Comparison of the performance of two Bayesian clustering methods for detecting multiple gene pools in mixed samples

ROBIN S. WAPLES[1] and MICHELE MASUDA[2]
[1]*Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd. East, Seattle WA 98112, USA, robin.waples@noaa.gov.*
[2]*Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, TSMRI, 17109 Point Lena Loop Road, Juneau, AK 99801, USA*

ABSTRACT

We used simulated data to compare performance of two genetic clustering programs (STRUCTURE and HWLER) to resolve mixtures of individuals from two different populations. The populations were separated by genetic differences indexed by F_{ST} values that ranged from < 0.01 to 0.06, and artificial mixtures included 100 individuals from each population. The number of genetic markers used (16 'microsat' loci) is comparable to what is commonly used in studies of cetaceans, including North Pacific common minke whales. Both methods correctly identified only a single gene pool under the control scenario (a single panmictic population), but both also detected only one gene pool with the most difficult two-population mixture (target F_{ST} = 0.007). STRUCTURE performed better than HWLER at slightly higher levels of genetic differentiation (target F_{ST} = 0.02 – 0.03), but HWLER was more reliable for target F_{ST} = 0.045 – 0.06. For the latter scenarios, STRUCTURE showed a tendency to overestimate the number of populations (3 or 4 rather than 2) and inconsistency among replicate runs using the same data. Follow-up studies would be useful to examine why STRUCTURE showed this erratic behavior at moderate levels of differentiation but not with what should be more difficult mixtures with lower F_{ST} values.

INTRODUCTION

For many cetaceans, it is difficult to collect samples on the breeding grounds, where it is reasonable to expect that individuals from only a single population or stock are subject to sampling. In these cases, samples, if any, must be collected from migrating individuals or on feeding grounds, and both these scenarios can produce samples that include a mixture of individuals from more than one population. When these are the only types of samples available, it can be very challenging to make robust inferences about the number of populations that exist and their biological characteristics.

North Pacific common minke whales (NPM) follow this pattern: breeding grounds are presumed to occur in the low latitudes in the Western Pacific but have not been identified, and the numerous biological samples that have been collected are primarily from animals migrating to and from their feeding grounds (Pastene et al. 2010; Wade et al. 2010). The difficult challenge is to estimate the number of contributing populations and their relative contributions in space and time, in the absence of any pure 'baseline' data from the breeding populations to anchor the analysis. Clustering methods are designed to address this type of problem. The most popular Bayesian clustering method for genetic data is STRUCTURE (Pritchard et al. 2000), which has been applied to a number of cetacean datasets. Under some circumstances, this program can produce impressive results, but its performance has been variable under the relatively low levels of genetic differentiation typically found within cetacean species (e.g., Kanda et al. 2010; see also discussion of BCB bowhead whale genetics papers in the report of the 2007 Scientific Committee meeting in Anchorage, AK). Another Bayesian clustering program (HWLER; Pella and Masuda 2006) can outperform STRUCTURE under at least some circumstances, and it is of considerable practical interest to explore this topic further. Here, we expand on a working paper presented at SC63 in Tromsø and compare the performance of STRUCTURE and HWLER with artificial mixtures of populations with low levels of genetic differentiation relevant to NPM and other cetaceans.

METHODS

Simulated data

Genetic data were simulated with EASYPOP (version 2.0.1; Balloux 2001). We generated data for two populations, each of constant size 1000, equal sex ratio, and random mating. Sixteen diploid loci were generated for each individual. Each dataset was created with a targeted F_{ST} between the two populations. Migration rates were varied to achieve target F_{ST} (0, 0.007, 0.02, 0.03, 0.045, and 0.06), and artificial mixtures were created with 100 individuals sampled from each population (200 total individuals). We analyzed 10 replicated datasets for each target F_{ST}.

F-*statistics*

Genetic differentiation between the two simulated populations was measured with Weir & Cockerham's θ (1984). We used FSTAT (version 2.9.3.2; Goudet 1995) to calculate θ and report the estimated θ as F_{ST}.

Estimating the number of populations

STRUCTURE

STRUCTURE (version 2.3.3; Pritchard et al. 2000) implements a Markov Chain Monte Carlo (MCMC) Gibbs sampler that assigns individuals to clusters based on their multilocus genotypes. The admixture model with correlated allele frequencies was assumed for the STRUCTURE runs, and all simulated individuals were input as unlabeled, i.e. origin unknown. Based on trace plots of various parameters for a small number of runs, we determined 200,000 MCMC samples to be sufficient for convergence of the MCMC chains and thereafter for expediency fixed the number of samples for all runs. The first one-half of chains was discarded as burn-in, and the remaining samples were used for approximating posterior distributions of parameters of interest.

STRUCTURE, which is based on the finite mixture model, requires the number of clusters or populations (k) be specified. In practice STRUCTURE is run on trial numbers of populations, which we set from 1 to 5. We followed the *ad hoc* procedure outlined in Pritchard et al. (2000) for approximating the posterior distribution of the number of clusters in the range of k specified by the user and report the posterior probabilities of k. Since the Gibbs sampler tends to not explore the parameter space well, 3 independent chains started from different initial values were run on each of the 10 replicated datasets and for each k (1-5). STRUCTURE also outputs, for each individual, posterior mean estimates (Q) of the fractions of the genome that are inherited from the inferred clusters. We report a few examples of estimated Q in the form of Q plots.

HWLER

Like STRUCTURE, HWLER (Pella and Masuda 2006) implements a Bayesian model-based clustering method that probabilistically assigns individuals to clusters based on their multilocus genotypes. HWLER, however, replaces the finite mixture model used by STRUCTURE with a Dirichlet process mixture model that allows the number of clusters to be open ended and not fixed. HWLER, therefore, does not require multiple runs for trial numbers of clusters but provides the posterior distribution of the number of clusters in a mixture sample from a single run.

We examined trace plots of the estimated numbers of clusters for a few runs and observed stabilization of the estimated numbers within 210,000 MCMC samples (the first one-half of each chain was discarded as burn-in). For convenience, we then fixed the number of MCMC samples for subsequent runs. An exception occurred for the datasets with target $F_{ST} = 0.03$. We observed convergence required more MCMC sampling and therefore ran HWLER for 2,100,000 samples with a burn-in discard of 1,050,000 samples. HWLER assumes mixture individuals are from separate breeding populations, i.e. no admixture was assumed.

Gibbs sampling alone has a tendency to not explore the parameter space well and may remain near a single mode. HWLER improves mixing by alternating sequentially-allocated merge-split (SAMS) sampling with Gibbs sampling, and we alternated two SAMS samples with five Gibbs scans. While the Gibbs scans probabilistically assign individuals one at a time to clusters, the SAMS sampling allows for probabilistic merging and splitting of groups of individuals. The problem of detecting too few clusters among similar individuals may be avoided with these group updates. See Pella and Masuda (2006) for details of their adaptation of the Gibbs and split-merge sampler of Jain and Neal (2004) to genetic data with improvements to the splitting by Dahl (2003). The improved mixing by HWLER makes multiple runs from different initial values unnecessary.

RESULTS

Realized heterozygosity at the last generation of the simulations (when samples were taken) ranged from 0.71 to 0.77 (Table 1), which is comparable to what has been reported for microsatellite data for North Pacific common minke whales (Kanda et al. 2010) and many other species. Under the panmixia scenario (true $k = 1$), both methods correctly identified a single gene pool in 100% of the replicates (Table 1). Conversely, both methods failed 100% of the time with the hardest two-population mixture (target $F_{ST} = 0.007$), in each case identifying only a single population. This result held for both the equilibrium migration scenario and one that involved a short period (16 generations) of complete isolation to produce the target F_{ST}.

Performance of the two methods differed for moderate ranges of target F_{ST} (0.02-0.06; Table 1). For target F_{ST} = 0.02-0.03, STRUCTURE found two populations in 100% of the replicates, while HWLER found only one population 100% of the time at target $F_{ST} = 0.02$ and 10% of the time at target $F_{ST} = 0.03$. Curiously, STRUCTURE had more difficulty with the mixtures with slightly larger F_{ST} (0.045-0.06), while HWLER got all of those correct. For these scenarios, all the mistakes by STRUCTURE were inferring too many gene pools (3 or 4 rather than 2). Furthermore, for many of the replicate datasets, results for STRUCTURE differed depending on the starting conditions (Appendix Table 1). These varying results in general did not change with longer runs for the few replicates we tested (data not shown). For example, we repeated the analyses with longer runs for 3 replicate datasets and target $F_{ST} = 0.045$ where STRUCTURE had estimated the number of populations as 2 to 4 (correct k = 2). We ran STRUCTURE for 1 million samples and discarded the first 900,000 as burn-in. Estimates of the most probable numbers of populations were the same as the estimates from the shorter runs except for two runs where the estimates increased by 1 (both incorrect). Furthermore, the variation in outcomes across runs with the same dataset was not always reflected in uncertainty in choice of k within each run. For example, for replicate 3 with target F_{ST} = 0.045 (actual $F_{ST} = 0.048$ for this replicate), the posterior probability for $k = 4$ was nearly 100% for runs 1 and 2 and was nearly 100% for $k = 3$ for run 3. In each case, therefore, the output from the program suggested confidence it had identified the true k, but the preferred value differed across runs and was (for this replicate) wrong in each case. For replicate 4 with target $F_{ST} = 0.06$, the three runs using the same data produced 3 different estimates of k, each having a nearly 100% posterior probability. Not all of the scenarios with heterogeneity among runs showed this same false sense of precision. For example, in run 2 of replicate 1 with target $F_{ST} = 0.045$, the posterior probability for k was split 35% for 2 populations and 65% for 3 populations (Appendix Table 1). Similarly, run 1 of replicate 8 for target $F_{ST} = 0.06$ also reflected considerable uncertainty in choice of k.

Further insights into these problematical mixtures are provided by Figure 1, which shows graphical results for one run for each of three replicate datasets for target $F_{ST} = 0.06$. Individuals in the plots are ordered by population (i.e. the first 100 individuals are from one population and the second 100 individuals are from the other population) and each thin vertical bar is one individual. In the top panel (a), STRUCTURE identified two populations (the correct number) (one 'red' and one 'green'), and most individuals were assessed to be of mostly "red" or mostly "green" origin, with a few individuals showing a mixed genetic heritage. This is the typical pattern expected in closely related populations, within which some individuals might be descendants of recent immigrants. Panel (b) shows results for a run where STRUCTURE identified 4 populations, but there are still only two general types of individuals: about half the total individuals appear to be roughly 50:50 mixtures of "yellow" and "green" genes, while the other half appear to be various mixtures of "red" and "blue" genes. Panel (c) shows an intermediate scenario for which $k = 3$ had the highest posterior probability: half the individuals appear to be mostly pure "red" genes, and the other half are mixtures of "green" and "blue". These scenarios where entire groups of individuals appear to be roughly equal mixtures of the same two gene pools are difficult to reconcile with the evolutionary ecology of natural populations--except in the case where one has sampled almost exclusively recent hybrids of two previously divergent populations. Such instances are relatively rare and typically could be identified by other means.

DISCUSSION

This evaluation of performance of the two Bayesian clustering methods for genetic data was rather limited, given the computation times involved and the constraint to produce results well in advance of the 2012 SC meeting in Panama. For example, we only considered n = 2 populations, 16 'microsatellite-like' gene loci, and artificial mixtures of 200 individuals in equal population proportions. Changing any or all of these key parameters could affect power to resolve mixtures of populations separated by a given level of genetic differentiation (F_{ST}).

Furthermore, we only used the admixture model for STRUCTURE and only considered the standard and admittedly *ad hoc* procedure for inferring *k* suggested by Pritchard et al. (2000). We weren't able to use a popular alternative approach (Evanno et al. 2005), because that method depends on changes in log likelihood between different values of *k* and therefore cannot provide any information about the relative likelihood for $k = 1$.

Nevertheless, we believe the results provide useful information for the Implementation Review for NPM, because:

- The levels of genetic differentiation considered (panmixia to $F_{ST} \sim 0.06$) encompasses the full range of stock differences implied by all the stock structure hypotheses under consideration;
- The number of loci used and level of genetic diversity are comparable to what is found in currently-available genetic datasets for these populations;
- Samples of S = 200 are not atypical for time-area stata for which management decisions are needed.

One major goal of this study was to extend results of a previous study (Pella and Masuda 2006), which suggested that under some circumstances at least HWLER might be able to detect stock structure when STRUCTURE could not. We didn't find that to be the case here. Neither method identified more than a single gene pool for the hardest two-stock problem (target $F_{ST} < 0.01$), Some of the variations of Stock Structure Hypothesis III, which postulate two J-like and/or two O-like stocks, could involve real genetic differences this small, so this result agrees with previous conclusions that this hypothesis will be very challenging to rigorously test without pure samples from breeding grounds. STRUCTURE, however, was able to reliably detect two gene pools at F_{ST} values in the range 0.02-0.03, which are generally considered relatively 'hard' mixtures to resolve. This result in general agrees with previous analyses of STRUCTURE (e.g., Waples and Gaggiotti 2006; Latch et al. 2006) and with theoretical work (Patterson et al. 2006) that shows that, for a given amount of information, a threshold of genetic differentiation exists, above which the problem is 'easy' but below which the mixture cannot be resolved.

One encouraging result was that neither program showed a tendency to overestimate the number of gene pools when the 'mixture' was actually a single population. However, STRUCTURE showed a surprising tendency to overestimate the number of gene pools when true $k = 2$ and differentiation was moderately strong ($F_{ST} = 0.045$-0.06). These 'wrong' estimates often varied across runs with the same data but most were associated with what appeared to be highly convincing posterior probabilities. This argues for caution in interpretation of results. Multiple runs with different starting conditions and examination of *Q* plots (Figure 1) can help identify these situations. HWLER did not have a problem identifying the correct number of populations for target $F_{ST} = 0.045$-0.06, but the threshold below which it could not reliably resolve mixtures ($F_{ST} \sim 0.03$) was higher than for STRUCTURE. It would be useful to conduct additional analyses to try to determine why STRUCTURE produced these erratic results for levels of genetic differentiation that were stronger than they were for mixtures the program could easily resolve.

It should be noted that, as migration rate increases and equilibrium F_{ST} approaches 0, it becomes problematical to identify a specific threshold that distinguishes one "population" from two or more. This is because no single quantitative definition of a biological population exists (Waples and Gaggiotti 2006). Instead, population differentiation occurs along a continuum, and the user must specify an operational definition that fits his/her particular objective. Thus, whether a particular estimated number of populations is "right" or "wrong" depends at least in part on the threshold level of differentiation that is of concern to the user.

ACKNOWLEDGMENTS

This work was supported in part by a grant of HPC resources from the Arctic Region Supercomputing Center and the University of Alaska, Fairbanks.

REFERENCES

Balloux, F. 2001. EASYPOP (version 1.7): a computer program for population genetics simulations. Journal of Heredity 92:301-302.

Dahl, D.B. 2003. An improved merge-split sampler for conjugate Dirichlet process mixture models. Department of Statistics, University of Wisconsin-Madison, Madison, Wisc. Tech. Rep. #1086.

Goudet J (1995) FSTAT version 1.2: a computer program to calculate *F*-statistics. Journal of Heredity 86:485–486.

Jain, S., and Neal, R.M. 2004. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of Computational and Graphical Statistics 13:158-182.

Kanda, N., Park, J.-Y., Goto, M., An, Y.-R., Choi, S.-G., Moon, D.-Y., Kishiro, T., Yoshida, H., Kato, H. and Pastene, L.A. 2010. Genetic analyses of western North Pacific minke whales from Korea and Japan based on microsatellite DNA. Paper SC/62/NPM11 presented to the IWC Scientific Committee, June 2010, Agadir, Morocco (unpublished). 13pp.

Latch EK, G Dharmarajan, JC Glaubitz, and OE Rhodes. 2006. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics 7: 295-302

Pastene, L.A., M. Goto, and N. Kanda. 2010. Progress in the development of stock structure hypotheses for western North Pacific common minke whales. Paper SC/62/NPM12 presented to the IWC Scientific Committee, June 2010, Agadir, Morocco (unpublished). 9 pp.

Patterson, N., A.L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLOS Genetics 2(12):e90.

Pella, J., and Masuda, M. 2006. The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baseline. Canadian Journal of Fisheries and Aquatic Sciences 63:576-596.

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Wade, P.R., R.L. Brownell, and T. Kasuya. 2010. A review of the biology of western North Pacific minke whales relevant to stock structure. Paper SC/62/NPM13 presented to the IWC Scientific Committee, June 2010, Agadir, Morocco (unpublished). 16 pp.

Waples, R. S., and O. Gaggiotti. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology 15:1419-1439.

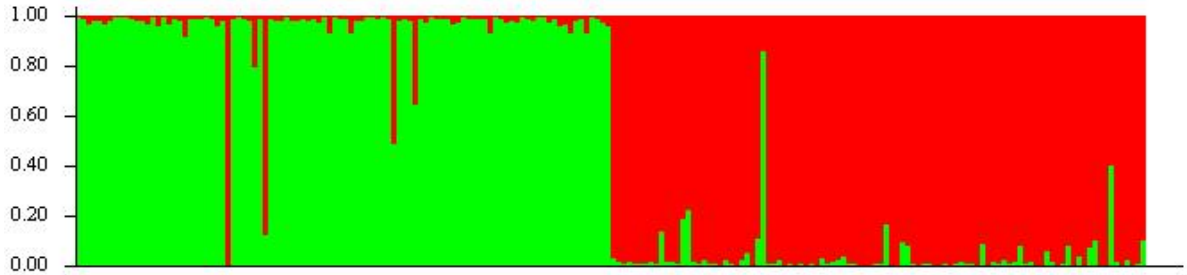
Weir BS, Cockerham CC. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–70.

Table 1. Number of estimated number of populations (k) falling in each class. For each target F_{ST}, 10 replicate datsets were generated, and for STRUCTURE each replicate was analyzed using 3 different starting points. H is average observed heterozygosity. Except as noted in footnotes, all scenarios were at mutation-migration-drift equilibrium with true $k = 2$. Simulated data included 16 'microsatellite-like' loci and each mixture included 100 individuals from each of the two modeled populations. Numbers in bold are estimates that agree with the true value of k.

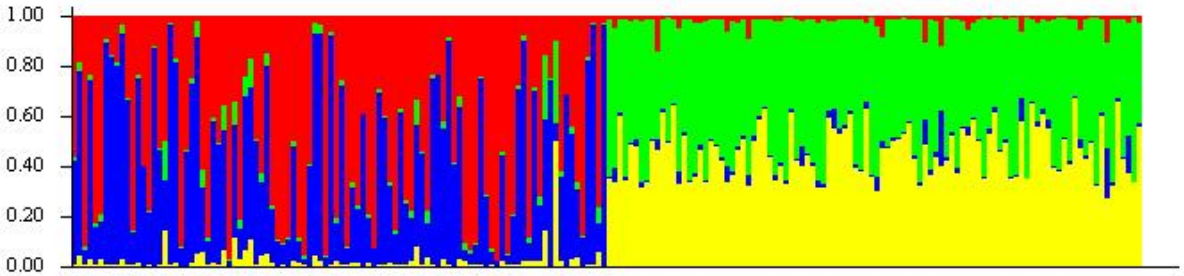
		Most probable k				
		STRUCTURE			HWLER	
Target F_{ST} (actual range)	H	1	2	other $\hat{k}$ (n)	1	2
0.06 (0.054-0.064)	0.73	-	**17**	3(5); 4(8)	-	**10**
0.045 (0.039-0.049)	0.71	-	**20**	3(5); 4(5)	-	**10**
0.03 (0.025-0.035)	0.72	-	**30**		1	**9**
0.02 (0.017-0.020)	0.76	-	**30**		10	-
0.007 (0.006-0.008)	0.76	30	-		10	-
0.007 (0.004-0.012)[a]	0.76	30	-		10	-
0 ($\lvert F_{ST}\rvert \leq 0.001$)[b]	0.77	**30**	-		**10**	-

[a] Target F_{ST} achieved after 16 generations of complete isolation
[b] Panmixia, so true $k = 1$.

a) Run 1 of replicate 1 (true $F_{ST} = 0.054$; $\hat{k} = 2$).



b) Run 1 of replicate 2 (true $F_{ST} = 0.061$; $\hat{k} = 4$)



c) Run 1 of replicate 3 (true $F_{ST} = 0.062$; $\hat{k} = 3$).

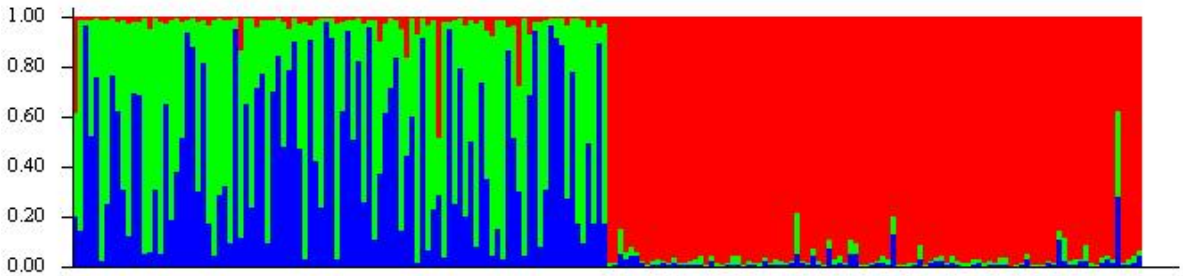


Figure 1. *Q* plots for 3 runs of STRUCTURE with target $F_{ST} = 0.06$. Each thin vertical bar represents one individual, and the colors indicate the estimated fractions of that individual's genes that are derived from each population. Individuals in the plots are ordered by population, so the first 100 individuals are from one population and the second 100 individuals are from the other population. Each panel shows the fit to the $\hat{k}$ value with the highest posterior probability for that run. Note that the colors are arbitrary and it is not meaningful to compare similar colors in different runs.

Appendix Table 1. Posterior probability for different numbers of populations ($\hat{k}$ = 2, 3, or 4) for STRUCTURE analyses for target F_{ST} = 0.045 and 0.06. For each replicate dataset, results are shown for each of 3 separate runs using different starting points. In all cases, posterior probabilities for $\hat{k}$ = 1 and $\hat{k}$ = 5 were << 0.001.

	Target F_{ST} = 0.045					Target F_{ST} = 0.06				
			k					*k*		
Replicate	Actual F_{ST}	Run	**2**	**3**	**4**	Actual F_{ST}	Run	**2**	**3**	**4**
1	0.047	1	**1.000**	0.000	0.000	0.054	1	**1.000**	0.000	0.000
		2	0.354	**0.646**	0.000		2	**1.000**	0.000	0.000
		3	**0.997**	0.003	0.000		3	**1.000**	0.000	0.000
2	0.039	1	0.000	0.000	**1.000**	0.061	1	0.000	0.000	**1.000**
		2	0.332	0.000	**0.668**		2	0.000	0.000	**1.000**
		3	**1.000**	0.000	0.000		3	**1.000**	0.000	0.000
3	0.048	1	0.000	0.000	**1.000**	0.062	1	0.100	**0.900**	0.000
		2	0.000	0.000	**1.000**		2	**0.900**	0.100	0.000
		3	0.000	**1.000**	0.000		3	**1.000**	0.000	0.000
4	0.049	1	0.000	**1.000**	0.000	0.060	1	0.000	**1.000**	0.000
		2	0.000	**1.000**	0.000		2	**1.000**	0.000	0.000
		3	0.000	**1.000**	0.000		3	0.000	0.000	**1.000**
5	0.044	1	**1.000**	0.000	0.000	0.058	1	**1.000**	0.000	0.000
		2	**1.000**	0.000	0.000		2	**1.000**	0.000	0.000
		3	**1.000**	0.000	0.000		3	**1.000**	0.000	0.000
6	0.044	1	**1.000**	0.000	0.000	0.061	1	**1.000**	0.000	0.000
		2	**1.000**	0.000	0.000		2	**1.000**	0.000	0.000
		3	**1.000**	0.000	0.000		3	**1.000**	0.000	0.000
7	0.046	1	**1.000**	0.000	0.000	0.062	1	**1.000**	0.000	0.000
		2	**1.000**	0.000	0.000		2	0.000	**1.000**	0.000
		3	0.002	0.000	**0.998**		3	**1.000**	0.000	0.000
8	0.041	1	**1.000**	0.000	0.000	0.063	1	**0.769**	0.231	0.000
		2	**1.000**	0.000	0.000		2	0.000	0.000	**1.000**
		3	**1.000**	0.000	0.000		3	0.000	0.000	**1.000**
9	0.048	1	**1.000**	0.000	0.000	0.064	1	0.000	0.000	**1.000**
		2	**1.000**	0.000	0.000		2	**1.000**	0.000	0.000
		3	**1.000**	0.000	0.000		3	0.000	0.000	**1.000**
10	0.044	1	**1.000**	0.000	0.000	0.059	1	0.000	**1.000**	0.000
		2	**1.000**	0.000	0.000		2	0.000	**1.000**	0.000
		3	**1.000**	0.000	0.000		3	0.000	0.000	**1.000**