Proposal: Extending dnadot with a Dynamic Beta-binomial Model for Robust Population Size Estimation

Mark Richardson

2025-09-01

1. Background: The Current `dotR` Framework

The `dnadot` method is a **Minimum Distance Estimation (MDE)** framework designed to estimate census population size (N_c) from genetic data. The core logic is as follows:

1. **Observe**: An empirical distribution of allele counts (O) is generated by analyzing many jackknife subsamples of the input data.
2. **Hypothesize**: A grid of possible parameter pairs, census size (N_c) and allele frequency (p), is created.
3. **Expect**: For each hypothesized (N_c, p) pair, a theoretical or **Expected** distribution of allele counts (E) is calculated.
4. **Compare**: A discrepancy metric (e.g., Wasserstein distance) is used to measure the "distance" between the Observed (O) and Expected (E) distributions.
5. **Estimate**: The (N_c, p) pair that minimizes this distance is chosen as the best estimate.

The key difference between the methods lies in the mathematical formulas used in steps 3 and 4.

The Expected Distribution: The Hypergeometric Model Both the `wasserstein` and `bhd` methods assume that the population is panmictic (randomly mating) and that sampling individuals is akin to drawing marbles from an urn without replacement. Therefore, the Expected distribution (E) is calculated using the hypergeometric probability mass function.

The probability of observing exactly k copies of a target allele in a subsample of n gene copies is given by:

$$P(k|n, K, N) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Where the parameters are defined as: * N: The total number of gene copies in the hypothesized population. This is calculated as $N = 2 \times N\_c$ for a diploid organism. * K: The total number of copies of the target allele in the hypothesized population. This is calculated as $K = N \times p$. * n: The total number of gene copies in the jackknife subsample. * k: The number of copies of the target allele, for which we are calculating the probability. For each point on the (N_c, p) grid, the `dotR` backend calculates this probability for all possible values of k (from 0 to n) to construct the full Expected distribution, E.

The Comparison Metrics The `wasserstein` Method: The Wasserstein distance (also known as the Earth Mover's Distance) is a powerful way to measure the difference between two probability distributions. Instead of comparing the probabilities at each point directly, it compares their **Cumulative Distribution Functions (CDFs)**. The CDF for a distribution at a point x is the sum of all probabilities up to that point: $CDF(x) =$
$sum\_i = 0^x P(i)$.

The Wasserstein distance, D_W, is calculated as the sum of the absolute differences between the CDFs of the Observed (O) and Expected (E) distributions across all possible allele counts k:

$$D\_W = sum\_k = 0^n |CDF\_O(k) - CDF\_E(k)|$$

Intuitively, this can be thought of as the total "work" required to transform one distribution into the other, making it a very robust measure of their difference. The `dnadot_snp` function finds the (N_c, p) pair that minimizes this D_W value.

The `bhd` Method (Buffered Histogram Discrepancy): The BHD method uses a more complex comparison that attempts to account for the inherent uncertainty in the observed data. It recognizes that an allele count from a single subsample is a noisy observation. Instead of treating each observed count k_obs as a single, discrete event, it "buffers" it by creating a small, localized distribution around it.

Buffering the Observation: For each jackknife subsample that yields an observed count k_obs, a triangular kernel is created. With a buffer width of w, the weight assigned to a bin k is given by:

$$weight(k, k\_obs, w) = max(0, w - |k - k\_obs| + 1)$$

This creates a "fuzzy" or smoothed representation of the single observation.

The final buffered observed distribution, O, is the average of these triangular kernels from all the jackknife subsamples. This results in a much smoother distribution than the raw histogram of counts. The BHD discrepancy, D_BHD, is then calculated as the sum of the absolute differences between this buffered observed distribution (O) and the standard expected hypergeometric distribution (E).

---

2. A Potential Solution: A Beta-binomial Model for Overdispersion

The v0.2.0 implementation of `dotR`relies on the **hypergeometric distribution** to calculate the Expected distribution. This is mathematically sound *if* the population is a single, panmictic unit. However, this assumption is often violated in natural populations due to cryptic structure, inbreeding, or a **Wahlund effect**. This causes the observed allele counts to be **overdispersed** (have higher variance) than predicted by the hypergeometric model, which can bias the estimate of N_c. To account for this overdispersion, we propose extending `dnadot` to use the **Beta-binomial distribution**.

Conceptual Framework: An "Urn of Urns"

The Beta-binomial model is more flexible and biologically realistic. It does not assume a single, fixed allele frequency (p). Instead, it models p as a random variable that is itself drawn from a Beta distribution.

- **Hypergeometric Model**: A single urn with a fixed proportion of red and blue marbles.
- **Beta-binomial Model**: A collection of many sub-urns, each with a slightly different proportion of red and blue marbles. A sample is drawn by first randomly selecting an urn, and then drawing marbles from it.

This "urn of urns" scenario directly models the variance in allele frequencies caused by population structure or non-random mating.

The Key Parameter: F_{IS} as the Measure of Variance

The Beta-binomial distribution introduces a third parameter, often denoted θ, which quantifies the overdispersion. In population genetics, when analyzing a single population sample, the most direct analogue for this parameter is the **inbreeding coefficient,** F_{IS}.

F_{IS} measures the deviation from HWE within a sample and is calculated as:

$$F_{IS} = 1 - \frac{H_o}{H_e}$$

Where:

- H_o is the observed heterozygosity.
- H_e is the expected heterozygosity ($2pq$).

A positive F_{IS} indicates a deficit of heterozygotes—the classic signature of population structure—and can be used to parameterize the Beta-binomial model.

---

3. Formal Mathematical Approach

The probability mass function of the Beta-binomial distribution for observing k successes in n trials is:

$$P(k|n, \alpha, \beta) = \binom{n}{k} \frac{B(k+\alpha, n-k+\beta)}{B(\alpha, \beta)}$$

Where B is the Beta function. For our purposes, we re-parameterize this using the mean allele frequency p and the overdispersion parameter θ (which we will estimate with F_{IS}):

- Mean allele frequency: $p = \frac{\alpha}{\alpha+\beta}$
- Overdispersion parameter: $\theta = F_{IS} = \frac{1}{\alpha+\beta+1}$

From these, we can solve for α and β:

- $\alpha = p\left(\frac{1-\theta}{\theta}\right)$
- $\beta = (1-p)\left(\frac{1-\theta}{\theta}\right)$

To account for this overdispersion, we will use the **Beta-binomial distribution**. Conceptually, while the hypergeometric model assumes sampling from a single urn with a fixed allele frequency, the Beta-binomial models sampling from an "urn of urns," where each sub-urn has a slightly different allele frequency. This directly models the variance in allele frequencies caused by population structure. The key parameter that quantifies this overdispersion is the inbreeding coefficient, F_{IS}.

The Initial Challenge: A Fixed F_{IS} Disconnects the Model from N_c

An initial approach would be to calculate a single, global average F_{IS} from the data and use this fixed value to parameterize the Beta-binomial distribution across the entire grid search.

However, this presents a fundamental conceptual problem. The hypergeometric model works because the variance of the sampling distribution is *directly* dependent on the hypothesized census size (N_c). A smaller N_c creates a more pronounced effect of sampling without replacement, increasing variance. The MDE framework finds the N_c that best explains this variance.

If we simply use a fixed, empirically observed F_{IS}, we are only describing the overdispersion that we see, but we fail to model *why* it exists. The link between the hypothesized N_c and the expected level of overdispersion is broken. The model would find the best-fitting allele frequency p, but the choice of N_c would become arbitrary.

---

3. The Path Forward: Modeling Drift to Dynamically Link N_c and F_{IS}

To make the Beta-binomial model a true tool for estimating census size, we must connect the hypothesized N_c to the expected amount of overdispersion. In population genetics, the mechanism that connects population size to the variance in allele frequencies is **genetic drift**.

A smaller population size leads to stronger genetic drift, which creates greater variance in allele frequencies among its conceptual subpopulations. This, in turn, leads to a higher expected F_{IS} value in a sample. Therefore, the key logical step is to replace a single, fixed F_{IS} with a **hypothesized F_{IS} that changes dynamically with each hypothesized N_c on the grid.**

The Bridge: Wright's Island Model

To create this dynamic link, we will use a classic population genetics model: **Wright's Island Model**. This model describes the equilibrium between genetic drift (which increases F_{IS}) and migration/gene flow (which decreases it). At equilibrium, the expected amount of differentiation is given by:

$$F_{IS} \approx \frac{1}{4N_e m + 1}$$

Where: - N_e is the **effective population size** of the subpopulations (which we are trying to estimate with N_c). - m is the **migration rate** between them.

This formula provides the crucial bridge, explicitly connecting the population size (N_e) to the overdispersion parameter (F_{IS}).

---

4. Proposed Implementation: A Two-Step Calibrated MDE

Our refined approach is a practical, two-step process that uses the empirical data to calibrate the theoretical model.

Reconciling Effective Size (N_e) and Census Size (N_c)

A critical distinction must be made. The goal of `dnadot` is to estimate the **census size (N_c)**. However, the Wright's Island model for genetic drift is formulated in terms of **effective population size (N_e)**. To ensure our final estimate is for N_c, our model makes a direct and explicit assumption: **we use the hypothesized census size (N_c) from the grid search as a proxy for the effective size (N_e) within the drift formula.**

For each point on the grid, we are testing the hypothesis: "If the census size were N_c, and assuming the effective size driving drift is approximately equal to N_c, does the resulting theoretical distribution match

our observation?" Because the grid being searched is composed of N_c values, the final estimate remains an estimate of census size.

Step 1: Calibrate the Model to Find the Migration Rate (m)

Before the main grid search, we perform a one-time calibration:

1. **Calculate Global Observed $\hat{F}_{IS}$**: As is standard practice, we will calculate a robust mean F_{IS} from all available loci. Let's call this $\hat{F}_{IS,global}$. This value represents the empirical reality our model must explain.
2. **Get an Initial N_c Guess**: We need a reasonable starting point for N_c. We will run the original, fast **hypergeometric `wasserstein` method** to get a ballpark estimate, $N_{c,initial}$. While potentially biased, it provides a biologically plausible anchor.
3. **Solve for the Migration Rate (m)**: With $\hat{F}_{IS,global}$ and $N_{c,initial}$, we can rearrange the island model formula to solve for the one remaining unknown, the effective migration rate `m`.

$$m = \frac{1 - \hat{F}_{IS,global}}{4 \times N_{c,initial} \times \hat{F}_{IS,global}}$$

   This calculation yields a migration rate that is perfectly consistent with our observed level of overdispersion and our initial estimate of population size. This `m` value will now be treated as a fixed constant for the main analysis.

Step 2: Run the Dynamic Beta-binomial Grid Search

With the calibrated migration rate `m`, we proceed with the MDE grid search. The logic is now fundamentally different from the fixed-F_{IS} approach:

1. The MDE framework iterates through each hypothesized census size, `N_try`, on the grid.
2. For **each `N_try`**, we calculate its corresponding **`hypothesized_Fis`** using our calibrated `m` and the island model formula:

$$hypothesizedF_{IS} = \frac{1}{4 \times \mathrm{N\_try} \times m + 1}$$

   This ensures that smaller hypothesized N_c values correctly lead to higher expected overdispersion.
3. The framework then iterates through the inner loop of hypothesized allele frequencies, `p_try`.
4. For each (N_c, p) pair on the grid, the Expected distribution of allele counts (E) is calculated using the Beta-binomial probability function, parameterized with the **`hypothesized_Fis`** specific to that N_c.
5. The Wasserstein distance between the Observed (O) and Expected (E) distributions is calculated, and the (N_c, p) pair that minimizes this distance is selected as the final estimate.

Conclusion

This revised approach transforms the Beta-binomial method from a simple descriptive model into a powerful inferential tool. It correctly asks the question: **"What census size (N_c) best explains the observed distribution of allele counts, fully accounting for the genetic drift and overdispersion that would occur in a population of that size?"** By using rich, multi-locus empirical data to calibrate a robust theoretical model, this method promises to deliver more accurate and biologically realistic estimates of census size, especially in populations that deviate from panmixia.