7

Genotyping for Human Whole-Genome Scans: Past, Present, and Future

James L. Weber[1]
Center for Medical Genetics
Marshfield Medical Research Foundation
Marshfield, Wisconsin 54449

Karl W. Broman
Department of Biostatistics
School of Hygiene and Public Health
Johns Hopkins University
Baltimore, Maryland 21205

I. SUMMARY

Efficient and effective whole-genome 10-cM short tandem repeat polymorphism (STRP) scans are now available. Doubling or tripling STRP density to an average spacing of 3–5 cM is readily achievable. However, if typing costs for diallelic polymorphisms can be brought close to, or preferably less than, one-third those of STRPs, then diallelics may gradually supplement or supplant STRPs in whole-genome scans. The power of higher density genome scans for gene map-

[1]To whom correspondence should be addressed.

0065-2660/01 $35.00

ping by association and for many other research and clinical applications is great. It would be wise to continue investing heavily for many years in genotyping technology.

II. INTRODUCTION: GENOTYPING PAST

In their landmark paper in 1980, Botstein, White, Skolnick, and Davis outlined the use of restriction fragment length polymorphisms (RFLPs) to map disease genes through linkage analysis (Botstein *et al.*, 1980). The breakthrough achieved by these authors was the concept that highly abundant DNA polymorphisms as opposed to protein polymorphisms or other phenotype-based markers could be utilized for whole-genome scans. Throughout the 1980s, hundreds of RFLPs were identified and combined into whole-genome linkage maps. Several important disease genes were mapped, including those for Duchenne muscular dystrophy, Huntington's disease, and cystic fibrosis (Gusella, 1986). Unfortunately, RFLPs were largely diallelic and therefore low in informativeness. Also, the methods required for analysis of RFLPs were relatively complicated and inefficient. Analysis involved digestion of genomic DNA with one or more restriction enzymes, separation of the resulting DNA fragments by size through electrophoresis on agarose gels, Southern blotting of the DNA fragments to membranes, and detection of specific DNA fragments on the membranes by hybridization to highly radioactive, cloned DNA probes.

In 1989 a new type of abundant, multiallelic DNA polymorphism, the short tandem repeat polymorphism (STRP) (also called microsatellite or simple sequence length polymorphism) was reported (Weber and May, 1989). STRPs are based on variations in the numbers of tandem repeats in relatively short (usually < 60 bp) runs of primarily mono-, di-, tri-, and tetranucleotide repeats. Many STRPs have heterozygosities in the range of 70–90%. Analysis of STRPs involved just two simple steps: PCR amplification of a short (70–400 bp) segment of genomic DNA, followed by sizing of the amplified fragment through electrophoresis on denaturing polyacrylamide gels. Because the PCR primers annealed to unique sequences flanking the runs of tandem repeats, each pair of primers was specific for a single locus in the genome. The only equipment required was a thermal cycler and electrophoresis apparatus. Since STRPs were more informative and easier to type than RFLPs, the former quickly supplanted the markers introduced earlier. Throughout the 1990s, about 10,000 human STRPs were identified and mapped. Linkage mapping successes for disease genes with STRPs were quickly achieved. Many hundreds of disease genes have since been mapped with the use of these markers.

III. GENOTYPING PRESENT

Today, mapping genes for monogenic disorders using STRPs is routine. When sufficient family material is available, a single experienced lab worker can map a monogenic disorder in less than a month—in optimal cases, over a weekend. However, for genetically more complex disorders, at least one to two orders of magnitude more DNA samples may be required for linkage mapping success. Typing STRPs on such a large scale has motivated the growth of large dedicated genotyping centers. Genotyping output at these centers has increased greatly over the last few years, and concomitantly genotyping costs have rapidly dropped. Table 7.1, for example, presents 1990s Marshfield output for 400-marker STRP scans.

Most of the whole-genome polymorphism scans carried out at Marshfield are supported by the National Heart, Lung, and Blood Institute (NHLBI) Mammalian Genotyping Service. Genotyping is offered for all types of disorders, not just those involving the heart, lung, or blood. Genotyping through the service is free; however, brief applications must be submitted which are subject to peer review and NHLBI staff evaluation. Capacity of the Mammalian Genotyping Service is currently about 5.5 million genotypes per year and is steadily increasing. The service is funded through September 2006. More information can be obtained from www.marshmed.org/genetics. The

Table 7.1. Marshfield Genotyping Output

Year	DNA samples with 400-marker genome scans	Total cost per genome scan[a]
1993	350[b]	$1,200
1994	674[b]	$920
1995	2,150[b]	$600
1996	3,600	$428
1997	7,700	$272
1998	11,400	$192
1999	14,200	$160

[a]Total cost is comprehensive and includes salaries, supplies, equipment, overhead, and miscellaneous expenses.
[b]In 1993–1995 much of the lab's genotyping was with CEPH families instead of for disease gene mapping. Therefore, for comparison purposes, total genotypes were divided by 400 to obtain equivalent numbers of DNA samples scanned.

Center for Inherited Disease Research (CIDR), an intramural program of the National Institutes of Health, offers a similar genotyping service (www.cidr.jhmi.edu).

A. Marker screening sets

Typically, human whole-genome polymorphism scans involve 350–400 STRPs with average sex–equal spacing of about 10 cM. Lower density screens are occasionally carried out, particularly for monogenic disorders. Since about 10,000 human STRPs have been identified and many more can now be easily developed from the human genomic sequence, the selection of a small subset of markers for the whole-genome scans is an important issue. STRPs differ greatly in quality. They vary widely in informativeness, amplification efficiency, and the ease by which the alleles can consistently be called (see also later). At Marshfield, we are currently putting the finishing touches on the tenth version of our whole-genome STRP screening set (see www.marshmed.org/genetics). Average marker heterozygosity in the Marshfield screening set is about 76%. The Marshfield set is comprised primarily of tri- and tetranucleotide STRPs, with dinucleotide STRPs only used at positions along the genetic map where a high-quality tri- or tetranucleotide STRP could not yet be found. Other labs utilize screening sets based primarily or exclusively on dinucleotide repeat STRPs (see, e.g., Reed *et al.*, 1994; www2.perkin–elmer.com/ab).

Marker spacing in the whole-genome screening sets is not uniform. An example of the marker spacing for chromosome 2 in our Marshfield Screening Set 10 is shown in Table 7.2. Because human linkage maps are based upon the typing of relatively few meioses in the CEPH families (Broman *et al.*, 1998), the estimated map distances have quite limited precision. There is also growing evidence that recombination rates along chromosomes differ among individuals (Yu *et al.*, 1996; Broman *et al.*, 1998). Therefore, although statistical geneticists often assume equal marker spacing in their simulations and theoretical work, in reality, screening set marker spacing will never be perfectly uniform and probably will always have a fair degree of uncertainty owing to individual differences in recombination patterns.

B. Genotyping quality

Clearly, polymorphism genotypes of relatively high quality are essential for successful completion of gene mapping projects. A summary of genotyping quality for large genotyping projects (> 700 samples) completed at Marshfield in 1998–1999 is shown in Table 7.3. Average genotyping completeness, after correction for samples that amplify poorly under our standard PCR conditions, was

Table 7.2. Marshfield Chromosome 2 Screening Set (from Set 10)

Locus	Marker	Heterozygosity	Map position (cM)[a]	Marker spacing (cM)
TPO	SRA	0.64	0	0
D2S1780	GATA72G11	0.71	10	10
D2S2952	GATA116B01	0.77	18	8
D2S1400	GGAA20G10	0.67	28	10
D2S1360	GATA11H10	0.82	38	10
D2S405	GATA8F07	0.67	48	10
D2S1788	GATA86E02	0.87	56	8
D2S1356	ATA4F03	0.76	64	8
D2S1352	ATA27D04	0.67	74	10
D2S441	GATA8F03	0.74	87	13
D2S1394	GATA69E12	0.71	91	4
D2S1790	GATA88G05	0.78	103	12
D2S2972	GATA176C01	0.73	114	11
D2S410	GATA4E11	0.81	125	11
D2S1328	GATA27A12	0.75	133	8
D2S1334	GATA4D07	0.81	145	12
D2S1399	GGAA20G04	0.82	152	7
D2S1353	ATA27H09	0.81	165	13
D2S1776	GATA71D01	0.76	173	8
D2S1391	GATA65C03	0.74	186	13
D2S1384	GATA52A04	0.76	200	14
D2S2944	GATA30E06	0.79	210	10
D2S434	GATA4G12	0.76	216	6
D2S1363	GATA23D03	0.77	227	11
D2S427	GATA12H10	0.70	237	10
D2S2968	GATA178G09	0.61	252	15
D2S2986	2QTEL47	0.68	265	13

[a]Based on sex-averaged map.

97.2%. Completeness is dependent upon the genotyping process, but it is also highly dependent upon the quality of the DNA samples. It is unfortunate but true that many groups involved in gene mapping projects take great care in phenotyping and analysis but skimp on the issues of DNA extraction and handling. This is a major mistake because projects cannot be successful without high-quality, accurately labeled DNA. PCR will not be effective unless the DNA is pure and at the correct concentration in the correct solute. Analysis will be substantially weakened if significant numbers of DNA samples are mislabeled. Substantial DNA quality problems are encountered with roughly 20% of the projects undertaken by the Mammalian Genotyping Service.

Table 7.3. Marshfield Genotyping Quality

Project	Completion date	Number of DNA samples	Average genotyping completeness (%)[a]	Estimated genotyping error rate (%)[b]	Average marker heterozygosity (%)[c]
A	7/17/98	1049	98.5	0.4	76
B	9/18/98	893	97.1	0.7	77
C	10/2/98	841	96.5	1.0	74
D	11/11/98	780	96.9	0.7	79
E	12/14/98	705	96.6	1.0	77
F	3/5/99	734	97.0	0.6	77
G	4/23/99	728	97.1	0.5	74
H	6/18/99	833	98.1	0.5	76
I	7/22/99	1068	97.3	0.6	77

[a]Completeness was calculated after all samples that had amplified especially poorly under standard PCR conditions (< 75 % complete).
[b]Error rates were determined by blind, duplicate, or triplicate genotyping of CEPH family individuals on different gels.
[c]Heterozygosity calculations excluded sex chromosome polymorphisms.

Genotyping error rate at Marshfield has averaged about 0.7% (Table 7.3). Note that this is genotype and not allele error rate. Since one of the two alleles is correct for most incorrect genotypes, allele error rate is approximately 60% of the genotyping error rate. Genotyping accuracy is monitored by blindly typing CEPH family DNA samples in duplicate or triplicate along with the remainder of the DNA samples. Family and individual numbering schemes for these control samples are disguised to match those of the remaining samples. The duplicated or triplicated CEPH family DNA samples are loaded on different gels, as opposed to loading in adjacent lanes of the same gel, so that error rates determined using these CEPH family samples are near the worst-case scenario. Marshfield genotyping error rates have been confirmed by collaborating labs that send their own blinded, duplicate DNA samples.

Genotyping accuracy is substantially improved when family structure is used as a final check on the allele calls. Under ideal conditions, such as the CEPH families with large sibships, genotyping accuracy improves to about 99.8%. Accuracy in this case refers to the consistency of allele calling within a single family. This is of course perfectly acceptable for linkage analysis, but consistency across families, gels, and time is required for association studies. Consistency requires the use of standard DNA with known screening set marker genotypes. At Marshfield, for example, amplified DNA from two of the CEPH family parents (133101 and 133102) is loaded about six times on each 200-lane gel.

Genotyping accuracy is also dependent upon specific laboratory processes. We have found, for example, that accuracy drops for the two or three lanes at the very edges of the gels, where there is often substantial skewing of fragment mobility compared with the interior portions of the gels. Error rates for the outer lanes are typically two to three times those for interior lanes. Also we have determined that there is substantial difference in accuracy among different classes of STRPs. Dinucleotide and noninteger (see later) STRPs have higher error rates ($\leq 2\%$) than tri- and tetranucleotide STRPs. This is the reason for the emphasis on tri- and tetranucleotide markers in the Marshfield screening sets. Finally, it is important to note that the foregoing discussion applies to genotyping as carried out specifically at Marshfield. Genotyping centers using different processes, different markers, and different equipment will likely show at least modest variation in quality from the Marshfield results.

C. Genotyping cost

As shown in Table 7.1, STRP genotyping costs at Marshfield have dropped dramatically over the last few years. Current costs are about $150 per 400-marker whole-genome scan or $0.38 per genotype (one STRP typed on one DNA sample). Superior markers, more experienced personnel, and economies of scale have all played important roles in the cost reductions, but the greatest factor has been improvements in technology. Dedicated genotyping instruments, especially including high-capacity water bath thermal cyclers and multidye fluorescence-based scanning electrophoretic instruments, have been designed and built. Our largest thermal cycler has a capacity of 600 microtiter plates per day. Our scanning fluorescence detectors (SCAFUDs) utilize 200-lane gels, and nearly all gels are used for four separate runs. SCAFUD throughput is currently over 16,000 genotypes per day. Sophisticated software packages have been generated for allele calling, for genotype checking, and for data storage and management. Laboratory process improvements include amplification of three to six markers simultaneously and the introduction of robotics for semiautomated sample handling.

Table 7.4 breaks down the genotyping costs at Marshfield by the steps in the genotyping process. Administration costs include the handling and managing of the DNA samples. These costs are unlikely to change greatly regardless of the type of marker or the approach used for genotyping. The PCR amplification step of the operation consumes most of the laboratory supplies for genotyping. Plastic microtiter plates, thermostable DNA polymerase, and fluorescent dye-labeled PCR primers currently comprise the great majority of the supply costs. The electrophoresis step is often cited as a drawback of utilizing STRPs. The costs of running the gels are not as high as often imagined, however: we utilize 200-lane gels and three marker dyes per gel run (and in the future more

Table 7.4. Marshfield 1998 Genotyping Costs by Operation

Operation	Cost (%)
Administration	14
Amplification	32
Electrophoresis	25
Scoring	29

than four), and we reuse each of the gels four times. The scoring step in the operation involves the greatest amount of labor because genotypes called by the computer must be manually checked. Overall, STRP genotyping remains a labor-intensive process, with about half the total cost devoted to salaries and fringe benefits. Labor costs could potentially be reduced substantially by conversion to a genotyping system in which allele calling is completely automated.

Low genotyping costs are dependent upon use of optimized markers in the whole-genome scans. Use of strongly amplifying and easily scored polymorphisms improves genotyping efficiency as well as quality. Substantial efficiencies are gained through purchase (or synthesis) of large quantities of fluorescent dye–labeled PCR primers and through the establishment of combinations of markers that amplify well together. These efficiencies are possible only with screening set markers, which are used in many different genome scans. The cost of typing non-screening-set markers, as in fine-mapping in a specific chromosome region to confirm and/or extend initial linkage mapping results, is roughly twice the cost of typing standard screening set markers. These factors have substantial implications for two-stage linkage mapping strategies in which low-density whole-genome scans are followed by fine-mapping by means of nonoptimized markers.

Genotyping quality is tightly connected to genotyping cost. By altering the genotyping process, as in the extreme example of typing each marker in duplicate, genotyping accuracy could be improved substantially. However, this improvement would be accompanied by significantly increased costs. Conversely, if quality were relaxed, then genotyping costs could be reduced. Automated STRP allele calling at Marshfield is currently about 94% accurate. Tedious and expensive manual editing of the genotypes is required to bring the error rate down below 1%. Through changes and improvements in the software and/or modified laboratory processes, it may, at least for some markers, be possible to get the automated genotyping accuracy up to 99%.

Throughput in whole-genome scans is becoming large enough to permit researchers to contemplate genotyping entire human populations. DeCode Genetics, for example, has plans to complete genome scans on essentially all

residents of Iceland (www.decode.is). At about $150 per 400-marker whole-genome scan, genotyping costs are becoming a small fraction of the total cost of a linkage mapping project. Except for phenotypes such as height and weight, which are unusually inexpensive to obtain, the costs of contacting, visiting, and phenotyping family members and of analyzing the genotype and phenotype data, usually greatly exceed the costs of genotyping. The possible scales of gene mapping projects are therefore largely limited by the phenotyping and analysis costs. This conclusion does not of course apply to whole-genome association studies, in which marker densities will generally be much greater than 400 per genome.

D. Genotyping limitations

Several difficulties with STRP genotyping affect the quality of the genotyping data and considerations for future progress in whole-genome scans. These include PCR artifacts such as strand slippage and weak/null alleles as well as the practice of using gel electrophoretic mobility to approximate true allele sequence. The problem of weak/null alleles also generally applies to typing of diallelic polymorphisms. Other limitations, including some not currently recognized, will undoubtedly plague any typing system for any class of polymorphisms.

Strand slippage (also called stuttering), an artifact seen in PCR with short tandem repeats, results in skipping of repeats during amplification and production of DNA fragments smaller in size than the original genomic fragment (see Figure 7.1). Strand slippage is highly dependent upon the repeat length. For mononucleotide repeats, strand slippage is so severe that despite the great abundance of these sequences in the human genome, they are only rarely used as polymorphic markers. For dinucleotides, strand slippage is manageable, and these markers can be scored accurately. However, in our many years of experience we have found that dinucleotide repeats are more difficult to score accurately than markers with higher repeat lengths. For trinucleotide and higher repeat lengths, strand slippage is minimal and is rarely a factor in genotyping. Despite considerable effort, no one has been able to devise a solution for strand slippage during PCR.

Weak or null alleles may occur in PCR when a second polymorphism occurs within one (or conceivably both) of the PCR primer annealing sites (see, e.g., Callen *et al.*, 1993). If the primer/template mismatch occurs near the 5′ end of the PCR primer, the effect may be only modest and the intensity of an allele with the mismatch may just be relatively weak compared to the other alleles. However, when the mismatch occurs near the 3′ end of primer, PCR can be disrupted entirely and only one of two alleles may be amplified, resulting in the scoring of the individual as a pseudo-homozygote. Whether a specific allele

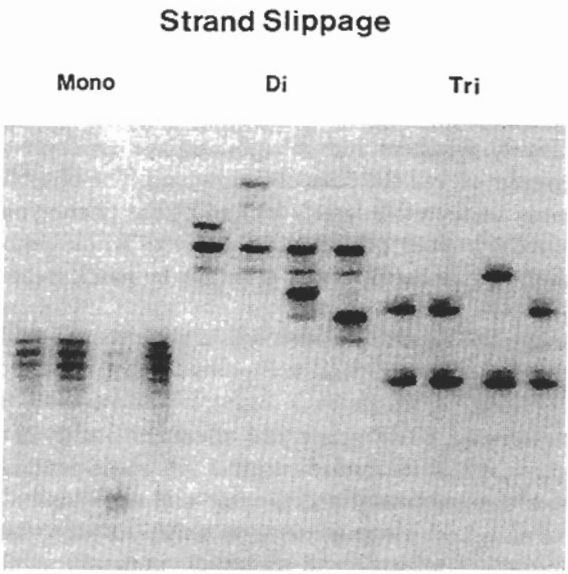

Figure 7.1. Electrophoretic profiles of amplified DNA from four unrelated individuals for each of three STRPs with mono, di, and trinucleotide repeats. Note that the relative amount of strand slippage decreases dramatically as the repeat length increases.

will be weak or null depends strongly on the PCR conditions. Detection of null alleles usually requires analysis of families rather than unrelated individuals. In nearly all cases, it should be possible to avoid weak/null alleles by shifting the offending PCR primer. Markers with frequent weak/null alleles are usually excluded from standard screening sets. Uncommonly large size differences between alleles may result in a relatively weak amplification for the longest alleles, but this effect is usually modest in comparison to that of mismatches between PCR primer and template.

Sizing of PCR products on denaturing acrylamide gels also limits STRP genotyping. STRP allele calling is nearly always based on the *mobility* of the amplified DNA fragment on the gels. Generally, gel mobility is a reasonably good indicator of the length of the PCR product and, therefore, of the numbers of tandem repeats within the allele. However, a number of situations exist in which mobility only approximates the true allele sequence (see, e.g., Primmer and Ellegren, 1998; Bergström *et al.*, 1999). This phenomenon is sometimes called homoplasy. Figure 7.2 shows several hypothetical examples of different alleles that will have indistinguishable mobilities on the gels and, therefore, will

Homoplasy

ATGGCACCTTTT	**$(AC)_{18}$**	**GGGCAATTGCTC**
ATGGCACCTTTT	**$(AC)_9(TC)(AC)_8$**	**GGGCAATTGCTC**
ATGGCACCTTTT	**$(AC)_{10}(AG)_8$**	**GGGCAATTGCTC**
ATGGCACCTT	**$(AC)_{19}$**	**GGGCAATTGCTC**

Figure 7.2. Sequences of hypothetical alleles all with the exact same nucleotide length and with likely indistinguishable mobilities on denaturing polyacrylamide gels. Note that in the last example the length difference is outside of the repeats.

all be assigned the same allele size. Note that in addition to imperfections in the array of tandem repeats, or even the presence of two or more different types of repeats, the insertion/deletion can lie outside the tandemly repeated region (Grimaldi and Crouau-Roy, 1997; Colson and Goldstein, 1999). For linkage studies in relatively small living families, homoplasy is unlikely to be a major factor. However, in association studies, homoplasy can be a major confounder, in as much as two alleles with exactly the same size and gel mobility can have very different ancestral histories.

Incomplete electrophoretic resolution of PCR products differing in length for various reasons (e.g., the use of short gels; long PCR products) can also limit STRP genotyping. For many STRPs, alleles differ in size only by integer multiples of the repeat length. So, for example, a tetranucleotide STRP might have alleles ranging from 8 to 14 full tandem repeats. However, at appreciable frequency, STRPs will also exhibit alleles that differ in size by other than integer multiples of the repeat lengths (Brinkman *et al.*, 1998). These "noninteger" alleles can be difficult to score consistently because they often differ in size from other alleles by only a single nucleotide. Noninteger alleles are known for di-, tri-, and tetranucleotide repeat markers but are most commonly recognized for the tetranucleotide repeat markers. The fraction of STRPs that have common noninteger alleles is uncertain. For the great majority of STRPs, only one or a few alleles have been sequenced. As much as possible, we have excluded STRPs with common noninteger alleles from our screening sets. However, noninteger alleles still arise unexpectedly for human populations that have not been typed.

In summary, an ideal STRP for genomic screening would have the following properties. It would amplify strongly with little strand slippage and would produce sharp (as opposed to diffuse and fuzzy) bands upon gel electrophoresis. It would be highly informative and would produce few if any nonin-

teger alleles. Accurate scoring by computers, without manual checking, would be possible. If STRP genotyping continues to be cost-competitive, then, eventually, each marker within a genome screen will have many alleles sequenced and will become at least reasonably close to this ideal.

In addition to genotyping limitations, sample labeling errors and pedigree structural errors detract substantially from the quality of gene mapping. In the Mammalian Genotyping Service, the rate of such pedigree structure or gender errors has ranged from near 0% in several projects, to an average of about 3% per project, to a high of 12% for one large project. In many cases, therefore, family structural errors substantially exceed genotyping errors. Although problematic individuals and families usually can be identified and excluded from subsequent analysis (see later), these problems still significantly reduce power. It cannot be emphasized strongly enough that for successful completion of a long-term, expensive, gene mapping project, careful recording of family structure and careful handling and labeling of the DNA samples are absolutely vital.

E. Error detection and effects

Prior to the analysis of data from a genome scan, pedigree and genotyping errors must be identified and resolved. Pedigree errors include sample mislabeling, errors in the entry of pedigree information into a computer database, nonpaternities, and unreported adoptions or twinning. Genotyping errors occur when observed genotypes do not correspond to the true underlying genetic information, as a result of a mistake in data entry or the misinterpretation of a pattern on a gel. A mutation at a marker locus may mimic a genotyping error; it is just as important to identify and resolve mutations as errors.

The effects of genotyping errors on the estimation of genetic maps have been well characterized (Buetow, 1991; Lincoln and Lander, 1992). Errors generally introduce apparent recombination events and thus lead to an expansion in the estimated maps. When the markers are more tightly spaced, the relative effect of errors greatly increases. The effects of genotyping errors on the power to detect disease susceptibility genes in linkage studies is not well understood. It is clear that power will generally decrease, though the extent of the effect has not been quantified. Analytic methods that use multipoint marker information will be more greatly affected by genotyping errors, in comparison to methods that make use of data on a single marker at a time. The effects of errors in pedigree information are also not well understood. A study on the effect of ignoring the relationship between two parents when they are first cousins (Mérette and Ott, 1996) showed that the effect of such errors may be considerable. More work clearly needs to be done to evaluate the effects of errors in genotypes and pedigree information (see, e.g., Rao, 1998).

While the detection and resolution of genotyping and pedigree errors are not yet completely automated, several computer programs are available to assist in the identification of errors. While the process is often tedious, if it is performed carefully, and preferably with the involvement of both the lab generating the data and the individuals ultimately responsible for data analysis, the resulting, more refined data set will have maximal power to detect disease genes.

The process of cleaning genotype data properly begins with the identification of pedigree errors. In many cases pedigree problems may be easily seen in checks for Mendelian inheritance. When parental data are missing, a more sophisticated approach may be necessary. Several computer programs are now available for verifying all pairwise relationships in a study (e.g., Boehnke and Cox, 1997; Göring and Ott, 1997; Broman and Weber, 1998). The relationship of each pair of individuals in a study, is inferred from their entire set of genotype data and then compared with the reported relationship. In most cases it is sufficient to consider the five relationships monozygotic twins, parent–offspring, full sibs, half-sibs, and other relationship or unrelated. One advantage to this type of approach is that the correct pedigree structure is often made clear, whereas when pedigree errors are observed by looking for large numbers of Mendelian inconsistencies, it can be tricky to determine what change in the pedigree structure will eliminate the problem. The approach of Boehnke and Cox (1997), implemented in the computer program RELPAIR, is especially valuable because it takes account of the known linkage relationship between the genetic markers and yet is very fast. Such a program should be used on the raw genetic data, prior to the resolution of any apparent genotyping errors, since apparently erroneous genotypes may provide important information about the relationship between individuals.

The detection of genotyping errors begins with the identification of genotypes that are inconsistent with Mendel's rules. One then determines the individual or individuals responsible for the problem. Generally one seeks the most parsimonious explanation for the problem, finding the fewest genotypes which must be removed to eliminate the inconsistency. Allele frequency information may be used to obtain probabilistic statements on which genotype is most likely in error. Several computer programs have been written that assist in the process of identifying and resolving Mendelian inconsistencies in genotype data. The programs PEDCHECK (O'Connell and Weeks, 1998) and a module in the Mendel package (Stringham and Boehnke, 1996) are especially good examples.

Ideally, one would go beyond such one-marker-at-a-time checks for genotyping errors, looking further for unlikely multiple recombination events that may indicate the presence of genotyping errors (Broman *et al.*, 1998). Unfortunately, the typical density at which most genome scans are performed

makes this a largely useless effort, since a double recombinant within 30 or 40 cM cannot be immediately ascribed to a genotyping error. In the future, if the use of diallelic markers becomes common, efforts to detect genotyping errors by looking for unlikely multiple recombination events will become much more important, since with fewer polymorphic markers, which have fewer possible genotypes, erroneous genotypes will be more likely to conform to Mendel's rules. For example, if both parents are heterozygous at a diallelic marker, their children may have any of the three possible genotypes, and so checks for Mendelian inheritance may fail to disclose any errors.

IV. GENOTYPING FUTURE

Human whole-genome polymorphism scans have important clinical in addition to research applications. These scans can be used to detect chromosomal aneuploidies and segmental aneusomies (Gusella, 1986; Rosenberg *et al.*, 2000). They can be used to propagate genetic information, such as the presence of a mutant gene through kindreds (Weber, 1994). They can be used to confirm putative biological relationships in patient families, and, if marker densities become high enough, they can be used to identify autozygous regions in individual patients (Broman and Weber, 1999) and to suggest from the presence of specific haplotypes which mutations an individual is likely to carry. In the long run, it may well turn out that clinical needs for the scans will outweigh needs for research applications. Perhaps someday, the whole issue of carrying out whole-genome polymorphism scans for research purposes will become irrelevant because these scans will have been routinely carried out on essentially all individuals for clinical purposes. Nevertheless, since this volume is devoted to the mapping of genes, our discussion focuses on this particular application.

For gene mapping, the two primary factors to consider for future genome scans are marker type and marker density. Several different types of polymorphisms with different properties and typing methodologies are potentially available. Scans with a broad range of average marker densities are similarly conceivable. Genotyping costs, of course, will have important bearing on both factors.

A. Marker type

Currently, the only types of polymorphisms that can be practically considered for whole-genome scans are diallelic base substitution or short insertion/deletion polymorphisms and multiallelic STRPs. Other types of polymorphisms such as minisatellites and complex chromosomal rearrangements such as large duplications and inversions are not sufficiently abundant or amenable to automation

to be used in genome scans. Informative STRPs (excluding mononucleotide repeats) occur on average roughly every 20 kb in the human genome (unpublished results). Together, diallelic base substitution and short insertion/deletion polymorphisms occur with reasonable informativeness about once every 1.0 kb (Cargill *et al.*, 1999; Halushka *et al.*, 1999). Base substitution polymorphisms appear to be approximately 10 times more abundant than the insertion/deletion polymorphisms (Kwok *et al.*, 1996; Wang *et al.*, 1998).

STRP genotyping costs have dropped substantially (Table 7.1) and are likely to continue decreasing in coming years. PCR reaction volumes are being reduced to minimize supply expenses. The number of fluorescent dyes that can be simultaneously detected in gel electrophoresis is steadily increasing. Capillaries and/or thinner slab gels may speed the time required for electrophoresis. Software for allele calling is steadily being improved, as is the quality of markers within the screening sets. Automated loading of electrophoresis platforms is being introduced into the process. All these and other improvements indicate that the genotype cost for STRPs will continue to fall. Currently these costs are at $0.40 per genotype. It is likely that over the next 3–5 years the cost can be brought down to about $0.25 per genotype. Nevertheless, there is no technology on the horizon that would permit the cost of STRP genotyping to decrease to a penny or less per genotype. These very low costs may, however, be achievable with diallelic polymorphisms.

If for gene mapping, the 1980s was the decade of RFLPs and the 1990s the decade of STRPs, then perhaps the first decade of the twenty-first century may belong to diallelic polymorphisms. Many have speculated that because these markers can be analyzed without gel electrophoresis, typing costs will be dramatically reduced. Although no one has yet achieved this feat, many groups in both the public and private sectors are devoting substantial resources to a wide spectrum of potentially promising approaches for typing diallelic polymorphisms. A number of promising closed-tube systems have recently been developed involving the Taqman assay (Livak *et al.*, 1995), molecular beacons (Tyagi *et al.*, 1998), the Invader assay (Lyamichev *et al.*, 1999), and thermal denaturation (Germer and Higuchi, 1999). These systems all have the attractive feature that samples are not handled after initial reaction setup. Approaches involving fluorescent microspheres (Fulton *et al.*, 1997; Michael *et al.*, 1998) and mass spectrometry (Ross *et al.*, 1998; Griffin *et al.*, 1999) have the potential to facilitate analysis of many polymorphisms simultaneously. Methods utilizing hybridization to dense microarrays of oligo probes or PCR products take advantage of the exceptional potential of miniaturization (Elango *et al.*, 1996; Wang *et al.*, 1998). In addition to these newer and more exotic approaches, a large group of effective though relatively expensive analysis methods currently exist (reviewed by Landegren *et al.*, 1998). These include direct sequencing of PCR products, restriction enzyme digestion of PCR prod-

ucts, single-stranded conformation polymorphism (SSCP) analysis, high-performance liquid chromatography, and allele-specific single-base polymerase extension of synthetic oligos.

That so many approaches are being explored for diallelic polymorphism analysis and so many dollars are being injected into this research bode well for the future. However, it is far too early to pick a winner or even a leader among the competing technologies. In addition, the big lead in efficiency enjoyed by STRP genotyping should not be discounted. Our many years of experience with genome scans have taught us that it is important to optimize each marker within screening sets. The cost required to optimize thousands of diallelic polymorphisms, even in highly efficient systems, will be large. Despite the long-term promise, arrival of efficient diallelic polymorphism systems for whole-genome scans may be farther away than most anticipate.

A number of groups have considered the question of how many lower informativeness diallelic polymorphisms will be required to match the information content of an average multiallelic STRP. Estimates have ranged from two to over five (Nickerson *et al.*, 1992; Kruglyak, 1997; Chapman and Wijsman, 1998). Answering this question requires consideration of many factors, including the average assumed informativeness of the markers, the type of family structure used for gene mapping, sample size, and genotyping error rates. STRPs show relatively little variation in informativeness among different populations. For example, in large studies carried out at Marshfield, average informativeness for various populations ranged from a low of 73% for a group of Native Americans to 80% for a collection of African Americans. Diallelics, in contrast, show much more population-specific variation in frequency (see, e.g., Gelernter *et al.*, 1999). Nearly all diallelics will be uninformative in a fraction of human populations. More theoretical work and also real gene mapping tests will be required to rigorously determine the relative information content of diallelics versus STRPs. However, a working estimate at this time is that three informative diallelics are required to match each informative STRP.

Besides informativeness and typing methods, STRPs and diallelic marker differ substantially in mutation rate. Most STRPs have mutation rates in the range of 10^{-3}–10^{-5} per gamete per generation (Weber and Wong, 1993; Brinkman *et al.*, 1998). In contrast, diallelic polymorphisms have much lower mutation rates on the order of 10^{-7}–10^{-9} per gamete per generation (Vogel and Motulsky, 1997). Mutation has little impact on linkage analysis when families with living members are tested, but the approximately 10,000-fold difference in mutation rates between STRPs and diallelics likely will have substantial impact on detection of linkage disequilibrium. Although disequilibrium can often be readily detected between closely linked STRPs and between STRPs and diallelic polymorphisms (Hurtley *et al.*, 1999; McPeek and Strahs, 1999),

shared haplotypes older than a few hundred years are usually affected by STRP mutation (Hästbacka *et al.*, 1992). Because STRPs mutate most often by the gain or loss of a single repeat, it may be possible to correct for mutation by considering windows of STRP alleles that differ by a single repeat (McPeek and Strahs, 1999). Nevertheless, the relatively high rate of STRP mutation will introduce complexity into disequilibrium analysis.

B. Marker density

For linkage analysis in families with living members, both for monogenic and more complex disorders, a 5- to 10- cM STRP scan appears satisfactory as a first step. Since gaps between markers in a 10-cM scan (see Table 7.2) can be fairly large, and an uninformative marker in such a gap can substantially decrease coverage of that portion of the genome, increasing STRP density to an average of one marker per 5 cM would be reasonable. However, for many other research and virtually all clinical applications of whole-genome scans, a much higher marker density would be preferred. The high-density polymorphism scan is analogous to a new more powerful telescope in astronomy. It will permit us to study such genetic phenomena as autozygosity, which were previously invisible (Broman and Weber, 1999).

In terms of gene mapping, high-density scans are particularly important in the detection of association. Most human genetic variation was established before humans migrated out of Africa, roughly 100,000 years ago (Hacia *et al.*, 1999). Disease alleles this old will tend to lie within quite short shared haplotypes in diverse, panmictic populations (Kruglyak, 1999). Detection of this ancient association may require marker densities up to 500,000 per genome, which is far beyond current technology. However, disease alleles that arose or were introduced into populations much more recently may be amenable to whole-genome association mapping. This approach has been very successfully applied to rare recessive diseases in isolated populations (Friedman *et al.*, 1995; Peltonen and Uusitalo, 1997; Sheffield *et al.*, 1998). Freimer, Sandkuijl, and colleagues have argued that densities as high as one marker every 3 cM may be sufficient for detection of association for complex disorders in isolated populations (Service *et al.*, 1999). Regardless of the validity of this last hypothesis, it is still clear that whole-genome polymorphism scans at marker densities of 1–3 cM (3000–1000 STRPs or equivalent) will find some useful application in gene mapping by association. These marker densities should be achievable within the next few years. It is also possible of course, through additional effort and cost, to type higher densities of polymorphisms in selected chromosomal regions—for example, as a follow-up to initial linkage mapping for complex disorders.

V. CONCLUSIONS

Efficient, reasonably effective whole-genome 10-cM STRP scans are now available. Doubling or tripling STRP density to an average spacing of 3–5 cM is readily achievable. However, if typing costs for diallelic polymorphisms can be brought close to or preferably below one-third of those of STRPs, then diallelics may gradually supplement or supplant STRPs in whole-genome scans. The power of higher density genome scans for gene mapping by association and for many other research and clinical applications is great. It would be wise to continue investing heavily for many years in genotyping technology.

References

Bergström, T. F., Engkvist, H., Erlandsson, R., Josefsson, A., Mack, S. J., Erlich, H. A., and Gyllensten, U. (1999). Tracing the origin of HLA-DRB1 alleles by microsatellite polymorphism. *Am. J. Hum. Genet.* **64,** 1709–1718.

Boehnke, M., and Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.* **61,** 423–429.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32,** 314–331.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation in human microsatellites: Influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* **62,** 1408–1415.

Broman, K. W., and Weber, J. L. (1998). Estimating pairwise relationships in the presence of genotyping errors. *Am. J. Hum. Genet.* **63,** 1563–1564

Broman, K. W., and Weber, J. L. (1999). Long homozygous chromosomal segments in the CEPH families. *Am. J. Hum. Genet.* **65,** 1493–1500.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., and Weber, J. L. (1998). Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63,** 861–869.

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

Callen, D. F., Thompson, A. D., Shen, Y., Phillips, H. A., Richards, R. I., Mulley, J. C., and Sutherland, G. R. (1993). Incidence and origin of "null" alleles in the $(AC)_n$ microsatellite markers. *Am. J. Hum. Genet.* **52,** 922–927.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G.Q., and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22,** 231–238.

Chapman, N. H., and Wijsman, E. M. (1998). Genome screens using linkage disequilibrium tests: Optimal marker characteristics and feasibility. *Am. J. Hum. Genet.* **63,** 1872–1885.

Colson, I., and Goldstein, D.B. (1999). Evidence for complex mutations at microsatellite loci in *Drosophila*. *Genetics 152,* 617–627.

Elango, R., Riba, L., Housman, D., and Hunter, K. (1996). Generation and mapping of *Mus spretus* strain-specific markers for rapid genomic scanning. *Mamm. Genome.* **7,** 340–343.

Friedman, T. B., Liang, Y., Weber, J. L., Hinnant, J. T., Barber, T. D., Winata, S., Arhya, I. N., and

Asher, J. H. Jr. (1995). A gene for congenital, recessive deafness *DFNB3* maps to the pericentromeric region of chromosome 17. *Nat. Genet.* **9,** 86–91.

Fulton, R. J., McDade, R. L., Smith, P. L., Kienker, L. J., and Kettman, J. R. Jr. (1997). Advanced multiplexed analysis with the FlowMetrix™ system. *Clin. Chem.* **43,** 1749–1756.

Gelernter, J., Cubells, J. F., Kidd, J. R., Pakstis, A. J., and Kidd, K. K. (1999). Population studies of polymorphisms of the serotonin transporter protein gene. *Am. J. Med. Genet.* **88,** 61–66.

Germer, S., and Higuchi, R. (1999). Single-tube genotyping without oligonucleotide probes. *Genome Res.* **9,** 72–78.

Göring, H. H. H., and Ott, J. (1997). Relationship estimation in affected sib pair analysis of late-onset diseases. *Eur. J. Hum. Genet.* **5,** 69–77.

Griffin, T. J., Hall, J. G., Prudent, J. R., and Smith, L. M. (1999). Direct genetic analysis by matrix-assisted laser desorption/ionization mass spectrometry. *Proc. Natl. Acad. Sci. USA* **96,** 6301–6306.

Grimaldi, M.-C., and Crouau-Roy, B. (1997). Microsatellite allelic homoplasy due to variable flanking sequences. *J. Mol. Evol.* **44,** 336–340.

Gusella, J. F. (1986). DNA polymorphism and human disease. *Annu. Rev. Biochem.* **55,** 831–854.

Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M., Brody, L. C., Wang, D., Lander, E. S., Lipshutz, R., Fodor, S. P. A., and Collins, F. S. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* **22,** 164–167.

Halushka, M. K., Fan, J.-B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22,** 239–247.

Håstbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., and Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. *Nat. Genet.* **2,** 204–211.

Huttley, G. A., Smith, M. W., Carrington, M., and O'Brien, S. J. (1999). A scan for linkage disequilibrium across the human genome. *Genetics* **152,** 1711–1722.

Kruglyak, L. (1997). The use of genetic map of biallelic markers in linkage studies. *Nat. Genet.* **17,** 21–24.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22,** 139–144.

Kwok, P.-Y., Deng, Q., Zakeri, H., Taylor, S. L., and Nickerson, D.A. (1996). Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. *Genomics* **31,** 123–126.

Landegren, U., Nilsson, M., and Kwok, P.-Y. (1998). Reading bits of genetic information: Methods for single-nucleotide polymorphism analysis. *Genome Res.* **8,** 769–776.

Lincoln, S. E., and Lander, E. S. (1992). Systematic detection of errors in genetic linkage data. *Genomics* **14,** 604–610.

Livak, K. J., Marmaro, J., and Todd, J. A. (1995). Towards fully automated genome-wide polymorphism screening. *Nat. Genet.* **9,** 341–342.

Lyamichev, V., Mast, A. L., Hall, J. G., Prudent, J. R., Kaiser, M. W., Takova, T., Kwiatkowski, R. W., Sander T. J., de Arruda, M., Acro, D. A., Neri, B. P., and Brow, M. A. (1999). Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat. Biotech.* **17,** 292–296.

McPeek, M. S., and Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* **65,** 858–875.

Mérette, C., and Ott, J. (1996). Estimating parental relationships in linkage analysis of recessive traits. *Am. J. Med. Genet.* **63,** 386–391.

Michael, K. L., Taylor, L. C., Schultz, S. L., and Walt, D. R. (1998). Randomly ordered addressable high-density optical sensor arrays. *Anal. Chem.* **70,** 1242–1248.

Nickerson, D. A., Whitehurst, C., Boysen, C., Charmley, P., Kaiser, R., and Hood, L. (1992). Identification of clusters of biallelic polymorphic sequence-tagged sites (pSTSs) that generate highly informative and automatable markers for genetic linkage mapping. *Genomics* **12,** 377–387.

O'Connell, J. R., and Weeks, D. E. (1998). PedCheck: A program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.* **6,** 259–266.

Peltonen, L., and Uusitalo, A. (1997). Rare disease genes—lessons and challenges. *Genome Res.* **7,** 765–767.

Primmer, C. R., and Ellegren, H. (1998). Patterns of molecular evolution in avian microsatellites. *Mol. Biol. Evol.* **15,** 997–1008.

Rao, D. C. (1998). CAT scans, PET scans, and genomic scans. *Genet. Epidemiol.* **15,** 1–18.

Reed, P. W., Davies, J. L., Copeman, J. B., Bennett, S. T., Palmer, S. M., Pritchard, L. E., Gough, S. C., Kawaguchi, Y., Cordell, H. J., Balfour, K. M., Jenkins, S. C., Powell, E. E., Vignal, A., and Todd, J. A. (1994). Chromosome-specific microsatellite sets for fluorescence-based, semi- automated genome mapping. *Nat. Genet.* **7,** 390–395.

Rosenberg, M. J., Vaske, D., Killoran, C. E., Ning, Y., Wargowski, D., Hudgins, L., Tifft, C. J., Meck, J., Blancato, J. K., Rosenbaum, K., Pauli, R. M., Weber, J., and Biesecker, L. G. (2000). The detection of chromosomal aberrations by a whole genome microsatellite screen. *Am. J. Hum. Genet.* **66,** 419–427.

Ross, P., Hall, L., Smirnov, I., and Haff, L. (1998). High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nat. Biotech.* **16,** 1347–1351.

Service, S. K., Temple Lang, D. W., Freimer, N. B., and Sandkuijl, L. A. (1999). Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am. J. Hum. Genet.* **64,** 1728–1738.

Sheffield, V. C., Stone, E. M., and Carmi, R. (1998). Use of isolated inbred human populations for identification of disease genes. *Trends Genet.* **14,** 391–396.

Stringham, H. M., and Boehnke, M. (1996). Identifying marker typing incompatibilities in linkage analysis. *Am. J. Hum. Genet.* **59,** 946–950.

Tyagi, S., Bratu, D. P., and Kramer, F. R. (1998). Multicolor molecular beacons for allele discrimination. *Nat. Biotech.* **16,** 49–53.

Vogel, F., and Motulsky, A.G. (1997). "Human Genetics: Problems and Approaches," 3rd ed. Springer-Verlag, Berlin.

Wang, D. G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L, Hsie, L, Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., and Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–1082.

Weber, J. L. (1994). Know thy genome. *Nat. Genet.* **7,** 343–344.

Weber, J. L., and May, P. M. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44,** 388–396.

Weber, J. L., and Wong, C. (1993). Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2,** 1123–1128.

Yu, J., Lazzeroni, L., Qin, J., Huang, M.-M., Navidi, W., Erlich, H., and Arnheim, N. (1996). Individual variation in recombination among human males. *Am. J. Hum. Genet.* **59,** 1186–1192.