Multipanel multicountry comparison

Introduction

The panels compared in this R markdown script include the GTseq panel (`GTseq`); the GTseq panel with CSP, TRAP, SERA2 and AMA1 removed (`GTseq_notCTSA`); a panel using only CSP, TRAP, SERA2 and AMA1 (`onlyCTSA`); and a panel based on the Sanger barcode (`sanger_barcode`). The marker and frequency data for `GTseq` and thus the data for panels derived from it (`GTseq_notCTSA` and `onlyCTSA`) are attached as example data along with the **paneljudge** package. The Sanger barcode data used to generate the results plotted in this R markdown script were processed by `./Process_Sanger_Barcode.R`. The relatedness results based on all four panels mentioned above were generated by `./Generate_results_multipanel_comparison.R`.

The frequencies of the GTseq panel and its derivations are based on sample data collected in Colombia, French Guiana, Senegal and Mali. At present, the frequencies of the Sanger panel are based on sample data collected in Colombia, French Guiana and Senegal, but not Mali. Sample data are not included in this example analysis nor the distribution of the **paneljudge** package.

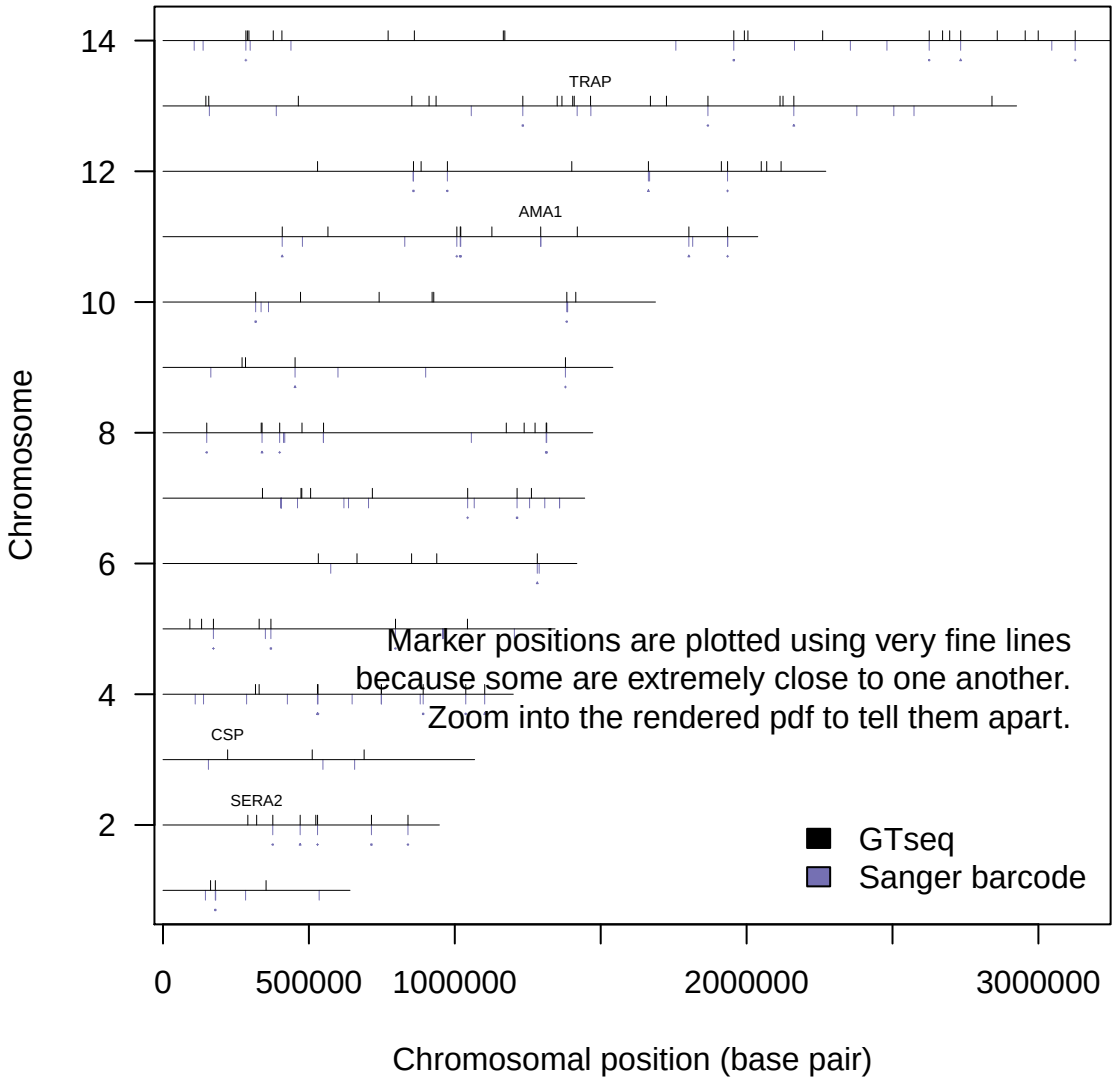
There is substantial overlap (code and text) between this Rmd file and the paneljudge vignette. To see the code, please open `multipanel_multicountry.Rmd`.

```
#> Loading required package: RColorBrewer
#> Loading required package: kableExtra
#> Warning: package 'kableExtra' was built under R version 3.5.2
#> Loading paneljudge
```

Comparative summaries and results

Marker counts and positions

For relatedness inference, panels with many, evenly spaced and highly diverse markers are informative (1). The example panel based on GTseq has 126 markers distributed across 14 of 14 chromosomes; the Sanger barcode panel has 118 markers distributed across 14 of 14 chromosomes:



Data on the GTseq panel are generated by sequencing amplicons spanning regions of the genome that highly diverse due to nearby single nucleotide polymorphisms (SNPs), which we call these regions microhaplotypes. These lengths are too small to perceive on the above plot of the microhaplotype marker positions, and some are very close together; see the paneljudge vignette for a summary of the distributions of region lengths and inter-region distances (minimum inter-region distance in the GTseq panel 295.5).

Data on the Sanger barcode are generated by sequencing amplicons spanning regions containing individual SNPs [is that true?]. Some of the Sanger barcode SNPs are so close to one another (minimum inter-region distance in the Sanger barcode panel 295.5) that they could be sequenced together as microhaplotypes versus individual SNPs. We do not have length information for the Sanger barcode [does it exist?]. However, GTseq regions contain 0 to 1 Sanger markers, totaling 44 across all 126 regions (they are marked below by a point whose symbol varies to facilitate distiction between markers that are extremely close to one another).

Otherwise stated, there appears to be substantial overlap between `GTseq` and `sanger_barcode`. Despite very close proximity to TRAP and AMA1, there are 0 sanger barcodes markers within the CTSA regions of the GTseq panel.

Marker diversities and effective cardinalities

Please see the vignette of the package **paneljudge** for a summary of how to interpret diversities and effective cardinalities and the function documentation (accessed via e.g. `?compute_eff_cardinalities()`) for details on how they are computed.

Table 1: Diversities GTseq

	Colombia	French Guiana	Mali	Senegal
Min.	0.00	0.00	0.00	0.00
1st Qu.	0.04	0.03	0.42	0.40
Median	0.41	0.22	0.52	0.53
Mean	0.33	0.25	0.52	0.51
3rd Qu.	0.50	0.44	0.70	0.70
Max.	0.74	0.70	0.93	0.95

Table 2: Diversities GTseq without CTSA

	Colombia	French Guiana	Mali	Senegal
Min.	0.00	0.00	0.00	0.00
1st Qu.	0.04	0.03	0.41	0.36
Median	0.40	0.20	0.52	0.52
Mean	0.33	0.24	0.51	0.49
3rd Qu.	0.50	0.44	0.69	0.67
Max.	0.74	0.70	0.93	0.93

Table 3: Diversities CTSA only

	Colombia	French Guiana	Mali	Senegal
Min.	0.04	0.19	0.76	0.82
1st Qu.	0.35	0.38	0.87	0.86
Median	0.48	0.49	0.91	0.91
Mean	0.41	0.45	0.88	0.89
3rd Qu.	0.53	0.56	0.91	0.94
Max.	0.64	0.63	0.93	0.95

Table 4: Diversities Sanger Barcode

	Colombia	French Guiana	Senegal
Min.	0.00	0.00	0.00
1st Qu.	0.01	0.03	0.28
Median	0.36	0.20	0.45
Mean	0.27	0.23	0.41
3rd Qu.	0.45	0.44	0.50
Max.	0.69	0.71	0.91

Table 5: Effective cardinalities GTseq

	Colombia	French Guiana	Mali	Senegal
Min.	1.00	1.00	1.00	1.00
1st Qu.	1.04	1.03	1.72	1.67
Median	1.69	1.29	2.08	2.11
Mean	1.70	1.47	2.82	2.84
3rd Qu.	2.00	1.80	3.31	3.30
Max.	3.92	3.28	13.95	20.33

Table 6: Effective cardinalities GTseq without CTSA

	Colombia	French Guiana	Mali	Senegal
Min.	1.00	1.00	1.00	1.00
1st Qu.	1.04	1.03	1.71	1.55
Median	1.66	1.26	2.07	2.09
Mean	1.69	1.46	2.59	2.53
3rd Qu.	2.00	1.77	3.19	3.08
Max.	3.92	3.28	13.70	14.34

Table 7: Effective cardinalities CTSA only

	Colombia	French Guiana	Mali	Senegal
Min.	1.04	1.23	4.21	5.42
1st Qu.	1.65	1.65	8.68	7.24
Median	1.92	1.96	10.64	11.91
Mean	1.92	1.97	9.86	12.39
3rd Qu.	2.19	2.28	11.82	17.06
Max.	2.80	2.74	13.95	20.33

Table 8: Effective cardinalities Sanger Barcode

	Colombia	French Guiana	Senegal
Min.	1.00	1.00	1.00
1st Qu.	1.02	1.03	1.38
Median	1.57	1.24	1.81
Mean	1.50	1.41	2.03
3rd Qu.	1.83	1.78	1.99
Max.	3.22	3.50	10.83

Table 9: Top 3 per country GTseq

	eff_cardinality	Diversity
Colombia.PF3D7_0827100	3.92	0.74
Colombia.P38	3.50	0.71
Colombia.PF3D7_1473700	3.31	0.70
French Guiana.PF3D7_1035700	3.28	0.70
French Guiana.PF3D7_1302900	3.17	0.68
French Guiana.PF3D7_0501800	3.12	0.68
Mali.zAMA1	13.95	0.93
Mali.PF3D7_1352900	13.70	0.93
Mali.zTRAP	11.10	0.91
Senegal.zCSP	20.33	0.95
Senegal.zAMA1	15.97	0.94
Senegal.PF3D7_1352900	14.34	0.93

Table 10: Top 3 per country GTseq not CTSA

	eff_cardinality	Diversity
Colombia.PF3D7_0827100	3.92	0.74
Colombia.P38	3.50	0.71
Colombia.PF3D7_1473700	3.31	0.70
French Guiana.PF3D7_1035700	3.28	0.70
French Guiana.PF3D7_1302900	3.17	0.68
French Guiana.PF3D7_0501800	3.12	0.68
Mali.PF3D7_1352900	13.70	0.93
Mali.P38	6.82	0.85
Mali.PF3D7_0615900	6.16	0.84
Senegal.PF3D7_1352900	14.34	0.93
Senegal.P38	6.06	0.84
Senegal.PF3D7_1475900	5.62	0.82

Table 11: Top 3 per country sanger barcode

	eff_cardinality	Diversity
Colombia.Pf3D7_14_v3_3126219	3.22	0.69
Colombia.Pf3D7_11_v3_408668	2.86	0.65
Colombia.CRT_371	2.83	0.65
French Guiana.Pf3D7_05_v3_350933	3.50	0.71
French Guiana.Pf3D7_04_v3_1102392_E808D_A	2.48	0.60
French Guiana.Pf3D7_06_v3_1282691_803K_A	2.27	0.56
Senegal.Pf3D7_13_v3_1466422	10.83	0.91
Senegal.Pf3D7_11_v3_1295068_E405K	9.22	0.89
Senegal.Pf3D7_13_v3_1419519	7.34	0.86

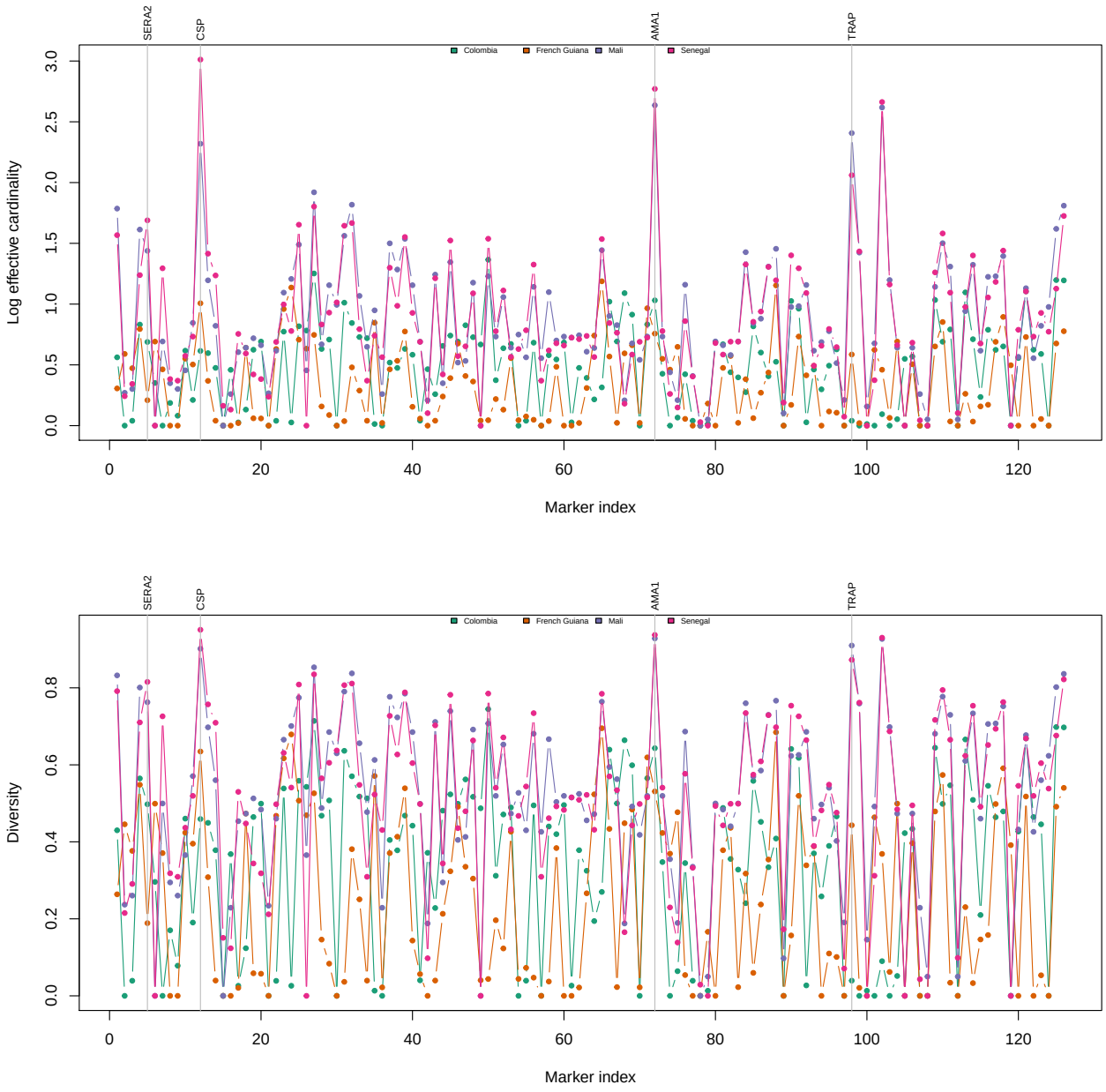


Figure 1: Marker summaries (diversities and effective cardinalities) for the GTseq panel

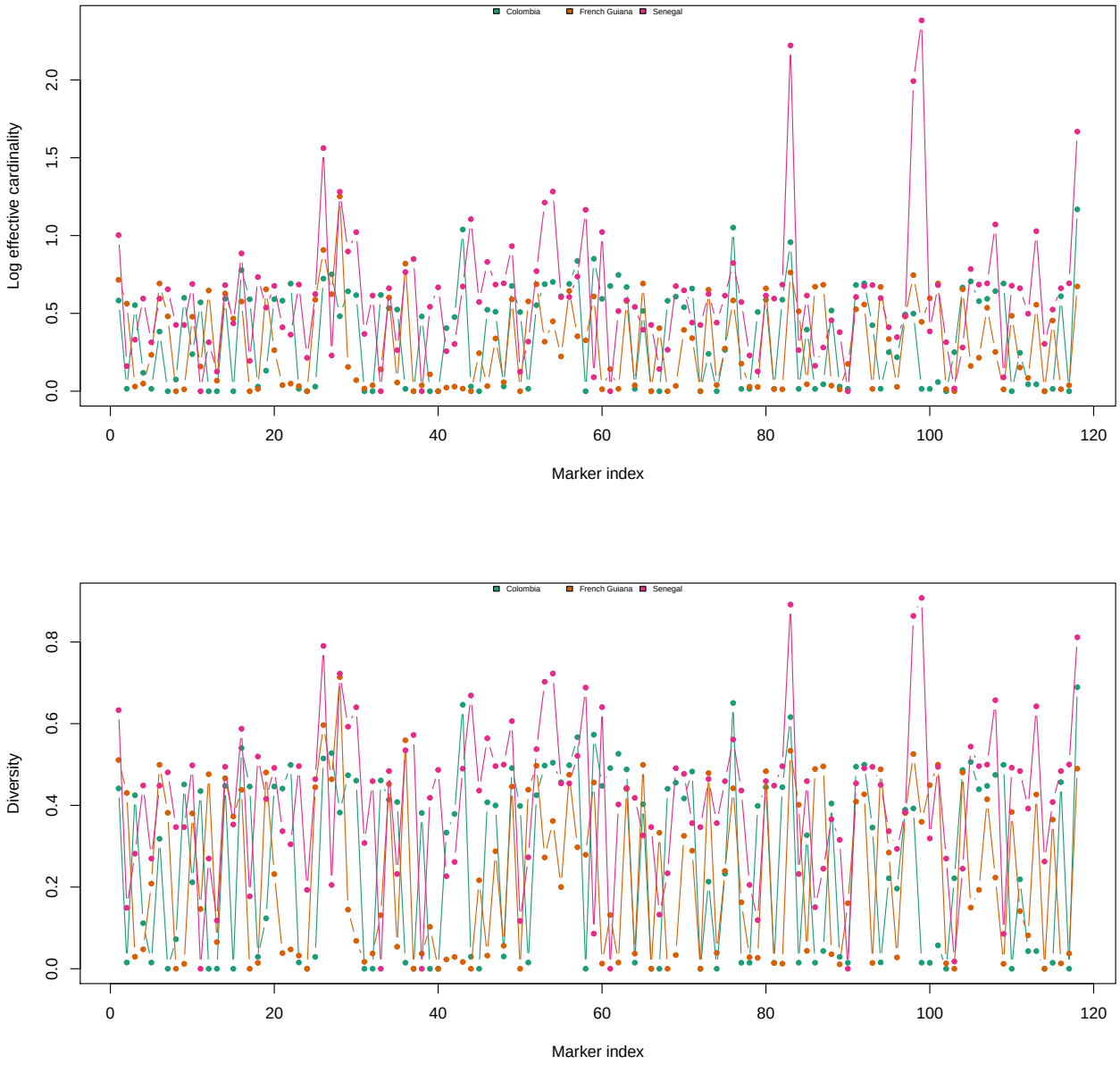


Figure 2: Marker summaries (diversities and effective cardinalities) for the Sanger barcode

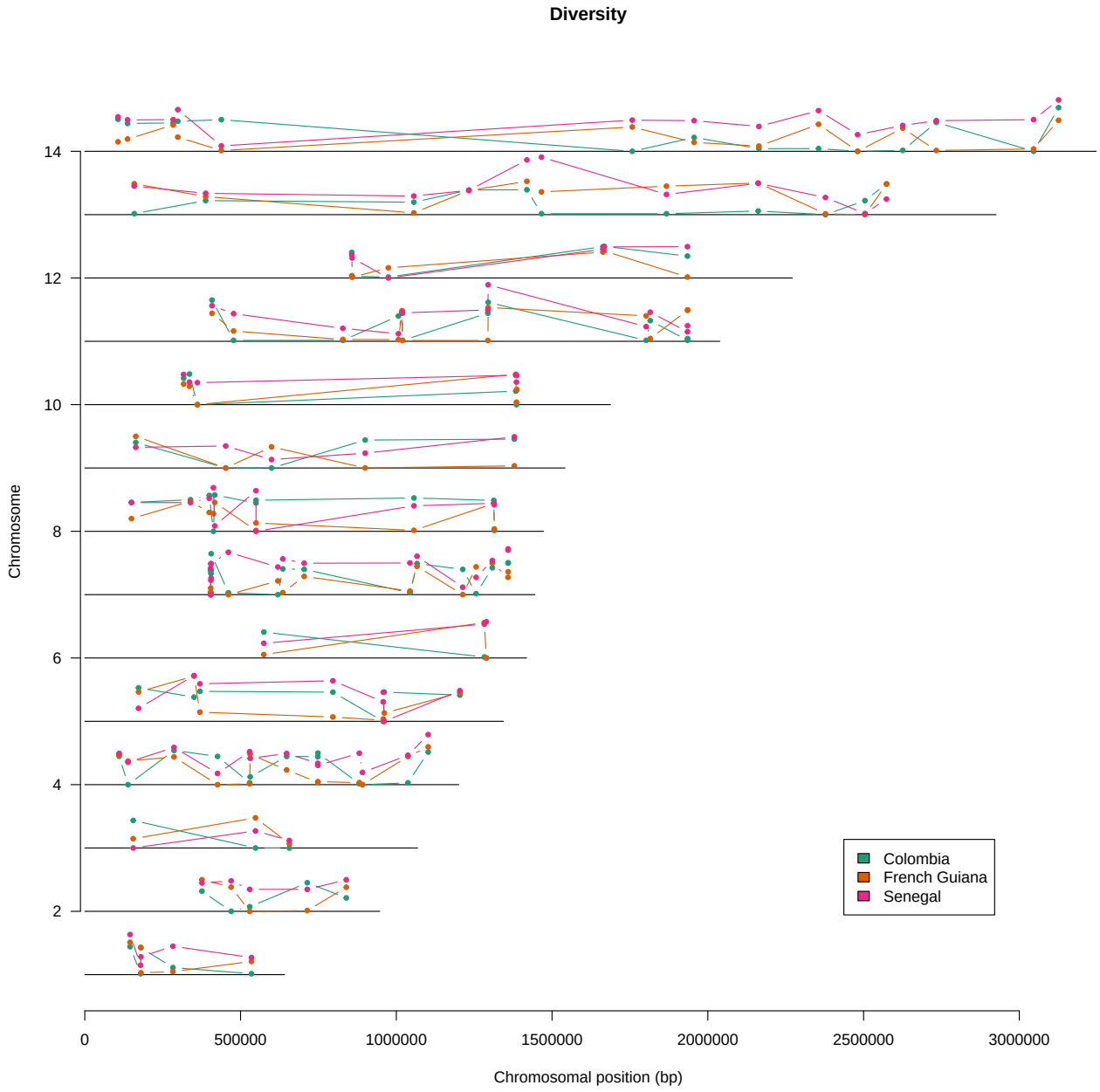
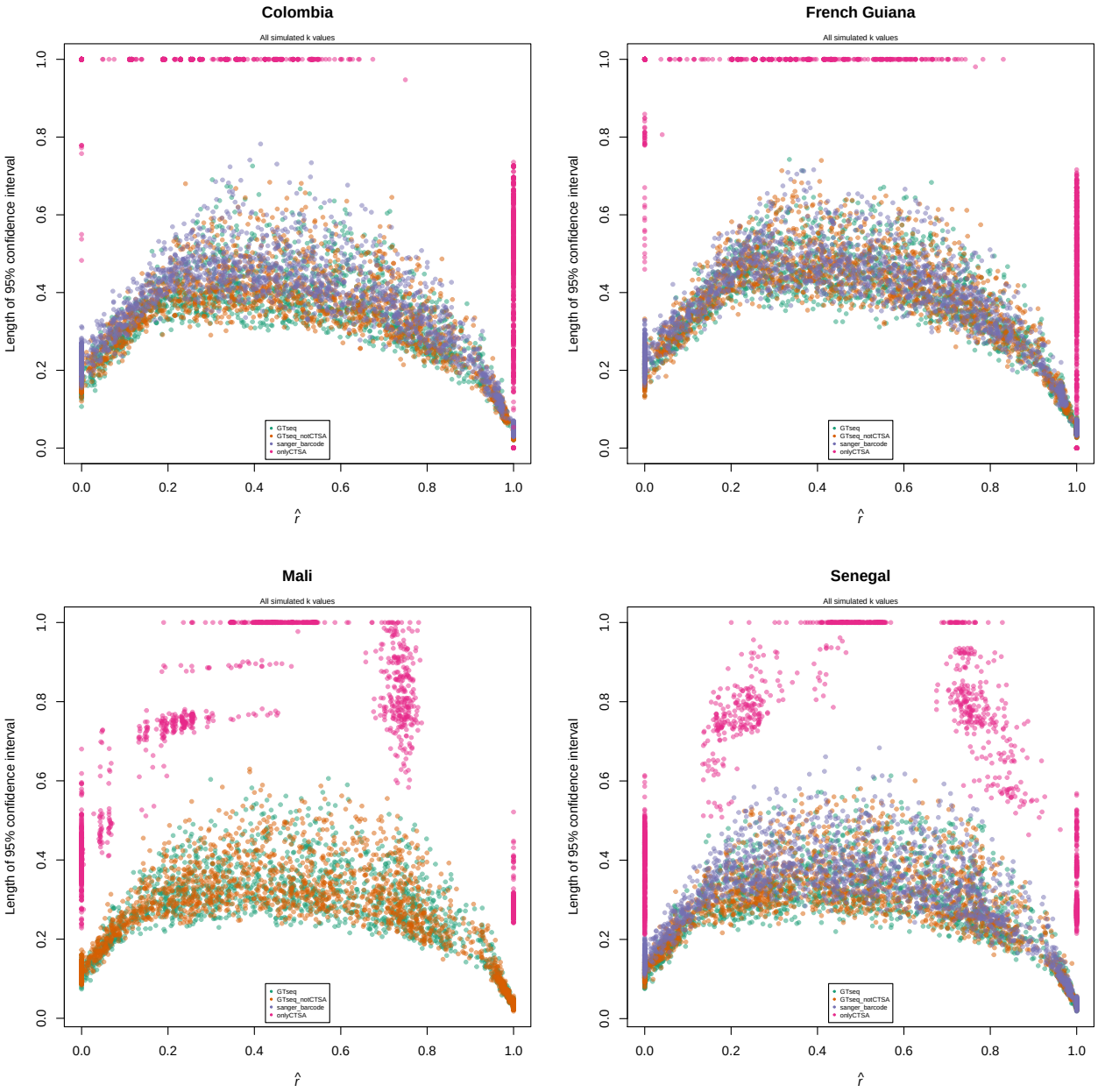
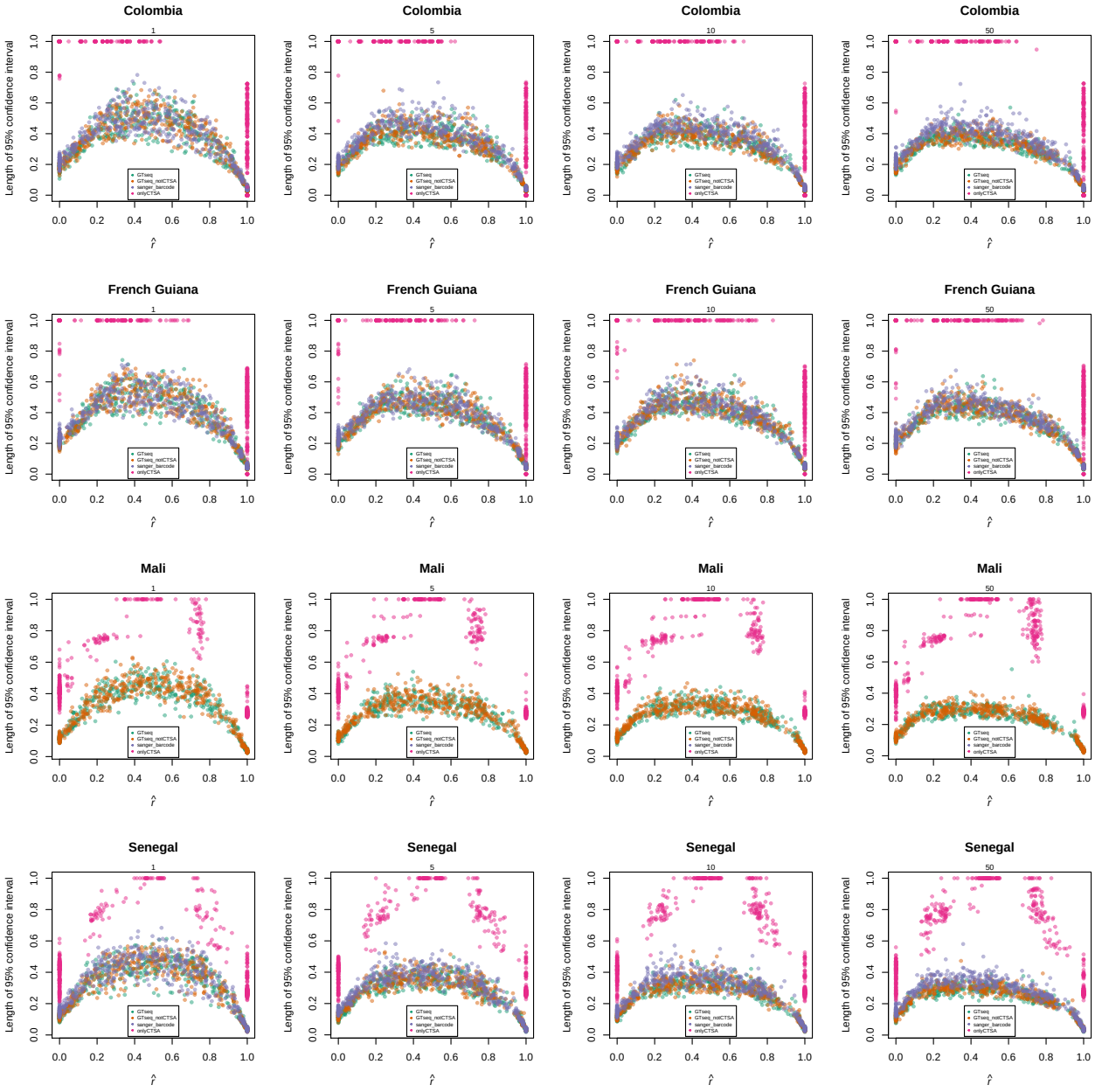
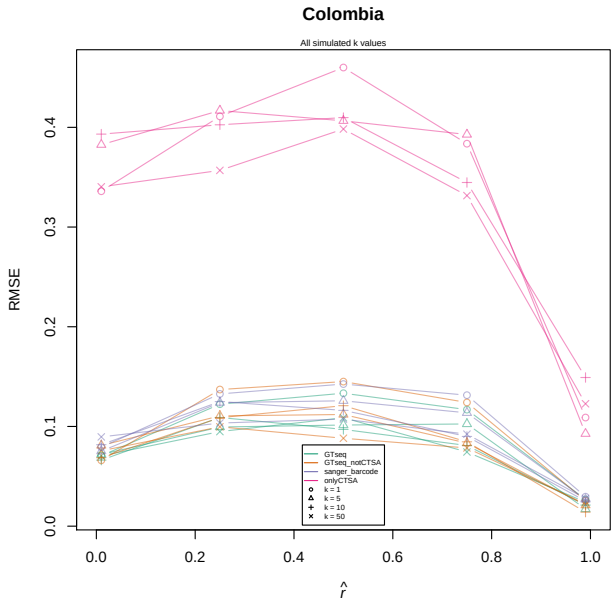


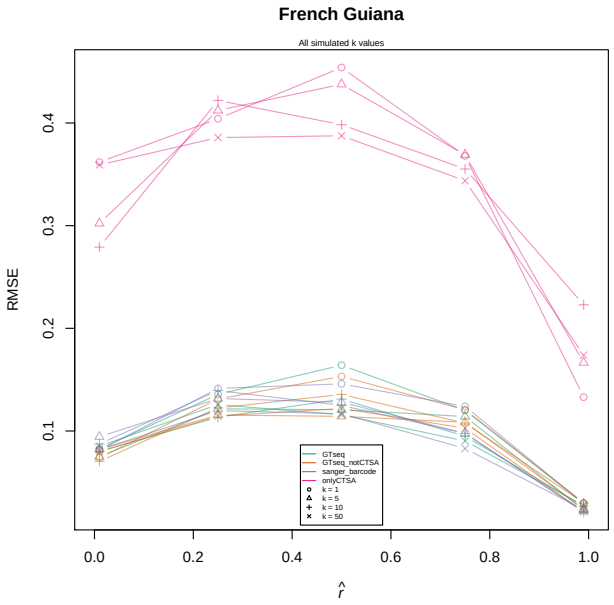
Figure 3: Marker diversities and positions for the GTseq panel

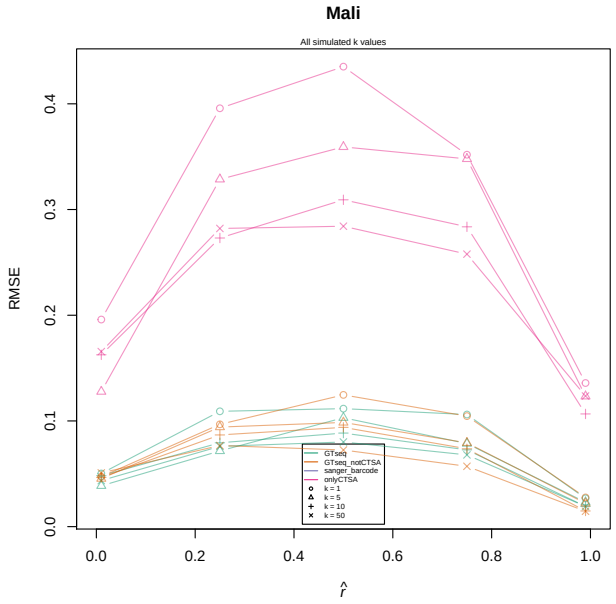
Relatedness estimates and their CIs based on simulated data

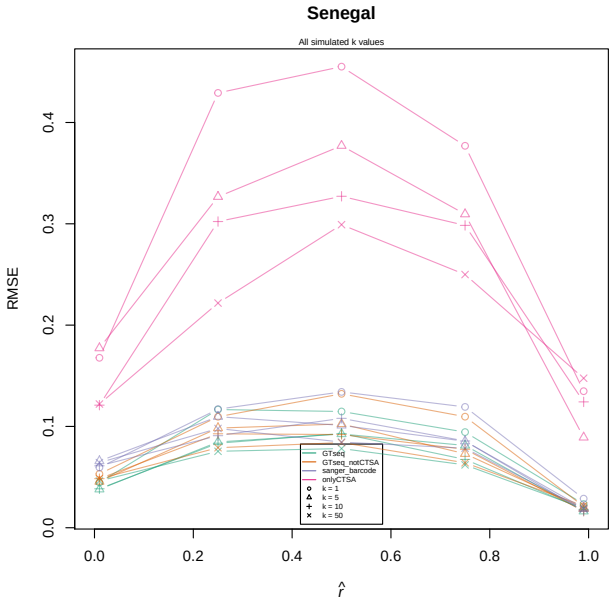


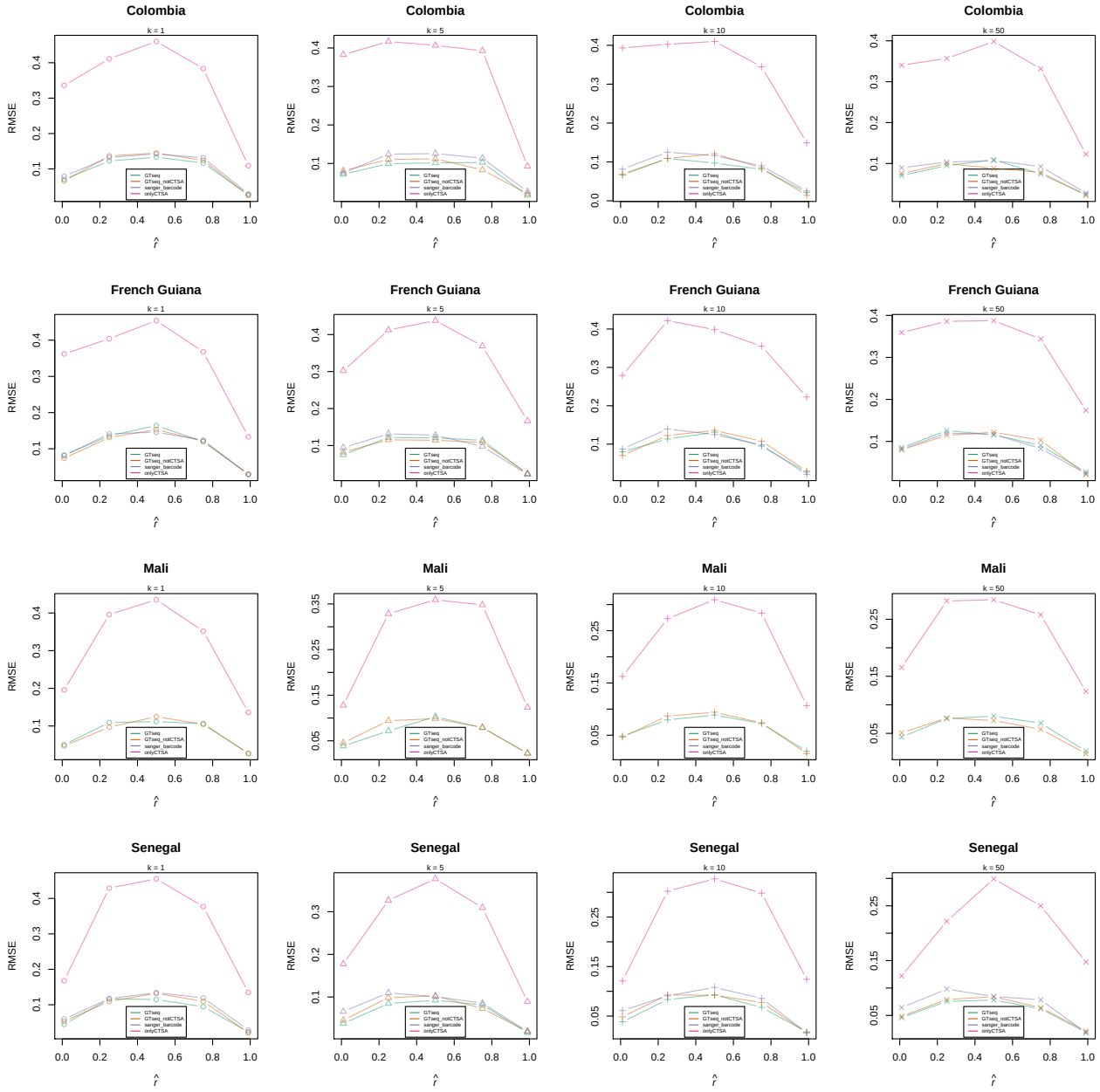

Colombia
1
5
10
50
French Guiana
Mali
Senegal
Length of 95% confidence interval
$\hat{r}$
GTseq
GTseq_notCTSA
sanger_barcode
onlyCTSA

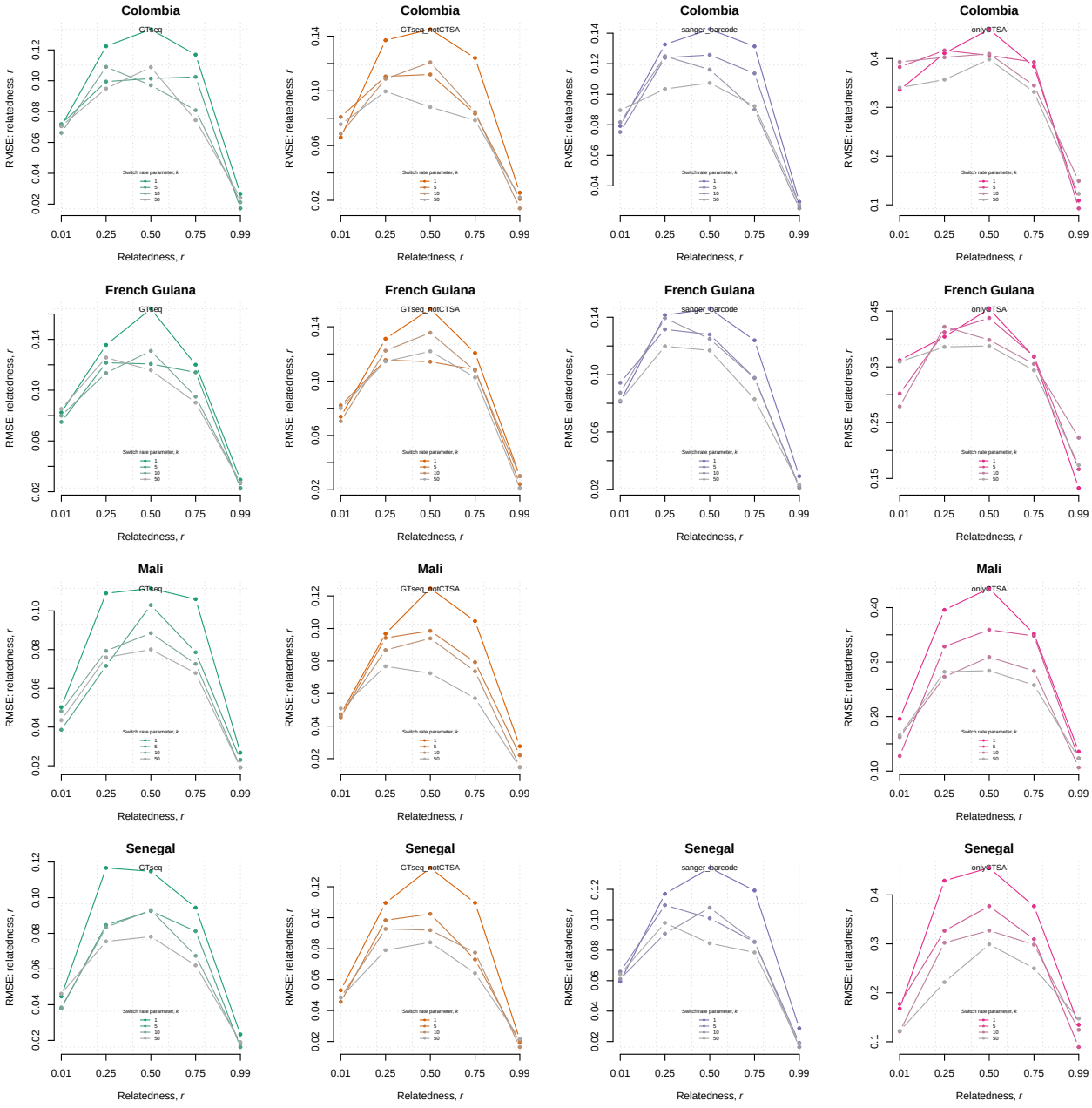

Colombia
All simulated k values
RMSE
r̂
GTseq
GTseq_notCTSA
sanger_barcode
onlyCTSA
k = 1
k = 5
k = 10
k = 50

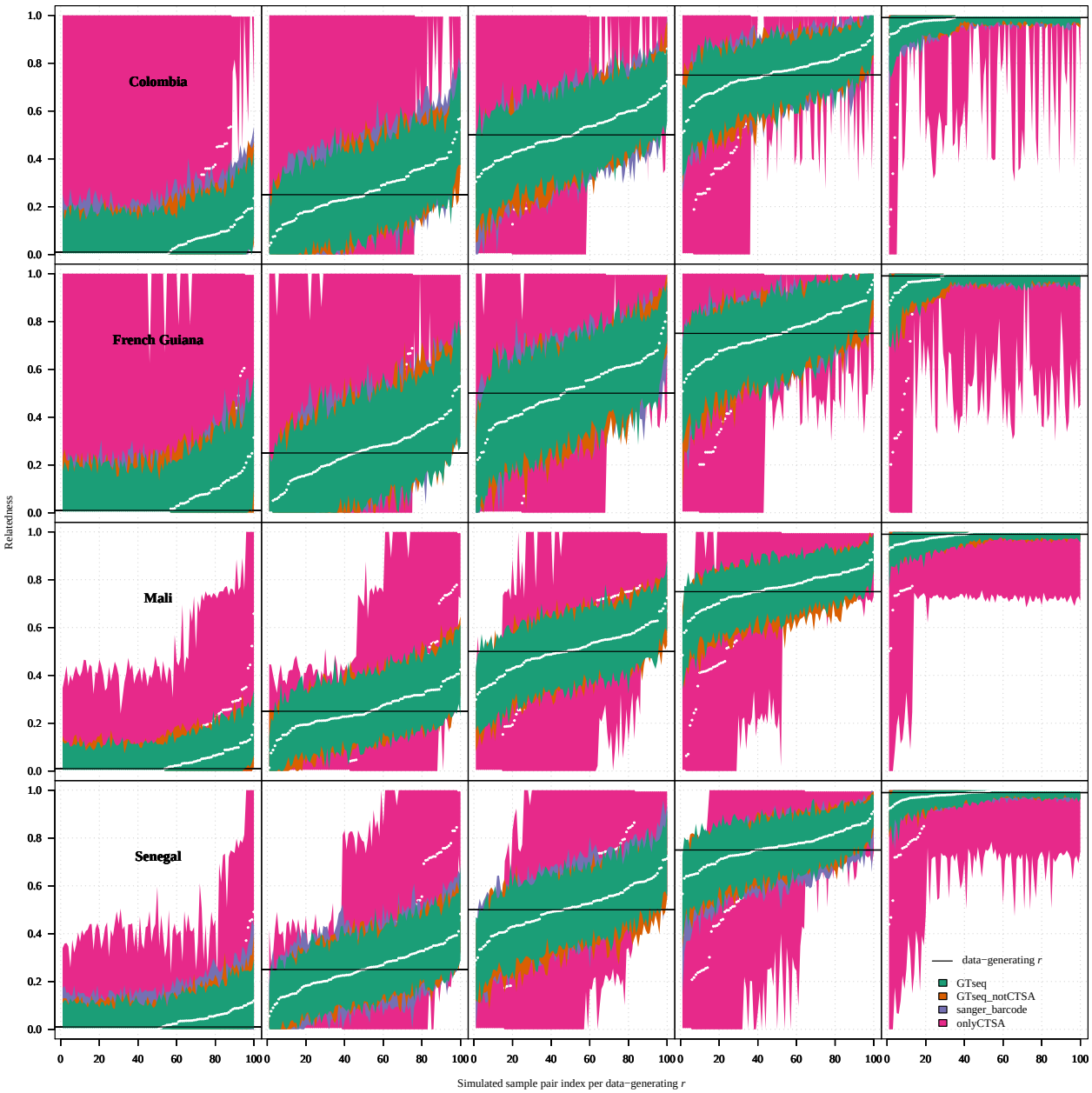

French Guiana
All simulated k values
RMSE
r̂
GTseq
GTseq_notCTSA
sanger_barcode
onlyCTSA
k = 1
k = 5
k = 10
k = 50


Mali
All simulated k values
RMSE
r̂
GTseq
GTseq_notCTSA
sanger_barcode
onlyCTSA
k = 1
k = 5
k = 10
k = 50


Senegal
All simulated k values
RMSE
r̂
GTseq
GTseq_notCTSA
sanger_barcode
onlyCTSA
k = 1
k = 5
k = 10
k = 50


Colombia
k = 1
k = 5
k = 10
k = 50
French Guiana
Mali
Senegal
RMSE
$\hat{r}$
GTseq
GTseq_notCTSA
sanger_barcode
onlyCTSA


Colombia
GTseq
Colombia
GTseq_notCTSA
Colombia
sanger_barcode
Colombia
onlyTSA
French Guiana
GTseq
French Guiana
GTseq_notCTSA
French Guiana
sanger_barcode
French Guiana
onlyTSA
Mali
GTseq
Mali
GTseq_notCTSA
Mali
onlyTSA
Senegal
GTseq
Senegal
GTseq_notCTSA
Senegal
sanger_barcode
Senegal
onlyTSA
RMSE: relatedness, r
Relatedness, r
Switch rate parameter, k
1
5
10
50



Based on the marker count multiplied by the mean effective cardinality (i.e. the effective cardinality summed over all markers), we can get a rough idea of how confidence intervals around relatedness estimates might scale, e.g. from large (least informative) to smaller (more informative), the panels considered here are ordered as follows: onlyCTSA, sanger_barcode, GTseq_notCTSA, GTseq.

References

1. Taylor AR, Jacob PE, Neafsey DE, Buckee CO. Estimating relatedness between malaria parasites. Genetics. 2019;212(4):1337–51.