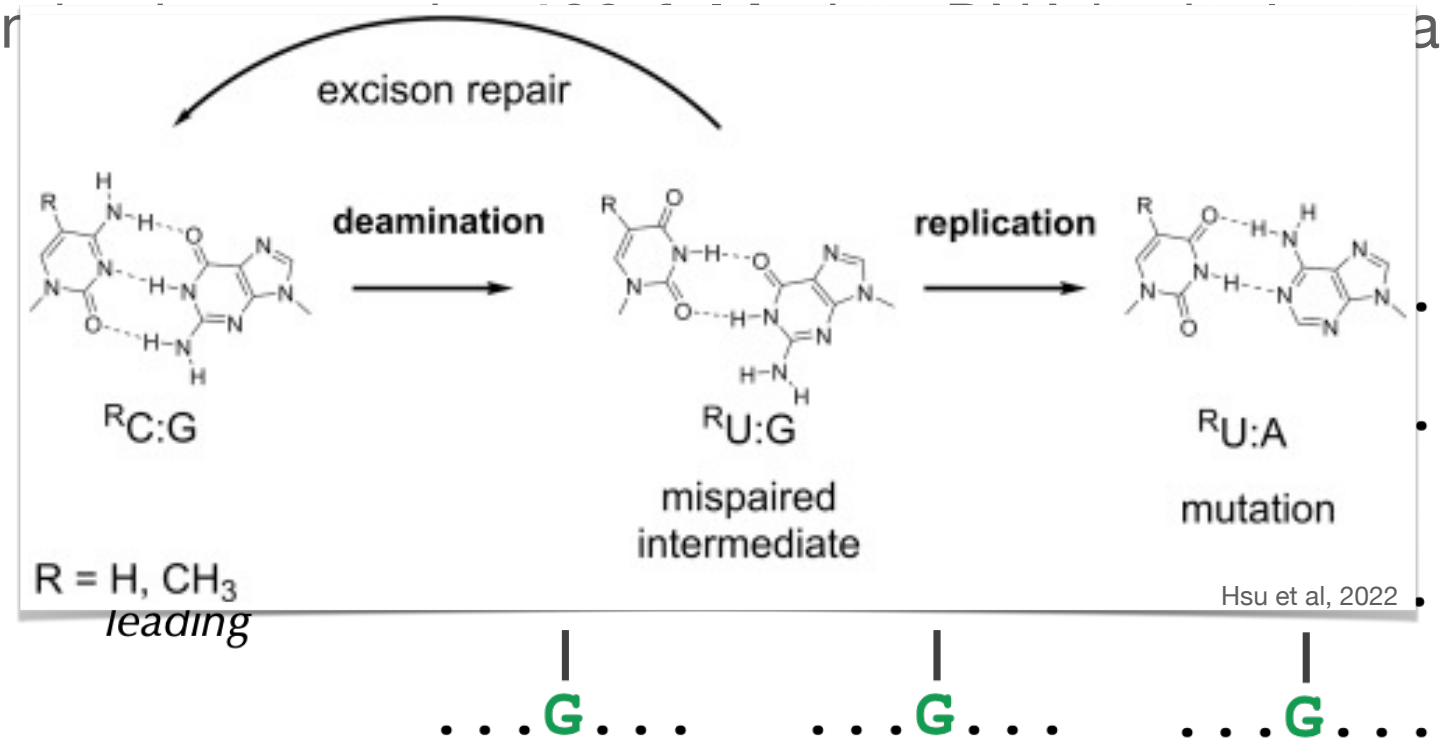
Deamination is the answer

Cytosine (C) rapidly mutates into thymine (T)* through deamination; deamination rates rise **100-fold** when DNA is single-stranded

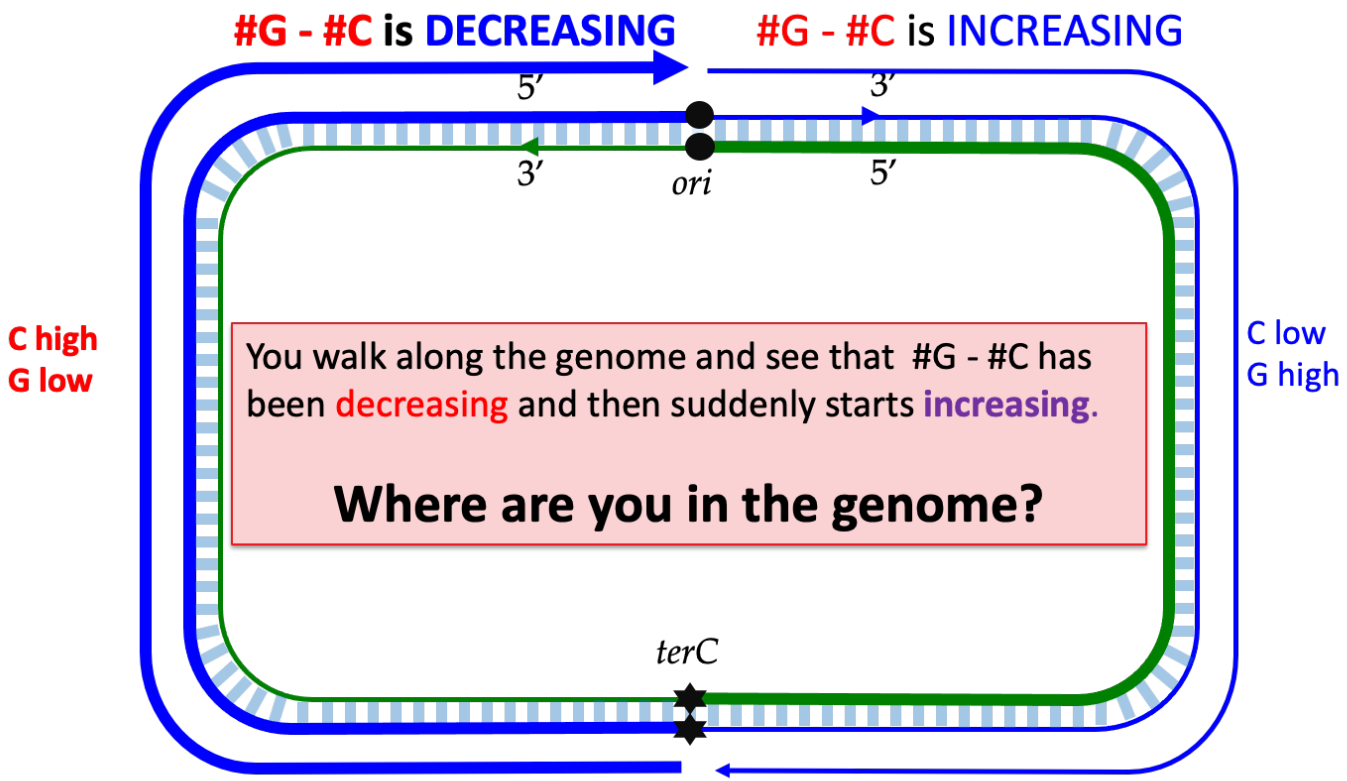


Deamination is the answer

Cytosine (C) rapidly mutates into thymine (T)* through deamination; dear anded



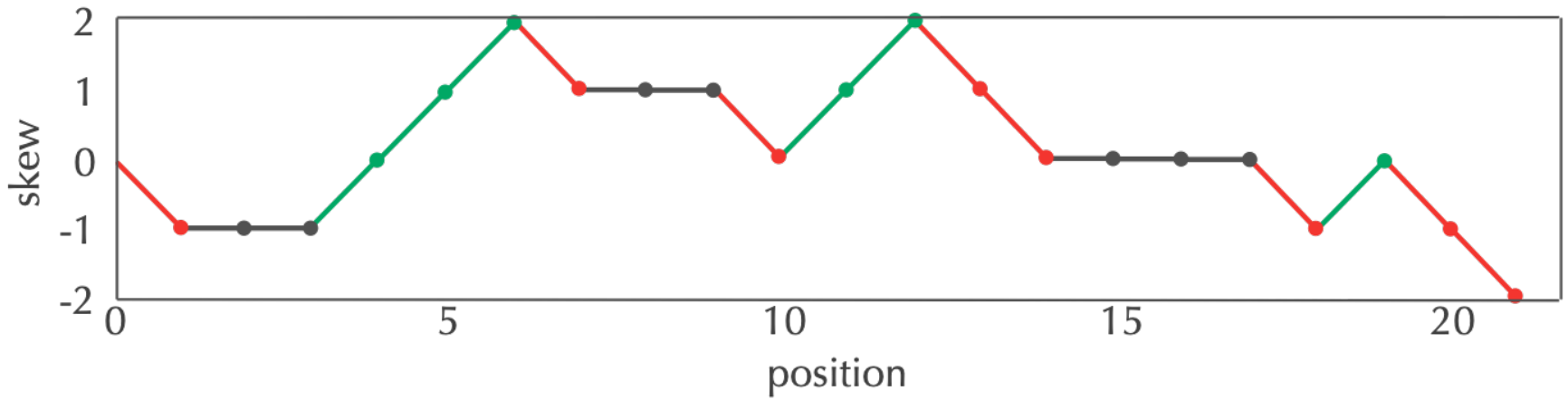
Intuition for where we'll see the G/C skew



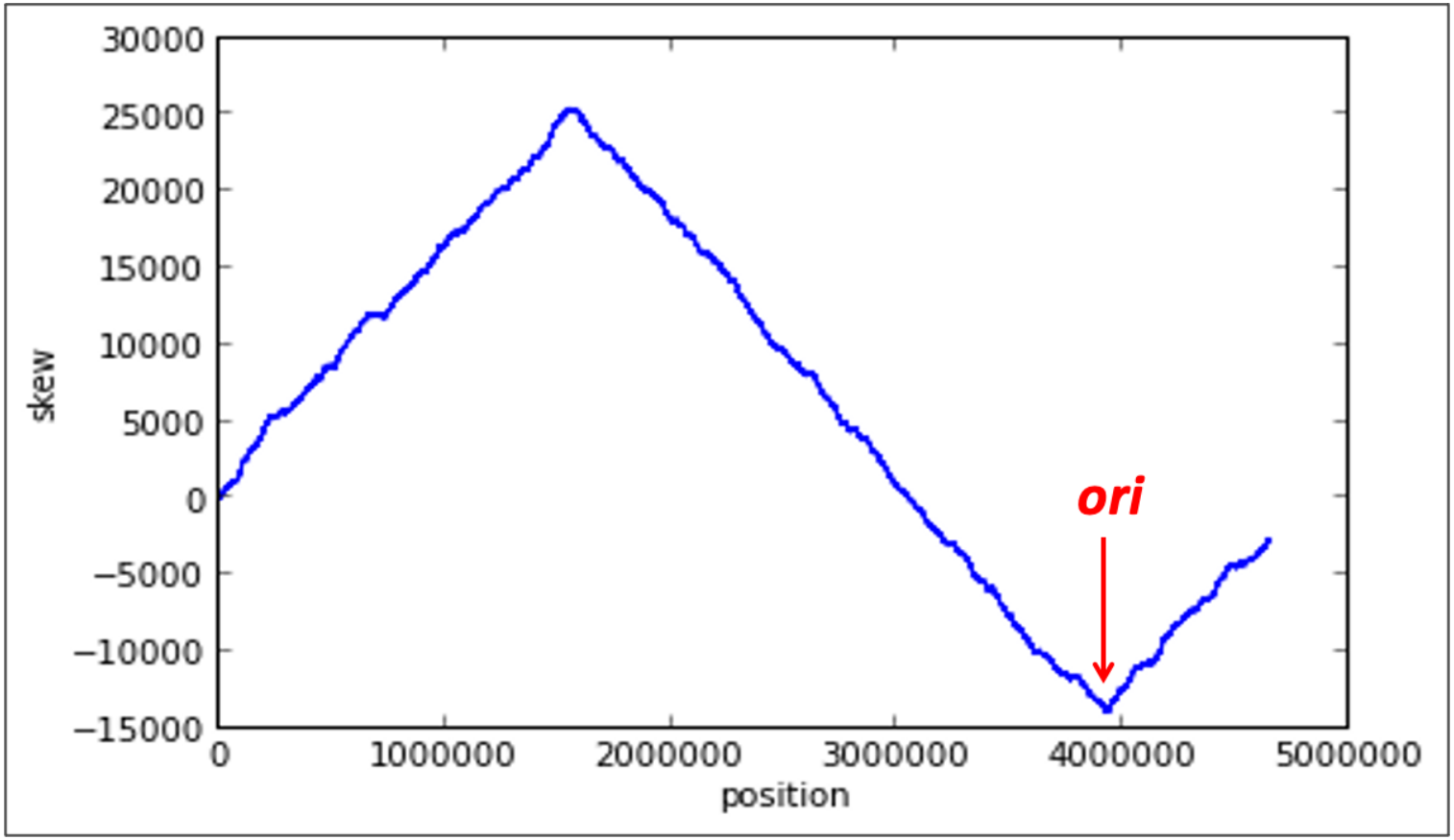
C high/G low → #G - #C is DECREASING as we walk along the LEADING half-strand

C low/G high → #G - #C is INCREASING as we walk along the LAGGING half-strand

Which we can turn into a diagram

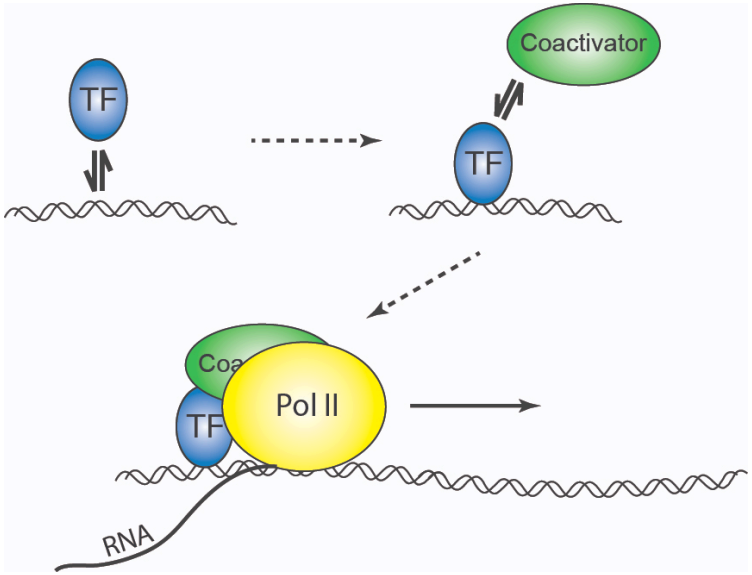


And use as a readout for OriC detection



Next week

- K-mers onto motifs!
- Gibbs sampling!
- PWM and more!
- No Thursday class!
- ‘Quiz’ on material so far, exercise today, class material switches



Motifs

t	a	a	c
G	T	c	t
c	c	g	G
a	c	t	a
A	G	G	T

PROFILE(*Motifs*)

A:	0.4	0.2	0.2	0.2
C:	0.2	0.4	0.2	0.2
G:	0.2	0.2	0.4	0.2
T:	0.2	0.2	0.2	0.4

Lets try group work