RESEARCH

Allele Frequency Distributions in Pooled DNA Samples: Applications to Mapping Complex Disease Genes

Sarah H. Shaw,[1,2] Minerva M. Carrasquillo,[1] Carl Kashuk,[1] Erik G. Puffenberger,[1] and Aravinda Chakravarti[1,3]

[1]Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106 USA

Genetic studies of complex hereditary disorders require for their mapping the determination of genotypes at several hundred polymorphic loci in several hundred families. Because only a minority of markers are expected to show linkage and association in family data, a simple screen of genetic markers to identify those showing linkage in pooled DNA samples can greatly facilitate gene identification. All studies involving pooled DNA samples require the comparison of allele frequencies in appropriate family samples and subsamples. We have tested the accuracy of allele frequency estimates, in various DNA samples, by pooling DNA from multiple individuals prior to PCR amplification. We have used the ABI 377 automated DNA sequencer and GENESCAN software for quantifying total amplification using a 5′ fluorescently labeled forward PCR primer and relative peak heights to estimate allele frequencies in pooled DNA samples. In these studies, we have genotyped 11 microsatellite markers in two separate DNA pools, and an additional four markers in a third DNA pool, and compared the estimated allele frequencies with those determined by direct genotyping. In addition, we have evaluated whether pooled DNA samples can be used to accurately assess allele frequencies on transmitted and untransmitted chromosomes, in a collection of families for fine-structure gene mapping using allelic association. Our studies show that accurate, quantitative data on allele frequencies, suitable for identifying markers for complex disorders, can be identified from pooled DNA samples. This approach, being independent of the number of samples comprising a pool, promises to drastically reduce the labor and cost of genotyping in the initial identification of disease loci. Additional applications of DNA pooling are discussed. These developments suggest that new statistical methods for analyzing pooled DNA data are required.

The availability of meiotic maps of highly polymorphic markers in the human has clearly been a boon to the mapping of Mendelian diseases, complex disorders, and quantitative phenotypes. This success has relied not only on the availability of high-resolution linkage maps and developments in rapid genotyping (Dib et al. 1996), but also on the development of statistical methods for interpreting data on the sharing of genotypes, and thus genomes, within and between families (for review, see Lander and Schork 1994). Complex genetic disorders and phenotypes likely arise from the allelic contributions of multiple genes possibly interacting with environmental and stochastic factors. The first challenge in the genetic dissection of a complex phenotype is in the identification of the candidate map locations of the genes underlying susceptibility/protective alleles via genetic markers; this is met by conducting a genome screen and identifying the relevant loci by recognition of excess/deficiency of allele sharing within families, as compared with that expected from segregation of unlinked genes. The second challenge is the fine-structure localization of any component gene to physical segments small enough to facilitate positional cloning or recognition of candidate genes; this is accomplished by saturating a target genomic segment with polymorphic markers and recognizing allelic associations in families. Because the genetic contribution of any specific gene in a multigenic trait may be small, both highly polymorphic markers and large numbers of families are necessary to detect allelic effects. All of these exercises require the individual genotyping of several hundred polymorphic loci in sev-

[2]Present address: Sequana Therapeutics, La Jolla, California 92037 USA.
[3]Corresponding author.
E-MAIL axc39@po.cwru.edu; FAX (216) 368-5857.

8:111–123 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

eral hundred families, a task that is still labor and cost intensive, and likely to remain so in the near future. Because, even for a complex phenotype, only a minority of all markers studied will show linkage as a result of association, a simple screening procedure that identifies the relevant marker loci from a larger set would be of great value. The studies described in this paper show that accurate, quantitative data on allele frequency in pools of DNA samples can be obtained under specified conditions.

We suggest that pooling specific classes (parents, offspring) of relatives in a collection of families and estimating allele frequencies within such pools can be used to rapidly screen the genome for linked markers by virtue of marker associations with phenotypes. Once identified, direct genotyping of a reduced marker set is necessary to enable statistical analysis of the genotype data. Finally, developments of new statistical methods to analyze pooled DNA data are necessary to identify the relevant markers with high accuracy, that is, low false positive and false negative rates. We show the utility of these methods for the identification of one Hirschsprung disease (HSCR) susceptibility locus (HSCR2) in the Old Order Mennonites of Lancaster County, Pennsylvania (Puffenberger et al. 1994a,b).

The idea of using pooled DNA samples to reduce the burden of genotyping is not new and, to our knowledge, was first suggested by Arnheim et al. (1985) in the context of case-control studies. These authors argued that alleles in linkage disequilibrium with a disease would be enriched (or deficient) in a pooled sample of affected individuals in comparison with a pooled control sample, and thus, beyond testing association for specific alleles, this principle could be used to search for associated alleles at specific genes, as they successfully did for HLA class II DR and DQ alleles in insulin dependent diabetes mellitus (IDDM) (Arnheim et al. 1985). Similar ideas have been prevalent in the plant genetics community where the search for linked markers in pools of progeny classified by phenotype, within a segregating cross, has been termed bulked segregant analysis (Michelmore et al. 1991). In these studies, the intensity of DNA hybridization (Southern blot) of a labeled probe, in two contrasting (presence vs. absense of some phenotype) pooled DNA samples, is used to recognize a linked (Michelmore et al. 1991) or associated marker (Arnheim et al. 1985). Quantitation of signals in blot hybridization experiments is possible but requires multiple controls and is difficult to standardize; however, qualitative comparisons, when signal intensities are very different, as arising from large allelic effects, are easier to detect. These experiments are feasible in experimental crosses in which no more than four different alleles can segregate within a cross, but can lead to considerable difficulty in outbred families, such as in humans, where many more marker alleles can segregate in a family collection. Not surprisingly, initial applications of this method in the human have been in genetic studies in isolated populations where allelic diversity is reduced (Puffenberger et al. 1994a,b; Sheffield et al. 1994; Carmi et al. 1995; Nystuen et al. 1996; Scott et al. 1996). The development of the PCR and the identification of microsatellite repeats as a common source of polymorphism led Pacek et al. (1993) to demonstrate that allele-specific signals could be quantitated in DNA pools. Pacek et al. (1993) showed that allele frequencies at loci with length polymorphisms could be estimated by quantitative analysis of the PCR products from pooled DNA samples. In particular, these authors demonstrated the accuracy of these estimates in pools of up to 1350 samples, by comparing their results with published estimates of allele frequencies.

The availability of high-resolution genetic maps of microsatellite markers in mice and humans led several investigators to suggest that DNA pooling could be used for genetic mapping. In particular, pools of DNA samples from ''affecteds'' and ''unaffecteds'' can be screened for genetic markers spanning a genome to identify loci with a marker allele that has a differential distribution (association) in the two pools. Although linkage does not lead to any permanent population association per se, linkage does lead to marker associations within a segregating cross or a large kindred or in isolated populations. In the mouse, both Mendelian (Asada et al. 1994; Taylor et al. 1994) and multigenic (Collin et al. 1996; Taylor and Phillips 1996) traits have been mapped by scanning the genome with pooled samples. In humans, Sheffield and colleagues have been prominent in applications of genome scanning in isolated populations by microsatellite marker analyses in DNA pools: In particular, the genes for Bardet–Biedl syndrome (Sheffield et al. 1994; Carmi et al. 1995), cerebellar ataxia (Nystuen et al. 1996), and autosomal recessive nonsyndromic hearing loss (Scott et al. 1996) have been mapped in this manner. The human studies on Mendelian recessive traits were carried out in genetically isolated populations with the expectation that affected individuals would be homozygous for a single marker allele at a closely linked locus. Thus, anonymous markers near the disease gene could be identified by visual inspection of either silver-stained or radioactively labeled microsatellite markers, and quantitation was not necessary. For complex phenotypes, however, quantitation is crucial because no single allele at any locus is necessary or sufficient for the disease phenotype. Additionally, for many population genetic and evolutionary studies the entire distribution of alleles is required.

We used the ABI 377 automated DNA sequencer and GENESCAN software for quantifying allele amplification at polymorphic microsatellite markers using 5′ fluorescently labeled forward PCR primers; allele frequencies were estimated from the relative values of the peak heights corresponding to each allele detected in the pooled sample. We have quantitatively tested the accuracy of allele frequencies estimated in this manner in comparison with that estimated by direct genotyping. We tested 11 microsatellite markers in two separate DNA pools and an additional four markers in a third DNA pool, using parents from the CEPH reference pedigrees. In addition, we extended this method to estimate the distribution of polymorphic alleles on transmitted and untransmitted chromosomes in parent-offspring trios in a large inbred Mennonite kindred segregating HSCR (Puffenberger et al. 1994a,b), as an approach to gene mapping using linkage disequilibrium. Thereby, we show how allele quantification can lead to direct statistical tests of linkage and association in pooled DNA samples.

RESULTS

We used a total of 14 polymorphic microsatellite markers in DNA pooling experiments with characteristics as provided in Table 1. Our first aim was to determine the fidelity of PCR for estimation of allele frequencies in pooled samples. Although Pacek et al. (1993) had shown that frequencies of alleles that differ even by one repeat motif (e.g., CA) could be accurately estimated, when compared with direct genotyping, the total variation expected in the PCR is unknown. Moreover, it is not unexpected that smaller-sized alleles would show preferential amplification. Although these investigators compared their results from pooling with some data obtained by direct genotyping, the same samples were not tested, and the standard comparison used published allele frequencies. Thus, we used the tetranucleotide marker VWF, within the von Willebrand factor gene, and assayed a DNA pool constructed from 76 unrelated samples (152 alleles) from the CEPH reference pedigrees. Table 2 shows estimates of each of the seven alleles at the VWF locus in 10 replicate PCR experiments on a single DNA pool, and the mean of the 10 replicates, obtained by quantitation on an ABI 377 automated sequencer, in comparison with that determined by direct genotyping of each individual in the DNA pool. These results show that there is little variation between replicate PCR experiments on the same pool because the standard deviation of the replicate measures (0.002–0.007) is, on average, one order of magnitude smaller than the sampling standard deviation of the allele frequency estimates (0.007–0.036) (Table 2). Thus, not only can the allele frequencies estimated from DNA pools be quantitative and accurate, but the results are reproducible in that replicate-to-replicate variation is small. To assess the accuracy of the pooled estimates with the true values, we compared the root-mean-square-error (RMSE) between each replicate estimate and the direct counts. These values ranged between 0.013 (pool 6) and 0.023 (pool 1); the average values of the replicates had a RMSE of 0.018, that is, any allele frequency x is in the range $x \pm 1.8\%$. Thus, although multiple replicates can be helpful in adding confidence to the estimates, they do not substantially increase the accuracy expected on averaging.

The results from marker VWF appear to generalize only to marker loci that amplify in a clean

Table 1. Polymorphic Markers Used for Genotyping

Locus	Repeat motif	Dye label	No. of alleles	Hetero-zygosity[a]
D13S792	tetra	HEX	7	0.57
D13S160	di	TET	9	0.81
D13S317	tetra	FAM, HEX, TET	7	0.82
D13S170	di	HEX	11	0.91
D13S921	tetra	HEX	7	0.69
D13S790	tetra	FAM	5	0.63
D13S764	tetra	TET	4	N.D.
GATA8G07	tetra	TET	6	0.87
D13S628	tetra	FAM	7	0.69
D13S281	di	HEX	4	0.62
D1S1660	tetra	FAM	7	0.83
D9S301	tetra	HEX	9	0.75
D10S1423	tetra	FAM	7	0.93
VWF	tetra	FAM	6	0.80

Shown are each polymorphic marker locus used, its repeat motif, the fluorescent dye used for labeling the forward primer, and the number of alleles and heterozygosity as obtained from the Genome Data Base.

[a](N.D.) No independent estimate available.

Table 2. Accuracy of Allele Estimates from DNA Pooling

Allele	Pool										Average	Direct
(bp)	1	2	3	4	5	6	7	8	9	10	(S.D.)	estimate
138	0.08	0.09	0.09	0.09	0.09	0.09	0.08	0.08	0.09	0.08	0.086 (0.003)	0.081 (0.022)
142	0.04	0.04	0.05	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.043 (0.002)	0.054 (0.018)
146	0.26	0.27	0.27	0.27	0.25	0.27	0.26	0.27	0.26	0.26	0.262 (0.005)	0.264 (0.036)
150	0.31	0.30	0.30	0.29	0.31	0.28	0.30	0.30	0.30	0.30	0.298 (0.007)	0.257 (0.035)
154	0.21	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.21	0.20	0.201 (0.004)	0.209 (0.033)
158	0.10	0.11	0.11	0.11	0.11	0.12	0.12	0.10	0.11	0.11	0.110 (0.005)	0.128 (0.027)
162	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.000 (0.000)	0.007 (0.007)

A tetranucleotide polymorphism within the VWF gene (12p13.3–12p13.2) was studied in 10 replicates of a single DNA pool constructed from 76 unrelated samples of CEPH parents. Quantitation of allele frequencies from relative fluorescence peak heights in each replicate, the average of these 10 replicates, and the corresponding estimate from direct genotyping are shown. Also shown in the last two columns are the estimated standard deviations from the 10 replicates and the direct counts, respectively.

manner, that is, without considerable PCR stutter. To assess the behavior of various marker loci, we studied two tetranucleotide (VWF, D13S317) and three dinucleotide (D13S160, D13S170, D13S281) polymorphisms in a pool of 54 (108 alleles) Mennonite parents (see Methods). In Figure 1, we present comparisons of allele frequencies at each of the five loci estimated from DNA pools and from direct genotyping. The values plotted for each pool are an average of 10 replicate PCR experiments. The pooled and direct estimates of allele frequencies are all highly correlated (r = 0.935, 0.998, 0.951, 0.885, and 0.970 for VWF, D13S317, D13S170, D13S160, and D13S281, respectively). The dinucleotide marker allele frequencies studied, however, show considerable variation around the expected value (RMSE = 0.014, 0.100, and 0.016 for D13S170, D13S160, and D13S281, respectively), both for low and high allele frequencies; moreover, the variation is quite marker locus dependent. The tetranucleotide markers provide data of greater accuracy because the correspondence between pooled and direct estimates are higher (RMSE = 0.018 and 0.010 for VWF and D13S317, respectively). Among the dinucleotide polymorphisms studied, D13S281 gave the most accurate allele frequency estimates, reflecting the fact that it has only four alleles and shows little PCR stuttering. In contrast, allele frequencies for D13S160 were remarkably inaccurate, reflecting its complex pattern and extreme stuttering on gel electrophoresis. In general, tetranucleotide markers appear to perform better in DNA pools, although pilot tests with a genome-wide set of microsatellite markers will be necessary to choose optimal loci.

Currently available automated DNA sequencers can detect signals on the basis of the fluorescence of multiple dye labels. We tested the three dyes FAM, HEX, and TET to determine whether a particular dye label provided more accurate allele frequency estimates. The data in Figure 2 show allele frequencies at D13S317 for both the transmitted (T) and untransmitted (U) chromosomes in 27 Mennonite HSCR trios. Allele frequencies from each T and U DNA pool of 54 chromosomes were estimated by use of three different dye-labeled forward PCR primers and compared with direct counts, as shown; pooled estimates arose from five replicates. Although each dye can estimate allele frequencies accurately (r = 0.970, 0.981, and 0.991 for TET, FAM, and HEX, respectively), the FAM dye appears to yield somewhat more accurate estimates (RMSE = 0.058, 0.028, and 0.034 for TET, FAM, and HEX, respectively). We have no simple explanation for this finding, which requires confirmation from other markers. Statistically, the results from the three dyes are not significantly different from one

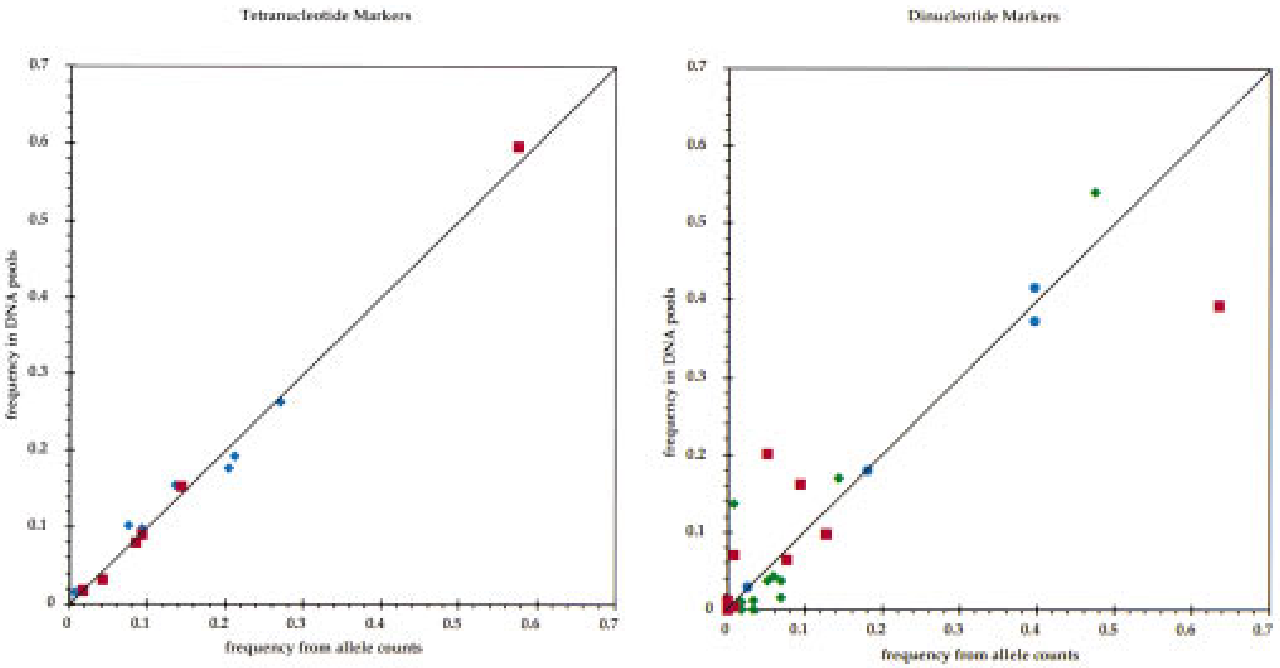

Figure 1 Accuracy of allele frequencies estimated from DNA pools. Allele frequencies were estimated from relative peak heights from the ABI 377 fluorescence chromatograms (frequency in DNA pools: *y*-axis values are averages of 10 replicates) and compared with allele frequencies estimated from genotyping individual samples (frequency from allele counts: *x*-axis). These data, from two tetranucleotide repeat markers, VWF (blue diamonds) and D13S317 (red squares) (*left*), and three dinucleotide repeat markers, D13S160 (red squares), D13S170 (green diamonds), and D13S281 (blue circles) (*right*), were obtained by studying 76 unrelated samples (parents) in CEPH.

another. Thus, for DNA pooling experiments all dyes behave satisfactorily.

We wished to test whether DNA pooling could be used as a preliminary screen to identify marker loci that were associated with a phenotype. Specifically, we wished to recapitulate the haplotype relative risk (HRR; Falk and Rubinstein 1987) test on parent–offspring trios. Consequently, we tested the accuracy of allele frequency estimates at 11 marker loci in a pool of 54 parental DNA samples and a pool of their 27 affected offspring. The samples were 27 trios in which each offspring (proband) had HSCR. These trios were ascertained from a large, inbred kindred segregating HSCR in the Old Order Mennonites of Lancaster County, Pennsylvania (Puffenberger et al. 1994a). We tested the markers D13S317, D13S790, D13S792, D13S921, GATA8G07, D13S628, and D13S764 known to map on human chromosome 13q22 and in a region that contains *EDNRB*, a HSCR susceptibility gene (HSCR2). We also included four control markers, the VWF gene (12p13.3–12p13.2), D1S1660, D9S301, and D10S1423, which are not known to be associated with HSCR in the Mennonites. For each marker we estimated the allele frequency distribution on T and U chromosomes as outlined in Methods.

A chromosome 13 radiation hybrid (RH) map containing *EDNRB* and flanking markers was constructed to test the maximal distance from *EDNRB* at which linkage disequilibrium could be detected. In addition, the map was also used to compare the sensitivity of direct genotyping versus DNA pooling for determining linkage disequilibrium. The resulting RH map is shown in Figure 3. The initial map was constructed by use of the microsatellite markers only; physical mapping data described in Puffenberger et al. (1994b) shows that *EDNRB* is located on the same YAC (754a2: 1.36 Mb) as D13S317. The marker D13S921 could not be placed on the RH map with high odds but is located in the interval between D13S170 and D13S790. All remaining markers are supported with at least 100:1 odds, except for D13S792, which is 20 times more likely to be in its most proximal position. In this RH panel, 1 cR_{9000} is equivalent to ~70 kb (Shaw et al. 1995), so that *EDNRB* is estimated to be located in an interval no more than ~20 cR_{9000} (1.36 Mb/70 kb/cR) on either side of D13S317.

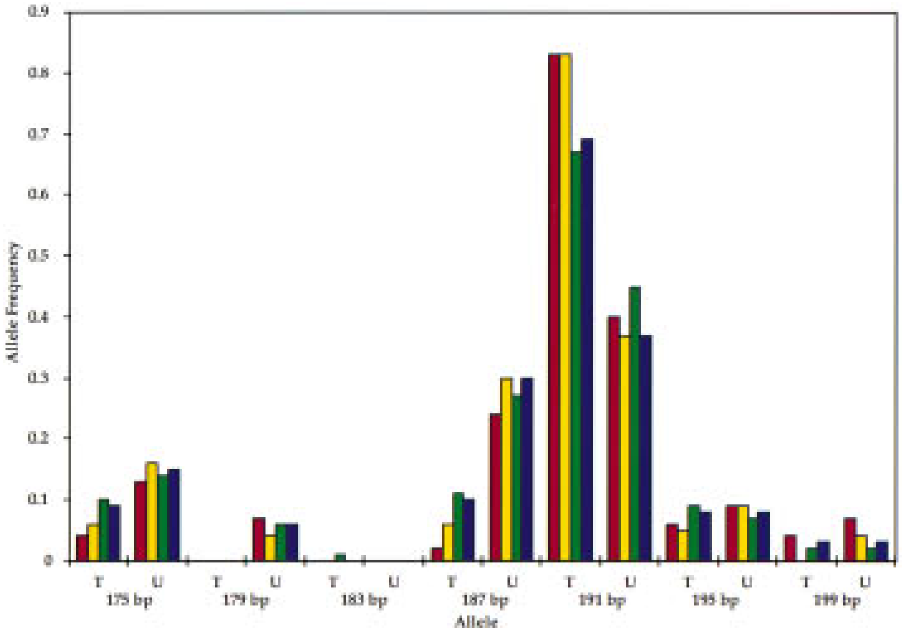

Figure 2 The effect of dye label on DNA pooling. Allele frequencies at D13S317 were estimated from DNA pools and compared with direct counts (red bars) as indicated in Fig. 1. This marker locus was examined in 54 Mennonite parents and their offspring. D13S317 T and U parental allele frequencies were estimated from amplification of a DNA pool by use of three different dye-labeled forward PCR primers [FAM (purple bars), HEX (gold bars), TET (green bars)], each the average of five replicate PCR experiments. Shown are allele frequencies for each allele in the sample and for T and U chromosomes separately.

Allele frequencies on T versus U chromosomes for each marker are shown in Figure 4 for the associated polymorphisms and in Figure 5 for the control markers. We show both the pool estimates and the direct count frequencies. This analysis was restricted to polymorphisms caused by tetranucleotide markers only. Visual inspection of the DNA pool data show clearly that the 191-bp allele at D13S317, 183-bp allele at D13S790, 287-bp allele at D13S792, 214-bp allele at D13S921, 197-bp allele at GATA8G07, 250-bp allele at D13S628, and the 314-bp allele at D13S764 are transmitted in excess to the HSCR affected offspring as compared with the control (U) data. This is confirmed by estimating the probability of each allele as being ancestral (Terwilliger 1995); the above listed alleles have probabilities exceeding 90% of being the ancestral alleles. The direct counts of T and U alleles were compared by use of a χ^2 test as outlined in Puffenberger et al. (1994a). The associated allele for each marker, the respective χ^2 value, and corrected *p*-value are shown in Table 3. Concordant with our previous mapping of HSCR in the Mennonites (Puffenberger et al. 1994b), all markers show statistically significant differences between T and U chromosomes. The marker GATA8G07, located ~194 cR distal to *EDNRB*, does not show statistically significant linkage disequilibrium for allele 197 bp because the true associated allele has high frequency (45%) on control chromosomes. In comparison, even the most distal marker, D13S628, shows a statistically significant linkage disequilibrium because the control allele frequency is much lower (23%). In a similar manner, we compared the T and U allele frequencies estimated from pooled DNA samples using the binomial proportions test with the normal deviate for assessing significance, because absolute counts cannot be obtained from DNA pools. The associated allele for each marker, the normal deviate statistic, and corrected *P* value are also shown in Table 3. The results are remarkably similar to those obtained from direct counts except the two most distal markers, GATA8G07 and D13S628, which did not show statistical significance. In particular, for the marker GATA8G07, the pool estimates were slightly more divergent (T vs. U is 0.64 vs. 0.42) than the direct counts (T vs. U is 0.60 vs. 0.45), but were still not significant. On the other hand, for the marker D13S628, the pool estimates were less divergent (T vs. U is 0.46 vs. 0.32) than the direct counts (T vs. U is 0.55 vs. 0.23) and therefore did not lead to a significant difference. Thus, small allele frequency differences, affected by additional variation from pooling, may not be detectable in small samples. Importantly, none of the control markers, D1S1660, D9S301, D10S1423, and VWF, showed statistically significant differences between T and U chromosomes in either the allele counts from direct genotyping or those estimated from DNA pools.

The results presented above show DNA pooling to give allele frequencies that are highly correlated with direct counts, and accurate as judged by RMSE values. To assess the statistical (numerical) relationship between pooled estimates (*y*) and direct counts

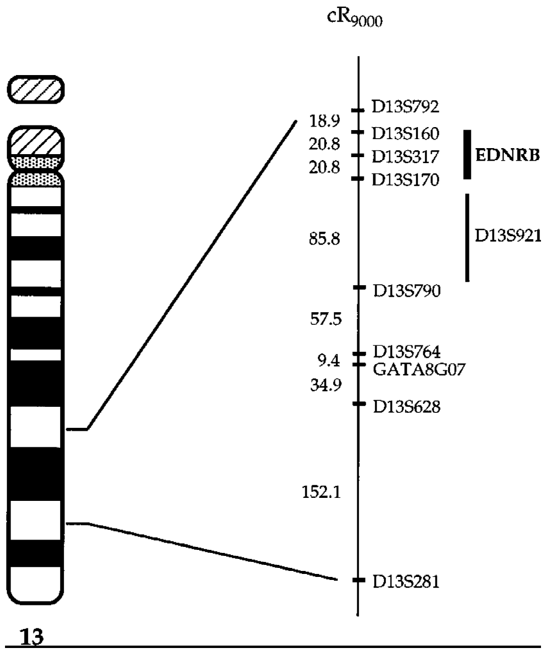

Figure 3 RH map of segment 13q22–13q31 on human chromosome 13. The map shows the location of all chromosome 13q markers used in DNA pooling for localizing a HSCR susceptibility gene. All uniquely placed markers are supported with odds of at least 100:1, except for D13S792 (supported with 20:1 odds); D13S921 maps to the interval between D13S170 and D13S790. The HSCR gene *EDNRB* is located within 1.4 Mb of D13S317 as determined from YAC physical mapping (Puffenberger et al. 1994b). In this chromosome 13 RH panel, 1 cR_{9000} is ~70 kb of DNA (Shaw et al. 1995).

(x), we performed linear regression analysis on all 131 alleles genotyped on both T and U chromosomes at the tetranucleotide markers. The results are shown in Figure 6. The linear regression is highly significant ($F = 1112.1$, 1 and 129 df, $P = 2.8 \times 10^{-65}$) with a correlation coefficient of 0.95. The fitted regression is shown by the solid line in Figure 6. This regression has an intercept of 0.011 (95% CI: -0.002–0.024), which is not significantly different from zero; the slope has an estimate of 0.933 (95% CI: 0.877–0.988) and shows a small, yet significant, departure from unity. If, however, the slope is assumed to be unity, then the errors (residuals) arising from pooling, that is, the values of $y - x$, are normally distributed with mean ~0 and a standard deviation of 0.05 ($\chi^2 = 3.28$, 4 df, $P = 0.51$). This alternative fit is shown by the broken line in Figure 6. Thus, for all practical purposes, the errors induced by DNA pooling are random and small. Whether these errors are acceptable or not depends on the particular application under consideration.

DISCUSSION

The results of this study show that quantitative data on allele frequencies can be obtained from a pool of DNA samples. Because allele frequencies are obtained from relative fluorescence peak heights, the procedure is statistical and thus introduces some error. As we have shown, this error is on average zero but subject to random variation that is small and may be simply assessed from the RMSE value. It appears that, to obtain accurate estimates, it may be necessary to utilize polymorphic markers that have a limited number of alleles (~5–8) and that can be amplified with little PCR artifacts. As expected, the tetranucleotide markers, on average but not individually, satisfy these requirements better than dinucleotide markers, although some dinucleotide markers perform equally well. For the 14 markers used in this study, the RMSE values ranged between 0.011 and 0.096. Thus, it may be necessary to perform pilot experiments on a large number of microsatellite polymorphisms and use the RMSE values to identify those markers that can be reliably used in DNA pooling experiments. Of course, the specific application desired determines the RMSE that can be tolerated.

DNA pooling can be an efficient genetic tool because it is sample size independent, that is, independent of the number of samples comprising the pool. The results presented here, and those of Pacek et al. (1993), show that one may obtain allele frequencies in pools of 100–1000 DNA samples. The crucial aspect is to maintain and increase the accuracy of the frequency estimates. It is well known that PCR may be biased towards greater efficiency of amplification of shorter rather than longer DNA templates, thus biasing estimates from DNA pools. Our data, however, do not show any large trend of frequencies from smaller alleles being overestimated and frequencies of larger alleles being underestimated (Fig. 6). These effects, if they exist, are minor and do not appear to significantly affect allele frequency estimation. We have also shown that all three fluorescent dyes commonly used with the automated sequencers can be used for accurate quantitation of peak heights. Additionally, better statistical tools are needed to estimate allele frequencies from DNA pools. Although we have used peak height as a natural statistic, one may use the area under the peak as well or use the total fluorescent trace for estimation. These aspects require a full investigation.

All DNA pooling experiments are exercises in estimating the allele frequency distribution of a

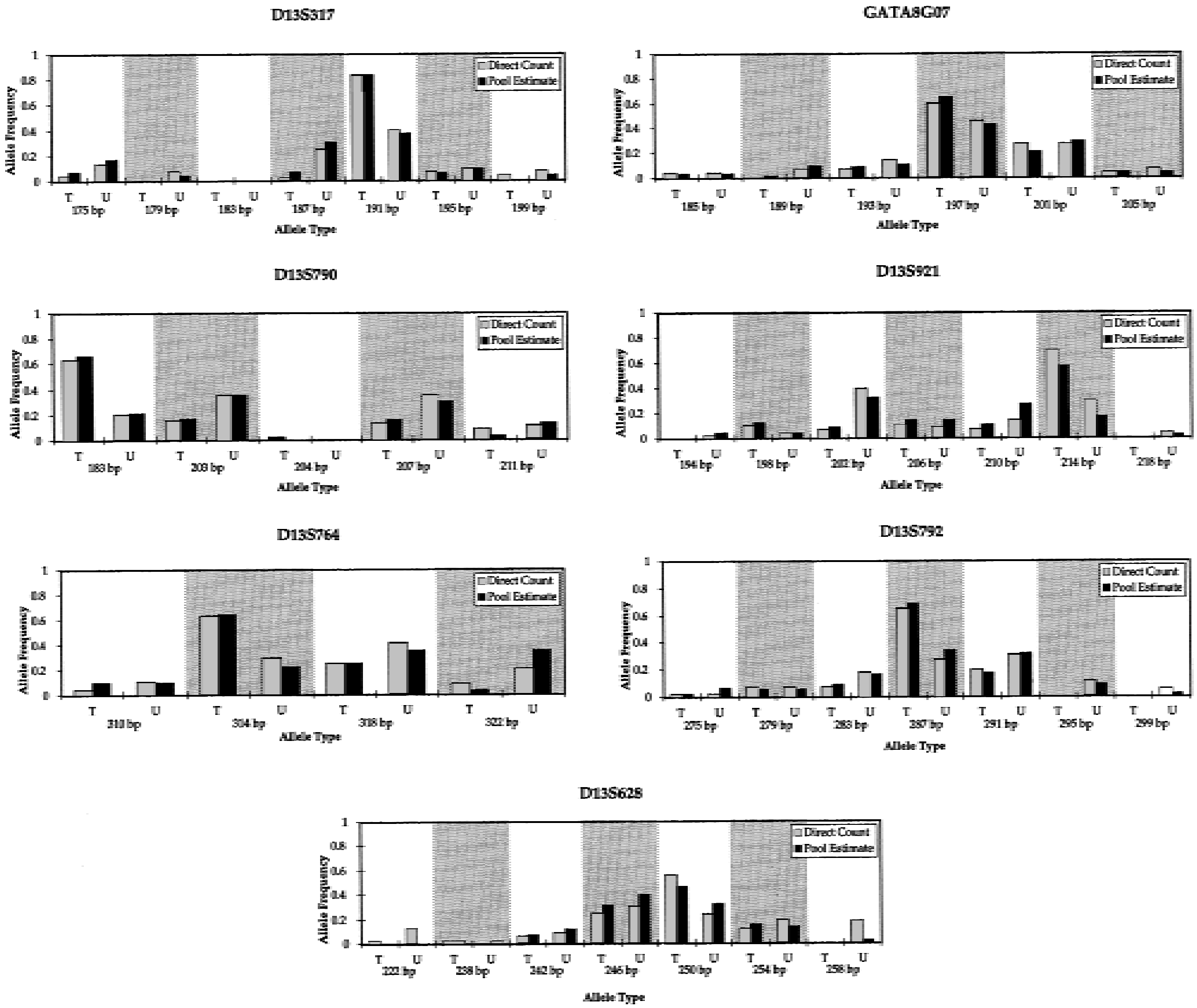

Figure 4 DNA pooling of HSCR-associated genetic markers. T and U parental allele frequencies estimated from direct genotyping and from two DNA pools (27 HSCR affected offspring and their parents) for markers on human chromosome 13q flanking the HSCR susceptibility gene *EDNRB.* For visual clarity, white and shaded stripes have been used to separate T and U frequencies on adjacent alleles. The legend follows that in Fig. 2.

sample; consequently, for gene mapping, we have to rely on frequency differences in alleles in family subsamples. Thus, the most direct applications of DNA pooling are in association studies in which cases and controls are compared (Arnheim et al. 1985). On the other hand, linkage mapping in a set of small independent families, such as affected sibling pairs, cannot benefit from DNA pooling because there is no expected association of alleles across families. The exception, and a major one, is the study of individuals in a large, segregating cross. The latter is clearly possible in any experimental cross in which pools can be constructed on the basis of the segregating phenotype in large human kindreds and in isolated populations. This exception arises because common descent within a cross, kindred or within an isolated population, leads to association of phenotypes at linked marker loci so that allele frequencies are expected to be different in pools of affected versus unaffected pools. It is this principle that allows us to use DNA pooling to map genes that are associated with linked markers either in pedigrees or in parent–offspring trios. Our results clearly show the utility of this method, as evidenced by our success in mapping a HSCR susceptibility gene.

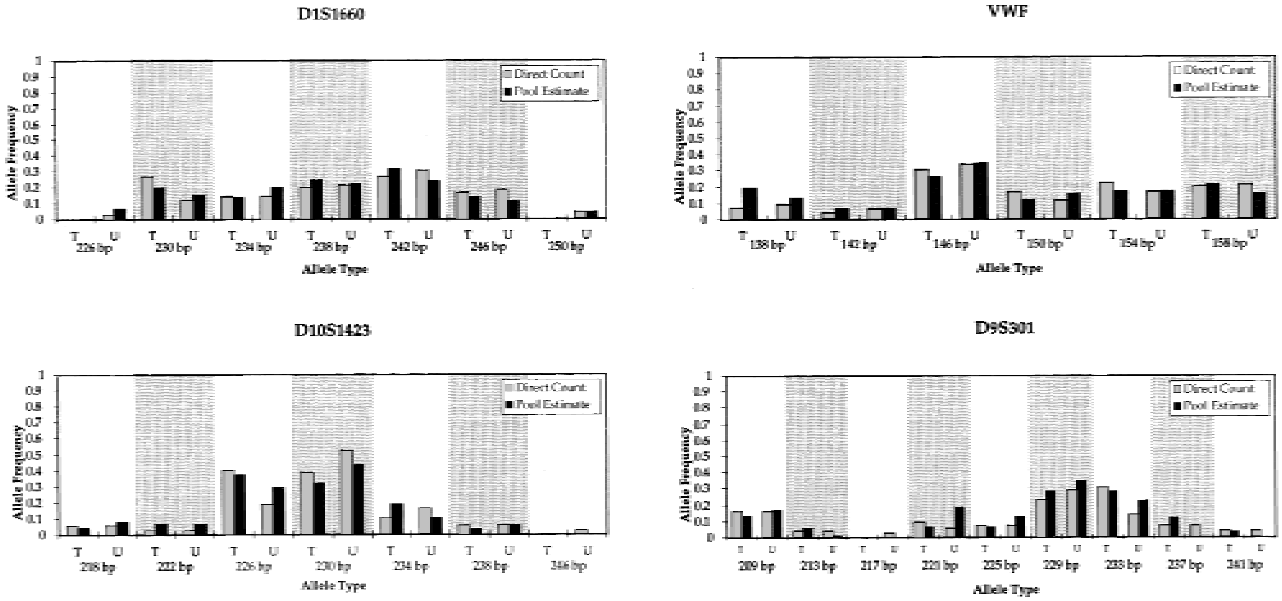

Figure 5 DNA pooling of HSCR control genetic markers. T and U parental allele frequencies determined from direct genotyping, estimated from two DNA pools (27 HSCR affected offspring and their parents) for control markers not linked to the HSCR susceptibility gene. For visual clarity, white and shaded stripes have been used to separate T and U frequencies on adjacent alleles. The legend follows that in Fig. 2.

Our results on HSCR2 show that disease susceptibility/protective loci can be identified by genome screening in trios sampled from an isolated population. Even in samples from outbred populations, genes may be mapped by association studies across the genome, provided there are associations between a phenotype and genetic markers in its vicinity. This is not wholly unexpected in humans, where large population expansions in the past several hundred generations and the young age of the species can lead to such associations. Consequently, mapping phenotypes by associations in parent–offspring trios by screening the human genome with a very dense set of markers has been proposed (Risch and Merikangas 1996). Although the TDT (transmission disequilibrium test) of Spielman et al. (1993) has been proposed for this purpose, the previously mentioned HRR test (Falk and Rubinstein 1987), which we used for the HSCR analysis, is also useful. The only impediment to this approach is the

Table 3. Detection of Allelic Association from DNA Pools

		Direct counts				Pool estimates			
Locus	Associated allele (bp)	T	U	χ^2	*P* value[a]	T	U	z	*P* value[a]
D13S792	287	30	12	13.0	<0.001	0.68	0.34	3.53	0.001
D13S317	191	40	18	18.6	<0.001	0.83	0.37	4.88	<0.001
D13S921	214	36	14	16.7	<0.001	0.56	0.16	4.33	<0.001
D13S790	183	30	9	17.8	<0.001	0.66	0.21	4.72	<0.001
D13S764	314	30	14	10.7	0.004	0.64	0.22	4.41	<0.001
GATA8G07	197	31	23	2.2	N.S.	0.64	0.42	2.3	N.S.
D13S628	250	29	13	10.1	0.007	0.46	0.32	1.49	N.S.

Seven 13q genetic markers near the HSCR susceptibility locus were used to assess linkage disequilibrium by direct genotyping and from quantitation in DNA pools. The statistical significance (*P* value) of differences in allele counts/frequencies on transmitted (T) vs. untransmitted (U) chromosomes in Mennonites segregating HSCR were assessed either by a χ^2 or a normal deviate test.
[a](N.S.) Not significant.

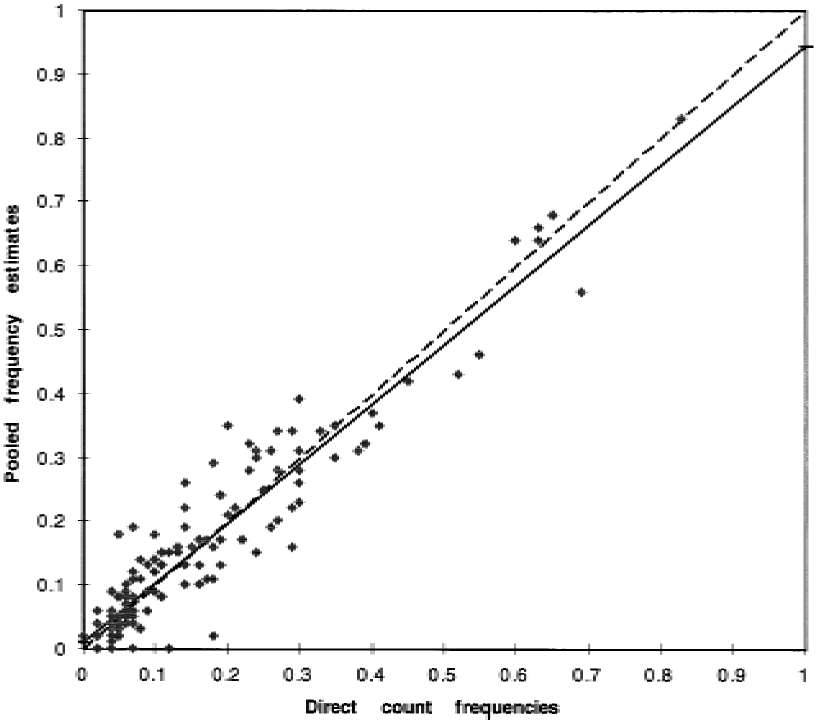

Figure 6 The relationship between pooled and direct allele frequencies. The *y*-axis shows the allele frequency estimate from a DNA pool, and the *x*-axis that from direct genotyping. The data are 131 alleles from all T and U chromosomes compared at the tetranucleotide markers in the Mennonites. The solid and the broken lines show the fitted linear regression and the line expected from perfect agreement, respectively.

identification of a sufficient number of polymorphic markers. Genome screening by association in trios is now feasible for some isolated populations, as we have recently shown for the Mennonites using ~500 microsatellite markers (E.G. Puffenberger, A. Lynn, C. Kashuk, and A. Chakravarti, unpubl.). The advent of biallelic markers that can be rapidly genotyped with a DNA chip (Chee et al. 1996) suggests that genome screening by association may be feasible for outbred populations as well. In fact, because allele-specific fluorescence can be quantitated very accurately on a DNA chip (Chee et al. 1996), DNA pooling may become a desired method for genome-wide genetic studies, as a very large number of markers can be evaluated.

In the example provided by this study, linkage disequilibrium could be detected from estimated transmitted and untransmitted parental allele frequencies. Individuals affected with HSCR in the Mennonite population are estimated to be 8 to 12 generations removed from a single ancestral couple. The disequilibrium could be detected in markers as far away as 185 cR_{9000} from the *EDNRB* gene. The marker D13S628, located 229 cR_{9000} distal from the *EDNRB* gene was significant when the direct genotype allele counts were used, but was not significant when estimated allele frequencies from DNA pools were used. Therefore, it appears that the DNA pooling strategy is not as sensitive as direct genotyping for detecting linkage disequilibrium for markers further away, however, at markers within 185 cR_{9000} (~13 Mb) in this population, the sensitivity is very consistent between the two methods. Thus, in general, DNA pooling can be very useful for identifying possibly associated alleles that may require confirmation by direct genotyping. The limitations of detecting linked loci from T versus U comparisons arise from the statistical test used or the sample size surveyed, and not from DNA pooling per se. In fact, because DNA pooling is sample size independent, this method can lead to increased statistical power.

There has been much recent discussion on the relative statistical validity of the TDT (Spielman et al. 1993) versus the HRR (Falk and Rubinstein 1987; Knapp et al. 1993) tests for gene mapping. The DNA pooling method uses the HRR test and cannot utilize the TDT test because the latter uses heterozygous parents only. Spielman and Ewens (1996) have reviewed this area to conclude that the HRR test is not statistically valid (i.e., leads to a higher false positive rate than that set by the χ^2 distribution) when singleton affected families are studied in a population with substructure (subdivision or recent admixture). Of course, the magnitude of the error introduced depends on the magnitude of human population substructure. Clearly, then, DNA pooling may not be useful for gene mapping when it is the only method used and in cases where the HRR test is invalid. For segregating crosses, large kindreds, and isolated populations, however, this problem cannot occur and our proposed methods are valid. We would argue that DNA pooling can and should be used in all situations followed by individual genotyping to sort out false from true positive results. This prescription is a practical matter and does not depend on the intrinsic validity of a particular statistical test.

In this paper we have concentrated largely on genetic mapping and association applications of DNA pooling. The primary aim of this method, however, is to obtain accurate and quantitative estimates of allele frequencies. We envision that DNA pooling can play a significant role in evolutionary studies of genetic distance and population affinities (for diverse applications, see Nei 1987), in which allele frequency distributions can be estimated at hundreds of loci in multiple populations with relative ease. Additionally, this method will also be useful for monitoring the genetic changes throughout a genome in experimental studies geared at elucidating the mechanisms of evolution (Buri 1956).

METHODS

Patient and Control Samples

For controls, we used 76 DNA samples from unrelated parents within the CEPH reference pedigrees; they were purchased from the Coriell Cell Repository (Camden, NJ). Twenty-seven individuals (probands) affected with HSCR and their respective parents were ascertained from a large, inbred kindred from the Old Order Mennonite community in Lancaster County, Pennsylvania. Derivative families who had migrated to other states within the USA were also ascertained. The identification of this kindred and ascertainment of individuals have been described previously in Puffenberger et al. (1994a). All DNA samples were collected under informed consent and processed as described previously (Puffenberger et al. 1994a).

DNA Pools Constructed

The concentration of each DNA sample was read on both a Beckman DU 640 (Schaumburg, IL) UV spectrophotometer and a Hoefer TKO 100 (San Francisco, CA) fluorometer. Each sample was diluted to a concentration of 10 µg/ml and reread to confirm the DNA concentration. Equal amounts of DNA from each sample constituting a pool were manually combined prior to PCR amplification. We constructed three DNA pools. The first DNA pool consisted of 27 individuals affected with HSCR. Previously, we have mapped and identified the major HSCR susceptibility locus in Mennonites (HSCR2) to human chromosome 13q22 and showed a missense mutation in residue 276 (W276C) in the gene encoding endothelin receptor B (*EDNRB*) (Puffenberger et al. 1994a,b); the genotypes of the 27 probands in the pooled sample were 13 W276C/W276C homozygotes, 10 W276C/+ heterozygotes, and 4 +/+ homozygotes, in which + denotes the wild-type allele. The second DNA pool consisted of all 54 parents of these 27 probands; the genotypes of these individuals were 3 W276/W276 homozygotes, 33 W276/+ heterozygotes, and 18 +/+ homozygotes. A third DNA pool consisted of 76 unrelated CEPH individuals (parents).

Polymorphic Markers

The two Mennonite DNA pools and the CEPH DNA pool were initially genotyped at three dinucleotide repeat polymorphisms (D13S160, D13S170, and D13S281) and two tetranucleotide repeat polymorphisms (VWF and D13S317). Subsequently, the two Mennonite DNA pools were genotyped at six additional tetranucleotide repeat polymorphisms all mapping to the *EDNRB* (HSCR2) region on chromosome 13q22. Four polymorphisms (VWF, D1S1660, D9S9301, and D10S1423) that are not linked to the *EDNRB* mutation in Mennonites were genotyped as control markers. Table 1 lists the genetic markers we used in this study, the fluorescent dye label used for each marker, and the observed number of alleles and heterozygosity of each marker as published in the Genome Data Base (Baltimore, MD).

PCR Amplification and Quantitation

Each sample was amplified with 50 ng of pooled DNA in a 25-µl reaction volume containing 1.5 mM $MgCl_2$, 10 mM Tris, 50 mM KCl, 200 µM of each dNTP, 5 pmoles of each primer (one primer 5′-labeled with either FAM, TET, or HEX fluorescent dye), and 0.6 units of *Taq* DNA polymerase (Fisher, Pittsburgh, PA); PCR amplification was performed for 30 cycles in a MJ Research model PTC 100 (Watertown, MA) thermocycler. Amplification of each pooled DNA sample, for a specific marker, was replicated 5–10 times as indicated. PCR products were diluted 1:5 to 1:10 prior to loading onto 4.25% polyacrylamide gels and were run for 2 hr at 3000 V on an ABI model 377 DNA sequencer and analyzed by use of GENESCAN software (Applied Biosystems, Inc., Foster City, CA). Direct genotyping was carried out either on an ABI model 377 sequencer by the method described above with 50 ng of genomic DNA from each individual, or with a ^{32}P-labeled PCR primer and manual genotyping methods as described in Puffenberger et al. (1994a,b).

Estimation of Allele Frequencies

Following electrophoresis, the GENESCAN output was used to estimate peak height for each allele. Allele frequencies were estimated from the relative proportion of the peak height for each allele versus the sum of peak heights for all alleles. Between 5 and 10 replicate PCR reactions were carried out per marker; allele frequencies presented as arithmetic averages were computed from these replicate runs.

Test for Linkage Disequilibrium

For tests of allelic association (linkage disequilibrium), we compared the distribution of alleles on parental chromosomes transmitted versus those untransmitted to probands in parent–offspring trios. This HRR (Falk and Rubinstein 1987) test on direct genotypes in the Mennonite trios was performed as described previously in Puffenberger et al. (1994a). In particular, for assessing allele frequency differences, a 2 × 2 contingency χ^2 test (1 df) was performed. A separate analysis was performed for each allele at every microsatellite marker with a Bonferroni correction to account for multiple tests (on the basis of the number of alleles at each locus).

To estimate allele frequencies on T and U chromosomes from DNA pools, note that all of the alleles in the proband pool represent the T class of alleles whereas all of the alleles in the parent pool represent the same T class of alleles in the offspring as well as an equal number of U class alleles. Thus, if T and U represent the allele frequency distributions of the transmitted and untransmitted alleles among the trios, then the pool of parental alleles is the compound distribution (T + U)/2. Thus, the proband distribution was estimated as the T distribution whereas the U distribution was estimated as twice the parental distribution [i.e., (T + U)/2] minus the proband distribution. The allele frequency differences were tested, under the null hypothesis that the transmitted and untransmitted frequencies were equal, by use of a two sample test for binomial proportions with the unit normal distribution for assessing significance. As with the χ^2 test for the direct genotypes, a Bonferroni correction was used to account for multiple testing.

RH Mapping of Chromosome 13 Markers

The chromosome 13 markers used in the study, in addition to a sequence-tagged site (STS) developed from *EDNRB* (Puffenberger et al. 1994b), were genotyped in a panel of 94 chromosome 13 RHs we had constructed previously, along with a positive control, a monochromosomal 13 hybrid (GM10898), and a negative hamster control (380-6) (methods used as described in Shaw et al. 1995). The genotype data were analyzed by use of multipoint maximum likelihood analysis and the computer program MultiMap version 2.0 (Matise et al. 1994) extended to RH mapping. Distances between markers were estimated in units of centiRays at 9000 rads; in this panel, ~1 cR_{9000} = 70 kb (Shaw et al. 1995).

Tests of Accuracy of Allele Frequencies in DNA Pools

To quantitate the difference between the estimated allele frequencies from DNA pools (denoted y_i for the *i*th allele at any locus) to that obtained by direct allele counts (denoted x_i for the *i*th allele at the same locus), we calculated both the standard product moment correlation (r_{xy} or r) and the RMSE defined as

$$RMSE = \sqrt{\sum_{i=1}^{k} (y_i - x_i)^2 / k}$$

where k is the number of alleles at the marker locus. Although r can be used to judge the linear relationship between y and x, the RMSE is a better descriptor of the similarity of estimated to direct count frequencies. Additionally, the RMSE is an estimate of the average difference in the two frequencies per allele. To study the relationship between y and x values, we also performed standard univariate linear regression analysis using the model $y = ax + b + \epsilon$, where ϵ is random error. We specifically tested the null hypothesis that $a = 1$ and $b = 0$, that is, y and x values differ only by random measurement error. We tested this latter model by fitting a normal distribution to the deviations from true values ($\epsilon' = y - x$) using a χ^2 test.

ACKNOWLEDGMENTS

We acknowledge the many contributions and discussions with members of the Chakravarti laboratory while this work was in progress. We also thank Drs. Myriam Fornage, Norman Arnheim, Joe Nadeau, and Lon Cardon for comments on the manuscript. This research was partially supported by funds from National Institutes of Health (NIH) grants HL 54466 and HD 28088 to A.C. S.H.S. was partly supported by funds from a National Alliance for Research on Schizophrenia and Depression fellowship, and M.M.C. by an NIH training grant (T32 GM 08056) to the Department of Genetics, Case Western Reserve University.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Arnheim, N., C. Strange, and H. Erlich. 1985. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: Studies of *HLA* class II loci. *Proc. Natl. Acad. Sci.* 82: 6970–6974.

Asada, Y., D.S. Varnum, W.N. Frankel, and J.N. Nadeau. 1994. A mutation in the *Ter* gene causing increased susceptibility to testicular teratomas maps to mouse chromosome 18. *Nature Genet.* 6: 363–368.

Buri, P. 1956. Gene frequency in small populations of mutant *Drosophila. Evolution* 10: 367–402.

Carmi, R., T. Rokhlina, A.E. Kwitek-Black, K. Elbedour, D. Nishimura, E.M. Stone, and V.C. Sheffield. 1995. Use of a DNA pooling strategy to identify a human obesity syndrome locus on chromosome 15. *Hum. Mol. Genet.* 4: 9–13.

Chee, M., R. Yang, E. Hubbell, A. Berno, X.C. Huang, D. Stern, J. Winkler, D.J. Lockhart, M.S. Morris, and S.P.A. Fodor. 1996. Accessing genetic information with high-density DNA arrays. *Science* 274: 610–614.

Collin, G.B., Y. Asada, D.S. Varnum, and J.N. Nadeau. 1996. DNA pooling as a quick method for finding candidate linkages in multigenic trait analysis: An example involving susceptibility to germ cell tumors. *Mamm. Genome* 7: 68–70.

Dib, C., S. Faure, C. Fizames, S. Marc, A. Vignal, R. Heilig, M. Lathrop, J. Morrissette, G. Gyapay, and J. Weissenbach. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380: 152–154.

Falk, C.T. and P. Rubinstein. 1987. Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* 51: 227–233.

Knapp, M., S.A. Seuchter, and M.P. Baur. 1993. The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. *Am. J. Hum. Genet.* 52: 1085–1093.

Lander, E.S. and N.J. Schork. 1994. Genetic dissection of complex traits. *Science* 265: 2037–2048.

Matise, T.C., M. Perlin, and A. Chakravarti. 1994. Automated construction of genetic linkage maps using an expert system (MultiMap): A human genome linkage map. *Nature Genet.* 6: 384–390.

Michelmore, R.W., I. Paran, and R.V. Kesseli. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci.* 88: 9828–9832.

Nei, M. 1987. *Molecular evolutionary genetics.* Columbia University Press, New York, NY.

Nystuen, A., P.J. Benke, J. Merren, E. Stone, and V.C. Sheffield. 1996. A cerebellar ataxia locus identified by DNA pooling to search for linkage disequilibrium in an isolated population from the Cayman Islands. *Hum. Mol. Genet.* 5: 525–531.

Pacek, P., A. Sajantila, and A.-C. Syvanen. 1993. Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. *PCR Methods & Applic.* 2: 313– 317.

Puffenberger, E.G., E.R. Kauffman, S. Bolk, T.C. Matise, S.S. Washington, M.A. Angrist, J. Weissenbach, K.L. Garver, M. Mascari, R. Ladda et al. 1994a. Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Hum. Mol. Genet.* 3: 1217–1225.

Puffenberger, E.G., K. Hosoda, S.S. Washington, K. Nakao, D. deWit, M. Yanagisawa, and A. Chakravarti. 1994b. A missense mutation on the endothelin-B receptor gene in multigenic Hirschsprung's disease. *Cell* 79: 1257–1266.

Risch, N. and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science* 273: 1516–1517.

Scott, D.A., R. Carmi, K. Elbedour, S. Yosefsberg, E.M. Stone, and V.C. Sheffield. 1996. An autosomal recessive nonsyndromic-hearing loss locus identified by DNA pooling using two inbred Bedouin kindreds. *Am. J. Hum. Genet.* 59: 385–391.

Shaw, S.H., J.E. Farr, B.A. Thiel, T.C. Matise, J. Weissenbach, A. Chakravarti, and C.W. Richard III. 1995. A radiation hybrid map of 95 STSs spanning human chromosome 13q. *Genomics* 27: 502–510.

Sheffield, V.C., R. Carmi, A. Kwitek-Black, T. Rokhlina, D. Nishimura, G.M. Duyk, K. Elbedour, S.L. Sunden, and E.M. Stone. 1994. Identification of a Bardet-Biedl syndrome locus on chromosome 3 and evaluation of an efficient approach to homozygosity mapping. *Hum. Mol. Genet.* 3: 1331–1335.

Spielman, R.S. and W.J. Ewens. 1996. The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* 59: 983–989.

Spielman, R.S., R.E. McGinnis, and W.J. Ewens. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52: 506–516.

Taylor, B.A. and S.J. Phillips. 1996. Detection of obesity QTLs on mouse chromosomes 1 and 7 by selective DNA pooling. *Genomics* 34: 389–398.

Taylor, B.A., A. Navin, and S.J. Phillips. 1994. PCR-amplification of simple sequence repeat variants from pooled DNA samples for rapidly mapping new mutations of the mouse. *Genomics* 21: 626–632.

Terwilliger, J.D. 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* 56: 777–787.

Received August 13, 1997; accepted in revised form January 7, 1998.

Allele Frequency Distributions in Pooled DNA Samples: Applications to Mapping Complex Disease Genes

Sarah H. Shaw, Minerva M. Carrasquillo, Carl Kashuk, et al.

Genome Res. 1998 8: 111-123
Access the most recent version at doi:10.1101/gr.8.2.111