RADseq *in silico* simulator

****
1. SUMMARY

The scripts described here were used to simulate RAD seq data for pairs of diploid individuals, sampled within the same population or in different populations.

ms (Hudson et al. 2002; http://home.uchicago.edu/rhudson1/source/mksamples.html) and seq-gen (Rambaut et al. 1997; http://tree.bio.ed.ac.uk/software/seqgen/) are used to simulate quadruplets of genome sequences, RAD_abc then simulates the retrieval of restriction sites and calculates the summary statistics associated with those RAD data.

****
2. PROGRAMS

run_simu_rad_abc.c uses a file containing a list of population parameter values (population parameter file) and runs ms, seq-gen and rad_abc.

rad_abc.c simulates RAD experiments for each sequence data (rad param file).

run_simu_rad_abc requires the installation of ms (Hudson et al. 2002) and seq-gen (Rambault et al. 1997) programs. These 2 programs should be accessible in the default path. Alternatively the exact path to these executables can be given as a parameter in the command line (see section 4.)

****
3. PARAMETER FILES

2 parameter files are required for the simulations. The first one (Population parameter file) defines the characteristics of the population(s) from which the individuals are drawn. To simulate the retrieval of RADseq data from 2 diploid individuals, sequence data corresponding to 4 haploid genomes are generated. The 2 individuals can be drawn from a structured population; the 2 pairs of haploid genomes come from 2 subpopulations n1 and n2, diverging since time t. To retrieve data corresponding to a single panmictic population, t is set to 0.

The second one (RAD parameter file) defines the coverage level in the simulated RADseq data, that is, the proportion of RAD loci (sequences flanking an intact restriction site) that are actually sequenced. In our manuscript, we set c1 = c2 = 1, so that all RAD loci are sequenced in both individuals, but this can be reduced, to simulate limited sequencing depth. This file also allows to increase the coverage probability of the second allele if the first one is sequenced, in order to investigate hypothetical cases in which the probability to validate sequence data for orthologs would not be independent at a post-sequencing stage. In our manuscript, we set cp = 0, so that there is no such effect.

For both parameter files, each line corresponds to a set of parameter values in the following order:

-Population parameter file (e.g. param_pop.txt):
N1 (Sets the initial effective size of subpopulation from which individual I1 is derived to N1 * N0, where N0 is the size of the ancestral population), N2 (Sets

the initial effective size of subpopulation from which individual I2 is derived to N2 * N0), t (divergence time between the 2 populations, in the case of structured populations) and theta (population-scaled mutation rate, theta = $4N_e u$)
In the example (simulation of genomes drawn from a panmictic population): N1=N2=1 and t=0.

-RAD parameter file (e.g. param_rad.txt):
c1 and c2 (proportions of loci covered in individuals 1 and 2, respectively), cp (additional probability to sequence the second allele when the 1st is sequenced).
In the example, as in the simulations used in our manuscript: c1=c2=1 and cp=0.

#######
4. COMMANDS:

Compilation:
cc -Wall -o RAD_abc RAD_abc_V3.3.c
cc -Wall -o run_simu_RAD_abc run_simu_RAD_abc_V3.3.c

Simulations:
./run_simu_RAD_abc out.txt param_pop.txt [ms=PATH_MS] [seqgen=PATH_SEQ-GEN]

note: [ms=PATH_MS] and [seqgen=PATH_SEQ-GEN] are optional parameters that can be used to specify the location of ms and seq-gen executables.

#######
5. OUTPUT DESCRIPTION

In the following, the 2 individuals whose genomes are simulated are called respectively I1 and I2. Homologous chromosomes a and b within these individuals are called respectively I1a and I1b and I2a and I2b.

A RAD locus is defined as the 100 bp-long sequence downstream (3') of the position of a restriction site, which has to be present in at least one of the four haplotypes. At a given RAD locus, the restriction site itself can be polymorphic. We will hereafter use the term "RAD haplotype" at a given locus to refer to haplotypes associated with an intact restriction site. A RAD locus is shared by two individuals if the restriction site is present in at least one haplotype of each individual.
If an individual contains only one RAD haplotype (i.e. the restriction site is absent in its second haplotype) then the RADseq experiment provides sequence reads corresponding to only one of its two alleles. Thus, the RADseq approach leads to consider this individual as monomorphic at this locus (i.e. I1a = I1b, or I2a = I2b).

The **true average heterozygosity (H_{true})** within one individual was computed according to this formula:

$$H_{true} = \frac{\sum_{k=1}^{k=n} d\_True\_ab_k}{\sum_{k=1}^{k=n} L_k} \qquad (1)$$

where n is the number of RAD loci in that individual, L_k is the length of locus k (here L=100 for all RAD loci), and $d\_True\_ab_j$ is the genetic distance at locus locus k (i.e. the number of heterozygous sites at this locus).

The **average RAD-inferred heterozygosity (H_{RAD})** within one individual was computed according to this formula:

$$H_{RAD} = \frac{\sum_{k=1}^{k=n} d\_RAD\_ab_k}{\sum_{k=1}^{k=n} L_k} \quad (2)$$

where $d\_RAD\_ab_k$ is the genetic distance observed in RAD data at locus locus k (with $d\_RAD\_ab_k$ = 0 if the individual contains only one RAD haplotype at this locus, or otherwise $d\_RAD\_ab_k$ = number of heterozygous sites at this locus).

The **average RAD-inferred nucleotidic diversity, pi_RAD**, was computed according to this formula:

$$pi\_RAD = \frac{\sum_{k=1}^{k=n} pi\_RAD_k}{\sum_{k=1}^{k=n} L_k} \quad (3)$$

for each RAD locus k, dxy_k is calculated as follow:

$$pi\_RAD_k = \frac{1}{h1*h2} * \sum_{i=1}^{i=h1} \sum_{j=1}^{j=h2} d\_RAD_{ij} \quad (4)$$

where d_RAD_{ij} is the genetic distance between allele i of individual I1 and allele j of individual I2, and h1 and h2 are the number of RAD haplotypes in individual I1 and I2 respectively.

---out.txt : Parameters and summary statistics corresponding to RADseq data

-1. n1	- Sets the initial effective size of subpopulation from which individual I1 is derived to N1 * N0 (where N0 is the size of the ancestral population)
-2. n2	- Sets the initial effective size of subpopulation from which individual I1 is derived to N1 * N0 (where N0 is the size of the ancestral population)
-3. t	- divergence time between the 2 populations in case of structured populations
-4. theta=pi_TRUE	- population-scaled mutation rate
-5. c1	- proportion of loci covered in individual I1
-6. c2	- proportion of loci covered in individual I2
-7. cp	-additional probability to sequence the second allele when the 1st is sequenced
-8. H_rad_i1	**-** RAD-inferred heterozygosity in individual I1 (see formula (2)).
-9. H_true_i1	**-** true heterozygosity at RAD loci (i.e. including non-RAD haplotypes) within individual i1 (see formula (1)).
-10. pmono_i1	- proportion of observed monomorphic loci within I1
-11. H_rad_polymorphic_i1	- RAD-inferred heterozygosity computed only on RAD loci that are polymorphic within I1 (i.e. containing at least one SNP).
-12. H_rad_i2	- cf H_rad_i1
-13. H_true_i2	- cf H_true_i1
-14. pmono_i2	- cf pmono_i1
-15. H_rad_polymorphic_i2	- cf. H_rad _polymorphic_i1
-16. pi_rad	**-** average distance between I1 and I2 at RAD haplotypes (see formula (3))
-17. pshared_1	**-** proportion of shared loci : (nb of shared loci)/(nb of loci I1)
-18. pshared_2	- proportion of shared loci : (nb of shared loci)/(nb of loci I2)
-19. pi_rad_polymorphic	- average distance between i1 and i2 at polymorphic RAD loci only (at least one SNP at this loci, i.e. at least 1 haplotype different from the 3 others)