Instructions for using Mutation_Age_estimation.R

Luke C. Gandolfo[*], Melanie Bahlo[†], and Terence P. Speed[‡]

This document explains how to use the R script Mutation_Age_estimation.R to produce mutation age estimates and confidence intervals using the method described in Gandolfo, Bahlo & Speed (2014)

1 Obtain the data

Firstly, use the following procedure to obtain the *genetic length* of ancestral segments in sampled individuals (see Figure 1):

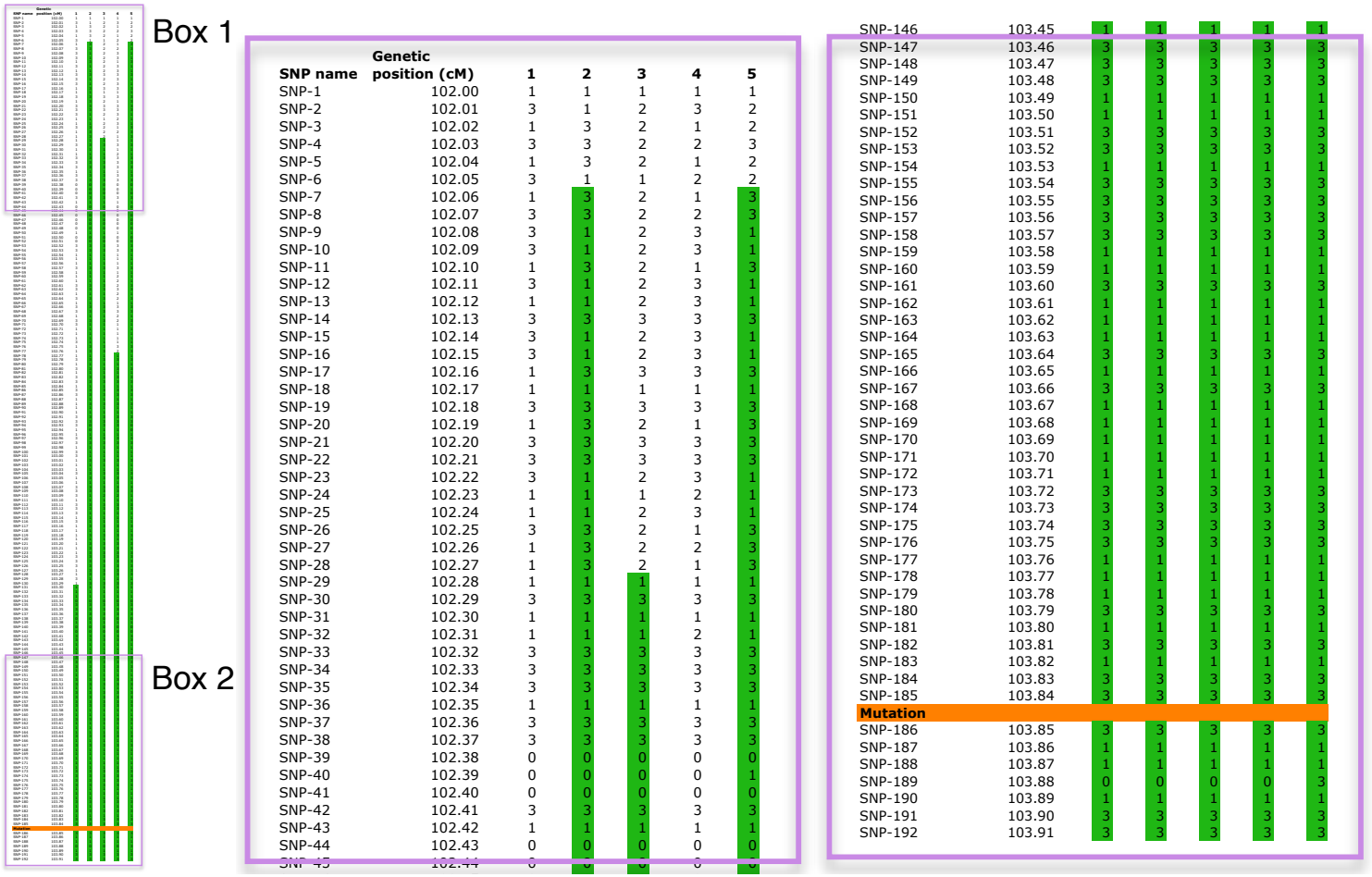
1. Line up your high density SNP data around the mutation locus.
2. Highlight *continuous* haplotype sharing between *two or more* individuals as you move away from the mutation locus. This sharing defines the ancestral segments.
3. Using a genetic map, record the genetic length of the segments. Specifically: record the segment "arm" lengths for each individual, i.e. the left arm length is the genetic distance between the mutation and the outermost marker heading in the direction of decreasing map position and the right arm length is the distance between the mutation and the outermost marker heading in the direction of increasing map position. Record the lengths in units of centiMorgans (cM).

To assess haplotype sharing for individuals with a single copy of the mutation (e.g. with a dominant mutation), or for individuals with copies of two distinct mutations at the same gene locus (i.e. with a compound heterozygous mutation), the data needs to be phased to resolve the underlying ancestral haplotype(s).[1] However, for individuals with two copies of the same recessive mutation, phasing is not required: ancestral segments are defined by continuous sharing of identical homozygous markers.[2]

[*] gandolfo@wehi.edu.au, or gandolfo@student.unimelb.edu.au
[†] bahlo@wehi.edu.au
[‡] terry@wehi.edu.au

[1] For an extensive discussion of phasing methods see the review by Browning & Browning (2011).

[2] To prepare the data in this case we recommend using the Perl script LINKDATAGEN (Bahlo & Bromhead, 2009), selecting the *complete data* option and then opening the resulting output in Excel.



SNP name	Genetic position (cM)	1	2	3	4	5
SNP-1	102.00	1	1	1	1	1
SNP-2	102.01	3	1	2	3	2
SNP-3	102.02	1	3	2	1	2
SNP-4	102.03	3	3	2	2	3
SNP-5	102.04	1	3	2	1	2
SNP-6	102.05	3	1	1	2	2
SNP-7	102.06	1	3	2	1	3
SNP-8	102.07	3	3	2	2	3
SNP-9	102.08	3	1	2	3	1
SNP-10	102.09	3	1	2	3	1
SNP-11	102.10	1	3	2	1	3
SNP-12	102.11	3	1	2	3	1
SNP-13	102.12	1	1	2	3	1
SNP-14	102.13	3	3	3	3	3
SNP-15	102.14	3	1	2	3	1
SNP-16	102.15	3	1	2	3	1
SNP-17	102.16	1	3	3	3	3
SNP-18	102.17	1	1	1	1	1
SNP-19	102.18	3	3	3	3	3
SNP-20	102.19	1	3	2	1	3
SNP-21	102.20	3	3	3	3	3
SNP-22	102.21	3	3	3	3	3
SNP-23	102.22	3	1	2	3	1
SNP-24	102.23	1	1	1	2	1
SNP-25	102.24	1	1	2	3	1
SNP-26	102.25	3	3	2	1	3
SNP-27	102.26	1	3	2	2	3
SNP-28	102.27	1	3	2	1	3
SNP-29	102.28	1	1	1	1	1
SNP-30	102.29	3	3	3	3	3
SNP-31	102.30	1	1	1	1	1
SNP-32	102.31	1	1	1	2	1
SNP-33	102.32	3	3	3	3	3
SNP-34	102.33	3	3	3	3	3
SNP-35	102.34	3	3	3	3	3
SNP-36	102.35	1	1	1	1	1
SNP-37	102.36	3	3	3	3	3
SNP-38	102.37	3	3	3	3	3
SNP-39	102.38	0	0	0	0	0
SNP-40	102.39	0	0	0	0	1
SNP-41	102.40	0	0	0	0	0
SNP-42	102.41	3	3	3	3	3
SNP-43	102.42	1	1	1	1	1
SNP-44	102.43	0	0	0	0	0
SNP-45	102.44	0	0	0	0	0

SNP-146	103.45	1	1	1	1	1
SNP-147	103.46	3	3	3	3	3
SNP-148	103.47	3	3	3	3	3
SNP-149	103.48	3	3	3	3	3
SNP-150	103.49	1	1	1	1	1
SNP-151	103.50	1	1	1	1	1
SNP-152	103.51	3	3	3	3	3
SNP-153	103.52	3	3	3	3	3
SNP-154	103.53	1	1	1	1	1
SNP-155	103.54	3	3	3	3	3
SNP-156	103.55	3	3	3	3	3
SNP-157	103.56	3	3	3	3	3
SNP-158	103.57	3	3	3	3	3
SNP-159	103.58	1	1	1	1	1
SNP-160	103.59	1	1	1	1	1
SNP-161	103.60	3	3	3	3	3
SNP-162	103.61	1	1	1	1	1
SNP-163	103.62	1	1	1	1	1
SNP-164	103.63	1	1	1	1	1
SNP-165	103.64	3	3	3	3	3
SNP-166	103.65	1	1	1	1	1
SNP-167	103.66	3	3	3	3	3
SNP-168	103.67	1	1	1	1	1
SNP-169	103.68	1	1	1	1	1
SNP-170	103.69	1	1	1	1	1
SNP-171	103.70	1	1	1	1	1
SNP-172	103.71	1	1	1	1	1
SNP-173	103.72	3	3	3	3	3
SNP-174	103.73	3	3	3	3	3
SNP-175	103.74	3	3	3	3	3
SNP-176	103.75	3	3	3	3	3
SNP-177	103.76	1	1	1	1	1
SNP-178	103.77	1	1	1	1	1
SNP-179	103.78	1	1	1	1	1
SNP-180	103.79	3	3	3	3	3
SNP-181	103.80	1	1	1	1	1
SNP-182	103.81	3	3	3	3	3
SNP-183	103.82	1	1	1	1	1
SNP-184	103.83	3	3	3	3	3
SNP-185	103.84	3	3	3	3	3
Mutation						
SNP-186	103.85	3	3	3	3	3
SNP-187	103.86	1	1	1	1	1
SNP-188	103.87	1	1	1	1	1
SNP-189	103.88	0	0	0	0	3
SNP-190	103.89	1	1	1	1	1
SNP-191	103.90	3	3	3	3	3
SNP-192	103.91	3	3	3	3	3

(a) (b) **Box 1** (c) **Box 2**

Figure 1: An invented SNP data set on five individuals, lined up around the mutation (locus highlighted *orange*), with continuous haplotype sharing between two or more individuals highlighted *green*, which we infer to be ancestral segments: (a) shows segment arms to the left of the mutation (i.e. heading in the direction of decreasing map position); (b) is a magnified view of Box 1; and (c) is a magnified view of Box 2. Numbers code for genotypes: 1 = AA, 2 = AB, 3 = BB, and 0 = missing. Here we assume each individual has two copies of the same recessive mutation, thus the ancestral segments are defined by continuous sharing of identical homozygous markers. The left segment arm lengths are the genetic distance between the mutation and the outermost markers (the right arm lengths are found similarly).

Note that mutations and errors in marker data should be obvious: continuous sharing interrupted by one of these features will typically be followed by another large run of continuous sharing. When these features are obvious, they should be ignored. These features may not be obvious near segment ends, but in that case they have only a minor influence. Missing allele or genotype calls should also be ignored.

2 Run the script

Now take the following steps to execute the script in R:

1. Enter the length data:

   `> l.lengths = c(`*`left arm length for individual 1,...,left arm length for individual n`*`)`

   `> r.lengths = c(`*`right arm length for individual 1,...,right arm length for individual n`*`)`

   Note that the components of each vector must correspond to the same individuals, i.e. the i^{th} position of each vector must refer to the same individual. Note also that the lengths must be in units of centiMorgans (cM).

2. Enter the desired parameters:

   `> confidence.coefficient =` *`number between 0 and 1`*

   `> chance.sharing.correction = TRUE or FALSE`

   Confidence interval coverage is specified with `confidence.coefficient` (e.g. for 95% confidence intervals use 0.95). Set `chance.sharing.correction = TRUE` to correct for chance sharing, or `FALSE` otherwise. If this is set to `TRUE`, the following extra parameters are needed:

   `> median.allele.frequency =` *`number between 0 and 1`*

   `> markers.on.chromosome =` *`number`*

   `> length.of.chromosome =` *`number`*

   These parameters refer to the chromosome with the mutation: the first specifies the median population frequency of the A marker alleles (or the B marker alleles) across the chromosome; the second specifies the number of markers on the chromosome; and the third specifies the genetic length (in cM) of the chromosome.

3. Enter the following to execute the script:

   `> source(Mutation_Age_estimation.R)`

   Note that the script needs to be in the R working directory (either place the script there, or set the working directory to where the script is located).

3 Example

Entering the following length data (from a hypothetical sample of 6 individuals) and parameters:

```
> l.lengths = c(1.2, 3.4, 1.6, 4.8, 4.8, 2.3)
> r.lengths = c(2.0, 3.8, 2.7, 2.1, 3.8, 1.8)
> confidence.coefficient = 0.95
> chance.sharing.correction = TRUE
> median.allele.frequency = 0.7
> markers.on.chromosome = 68378
> length.of.chromosome = 290,
```

and running the script

```
> source(Mutation_Age_estimation.R),
```

produces the following output:

```
   [1] Assuming an 'independent' genealogy:  age estimate = 27.5 generations,
with confidence interval (15.5,49.2)
   [1] Assuming a 'correlated' genealogy:  age estimate = 22.3 generations,
with confidence interval (7.9,67.5)
```

References

Bahlo, M. & Bromhead, C. J. (2009). Generating linkage mapping files from Affymetrix SNP chip data. *Bioinformatics*, *25*(15), 1961–1962.

Browning, S. R. & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, *12*(10), 703–714.

Gandolfo, L. C., Bahlo, M., & Speed, T. P. (2014). Dating rare mutations from small samples with dense marker data. *Genetics*, *197*(4), 1315–1327.