MAPS: recent Migration And Population Size surface estimation

Hussein Al-Asadi, Desislava Petkova, John Novembre, Matthew Stephens

1 Methods

MAPS assumes a population genetic model consisting of a d x d grid of demes with symmetric migration and per the stepping stone assumption, only neighboring demes are connected. Like EEMS, the density of the grid is prespecified by the user with the consideration that the computational complexity is $O(d^3)$. We use Bayesian inference to estimate the MAPS parameters: m and q. Its key components are the likelihood, which measures how well the parameters explain the observed data, and the prior, which captures the expectation that m and q have some spatial structure (in particular, the idea that nearby edges will tend to have similar migration and coalescent rates).

1.1 The Likelihood

Let $N^u_{i,j}$ be the number of segments greater than u centimorgans between individuals i and j that have not experienced any intervening recombination. These UnRecombined segments (UnR) are commonly referred to as "IBD" segments; however, we do not refer to them as IBD segments to avoid confusing them with the myriad other definitions of IBD. As u increases, UnR segments have shorter coalescent times and become increasingly sensitive to recent population structure and in practice, become a function of only recent population structure. In this study, we focus on characterizing population structure in the recent past (e.g. < 1500 years) and accordingly focus on UnR segments with short coalescent times which we arbirtarily designate as UnR segments greater than 2cM. Fortuitously, methods for calling UnR segments at this range (aka IBD calling) have high power and low false positive rates (Browning & Browning, 2007; Lowe et al, 2008; Raph & Coop, 2008 and Palamara et al., 2012).

Let $\theta = (m, q)$ encode the parameters of the demographic model, which are the migration rates and population sizes of a d by d grid of demes. Like Petkova et al 2015, we only allow migration between adjacent demes to enfore the isolation by distanc assumption which is widely observed in data.

First we compute the expectation of $N^u_{i,j}$. Following Palamara et al. 2012,

$$E[N^u_{i,j}] \approx \frac{L\, E[f|\theta]}{E[s|\theta]} \tag{1}$$

where $E[f|\theta]$ is the expected fraction of the genome that lies in UnR segments greater than ucM, L the length of the genome in units of cM, and $E[s|\theta]$ the expected size of a UnR segment conditional on it is at least ucM. To gain intuition, it is helpful to see that equation (1) equates to saying that the average size of each UnR segment ($E[s|\theta]$) multiplied by the average number of segments ($E[N^u_{i,j}]$) is approximately the total length of the genome that lies in UnR segments ($L\, E[f|\theta]$).

Furthermore, following Palamara et al. 2012,

$$E[s|\theta] = \frac{\int_u^\infty p(l|\theta)dl}{\int_u^\infty p(l|\theta)/l\, dl} \tag{2}$$

$$E[f|\theta] = \int_u^\infty p(l|\theta)dl \tag{3}$$

where $p(l|\theta)$ is the probability that a randomly chosen base pair in the genome is contained with an UnR segment of at least ucM. Combining both equations together,

$$E[N^u_{i,j}|\theta] \approx L \int_u^\infty p(l|\theta)/l\, dl$$

Expanding $L \int_u^\infty p(l|\theta)/l\, dl$,

$$L \int_u^\infty p(l|\theta)/l\, dl = L \int_u^\infty \int_0^\infty p(l,t|\theta)/l\, dt\, dl = L \int_0^\infty p(t|\theta) \int_u^\infty p(l|t)/l\, dl\, dt$$

where $p(l,t|\theta) = P(l|t,\theta)p(t|\theta) = p(l|t)p(t|\theta)$.

$p(l|t)$ is the probability of observing a UnR segment of exactly length l given a coalescent event at time t, this quantity can be derived by considering the probability of no recombination for t generations (see Hein et al, 2005) for a pair of individuals. Namely, given a site a, the probability of no recombination in t generations for x base pairs to the right of a is simply e^{-rtx} where r is the recombination rate for one of the pair and e^{-2rtx} for both individuals. Consider the length of no recombination to left of a (X_L) and to the right of a (X_R), then $l = X_L + X_r$ is a convolution of two exponential random variables. Therefore,

$$p(l|t) = 4r^2t^2le^{-2trl} \tag{4}$$

Thus, the inner integral $\int_u^\infty p(l|t)/l\,dl$ can be computed analytically as $2rtle^{-2tru}$. The next quantity to consider is $p(t|\theta)$, where θ encodes the demographic model. Under the structured coalescent,

$$p(T_{i,j} = t|\theta) = \sum_k q_k(e^{-Rt})_{(I(i),I(j))\to(k,k)} \tag{5}$$

which can be derived by considering all possible demes in which lineages i and j can coalesce. q_k is the coalescent rate of deme k, $I(i)$ indicates the deme which individual i belongs to, and R is the instantaneous rate matrix of the structured coalescent process for a sample size of two, see (Wakeley, 2008) for a detailed explanation of R. R is a matrix of size $O(d^2)$ by $O(d^2)$ and as a result becomes practically impossible to exponentiate for large d. Instead, we approximate the coalescent process as a random walk with absorption where,

$$p(T_{i,j} = t|\theta) \approx \sum_k q_k(e^{-Mt})_{I(i)\to k}(e^{-Mt})_{I(j)\to k}$$

M is a d by d matrix encoding a random walk on d vertices such that,

$$M_{\alpha,\beta} = m_{\alpha,\beta} \tag{6}$$

$m_{\alpha,\beta}$ is the migration rate between individual deme α and deme β and $m_{\alpha,\beta} = 0$ if demes α and β are unconnected. Or equivalently,

$$P(t) \approx e^{-Mt}Qe^{-Mt}$$

where $Q = diag(q_1, ..., q_d)$ and $P(t)_{I(i),I(j)} = p(T_{i,j} = t|\theta)$

We found this random walk approximation to work very well (see Supplementary Fig 1). Effectively, we are assuming that coalescent events before time t do not occur which is approximately true for small t and small q. Having all the components of equation (1), we are left to evaluate a single integral which we do by gaussian quadrature. The quadrature routine only evaluates $p(T_{i,j} = t|\theta)$ at small t because the probability of observing long UnR segments with deep coalescent times is effectively zero, and therefore do not contribute any mass to the integral.

Now that we can efficiently compute $E[N^u_{i,j}]$, we can compute the distribution $N^u_{i,j}$

$$N^u_{i,j}|\theta \sim Pois(E[N^u_{i,j}|\theta]) \tag{7}$$

which we found to be an excellent fit to the POPRES data, consistent with Ralph & Coop 2008 and Palamara et al 2012.

The composite likelihood of the data is,

$$p(D = N^u_{1,2}, N^u_{1,3}....N^u_{n-1,n}|\theta) = \prod_{i<j} p(N^u_{i,j}|\theta)$$

Composite likelihoods have been widely used in population genetics and have been used in Bayesian population genetic methods (see Lawson et al 2012 for an example). However, using the composite likelihood lead to very poor mixing because the composite likelihood assumes there are $\binom{n}{2}$ independent data points and as a result becomes overly concentrated around the mle. To improve mixing, we downweighted the composite likelihood so that there are actually $(n-1)$ independent data points. As a result, the likelihood became less sharp around the mle with the mle remaining identical. More specifically, the full composite log-likelihood (which we assumed to be Poisson) is,

$$l(\lambda) = \binom{n}{2}(log(\lambda)\bar{x} - \lambda)$$

which we then altered to

$$(n-1)(log(\lambda)\bar{x}-\lambda)$$

to reflect that we really just have $O(n)$ data points.

1.2 Computing $E[N_{i,j}^R]$

Previously, we computed the expected number of UnR segments greater than u cM. It is also helpful to visualize UnR segments in a range $R = (u, v)$, this is simply,

$$E[N_{i,j}^R] = E[N_{i,j}^u] - E[N_{i,j}^v]$$

1.3 The Prior

We use the EEMS prior in Petkova et al 2015, except that we estimate an additional overall mean rate for the coalescent rates wheras Petkova et al fixed the analogous parameter to one.

1.4 Markov chain Monte Carlo estimation

Like EEMS, MAPS uses MCMC to estimate the migration and coalescent rates by sampling from their posterior distribution given the data.

1.5 Computational time

The computational cost of MAPS is cubic in the size of the population grid, and the current implementation does not scale well beyond 400 demes.