**Infer Caenorhabditis becei founder phases from RILs**

**Crossing scheme:**

* Two female founders, A and B, were individually crossed with the same male founder, M. Following each initial cross (cross A and cross B), the progeny underwent expansion for 5 generations of outcrossing, followed by X generations of sib-mating to produce inbred lines.

**Data**:

* Pool of worms from the initial cross A and cross B were sequenced, as well as a pool of worm originating from the founder M and the C. becei reference strain. The pool-sequencing data were used to infer the diploid genotypes of each founder at each SNP. XXXX METHOD BY JOSE
* RILs genotype XXXX METHOD BY JOSE
* To infer the founders’phased haplotypes, only the hard-filtered SNP were used XXXX METHOD BY JOSE

**Rational:**

The RILs comprise blocks of founder haplotypes, which are expected to remain mostly unbroken and represent the predominant haplotype blocks over short enough genetic distances. Within a given genomic window, the attribution of haplotypes to different founders can be determined by considering the following:

1. Founder M haplotype blocks should be present among the shared haplotype blocks within the RILs panels A and B.
2. The most probable attribution of haplotype blocks is the one that minimizes the disparity with the inferred diploid genotype of the founder, as determined by pool sequencing
3. The correct phase of haplotypes between the different window should minimize the number of breaks in the rils.

This method have to intrinsic limitations:

1. If the linkage is excessively disrupted within the RILs, the inference of unbroken founder haplotypes becomes challenging or impossible
2. If one of the founder haplotypes has been eliminated or purged from the RILs, it cannot be reliably identified.

**Script:**

The inputs of the script are a **rilsA** and **rilsB** genotype matrix [snp, lines] containing with value 0 or 1 for homozygous reference or alternative allele. Heterozygous genotypes are set as NA values. For each founder a two-column matrix [snp x genomes] gives the unphased haplotypes (**founderA, founderB, founderM**). A data.frame, **info**, contain the SNP info (colums = ID, CHROM, POS, cM).

There is a filtering step to exclude:

* SNP with more than 5% NA among all **rilsA** and **rilsB**
* SNP that are fixed
* SNP that are in high LD (r > 0.9 or < -0.9) with less than than 3 other SNP within a 500 SNP window. These SNP are identified with the function **Search.LowLDSNP(rils, winsizeLD = 500, LDth=0.9, min.nHighlink=3)**.

The founder phased are obtained through two main function: **InferFoundersHaploBlocks** and **phaseHaplotypes**. The first function infer the major haplotype in RILs within a sliding window and attributed them to the three different founders. The second function phase the different haplotype blocks in the different window.

1. **InferFoundersHaploBlocks**

InferFoundersHaploBlocks(founderA, founderB, founderM,, rilsA, rilsB, info, nsnpWin = 100, maxSizeCM=3)

* founderA, founderB, founderM,, rilsA, rilsB, info are the input data described above
* nsnpWin is the window size
* maxSizeCM is the maximal genetic size (cM) of a window that a window should have, if the window is bigger, it will be split.

1. Overlaping windows of size nsnpWin (if >maxSIzeCM) are stored into a list where the number correspond to the ordered position of SNPs.

i.e., list(1:100, 51:150, 101:200, …)

1. For each windows:

We infer the major haplotypes for cross A and cross B using the following line of code:

cutree(hclust(dist(t(gx))), h = 0)

Here, gx represents either rilsA[win,] or rilsB[win,]. This code clusters the RILs by haplotypes. Haplotypes that are represented in over 10 RILs are considered major haplotypes. If there are more than four major haplotypes, we select the four most frequent haplotypes, as we anticipate a maximum of 4 haplotypes per RIL panel.

The inferred major haplotypes that are identical between the cross A and cross B are considered as shared haplotypes, other are considered as unique to cross A or cross B.

For each founder, all the possible combinations of relevant RILs haplotype are tested and their genotype distance to the poolseq inferred genotype is caculated: , where g is the diploid genotype (0,0.5 or 1) at each snp. For each founder, the combination minimizing the genotype distance is chosen. The relevant RILs haplotypes for founder M are the shared haplotype between rilsA and rilsB. The relevant RILs haplotypes for founder A/B are the shared haplotype or the haplotype which are unique to cross A/B.

For each founder, we test all possible combinations of relevant RILs haplotypes and calculate their genotype distance to the poolseq-inferred genotype using the formula:

,

Here, represents the diploid genotype (0, 0.5, or 1) at each SNP. For each founder, the combination minimizing the genotype distance is chosen.

For founder M, the relevant RILs haplotypes to be tested are the shared haplotypes between rilsA and rilsB. For founders A/B, the relevant RILs haplotypes are either the shared haplotypes or the haplotypes unique to cross A/B.

1. Check that the overlapping intervals are compatible

Once founder haplotypes are inferred for each window, we verify that the different windows are compatible with each other at the overlapping snps. Here, compatible means that for a given founder, the two window at the shared snps should be identical in cis or in trans.

If some window are “not compatible”, the algorithm try to extend the window and check if the problem is resolved.

1. **phaseHaplotypes**

phaseHaplotypes(founderhaplotypes, info)

* founderhaplotypes: a list returned by InferFoundersHaploBlocks
* info: see input data presented above

The phasing occurs in two steps:

1. The first step capitalizes on the estimation of founder haplotype blocks across overlapping windows. Matching shared SNPs between two windows for a given founder is straightforward. The algorithm iterates from window 1 to n, comparing window ith to window jth to phase the latter with the previous windows. However, this method is not efficient when window ith is homozygous, as it lacks phase information of windows 1 to ith. After this initial step, we obtain "blocks" of phased windows separated by homozygous chunks. Haplotypes are phased within blocks but not between blocks.
2. The second step involves phasing together the "phased blocks." For a given founder, the phase of block jth is determined by counting the number of cis and trans combinations with the ith block in the RILs. The correct phase (cis or trans) is assumed to be the most common one among the founders.