Estimating Mutation Parameters and Population History Simultaneously from Temporally-Spaced Genome Data

Arman Bilge, Tanja Stadler, Matthew Kearse, and Alexei J. Drummond

email: abil933@aucklanduni.ac.nz

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Motivation and Primary Challenges

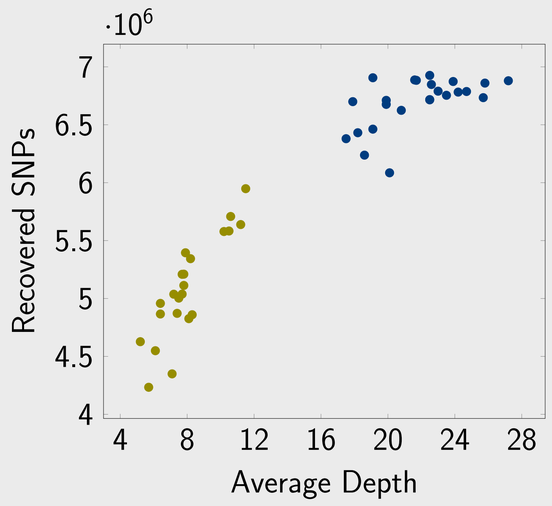
- Very feasible to sequence entire genomes
- More recently, even possible to recover **ancient genomes**
- **Temporally-spaced genome data**
- Opportunity to do **inference previously only possible for fast-evolving organisms** (e.g., viruses)
  - **mutation rate**
  - **population size** through time

But...

- Difficult to phase diploid genomes
- Recombination
- **Low coverage and sequencing error**, especially for ancient genomes
- **Cannot use existing Bayesian phylogenetic methods**

Sequencing Depth and Error

- Assume that all individuals have **about same number of SNPs**
- Average sequencing depth of sample is **correlated** with observed SNPs
- Our dataset approaches complete SNP recovery at **22x coverage**
- Ancient genomes are sequenced at lower depth and thus **missing many SNPs**
- Leads to **systematic bias** in estimates



Overview of Methodology

- Want to estimate mutation and population parameters θ from **pileup data**
- pileup data is **unsummarized, aligned reads** for each sampled individual
- To compute posterior need to **marginalize over individual's genotypes**
- **Computationally intractable** so use **importance sampling**
- The importance distribution assumes independence of individual's genotypes

$$P(\theta \mid D) = \sum_G P(\theta \mid G) P(G \mid D)$$
$$= \lim_{n\to\infty} \sum_{i=1}^{n} P\left(\theta \mid G^{(i)}\right) \frac{P\left(G^{(i)} \mid D\right)}{\hat{P}\left(G^{(i)} \mid D\right)}, G^{(i)} \sim \hat{P}(\cdot \mid D)$$

- Finally, use **standard MCMC to sample parameters** for a given genotype

$$P(\theta \mid G) \propto P(G \mid \theta) P(\theta)$$

Sampling an Individual's Genotype

- Want the genotype $g_1, g_2 \in \{A, C, G, T\}$ of individual at a position in its genome
- Data is the observed base calls at this position with their Phred quality scores

$$D = (b_q : b \in \{A, C, G, T\}, q \in \mathbb{N})$$

- Number of base calls $|D|$ is the sequencing depth at this position
- Sample genotype from the posterior distribution

$$P(g_1, g_2 \mid D) = \frac{P(D \mid g_1, g_2) P(g_1) P(g_2)}{P(D)}$$

- To compute $P(D \mid g_1, g_2)$ assume base calls are multinomially distributed with probabilities

$$P(b_q \mid g_1, g_2) = \frac{1}{2} P(b_q \mid g_1) + \frac{1}{2} P(b_q \mid g_2)$$

- Using the definition of a Phred quality score (assuming equal error rates for all bases)

$$P(g_i \mid b_q) = \begin{cases} 1 - 10^{\frac{-q}{10}} & \text{if } b = g_i \\ \frac{1}{3} 10^{\frac{-q}{10}} & \text{if } b \neq g_i \end{cases}$$

we have

$$P(b_q \mid g_i) = \frac{P(g_i \mid b_q) P(b_q)}{P(g_i)}$$

- Use empirical estimates for $P(g_i)$ and $P(b_q)$

Genotype Probability

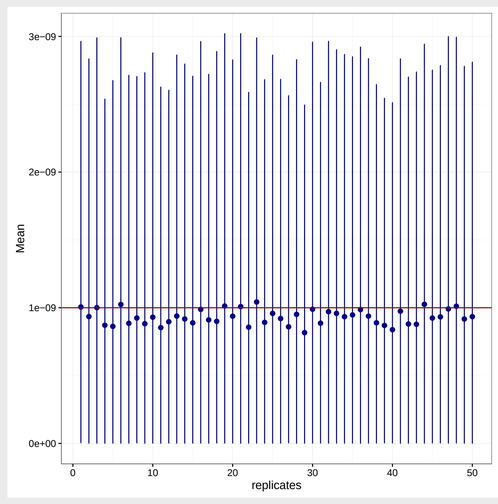
- Assumes that **sites are unlinked**; i.e., phylogenetically independent
  - **Models recombination** with **no dependence on correct phasing**
- Assumes a site is **biallelic**; i.e., has only two possible nucleotide states
  - Often true in practice
  - Can be handled rigorously with **ascertainment bias correction**
- Want the probability of all the individuals' genotypes by marginalizing over all phylogenies

$$P(G \mid \theta) = \int_T P(G \mid T, \theta) P(T \mid \theta) \, \mathrm{d}T$$

- Integral is **computed numerically** using similar technique to SNAPP [2]
- **Divide time into intervals** using sampling times and population change times
- Each interval i can be described by a **linear system of differential equations** $\mathbf{Q}_i$
- Solve each system by taking the **matrix exponential** $\exp \mathbf{Q}_i$ [1]
- **Can be done efficiently** by caching and reusing matrix exponentials

Simulation Study

- 8 diploid taxa, including 4 ancient individuals up to 50k years old
- Mutation rate $\mu = 10^{-9}$ s/s/yr
- HKY model with $\kappa = 5$
- Constant size population with $N_e = 3 \times 10^6$
- Simulated 50 datasets of 10^4 total sites
- Attempted to infer parameters
- **True values always within 95% HPD**
- Mean $\hat{\mu}$ within $\pm 1.8 \times 10^{-10}$ of true μ



Summary

- Fully Bayesian inference of mutation and population parameters from raw sequencing data
- Considers both **biological processes** and **practical problems**
- Critically, **avoids systematic bias** due to low coverage of ancient genomes
- **Combats intractability** using a variety of numerical and Monte Carlo techniques
- Looks promising but needs **comprehensive simulation study**
- Applying to a very exciting dataset!

Acknowledgements

- David Bryant and Remco Bouckaert
- Dong Xie
- Sankar Subramanian, Matthew Parks, and David Lambert
- New Zealand eScience Infrastructure
- SMBE16 Conference

References

1. AH Al-Mohy and NJ Higham. *SISC* 33.2 (2011). doi:10.1137/100788860
2. D Bryant et al. *Mol Biol Evol* 29.8 (2012). doi:10.1093/molbev/mss086
3. M Li et al. *Nucleic Acids Res* 32.17 (2004). doi:10.1093/nar/gkh850

Interested?

Download this poster.

doi:10.5281/zenodo.56495

Fork the source code.

git.io/vo7HR