October 15, 2013

1 usage:

./stats_pop -m msfile

Options:

-m filename with ms-like data

-a anc/maf (whether known and unknown ancestral allele)

-f l u (lower and upper limint in KB for filtering pairwise SNP to calculate the statistics of LD and recombination. default is 5 to 10kb)

-p 1 or 2 (1 for phased, 2 for unphased data)

-h help

2 Output ~ similar to input for EstL

The header of the output is required by EstL, e.g.:

- Known ancestral allele
- UnPhased data
- Filter for pairwise LD: 5.000000 15.000000 kb

Then the header for all the summary statistics calculated followed by the statistics for each region.

2.1 List of summaries

- r, region number
- n_1 and n_2, sample size in sample 1(2). $n = n_1 + n_2$
- L, length of the region in bp $L = z_2$

2.1.1 Summaries of the polymorphisms

- S, the total number of segregating sites found in the data set. $S = S_1 + S_2 + S_s + S_f$.
- S_1 and S_2, the number of segregating sites with the derived alleles polymorphic uniquely in the sample from population 1 (2).
- S_s, the number of segregating sites with the derived allele shared by both population samples. $S_s = S_{ss} + S_{sf1} + S_{sf2}$ and $S_s = S_{sl} + S_{sh}$.
  - S_{ss}, the number of segregating sites with the derived allele shared AND polymorphic in both population samples.
  - $S_{sf1}$and S_{sf2}, the number of segregating sites with the derived allele shared AND polymorphic ONLY in the sample from population 2 (or 1) (i.e., fixed in the sample from population 1 or 2).
  - $S_{sl}$and S_{sh}, the number of segregating sites with the derived allele shared AND at low (high) frequency (i.e., $\leq 10\%$ ($> 10\%$)).
- S_f, the number of segregating sites with the alleles fixed between the population sampled. $S_f = S_{f1} + S_{f2}$.
  - $S_{f1}$and S_{f2}, the number of segregating sites with the derived allele fixed in the sample from population 1 (2).
- S_o, the number of singletons found in the data set. I.e.: $S_o = S_{o1} + S_{o2} \in [0, S]$ where the frequency of the derived allele is $1/(n)$.

- $S_{o1}$and S_{o2}, the number of singletons found in the sample from population 1 (2). I.e., $S_{o1} \in [0, S_1]$ where the frequency of the derived allele is $1/n_1$ ($S_{o2} \in [0, S_2]$ where the frequency of the derived allele is $1/n_2$).

2.1.2 Frequencies of the polymorphisms

The mean derived allele frequencies for seven types of segregating sites described above.

- $F(S)$The mean derived allele frequency over n haplotypes for the S segregating sites in the data set.
- $F(S_1)$and $F(S_2)$ The mean derived allele frequency over n_1 (n_2) haplotypes for the segregating sites in S_1 (S_2).
- $F(S_s)$, $F(S_{ss})$, $F(S_{sf1})$, $F(S_{sf2})$, $F(_{S_s l})$ and $F(S_{sh})$ The mean derived allele frequency over n haplotypes for the segregating sites in S_s, S_{ss}, S_{sf1}, S_{sf2}, S_{sl} and S_{sh}, respectively.

2.1.3 Measures of differentiation

- F_{ST}, the level of differentiation between the two population samples. F_ST=1-(H_w1+H_w2)/H_b (Wright 1931; Hudson et al. 1992).
- H_b, the mean pairwise differences between both population samples.
- H_{w1} and H_{w2}, the mean pairwise differences within the sample from population 1 (2).
- S_{nn}, the nearest-neighbor statistic, which measures how often the nearest neighbors of haplotypes are found in the same population sample.
  - $S_n n = \Sigma[X_i/n] = \Sigma[W_i/(T_i * n)]$, $i \in [1, n]$ (Hudson 2000).
  - W_i: The number of nearest neighbors to haplotype i that are from the same population sample as haplotype i.
  - T_i: The number of nearest neighbors of haplotype i.
  - X_i: The fraction of nearest neighbors of haplotype i that are from the same population sample as haplotype i.

2.1.4 Estimators of the population mutation rate

- π, π_1 and π_2, The unbiased estimator of the population mutation rate calculated from the average nucleotide diversity (or heterozygosity) for the full sample and samples from population 1 and 2, π (Nei and Li 1979, but equations 10 and 12 of Tajima 1989).
  - $\pi = \Sigma[2 * p * (1 - p) * n/(n - 1)]$for the segregating sites in S.
  - $\pi_1 = \Sigma[2 * p * (1 - p) * n_1/(n_1 - 1)]$for the segregating sites in S_1.
  - $\pi_2 = \Sigma[2 * p * (1 - p) * n_2/(n_2 - 1)]$for the segregating sites in S_2.
- θ_W, θ_{W1} and θ_{W2}. The unbiased estimator of the population mutation rate calculated from the number of segregating sites, $\backslash theta_W$ (Watterson 1975).
  - $\theta_W = \Sigma[1/(1/n)]$for the segregating sites in S.
  - $\theta_W 1 = \Sigma[1/(1/n_1)]$for the segregating sites in S_1.
  - $\theta_W 2 = \Sigma[1/(1/n_2)]$for the segregating sites in S_2.
- θ_H, $\theta_H 1$ and $\theta_H 2$ The unbiased estimator of the population mutation rate weighted by the homozygosity of the derived allele, θ_H (Fay and Wu 2000).
  - $\theta_H = \Sigma[2 * (p^2)/[n(n - 1)]]$for the segregating sites in S.
  - $\theta_H 1 = \Sigma[2 * (p^2)/[n_1(n_1 - 1)]]$for the segregating sites in S_1.
  - $\theta_H 2 = \Sigma[2 * (p^2)/[n_2(n_2 - 1)]]$for the segregating sites in S_2.

2.1.5 Tests of neutrality

- D, D_1 and D_2 The test of neutrality from the difference between π and θ_W, Tajima's D (Tajima 1989).
  - $D = (\pi - \theta_W)/\sqrt{(var(\pi - \theta_W))}$.
  - $D_1 = (\pi_1 - \theta_{W1})/\sqrt{(var(\pi_1 - \theta_{W1}))}$.
  - $D_2 = (\pi_2 - \theta_{W2})/\sqrt{(var(\pi_2 - \theta_{W2}))}$.
- H, H_1 and H_2 The test of neutrality from the difference between π and θ_H, Fay & Wu's H (Fay and Wu 2000).
  - $H = \pi - \theta_H = -2 * p(2 * p - 1)n/(n - 1)$for the segregating sites in S.
  - $H_1 = \pi_1 - \theta_{H1} = -2 * p(2 * p - 1)n_1/(n_1 - 1)$for the segregating sites in S_1.
  - $H_2 = \pi_2 - \theta_{H2} = -2 * p(2 * p - 1)n_2/(n_2 - 1)$for the segregating sites in S_2.
- $D^\star$, $D_1^\star$ and $D_1^\star$ The test of neutrality from the difference between the number of singletons and the number of segregating sites, Fu & Li's $D^\star$ (Fu & Li 1993).
- D$^\star = (S - S_o)/\sqrt{(var(S - S_o))}$.
- $D_1^\star = (S_1 - S_{o1})/\sqrt{(var(S_1 - S_{o1}))}$.
- $D_2^\star = (S_2 - S_{o2})/\sqrt{(var(S_2 - S_{o2}))}$.

2.1.6 Measure of recombination

These numbers are calculated between SNPs separated by the distance specified qith option (-f)

- D', D_1' and D_2', measure of LD in total and population samples.
- r^2, r_1^2 and r_2^2, correlation coeafficient total and population samples.
- n_H, n_{H1} and n_{H2}Number of haplotype
- R_m, R_{m1} and R_{m2} Minimun number of recombiantion events (Hudson etal 1992).

References

Becquet, C., and M. Przeworski. 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Res. 17:1505-1519.

Fay, J. C., and C. -I. Wu. 2000. Hitchhiking Under Positive Darwinian Selection. Genetics 155:1405-1413.

Fu, Y. -X. and W. -H. Li. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

Hudson, R. R., 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23:183-201.

Hudson, R. R., 1990. Gene genealogies and the coalescent process, in D. Futuyma and J. Antonovics (eds), Oxford Surveys in Evolutionary Biology, Vol. 7:1-44.

Hudson, R. R., Slatkin, M., and Maddison, W. P., 1992. Estimation of levels of gene flow from DNA sequence data. Genetics, 132:583-589.

Hudson, R. R., 2000. A New Statistic for Detecting Genetic Differentiation. Genetics 155:2011-2014.

Hudson, R. R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338. .

Nei, M., and W. H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76:5269-5273.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Wakeley, J., and J. Hey. 1997. Estimating ancestral population parameters. Genetics 145:847-855.

Watterson, G. A., 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256-276.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16:97-159.