Pacific coast 250-SNP barcode analysis

Aimee Taylor

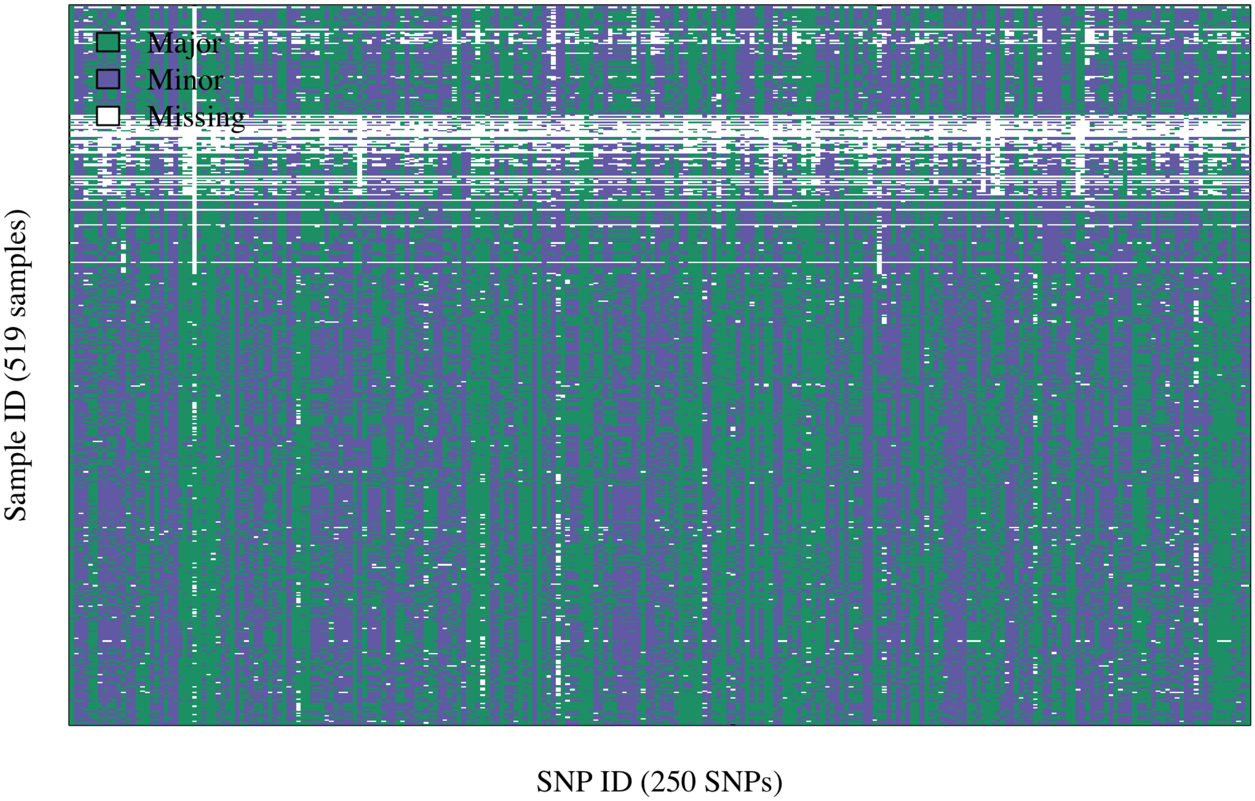

Figure 1: Plot of 250-SNP data, excluding samples with no 250-SNP data.

Table 1: Yearly sample counts per city of the samples with no 250-SNP 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 250-SNP data.

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

Methods

Pre-processing of data

The extended 250 SNP data set (snpdata.RData, formated in Format_snpdata_extended.R) includes 526 samples. Of these, 7 samples have no 250-SNP barcode data. Their sample ids are Pf054, Pf057, Pf058, Pf063, Pf064, Pf070, Pf073. The year and place of sampling of these 7 samples are summarised in Table 1.

Removal of the 7 samples with no data, leaves 519 samples with data on 1 to 250 SNPs (Figure 1). The year and place of sampling of the 519 samples with some 250-SNP data are summarised in Table 2. The full set of metadata are provided in metadata.RData, formatted in Format_metadata_extended.R.

Generation of relatedness estimates and confidence intervals

Relatedness inference between the 519 samples with some 250-SNP data was attemped for all pairs that shared at least one SNP of data (Generate_mles_CIs_extended.R). Using this very tolerant specification, 134258 relatedness estimates were computed based on data on 1 to 250 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 has very little bearing on the samples that feature in Taylor et al. 2020, because there are few missing SNPs among these samples (Figure 2). It has been corrected in the current analysis, where it has a large impact among the samples that did not feature in Taylor et al. 2020 due to many missing SNPs (Figure 2).

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 relatedess 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 sparasity.

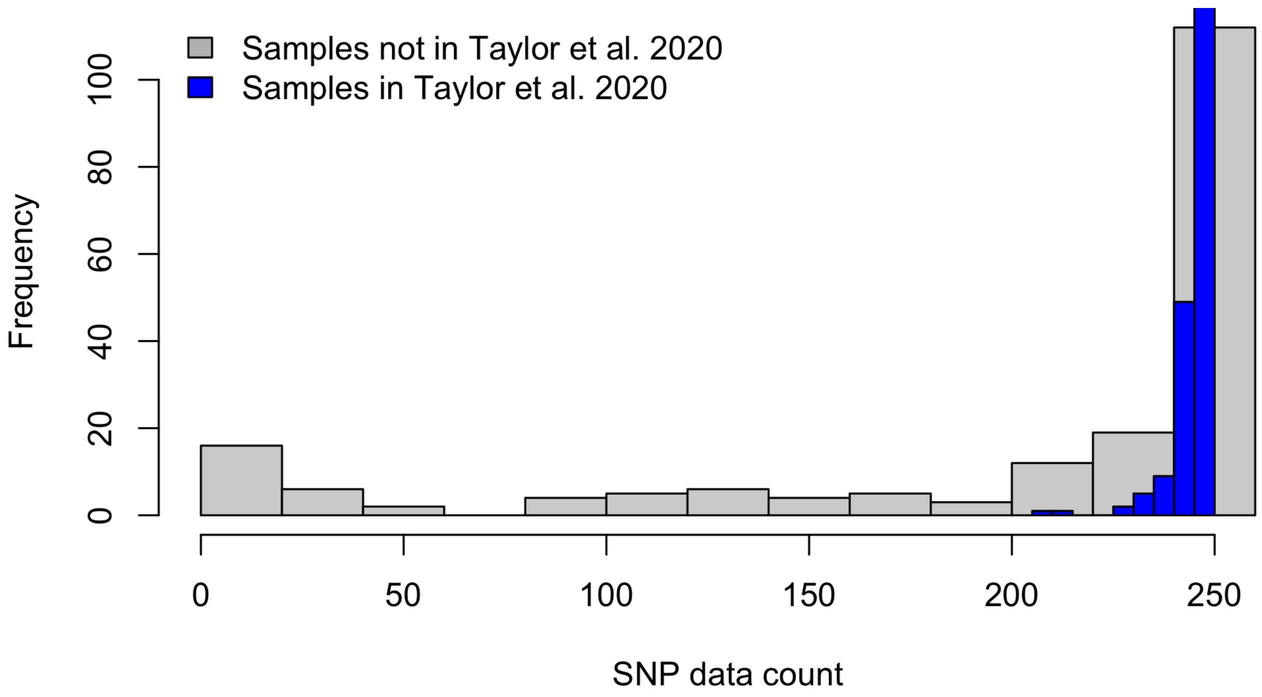

Figure 2: Histogram of SNP data counts

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

	2015	2013	2014	2016	Total
Guapi	4	0	0	1	5
Esmeraldas	0	3	1	0	4
SanLorenzo	2	1	5	0	8
Sucumbios	1	0	1	0	2
TobarDonoso	3	0	0	0	3
Total	10	4	7	1	22

Removal of the relatedness estimates where the parametric bootstrap has clearly failed, leaves 134258 relatedness estimates based on data on 1 to 250 SNPs shared between pairs among 519 samples. Meta data are added to these estimates in Format_mles_CIs_extended.R.

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.

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

	2005	2003	2004	2006	2000	2001	2002	1997	1999	2007	1993	1994	2015	2014	2016	2013	2017	Total
Tumaco	23	59	0	0	2	10	11	0	2	25	0	0	0	0	0	0	0	132
Guapi	0	66	0	0	1	0	0	0	1	0	0	0	31	5	79	0	1	184
Buenaventura	15	0	12	10	0	0	0	0	5	0	4	1	0	0	0	0	0	47
Quibdo	6	0	14	13	6	1	0	2	0	22	0	0	0	0	0	0	0	64
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	1	0	3	0	0	4
Venezuela	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2
StaBarbara_Iscuande	0	0	0	0	0	0	0	0	0	0	0	0	2	1	5	0	1	9
Esmeraldas	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	23	0	24
SanLorenzo	0	0	0	0	0	0	0	0	0	0	0	0	3	6	0	4	0	13
Orellana	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Sucumbios	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2
TobarDonoso	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Total	44	125	26	23	9	23	13	2	8	47	4	1	39	16	88	27	2	497

Removal of 22 samples with one or more missing relatedness estimates, leaves 123256 relatedness estimates based on data on 2 to 250 SNPs shared between pairs among 497 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 497 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.

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 46 clonal components reported in Taylor et al. 2020 (46 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 46 CCs and the new-data clonal components and between the 46 CCs and the new samples, where the latter set of new samples includes the samples that have one or more missing relatedness estimates. Heatplots and graphs of relatedness between the 46 CCs and the new-data clonal components and the new samples are generated in Plot_relatedness_to_CCs.R.

To see how the new samples cluster with the 46 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 46 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.

Connectivity analyses

A brief connectivity analysis (Generate_and_plot_connectivity.R) to see if there are links between ports suggests there are not strong links between ports in Colombia (Buenaventura and Tumaco) and Ecudaor (Esmeraldas). However, the data from Esmeraldas essentially come 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 probablity 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.