LDpop: 2-locus likelihoods for recombination map estimation under variable population size

John Kamm[1,*], Jeffrey Spence[2,*], Jeffrey Chan[1], and Yun S. Song[1,3]

[1]EECS, [2]Computational Biology @ UC Berkeley; [3]Math, Biology @ UPenn
[*]contributed equally

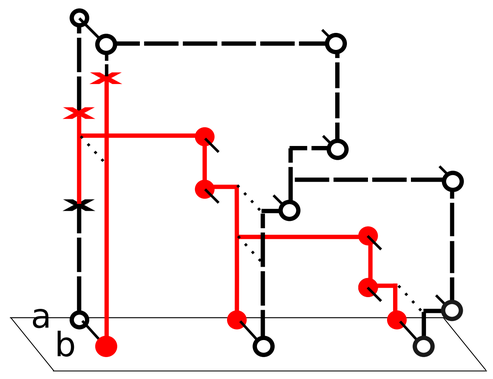
https://github.com/popgenmethods/ldpop

Pairwise likelihoods

- Recombination maps often estimated by combining likelihoods at pairs of SNPs within a composite likelihood:

$$\mathcal{L}(\rho) = \prod_{a,b:\ 0<b-a<W} \mathbb{P}(\mathbf{n}_{a,b}; \rho_{a,b})$$

- Previous methods to compute $\mathbb{P}(\mathbf{n}_{a,b}; \rho_{a,b})$ assume constant population size, leading to inaccuracies such as spurious recombination hotspots.
- Our new method **LDpop** computes pairwise likelihoods under changing population size, **improving the accuracy** of inferred maps.



$\mathbb{P}(\mathbf{n}_{a,b}; \rho_{a,b})$ the likelihood at loci a, b under map $\boldsymbol{\rho}$

$\rho_{a,b}$ = recombination distance

$\mathbf{n}_{a,b} = \{n_{\circ\bullet} = 1, n_{\bullet\circ} = 2\}$

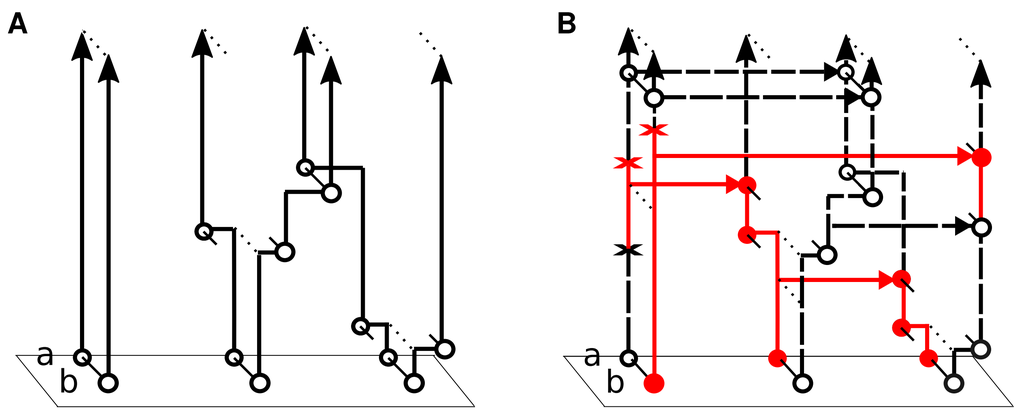
An exact formula for $\mathbb{P}(\mathbf{n}; \rho)$

We compute $\mathbb{P}(\mathbf{n}_{a,b}; \rho_{a,b})$ by constructing a process $\{\tilde{\mathbf{M}}_t\}$ that contains the 2-locus coalescent *embedded* within it.

- $\{\tilde{\mathbf{M}}_t\}$ constructed in 2 steps: step **A** constructed *backwards-in-time*, step **B** constructed *forwards-in-time*.
- $\mathbb{P}(\tilde{\mathbf{M}}_0)$ given by a product of sparse matrix exponentials:

$$\left[\mathbb{P}(\tilde{\mathbf{M}}_0 = \mathbf{n}; \rho)\right]_{\mathbf{n}} = \left(\prod_{d=1}^{D} e^{\tilde{\Lambda}_d}\right) \mathbf{v}$$

where D = number of population size changes.

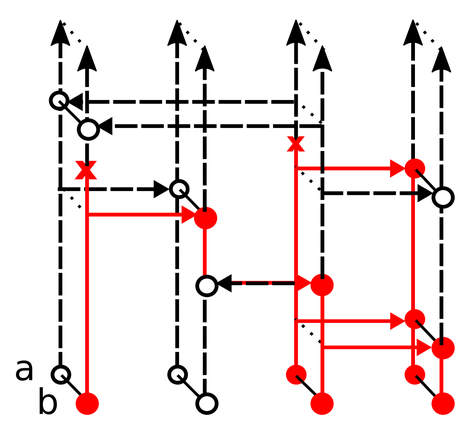


A fast approximation $\hat{\mathbb{P}}(\mathbf{n}; \rho)$

- We also develop a fast approximation based on a Moran model with $N \geq n$ particles.
- Computed with a similar product of sparse matrix exponentials,

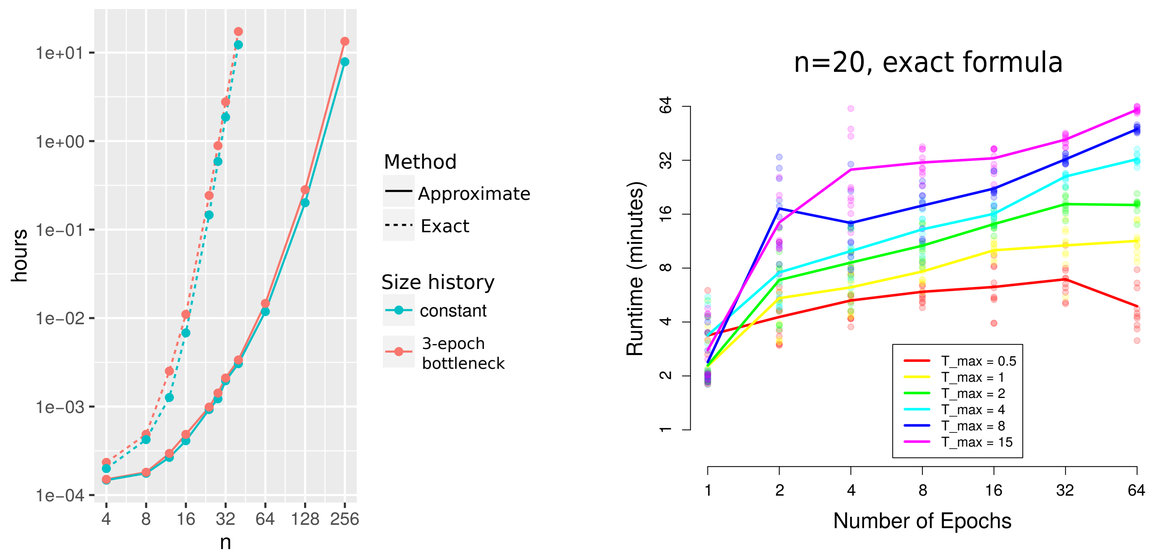
$$[\hat{\mathbb{P}}(\mathbf{n}; \rho)]_{\mathbf{n}} = \left(\prod_{d=1}^{D} e^{\hat{\Lambda}_d}\right) \mathbf{v}$$

but the matrices $\hat{\Lambda}_d$ are much smaller.



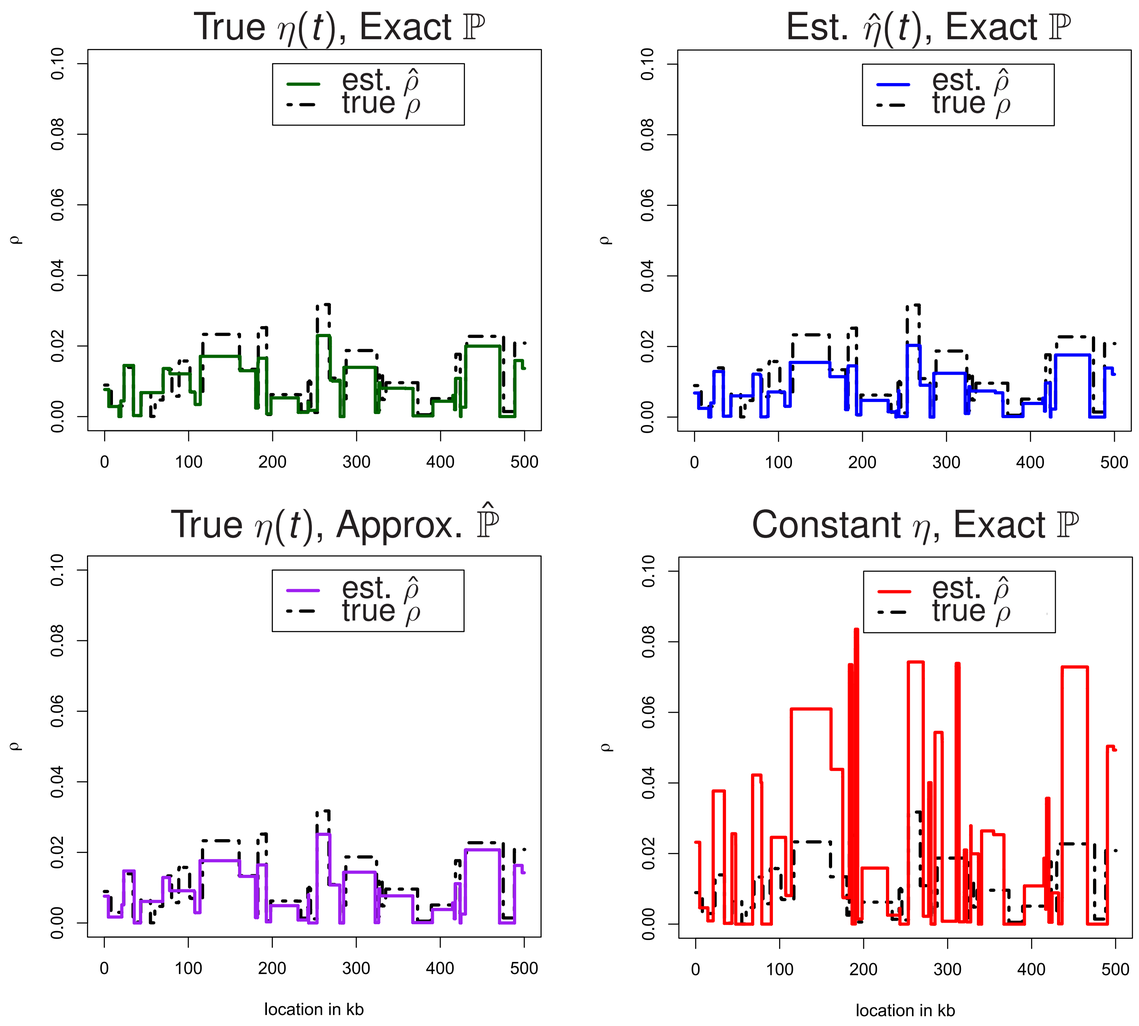
The error in the approximation disappears as $N \to \infty$. In practice, the approximation is good even when $N = n$.

Runtime to compute lookup tables



Time to compute $[\mathbb{P}(\mathbf{n}; \rho)]_{|\mathbf{n}|=n,\rho\in\{0,1,\ldots,100\}}$ with 24 CPUs.

Simulation study: using LDpop improves accuracy

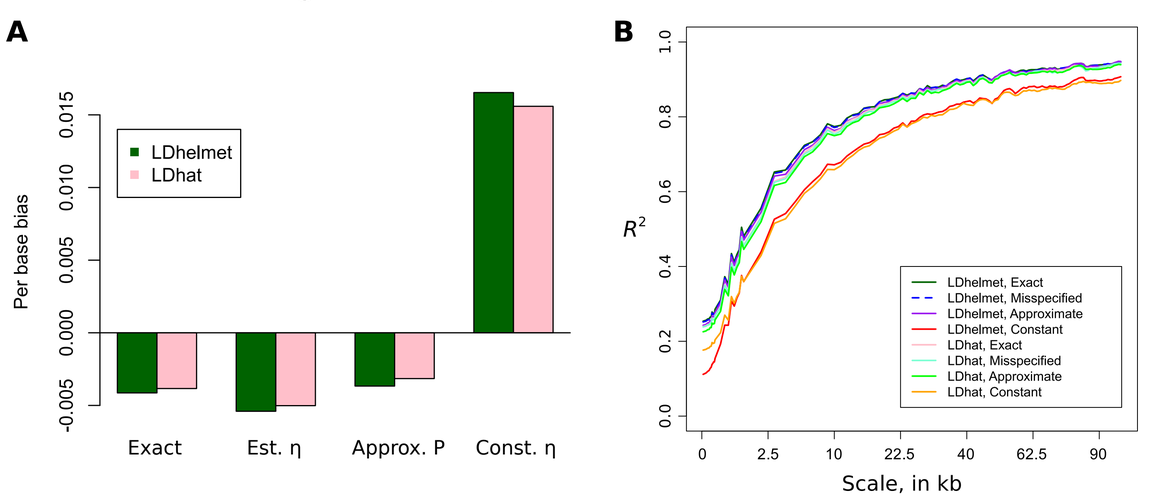


Simulated datasets with $n = 20$ haplotypes under demography $\eta(t)$

$$\eta(t) = \begin{cases} 10^5, & t < 10^3 \text{ gens ago,} \\ 10^2, & 1.16 \times 10^3 > t \geq 10^3 \text{ gens ago,} \\ 10^3, & t \geq 1.16 \times 10^3 \text{ gens ago} \end{cases}$$

Inferred $\hat{\boldsymbol{\rho}}$ by computing lookup tables $[\mathbb{P}(\mathbf{n}; \rho)]_{\mathbf{n},\rho}$ under LDpop and providing them to composite likelihood methods LDhat, LDhelmet.

- Using the true or estimated demography $\eta(t)$ is more accurate than assuming constant η.



Per-base bias of $\hat{\rho}$ (left), and R^2 of ρ with $\hat{\rho}$ at different scales (right).

This research is supported by NIH grant R01-GM108805, NIH training grant T32-HG000047, and a Packard Fellowship for Science and Engineering.

{jkamm,spence.jeffrey,chanjed,yss}@berkeley.edu