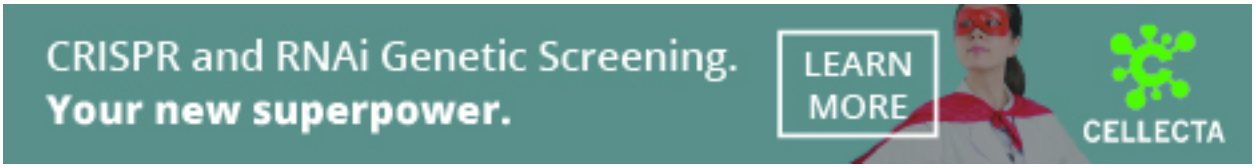
Batched analysis of genotypes.

C LeDuc, P Miller, J Lichter, et al.

Genome Res. 1995 4: 331-336



Batched Analysis of Genotypes

C. LeDuc, P. Miller, J. Lichter, and P. Parry

Sequana Therapeutics, Inc., La Jolla, California 92037

Polymorphic microsatellite markers are widely used in molecular analyses. The range of allele sizes and the allele frequencies within a population are important characteristics of the marker. Their determination previously has involved genotyping a large number of individuals. We have developed a technique for defining these characteristics by coamplification of many samples in a DNA pool. Groups of 32 and 42 DNA samples were genotyped and results were compared with those from individual genotype determinations. To improve the accuracy in the estimation of allele frequencies, arithmetic removal of stutter bands was carried out and the consistency of each marker was characterized. This approach was also applied to a group of 94 individuals. All of the work has been done using nonradioactive methods. Potential applications of this technique are in population genetics, high throughput genotyping, and loss of heterozygosity studies.

Microsatellite dinucleotide markers are proving to be very powerful tools in the identification of human disease genes. Dinucleotide repeats are highly polymorphic,[1,2] having up to 24 alleles, and can be amplified by using primers in the unique flanking sequences and PCR. Following amplification, products from several different loci can be multiplexed prior to electrophoresis. Using fluorescently labeled primers, ABI 373a technology, and GENESCAN software, up to 24 PCR products, labeled with one of three different dyes, can be electrophoresed together in a single lane of an acrylamide gel.[3] The process of applying the amplified microsatellite fragment to an acrylamide gel, followed by the analysis of peaks to determine allele size, is termed genotyping.

Chromosome-specific sets of markers, allowing high density multiplexing, have been reported and used to perform whole genome scans for linkage analysis.[4] Sets of markers were chosen with reference to published reports such as CHLC[5] and CEPH,[6–8] which provide an allele size range (the difference between maximum and minimum allele size, in base pairs) and a number of alleles for a given marker. However, these ranges have been determined on a limited number of individuals,[6,7] and do not reflect the total number of potential alleles for each marker. This is particularly an issue when microsatellites are applied to a larger or more diverse population than that in which the marker was initially characterized. We have developed a method that will allow the determination of allele ranges on many individuals using a small number of PCR reactions. This circumvents the need to generate large numbers of individual genotypes prior to designing multiplex sets of markers. The method is referred to as batched analysis of genotypes (BAGS).

We have investigated the possibility of using the same approach to determine the number of occurrences of each allele in a population of coamplified DNA samples. An artifact typically generated by PCR amplification of microsatellites is shadow bands (when detected by autoradiography) or stutter peaks (when detected by GENESCAN software). These are PCR products smaller than the major amplification product, generally by increments of the repeat unit, and each amplified allele typically shows these stutter bands. Determining the frequency of each allele in a population of coamplified DNA samples is confounded by these artifacts. We developed and applied statistical methods to account for the stutter band. Accurate prediction of the allele frequencies was possible using $\geq$32 coamplified samples.

The combination of the above techniques allowed rapid characterization of a marker in a given group of DNA samples. In addition to enabling the design of efficient multiplex sets for high throughput genotyping, there are applications of this approach in the field of population genetics. The determination of allele frequencies in large populations, among different ethnic groups and interspecifically, allows inference of ancestral relationships, evolution, and systematics.[9] The issue of determining the number of alleles in a mixed population of cells is another potential application. The loss of heterozygosity (LOH) or change in allelic composition in a tumor sample may be detected without the need for extensive purification of the tumor material from surrounding tissue.

MATERIALS AND METHODS

PCR primer pairs were synthesized according to published sequences flanking polymorphic CA repeats.[3] DNA amplification was performed using the

4:331–336©1995 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/95 $5.00

PTC100 thermocycler (MJ Research, San Francisco, CA). The forward primer of each pair was labeled with one of three fluorescent dyes, FAM, HEX, or TET (Applied Biosystems, Foster City, CA), to enable detection. PCR was carried out in a final reaction volume of 20 μl. Amplification occurred during 35 cycles each of 94°C for 30 sec, 30 sec at the primer specific annealing temperature, followed by 30 sec at 72°C. The reaction includes 0.75 unit of *Taq* polymerase (Perkin-Elmer Cetus, Norwalk, CT). PCR product was subjected to electrophoresis on an ABI 373a, using a 12-cm well-to-read denaturing 6% polyacrylamide gel, followed by analysis using Applied Biosystems GENESCAN 1.2 software.

After amplification of a pool of DNA samples, the following method was applied to determine the number of occurrences of each allele in the DNA pool and the allele frequencies. The peak height of each allele was measured and recorded. Each allele in this analysis is preceded by a stutter band. Because the main stutter band is less intense and differs in size by one repeat unit, the stutter peak height was measurable for the smallest allele and for alleles separated from the next smaller by more than one repeat unit. The stutter ratio (r) was calculated as the ratio of stutter peak height to smallest allele peak height.

The peak heights of other alleles were biased by the stutter of adjacent alleles, which were one or more repeat units larger because the DNA of the stutter artifact is aggregated into the signal of the allele. The stutter ratio was applied as a correction factor to derive an unbiased measure of height.

The procedure for stutter band removal can be summarized as (1) the peak of the left-most allele (x) was measured; (2) the peak of the adjacent (immediately preceding) stutter band (y) was measured; (3) the ratio $r=y/x$ was calculated; and (4) peak heights were adjusted by multiplication of the adjacent peak by r and subtraction of a correction.

A first-order approximation, using only a single term from the nearest neighbor, was applied in the initial calculations. The general solution can be found by solving a system of linear equations, by recursion (see Table 1). The effect of the higher order terms is a function of the size of r and the distribution of allele frequencies. The r was measured in a series of five replicates to assess the stability of the stutter peak. Each amplification reaction was prepared separately and the experiment was repeated using two different microsatellite markers.

TABLE 1 General Solution for Stutter Correction

Let

$\hat{a}_1$ = observed peak height at smallest allele
a_1 = actual peak height
$\hat{a}_2$ = observed peak height at second allele
a_2 = actual peak height
$\hat{a}_n$ = observed peak height nth allele
a_n = actual peak height
p = number of alleles
r = stutter ratio

Then,

$$a_1 = \hat{a}_1 - r \times a_2$$
$$= \hat{a}_1 - r \times [\hat{a}_2 - r \times a_3]$$

By recursion,

$$a_n = \hat{a}_n + \sum_{m=n+1}^{p} [(-1^m) \times r^m \times \hat{a}_m]$$

RESULTS

In developing BAGS as a technique for genotyping, we have considered determination of allele range, ability to estimate allele frequency, stutter stability, and other applications. These will be described in turn.

Allele Range Determination

The quality of the PCR product of each marker was determined according to the sharpness and density of the band on an agarose gel. Once the conditions for PCR amplification from a single DNA sample were established, those for coamplification of multiple samples were defined. The DNA concentration in the reaction, the reaction volume, and the amount of PCR product loaded on the gel were varied systematically and, to have an accurate measure of peak height, the amplitude had to fall within the detection range of 50–4000 units on the ABI 373a. To assess the amplification efficiency, PCR reactions were also carried out using a range of 2–42 pooled and coamplified DNA samples, in increments of four. Conditions under which the less-frequent alleles were clearly detectable were established for the experimental markers. Once the amplification conditions were fixed, a series of comparisons between individual and coamplified DNA samples were conducted.

Each DNA sample was genotyped prior to pooling to determine the absolute range and number of alleles within the DNA pool. DNAs from 42 individuals were mixed and coamplified in a single 20-μl reaction, using a total of 20 ng of DNA. To test the amplification consistency, each experiment was replicated with four additional amplifications, with each one using the same DNA pool. The replicate PCR reactions were prepared separately and each amplification was performed in a different PCR machine. The electropherograms were analyzed by a consistent combination of manual and automatic methods using the GENESCAN software.

The results using the highly polymorphic marker D14S80 [GenBank data base accession no., Z1706, heterozygosity (het.) =0.83] are discussed. Forty-two DNA samples were pooled, coamplified by PCR (annealing temperature 56°C), and genotyped. Individual genotyping identified 10 different alleles at this site in this population. Using the BAGS method, nine of these alleles were detected with an allele range of 133.52–155.09 bp (Fig. 1). In this case, the largest allele of 156.89 bp, which occurs only once in the population from a single, heterozygous individual, is disguised by the preceding peak. To account for this in practice, 4 bp is routinely added to the upper end of the observed allele range. It is significant that the smallest allele was also present only once in the sample, from a single, heterozygous individual. This single allele on the low end of the distribution was detected by the BAGS procedure.

Allele Frequency

A statistical analysis was carried out to test the relationship between peak height and the number of occurrences of each allele. Two markers were used to assess the sensitivity of the method, D14S80 and D11S903 (GenBank data base accession no., Z16529, het.=0.74; PCR annealing temperature 56°C). PCR quality was measured, as before, by the presence of a defined, single band on an agarose gel. The two markers were se-

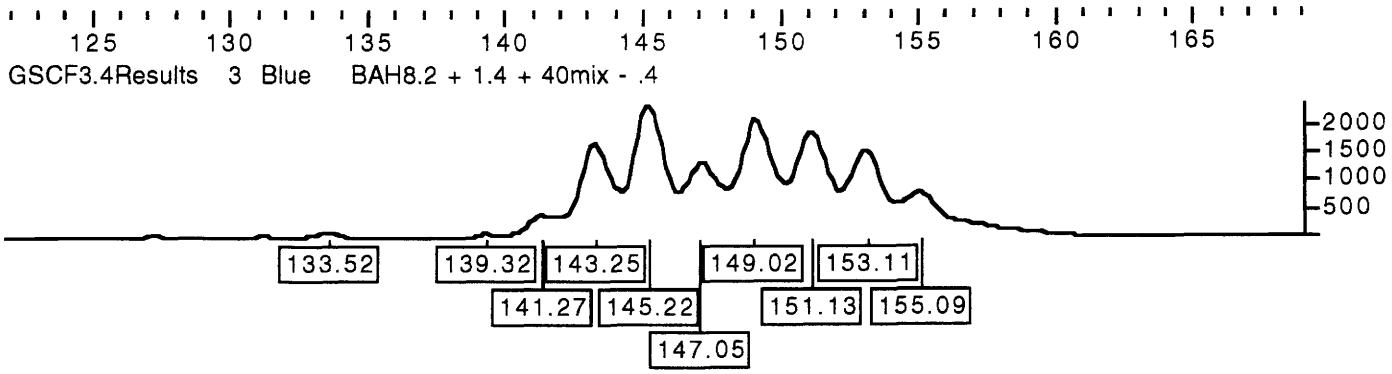


FIGURE 1 As increasing numbers of DNA samples were pooled and coamplified, the amount of PCR product loaded on the gel was varied to maintain detection of the rare alleles. The results using the primer D14S80 are illustrated. By loading 0.8 μl of PCR product the allele sized at 133.52 bp, which occurs once in the population of 42 DNA samples (84 alleles), is detected.

lected to differ in quality, with the first being a high quality primer, and the second, less so. In each case, 32 DNA samples were coamplified. Three pools (3×32 samples) could process the DNA from a 96-well microtiter plate.

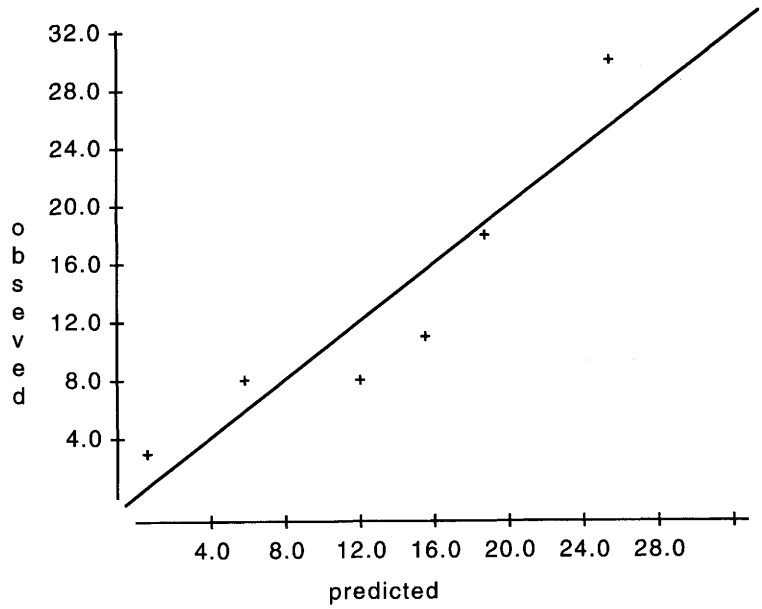
Because the PCR product contains many alleles, often separated by only one repeat unit, removal of the stutter bands was necessary to quantify the number of occurrences of any one allele. This allowed measurement of the frequency of each allele and mitigated against false-positive observations. The total number of individuals in a pool was known, so the adjusted height could be converted into absolute numbers of individuals by solving the simple linear equation, with a single unknown: $\Sigma[a(i)\times X]=Y$, where Y=total number in pool; $a(i)$= adjusted height at allele i; and X=the conversion factor.

The results were studied by regression and analysis of variance (ANOVA). The control for this analysis was that each DNA sample was genotyped individually using each of the markers, so that the absolute number of occurrences of each allele in the population was known. In the regression model, this number was taken as the independent variable, and the dependent variable was the number predicted from adjusted peak heights. The regression provided an estimate of the number of occurrences of an allele and a test of significance. ANOVA was carried out to determine the significance of the variability between replicates.

The regression results are presented in summary form. The regression analysis for one replicate of D11S903 is shown in Figure 2, along with a plot of the number of individuals against peak height. The predicted and observed numbers of each allele are listed in Table 2. The regression was significant at $P<0.001$, with an R^2 of 0.852.

Four replications of the experiment were performed, and the results of the regression analysis for each marker are summarized in Table 3. The R^2 is interpreted as the percent of variance predicted by the independent variable. D11S903 shows greater variability attributable to overloading the gels. In contrast, the pattern of D14S80 was highly stable.

The allele frequency estimates are summarized by plotting the predicted



R squared = 85.2%
sd = 4.16 with 6-2=4 df

Source	SS	df	Mean Square	F-ratio
Regression	398.53	1	398.53	22.9
Residual	69.46	4	17.36	

Variable	Coeff	se	t-ratio	Prob
Constant	-11.42	5.38	-2.13	0.101
predicted ht.	0.012	0.0025	4.79	0.009

FIGURE 2 The regression of the number of alleles predicted by BAGS vs. the number observed by determining each genotype independently.

TABLE 2 Numbers of Individuals, Observed vs. Predicted

Allele	Observed	Predicted
a	30	26
f	8	11
e	8	6
d	3	1
b	18	19
c	11	14

numbers of individuals at each allele. The graph shows the actual number, by individual PCR, in the foreground and the replicates in the adjacent plots. Marker D11S903 is shown in Figure 3a and D14S80 in Figure 3b.

ANOVA was carried out to determine the variation between replicates, controlling for the variables of marker and allele number. Statistical significance was obtained for D11S903 (Table 4) indicating significant variation between replicates; this was the result of gel overloading. The normal operating range of the optical sensor was exceeded in some of the replicates and illustrates the importance of adjusting the quantity of DNA loaded on the gel prior to these analyses. The more robust marker D14S80 did not show a statistically significant difference between replicates (not shown).

Consistency of the Stutter Band Within a Marker

The purpose of this experiment was to clarify the degree of stability of the stutter pattern for any given marker. The following analysis gave a confidence range for the stutter correction and the derived calculation of allele frequency. The procedure was to perform five replicate coamplifications of pooled DNA samples, with two markers, D14S72 (GenBank accession no., Z16835, het. = 0.76; annealing temperature 56°C) and D1S228 (GenBank accession no., Z16878, het. = 0.82; annealing temperature 58°C). One marker was used to amplify two DNA pools of different composition, one of 32 and one of 94 individuals.

TABLE 3 Regression of Actual vs. Pooled Estimates, Five Replicates

Marker	R^2	Significance
D11S903	0.83–0.98	$P < 0.01$–0.0001
D14S80	0.89–0.91	$P < 0.001$

The stutter band is relatively stable within the replicates (Table 5), with a standard deviation of similar magnitude to that of the predicted allele frequency estimates (Fig. 2). The total amount of stutter ratio is dependent on the height of the adjacent allele. However, the number of coamplified samples was associated with an alteration of the stutter ratio, in the case of D14S72. Therefore, measurement of the stutter ratio, for each DNA pool of different complexity, is suggested.

Applications

The BAGS method was applied to three subsamples from a population survey made up of Caucasians of more than six different European ethnic groups. Figure 4 shows the results of amplification of three human DNA pools, each comprising 32 individuals, as amplified by D1S243 primers (GenBank accession no., Z16979; annealing temperature 60°C). The allele range was determined to be 135–175 bp. It is interesting that there are differences in allele frequencies between pools one and three, although no selection criteria were used in composing the groups.

TABLE 4 ANOVA of Marker D11S903

Source	df	Sum of squares	Mean square	F ratio	P
Replicates	1	1014	1014	12.7	0.02
Error	5	399.1	79.8		
Total	5	399.1			

Determining marker heterozygosity is especially important in developing new markers for use in a genotyping project.[1] The Hardy–Weinberg distribution allows prediction of heterozygosity using allele frequency, provided that the alleles are in equilibrium. Previously this determination was made by individually genotyping 94 unrelated individuals and a positive and negative control in a 96-well microtiter plate. The BAGS method has been used to examine these individuals in 3, rather than 96, reactions and allows rapid determination of those markers that are sufficiently heterozygous to be useful for population-wide genotyping.

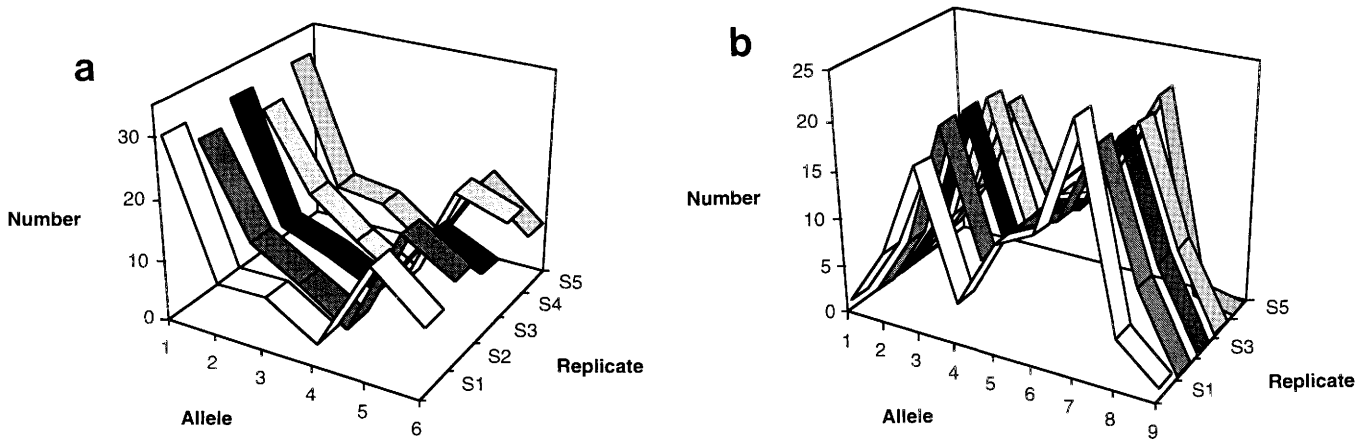


FIGURE 3 Comparison of the predicted number of each allele amplified by D11S903 (*a*) and D14S80 (*b*). In each case, four replicate amplifications and analyses were performed and are compared to the absolute number of alleles in the study population. The replicates are shaded and the control is shown by the white ribbon.

TABLE 5 Stutter Band Stability

Marker	Number of DNA sampled pooled	Stutter ratio	Mean and standard deviation
D14S72	32	0.256	0.278 (0.029)
		0.317	
		0.248	
		0.297	
		0.274	
	94	0.193	0.199 (0.015)
		0.199	
		0.180	
		0.203	
		0.222	
D1S228	32	0.375	0.391 (0.019)
		0.402	
		0.366	
		0.399	
		0.412	

DISCUSSION

This paper describes a method for batched analysis of genotypes. The method allows the determination of allele ranges for microsatellite markers by coamplification of a complex mixture of DNA samples. DNAs from 94 individuals have been coamplified. The number of alleles and the range of allele sizes have been determined successfully. In addition, we have been able to estimate the frequency of each allele in the population. The level of precision obtained was consistent with the characteristics of quantitative PCR.[10] Similar experiments have been reported previously; for example, Pacek et al.[11] showed the results of the coamplification of 1000 DNA samples with a tetranucleotide repeat and a variable number tandem repeat (VNTR) marker. They were able to make a prediction of the number of occurrences of each allele in their population. However, tetranucleotide repeat and VNTR markers do not generate complex stutter bands during amplification and so are simpler systems than the more frequent CA repeats. Darvasi et al.[12] statistically characterized the pooling method using dinucleotide repeat polymorphisms but were limited by the stutter artifact. With stutter band removal, we have contributed to making the method more useful with common dinucleotide repeats.

Population and evolutionary genetics may particularly benefit from the ability to determine allele ranges and frequencies in species, and populations within species, which are as yet uncharacterized in this respect. Recently the utility of microsatellite alleles in the construction of human population trees[13] has been shown. Genotyping individuals in a batch will limit the types of analyses possible; for example, the pairwise comparisons described by Bowcock et al.[13] are not possible. However, there is tremendous power to construct evolutionary trees based on the allele frequencies of several marker systems in a large number of individuals from different populations.

This approach can also be used in species other than human. We have successfully used human primer pairs to amplify baboon DNA. Following optimization of the amplification conditions, primers were used to amplify three subsamples, each of 32 unrelated animals. On the basis of the BAGS results, we have identified microsatellite markers that have high heterozygosities. So, by performing 3 PCR reactions as opposed to 96 we are able to include or exclude markers from a baboon genome wide polymorphic marker set.

Recently, Sheffield et al.[14] used a similar pooling approach in homozygos-

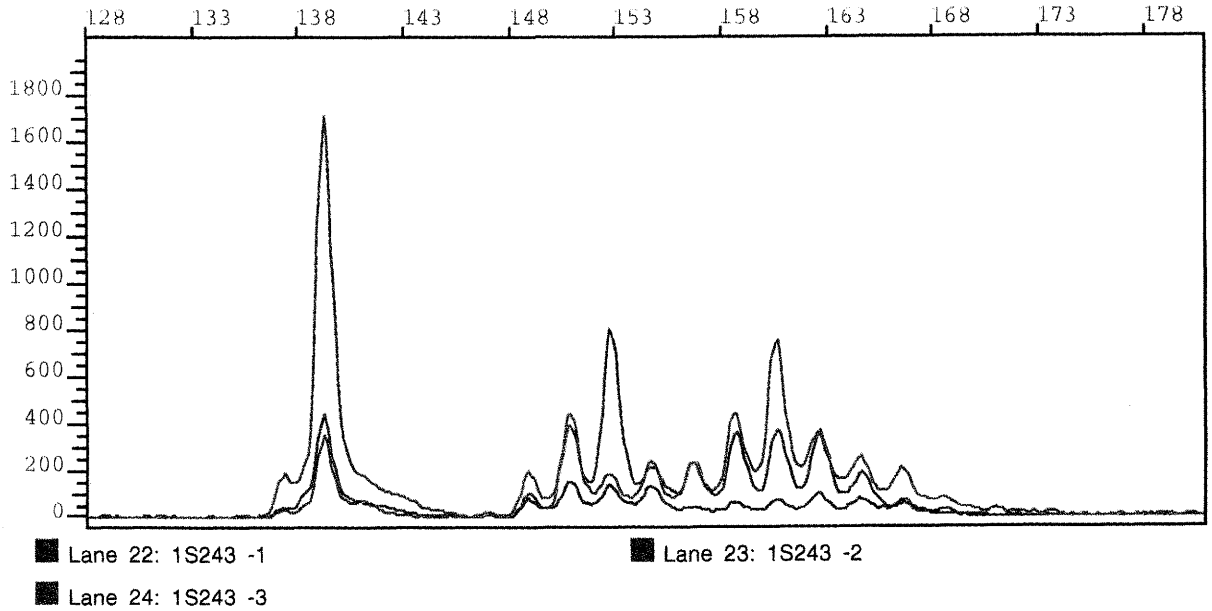


FIGURE 4 Three human DNA pools, each comprising 32 individuals, coamplified with primers at D1S243. The allele range is determined as 135–175 bp. There are apparent differences in the frequencies of the alleles between pools 1 (*blue*) and 3 (*red*).

ity mapping of a rare autosomal recessive trait. Although it incorporated a nonfluorescent procedure and aggregate of only 12 individuals from one family, the utility of pooling for linkage analysis requiring allele frequency (and allele sharing) data was demonstrated. This method could also be used for the detection of allele associations in groups of individuals with common diseases. Comparing the allele frequencies in a population of individuals with disease to the frequencies in an ethnically matched group of individuals without disease allows for the calculation of allelic association. With common diseases, the number of individuals required to detect associations can be quite large.[15–17] This technique will allow the rapid analysis of large numbers of individuals.

Tumor development is a multistep process that depends on the accumulation of mutations in multiple genes. If a tumor is considered to be a population of cells, some of which differ in their genetic complement from each other and from the normal surrounding tissue, then the BAGS approach may be a simple and rapid means of looking for differences in the genetic status at a given locus. LOH is a common genetic aberration observed in tumor samples. The laboratory techniques employed most commonly in LOH studies depend on the separation of the tumor from the surrounding tissue to allow extensive enrichment of tumor cells.[18] This particularly arduous task may be avoided by using the analyses that have been described above on the mixed population of tumor and normal cells. Hereditary nonpolyposis coli is associated with a high frequency of variation in the length of microsatellite repeats,[19] which may be recognized using BAGS to detect a change in relative allele sizes, circumventing the need for tumor purification.

BAGS is very efficient in terms of cost of materials, equipment use, and time. Range determination, assessment of amplification quality, and numbers of alleles are accomplished in a small number of reactions. Allele frequency determination shows promising results and may have significant applications.

ACKNOWLEDGMENTS

We thank John Todd for his initial stimulating discussions, Pirjo Vaisanen for DNA manipulation, Matteo McCarthy for oligonucleotide synthesis, Alan Roter for database maintenance, and Jeff Rogers for discussions on baboons. This research was supported by Sequana Therapeutics, Inc.

REFERENCES

1. Weber, J.L. 1990. Human polymorphism based on length variations in simple-sequence tandem repeats. In *Genome analysis*, Vol 1, pp. 159–181. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
2. Weber, J.L. and C. Wong. 1993. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2:** 1123–1128.
3. Reed, P.W., J.L. Davies, J.B. Copeman, S.T. Bennett, S.M. Palmer, L.E. Pritchard, S. Gough, Y. Kawaguchi, H. Cordell, K. Balfour, S.C. Jenkins, E.E. Powell, A. Vignal, and J.A. Todd. 1994. Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. *Nature Genet.* **7:** 390–395.
4. Davies, J., Y. Kawaguchi, S.T. Bennett, J.B. Copeman, H.J. Cordell, L.E. Pritchard, P.W. Reed, S.C.L. Gough, S.M. Palmer, K.M. Balfour, S.C. Jenkins, B. Rowe, M. Farrall, A.H. Barnett, S.C. Bain, and J.A. Todd. 1994. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371:** 130–136.
5. Murray, J., K.H. Buetow, J.L. Weber, S. Ludwigsen, T. Scherpbier-Heddema, F. Manion, J. Quillen, V.C. Sheffield, S. Sunden, G. Duyl, J. Weissenbach, G. Gyapay, C. Dib, J. Morrissette, G.M. Lathrop, A. Vignal, R. White, N. Matsuname, S. Gerken, R. Melis, H. Albertsen, R. Plaetke, S. Odelberg, D. Ward, J. Dausset, D. Cohen, and H. Cann. 1994. A comprehensive human linkage map with centimorgan density. *Science* **265:** 2049–2055.
6. Weissenbach, J., G. Gyapay, C. Dib, A. Vignal, J. Morissette, P. Millasseau, G. Vayssiex, and M. Lathrop. 1992. A second-generation linkage map of the human genome. *Nature* **359:** 67–86.
7. Gyapay, G., J. Morissette, A. Vignal, C. Dib, C. Fizames, P. Millasseau, S. Marc, G. Bernardi, M. Lathrop, and J. Weissenbach. 1994. The 1993-94 Genethon human genetic linkage map. *Nature Genet.* **7:** 246–339.
8. 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.
9. Cavalli-Sforza, L., P. Menozzi, and A. Piazza. 1994. *History and geography of the human genes*. Princeton University Press, Princeton, NJ.
10. Ferre, F. 1994. Quantitative PCR: An Overview. In *The polymerase chain reaction* (ed. K. Mullis, F. Ferre, and R. Gibbs), pp. 69–88. Birhauser Press, Boston, MA.
11. Pacek, P., A. Samantila, and A. 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.
12. Darvasi, K.A., Y. Plotski, and M. Soller. 1994. Determining relative microsatellite allele frequencies in pooled DNA samples. *PCR Methods Applic.* **4:** 13–18.
13. Bowcock, A., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368:** 455–457.
14. Sheffield, V.C., R. Carmi, A. Kwitek-Black, T. Rokhlina, D. Nishimura, G. Dyuk, K. Elbedour, S. 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.
15. Cox, N.J. and G.I. Bell. 1989. Perspectives in diabetes, disease associations, chance, artifact, or susceptibility genes? *Diabetes* **38:** 947–950.
16. Olson, J.M. and E.M. Wijsman. 1994. Design and sample-size considerations in the detection of linkage disequilibrium with a disease locus. *Am. J. Hum. Genet.* **55:** 574–580.
17. Schaid, D.J. and S.S. Sommer. 1994. Comparison of statistics for candidate-gene association studies using cases and parents. *Am. J. Hum. Genet.* **55:** 402–409.
18. Golez, S., S. Hamilton, and B. Vogelstein. 1985. Purification of DNA from formaldehyde fixed and paraffin embedded human tissue. *Biochem. Biophys. Res. Commun.* **130:** 118–126.
19. Aaltonen, L.A., P. Peltomaki, F.S. Leach, P. Sistonen, L. Pylkkanen, J.-P. Mecklin, H. Jarvinen, S.M. Powell, J. Jen, S.R. Hamilton, G.M. Petersen, K.W. Kinzler, B. Vogelstein, and A. de la Chapelle. 1993. Clues to the pathogenesis of familial colorectal cancer. *Science* **260:** 812–816.

Received February 28, 1995; accepted in revised form April 21, 1995.