# Predicting the Genomic Resolution of Bulk Segregant Analysis

**Runxi Shen[1], Philipp W. Messer[1]**

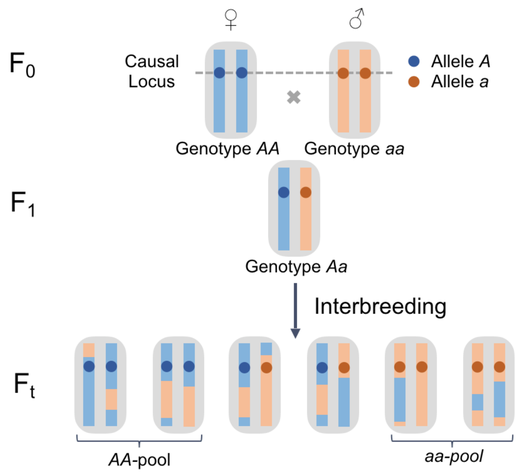
**1. Department of Computational Biology, Cornell University, Ithaca, NY 14853, United States**

## Abstract

Bulk segregant analysis (BSA) is a genetic mapping technique for identifying the loci that underlie phenotypic variation. The basic principle of this method is to select two pools of individuals from the opposing tails of the phenotypic distribution for the trait of interest. These pools are then each sequenced and scanned for alleles that show characteristically diverged frequencies between the pools, indicating that they could be responsible for the observed trait differences. BSA has already been successfully applied for the mapping of quantitative trait loci (QTLs) in organisms ranging from yeast to crops. However, these studies have typically suffered from rather low genomic resolution, and we still lack a detailed understanding of how this resolution is affected by experimental parameters. Here, we use coalescence theory to derive analytical results for the expected mapping resolution of BSA. We first show that in an idealized population without genetic drift the expected mapping resolution is inversely proportional to the recombination rate, the number of generations of interbreeding, and the number of genomes sampled, as intuitively expected. In a finite population, coalescence events in the genealogy of the sample reduce the number of potentially informative recombination events during interbreeding, and thus the achievable mapping resolution. This is incorporated in our theory by introducing an effective population size parameter, specified by the pairwise coalescence rate in the interbreeding population. We show that the mapping resolution predicted by our theory is in excellent accordance with numerical simulations. Our framework can enable researchers to assess the expected power of a given BSA experiment, and to test how experimental setup could be tuned to optimize mapping resolution.

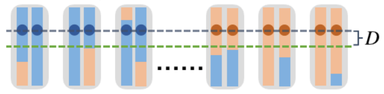
## Bulk Segregant Analysis (BSA)

BSA is a mapping approach that combines certain ideas from linkage mapping and GWAS. It starts from two parental strains of contrasting phenotypes. These strains are then crossed to generate an $F_1$ population, which is further interbred for several generations while maintaining a sufficiently large population size to allow recombination to break up linkage from the two parental strains. In the final generation, two pools of individuals are selected from the tails of the phenotypic distribution. The alleles responsible for trait differences (as well as any alleles linked to them) should then show characteristic frequency differences between the two pools, while alleles at other loci should still be segregating at similar frequencies to those expected in the $F_1$.



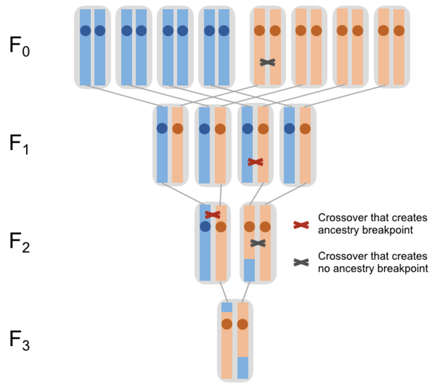

## BSA Genomic Resolution

**Ancestry breakpoint:** the point where ancestry changes along a chromosome, e.g. the ancestry from the *AA* strain in blue switches to the *aa* strain in orange, or vice versa.
***D*:** the distance to the closest ancestry breakpoint downstream of the QTL, observed among all chromosomes in the sample.




The distance ***D*** to the closest ancestry breakpoint located downstream of the QTL in a sample of 2*s* gametes from the $F_1$ (representing a sample of *s* diploid individuals from the $F_2$) will then be an exponential random variable with cumulative density function:

$$P(D \leq d) = 1 - e^{-2rsd},\ with\ E[D] = \frac{1}{2rs}. \quad [1]$$




## Infinite Population vs Finite Population Model

**Infinite Population:** extend the processes from above to *s* sampled individuals from the $F_t$. As we neglect drift here, all lineages are independent of each other. The expect *D* would thus be:

$$E[D] = \frac{1}{rst}. \quad [2]$$

**Finite Population:** suppose we have a diploid population of size *N* under the Wright-Fisher model, and we denote the number of haploid lineages as *x*(*n*) at generation *n*. Hence, in the final generation $F_t$ with *s* diploid samples, we have *x*(*t*)=2*s* (see figure in the third column).

*Recursive Exact Solution*: the expected number of lineages in generation *n* can be estimated based on results of occupancy distributions using a recursive expression:

$$E[x(n-1)\,|\,x(n) = i] = N - N(1 - \frac{1}{N})^i. \quad [3]\ (Maruvka\ et\ al.\ 2011)$$

With the above expression, we can then sum up the expected number of lineages in each generation to calculate the expected total length (*T*) of the genealogy, where all the ancestry breakpoints could be generated, and estimate the expected *D* in the finite population model as:

$$E[T] = \sum_{i=3}^{t} \frac{E[x(n)]}{2} + E[x(2)] \Rightarrow E[D] = \frac{1}{rE[T]}. \quad [4]$$

The factor of 1/2 in the above expression characterizes the probability of heterozygosity at any given genomic position in generation *n*>2, where ancestry breakpoints could be generated. The only exception is for the lineages in generation *n*=2, whose parents are all heterozygous at any genomic position from the BSA experiment setup.

*Approximate Closed-form Solution*: by expressing the recursive solution in a differential equation, we can also solve it for a deterministic approximation of the expected number of lineages at generation *n*:

$$E[x(n)] = \frac{2s}{(2s - (2s-1)e^{-\frac{t-n}{4N}})}. \quad [5]\ (Maruvka\ et\ al.\ 2011)$$

The expected total length (*T*) of the genealogy in this case can then be integrated from 0 to *t*:

$$E[T] = \frac{1}{2}\int_0^t E[x(n)] \Rightarrow E[D] = \frac{1}{rE[T]} = \frac{1}{2rNln(2s(e^{\frac{t}{4N}} - 1) + 1)}. \quad [6]$$

## Model Comparison

We now take a closer look at the expected mapping resolution derived in Eq. [6] and discuss how it relates to the infinite population model. When $t << 2N$, it specifies a regime where the probability that a given pair of lineages coalesces over the course of the experiment is still small. Under this assumption, we can perform a Taylor approximation of the exponential in Eq. [6]:

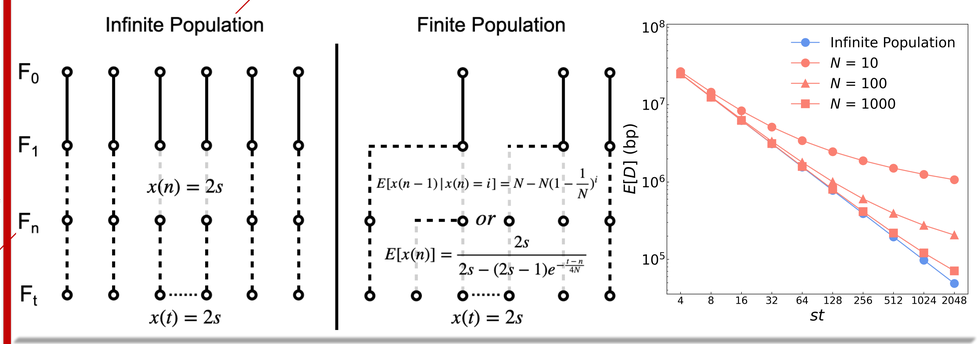
$$ln(2s(e^{\frac{t}{4N}} - 1) + 1) \approx ln(\frac{st}{2N} + 1) \Rightarrow E[D] \approx \frac{1}{2rNln(\frac{st}{2N} + 1)}. \quad [7]$$

Eq. [7] shows how the infinite and finite population models differ from each other. In the infinite population model, mapping resolution was simply inversely proportional to each of the recombination rate, sample size, and length of the experiment. In the finite population model, mapping resolution is still inversely proportional to the recombination rate, but the effects of sample size and experiment length are now attenuated by a logarithm.

We further note that *s* and *t* enter Eq. [7] only in the form of the product *st*. Thus, varying each of these two parameters by the same factor is expected to produce a similar impact on the expected mapping resolution (as long as $t << 2N$ still holds). Eq. [7] also shows us where these effects start to become relevant. If $st << 2N$, we can further approximate:

$$ln(\frac{st}{2N} + 1) \approx \frac{st}{2N} \Rightarrow E[D] \approx \frac{1}{rst}. \quad [8]$$

Thus, the infinite and finite population models nicely converge in this regime.




## Solution Analysis & Numerical Validation

We conducted forward-in-time, individual-based simulations of a BSA experiment to evaluate the accuracy of our analytical results. We modeled a trait determined by a single QTL, located at the center of a chromosome of length 100 Mbp with a uniform recombination rate of $r = 10^{-8}$ per bp and generation. The free parameters of our simulation model are the sample size (*s*), the population size (*N*), and the length of the BSA experiment (*t*). Simulations were implemented in SLiM, using tree sequence recording to track ancestry segments along each chromosome (this allowed us to directly detect the true location of ancestry breakpoints).