Maize Population Genetics — Fermi Math Cheatsheet

THE "RULE OF THUMB" TOP 10

Quantity	Fermi Number	Exact / Literature Range
Haploid genome size (G)	~2 Gb (2×10^9 bp)	2.13 Gb in 10 chromosomes[1] ($2n = 20$)[2]; up to 3.1 Gb in landraces[3,4]
Protein-coding genes	$\sim 4 \times 10^4$	39,756 in B73[1]; ~103,000 pan-genes (~1/3 core)[5]
Mutation rate (μ)	$\sim 3 \times 10^{-8}$ bp^{-1} gen^{-1}	$3–7.3 \times 10^{-8}$ (~65–155 new mutations/gamete)[6,7]
Recombination rate (r)	$\sim 10^{-8}$ bp^{-1} (~1 cM/Mb)	0.7 cM/Mb mean; ~1,500 cM map (~15 COs/gamete)[8–10]
Nucleotide diversity (π)	~1.5% (1.5×10^{-2})	1.4×10^{-2} genome-wide; 1.2×10^{-2} at genes, 2.1×10^{-2} far from genes[11]
Pop-scaled mutation ($\theta = 4N_e\mu$)	~0.015	$\theta_W \approx 0.015$ across 26 NAM founders (1 SNP / 18 bp)[11]
Pop-scaled recomb. ($\rho = 4N_e r$)	~0.004 bp^{-1} (~4 kb^{-1})	$\rho/\theta = r/\mu \approx 0.1–0.35$ (ignores gene conversion)
Effective pop. size (N_e)	$\sim 10^5$ (long-term)	$0.7–1.8 \times 10^5$ from π_0[11]; bottleneck ~8k (9 kya), modern ~10k[12]
Census size (N_c) & density	$\sim 10^{12}$ plants (~3×10^4/ha)	$\sim 10^{12}$ open-pollinated plants[13]; $2–8 \times 10^4$ plants/ha (dryland to irrigated)[14]
Fecundity & mating	~500 kernels/ear; ~95% outcrossed	300–1,000 kernels/ear[14] (400–600 typical[15]); $5–50 \times 10^6$ pollen grains/plant/season[16]; ~5% selfing[14]; 1 yr/gen[17]

GENOMIC LENGTH SCALES (SMALLEST → LARGEST)

- **Conserved noncoding sequence (CNS): ~30 bp** (median 12–40 bp across species; 73% upstream of their gene, ~25% >25 kb away)[18]; 27 bp–15 kb in maize[19]
- **Exon / intron: ~150 bp** each (160 bp exon, 140 bp intron medians; 90% of introns < 1.4 kb)[1]
- **TIR transposon copy: ~300 bp** (310 bp median)[20]
- **Coding sequence (CDS) per gene: ~900 bp** (880 bp median; 3 exons/gene median; 23% single-exon)[1]
- **Pre-mRNA gene span: ~3 kb** (2.7 kb median; 90% < 9 kb)[1]
- **Intact TE copy: ~10 kb** (LTR retrotransposons 11 kb; Helitrons 7 kb medians)[20]
- **Intergenic distance: ~25 kb** (26 kb median; 10–90%: 0.8–136 kb)[1]
- **Extent of LD (distance to mean $r^2 = 0.1$): ~10–50 kb** in annual teosintes (e.g. *parviglumis* $n = 70$, *mexicana* $n = 81$) vs. **~200 kb** in 507 maize inbreds[21]; within genes, $r^2 < 0.1$ by 1.5 kb in 102 diverse inbreds (6 genes)[22]
- **Distal regulatory elements: ~2–100 kb** from genes (~1/3 of accessible chromatin >2 kb from a gene[23]; *tb1* ~60 kb[24], *Vgt1* 70 kb[25], *b1* 100 kb[26])
- **Domestication sweep: ~300 kb** (mean selected region)[27]
- **Centromere (CENH3 domain): ~2 Mb** (1.8 Mb)[28]
- **Chromosome length: ~200 Mb** (152–308 Mb; ~16 genes/Mb median, 35/Mb near tips vs. 10/Mb mid-chromosome)[1]

GENOME COMPOSITION, VARIATION & SELECTION

- **Genome fractions: ~85% TEs** (~74% LTR, ~8.5% DNA TEs)[5,29]; 200k intact copies in 29k families; *huck*, *ji*, *cinful-zeon*, *opie* 7–11% each[20,30]; **~8% genic**; **~2% CDS**[1]; **42% callable** across NAM assemblies[11].
- **Conserved noncoding sequence: ~5%** of the genome (~107 Mb conserved across Andropogoneae, 42 Mb in all species; ~half in UTRs and introns; 87% of genes have a CNS within 2 kb upstream)[19]; 80% of maize CNSs carry a predicted TF binding site[18].
- **Protein evidence:** peptides detected for **66%** of annotated B73 proteins across public mass-spec data[31]; proteins from ~18,500 genes in a 33-tissue atlas[32].
- **Paleopolyploidy:** ancient allotetraploid (subgenomes split ~12 MYA[33]; polyploidy shared with *Tripsacum*[21]); **≥50%** of duplicate genes lost[34].
- **Polymorphism & SVs: ~1 SNP / 70 bp** between two random haplotypes[11]; **>83M SNPs** (1,218 lines)[35]; **~8×10^5 SVs** (>100 bp) in 26 NAM founders (50% < 5 kb)[5].
- **Divergence & admixture:** landraces retain **~83%** of *parviglumis* π ($F_{ST} \approx 0.11$)[27]; *parviglumis* vs. *mexicana* $F_{ST} \approx 0.33$[36]; modern maize carries **~18%** (15–25%) *mexicana* ancestry[37].
- **Selection & constraint:** background selection reduces π **~43%** at genes ($B = 0.57$; recovers by 0.1 cM) and **~26%** genome-wide (mean $B = 0.74$)[10,11]; only ~10% of the genome (>1 cM/Mb) nears neutral π[38]. Deleterious rate **~3×10^{-10} bp^{-1}** ($U_d \approx 0.6$ per gamete; $u/\mu \approx 0.5–1\%$, an equivalent target of ~10–20 Mb)[11]. Domestication $s \approx 0.015$[27].

EVOLUTIONARY TIMELINE (APPROXIMATE)

- **~12 MYA:** maize–sorghum & subgenome split[33] • **<3 MYA:** LTR expansion[39] (median intact LTR ~0.8 MY[6,20])
- **~650 kya:** *Zea–Tripsacum*[21] → **~120 kya:** *Z. luxurians*[21] → **~30–60 kya:** *parviglumis–mexicana* split[21,40]
- **~9 kya:** Balsas domestication[41] (microfossils 8.7 kya[42]; cobs 6.25 kya[43]) → **~6.5 kya:** S. America[44] → **~5 kya:** *mexicana* admixture[37] → **~4.1 kya:** US Southwest[45] → **1493 CE:** Europe[46]

NOTES ON DERIVED & COMPUTED VALUES

- **Genome size & DNA mass:** landrace 2C = 4.62–6.29 pg[3,4]; haploid size = 2C/2 × 0.978 Gb/pg = 2.26–3.08 Gb (B73: 2.13 Gb ≈ 2.18 pg per 1C).
- **Mutations & crossovers per gamete:** $\mu \times 2.13 \times 10^9$ bp for $\mu = 3–7.3 \times 10^{-8}$[6,7]; 7.3×10^{-8} is from unpublished mutation-accumulation lines. The NAM consensus map has 1,478 markers at 1-cM spacing across 10 chromosomes (~1,470 cM)[8,9]; map/100 ≈ 15 COs/gamete; 1,470 cM / 2,132 Mb ≈ 0.69 cM/Mb.
- **Gene & TE annotations:** canonical transcripts of 39,756 protein-coding genes in `Zm-B73-REFERENCE-NAM-5.0_Zm00001eb.1.gff3` [1]; density in 2,128 1-Mb windows ("tips"/"mid-chromosome" = outer/inner 10% of each chromosome half). Intact TEs (198,875) from `Zm-B73-REFERENCE-GRAMENE-4.0.TEs.1.0.gff3` (2,106 Mb v4)[20,30]; nested TEs overlap.
- **Diversity, θ_W, N_e, B:** across 904,011,088 callable sites (42% of B73 v5; ≤6 of 25 non-B73 NAM founders missing)[11]. Fitting $\pi = \pi_0 e^{-u/r}$ gives $\pi_0 = 0.0189$ (single-locus asymptote 0.0212) and mean $B = 0.74$[10,11]. $\theta_W = (51{,}554{,}740 / 904{,}011{,}088) / a_{26}$ ($a_{26} = 3.82$). $N_e = \pi_0 / 4\mu$; equivalent target fraction = u/μ. Masked regions (56%) are high-recombination and low-constraint, so π, N_e and u are probably underestimates.
- **Demography & census:** ARG N_e trajectory (Magalang, Fairbanks et al., unpublished) read from a plot (±10–20%)[12]; ~85% bottleneck over ~2,500 years, ending near the earliest archaeological maize at Xihuatoxtla[42]. Census: 47.9M ha of open-pollinated maize × ~25,000 plants/ha[13].
- **LTR ages:** $T = K/2\mu$ ($K = 1 -$ LTR identity; $\mu = 3.3 \times 10^{-8}$/bp/yr[6]; 128,481 elements[20]; median identity 94.9%; 62% < 1 MY). The 85% identity cutoff truncates ages at ~2.3 MY.

REFERENCES

1. MaizeGDB. B73 Zm-B73-REFERENCE-NAM-5.0 assembly and Zm00001eb.1 gene annotation (GFF3). https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/ (2020).
2. Gaut, B. S. *et al.* Maize as a model for the evolution of plant nuclear genomes. *Proceedings of the National Academy of Sciences USA* **97**, 7008–7015 (2000).
3. Realini, M. F. *et al.* Intra-specific variation in genome size in maize: Cytological and phenotypic correlates. *AoB Plants* **8**, plv138 (2016).
4. Bilinski, P. *et al.* Parallel altitudinal clines reveal trends in adaptive evolution of genome size in *Zea mays*. *PLoS Genetics* **14**, e1007162 (2018).
5. Hufford, M. B. *et al.* De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
6. Clark, R. M. *et al.* Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. *Molecular Biology and Evolution* **22**, 2304–2312 (2005).
7. Yang, N. *et al.* Contributions of *Zea mays* subspecies *mexicana* haplotypes to modern maize. *Nature Communications* **8**, 1874 (2017).
8. Ogut, F. *et al.* Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. *Heredity* **114**, 552–563 (2015).
9. McMullen, M. D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).
10. Ross-Ibarra, J. FineMap: A composite genetic map of maize. https://doi.org/10.5281/zenodo.19639077 (2026).
11. Liu, B. *et al.* Genome-wide selection on transposable elements in maize. *bioRxiv* https://doi.org/10.1101/2025.09.16.676665 (2025).
12. Ross-Ibarra, J. PEQG keynote 2026. https://www.youtube.com/watch?v=ntJOdZd0mwc&t=2470s (2026).
13. Beissinger, T. M. *et al.* Recent demography drives changes in linked selection across the maize genome. *Nature Plants* **2**, 16084 (2016).
14. Office of the Gene Technology Regulator. *The biology of Zea mays l. Ssp mays (maize or corn)*. https://www.ogtr.gov.au/sites/default/files/files/2021-07/the_biology_of_zea_mays_l._ssp_mays_maize_or_corn.pdf (2008).
15. Nielsen, R. L. Ear size determination in corn. https://www.agry.purdue.edu/ext/corn/news/timeless/EarSize.pdf (2007).
16. Hofmann, F. *et al.* Maize pollen deposition in relation to distance from the nearest pollen source under common cultivation: Results of 10 years of monitoring (2001 to 2010). *Environmental Sciences Europe* **26**, 24 (2014).
17. Wang, L. *et al.* The interplay of demography and selection during maize domestication and expansion. *Genome Biology* **18**, 215 (2017).
18. Amundson, K. R. *et al.* A deep-time landscape of plant cis-regulatory sequence evolution. *Science* https://doi.org/10.1126/science.adt8983 (2026).
19. Song, B. *et al.* Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome Research* **31**, 1245–1257 (2021).
20. MaizeGDB. B73 RefGen_v4 structural transposable element annotation, TEs.1.0 (GFF3). https://www.maizegdb.org (2016).
21. Chen, L. *et al.* Genome sequencing reveals evidence of adaptive variation in the genus *Zea*. *Nature Genetics* **54**, 1736–1745 (2022).
22. Remington, D. L. *et al.* Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences USA* **98**, 11479–11484 (2001).
23. Ricci, W. A. *et al.* Widespread long-range cis-regulatory elements in the maize genome. *Nature Plants* **5**, 1237–1249 (2019).
24. Studer, A. *et al.* Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nature Genetics* **43**, 1160–1163 (2011).
25. Salvi, S. *et al.* Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proceedings of the National Academy of Sciences USA* **104**, 11376–11381 (2007).
26. Stam, M. *et al.* Differential chromatin structure within a tandem array 100 kb upstream of the maize *b1* locus is associated with paramutation. *Genes & Development* **16**, 1906–1918 (2002).
27. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nature Genetics* **44**, 808–811 (2012).
28. Wang, K. *et al.* Maize centromeres expand and adopt a uniform size in the genetic background of oat. *Genome Research* **24**, 107–116 (2014).
29. Schnable, P. S. *et al.* The B73 maize genome: Complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
30. Jiao, Y. *et al.* Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
31. Wijk, K. J. van *et al.* The *Zea mays* PeptideAtlas: A new maize community resource. *Journal of Proteome Research* **23**, 3984 (2024).
32. Walley, J. W. *et al.* Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814–818 (2016).
33. Swigoňová, Z. *et al.* Close split of sorghum and maize genome progenitors. *Genome Research* **14**, 1916–1923 (2004).
34. Lai, J. *et al.* Gene loss and movement in the maize genome. *Genome Research* **14**, 1924–1931 (2004).
35. Bukowski, R. *et al.* Construction of the third-generation *Zea mays* haplotype map. *GigaScience* **7**, gix134 (2018).
36. Pyhäjärvi, T. *et al.* Complex patterns of local adaptation in teosinte. *Genome Biology and Evolution* **5**, 1594–1609 (2013).
37. Yang, N. *et al.* Two teosintes made modern maize. *Science* **382**, eadg8940 (2023).
38. Tittes, S. *et al.* Not so local: The population genetics of convergent adaptation in maize and teosinte. *bioRxiv* https://doi.org/10.1101/2021.09.09.459637 (2021).
39. SanMiguel, P. *et al.* The paleontology of intergene retrotransposons of maize. *Nature Genetics* **20**, 43–45 (1998).
40. Ross-Ibarra, J. *et al.* Historical divergence and gene flow in the genus *Zea*. *Genetics* **181**, 1399–1413 (2009).
41. Matsuoka, Y. *et al.* A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences USA* **99**, 6080–6084 (2002).
42. Piperno, D. R. *et al.* Starch grain and phytolith evidence for early ninth millennium B.P. Maize from the Central Balsas River Valley, Mexico. *Proceedings of the National Academy of Sciences USA* **106**, 5019–5024 (2009).
43. Piperno, D. R. & Flannery, K. V. The earliest archaeological maize (*Zea mays* L.) from highland Mexico: New accelerator mass spectrometry dates and their implications. *Proceedings of the National Academy of Sciences USA* **98**, 2101–2103 (2001).
44. Kistler, L. *et al.* Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* **362**, 1309–1313 (2018).
45. Merrill, W. L. *et al.* The diffusion of maize to the southwestern United States and its impact. *Proceedings of the National Academy of Sciences USA* **106**, 21019–21026 (2009).
46. Tenaillon, M. I. & Charcosset, A. A European perspective on maize history. *Comptes Rendus Biologies* **334**, 221–228 (2011).

Worked Fermi Math Examples in Maize Genetics

Back-of-the-envelope calculations showing how the Top 10 constants and genomic length scales combine to answer practical biological, evolutionary and experimental questions.

1. How many paired-end short reads are needed for 10× sequencing coverage?

$\frac{10 \times (2 \times 10^9 \text{ bp})}{2 \times 150 \text{ bp/pair}} = \frac{2 \times 10^{10} \text{ bp}}{300 \text{ bp/pair}} \approx$ **7×10^7 read pairs** (70 million pairs)

Only ~42% of the genome (~900 Mb) aligns cleanly across assembled NAM genomes; most of the rest is repetitive TE sequence. Expect usable ~10× depth over roughly that fraction, while reads from high-copy TE families pile up or map ambiguously.

2. What is the global mutation supply per site each crop generation?

(10^{12} OP plants) × (2 gametes/plant) × (3×10^{-8} mutations/bp/gamete) ≈ **6×10^4 transmitted substitutions per site per generation**

This is the expected number of new mutations transmitted at a site in one crop generation, averaged across sites, under the stated census and mutation rate. For adaptation, the useful next calculation is $2N\mu_{target}$ in the population under selection, where μ_{target} is the rate to the relevant allele or set of alleles.

3. How many differences do you expect in a typical gene between two random inbred lines?

(3×10^3 bp gene span) × ($\pi_{genic} \approx 1.2 \times 10^{-2}$) ≈ **36 expected differences per gene**

A rough estimate for a 3 kb gene at the reported genic π. The ~11 differences you would get for the ~900 bp CDS is an equal-diversity benchmark; for a coding estimate, use CDS-specific π. The genic mean is not an upper bound for any individual gene.

4. How long until recombination splits a 100 kb segment?

r_{block} = (0.1 Mb) × (1 cM/Mb) = 0.1 cM = 10^{-3} crossovers/gen → $T \approx \frac{1}{r_{block}}$ ≈ **1,000 generations**

This is the mean waiting time for a recombination event across 100 kb along one lineage at 1 cM/Mb; at the measured mean of 0.7 cM/Mb it is ~1,400 generations.

5. How many new deleterious mutations enter the genome per gamete?

U_d = (2×10^9 bp) × (3×10^{-10} deleterious mutations/bp) ≈ **0.6 deleterious mutations/gamete** under the fitted model

Dividing the modeled deleterious rate by the total substitution rate gives $u/\mu \approx 0.5$–1%, an equivalent target size of ~10–20 Mb under the model. This does not identify a physical set of bases at which every possible substitution is deleterious.

6. How many crossovers occur in a 100-plant F_2 mapping population, and how far apart are they?

(100 F_2 plants) × (2 gametes/plant) × (15 COs/gamete) = 3,000 COs → $\frac{2.13 \times 10^9 \text{ bp}}{3,000 \text{ COs}}$ ≈ **0.7 Mb between breakpoints**

About 0.7 Mb is the mean physical spacing between pooled crossover breakpoints. Bin size sets the limit on mapping resolution; QTL support intervals from 100 plants are usually far wider.

7. How big a field plot is needed to grow 10,000 plants for a selection experiment?

$\text{Area} = \frac{10^4 \text{ plants}}{3 \times 10^4 \text{ plants/ha}}$ ≈ **0.33 hectares** (~0.8 acres, or a ~55 m × 60 m plot)

At ~500 kernels per ear, harvesting the top 1% (100 ears) yields ~50,000 seeds, more than enough to replant the next cycle at the same population size.