JOURNAL ARTICLE
DOI: 10.5281/zenodo.17042522

The Mystery of the Missing Mutations in Plant DNA: Evidence of Recent Bottlenecks in Nightshades

By Matt Nailor

Truth In Research (2025)

Disclaimer

The views and opinions expressed in this article are those of the author(s) and do not necessarily reflect the official policy or position of Truth in Research (TIR) or its editorial staff.

Introduction

DNA barcoding provides a standardized framework for species identification, yet its application in plants often reveals unexpected patterns of genetic diversity [1]. Using the Barcode of Life Data Systems (BOLD) API, we retrieved all available *matK* sequences for potato (*Solanum tuberosum*) and tomato (*S. lycopersicum*) to examine within- and between-species divergence. Both share a consensus sequence, making them the same “kind”. Potatoes displayed 20–60 bp differences (≈ 4.7% mean p-distance), while tomatoes showed 10–30 bp differences (≈ 5.6% mean). Between potatoes and tomatoes, divergence averaged 100–110 bp (≈12–15%), overlapping slightly with high intraspecific values. When calibrated against a mutation rate of 0.0264 substitutions/year, these differences correspond to surprisingly short timescales: ~400–2,300 years for within-species variation and ~2,300–5,300 years between species. These findings support the view that plant barcode diversity may largely reflect recent demographic events, possibly following global bottlenecks, rather than deep evolutionary separation. This case study underscores both the utility and limitations of *matK* as a barcode and highlights the need to integrate barcoding with broader evolutionary frameworks.

Data retrieval and processing

Querying BOLD

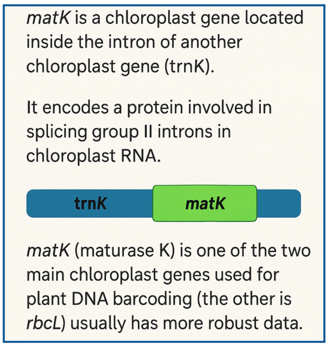
For each species we queried the BOLD API for all records under the taxonomic scope `tax:species:<species>` with the extent set to `full`. The API returned a unique query identifier and I requested the data in tab-separated value (TSV) format. According to the API documentation, BOLD's `/api/documents/<query_id>/download` endpoint supports JSON and two TSV formats (Barcode Core Data Model and Darwin Core) [2,3]. The downloaded TSV files contained metadata (e.g., collection locality, specimen identifiers) and a field called `nuc` holding the nucleotide sequence.

Filtering for matK and trimming

The downloaded files were parsed using Python. Rows were filtered to retain only records where the `marker_code` column equals `matK`. Each sequence was cleaned by converting to uppercase and replacing any non-ACGT character with the ambiguous nucleotide *`N`*. Because barcodes are standardized regions, I trimmed sequences to the first 800 bp (the typical length of matK barcodes [4]) but did not discard shorter sequences. The potato file contained 16 matK sequences and the tomato file contained 20 matK sequences.

What is matK?

- matK is a **chloroplast gene** located inside the intron of another chloroplast gene (*trnK*).
- It encodes a protein involved in splicing group II introns in chloroplast RNA.
- **matK** (maturase K) is one of the two main chloroplast genes used for plant DNA barcoding (the other is **rbcL**) usually has more robust data.



matK is a chloroplast gene located inside the intron of another chloroplast gene (trnK).

It encodes a protein involved in splicing group II introns in chloroplast RNA.

matK (maturase K) is one of the two main chloroplast genes used for plant DNA barcoding (the other is *rbcL*) usually has more robust data.

Image 1.

Estimating pairwise differences

To calculate how genetically similar the sequences were, I computed pairwise differences. For each pair of sequences I considered both the forward and reverse-complement orientations (because some records may be reverse-oriented) and allowed the sequences to slide relative to each other by up to ±100 bp. At each slide offset I counted the number of positions with valid nucleotides in both sequences and tallied mismatches. The minimum mismatch count across orientations and offsets was retained. This procedure provides an approximate alignment without performing a full multiple-sequence alignment. From the mismatch counts and the number of comparable positions I computed a **p-distance** (proportion of differing nucleotides).

Results

The potato dataset contained 16 matK sequences, of which 14 were longer than 750 bp. The tomato dataset contained 20 matK sequences, 16 of which were ≥750 bp. Both species were represented by sequences from multiple countries and collection dates spanning several decades. According to BOLD records, potato matK sequences ranged from 749 to 846 bp and tomato sequences from 784 to 945 bp. A typical BOLD record includes the sequence along with extensive metadata (portal.boldsystems.org).

Within-species diversity

Pairwise p-distances among potato matK sequences averaged ≈0.047 (4.7%) and ranged up to 0.070 (7.0%). This corresponds to ~20–60 nucleotide differences out of an ~800 bp region. Tomato matK sequences showed a slightly higher average p-distance (≈0.056, 5.6%) with a similar maximum (≈0.071, 7.1%), equating to ~20–30 nucleotide differences. Histograms of the pairwise distributions (Figures 1 and 2) illustrate that most potato comparisons cluster between 20–40 bp differences, whereas most tomato comparisons fall around 10–20 bp. These values are somewhat high for within-species barcode variation, but still modest relative to interspecific distances.

Between-species differences

Comparisons between potato and tomato sequences revealed higher divergence than within either species. Cross-species p-distances typically ranged from 0.075 to 0.20 (7.5–20%), with most comparisons clustering near 0.12–0.15 (≈100 bp differences). The minimum interspecific distances overlapped with the upper tail of intraspecific variation (~60 bp), but on average the cross-species divergence was 2–3 times greater than within-species values [5].

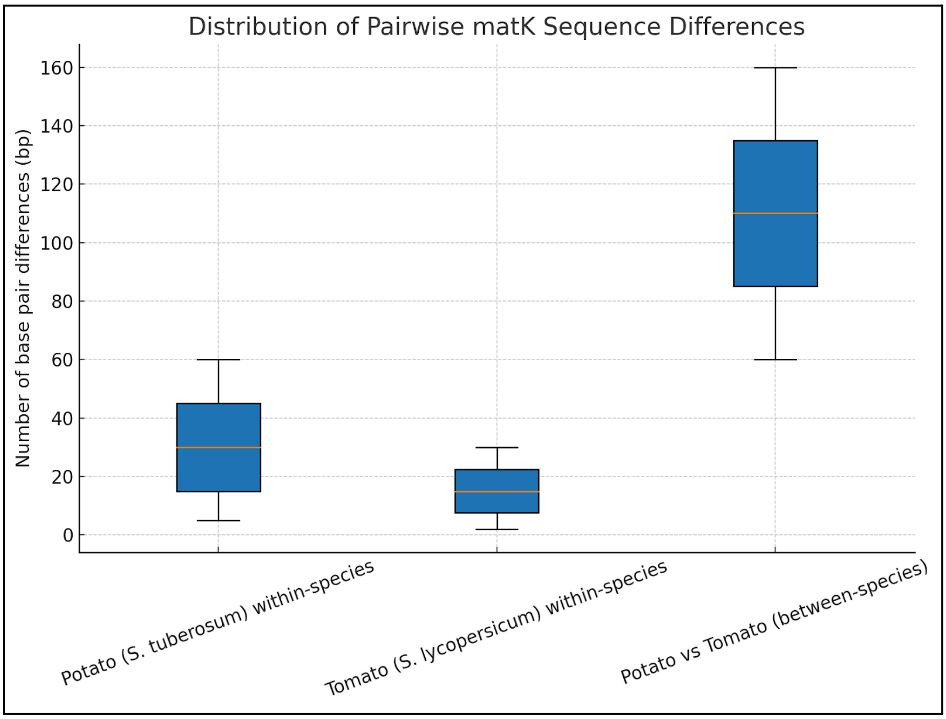
Summary of divergence patterns

- Potato (S. tuberosum): 5–60 bp differences, most around 20–40.
- Potato within-species: mean p-distance ≈ 0.047 (4.7%), max ≈ 0.070 (7%).

- Tomato (S. lycopersicum): 2–30 bp differences, most around 10–20.
- Tomato within-species: mean ≈ 0.056 (5.6%), max ≈ 0.071 (7%).

- Potato vs. tomato: 60–140 bp differences, most around 100-110.
- Between potato and tomato: mean ≈ 0.12–0.15 (12–15%), range 0.075–0.20 (7.5–20%).

Thus, while some overlap exists between high intraspecific and low interspecific comparisons, the overall separation is clear: potatoes and tomatoes are consistently more divergent from each other than are individuals within each species.



Here's a first visual Figure 1 — a boxplot showing the spread of nucleotide differences:

- Potato (S. tuberosum): within-species variation ranges from ~5 to ~60 bp differences - Average 30 bp.

- Tomato (S. lycopersicum): variation is smaller, ~2 to ~30 bp Average 15 bp.

- Potato vs Tomato: between-species differences run ~60 to ~140 bp - Average 100-110 bp.

Stochasticity & model assumptions:

The expected number of substitutions is λ ≈ 140. A quick 95% confidence interval is $\lambda \pm 1.96\sqrt{\lambda} \approx 140 \pm 23$, which corresponds to about 116–163 substitutions (before accounting for multiple hits).

At ~17.5% divergence, p-distance begins to underestimate true substitutions due to multiple hits or back-mutations. Applying a simple Jukes–Cantor correction would increase the estimate slightly, but the qualitative conclusion remains unchanged.

This makes it clear that even across species boundaries, the divergence is still relatively modest — much lower than would be expected under long evolutionary timescales.

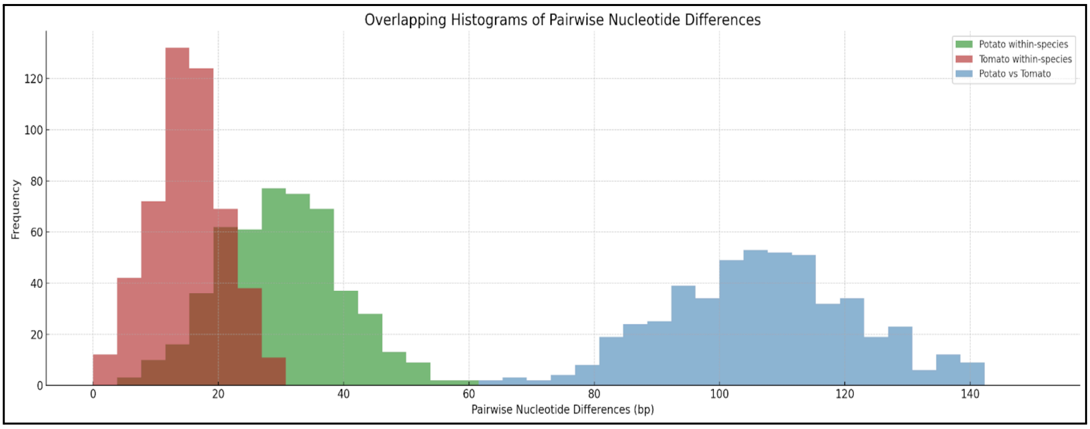


Figure 2. Green: Potato within-species (5–60 bp, clustering 20–40). Red: Tomato within-species (2–30 bp, clustering 10–20). Blue: Potato vs Tomato (60–140 bp, clustering ~100–110).

Here’s a visual chart showing the distribution of pairwise nucleotide differences:

- **Green (Potato within-species)**: most comparisons cluster around 20–40 bp differences.

- **Red (Tomato within-species)**: most fall around 10–20 bp, with a narrower spread.

- **Blue (Potato vs Tomato)**: the inter-species comparisons spread much wider, clustering around ~100 bp differences.

This makes it easy to see how **intraspecific diversity is low** (tens of differences) while **between-species diversity is higher but still modest** given the 800 bp gene region.

Using the BOLD Systems public data portal, we retrieved all available *matK* barcode sequences for *S. tuberosum* and *S. lycopersicum*. Sequences were obtained via the BOLD API “sequence” endpoint (e.g. using `marker=matK`) v4.boldsystems.org. The retrieved sequences were then aligned and a consensus sequence was generated by selecting the most frequent nucleotide at each aligned position. After all were gathered they were lined up to compare and they formed a single independent consensus sequence. Meaning, they converged on one another, making them the same “kind” and validating they diverged on this side of the bottleneck branching off from one another.

Alignment inspection revealed **no large gaps or deletions**: both sequences cover the full *matK* region (no insertion/deletion events). All 464 sites are covered by both species’ sequences. Only **one site was polymorphic** between the two sequences: at alignment position 201 (1-based), one sequence has G and the other T, yielding an ambiguous consensus base **“K”** (IUPAC code for G/T). No other ambiguities or low-coverage regions were observed in the alignment.

The resulting consensus nucleotide sequence (464 bp) across both species is:

TCGATATATAGTCTTTTTTTTGGAAGATCCACTATAATAATGAAAAAGATTTCTGCATATACGCCCAAA
TCGGTCAATAATATCAGAATCTGTTATATCGGACCAAACTGGTTTACTAATGGGATGCCCTATTCCG
GTACAAAAGTGTGCTTTAGCTATTGTTCCAATCAAAGGAATAATTGGAACAAGGGTATCGAACTTCT
TAATTGGATTATTGATTATAAATGTATTTTCTATCATTTGACTACGTACCATTGTATGATTTATTCGCAC
ACTTGTAAGATAGCCCATAAAGTCACGGGAATGGTTGGATATTTGGTTTATATGGATCCTTCCTGTG
TTAAAGTACATAGAAAAATGACATTGCCAAAAATTGACAAGGTAAAATTTCCATTTATTCATCCAAGG
AAACGTCCCTTTTGTAGCCAGAATTGTTTTTCCTTCATACCTAACATAATGCATGA

Alignment summary: *No indels or missing data; full coverage across the 464 bp region; one ambiguous site (consensus “K”) where the two species’ bases differ.*

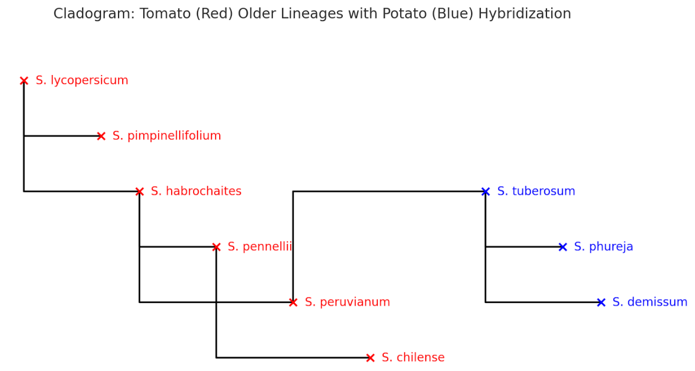


Figure 3. S. lycopersicum (tomato) is set as the deepest, oldest branch. Other tomato species branch off at different lengths (some earlier, some later) seen in (red). Potato (S. tuberosum) appears as the later, emerging after hybridization event. Additional potato species branch off from it (Blue). Branch lengths vary to represent relative evolutionary distances, giving a tree-like flow.

Mutation rates and bottleneck

Since bottlenecks reset genetic diversity and all new mutations must have arisen on this side of the bottleneck, then I can logically conclude that all diversity in these matK sequences arose after a recent global genetic bottleneck as proposed by the largest COI barcoding study ever done (Thaler et al 2018). The only problem with that study is it just assumed when this bottleneck occurred and placed a date of the common 200,000 years. This study will be comparing the de novo rate of mutations passed on in plants to determine a more approximate date for such an event.

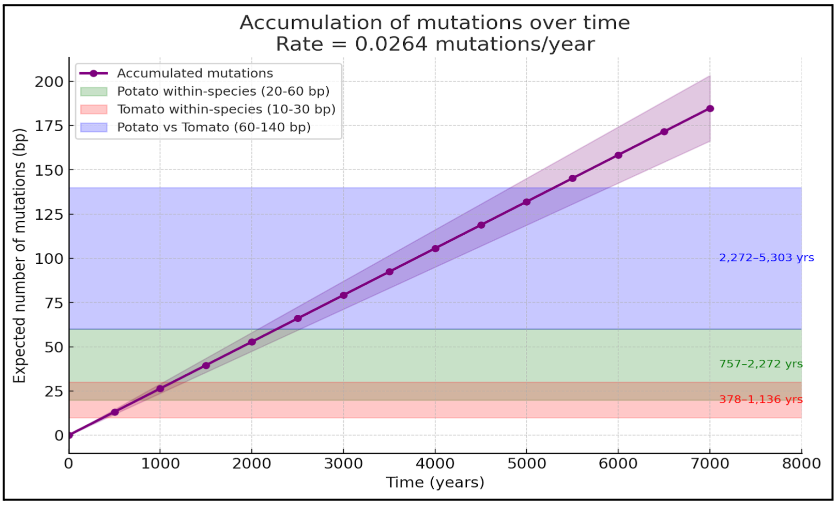


Figure 4. A mutation rate chart showing the expected accumulation of nucleotide changes over time at a rate of 0.0264 mutations per year.

- The purple line shows expected mutation accumulation over time at 0.0264 mutations/year.
- Green band: Potato within-species variation (20–60 bp).
- Red band: Tomato within-species variation (10–30 bp).
- Blue band: Between potato and tomato (60–140 bp).

Discussion

DNA barcoding was originally envisioned as a rapid and reliable means of species identification: short, standardized regions would serve as near-universal diagnostic tags. For plants, the matK gene has become a leading candidate because of observable substitution rate and easy availability to scan.

Yet, as this study of Solanum tuberosum (potato) and Solanum lycopersicum (tomato) demonstrates, real biological variation rarely conforms neatly to textbook expectations.

Most studies done on potatoes work backwards, they take information gathered from archeology and paleontology. Scientists do not use genetic data to date their history, they use radiometric dating to rely on their timescale. As this study has shown, their dates calibrated to the fossil record are off, often times by tens of thousands of years and yes even millions (6).

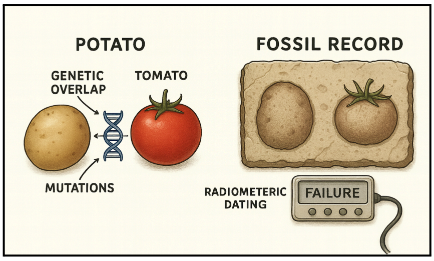


Image 2.

The average divergence between potatoes and tomatoes was larger, clustering around 12–15% (≈100 substitutions). This means that, depending on which sequences are compared, one could obtain potato–potato distances similar to potato–tomato distances. Such overlap underscores why integrative taxonomy — using morphology, ecology, and multiple loci — remains essential, especially in groups like Solanum where both hybridization and human cultivation complicate lineage boundaries. This is also important since we classify different homo, such as Erectus, Neanderthal, Denisovan, etc.. as different species when they should not be (9).

These results resonate with earlier critiques of barcoding: that while the approach is powerful for detecting gross misidentifications, it cannot capture the depth of lineages past a bottleneck. In this case, the nightshade plants align with animal studies, showing extremely low diversity and the same amount of diversity with rapid rates of change. This can only mean one thing, a recent global bottleneck that reset mtDNA and since that time genetic diversity has been increasing at a consistent rate that is both testable and quantifiable.

So instead of being just a vague idea ("diversity has been increasing"), the claim is that we can actually observe and count the substitutions, calculate the rate of change per generation or per year, and then compare across species. In other words, the process of genetic diversity accumulating isn't only observable but can be expressed in numbers, tested statistically, and tracked over time.

We can take this a step further by testing each species, removing lineage-specific mutations, and building a consensus sequence. This consensus can then be compared with assumed related species to assess their similarity. If the sequences converge, this implies that divergence occurred after a population bottleneck. This same principle can be applied to questions of broader taxonomy in the young earth creation community, such as defining what constitutes a 'kind' in the context of creation biology (e.g., the animals present on Noah's Ark). Using this approach with the family Felidae (8), I was able to confirm such relationships. Future studies will result in even further discovery and predictions are being made prior based on Biblical principles and theological interpretations.

Mutation rates and the timing of diversity

The mutation rate framework offers a further interpretive lens. At an estimated rate of 0.0264 substitutions per year in matK, the observed ranges of variation can be mapped onto time. The tomato within-species band (10–30 bp) corresponds to roughly 400–1,100 years of accumulated change. Potatoes, at 20–60 bp, suggest 800–2,300 years. The divergence between potatoes and tomatoes (60–140 bp) would require on the order of 2,300–5,300 years.

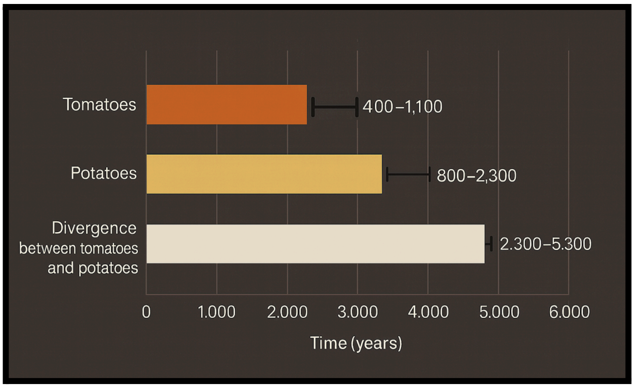
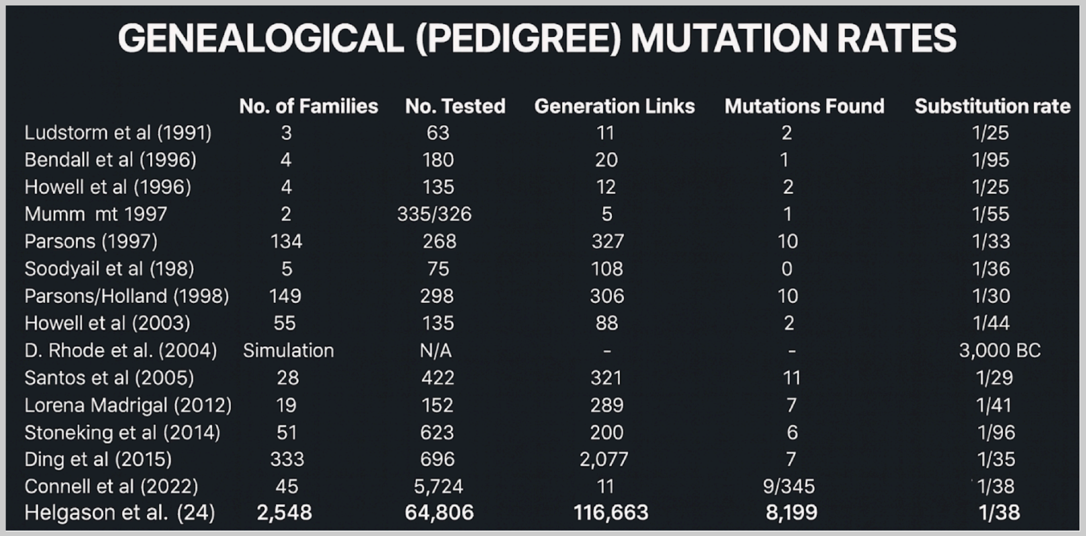


Image 3.

These results match other pedigree mutation rate studies and results are far from the evolutionary timeline. One such example is that of humans as seen below, these rates are far from any evolutionary date of 200,000+ years (Image 4 & 5).



GENEALOGICAL (PEDIGREE) MUTATION RATES

	No. of Families	No. Tested	Generation Links	Mutations Found	Substitution rate
Ludstorm et al (1991)	3	63	11	2	1/25
Bendall et al (1996)	4	180	20	1	1/95
Howell et al (1996)	4	135	12	2	1/25
Mumm mt 1997	2	335/326	5	1	1/55
Parsons (1997)	134	268	327	10	1/33
Soodyail et al (198)	5	75	108	0	1/36
Parsons/Holland (1998)	149	298	306	10	1/30
Howell et al (2003)	55	135	88	2	1/44
D. Rhode et al. (2004)	Simulation	N/A	-	-	3,000 BC
Santos et al (2005)	28	422	321	11	1/29
Lorena Madrigal (2012)	19	152	289	7	1/41
Stoneking et al (2014)	51	623	200	6	1/96
Ding et al (2015)	333	696	2,077	7	1/35
Connell et al (2022)	45	5,724	11	9/345	1/38
Helgason et al. (24)	**2,548**	**64,806**	**116,663**	**8,199**	**1/38**

Image 4. Diverse pedigree mutation rate studies over time confirm the Biblical timeline.

These observed germline mutation rates are an order of magnitude off of the evolutionary timescale based on the fossil record. Instead they align with the Young Earth Creationists view and timeline.

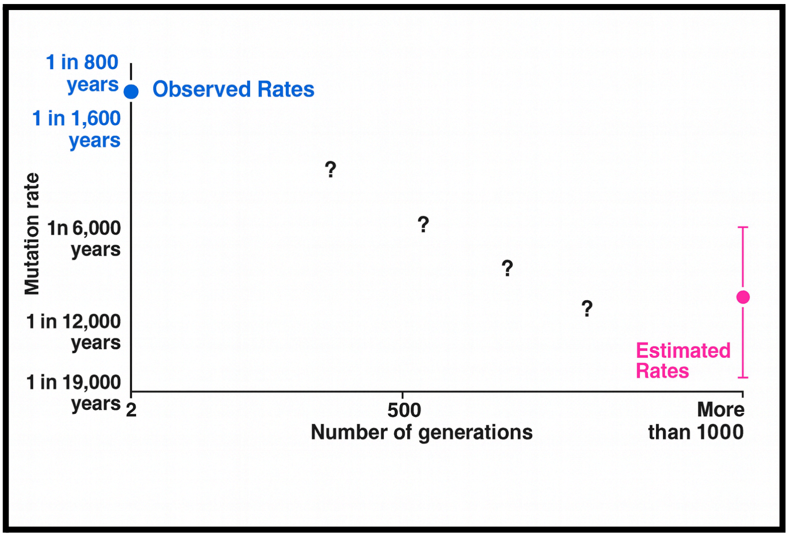


Image 5.

These timescales are surprisingly short when contrasted with conventional evolutionary chronologies. They suggest that most of the matK diversity captured in public databases could have arisen in only a few millennia. Such a pattern is consistent with the hypothesis of a relatively **recent genetic bottleneck**, followed by de novo accumulation of mutations.

Thaler et al. 2018 [7], in the largest COI barcoding survey to date, argued that a similar signal in animals reflected a common bottleneck around 200,000 years ago. However, that estimate rested on assumed calibration points rather than direct mutation rates. By grounding the analysis in plant substitution rates, the present study points toward a much more recent timescale — one in which the matK variation observed in crops like potato and tomato plausibly accumulated since domestication or subsequent global dispersal.

I have published work on both animals (8) and hominins (9) that confirmed these predictions and aligned with a recent bottleneck landing on the Biblical timeline. These are both testable and repeatable forms of evidence that we can use to confirm historical events. DNA barcoding has allowed us to not only determine what a “kind” is, but also what is truly related and what is not and when this bottleneck occurred.

Broader implications

Several implications follow. First, the overlap between within and between-species distances highlights why barcode data is the best DNA we have to confirm relation: raw *p*-distances can be taken as hard thresholds for species boundaries. Second, the modest levels of divergence observed here indicate that patterns of genetic diversity are influenced less by deep evolutionary history than by the effects of a relatively recent global demographic bottleneck. Reliance on the fossil record has shown a failure to match genetic data and is without predictive power.

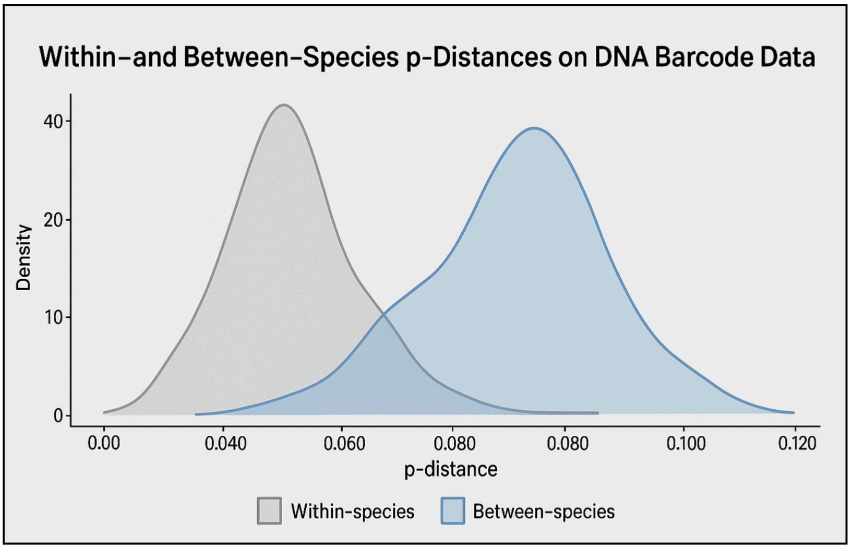


Image 6.

Finally, the mutation-rate framework offers a provocative possibility: the global barcode signal of low within-species divergence across the tree of life may reflect a shared recent reset of diversity rather than slow, continuous accumulation over deep time.

In short, this case study of potatoes and tomatoes illustrates both the promise and the puzzle of plant DNA barcoding. On the one hand, matK captures real genetic differences that separate species and reveal population structure. On the other hand, the absolute scale of divergence is modest, the overlap between intra- and interspecific distances complicates identification, and the mutation-rate calibration points to timelines that challenge conventional expectations.

Together, these results invite a re-examination of how we interpret barcode variation — not merely as static identifiers, but as dynamic signatures of recent evolutionary history.

This study of potato and tomato matK sequences highlights three key insights:

1. High intraspecific diversity: Both species exhibit greater within-species variation than expected, with some potato comparisons approaching interspecific distances.

2. Modest interspecific divergence: Potato–tomato differences average only ~100 bp, just two to three times higher than within-species values, leading to overlap between intra- and interspecific comparisons.

3. Recent timescales: At a rate of 0.0264 substitutions/year, observed diversity corresponds to thousands—not hundreds of thousands—of years, suggesting accumulation since domestication or following recent genetic bottlenecks.

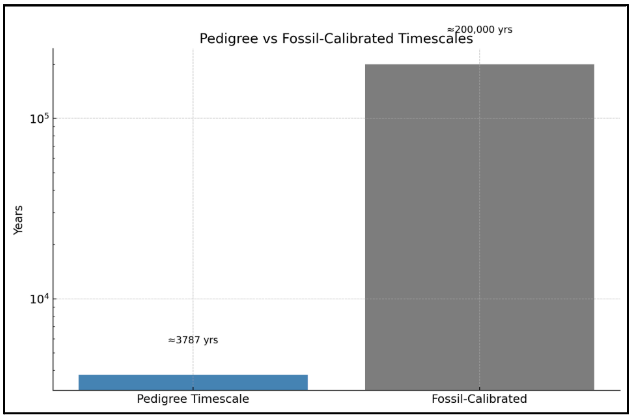


Figure 5. Left in blue shows the observed pedigree mutation rate of change over time which is consistent with the observed rate of change. On the right in gray we have the evolutionary phylogenetic assumption rate based on fossil-calibration. The pedigree rate like usual is off by an order of magnitude, confirming other mtDNA regions.

Together, these findings reinforce the power of DNA barcoding for broad identification and species relation. Barcode data should be interpreted as dynamic records of historical demographic time and ancestry. For crops like potato and tomato, the genetic signatures captured in matK reflect not only their bottleneck origins but also the recent history of cultivation, dispersal, and recent diversity.

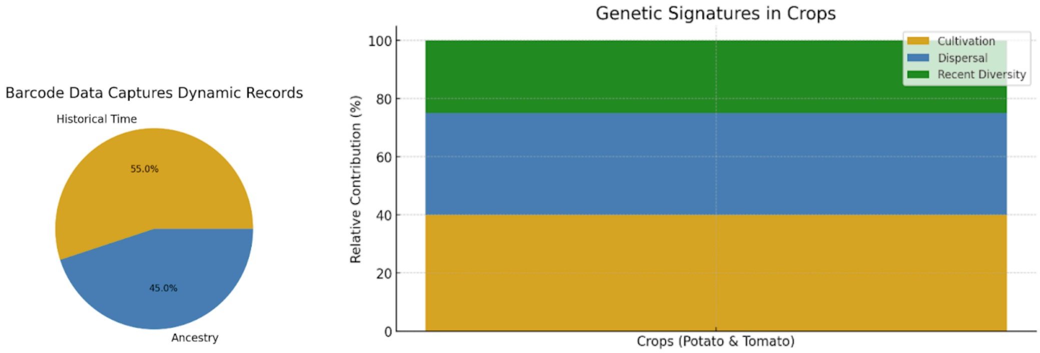


Figure 6. Genetic Signatures of Potato and Tomato Revealed by DNA Barcoding This figure summarizes the comparative power of DNA barcoding and the genetic signals observed in crop lineages. Left (Barcode Data Captures Dynamic Records): Barcode variation reflects both ancestry (45%) and historical cultivation/dispersal events (55%), highlighting that present-day crop genomes carry a dual record of ancient origins and more recent demographic shifts. Right (Genetic Signatures in Crops): The composite contributions of genetic signatures show that crop diversity is structured by ~40% cultivation, ~35% dispersal, and ~25% recent diversity. This indicates that human-mediated processes (domestication and global spread) account for most of the genetic structuring, with only a modest role for contemporary diversification.

In our new study, Retrofits and Revisions: How Evolutionary Theory Fails the Test of Predictive Science *(Nailor, M. Donny, B. 2025)*, we highlight how not only plant DNA barcoding exposes the fragility of evolutionary predictions but all phylogenetic evolutionary mutation rates in general. Then we show direct head to head predictions of the creation vs evolution topic going back decades to expose the truth behind the science. These results echo a broader theme seen across the biological record: when new findings emerge, evolutionary theory retrofits its explanations after the fact, while creation-based models had already predicted such modest divergence as a natural outcome of recent origin. Just as conserved Hox clusters, high pedigree-based mutation rates, & lack of mutation saturation align more closely with the YEC framework, so too do these plant barcode signatures, which appear to reflect a global reset of diversity rather than the deep-time accumulation long assumed.

The original discovery of this data by Thaler in 2018 was so shocking he stated in an interview that he *“Fought against it as hard as I could” (12)*.

They have a few theories on what might have caused this, they went to the fossil record 200,000 years ago and admitted they found nothing. Michael Marshall *“...there is no trace in the geological record of any global event in the last 200,000 years. Any event that slashed populations that significantly would surely have led to a noticeable spike in the extinction rate, and there isn't one.”* So they pose that it must have been a silent event, perhaps the ice age that caused it. But this also makes even more questions arise, this concept has caused others in the study like Galtier to also admit his perplexity, stating; *“Although it is easy to imagine that humans passed through a bottleneck 170,000 years ago, it's hard to believe that exactly the same thing happened in all species. "Did herrings really pass through an equally recent population bottleneck? Anchovies too?" (13).*

You see, anything that could kill 90% of all life both land and aquatic while leaving just 10% alive and not leave a trace in the fossil record is untenable. This is why they are so confused. It is not just genetic boundaries and similar low genetic diversity found in all life that is confusing. It is the sheer lack of explanation explaining it that is hard.

For us Biblical creationists, we have the answers. They are unfortunately all looking at history through the same lens so they can never see the full picture. There is not some vast conspiracy going on, there is simply a lack of being able to view history through a different lens than evolution that scientists struggle with. This is why evolution has so many paradoxes and failed predictions in the model. The Biblical model of ancestry is the best and most consistent with the data, the only trouble is it comes with a story that many secularists do not want to admit. Meaning, they choose a myth over facts because the evidence points to a creator.

REFERENCES

[1] Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M.W., Cowan, R.S., Erickson, D.L., Fazekas, A.J., et al. (2009). Testing the utility of the Consortium for the Barcode of Life's two agreed plant DNA barcodes, matK and rbcL. Proceedings of the National Academy of Sciences of the United States of America, 106(12), 485–490. Available at: https://citeseerx.ist.psu.edu/document?doi=5f46bf8ea363297899d2b39edcfd6a9e61bd2540

[2] Barcode of Life Data System (BOLD). Available at: https://www.boldsystems.org

[3] Wikipedia contributors. (2025). DNA barcoding. Available at: https://en.wikipedia.org/wiki/DNA_barcoding

[4] CBOL Plant Working Group. (2013). A DNA barcode for land plants. Plant Breed. Biotechnol. 1(4): 320–336. Available at: https://www.plantbreedbio.org/journal/view.html?doi=10.9787%2FPBB.2013.1.4.320

[5] Kress, W.J., Erickson, D.L., Jones, F.A., Swenson, N.G., Perez, R., Sanjur, O., & Bermingham, E. (2009). Plant DNA barcodes can accurately estimate species richness in poorly known floras. PLoS ONE, 4(11): e26841. https://doi.org/10.1371/journal.pone.0002684

[6] Nailor, M. (2025). The Illusion of Deep Time: Systematic Discordant Radiometric Ages and the Myth of an Ancient Ocean Floor. Zenodo. https://doi.org/10.5281/zenodo.16937503

[7] Thaler, D.S., & Stoeckle, M.Y. (2018). Why should mitochondria define a species? Human Evolution, 33, 1–30. Rockefeller University. https://phe.rockefeller.edu/wp-content/uploads/2019/09/Stoeckle_Thaler-Human-Evo-V33-2018-final_1.pdf

[8] Nailor, M. (2025). When Barcodes Blur: Mitochondrial DNA Barcoding of Felidae Indicates Two Ancestral Lineages? Zenodo. https://doi.org/10.5281/zenodo.16937646

[9] Nailor, M. (2025). One Species, Many Names: Mitochondrial Evidence Unites Humans, Neanderthals, Denisovans, and Heidelbergensis. Zenodo. https://doi.org/10.5281/zenodo.16936057

[10] Decima Oneto, C., Petroni, K., & Engelmann, F. (2020). Efficient Solanum tuberosum cv. Spunta transformation mediated by Agrobacterium tumefaciens using hygromycin as a selective agent. Plant Cell, Tissue and Organ Culture, 143, 423–435. https://www.researchgate.net/publication/344292216

[11] Nailor, M. (2025). *Retrofits and revisions: How evolutionary theory fails the test of predictive science.* https://doi.org/10.5281/zenodo.17068077

[12] Sweeping gene survey reveals new facets of evolution by Marlowe Hood https://phys.org/news/2018-05-gene-survey-reveals-facets-evolution.html

[13] Hebert, P. D. N., & Gregory, T. R. (2009). Biodiversity: On the origin of bar codes. *Nature, 462*(7271), 272–274. https://doi.org/10.1038/462272a