Best configuration analysis

Nick Sard

July 18, 2019

Step 1) Example BestConfig file

off*	P1	P2
o1_m1_d1	P1_1	P2_1
o2_m1_d1	P1_1	P2_1
o1_m1_d2	P1_1	P2_2
01_m2_d2	P1_10	P2_3
01_m10_d10	P1_10	P2_10
…	…	…

*Note that parent IDs that produced each offspring are embedded in each offspring's name.

Step 2) Using information from BC file to determine assignment accuracy and errors

Dyads		Inferred parents					Known parents				
off1	off2	P1	P2	P1	P2	IR	P1	P2	P1	P2	KR
o1_m1_d1	o2_m1_d1	P1_1	P2_1	P1_1	P2_1	FS	m1	d1	m1	d1	FS
o1_m1_d1	o1_m1_d2	P1_1	P2_1	P1_1	P2_2	HS	m1	d1	m1	d1	HS
o1_m1_d1	01_m10_d10	P1_1	P2_1	P1_10	P2_10	UR	m1	d1	m10	d10	UR
o1_m2_d2	01_m10_d10	P1_10	P2_3	P1_10	P2_10	HS	m2	d2	m10	d10	UR
…	…	…	…	…	…	…	…	…	…	…	…

Step 3) Summarize counts of known and inferred dyads within a BestConfig file in a matrix

Known dyads \ Inferred dyads	FS	HS	UR	Total
FS	140	0	0	140
HS	2	302	14	318
UR	0	6	9436	9442
Total	142	308	9450	10000

Calculating the accuracy and false negative rate using information from step 3.

Definitions:

Accuracy – The proportion of inferred dyads (FS, HS, or UR) that were correctly assigned.

False negative rate – The proportion of known dyads (FS, HS, or UR) that were incorrectly inferred.

Example:

Accuracy for HS in step 3 was 302/308 = 98% and the false negative rate for HS was 16/318 = 5%. That is, of all the HS inferred, 98% were correct; however, 5% of known HS were falsely inferred as another relationships (2 were incorrectly inferred as FS and 14 were inferred as UR).

R Markdown

There are there major steps when analyzing a COLONY BestConfig (BC) file generated using simulated offspring genotypes where the user provides no parental genotypes. When COLONY does not have candidate parental genotypes it will infer a parent, and its associated genotypes. These parents are given generic names (e.g., *1 or #3). Given that the user does not know which column represents each sex, each parent (P) column is given a generic column name (P1, P2), and the inferred parent IDs are changed slightly to reflect the header columns (e.g.,* 1 and #3 are changed to P1_1 and P2_3). It is also important to note that the offspring ID column (off) contains information about the known parents that produced each offspring, which is embedded in each offspring ID. For instance, the user knows that Mom 1 (m1) and Dad 1 (d1) produced offspring "o1_m1_d1". Step 2 consists of evaluating all unique pairwise comparisons among all offspring genotyped (i.e., dyads). For each dyad (i.e. row) the user can determine the inferred relationship (IR) and known relationship (KR) between the two individuals by counting how many parents the two individuals share. That is, full-siblings (FS, 2 parents shared), half-siblings (HS, 1 parent shared), and unrelated (UR, 0 parents shared) dyads can be determined using both the Inferred and Known Parent IDs. Using information in step three the accuracy and false negative rate can be calculated for each simulation. See definitions and example in Figure for more details.