The American Journal of Human Genetics, Volume *108*

Supplemental Data

Distinguishing pedigree relationships via multi-way identity by descent sharing and sex-specific genetic maps

Ying Qiao, Jens G. Sannerud, Sayantani Basu-Roy, Caroline Hayward, and Amy L. Williams

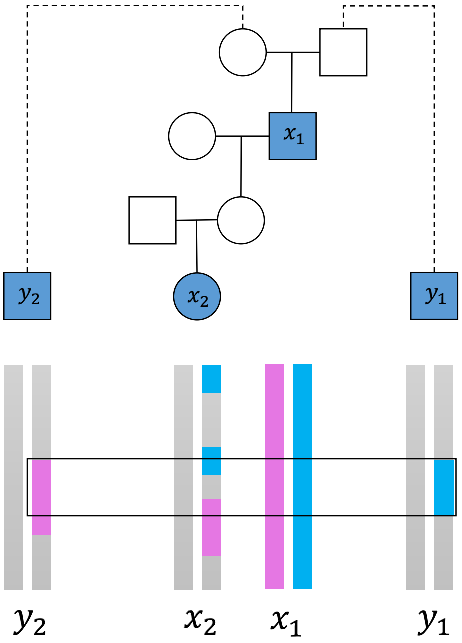


Figure S1: Example IBD sharing between a GP pair and their mutual relatives on both the maternal and paternal sides of the grandparent. The mutual relatives y_1 and y_2 are related to the GP pair x_1 and x_2 through the grandparent's mother and father, respectively. The blue or purple regions represent either one haplotype of x_1 or IBD segments other individuals share with those haplotypes. The black box outlines the regions CREST deems as being IBD2 between x_1 and the mutual relatives.

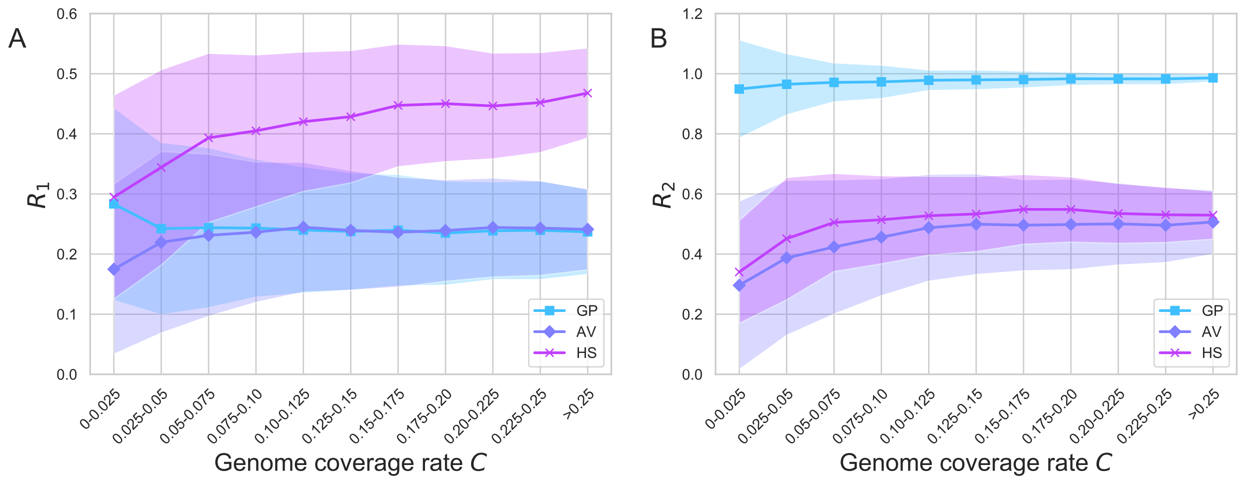


Figure S2: The variance of ratios R_1 and R_2 decrease as the genome coverage rate increases. (A) R_1 and (B) R_2 values across bins of genome coverage rates. The dots show the mean value in each bin, and the shaded regions span one standard deviation from the mean. Results are from simulated data, with IBD segments detected in genotype data, for all three types.

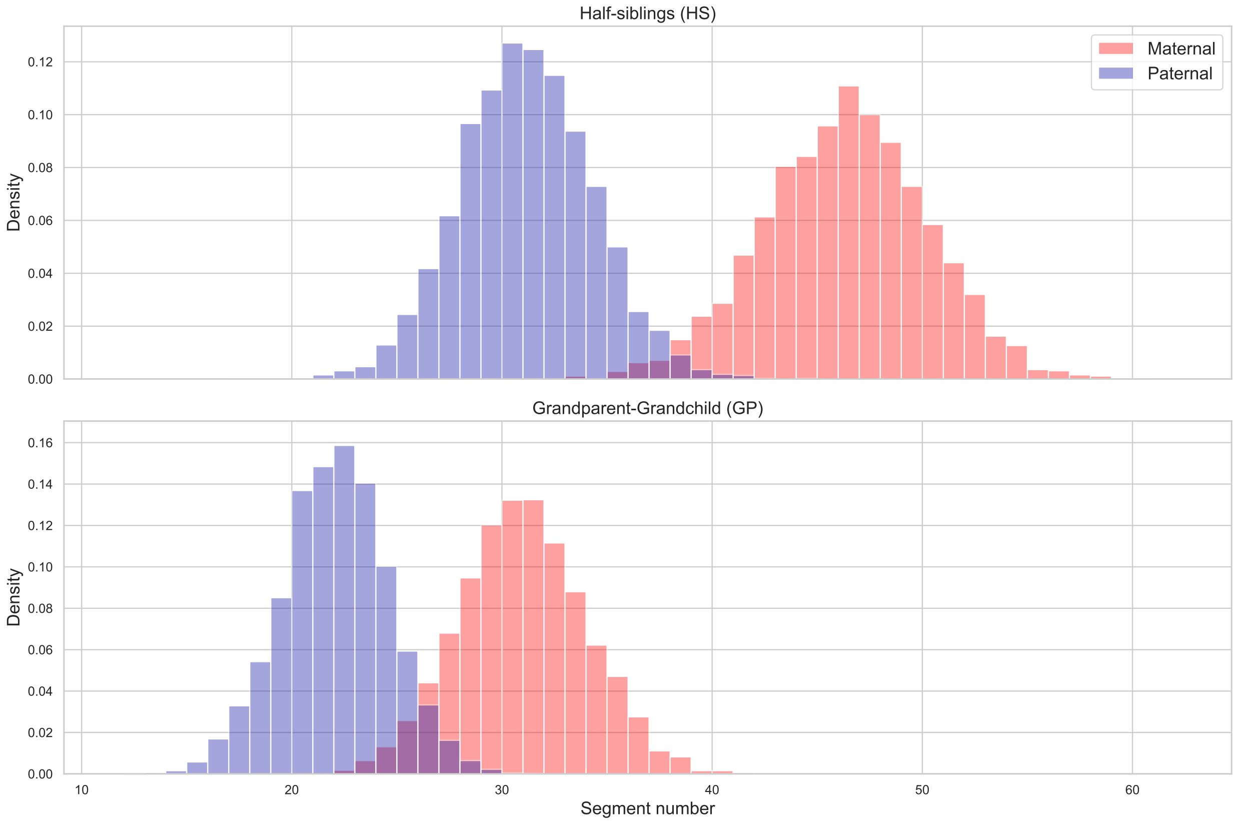


Figure S3: Segment number distributions for simulated HS and GP pairs. Histograms of total IBD segment numbers shared between HS (top) and GP (bottom) pairs using segments detected by IBIS in simulated relatives. Paternal relatives fall within the blue distributions, and maternal within the red.

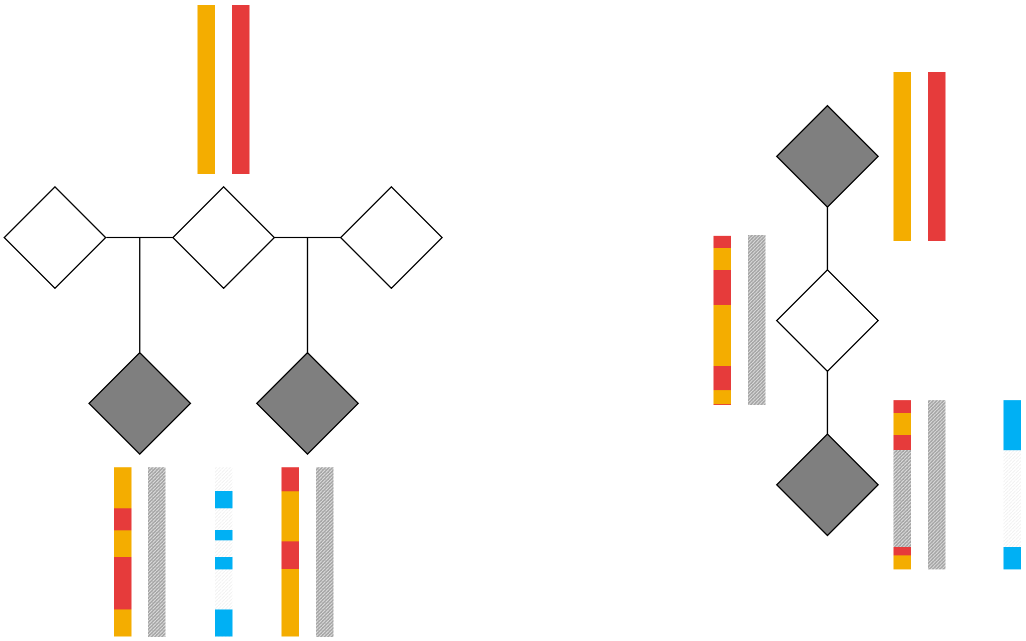


Figure S4: Example unphased IBD segments in HS and GP pairs. Haplotypes are presented as varicolored pairs of bars beneath or above diamonds that represent unsexed individuals. Gray stripes in these bars signify a region that is not IBD to other haplotypes. Haplotypes (red, gold) in the HS parent (left) and grandparent (right) are transmitted with crossovers (switches in color). Unphased IBD segments (blue) between the HS and GP pair appear at the bottom. In GP pairs, internal crossovers from the grandparent-to-parent meiosis do not produce IBD segment boundaries: only the crossovers in the parent-to-grandchild meiosis interrupt IBD sharing. Crossovers the parent transmits change the IBD status, from IBD0 to IBD1 and vice versa, for both relationship types.

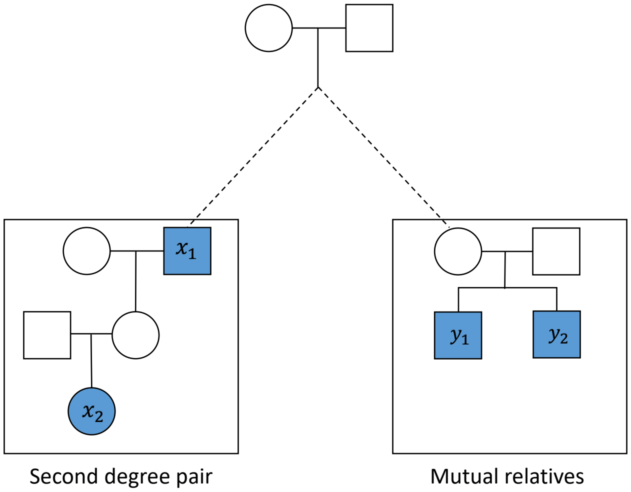


Figure S5: The structure of the simulated pedigrees used to evaluate CREST's relationship type classification. The left side shows an example target second degree pair, which is either a GP, AV, or HS pair. The right side depicts example mutual relatives, which include one or more individuals that are related to the second degree pair and to each other (Methods). Genotyped samples are shown as filled shapes. The dashed line connects the second degree pair and the mutual relatives to their unknown MRCA.

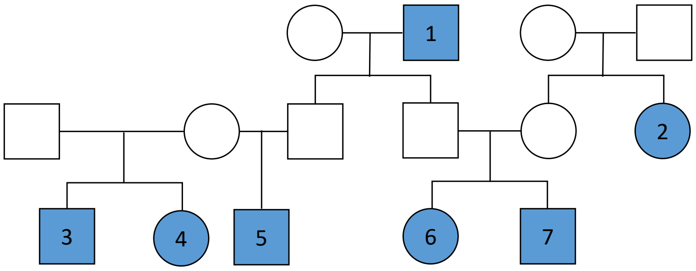


Figure S6: Example pedigree with individuals contained in multiple second degree pairs. Sample 1 and each of samples 5, 6, and 7 are three GP pairs, while sample 2 and samples 6 and 7 are two AV pairs. The real data results average the sensitivity and specificity among all samples with the same genetically older sample for these types, so the three GP pairs would each contribute a count of $\frac{1}{3}$ to the GP metrics, and the two AV pairs would contribute $\frac{1}{2}$ to the AV metrics. In turn, sample 5 and samples 3 and 4 form two HS pairs with the same common parent, and the real data results similarly include average scores for such pairs, in this case weighting each by $\frac{1}{2}$. Note that sample 5 is a member of both a GP and HS pair, and the results consider each type separately, incorporating the average metrics for all pairs within each type.

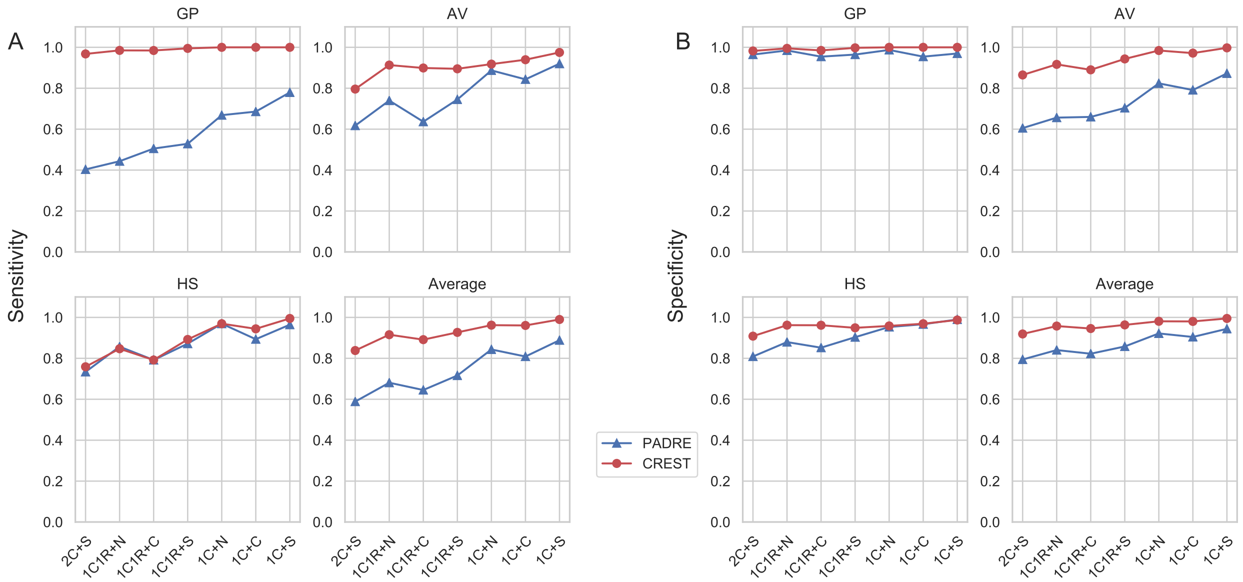


Figure S7: Performance of CREST and PADRE for second degree relationship type classification of pairs both tools classify. (A) The sensitivity and (B) specificity of CREST and PADRE for inferring GP, AV, and HS relationship types, along with the average of these rates across the three relationships. The x-axis indicates the mutual relatives types included in the analysis, with each target relationship type and mutual relative combination including data only for those pairs (out of 200 per data point) that both PADRE and CREST classify.

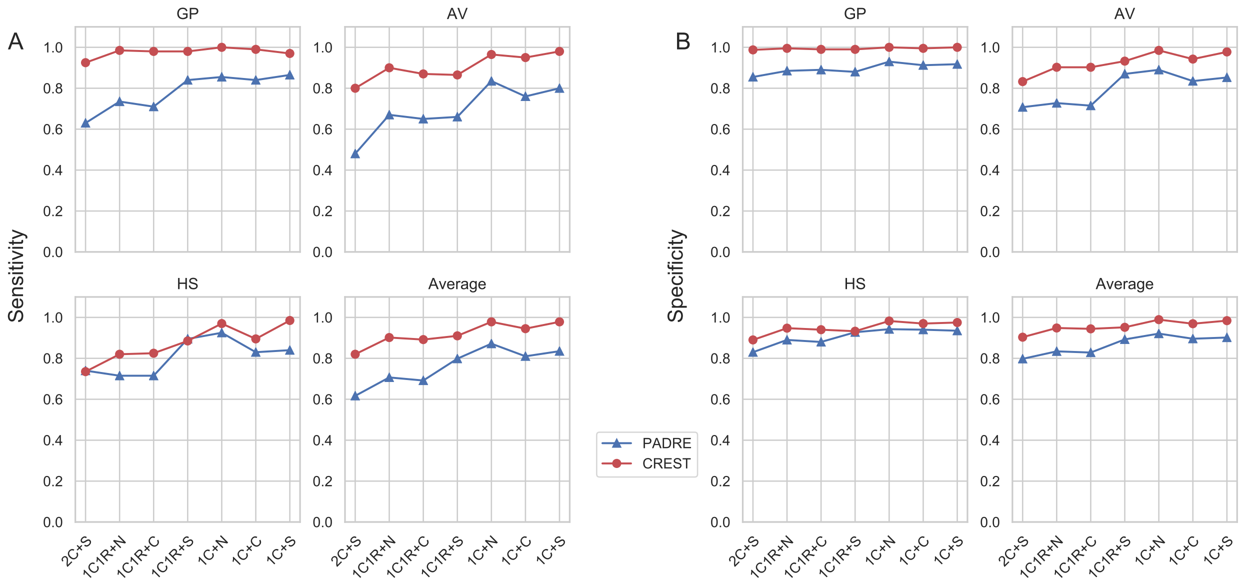


Figure S8: Performance of CREST and PADRE for second degree relationship type classification where PADRE used perfect haplotypes. (A) The sensitivity and (B) specificity of CREST and PADRE for inferring GP, AV, and HS relationship types, along with the average of these rates across the three relationships. The x-axis indicates the mutual relatives types included in the analysis, with each target relationship type and mutual relative combination including data from simulated phased haplotypes of 200 pairs.

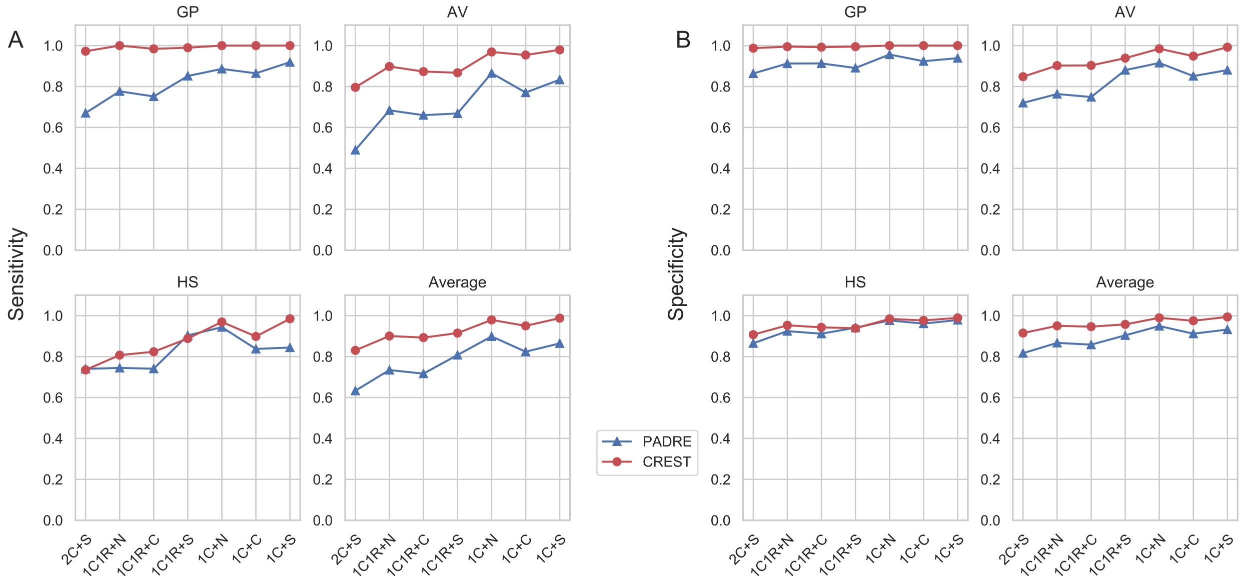
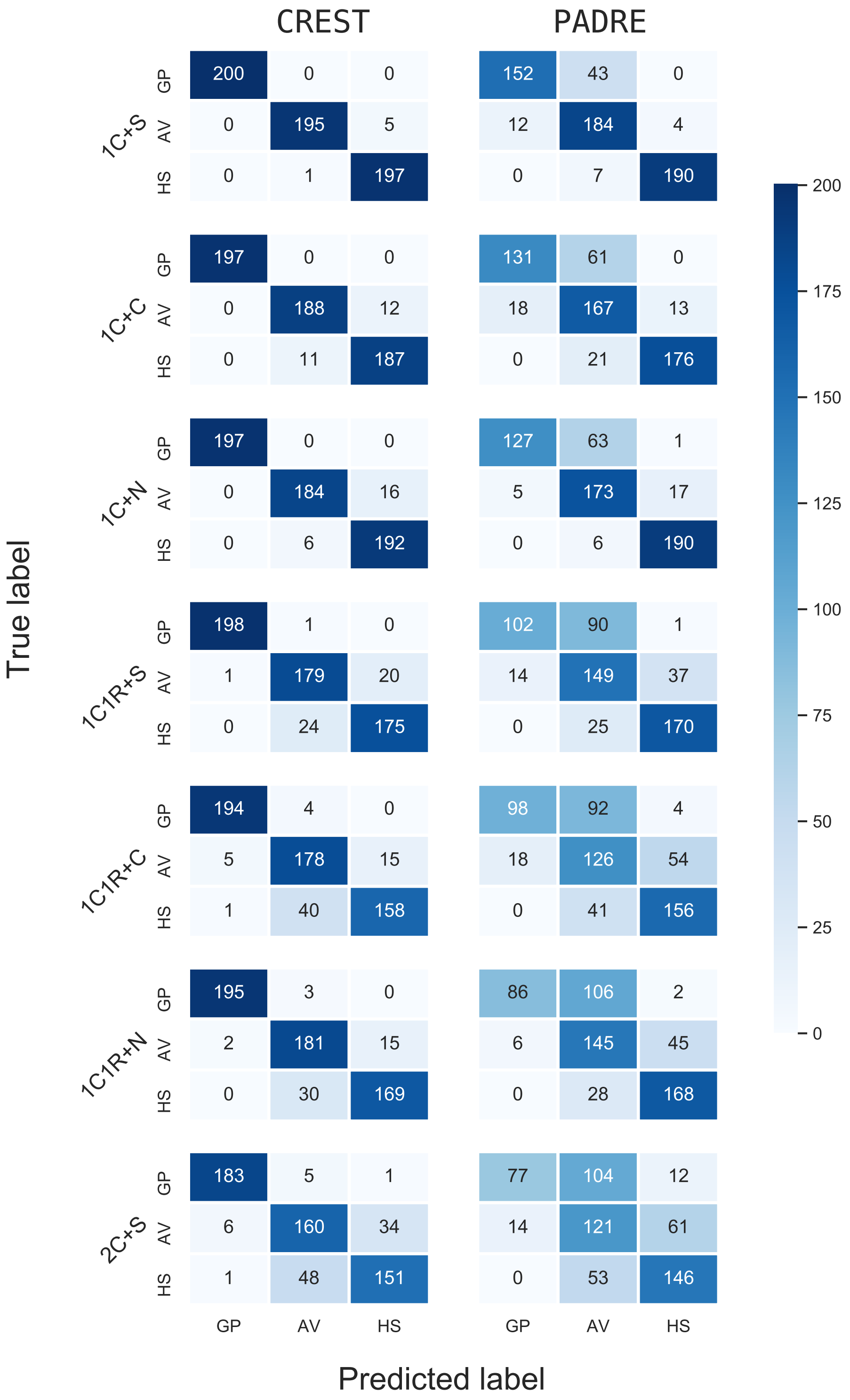
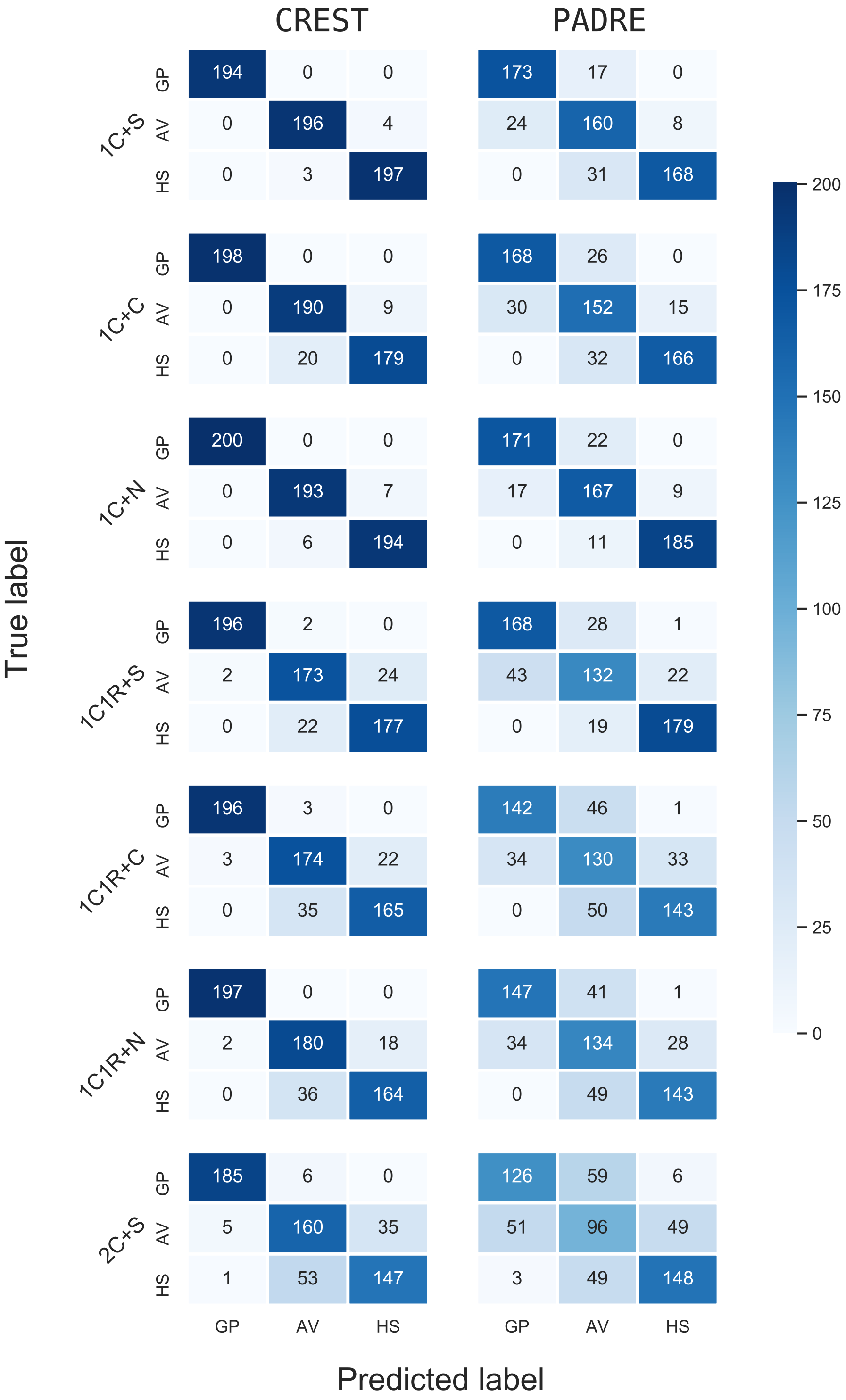


Figure S9: Performance of CREST and PADRE for second degree relationship type classification of pairs both tools classify and where PADRE used perfect haplotypes. (A) The sensitivity and (B) specificity of CREST and PADRE for inferring GP, AV, and HS relationship types, along with the average of these rates across the three relationships. The x-axis indicates the mutual relatives types included in the analysis, with each target relationship type and mutual relative combination including data only for those pairs (out of 200 per data point) that both PADRE and CREST classify.



		CREST			PADRE		
		GP	AV	HS	GP	AV	HS
1C+S	GP	200	0	0	152	43	0
	AV	0	195	5	12	184	4
	HS	0	1	197	0	7	190
1C+C	GP	197	0	0	131	61	0
	AV	0	188	12	18	167	13
	HS	0	11	187	0	21	176
1C+N	GP	197	0	0	127	63	1
	AV	0	184	16	5	173	17
	HS	0	6	192	0	6	190
1C1R+S	GP	198	1	0	102	90	1
	AV	1	179	20	14	149	37
	HS	0	24	175	0	25	170
1C1R+C	GP	194	4	0	98	92	4
	AV	5	178	15	18	126	54
	HS	1	40	158	0	41	156
1C1R+N	GP	195	3	0	86	106	2
	AV	2	181	15	6	145	45
	HS	0	30	169	0	28	168
2C+S	GP	183	5	1	77	104	12
	AV	6	160	34	14	121	61
	HS	1	48	151	0	53	146

Figure S10: The confusion matrices from the CREST and PADRE classification results. Analyses of CREST and PADRE include 200 pairs of GP, AV, and HS over different pedigree structures. Labels on the left indicate the mutual relatives in the pedigree structures. The row of each matrix gives the true relationship type and the column is the predicted relationship type. Since a few pairs failed classification by CREST or PADRE (Results), the sums of each row are not always 200.



		CREST			PADRE		
		GP	AV	HS	GP	AV	HS
1C+S	GP	194	0	0	173	17	0
	AV	0	196	4	24	160	8
	HS	0	3	197	0	31	168
1C+C	GP	198	0	0	168	26	0
	AV	0	190	9	30	152	15
	HS	0	20	179	0	32	166
1C+N	GP	200	0	0	171	22	0
	AV	0	193	7	17	167	9
	HS	0	6	194	0	11	185
1C1R+S	GP	196	2	0	168	28	1
	AV	2	173	24	43	132	22
	HS	0	22	177	0	19	179
1C1R+C	GP	196	3	0	142	46	1
	AV	3	174	22	34	130	33
	HS	0	35	165	0	50	143
1C1R+N	GP	197	0	0	147	41	1
	AV	2	180	18	34	134	28
	HS	0	36	164	0	49	143
2C+S	GP	185	6	0	126	59	6
	AV	5	160	35	51	96	49
	HS	1	53	147	3	49	148

Figure S11: The confusion matrices from the CREST and PADRE classification results where PADRE used perfect haplotypes. Analyses of CREST and PADRE include 200 pairs of GP, AV, and HS over different pedigree structures. Labels on the left indicate the mutual relatives in the pedigree structures. The row of each matrix gives the true relationship type and the column is the predicted relationship type. Since a few pairs failed classification by CREST or PADRE (Results), the sums of each row are not always 200.

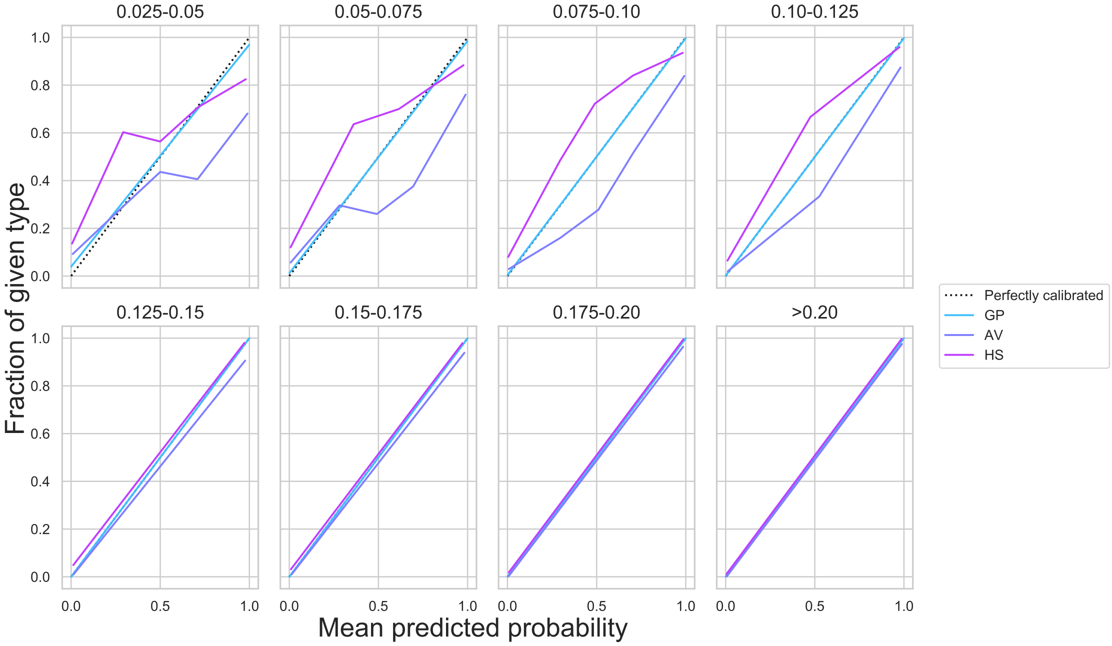


Figure S12: The calibration curves for classifying second degree relatives over different coverage rates. In each plot, the analysis includes 1,000 pairs of each type. The x-axis shows the per-bin mean predicted probability and the y-axis indicates the proportion of pairs that are of the given type in the corresponding bin. We used five bins where possible, but reduced the number of bins if needed to ensure that each bin includes at least 50 pairs.In all cases, bins are uniformly spaced.

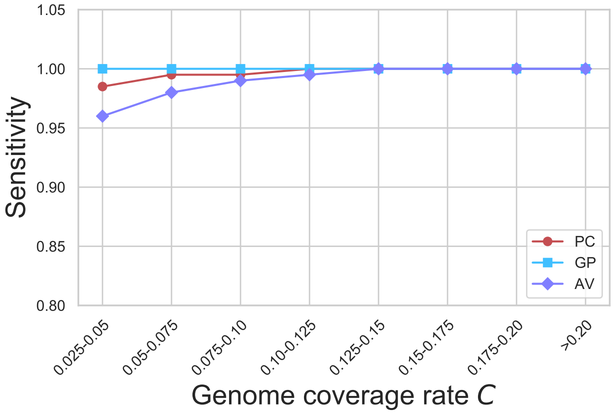


Figure S13: CREST accurately infers the directionality of GP, AV, and PC pairs. Plot shows the sensitivity across bins of genome coverage rates for 200 pairs in each bin.

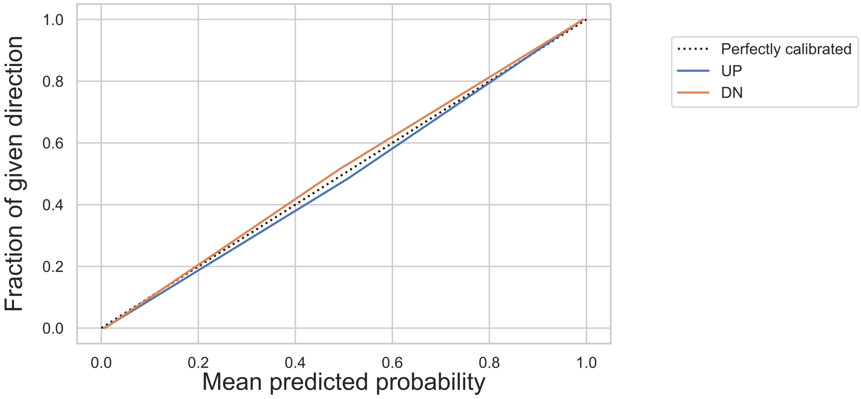


Figure S14: The calibration curves for inferring the relationship directionality of GP, AV, and PC pairs. The x-axis shows the per-bin mean predicted probability and the y-axis indicates the proportion of pairs that are of the given direction in the corresponding bin.The analysis includes 300 pairs of each direction. Plot includes three uniformly spaced bins.

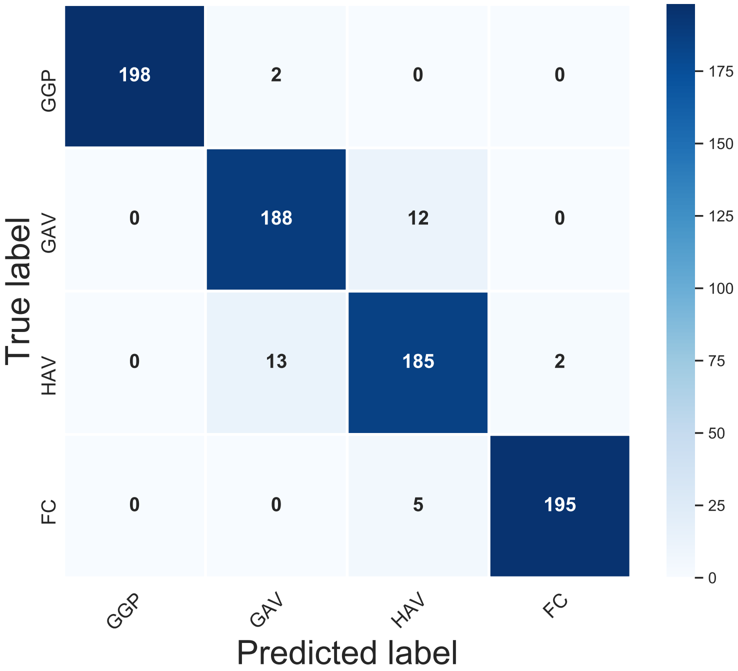


Figure S15: The confusion matrix for classifying third degree relatives. The rows correspond to the true relationship type and the column is the predicted type. The analysis includes 200 simulated pairs of each type (didn't use inferred degrees).

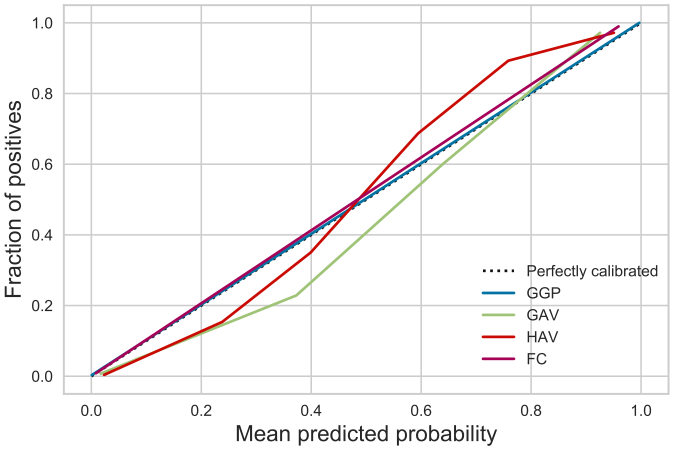


Figure S16: The calibration curves for classifying third degree relatives. The x-axis shows the per-bin mean predicted probability and the y-axis indicates the proportion of pairs that are of the given type in the corresponding bin. The analysis includes 200 pairs of each type. Plot includes five uniformly spaced bins.

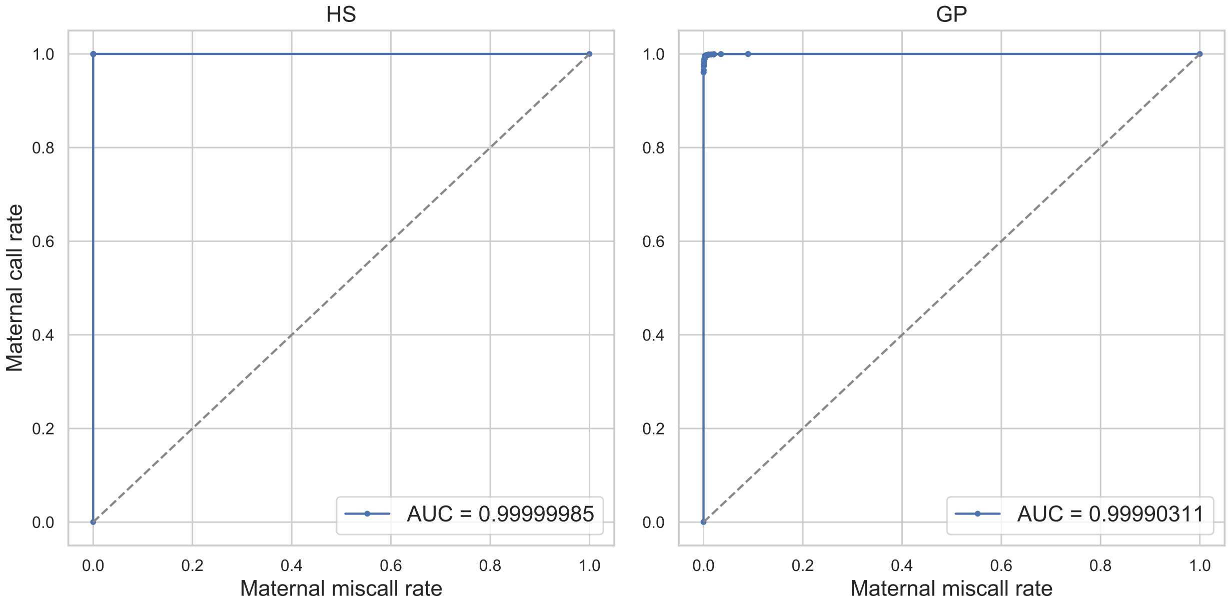


Figure S17: Receiver operating characteristic (ROC) curves for sex inference on simulated HS and GP pairs. Plots of the ROC curves for simulated data from Ped-sim, with area under the curve (AUC) values. Maternal call and miscall rates refer to the proportion of pairs correctly and incorrectly classified as maternal. The choice to plot maternal performance as opposed to paternal performance is arbitrary and does not alter the properties of the ROC.

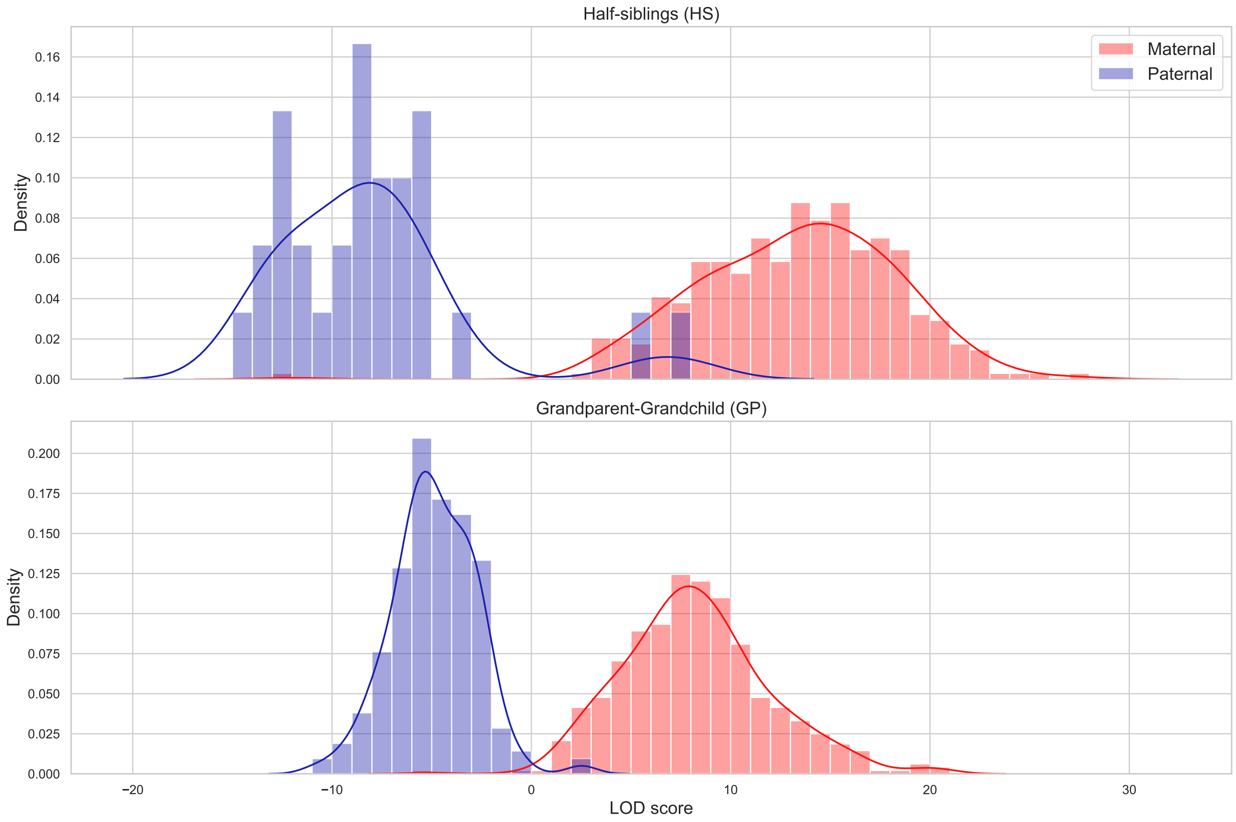


Figure S18: Preliminary LOD score histograms from the Generation Scotland data. Histograms of LOD scores for HS (top) and GP (bottom) pairs include several pairs that are extreme outliers for their reported relationships. We later determined these to be incorrectly labeled in the dataset. Visible anomalies corresponding to removed pairs include: a reported maternal HS pair at $LOD \approx$ -12, later determined to be paternal HS; a reported paternal HS pair at $LOD \approx 5.7$, of uncertain (possibly avuncular) true relationship; a reported paternal HS pair at $LOD \approx 7.9$ later determined to be maternal HS; a reported maternal GP peak at $LOD \approx$ -5.1, later determined to be paternal GP.

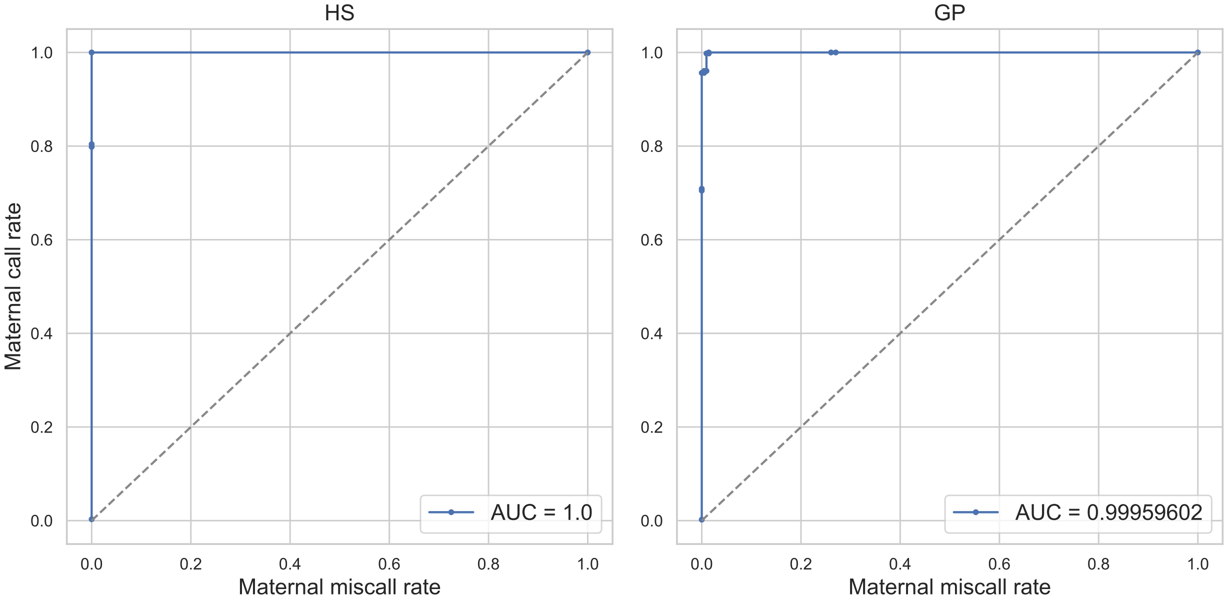


Figure S19: Receiver operating characteristic (ROC) curves for sex inference on Generation Scotland HS and GP pairs. Plots of the ROC curves for pairs from Generation Scotland, with area under the curve (AUC) values. Maternal call and miscall rates refer to the proportion of pairs correctly and incorrectly classified as maternal. The choice to plot maternal performance as opposed to paternal performance is arbitrary and does not alter the properties of the ROC.

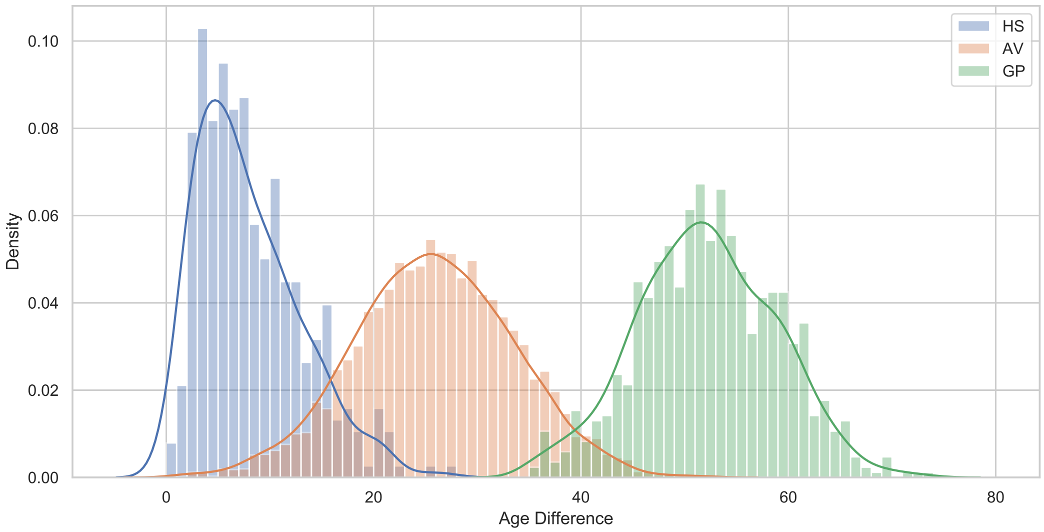


Figure S20: The distribution of age differences of second degree relatives in GS dataset. Histograms of the absolute value of age differences of all GP, AV, and HS pairs in the GS dataset.