The American Journal of Human Genetics, Volume *105*

Supplemental Data

Characterization of Prevalence and Health Consequences of Uniparental Disomy in Four Million Individuals from the General Population

Priyanka Nakka, Samuel Pattillo Smith, Anne H. O'Donnell-Luria, Kimberly F. McManus, 23andMe Research Team, Joanna L. Mountain, Sohini Ramachandran, and J. Fah Sathirapongsasuti

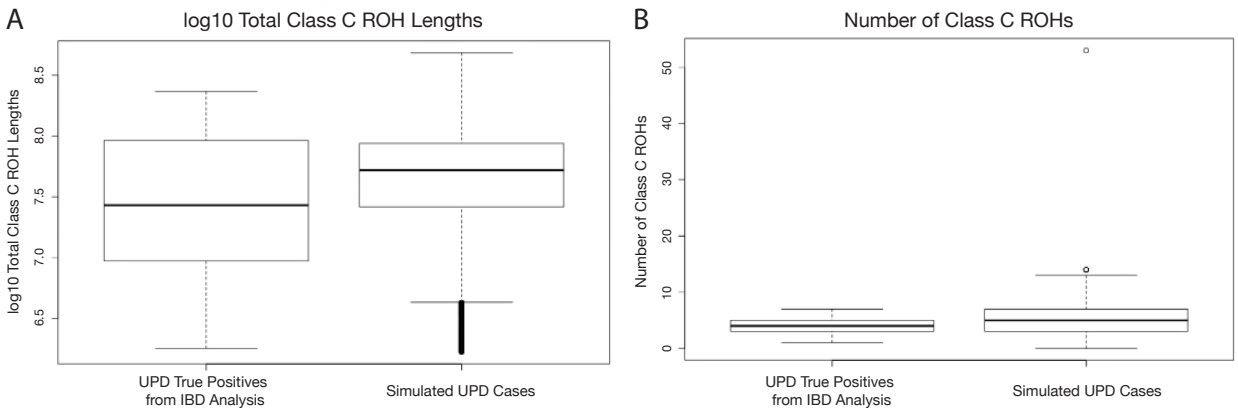

Figure S1. Boxplots comparing A) log10 total Class C ROH lengths of UPD true positives in the 23andMe dataset, which are identified using IBD analysis (left) and of simulated UPD cases (right) (t-test *p*-value = 0.94) and B) total number of Class C ROHs of UPD true positives (left) and of simulated UPD cases (right) (t-test *p*-value = 0.04). Because the number of Class C ROH differed significantly between real UPD cases and simulated UPD cases (panel B, t-test *p*-value < 0.05), we did not train on this variable in our classifiers and instead train on total Class C ROH length (panel A, t-test *p*-value > 0.05).

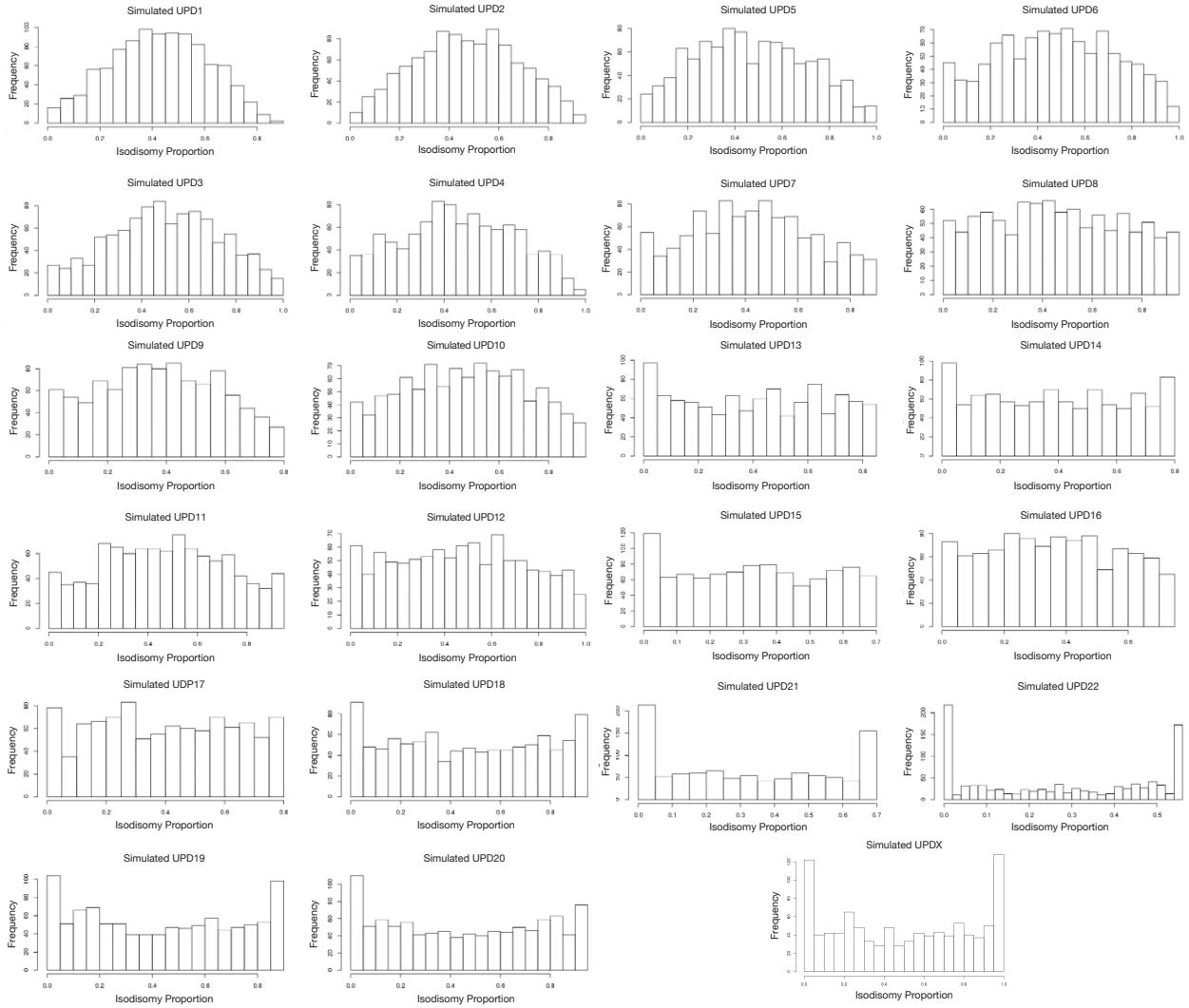

Figure S2. **Distributions of isodisomy proportion in simulated UPD cases for each of 23 chromosomes.** We simulated 1000 cases of UPD for each chromosome for each cohort based on genotype data from 23andMe. We see that, though the distribution of cases varies between chromosomes, every subtype of UPD (hetUPD, in which 0% isodisomy occurs; isoUPD, in which 100% isodisomy occurs; and partial isoUPD, in which an intermediate proportion between 0 and 100% isodisomy occurs) is produced on each chromosome by our simulation method.

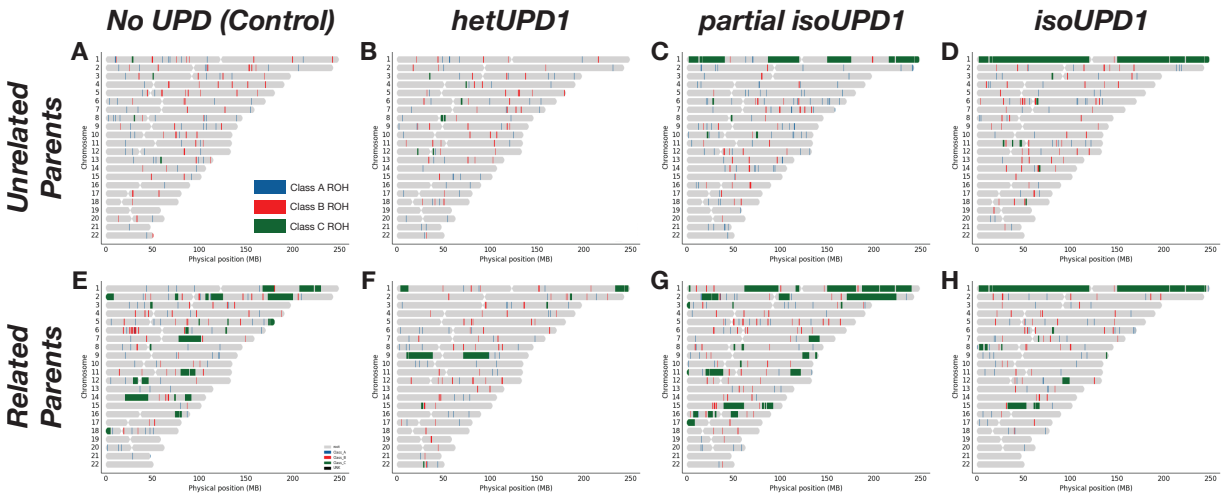

Figure S3. Example ideograms showing ROH locations (Class A ROH in blue, Class B ROH in red and Class C ROH in green). These were drawn from six simulated UPD cases and two simulated controls to illustrate the classification problem our ROH-based supervised classifiers faced. A) Ideogram of ROH in a control with unrelated parents, B) hetUPD of chromosome 1 with unrelated parents, C) partial isoUPD of chromosome 1 with unrelated parents, D) isoUPD of chromosome 1 with unrelated parents, E) control with related parents, F) hetUPD of chromosome 1 with related parents, G) partial isoUPD of chromosome 1 with related parents, and H) isoUPD of chromosome 1 with related parents. These figures show that long ROH can occur in partial isoUPD and isoUPD cases as well as individuals with related parents, and further illustrate that hetUPD cannot be identified based on ROH.

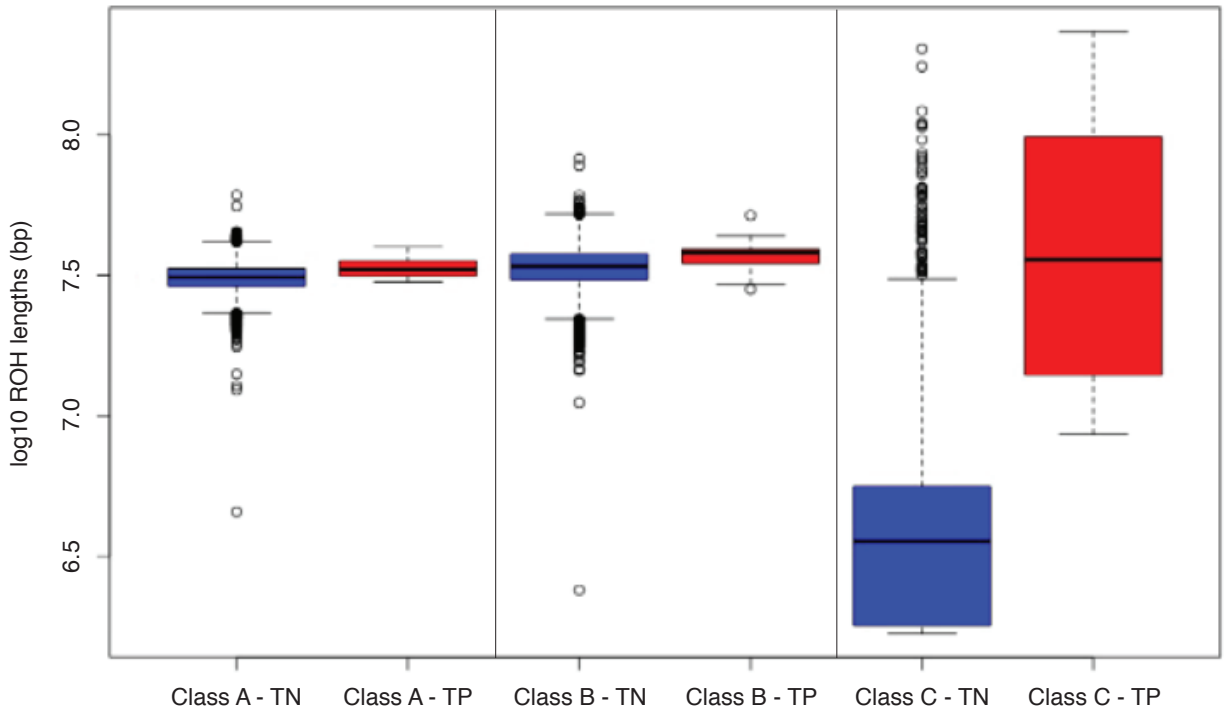

Figure S4. **Boxplots comparing log10 ROH lengths (in bp) between UPD true negatives (TN, blue) and UPD true positives (TP, red) in the 23andMe dataset across the three length classes of ROH, from left to right: 1) Class A, the shortest ROH; 2) Class B, intermediate length ROH; 3) Class C, the longest ROH.** ROH length class boundaries for each cohort are determined by GARLIC using gaussian mixture modeling (Table S1). Only Class C ROH lengths differ significantly between UPD true negatives and true positives (t-test *p*-value < 0.05).

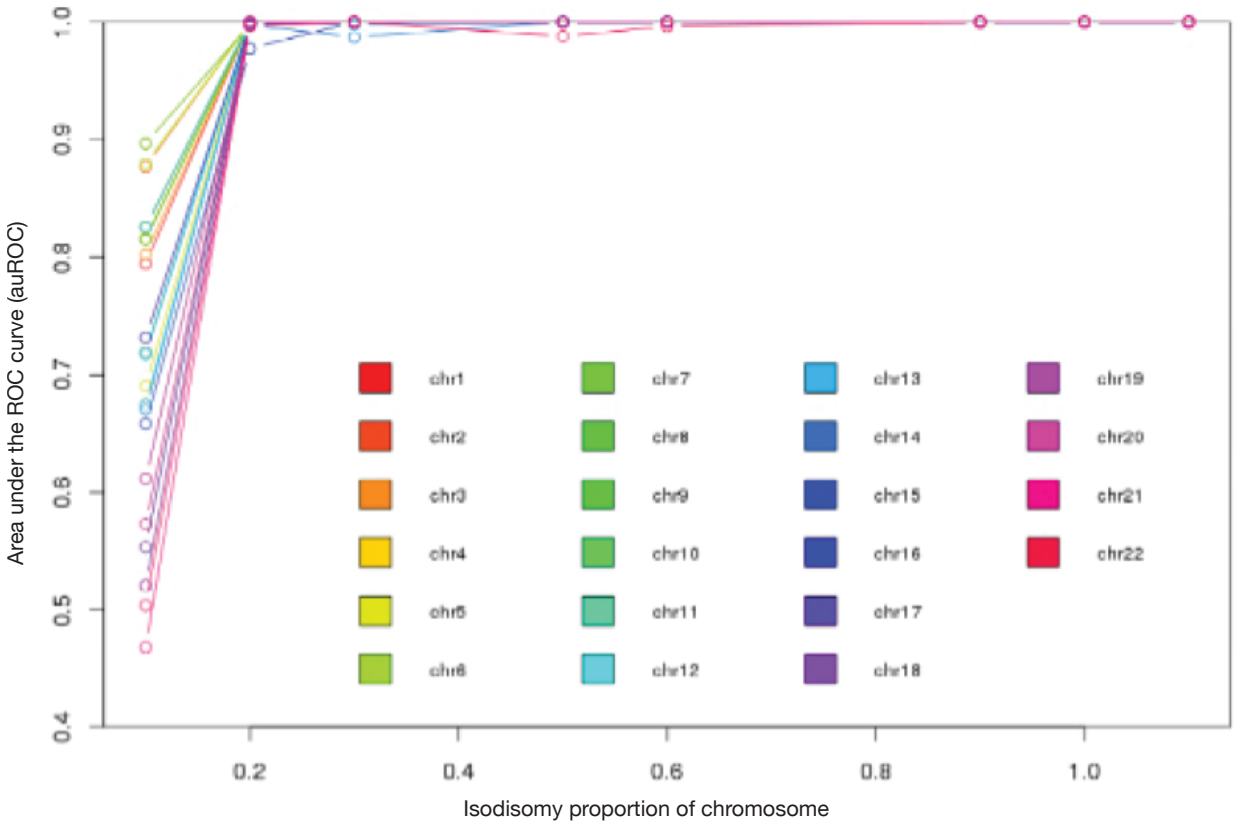

Figure S5. **Area under the ROC curve (auROC) versus isodisomy proportion on the simulated chromosome.** We found that auROC increases with isodisomy proportion on the simulated chromosome. Our classifiers perform best when isodisomy spans at least 20-50% of the chromosome.

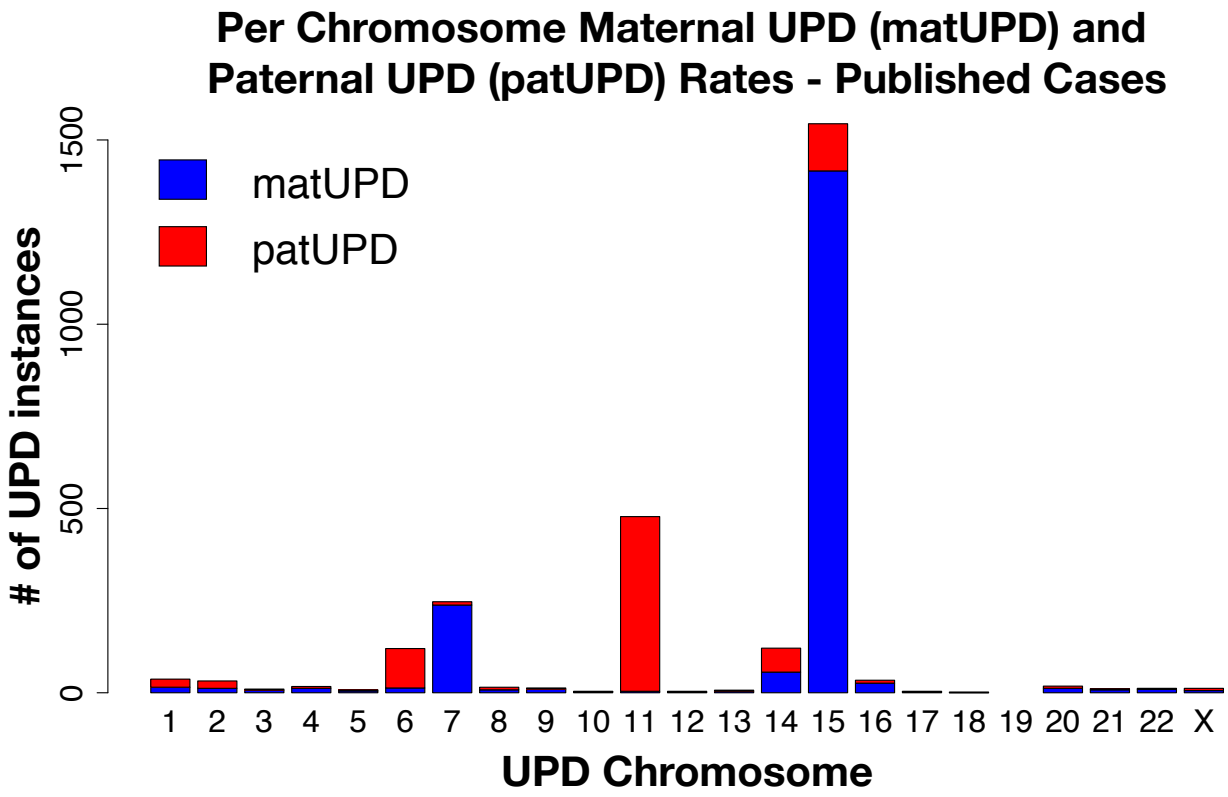

Figure S6. **The per chromosome distribution of clinical UPD cases published in the literature to date**[1] **(see Web Resources, accessed 11/29/18).** More than one case has been observed on each autosome except 19 and the X chromosome. Published UPD cases seem to cluster on chromosomes 6, 7, 11, 14 and 15, which contain clusters of imprinted genes that cause clinical phenotypes (Figure S7). There are 1869 matUPD cases in total and 881 patUPD cases in total, suggesting that matUPD is about twice as common as patUPD.

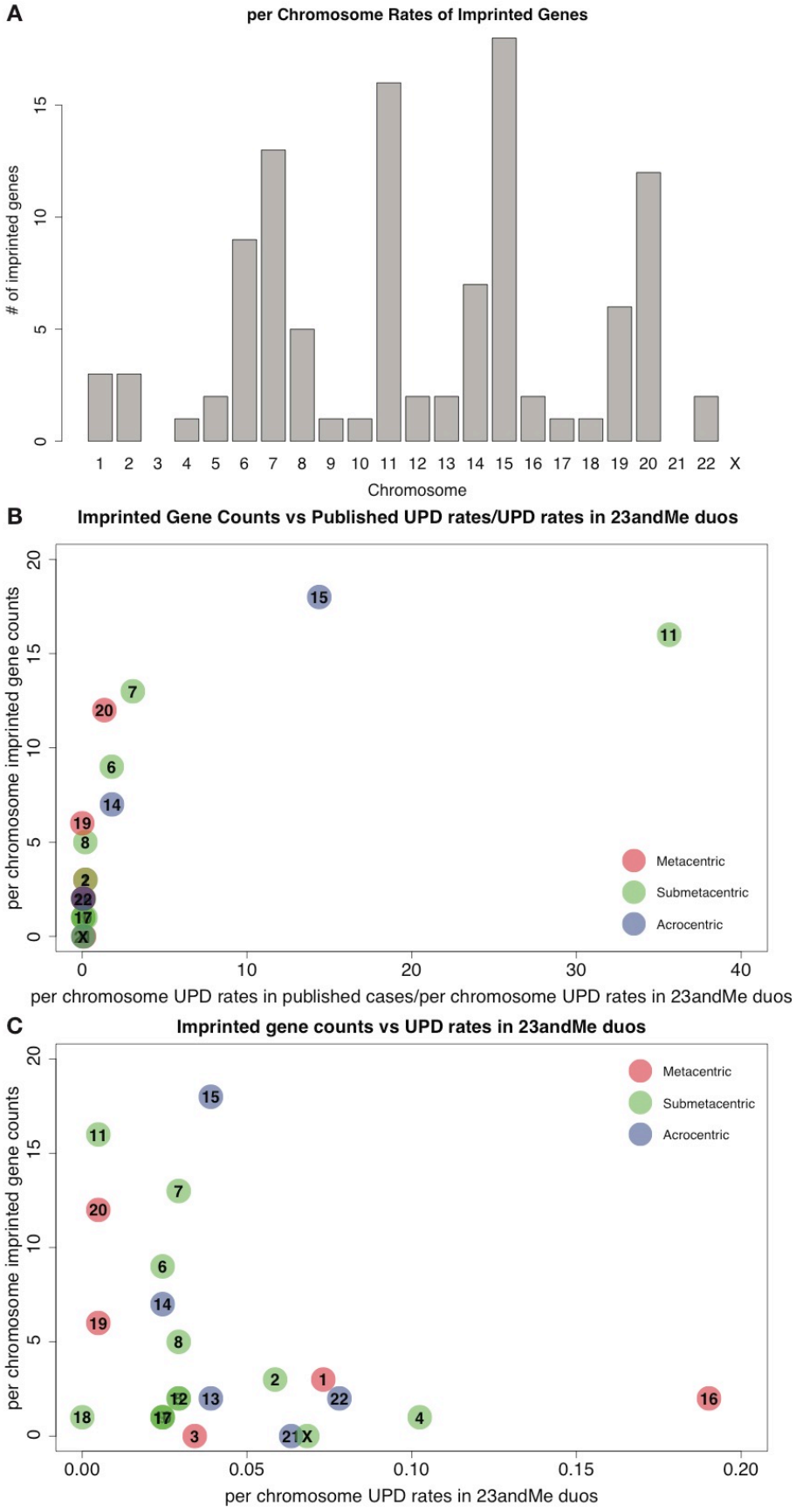
A
per Chromosome Rates of Imprinted Genes
# of imprinted genes
0
5
10
15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome
B
Imprinted Gene Counts vs Published UPD rates/UPD rates in 23andMe duos
per chromosome imprinted gene counts
0
5
10
15
20
0 10 20 30 40
per chromosome UPD rates in published cases/per chromosome UPD rates in 23andMe duos
Metacentric
Submetacentric
Acrocentric
C
Imprinted gene counts vs UPD rates in 23andMe duos
per chromosome imprinted gene counts
0
5
10
15
20
0.00 0.05 0.10 0.15 0.20
per chromosome UPD rates in 23andMe duos
Metacentric
Submetacentric
Acrocentric

Figure S7. Correlation between imprinted gene counts and published UPD rates and UPD rates in the 23andMe duos. A) The per chromosome distribution of imprinted gene counts (see Web Resources, accessed 06/28/19). B) We find that the ratio of per chromosome UPD rates from published cases to the per chromosome UPD rates from 23andMe duos is significantly correlated with the number of imprinted genes on each chromosome (Pearson's correlation = 0.70, *p*-value = 0.0003). C) We also find that the per chromosomes rates of UPD from 23andMe duos are not significantly correlated with the number of imprinted genes (Pearson's correlation = -0.32, p-value = 0.13). Chromosomes are colored by centromeric type: metacentric chromosomes are shown in red, submetacentric chromosomes in green and acrocentric chromosomes in blue.

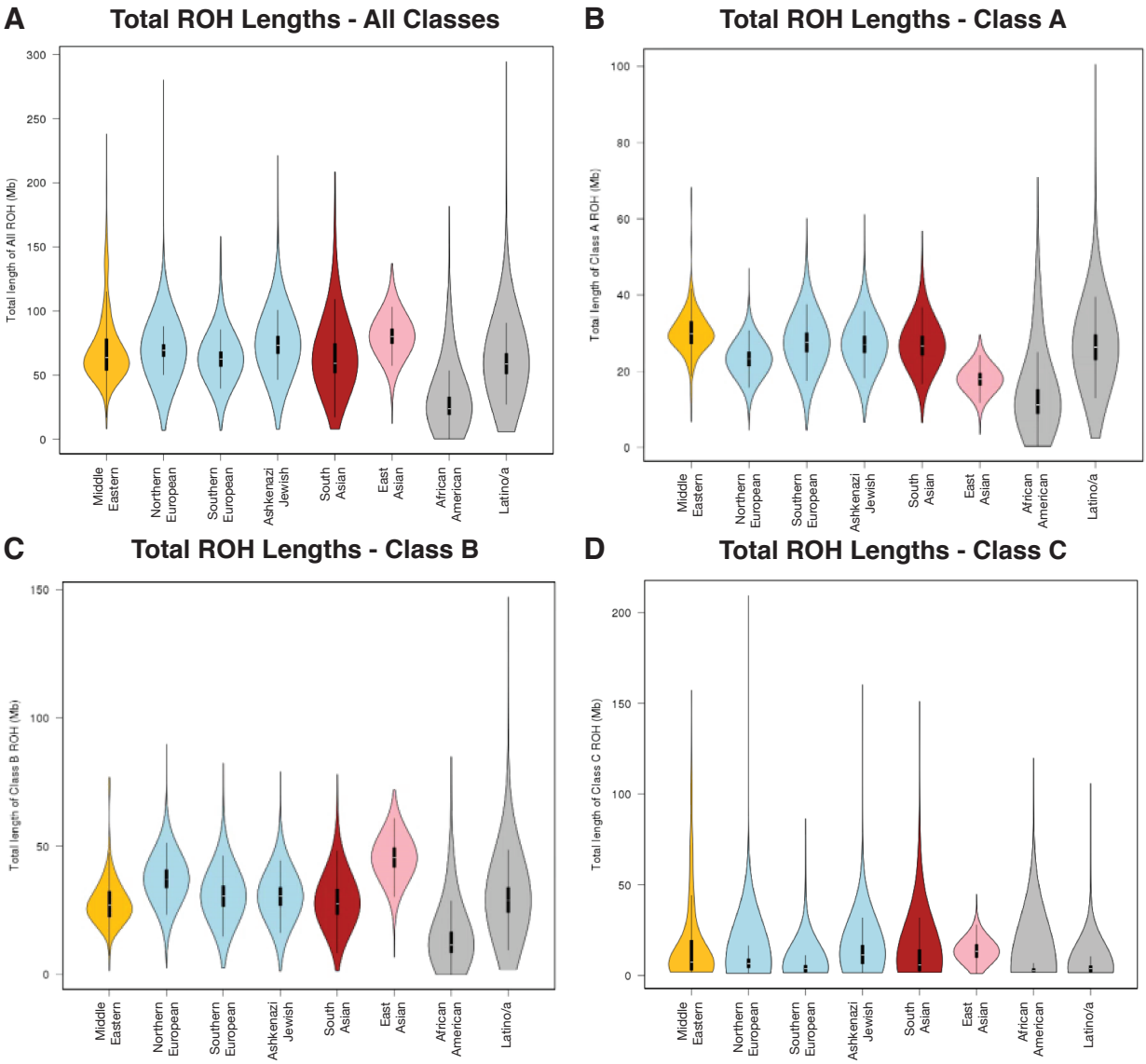

Figure S8. **Violin plots of ROH length distributions (in Mb) for eight cohorts in the 23andMe database.** The cohorts are colored by continental ancestry group: Middle Eastern (yellow), European (blue), South Asian (red), East Asian (pink), and admixed (grey). ROH were identified using GARLIC, which divides inferred ROH into three classes based on length. These plots show A) all classes combined, B) class A, the shortest ROH, C) class B, intermediate length ROH, and D) class C, the longest ROH. These plots recapitulate patterns of ROH length distributions seen in published analyses of ROH across global populations.[2,3]

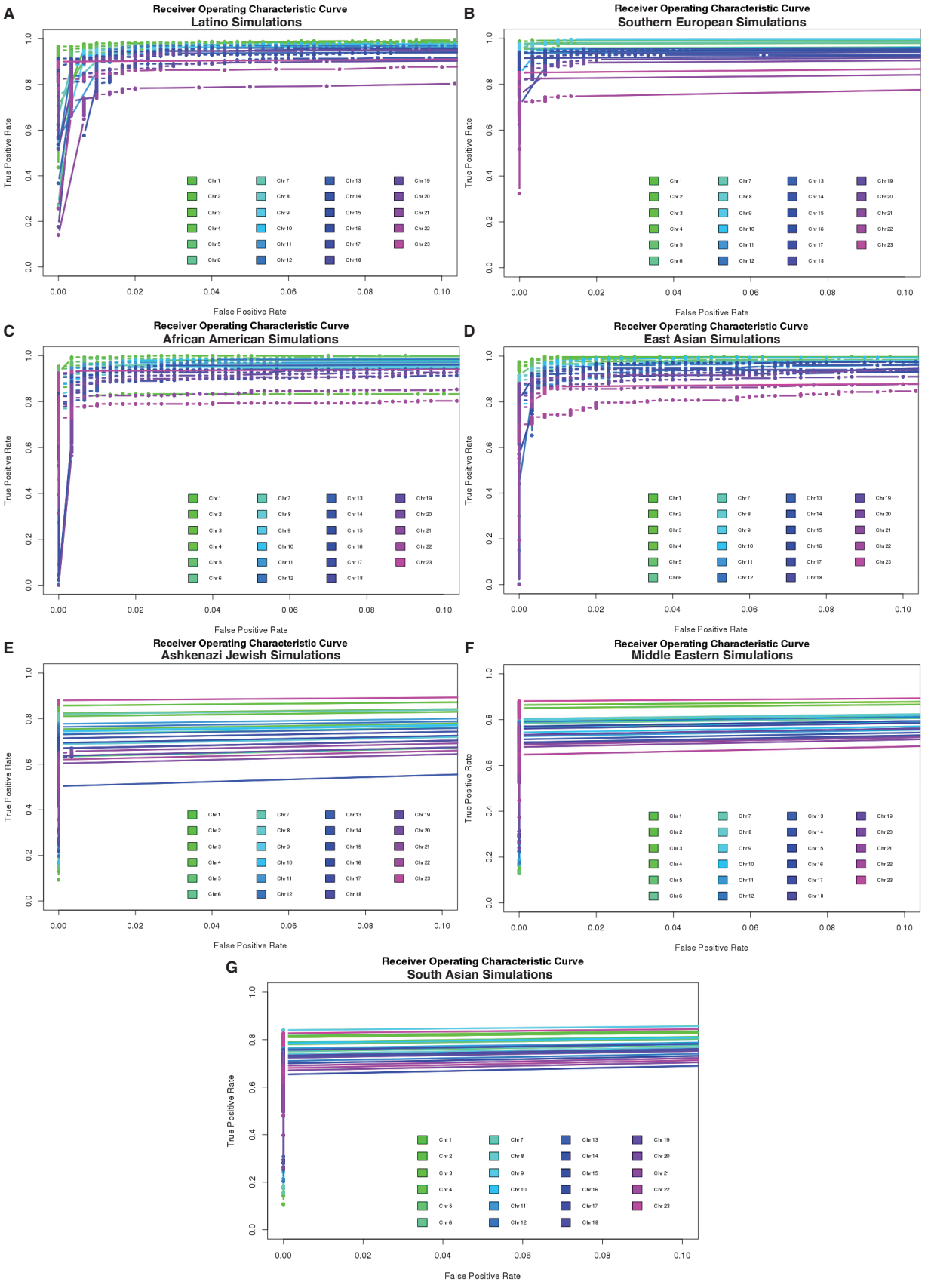
A
Receiver Operating Characteristic Curve
Latino Simulations
B
Receiver Operating Characteristic Curve
Southern European Simulations
C
Receiver Operating Characteristic Curve
African American Simulations
D
Receiver Operating Characteristic Curve
East Asian Simulations
E
Receiver Operating Characteristic Curve
Ashkenazi Jewish Simulations
F
Receiver Operating Characteristic Curve
Middle Eastern Simulations
G
Receiver Operating Characteristic Curve
South Asian Simulations
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Chr 6
Chr 7
Chr 8
Chr 9
Chr 10
Chr 11
Chr 12
Chr 13
Chr 14
Chr 15
Chr 16
Chr 17
Chr 18
Chr 19
Chr 20
Chr 21
Chr 22
Chr 23

Figure S9. **Receiver operating characteristic curves for 23 UPD classifiers (one per chromosome) trained on simulated individuals based on** A) Latino cohorts (TPR between 0.76 and 0.98 when FPR is fixed at 0.01), B) South European cohorts (TPR between 0.75 and 0.99 when FPR is fixed at 0.01), C) African American cohorts (TPR between 0.79 and 0.99 when FPR is fixed at 0.01), D) East Asian cohorts (TPR between 0.74 and 1 when FPR is fixed at 0.01), E) Ashkenazi Jewish cohorts (TPR between 0.51 and 0.899 when FPR is fixed at 0.01), F) Middle Eastern cohorts (TPR between 0.62 and 0.88 when FPR is fixed at 0.01), and G) South Asian cohorts (TPR between 0.63 and 0.84 when FPR is fixed at 0.01). Plots A-D show ROC curves with auROC > 0.9, which lead to successful classification in real data, whereas plots E-G show classifiers that perform relatively poorly on simulated testing data (auROC < 0.9). The cohorts in plots E-G, Ashkenazi Jewish, Middle Eastern and South Asian, are known to have practiced endogamy and so high levels of consanguinity may be confounding UPD detection for these classifiers.

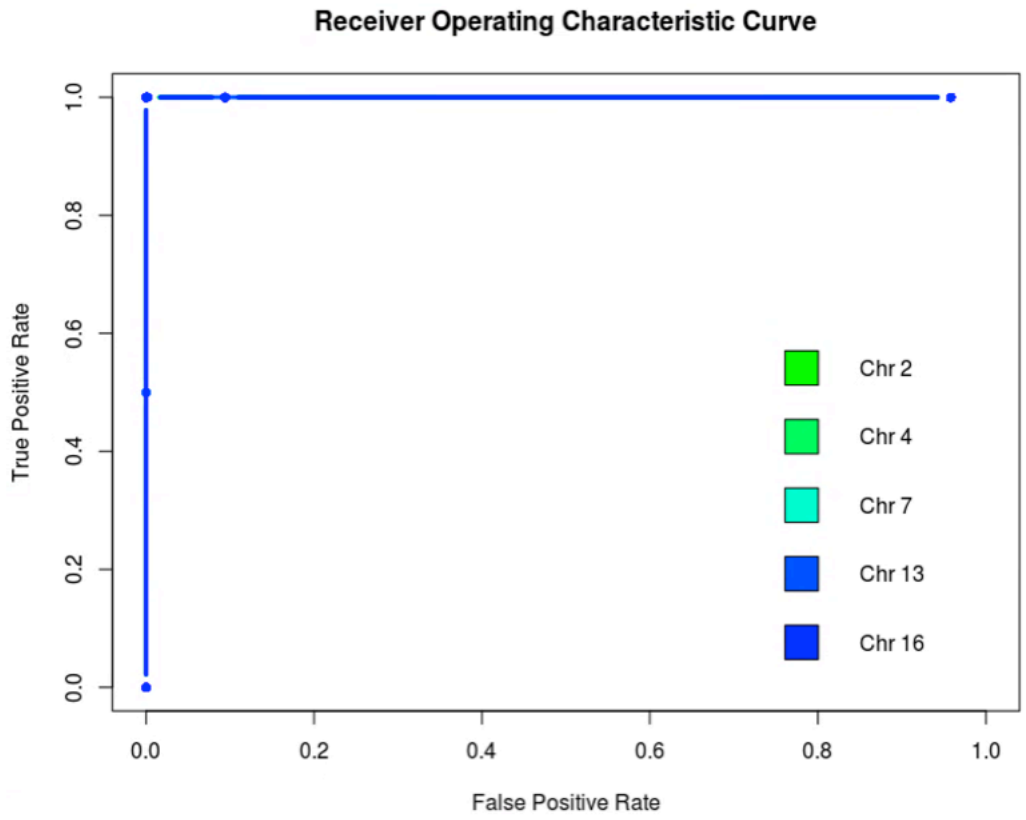

Figure S10. Receiver operating characteristic curves for 5 UPD classifiers in northern European true positives and true negatives from IBD-based UPD detection. We further validated our classifiers using true positives and true negatives from IBD-based UPD detection, specifically, we analyzed northern European true positives with ROH spanning at least 20% of the UPD chromosome and northern European true negatives. All 5 ROC curves shown here have auROC > 0.95 and using our chosen probability cutoff of 0.9, we identified 85% of the northern European true positives (TPR) and we did not classify any of the northern European true negatives as putative UPD cases (FPR).

Figure S11. Ideograms of ROH for 22 ROH-based UPD cases we identified in the UK Biobank. We applied our ROH-based classifiers to 431,094 northern European individuals from the UK Biobank and identified 172 putative cases of UPD across 21 autosomes and the X

chromosome (we did not classify any UPD cases on chromosome 20). This figure shows ideograms of ROH for 22 of the 172 putative cases, randomly drawn to illustrate ROH patterns of UPD cases of each chromosome for which we classify UPD; blue rectangles along the chromosomes represent Class A ROH, red rectangles represent Class B ROH, and green rectangles represent Class C ROH.

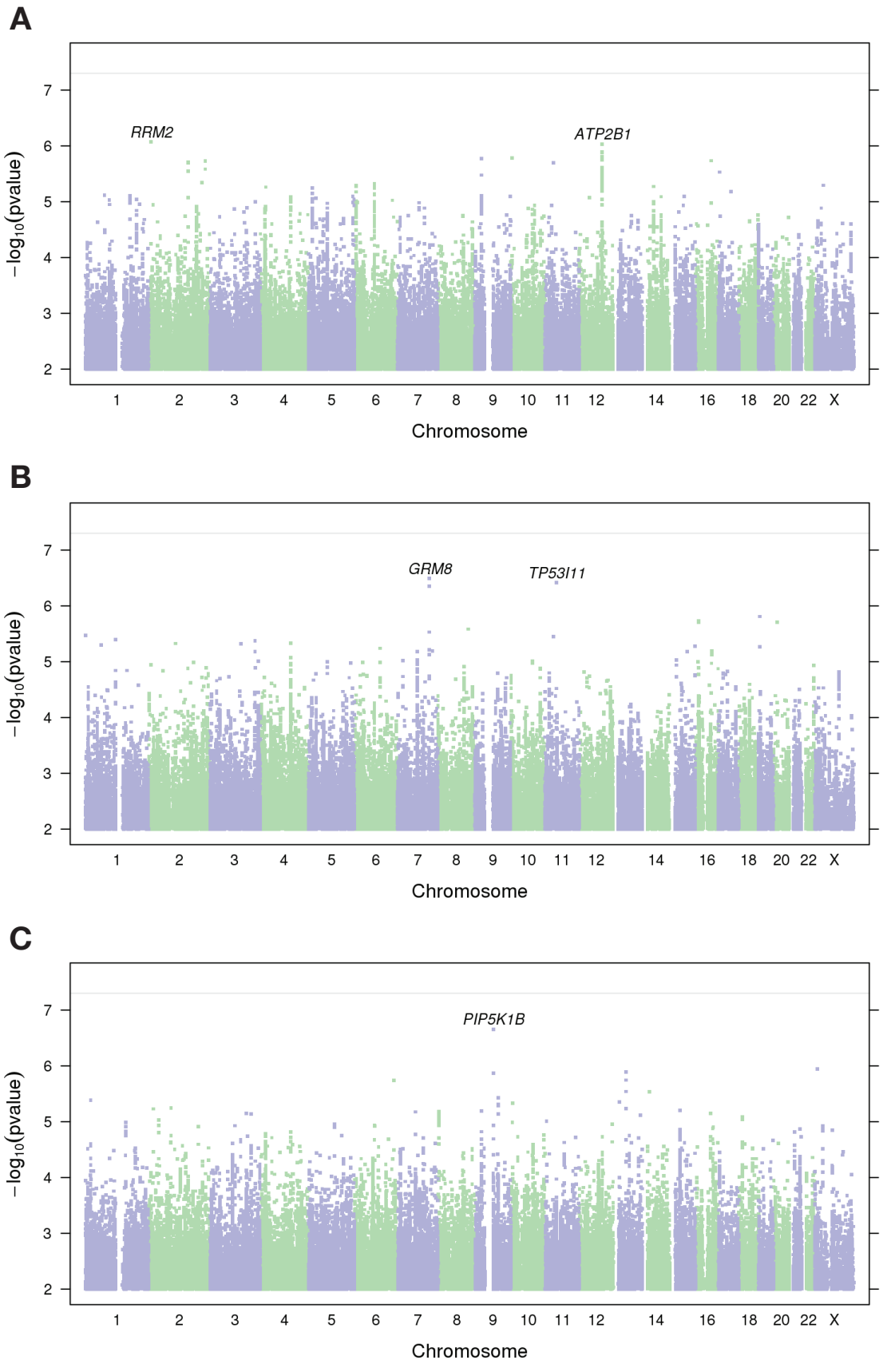
A
RRM2
ATP2B1
−log10(pvalue)
Chromosome
B
GRM8
TP53I11
−log10(pvalue)
Chromosome
C
PIP5K1B
−log10(pvalue)
Chromosome

Figure S12. Manhattan plots of GWAS of UPD incidence, stratified by parental sex. (A) GWAS of all parents of UPD cases (both mothers and fathers), adjusted for sex; (B) GWAS of mothers of UPD cases only; and (C) GWAS of fathers of UPD cases only. There are no variants reaching genome-wide significance and the few hits reaching suggestive association level (*p*-value < $1x10^{-6}$) are likely false positives based on gene annotations.

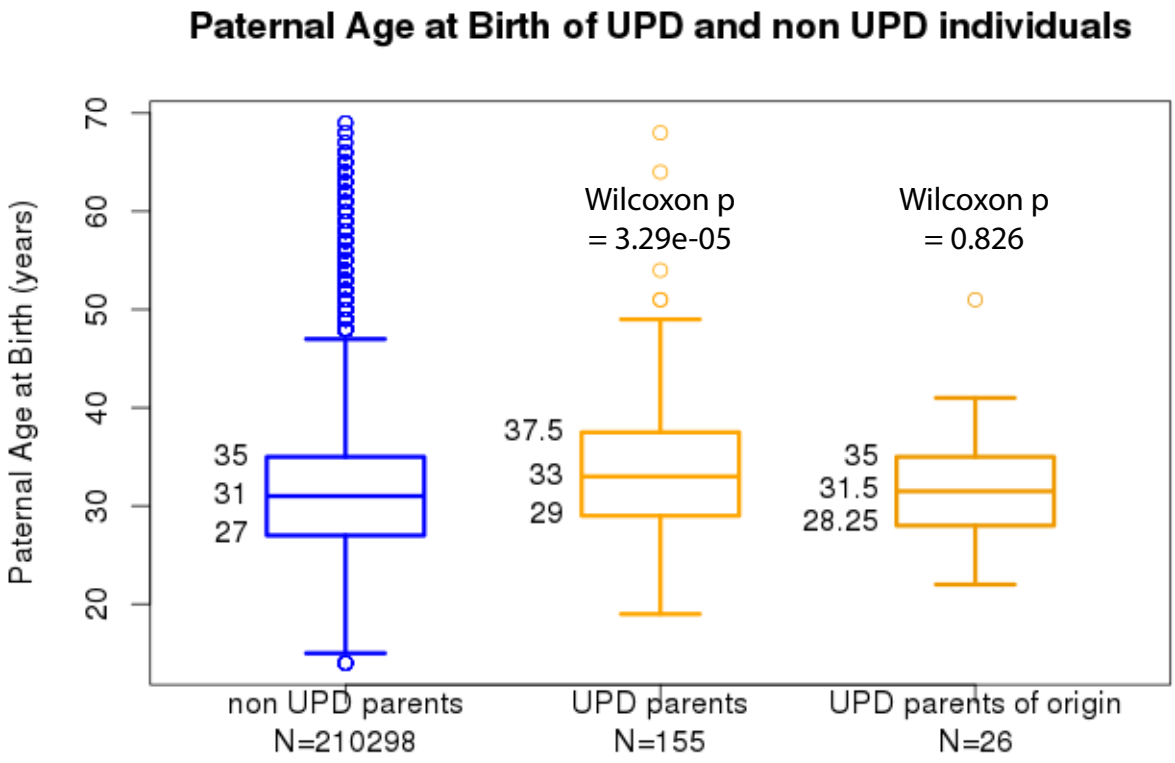

Figure S13. **The age distribution of fathers of UPD true negatives (blue) and that of fathers who are parents of origin of UPD true positives (patUPD cases, yellow) in the 23andMe dataset.** Fathers of UPD cases are significantly older than fathers of true negatives (Wilcoxon *p*-value = 3.29 x 10^{-5}). However, we do not observe a significant difference in

paternal age when we restrict analysis to fathers who are parents of origin of UPD children (Wilcoxon *p*-value = 0.286).

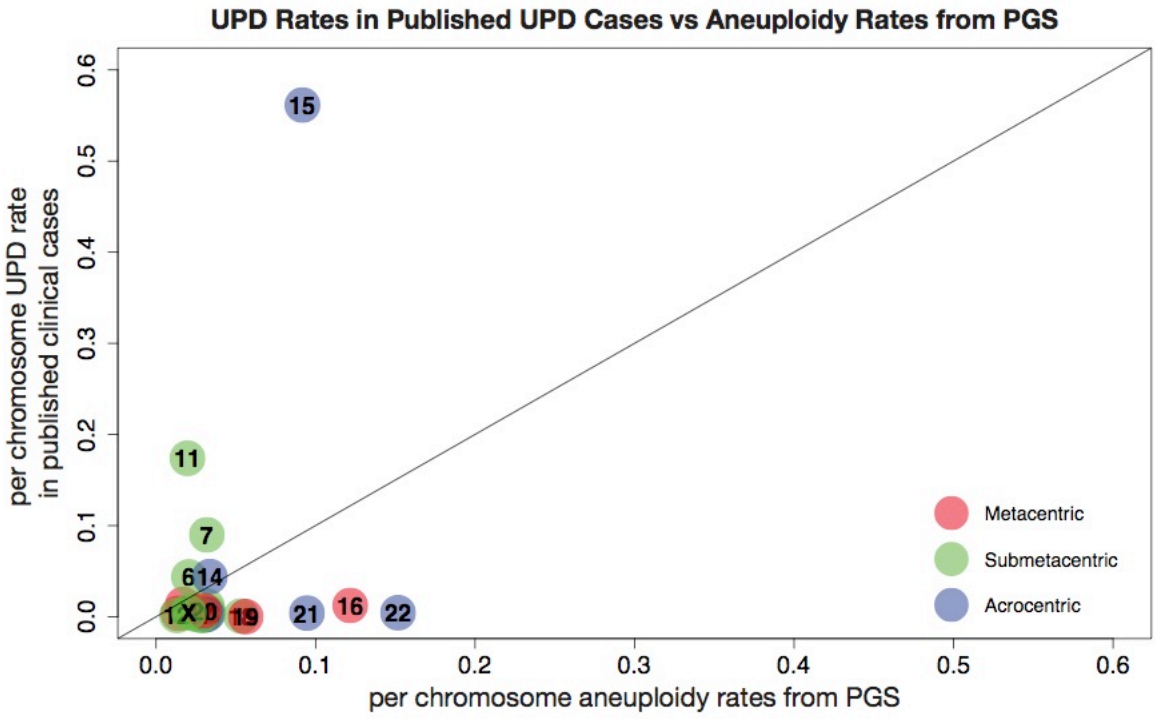

Figure S14. Correlation between per-chromosome UPD rate in published clinical cases[1] **(see Web Resources) and per-chromosome aneuploidy rate in published pre-implantation embryo data.**[4] Chromosomes are colored by centromeric type: metacentric chromosomes are shown in red, submetacentric chromosomes in green and acrocentric chromosomes in blue. In contrast to our results of correlation between per chromosome UPD rates in the 23andMe dataset and those from PGS data (Figure 4A), these two rates are not significantly correlated (Pearson's correlation = 0.2; *p*-value = 0.34). This is expected since clinical cases are likely biased towards chromosomes causing serious medical phenotypes.

Population	Class A Lengths (bp)	Class B Lengths (bp)	Class C Lengths (bp)
Northern European (23andMe)	16,639-476,876	476,877-1,244,859	1,244,867-53,180,213
Northern European (UK Biobank)	2,308-808,025	808,026-2,398,796	2,398,803-137,730,942
Southern European	24,862-554,584	554,589-1,590,178	1,591,210-18,452,311
African American	27,649-591,299	591,321-1,720,202	1,721,681-47,567,224
Ashkenazi Jewish	19,879-571,582	571,595-1,578,587	1,578,882-40,139,153
East Asian	15,640-385,089	385,093-963,677	963,692-17,530,267
Latino	10,032-561,786	561,789-1,616,663	1,616,842-38,806,880
Middle Eastern	30,028-617,344	617,354-1,882,483	1,883,339-53,080,779
South Asian	25,273-585,458	585,514-1,735,929	1,736,607-49,036,678

Table S1. **Boundaries (in base pairs) for each of three length classes of ROH across nine cohorts in the 23andMe database and the UK Biobank.** Class boundaries were calculated using Gaussian mixture modeling on ROH length distributions in UPD true negatives from each of the eight cohorts in the 23andMe dataset and from all northern Europeans in the UK Biobank.

Quantitative phenotype	Mean phenotype in UPD true negatives	Standard deviation (SD) of phenotype in UPD true negatives	Mean phenotype in UPD true positives	SD of phenotype in UPD true positives	Sample sizes in UPD true negatives	Sample sizes in UPD true positives
amount_of_stress	2.58	1.2	2.64	0.9	20790	25
beighton_hypermobility_qtl	2.11	2.1	4	2.8	6182	<5
birthweight	121.11	21.7	124.5	14.8	3441	<5
bmi	25.87	5.6	24.97	4.6	110256	97
bmi_qnorm	-0.2	1	-0.35	0.9	110255	97
cup_size	4.96	1.6	4.5	0.7	3522	<5
dass_any	2.71	3	3.28	3.5	30326	32
empathy_qt	46.18	13.9	47.67	17.9	4795	<5
height	67.87	3.9	67.07	4	110285	97
height_qnorm	0.04	1	-0.15	1.1	110284	97
iqb.age_started_reading	4.08	1.3	4.22	1.2	16867	18
mind_in_eyes_qnorm	0.14	1	-1.09	1	11972	5
perceived_stress_qt	15.75	7.3	15.53	7.9	16103	17
personality_activity	4.69	2	5.18	1.6	16086	11
personality_aesthetics	8.95	2.7	9	2.4	15869	11
personality_agreeableness	24.7	5.8	27	4.9	15856	11
personality_altruism	11.31	2.9	12.45	2.3	15962	11
personality_anxiety	7.55	3.9	8.36	3.9	15913	11
personality_assertiveness	9.42	5.1	9.27	4.3	15953	11
personality_compliance	7.69	2.5	7.82	2.6	16055	11
personality_conscientiousness	24.68	6.2	22.73	5.8	15856	11
personality_depression	3.82	2.2	4.64	2.3	15978	11
personality_extraversion	16.34	7.4	16.55	6.3	15903	11
personality_ideas	14.67	3.3	13.73	3.7	15888	11
personality_neuroticism	15.35	7	16.73	7	15897	11
personality_openness	29.64	6	28.18	5.6	15842	11
personality_order	4.73	2.2	4.36	2.1	16064	11

personality_self_discipline	13.55	3.6	12.36	3.6	15870	11
self_esteem	6.24	1.9	5.57	2.3	18912	21
shoe_size	10.74	1.4	10.83	1.2	20922	18
shoe_size_normalized	-0.08	1	-0.11	0.9	19238	16
shoe_size_qnorm	-0.07	1	-0.07	0.9	19238	16
systemizing_qt	72.19	21.6	44.67	16.3	4983	<5
vocab_tx	-1.73	0.5	-1.69	0.5	12754	14
weight	169.94	42.6	160.3	36.1	110285	97
weight_qnorm	-0.18	1	-0.41	1	110277	97

Table S3. Summary statistics and sample sizes for 36 quantitative phenotypes tested in PheWAS. We tested for association between UPD of each of the chromosomes with at least one UPD case and 36 quantitative phenotypes across five categories (cognitive, personality, morphology, obesity and metabolic traits). Where possible, we also tested for association between matUPD and patUPD of each of the chromosomes separately.

UPD type	Phenotype	Effect Size (95% CI)	P-values (Uncorrected)
patUPD1	Any Bariatric Surgery	3.79 (1.39 6.20)	0.0020
UPD1	Type 2 Diabetes	3.19 (0.87 5.50)	0.0070
UPD3	Hyperglycemia	4.73 (2.32 7.13)	0.0001
UPD3	Type 2 Diabetes	4.04 (1.08 7.00)	0.0076
UPD6	Hyperglycemia	4.38 (1.60 7.15)	0.0020
UPD6	Type 2 Diabetes	3.90 (1.12 6.68)	0.0060
UPD6	**Weight**	**-2.02 (-3.38 -0.65)**	**0.0038**
UPD6	**Height**	**-1.99 (-3.40 -0.59)**	**0.0055**
UPD7	Self-rated attractiveness	-4.20 (-7.08 -1.32)	0.0042
UPD7	Life satisfaction	-3.75 (-6.53 -0.97)	0.0081
UPD8	Birth weight	-5.31 (-8.21 -2.41)	0.0003
UPD8	Autism	3.73 (1.36 6.11)	0.0021
UPD8	Memory Loss	4.19 (1.37 7.02)	0.0036
UPD8	Altitude Sickness	-4.11 (-6.90 -1.32)	0.0038
UPD8	Likes to play with ideas	-8.36 (-14.59 -2.13)	0.0086
UPD8	Autism Spectrum	3.14 (0.78 5.49)	0.0090
matUPD15	Autism Spectrum	5.47 (2.42 8.51)	0.0004
UPD15	Self-rated math ability	-3.53 (-6.09 -0.96)	0.0071
patUPD16	Type 2 Diabetes	7.72 (4.72 10.73)	4.596×10^{-7}
patUPD16	Hyperglycemia	6.05 (3.22 8.88)	2.778×10^{-5}
patUPD16	High Cholesterol	3.29 (0.84 5.74)	0.0084
matUPD21	Feels left out of social activity	-3.52 (-5.93 -1.11)	0.0042
UPD22	**Autism Spectrum**	**3.61 (1.93 5.30)**	**2.557×10^{-5}**

Table S4. Phenotypes significantly associated with UPD of chromosomes 1, 3, 6, 7, 8, 15, 16, 21 and 22 (*p*-value < 0.01). Traits with at least two cases (or two measurements for quantitative traits) are shown in bold. We tested for association between UPD of each of the chromosomes and 208 phenotypes (Tables S2-3) across five categories (cognitive, personality,

morphology, obesity and metabolic traits). Where possible, we also tested for association between matUPD and patUPD of each of the chromosomes separately. Effect sizes shown are odds ratios.

Supplemental Methods

Genome-Wide Association Study

For the genome-wide association study (GWAS) of UPD, we restricted participants to a set of individuals who have European ancestry determined through an analysis of local ancestry described in the Methods section. A maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm.[5] Individuals were defined as related if they shared more than 700 cM IBD, including regions where the two individuals share either one or both genomic segments IBD. This level of relatedness (roughly 20% of the genome) corresponds approximately to the minimal expected sharing between first cousins in an outbred population. When selecting individuals for case/control phenotype analyses, the selection process is designed to maximize case sample size by preferentially retaining cases over controls. Specifically, if both an individual case and an individual control are found to be related, then the case is retained in the analysis.

Imputation panels created by combining multiple smaller panels have been shown to give better imputation performance than the individual constituent panels alone.[6] To that end, we combined the May 2015 release of the 1000 Genomes Phase 3 haplotypes[7] with the UK10K imputation reference panel[8] to create a single unified imputation reference panel. To do this, multiallelic sites with N alternate alleles were split into N separate biallelic sites. We then removed any site whose minor allele appeared in only one sample. For each chromosome, we used Minimac3[9] to impute the reference panels against each other, reporting the best-guess genotype at each site. This gave us calls for all samples over a single unified set of variants. We then joined these

together to get, for each chromosome, a single file with phased calls at every site for 6,285 samples. Throughout, we treated structural variants and small indels in the same way as SNPs.

In preparation for imputation we split each chromosome of the reference panel into chunks of no more than 300,000 variants, with 10,000 variants overlapping on each side. We used a single batch of 10,000 individuals to estimate Minimac3 imputation model parameters for each chunk. To generate phased participant data for the v1 to v4 platforms, we used an internally-developed tool at 23andMe, Inc., Finch, which implements the Beagle graph-based haplotype phasing algorithm[10], modified to separate the haplotype graph construction and phasing steps. Finch extends the Beagle model to accommodate genotyping error and recombination, in order to handle cases where there are no consistent paths through the haplotype graph for the individual being phased. We constructed haplotype graphs for all participants from a representative sample of genotyped individuals, and then performed out-of-sample phasing of all genotyped individuals against the appropriate graph. For the X chromosome, we built separate haplotype graphs for the non-pseudoautosomal region and each pseudoautosomal region, and these regions were phased separately. For the 23andMe participants genotyped on the Illumina Global Screening Array-based platform (see “Genotyping and Quality Control” section), we used a similar approach, but using a new phasing algorithm, Eagle2.[11]

We imputed phased participant data against the merged reference panel using Minimac3, treating males as homozygous pseudo-diploids for the non-pseudoautosomal region.
We computed association test results for the genotyped and the imputed SNPs. We assessed association by logistic regression assuming additive allelic effects. For tests using imputed data, we used the imputed dosages rather than best-guess genotypes. We also included covariates for age, gender, the top five principal components to account for residual population structure,

and indicators for genotype platforms to account for genotype batch effects. The association test *p*-value we report was computed using a likelihood ratio test, which in our experience is better behaved than a Wald test on the regression coefficient. For quantitative traits, association tests were performed by linear regression. Results for the X chromosome were computed similarly, with male genotypes coded as if they were homozygous diploid for the observed allele.

Principal component analysis was performed independently for each ancestry, using ~65,000 high quality genotyped variants present in all five genotyping platforms. It was computed on a subset of one million participants randomly sampled across all the genotyping platforms. PC scores for participants not included in the analysis were obtained by projection, combining the eigenvectors of the analysis and the SNP weights.

Supplemental References

1. Liehr, T. (2010). Cytogenetic contribution to uniparental disomy (UPD). Mol. Cytogenet. *3*, 8.

2. Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., and Li, J.Z. (2012). Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum. Genet. *91*, 275–292.

3. Kang, J.T.L., Goldberg, A., Edge, M.D., Behar, D.M., and Rosenberg, N.A. (2017). Consanguinity Rates Predict Long Runs of Homozygosity in Jewish Populations. Hum. Hered. *82*, 87–102.

4. Rodriguez-Purata, J., Lee, J., Whitehouse, M., Moschini, R.M., Knopman, J., Duke, M., Sandler, B., and Copperman, A. (2015). Embryo selection versus natural selection: how do outcomes of comprehensive chromosome screening of blastocysts compare with the analysis of products of conception from early pregnancy loss (dilation and curettage) among an assisted reproductive technology population? Fertil. Steril. *104*, 1460-1466.e12.

5. Henn, B.M., Hon, L., Macpherson, J.M., Eriksson, N., Saxonov, S., Pe'er, I., and Mountain,

J.L. (2012). Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. PLoS One *7*, e34267.

6. Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.-F., et al. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat. Commun. *6*, 8111.

7. Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

8. Consortium, T.U. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.

9. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.

10. Browning, S.R., and Browning, B.L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. *81*, 1084–1097.

11. Loh, P.-R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816.