General

Parentage assignment by sequoia is divided into three steps:

- Filtering to create short lists of candidate dams & sires
- Filtering and selection of the most-likely parent pair
- Filtering and selection of the most-likely single parent (if no parent pair assigned)

When running sequoia with `--log`, for each individual the count of candidate parents remaining after each filtering step is written to a log file, as well as some other information to assist in problem finding.

Parentage assignment happens in several iterations, and each iteration creates its own log, named `AssignmentLog01.txt`, `AssignmentLog02.txt`, etc. With high quality data, almost all assignments happen in the first iteration. The subsequent iterations are typically only used when there's individuals with unknown sex and/or birth year or poor genetic data quality, which get 'puzzled in' once the pedigree is nearly complete.

Very occasionally an assignment may 'wobble', and get dropped one iteration to be re-assigned the next. This happens when two mutually exclusive pedigree configurations are nearly equally likely. It may be due to poor genetic data quality and/or a weird, unlikely (but correct) pattern of Mendelian inheritance.

Read file into R

The lines of the log file are of unequal length. To deal with this, use `fill=TRUE`:

```
read.table('AssignmentLog01.txt', header=TRUE, fill=TRUE, na.strings=c('NA', '-999.0'))
```

Explanation of columns in assignment log

Make short list

The first 5 columns are related to the filtering steps to create short lists of candidate dams & sires:

- **G__row**: focal individual's rownumber in genotype matrix
- **id**: focal individual
- **dam__in, sire__in**: parents assigned at the start of this iteration, i.e. assigned in one of the previous iterations
- **n__cand__dam, n__cand__sire**: Number of candidate dams/sires of individual i that pass the initial filtering criteria. Candidate parents of unknown sex are included in both the shortlist of candidate dams and candidate sires.

The filtering criteria that determine whether individual j gets on the short list of candidate parents for focal individual i are:

- not offspring of i
- OH count <= MaxMismatchOH
- LLR(PO/U) > T_filter
- older than i, or age difference unknown
- if age difference unknown: not a descendant of i (check up to 6 generations back)
- if age difference known: age difference has non-zero probability in age prior

The OH counts are calculated once for all possible pairs of individuals at the start of the program. Values LLR(PO/U) are also calculated at the start, for all pairs with OH count <= MaxMismatchOH. Thus, all steps except for the descendant check are simple look-ups in existing matrices, and therefore very fast.

Parent pair

The next 7 columns are related to filtering and assigning a parent pair. This is done in four steps, and columns **n__pairs__s1** to **n__pairs__s4** give the number of candidate pairs remaining after each filtering step. The initial number of pairs is n_cand_dam X n_cand_sire.

Within each step, later sub-steps are only calculated for those candidate pairs that passed the previous sub-steps. Since these sub-steps are done within the same loop, it is not as straightforward (but not impossible) to get counts after specific sub-steps.

step 1 Only relevant for monogamous or hermaphrodite breeding sytems.

- if monogamous breeding system: valid pair if $d + s$ either already got assigned offspring as a pair, or neither is yet assigned as parent
- invalid pair if both d and s are of unknown sex, because unable to tell which is the dam and which the sire. Exception: hermaphrodite breeding system & i's inbreeding coefficient suggests it is the product of selfing and $d{=}{=}s$.
- if hermaprodite: invalid pair if i is product of selfing and $d \neq s$

step 2

- Count of trio Mendelian mismatches < MaxMismatchME. E.g. if at a locus i is AB and both d and s are AA, this is no mismatch for either i–d or i–s, but is a mismatch for the trio i–d–s.
- LLR(both parent vs. both unrelated) > 2 * T_filter
- LLR(both parent vs. only dam or only sire) > T_assign

step 3

- if birth year of i is unknown: d and s must be of compatible age
- LLR(both parent vs. next-most-likely configuration of close relationships) > T_assign ('LLRZpair')

step 4

- if i, d or s has unknown birth year: configuration $i = d$ x s is more likely than d offspring and s parent of i or v.v. by margin 2*T_assign (only configurations possible based on known ages considered)

Final selection If >1 pairs remaining after step 4, final selection is based on - assign the best dam + sire - loop over all candidate dams & sires (loop-within-loop), and if (currently assigned dam) + (this candidate sire) or v.v. are a plausible pair (passed all filtering steps), then - calculate likelihoods for many different configuration of close relationships for the quartet focal individual + current dam + current sire + this candidate sire/dam - if likelihood for current dam + this candidate sire to be parent-pair is higher than for any alternative mini-pedigree, replace 'current sire' by 'this candidate sire' (or dam) This cumbersome approach is redundant in simple, high quality datasets, but reduces incorrect assignments in complicated situations.

- **mx_LR_pair**: Highest LLR(both parent vs. next-most-likely) for the most-likely pair (which may or may not be assigned). Missing value code = -999.0
- **dam_pair, sire_pair**: assigned parent-pair.

IDs of most-likely-but-not-assigned parent-pairs are currently not reported (to avoid any possible confusion), but would be easy to add.

Duplicated candidate parent

The next 3 columns relate to a check whether there are any potential duplicated genotypes among the candidate parents. This check is only run with `--log` and does not affect assignment in any way. It is only run between dams and between sires, i.e. it assumes that the sex of candidate parents is correct (or unknown).

- **dup_LR**: maximum LLR(duplicate vs (PO/FS/U)) among these pairs. Missing value = -999.0; LLR not calculated if count of different SNPs for the two individuals > MaxMismatchDUP.
- **dup_sex**: sex of duplicates: 1=dams, 2=sires, 0=NA
- **dup_id_A, dup_id_B**: ids of the probably duplicated candidate parents

Single parent

The last 6 columns are related to finding the most likely single parent, from among the original short lists. This is done in three steps, and columns **n_single_s1** to **n_single_s3** give the number of candidate parents (dam + sire) after each filtering step. The initial number is n_cand_dam + n_cand_sire.

Step 1

- known sex (unless hermaphrodites)
- if monogamous breeding system: candidate does not have a mate yet
- LLR(parent vs otherwise related) > T_assign. This 'LLRZsingle' is calculated as by-product of the trio/quartet likelihoods for parent-pairs; this is the lowest value across potential co-parents. It is the LLR between the most-likely configuration in which j is parent vs most-likely in which j is otherwise related.

Step 2

- for all possible pairs of candidate single parents (same sex + opposite sex), calculate LLRZsingle (i.e. including conditional on the other candidate being the parent). Again, LLRZsingle must be > T_assign.

Step 3

- if age difference between i and j is not (exactly) known: check if i might be parent of j instead

Final selection If exactly 1 candidate remains after filtering step 3 the assignment is made, otherwise not.

- **mx_LR_single**: highest LLR(parent vs otherwise related), whether assigned or not. Missing value = -999.0
- **dam_single, sire_single**: assigned single parent.

Potential reasons for non-assignment & their fixes

- True parent is not genotyped
- True parent is not shortlisted
  - Age difference not allowed by age prior and/or is incorrect
  - OH count > MaxMismatchOH (–> increase MaxMismatchOH)
  - LLR(PO/U) < T_filter (–> decrease T_filter and/or increase Err)
- Likelihood not conclusive that related as parent rather than e.g. full sibling or grandparent: LLR(PO/other) < T_assign (–> decrease T_assign and/or increase Err)
- True parent pair not deemed a 'compatible' parent pair. Either may be assigned as single parent, or neither when they are approximately equally likely. (–> Probably genotyping errors: re-genotype)
- Multiple parent(s/ pairs) are shortlisted, but none is decisively the most likely
  - Special case: the true parent's genotype is duplicated