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