Pacific coast SNP-barcode analysis

Aimee Taylor

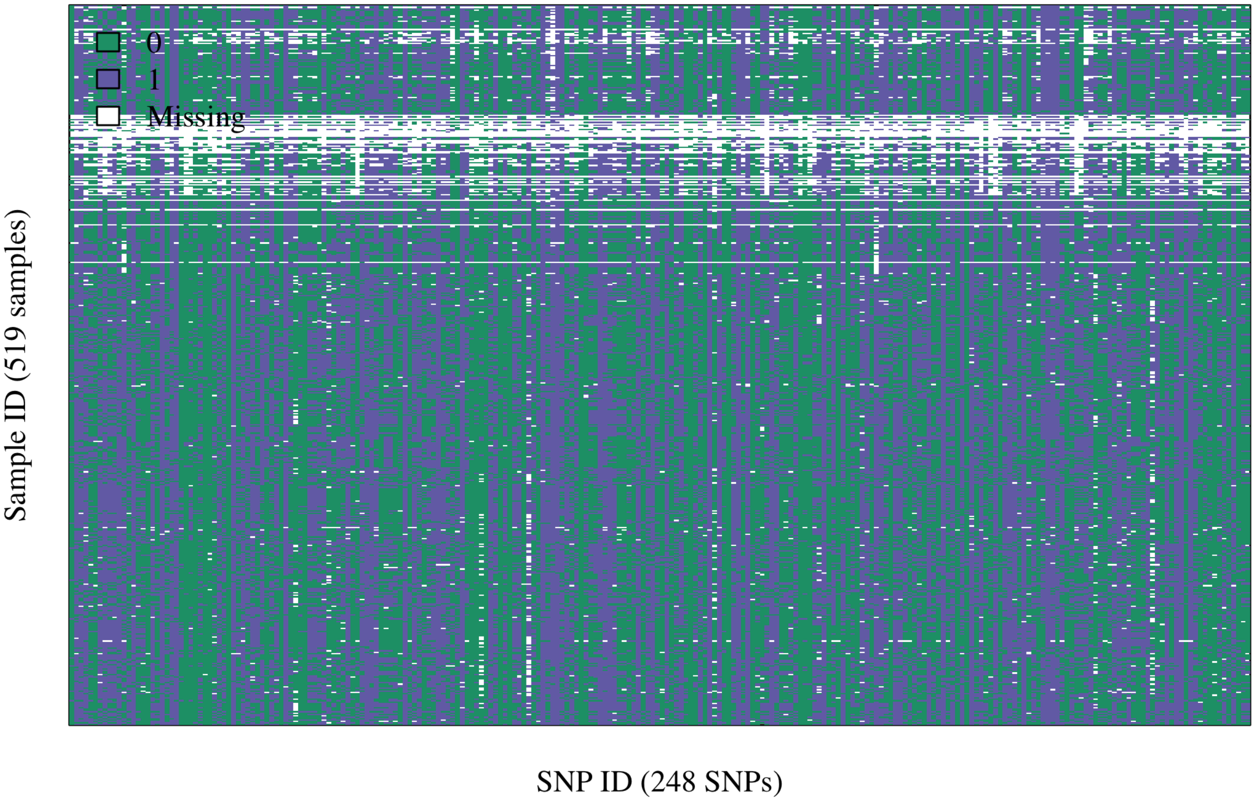

Figure 1: Plot of SNP-barcode data, excluding samples with no SNP-barcode data.

Table 1: Yearly sample counts per city of the samples with no SNP-barcode data.

	2014	2015	Total
Orellana	1	0	1
SanLorenzo	4	1	5
TobarDonoso	0	1	1
Total	5	2	7

Table 2: Yearly sample counts per city of the samples with some SNP-barcode data.

	1993	1994	1997	1999	2000	2001	2002	2003	2004	2005	2006	2007	2013	2014	2015	2016	2017	Total
Buenaventura	4	1	0	5	0	0	0	0	12	15	10	0	0	0	0	0	0	47
Esmeraldas	0	0	0	0	0	0	0	0	0	0	0	0	26	2	0	0	0	28
Guapi	0	0	0	1	1	0	0	66	0	0	0	0	0	5	35	80	1	189
Orellana	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Quibdo	0	0	2	0	6	1	0	0	14	6	13	22	0	0	0	0	0	64
SanLorenzo	0	0	0	0	0	0	0	0	0	0	0	0	5	11	5	0	0	21
StaBarbara_Iscuande	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	5	1	9
Sucumbios	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0	4
Tado	0	0	0	0	0	12	2	0	0	0	0	0	0	0	0	0	0	14
Timbiqui	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	4
TobarDonoso	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	4
Tumaco	0	0	0	2	2	10	11	59	0	23	0	25	0	0	0	0	0	132
Venezuela	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2
Total	4	1	2	8	9	23	13	125	26	44	23	47	31	23	49	89	2	519

Methods

This summary describes analyses done for the extended data after have done the marker reordered analysis of the samples that feature in Taylor et al. 2020 (see README.md). The 325 samples that feature in Taylor et al. 2020 are those deemed monoclonal in Echeverry et al. 2013. For brevity, we refer to these as the samples that feature in Taylor et al. 2020.

Pre-processing of data

The raw SNP-barcode data set (Diego-Vladimir-Fabian_GG3D7_Recode_V2_07March2021.txt formatted to snpdata.RData in Format_snpdata_extended.R) of 526 samples and 250 markers, includes 2 markers with no data (one on chromosome 4 was deleted across all samples because it was missing among samples that did not feature in Taylor et al. 2020; another on chromosome 8 was deleted across all samples because it was deemed unreliable due to some inconsistencies across WGS and re-coded Golden gate data on some samples that feature in Taylor et al. 2020), and 7 samples with no data. The year and place of sampling of these 7 samples are summarised in Table 1. Their sample ids are Pf054, Pf057, Pf058, Pf063, Pf064, Pf070, Pf073.

Removal of the 7 samples and 2 markers with no data, leaves 519 samples with data on 1 to 248 SNPs (Figure 1). The year and place of sampling of the 519 samples with some SNP-barcode data are summarised in Table 2. The full set of metadata are provided in metadata.RData, which was formatted in Format_metadata_extended.R.

Generation of relatedness estimates and confidence intervals

Relatedness inference between the 519 samples with some SNP-barcode data was attempted for all pairs that shared at least one SNP of data (Generate_mles_CIs_extended.R). Using this as-tolerant-as-possible specification, 134263 relatedness estimates were computed based on data on 1 to 248 SNPs shared between pairs among 519 samples.

Note that the confidence intervals presented in Taylor et al. 2020 (and in Taylor et al. Genetics 2019) were computed assuming data were available on all 250 SNPs, i.e. unlike the real data that had some missing SNPs, the data simulated during the parametric bootstrap did not have any missing SNPs.

This oversight was corrected in the marker reordered analysis of Taylor et al. 2020, where it likely has negligible impact because among the samples that feature in Taylor et al. 2020 there are few missing SNPs; it has also been corrected in the current analysis, where it would have had a large impact due to many missing SNPs among the samples that did not feature in Taylor et al. 2020 (Figure 2).

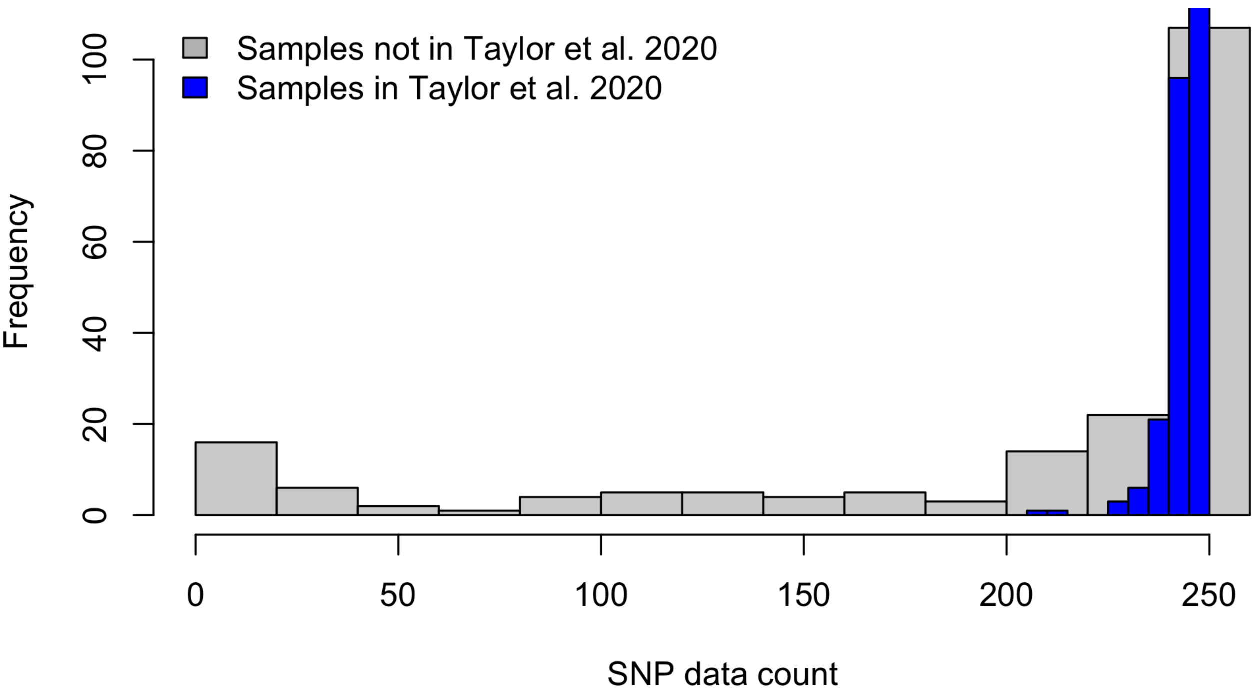

Figure 2: Histogram of SNP data counts

Post-processing of relatedness estimates

Removal of relatedness estimates where the parametric bootstrap clearly failed

Confidence intervals around relatedness estimates, which are maximum likelihood estimates (mles), were computed using the parametric bootstrap (we cannot use the standard bootstrap because we cannot sample loci with replacement and we cannot use standard asymptotic theory because of the finite length of the genome; see Taylor et al. Genetics 2019). Intuitively, for a given value of relatedness, one expects confidence intervals to be large when the data are sparse. However, when the data are sparse and the relatedness estimate is close to zero or one (especially one), the parametric bootstrap fails (explained below by analogy with a fair coin that is flipped only once; Box 1). In Filter_mles_CIs_extended.R, we identify and remove the relatedness estimates where the parametric bootstrap has clearly failed. We do this by plotting the confidence interval width per relatedness estimate against the number of SNPs with data per relatedness estimate and looking for points (relatedness estimates) that deviate from the expected trend between confidence interval width and data sparsity.

Removal of the relatedness estimates where the parametric bootstrap has clearly failed, leaves 134263 relatedness estimates based on data on 1 to 248 SNPs shared between pairs among 519 samples. Metadata are added to these estimates in Format_mles_CIs_extended.R.

Removal of uninformative estimates

Many estimates have 95% confidence intervals that extend all the way from zero to one. Removal of these relatedness estimates, which we consider to be uninformative, leaves 128657 relatedness estimates based on data on 1 to 248 SNPs shared between pairs among 512 samples. Metadata are added to these estimates in Format_mles_CIs_extended.R.

In addition, in the plots of the connectivity analyses, fractions are limited to those with numerators of at

Table 3: Yearly sample counts per city of the samples ommitted from graphs due to one or more missing relatedness estimate.

	2013	2014	2015	Total
Esmeraldas	3	1	0	4
Guapi	0	0	4	4
Orellana	0	1	0	1
SanLorenzo	1	5	2	8
Sucumbios	0	2	1	3
TobarDonoso	0	0	3	3
Total	4	9	10	23

Table 4: Yearly sample counts per city of the samples included in graphs

	1993	1994	1997	1999	2000	2001	2002	2003	2004	2005	2006	2007	2013	2014	2015	2016	2017	Total
Buenaventura	4	1	0	5	0	0	0	0	12	15	10	0	0	0	0	0	0	47
Esmeraldas	0	0	0	0	0	0	0	0	0	0	0	0	23	1	0	0	0	24
Guapi	0	0	0	1	1	0	0	66	0	0	0	0	0	5	31	80	1	185
Quibdo	0	0	2	0	6	1	0	0	14	6	13	22	0	0	0	0	0	64
SanLorenzo	0	0	0	0	0	0	0	0	0	0	0	0	4	6	3	0	0	13
StaBarbara_Iscuande	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	5	1	9
Sucumbios	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Tado	0	0	0	0	0	12	2	0	0	0	0	0	0	0	0	0	0	14
Timbiqui	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	4
TobarDonoso	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
Tumaco	0	0	0	2	2	10	11	59	0	23	0	25	0	0	0	0	0	132
Venezuela	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2
Total	4	1	2	8	9	23	13	125	26	44	23	47	27	14	39	89	2	496

least 10 observations per yearly partition and at least 10 samples per city for the city partitions.

Removal of samples with one or more missing relatedness estimates

In various scripts we generate graphs whose edges are weighted by relatedness estimates. These graphs do not support samples that are missing estimates of relatedness with one or more of the other samples (igraph::graph_from_incidence_matrix() with weighted = T returns an error if a vertex with one or more missing edges is within the incidence matrix; igraph::graph_from_adjacency_matrix imputes missing edges if weighted = T and vertices with one or more missing edges are within the adjacency matrix). As such, in Generate_sids_remv.R, we generate a list of samples to remove before generating weighted graphs. We do this by computing the number of NA relatedness estimates per sample, removing the sample with the highest count and iterating. Although there is a strong association between per-sample NA relatedness counts and per-sample marker data count, the removed sample doesn't always have fewest data.

Removal of 23 samples with one or more missing relatedness estimates, leaves 122760 relatedness estimates based on data on 12 to 248 SNPs shared between pairs among 496 samples. The year and place of sampling of the samples with one or more missing relatedness estimates are summarised in Table 3. The year and place of sampling of the remaining 496 samples (the samples used to generate graphs) are summarised in Table 4.

Graph analyses

Unless otherwise stated, the graph analyses exclude samples with one or missing relatedness analyses, which were identified in Generate_sids_remv.R (see above).

To see how the new samples (those that didn't feature in Taylor et al. 2020) and the old samples (those that did feature in Taylor et al. 2020) are related to one another, we plot a graph of relatedness in Plot_relatedness_graph.R.

To see how the new samples relate to the clonal components of the marker-reordered analysis of the samples that feature in Taylor et al. 2020 (45 CCs), we generate clonal components using the new samples (in Generate_components.R), and then (in Generate_relatedness_to_CCs.R) we compute the average relatedness between the 45 CCs and the new-data clonal components; between the 45 CCs and the new samples, where the latter set of new samples includes the samples that have one or more missing relatedness estimates; and between the clonal components based on all data combined. The results are plotted in Plot_relatedness_to_CCs.R.

To see how the new samples cluster with the 45 CCs, we generate clonal components using all the data together (in Generate_components.R), and then (in Compare_components.R) we take each of the 45 CCs, categorising them as either identical to one of the all-data clonal components, extended (nested within one of the all-data clonal components), or broken apart across one or more of the all-data clonal components. The results of clonal propagation are plotted in Plot_extended_components.R and Plot_cc_7.R.

Connectivity analyses

A brief connectivity analysis (Generate_LonLats.R, Generate_fraction_highly_related.R, Plot_fraction_highly_related.R and Generate_regression_trends.R) suggests access by ports and along the coast increases the probability of high relatedness to a greater extent to that of which distance in space and time decreases it. It also shows that the data from Esmeraldas essentially comes from a single clonal expansion.

Box 1

Consider a fair coin, whose probability of heads, $p = 0.5$, we want to estimate with confidence intervals using a Binomial model and the parametric bootstrap. The parametric bootstrap works by computing an estimate of p (e.g. a maximum likelihood estimate, mle) based on some observed real data; plugging that estimate, $\widehat{p}$, into the model; simulating data many times under the model with $\widehat{p}$ plugged-in; re-estimating p using the simulated data; and using a summary of the distribution of the many estimates of p based on simulated data to construct a confidence interval for $\widehat{p}$.

In the extremely sparse setting where the coin is flipped only once, the mle of p will either be zero (if the flip returns a tail) or one (if the flip returns a head). The data simulated under the model (the Binomial distribution with $n = 1$ flip) with probability set equal to either $\hat{p} = 0$ or 1 will either be all tails (if $\hat{p} = 0$) or all heads (if $\hat{p} = 1$) and the estimates of p based on simulated data will either be all zero or all one, respectively. Otherwise stated, there will be no diversity among estimates of p based on data simulated under a coin model whose probability is set equal to the mle based on a single flip.

Similarly, when the relatedness estimate, $\widehat{r}$, is one and there are data on only one SNP, the parametric bootstrap generates data simulated from a single locus that is IBD with probability equal to $\widehat{r} = 1$. Unless there are genotyping errors, all the simulated data will be IBS and all the relatedness estimates based on the simulated data will be equal to that based on the observed data. When $\widehat{r} = 0$ data are simulated from a single locus that is IBD with probability equal to $\widehat{r} = 0$. Depending on the allele frequencies, the observed data can either be IBS or not, and so relatedness estimates based on simulated data can either be one value that is different to the observed data or another value that is the same as the observed data.

Would-be sample counts with hard snp filters

In summary, the the above analyses were based on 512 samples in the non-graph analyses, with 128657 of 512-choose-2 (1.30816×10^5) sample comparisons; and 496 samples in the graph analyses with all 496-choose-2 (1.2276×10^5, 122760) sample comparisons.

If confidence intervals were not computed, a hard SNP cut off would be required, e.g. only compute relatedness estimates for sample pairs with data on at least 50, or 100 SNPs.

If the cut of was data on at least 50 SNPs, there would be 495 samples in the non-graph analyses, with 122217 of 495-choose-2 (1.22265×10^5) sample comparisons; and 489 samples in the graph analyses with all 489-choose-2 (1.19316×10^5) sample comparisons.

If the cut of was data on at least 100 SNPs, there would be 490 samples in the non-graph analyses, with 119451 of 495-choose-2 (1.19805×10^5) sample comparisons; and 476 samples in the graph analyses with all 476-choose-2 (1.1305×10^5) sample comparisons.

Relative to confidence-interval based non-graph and graph analyses respectively, the hard cut-offs amount to data losses of 4.62% and 1.41% using a 50-SNP cut-off, and of 5.59% and 4.03% using a 100-SNP cut-off.

Relative to confidence-interval based non-graph and graph analyses respectively, these data losses result in the dropout of cities (Orellana) and () using a 50-SNP cut-off; and (Orellana) and (Sucumbios, TobarDonoso) using a 100-SNP cut-off.

In the plots of the connectivity analyses fractions are limited to those with numerators of at least 10 observations per yearly partition and at least 10 samples per city for the city partitions.