SNP-based age analysis methodology: a summary

Summarised description of the age analysis pipeline — Iain McDonald, June 2017

1 Abstract

This document provides a brief description of the mathematical processes and assumptions behind the age analysis pipeline seen on this site.

2 Assumptions and inputs

Mathematically, it is assumed that mutations occur randomly throughout the Y chromosome, and that that random process can be defined by a mutation rate (μ), which represents the average number of years between simple nucleotide polymorphism (SNP) mutations within some region of the chromosome. While in principle, this mutation rate can be a function of many factors (e.g. time, geography, geology, generation length, ethnic group, cultural factors, etc.) numerous studies have sampled different geographical regions, different periods and father–son pairs of different generational gaps, showing that the mutation rate in *Homo sapiens sapiens* is fixed to within $\sim \pm 7.5\%$. A summary is provided in Jobling & Tyler-Smith (2017).

The mutation rate is known to vary slightly within different parts of the chromosome. For this study, we have chosen the following regions (Build 37 positions in BED format):

chrY 2649599 2917999
chrY 6616499 7446499
chrY 9166089 9466025
chrY 9757201 10034899
chrY 13870499 16095999
chrY 16167425 18217274
chrY 18537844 19622082
chrY 20995635 22216399
chrY 22512899 23693155
chrY 23996337 24050183

These regions are confined to the non-palindromic, euchromic regions of the Y chromosome.

The final mutation rate we apply to these regions varies as fresh data comes in. As of June 2017, a rate of 8.186×10^{-10} bp^{-1} yr^{-1} is used, with a 95% confidence interval (c.i.) of 7.589–8.771 $\times 10^{-10}$ bp^{-1} yr^{-1}. Over the regions indiciated above, this translates to $\mu = 186.80$ (174.33–201.47) years per SNP mutation. This mutation rate is derived from a number of literature sources and from direct estimates from the BigY data itself. An estimate of $\mu_{BigY} = 206.50$ (165.58–260.34) yr SNP^{-1} is derived from the 44 400 years of paper-trail genealogies which have been triangulated by pairs of BigY tests: the rates match within the expected uncertainties.

3 Tree formation

The age estimation method relies on counting mutations in the aforementioned regions between nodes on the haplotree, defined by clades. Hence, it is important to create an accurate tree, but not necessarily a complete one.

To create a tree, the VCF files are scanned to compile a list of mutations that occur in one or more tests. This list is parsed, and each mutation is called either positive or non-positive in each test, by referring to that test's variant call (VCF) file. Each non-positive is reclassified as explicitly negative or uncalled by reference to the coverage (BED) file. Mutations are sorted by frequency, which creates the basic tree order, and tests are sorted by the order of positive SNPs in this sequence, which groups tests by clade.

At this point, uncalled SNPs can be derived positive or negative by reference to the rest of the tree, by means a series of implications. In the same way that anyone who is part of R1ba1a2-M269 should be part of R1b-M343, implications can be made that any persons with one or more defining SNPs of a child clade should automatically be positive for the parent clade, even if the relevant SNPs are uncalled. Occasionally, particularly for individual testers' single novel variants (SNVs or 'singletons'), these positive or negative calls cannot be made unambiguously. In these

cases, references are made to the BAM files where possible. As a last resort, placement can be guessed on the basis of the relative ratio of SNPs defining the parent clade versus the child clade. Uncertain placement adds uncertainty to the subsequent age calculations, which is one reason that the regions of the chromosome used must be chosen carefully to by-and-large include only well-covered regions of the chromosome.

Any clades which cannot be marked in the tree can either be labelled as recurrent or 'bad'. The dividing line between a recurrent SNP and a false result should be done on the basis of repeatability, but in publicly contributed data it is difficult to get people to pay for the same test twice. Hence the dividing line between the two can be considered a grey area. Generally, recurrent SNPs are not counted in mutation rate calculations, so they should not be counted in the age analyses that depend on those mutation rates. However, every base pair is recurrent at some level, so some cut-off in recurrency is required to divide 'bad' SNPs from recurrent SNPs. Here, a rough threshold of one recurrency per 1000 tests is adopted, although this is not always followed if the base pair is generally poorly read, or if it forms a Sanger-sequencing-confirmed clade.

Clades can be artificially inserted into the tree by inserting their defining mutations into the parsed list, and inserting the relevant implications to identify clades in which they are positive. Inserting clades which are not recovered by BigY can be done for output purposes. This is done in the U106 tree for Z301, DF98 and Z17, which are not called in BigY and do not have equivalent SNPs which are called. However, each addition artificially lengthens that branch by 0.607 mutations (∼113 years) due to the statistical methods employed, resulting in an over-estimated age. The mere presence of an unrecovered SNP implies a non-zero lengthening is needed, at least by one generation, but since the SNP is unrecovered, this ∼113 years will be an over-estimate. The proper addition depends on the resolution of the test the mutation's location was determined in: for example, an FGC YElite 2.1 test, at 14 million base pairs compared to BigY's 8.6 million base pairs, should receive an addition of $113 \times 8.6/14 \approx 70$ years.

Insertions and deletions, while valid mutations, are not included in the age analysis calculation as they are generally less well recovered, and are not incorporated in the underlying mutation rate calculations. Similarly multi-nucleotide polymorphisms (MNPs) are not generally counted either, so should be masked from the original data, however these are sufficiently rare that they are not a major concern.

4 The age estimation

The basis of the model described here is the same as that used by the commercial company YFull (Adamov et al. 2015). However, this model is amended to produce a temporally more-robust tree structure.

Mutations should accumulate according to Poisson statistics. On average, over a given number of years (t), a given number of mutations are expected (λ), as given by $\lambda = t/\mu$. The probability (P) of observing any given number of mutations (k) in that time is described by Poissonian statistics as:

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{1}$$

Conversely, the time between two nodes on a tree is nominally given simply by $t = k\mu$, Poissonian statistics becomes important for small values of k. The probability of those two nodes being separated by a time less than M years is therefore given by:

$$P(M) = \frac{\int_0^M \left(\frac{(t/\mu)^k \exp^{-t/\mu}}{k!} \right) \mathrm{d}t}{\int_0^\infty \left(\frac{(t/\mu)^k \exp^{-t/\mu}}{k!} \right) \mathrm{d}t}. \tag{2}$$

This can be numerically solved for $P(M) = 0.5$, 0.025 and 0.975 to identify the mean, and the lower and upper boundary of the 95% c.i., respectively.

In this way, clades can be built up by multiplying these probability distribution functions (PDFs) together and renormalising. For example, for two related individuals, having k_1 and k_2 mutations, share a most-recent common ancestor (MRCA) born M years before their average birth date, as described by the following PDF:

$$P(M) = \frac{\int_0^M \left(\frac{(t/\mu)^{k_1} \exp^{-t/\mu}}{k_1!} \times \frac{(t/\mu)^{k_2} \exp^{-t/\mu}}{k_2!} \right) \mathrm{d}t}{\int_0^\infty \left(\frac{(t/\mu)^{k_1} \exp^{-t/\mu}}{k_1!} \times \frac{(t/\mu)^{k_2} \exp^{-t/\mu}}{k_2!} \right) \mathrm{d}t}, \tag{3}$$

which we can simplify using the substitution:

$$p_i(M) = \frac{(t/\mu)^{k_i} \exp^{-t/\mu}}{k_i!}, \tag{4}$$

to become:

$$P(M) = \frac{\int_0^M p_1(M)p_2(M)\mathrm{d}t}{\int_0^\infty p_1(M)p_2(M)\mathrm{d}t}. \quad (5)$$

Similarly, any number of individuals (c) can be multiplied together in this way, such that:

$$P(M) = \frac{\int_0^M \prod_{i=1}^c p_i(M)\mathrm{d}t}{\int_0^\infty \prod_{i=1}^c p_i(M)\mathrm{d}t}. \quad (6)$$

The time since the MRCA (TMRCA) of the parent clade is simply the sum of the TMRCA of the child clade, plus the time between the child clade and the parent clade. Hence, the PDF of the time to MRCA of the parent clade ($P_P(M_P)$) can be calculated from the PDF of the TMRCA for the child clade ($P_C(M_C)$) plus the PDF of the inter-clade period ($P_I(M_I)$):

$$P(M_P) = P_C(M_C) + P_I(M_I) \text{ where } M_P = M_C + M_I, \quad (7)$$

remembering that both P_C and P_I are temporal distribution functions, so must be added with convolution. This is computationally expensive, as M_P must be evaluated for every pair of M_C and M_I. Since Equation (2) can be used between any two nodes of the tree, it can be seen that Equation (6) can immediately be generalised to include the PDF of child clades in this fashion.

In this way, a tree can be built up until a PDF is created for the head node of the tree, from which all results descend. Although not implemented in the current (June 2017) version, during this process the PDF of each clade could be restricted using prior information, such as genealogical paper trails or ancient DNA carbon dating. The difficulty here is that the PDF counting is done in terms of mutations, rather than calendar years. Hence, the uncertainty in μ will cause a resulting spread in the PDF. This kind of restriction can therefore only be done to first order, but could be included as an extra multiplicative PDF in Equation (6).

Also not included in the current (June 2017) version is a renormalisation of the tree length. In principle, knowing the age of the head node allows a unique mutation rate to be established for each child clade (μ_i'), which accounts in part for the random nature of mutations providing some clades with fewer SNPs and some with more. This would be in terms of a normalisation by which:

$$\mu_i'/\mu = < K_i > /\lambda_i, \quad (8)$$

where $< K_i >$ is the clade-weighted average number of mutations found in that line, and λ_i is the expected number given by $\lambda_i = M_i/\mu$. This is not currently implemented because it is more computationally expensive, as the temporal grid spacing which is used to perform the above integrations must be interpolated over, and due to the complexity resulting from uncertainty propagation in μ, outlined below.

The final amendment we make to the overall branching points is to use the PDF of each parent clade to further constrain the PDF of each child clade. This ensures that causality is adhered to, and that no child clade should be older than its parent. This is done simply by inverting Equation (7), such that:

$$P(M_C) = P_P M_P - P_I M_I \text{ where } M_C = M_P - M_I, \quad (9)$$

and using the resulting PDF as a further multiplicative factor in Equation (6). In this way, each clade's TMRCA PDF is constrained in a systematic and self-similar manner by every associated branch, whether it is a single test, a child clade, or the clade's parent.

The final step is now to take into account the uncertainty in the mutation rate. The calculation thus far has been in terms of the primary measured variable: the number of mutations observed between two tree nodes, and converted to age by simply multiplying by μ. However, μ has its own PDF, which we have so far neglected since μ is a simple scaling variable. Incorporating this PDF earlier (e.g. in Equation (6)), would lead to unnecessarily poor definition in the resulting PDF. Instead, it is incorporated at this last stage by convolving the TMRCA PDF of each clade by a normalised, scaled mutation-rate PDF. This latter PDF is described by two half Gaussians, split at the modal value, with characteristic widths set such that the 2.5th and 97.5th centiles of the integrated function return the 95% confidence interval in μ.

The final best-estimate age and 95% c.i. can be read from the convolved PDF as the 50th, 2.5th and 97.5th centiles of the resulting distribution. This is converted to a date by subtraction (with convolution) from the mean age of BigY testers, dated at present (June 2016) from a compilation of 121 testers as 1950 AD +/- 15 years.

5 Comparison to Adamov et al.

The primary disadvantage of this method compared to YFull's implementation of Adamov's method is that we are reliant on the derivative VCF data from Family Tree DNA, which dictates a binary outcome for each SNP. In reality, each call comes with a certainty weighting that can be derived from the quality statistics and individual reads present in the BAM file which allows a more definite tree to be built up.

Conversely, there are a number of advantages to this method, which can be summarised as follows:

- Preserving the PDF ensures that uncertainties from child clades are properly propagated up through the tree.
- Multiplying the PDFs of each clade to obtain an age for a clade provides a more robust method of weighting the contribution of each clade.
- Reprocessing the tree from a top-down perspective prevents problems with causality: no manual correction is needed to make child clades younger than their parents, and a more accurate age estimate is provided.

6 References

Adamov, D., Gurianov, V.M., Karzhavin, S., Urasin, V., 2015, Russian Journal of Genetic Genealogy, 7, 1
Jobling, M. & Tyler-Smith, C., 2017, Nature Reviews Genetics, 36