Discrimination between ancestral polymorphism and recent introgression as causes of shared polymorphic sites between species

Miguel Navascués, Frantz Depaulis

Distinguishing shared ancestral polymorphism from recent introgression

Miguel Navascués & Frantz Depaulis
CNRS UMR 7625 Écologie et Évolution
m.navascues@gmail.com

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ PIERRE ET MARIE CURIE / ÉCOLE NORMALE SUPÉRIEURE
This work was funded by the ANR Biodiversité project TRANSBIODIV

Studying the genetic diversity of two species (or populations) it is possible to derive measures of genetic differentiation, namely, F-statistics. From such statistics it is possible to estimate time of divergence (*T*) or migration rate (*m*) by assuming two alternative extreme population structure models: isolation or migration[1] (figure 1). In this work we take advantage of recombination in DNA sequences to distinguish whether shared polymorphic sites are due to ancestral polymorphism (i.e. isolation model) or by recent introgression (i.e. migration model). Ancestral shared polymorphism is expected to have had more chances to recombine than shared polymorphism produced by recent introgression. Thus, linkage disequilibrium (LD) between shared site might indicate introgression[2].

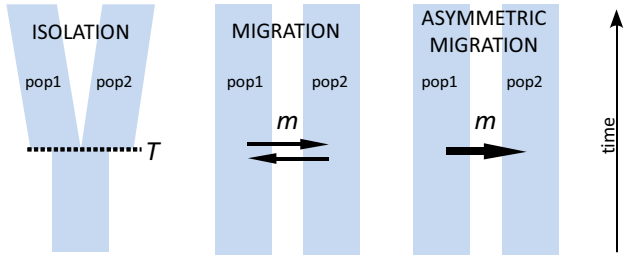

Figure 1: Population structure models. Blue shapes represent the population history along time.

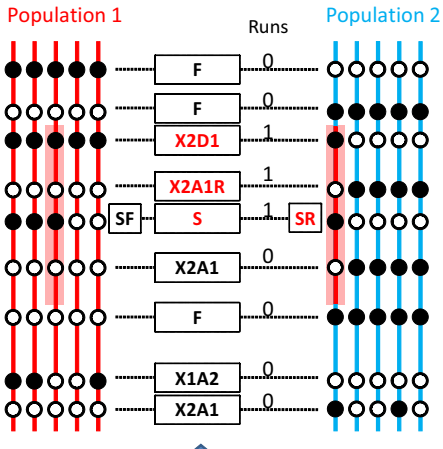

Figure 2: Alignment of DNA sequences from two populations (red and blue). Circles represent polymorphic sites: white represent the ancestral state and black the derived state. An introgressed block in population 2 (from population 1) is represented. F: fixed differences; S: shared polymorphism (derived state rare in population 2); X2A1: exclusive polymorphism in population 2 with ancestral state fixed in population 1; X2D1: exclusive polymorphism in population 2 with derived state fixed in population 1; X2A1R: X2A1 with ancestral state in low frequency in population 2. An instance of run coding for asymmetric migration is shown.

To avoid sensitivity to frequencies of classical LD measures, we propose to study LD by means of statistical test of randomness of categories of segregating sites along an alignment (figure2): i.e. the runs tests[3,4]. A run is defined as a succession of consecutive alike sites succeeded by unlike sites.

Tests for detection of introgression:

- Randomness of runs of fixed (F) differences, using shared polymorphic sites (S) and exclusive (X) polymorphisms as comparison.
- Randomness of runs of shared (S) polymorphic sites, using fixed (F) sites and exclusive (X) polymorphisms as comparison.

Test for detection of asymmetrical introgression (needs mutations oriented with an outgroup):

- Randomness of runs of polymorphisms expected in introgression 1→2: X2D1, X2A1R, SR (see figure 2), using the rest of S and X as comparison.

Coalescent simulations were used to produce dataset for the different models (figure 1) and the tests were applied on the output to assess the power and robustness of this tests.

CONCLUSIONS: The results (figure 3) show that this approach can be used to distinguish recent introgression from ancestral polymorphism. The power of the tests increased with the rate of recombination (results not shown).

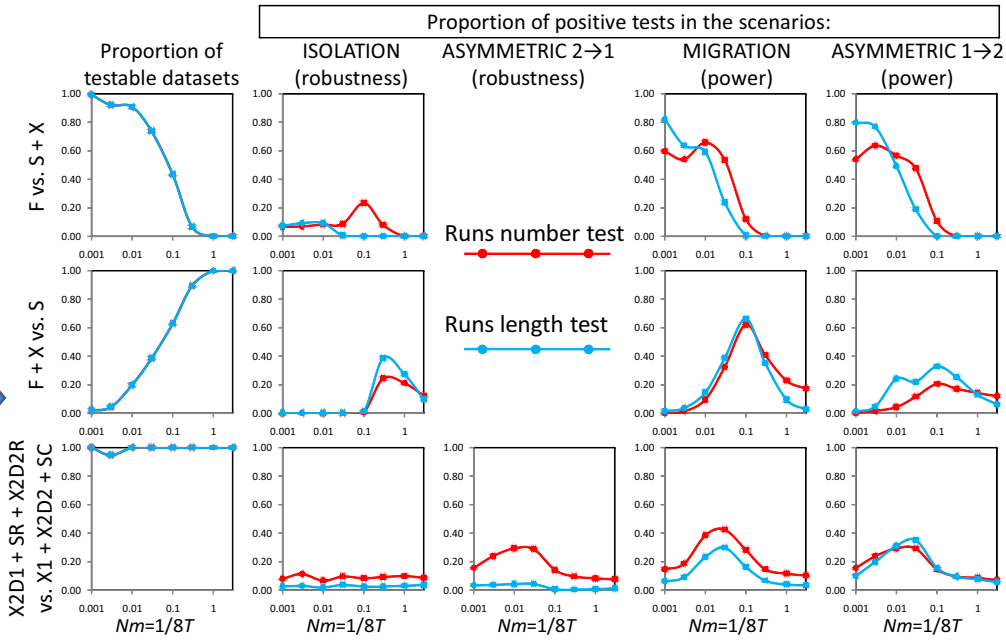

Figure 3: Results of runs test on simulated data sets under the three models (in columns). For each model eight levels of genetic differentiation (***T***ime of divergence/***m***igration rate) are considered. First two rows are test to detect introgression in any direction and the third row is focus in detecting asymmetrical introgression from population 1→2. A low rate of positive test under null models (isolation or asymmetric 2→1) indicate a robust test. A high rate of positive test under the model of migration or asymmetric migration 1→2 indicate a high power of the test.

REFERENCES

[1] Wakeley (1996) *Theoretical Population Biology*, 49:369–386.
[2] Machado et al. (2002) *Molecular Biology and Evolution*, 19:472–488.
[3] Takahata (1994) *Immunogenetics*, 39:146–149.
[4] Sneath (1998) *Bioinformatics*, 14:608–616.