European Journal of Human Genetics (2004) 12, 925–934
© 2004 Nature Publishing Group All rights reserved 1018-4813/04 $30.00
www.nature.com/ejhg

ARTICLE

Accurate determination of microsatellite allele frequencies in pooled DNA samples

Hugo G Schnack*[,1,5], Steven C Bakker[2,5], Ruben van 't Slot[3], Bart M Groot[4], Richard J Sinke[2], Rene S Kahn[1] and Peter L Pearson[2]

[1]*Department of Psychiatry, University Medical Center Utrecht, The Netherlands;* [2]*Department of Medical Genetics, University Medical Center Utrecht, The Netherlands;* [3]*Department of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands;* [4]*Institute of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands*

Pooling of DNA samples instead of individual genotyping can speed up genetic association studies. However, for microsatellite markers, the electrophoretic pattern of DNA pools can be complex, and procedures for deriving allele frequencies are often confounded by PCR-induced stutter artefacts. We have developed a mathematical procedure to remove stutter noise and accurately determine allele frequencies in pools. A stutter correction model can be reliably derived from one standard 'training set' of the same 10 individual DNA samples for each marker, which can also include heterozygous patterns with partially overlapping peaks. Compared with earlier methods, this reduces the number of genotypes needed in the training set considerably, and allows standardization of analyses for different markers. Moreover, the use of a procedure that fits all data simultaneously makes the method less sensitive to aberrant data. The model was tested with 34 markers, 18 of which were newly defined from human sequence data. Allele frequencies derived from stutter-corrected DNA pool patterns were compared with the summed individual genotyping results of all the individuals in the pools ($n = 109$ and $n = 64$). We show that the model is robust and accurately extracts allele frequencies from pooled DNA samples for 32 of the 34 microsatellite markers tested. Finally, we performed a case–control study in celiac disease and found that weakly associated disease alleles, identified by individual genotyping, were only detectable in pools after stutter correction. This efficient method for correcting stutter artefacts in microsatellite markers enables large-scale genetic association studies using DNA pools to be performed.

European Journal of Human Genetics (2004) **12,** 925–934. doi:10.1038/sj.ejhg.5201234
Published online 11 August 2004

Keywords: microsatellite marker pool stutter PCR

Introduction

It has been cogently argued that population-based genetic association studies will have a greater power than linkage studies to localize genes contributing moderately or only a little to the phenotype of complex diseases.[1] However, the detection of association, or linkage disequilibrium between a genetic marker and a disease locus in outbred populations is only possible over small genetic distances.[2–6] For screening large genomic regions, or even comprehensive whole-genome association studies, this implies that hundreds or thousands of markers have to be genotyped for each subject. Such studies are barely feasible using currently available genotyping technology.

Pooling of DNA samples for genetic marker analysis is a method to reduce the amount of genotyping required in allelic association studies.[7–16] This technique involves combining equal amounts of DNA from patients and controls into separate pools, and comparing the pools for

*Correspondence: Dr Hugo G Schnack, Department of Psychiatry, A01.126, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands. Tel.: +31 30 2507130;
Fax: +31-30-2505443; E-mail: hschnack@azu.nl

[5]Both authors contributed equally to this work

Received 1 October 2003; revised 29 April 2004; accepted 5 May 2004

differences in allele distributions of genetic markers. In the absence of haplotype information, which is the situation encountered in a typical association study based on pooled case–control comparisons, the biallelic variation of single-nucleotide polymorphisms (SNPs) contains far less polymorphic information than microsatellite markers. Therefore, microsatellites provide a more powerful tool on a marker-by-marker basis than SNPs.[17,18] However, in the case of microsatellite markers, the overall genotype patterns of pooled samples are often distorted by PCR artefacts such as stutter and preferential amplification, which prevent an accurate determination of the allele frequencies by simple procedures. Several methods have been proposed to handle these artefacts. Some studies compared summed differences in patterns between two pools without correction for PCR artefacts, and without allotting the individual allelic contributions to the differences.[8–11]

A fundamentally different way to compare pool patterns is to correct the pool signal for predicted PCR artefacts, in order to derive more accurate estimates of the allele frequencies. Advantages of this approach are that it allows the comparison of frequencies for individual marker alleles, and that results from different experiments can be summated and analyzed using regular statistics such as χ^2 tests, since the entire pool signal is deconvoluted into individual allele counts.[12] All recent correction methods use information derived from a training set of individual genotype patterns to obtain information about the stutter behavior of the marker under investigation. One approach is to build a matrix of stutter patterns for individual alleles.[7,14,15] This requires a set of well-distributed homozygous or well-separated (nonoverlapping) heterozygous individual genotype patterns, and interpolation or extrapolation has to be invoked to complete the matrix for missing alleles. These methods are sensitive to one or more nonrepresentative patterns caused by, for example, measurement errors.

Alternatively, a stutter model can be derived from individual genotypes, which is used to correct for stutter and permits interpolation of stutter for allele sizes not encountered in the training set.[12] The advantage of a model is that it partly removes the influence of aberrant patterns. On the other hand, it interprets the stutter peaks according to a fixed behavior, which can yield a less accurate description. The model approaches presented thus far also require well-distributed homozygous or well-separated heterozygous individual patterns for each marker to define the model parameters. In both types of correction procedure, a rather large set (at least 20 to 50) of individual patterns has been considered necessary[7,8] to provide sufficient data to obtain the necessary stutter information.[12] The search for and analysis of informative marker data often make these approaches tedious and highly interactive.

We have developed a stutter correction method that fits a model to one small set of genotype data from 10 individuals. This standard training set is identical for all markers, and can be of any allelic composition, since it does not need to include particularly defined homozygous or heterozygous individuals. The accuracy of the stutter correction model has been tested on 34 different microsatellite markers and used in a case–control study for celiac disease.

Materials and methods

Definitions

Uncorrected pool: allele frequencies derived from pool signals uncorrected for stutter.

Corrected pool: allele frequencies derived from pool signals corrected for stutter.

True pool: allele frequencies obtained by individual genotyping of all samples present in a pool, and summing the allele counts.

Preparation of DNA pools, marker selection, PCR, and analysis

Genomic DNA was obtained from peripheral blood lymphocytes using established procedures. Stock solutions were diluted to approximately 25 ng/μl, vortexed gently, and measured with Pico Green (Molecular Probes, Leiden, the Netherlands) on a Genios plate reader (Tecan, Männedorf). Subsequently, samples were diluted to 10 ng/μl and final concentrations were measured in triplicate. Each sample was tested for adequate PCR amplification. Volumes containing 100 ng of DNA from individual samples were pooled. Pools, as well as a set of 10 random individual samples, were purified by phenol extraction, and diluted with water to 10 ng/μl. Characterized microsatellite markers were obtained from the Genome Database (GDB) and Marshfield database. New microsatellite markers were identified by searching a 4 Mb ADHD linkage region on chromosome 15 for microsatellite repeats[19] using the Tandem Repeat Finder program (TRF). PCR primers flanking the repeats were designed with the Primer3 program (sequences are available on request). A so-called pig-tail sequence extension was added to one of the primers in order to reduce plus-A artefact during PCR.[20] The other primer was labeled with 6-FAM, HEX, or NED fluorescent dyes (Biolegio, Malden, the Netherlands, and Applied Biosystems, Foster City, CA, USA).

Individual samples and triplicate pools were amplified simultaneously as described elsewhere,[19] but with 27 instead of 33 cycles. Up to three products were pooled, and analyzed on an ABI 3700 sequencer.[19] Sample files were analyzed using Genescan 3.5 and Genotyper 3.6 for Windows NT and the heights of all peaks were labeled.

Samples with allelic peak heights below 200 or above 6000 were not labeled. A computer program called PoolFitter (freely available from our web site), which is a user interface invoking our stutter correction algorithm, then processed the tables with allele sizes and peak heights (see below). The pool patterns were corrected for stutter by applying the model parameters derived from the individual genotypes (see below). For marker D7S2422 only, preferential amplification of shorter alleles was compensated in the PoolFitter program, by dividing the peak heights of both individual data and pooled data before model fitting by a function fitted to the corrected heterozygous patterns without compensation for preferential amplification. Estimates from corrected and uncorrected pool patterns (averages of triplicate measurements) were compared with true pools using the program CLUMP.[21]

The model

The basic concept is that, for pooled DNA, any electrophoretic microsatellite marker pattern (See Figure 1a) is the sum of its constituent parts comprising a mixture of homozygous and heterozygous individual patterns. Peaks may represent individual alleles, or individual alleles plus a stutter component, or only stutter. We describe a pattern by $Y(a)$, where Y is the height of the signal at fragment length a. In all figures, the signal height has been scaled to facilitate comparison with calculated quantities later on. The length a can assume discrete values differing by multiples of the repeat length Δa. For a dinucleotide marker, $\Delta a = 2$ base pairs. Looking at a pattern for a single allelic peak at length a_0, one expects stutter peaks at $a_0 - \Delta a$, $a_0 - 2\Delta a$, etc. and possibly 'up-stutter' peaks at $a_0 + \Delta a$, etc. In general, peaks are located at $a = a_0 - m\Delta a$, with m integer. The modeled peaks are described by $y(a)$, representing the peak height at length a. This peak $y(a)$ can have contributions from an allelic peak $y_0(a)$, and from stutter peaks of alleles located m base pairs away: $y(a) = y_m(a + m\Delta a)$. Thus, the index m refers to the order of the (stutter) peak: $y_0(.)$ is the main, allelic peak, $y_1(.)$ is the first stutter peak, and so on. The argument of $y_m(.)$ refers to the location of the main, allelic peak. Our aim is to describe the total set of peaks by a model with as few parameters as possible. The values of the parameters will depend on the marker, PCR conditions, settings of the electrophoresis apparatus, etc. Knowing the stutter parameters, it is possible to deconvolve the measured signal $Y(a)$ of a DNA pool into the contributions of individual alleles and hence calculate the frequency of each allele in the pool. The ratio between allele and stutter signal appears to be marker specific, and since pool patterns do not contain enough information to obtain reliable estimates for the stutter parameters used in the model, these

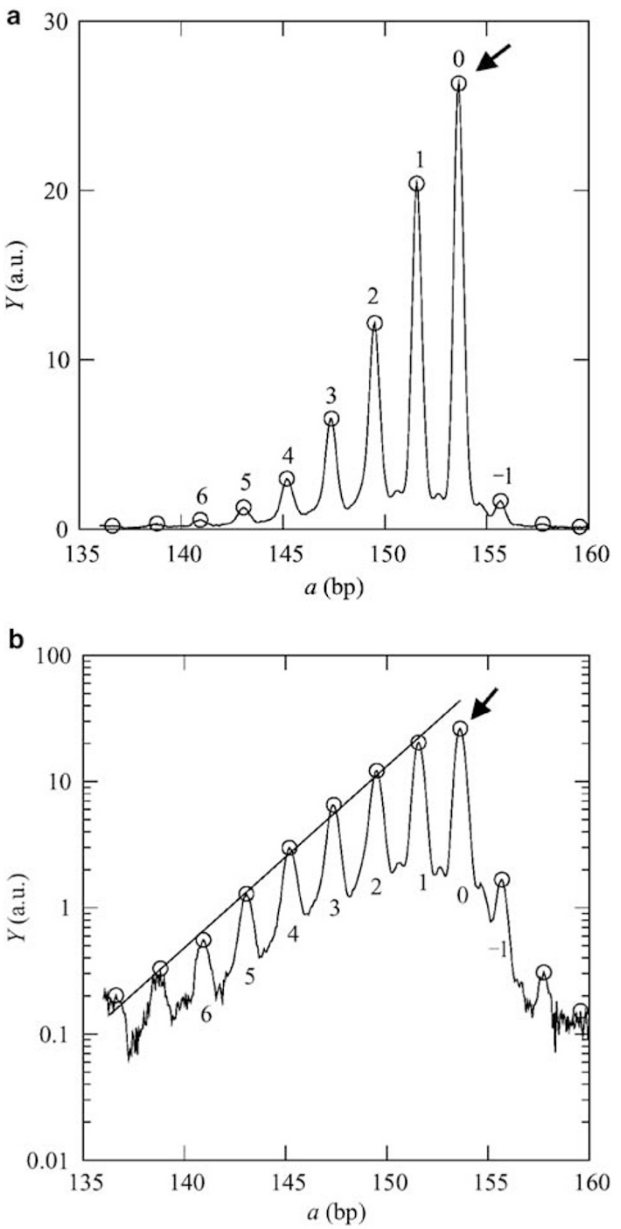

Figure 1 (**a**) Typical individual electrophoretic pattern after removal of background signal (line). The marker is the dinucleotide repeat marker DRD5. The individual is homozygous for the allele of size $a = 153$ base pairs (indicated by the arrow). The circles indicate the tops of the peaks, which are used as measures of the amount of stutter and allele signal present. The numbers refer to the indexing of the peaks: 0 is the allelic peak, 1 the first stutter peak, and so on. Since the absolute heights of the peaks are not important, Y has been scaled to a 100% scale in most figures. (**b**) The same electrophoretic pattern as in Figure 1a, plotted on a logarithmic scale (line and circles; the numbers refer to the indexing of the peaks). The straight line indicates the exponential relationship between successive stutter peaks. The arrow indicates the main, allelic, peak at $a = 153$.

parameters have to be derived by fitting the model to a number of individual test patterns for each marker.

Stutter pattern of a homozygous individual

It can be demonstrated that the heights of stutter peaks decay exponentially with the number of stutters, as clearly shown in a logarithmic plot (Figure 1b), in which a straight line can be drawn through the tops of the stutter peaks. A few simple theoretical assumptions about the nature of DNA amplification predict this exponential behavior.[22] From many of such plots we found empirically that the ratios of the heights of successive stutter peaks are roughly constant for all samples of the same marker and amplification condition, but that the constant differs between markers and conditions. We denote this constant by the ratio *r*. The first stutter peak is usually found to be proportionally higher compared to the main peak; in Figure 1b this is observed as a deviation of the stutter straight line with the top of the allelic peak. We therefore use a different ratio to describe the relationship between the first stutter peak and the allele peak:

$$y_m(a)/y_{m-1}(a) = r, \quad m = 2, 3\ldots \tag{1a}$$

$$y_1(a)/y_0(a) = \lambda r, \quad m = 1 \tag{1b}$$

with $0<r<1$, and $\lambda>1$, normally. This is for the 'normal' downward stutter. For the upward stutter, we take only one peak into account, as it is rare to see more up-stutter peaks; however, the model can easily be extended to more, if necessary.

$$y_{-1}(a)/y_0(a) = \mu, \quad m = -1 \tag{1c}$$

with $0<\mu\ll 1$, normally.

For all other positions, that is, positions at larger lengths:

$$y_m(a) = 0, \quad m = -2, -3, \ldots \tag{1d}$$

We can combine and rewrite Equations (1a)–(1d) as follows:

$$y_m(a) = \begin{cases} y_0(a), & m = 0\text{: main peak} \\ y_0(a)\lambda r^m, & m = 1, 2, 3\ldots\text{: stutter peaks} \\ y_0(a)\mu, & m = -1\text{: up}-\text{stutter} \\ 0, & m = -2, -3, \ldots \end{cases} \tag{2}$$

The fragment length of (stutter) peak $y_m(a)$ is

$$a_m = a - m\Delta a. \tag{3}$$

It is usually observed that stutter is more severe for longer alleles than shorter alleles. This can be understood, at least qualitatively, by realizing that a larger number of repeats offers more chances for the PCR process to stutter. We therefore introduce an *a* dependence of the stutter ratio *r*:

$$r = \exp(b_0 + b_1 a). \tag{4}$$

For positive values of b_1, this formula yields an increasing stutter for increasing *a*.

The true amount of signal at the allelic peak, that is, if no allele signal had been dissipated into stutter peaks, is represented by

$$y_t(a) = \sum_{m=-1}^{\infty} y_m(a). \tag{5}$$

We have now described the set of stutter peaks by four parameters: b_0, b_1, λ, μ. In the trivial case of a pattern of a homozygous individual, Equation (2) can be fitted directly to the $Y(a)$ data, with $y_0(a_0)$ as a fifth fit parameter. The length of the allele is directly read from the pattern: a_0; $y_0(a_0)$ is just the height of the measured main peak $Y_0(a_0)$; μ is the ratio of the up-stutter peak at $a_0+\Delta a$ and the main peak. The factor *r* is determined by the logarithm of the heights of the stutter peaks $y_1(a_0)$, $y_2,(a_0)\ldots$ the heights of which are directly taken from the measured peaks $Y(a_0-\Delta a)$, $Y(a_0-2\Delta a)$,… Then λ follows from $y_1(a_0)/y_0(a_0)$, with the value of *r* inserted in Equation (2). In this example, *r* is kept constant; in a more realistic situation involving alleles of several lengths (such as in a pooled DNA sample), b_0 and b_1 can be fitted instead of a constant *r*.

Stutter pattern of a heterozygous individual

In the case of a heterozygous individual, there are more measured peaks to fit, and there is one extra fit parameter, namely the y_0 of the second allelic peak. We will refer to the two y_0s as 0S and 0L, located at a_{0S} and a_{0L}, respectively, with $a_{0S}< a_{0L}$ (see inset of Figure 2). Heterozygous patterns often overlap to a large extent, and pool patterns always

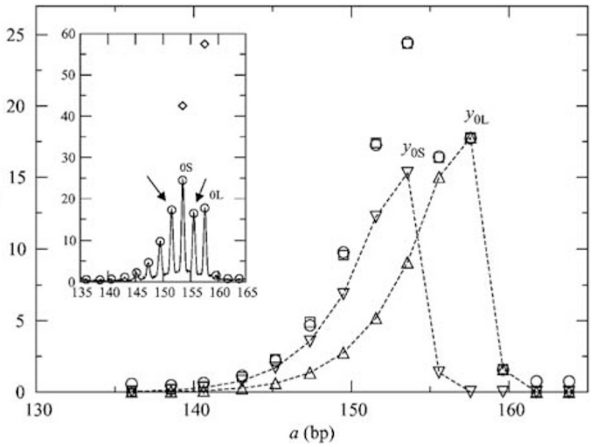

Figure 2 Model fitted to a heterozygous individual electrophoretic pattern (see inset for the original pattern (line) with peaks (circles); allelic peaks 0S and 0L are indicated). The marker is DRD5. The circles represent the data peaks; the squares show the fit. The two types of triangles and the dashed lines indicate the individual peak patterns of the two alleles (y_{0S} and y_{0L}) that comprise the measured signal. The diamonds in the inset represent the corrected, that is, estimated frequencies of the two alleles. These values are obtained by summing all peaks for each separate allele and normalizing the total sum of the two alleles to 100.

do. For two alleles close together, the measured peak heights in the overlapping region are the sum of two contributing peaks, one for the (shorter) S-allele, and one for the (longer) L-allele. For instance, the peak at the left arrow in the inset of Figure 2 is made up of the first stutter peak of the S-allele and the third stutter peak of the L-allele and is represented by

$$\begin{aligned} y(a=151) &= y_1(a_{0S}) + y_3(a_{0L}) \\ &= y_0(a_{0S})\lambda r + y_0(a_{0L})\lambda r^3, \end{aligned} \tag{6}$$

with $a_{0S} = 153$ and $a_{0L} = 157$. This effect makes the fit procedure more challenging and real solution algorithms have to be invoked. We used the Levenberg–Marquardt method.[23] The result of such a fit is shown in Figure 2. The model fits the data well, and the relatively large contribution of the stutter peaks of the L-allele to both the allelic peak and stutter peaks of the S-allele is clearly seen.

Pattern of a pooled sample

The generalization to fitting a pattern of a pooled DNA sample containing alleles of n individuals is straightforward. At every measured fragment length a, the following peaks can contribute to $y(a)$, depending on the presence of alleles in the pooled sample:

(a) the allelic peak $y_0(a)$ of the allele at $a_0 = a$;
(b) the up-stutter peak $y_{-1}(a - \Delta a)$ of the allele just left of it, at $a_0 = a - \Delta a$;
(c) the first-order stutter peak $y_1(a + \Delta a)$ of the allele just right of it, at $a_0 = a + \Delta a$;
(d) higher-order stutter peaks $y_m(a + m\Delta a)$ of alleles more to the right ($m = 2,3,\ldots$).

In a formula, this can be written as

$$y(a) = \sum_{m=-1}^{\infty} y_m(a + m\Delta a), \tag{7}$$

with $y_m(a + m\Delta a) = y_0(a + m\Delta a)\lambda \exp((b_0 + b_1 a)m)$ for $m \geq 1$ the mth stutter peak of the allele at $a_0 = a + m\Delta a$, and $y_{-1}(a - \Delta a) = y_0(a - \Delta a)\mu$ the up-stutter peak of the allelic peak just left, at $a_0 = a - m\Delta a$. The arguments of $y(.)$ in Equation (7) must lie in the measured range of allele lengths $(a_{\min}, a_{\max})$.

The $y_t(a)$s are now calculated as follows:

$$y_t(a) = \sum_{m=-1}^{\infty} y_m(a). \tag{8}$$

These values are proportional to the number of individuals n contributing to that allele a. To obtain estimates of the true allelic frequencies $F(a)$ in the pool, one calculates:

$$F(a) = 2ny_t(a) / \sum_{a'} y_t(a'), \tag{9}$$

where the summation is carried out over the full range $(a_{\min}, a_{\max})$ of the pool pattern.

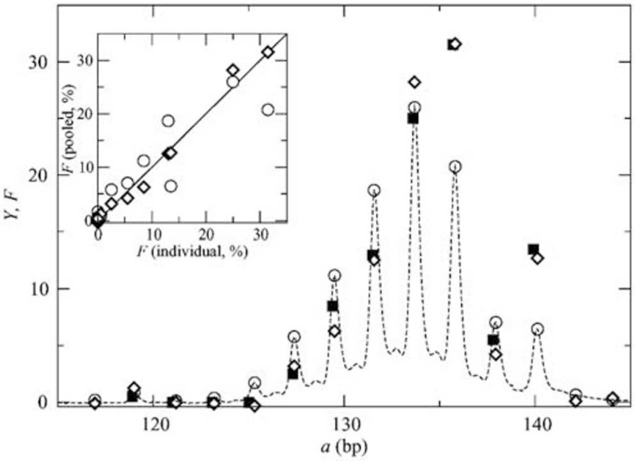

Figure 3 Electrophoretic pattern of a pool of 100 individuals for marker D6S273 (dashed line). The circles mark the peaks used in the analysis. Diamonds indicate the estimated true allelic frequency F after compensation for stutter, as calculated from the model. The model parameters were derived from 10 randomly chosen individual marker patterns. The fit parameters (b_0, b_1, λ, μ) were inserted into the pool fit, from which the y_ts were derived. For most alleles, quite a big difference is observed between the original and corrected peak heights. The squares represent the summed individual genotypes. The inset shows the relationship between summed individual frequencies and uncorrected (circles) and corrected (diamonds) frequencies from the pool. The straight line represents the identity line: symbols on this line represent alleles for which the frequencies estimated from the pool equal the summed individual frequencies, showing perfect agreement.

To correct a pattern of a pooled DNA sample, one has to fit Equation (7) to the measured data $Y(a)$. Values for the four model parameters b_0, b_1, λ, μ could be found from fitting the model to the genotype patterns of a small number of representative individuals one at a time, and deriving n_i values for each of the fit parameters. These n_i values could then be averaged to obtain a good estimate for each of the parameters. A much more efficient way is to perform the model fitting to all individual patterns simultaneously. The total number of data points is $n_i m_i$, with m_i the average number of measured peaks per individual. The total number of fit parameters is $4 + n_i(1+h)$, with h the calculated heterozygote frequency of the marker ($0 \leq h \leq 1$). A set of $n_i = 10$ individuals, each with on average $m_i = 7$ data points and a heterozygote frequency of $h = 0.5$, requires fitting a model with 19 parameters to a combined data set of 70 data points, which, as shown in Figure 3 for marker D6S273, yields a very stable fit.

Comparison with a deconvolution method

The most robust method previously published is the deconvolution method described by Perlin *et al.*[15] Like our method, it uses a set of individual patterns to obtain the stutter behavior. The main difference between our method and Perlin *et al*'s is the fact that we fit a model to the data to describe the stutter behavior, which makes our method potentially much more robust, and thus requiring fewer individual patterns to train the method. This has been tested below.

Results

Minimum number of genotypes required in the training set

For marker D6S273, we investigated the influence of training set size on the reproducibility of the results by fitting models based on sets varying in size from 2 to 30 individuals, which were taken at random from the n individuals in the pool. For each chosen set size, a random selection of individuals was taken 20 times to derive the model parameters and to correct the pool data. Figure 4 shows the effect of training set size plotted against the spread in the corrected peak height of one of the alleles ($a=127$; see Figure 3) in the pool. We chose to show this allele because of its low frequency (3%), in which adequate correction is crucial. For sets smaller than about five individuals, the variation in the results was relatively large, but for 10 individuals or more, the gain in reproducibility was limited. The effect of training set size was also tested for two other dinucleotide markers, with similar results (not shown). A set of $n_i=10$ was found to give reliable results with a coefficient of variation of about 1%.

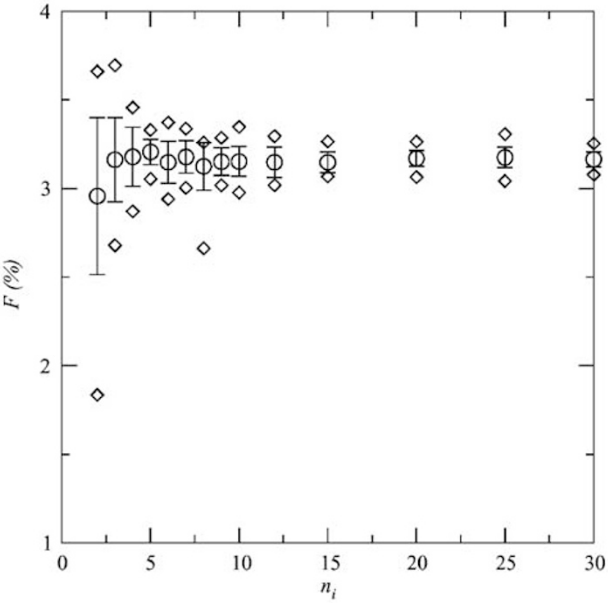

Figure 4 The peak at $a=127$ of Figure 3 calculated from fits with individual sets of $n_i=2$, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 25, and 30 individuals. For each value of n_i, 20 fits with randomly comprised sets were carried out. The values of the allele frequency F derived following correction are plotted on the vertical axis: mean (circles) and standard deviation (error bars) and the smallest and largest values (diamonds). Small sets already provide reliable frequencies with mean values of 3.2. The frequency found from summing the individual genotypes in the pool is 2.5. The frequency read from the uncorrected pool pattern was 5.8.

This test was also carried out for Perlin *et al*'s method.[15] For all alleles and values of n_i, the variation in estimated frequencies from this method was at least three times as large. For $n_i=30$, the variation was still twice as high as our method's variation for $n_i=10$.

Robustness to atypical training sets and measurement errors

Using data for marker D6S273, the robustness of the algorithm was checked in the following way: various sets of $n_i=10$ individuals were used as a training set to fit the model to the pooled data depicted in Figure 3: (i) a set of 10 homozygous individuals; (ii) a set of 10 heterozygous individuals; (iii) a set of 10 individuals whose alleles were closely packed together in a certain region, leaving part of the allelic range of the pool uncovered; (iv) a set of eight regular individuals plus two measurement errors: patterns containing an allele exhibiting a completely different stutter behavior to the others (but with peaks in the same molecular weight range).

All tests yielded good results that hardly differed from the 'normal' pool fit results of Figure 3.

A comparison of test (ii) with test (i) shows that no prechosen homozygous (or well-separated heterozygous) individual patterns are needed to derive good parameter estimates. Further, test (iii) shows that there is no need for training data to cover the full molecular range of alleles. Only in the extreme case of having only data points at one extreme of the molecular range in the training set do the pool results at the other end become less reliable. Test (iv) simulates the presence of measurement errors. If one or two of the 10 individual patterns are dissimilar to the others, for example, because of an artefact in the PCR process or a measurement error, the fit procedure does not appear to be misguided. The test showed that the fits derived from a training set of eight normal and two abnormal patterns were nearly as good as those based on 10 good patterns.

Validation of the model

For 34 different microsatellite markers, correction models were derived from the same training set of 10 individuals, and both uncorrected and corrected pools were compared with the true pools (for definitions see Materials and methods section). An example is shown in Figure 3. In total, five genotypes from four different markers could not be determined reliably, and these were discarded. Correlation coefficients of uncorrected and corrected pools *vs* true

Table 1 Statistical comparison of allele frequencies obtained by individual genotyping and frequency estimates from uncorrected and corrected pool patterns

				Uncorrected		*Corrected*	
Marker	*Type*	*Alleles*	*Het.*	*r*	*P*	*r*	*P*
D11S1338	di	7	0.72	0.92	$<10^{-3}$	0.99	0.89
D11S1760	di	10	0.77	0.44	$<10^{-3}$	0.61	$<10^{-3}$
D11S3178	di	10	0.67	0.84	$<10^{-3}$	0.99	0.92
D11S3179	di	7	0.70	0.93	$<10^{-3}$	0.99	0.93
D3S3585	di	7	0.58	0.92	$<10^{-3}$	1.00	0.80
D3S3665	di	6	0.55	0.93	$<10^{-3}$	0.99	0.80
D4S1582	di	7	0.78	0.89	$<10^{-3}$	0.98	0.88
D5S2005	di	7	0.66	0.91	$<10^{-3}$	0.99	0.10
D6S273	di	6	0.70	0.84	$<10^{-2}$	0.99	0.95
D6S291	di	8	0.72	0.92	$<10^{-3}$	0.99	0.20
D7S2422	di	15	0.83	0.78	$<10^{-3}$	0.93	0.01
DRD5	di	13	0.79	0.51	$<10^{-3}$	0.93	0.21
RH27315	di	5	0.64	0.86	$<10^{-3}$	1.00	0.87
D19S400	tetra	9	0.84	0.98	0.96	0.98	0.97
GAAT	tetra	6	0.69	0.98	0.45	0.98	0.44
TH01	tetra	7	0.77	0.99	0.85	0.99	0.85
Average		8	0.71	0.85		0.96	
kk3	di	6	0.75	0.81	0.06	0.94	0.75
kk7	di	6	0.74	0.90	0.29	0.99	0.93
kk9	di	5	0.77	0.81	0.02	1.00	1.00
kk11	di	6	0.73	0.81	0.02	0.99	0.99
kk16	di	6	0.57	0.86	0.07	1.00	0.69
kk20	di	7	0.55	0.95	0.19	0.99	0.19
kk24	di	12	0.82	0.78	$<10^{-3}$	0.99	0.49
kk26	di	7	0.77	0.80	0.05	0.96	0.90
kk28	di	14	0.79	0.76	0.08	0.95	0.77
kk31	di	14	0.86	0.77	0.78	0.96	0.96
kk37	di	9	0.72	0.85	0.19	0.99	1.00
kk42	di	6	0.31	0.90	$<10^{-3}$	1.00	0.63
kk43	di	9	0.72	0.82	0.01	0.99	0.91
kk45	di	11	0.70	0.79	0.02	0.99	0.94
kk56	di	9	0.76	0.83	0.11	0.99	0.82
kk58	di	9	0.84	0.83	0.67	0.95	0.97
kk61	di	9	0.78	0.88	0.26	0.99	0.98
kk62	tetra	7	0.75	0.87	0.28	0.85	0.16
Average		8	0.72	0.83		0.97	

Upper half of the table: values for characterized markers, analyzed in pooled DNA from 109 individuals. Lower half of the table: values for 'home-made' markers, analyzed in pooled DNA from 64 individuals. Di = dinucleotide repeat marker, tetra = tetranucleotide repeat marker. Alleles = number of marker alleles, determined by individual genotyping of pool samples. Het. = marker heterozygosity. r = correlation coefficient of individual genotyping results *vs* estimates from uncorrected as well as stutter corrected pools. P = P-value of χ^2 tests after combining alleles with expected low values, as calculated by the CLUMP algorithm.

pools for all 34 markers are given in Table 1. A graphical representation of the data for 16 characterized markers can be found on our web site (Figure C).

The only markers in which uncorrected pools approached true pools were the four tetranucleotide markers. For the dinucleotide markers, uncorrected pools were generally very different from the true pools, whereas corrected and true pools did not differ significantly, with the exception of markers D11S1760, D7S2422, and kk9. For marker D11S1760, there was a large overestimation of the frequency of the shortest allele in both uncorrected and corrected pools. Analysis of all individual genotype patterns for this marker revealed that stutter did not increase with allele length in a regular fashion (see web Figure D), which is an underlying assumption in the correction model. Marker D7S2422 showed a systematic overestimation of the peak height of the shorter allele in heterozygotes in the PoolFitter program, which persisted after correction for stutter. This suggested preferential amplification of shorter alleles, and after applying a simple compensation in the program, the differences between corrected and true pools were no longer significant (data not shown). No evidence for preferential amplification was found in the other markers (see web Figure E). Marker kk9 had two extra alleles (together accounting for 18% of all alleles in the true pool), with a size exactly between alleles at the regular 2 bp intervals. These aberrant alleles were discarded from the analysis, since the correction method ignores alleles at irregular intervals.

Case–control study

We investigated the application of the correction method in a case–control study in celiac disease (CD). DNA from 50 CD patients and 100 healthy controls was combined into two pools. Five microsatellite markers that had previously been used in association studies of CD patients were blinded and analyzed in CD and control pools. For three markers, allele frequencies did not differ significantly between cases and control pools, in either individual genotyping or pooled analysis. The other two markers showed significant differences between cases and controls. In each marker, one allele was very strongly associated, and already detectable in uncorrected pools, but after stutter correction, both markers also showed a much weaker but significant association with a second allele (see Figure 5). Both weak and strong associations were also demonstrable in the summed individual analysis.

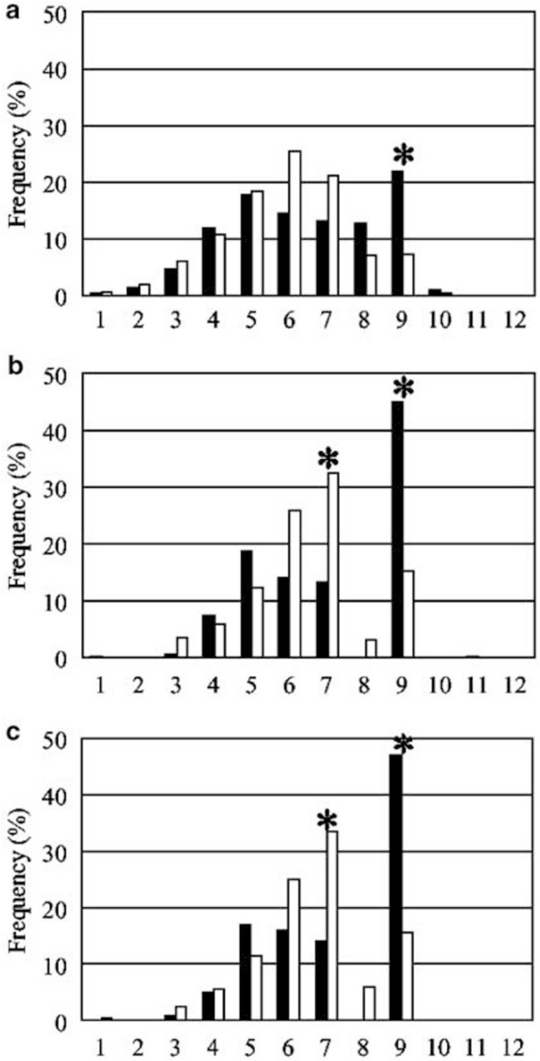

Figure 5 Comparison of allele frequencies in pools of celiac disease patients (black bars) and healthy controls (white bars) for marker D6D273. (**a**) Uncorrected pools, (**b**) corrected pools, (**c**) true pools. *X*-axis: allele number (increasing size). *Y*-axis: frequency of individual alleles (%). Significant differences for single alleles ($P<0.05$, not corrected for testing multiple alleles) are indicated with (*).

Discussion

Although the use of pooled DNA enormously reduces the amount of genotyping in comparing cases and controls, it suffers from the inability to generate haplotype information. As a result, microsatellite markers, with their high information content, are much more suitable than SNPs for use in pooled DNA samples. The number of potentially polymorphic microsatellites in the genome is much higher than the number of characterized markers in public databases. For example, in 11 schizophrenia candidate genes, we have tested 19 polymorphic microsatellites, eight of which were intragenic, while flanking markers were on average at 45 kb distance from the gene (max 130 kb). In a schizophrenia candidate region, we found nearly 250 potentially polymorphic microsatellites with an average spacing of 55 kb (max 168 kb). However, the widespread application of microsatellite markers in DNA pooling may have been prevented by uncertainties induced by stutter artefacts and the consequent distortion of allele frequency estimates.

We have developed a novel method, which enables accurate extraction of allele frequencies from microsatellite pool signals. A prerequisite for the application to large studies is that the correction method does not entail much additional analysis time. Our method meets this requirement, since the same training set of only 10 independent DNA samples plus the pool samples is required to carry out an analysis for a given marker. An apparent advantage of our approach is that there is no requirement for stutter and allelic peak signals of heterozygous individuals to be clearly separated, which greatly reduces the number of individuals required. There was little gain in accuracy when more than 5-10 individual genotypes were used, and accordingly, we choose one set of the same 10 independent individuals for all analyses, to allow for occasional dropouts. Other advantages of our fit algorithm are the simultaneous fitting of all data, which decreases the sensitivity to aberrant data, and that the size distribution of alleles and stutter, or alleles with an anomalous stutter height, had little influence on the predictive accuracy of the model.

The model was tested on DNA pool patterns with 34 different microsatellite markers, 18 of which were newly defined from human sequence data, since well-characterized markers could have been selected for their accuracy in genotyping. Our results with tetranucleotide markers confirm previous reports that stutter is low in these markers (generally $<5\%$) and that no stutter correction is required.[7,12,24] Significantly, for the two dinucleotide markers in which correction remained inaccurate, the presence of an aberration was readily detected in the PoolFitter program, even though it could not correct the stutter distortion.

In a case–control study involving celiac disease, marker alleles that were weakly associated in individual genotyping were also found to be associated in the pool analysis, but only after stutter correction. These two exceptionally strongly associated markers in the HLA region would have been detected even without correction, but would have been missed if only the weakly associated alleles had been present. This clearly demonstrates the benefit of stutter correction in DNA pooling.

Taken together, stutter correction generally resulted in accurate estimates of true allele frequencies in DNA pools. Compared with methods that use uncorrected pool patterns, several important advantages are apparent. The recently proposed ΔAIP and ΔTAC methods compare overall differences in peak area or peak height between pool patterns.[8,9] However, both methods assume a single fixed stutter profile for all markers and simulate large numbers of pool patterns to determine what proportion by chance will deviate significantly. Since the heights as well as the number of stutter peaks can differ greatly between markers, these methods raise the question whether realistic significance levels can be calculated in this way. In any case, such an approach prevents ascribing differences between pools to single alleles and summing results from different subpools or different experiments.[12] These drawbacks are not evident in our method.

We found that technical measures, such as reducing the number of PCR cycles, and adding pig-tail sequences to primers to eliminate plus-A artefacts, and separation on a capillary sequencer instead of a slab gel machine,[21] consistently improved the accuracy of DNA pool measurements. However, the nature of DNA pooling will inevitably result in some loss of sensitivity compared to individual genotyping.[25] Furthermore, a four-parameter model is not a perfect description of reality.

Despite these and other limitations, such as the lack of haplotype information, until cheap and rapid large-scale individual genotyping of markers for single individuals becomes technically feasible, DNA pooling methods allow efficient initial screening of candidate regions, and candidate gene systems. In pooled DNA, microsatellites are much more informative than single SNPs. In a second phase, associated microsatellites could then be followed-up by individual genotyping of high-density SNP markers, and haplotype analysis. Even if cases and controls were divided into pools of only 100 individuals each, as recently advocated,[16,26] and all amplified in triplicate, DNA pooling decreases genotyping by a factor of 30 in studies involving 500 cases and 1000 controls.

Our results confirm that the accuracy of analyzing corrected pool patterns generated from microsatellites approaches that of individual genotyping. Particularly in complex disorders, where the association of marker alleles with disease loci is likely to be only moderate or weak, a gain in sensitivity with stutter correction in pooled analyses justifies the limited amount of extra genotyping required to create a small training set.

In conclusion, we have demonstrated that accurate estimates of microsatellite allele frequencies from DNA pools are feasible with a novel stutter correction method requiring one standard training set of only 10 additional individual genotypes. This method opens the way for realistic large-scale genetic association studies using microsatellite markers.

Acknowledgements

We acknowledge the special contribution of Lodewijk Sandkuijl in formulating some of the basic concepts embodied in this work; he died shortly before completion of the manuscript. We also thank Martine van Belzen for providing the celiac disease samples and individual genotypes, and to Jackie Senior for critically reading the manuscript.

Electronic database information

The PoolFitter program and more illustrating figures are available at our website: http://www.smri.nl/microsatellites
CLUMP (DOS version): http://www.mds.qmw.ac.uk/statgen/dcurtis/software.html
Primer3: http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
Tandem Repeat Finder: http://c3.biomath.mssm.edu/trf.html
Genome Database (mirror site): http://gdbwww.dkfz-heidelberg.de/
Marshfield Center for Medical Genetics: http://research.marshfieldclinic.org/genetics/

References

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 Dunning AM, Durocher F, Healey C *et al*: The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet* 2000; **67**: 1544–1554.
3 Jorde LB: Linkage disequilibrium and the search for complex disease genes. *Genome Res* 2000; **10**: 1435–1444.
4 Abecasis GR, Noguchi E, Heinzmann A *et al*: Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 2001; **68**: 191–197.
5 Innan H, Padhukasahasram B, Nordborg M: The pattern of polymorphism on human chromosome 21. *Genome Res* 2003; **13**: 1158–1168.
6 Salisbury BA, Pungliya M, Choi JY, Jiang RH, Sun XJ, Stephens JC: SNP and haplotype variation in the human genome. *Mutat Res* 2003; **526**: 53–61.
7 Barcellos LF, Klitz W, Field *et al*: Association mapping of disease loci, by use of a pooled DNA genomic screen. *Am J Hum Genet* 1997; **61**: 734–747.

8 Collins HE, Li H, Inda SE *et al*: A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. *Hum Genet* 2000; **106**: 218–226.

9 Daniels J, Holmans P, Williams N *et al*: A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. *Am J Hum Genet* 1998; **62**: 1189–1197.

10 Fisher PJ, Turic D, Williams NM *et al*: DNA pooling identifies QTLs on chromosome 4 for general cognitive ability in children. *Hum Mol Genet* 1999; **8**: 915–922.

11 Plomin R, Hill L, Craig IW *et al*: A genome-wide scan of 1842 DNA markers for allelic associations with general cognitive ability: a five-stage design using DNA pooling and extreme selected groups. *Behav Genet* 2001; **31**: 497–509.

12 Kirov G, Williams N, Sham P, Craddock N, Owen MJ: Pooled genotyping of microsatellite markers in parent-offspring trios. *Genome Res* 2000; **10**: 105–115.

13 LeDuc C, Miller P, Lichter J, Parry P: Batched analysis of genotypes. *PCR Methods Appl* 1995; **4**: 331–336.

14 Lipkin E, Mosig MO, Darvasi A *et al*: Quantitative trait locus mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: analysis of milk protein percentage. *Genetics* 1998; **149**: 1557–1567.

15 Perlin MW, Lancia G, Ng S-K: Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. *Am J Hum Genet* 1995; **57**: 1199–1210.

16 Sham P, Bader, JS, Craig I, O'Donovan M, Owen M: DNA pooling: a tool for large scale association studies. *Nat Rev Genet* 2002; **3**: 862–869.

17 Sham PC, Zhao JH, Curtis D: The effect of marker characteristics on the power to detect linkage disequilibrium due to single or multiple ancestral mutations. *Ann Hum Genet* 2000; **64**: 161–169.

18 Morris RW, Kaplan NL: On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 2002; **23**: 221–233.

19 Bakker SC, van der Meulen EM, Buitelaar JK *et al*: A whole-genome scan in 164 Dutch sib pairs with attention-deficit/hyperactivity disorder: suggestive evidence for linkage on chromosomes 7p and 15q. *Am J Hum Genet* 2003; **72**: 1251–1260.

20 Brownstein MJ, Carpenter JD, Smith JR: Modulation of non-templated nucleotide addition by *Taq* DNA polymerase: primer modifications that facilitate genotyping. *BioTechniques* 1996; **20**: 1004–1010.

21 Sham PC, Curtis D: Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann Hum Genet* 1995; **59** (Part 1): 97–105.

22 Miller MJ, Yuan B-Z: Semiautomated resolution of overlapping stutter patterns in genomic microsatellite analysis. *Anal Biochem* 1997; **251**: 50–56.

23 Press WH, Teukolsky SA, Vettering WT, Flannery BH: *Numerical recipes in C – the art of scientific computing*, 2nd edn. Cambridge: Cambridge University Press, 1992.

24 Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A: Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res* 1998; **8**: 111–123.

25 Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG: Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann Hum Genet* 2002; **66**: 393–405.

26 Sawcer S, Maranian M, Setakis E *et al*: A whole genome screen for linkage disequilibrium in multiple sclerosis confirms disease associations with regions previously linked to susceptibility. *Brain* 2002; **125**: 1337–1347.