Allele Phaser Methods

The allele phaser determines for each locus, the alleles that are statistically supported by the assembled reads. Because the determination is made for each individual separately, no population assumptions are required (e.g. Hardy Weinberg). Phasing is made possible by reads (and read pairs) that overlap with two or more polymorphic sites. The following steps are taken:

1) Identify the polymorphic sites in the assembly for a given locus.

2) Note the base calls for reads overlapping with each polymorphic site.

3) Load information from the assembled reads into a linked list, storing only base calls and positions of polymorphic sites (the monomorphic sites are added back in below). The positions correspond to the position in the consensus sequence generated from the assembly.

4) Identify the paired reads and collapse them into a single node in the linked list.

5) Construct alleles containing only the polymorphic sites and initialize the alleles by drawing bases from a uniform prior.

6) Estimate the Bayesian posterior distribution of alleles sets by Markov-Chain Monte Carlo (MCMC). Repeat steps 7 thru 10 below nGens times, taking samples every 100 generations, starting with generation nGens/2 (50% of generations removed as burn-in).

7) Propose a change to the allelic states by either a recombination move or a mutation move. The recombination move (chosen probSwap proportion of the time), involves randomly choosing two alleles, and exchanging the bases (between the alleles) found after a randomly chosen position within the reads. The mutation move (chosen 1-probSwap portion of the time), involves randomly drawing from a uniform prior a new base for a randomly chosen base position for a randomly chosen allele.

8) Compute the log likelihood of observing the reads given the allelic states of the current generation, i,

, where

, where

, where *R* is the set of observed reads, *A* is the set of currently assumed alleles, *S* is the set of polymorphic sites for which *r* overlaps, and *M* is a user-specified 4x4 matrix containing the assumed probabilities of sequencing error. The following matrix, which corresponds to a 0.75% sequencing error, is assumed by default:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | True Base | | | |
|  |  | A | T | C | G |
| Sequenced Base | A | 0.9925 | 0.0025 | 0.0025 | 0.0025 |
| T | 0.0025 | 0.9925 | 0.0025 | 0.0025 |
| C | 0.0025 | 0.0025 | 0.9925 | 0.0025 |
| G | 0.0025 | 0.0025 | 0.0025 | 0.9925 |

Future versions of the program could include utilization of the quality scores generated by the sequencer, but this would come at a substantial computational cost. More work is required to determine to what that degree the accuracy of allele estimates would benefit from the utilization of the quality scores.

9) Compute the log of the Hastings ratio, which is equal to the difference in log likelihood scores between generations (lnLi - lnLi-1), when the prior is uniform and the proposals are symmetric.

10) Accept the move if the log of the Hastings ratio is greater than the log of a randomly chosen number between 0 and 1.

11) After the posterior distribution has been estimated, generate full-length alleles by replacing the polymorphic bases in the consensus sequence with the bases obtained from alleles sampled in the posterior distribution. Uncertainty in the posterior distribution is accommodated by one of two methods. The first method generates fully-phased alleles with no ambiguities by drawing alleles randomly from the posterior distribution. The second method generates partially-phased alleles with no ambiguities in the largest region containing 95% posterior confidence (adjacent polymorphic sites are associated in 95% of the posterior distribution), and ambiguities for the remainder of the polymorphic sites (ambiguities are determined by the posterior probability of each base).

Storeria-Specific Parameters

nGens = 20000

probSwap = 0.5

nAlleles = 2

sampleFrequency = 100

burnin = 10000 generations

probMatrix = default, pSequencingError = 0.0025

ambiguity resolution method = allelesWAmbigs

Allele Phaser Program Notes (for ReadMe file?)

Program Input:

fileStem e.g. I13440,

program uses \*\_polys.txt, \*\_conSeqs.fasta, and \*\_readsInFiles.txt

nGens e.g. 10000

number of MCMC generations

probSwap e.g. 0.5

proportion of times mcmc-proposed change is a swap of allele bases

nAlleles e.g. 2

Number of alleles assumed

verbose e.g. VERBOSE

optional last parameter indicating that verbose output is desired

ProbMatrix.txt

Matrix specifying the probabilities of sequencing error

\*\_conSeqs.fasta Consensus sequences with >L1.1 header format

\*\_polys.txt File detailing the polymorphic sites

e.g. 4 1 35970 331 202 C E

locus sup read aPos rPos base qual

Program Output

\*\_polySummary Polymorphic sites and connections between them

e.g. 5<=(10)=>25<=(99)=>135<=(0)=>450

\*\_posteriors\_\*.txt Posterior probability distribution of allele sets

\*\_allelesFromPost Fully resolved alleles, drawn randomly from the post.

\*\_allelesWAmbigs Alleles resolved via 95% confidence, otherwise ambig

\*\_posteriorSummary 101 x 101 matrix with counts of nHetSites x postProb

Starting Conditions

Equal base frequencies (alleles drawn from null distribution)

Data Structure: AllelePhaserRead for linked list

readID read number in fastq file, used to account for paired reads

weight currently 1, should improve to be count of identical b[], s[]

b[] array indicating bases observed in polymorphic sites

s[] array indicating site positions of bases in b[]

length number of bases in b[]

next link to next read in list