Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives

Noah A. Rosenberg

Department of Human Genetics, Bioinformatics Program, and the Life Sciences Institute, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Ave., Ann Arbor, MI 48109-2218 USA, (734) 615-9556 (phone), (734) 615-6553 (fax), rnoah@umich.edu

Summary

The HGDP-CEPH Human Genome Diversity Cell Line Panel is a widely-used resource for studies of human genetic variation. Here, pairs of close relatives that have been included in the panel are identified. Together with information on atypical and duplicated samples, the inferred relative pairs suggest standardized subsets of the panel to be employed in future population-genetic studies.

Introduction

The HGDP-CEPH Human Genome Diversity Cell Line Panel (henceforth the "diversity panel") is a collection of 1064 DNA samples from individuals distributed around the world (Cann *et al.* 2002). The DNA samples in the diversity panel are publicly available for studies of genetic variation, and they now form the basis for a sizeable body of human genetics research (Cavalli-Sforza, 2005).

Analyses of the diversity panel performed since the initial article of Cann *et al.* (2002) have revealed much information that is of use to investigators who are currently designing studies that utilize this valuable resource. Here, descriptions are provided of atypical (and potentially mislabeled) DNAs, duplicated DNAs, and pairs of close relatives included in the diversity panel. The likely mislabelings and duplicates have been previously known (Rosenberg *et al.* 2002; Mountain & Ramakrishnan, 2005), and the relative pairs are reported here for the first time.

Three standardized subsets of the original diversity panel are recommended here for future applications of the panel in most types of population-genetic studies. For convenience these subsets are denoted H1048, H971 and H952. H1048 contains no duplicates or individuals that are extremely atypical for their populations, H971 additionally contains no two individuals with a first-degree relationship (parent/offspring or full siblings), and with a few possible exceptions, H952 further contains no two individuals with a second-degree relationship (half siblings, avuncular, or grandparent/grandchild).

To explain how the standardized data sets have been constructed, I begin from a set of 1066 samples – the 1064 in the diversity panel, and two from outside the panel – each of which has been genotyped for one or more genome-wide sets of loci by the Mammalian Genotyping Service at the Center for Medical Genetics, Marshfield Medical Research Foundation. Based on the collections of samples that have been excluded from consideration in various settings, the 1066 samples can be viewed as consisting of nine disjoint subsets (Supplementary Tables 1 and 2).

Atypical and duplicated samples

Atypical samples

Among 1056 samples that we analyzed previously (Rosenberg *et al.* 2002), we identified two samples with genotypes that were extremely atypical for their populations (Supplementary Table 1). For these two samples, it is likely that mislabeling or DNA contamination occurred between the time of sample collection and the time of assembly of the diversity panel. Similar analysis of all 1066 samples has not suggested that mislabeling or contamination occurred in the remaining samples not included in the Rosenberg *et al.* (2002) study (results not shown).

Duplicated samples

Duplicates among the samples were first noticed by Joanna Mountain and James Weber, who independently identified 13 pairs with a high degree of allele sharing. These duplicates were initially reported as personal communications to Howard Cann and were later published by Mountain & Ramakrishnan (2005). Separate analysis of the genotypes from Rosenberg *et al.* (2002) using the proportion-of-shared-alleles (PSA) distance (Mountain & Cavalli-Sforza, 1997)

reveals the same duplicate pairs as those reported in Mountain and Ramakrishnan (2005): the 13 pairs of individuals in Supplementary Table 3 have PSA distance <0.02, whereas no other pair, among 557,040 total pairs, has PSA distance <0.20. As with the likely mislabelings, consideration of all 1066 samples whose genotypes were available does not yield any additional duplicates (results not shown). Note that although pairs with unusually low PSA distance are described as duplicate samples, sample duplications are indistinguishable from monozygotic twins. If genotypic differences between samples are to be attributed specifically to genotyping error or to mutation, it is important whether duplicates are sample duplications or twins. Laboratory duplication seems a more likely explanation in view of the low prevalence of monozygotic twinning worldwide, the care taken in recruiting individuals by the diversity panel investigators, and the various opportunities for error after collection.

Construction of recommended subset H1048

Exclusion from the 1064 samples in the diversity panel of the two atypical samples and of one member of each duplicate pair – both members, for one instance in which the duplicates had different population labels – yields the subset H1048, consisting of 1048 samples (Supplementary Table 4). This subset of the diversity panel is the one considered by Rosenberg *et al.* (2005), and with the exception that Ramachandran *et al.* (2005) excluded the Surui, it is also the subset utilized by Ramachandran *et al.* (2005).

Relative Pairs

The existence of pairs of relatives in the diversity panel was noted by Cann *et al.* (2002) for four populations (Karitiana, Maya, Pima, and Surui), with specific reports about which individuals were related (Mountain & Ramakrishnan, 2005; Howard Cann, pers. comm.). The hierarchical population structure analysis of Ekins *et al.* (2006) further suggested the presence in the diversity panel of many additional groups of related individuals.

To search systematically for relative pairs, for each of the 548,628 pairs of individuals in H1048, allele sharing and RELPAIR 2.0.1 (Boehnke & Cox, 1997; Epstein *et al.* 2000) were employed together with the genome-wide microsatellite genotypes studied by Rosenberg *et al.* (2002), Ramachandran *et al.* (2005), and Rosenberg *et al.* (2005). The formal RELPAIR analysis was used to verify first-degree relationships obtained from the exploratory allele-sharing analysis, as well as to identify higher-order relationships.

Allele-sharing analysis

For each pair of individuals, the proportions of the loci at which the individuals shared 0, 1, and 2 alleles identical in state (IIS) – denoted p_0, p_1, and p_2, respectively – were determined. Among the 783 loci considered by Ramachandran *et al.* (2005) and Rosenberg *et al.* (2005), only loci for which neither individual was missing genotypes were included.

Low values of p_0 indicate likely parent/offspring pairs, because in parent/offspring pairs, p_0 can differ from 0 only as a result of genotyping errors or mutations. In these data, as can be inferred from the level of allele sharing among duplicate samples (Supplementary Table 3), error and mutation have a combined rate of no more than approximately 0.01. The 69 pairs with the smallest values of p_0 were hypothesized to be parent/offspring pairs. Of these pairs, the 64 with

the smallest p_0 appeared to be clear parent/offspring pairs, with $p_0<0.012$. The next 5 pairs all involved African individuals, with $p_0<0.026$ and $p_1>0.73$ for each pair. Given the high heterozygosity in Africa in this data set (Rosenberg *et al.* 2002; Ramachandran *et al.* 2005), it is unlikely for a pair of African individuals to have such a large value of p_1 without being close relatives. The 70th pair had $p_0=0.035$, an improbable value for a parent/offspring pair, given a combined genotyping error and mutation rate below ~0.01. This pair was a pair of Pima individuals, and had $(p_0, p_1, p_2) = (0.035, 0.457, 0.508)$. As 15 Pima pairs were among the 69 pairs with smallest p_0, and all of these had $p_0<0.007$ and $p_1>0.51$, it was concluded that the individuals in this 70th pair were not likely to be parent and offspring, although they were likely to be relatives. Indeed, the high value of p_2 suggested that this pair of individuals, Pima 1048 and 1050, was a full sib pair. Of the 69 hypothesized parent/offspring pairs in the diversity panel, 31 were in populations for which the existence of pairs of close relatives had not previously been known.

Large values of p_2 indicate likely full sib pairs: because full sibs share both alleles at a locus identically by descent (IBD) for 25% of loci on average, p_2 is likely to be at least 0.25 to 0.30 for full sibs – greater in populations with high homozygosity, due to the increased likelihood for alleles to be shared IIS without being IBD. Excluding the Native Americans, who are more homozygous (Rosenberg *et al.* 2002; Ramachandran *et al.* 2005), and the previously hypothesized parent/offspring pairs, there were 18 pairs with $p_2>0.34$ and no other pairs with $p_2>0.26$. These 18 pairs were hypothesized to be full sib pairs.

Because of their greater homozygosity, in Native Americans, p_2 must be larger for inference of a full sib relationship. In the Colombian population, among pairs not hypothesized to have a parent/offspring relationship, one had $p_2=0.43$, and no others had $p_2>0.35$; in Maya, one such pair had $p_2=0.42$, and no others had $p_2>0.28$; in Pima, six pairs had $p_2>0.42$, and no others had $p_2>0.33$. These eight pairs were also hypothesized to be full sib pairs.

In Karitiana and Surui, homozygosity is larger than in the other Native American populations (Rosenberg *et al.* 2002; Ramachandran *et al.* 2005). The overall level of relationship is also thought to be greater, so that p_2 must be larger than in other Native Americans for inference of full sib relationships. In Karitiana, six pairs not hypothesized to have a parent/offspring relationship had $p_2>0.49$, and no others had $p_2>0.43$. In Surui, 14 such pairs had $p_2>0.48$, and no others had $p_2>0.44$. These 20 pairs were thus hypothesized to be full sib pairs.

In summary, the allele-sharing analysis suggested 69 parent/offspring and 46 full sib pairs. The 864 pairs with the smallest values of p_0 and the 669 pairs with the largest values of p_2 each involved a pair of individuals from the same population, and no inter-population pair had $p_0<0.25$ or $p_2>0.24$. It was therefore determined to be improbable that any pair of close relatives had different population labels. Consequently, the RELPAIR analysis proceeded by searching for relative pairs separately within each of the predefined populations.

RELPAIR analysis

Identification of relative pairs via the software package RELPAIR uses a Markov chain on underlying states of IBD status, proceeding sequentially along chromosomes to evaluate the probability of the set of genotypes for a pair of individuals, conditional on their relationship, known allele frequencies in their population, and a known genotyping error rate (Boehnke & Cox, 1997; Epstein *et al.* 2000). The error rate can be viewed as subsuming mutations, although the effects of error and mutation on the probability of a genotype configuration for a given level

of relationship are not strictly equivalent. Eight different relationships are examined by RELPAIR: monozygotic twins (MZ), full siblings (FS), parent/offspring (PO), half siblings (HS), grandparent/grandchild (GG), avuncular (AV), first cousin (CO), and "unrelated" (UN). If the likelihood of one of these relationships exceeds the likelihood of each of the others by a multiplicative factor greater than a predefined critical value, the pair of individuals is inferred to have that relationship.

In the RELPAIR analysis, 772 autosomal microsatellite genotypes were used, a subset of the 783 considered in the allele-sharing analysis. RELPAIR makes use of genetic map positions, whereas allele sharing does not require this information. Thus, each of the 11 loci excluded from the RELPAIR analysis was omitted as a result of either an uncertainty in its map position, or of an error that led to a failure to record the map position (Supplementary Table 5).

The putative relationship was set to "unrelated" for all pairs of individuals. Pairs for which the inferred relationship differed from "unrelated" were identified, as were pairs for which it was not possible to confidently infer a specific relationship because two or more distinct relationships (other than "unrelated") had high likelihoods. For each pair of individuals, allele frequencies were set to the count estimates in their predefined population. The genotyping error rate was set to 0.008, as this was close to the average PSA distance across the 13 duplicate pairs for the 377 loci in the Rosenberg *et al.* (2002) data (Supplementary Table 3). The critical value was set to 100.

The relationships inferred via RELPAIR for each of the geographic regions in Rosenberg *et al.* (2002) are summarized in Supplementary Tables 6-12, with separate tables for some Native American populations in which large numbers of relative pairs were identified (Supplementary Tables 13-15). Other than a few discrepancies in Karitiana and Surui, the RELPAIR analysis agreed precisely with the hypotheses based on allele-sharing analysis for parent/offspring and full sib relationships (Supplementary Table 16). In Karitiana and Surui, when allele sharing and RELPAIR disagreed on inferences of first-degree relationships, allele sharing was taken to be more reliable. The RELPAIR algorithm utilizes allele frequencies among unrelated individuals in order to probabilistically attribute identity in state to identity by descent. With a small number of relative pairs present in a data set, the occurrence of a few sets of alleles that are identical by descent does not have a major influence on the required estimates of allele frequencies. However, with many relative pairs, such as in Karitiana and Surui, the estimates of allele frequencies among "unrelateds" are poor, and probabilistic attribution of identity in state to identity by descent cannot be performed accurately.

Inferred relative pairs for different levels of relationship are listed in Supplementary Tables 17-19, and a list of parent/parent/offspring trios is given in Supplementary Table 20. The close agreement of RELPAIR and allele sharing in estimating parent/offspring and full sib relationships (Supplementary Tables 16-18) suggests that in all populations, with the possible exceptions of Karitiana and Surui, the pairs in Supplementary Tables 17 and 18 constitute all first-degree relative pairs in the diversity panel. Greater uncertainty exists in the inference of second-degree relationships, but it is likely that Supplementary Table 19 contains all or nearly all second-degree relative pairs outside of Karitiana and Surui, with the possible inclusion of a few distantly related pairs erroneously inferred to be second-degree relatives.

Conditional on the relationships in Supplementary Tables 17-19, Figure 1 displays the levels of allele sharing for pairs of individuals from different regions, pairs from different populations in the same region, and for various levels of relationship for pairs of individuals from the same population. Because of the higher homozygosity of Native Americans, Figure 1 restricts

attention to pairs in which neither individual is a Native American, and each of Supplementary Figures 2-6 considers pairs in which one or both individuals is a member of a specific Native American population. In each figure, distinct clusters of points are present, corresponding to pairs with different levels of relationship (incorporating the pairs involving Native Americans into Figure 1 would cause these clusters to be obscured). Additionally, the figures clearly illustrate that the diversity panel contains no close relative pairs from different populations.

The plot of allele sharing in Figure 1, a variant of a graphical display method commonly used for verifying putative relationships (Abecasis *et al.* 2001), provides an illustration of a well-known property of human populations: from Figure 1, it can be observed that pairs of individuals from the same population tend to share only a slightly greater proportion of their alleles than do pairs from different populations in the same region, who in turn tend to share only a slightly greater proportion of their alleles than do pairs from different regions. Averaging across all pairs in H1048, except pairs involving Karitiana or Surui and pairs with relationship closer than CO, the levels of allele sharing ($p_2+p_1/2$, or $[1+p_2-p_0]/2$) for two individuals from the same population, two individuals from different populations in the same region, and two individuals from different regions, are 0.387, 0.377, and 0.343, respectively. If the average pairwise genetic difference for two individuals from different regions is partitioned into components for the average difference for two individuals from the same population, the average difference for two individuals from different populations in the same region beyond that of two individuals from the same population, and the average difference for two individuals from different regions beyond that of two individuals from different populations in the same region, these components equal (1-0.387)/(1-0.343)=0.933, (0.387-0.377)/(1-0.343)=0.016, and (0.377-0.343)/(1-0.343)=0.051, respectively. With the subset of the data considered here corresponding to the data of Rosenberg *et al.* (2002), partitions of genetic variation into similar components via alternative methods previously yielded similar values (Rosenberg *et al.* 2002, 2003; Excoffier & Hamilton, 2003).

Construction of recommended subsets H971 and H952

The recommended subsets H971 and H952 were constructed from H1048 by avoiding inclusion of first- and second-degree relative pairs, respectively. The following principles were used in deciding which individuals to exclude from H1048 in developing the data sets H971 and H952:

1. CO relationships inferred by RELPAIR were not considered close enough to require exclusion of any individuals from the data set. Because CO relationships are the most distant relationship investigated by RELPAIR other than "unrelated," many relationships such as great-aunt/great-nephew, second cousins, and so forth may lead to high likelihoods for CO.

2. If RELPAIR found that the most likely relationship for a pair of individuals was CO, but that the likelihood ratio for CO and the relationship with the second-highest likelihood did not exceed the critical value, the relationship was not considered close enough to require exclusion of any individuals from the data set.

3. If two or more relationships inferred by RELPAIR were incompatible when considering several pairs of individuals (for example, if two individuals were inferred to be full sibs, and a third individual was inferred to be the half sib of one of them but not of the other), first-degree relationships were treated as accurate and second-degree relationships as less certain. In all cases in which three or more individuals were linked in the same pedigree – with a few exceptions in Karitiana and Surui – no incompatibilities were observed between different inferences about first-degree relationships. In other words, with some exceptions in

Karitiana and Surui, the pedigrees constructed by assembly of PO and FS pairs were always consistent both with the inferred set of first-degree pairs and with its complement. As distinguishing among higher-order relationships is often difficult, pedigrees were generally consistent with at least some inferred AV, HS, GG, and CO relationships, but sometimes conflicted with others.

4. In populations for which the number of relationships was particularly large in comparison with sample size – Karitiana and Surui – RELPAIR inference was particularly difficult, and the allele-sharing analysis was used to assist in decisions about which individuals to exclude. In these populations, as noted above, when a discrepancy was observed between allele sharing and RELPAIR in inferences of PO or FS relationships, the estimate based on the allele-sharing analysis was used (Supplementary Tables 17 and 18).

5. Individuals were excluded so as to minimize the number of required exclusions. Given equal levels of inferred relationship, the individual with the higher sample identification number was excluded. An exception to this rule was made for Druze 570. Although this sample had the lower identification number in a relative pair, it was excluded due to its large amount of missing data in a study currently in progress (data not shown).

In Karitiana and Surui it is difficult to be certain that after the exclusions in Supplementary Tables 13 and 15 are made, no relative pairs closer than first cousins are present. Thus, even with the recommended subsets H971 and H952, particular caution should be exercised in interpretation of patterns of genetic variation in these two populations.

Conclusions

This article has described three subsets of the HGDP-CEPH Human Genome Diversity Panel that are recommended for future use (Supplementary Tables 21-24). Data set H1048 consists of the original HGDP-CEPH panel, excluding one member of each duplicate pair (both members in one case) and two extremely atypical individuals. Data set H971 excludes 77 individuals from H1048 in order to avoid including first-degree relative pairs, and data set H952 excludes an additional 19 individuals from H971 to avoid second-degree relatives. It is believed that H952 contains no pairs of relatives closer than first cousins, with possible exceptions in Karitiana and Surui.

Note that samples not in the recommended subsets might also be useful in specialized contexts. For example, the duplicates might be of use in genotyping assays that frequently have sample failures, or in the measurement of genotyping error rates; the parent/offspring pairs might assist in resolving unknown haplotype phase or in estimating mutation rates. More generally, the relative pairs might be useful in identifying relatives among other individuals genotyped for the same markers as those typed in the diversity panel.

Acknowledgments

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O. C., & Cardon, L. R. (2001) GRR: graphical representation of relationship errors. *Bioinformatics* **17**, 742-743.

Boehnke, M. & Cox, N. J. (1997) Accurate inference of relationships in sib-pair linkage studies. *Am J Hum Genet* **61**, 423-429.

Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., *et al.* (2002) A human genome diversity cell line panel. *Science* **296**, 261-262.

Cavalli-Sforza, L. L. (2005) The Human Genome Diversity Project: past, present and future. *Nature Rev Genet* **6**, 333-340.

Ekins, J. E., Ekins, J. B., Layton, L., Hutchison, L. A. D., Myres, N. M., & Woodward, S. R. (2006) Inference of ancestry: Constructing hierarchical reference populations and assigning unknown individuals. *Hum Genomics* **4**, 212-235.

Epstein, M. P., Duren, W. L. & Boehnke, M. (2000) Improved inference of relationship for pairs of individuals. *Am J Hum Genet* **67**, 1219-1231.

Excoffier, L. & Hamilton, G. (2003) Comment on "Genetic structure of human populations." *Science* **300**, 1877.

Mountain, J. L. & Cavalli-Sforza, L. L. (1997) Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am J Hum Genet* **61**, 705-718.

Mountain, J. L. & Ramakrishnan, U. (2005). Impact of human population history on distributions of individual-level genetic distance. *Hum Genomics* **2**, 4-19.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. & Cavalli-Sforza, L. L. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* **102**, 15942-15947.

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K. & Feldman, M. W. (2005) Clines, clusters, and the effect of study design on the inference of human population structure. *PloS Genet* **1**, 660-671.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. & Feldman, M. W. (2002) Genetic structure of human populations. *Science* **298**, 2381-2385.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. & Feldman M. W. (2003) Response to comment on "Genetic structure of human populations." *Science* **300**, 1877.
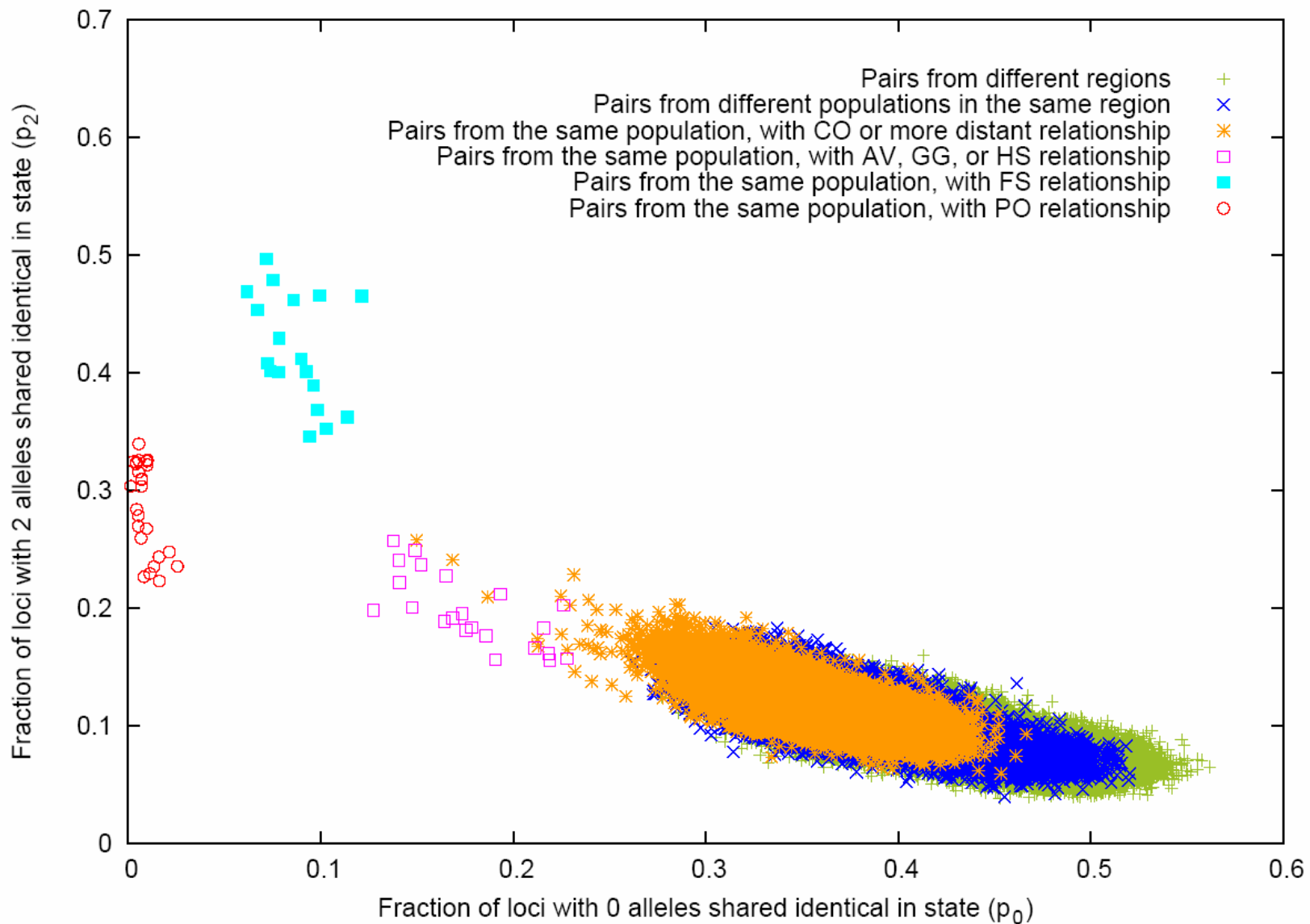
Figure 1. Allele sharing for pairs of individuals in which neither member of the pair is a Native American. The plot contains 25 parent/offspring pairs, 18 full sib pairs, and 22 pairs with second-degree relationships. The five pairs with CO or more distant relationship with the smallest values of p_0 were inferred to be first cousin pairs, and may indeed have CO relationships: Melanesian 491 and 663 plotted at (0.150, 0.258), Melanesian 823 and 825 at (0.169, 0.242), Naxi 1339 and 1342 at (0.187, 0.210), Kalash 274 and 313 at (0.213, 0.174), and Druze 562 and 594 at (0.213, 0.168).

Supplementary information for "Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives"

The phrase "inferred to be cousins" in Supplementary Tables 6-15 means "inferred to be first cousins or other distant relatives." First cousins are not included in "inferred relative pairs" in the tables; the phrase "no other relationships" in the tables, however, means "no other FS, PO, HS, GG, AV, or CO relationships." A pair is listed in the tables if the likelihood ratio for the most likely relationship in comparison to "unrelated" exceeds 100. If the most likely relationship for a pair is HS, AV, GG, or CO, and if the likelihood ratio for this relationship and "unrelated" exceeds 100, other relationships (among HS, AV, GG, and CO) are also mentioned as secondary possibilities if their likelihoods are 10% or more of the likelihood of the most likely relationship (regardless of whether or not the likelihood ratios for these additional possibilities and "unrelated" exceed 100). The threshold of 10% was chosen for convenience; with a threshold considerably smaller than 10%, the tables would become unwieldy. For a given pair, if several alternative relationships are listed, the list proceeds in decreasing order of the likelihoods of the relationships. If CO is the most likely relationship for a pair of individuals, other relationships are only mentioned for that pair if they affect a decision about exclusion that utilizes inferred relationships for other pairs. Samples are indicated by identification numbers that were assigned by CEPH and that range from 1 to 1419.

Supplementary Web Resources

Marshfield Human Diversity Panel website, http://research.marshfieldclinic.org/genetics/Freq/FreqInfo.htm
Rosenberg Lab website, http://rosenberglab.bioinformatics.med.umich.edu
Rosenberg USC Diversity Panel website, http://www.cmb.usc.edu/people/noahr/diversity.html

Supplementary References

Ramachandran, S., Rosenberg, N. A., Zhivotovsky, L. A. & Feldman, M. W. (2004) Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. *Hum Genomics* **1**, 87-97.
Rosenberg, N. A. (2005) Algorithms for selecting informative marker panels for population assignment. *J Comput Biol* **12**, 1183-1201.
Rosenberg, N. A. & Calabrese, P. P. (2004) Polyploid and multilocus extensions of the Wahlund inequality. *Theor Pop Biol* **66**, 381-391.
Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. (2003b) Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* **73**, 1402-1422.
Zhivotovsky, L. A., Rosenberg, N. A. & Feldman, M. W. (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am J Hum Genet* **72**, 1171-1186.

Supplementary Table 1. Nine disjoint subsets into which 1066 samples can be subdivided.

Subset number	Samples included	Explanation
1	Japanese 1026	The individual is not in the diversity panel.
2	She 1331	The genotypes for this individual were excluded from data sent from the Mammalian Genotyping Service to Marcus Feldman in March 2002, and were therefore not included in the data analyzed by Rosenberg *et al.* (2002). However, this individual is in the diversity panel and his genotypes do appear in the files of microsatellite genotypes posted on the Marshfield Human Diversity Panel website (the individual is male).
3	Biaka Pygmy 980 Japanese 770	These individuals were found by Rosenberg *et al.* (2002) to be extremely atypical and potentially mislabeled.
4	Herero 1028 Herero 1035 Ovambo 1031 Pedi 993 Sotho 994 Tswana 1030 Tswana 1034 Zulu 1033	These individuals are from populations in which only one or two individuals was included in the diversity panel.
5	Nilote 1410	This individual is the sole representative of his population and is not in the diversity panel (the individual is male).
6	Bedouin 652 Biaka Pygmy 1087 Biaka Pygmy 1092 Biaka Pygmy 981 Druze 589 Han 1022 Hezhen 1235 Italian 1154 Japanese 1025 Melanesian 826 Melanesian 659 Melanesian 979	Each of these samples is a duplicate of the sample in the corresponding position in the list in set 7.
7	Bedouin 650 Biaka Pygmy 452 Biaka Pygmy 457 Biaka Pygmy 472 Druze 583 Han 813 Hezhen 1233 Italian 1149 Japanese 762 Melanesian 657 Melanesian 658 Melanesian 660	Each of these samples is a duplicate of the sample in the corresponding position in the list in set 6. The individuals in set 7 are the ones with the smaller identification numbers in their duplicate pairs.
8	Hazara 111 Pathan 220	These samples, from Pakistan, are duplicates of each other but are listed with different population labels.
9	All 1027 individuals not in subsets 1-8.	

11

Supplementary Table 2. Combinations of subsets from Supplementary Table 1 that are studied in various settings.

Description	Subsets from Supplementary Table 1 that are included
Included in HGDP-CEPH Human Genome Diversity Cell Line Panel	2 3 4 6 7 8 9
Genotyped by Marshfield for microsatellites from screening set 10	1 2 3 4 5 6 7 8 9
Analyzed in Rosenberg *et al.* (2002)	1 3 6 7 8 9
Genotyped by Marshfield for microsatellites from screening sets 13 and 52	2 3 4 6 7 8 9
Genotyped by Marshfield for indel markers from screening set 100	1 2 3 4 5 6 7 8 9
Included in H1048 (see Supplementary Table 4)	2 4 7 9

The samples analyzed in the Rosenberg *et al.* (2002) paper are identical to those analyzed in Rosenberg *et al.* (2003), Rosenberg *et al.* (2003b), Zhivotovsky *et al.* (2003), Ramachandran *et al.* (2004), Rosenberg & Calabrese (2004), and Rosenberg (2005). The exact data used in the Rosenberg *et al.* (2002) paper are located on the Rosenberg Lab website and were previously located on the Rosenberg USC Diversity Panel website. An article that refers to Rosenberg *et al.* (2002) and to either of these websites very likely used this same set of individuals. An article that references the Marshfield Human Diversity Panel website would likely have used a slightly different combination of individuals. In the table, "Marshfield" refers to the Mammalian Genotyping Service at the Center for Medical Genetics, Marshfield Medical Research Foundation.

Supplementary Table 3. Duplicate pairs, adapted from Mountain & Ramakrishnan (2005, Table 1).

Member of duplicate pair retained in H1048	Member of duplicate pair excluded from H1048	Proportion-of-shared-alleles distance between the pair
Bedouin 650	Bedouin 652	0.004
Biaka Pygmy 452	Biaka Pygmy 1087	0.016
Biaka Pygmy 457	Biaka Pygmy 1092	0.006
Biaka Pygmy 472	Biaka Pygmy 981	0.006
Druze 583	Druze 589	0.014
Han 813	Han 1022	0.008
Hezhen 1233	Hezhen 1235	0.004
Italian 1149*	Italian 1154	0.003
Japanese 762	Japanese 1025	0.009
Melanesian 657	Melanesian 826	0.003
Melanesian 658	Melanesian 659	0.007
Melanesian 660	Melanesian 979	0.006
	Hazara 111 Pathan 220	0.017

* This corrects a typographical error in Mountain & Ramakrishnan (2005).

This analysis is based on the 377 microsatellite loci studied by Rosenberg *et al.* (2002) and utilizes the proportion-of-shared-alleles distance as described in Mountain & Cavalli-Sforza (1997).

Supplementary Table 4. The H1048 data set.

Subset (from Supplementary Table 1)	Reason for exclusion from H1048
1	Not in the diversity panel
3	Correct population labels are unknown
5	Not in the diversity panel
6	Duplicates; the convention is to discard duplicates with larger identification numbers
8	Duplicates; the correct population label is unknown

Supplementary Table 5. The 11 of 783 loci from Ramachandran *et al.* (2005) and Rosenberg *et al.* (2005) that were not used in the RELPAIR analysis.

Locus	Reason for exclusion
D20S201*	Uncertain/unknown map position
D11S4463*	Uncertain/unknown map position
ATA43C09M	Bioinformatics error
GATA12A08P	Bioinformatics error
GATA143C02	Uncertain/unknown map position
GATA71E06	Uncertain/unknown map position
GTTT002P	Bioinformatics error
TAT028P	Bioinformatics error
TTA008P	Bioinformatics error
TTTA075P	Bioinformatics error
TTTTA002	Uncertain/unknown map position

* These loci were omitted in Rosenberg & Calabrese (2004) for the same reason. The other nine loci in the table are among those that have been genotyped more recently and that were not considered in Rosenberg *et al.* (2002). The bioinformatics errors that caused loci to be excluded were generally of an inconsequential nature, such as typographical errors that led to a loss of information about map position. These errors were discovered only after the analysis was performed, and the exclusion of these loci is not expected to substantially influence the calculations.

Supplementary Table 6. Inferred relative pairs for (sub-Saharan) Africa.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Bantu (S. Africa)		No relationships in this population.		
Bantu (Kenya)	(1411, 1413) FS	No other relationships involving 1411, 1413. (1412, 1418) are inferred to be cousins. No other relationships in this population.	1413	1413
Mandenka	(913, 919) HS or AV (915, 916) AV or HS	No other relationships involving 913, 919. No other relationships involving 915, 916. (908, 1285) are inferred to be cousins. No other relationships in this population		919 916
Yoruba	(920, 921) FS (922, 923) FS (922, 925) PO (923, 925) PO	No other relationships involving 920, 921. No other relationships involving 922, 923, 925. It seems safe to infer that 925 (f) is a parent and that 922 (f) and 923 (m) are her offspring. No other relationships in this population.	921 922 923	921 922 923
San	(987, 988) PO	No other relationships in this population.	988	988
Mbuti Pygmy	(982, 983) PO (468, 471) PO (468, 984) AV or HS	No other relationships involving 982, 983. No other relationships involving 468, 471, 984. No other relationships in this population.	983 468	983 468
Biaka Pygmy	(473, 1089) PO (466, 1088) FS (1085, 1088) AV, CO, or HS (465, 1085) HS, AV, or CO (477, 1093) PO (457, 1093) PO (1084, 1093) FS (477, 1084) GG, HS, CO, or AV (457, 477) GG or HS (457, 1084) AV, HS, or CO (451, 464) PO (472, 1091) AV, HS, or CO (448, 461) AV or HS	No other relationships involving 473, 1089. (466, 1085), (465, 466), (465, 1088), (455, 1085) are inferred to be cousins. No other relationships involving 455, 465, 466, 1085, 1088. It is likely that 1093 (m) is a parent of 457 (m) and 477 (m), who are half sibs, and that 1084 (f) is a full sib of 1093. No other relationships involving 457, 477, 1084, 1093. (451, 1091), (451, 472) are inferred to be cousins. (451, 1091) has a higher relative likelihood for grandparent/grandchild than (451, 472) has for any non-cousin relationship. No other relationships involving 451, 464, 472, 1091. (448, 460) are inferred to be cousins. No other relationships involving 448, 460, 461. (453, 479) are inferred to be cousins. No other relationships in this population.	1089 1088 477 1093 451	1089 1088 1085 477 1093 1084 451 1091 448

Supplementary Table 7. Inferred relative pairs for Europe.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Orcadian	(794, 801) PO	No other relationships in this population.	801	801
Adygei		No relationships in this population.		
Russian		No relationships in this population.		
Basque		No relationships in this population.		
French	(511, 532) FS	No other relationships in this population.	532	532
Italian		No relationships in this population.		
Sardinian		No relationships in this population.		
Tuscan		No relationships in this population.		

Supplementary Table 8. Inferred relative pairs for the Middle East.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Mozabite	(1280, 1281) FS	No other relationships in this population.	1281	1281
Bedouin	(616, 633) PO (617, 635) AV or HS	No other relationships involving 616, 633. (617, 619) are inferred to be cousins. No other relationships involving 617, 619, 635. Nine additional cousin pairs – (610, 612), (614, 615), (614, 626), (614, 642), (615, 626), (615, 628), (618, 701), (622, 642), (630, 631) – but otherwise no other relationships in this population.	633	633 617
Druze	(571, 592) PO (569, 603) FS (568, 585) HS or AV (590, 605) FS (570, 591) AV or HS	No other relationships involving 571, 592. (569, 585), (585, 603), (577, 585), (568, 577) are inferred to be cousins. No other relationships involving 568, 569, 577, 585, 603. (581, 605), (573, 605), (581, 590), (558, 590) are inferred to be cousins. (581, 604) are inferred to be cousins. No other relationships involving 558, 573, 581, 590, 605. No other relationships involving 570, 591. Nine additional cousin pairs – (557, 565), (557, 578), (557, 594), (559, 584), (562, 594), (564, 594), (567, 588), (575, 583), (575, 604) – but otherwise no other relationships in this population.	592 603 605	592 603 585 605 570
Palestinian	(694, 695) FS (681, 684) HS or AV (682, 743) AV or HS (723, 743) AV or HS (726, 728) AV or HS (693, 742) AV or HS	No other relationships involving 694, 695. (681, 734) are inferred to be cousins. No other relationships involving 681, 684, 734. (682, 723) are inferred to be cousins. No other relationships involving 682, 723, 743. No other relationships involving 726, 728. (679, 693), (679, 742) are inferred to be cousins. No other relationships involving 679, 693, 742. Ten additional cousin pairs – (675, 737), (677, 724), (678, 735), (683, 690), (688, 727), (691, 746), (696, 730), (697, 733), (724, 725), (732, 735) – but otherwise no other relationships in this population.	695	695 681 743 728 742

Supplementary Table 9. Inferred relative pairs for Central/South Asia.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Balochi	(82, 84) FS	No other relationships in this population.	84	84
Brahui		No relationships in this population.		
Makrani		(154, 157) are inferred to be cousins. No other relationships in this population.		
Sindhi	(167, 203) PO	No other relationships involving 167, 203. (173, 175) are inferred to be cousins. No other relationships in this population.	203	203
Pathan		No relationships in this population.		
Burusho		No relationships in this population.		
Hazara	(106, 113) FS (112, 128) HS or AV	No other relationships involving 106, 113. No other relationships involving 112, 128. Five additional cousin pairs – (102, 105), (102, 108), (104, 118), (105, 108), (121, 122) – but otherwise no other relationships in this population.	113	113 128
Uygur		No relationships in this population.		
Kalash	(288, 292) PO (321, 326) HS, AV, CO, or GG	(292, 328) are inferred to be cousins. No other relationships involving 288, 292, 328. (286, 321), (286, 319) are inferred to be cousins. No other relationships involving 286, 319, 321, 326. Two additional cousin pairs – (267, 277), (274, 313) – but otherwise no other relationships in this population.	292	292 321

Supplementary Table 10. Inferred relative pairs for East Asia.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Han		No relationships in this population.		
Han (N. China)		No relationships in this population.		
Dai		No relationships in this population.		
Daur		No relationships in this population.		
Hezhen		No relationships in this population.		
Lahu	(1321, 1325) FS (1323, 1324) PO	No other relationships involving 1321, 1325. No other relationships involving 1323, 1324. No other relationships in this population.	1325 1324	1325 1324
Miao		No relationships in this population.		
Oroqen	(1203, 1210) FS	No other relationships in this population.	1210	1210
She		No relationships in this population.		
Tujia		No relationships in this population.		
Tu		No relationships in this population.		
Xibo		No relationships in this population.		
Yi		No relationships in this population.		
Mongola		No relationships in this population.		
Naxi	(1340, 1343) FS	No other relationships involving 1340, 1343. (1339, 1342) are inferred to be cousins. No other relationships in this population.	1343	1343
Cambodian	(713, 718) PO	No other relationships in this population.	718	718
Japanese		No relationships in this population.		
Yakut		No relationships in this population.		

Supplementary Table 11. Inferred relative pairs for Oceania.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Melanesian	(660, 789) PO (660, 824) PO (788, 789) PO (788, 824) PO (789, 824) FS	No other relationships involving 660, 788, 789, 824. 660, 788, 789, 824 is a family with parents 660 (f) and 788 (m), and offspring 789 (m) and 824 (m).	789 824	789 824
	(655, 657) PO (656, 657) PO	No other relationships involving 655, 656, 657. 655 (m) and 656 (f) are parents and 657 (f) is their offspring.	657	657
	(658, 978) FS (658, 664) PO (664, 978) GG	No other relationships involving 658, 664, 978. 978 cannot be both the full sib of 658 and the grandparent or grandchild of 664. The likelihood of an avuncular relationship for (664, 978) is small but not negligible in comparison with a grandparent/grandchild relationship. It is likely that 658 (f) is a parent of 664 (f) and that 978 (f) is the full sib of 658 and the aunt of 664.	658	658 978
	(490, 662) PO (490, 663) PO	(491, 663) are inferred to be cousins. No other relationships involving 490, 491, 662, 663. 662 (m) and 663 (f) are the parents of 490 (m).	490	490
	(661, 825) FS (661, 823) GG, HS, or CO	(823, 825) are inferred to be cousins. No other relationships involving 661, 823, 825. If (661, 825) are full sibs, then 823 must have the same relationship to both 661 and 825. Avuncular and half sibs both have likelihoods >10% of the likelihood of cousins for (823, 825). Half sibs and cousins both have likelihoods >10% of the likelihood of grandparent/grandchild for (661, 823). No other relationships in this population.	825	825 823
Papuan		No relationships in this population.		

Supplementary Table 12. Inferred relative pairs for Colombian and Maya.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Colombian	(709, 710) PO	(705, 709), (707, 709), (705, 707), (705, 708) are inferred to be cousins. No other relationships involving 705, 706, 707, 708, 709, 710.	709	709
	(707, 708) PO		707	707
	(705, 706) PO		705	705
	(793, 970) PO	No other relationships involving 703, 793, 970. 703 (m) and 970 (f) are parents and 793 (f) is their offspring.	793	793
	(703, 793) PO			
	(702, 792) FS	No other relationships involving 702, 792.	792	792
	(704, 827) PO	No other relationships involving 704, 827.	827	827
		No other relationships in this population.		
Maya	(862, 867) PO	(862, 866), (858, 867) are inferred to be cousins.	867	867
	(858, 866) PO		866	866
	(866, 867) AV or HS			
	(876, 878) FS	No other relationships involving 876, 878.	878	878
	(854, 874) HS, AV, or GG	(865, 874), (873, 874), (865, 873) are inferred to be cousins. No other relationships involving 854, 873, 874.		874
		Two additional cousin pairs – (859, 865), (868, 869) – but otherwise no other relationships in this population.		

In the Maya population, previously reported family information (Howard Cann, pers. comm.) suggested certain relative pairs. A reported HS relationship between 858 and 865 was not supported by the analysis. The other reported relationships – PO relationship between 858 and 866, and FS relationship between 876 and 878 – were confirmed. The reported polarity of the PO relationship, with 866 as the parent and 858 as the offspring, was consistent with the analysis, but could not be confirmed. All other inferred relationships were not among those that were previously reported.

Supplementary Table 13. Inferred relative pairs for Karitiana.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Karitiana	(1016, 1018) PO (1007, 1016) FS (999, 1007) PO (997, 999) PO (1004, 1012) PO (1004, 1007) FS (1007, 1018) GG or CO (1004, 1018) GG (1004, 1016) FS (1005, 997) FS (1005, 1006) HS, AV, or GG (1004, 1017) GG, CO, HS, or AV (1014, 1017) PO	It is likely that 997, 999, 1004, 1007, 1012, 1016, 1018 are an extended family, that 1016 (f) is a parent of 1018 (f), 1004 (m) is a parent of 1012 (m), 997 (m) and 1007 (f) are the parents of 999 (f), and that 1004, 1007, 1016 are full sibs. 1005 is the full sib of 997 and 1006 is related by some second-degree relationship to 1005. 1014 and 1017 are somehow linked into this pedigree. No other non-cousin relationships involving 997, 999, 1004, 1014, 1005, 1006, 1006, 1012, 1016, 1017, 1018.	997 1004 1005 1007 1016 1017	997 1004 1005 1007 1016 1017
	(998, 1000) FS (998, 1011) FS (998, 1008) FS (1008, 1011) FS (1000, 1008) GG, FS, or HS (1000, 1011) CO, HS, AV, or GG	No other first-degree relationships involving 998, 1000, 1008, 1011. (1008, 1011) was inferred to be a parent/offspring pair according to allele sharing but a full sib pair according to RELPAIR. The inference from allele sharing is likely to be more reliable. According to the reported relationships (Howard Cann, pers. comm.), 998 (m), 1000 (m), and 1008 (f) are full sibs and 1011 (f) is the offspring of 1008.	1000 1008 1011	1000 1008 1011
	(995, 996) FS	No other non-cousin relationships involving 995, 996. No other first-degree relationships in this population.	996	996

In the Karitiana population, there are many relative pairs, and for convenience, some second-degree relative pairs are not listed in the table. The relationships in Karitiana differ substantially from the proposed list based on reported family relationships (Howard Cann, pers. comm.). That list does have the FS relationship between 1004, 1007, and 1016, as well as all of the inferred PO relationships except between 1014 and 1017. However, a reported PO relationship between 995 and 1014 was not confirmed, nor was a reported HS relationship of 995 to 1004, 1007, and 1016. The analysis was consistent with a reported HS relationship for 1005 and 1006. It was also consistent with the reported FS relationship of 998, 1000, and 1008; the allele sharing but not the RELPAIR analysis was consistent with the report that 1008 is a parent of 1011. The reported HS relationship of 1010 and 1011 was not supported by RELPAIR, nor was the reported FS relationship of 996 and 1017.

Supplementary Table 14. Inferred relative pairs for Pima.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Pima	(1047, 1049) PO	No other non-cousin relationships involving 1037, 1038, 1039, 1040, 1047, 1048, 1049, 1050, 1052. 1047 (m) and 1048 (f) are the parents of 1049 (f). 1037 (m) and 1038 (f) are the parents of 1039 (m) and 1040 (m). 1038 (f), 1048 (f), and 1050 (m) are full sibs and are the offspring of 1052 (m). Several other cousin and grandparent/grandchild relationships involving this pedigree are also inferred. Not all of these relationships are consistent with the inferred pedigree but the inferences about first-degree relationships are taken to be more reliable.	1048	1048
	(1048, 1049) PO		1049	1049
	(1037, 1039) PO		1052	1052
	(1037, 1040) PO		1038	1038
	(1038, 1039) PO		1039	1039
	(1038, 1040) PO		1040	1040
	(1039, 1040) FS			
	(1048, 1050) FS			
	(1048, 1038) FS			
	(1038, 1050) FS			
	(1048, 1052) PO			
	(1050, 1052) PO			
	(1038, 1052) PO			
	(1054, 1055) PO	(1054, 1056) are inferred to be cousins. No other relationships involving 1054, 1055, 1056.	1054	1054
	(1043, 1046) PO	(1043, 1044), (1043, 1053), (1042, 1046) are inferred to be cousins. No other relationships involving 1043, 1044, 1045, 1046. 1043 (m) and 1044 (f) are the parents of 1045 (m) and 1046 (f).	1045	1045
	(1044, 1046) PO		1046	1046
	(1043, 1045) PO			
	(1044, 1045) PO			
	(1045, 1046) FS			
	(1041, 1042) PO	No other non-cousin relationships involving 1041, 1042, 1053.	1042	1042
	(1041, 1053) GG, HS, or AV			
	(1060, 1061) FS	(1037, 1061) are inferred to be cousins. No other relationships involving 1060, 1061. No other non-cousin relationships in this population.	1061	1061

In this population, there are many relationship pairs, and for convenience, only first-degree relatives are listed for the large pedigree that includes 1037, 1038, 1039, 1040, 1047, 1048, 1049, 1050, 1052. All of the previously reported relationships in Pima (Howard Cann, pers. comm.) were confirmed, except that the polarity of two PO relationships could not be inferred: the reported relationships listed 1041 as a parent of 1042 and 1054 as a parent of 1055. Several relationships in addition to those previously reported were also identified. In particular, these included the FS relationship between 1038, 1048, and 1050 and the PO relationships between 1038 and 1052, and between 1048 and 1052.

Supplementary Table 15. Inferred relative pairs for Surui.

Population	Inferred relative pairs	Comments	Individuals excluded from H971	Individuals excluded from H952
Surui	(839, 840) FS (839, 841) FS (839, 842) FS (839, 850) FS (840, 841) FS (840, 842) FS (840, 850) FS (841, 842) FS (841, 850) FS (842, 850) FS (837, 839) PO (837, 840) PO (837, 841) GG (837, 842) PO (837, 850) PO (838, 839) PO (838, 840) PO (838, 841) AV, HS, or GG (838, 842) GG, HS, or AV (838, 850) GG or HS	It is likely that 837, 838, 839, 840, 841, 842, 850 is a family with parents 837 (m) and 838 (f) and offspring 839 (m), 840 (f), 841 (f), 842 (m), 850 (f). All ten parent/offspring pairs in this family are identified by allele sharing, although only six are identified by RELPAIR. All ten full sib relationships are identified both by allele sharing and by RELPAIR. Reported information (Howard Cann, pers. comm.) had suggested that 837 (m) is a parent of 839, 840, 841, 842, 850.	839 840 841 842 850	839 840 841 842 850
	(838, 851) FS (833, 851) PO (830, 851) PO (843, 848) PO (846, 848) PO (833, 834) FS (834, 835) FS (844, 852) FS (844, 847) FS	(830, 833) are inferred to be cousins. Reported information (Howard Cann, pers. comm.) suggested that they are half sibs. It is likely but not certain that 851 (m) is the full sib of 838 (f) and a parent of 830 (f) and 833 (f). (843, 846) are inferred to be unrelated. Therefore it seems likely that 843 (m) and 846 (f) are the parents of 848 (f). (834, 835) and (833, 834) but not (833, 835) are inferred to be FS, inconsistent with the transitivity of the FS relation. Neither 834 nor 835 was inferred to have a PO relationship with 851, a putative parent (offspring?) of 833. 834, however, but not 835, was inferred to have a GG relationship with 851. Other questionable relationships include: (850, 851) are inferred to be GG or CO; (844, 847) and (844, 852) but not (847, 852) are inferred to be FS, when (844, 847) was a putative PO pair by allele sharing; (832, 844) are inferred to be GG, CO, or HS; (835, 843) GG, HS, or AV; (830, 832) are inferred to be AV or HS – reported information suggested that these individuals are HS. No other first-degree relationships in this population.	851 830 848 834 835 844 847	851 830 833 848 834 835 844 847

In the Surui population, there are many relative pairs, and for convenience, some second-degree relative pairs are not listed in the table. Most of the reported family relationships in Surui (Howard Cann, pers. comm.) were confirmed. In particular, 837 was seen to be a parent of 839, 840, 841, 842, and 850. 851 was seen to have PO relationships with 830 and 833, but 830 and 833 were not seen to have a relationship that could enable inference of polarity; to be very conservative, 830 was excluded along with 851 from data set H971. The analysis was consistent with the reported HS relationship for 830 and 832, but AV had a higher likelihood than HS. Notable previously unknown relationships include the fact that 838 is likely to be the other parent of 839, 840, 841, 842, and 850.

Supplementary Table 16. Concordance of allele-sharing and RELPAIR analyses.

Population	Sample size in H1048	Number of PO pairs (allele sharing)	Number of PO pairs (RELPAIR)	Number of FS pairs (allele sharing)	Number of FS pairs (RELPAIR)
Adygei	17	0	0	0	0
Balochi	25	0	0	1	1
Bantu (Kenya)	12	0	0	1	1
Bantu (S. Africa)	8	0	0	0	0
Basque	24	0	0	0	0
Bedouin	48	1	1	0	0
Biaka Pygmy	32	4	4	2	2
Brahui	25	0	0	0	0
Burusho	25	0	0	0	0
Cambodian	11	1	1	0	0
Colombian	13	6	6	1	1
Dai	10	0	0	0	0
Daur	10	0	0	0	0
Druze	47	1	1	2	2
French	29	0	0	1	1
Han	34	0	0	0	0
Han (N. China)	10	0	0	0	0
Hazara	24	0	0	1	1
Hezhen	9	0	0	0	0
Italian	13	0	0	0	0
Japanese	29	0	0	0	0
Kalash	25	1	1	0	0
Karitiana	24	6	5	6	9
Lahu	10	1	1	1	1
Makrani	25	0	0	0	0
Mandenka	24	0	0	0	0
Maya	25	2	2	1	1
Mbuti Pygmy	15	2	2	0	0
Melanesian	19	9	9	3	3
Miao	10	0	0	0	0
Mongola	10	0	0	0	0
Mozabite	30	0	0	1	1
Naxi	10	0	0	1	1
Orcadian	16	1	1	0	0
Oroqen	10	0	0	1	1
Palestinian	51	0	0	1	1
Papuan	17	0	0	0	0
Pathan	24	0	0	0	0
Pima	25	15	15	6	6
Russian	25	0	0	0	0
San	7	1	1	0	0
Sardinian	28	0	0	0	0
She	10	0	0	0	0
Sindhi	25	1	1	0	0
Surui	21	15	10	14	15
Tu	10	0	0	0	0
Tujia	10	0	0	0	0
Tuscan	8	0	0	0	0
Uygur	10	0	0	0	0
Xibo	9	0	0	0	0
Yakut	25	0	0	0	0
Yi	10	0	0	0	0
Yoruba	25	2	2	2	2
Total	**1048**	**69**	**63**	**46**	**50**

Supplementary Table 17. 69 inferred parent/offspring pairs in H1048.

Population	First individual			Second individual			Method of inference: allele sharing (A) or RELPAIR (R)
	Identification number	Sex	Parent (P), offspring (O), or uncertain (U)	Identification number	Sex	Parent (P), offspring (O), or uncertain (U)	
Bedouin	616	M	U	633	F	U	A,R
Biaka Pygmy	451	M	U	464	M	U	A,R
Biaka Pygmy	457	M	O	1093	M	P	A,R
Biaka Pygmy	473	M	U	1089	M	U	A,R
Biaka Pygmy	477	M	O	1093	M	P	A,R
Cambodian	713	F	U	718	F	U	A,R
Colombian	703	M	P	793	F	O	A,R
Colombian	704	F	U	827	F	U	A,R
Colombian	705	M	U	706	F	U	A,R
Colombian	707	F	U	708	F	U	A,R
Colombian	709	M	U	710	M	U	A,R
Colombian	793	F	O	970	F	P	A,R
Druze	571	F	U	592	F	U	A,R
Kalash	288	M	U	292	M	U	A,R
Karitiana	997	M	P	999	F	O	A,R
Karitiana	999	F	O	1007	F	P	A,R
Karitiana	1004	M	P	1012	M	O	A,R
Karitiana	1008	F	U	1011	F	U	A
Karitiana	1014	F	U	1017	F	U	A,R
Karitiana	1016	F	P	1018	F	O	A,R
Lahu	1323	F	U	1324	F	U	A,R
Maya	858	F	U	866	F	U	A,R
Maya	862	F	U	867	F	U	A,R
Mbuti Pygmy	468	M	U	471	F	U	A,R
Mbuti Pygmy	982	M	U	983	M	U	A,R
Melanesian	490	M	O	662	M	P	A,R
Melanesian	490	M	O	663	F	P	A,R
Melanesian	655	M	P	657	F	O	A,R
Melanesian	656	F	P	657	F	O	A,R
Melanesian	658	F	U	664	F	U	A,R
Melanesian	660	F	P	789	M	O	A,R
Melanesian	660	F	P	824	M	O	A,R
Melanesian	788	M	P	789	M	O	A,R
Melanesian	788	M	P	824	M	O	A,R
Orcadian	794	F	U	801	F	U	A,R
Pima	1037	M	P	1039	M	O	A,R
Pima	1037	M	P	1040	M	O	A,R
Pima	1038	F	P	1039	M	O	A,R
Pima	1038	F	P	1040	M	O	A,R
Pima	1038	F	O	1052	M	P	A,R
Pima	1041	F	U	1042	M	U	A,R
Pima	1043	M	P	1045	M	O	A,R
Pima	1043	M	P	1046	F	O	A,R
Pima	1044	F	P	1045	M	O	A,R
Pima	1044	F	P	1046	F	O	A,R
Pima	1047	M	P	1049	F	O	A,R

Pima	1048	F	P	1049	F	O	A,R
Pima	1048	F	O	1052	M	P	A,R
Pima	1050	M	O	1052	M	P	A,R
Pima	1054	F	U	1055	M	U	A,R
San	987	M	U	988	M	U	A,R
Sindhi	167	M	U	203	M	U	A,R
Surui	830	F	U	851	M	U	A,R
Surui	833	F	U	851	M	U	A,R
Surui	837	M	P	839	M	O	A,R
Surui	837	M	P	840	F	O	A,R
Surui	837	M	P	841	F	O	A
Surui	837	M	P	842	M	O	A,R
Surui	837	M	P	850	F	O	A,R
Surui	838	F	P	839	M	O	A,R
Surui	838	F	P	840	F	O	A,R
Surui	838	F	P	841	F	O	A
Surui	838	F	P	842	M	O	A
Surui	838	F	P	850	F	O	A
Surui	843	M	P	848	F	O	A,R
Surui	844	M	U	847	M	U	A
Surui	846	F	P	848	F	O	A,R
Yoruba	922	F	O	925	F	P	A,R
Yoruba	923	M	O	925	F	P	A,R

The relationship for Melanesians 658 and 664 is listed as being of uncertain polarity, but it is likely that 658 is a parent and 664 is her offspring. It is also likely that Surui 844 is a parent and 847 is his offspring.

Supplementary Table 18. 46 inferred full sib pairs in H1048.

Population	First individual		Second individual		Method of inference: allele sharing (A) or RELPAIR (R)
	Identification number	**Sex**	**Identification number**	**Sex**	
Balochi	82	M	84	M	A,R
Bantu (Kenya)	1411	M	1413	M	A,R
Biaka Pygmy	466	M	1088	M	A,R
Biaka Pygmy	1084	F	1093	M	A,R
Colombian	702	F	792	M	A,R
Druze	569	F	603	M	A,R
Druze	590	F	605	M	A,R
French	511	M	532	F	A,R
Hazara	106	M	113	M	A,R
Karitiana	995	F	996	F	A,R
Karitiana	998	M	1000	M	A,R
Karitiana	998	M	1008	F	A,R
Karitiana	1004	M	1007	F	A,R
Karitiana	1004	M	1016	F	A,R
Karitiana	1007	F	1016	F	A,R
Lahu	1321	M	1325	F	A,R
Maya	876	F	878	M	A,R
Melanesian	658	F	978	F	A,R
Melanesian	661	F	825	F	A,R
Melanesian	789	M	824	M	A,R
Mozabite	1280	F	1281	F	A,R
Naxi	1340	M	1343	M	A,R
Orogen	1203	M	1210	M	A,R
Palestinian	694	F	695	F	A,R
Pima	1038	F	1048	F	A,R
Pima	1038	F	1050	M	A,R
Pima	1039	M	1040	M	A,R
Pima	1045	M	1046	F	A,R
Pima	1048	F	1050	M	A,R
Pima	1060	M	1061	M	A,R
Surui	833	F	834	M	A,R
Surui	834	M	835	M	A,R
Surui	838	F	851	M	A,R
Surui	839	M	840	F	A,R
Surui	839	M	841	F	A,R
Surui	839	M	842	M	A,R
Surui	839	M	850	F	A,R
Surui	840	F	841	F	A,R
Surui	840	F	842	M	A,R
Surui	840	F	850	F	A,R
Surui	841	F	842	M	A,R
Surui	841	F	850	F	A,R
Surui	842	M	850	F	A,R
Surui	844	M	852	F	A,R
Yoruba	920	F	921	F	A,R
Yoruba	922	F	923	M	A,R

Surui pairs (833, 834) and (834, 835) but not (833, 835) were inferred to be full sibs. For at least one of these three pairs, the relationship must have been incorrectly inferred.

Supplementary Table 19. 34 inferred second-degree relative pairs in H1048.

Population	First individual			Second individual			Method of inference: allele sharing (A) or RELPAIR (R)
	Identification number	**Sex**	**Half sib (H), aunt or uncle (A), niece or nephew (N), grandparent (G), grandchild (C), or uncertain (U)**	**Identification number**	**Sex**	**Half sib (H), aunt or uncle (A), niece or nephew (N), grandparent (G), grandchild (C), or uncertain (U)**	
Bedouin	617	M	U	635	F	U	R
Biaka Pygmy	448	M	U	461	M	U	R
Biaka Pygmy	457	M	N	1084	F	A	R
Biaka Pygmy	457	M	H	477	M	H	R
Biaka Pygmy	465	M	U	1085	F	U	R
Biaka Pygmy	472	M	U	1091	M	U	R
Biaka Pygmy	477	M	N	1084	F	A	R
Biaka Pygmy	1085	F	U	1088	M	U	R
Druze	568	F	U	585	F	U	R
Druze	570	F	U	591	F	U	R
Hazara	112	M	U	128	M	U	R
Kalash	321	M	U	326	M	U	R
Mandenka	913	M	U	919	M	U	R
Mandenka	915	F	U	916	F	U	R
Maya	854	F	U	874	F	U	R
Maya	866	F	U	867	F	U	R
Mbuti Pygmy	468	M	U	984	M	U	R
Melanesian	661	F	U	823	M	U	R
Melanesian	664	F	U	978	F	U	R
Palestinian	681	F	U	684	F	U	R
Palestinian	682	F	U	743	F	U	R
Palestinian	693	F	U	742	F	U	R
Palestinian	723	M	U	743	F	U	R
Palestinian	726	M	U	728	M	U	R
Pima	1038	F	A	1049	F	N	R
Pima	1039	M	N	1048	F	A	R
Pima	1039	M	N	1050	M	A	R
Pima	1039	M	C	1052	M	G	R
Pima	1040	M	N	1048	F	A	R
Pima	1040	M	N	1050	M	A	R
Pima	1040	M	C	1052	M	G	R
Pima	1041	F	U	1053	F	U	R
Pima	1049	F	N	1050	M	A	R
Pima	1049	F	C	1052	M	G	R

The type of second-degree relationship is regarded as certain only if a single type of relationship is compatible with the pairs that appear in Supplementary Tables 17 and 18 (and with the lack of appearance of any other pairs in those tables). Due to the considerable uncertainty in Karitiana and Surui, no pairs are listed for these populations. Some pairs listed with uncertain relationship might not actually be second-degree relatives, and some second-degree pairs might not have been identified. Melanesians 664 and 978 are listed as having an uncertain relationship, but it is likely that 978 is an aunt and that 664 is her niece.

Supplementary Table 20. 17 inferred parent/parent/offspring trios in H1048.

Population	Identification number of father	Identification number of mother	Identification number of offspring	Sex of offspring	Method of inference: allele sharing (A) or RELPAIR (R)
Colombian	703	970	793	F	A,R
Karitiana	997	1007	999	F	A,R
Melanesian	655	656	657	F	A,R
Melanesian	662	663	490	M	A,R
Melanesian	788	660	789	M	A,R
Melanesian	788	660	824	M	A,R
Pima	1037	1038	1039	M	A,R
Pima	1037	1038	1040	M	A,R
Pima	1043	1044	1045	M	A,R
Pima	1043	1044	1046	F	A,R
Pima	1047	1048	1049	F	A,R
Surui	837	838	839	M	A,R
Surui	837	838	840	F	A,R
Surui	837	838	841	F	A
Surui	837	838	842	M	A
Surui	837	838	850	F	A
Surui	843	846	848	F	A,R

H1048 includes several sets that contain two parents and two or more of their offspring: Melanesians 660, 788, 789, 824; Pima 1037, 1038, 1039, 1040; Pima 1043, 1044, 1045, 1046; and Surui 837, 838, 839, 840, 841, 842, 850.

Supplementary Table 21. Numbers of individuals excluded from H1048 in H971 and H952.

Population	Sample size in H1048	Number of individuals excluded from H1048 in H971	Sample size in H971	Number of individuals excluded from H1048 in H952	Sample size in H952
Adygei	17	0	17	0	17
Balochi	25	1	24	1	24
Bantu (Kenya)	12	1	11	1	11
Bantu (S. Africa)	8	0	8	0	8
Basque	24	0	24	0	24
Bedouin	48	1	47	2	46
Biaka Pygmy	32	5	27	9	23
Brahui	25	0	25	0	25
Burusho	25	0	25	0	25
Cambodian	11	1	10	1	10
Colombian	13	6	7	6	7
Dai	10	0	10	0	10
Daur	10	0	10	0	10
Druze	47	3	44	5	42
French	29	1	28	1	28
Han	34	0	34	0	34
Han (N. China)	10	0	10	0	10
Hazara	24	1	23	2	22
Hezhen	9	0	9	0	9
Italian	13	0	13	0	13
Japanese	29	0	29	0	29
Kalash	25	1	24	2	23
Karitiana	24	10	14	10	14
Lahu	10	2	8	2	8
Makrani	25	0	25	0	25
Mandenka	24	0	24	2	22
Maya	25	3	22	4	21
Mbuti Pygmy	15	2	13	2	13
Melanesian	19	6	13	8	11
Miao	10	0	10	0	10
Mongola	10	0	10	0	10
Mozabite	30	1	29	1	29
Naxi	10	1	9	1	9
Orcadian	16	1	15	1	15
Oroqen	10	1	9	1	9
Palestinian	51	1	50	5	46
Papuan	17	0	17	0	17
Pathan	24	0	24	0	24
Pima	25	11	14	11	14
Russian	25	0	25	0	25
San	7	1	6	1	6
Sardinian	28	0	28	0	28
She	10	0	10	0	10
Sindhi	25	1	24	1	24
Surui	21	12	9	13	8
Tu	10	0	10	0	10
Tujia	10	0	10	0	10
Tuscan	8	0	8	0	8
Uygur	10	0	10	0	10
Xibo	9	0	9	0	9
Yakut	25	0	25	0	25
Yi	10	0	10	0	10
Yoruba	25	3	22	3	22
Total	**1048**	**77**	**971**	**96**	**952**

Supplementary Table 22. The 77 individuals included in H1048 but not in H971.

Population	Individuals in H1048 but not in H971
Balochi	84
Hazara	113
Sindhi	203
Kalash	292
Biaka Pygmy	451
Mbuti Pygmy	468
Biaka Pygmy	477
Melanesian	490
French	532
Druze	592
Druze	603
Druze	605
Bedouin	633
Melanesian	657
Melanesian	658
Palestinian	695
Colombian	705
Colombian	707
Colombian	709
Cambodian	718
Melanesian	789
Colombian	792
Colombian	793
Orcadian	801
Melanesian	824
Melanesian	825
Colombian	827
Surui	830
Surui	834
Surui	835
Surui	839
Surui	840
Surui	841
Surui	842
Surui	844
Surui	847

| Population	Individuals in H1048 but not in H971 (continued)
Surui	848
Surui	850
Surui	851
Maya	866
Maya	867
Maya	878
Yoruba	921
Yoruba	922
Yoruba	923
Mbuti Pygmy	983
San	988
Karitiana	996
Karitiana	997
Karitiana	1000
Karitiana	1004
Karitiana	1005
Karitiana	1007
Karitiana	1008
Karitiana	1011
Karitiana	1016
Karitiana	1017
Pima	1038
Pima	1039
Pima	1040
Pima	1042
Pima	1045
Pima	1046
Pima	1048
Pima	1049
Pima	1052
Pima	1054
Pima	1061
Biaka Pygmy	1088
Biaka Pygmy	1089
Biaka Pygmy	1093
Oroqen	1210
Mozabite	1281
Lahu	1324
Lahu	1325
Naxi	1343
Bantu (Kenya)	1413

Supplementary Table 23. The 96 individuals included in H1048 but not in H952.

| Population	Individuals in H1048 but not in H952
Balochi	84
Hazara	113
Hazara	128
Sindhi	203
Kalash	292
Kalash	321
Biaka Pygmy	448
Biaka Pygmy	451
Mbuti Pygmy	468
Biaka Pygmy	477
Melanesian	490
French	532
Druze	570
Druze	585
Druze	592
Druze	603
Druze	605
Bedouin	617
Bedouin	633
Melanesian	657
Melanesian	658
Palestinian	681
Palestinian	695
Colombian	705
Colombian	707
Colombian	709
Cambodian	718
Palestinian	728
Palestinian	742
Palestinian	743
Melanesian	789
Colombian	792
Colombian	793
Orcadian	801
Melanesian	823
Melanesian	824
Melanesian	825
Colombian	827
Surui	830
Surui	833
Surui	834
Surui	835
Surui	839
Surui	840
Surui	841
Surui	842
Surui	844
Surui	847

Population	Individuals in H1048 but not in H952 (continued)
Surui	848
Surui	850
Surui	851
Maya	866
Maya	867
Maya	874
Maya	878
Mandenka	916
Mandenka	919
Yoruba	921
Yoruba	922
Yoruba	923
Melanesian	978
Mbuti Pygmy	983
San	988
Karitiana	996
Karitiana	997
Karitiana	1000
Karitiana	1004
Karitiana	1005
Karitiana	1007
Karitiana	1008
Karitiana	1011
Karitiana	1016
Karitiana	1017
Pima	1038
Pima	1039
Pima	1040
Pima	1042
Pima	1045
Pima	1046
Pima	1048
Pima	1049
Pima	1052
Pima	1054
Pima	1061
Biaka Pygmy	1084
Biaka Pygmy	1085
Biaka Pygmy	1088
Biaka Pygmy	1089
Biaka Pygmy	1091
Biaka Pygmy	1093
Oroqen	1210
Mozabite	1281
Lahu	1324
Lahu	1325
Naxi	1343
Bantu (Kenya)	1413

Supplementary Table 24. Regional sample sizes for data sets H1048, H971, and H952.

| Region	H1048	H971	H952
Africa	123	111	105
Europe	160	158	158
Middle East	176	170	163
Central/South Asia	208	204	202
East Asia	237	232	232
Oceania	36	30	28
America	108	66	64
Total	**1048**	**971**	**952**
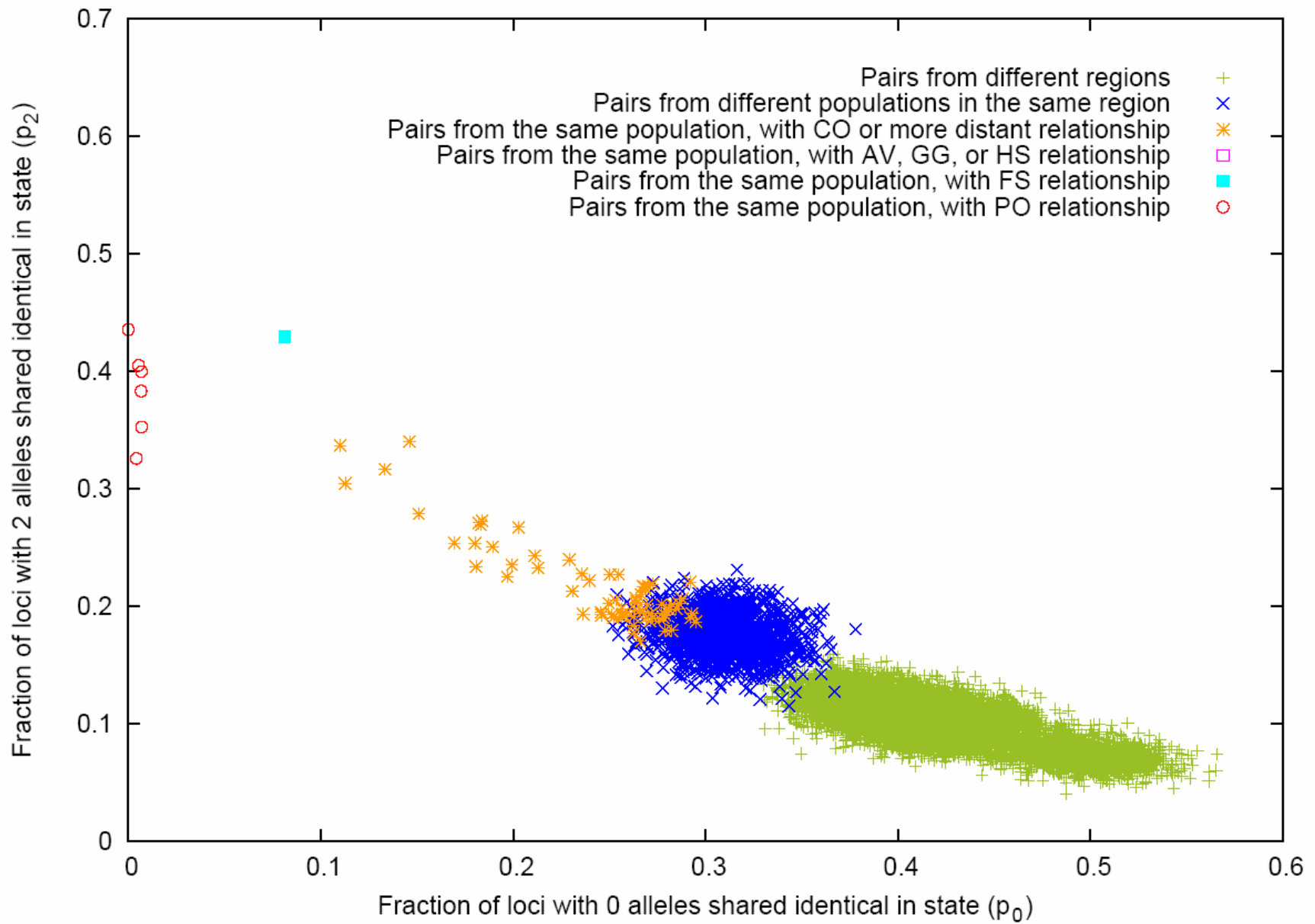
Supplementary Figure 2. Allele sharing for pairs of individuals in which at least one member of the pair is from the Colombian population. The plot contains six parent/offspring pairs and one full sib pair.
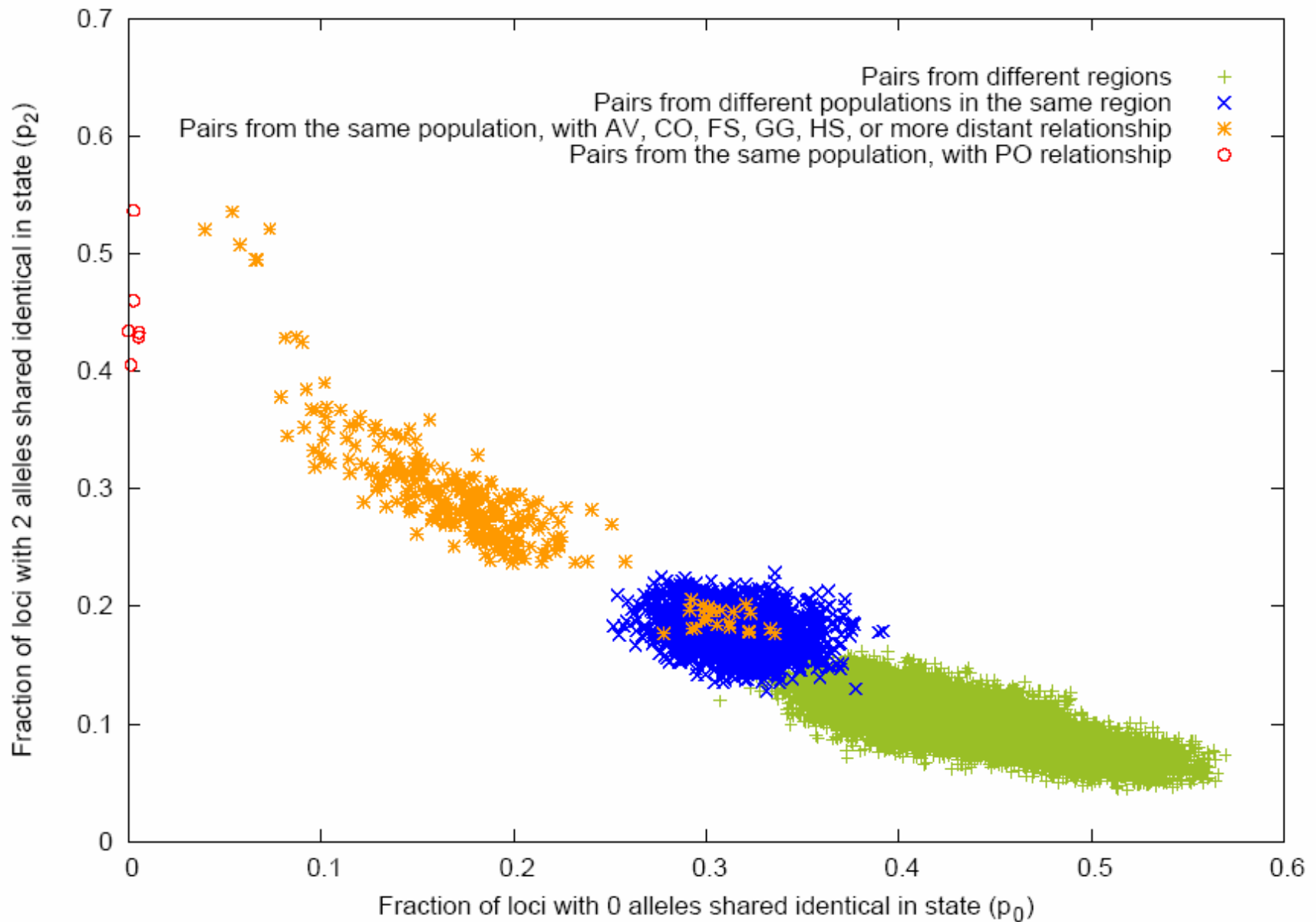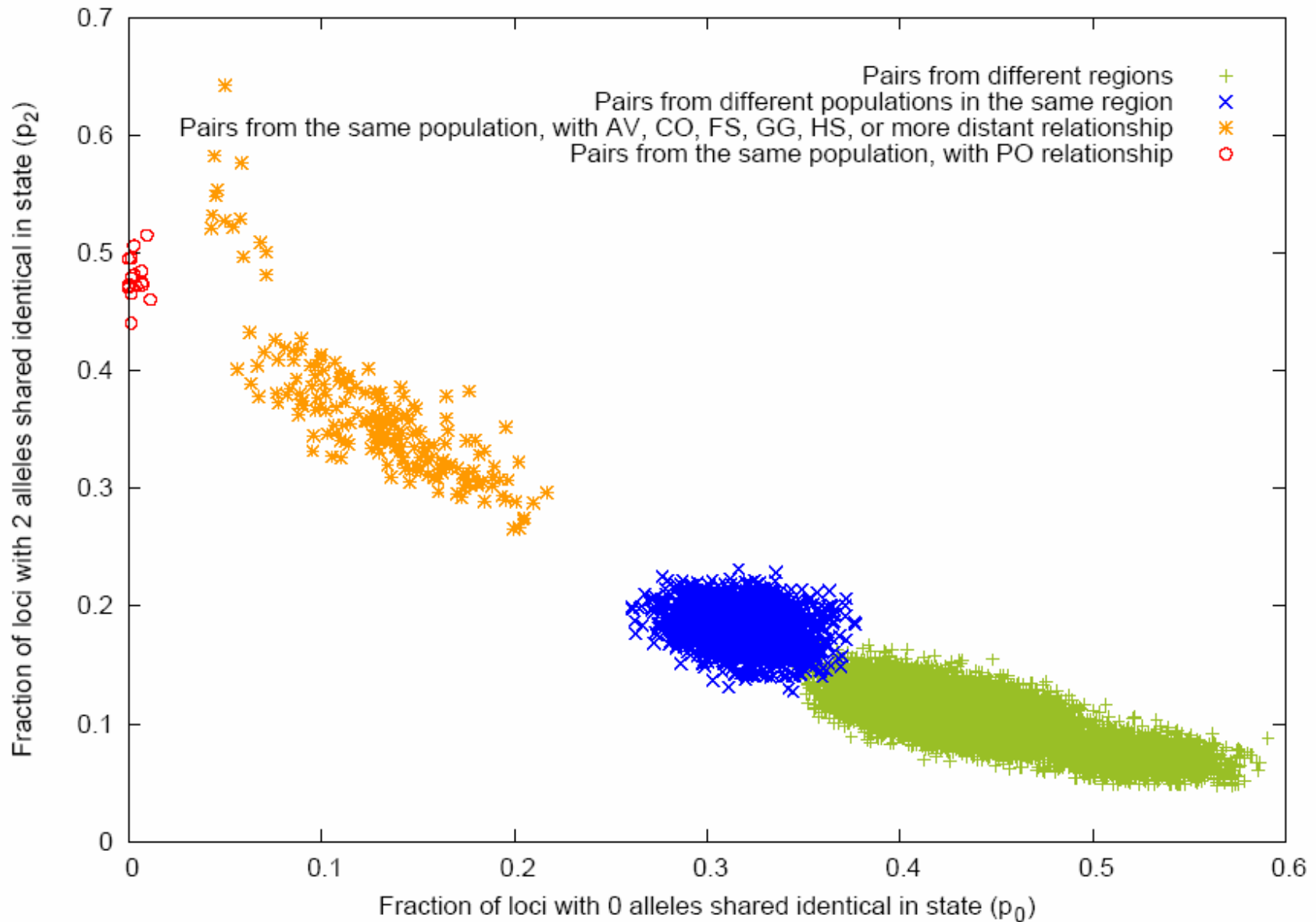
Supplementary Figure 3. Allele sharing for pairs of individuals in which at least one member of the pair is from the Karitiana population. Because the confidence placed in specific relationship inferences was lower in this population than in other populations, all pairs from the Karitiana population except for parent/offspring pairs were plotted with the same symbol. The plot contains six parent/offspring pairs.
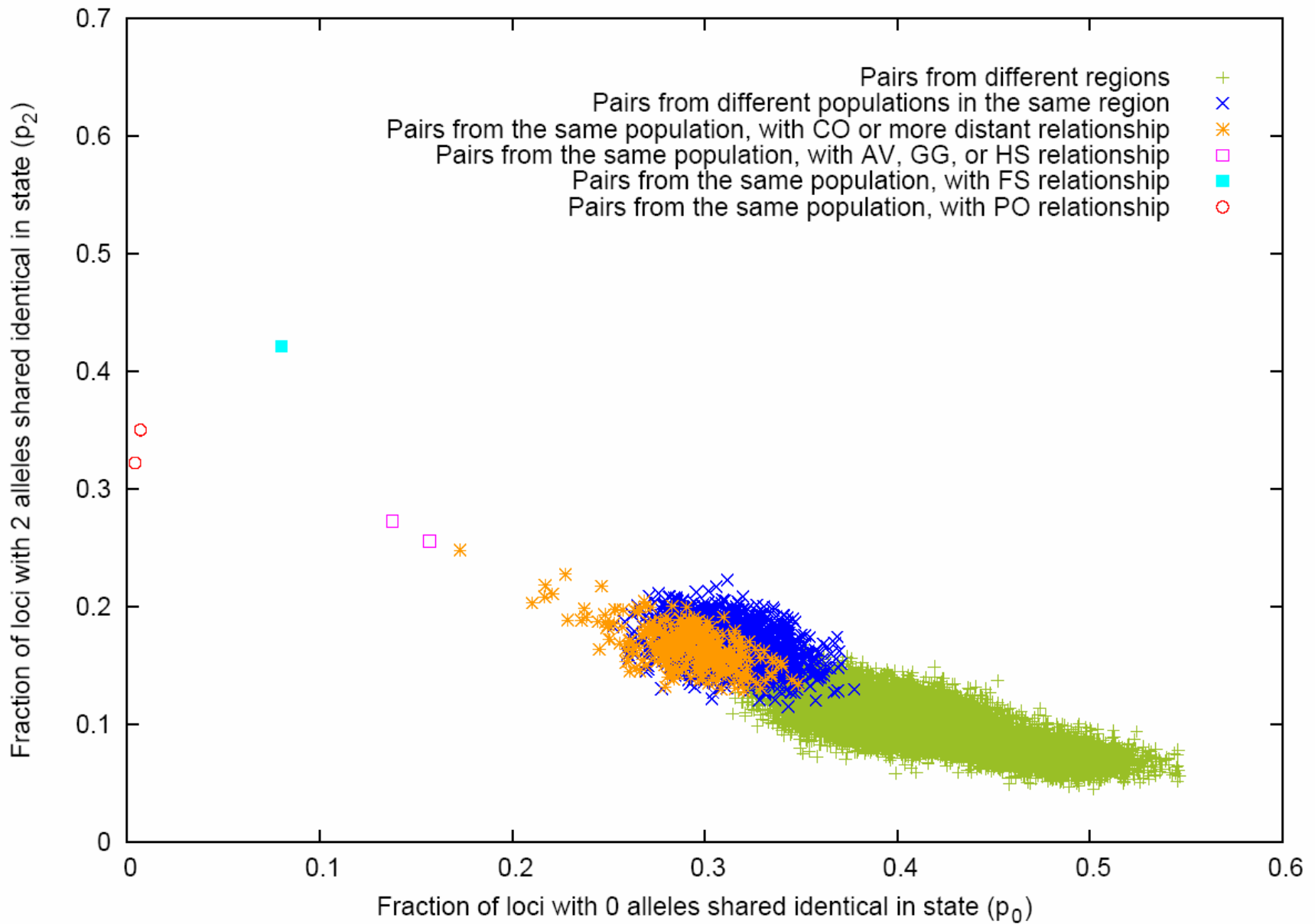
Supplementary Figure 4. Allele sharing for pairs of individuals in which at least one member of the pair is from the Maya population. The plot contains two parent/offspring pairs, one full sib pair, and two pairs with second-degree relationships.
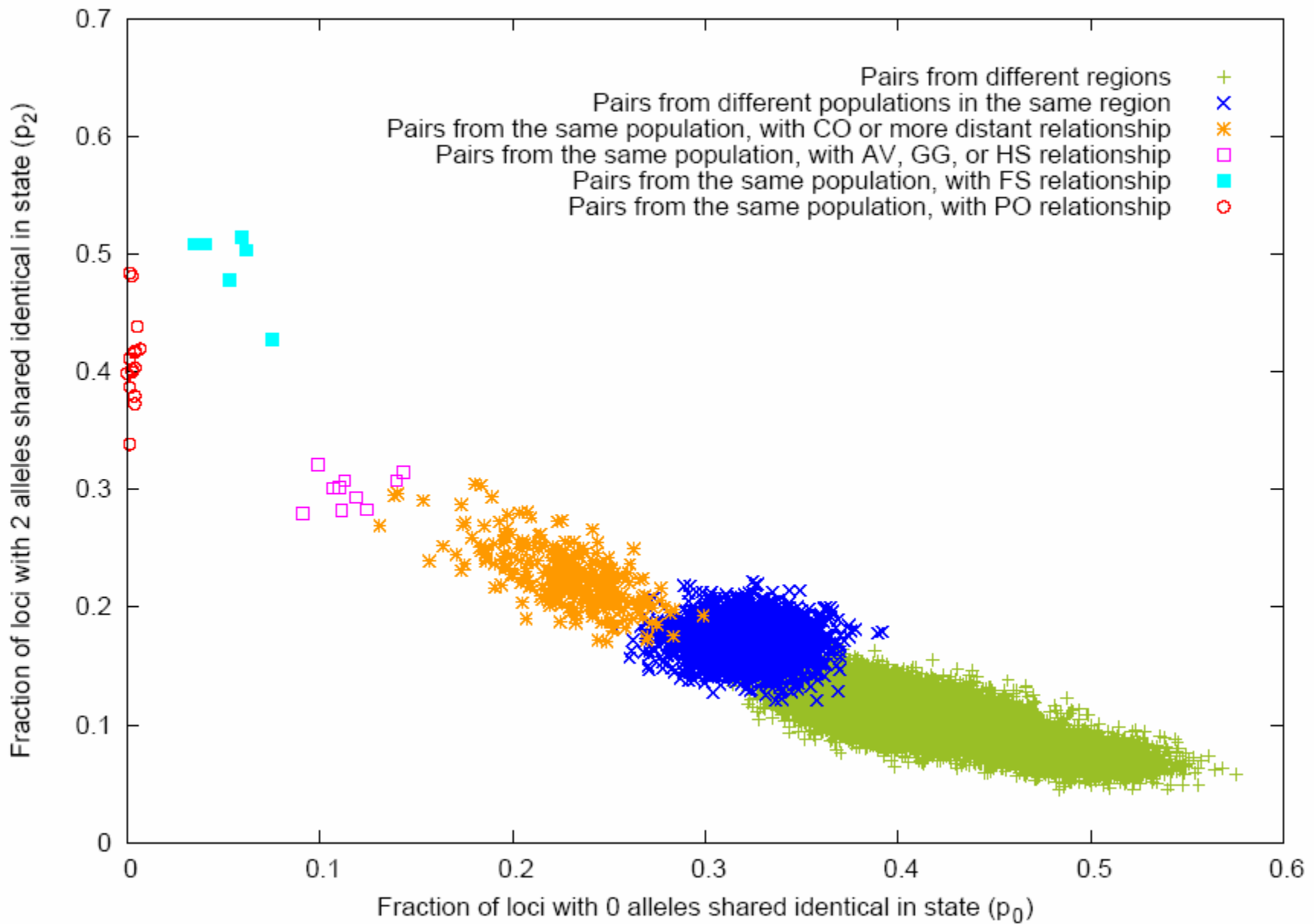
Supplementary Figure 5. Allele sharing for pairs of individuals in which at least one member of the pair is from the Pima population. The plot contains 15 parent/offspring pairs, six full sib pairs, and 10 pairs with second-degree relationships.

Supplementary Figure 6. Allele sharing for pairs of individuals in which at least one member of the pair is from the Surui population. Because the confidence placed in specific relationship inferences was smaller in this population than in other populations, all pairs from the Surui population except for parent/offspring pairs were plotted with the same symbol. The plot contains 15 parent/offspring pairs.