SUPPLEMENTAL MATERIAL

Supplemental Text

Procedure for reporting ancestry in the PMRP

Participants in the PMRP filled out a questionnaire. The question related to ancestry was: “How would you describe your ancestry or ethnic origin? (You may list more than one if applicable.)” Participants were given the following options: Czech, Dutch, English, French/French-Canadian, German, Irish, Norwegian, Swedish, and Other. The listed nationalities were based on frequency of the prevalence of these groups in the sampled locations in recent census data[11].

MS command to simulate a "continent" of source populations

We used MS embedded in MARKSIM to produce simulated SNP chip data for a hypothetical source continent. The continent comprises five populations, from which we sampled 200 individuals each. We used the following MS command in the MARKSIM *params* file:

```
./ms 1000 1 -t 1 -I 5 200 200 200 200 200 -ma x 225 185 160 140 225 x 190 170 155 185 195 x
195 175 160 175 195 x 210 140 155 175 210 x -T >ms_outfile
```

Swiss Ancestry and SVM performance

Note that all Swiss individuals were grouped together in the West region despite the fact that German (Central region) and Italian (Mediterranean region) ancestry is present in Switzerland. We found that removing Swiss individuals or placing them in another region decreased SVM model performance.

Supplemental Tables

Table S1. Datasets used in this study.

Dataset	**Number of individuals**	**Notes**	**Applications**
$PMRP_{unabridged}$	3,903 (2,009 insular, 1894 admixed)		projected set in projection PCA
$PMRP_{abridged}$	2,001 (544 insular, 1457 admixed)	compared to $PMRP_{unabridged}$, 847 individuals reporting insular German ancestry and 1,055 close relatives were removed	included in combined PCA
$PMRP_{insular}$	2,009 (2,009 insular)	compared to $PMRP_{unabridged}$, 1,894 individuals reporting admixed ancestry were removed	tested self-reports of this set using $SVM_{projection}$; estimate FST and genetic-geograhpic correlation
$PMRP_{insular-abridged}$	544 (544 insular)	compared to $PMRP_{abridged}$, 1,455 individuals reporting admixed ancestry were removed	tested self-reports of this set using $SVM_{combined}$
$POPRES_{europe}$	1,247 (1,247 insular)	compared to the original POPRES sample of Europeans, randomly removed 702 individuals sampled from England and 1300 individuals sampled from Switzerland	training data for $SVM_{combined}$ and $SVM_{projection;}$ included in combined and projection PCA; estimate F_{ST} and genetic-geographic correlation

Table S2. Immigration schedule: number of immigrants from each source population each generation.

Generation	# pop A	# pop B	# pop C	# pop D	# pop E
0	400	200	200	150	50
1	200	100	100	50	25
2	100	50	50	25	10
3	50	20	20	10	5
4	25	10	10	2	2
5	10	5	5	2	2
>=6	2	2	2	2	2

Table S3. Results from fitting training data to SVM models from combined and projection PCA and η = 0.1 or 0.4. *fitted MC rate* is misclassification error observed in the fitted training data shown in the table. *est. fitted rate* is misclassification error estimated from 10-fold cross-validation.

(a)

COMBINED, nu = 0.1

	central	british isles	mediterranea	northeast	iberia	scandinavia	southeast	western
central	42	2	0	2	0	2	0	4
british isles	18	228	0	0	0	1	0	7
mediterranea	0	0	176	0	1	0	0	1
northeast	2	0	0	34	0	0	1	0
iberia	0	0	4	0	237	0	0	4
scandinavia	1	0	0	0	0	5	0	0
southeast	0	0	1	5	0	0	48	0
western	23	7	1	0	1	0	0	241
n	86	237	182	41	239	8	49	257
fitted MC rate	0.500	0.038	0.033	0.171	0.008	0.375	0.020	0.062
est. MC rate	0.491	0.036	0.041	0.229	0.012	0.500	0.095	0.071

(b)

COMBINED, nu = 0.2

	central	british isles	mediterranea	northeast	iberia	scandinavia	southeast	western
central	39	1	0	2	0	2	0	2
british Isles	15	207	0	0	0	0	0	4
mediterranea	0	0	158	0	0	0	0	0
northeast	1	0	0	32	0	0	2	0
iberia	0	0	2	0	213	0	0	0
scandinavia	2	1	0	0	0	10	0	0
southeast	0	0	1	3	0	0	39	0
western	19	207	0	0	0	0	0	222
n	*76*	*211*	*161*	*37*	*212*	*12*	*41*	*228*
fitted MC rate	*0.49*	*0.02*	*0.02*	*0.14*	*0*	*0.17*	*0.05*	*0.03*
est. MC rate	*0.45*	*0.02*	*0.02*	*0.15*	*0*	*0.35*	*0.03*	*0.04*

(c)

COMBINED, nu = 0.4

	central	british isles	mediterranea	northeast	iberia	scandinavia	southeast	western
central	36	0	0	0	0	2	0	1
british isles	9	157	0	0	0	0	0	0
mediterranea	0	0	121	0	0	0	0	0
northeast	0	0	0	28	0	0	0	0
iberia	0	0	0	0	159	0	0	0
scandinavia	0	0	0	0	0	4	0	0
southeast	0	0	0	0	0	0	33	0
western	13	0	0	0	0	0	0	170
n	58	157	121	28	159	6	33	171
fitted MC rate	0.379	0.000	0.000	0.000	0.000	0.333	0.000	0.006
est. MC rate	0.372	0.000	0.006	0.000	0.000	0.500	0.017	0.004

(d)

PROJECTION, nu = 0.1

	central	british isles	mediterranea	northeast	iberia	scandinavia	southeast	western
central	51	0	0	2	0	5	0	4
british isles	16	231	0	0	0	1	0	6
mediterranea	0	0	177	0	0	0	0	1
northeast	2	0	0	35	0	0	3	0
iberia	0	0	0	0	234	0	0	0
scandinavia	0	0	0	0	0	5	0	0
southeast	0	0	0	6	0	0	51	0
western	15	5	2	0	0	0	0	247
n	84	236	179	43	234	11	54	258
fitted MC rate	0.393	0.021	0.011	0.186	0.000	0.545	0.056	0.043
est. MC rate	0.443	0.033	0.010	0.272	0.000	0.667	0.070	0.048

(e)

PROJECTION, nu = 0.2

	central	british isles	mediterranea	northeast	iberia	scandinavia	southeast	western
central	54	0	0	1	0	5	0	2
british Isles	10	207	0	0	0	0	0	0
mediterranea	0	0	159	0	0	0	0	0
northeast	0	0	0	32	0	0	3	0
iberia	0	0	0	0	208	0	0	0
scandinavia	0	0	0	0	0	6	0	0
southeast	0	0	0	5	0	0	44	0
western	10	3	0	0	0	0	0	227
n	*74*	*210*	*159*	*38*	*208*	*11*	*47*	*229*
fitted MC rate	*0.27*	*0.01*	*0*	*0.16*	*0*	*0.46*	*0.06*	*0.01*
est. MC rate	*0.29*	*0.02*	*0*	*0.22*	*0*	*0.78*	*0.06*	*0.01*

(f)

PROJECTION, nu = 0.4

	central	british isles	mediterranea	northeast	iberia	scandinavia	southeast	western
central	52	0	0	0	0	2	0	0
british isles	3	157	0	0	0	0	0	0
mediterranea	0	0	119	0	0	0	0	0
northeast	0	0	0	27	0	0	1	0
iberia	0	0	0	0	156	0	0	0
scandinavia	0	0	0	0	0	5	0	0
southeast	0	0	0	0	0	0	35	0
western	0	0	0	0	0	0	0	171
n	55	157	119	27	156	7	36	171
fitted MC rate	0.055	0.000	0.000	0.000	0.000	0.286	0.028	0.000
est. MC rate	0.067	0.000	0.000	0.056	0.000	0.333	0.050	0.000

Table S4. Wisconsin ancestry prediction from projection and combined PCA and η = 0.1 or 0.4. Misreported is the fraction of n estimated to be misreports after correction for the MCE of the model. Results for η = 0.2 are in Table 1 of the main text.

(a)

COMBINED, nu = 0.1

	Czech	Dutch	English	French	German	Irish	Norwegian	Polish	Swedish
central	8	4	12	3	162	1	2	5	2
british isles	0	11	63	5	6	29	14	0	1
mediterranean	0	0	0	0	0	0	0	0	0
northeast	21	1	3	2	9	0	0	80	0
iberia	0	0	0	1	0	0	0	0	0
scandinavia	0	4	1	0	9	0	43	0	18
southeast	0	0	0	0	0	0	0	0	0
western	0	2	2	9	12	0	1	0	0
n	29	22	81	20	198	30	60	85	21
mri	8	18	18	11	36	1	17	5	3
MCE	0.229	0.491	0.036	0.071	0.491	0.036	0.5	0.229	0.5
mri(corr)	6.168	9.162	17.352	10.219	18.324	0.964	8.5	3.855	1.5
misreported	0.21268966	0.41645455	0.21422222	0.51095	0.09254545	0.03213333	0.14166667	0.04535294	0.07142857

(b)

COMBINED, nu = 0.4

	Czech	Dutch	English	French	German	Irish	Norwegian	Polish	Swedish
central	3	5	15	4	148	2	3	2	2
british isles	0	11	58	6	4	28	6	0	1
mediterranean	0	0	0	0	0	0	0	0	0
northeast	25	1	4	2	27	0	0	83	0
iberia	0	0	0	1	0	0	0	0	0
scandinavia	1	5	3	0	13	0	50	0	18
southeast	0	0	0	0	0	0	0	0	0
western	0	0	1	7	6	0	1	0	0
n	29	22	81	20	198	30	60	85	21
mri	4	17	23	13	50	2	10	2	3
MCE	0	0.372	0	0.004	0.372	0	0.5	0	0.5
mri(corr)	4	10.676	23	12.948	31.4	2	5	2	1.5
misreported	0.13793103	0.48527273	0.28395062	0.6474	0.15858586	0.06666667	0.08333333	0.02352941	0.07142857

(c)

PROJECTION, nu=0.1

	Czech	Dutch	English	French	German	Irish	Norwegian	Polish	Swedish
central	14	13	23	2	992	4	37	10	15
british isles	1	10	67	8	159	39	34	0	9
mediterranea	0	0	0	1	1	0	0	2	0
northeast	22	0	4	0	143	0	0	102	1
iberia	0	0	0	0	0	0	0	0	0
scandinavia	0	0	1	0	25	0	8	1	1
southeast	1	0	0	0	3	0	0	0	0
western	0	2	5	12	235	1	1	0	0
n	38	25	100	23	1558	44	80	115	26
mri	16	12	33	11	566	5	72	13	25
MCE	0.272	0.443	0.033	0.048	0.443	0.033	0.667	0.272	0.667
mri(corr)	11.648	6.684	31.911	10.472	315.262	4.835	23.976	9.464	8.325
misreported	0.30652632	0.26736	0.31911	0.45530435	0.20235045	0.10988636	0.2997	0.08229565	0.32019231

(d)

PROJECTION, nu =0.4

	Czech	Dutch	English	French	German	Irish	Norwegian	Polish	Swedish
central	14	16	38	7	1059	7	17	5	6
british isles	1	5	56	6	81	37	20	0	5
mediterranea	0	0	0	0	1	0	0	0	0
northeast	23	1	4	0	175	0	0	105	1
iberia	0	0	0	0	0	0	0	0	0
scandinavia	0	3	1	0	132	0	43	5	14
southeast	0	0	0	0	0	0	0	0	0
western	0	0	1	10	110	0	0	0	0
n	38	25	100	23	1558	44	80	115	26
mri	15	9	44	13	499	7	37	10	12
MCE	0.056	0.067	0	0	0.067	0	0.333	0.056	0.333
mri(corr)	14.16	8.397	44	13	465.567	7	24.679	9.44	8.004
misreported	0.37263158	0.33588	0.44	0.56521739	0.29882349	0.15909091	0.3084875	0.08208696	0.30784615

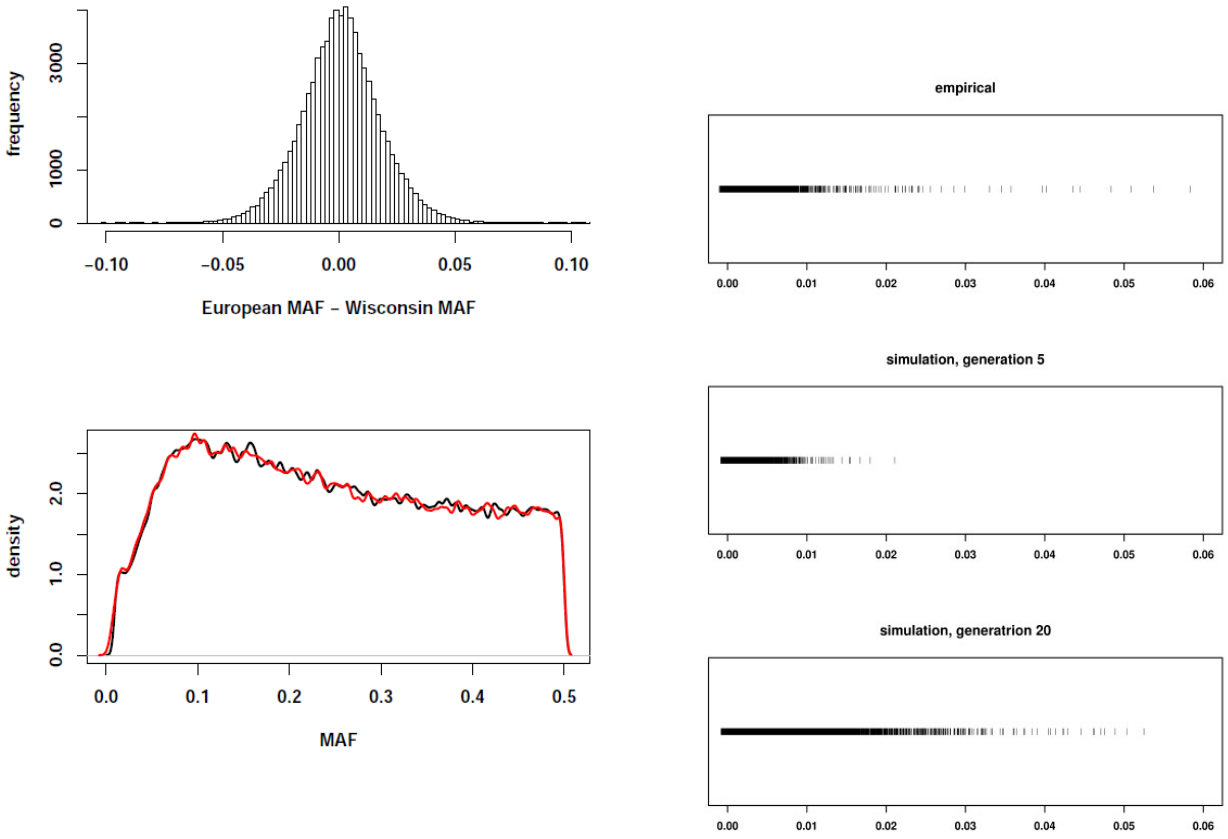

Figure S1. Upper left-hand panel shows the distribution of the European minor allele frequency (MAF) less the Wisconsin MAF for all 75,293 SNPs used in the analyses. Note the symmetrical distribution as well as the absence of extreme values. Lower left-hand panel shows the density estimate of MAFs for the PRMP (red) and POPRES (black) datasets at all 75,293 SNPs. Note the high similarity and obvious ascertainment bias towards high-frequency variants. Upper right-hand panel shows the distribution of single-SNP F_{ST} values for all 75,293 SNPs. Note the lack of extreme value SNPs. Eight SNPs with F_{ST} >0.1 were not included in analyses. Middle and lower right-hand panels show the distribution of F_{ST} values in simulations at 5 and 20 generations post-colonization. Note that the empirical distribution is somewhere in between these two simulated distributions.

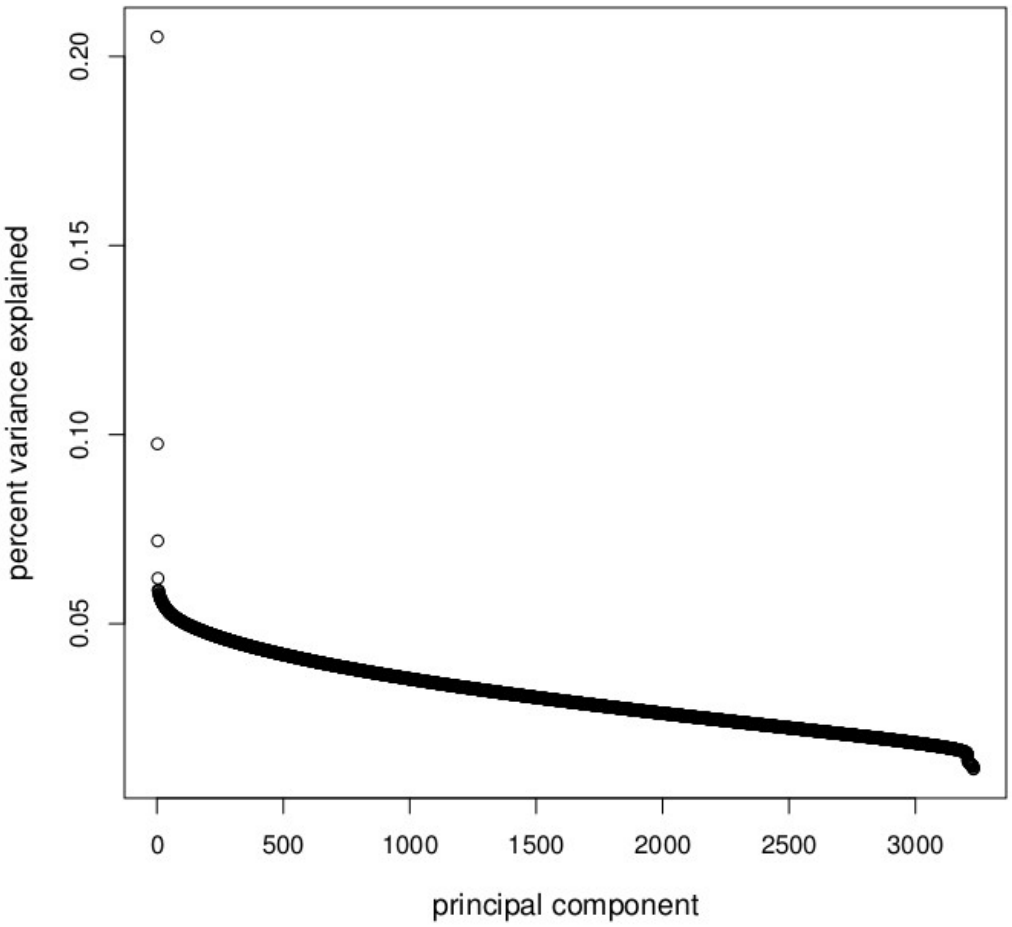

Figure S2. Percent variance explained by each PC under combined PCA.

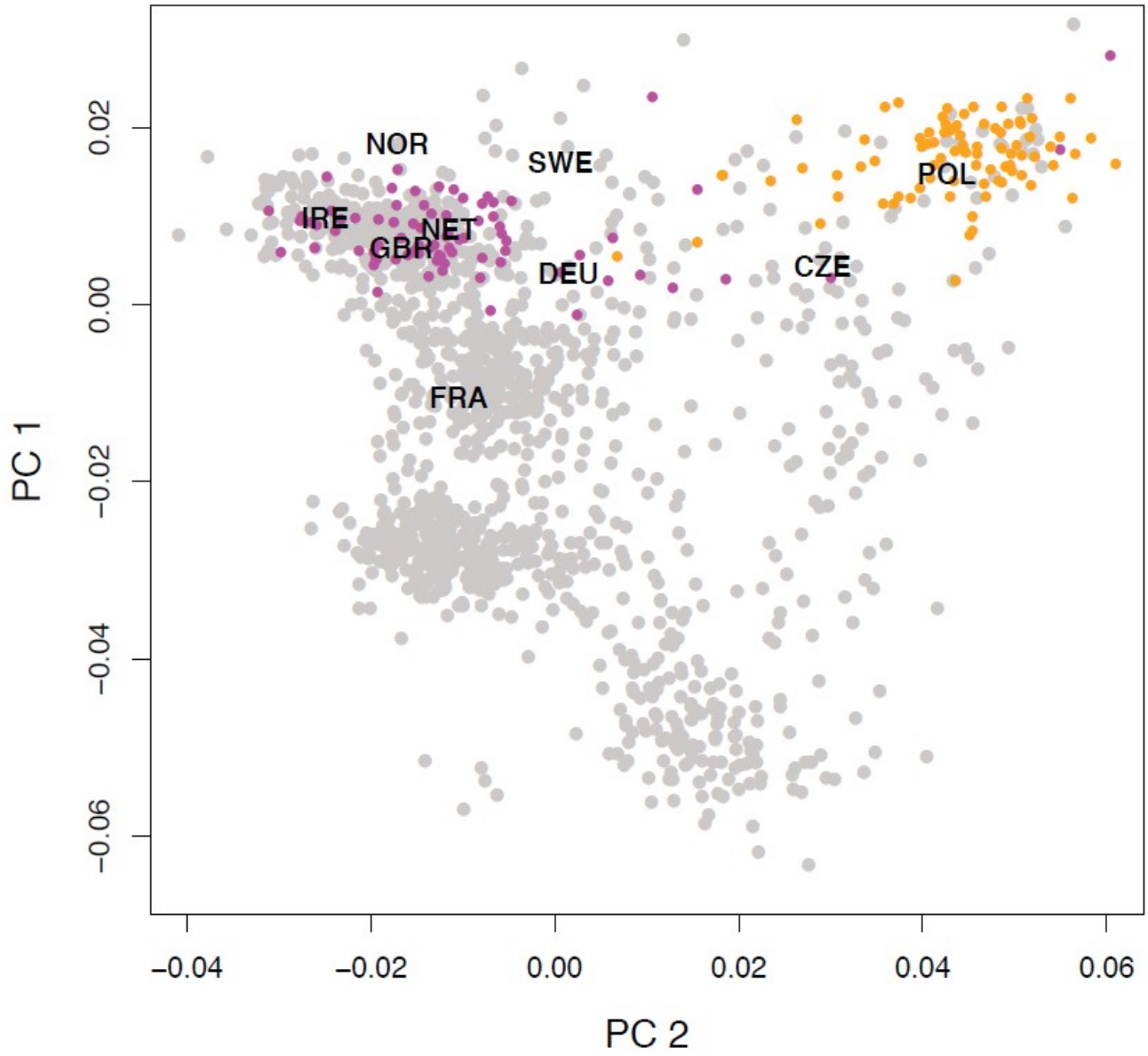

Figure S3. Self-reported ancestries of Wisconsinites (magenta: insular English ancestry reported; orange: insular Polish ancestry repored) form mostly concentrated clusters in appropriate positions upon the underlying map of Europe (gray circles). However, a number of visually obvious outliers are present and separation between Northwestern European countries is minimal. Bold face country codes mark the mean positions of Europeans sampled from those countries (IRE: Ireland; GBR: England; FRA: France; NET: Netherlands; NOR: Norway; SWE: Sweden; DEU: Germany; CZE: Czech Republic; POL: Poland).

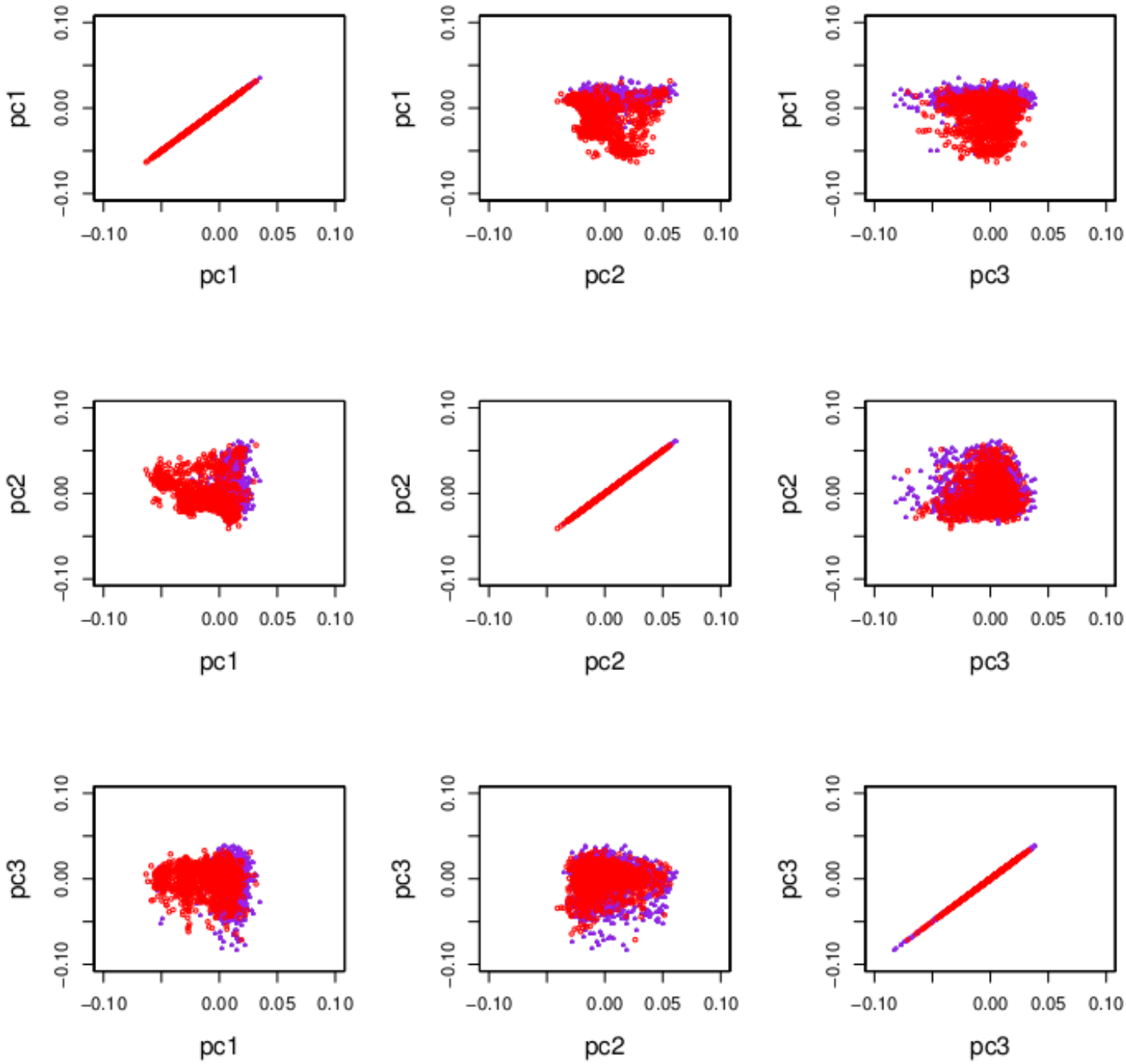

Figure S4. Results from combined PCA; biplots of PCs 1-3. Purple circles: $PMRP_{abridged}$ individuals; red circles: $POPRES_{europe}$ individuals. Higher dispersion of PMRP individuals is not evident in these plots as it is with higher-order PCs.

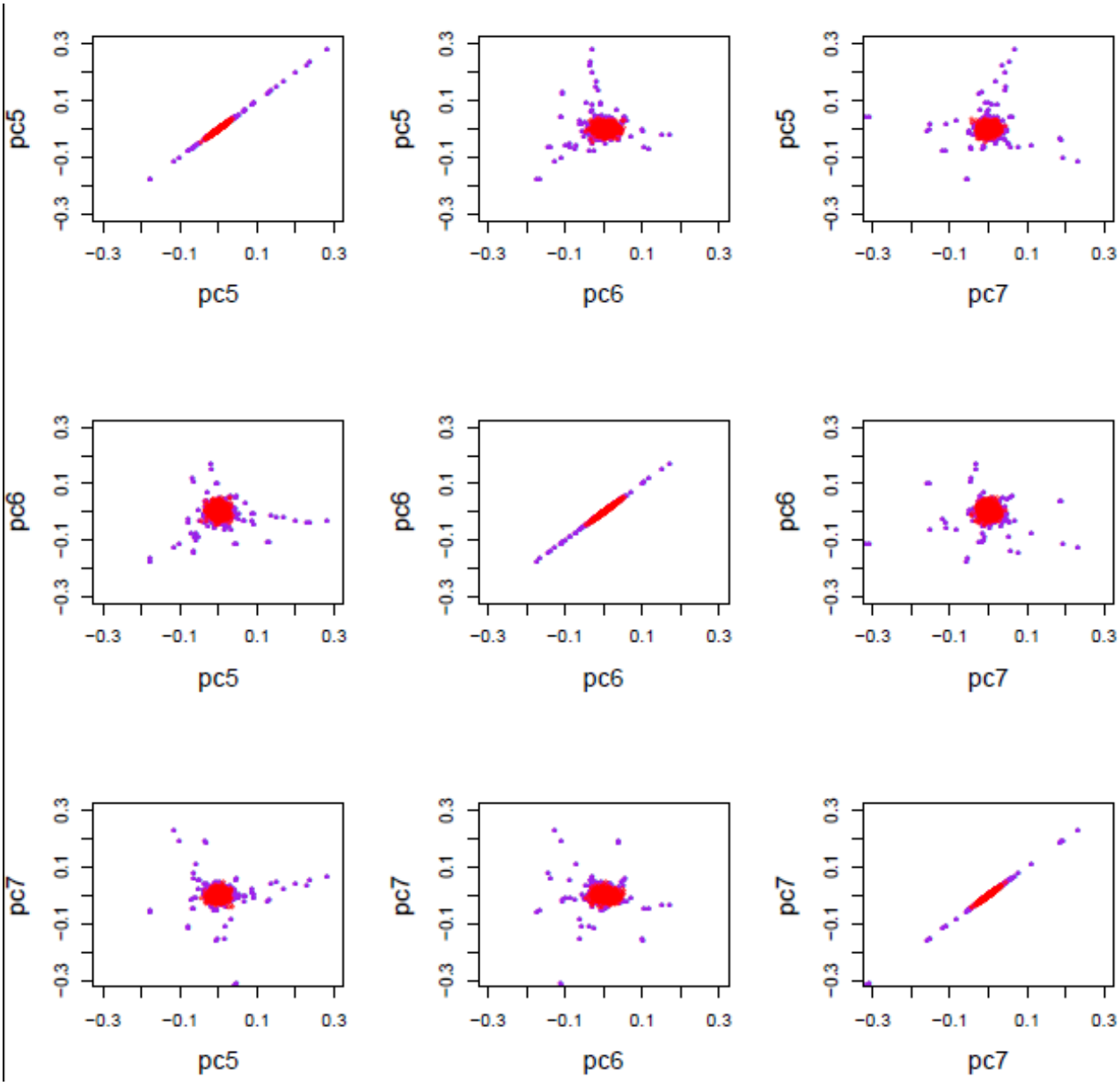

Figure S5. Results from combined PCA; biplots of PCs 4-7. Purple circles: $PMRP_{abridged}$ individuals; red circles: $POPRES_{europe}$ individuals.

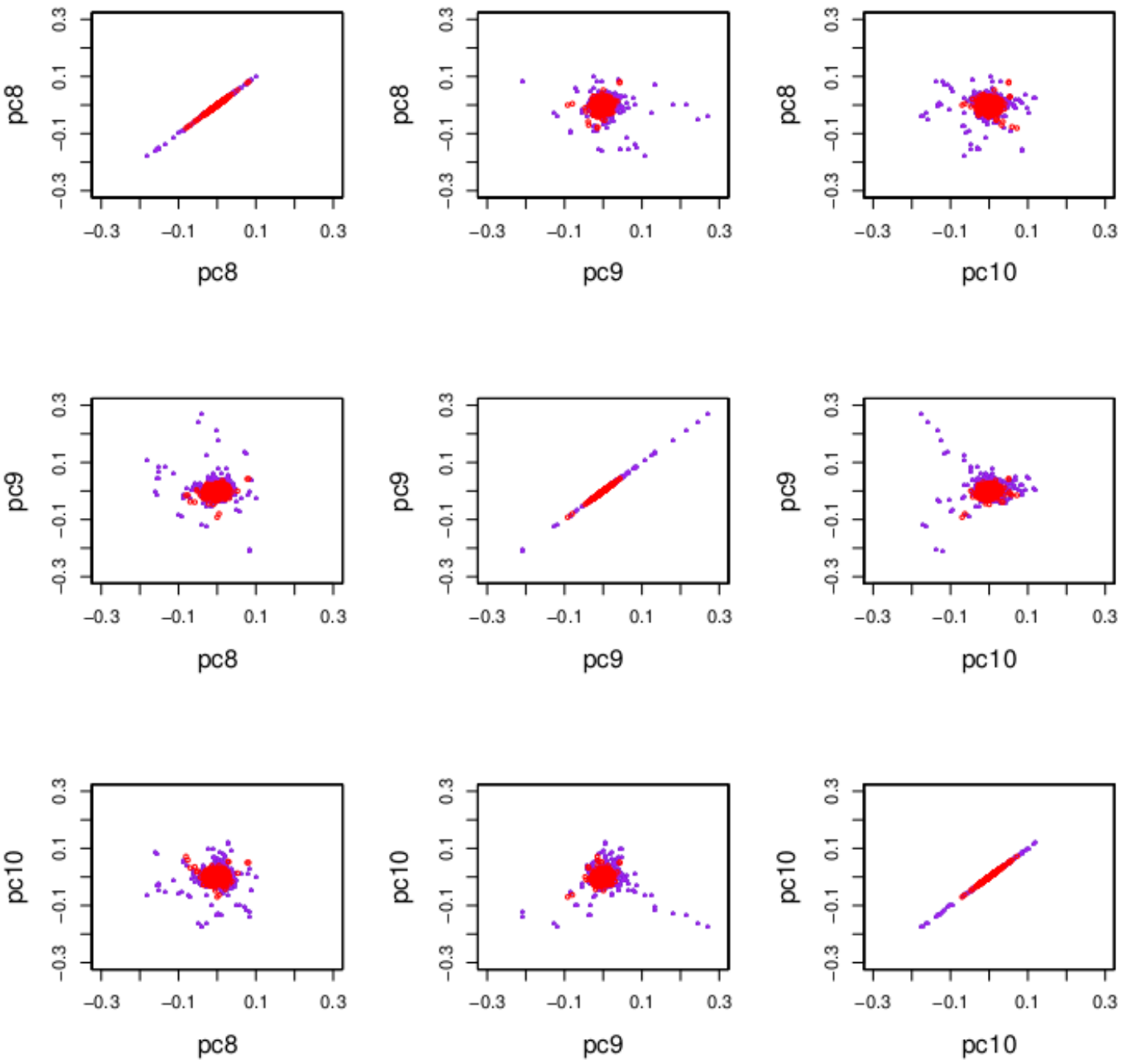

Figure S6. Results from combined PCA; biplots of PCs 8-10. Purple circles: $PMRP_{abridged}$ individuals; red circles: $POPRES_{europe}$ individuals.

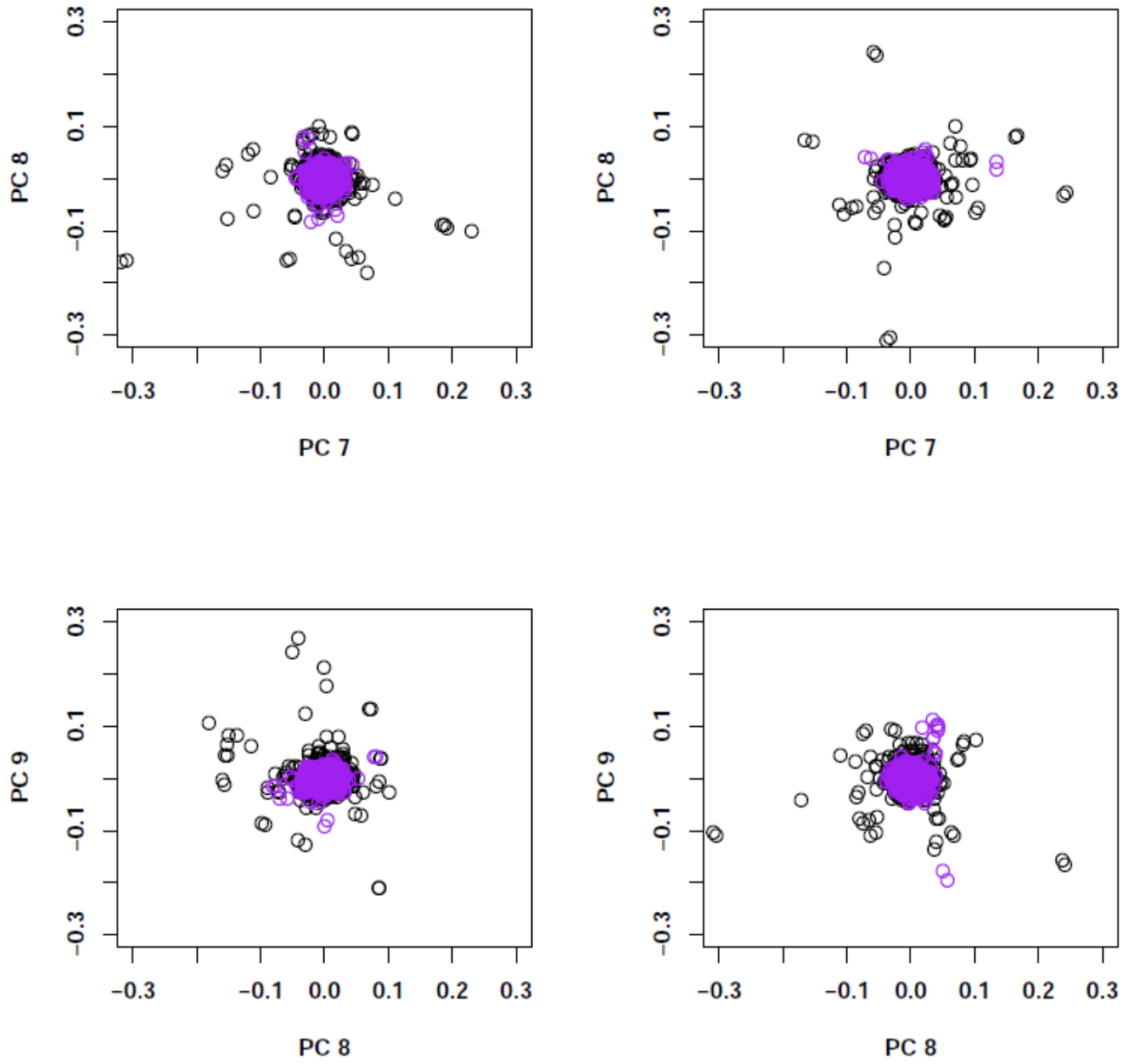

Figure S7. PC7/PC8 and PC8/PC9 biplots from analyses in which the original datasets (left-hand side) were used and analyses where each member of the PMRP dataset had a random 1861 genotypes deleted to equalize missing data rates (right-hand side). PMRP individuals are plotted in black while POPRES individuals are plotted in purple.

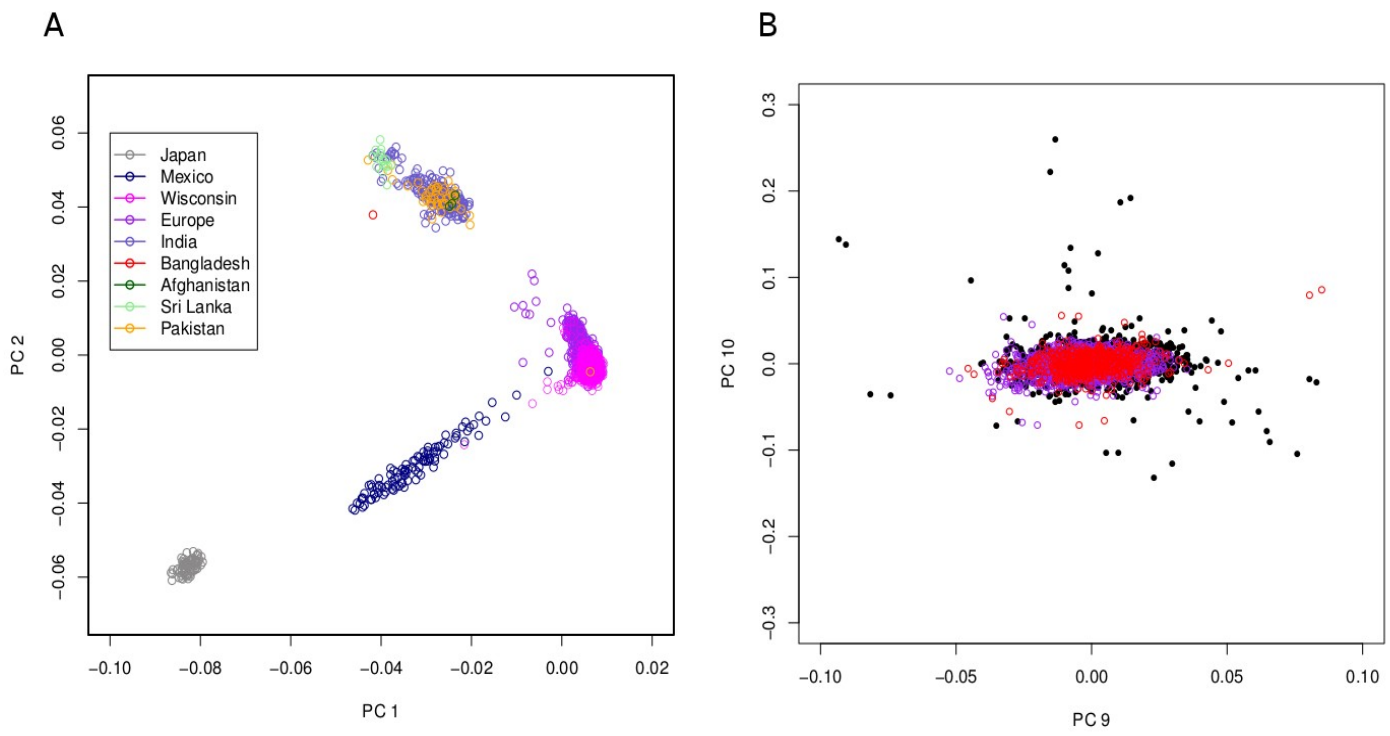

Figure S8. Results from combined PCA of $PMRP_{abridged}$ + $POPRES_{europe}$ + other samples from around the world. (A) On the PC1 / PC 2 biplot PMRP individuals (magenta circles) are coincident with and equally variable to Northern Europeans (purple circles). (B) On the PC 9 / PC 10 biplot, PMRP individuals show greater dispersion ($PMRP_{abridged}$:black circles; $POPRES_{europe}$: purple circles; elsewhere :red circles).

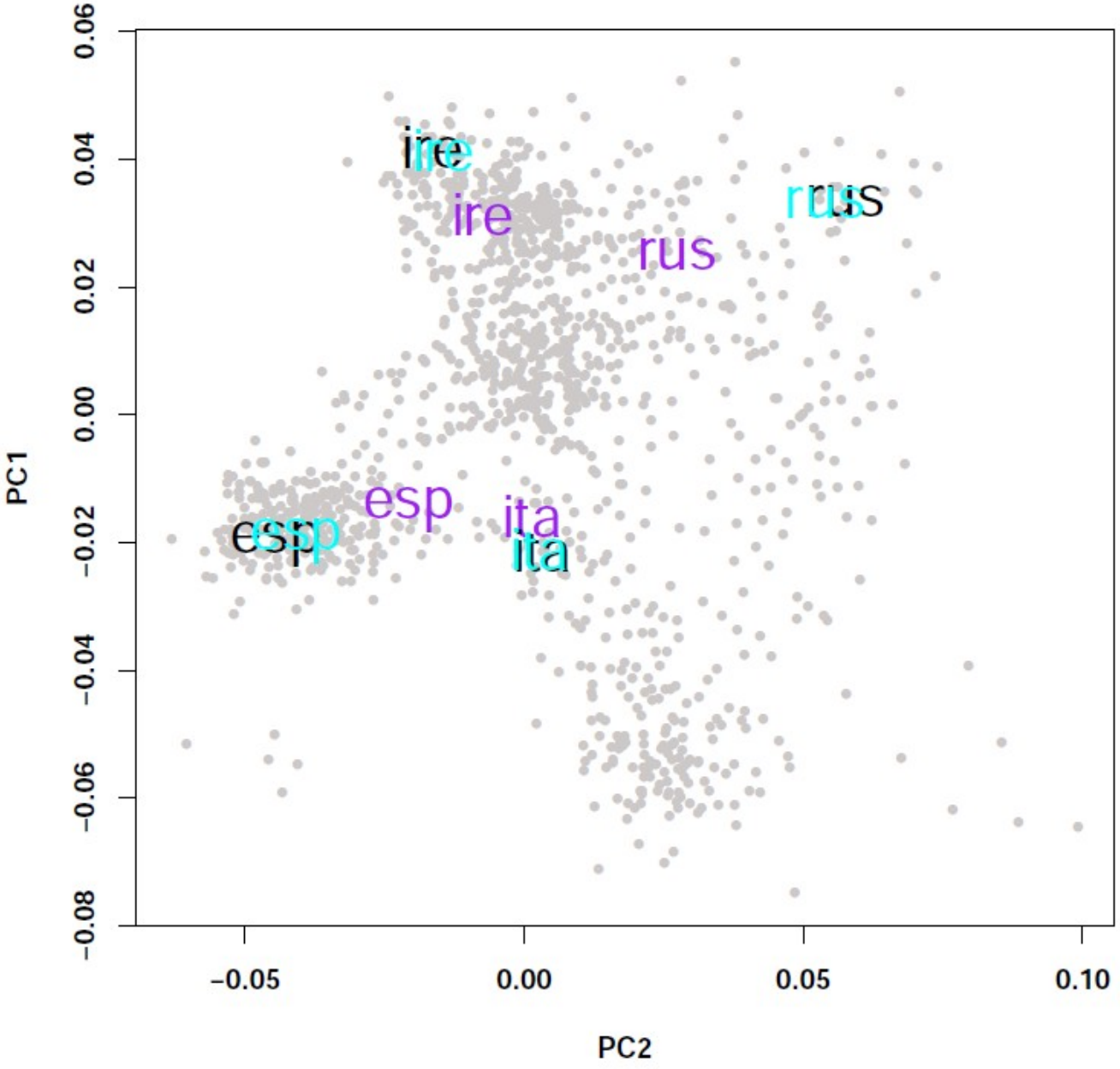

Figure S9. Demonstration of the efficacy of Lee et al's[23] projection bias correction. We performed PCA on $POPRES_{europe}$ alone, but left out four individuals who were projected to the top two PCs (one from Spain, ESP, one from Italy, ITA, one from Russia, RUS, and one from Ireland, IRE). Removal of these individuals did not affect the underlying map of Europe (gray dots). On the figure we plot the positions of the four individuals before bias correction (purple), after bias correction (cyan), and when included in combined PCA (black). As evidenced by the coincidence of cyan and black labels, bias correction largely removes the projection bias.

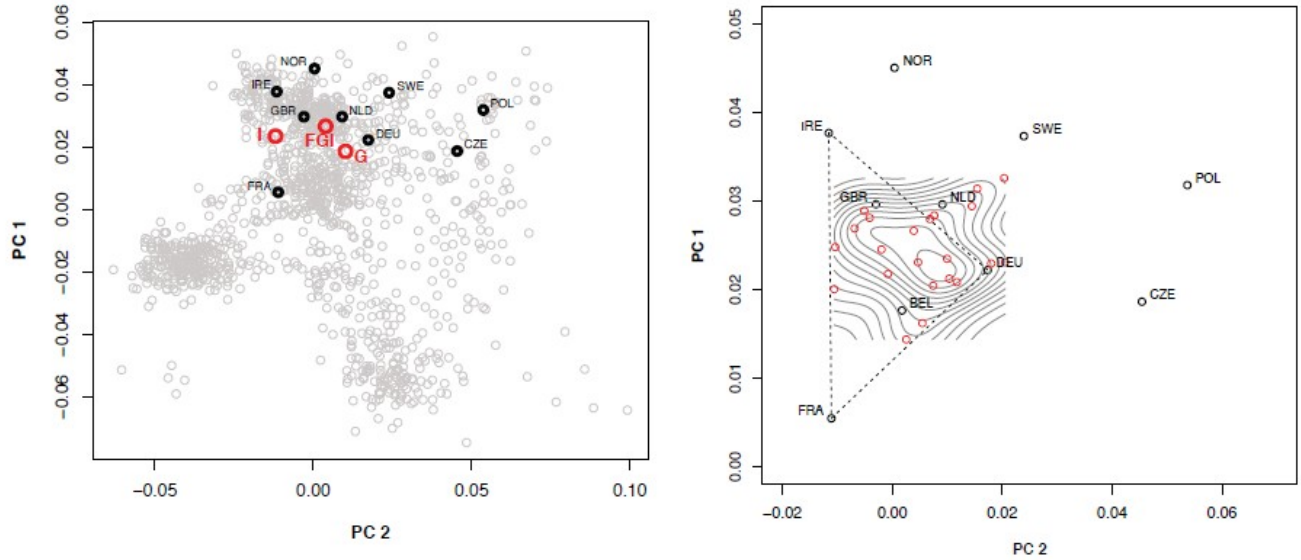

Figure S10. (left) A mother, father, and daughter from the PMRP sample (open, red circles) projected onto the map of European genetic distance. Open black circles mark the mean positions of individuals from the indicated countries in the $POPRES_{europe}$ dataset (abbreviations as in Fig. S2). The self-reported ancestry of the daughter (French, German, Irish) conflicts with those reported by her mother (Irish) and father (German). The position of the mother midway between the means of Ireland and France suggests she failed to report her French ancestry. (right) The distribution of $PMRP_{abridged}$ individuals who self-reported French-German-Irish (FGI) ancestry. The underlying contour is a two-dimensional kernel density estimate of this distribution. Note that FGI individuals tend to fall within a triangle with the three source countries at its vertices, which places them closest to English and Belgian means.

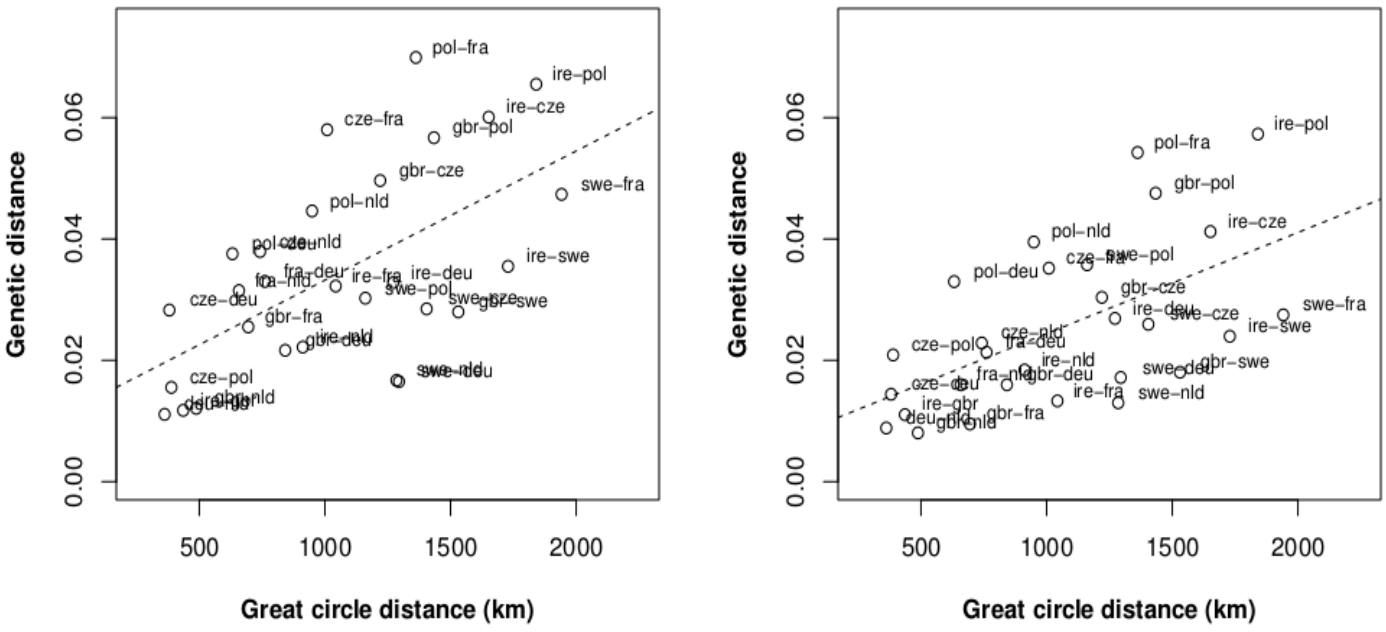

Figure S11. Genetic vs. great-circle distance between (left) Europeans in nine European countries and (right) Wisconsinites claiming insular ancestry from the same nine countries. The dashed line is the best-fit line from standard linear regression. Note that the majority of comparisons falling above the trend-line include either the Czech Republic or Poland.

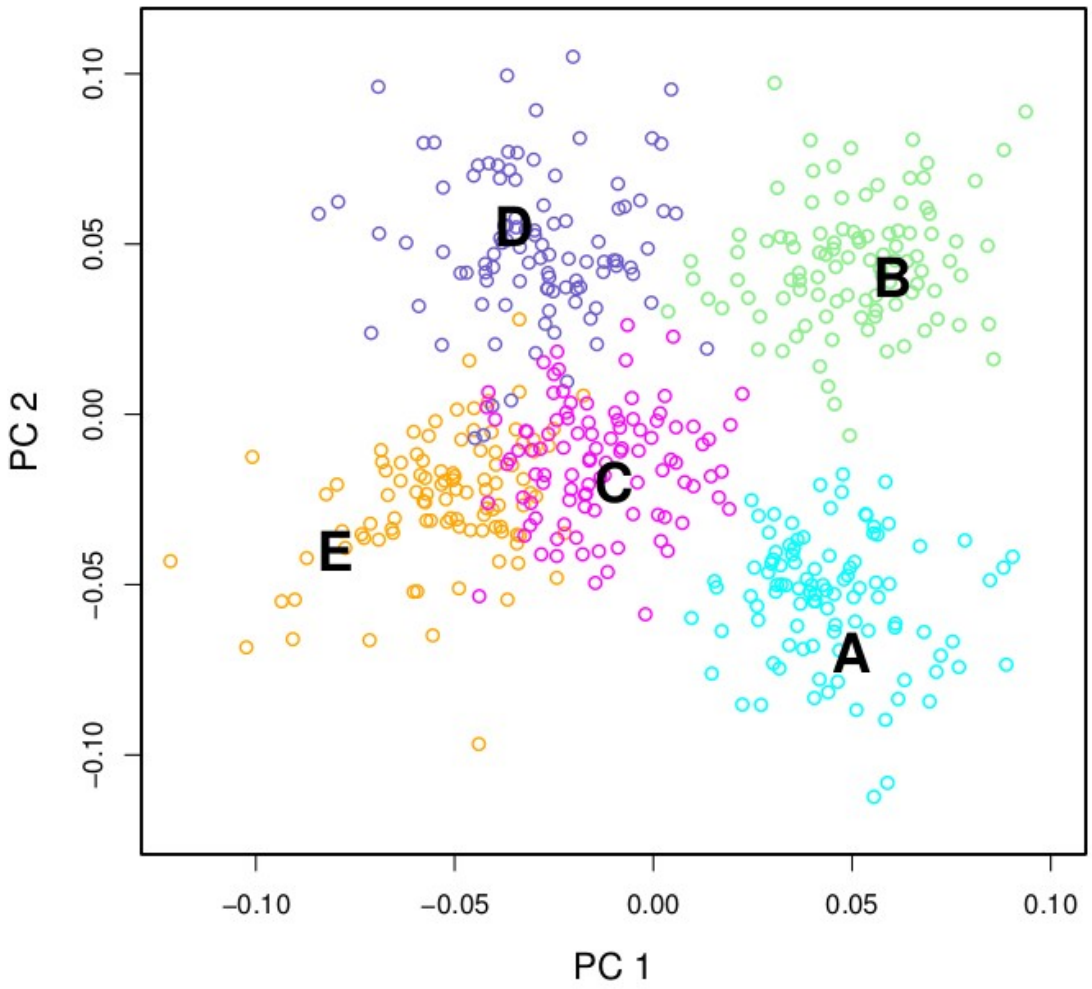

Figure S12. Results from PCA of the five simulated source populations.

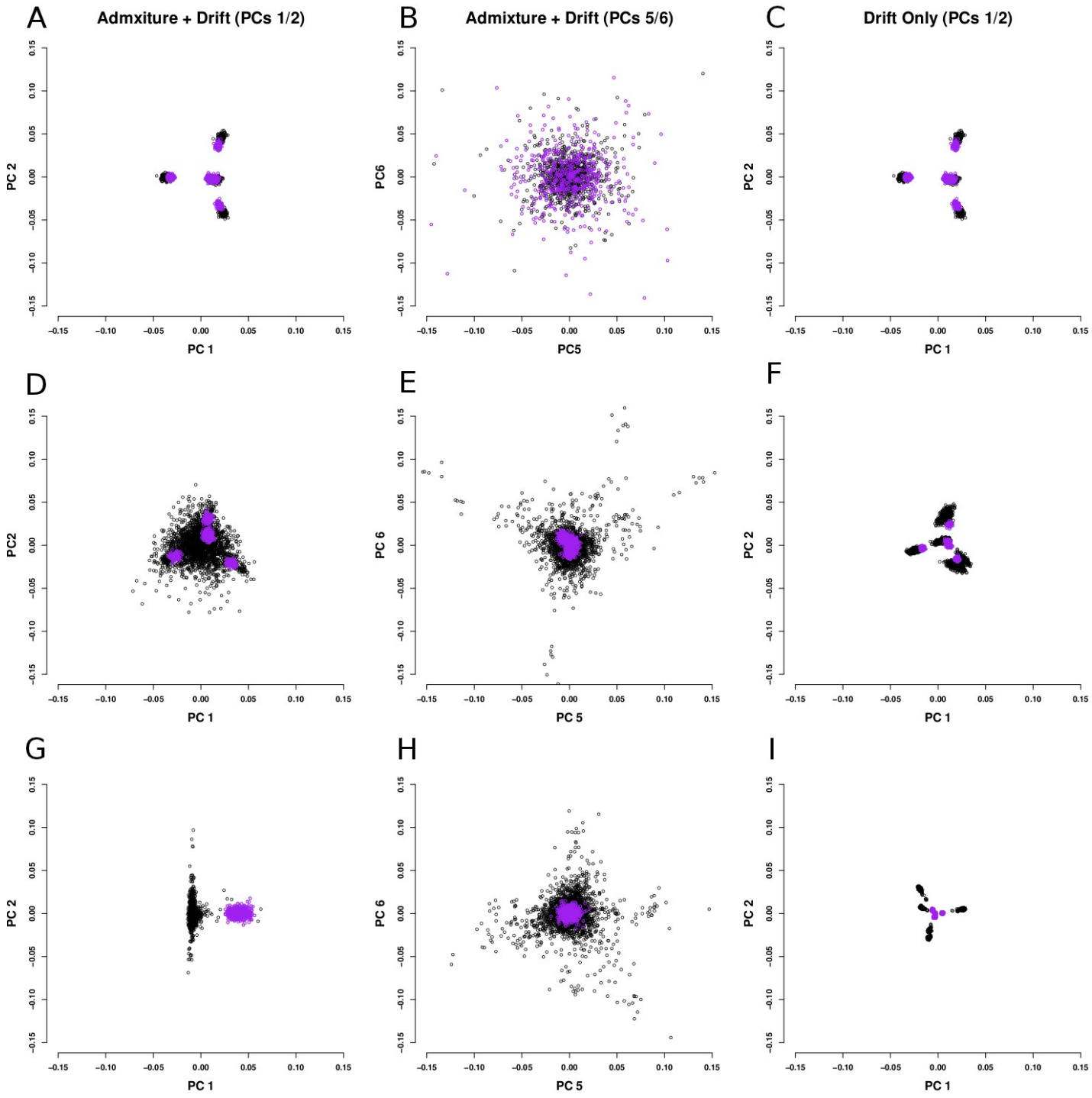

Figure S13. Results from combined PCA of simulated genotype data. Left and middle columns display results from the admixture plus drift scenario, while the right column displays results from the drift only scenario. Top row: initial colonization; middle row: 5 generations after colonization; bottom row: 40 generations after colonization. Purple dots are source individuals while black dots are sink individuals.

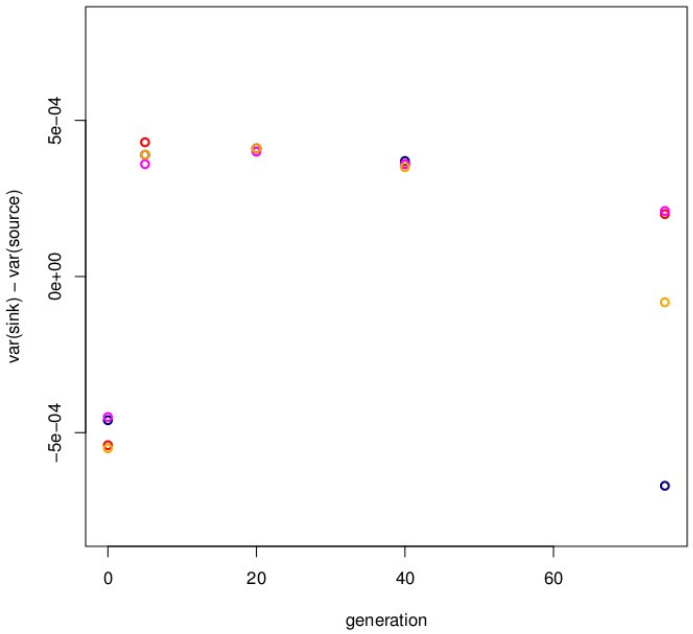

Figure S14. The metric dV_{PCx} = var(sink) - var(source) is averaged across all five sink-source pairs in simulations of admixture and drift at generations 0, 20, 40, and 75 for PC5 (navy), PC6 (orange), PC7 (red), and PC8 (magenta). In each case, var(·) is the variance of PC_x scores among individuals of the population in question. Likely due to founder effects, var(source) is greater than var(sink) at founding generation 0. However, by generation 5, the situation is reversed and dV_{PCx} remains positive for all sink-source pairs for many generations. This result suggests dV_{PCx} provides information regarding nascent divergence, which might be exploited in a permutation test wherein a null distribution of dV_{PCx} is obtained by permuting population labels thousands of times and recalculating dV_{PCx} each time. Further simulation would be required to determine the power of this test, which would likely be improved by including variance data from multiple higher-order PCs. A practical limitation to the suggested test is that *a priori* knowledge of potential source populations is required. Nevertheless, a simple test for nascent divergence should help both human and population geneticists characterize the recent history and genetic composition of focal populations.