Efficient inference of population size changes and mutation rates using the site frequency spectrum from large samples

Anand Bhaskar[1] and Yun S. Song[1,2]

[1]Computer Science Division and [2]Department of Statistics, University of California, Berkeley

Motivation

Null models in population genetics are used for, among many other things,

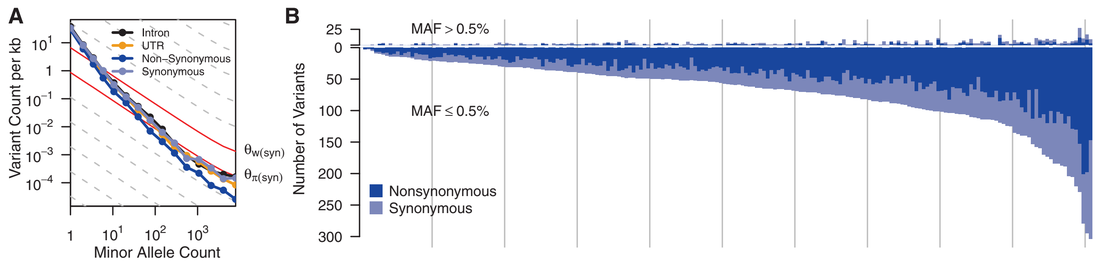
- Finding genomic regions under selection
- Genome-wide association studies
- Reconstructing demography
- Forensic applications

Commonly used null model assumptions:

- Constant effective population size
- Constant mutation rate across loci
- No population substructure

Several recent large-sample datasets have found an excess of rare variants compared to predictions from previously inferred demographic models

- Coventry et al. (2010) - 13,715 individuals at 2 genes
- Tennessen et al. (2012) - 2,440 individuals at 15,585 genes
- Nelson et al. (2012) - 14,002 individuals at 202 genes



(Figure from Nelson et al., 2012)

Recent exponential population growth can explain the abundance of rare variants

Need large sample sizes of tens of thousands of individuals to infer such growth

Problem

Data: Sample of n haplotypes at several unlinked loci

Summarize the data by the site frequency spectrum (SFS) at each locus

$\boldsymbol{o} = (o_1, o_2, \ldots, o_{n-1})$, where o_i = number of SNVs with i copies of the derived allele (can use the folded SFS if the ancestral allele is not known)

Goal: Given SFS from a *large sample* at many loci, infer *historical population size changes*

Previous approaches:

- Full likelihood approaches based on the sequentially Markovian coalescent (Li and Durbin, 2011; Sheehan et al., 2012). Cannot scale to large sample sizes needed to infer recent population expansion events
- Monte-Carlo coalescent-based methods for the SFS (Coventry et al., 2010). Unsuitable for gradient-based optimization, requiring grid search for the population size changes and mutation rates
- Diffusion-theoretic methods for the SFS. Handle rich class of demographic models, but can suffer from discretization accuracy issues, especially for large population sizes

Our method

Population size function $\{N(t), t \geq 0\}$ is not identifiable in general from SFS data.

Restrict attention to population size functions $N(t)$ in a restricted family of functions $\mathcal{F}$.

We consider the family $\mathcal{F}$ of piecewise exponential functions with M pieces.

- Time intervals for pieces $[t_i, t_{i+1})$, $0 \leq i < M$, $t_0 := 0, t_M := \infty$
- Population growth rates β_{i+1} in piece $[t_i, t_{i+1})$, $0 \leq i < M-1$
- $\beta_M := 0$ (constant ancestral population size)
- $N(t) = N(t_i)\exp(-\beta_{i+1}(t - t_i))$ for $t \in [t_i, t_{i+1})$

This family of functions can capture arbitrary number of epochs of exponential population growth, bottlenecks, and intervals of constant population size.

For a sample of size n and demographic model $\Phi \in \mathcal{F}$, define:

- $T^{\Phi}_{n,i}$ – waiting time in the coalescent while there are i lineages given that there are n lineages at the current time $t = 0$, $2 \leq i \leq n$
- $\tau^{\Phi}_{n,i}$ – total length of edges subtending i leaves in coalescent tree, $1 \leq i \leq n-1$
- L^{Φ}_n – total length of edges in coalescent tree

Mutations on edges subtending i lineages contribute to the i^{th} entry of the SFS.

Under the Poisson random field approximation of sites being completely unlinked within a locus, the log-likelihood can be written as,

$$\log \mathbb{P}(\boldsymbol{o} \mid \Phi, \theta) = \sum_{i=1}^{n-1} o_i \left(\log \mathbb{E}\left[\tau^{\Phi}_{n,i}\right] + \log \theta\right) - \frac{\theta}{2}\mathbb{E}\left[L^{\Phi}_n\right]$$

MLE for θ is θ^*, given by a generalization of Watterson's estimator to variable population models,

$$\theta^* = \frac{2\sum_{i=1}^{n-1} o_i}{\mathbb{E}\left[L^{\Phi}_n\right]}$$

MLE for Φ is Φ^* which minimizes the KL-divergence of the expected SFS from the observed SFS,

$$\Phi^* = \arg\min_{\Phi} \mathrm{KL}\left(\frac{\boldsymbol{o}}{|\boldsymbol{o}|} \middle\| \frac{\mathbb{E}\left[\boldsymbol{\tau}^{\Phi}_n\right]}{\mathbb{E}\left[L^{\Phi}_n\right]}\right)$$

Can compute $\mathbb{E}\left[\tau^{\Phi}_{n,i}\right]$ and $\mathbb{E}\left[L^{\Phi}_n\right]$ exactly and efficiently for models $\Phi \in \mathcal{F}$ using analytical theory of the SFS.

Gradient of log-likelihood with respect to model parameters can be calculated using automatic differentiation.

Computational details

Polanski and Kimmel (2003) showed that for arbitrary population size functions $N(t)$ and mutation rate θ,

$$\mathbb{E}[\tau_{n,i}] = \frac{\theta}{2}\sum_{k=2}^{n} a_{n,i,k}\mathbb{E}[T_{k,k}]$$

$$\mathbb{E}[L_n] = \frac{\theta}{2}\sum_{k=2}^{n} b_{n,k}\mathbb{E}[T_{k,k}],$$

where $a_{n,i,k}$ and $b_{n,k}$ are *universal coefficients* that do not depend on $N(t)$. All the dependence on the population model is captured in $\mathbb{E}[T_{k,k}]$,

$$\mathbb{E}[T_{k,k}] = \int_0^{\infty} t\,\frac{\binom{k}{2}}{N(t)}\exp\left(-\int_0^t \frac{\binom{k}{2}}{N(s)}ds\right)dt$$

Each coefficient $a_{n,i,k}$ and $b_{n,k}$ can be computed in $O(1)$ time by dynamic programming.

For functions $N(t)$ in our family of models $\mathcal{F}$, the integral in $\mathbb{E}[T_{k,k}]$ can be solved in terms of exponential functions and the exponential integral special function $\mathrm{Ei}(x)$, and can be computed in $O(M)$ time.

Evaluation

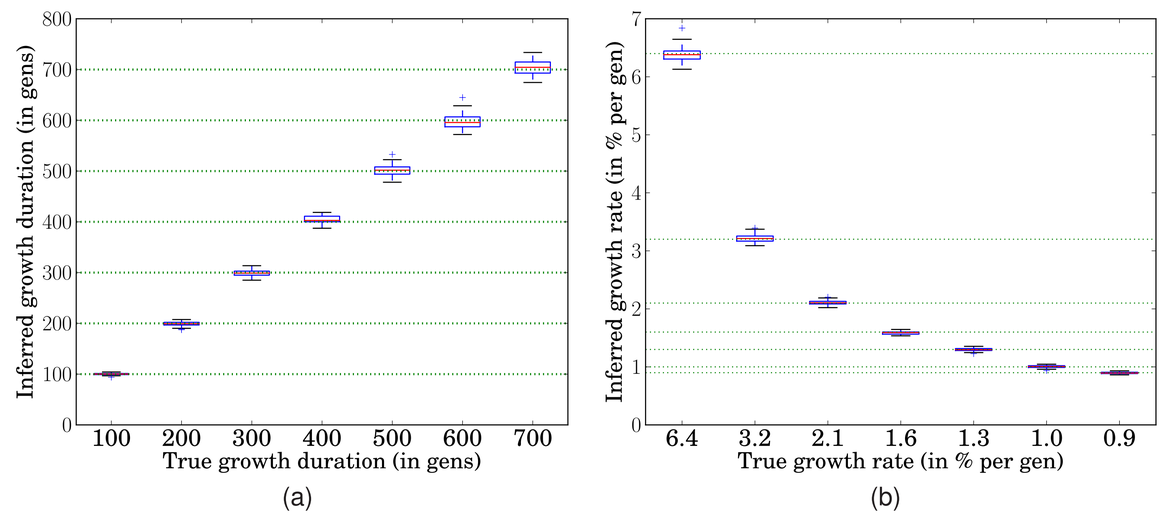


Figure: Box plots of the inferred (a) duration and (b) rate of exponential growth for 100 simulated datasets with 10,000 individuals at 200 loci for 7 population expansion scenarios. The dotted green lines indicate the true values for the inferred parameters.

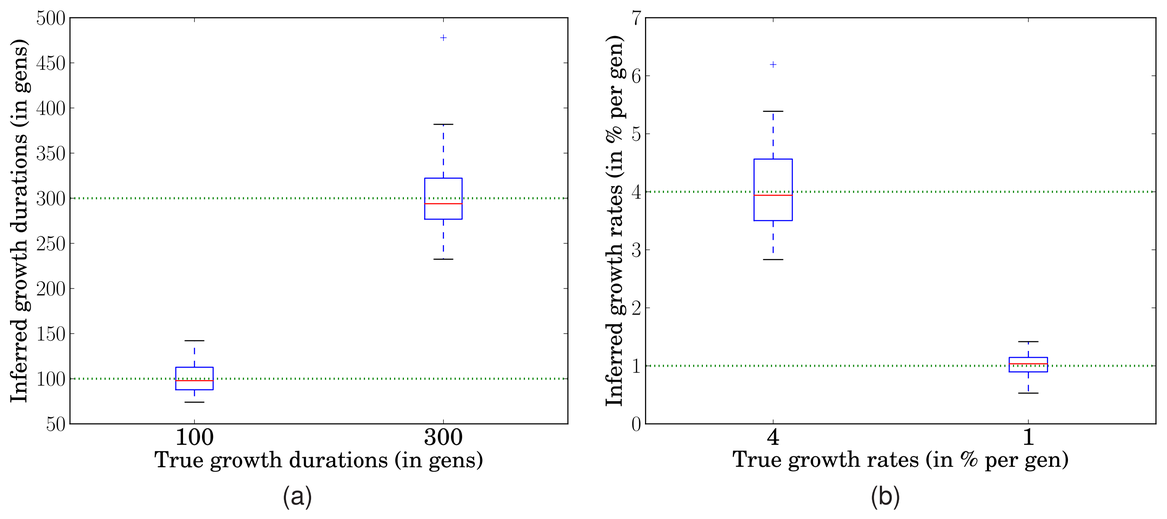
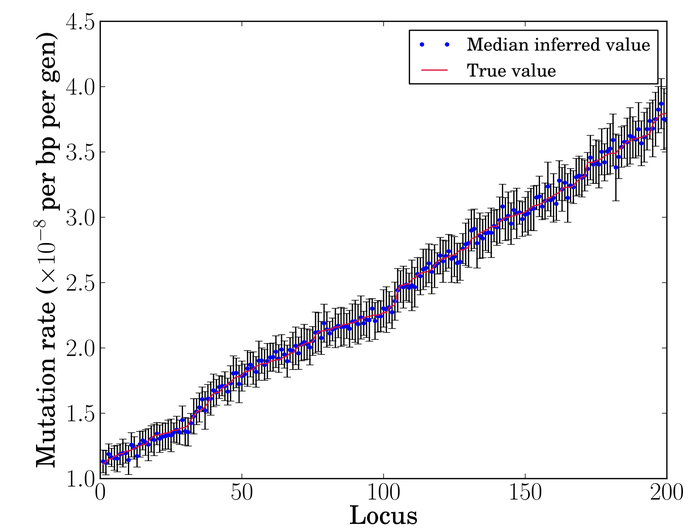


Figure: Box plots of the inferred (a) durations and (b) exponential growth rates for 50 simulated datasets with 10,000 individuals at 200 loci, with 2 epochs of recent exponential growth. The dotted green lines indicate the true values for the inferred parameters.



Inferred mutation rates at 200 loci with a recent epoch of exponential growth of 100 generations at 6.4% per generation. The loci are sorted by their true mutation rates.

Future research directions

- Incorporating multiple subpopulations and migration events
- Tradeoff in parameter estimation uncertainty as a function of sample size
- Model selection

References

Coventry, A., Bull-Otterson, L. M., Liu, X., et al. 2010. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, 1 131.

Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357) 493–496.

Nelson, M. R., Wegmann, D., Ehm, M. G., et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090) 100–104.

Polanski, A. and Kimmel, M. September 2003. New explicit expressions for relative frequencies of Single-Nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1) 427–436.

Sheehan, S., Harris, K., and Song, Y. S. 2013. Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, Accepted pending minor revision.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090) 64–69.

This research is supported in part by an NIH grant R01-GM094402, and a Packard Fellowship for Science and Engineering

{bhaskar,yss}@eecs.berkeley.edu