Preliminary exploration of P. falciparum relatedness on the Colombian coast

Notes to self

To reproduce all, run

1. DataFormat.R
2. Geo_dist_sites_colombia.R
3. Generate_Fst_Colombia.R
4. Generate_Result_Colombia.R
5. child scripts for separate analyses
6. this script for all analyses combined

To-do list

- Incorporate maps here
- Consider modelling as a mixture distribution Beta distributions (one for each of the n choose 2 site comparisons). The number of parameters would be
  - Beta distributions: 2 shape parameters * choose(number sites,2)
  - Mixing: 1 vector length choose(number sites,2)

Association of genetic signal to human mobility

- GIS analysis (cost surface creation)
- Add travel time to regression analyses (sent message to Marine Traffic and to Harvard CGA).
- Consider random design regression analysis to account for uncertainty in travel time

Is genetic connectivity between Buenaventura and Tamaco the result of single or multiple travel events?

Following discussions with Dan 8th Aug 2017, extreme scenarios inc.

1. single travel event single clone
2. single travel event with multiple clones
3. multiple travel single clone
4. multiple travel events differnt clones

In terms of malaria control, we care about single versus multiple travel events. If multiple different haplotypes underpin comparisons that are IBD, we can refute the null hypothesis that there was a single travel event with a single clone (calculation of LD might come into this). We can then compare the number of unique haplotypes in source populations. Perhaps still worth considering signal between Buenaventura and Tumaco wrt expected breakdown expected over time (based on regression).

Reason for no-longer pursuing STRUCTURE on the viterbi output: In terms of the viterbi output, the haplotype from a single travel event could give rise to multiple independent lineages of breakpoints. In which case, muliple clusters inferred using STRUCTURE run on the viterbi output would not refute the the null hypothesis that there was a single travel event. Even if there were a single lineage, STRUCTURE

may incorrectly identify nested clusters as independent (i.e. would consider a cluster whose breakpoints are compatible with mieotic errosion of a parent cluster to be independent of said parent cluster).

Following convo with Caroline on 8th. Could use an epi explanation to refute the null. For example, how big would the R0 need to be for the situation to occur. She also drew stripes based on just scalar expectations: repeat and make plot wider. That should suffice

Error log

On Friday 1st of Dec, I realised the lack of correlation between IBS and IBD was mainly due to not transposing the results of apply() and then also due to re-ordering SNPs for hmmIBD but not in the SNP data used to calculate FST and IBS. This resulted in meaningless FST and IBS estimates, which did not agree with previously FST published results nor correlate with IBD. The IBD analyses were unchanged. With a well chosen cut-off, IBS now performs as well as IBD in the Colombian setting.

For skype with Diego:

- Why "As expected, the lowest differentiation between the provinces was found between Valle-Nariño (FST = 0.023) while the highest was for Cauca–Chocó (FST = 0.117) (Additional file 4)."
- Confirm missing '–'
- Apologise for not seeing the lon/lat
- How to consolidate story if reporting FST?
- What is the probability of being asymptomatic and therefore missing nodes?

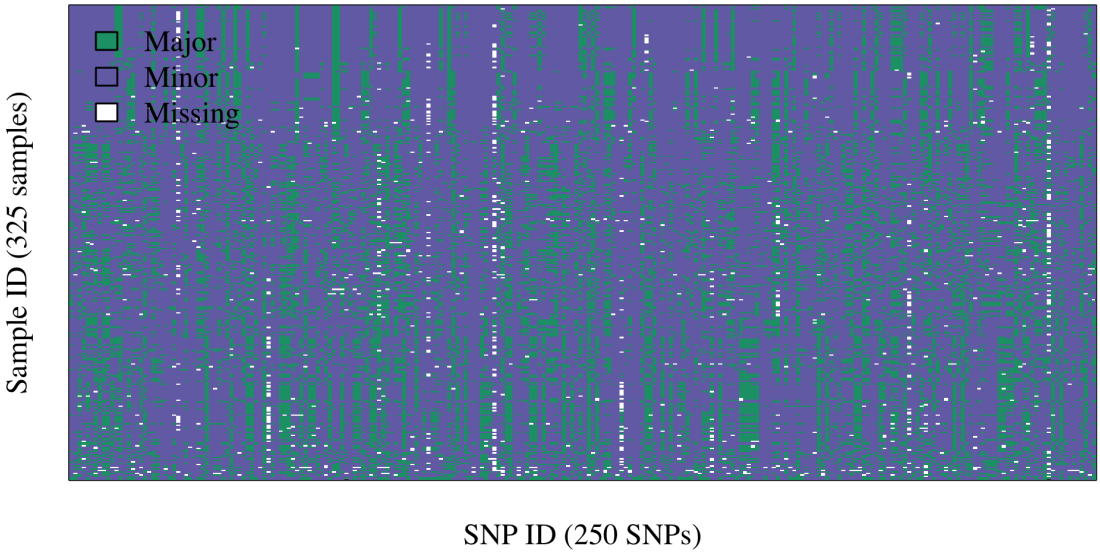


Figure 1: Plot of raw data.

Summary of the data

The data set includes 325 samples collected from 5 cities on the Colombian Pacific coast between 1993 and 2007 (Table 1). Samples were genotyped at 250 SNPs (Figure 1), with minor allele frequencies ranging from 0.006 to 0.495 (Figure 2).

Table 1: Yearly sample counts per city

	1993	1994	1997	1999	2000	2001	2002	2003	2004	2005	2006	2007	Total
Tumaco	0	0	0	2	2	10	11	59	0	23	0	25	132
Guapi	0	0	0	1	1	0	0	66	0	0	0	0	68
Buenaventura	4	1	0	5	0	0	0	0	12	15	10	0	47
Quibdo	0	0	2	0	6	1	0	0	14	6	13	22	64
Tado	0	0	0	0	0	12	2	0	0	0	0	0	14
Total	4	1	2	8	9	23	13	125	26	44	23	47	325

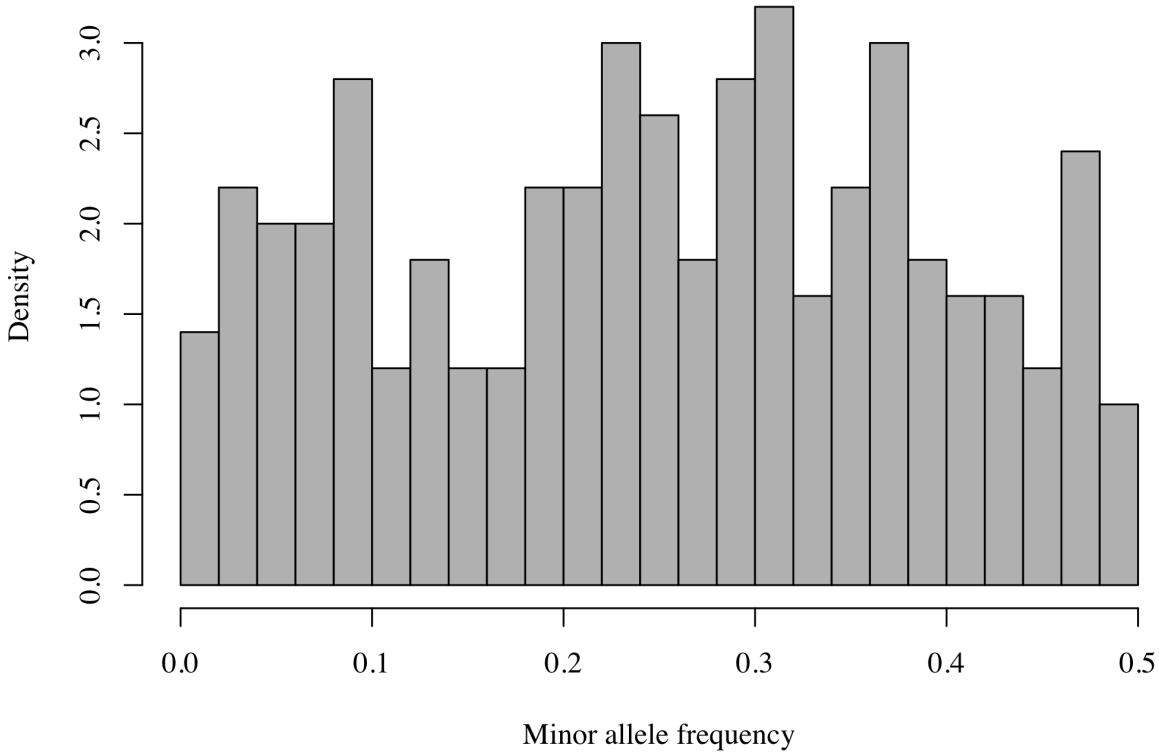


Figure 2: Minor allele frequency distribution of all 250 SNPs.

Summary of FST analyses

Previously published FST analyses

Previously published FST estimates were generated at the level of the state using GENALEX6 (Figure 3). Echeverry et al. note that "As expected, the lowest differentiation between the provinces was found between Valle-Nariño (FST = 0.023) while the highest was for Cauca–Chocó (FST = 0.117)."

Re-estimated FST results

City level

Using Hudson's estimator, I recover estimates that do not vary systematically way with distance (Figure 4). P-values generated by permutation of city labels suggest significance (Table 2), as do confidence intervals generated using an independent method of bootstrapping over SNPs (Figure 4).

The point estimates agree almost perfectly with estimates generated using Weir and Hill's and Weir and Cockerham's estimators (top row, Figure 5). The latter are not strongly sensitive to unbalanced sample sizes (bottom left, Figure 5). Of those that can be compared with previously published results at the level of the state, estimates do not agree in absolute terms but are highly correlated (bottom right, Figure 5).

State level

To explore the discrepancy between the previously published results and my estimates, I re-estimated FST at the state level, and compared results with estimates generated using a independent package, DivRsity, which reportedly uses Weir & Cockerham's 1984 estimator. DivRsity estimates almost perfectly match those generated using Hudson's, Wier and Hill's and Wier and Cockerham's estimators (top row and bottom left, Figure 6). Absolute values do not agree with previously published results, but the estimates are highly correlated (bottom right, Figure 6). Although previously published and re-estimated FST estimates do not agree, because they are highly correlated, the trend with respect to distance is the same based on previously published versus re-estimated FST estimates at the level of the state (Figures 3 and 7, respectively).

Table 2: P-values of FST estimated using the Hudson estimator

	p-value
Quibdo Tado	0.001
Guapi Tumaco	0.000
Buenaventura Tado	0.000
Buenaventura Guapi	0.000
Buenaventura Quibdo	0.000
Buenaventura Tumaco	0.004
Guapi Tado	0.000
Guapi Quibdo	0.000
Tado Tumaco	0.000
Quibdo Tumaco	0.000

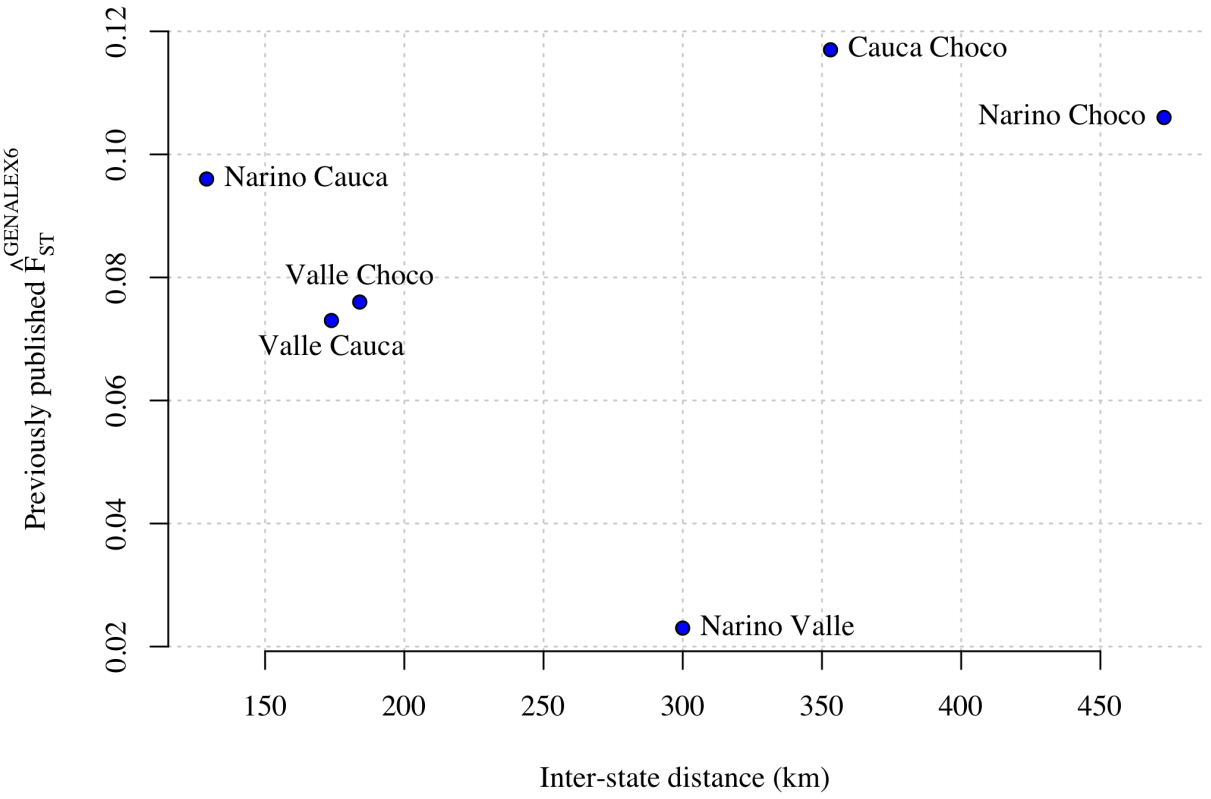


Figure 3: Previously published FST estimates generated using GENALEX 6. No error bars available. Provinces from North to South (with cities in parentheses): Chocó (Tadó and Quibdó), Valle (Buenaventura), Cauca (Guapi), Nariño (Tumaco).

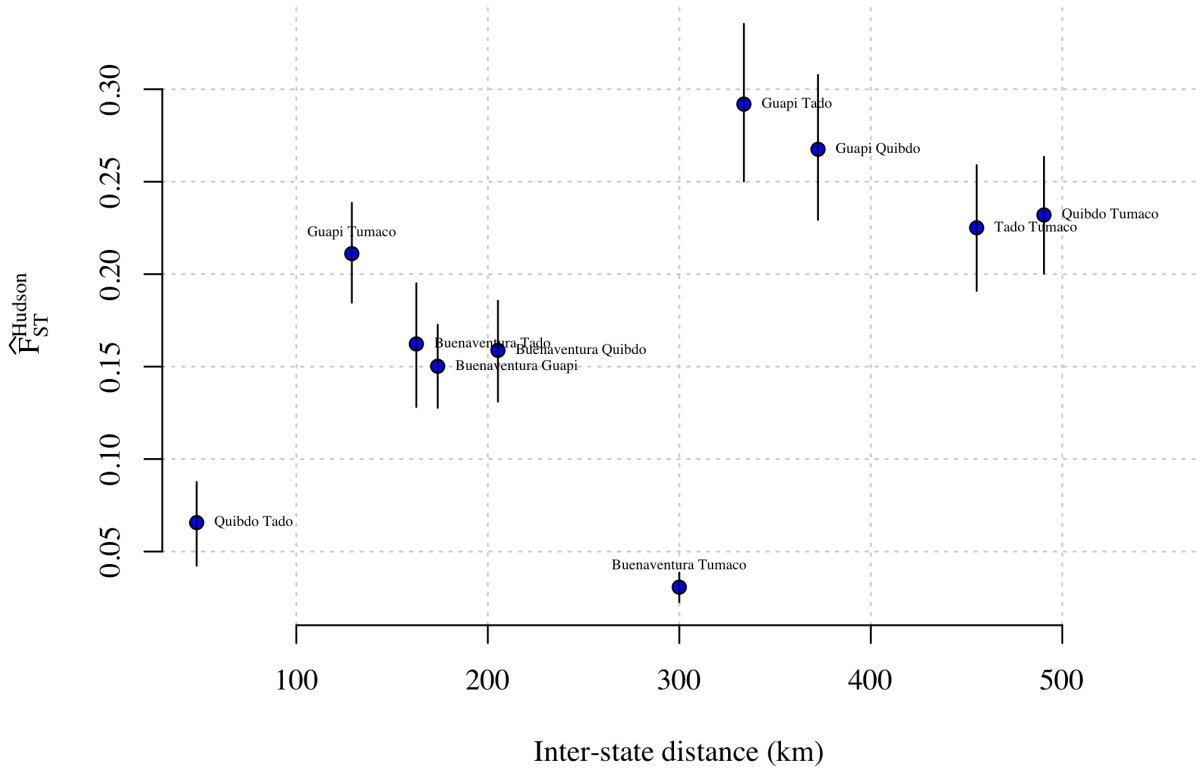


Figure 4: FST estimated at the level of the city using Hudson's estimator.

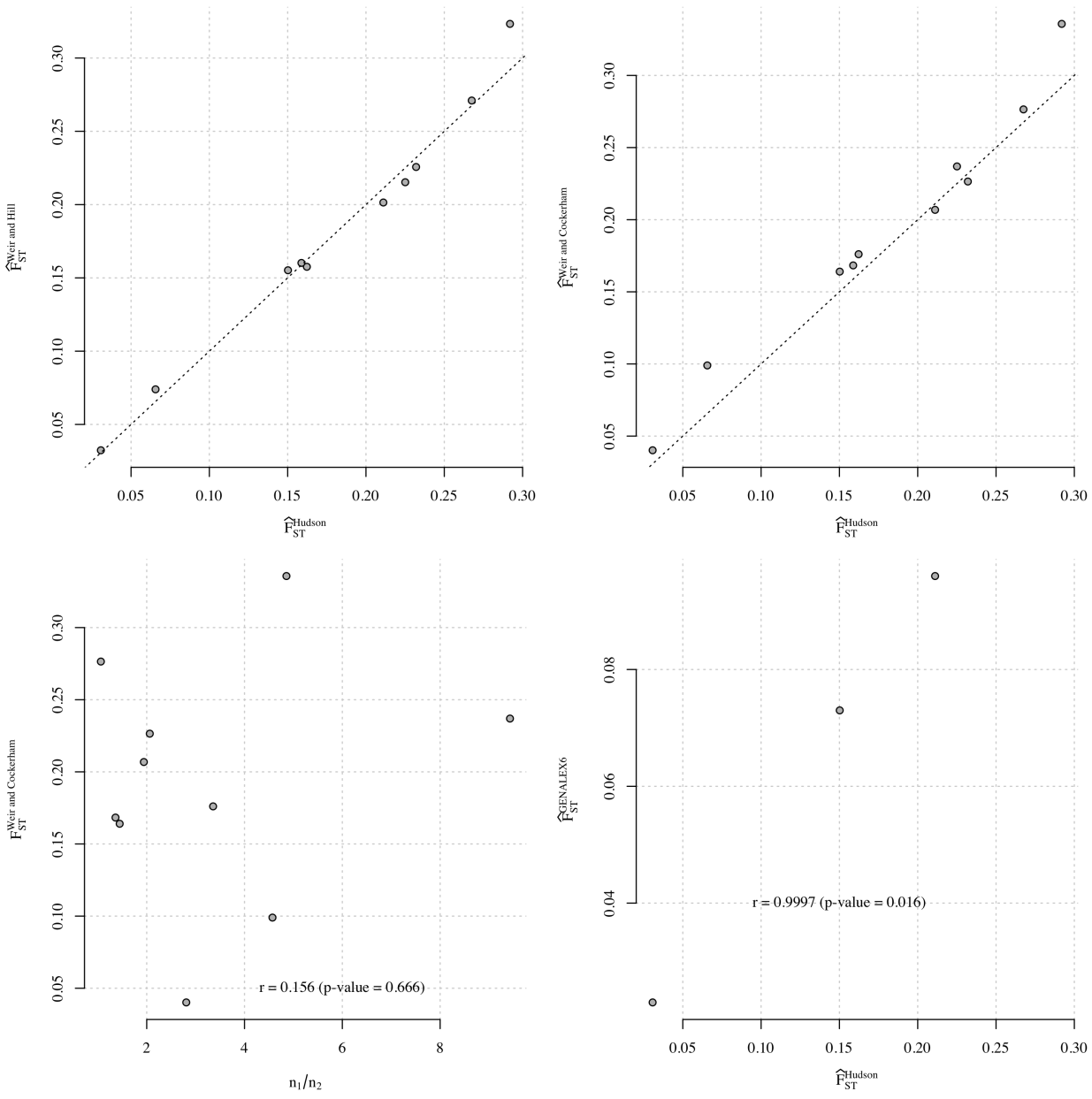


Figure 5: City-level FST estimates generated using different estimators.

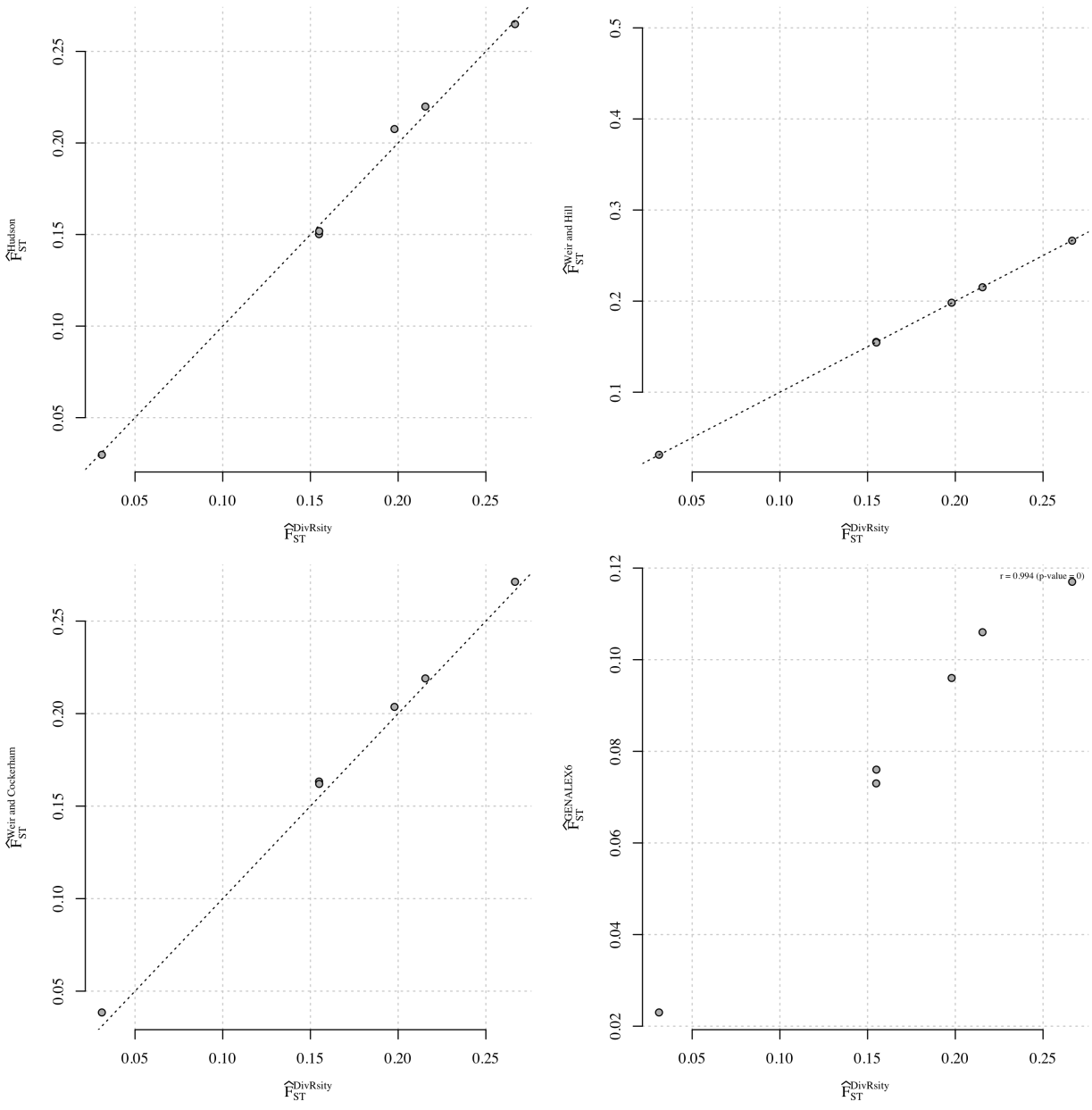


Figure 6: State-level FST estimates generated using different estimators.

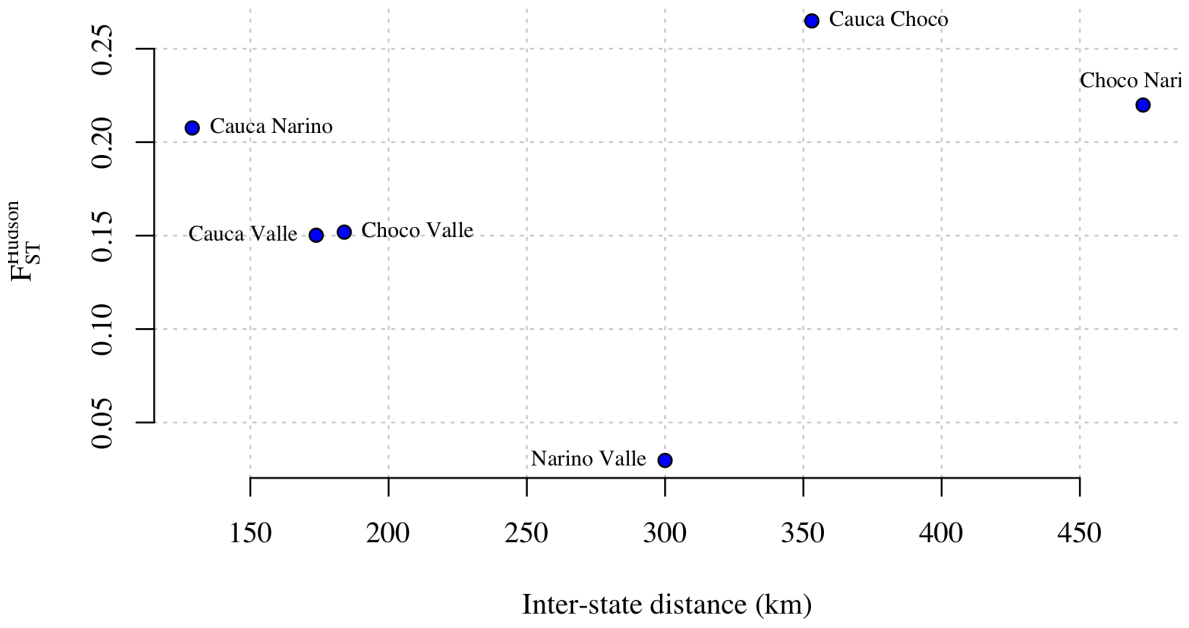


Figure 7: State-level FST estimates generated using Hudson's estimator plotted with respect to inter-state distance (km).

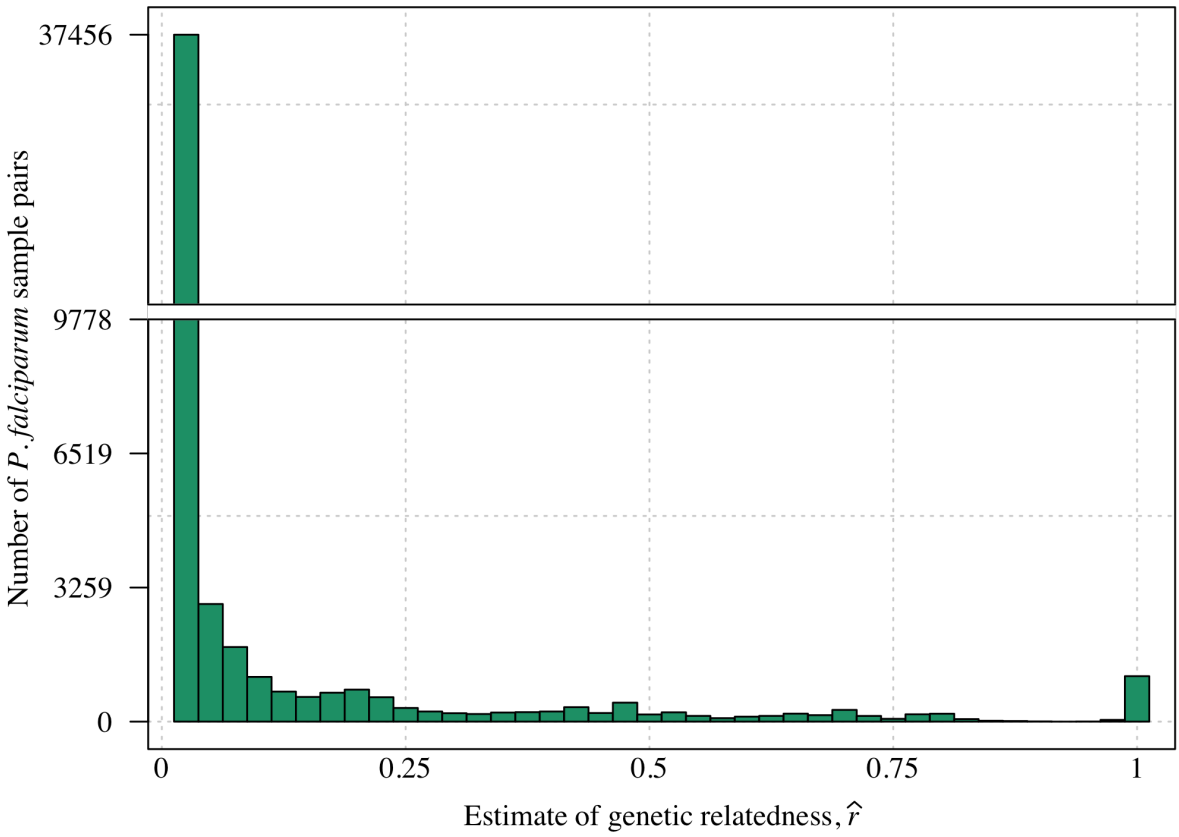


Figure 8: Estimates of $\hat{r}$ for all pairwise comparisons of 325 samples.

Summary of IBD analyses

In the following I have explored the trend in parasite relatedness (based on identity by descent, IBD, estimated using hmmIBD [REF]) with inter-city distance (km), with a view to comparison with the Thai-Myanmar border, where, on average, the log-odds of relatedness decrease by 0.02 with every kilometer between collection sites and week between collection dates. I have treated het calls and those labelled '–' as missing since all the samples were previously categorized as single-genotype.

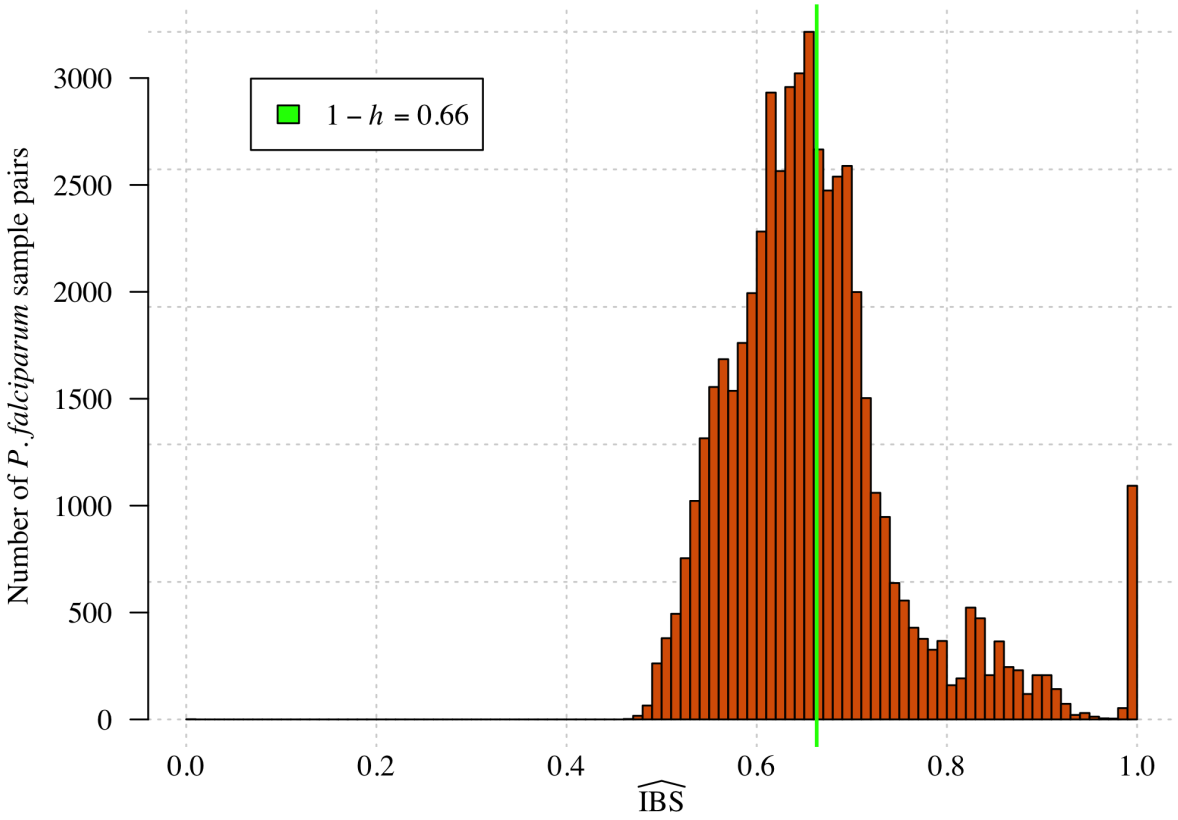
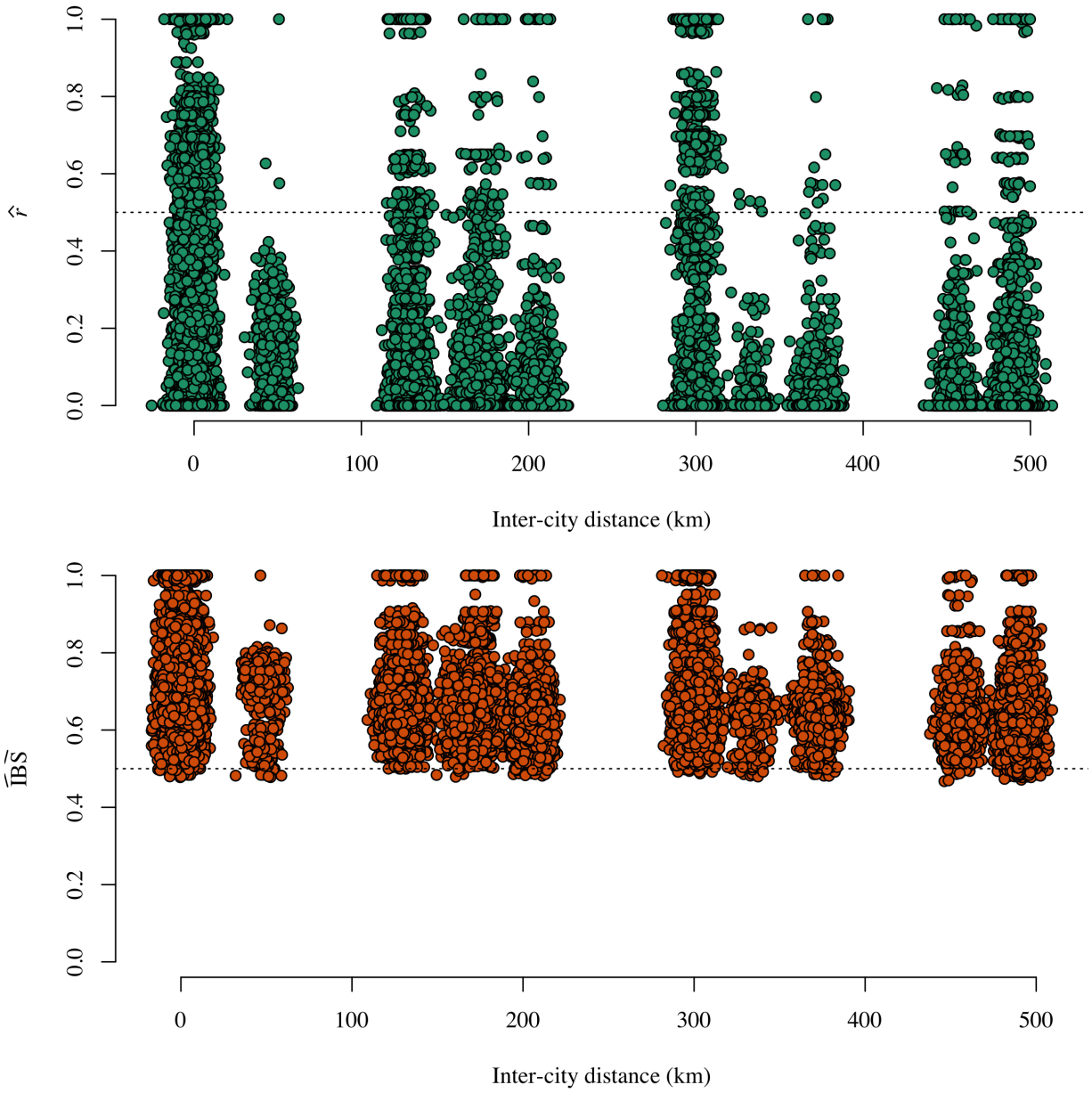


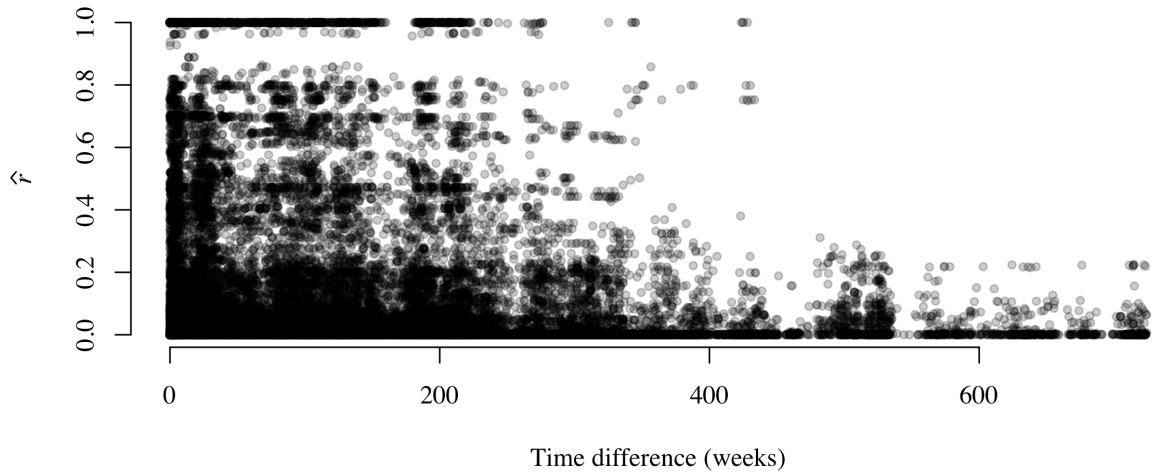
Figure 9: Estimates of π_{IBS} for all pairwise comparisons of 325 samples.

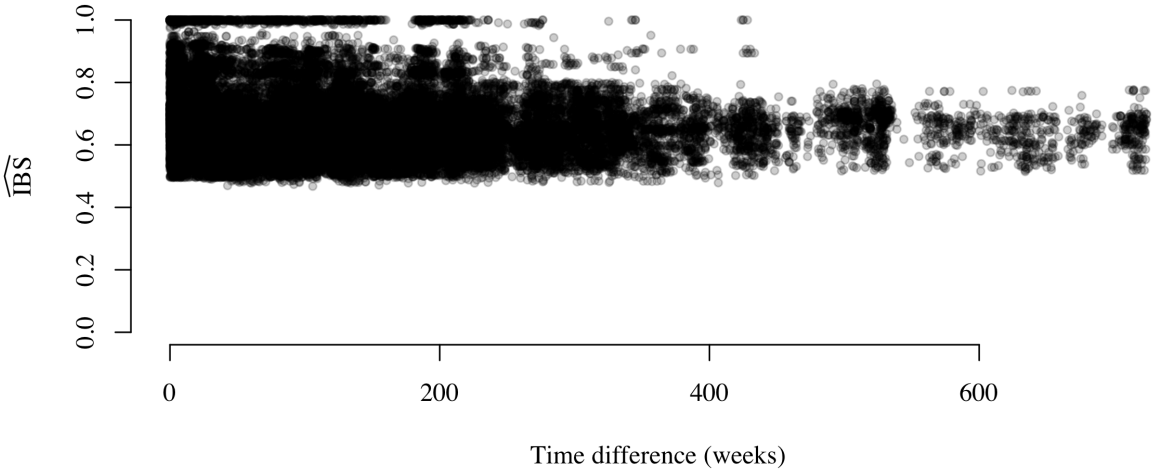
Plots of $\widehat{r}$ and $\widehat{\text{IBS}}$

Plot of $\widehat{r}_m$ and $\widehat{\text{IBS}}_m$ against distance

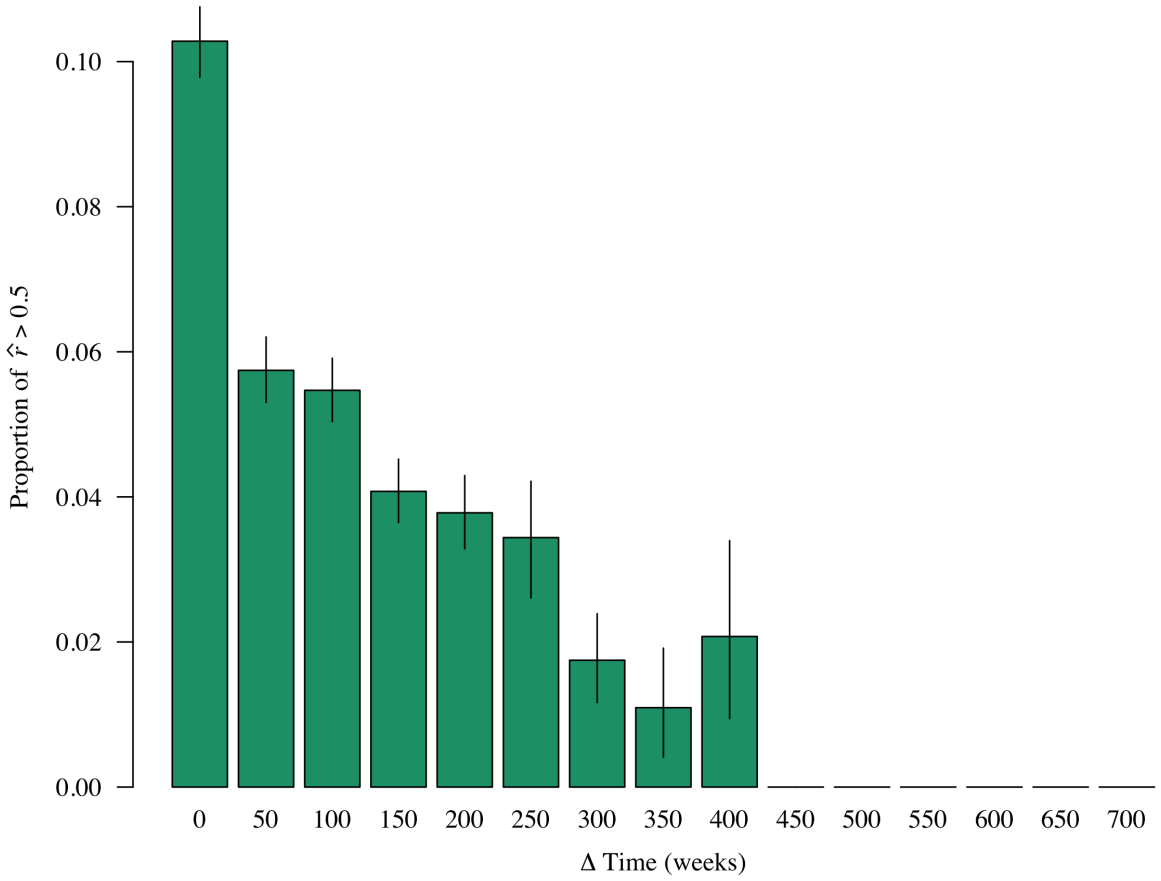


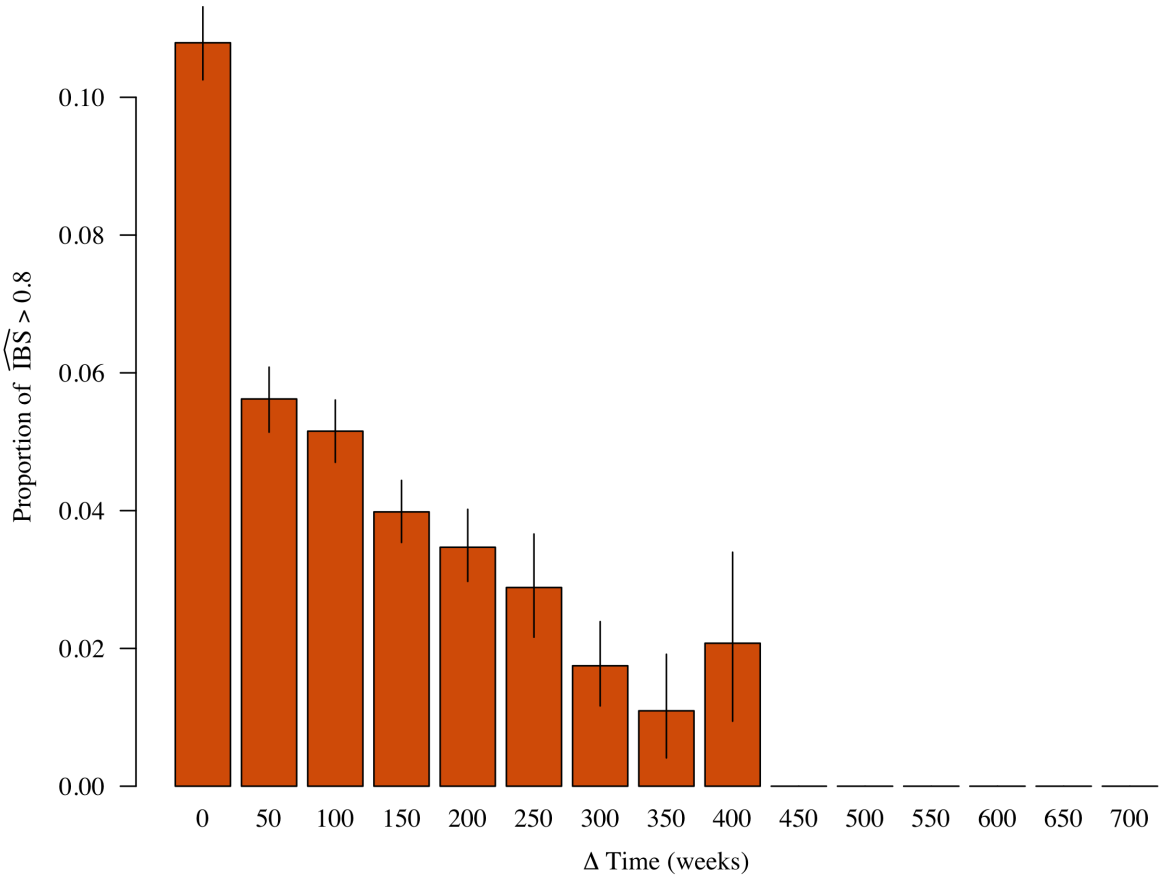
Plot of $\widehat{r}_m$ and $\widehat{\text{IBS}}_m$ against time



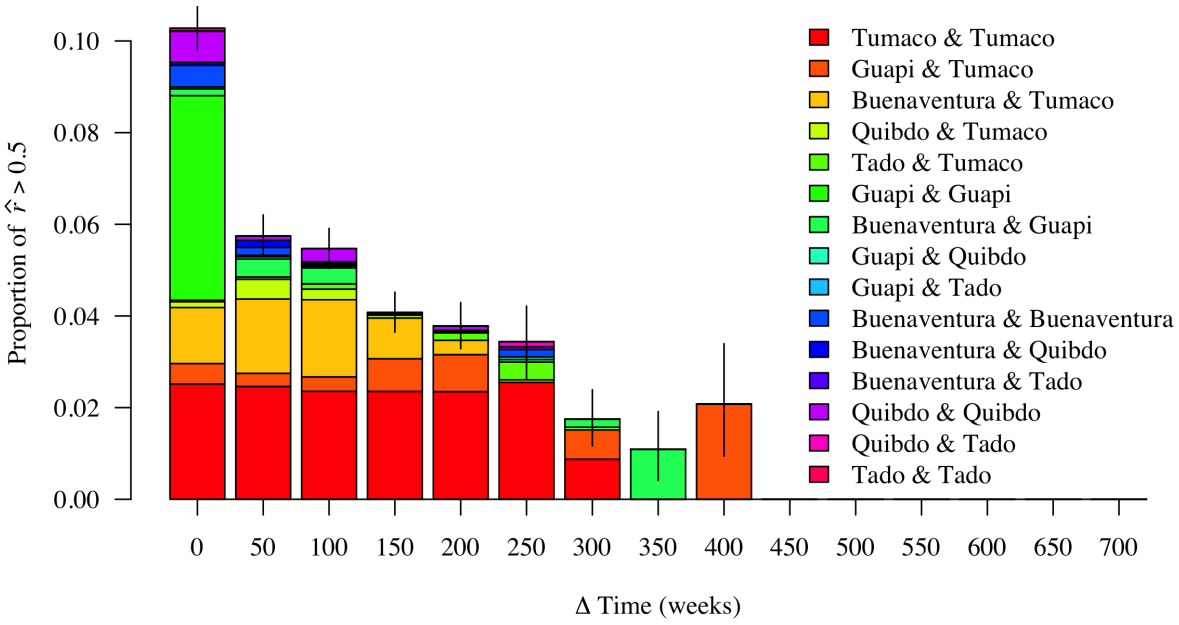


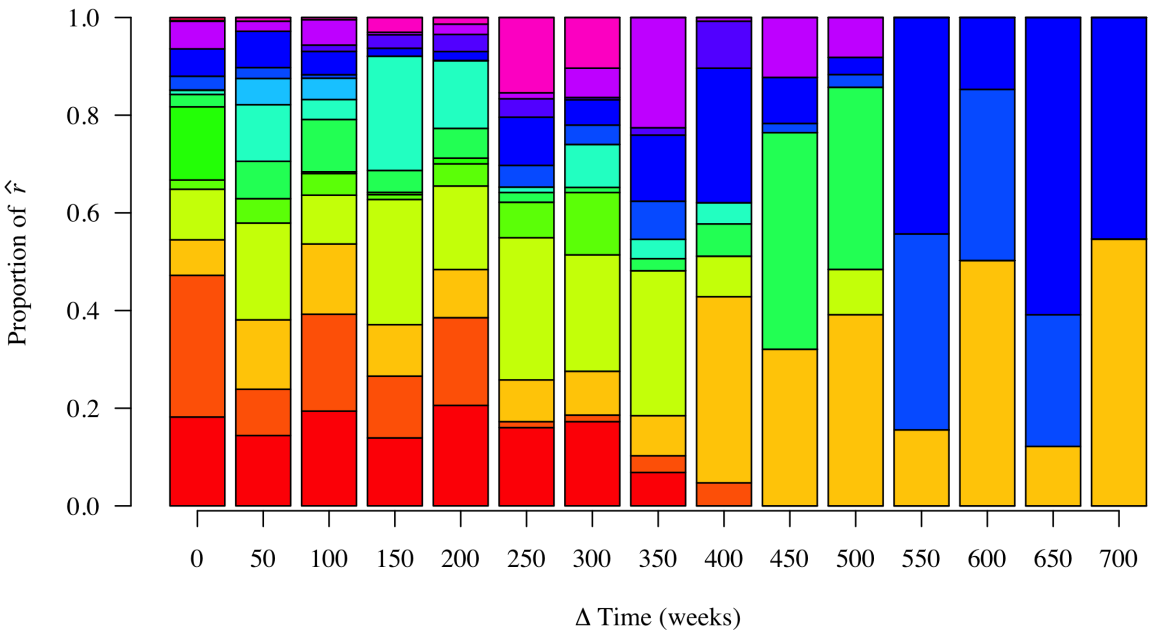
Plot of proportions $\widehat{r}_m > 0.5$ and $\widehat{\mathbf{IBS}}_m > 0.84$ against time



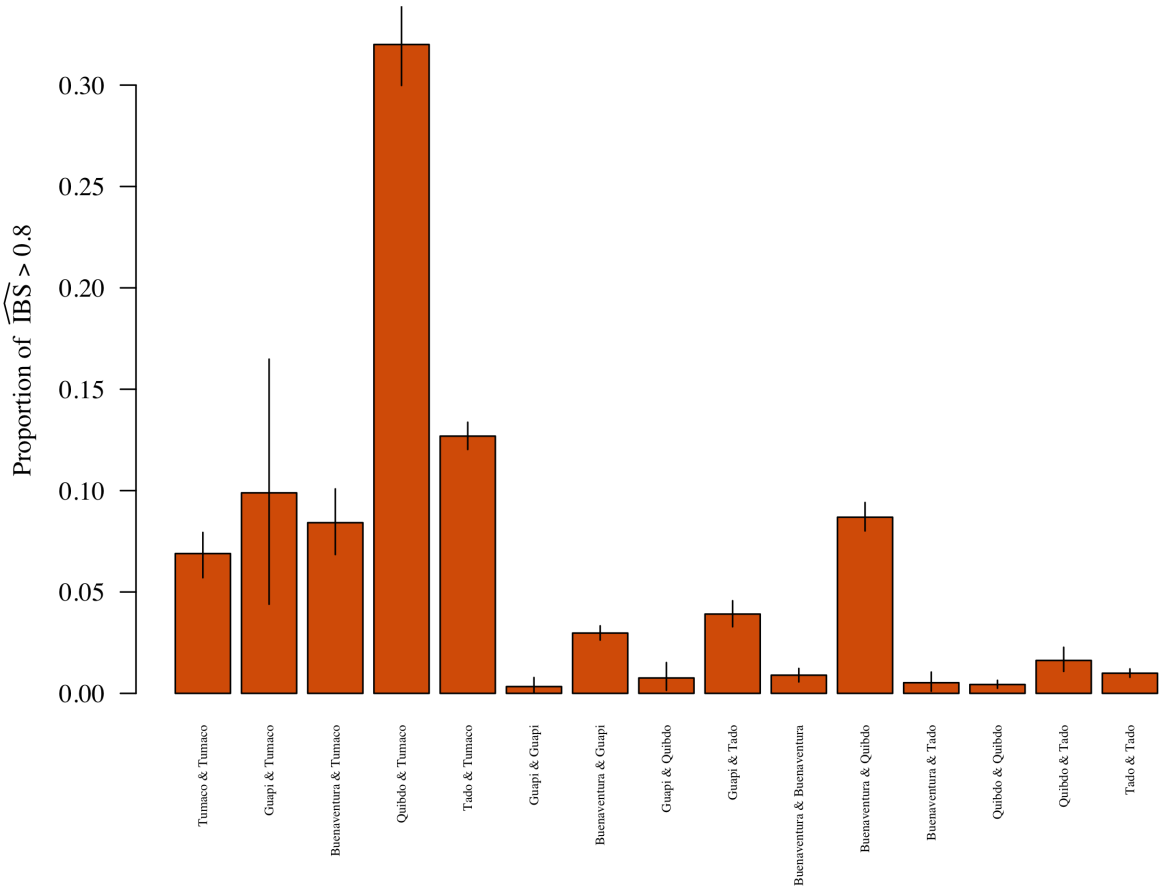

0.10
0.08
0.06
0.04
0.02
0.00
Proportion of $\widehat{\text{IBS}} > 0.8$
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
Δ Time (weeks)

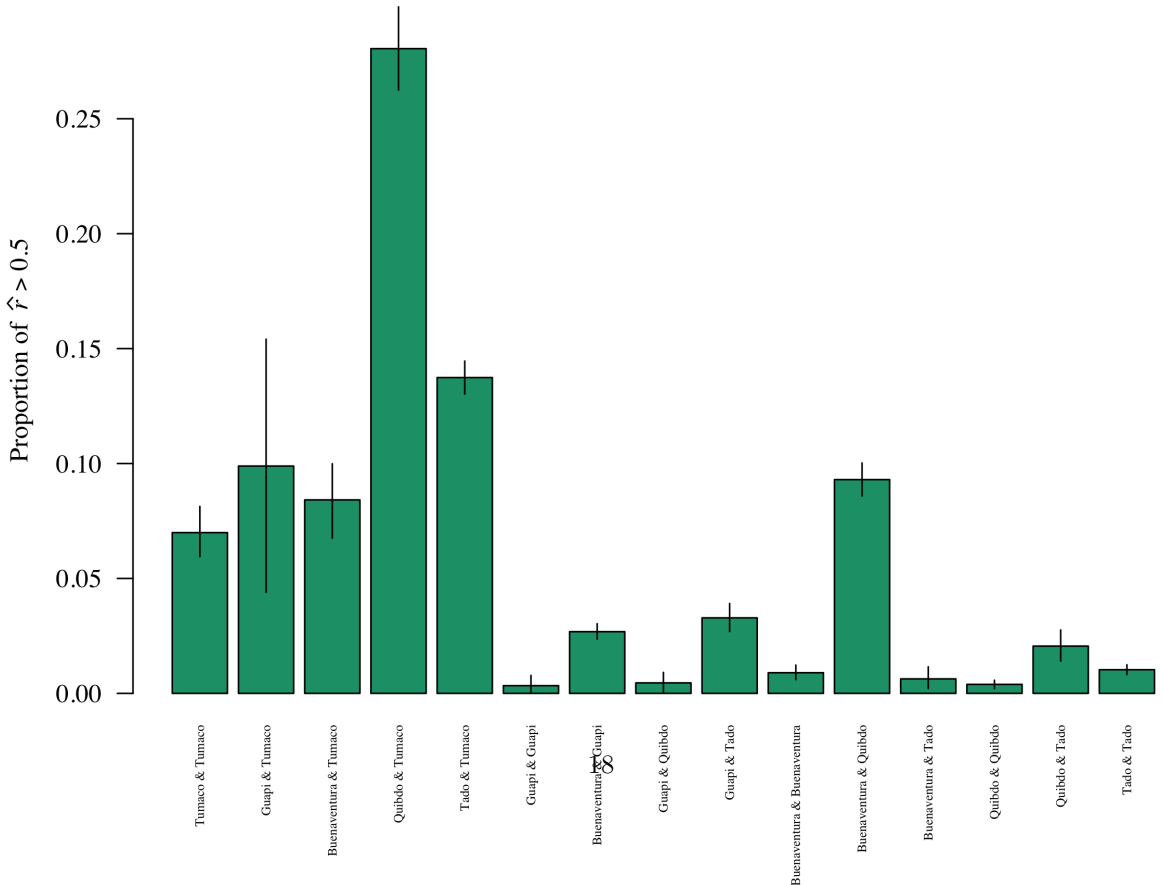
Plot of proportions $\widehat{r}_m > 0.5$ in time broken down by site comparison



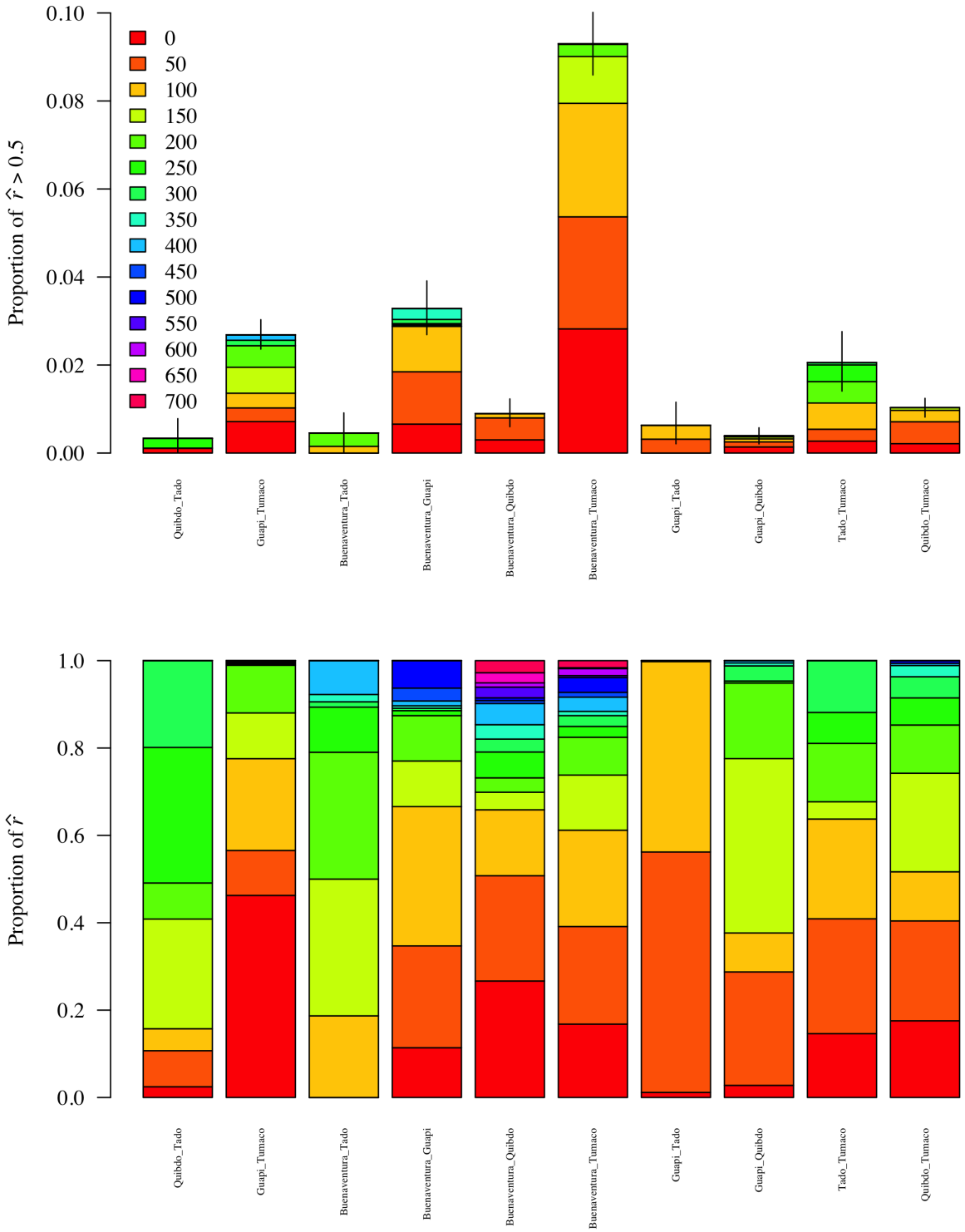


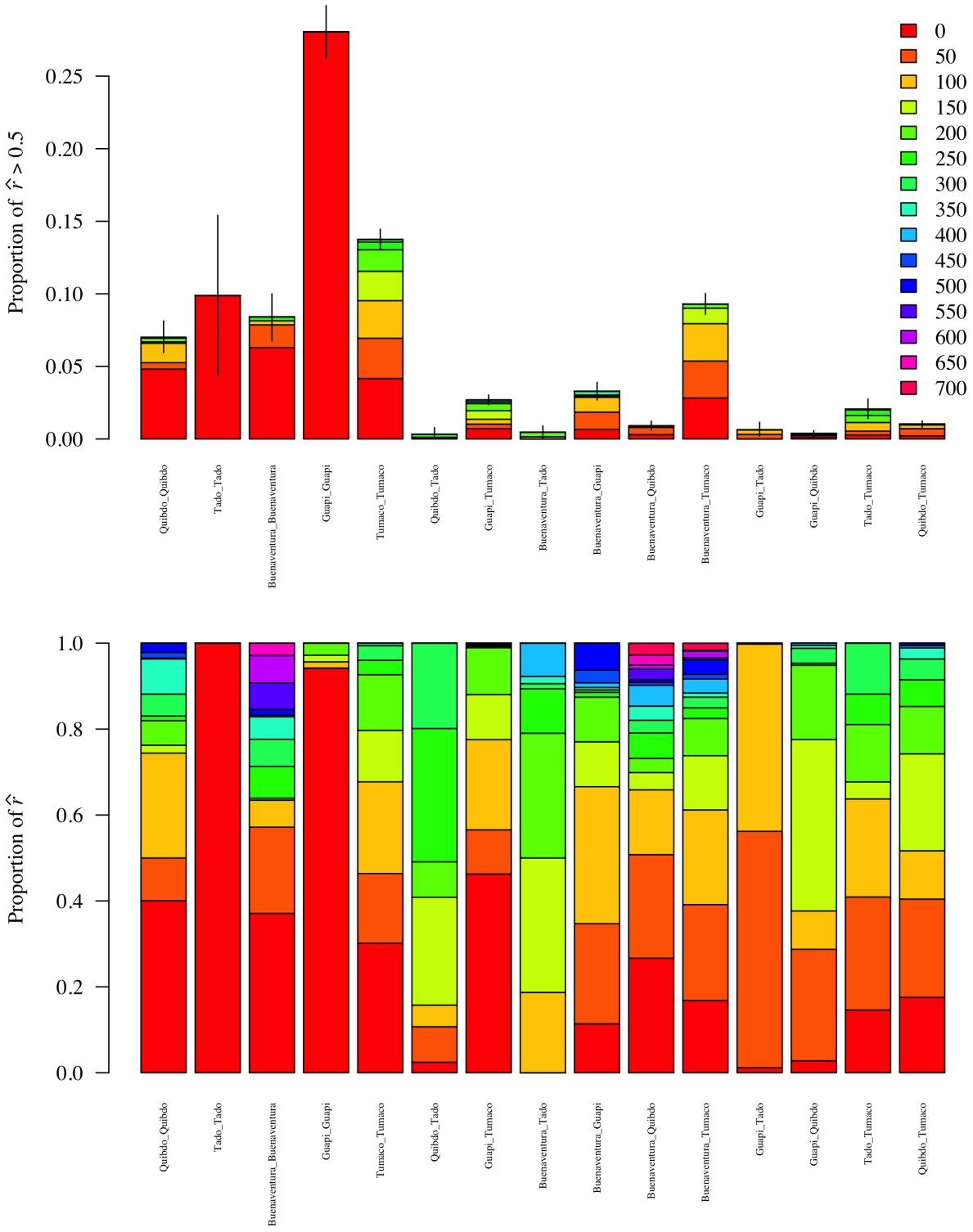
Plot of proportions $\widehat{r}_m > 0.5$ and $\widehat{\text{IBS}}_m > 0.84$ against city comparisons



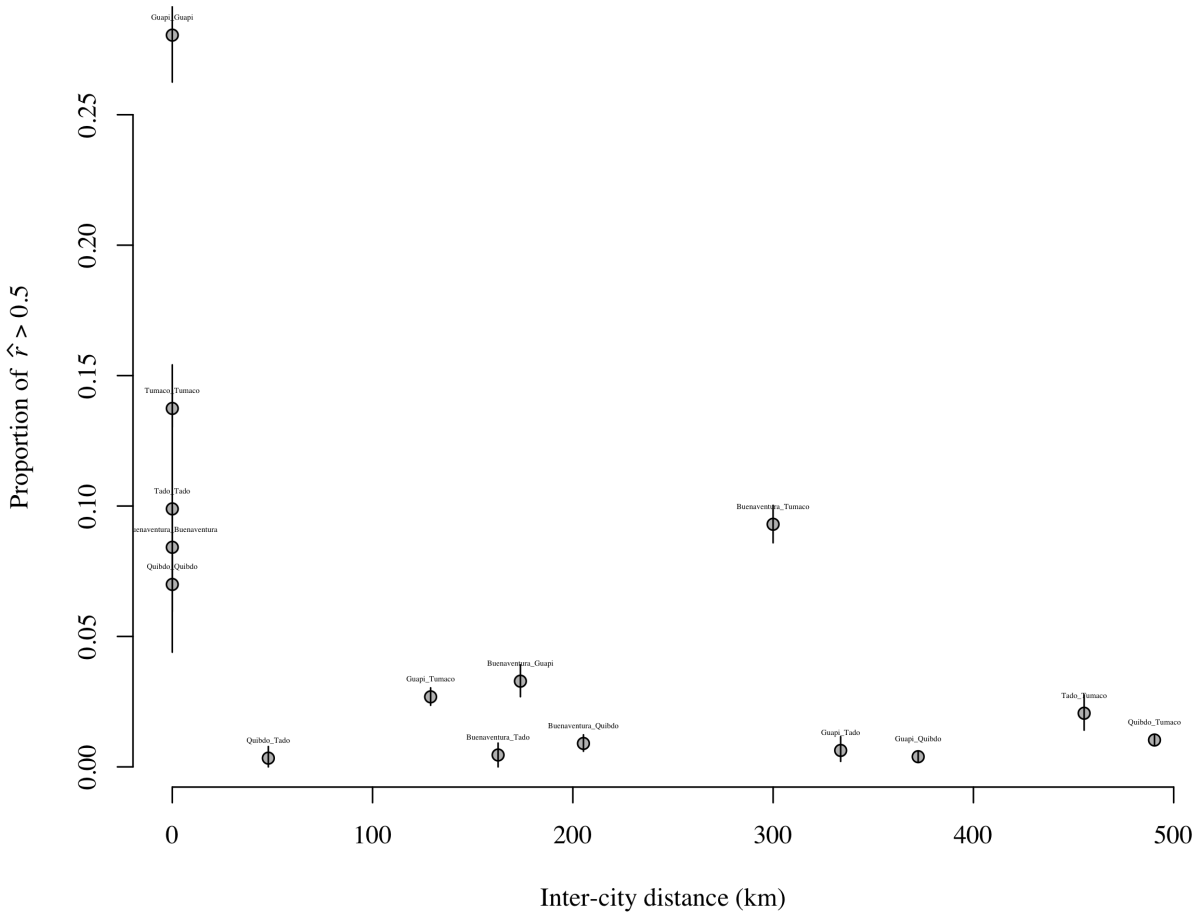


Plot of proportions $\widehat{r}_m > 0.5$ in time broken down by time between collection dates (weeks)

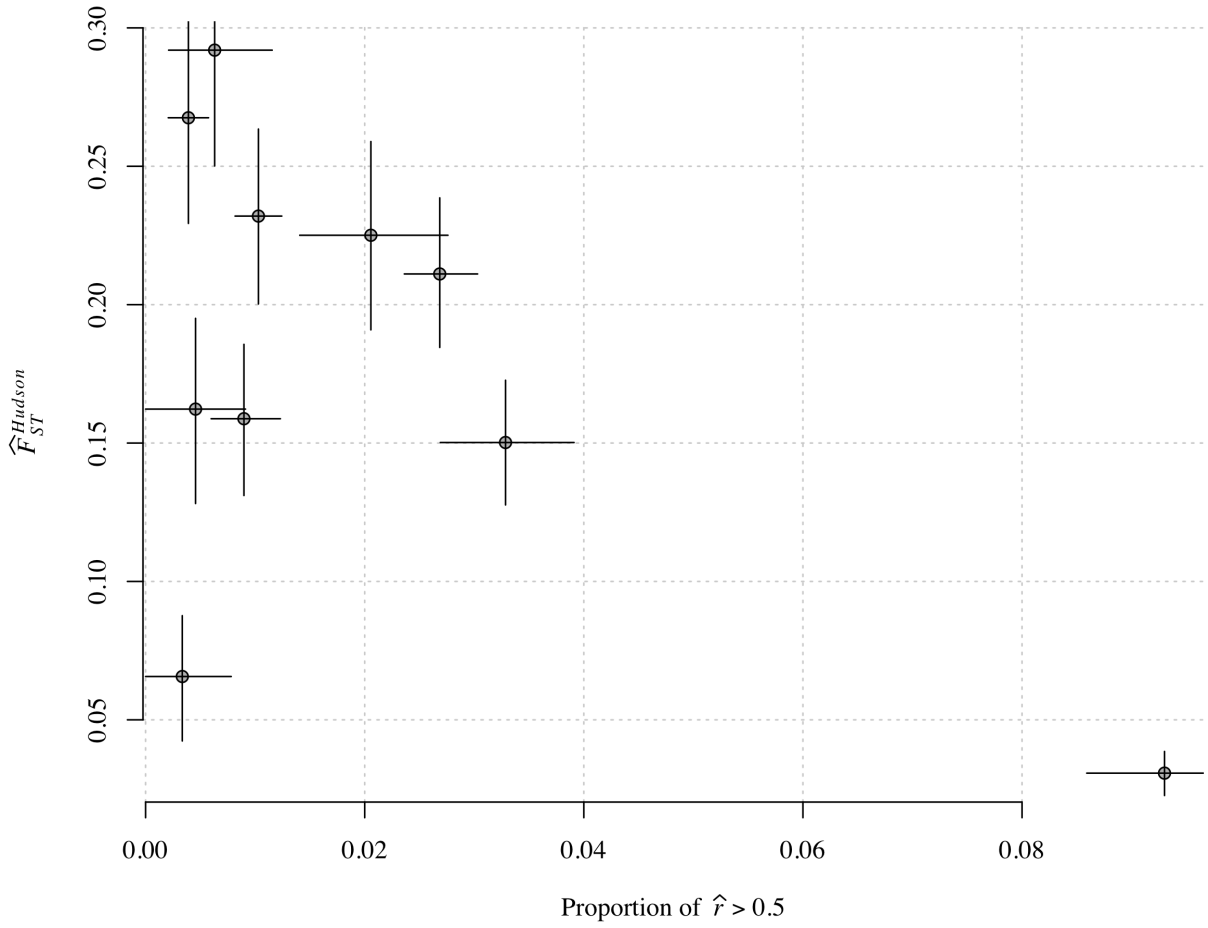



Proportion of $\hat{r} > 0.5$
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
Quibdo_Quibdo
Tado_Tado
Buenaventura_Buenaventura
Guapi_Guapi
Tumaco_Tumaco
Quibdo_Tado
Guapi_Tumaco
Buenaventura_Tado
Buenaventura_Guapi
Buenaventura_Quibdo
Buenaventura_Tumaco
Guapi_Tado
Guapi_Quibdo
Tado_Tumaco
Quibdo_Tumaco
Proportion of $\hat{r}$
1.0
0.8
0.6
0.4
0.2
0.0
Quibdo_Quibdo
Tado_Tado
Buenaventura_Buenaventura
Guapi_Guapi
Tumaco_Tumaco
Quibdo_Tado
Guapi_Tumaco
Buenaventura_Tado
Buenaventura_Guapi
Buenaventura_Quibdo
Buenaventura_Tumaco
Guapi_Tado
Guapi_Quibdo
Tado_Tumaco
Quibdo_Tumaco

Plot of proportions $\widehat{r}_m > 0.5$ against inter-city distance



Plot of city level FST estimates against proportions $\widehat{r}_m > 0.5$



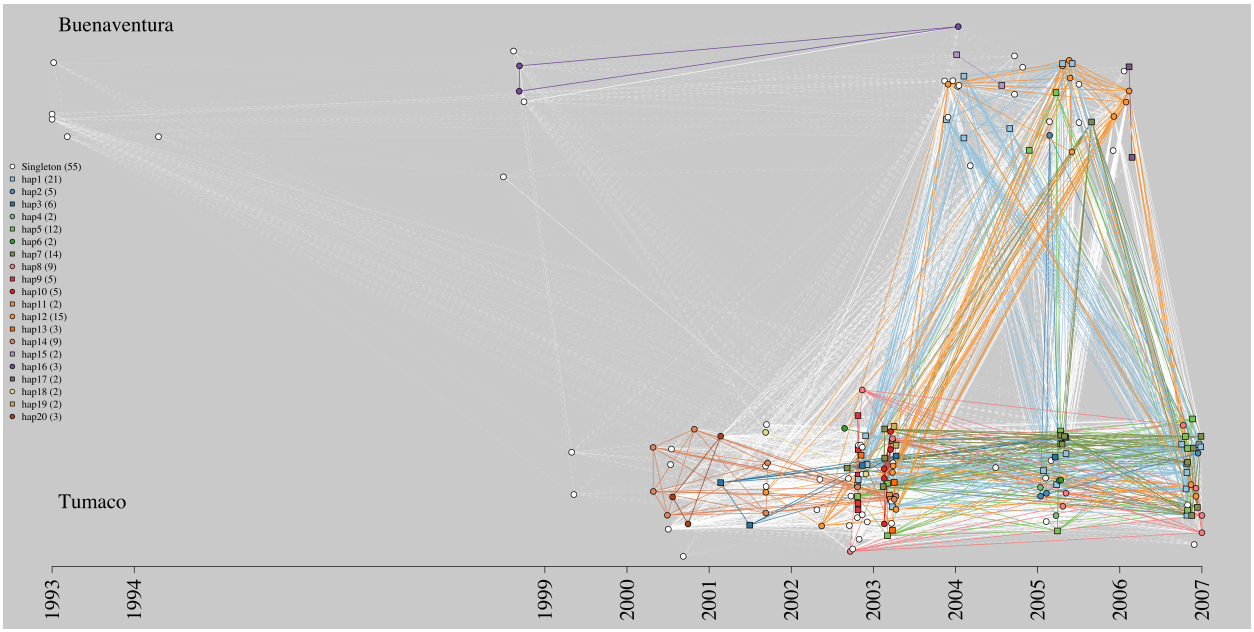
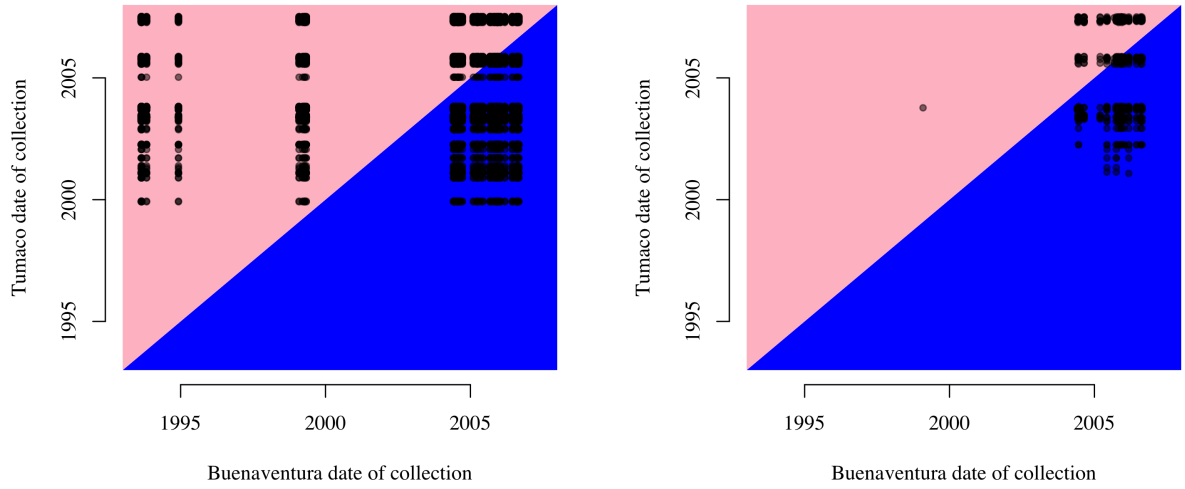


Figure 10: Edges with IBD between 0 and 100% (non inclusive) are white, while those with 100% IBD are coloured by haplotype labels. Edges with IBD $< 50\%$ are dashed, while those with IBD $> 50\%$ are solid.

Are all Buenaventura Tomaco disproportionally from 2005? Was it a clonal outbreak?



Deep-dive into Buenaventura and Tumaco

We see a lot of genetic relatedness within and between sites and want to know if there is A) ongoing or one-time gene flow and B) two-way or source to sink gene flow.

Assuming travel between cites Tumaco and Buenaventura is symmetrical (try to confirm with expert knowledge

/ maritime traffic data), the site with higher incidence (Tumaco) is the most likely source.

Vertical lines are compatible with a transmission event, but could also arise due to similar haplotypes being detected at the same time across sites. Diagonal lines are not compatible with direct travel.

The fact that there are multiple repeat haplotypes suggests gene flow was not a one-time event. The fact that each site has some private repeat haplotypes suggests that there is potential for two-way flow.

Assuming Tumaco to be the source, Figure 10 suggests haplotypes 5 and 7 were transmitted from Tumaco to Buenaventura sometime after haplotypes 1 and 12 (circa years 2005 and 2003, respectively). Haplotype 2 was not detected in either site before 2005, we thus speculate that it was transmitted across sites sometime around 2005. Upon inspection of Figure 10 it is important to note however that absence of evidence is not evidence of absence due to non-exhaustive sampling of infections (e.g. all haplotypes could have been present in either site prior to their detection due to unsampled infections).

The number of differently coloured repeat haplotypes might give a false impression of the number of independent pieces of information, however. A network of relatedness between the repeat haplotypes (white edge IBD = 0.001, darkest edge IBD = 0.83) was thus plotted (Figure 11).

Of the haplotypes shared across Buenaventura and Tumaco, haplotypes 5, 7 and 12 are all highly related to one another, so might originate from a single parasite. On the contary, haplotypes 1 and 2 are not highly related to the above cluster nor one-another. Together these results suggest three travel events.

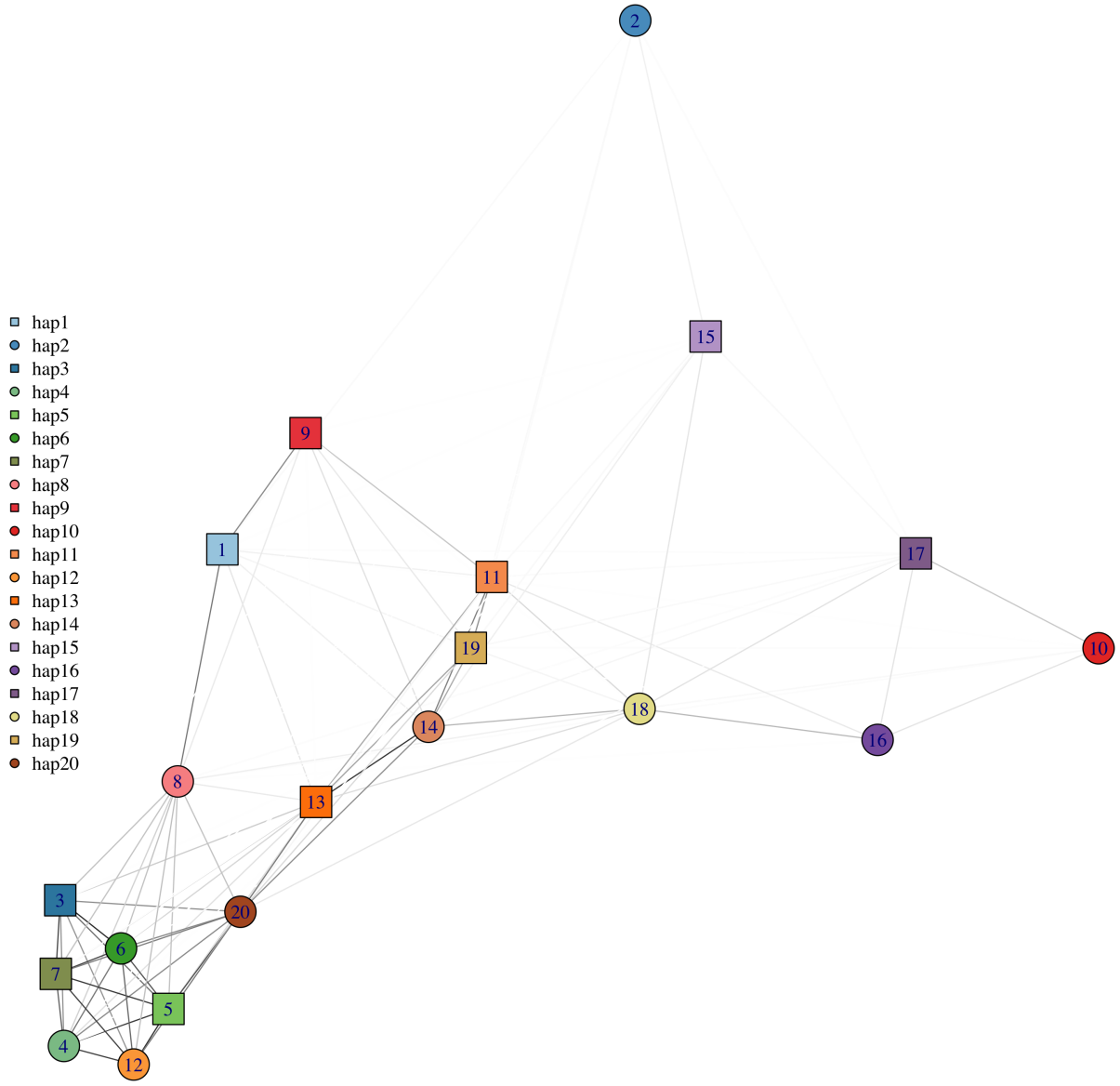


Figure 11: Relateness across repeat haplotypes.

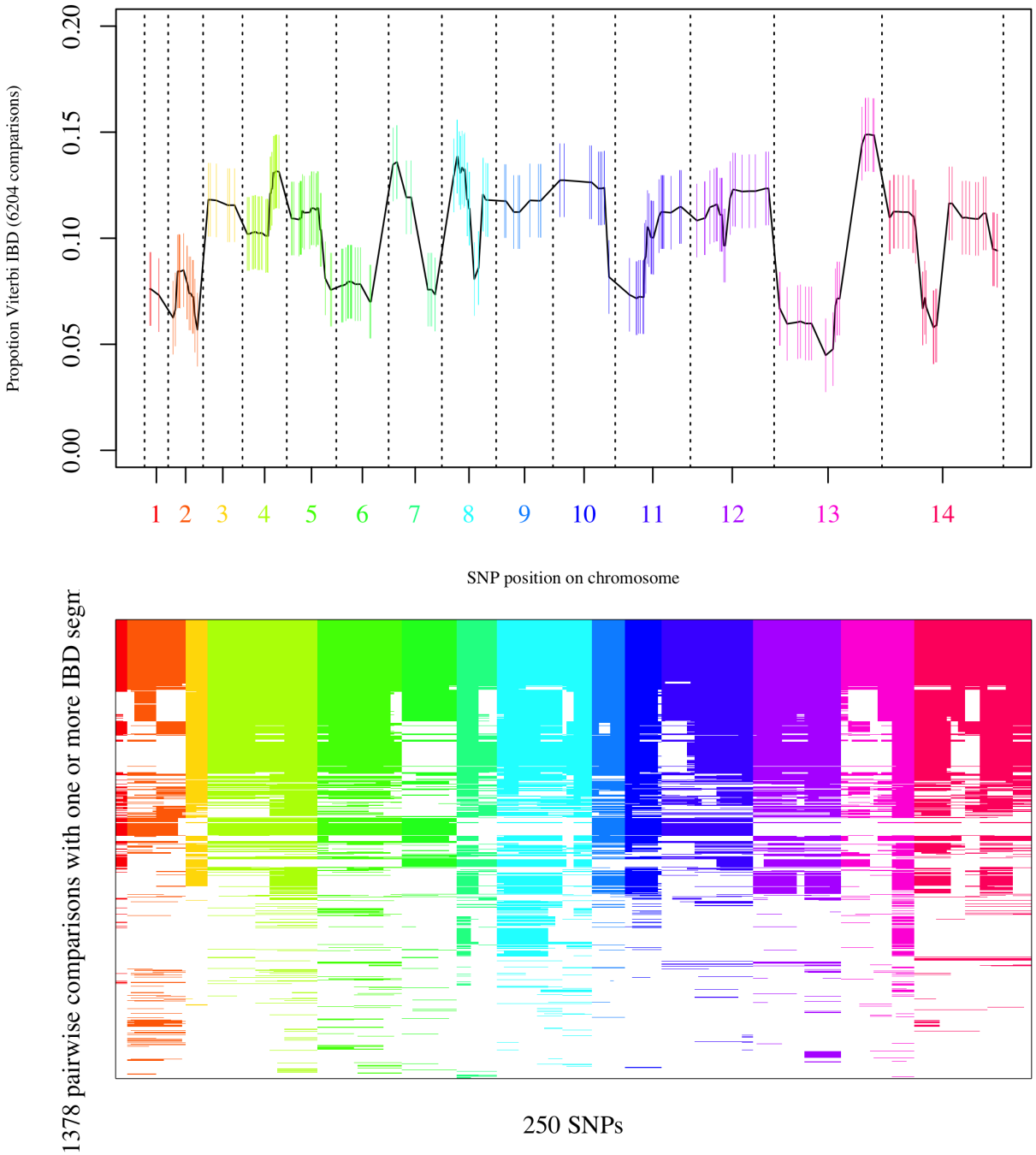


Figure 12: Deep dive into comparisons between Buenaventura and Tumaco. Top plot shows proportion of sample comparisons with IBD segment based on Viterbi path at different positions along the genome. Bottom plot shows heat map of IBD segments (coloured according to chromosome) for those sample comparisons with one or more IBD segement. Sample comparisons are order with respect to decreasing genominc fraction IBD estimates.

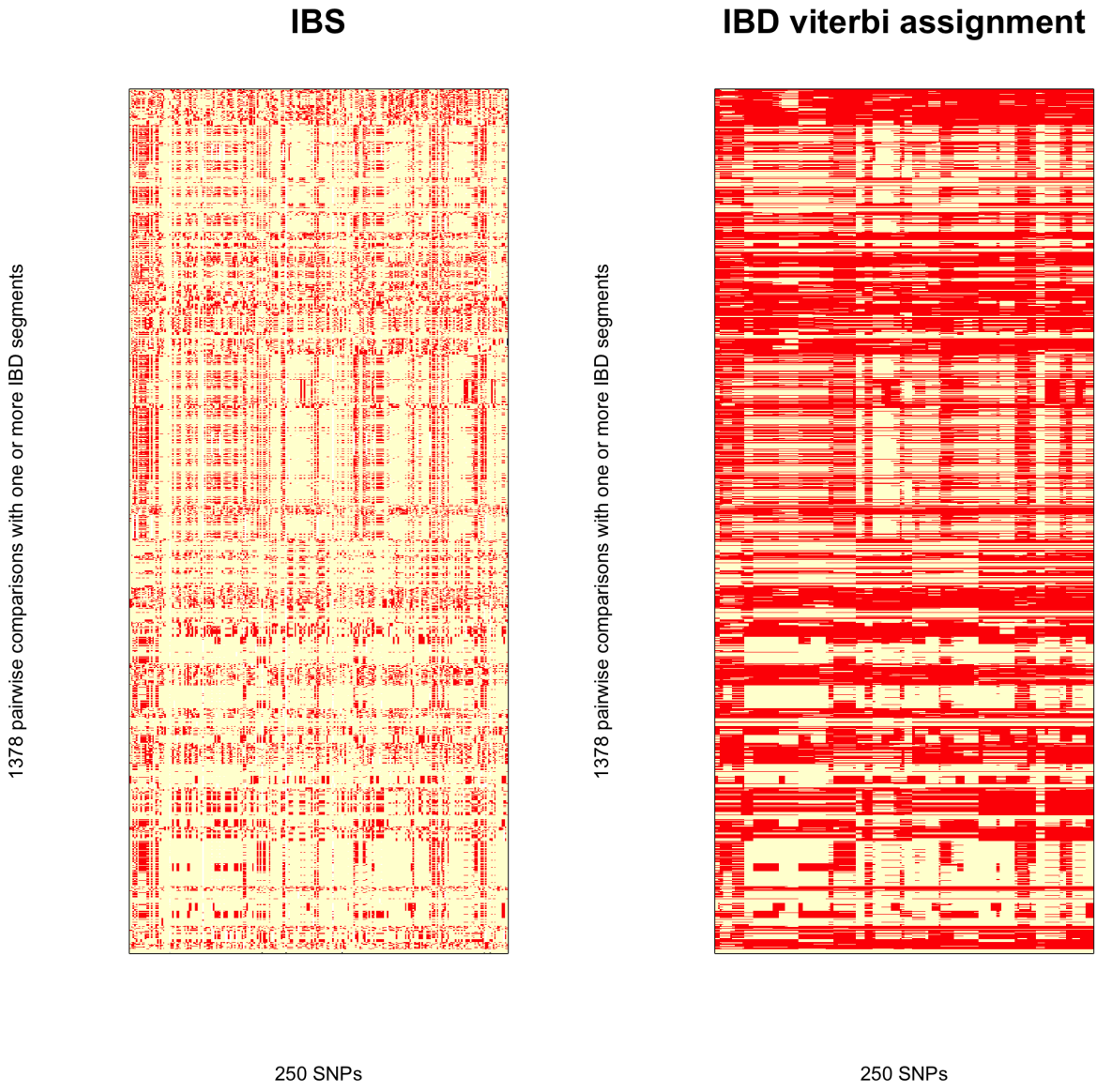


Figure 13: Comparison of IBS and IBD viterbi assignments for comparisons between Buenaventura and Tumaco with at one or more IBD assignment. Based on results of most likely sequence under independece (IBS vs IBD project), the fact that the two are different suggests SNPs are not independent.

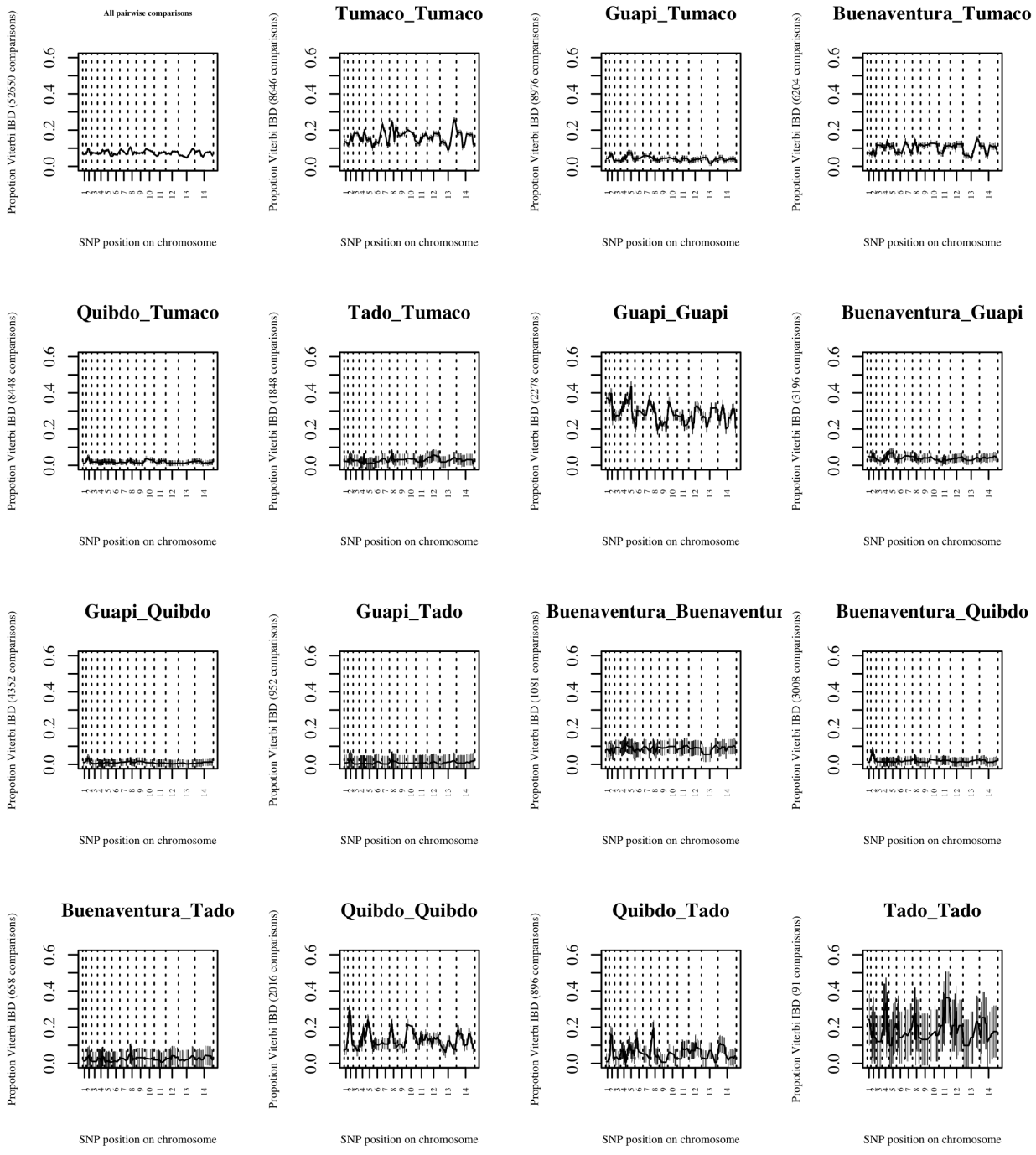


Figure 14: Plots show proportion of sample comparisons with IBD segment based on Viterbi path at different positions along the genome for different sample comparisons. Error bars denote 95 percent confidence intervals and were calculated using Hoeffding‘s inequality.

Regression analyses thus far

In this section we compare six models fit to raw IBD estimates (using a ordinary linear model, OLM) and dichotomized IBD estimates (using a generalised linear model, GLM). This six models are,

- Model 1: ~ geo_dist + time_dist
- Model 2: ~ geo_dist * time_dist
- Model 3: ~ geo_dist + time_dist + Tumaco + Guapi + Buenaventura + Quibdo + Tado
- Model 4: ~ geo_dist * time_dist + Tumaco + Guapi + Buenaventura + Quibdo + Tado
- Model 5: ~ geo_dist + time_dist + Within
- Model 6: ~ geo_dist + time_dist * Within

GLMs

The glm models without interaction terms (1,3 and 5) are preferable to those with (2,4,6). Of those without interaction terms, model 3 has the most favourable AIC. In terms of binned residules, model 1 is poor; models 3 and 5 are comparable, but many unexplained anomalies remain (Figure 15). If we fit to across site comparisons only using model 1 (models 3 and 5 irrelvant when intra site comparisons are removed), regression coefficients remain significant and negative (AIC is not comparable since data are different); there are more binned residual outliers, however. We thus henceforth discount models fit to partial data.

OLM

All the lm models (1 to 6) seem to be viable candidates, with more as-expected coefficient signs and significant coefficients than those generated under the GLMs. According to the residual squared error, M3 seems to be the best model. It is far from fitting the linear model assumptions, however (Figure 16).

Conclusion

To conclude, model 3 appears to be the best of all linear models and GLMs (regression coefficients, Table 3). Neither fit the data perfectly however (the violations of the linear model seem worse, but this is arguably because of a lack of plots for the GLM). I suspect the addition of travel distance, will improve model fit.

Table 3: Regression coefficient estimates with p-values in parentheses.

	Generalised linear model	Ordinary linear model
(Intercept)	-2.86e+00 (0.00e+00)	8.58e-02 (4.21e-275)
geo_dist	-9.21e-04 (2.52e-05)	-8.42e-05 (2.64e-34)
time_dist	-3.29e-03 (8.85e-48)	-8.97e-05 (2.17e-35)
TumacoTRUE	1.37e+00 (1.37e-77)	9.15e-02 (1.03e-196)
GuapiTRUE	2.00e+00 (1.28e-130)	2.16e-01 (0.00e+00)
BuenaventuraTRUE	9.35e-01 (4.17e-13)	2.78e-02 (9.59e-06)
QuibdoTRUE	6.28e-01 (1.09e-08)	9.64e-02 (3.44e-88)
TadoTRUE	6.99e-01 (5.05e-02)	1.51e-01 (1.09e-13)

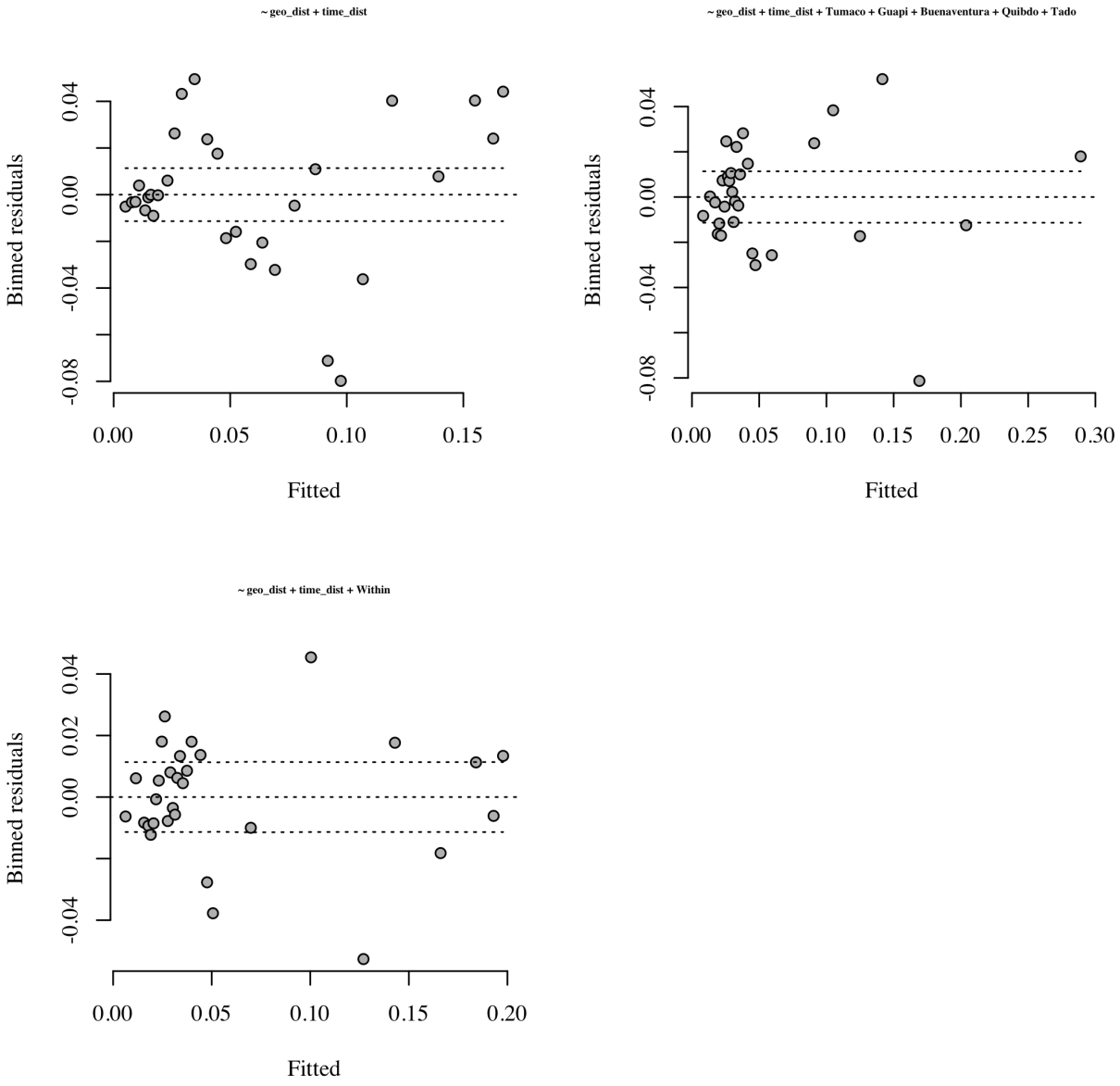


Figure 15: Plots of binned residuals for candidate models fit to dichotomized IBD estimates.

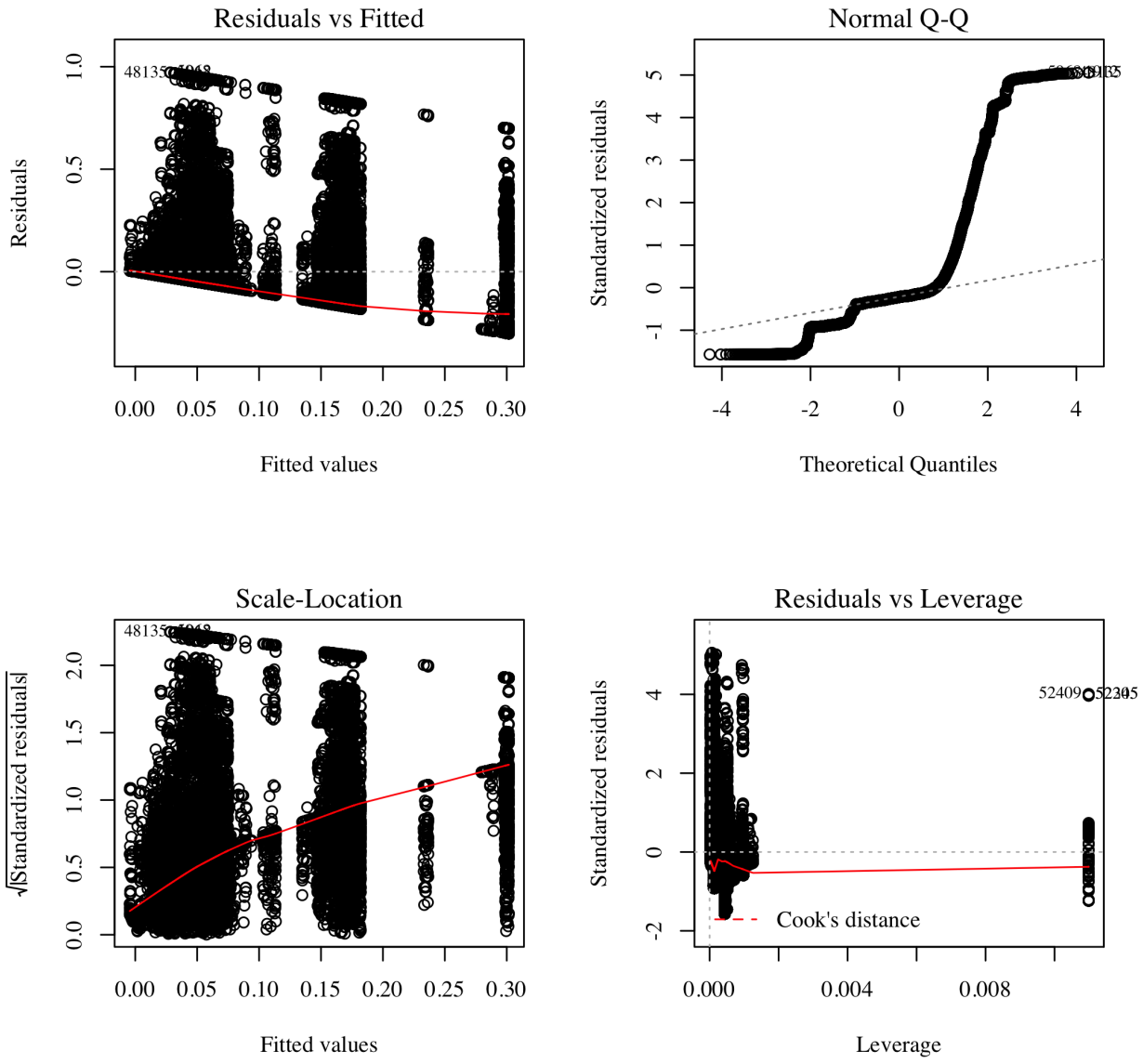


Figure 16: Plots of candidate linear least square models fit.