Effects of population structure on match probabilities.

Bruce Weir

bsweir@uw.edu

with John Buckleton, James Curran,
Jérôme Goudet and Alexandre Thiery

Match Probabilities

A genetic profile G has been found to match between a person of interest (POI) and an item of evidence (CS). What is the probability that someone other than the person of interest would also have profile G? This is the *match probability* $\Pr(G|G)$.

The first approach is to assume Hardy-Weinberg equilibrium. If the matching profile has genotype ab, then the profile probability is $2p_a p_b$. If the POI and the contributor to CS have independent profiles, then this is also the match probability. If we have sample allele frequencies $\tilde{p}_a, \tilde{p}_b$ then the match probability estimate is $2\tilde{p}_a\tilde{p}_b$.

Population structure destroys HWE

The issue about population structure is that we generally don't have data from the relevant population, but instead we have data from some larger group such as African American, Asian, Caucasian or Hispanic. What is the problem?

If a population consists of a number of subpopulations, each in HWE but with different allele frequencies, there will be a departure from HWE at the population level. Suppose there are two equal-sized subpopulations, each in HWE but with different allele frequencies, then

	Subpopn 1	Subpopn 2	Total Population
p_a	0.6	0.4	0.5
p_b	0.4	0.6	0.5
P_{aa}	0.36	0.16	$0.26 > (0.5)^2$
P_{ab}	0.48	0.48	$0.48 < 2(0.5)(0.5)$
P_{bb}	0.16	0.36	$0.26 > (0.5)^2$

Population structure destroys HWE

If we use whole-population sample allele frequencies and assume Hardy-Weinberg Equilibrium, then we won't be giving valid estimates for the profile probability or the match probability in either the whole population or in any subpopulation.

The remedy is to use the "theta correction" that lies behind the Balding-Nichols expressions that were cited by the National Research Council report in 1996:

$$\Pr(aa|aa) = \frac{[2\theta + (1-\theta)p_a][3\theta + (1-\theta)p_a]}{(1+\theta)(2+2\theta)}$$

$$\Pr(ab|ab) = \frac{2[\theta + (1-\theta)p_a][\theta + (1-\theta)p_b]}{(1+\theta)(2+2\theta)}$$

These reduce to the HWE results when $\theta = 0$. So what is θ?

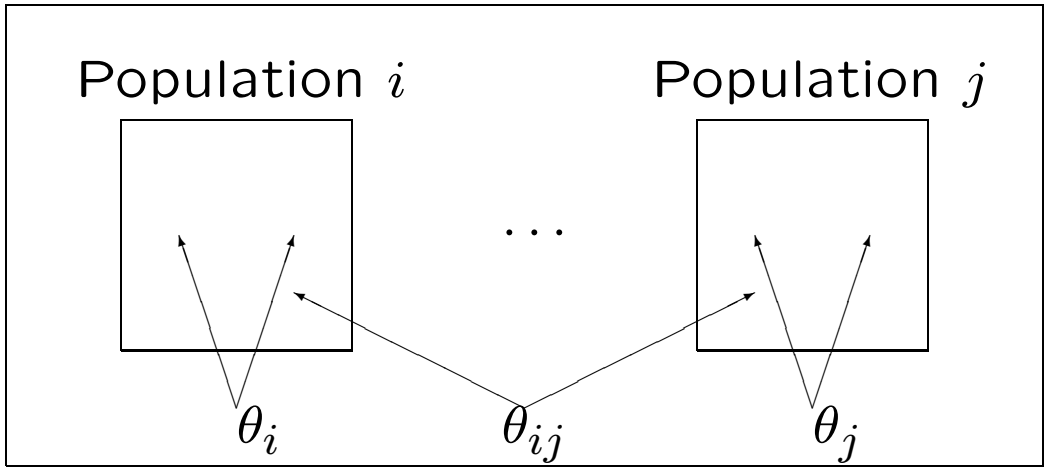
What is θ?

Two ways of thinking about θ.

It measures the extra degree of relatedness of individuals because they belong to the same population. We think of this reflecting long-term evolutionary history, but can use the same logic for people related because they are in the same family: first cousins have a θ value of 0.0625.

θ also helps to measure the variance of allele frequencies over populations.

What is θ?

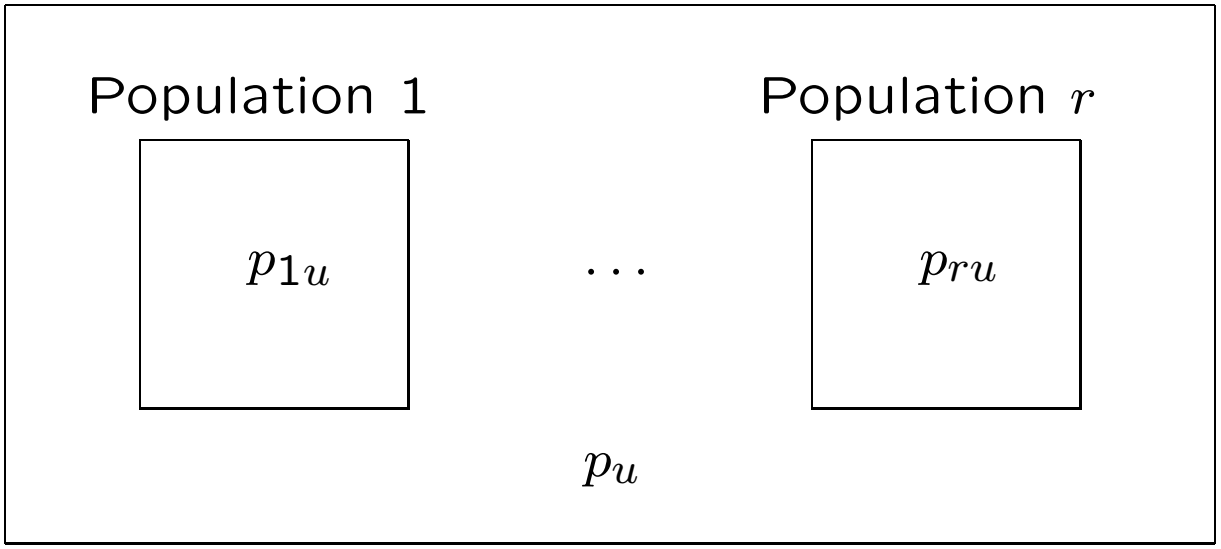


θ's are statements about pairs of alleles: the probabilities the pairs are identical by descent.

θ_W is the average of the within-population coancestries θ_i.

θ_B is the average of the between-population-pair coancestries θ_{ij}.

Allele u Frequencies



Allele frequencies p_{iu} for type u differ among populations i. The average over all populations is p_u. The variances and covariances of sample allele frequencies depend on the θs.

Sample Heterozygosities

Population genetic literature makes use of "heterozygosities"

$$\begin{aligned}\tilde{H}_i &= \frac{n_i}{n_i - 1}\sum_u \tilde{p}_{iu}(1 - \tilde{p}_{iu}) = \frac{n_i}{n_i - 1}(1 - \sum_u \tilde{p}_{iu}^2) \\ &\approx 1 - \sum_u \tilde{p}_{iu}^2\end{aligned}$$

$$\tilde{H}_{ij} = 1 - \sum_u \tilde{p}_{iu}\tilde{p}_{ju}$$

Unweighted averages of these quantities over populations are

$$\begin{aligned}\tilde{H}_W &= \frac{1}{r}\sum_{i=1}^{r} \tilde{H}_i \\ \tilde{H}_B &= \frac{1}{r(r-1)}\sum_{i \neq j} \tilde{H}_{ij}\end{aligned}$$

Sample Match Proportions

If a sample of n_i alleles from population i has n_{iu} copies of allele u, then the proportion of pairs of alleles that match within this population is

$$\tilde{M}_i = \frac{1}{n_i(n_i-1)} \sum_u n_{iu}(n_{iu}-1) = 1 - \tilde{H}_i$$

Similarly, the proportion of pairs of alleles, one from each of populations i and j, that match is

$$\tilde{M}_{ij} = \frac{1}{n_i n_{ij}} \sum_u n_{iu} n_{ju} = 1 - \tilde{H}_{ij}$$

Sample match proportions and heterozygosities may therefore be used interchangeably: $\tilde{M} = 1 - \tilde{H}$.

Population Genetic Model

Under our model, we can predict the values of sample match proportions:

$$\begin{aligned}
\mathcal{E}(\tilde{M}_i) &= \theta_i + (1-\theta_i)M_T \\
\mathcal{E}(\tilde{M}_{ij}) &= \theta_{ij} + (1-\theta_{ij})M_T
\end{aligned}$$

$$\begin{aligned}
\mathcal{E}(\tilde{M}_W) &= \theta_W + (1-\theta_W)M_T \\
\mathcal{E}(\tilde{M}_B) &= \theta_B + (1-\theta_B)M_T
\end{aligned}$$

where $M_T = \sum_u p_u^2$.

These suggest that we can manipulate sample matching proportions to make statements about θ_W and θ_B.

Estimators of β's

We can estimate θ_i or θ_W only relative to θ_B. Matching within a population has meaning only when compared to matching between populations. The estimates do not require us to know the true allele frequencies.

$$\tilde{\beta}_i = \frac{\tilde{M}_i - \tilde{M}_B}{1 - \tilde{M}_B} \quad , \quad \mathcal{E}(\tilde{\beta}_i) = \beta_i = \frac{\theta_i - \theta_B}{1 - \theta_B}$$

$$\tilde{\beta}_W = \frac{\tilde{M}_W - \tilde{M}_B}{1 - \tilde{M}_B} \quad , \quad \mathcal{E}(\tilde{\beta}_W) = \beta_W = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

The estimates are "relative to" θ_B, but they allow comparisons among populations. Estimation of θ_B is not possible: θ_B can also refer to the ancestral/reference population.

Predicted Values of the θ's: Pure Drift

This treatment is based on the covariance structure of sample allele/profile frequencies, rather than a specific evolutionary model.

However, in the case of pure drift, where population i has constant size N_i and there is random mating, t geneations after the population becomes isolated

$$\beta_W(t) \;=\; 1 - \frac{1}{r}\sum_i \left(1 - \frac{1}{2N_i}\right)^t \approx \frac{t}{2N_h}$$

Worldwide Autosomal-STR Survey

Buckleton and Curran have compiled a survey of 250 published papers showing allele frequencies at 24 forensic STR markers from 446 populations in 8 ancestral groups. Represents data from 494,473 individuals.

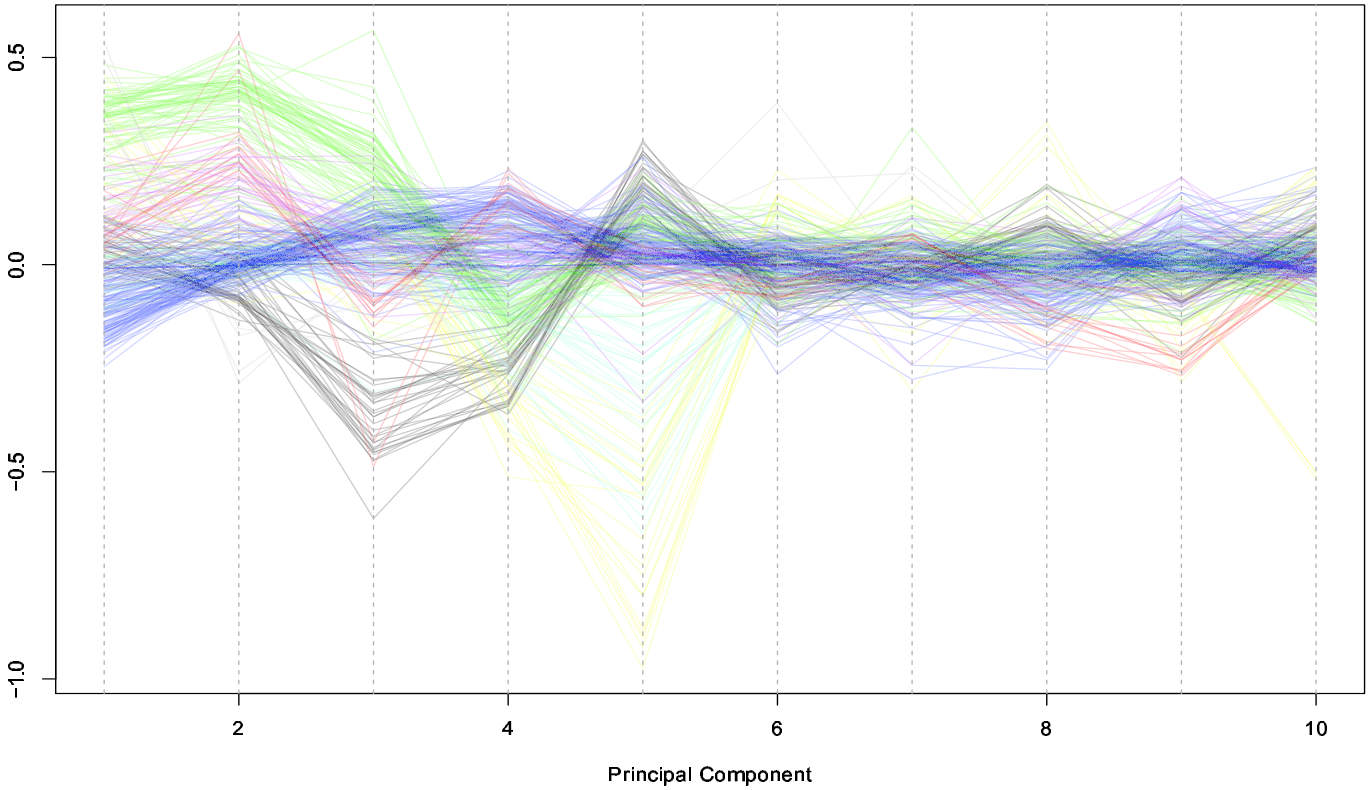
The ancestral groups were identified by a combination of clustering and geographic criteria.

Moment estimates were obtained for each locus l in each population i from

$$\hat{\beta}_{il} = \frac{\tilde{M}_{il} - \tilde{M}_{Bl}}{1 - \tilde{M}_{Bl}}$$

STR Survey: Principal Components Analysis

(Black: Australian Aborigine; Blue: European; Green: Asian; ..)



STR Survey: $\widehat{\beta}_W$ Values for Groups and Loci

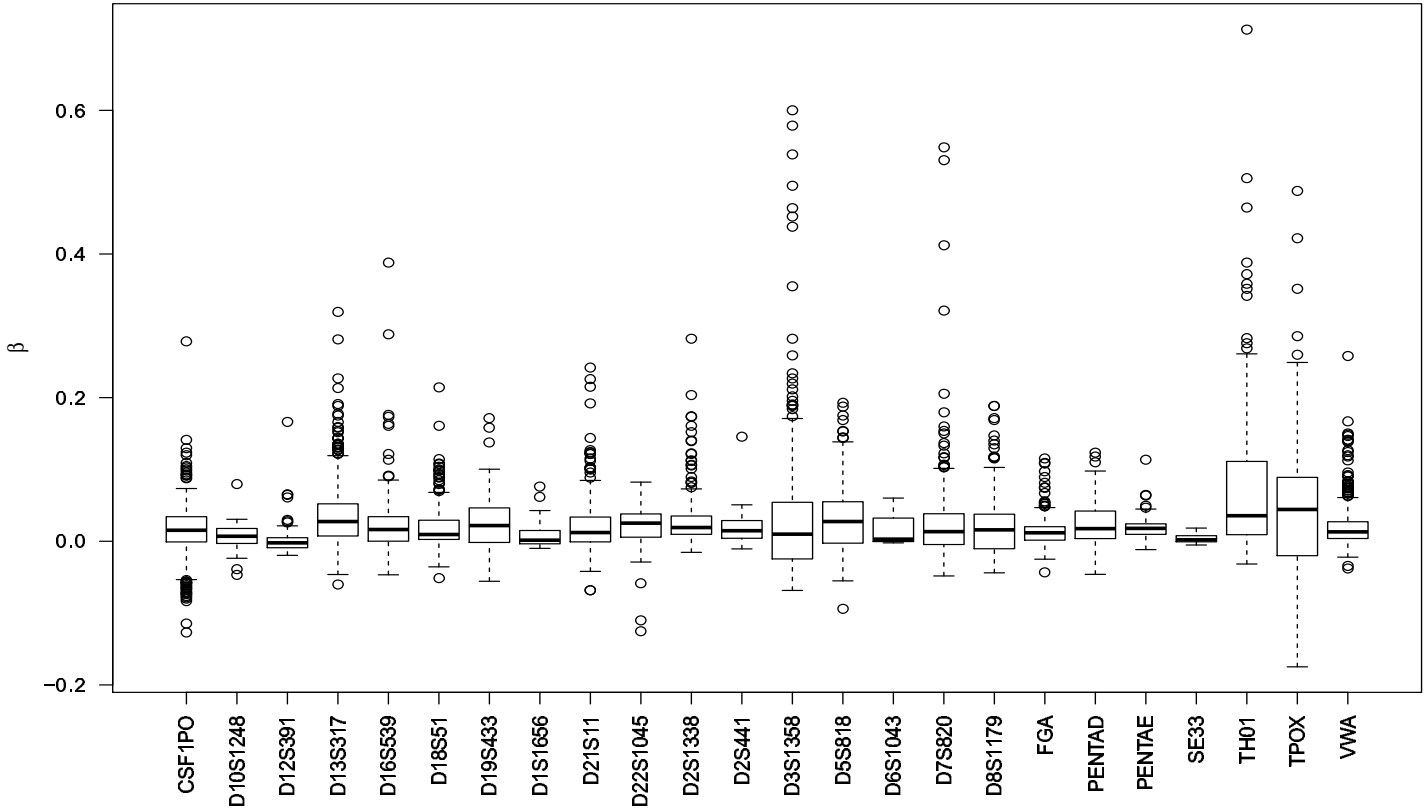
Locus	Afr	Aus	Asian	Cauc	Hisp	IndPK	NatAm	Poly	Aver.
CSF1PO	0.003	0.002	0.008	0.008	0.002	0.007	0.055	0.026	0.011
D1S1656	0.000	0.000	0.000	0.002	0.003	0.000	0.000	0.000	0.011
D2S441	0.000	0.000	0.002	0.003	0.021	0.000	0.000	0.000	0.020
D2S1338	0.009	0.004	0.011	0.017	0.013	0.003	0.023	0.005	0.031
D3S1358	0.004	0.010	0.009	0.006	0.012	0.040	0.079	0.001	0.025
D5S818	0.002	0.013	0.009	0.008	0.014	0.018	0.044	0.007	0.029
D6S1043	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016
D7S820	0.004	0.021	0.010	0.007	0.007	0.046	0.030	0.005	0.026
D8S1179	0.003	0.007	0.012	0.006	0.002	0.031	0.020	0.008	0.019
D10S1248	0.000	0.000	0.000	0.002	0.004	0.000	0.000	0.000	0.007
D12S391	0.000	0.000	0.000	0.003	0.020	0.000	0.000	0.000	0.010
D13S317	0.015	0.016	0.013	0.008	0.014	0.025	0.050	0.014	0.038
D16S539	0.007	0.002	0.015	0.006	0.009	0.005	0.048	0.004	0.021
D18S51	0.011	0.012	0.014	0.006	0.004	0.010	0.033	0.003	0.018
D19S433	0.009	0.001	0.009	0.010	0.014	0.000	0.022	0.014	0.023
D21S11	0.014	0.012	0.013	0.007	0.006	0.023	0.067	0.018	0.021
D22S1045	0.000	0.000	0.007	0.001	0.000	0.000	0.000	0.000	0.015
FGA	0.002	0.009	0.012	0.004	0.007	0.016	0.021	0.006	0.013
PENTAD	0.008	0.000	0.012	0.012	0.002	0.017	0.000	0.000	0.022
PENTAE	0.002	0.000	0.017	0.006	0.003	0.012	0.000	0.000	0.020
SE33	0.000	0.000	0.012	0.001	0.000	0.000	0.000	0.000	0.004
TH01	0.022	0.001	0.022	0.016	0.018	0.014	0.071	0.017	0.071
TPOX	0.019	0.087	0.016	0.011	0.007	0.018	0.064	0.031	0.035
VWA	0.009	0.007	0.017	0.007	0.012	0.022	0.028	0.005	0.023
All Loci	0.006	0.014	0.010	0.007	0.008	0.018	0.043	0.011	0.022
CODIS	0.009	0.015	0.013	0.007	0.009	0.021	0.046	0.011	0.027

Expected Variation in θ

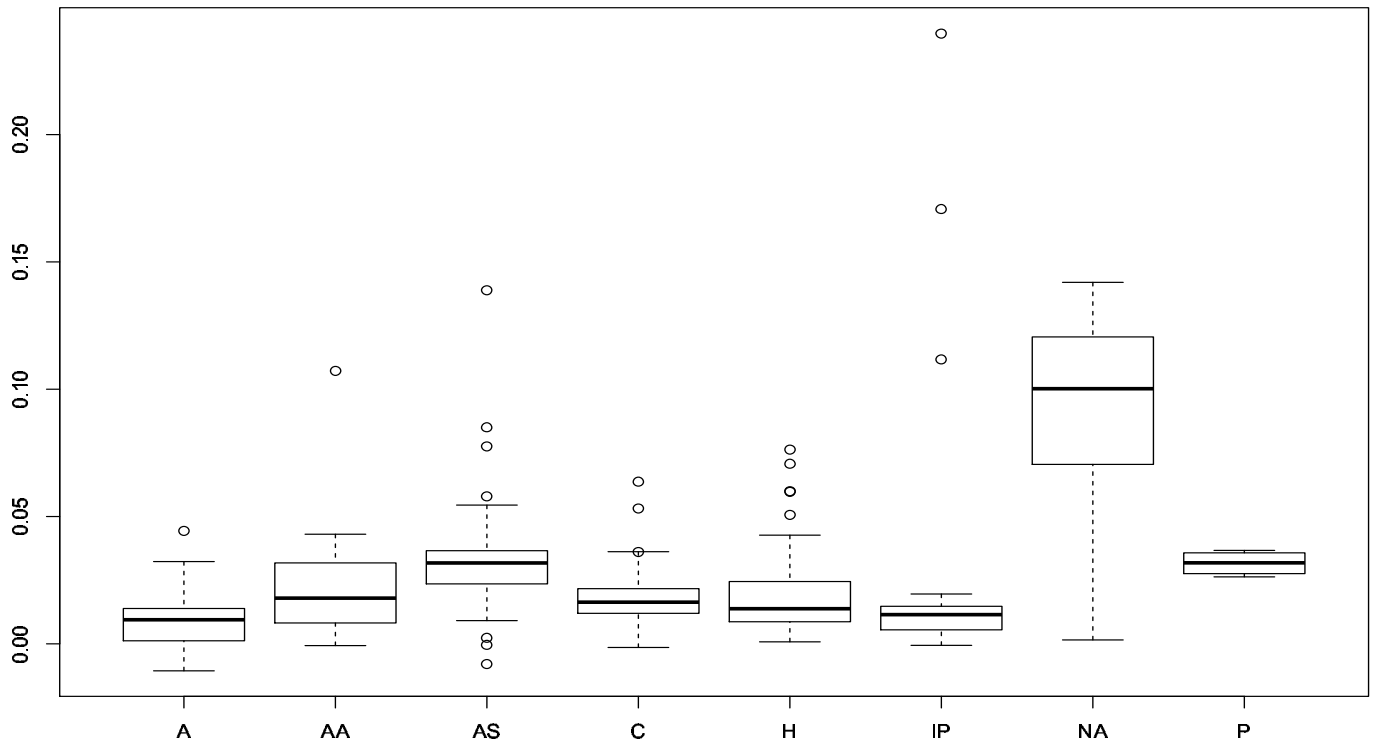
At drift/mutation equilibrium, the values of θ and its coefficient of variation are:

		$\mu = 10^{-3}$	$\mu = 10^{-4}$	$\mu = 10^{-5}$
$N = 10^3$	θ	0.2000	0.7143	0.9615
	CV	0.4364	0.3131	0.1136
$N = 10^4$	θ	0.0244	0.2000	0.7143
	CV	0.2104	0.4364	0.3131
$N = 10^5$	θ	0.0025	0.0244	0.2000
	CV	0.0702	0.2105	0.4364

Variation Among Populations for Each Locus



Variation Among Loci for Each Group



STR Mutation

Population genetic theory predicts the value of θ for STRs in a population of specified size and haplotypes of specified mutation rates.

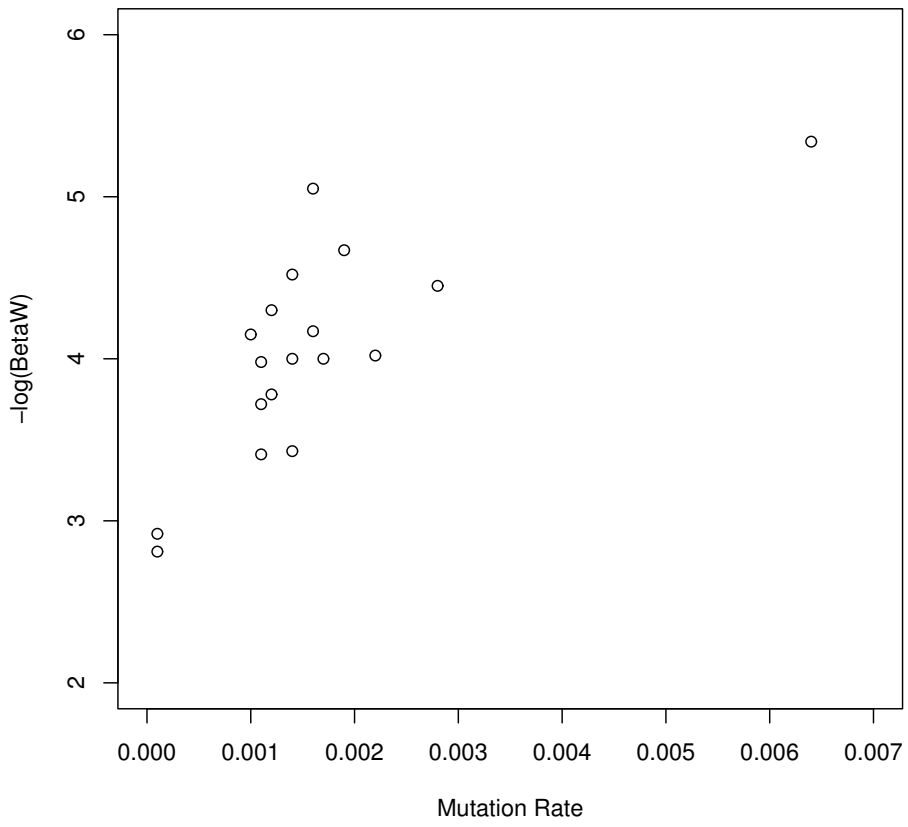
For autosomes:

$$\theta = \frac{1}{\sqrt{1 + 8N\mu}}$$

On a log-scale

$$\begin{aligned} -\ln(\theta) &= \ln(1 + 8N\mu) \approx 8N\mu \\ &\propto \mu \end{aligned}$$

STR Mutation

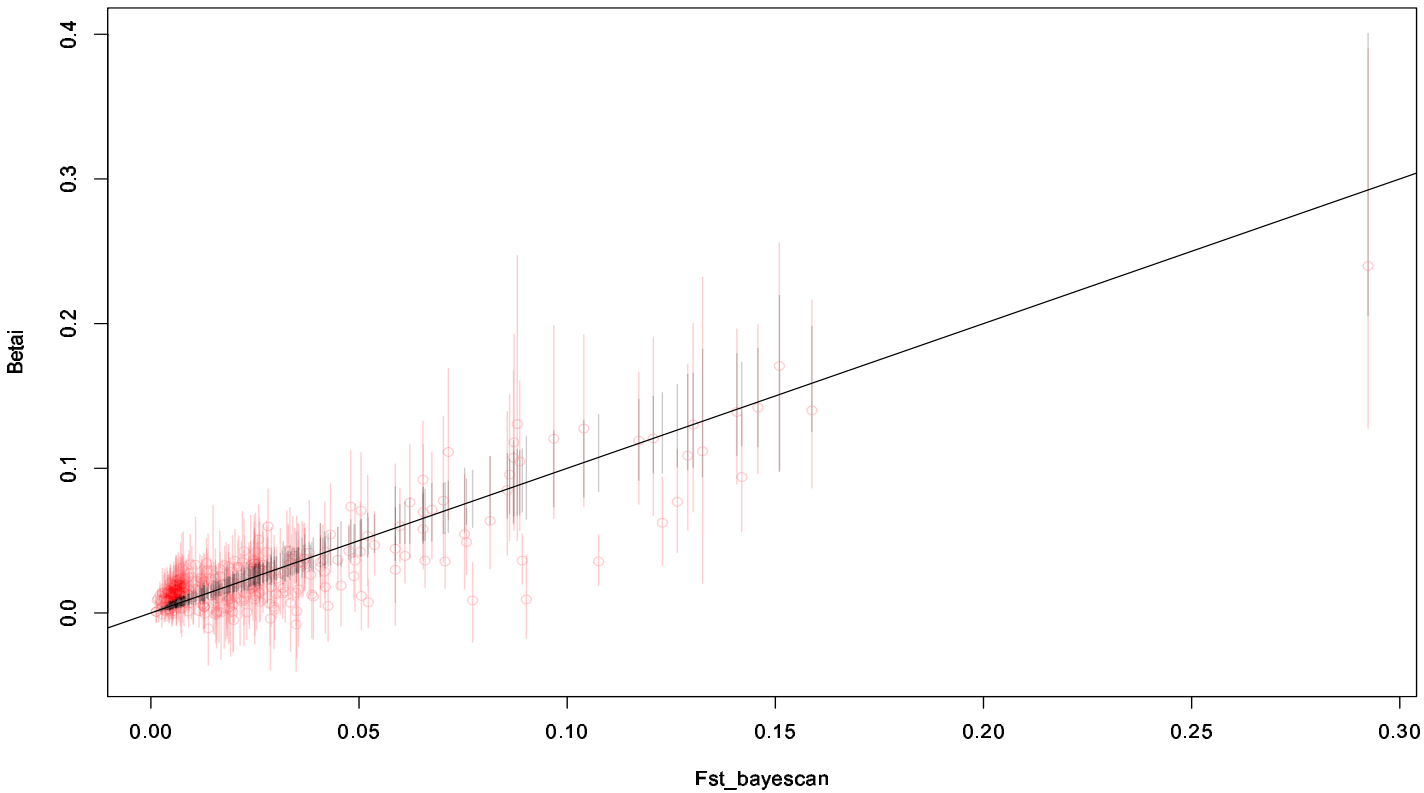


Bayesian Estimation

Moment estimators use only the means and variances/covariances of allele frequencies over populations. They can give estimates with large variances so that estimates can be negative.

An alternative approach is to assume the form of the distribution of allele frequencies over populations: the most appropriate is the Dirichlet distribution. The BayesScan software builds Bayesian estimates on this assumption.

STR Survey: Moment vs Bayes Estimates Over Loci for Populations

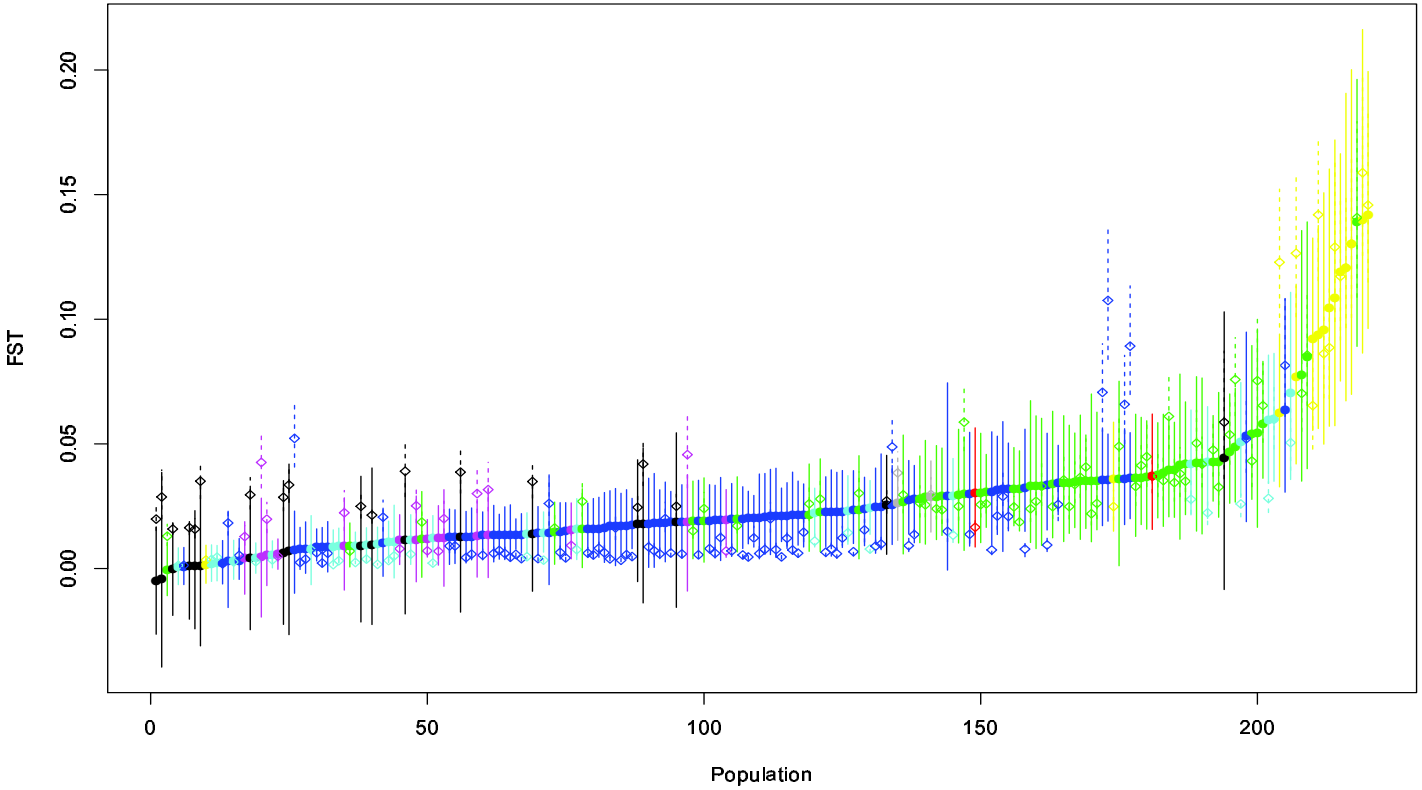


Andaman Islanders

The outlier value of $\widehat{\beta}_i$ is for a sample from the Andaman Islands, a district of India in the southeastern part of the Bay of Bengal.

Andamanese people are pygmies and have been isolated: current population size is 400-450 (Wikipedia).

STR Survey: Estimates Over Loci for Populations



Match Probabilities

If a suspected and actual offender (when different) are both in the same subpopulation i then the probability they both have allele u is $\breve{p}_{iu}^2$ if alleles are assumed to be independent.

If the relevant allele frequencies $\breve{p}_{iu}$ are not known, then the joint probability can be expressed as

$$P_{uu_i} = \theta_i p_u + (1 - \theta_i) p_u^2$$

where p_u is the total allele frequency. This expression is an average over evolutionary replicates of population $i$of interest, for which there is not a sample allele frequency. There is, however, a sample frequency $\tilde{p}_u$ from a larger population (or collection of populations). The appropriate estimate of the joint probability P_{uu_i} is

$$\widehat{P}_{uu_i} = \beta_i \tilde{p}_A + (1 - \beta_i) \tilde{p}_A^2$$

The quantity β_i can be negative.

Match Probabilities

It may be appropriate to average over all subpopulations i:

$$\widehat{P}_{uu_W} \;=\; \beta_W \tilde{p}_u + (1-\beta_W)\tilde{p}_u^2$$

where β_W is the average of the β_i's. It is positive.

The joint frequency can be converted to a matching probability, assuming $\tilde{p}_u \neq 0$:

$$\widehat{P}_{u|u_W} \;=\; \beta_W + (1-\beta_W)\tilde{p}_u$$

and then averaged over u to give a within-population matching estimate:

$$\widehat{M}_W = \sum_u \tilde{p}_u \widehat{P}_{u|u_W} \;=\; \beta_W + (1-\beta_W)\sum_u \tilde{p}_u^2$$

Y-chromosome

The same general approach can be applied to Y-STR data except that alleles u now refer to haplotypes.

NIST Y-STR Single-locus Estimates

Locus	$\tilde{M}_W$	$\tilde{M}_B$	$\hat{\beta}_W$
DYS19	0.32571062	0.24309148	0.10915340
DYS385a/b	0.07982377	0.04427420	0.03719640
DYS389I	0.41279418	0.38319082	0.04799436
DYS389II	0.26072434	0.23741323	0.03056847
DYS390	0.28981997	0.18813203	0.12525182
DYS391	0.52191425	0.48517426	0.07136392
DYS392	0.39961865	0.35168087	0.07394164
DYS393	0.50285122	0.48769253	0.02958906
DYS437	0.46400112	0.38595032	0.12710828
DYS438	0.36817530	0.23212655	0.17717601
DYS439	0.35507469	0.34990863	0.00794667
DYS448	0.30091326	0.22640195	0.09631787
DYS456	0.33444029	0.32578009	0.01284478
DYS458	0.21642167	0.19701369	0.02416976
DYS481	0.18867019	0.14121936	0.05525373
DYS533	0.39365769	0.37177174	0.03483757
DYS549	0.33976578	0.30691346	0.04740003
DYS570	0.21298105	0.20775666	0.00659442
DYS576	0.20955290	0.18125443	0.03456321
DYS635	0.27720127	0.20653182	0.08906400
DYS643	0.28394262	0.20058158	0.10427710
Y-GATA-H4	0.40667782	0.39899963	0.01277568

NIST Y-STR Haplotype Estimates

No. Loci	Added Locus	$\tilde{M}_W$	$\tilde{M}_B$	$\hat{\beta}_W$
1	DYS_438	0.37903281	0.27283973	0.14603806
2	DYS_392	0.22353526	0.10233258	0.13501958
3	DYS_19	0.11294942	0.05471374	0.06160639
4	DYS_390	0.05923470	0.02393636	0.03616398
5	DYS_643	0.04798422	0.02456341	0.02401059
6	YGATA_C4	0.03119210	0.01541060	0.01602851
7	DYS_533	0.01979150	0.00777794	0.01210774
8	DYS_393	0.01482393	0.00650531	0.00837309
9	DYS_456	0.01073170	0.00396487	0.00679377
10	DYS_438	0.00889934	0.00287761	0.00603912
11	DYS_549	0.00524369	0.00123093	0.00401770
12	DYS_481	0.00317518	0.00055413	0.00262250
13	DYS_389I	0.00240161	0.00031517	0.00208710
14	DYS_391	0.00200127	0.00017039	0.00183119
15	DYS_576	0.00106995	0.00005877	0.00101124
16	DYS_ 389II	0.00089896	0.00004205	0.00085695
17	DYS_385	0.00065020	0.00002729	0.00062293
18	YGATA_H4	0.00063652	0.00002427	0.00061227
19	DYS_448	0.00055062	0.00000713	0.00054349
20	DYS_458	0.00051100	0.00000423	0.00050677
21	DYS_570	0.00043010	0.00000423	0.00042587
22	DYS_439	0.00038612	0.00000423	0.00038189

Predicted PP23 Matches