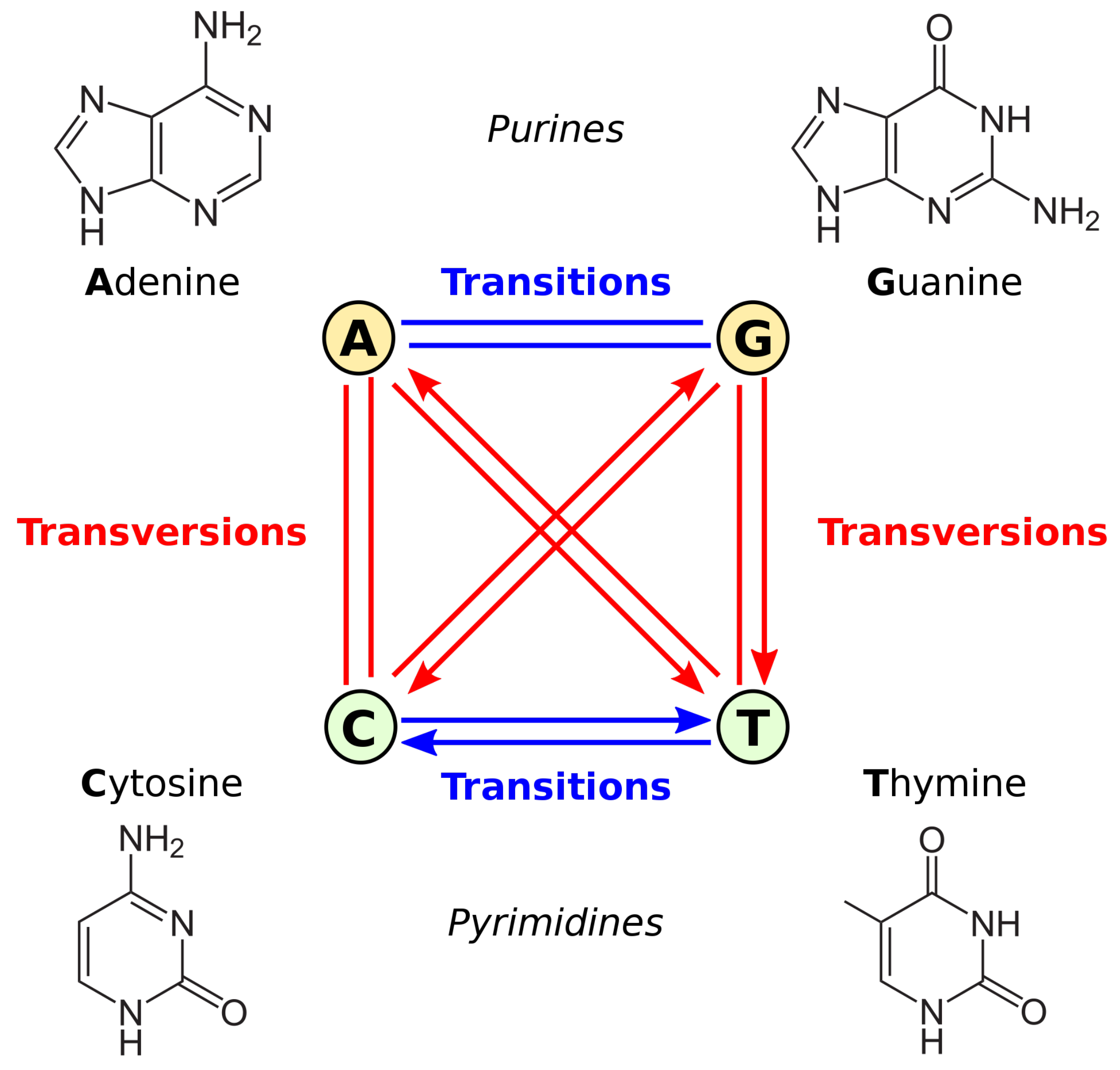
"Mistakes were made..."

- DNA replication is an imperfect process; errors (**mutations**) occur during replication
- Types of mutations
    - **Point mutations**
        - **Substitutions** - one DNA nucleotide is substituted during replication
            - **Transitions** - purine (A,G) for purine (G,A) or pyrimidine (C,T) for pyrimidine (T,C)
            - **Transversions** - purine (A,G) for pyrimidine (C,T) or pyrimidine (C,T) for purine (A, G)
    - **Insertions/Deletions (indels)** - loss or gain of one ore more bases
    - **Genome rearrangements**
        - **Translocations** - a part of a chromosome ends up in a different genomic region
        - **Inversions** - a region of a chromosome is "inverted" relative to its prior orientation
        - **Duplication** - a region of a chromosome is duplicated somewhere else in the genome

DNA substitutions

- Transitions:
  - example: `ATGCGAAAT -> ATGCGAGAT`
- Transversions:
  - example: `ATGCGAAAT -> ATGCGACAT`

Image source: wikipedia



Transition and tranversion rates in *Arabidopsis thaliana*

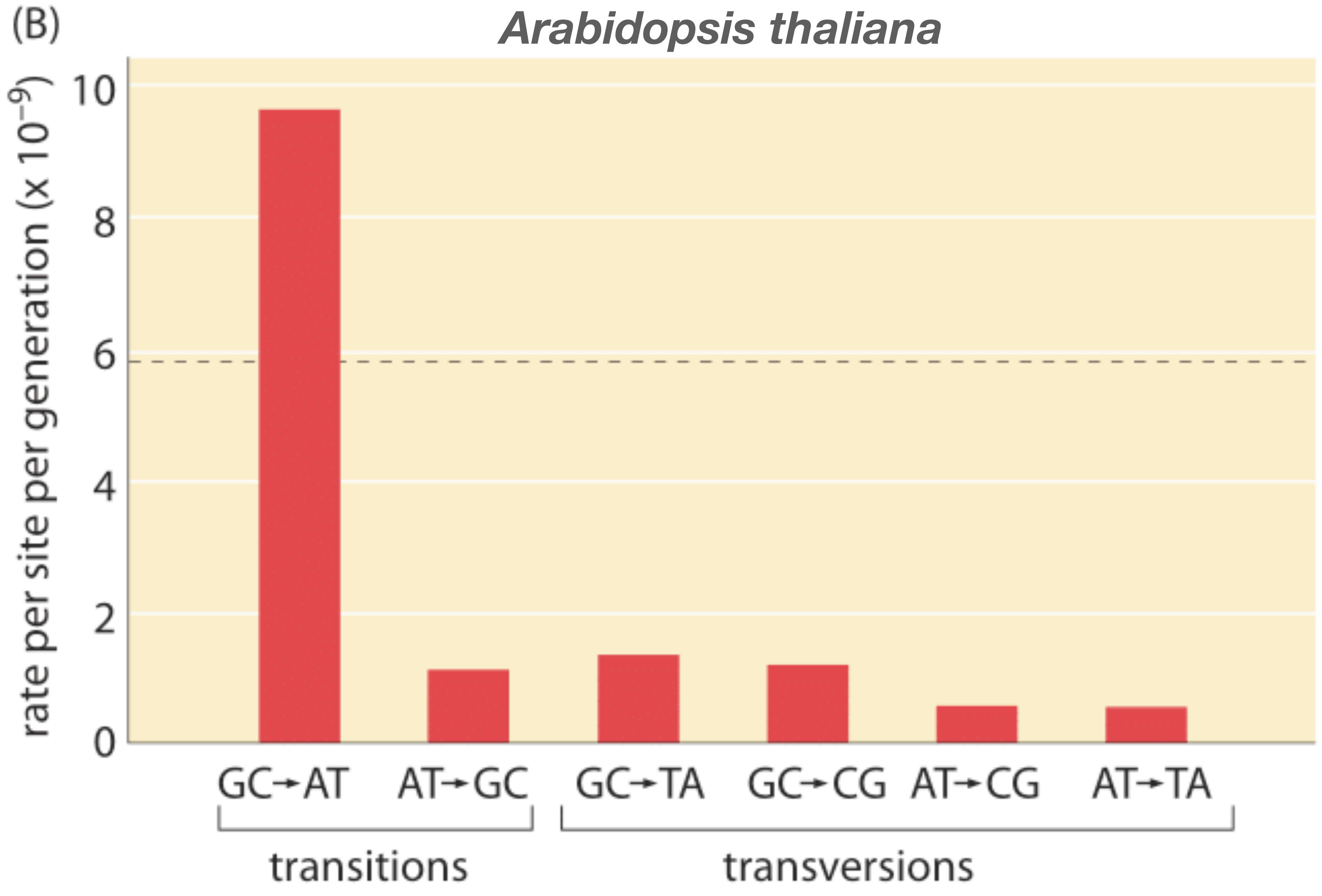


image from Cell Biology by the Numbers

Genome-wide Mutation Rates (substitutions)

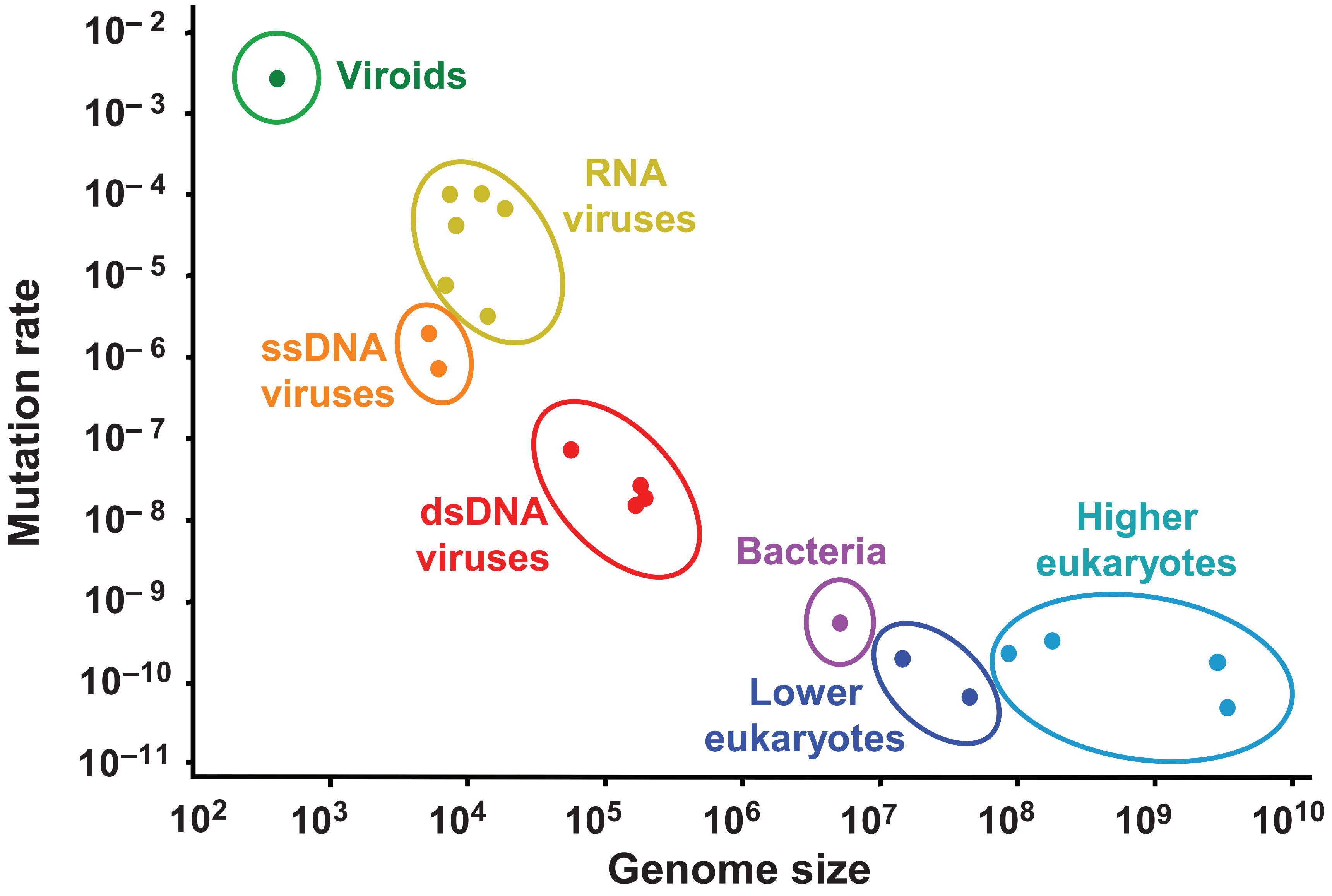


Fig. 1. Per-site mutation rate versus genome size for CChMVd and other biological entities [reviewed in (*2*) and updated with more recent data from (*3*)]. RNA viruses (left to right) are tobacco mosaic virus, human rhinovirus, poliovirus, vesicular stomatitis virus, bacteriophage Φ6, and measles virus. Single-stranded DNA viruses are bacteriophage ΦX174 and bacteriophage m13. Double-stranded DNA viruses are bacteriophage λ, herpes simplex virus, bacteriophage T2, and bacteriophage T4. Bacteria is *Escherichia coli*. Lower eukaryotes are *Saccharomyces cerevisiae* and *Neurospora crassa*. Higher eukaryotes are *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens*. When several estimations were available, the mean value is shown.

from Gago et al. 2009, Science

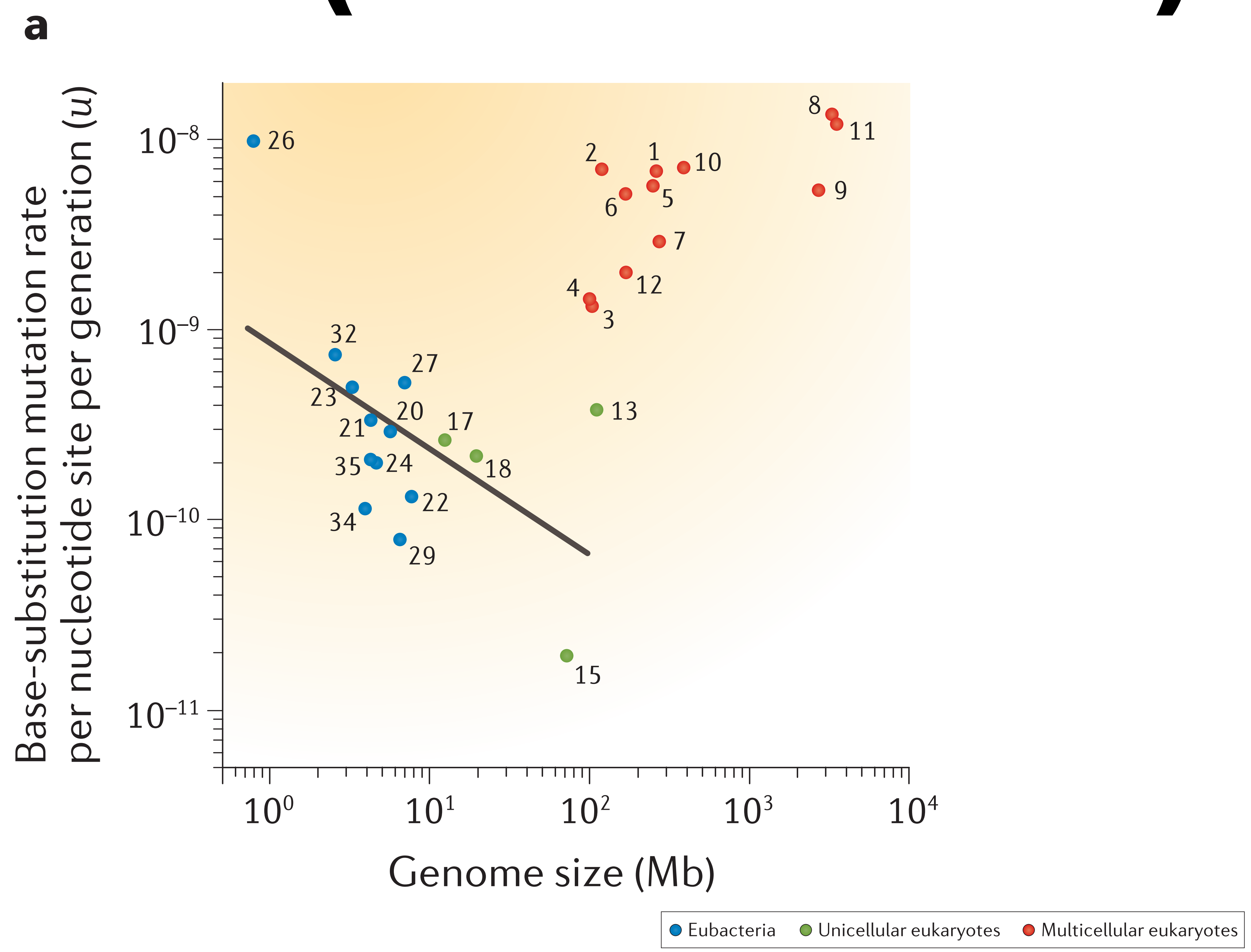


Figure 3 | **Scaling relationships involving the base-substitution mutation rate. a** | The relationship of the base-substitution mutation rate per nucleotide site per generation (u) with total haploid genome size is given for the full set of species for which data are available from mutation-accumulation whole-genome sequencing (MA-WGS) or pedigree analyses. The regression line only incorporates the data for unicellular species. **b** | The regression of u on the estimated effective population size (N_e). To increase the sample size here, the mutation rates of three bacteria (data points 25, 30 and 33) and two unicellular eukaryotes (data points 16 and 19) are based on reporter-construct estimates. **c** | The regression of the total (genome-wide) mutation rate in protein-coding DNA per generation (U_P) on N_e. The solid line is the regression fitted to the full data set, whereas the dashed lines are reference lines with slopes equal to –1.0. The arrows are the approximate degree to which the multicellular eukaryote measures are likely to move upwardly if all sites under selection are accounted for (as described in the text). All plotted data are in Supplementary information S1 (table). Numbered data points correspond to the following species: 1, *Apis mellifera*; 2, *Arabidopsis thaliana*; 3, *Caenorhabditis briggsae*; 4, *Caenorhabditis elegans*; 5, *Daphnia pulex*; 6, *Drosophila melanogaster*; 7, *Heliconius melpomene*; 8, *Homo sapiens*; 9, *Mus musculus*; 10, *Oryza sativa*; 11, *Pan troglodytes*; 12, *Pristionchus pacificus*; 13, *Chlamydomonas reinhardtii*; 14, *Neurospora crassa*; 15, *Paramecium tetraurelia*; 16, *Plasmodium falciparum*; 17, *Saccharomyces cerevisiae*; 18, *Schizosaccharomyces pombe*; 19, *Trypanosoma brucei*; 20, *Agrobacterium tumefaciens*; 21, *Bacillus subtilis*; 22, *Burkholderia cenocepacia*; 23, *Deinococcus radiodurans*; 24, *Escherichia coli*; 25, *Helicobacter pylori*; 26, *Mesoplasma florum*; 27, *Mycobacterium smegmatis*; 28, *Mycobacterium tuberculosis*; 29, *Pseudomonas aeruginosa*; 30, *Salmonella enterica*; 31, *Salmonella typhimurium*; 32, *Staphylococcus epidermidis*; 33, *Thermus thermophilus*; 34, *Vibrio cholera*; 35, *Vibrio fischeri*.

from Lynch et al. 2016, Nat Rev Genet

DNA insertions/deletions

- Insertions:
  - example: `ATGCGAAAT -> ATGCGAAAAT`
- Deletions
  - example: `ATGCGAAAT -> ATGCGAAT`

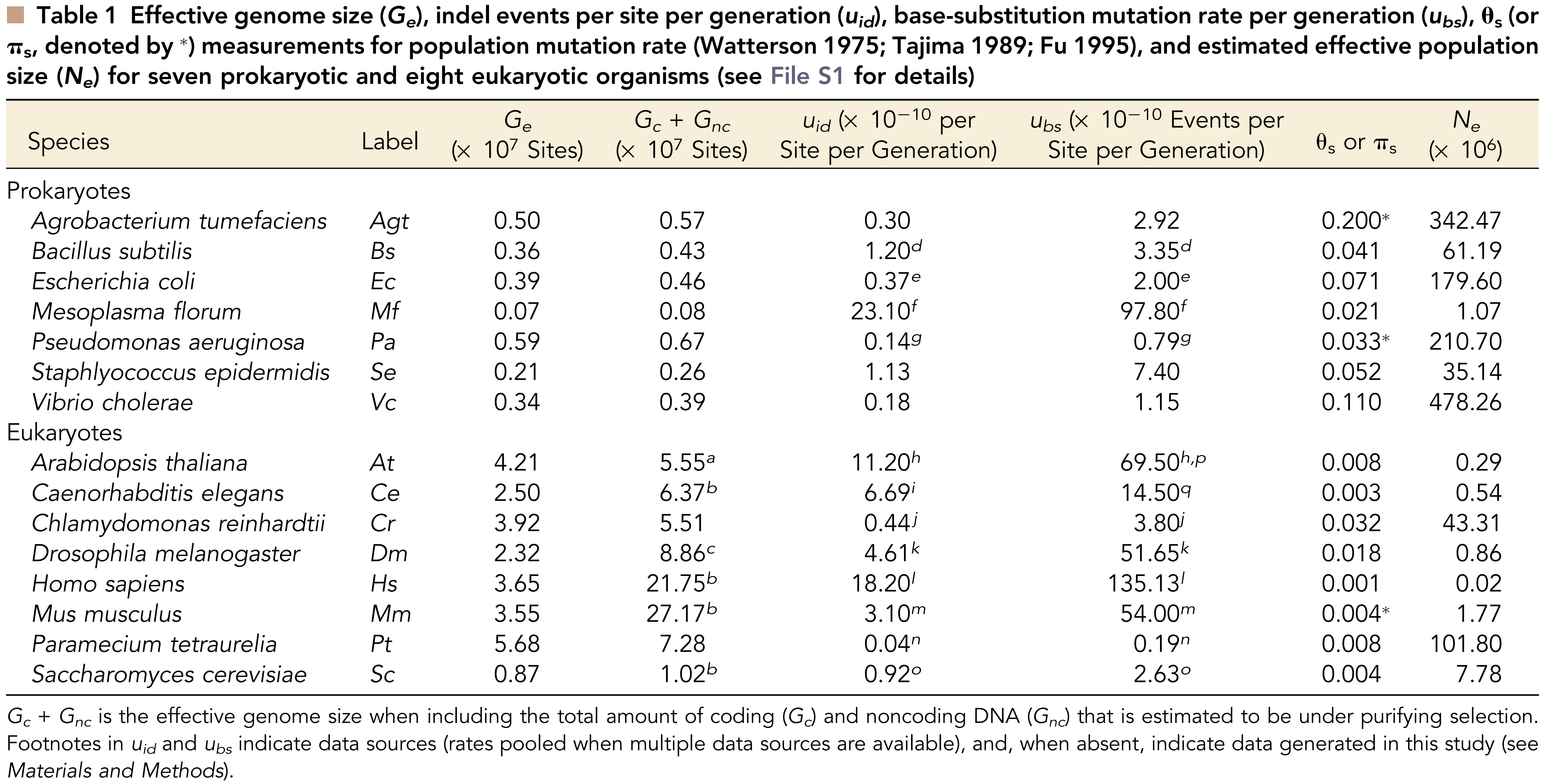


Table 1 Effective genome size (G_e), indel events per site per generation (u_{id}), base-substitution mutation rate per generation (u_{bs}), θ_s (or π_s, denoted by *) measurements for population mutation rate (Watterson 1975; Tajima 1989; Fu 1995), and estimated effective population size (N_e) for seven prokaryotic and eight eukaryotic organisms (see File S1 for details)

Species	Label	G_e ($\times$ 10^7 Sites)	G_c + G_{nc} ($\times$ 10^7 Sites)	u_{id} ($\times$ 10^{-10} per Site per Generation)	u_{bs} ($\times$ 10^{-10} Events per Site per Generation)	θ_s or π_s	N_e ($\times$ 10^6)
Prokaryotes							
Agrobacterium tumefaciens	*Agt*	0.50	0.57	0.30	2.92	0.200*	342.47
Bacillus subtilis	*Bs*	0.36	0.43	1.20^{d}	3.35^{d}	0.041	61.19
Escherichia coli	*Ec*	0.39	0.46	0.37^{e}	2.00^{e}	0.071	179.60
Mesoplasma florum	*Mf*	0.07	0.08	23.10^{f}	97.80^{f}	0.021	1.07
Pseudomonas aeruginosa	*Pa*	0.59	0.67	0.14^{g}	0.79^{g}	0.033*	210.70
Staphlyococcus epidermidis	*Se*	0.21	0.26	1.13	7.40	0.052	35.14
Vibrio cholerae	*Vc*	0.34	0.39	0.18	1.15	0.110	478.26
Eukaryotes							
Arabidopsis thaliana	*At*	4.21	5.55^{a}	11.20^{h}	69.50^{h,p}	0.008	0.29
Caenorhabditis elegans	*Ce*	2.50	6.37^{b}	6.69^{i}	14.50^{q}	0.003	0.54
Chlamydomonas reinhardtii	*Cr*	3.92	5.51	0.44^{j}	3.80^{j}	0.032	43.31
Drosophila melanogaster	*Dm*	2.32	8.86^{c}	4.61^{k}	51.65^{k}	0.018	0.86
Homo sapiens	*Hs*	3.65	21.75^{b}	18.20^{l}	135.13^{l}	0.001	0.02
Mus musculus	*Mm*	3.55	27.17^{b}	3.10^{m}	54.00^{m}	0.004*	1.77
Paramecium tetraurelia	*Pt*	5.68	7.28	0.04^{n}	0.19^{n}	0.008	101.80
Saccharomyces cerevisiae	*Sc*	0.87	1.02^{b}	0.92^{o}	2.63^{o}	0.004	7.78

G_c + G_{nc} is the effective genome size when including the total amount of coding (G_c) and noncoding DNA (G_{nc}) that is estimated to be under purifying selection. Footnotes in u_{id} and u_{bs} indicate data sources (rates pooled when multiple data sources are available), and, when absent, indicate data generated in this study (see *Materials and Methods*).

Table from Sung et al 2016, G3

Some problems

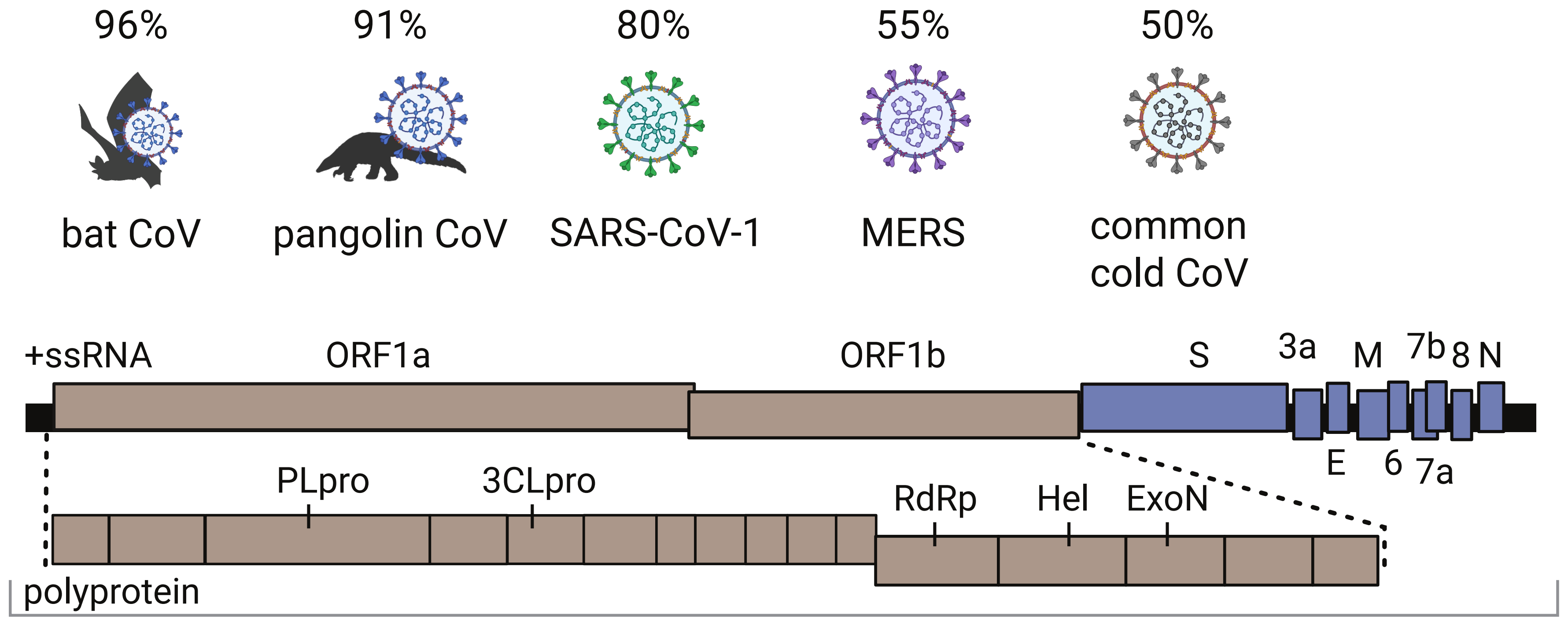
- A culture of E *coli* can go from a density of 1 cell/ml to ~10^9 cells/ml overnight (saturated culture).
  - What would you estimate the minimum number of DNA replications (per ml) that occur during this process? Explain your reasoning.
- The *E coli* genome is approximately 5 Mb in size. The per base mutation rate in *E coli* is ~10^{-10} mutations/base/replication.
    - How many many genomic mutations would you expect to see in one genome replication?
- In a 10 ml saturated culture of *E coli*, all descended from a single cell, how many novel mutations would you expect to observe? How does the number of expected mutations compare to E. coli's genome size?
- How many point mutations per genome per generation do you expect to observe in humans? How many indels per genome do you expect to observe?

Some problems

- Given the estimate of viral mutation rate in the figure to the right, what is the probability that **no genomic mutations** occur during a single replication event in COVID-19?
- The "burst size" of a virus is the number of virions produced from infection of a single cell. How many COVID-19 mutations would you expect to observe per infected cell (at "burst")?
- What is the probability of observing **no viral mutations in the viral population** at burst?
- Would you expect every cell to show the same number of mutations? Why or why not?
- Sketch or outline how you might setup a simulation in Python to model the accumulation of viral mutations in a single cell during viral replication.

Genome

Nucleotide identity to SARS-CoV-2



Length: ≈30kb; β-coronavirus with 10-14 ORFs (24-27 proteins)

Evolution rate: ~10^{-3} nt^{-1} yr^{-1} (measured for SARS-CoV-1)

Mutation rate: ~10^{-6} nt^{-1} $cycle^{-1}$ (measured for MHV coronavirus)

Replication Timescales

in tissue-culture

Virion entry into cell: ~10 min (measured for SARS-CoV-1)

Eclipse period: ~10 hrs (time to make intracellular virions)

Burst size: ~10^3 virions (measured for MHV coronavirus)

from Bar-on et al. 2020, eLife

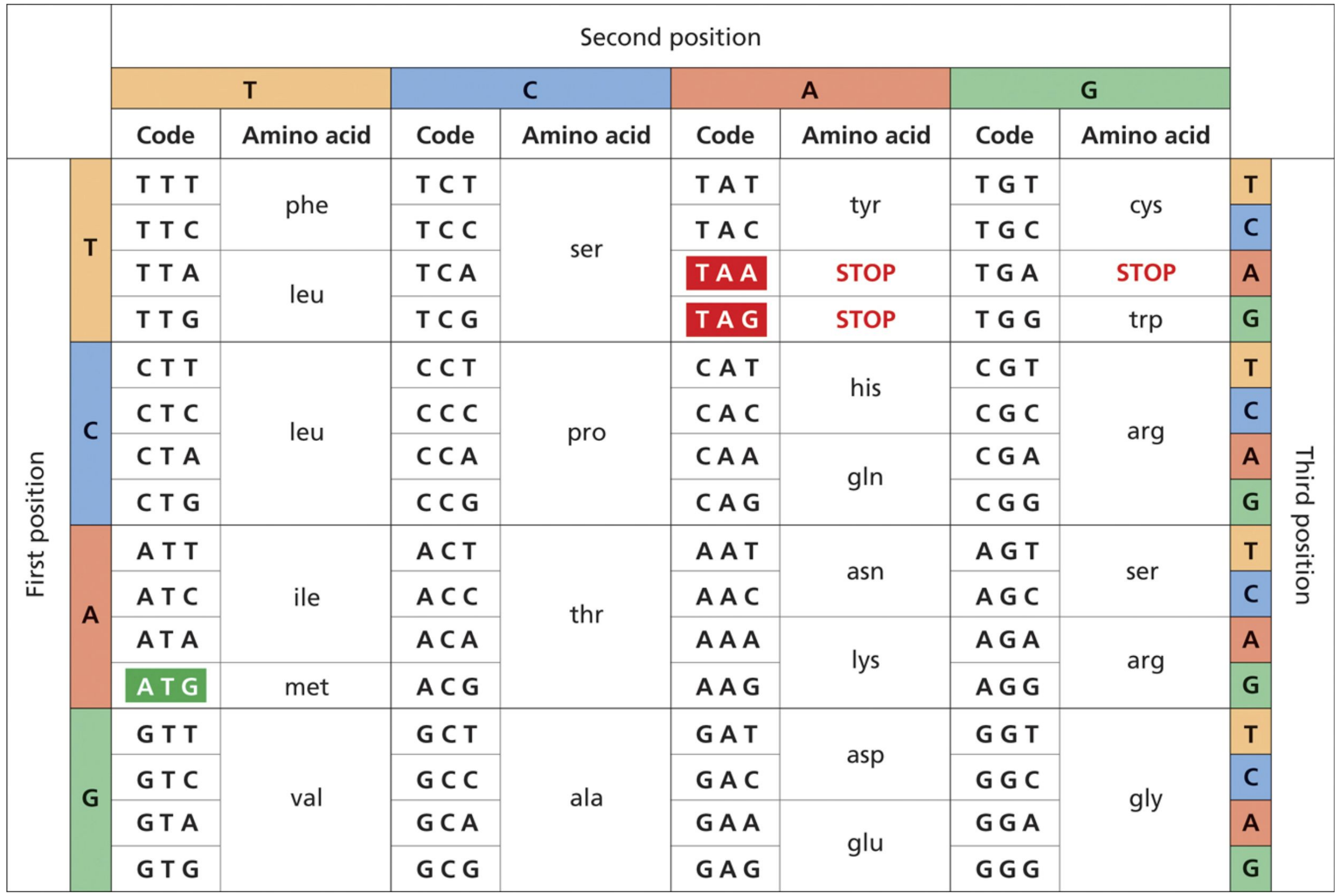
Consequences of mutations

- Coding regions
    - Substitutions
        - Synonymous - substitution mutations that don't change amino acid sequence
        - Non-synonymous - substitution mutations that change amino acid sequence (missense mutations)
    - Indels
        - Frame shifts - insertion or deletion usually leads to change of amino acid sequence because the "reading frame" shifts
    - Both substitutions and indels mutations can cause nonsense mutations (premature stop codons) and nonstop mutation (mutations in stop mutations that lead to continued translation past typical stop)
    - Mutations in splice sites can effect splicing
- Non-coding regions
    - Harder to predict consequences a priori because relationship between sequence and function is not as straightforward as it is in coding regions

Some problems

- What fraction of mutations at the 3rd position in codons are synonymous?

- Which amino acid have codons that exhibit the greatest degeneracy at the 3rd position?



	Second position								
	T		C		A		G		
First position	Code	Amino acid	Code	Amino acid	Code	Amino acid	Code	Amino acid	Third position
T	TTT	phe	TCT	ser	TAT	tyr	TGT	cys	T
	TTC		TCC		TAC		TGC		C
	TTA	leu	TCA		TAA	STOP	TGA	STOP	A
	TTG		TCG		TAG	STOP	TGG	trp	G
C	CTT	leu	CCT	pro	CAT	his	CGT	arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	ile	ACT	thr	AAT	asn	AGT	ser	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	lys	AGA	arg	A
	ATG	met	ACG		AAG		AGG		G
G	GTT	val	GCT	ala	GAT	asp	GGT	gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	glu	GGA		A
	GTG		GCG		GAG		GGG		G

from Momand and McMurdy 2016