doi: 10.1111/j.1469-1809.2006.00285.x

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, Phone: (734) 615-9556, Fax: (734) 615-6553. E-mail: 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 for use 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 mislabelled) DNAs, duplicated DNAs, and pairs of close relatives included in the diversity panel. The likely mislabellings and duplicates have previously been reported (Rosenberg *et al.* 2002; Mountain & Ramakrishnan, 2005), and the full lists of relative pairs are reported here for the first time.

Three standardized subsets of the original diversity panel are recommended 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 construct these the standardized data sets I began with 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 mislabelling 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 did not suggest that mislabelling 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 by 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), revealed the same duplicate pairs as those reported by 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 mislabellings consideration of all 1066 samples whose genotypes were available did 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 to know 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 errors 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 – or both members for the one instance in which the duplicates had different population labels – yielded 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 had 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 was a pair of Pima individuals with $(p_0, p_1, p_2) = (0.035, 0.457, 0.508)$. 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 identical 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 the Maya one such pair had $p_2 = 0.42$ and no others had $p_2 > 0.28$; in the 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 the 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 (Kidd *et al.* 1993; Calafell *et al.* 1999) so that p_2 must be larger than in other Native Americans for inference of full sib relationships. In the 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 the 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 the 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 the Karitiana and Surui (Kidd *et al.* 1993; Calafell *et al.* 1999), 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 the 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 the 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

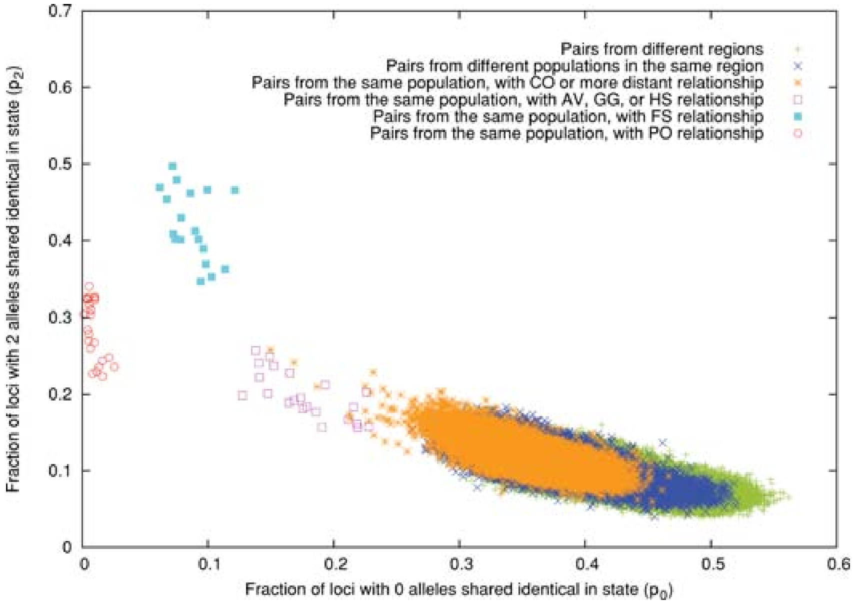

Figure 1 Allele sharing for pairs of individuals from H1048 in which neither member of the pair is a Native American. The plot contains 25 parent/offspring pairs, 18 full sib pairs, and 22 pairs with second-degree relationships. The five pairs with CO or a more distant relationship with the smallest 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).

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-degree relative pairs and of both 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 the Karitiana and Surui – no incompatibilities were observed between different inferences about first-degree relationships. In other words, with some exceptions in the 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 between 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 exclusions required. 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 the 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 the 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

I am very grateful to Howard Cann and Mingyao Li for thorough readings of an earlier draft of the manuscript. This work was supported by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences.

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.

Calafell, F., Shuster, A., Speed, W. C., Kidd, J.-R., Black, F. L. & Kidd, K. K.. (1999) Genealogy reconstruction from short tandem repeat genotypes in an Amazonian population. *Am. J. Phys. Anthropol.* **108**, 137–146.

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.

Kidd, J. R., Pakstis, A. J. & Kidd, K. K.. (1993) Global levels of DNA variation. pp. 21–30 in. *Procedings from the Fourth International Symposium on Human Identification*. Promega Corporation.

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.

Received: 4 November 2005
Accepted: 13 February 2006

Supplementary Material

The following supplementary material is available for this article online:

Tables S1.–S24.

Figures S1.–S6.