GENOME METHODS

Methods for Precise Sizing, Automated Binning of Alleles, and Reduction of Error Rates in Large-Scale Genotyping Using Fluorescently Labeled Dinucleotide Markers

Soumitra Ghosh,[1,6] Zarir E. Karanjawala,[1] Elizabeth R. Hauser,[2] Delphine Ally,[1] Julie I. Knapp,[1] Joseph B. Rayman,[1] Anjene Musick,[1] Joyce Tannenbaum,[1] Catherine Te,[1] Shane Shapiro,[1] William Eldridge,[1] Tiffany Musick,[1] Colin Martin,[1] Jeffrey R. Smith,[1] John D. Carpten,[1] Michael J. Brownstein,[4] John I. Powell,[3] Raymond Whiten,[1] Peter Chines,[1] Stella J. Nylund,[5] Victoria L. Magnuson,[1] Michael Boehnke,[2] Francis S. Collins,[1] and the FUSION (Finland–U.S. Investigation of NIDDM Genetics) Study Group

[1]Positional Cloning Section, Laboratory of Gene Transfer, National Center for Human Genome Research, [3]Computational Bioscience and Engineering Laboratory, Division of Computer Research and Technology, and [4]National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892; [2]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109-2029; [5]Diabetes and Genetic Epidemiology Unit, Department of Epidemiology and Health Promotion, National Public Health Institute, Helsinki, Finland

Large-scale genotyping is required to generate dense identity-by-descent maps to map genes for human complex disease. In some studies the number of genotypes needed can approach or even exceed 1 million. Generally, linkage and linkage disequilibrium analyses depend on clear allele identification and subsequent allele frequency estimation. Accurate grouping or categorization of each allele in the sample (allele calling or binning) is therefore an absolute requirement. Hence, a genotyping system that can reliably achieve this is necessary. In the case of affected sib-pair analysis without parents, the need for accurate allele calling is even more critical. We describe methods that permit precise sizing of alleles across multiple gels using the fluorescence-based, Applied Biosystems (ABI) genotyping technology and discuss ways to reduce genotyping error rates. Using database utilities, we show how to minimize intergel allele size variation, to combine data effectively from different models of ABI sequencing machines, and automatically bin alleles. The final data can then be converted into a format ready for analysis by statistical genetic packages such as MENDEL.

There are many studies in progress worldwide to elucidate the genetic determinants of complex diseases (Davies et al. 1994). After careful study design and power calculations, it is usually necessary to perform a total genome scan (Lander and Schork 1994; Ghosh and Collins 1996; Hauser et al. 1996). This approach assumes no prior information as to the location, function, or number of genes contributing to disease (Collins 1995). Dense marker coverage for all chromosomes is necessary, typically at a resolution of 10–20 cM. Microsatellite markers are useful for this purpose, as they are abundant, fairly evenly spread throughout the genome, easily amplified by PCR, and highly polymorphic. As dinucleotide repeat markers are more abundant than microsatellites with even larger motifs, the former have been the most commonly used reagents (Cooperative Human Linkage Center (CHLC) 1994; Gyapay et al. 1994; Dib et al. 1996). However, given that a

[6]Corresponding author.
E-MAIL sghosh@alw.nih.gov; FAX (301) 480-9667.

10-cM map resolution is now also achievable with tri- and tetranucleotide repeats, it is becoming commonplace to type with these markers alone (Dubovsky et al. 1995; The Utah Marker Development Group 1995; Gastier et al. 1995; Sheffield et al. 1995). The advantage of tri- and tetranucleotide repeats over dinucleotides is the greater ease of the scoring of alleles due to less complex patterns of overlapping allele peaks. The disadvantage is that usually fewer markers of this type can be coelectrophoresed together in a single gel lane owing to their larger interallelic distances. As our primary goal was to maximize genotyping throughput, we chose to type with dinucleotide markers.

Our experiences come mainly from the FUSION [Finnish–U.S. Investigation of NIDDM (noninsulin-dependent diabetes mellitus) Genetics] study in which >900,000 genotypes are being generated using 365 markers to type nearly 2500 individuals. During the last year we have been expanding our genotyping capabilities by purchasing new model sequencers and partially automating the data generation and analysis steps. Specifically, based on our experience, we describe an approach for managing large amounts of genotype data using dinucleotide markers divided into panels and electrophoresed on different types of model sequencers. Each panel comprises a set of markers that can be coelectrophoresed together in a single lane. These panels are mostly from the ABI PRISM Linkage Mapping Set (Applied Biosystems Division/Perkin-Elmer, Foster City, CA) but with some changes because we wanted a denser map in regions of the genome containing possible NIDDM susceptibility genes.

Several groups have described the use of the ABI system (Ziegle et al. 1992; Reed et al. 1994; Schwengel et al. 1994), but none to our knowledge have described programmatic utilities necessary to efficiently process large amounts of data. We also discuss in detail laboratory techniques used to carry out large-scale genotyping but primarily focus on ways to increase the precision of allele sizing and to reduce genotyping error rates. Increased precision for allele sizing leads to more accurate allele calling or binning.

RESULTS

Allele-Sizing Algorithms and Binning

Sizing of alleles with three-dye (blue, green, yellow) fluorescent labeling depends on the coelectrophoresis of a standard molecular weight ladder in each lane that is labeled with a fourth dye, usually red. Allele sizes for microsatellite fragments are estimated using interpolation after fitting a calibration curve to the mobility data for the size standards. In this paper we define binning or allele calling to be the unique categorization of each allele of a marker by the assignment of an integer label. This integer label is the mean size in base pairs of alleles for that particular allele category, rounded off to the nearest whole number. Thus, the characteristics of each bin can be described by the bin label, mean, range, and standard deviation. This definition of binning is different from that applied to variable numbers of tandem repeats (VNTRs), where many alleles of different sizes are grouped together in one bin because they cannot be resolved. Precise binning is particularly critical for late-onset diseases like NIDDM, as the availability of parents for typing is usually limited, but is also important for any study involving multiple pedigrees in which not all family members are typed.

Prior to initiating large-scale genotyping we tested the five specific built-in calibration methods of the ABI GENESCAN software (local Southern, global Southern, second-order least squares, third-order least squares, or cubic spline) and confirmed that the local Southern, as recommended by ABI, was the method of choice (see Fig. 1 for an example of the local Southern compared with the second-order least squares method). The local Southern method uses molecular weight standards flanking the microsatellite product to estimate parameters for a model that assumes the mobility of DNA fragments in a gel to be inversely proportional to size in base pairs (Southern 1979). Figure 1 shows that the calibration method employed does have a bearing on the precision of allele sizing in contrast to what was stated in an earlier study by Maynard et al. (1992). All subsequent genotypes were sized using local Southern calibration and the ABI PRISM GENESCAN-500 (GS-500) size standards (see Methods).

External Adjustment Across Two 373 Sequencing Machines

Rationale

After typing large numbers of DNA samples ($n > 2000$) across two different 373 sequencers, it was noticed that the bin ranges for some markers became unacceptably large (≥ 1 bp). Following Mansfield et al. (1994), where the automated laser fluorescence sequencer (ALF) system (Pharmacia) was described, we initiated the practice of running a uniform control DNA [Centre d'Etude du Polymorphisme Humain (CEPH) 1347-02] sample on each

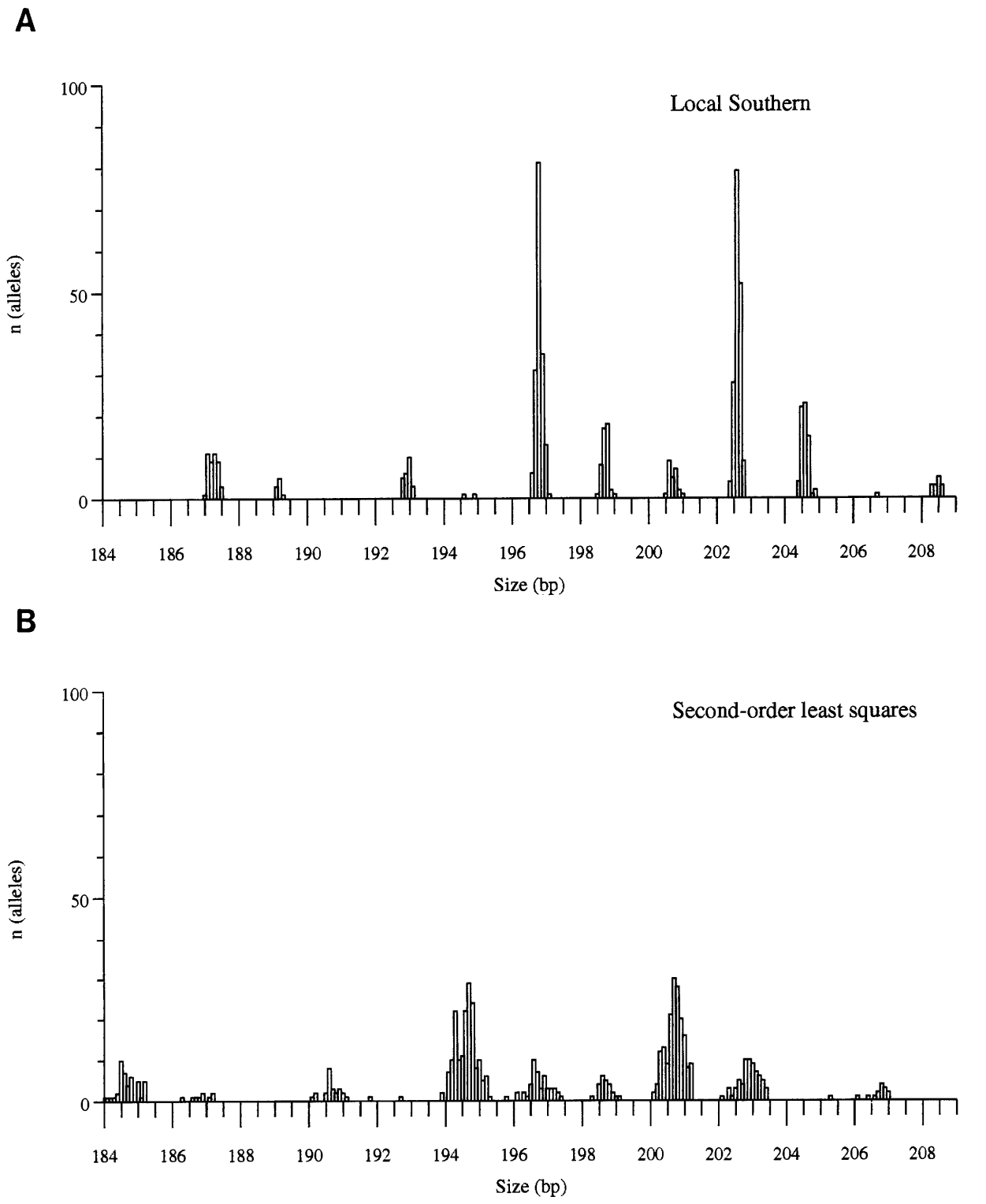

Figure 1 D1S468 frequency histograms for 572 microsatellite alleles run on a 373 and sized using either local Southern (*A*) or second-order least squares (*B*) algorithms.

gel in the center, lane 17. This sample was typed with the same markers as for our study DNAs. Using allele sizing information obtained from this control DNA, external adjustment (which compensates for

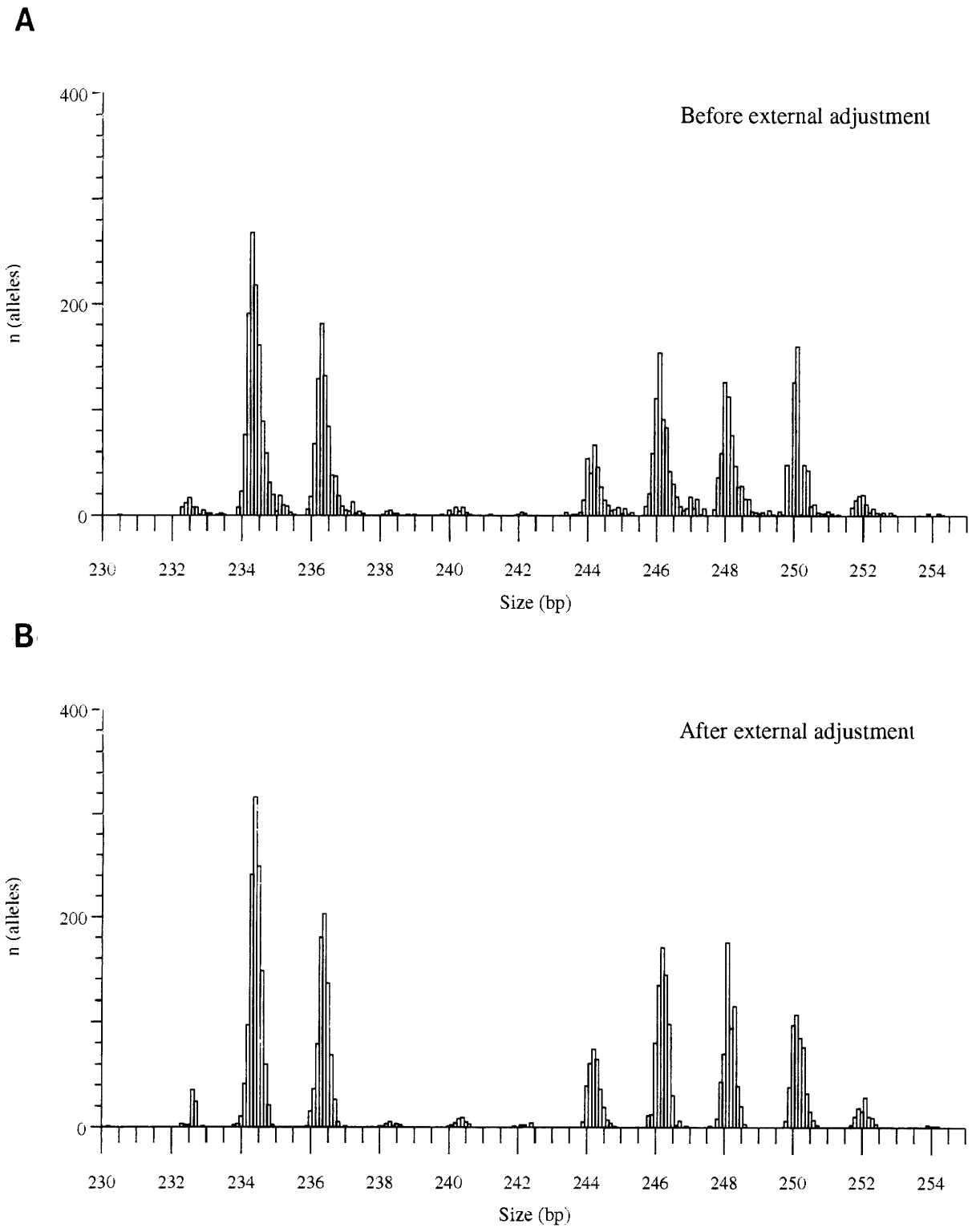

Figure 2 Illustrative example for external adjustment to correct for intergel variation: Frequency histograms for D1S220 prior to (*A*) and after (*B*) adjustment based on 2098 genotypes run on two different 373 sequencers (n = 4196 alleles).

intergel size variation) was applied to sample data to increase precision in allele sizing. Initially, external adjustment was shown to reduce the ranges of bins and increase the interbin distances for the set of 10 markers studied in the ABI PRISM Linkage Mapping Set, panel 1, that were coelectrophoresed on two different 373 model sequencers (see Fig. 2 for an example). Subsequent experience with several panels has confirmed the value of the external adjustment step in genotyping on the ABI system.

Algorithm

External adjustment (as opposed to internal adjustent with GS-500 size standards) is used to reduce the bin sizes and to increase interbin distances before efficient, automated binning. Although the results are similar, the external adjustment algorithm developed here is unlike the one described in Mansfield et al. (1994), which applied the difference between estimated allele sizes and the real sizes to all alleles for a specific marker on a gel. In contrast, our external adjustment algorithm compares the sizes of marker alleles from a CEPH (1347-02) standard, averaged over all electrophoresis runs, with the 1347-02 genotype run on each gel. The difference between these two values is subtracted from all sample alleles for that particular marker/gel combination. If the CEPH standard genotype is heterozygous, means are computed for each of the two alleles across the total number of gels. The deviations for the two alleles on a specific gel from their corresponding mean values are then averaged and the final value applied to all sample marker alleles, as outlined above. Thus, our algorithm is not dependent on whether the control sample genotype is homozygous or heterozygous. It also normalizes to mean sizes as opposed to real sizes and adjusts each marker's allele sizes independently of other markers.

Binning of Data

Rationale

At the time we began the project (June 1995) we were unaware of methods for automatically binning alleles using the ABI system. However, it is possible by using the ABI GENOTYPER software to find means and standard deviations for all allele categories one-by-one and then assigning them a bin label (e.g. bin = mean $\pm$ 2 S.D.). Unfortunately, this process is time-consuming, tedious, memory intensive, and somewhat error-prone, as it remains largely a manual process. Furthermore, we wanted to be able to bin the whole data set simultaneously from nearly 2500 individuals rather than using a subset to initially define the bin ranges. This is because we could potentially miss rarer alleles by using the latter approach. Finally, we wanted the added flexibility of binning alleles whose mean sizes differed by only 1 bp and also bin markers alleles with wide size ranges. The algorithm that we designed has these features.

Algorithm

For each marker, alleles for the total data set are sorted according to size. A userspecified "tolerance level" is selected that represents the minimum allowable distance between adjacent bins in base pairs. A good starting value is 0.4 bp. When the size difference between two sequentially sized alleles is greater than the set tolerance level, a new bin is created. The procedure is performed for each marker until all alleles are binned. The smallest and largest sized alleles for any marker represent the start of the first bin and the end of the last bin, respectively. Clearly, the assumption is that within a true bin no two sequentially sized alleles will have a size difference greater than or equal to the set tolerance level. After grouping the alleles, the means, standard deviations, and ranges for all the bins are calculated. Next, the bin labels, which are the mean sizes rounded up to nearest whole numbers, are assigned for each group. For alleles 1 base apart the tolerance level is normally set at a value near 0.2 bp for error-free binning to proceed. We have written software to carry out binning and adjustment automatically (ABAS or Automated Binning and Adjustment Software).

External Adjustment and Binning Across both 373 and 377 Sequencers

Rationale

The recent purchase of a model 377 sequencer prompted us to directly compare sizing between the 377 and 373 for the same marker alleles ranging from 100 to 300 bp using data from 12 dinucleotide markers in a modified version of the ABI PRISM Linkage Mapping Set, panel 3. The most common allele category or bin was studied for each marker. A comparison between electrophoresis runs on 373 and 377 sequencers revealed that means for allele

Table 1. Comparison of sizing on ABI 373 and ABI PRISM 377 Sequencer

Marker	Machine	Bin label (in bp)	*n* (alleles)	Mean	Standard deviation
D2S326	373	102	307	101.68[a]	0.16[b]
	377	100	399	100.04	0.09
D2S362	373	105	611	104.73[a]	0.17[b]
	377	103	416	103.03	0.10
IRS1	373	132	843	132.17[a]	0.14[b]
	377	131	1070	130.92	0.09
D2S336	373	136	484	135.80[c]	0.09[b]
	377	135	686	134.90	0.08
D2S206	373	146	564	145.66[a]	0.12[b]
	377	145	725	144.83	0.08
D2S325	373	166	488	165.55[a]	0.16[b]
	377	164	636	164.41	0.08
D2S113	373	207	671	207.04[a]	0.11[b]
	377	207	762	206.91	0.07
D2S117	373	208	377	207.83[a]	0.11[b]
	377	207	455	207.45	0.08
D2S142	373	237	513	237.06[a]	0.20[b]
	377	237	607	237.15	0.11
D2S164	373	278	615	277.79[a]	0.20[b]
	377	277	762	277.41	0.08
D2S126	373	298	387	297.61[a]	0.13[b]
	377	297	542	297.42	0.07
D2S367	373	299	309	299.08[a]	0.16[b]
	377	299	390	298.85	0.07

The *n* value represents the number of alleles observed across all gels for the most common allele bin for each marker. The bin categories for the same marker may differ across the two types of sequencers as shown. Sizes were computed using the local Southern method.

[a] $P < 0.0001$ for testing the hypothesis that the two means are equal using the two-sample *t*-test allowing for unequal variances.

[b] $P < 0.0001$ for testing the hypothesis that the variances are equal using the two-sample F-statistic.

sizes across all molecular weight ranges were significantly different at the nominal value of $P = 0.0001$ (Table 1). It was also noted that all alleles were sized to greater precision on the 377 sequencer. In general, the differences between the means for the same alleles on the 373 versus 377 were less pronounced with higher molecular weights.

Because the sizing of some marker alleles differed significantly across 373 and 377 model sequencers, it was necessary to apply external adjustment to panel data electrophoresed on both types of machines. The aim was to bin alleles accurately belonging to the same category (see Fig. 3 for an example).

Evaluation

To check that all data have been appropriately adjusted, database programs within ABAS can flag alleles that have either zero adjustment or improper adjustment. Maladjusted alleles are detected by scrutinizing outliers where the adjustment is beyond three standard deviations from the mean adjustment value, for that marker, across all gels.

Terminal Transferase Activity of *Taq* DNA Polymerase and Allele Systems 1 bp Apart

Within the last year, we have routinely used a method developed by FUSION that increases the fraction of PCR products with a 3′ terminal nucleotide extension (“allele-plus-A”) to >70% for all markers tested (Magnuson et al. 1996). Modification of the 5′-end nucleotide of the reverse, nonfluorescent primer to G (guanine) promotes the nontemplated addition of adenine by *Taq* DNA polymerase

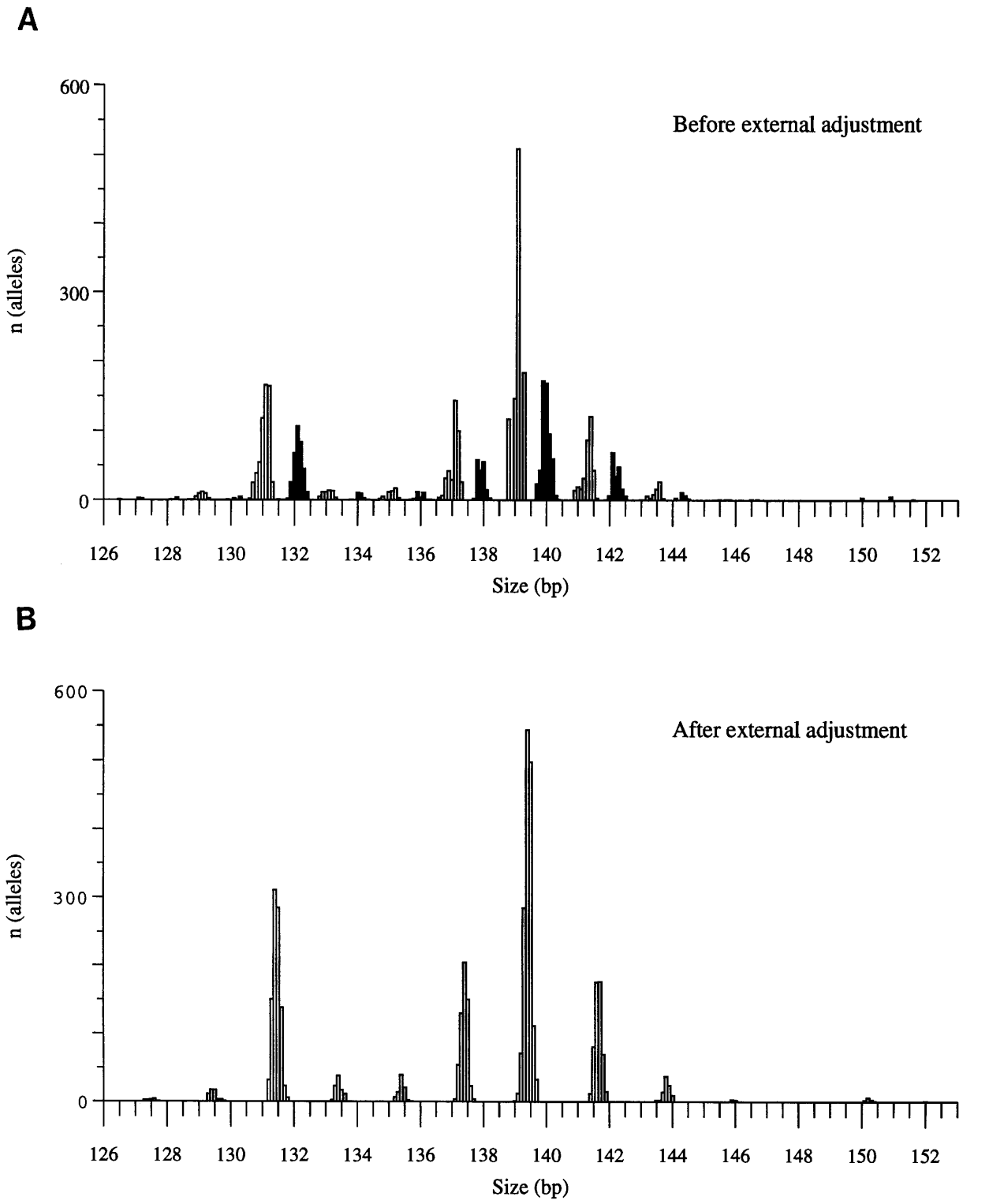

Figure 3 Example of external adjustment using marker D8S504 across both 373 (solid bars) and 377 (open bars) model sequencers prior to (*A*) and after (*B*) adjustment (open bars) based on 1974 genotypes (n = 3948 alleles.)

onto the complementary, fluorescently labeled strand. Because of this consistent shift in allelic profile towards allele-plus-A, it has been possible to virtually eliminate the incorrect calling of alleles by GENOTYPER in cases where previously an approximate equal amount of allele and allele-plus-A prod-

uct was present (Magnuson et al. 1996). Furthermore, the modification makes it easier to detect markers with alleles that are truly 1 bp apart.

Among 121 markers mainly from the ABI PRISM Linkage Mapping Set, 17 markers (14%) had alleles differing by 1 bp: Six (5%) of them (D1S214, D1S499, D2S396, D7S486, D7S550, D15S127) had 1-bp alleles that were also seen in CEPH families assayed by Genethon; 10 (8.3%) markers (D1S199, D1S234, D1S423, D2S113, D2S142, D2S368, D8S258, D12S306, D15S123, D15S126) had 1-bp alleles that were common among Finnish samples seen during genotyping; and 1 (0.83%) marker (D2S139) had a 1-bp allele system that was rare or family-specific, detected only after binning.

Of the 121 markers studied, 105 were used to assay Finnish samples (however, see below in Evaluation of the Overall Genotyping Process for a note on marker D8S272, as this marker is not included in the count). Sixteen markers were removed because of poor quality of which seven were due to detected 1-bp allele systems. Our preference is not to type dinucleotide markers that give alleles 1 bp apart (according to the sizes listed in the public databases) as these are more difficult to bin. In addition, markers containing compound or complex repeat structures also have been avoided.

Ten markers of the 105 markers (9.5%) finally typed on the FUSION sample exhibited 1-bp allele systems. The total number of alleles for the marker and the frequency of the most common 1-bp allele (by gene counting) are given, respectively, in parentheses: D1S199 (15, 28/4048); D7S550 (17, 46/3938); D1S234 (15, 14/3834); D2S113 (16, 11/4428); D2S142 (10, 10/4530), D2S368 (22, 1568/4282); D8S258 (8, 8/4046); D12S306 (17, 36/4174); D15S123 (13, 93/4504); and D2S139 (13, 5/4564). The reduction of the tolerance level to ~0.2 bp allows automated binning of alleles that are 1 bp apart or more, provided that allele bins do not overlap in size range (see previous section). Using this approach, it has been possible to successfully bin the above 10 markers where many alleles are 1 bp apart.

Data Flow in the Laboratory

We have developed a set of utilities for efficient management of genotyping data in the FOXPRO relational database environment (see Fig. 4). We run these utilities subsequent to the normal editing process using ABI commercial software. After the raw data have been sized by GENESCAN version 1.2.2-1 (or by v. 2.02 for the 377), the data are independently checked by two technicians using GENOTYPER (v. 1.1r8). Two text files are created (GENOTYPER Primary Check and GENOTYPER Secondary Check); (Fig. 4). These originate from the same result (373) or sample file (377). Manual checking is necessary, as GENOTYPER occasionally labels noise bands and other peaks that represent chromatic interference from a separate color ("bleedthrough"). Furthermore, the filtering algorithms within GENOTYPER often remove labels from true alleles.

After the independent manual checkings, the fields in the two text files (GENOTYPER Primary Check and GENOTYPER Secondary Check) are compared using a matching algorithm. An output listing the differences between the two independent scorings is generated. Genotyping data from the most recent 92 runs (during a 2-week period) in our laboratory reveals a discrepancy rate of 2.6% (838/32,035). Most discrepancies are usually attributable to a difference of opinion in scoring (549 or 1.7%) and are resolved by discussion. The rest (289 or 0.9%) are attributable to errors on the part of one person. In these cases, the alternate person has usually made the correct judgement, which is formally accepted. Differences between the two GENOTYPER files (primary and secondary) are ultimately reconciled and a composite final text file is manually created.

For each electrophoresis run, only one corresponding text file is eventually taken through the FOXPRO programs. Once this final text file enters the database environment all manual editings are filtered and stored for future reference. During the binning process concurrent examination of a histogram showing allele sizes is a particularly useful step in detecting outliers in the binning process. External adjustment is then applied to the data before binning and the data converted into a format ready for statistical analysis (Fig. 4). For a more detailed description of special editing rules please contact us.

Evaluation of the Overall Genotyping Process

Every allele profile generated in the FUSION project is scored independently by two different genotypers (see above). Fifty DNAs are present in duplicate in the total sample set, which is regularly genotyped in the FUSION project. The genotypers are blind to the identity of the duplicate samples allowing for an estimate of genotyping error rates for each panel of markers. Genotype-specific error rates are calculated as the percentage of discrepant genotypes among the total number of genotypes in fully typed pairs and thus assumes that at least one of the duplicate

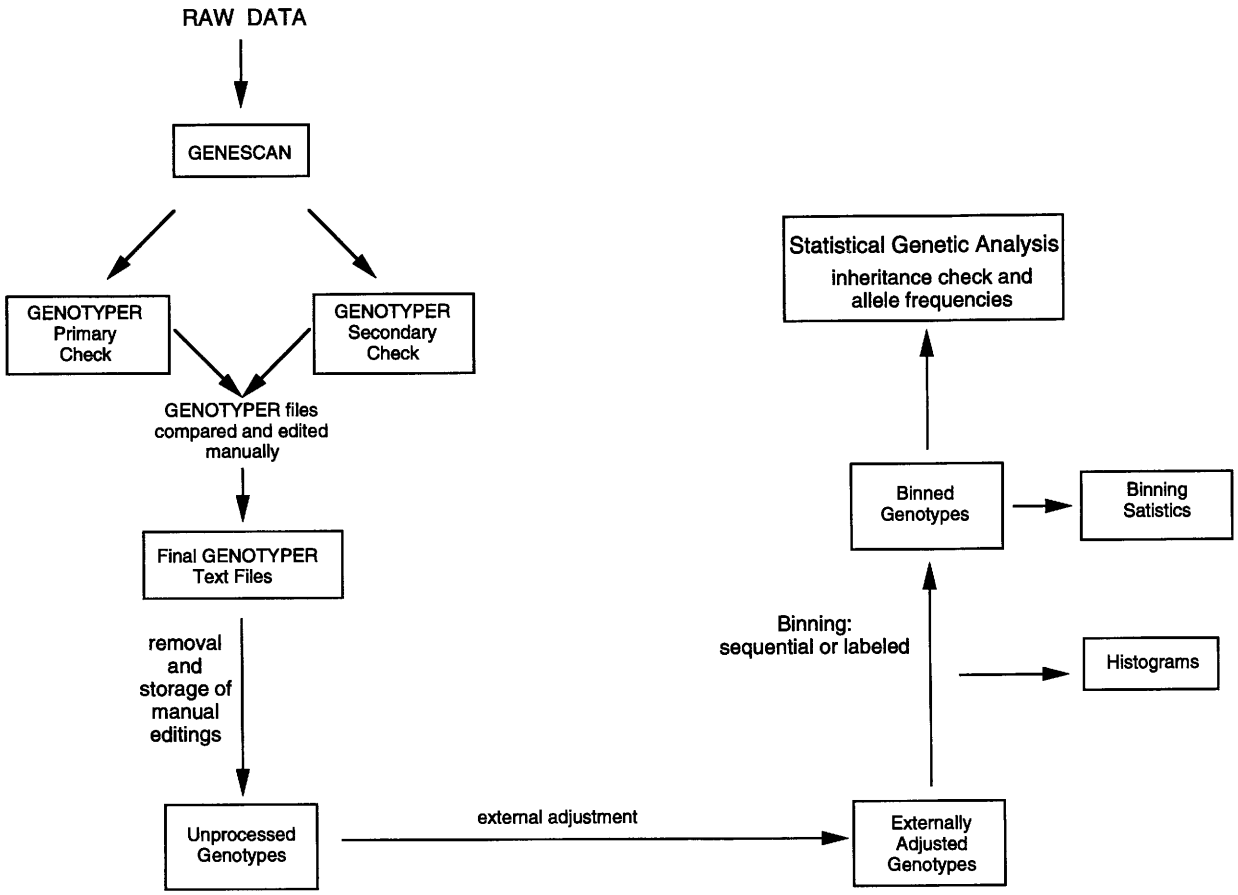

Figure 4 Data flow: from raw genotype to statistical genetic analysis files (see text for description).

genotypes is correct. Likewise allele-specific error rates are calculated as the percentage of discrepant alleles among the total nonmissing numbers of alleles for pairs of genotypes with complete information (see Table 2). It should be noted that the genotype-specific error rate defined here is approximately half the discrepancy rate, which is the proportion of fully typed duplicate pairs that are discordant for binned genotypes. Furthermore, because in most discrepant cases just one allele of a retyped genotype is different, the allele-specific error rates are very close to one-half the genotype-specific error rates. Because the FUSION data set contains a large number of sib-pairs without parents that do not provide information about non-Mendelian inheritance (NMI), we compute error rates for sib-pairs and for larger families separately, in addition to the overall error rate. In addition, we prefer to discuss allele-specific error rates, as each allele is binned independently of others in ABAS. However, because genotypes are the currency of genetic analysis, genotype error rates are also given in Table 2.

There is no fully satisfactory way to evaluate the accuracy and reproducibility of our binning and external adjustment algorithms, in themselves. This is because sample switches, sample loading errors, PCR artifacts, sizing inaccuracies, and editing errors all contribute to the lack of reproducibility of data or to the generation of NMI. However, after removing data in which at least one genotype is missing from the corresponding duplicate samples, the allele-specific error rates on typing duplicate samples are 0.30% (11/3632) for sib-pairs without parents and 0.20% (32/16,084) when more extended family information is available (see Table 2). Of the 108 markers assayed, 106 were unique markers, as 2, D7S484 and D7S517, were typed twice for quality control purposes.

The error rates are inflated because of one marker in panel 11, D8S272, where a primer sample mixup caused the wrong primer pair to be used for generating genotypes in a proportion of the individuals in the data set. When the duplicate samples were compared this problem became obvious as the same allelic PCR product from the correct and in-

Table 2. Error Rates for Nine Panels Genotyped During the FUSION Study

Panel	No. of markers	Overall				Sib-pairs				Extended families			
		genotypes (no.)	missing (%)	error G (%)	error A (%)	genotypes (no.)	missing (%)	error G (%)	error A (%)	genotypes (no.)	missing (%)	error G (%)	error A (%)
1	10	500	4.0	0.42	0.21	90	4.4	1.74	0.87	410	3.9	0.13	0.06
2	10	500	4.6	0.42	0.21	90	4.4	1.16	0.58	410	4.6	0.26	0.13
3	12	600	5.0	0.09	0.05	108	0.9	0.00	0.00	492	5.9	0.11	0.05
4	15	750	4.9	0.00	0.00	135	2.2	0.00	0.00	615	5.5	0.00	0.00
11	15	750	10.5	1.04	1.04	135	13.3	1.28	1.28	615	9.9	0.99	0.99
12	9	450	16.0	0.53	0.33	81	12.3	0.00	0.00	369	16.8	0.65	0.41
18	14	700	7.4	0.00	0.00	126	10.3	0.00	0.00	574	6.8	0.00	0.00
21	11	550	14.4	0.00	0.00	99	7.1	0.00	0.00	451	16.0	0.00	0.00
22	12	600	13.2	0.10	0.05	108	3.7	0.00	0.00	492	15.2	0.12	0.06
Total	108	5400	8.7	0.28	0.22	972	6.6	0.44	0.30	4428	9.2	0.25	0.20
Total without D8S272	107	5350	8.7	0.14	0.08	963	6.5	0.28	0.14	4387	9.2	0.11	0.06

correct primer pairs differed by 2 bp in length prompting us not to analyze D8S272 any further. As this was an unusually rare occurrence, the removal of D8S272 brings the overall allele-specific error rates down from 0.22% (43/19,716) to 0.08% (15/19,532) across 105 markers (Table 2).

Mendelian inheritance checks are performed during allele frequency estimations with the USERM13 program of the MENDEL package (Lange et al. 1988; Boehnke 1991). Because available parents are rare in the FUSION sample, detecting cases of NMIs is more difficult. Nevertheless, for those families where some extended family information is available, the number of NMIs detected per meioses for the five panels typed most recently was 4.53×10^{-4} (28/61,760) across 64 markers. This value is an estimation from the last five panels (modified panels 3, 4, 18, 21, and 22, consisting of 64 markers) typed in our laboratory and represents the only data presently available where amplification products have been electrophoresed across two different types of model sequencers. The estimate of 4.53×10^{-4} compares favorably with accepted mutation rates for microsatellites (10^{-3}–10^{-6}), attesting to the fact that our binning process is robust.

The percentage of missing genotypes attributable to either random PCR failure (as opposed to all 96 samples not amplifying) or to manual removal following first-pass typing is, on average, 8.34% (12,903/154,802) for the last 64 markers.

DISCUSSION

Errors in genotyping can give rise to inflated map lengths and reduced power to detect linkage (Buetow 1991). Therefore, methods to minimize genotyping errors are important. The aim of the present study was to devise precise sizing conditions for dinucleotide repeat markers across multiple gels using the fluorescent-based ABI technology. Increased precision in sizing should lead to more accurate binning of alleles. Independent scoring of each genotype, allele frequency estimation, checking for Mendelian inheritance, comparison with expected Hardy–Weinberg frequencies, comparison of map distances with published values, and looking for closely spaced multiple recombinants are part of our scheme to achieve the goal of highly reproducible and precise sizing of alleles. Careful genotyping is particularly important in studies of small families such as affected sib-pairs, in which errors may not be readily detectable as incompatibilities with Mendelian inheritance.

The methods presented in this paper use straightforward algorithms to statistically adjust and bin thousands of genotypes in a single step. The process is fully automated and almost error-free when modified primer sets are used, thereby reducing ambiguous allele calls using GENOTYPER. The external adjustment algorithm applies an average correction to all alleles for a particular marker from

a specific electrophoresis run and thus can be used even when the control sample is a heterozygote. In addition, by adjusting to the mean of the distribution of control allele sizes, it is possible to increase interbin distances and reduce bin sizes for all alleles, which increases binning efficiency (see Figs. 2 and 3). Finally, knowledge of the true allele sizes of the control sample genotypes is not necessary, which allows any DNA sample to be used as a control provided it is the same one each time.

The binning algorithm in ABAS does not require the determination of allele bin statistics prior to binning (as does GENOTYPER, v. 1.1.1 and test verions of 2.0) and can bin data from >100 electrophoresis runs (>40,000 genotypes) at once, thus allowing simultaneous automated binning for all alleles in the same data set. It is more robust when using larger as compared to smaller data sets, simply because the algorithm assumes that that no two alleles within the same bin differ in size greater than the set tolerance level. However, the use of the variable tolerance level gives the algorithm the added flexibility of being capable of binning alleles from smaller data sets (by increasing the tolerance levels) and correctly binning alleles that differ by 1 bp (by decreasing the tolerance level)

Maynard et al. (1992) hypothesized that by using an internal molecular weight standard in every lane, any inhomogeneity in running conditions (e.g., heat distribution, sample amount, salt concentration) between and within gels could be reduced. Using the local Southern method to interpolate the allele sizes of dinucleotide markers from the GS-500 size standard does give good precision within a given gel. Nevertheless, results described here indicate that there are still unknown factors that give rise to intergel allele size variation (even on a single type of sequencer) when using the local Southern method, which are not compensated for by "internal adjustment". Furthermore, sizing of alleles on the 373 sequencer is less precise than sizing on the 377 sequencer and therefore the latter is far superior for large-scale genotyping. There are also differences in electrophoretic conditions, gel composition, gel thickness, and detection that can lead to systematic differences of up to nearly 2 bp in allele sizing on a 373 sequencer as compared to a 377 sequencer (see Table 1; Methods). Such major discrepancies could potentially compromise the success of a large-scale genotyping project when both machines are employed.

By typing the same genomic control sample on every gel and adjusting for variations in observed allele sizing of this control, it was possible to increase precision in sizing and to advance substantially toward the goal of fully automated and error-free binning. Intragel size variation, which did not seem to pose a significant problem, is not affected by this external adjustment. Significantly, application of external adjustment made it possible to use two different models of sequencers to generate data that could then be pooled. This is important, as many laboratories now have both 373s and 377s in operation. Thus, by using external adjustment, one panel of markers can be typed at any one time regardless of the model or number of sequencers in a laboratory, thereby expediting data generation. Furthermore, the typing of an invariant sample makes it possible to potentially combine data between two or more independent genotyping studies. This is critical for complex diseases where there may not be enough power in any one study to detect weak gene effects.

It might be possible to increase the number of size standard fragments in each lane to enhance precision in sizing. However, we have found that the methods adopted here are adequate for most purposes, as shown by our low error rates. Another alternative would be to use microsatellite size standards. This may reduce sizing variation because ambient conditions will now affect migration of alleles and size standards equally. To date, such size standards have not become widely available.

Accurate binning of alleles is critical prior to statistical analysis. For dinucleotide markers, where alleles normally differ by multiples of 2 bp, ideal bins should have small ranges (<0.8 bp) and large interbin distances (>1 bp). With alleles 1 bp apart, there can be a potential problem with overlapping bins. Because we have observed very few cases where overlapping bins are a problem after external adjustment, the algorithm presented here is sufficient for simultaneous large-scale binning.

Our overall error rates have improved with fuller automation and with increasing experience. Interestingly, none of these errors were with markers exhibiting 1 bp allele systems. Recently, we have been maintaining an allele-specific error rate of 0.017% (2/11,692) or a genotypic-specific error rate of 0.034% (2/5846) for the last 64 markers typed in the laboratory across two different model sequencers. It is possible that these are underestimates, as most of our sample is composed of sib-pairs, whereas the duplicate samples are largely from extended families. However, it is unlikely that these estimates are very different from the true values. Extended families provide more information to assign correct genotypes, especially in cases where electropherogram profiles contain constantly sized noise peaks that may overlap with true alleles.

We have estimated that the overall costs of genotyping and marker optimization using fluorescent technology in our laboratory are $2.28 per genotype. This estimate includes costs of reagents, laboratory equipment, salaries, and overhead. However, based on the high quality and reproducibility of the data, we feel that this cost is worthwhile.

In summary, we have shown that with some modifications and the application of simple programs, it is possible to generate high-quality data for subsequent statistical analysis using fluorescent genotyping technology, dinucleotide microsatellite markers, and a largely sib-pair sample set. Our methods complement the ABI software programs and are particularly suitable for large-scale genotyping projects that employ different types of model sequencers. Together with other recent advances (Perlin et al. 1995; Smith et al. 1995; Brownstein et al. 1996; Magnuson et al. 1996), these developments bring us several steps closer to the goal of fully automatable, accurate, high-throughput genotyping of microsatellite markers.

METHODS

DNA Isolation

CEPH family 884 DNA was purchased from Bios laboratories (New Haven, CT). The CEPH 134702 DNA was purchased from Coriell Institute for Medical Research (Camden, NJ). All other DNA samples were isolated from 30 ml of whole blood using a salting-out procedure (GENTRA DNA Isolation Kit, Minneapolis, MN). Each sample was diluted to 10 ng/µl for amplification before being frozen in 96-well deep plates. Prior to PCR, DNAs from the deep-well plates were aliquoted into 96-well MicroAmp plates (Perkin Elmer, Norwalk, CT) using a HYDRA 96 (Robbins Scientific, Sunnyvale, CA) microdispenser.

Dinucleotide Repeat Markers

The forward strand of each primer pair was labeled with one of three phosphoramidites: 6-FAM (blue), HEX (yellow), and TET (green). Two strategies were developed chronologically to deal with the allele-plus-A problem described above (Smith et al. 1995; Magnuson, et al. 1996). The first strategy involved the use of PCR cycling conditions to attempt to control whether predominately allele or allele-plus-A PCR products were generated. Primers for markers D1S220, D1S235, D1S255, D1S249, D1S197, D1S206, D1S484, D1S196, D1S213, D1S234, D1S199, D1S238, D1S252, D1S413, D1S425, D1S207, D1S209, D1S218, D1S468, D1S498, D7S484, D7S510, D7S513, D7S516, D7S517, D7S530, D7S550, D7S640, D7S657, D7S669, D8S258, D8S260, D8S272, D8S504, D8S514, D7S507, D8S550, D7S515, D8S283, D7S493, D8S270, D8S284, D8S285, and D7S531 were from ABI PRISM Linkage Mapping Set, panels 1, 2, 11, and 12. These markers were assayed under two-step, or three-step + 90-min extension at 72°C PCR protocols described by Smith et al. (1995) to drive PCR reactions for a given marker to either allele, or allele-plus-A, PCR products, respectively.

The second strategy of reverse primer modification was adopted for all subsequent assays to drive all reactions to allele-plus-A PCR products (Magnuson et al. 1996). For the following markers from ABI PRISM Linkage Mapping Set, panels 3, 4, 18, 21, and 22 or those selected from the literature, all reverse primers were modified by substituting a G for the 5′-end nucleotide [except where A was substituted (*)]: D1S228, D1S198, D1S418, D2S362, IRS1, D2S113, D2S367, D2S326, D2S206, D2S117, D2S142, D2S126, D2S336, D2S325, D2S164, D2S286, D2S162, D2S121, D2S152, D2S305, D2S139, D2S319, D2S168, D2S151, D2S368, D2S383, D2S165, D2S337, D8S276, D12S1718, D12S76, D12S1349, D12S306, D12S105, D12S1679, D12S357, D12S367, D12S351, D12S78, D12S324, D12S79, D12S86, D16S405, D16S401*, D16S411, D15S130, D16S520*, D15S165, D15S131, D16S503, D16S511*, D15S117, D15S205, D16S415, D16S515, D16S423*, D15S123, D1S508, D16S403, D15S120, D15S128, and D16S516. All markers were assayed under the three-step + 10-min extension at 72°C protocol as described below.

PCR Amplification and Pooling

All PCR reactions were carried out using 60 ng of template DNA (10 ng/µl) in a 15-µl total reaction volume. All reactions were optimized at a final $MgCl_2$ concentration of 1.5 mM in Perkin Elmer PCR buffer II [10 mM Tris-HCl (pH 8.3) and 50 mM KCl]. Each reaction contained 333 nM each of forward and reverse primer, 250 µM of each dNTP (dATP, dCTP, dGTP, dTTP) (Pharmacia Biotech, Piscataway, NJ), and 0.6 units of AmpliTaq DNA polymerase (Perkin Elmer, Norwalk, CT) (5 U/µl). The primer-specific PCR mixes were aliquoted into the appropriate MicroAmp plates using a MultiPROBE 204DT Robotic Liquid Handling System (Packard Instrument Company, Downers Grove, IL) in Robbins Scientific (Sunnyvale, CA) tubes and thermocycled on a Perkin Elmer (Norwalk, CT) 9600 GeneAmp Thermocycler. Reactions for panels 1, 2, 11, and 12 were optimized to thermocycle under one of two PCR protocols: (1) two-step PCR, 95°C for 5 min, followed by 10 cycles of 94°C for 15 sec and 55°C for 15 sec, followed by an additional 23 cycles of 89°C for 15 sec and 55°C for 15 sec; or (2) three-step PCR + 90 min extension at 72°C, 95°C for 5 min, followed by 10 cycles of 94°C for 15 sec, 55°C for 15 sec, and 72°C for 30 sec, followed by an additional 20 cycles of 89°C for 15 sec, 55°C for 15 sec, and 72°C for 30 sec, plus a final extension step of 72°C for 90 min. Reactions for panels 3, 4, 18, 21, and 22 were optimized to thermocycle under one PCR protocol, three-step PCR + 10 min extension at 72°C: 95°C for 5 min, followed by 10 cycles of 94°C for 15 sec, 55°C for 15 sec, and 72°C for 30 sec, followed by an additional 20 cycles of 89°C for 15 sec, 55°C for 15 sec, and 72°C for 30 sec, plus a final extension step of 72°C for 10 min.

Markers were pooled (typically, total volume 40 µl) using a HYDRA 96 (ideally not the same one used for DNA aliquoting to avoid cross-contamination) adjusting individual marker volumes (range = 1.0–8.0 µl) to give electropherogram peak intensities of 1000–2000 fluorescence units. Pooled products were from a panel of markers amplified from a single DNA template. Next, 1.5 µl of the PCR pools was added to 3.5 µl of loading dye cocktail. The dye cocktail comprised a mixture of formamide, blue dextran loading buffer, and DNA standard (GS-500, ABI Divisions/PerkinElmer Foster City, CA)

at the ratio of 5:1:1, respectively. The GS-500 size standard was TAMRA-labeled (red) with size range from 35 to 500 bp.

Gel Electrophoresis

373 Sequencer

After heat denaturation, each pooled product/dye mix (3.5 µl) consisting of all markers was loaded onto a single lane of the gel. Samples were electrophoresed at 15 W (constant) in a 0.3-mm-thick 7% polyacrylamide gel (Bio-Rad, Hercules, CA) using 12-cm well-to-read (wtr) glass plates. The gel was run (filter set B) for a maximum of 4 hr to allow the 350-bp TAMRA-labeled fragment to be detected, using GS ABI 672 data collection software version 1.1.

377 Sequencer

After heat denaturation and pre-run gel, 1.8 µl of microsatellite product/dye mix was electrophoresed in a 0.2-mm-thick 4.5% polyacrylamide gel using 36-cm wtr glass plates. The gel was run for a maximum of 3 hr to allow the 350-bp TAMRA-labeled fragment to be detected. Data were gathered using ABI PRISM 377 Collection version 1.1 software.

Analysis of PCR Products on 373 and 377 ABI Sequencers

Genotype data were analyzed using GS Analysis [v. 1.2.2-1 (373) and v. 2.02 (377)] and the local Southern sizing method. The sized microsatellite data were processed through GENOTYPER (v. 1.1r8).

Importing Data into a FOXPRO Table and ABAS

Genotype information was imported into a FOXPRO (v. 2.6, Microsoft, Redmond, WA) table. This table tracks information for each specific genotype. It contains the file name, raw alleles, processed alleles, adjusted size values, and final bin labels. This table is readily transferable into a specified format to data files for analysis in statistical genetics packages such as MENDEL (Lange et al. 1988). Please contact us at the addresses given to learn how to transfer a modified copy of the software ABAS, which is freely available.

Statistical Analysis

The SAS (v. 6.09, SAS Institute Incorporated, Cary, NC) statistical analysis package procedure, PROC TTEST, was used to perform tests of the equality of the means and variances of allele sizes for the 373 and 377 sequencers.

ACKNOWLEDGMENTS

We thank Dennis Gilbert, Adam Lowe, Janet Ziegle, and the ABI Division of Perkin Elmer for introducing us to the field of fluorescence-based genotyping. Programming assistance by Rajesh Mahadwar, Gunther Birznieks, and David Johns is greatly appreciated. We acknowledge the FUSION team in Helsinki for collecting the family material and performing the DNA extraction to make this study possible: J. Eriksson, K. Kohtamaki, L. Toivanen, J. Tuomilehto, E. Tuomilehto-Wolf, T. Valle, and G. Vidgren.

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

Boehnke, M. 1991. Allele frequency estimation from data on relatives. *Am. J. Hum. Genet.* **48:** 22–25.

Brownstein, M.J., J.D. Carpten, and J.R. Smith. 1996. Modulation of non-templated nucleotide addition by Taq DNA polymerase: Primer modifications that facilitate genotyping. *BioTechniques* **20:** 1004–1010.

Buetow, K.H. 1991. Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am. J. Hum. Genet.* **49:** 985–994.

Collins, F.S. 1995. Positional cloning: From perditional to traditional. *Nature Genet.* **9:** 347–350.

Cooperative Human Linkage Center (CHLC). 1994. A comprehensive human linkage map with centimorgan density. *Science* **265:** 2049–2054.

Davies, J.L., Y. Kawaguchi, S.T. Bennett, J.B. Copeman, H.J. Cordell, L.E. Pritchard, P.W. Reed, S.C.L. Gough, S.C. Jenkins, S.M. Palmer, K.M. Balfour, B.R. 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.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach. 1996. A comprehensive genetic map of genetic map of the human genome based on 5,264 microsatellite. *Nature* **380:** 152–154.

Dubovsky, J., V.C. Sheffield, G.M. Duyk, and J.L. Weber. 1995. Sets of tandem repeat polymorphisms for efficient linkage screening of the human genome. *Hum. Mol. Genet.* **4:** 449–452.

Gastier, J.M., J.C. Pulido, S. Sunden, T. Brody, K.H. Buetow, J.C. Murray, T.J. Hudson, V.C. Sheffield, and G.M. Duyk. 1995. Survey of trinucleotide repeats in the human genome: Assessment of their utility as genetic markers. *Hum. Mol. Genet.* **4:** 1829–1836.

Ghosh, S. and F.S. Collins. 1996. The geneticist's approach to complex disease. *Annu. Rev. Med.* **47:** 333–353.

Gyapay, G., J. Morissette, A. Vignal, C. Dib, C. Fizames, P. Millasseau, S. Marc, G. Barnardi, M. Lathrop, and J. Weissenbach. 1994. The 1993–94 Genethon human genetic linkage map. *Nature Genet.* **7:** 246–249.

Hauser, E.R., M. Boehnke, S.-W. Guo, and N. Risch. 1996. Affected-sib-pair interval mapping and exclusion for complex genetic traits: Sampling considerations. *Genet. Epidemiol.* **13:** 117–137.

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

Lange, K., D. Weeks, and M. Boehnke. 1988. Programs for pedigree analysis: MENDEL, FISHER and dGENE. *Genet. Epidemiology* **5:** 471–472.

Magnuson V.L., D.S. Ally, S.J. Nylund, Z.E. Karanjawala, J.B. Rayman, J.I. Knapp, A.L. Lowe, S. Ghosh, and F.S. Collins. 1996. Substrate nucleotide-determined non-templated addition of adenine by Taq DNA polymerase: Implications for PCR-based genotyping and cloning. *BioTechniques* **21:** 700–709.

Mansfield, D.C., A.F. Brown, D.K. Green, A.D. Carothers, S.W. Morris, H.J. Evans, and A.F. Wright. 1994. Automation of genetic linkage analysis using fluorescent microsatellite markers. *Genomics* **24:** 225–233.

Maynard, P.E., K.P. Corcoran, J.S. Ziegle, J.M. Robertson, L.B. Hoff, and M.N. Kronick. 1992. The use of fluorescence detection and internal lane standards to size PCR products automatically. *Appl. Theor. Electrophoresis* **3:** 1–11.

Perlin, M.W., G. Lancia, and S.-K. Ng. 1995. Toward fully automated genotyping: genotyping microsatellite markers by deconvolution (GMBD). *Am. J. Hum. Genet.* **57:** 1199–1210.

Reed, P.W., J.L. Davies, J.B. Copeman, S.T. Bennett, S.M. Palmer, L.E. Pritchard, S.C.L. Gough, Y. Kawaguchi, H.J. Cordell, K.M. 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.

Schwengel, D.A., A.E. Jedlicka, E.J. Nanthakumar, J.L. Weber, and R.C. Levitt. 1994. Comparison of fluorescence-based semi-automated genotyping of multiple microsatellite loci with autoradiographic techniques. *Genomics* **22:** 46–54.

Sheffield, V.C., J.L. Weber, K.H. Buetow, J.C. Murray, D.A. Even, K. Wiles, J.M. Gastier, J.C. Pulido, C. Yandava, S.L. Sunden, G. Mattes, T. Businga, A. McClain, J. Beck, T. Scherpier, J. Gilliam, J. Zhong, and G.M. Duyk. 1995. A collection of tri- and tetranucleotide repeat markers used to generate high quality, high resolution human genome-wide linkage maps. *Hum. Mol. Genet.* **4:** 1837–1844.

Smith, J.R., J.D. Carpten, M.J. Brownstein, S. Ghosh, V.L. Magnuson, D.A. Gilbert, J.M. Trent, and F.S. Collins. 1995. Approach to genotyping errors caused by nontemplated nucleotide addition by Taq DNA polymerase. *Genet. Res.* **5:** 312–317.

Southern, E.M. 1979. Measurement of DNA length by gel electrophoresis. *Anal. Biochem.* **100:** 319–323.

The Utah Marker Development Group. 1995. A collection of ordered tetranucleotide-repeat markers from the human genome. *Am. J. Hum. Genet.* **57:** 619–628.

Ziegle, J.S., Y. Su, K.P. Corcoran, L. Nie, P.E. Mayrand, L.B. Hoff, L.J. McBride, M.N. Kronick, and S.R. Diehl. 1992. Application of automated DNA sizing technology for genotyping microsatellite loci. *Genomics* **4:** 1026–1031.

Received July 29, 1996; accepted in revised form January 8, 1997.

Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. FUSION (Finland-U.S. Investigation of NIDDM Genetics) Study Group.

S Ghosh, Z E Karanjawala, E R Hauser, et al.

Genome Res. 1997 7: 165-178
Access the most recent version at doi:10.1101/gr.7.2.165