Excerpt from Fragoso, Christopher A., et al. "Genetic Architecture of a Rice Nested Association Mapping Population." *G3: Genes, Genomes, Genetics* (2017): g3-117.

BP-Impute algorithm

BP-Impute is a hidden Markov model method, like LB-Impute {Fragoso, 2016 #92}, but features a few key differences. A Markov chain is constructed from either end of an ambiguous breakpoint region. Each chain from either direction, however, is constrained to the hidden state of the last marker imputed by LB-Impute. BP-Impute works under the assumption that there is only one transition in parental state in the ambiguous interval. The initial probability of the high-confidence LB-Impute marker's parental state is 1. Transition probabilities are calculated from the proportion of recombined lines within the missing interval, for the entire population. This is a rather naïve measure of transition probabilities that will be improved in future versions. The probability calculation for one marker, for one homozygous parental state, in a left to right Markov chain, is demonstrated in Equation 1:

$$P_{right}(state_t = parent_A) = P(state_{t-1} = parent_A) * \left(1 - P(recombination_{Interval})\right) * P(homozygous\ emission \mid state_t = parent_A) \quad \text{Eq. 1}$$

Sequenced (yet unimputed) markers within the ambiguous regions may be incorporated into the model in two ways. One way is to view the read depth as an emission from the constrained parental state. This may include valuable additional information to fine tune the breakpoint, as aligned reads may greatly support one parental state over the other. For this assumption, we use the same binomial emission model as with LB-Impute. Second, sequenced markers could be assumed to have high confidence genotypes, just as with the LB-Impute imputed marker set. These markers then also serve as "anchors" to the Markov chains. This is particularly useful at the distal ends of chromosomes, which LB-Impute often leaves unimputed.

The genotype probabilities from each chain, after being normalized to sum at 1, may then be used to weight each parental genotype, and a weighted average genotype is produced. The probabilities are then divided by 2 so that the maximum value is 1. For assigning discrete genotypes to the probabilities, a separate R script is used (assign_genotypes.R) that employs least squares to identify the breakpoint. Then, markers on either side of the breakpoint are assigned the proper genotype. If the probabilities are exactly halfway in between two parental states, the breakpoint is randomly assigned.

BP-Impute example commands

Rscript bpimpute_v33.R IR64xAzucena_reseq_filtered_parimpute_parfilt_imputed_keep.vcf 43 ID152bH10.P2_CGTACG.GTGGCC ID152bH11.P2_GAGTGG.GTGGCC 0.1 trim yeshet;

Arguments:
1. Path to vcf file
2. Number of "##" lines to skip to vcf file
3. Name of common RIL parent in vcf file
4. Name of diversity donor parent in vcf file
5. Heterozygosity threshold for removing RIL line from file, value between 0-1 (default 0.1)
6. Trim chromosomes by most distal, shared marker in population, trim OR notrim (default trim)
7. Keep heterozygous genotypes in RILs or set as missing and impute as homozygous, yeshet OR nohet (default yeshet)

Assign genotypes example commands

Rscript assign_genotypes_v14.R IR64xAzucena_weighted_genos_trim_yeshet.txt IR64xAzucena_imputed_binary_trim_yeshet.txt;

Arguments:
1. Path to BP-Impute genotypes file
2. Path to BP-Impute imputed boolean file