Supplementary Material

Supplementary Text

Estimating recent ancestry in admixed populations

Analysis of our European samples demonstrates that ERSA performs well in a homogeneous population with no history of recent admixture from a more distantly related population. Because we do not have pedigree data for an admixed population, we cannot directly evaluate ERSA's performance in the presence of admixture. Impacts of admixture on ERSA's performance would most likely be mediated through effects on the expected distributions of the number and lengths of IBD segments shared between unrelated individuals. Admixture should reduce the number and lengths of such segments. The reasoning for this expected reduction is as follows. The detection of IBD segments is based largely on long runs of consecutive loci at which the genotypes are consistent with identity-by-state (IBS). Admixture will introduce alleles that are frequently IBS among pairs of individuals in the population due to shared ancestry. However, in the absence of founder effect, given that two admixed individuals are of identical ancestry at a particular genomic segment, they are no more likely to share long runs of IBS than individuals chosen at random from the appropriate reference population. When individuals are not required to share ancestry at any particular genomic segment (as would be the case for ascertainment for a shared genetic disease), it results in an expectation of fewer and smaller shared segments among unrelated individuals relative to at least one of the reference populations.

We evaluated this prediction by comparing individuals from a sample of 25 Bolivian individuals genotyped on Affymetrix SNP 6.0 arrays (Xing *et al.* 2010). We identified substantial European admixture (19-41%; data not shown) in 9 Bolivians using the Admixture software (Alexander *et al.* 2009) and divided the Bolivian population into groups with and without admixture. All non-admixed Bolivians were estimated to have < 0.1% admixture. We then applied the same process used to identify shared segments in our European sample, *i.e.*, using Beagle (Browning and Browning 2010) to phase and impute the data and GERMLINE (Gusev *et al.* 2009) to identify all shared segments longer than 2.5 cM. On average, the admixed Bolivians shared 43 segments (95% C.I. 41-45 segments) with an average size of 3.5 cM (95% C.I. 3.4-3.7 cM), compared to 88 segments (95% C.I. 86-92 segments) with an average size of 4.2 cM (95% C.I. 4.1–4.3 cM) in non-admixed Bolivians, consistent with our predictions.

In comparisons of distantly-related admixed individuals, the smaller expected number and size of background segments could slightly improve ERSA's detection power: short but meaningful shared IBD segments could become statistically significant when compared to a shorter background size distribution. In comparisons of distantly-related individuals with ancestries mostly confined to one of the reference populations, however, the admixed population background distributions would be incorrect. Using them might cause ERSA to suffer a slightly increased false positive rate or a bias towards overestimating the degree of relationship due to the misattribution of some short background segments to a distant relationship.

Supplemental Methods

Inferring first-degree relationships

Many existing methods for detecting IBD segments do not distinguish segments that overlap on homologous chromosomes, and rather than consider them to be separate, merge them into one (see Figure S5). For two or more degrees of relationship, Eqs. 7 and 8 provide close approximations to the results of this procedure (Thomas *et al.* 2008). However, in the case of full siblings, Eq. 7 systematically overestimates the number of detected shared segments, and Eq. 8 systematically underestimates the length of the merged segment. Therefore, for $d = 2$ and $a = 2$, we adjust the calculation for N_A and F_A to account for shared segments that have been bioinformatically merged:

S1. $$N_A(n \mid d=2, a=2) = \frac{e^{-\frac{3}{4}c+2dr\frac{3}{4}\cdot\frac{1}{4}}\left[\frac{3}{4}c+2dr\frac{3}{4}\cdot\frac{1}{4}\right]^n}{n!}.$$

S2. $$F_A(i \mid d=2, a=2, t) = \frac{(i-t)^{\hat{k}-1} e^{-d(i-t)/100}}{(100/d)^{\hat{k}}(\hat{k}-1)!},$$

where $\hat{k}$ is the maximum likelihood estimate for the number of merged segments. Because Eq. S2 introduces additional estimated parameters into the full-sibling model, ERSA only reports the full-sibling model as the maximum likelihood estimate if it is significantly more likely than all other models at the 0.05 level.

Many existing IBD methods are also unable to detect the recombination breakpoints between parent-offspring pairs and usually report the length of each entire chromosome as a shared segment (Gusev *et al.* 2009; Thomas *et al.* 2008). With this detection scheme, a probabilistic description of the number and size of shared segments is no longer appropriate. Therefore, to identify parent-offspring relationships, we consider a different statistic, the total proportion of the genome shared between the two individuals. We reject a sibling relationship in favor of a parent-offspring relationship when the proportion of the genome shared exceeds a specified significance level for siblings (default is 0.01). ERSA includes options to bypass Eqs. S1, S2, and/or the parent-offspring option for situations where the overlapping segments can be accurately identified.

Parameters *d* and *a* in the likelihood ratio test

Although d and a are specified as two separate parameters in the likelihood ratio test, our analyses indicate that allowing a to vary has almost no effect on the distribution of likelihood scores under the null hypothesis. To demonstrate this behavior, we evaluated the likelihood scores for pairs of individuals from two closely-related populations, the CHB (45 Han Chinese in Beijing) and JPT (45 Japanese in Tokyo) samples, using the HapMap phase 2 SNP genotype data (HapMap Consortium 2005). For each pair of individuals, we calculated the maximum likelihood for two alternative models (L_1 and L_2). In model 1, a is allowed to vary, and in model 2, a is fixed equal to 2 (d is estimated in both). To evaluate the effect of allowing a to vary, we calculated a likelihood ratio test (LRT) statistic for the two models ($-2\ln[L_1/L_2]$; Figure S6, blue "Observed" line). For comparison, we also plot the expected cumulative distribution of a χ^2 with one degree of freedom (red). As the cumulative distribution illustrates, all of the observed LRT values are less than 10^{-8}, indicating that there is very little difference between the likelihoods of the two models. Thus d and a can be treated as a single parameter when applying the χ^2 approximation to the likelihood ratio test statistic.

Proof of Equation 9

Given a set of shared segment lengths between two individuals, s, we are interested in identifying the subset of these segments, m, containing the n_A elements that are most likely to have been inherited from recent ancestor(s). Eq. 9 assumes that that m is equal to the largest n_A elements in s. Here, we show why this assumption holds. Let $\theta_1 = 100/d$, which is the expected length of a shared segment inherited from a recent ancestor. Let $\theta_2 = \theta$, which is the expected length of a shared segment if it is not inherited from a recent ancestor. If the average time to the most recent common ancestor between individuals in the population is greater than $d/2$, then $\theta_1 > \theta_2$. If $\theta_1 < \theta_2$, then individuals selected at random from the population are more closely related than the relationship we are trying to detect, and therefore there is no power to detect a relationship.

To demonstrate that m is equal to the set containing the largest n_A elements of s, consider two mutually exclusive subsets of s, z_P and z_A, with z_A containing n_A elements. Let x_1 equal the largest element in z_P and x_2 equal the smallest element in z_A. Let y_P and y_A respectively equal the sets z_P and z_A, with the exception that x_1 and x_2 are swapped. As long as $x_1 > x_2$, the likelihood of z_P and z_A is less than the likelihood of y_P and y_A:

$$L_R(n_p, n_a, y_a, y_p \mid d, a, t) < L_R(n_P, n_a, z_A, z_P \mid d, a, t).$$

The components of L_R are N_A, N_P, S_A, and S_P. Because N_A and N_P depend only on n_P and n_A, the above condition simplifies to:

$$S_P(y_P \mid t) S_A(y_A \mid d, a, t) < S_P(z_P \mid t) S_A(z_A \mid d, a, t).$$

The elements in both z_P and z_A, and y_P and y_A are equal, with the exception of x_1 and x_2. Therefore, by Eq. 6, the inequality becomes

$$F_P(x_2 \mid t) F_A(x_1 \mid d, a, t) < F_P(x_1 \mid t) F_A(x_2 \mid d, a, t),$$

which (by Eqs. 3 and 7) is equal to

$$\frac{1}{\theta_1} e^{-\frac{x_2}{\theta_1}} \frac{1}{\theta_2} e^{-\frac{x_1}{\theta_2}} < \frac{1}{\theta_1} e^{-\frac{x_1}{\theta_1}} \frac{1}{\theta_2} e^{-\frac{x_2}{\theta_2}}.$$

This simplifies to

$$\frac{x_2 - x_1}{\theta_2} < 0 < \frac{x_1 - x_2}{\theta_1}.$$

Q.E.D.

Supplementary Figure Legends

Figure S1: ERSA's power and accuracy for one-ancestor relationships. Figures 3 and 4 in the main text display results for all known two-ancestor relationships in the pedigree where the two inheritance paths are the same length, such as full siblings and full cousins. This figure displays the equivalent results for all relationships with exactly one known one-ancestor relationships, i.e. half siblings and half cousins. (**A**) Known vs. estimated degree of relationship. (**B**) Number of pairs in the pedigree with the specified known degree of relationship. (**C**) Power to detect a significant relationship at the $\alpha = 0.001$ significance level plotted against the maximum theoretical power, calculated from Eq. 7 with $a = 1$ and $t = 0$).

Figure S2: Known vs. estimated degree of relationship for individuals that share exactly two common ancestors and where both paths connecting the pair have the same length, using (**A**) ERSA with $\alpha = 0.05$ based on IBD segments estimated by GERMLINE (Gusev *et al.* 2009) IBD segments; (**B**) ERSA with $\alpha = 0.001$ and GERMLINE IBD segments (same as Figure 3 of the main text); (**C**) ERSA with $\alpha = 0.05$ and Beagle 3.3 fastIBD (http://faculty.washington.edu/browning/beagle/beagle.html) segments; (**D**) GBIRP and 10,028 evenly-spaced SNPs with MAF > 0.4, with a LOD threshold of 2.34 for significance (as in Stankovich *et al.* 2005); and (**E**) RELPAIR with 9,990 evenly-spaced

SNPs and requiring a likelihood ratio > 10 for significance (the default in RELPAIR; Epstein *et al.* 2000). (**F**) The number of pairs in each relationship class. For GBIRP analysis, SNP data was thinned (following Berkovic *et al.* 2008) after phasing and imputation as described in the main text Methods, then written to GBIRP-readable .data format files (fdist, ffreq, fhaplos, and fLastMarkers; http://bioinf.wehi.edu.au/software/GBIRP/main.html), with allele frequencies estimated from the entire sample of 169 individuals. GBIRP analyses were performed with various numbers of markers (from 1,000 to 50,000) with different minimum MAF values (from 0.1 to 0.4); the optimal results are shown.

Figure S3: Performance of ERSA's nominal (**A**) 95% and (**B**) 99% confidence intervals (C.I.). The proportion of pairs for which the nominal C.I. contains the known value is plotted vs. the known relationship (degree of relationship for a pair of individuals that share two common ancestors, where both paths through those ancestors have the same length, with $a = 2$).

Figure S4: Realized vs. expected sums of shared IBD segment lengths between pairs of related individuals sharing exactly two ancestors. The dotted lines enclose the middle 90% of observed values. The expectation for the sum of IBD segment lengths (dashed line) is adjusted to account for the fact that IBD segments detected by GERMLINE do not distinguish between haploid and diploid sharing and for the expected overlap of IBD segments in siblings.

Figure S5: Bioinformatic merging of shared segments in full siblings. Two homologous autosomal chromosomes are shown for two parents, each colored differently. Meiosis and recombination occurs and two sibling offspring inherit recombinant chromosomes. Although the siblings share three distinct IBD segments, two of these segments overlap and are thus merged bioinformatically (by GERMLINE or BEAGLE) into a single shared segment (black bar, far right). Eq. S1 and S2 account for this process of bioinformatic merging.

Figure S6: The effect of allowing *a* to vary under the null model. The cumulative probability for values of the observed LRT statistic comparing models with *a* free to vary or fixed equal to 2 is shown in blue. The cumulative distribution for a χ^2 distribution with one degree of freedom is shown in red for comparison.

Supplementary References

Alexander, D.H., Novembre, J., and Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19:** 1655-1664.

Berkovic, S.F., Dibbens, L.M., Oshlack, A., Silver, J.D., Katerelos, M., Vears, D.F., Lullmann-Rauch, R., Blanz, J., Zhang, K.W., Stankovich, J. *et al.* 2008. Array-based gene discovery with three unrelated subjects shows SCARB2/LIMP-2 deficiency causes myoclonus epilepsy and glomerulosclerosis. *Am J Hum Genet* **82:** 673-684.

Browning, S.R. and Browning, B.L. 2010. High-Resolution Detection of Identity by Descent in Unrelated Individuals. *The American Journal of Human Genetics* **86:** 526-539.

Epstein, M.P., Duren, W.L., and Boehnke, M. 2000. Improved Inference of Relationship for Pairs of Individuals. *The American Journal of Human Genetics* **67:** 1219-1231.

Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. 2009. Whole population, genome-wide mapping of hidden relatedness. *Genome Research* **19:** 318-326.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299-1320.

Stankovich, J., Bahlo, M., Rubio, J.P., Wilkinson, C.R., Thomson, R., Banks, A., Ring, M., Foote, S.J., and Speed, T.P. 2005. Identifying nineteenth century genealogical links from genotypes. *Hum Genet* **117:** 188-199.

Thomas, A., Camp, N.J., Farnham, J.M., Allen-Brady, K., and Cannon-Albright, L.A. 2008. Shared Genomic Segment Analysis. Mapping Disease Predisposition Genes in Extended Pedigrees Using SNP Genotype Assays. *Annals of Human Genetics* **72:** 279-287.

Xing, J., Watkins, W.S., Shlien, A., Walker, E., Huff, C.D., Witherspoon, D.J., Zhang, Y., Simonson, T.S., Weiss, R.B., Schiffman, J.D. *et al.* 2010. Toward a more uniform sampling of human genetic diversity: A survey of worldwide populations by high-density genotyping. *Genomics* **96:** 199-210.

Supplementary Tables

Table S1: Data of Figure S2 and Figure 3.

ERSA + GERMLINE, $\alpha = 0.05$

	Known degree of relationship														
Estimated degree	1	2	3	4	5	6	7	8	9	10	11	12	13	14	None known
None detected						6	14	53	180	263	339	334	103	6	6584
9						10	20	15	63	48	36	10	7		133
8					1	25	41	39	94	64	28	16	8	1	184
7					16	75	65	38	38	15	4	1			25
6					102	126	28	6	4	1					3
5				28	164	29									1
4		1	19	85	7										
3		3	75	4											
2	3	23													
1	12	5	1												

ERSA + GERMLINE, $\alpha = 0.001$ (data of Figure 3 of the main text)

	Known degree of relationship														
Estimated degree	1	2	3	4	5	6	7	8	9	10	11	12	13	14	None known
None detected						10	21	57	213	296	360	350	110	7	6829
9						7	15	14	44	34	23	5	4		33
8					1	24	39	36	80	46	20	5	4		46
7					16	75	65	38	38	14	4	1			18
6					102	126	28	6	4	1					3
5				28	164	29									1
4		1	19	85	7										
3		3	75	4											
2	3	23													
1	12	5	1												

ERSA + BEAGLE, $\alpha = 0.001$

	Known degree of relationship														
Estimated degree	1	2	3	4	5	6	7	8	9	10	11	12	13	14	None known
None detected		2	2		17	64	74	105	323	360	397	361	118	7	6907
9						4	4	4	4	7	2				5
8					3	17	27	18	22	13	7				8
7		1			14	55	39	15	25	11	1				5
6		1		1	48	87	22	8	5						3
5				7	137	39	2	1							2
4			3	68	71	5									
3	3		68	41											
2	12	28	22												
1															

GBIRP, LOD > 2.34

	Known degree of relationship														
Estimated degree	1	2	3	4	5	6	7	8	9	10	11	12	13	14	None known
None detected		1		4	63	149	127	123	353	378	405	359	116	7	6905
9															
8					2	19	10	9	12	6	2	2	2		18
7		1	2	1	33	47	23	15	14	7					6
6		2	3	14	120	50	8	4							
5			15	74	68	6									1
4	1	5	62	24	4										
3	14	23	13												
2															
1															

RELPAIR, likelihood ratio > 10

	Known degree of relationship														
Estimated degree	1	2	3	4	5	6	7	8	9	10	11	12	13	14	None known
None detected					40	164	150	147	376	391	405	361	118	7	6924
3+			90	117	250	107	18	4	3		2				6
2		20	2												
1	15	12	3												

Table S2: Number of pairs in each relationship degree class (data of lower panel of Figure 3 in the main text.)

Known degree of relationship	1	2	3	4	5	6	7	8	9	10	11	12	13	14	None known
Number of pairs	15	32	95	117	290	271	168	151	379	391	407	361	118	7	6930

Table S4: Percent detection power for various methods (data of Figure 3 in the main text)

Degree of relationship known	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Maximum Theoretical Power	100	100	100	100	100	99.98	99.14	92.94	76.85	55.08	35.25	20.91	11.83	6.5
ERSA + GERMLINE, *a* = 0.05	100	100	100	100	100	97.79	91.67	64.9	52.51	32.74	16.71	7.48	12.71	14.29*
ERSA + GERMLINE, *a* = 0.001	100	100	100	100	100	96.31	87.5	62.25	43.8	24.3	11.55	3.05	6.78	0
ERSA + BEAGLE, *a* = 0.001	100	93.75	97.89	100	94.14	76.38	55.95	30.46	14.78	7.93	2.46	0	0	0
GBIRP	100	96.88	100	96.58	78.28	45.02	24.4	18.54	6.86	3.32	0.49	0.55	1.69	0
RELPAIR	100	100	100	100	86.21	39.48	10.71	2.65	0.79	0	0.49	0	0	0

* For very distant relationships, estimated power sometimes exceeds the maximum expected power. This is likely due to the existence of some undocumented distant relationships, since our pedigrees are not complete at such depths, as well as to false positive results.

Table S5: ERSA + GERMLINE, α = 0.001, one-ancestor model and data set (data of Figure S1)

	Known degree of relationship											
Estimated degree	1	2	3	4	5	6	7	8	9	10	11	None known
None detected							14	57	50	38	7	6826
9							6	13	13	6	1	33
8						4	24	27	17	6	1	45
7					5	29	58	34	12	2		22
6					16	59	29	4				2
5				2	44	21	1					
4				4	2							
3		3	2									
2	1	4										
1	10											
Number of Pairs	11	7	2	6	67	113	132	135	92	52	9	6930
Estimated Power	100	100	100	100	100	100	89.39	57.78	45.65	26.92	22.22	

Table S6: Estimates of significant recent ancestry ($\alpha = 0.001$) among pairs of parent individuals in the HapMap CEU dataset.

Individual 1	Individual 2	Estimated number of shared ancestors (a)	Estimated degree of relationship	99.9% Confidence Interval for the degree of relationship		–lnL(Related)	–lnL(Unrelated)
				$a = 2$	$a = 1$		
NA12154	NA12892	2	9	6-21	6-21	12.90	19.98
NA06985	NA12812	1	7	5-13	5-13	23.86	67.49
NA06993	NA07022	2	4	3-6	3-6	81.95	499.50
NA11995	NA12145	2	8	5-16	5-16	16.74	26.85
NA11840	NA12717	2	8	6-16	5-16	15.70	30.77
NA12056	NA12872	2	8	5-13	5-13	18.67	27.12
NA07034	NA12145	1	9	6-19	5-19	16.33	37.98
NA12146	NA12812	2	8	5-19	5-19	21.11	30.25
NA11881	NA12762	2	8	5-17	5-17	14.62	23.63
NA06993	NA07056	2	4	3-6	3-6	85.14	510.44
NA11993	NA12239	2	8	6-18	5-18	17.78	27.13
NA11829	NA12815	2	7	5-13	5-13	22.46	32.26
NA07034	NA11882	2	6	5-8	4-8	33.72	139.83
NA07000	NA12057	2	8	5-18	5-18	23.27	42.08
NA12155	NA12264	2	4	3-5	3-5	103.79	631.83
NA12006	NA12155	2	9	6-20	6-20	10.12	19.43
NA07034	NA12750	2	8	5-19	5-19	20.75	41.10
NA12236	NA12716	1	9	5-17	5-17	18.32	60.64
NA06994	NA07000	1	9	6-17	5-18	13.29	49.92
NA07022	NA07056	2	8	5-18	5-18	19.80	35.36
NA12043	NA12760	2	8	6-18	5-18	12.42	19.73
NA11994	NA12146	2	8	5-19	5-19	15.21	24.71
NA06994	NA12892	2	5	4-7	4-6	65.19	296.69