

Mathematical Validation of the Jacob Ancestry Simulator

Independent benchmark replication of Rohde, Olson & Chang (2004)

Validation report and reproducibility appendix | 31 August 2026

Executive result

An independently written implementation reproduced all 20 cells of Rohde et al. Tables 1 and 2 using the same 100 Monte Carlo runs per cell. Across the 40 reported MRCA/IA means, the mean absolute difference from the published means was 0.160 generations; the largest difference was 0.51 generations. Every replicated mean fell within one published run-level standard deviation of the corresponding published mean. Separate brute-force unit tests also passed.

Critical scope limitation

This report validates the mathematical ancestry mechanism and an independent implementation of the published Rohde-style benchmark. It does not prove that the biblical Jacob was historical, that his lineage survived, that historical migration rates equaled any chosen simulator parameters, or that the current production website code is byte-for-byte identical to this validation implementation. Those are separate historical, demographic, and software-audit questions.

Prepared as a reproducible technical appendix for scholarly review

1. What is being tested?

The question is not whether a simulation can be made to produce rapid common ancestry. The stronger test is whether an independently written program reproduces benchmark results for which the mathematical behavior is already known and peer reviewed.

Rohde, Olson, and Chang (Nature 431, 2004) studied genealogical ancestry, not merely genetic ancestry. In the one-population model, each person has two parents and the population is randomly mating. Chang's earlier theorem gives, asymptotically:

$$T_N / \log_2(N) \rightarrow 1$$

where T_N is the number of generations back to the genealogical most recent common ancestor (MRCA). The identical-ancestors (IA) point satisfies:

$$U_N / [1.77 \log_2(N)] \rightarrow 1$$

For the graph-structured model in Supplementary Methods A, let R be the graph radius, D the graph diameter, and Δ a graph-structure correction between 0 and 1. Then:

$$T_N / [(R + \Delta) \log_2(N)] \rightarrow 1$$

$$U_N / [(D + 1.77) \log_2(N)] \rightarrow 1$$

Consequently, doubling N should asymptotically increase T_N by $R + \Delta$ generations and U_N by $D + 1.77$ generations. This provides a direct calculator-level test independent of historical claims about Jacob.

2. Published benchmark structure

Rohde et al. reported Tables 1 and 2 for five population sizes (1,000 through 16,000), with 100 simulation runs per cell. Neighboring subpopulations exchanged one pair of migrants per edge per generation. The last columns of the published tables report the theoretical increment expected when population size doubles.

Graph	G	R	D	Delta	Theory ΔT / doubling	Theory ΔU / doubling
One node	1	-	0	-	1.0	1.77
Three fully connected	3	1	1	0.5	1.5	2.77
Five fully connected	5	1	1	0.75	1.75	2.77
Ten-node Fig. 1	10	3	5	0.0	3.0	6.77

Notation note: in the Nature paper, OCR can make the Greek Delta in $R + \Delta$ look like the Latin letter D. Delta is distinct from D, the graph diameter.

3. Independent implementation

The replication code was written independently rather than copied from the authors' original program. It tracks, for every past individual, exactly which present-day individuals descend from that person. A bit is assigned to each present-day person. Moving one generation backward, a parent receives the bitwise union of the descendant sets of every child that selected that parent.

Definitions used by the program:

- MRCA time T_N : the first past generation in which at least one individual is an ancestor of every present-day individual.
- IA time U_N : the first past generation in which every person is either an ancestor of everyone in the present population or an ancestor of nobody in the present population.
- Migration benchmark: for each undirected graph edge in each generation, one resident in each direction is assigned origin in the neighboring subpopulation; both parents are then sampled uniformly from that origin subpopulation. Parent choices are independent and with replacement.
- Replicates: 100 runs for each of 4 graph structures x 5 population sizes = 2,000 independent runs, matching the published number of runs per cell.
- Randomness: deterministic recorded seeds are used so every run can be repeated exactly.

Core bookkeeping, in pseudocode:

```
descendants[present_i] = bit(i)

for generation = 1, 2, ...:

    choose two parents for each child under the graph/migration rule

    previous[parent1] |= descendants[child]

    previous[parent2] |= descendants[child]

    if any previous[x] == EVERYONE and  $T_N$  not set: set  $T_N$ 

    if  $T_N$  set and every previous[x] is 0 or EVERYONE: set  $U_N$  and stop
```

4. Replication of Rohde et al. Tables 1 and 2

Across all 40 MRCA/IA means, the mean absolute error was 0.160 generations. MRCA RMSE was 0.174 generations; IA RMSE was 0.245 generations. The largest absolute difference was 0.51 generations. All replicated means were within one published run-level standard deviation of the published means.

One node

N	T_{pub}	T_{repl}	ΔT	U_{pub}	U_{repl}	ΔU
1000	10.8	10.71	-0.09	20.8	21.12	+0.32
2000	11.8	11.75	-0.05	22.6	22.52	-0.08
4000	12.8	12.85	+0.05	24.6	24.59	-0.01
8000	13.9	13.85	-0.05	26.5	26.47	-0.03
16000	14.8	14.96	+0.16	28.3	28.28	-0.02

Published values are means from Rohde et al.; standard deviations are available in the machine-readable CSV and original Tables 1-2. Replication uses 100 new runs per cell.

Three fully connected nodes

N	T_{pub}	T_{repl}	ΔT	U_{pub}	U_{repl}	ΔU
1000	14.0	13.91	-0.09	27.4	27.15	-0.25
2000	15.6	15.63	+0.03	30.3	30.38	+0.08
4000	17.1	17.15	+0.05	33.4	33.12	-0.28
8000	18.9	18.80	-0.10	36.2	36.38	+0.18
16000	20.3	20.49	+0.19	38.9	39.24	+0.34

Published values are means from Rohde et al.; standard deviations are available in the machine-readable CSV and original Tables 1-2. Replication uses 100 new runs per cell.

Five fully connected nodes

N	T pub	T repl	ΔT	U pub	U repl	ΔU
1000	14.0	13.95	-0.05	25.9	26.00	+0.10
2000	15.8	15.90	+0.10	28.9	28.89	-0.01
4000	17.8	17.72	-0.08	32.1	31.83	-0.27
8000	19.6	19.59	-0.01	35.3	35.04	-0.26
16000	21.5	21.51	+0.01	37.9	38.07	+0.17

Published values are means from Rohde et al.; standard deviations are available in the machine-readable CSV and original Tables 1-2. Replication uses 100 new runs per cell.

Ten-node graph shown in Fig. 1

N	T pub	T repl	ΔT	U pub	U repl	ΔU
1000	21.1	21.54	+0.44	46.3	46.65	+0.35
2000	24.3	24.07	-0.23	53.0	52.49	-0.51
4000	27.6	27.67	+0.07	59.8	59.83	+0.03
8000	30.5	30.42	-0.08	66.8	66.41	-0.39
16000	33.8	33.32	-0.48	73.6	73.31	-0.29

Published values are means from Rohde et al.; standard deviations are available in the machine-readable CSV and original Tables 1-2. Replication uses 100 new runs per cell.

5. Calculator-level doubling test

Because $\log_2(2N) = \log_2(N) + 1$, the asymptotic theory predicts a fixed increase each time population size doubles. This is an especially useful audit because a reviewer can check it with a calculator without trusting any code.

Graph	Theory ΔT	Published avg ΔT	Replication avg ΔT	Theory ΔU	Published avg ΔU	Replication avg ΔU
One node	1.00	1.000	1.062	1.77	1.875	1.790
Three fully connected	1.50	1.575	1.645	2.77	2.875	3.023
Five fully connected	1.75	1.875	1.890	2.77	3.000	3.018
Ten-node graph Fig. 1	3.00	3.175	2.945	6.77	6.825	6.665

Example: ten-node world graph. Published MRCA means are 21.1, 24.3, 27.6, 30.5, 33.8. The four increases are 3.2, 3.3, 2.9, 3.3, averaging 3.175 generations; theory predicts 3.00. Published IA increases average 6.825 generations; theory predicts 6.77. The independent replication averages 2.945 and 6.665, respectively.

6. Independent bookkeeping and failure-condition tests

A benchmark can match by accident if the ancestry bookkeeping is wrong in a compensating way. Therefore, a second test suite checks the lineage engine itself using deliberately small pedigrees that can be verified by a completely different, slow graph-traversal method.

- Bitset vs. explicit traversal: 250 randomly generated pedigrees ($N = 2-12$, depth = 1-10) were evaluated by both methods. Every ancestor-descendant set matched exactly.
- Zero-migration falsification test: two disconnected components of 40 individuals were simulated for 80 generations in 25 trials. A global common ancestor never appeared, as logic requires.
- Seed reproducibility: the same seed produced the identical (T_N , U_N) result (8, 17) in repeated $N = 250$ runs.

- MRCA persistence invariant: in 25 random $N = 100$ pedigrees, once a common ancestor appeared, at least one common ancestor remained in every earlier generation, as ancestry logic requires.

Unit-test console output:

```
bitset_vs_explicit_trials= 250  
disconnected_zero_migration_trials= 25  
seed_reproducibility_result= (8, 17)  
mrca_persistence_trials= 25  
  
ALL TESTS PASSED
```

7. What the result establishes - and what it does not

Established by these tests

The central genealogical-convergence mechanism is computationally reproducible. An independently written program recreates the published simple-graph results to a small fraction of a generation on average, and the observed doubling behavior tracks the analytical predictions. The ancestry-set bookkeeping also passes brute-force and logical failure tests.

Not established by these tests

The tests do not independently establish the historical existence of Jacob, the survival of his biological lineage, any specific ancient migration rate, or a numerical percentage of modern humans descended from him. Those conclusions depend on historical premises and demographic parameterization. A further software-audit step is also required before saying that the exact deployed website code has itself passed this test harness.

This distinction matters for scholarly presentation. A defensible statement is: "The simulator uses a genealogical mechanism whose benchmark behavior reproduces Rohde et al. and the associated asymptotic theory. Jacob-specific outputs are conditional model results under stated historical and demographic assumptions."

8. Suggested independent-review checklist

1. Verify the definitions of T_N (genealogical MRCA) and U_N (IA point) against Rohde et al. and Chang.
2. Verify the graph parameters R , D , and Δ for the three-, five-, and ten-node benchmarks.
3. Run `rohde_full_validation.py` independently and compare the generated CSV with the published Tables 1 and 2.
4. Run `jacob_validation_unit_tests.py` and inspect the brute-force traversal test.
5. Check the SHA-256 hashes in `validation_manifest.json` to ensure the reviewed files are unchanged.
6. Audit the production Jacob simulator code separately and confirm that its ancestry-propagation routine is equivalent to the validated routine.
7. Review historical inputs (population sizes, migration routes/rates, fertility, endogamy, dates) separately from the mathematics.

8. Require sensitivity analysis: show how Jacob-specific outcomes change under conservative, moderate, and liberal connectivity assumptions.

Appendix A. Hand calculations

A1. One population

For $N = 1,000$, $\log_2(1,000) = 9.966$. The asymptotic MRCA prediction is therefore about 10 generations. Rohde et al. report 10.8; the independent replication gives 10.71. For the IA point, $1.77 \times 9.966 = 17.64$ as the leading asymptotic term; finite- N effects raise the simulation mean to about 21 generations. Supplementary Methods A explicitly notes that lower-order additive offsets are not determined by the leading asymptotic theorem.

A2. Complete graphs

For a complete graph with $G > 1$ nodes, $R = 1$ and $D = 1$. Supplementary Methods A gives $\Delta = (G - 2)/(G - 1)$. Thus:

- $G = 3$: $\Delta = 1/2$, so $R + \Delta = 1.5$.
- $G = 5$: $\Delta = 3/4$, so $R + \Delta = 1.75$.
- For both graphs, $D + 1.77 = 2.77$.

A3. Ten-node graph

The Nature Figure 1 graph has radius $R = 3$ and diameter $D = 5$. Its Δ value is 0, so the expected MRCA increase per doubling is 3.00 generations. The IA increase is $5 + 1.77 = 6.77$ generations. These are exactly the final-column predictions printed in the Nature tables.

Appendix B. Reproducibility package

Reference run environment: Python 3.13.5, NumPy 2.3.5. The benchmark used 2,000 runs and completed in 264.1 seconds on the validation environment using 4 worker processes.

File	Purpose	SHA-256
rohde_full_validation.py	Full 20-cell benchmark replication	1010b060778e2199816a56f20ef3057310ce5bbe5b9465ba5938dd061d5fb184
rohde_full_validation_results.csv	Cell means/SDs vs. published values	86f9466a2c33d2c8cc81f40d4d77c6dd6bf15c74080ddb28513b5e55352960f8
rohde_full_validation_raw_runs.csv	All 2,000 individual simulation outputs	1ba4264c5d1a9f8306dc3c40924837bf006021a814535dc057140638bd8038f5
jacob_validation_unit_tests.py	Brute-force and logical invariants	4e1a1e4cf8b5b956499b129255bd416f95d3a14a2106b99dd5f1f43d76a8c24d
jacob_validation_unit_test_results.txt	Unit-test console output	afbbd54a42364004c890776b6f0c792cef09fcee72497fd6e4221776675bf3

Minimal commands:

```
python rohde_full_validation.py
```

```
python jacob_validation_unit_tests.py
```

References

1. Rohde, D. L. T., Olson, S., & Chang, J. T. (2004). Modelling the recent common ancestry of all living humans. *Nature* 431, 562-566. <https://doi.org/10.1038/nature02842>
2. Rohde, D. L. T., Olson, S., & Chang, J. T. Supplementary Methods A: Further Explanation and Derivations of Mathematical Results.
3. Rohde, D. L. T., Olson, S., & Chang, J. T. Supplementary Methods B: Further Details of the Computational Model.
4. Chang, J. T. (1999). Recent common ancestors of all present-day individuals. *Advances in Applied Probability* 31, 1002-1026. <https://doi.org/10.1017/S0001867800009587>
5. Hein, J. (2004). Human evolution: Pedigrees for all humanity. *Nature* 431, 518-519. <https://doi.org/10.1038/431518a>
6. Kelleher, J., Etheridge, A. M., Veber, A., & Barton, N. H. (2016). Spread of pedigree versus genetic ancestry in spatially distributed populations. *Theoretical Population Biology* 108, 1-12. <https://doi.org/10.1016/j.tpb.2015.10.008>
7. Ralph, P., & Coop, G. (2013). The Geography of Recent Genetic Ancestry across Europe. *PLoS Biology* 11(5), e1001555. <https://doi.org/10.1371/journal.pbio.1001555>