

Genomic Mismatch Scanning Identifies Human Genomic DNA Shared Identical by Descent

Vivian G. Cheung* and Stanley F. Nelson†¹

*Department of Pediatrics and Neurology, Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, Pennsylvania 19104; and †Department of Pediatrics, Biological Chemistry and Psychiatry, University of California, Los Angeles, California 90095

Received July 8, 1997; accepted October 21, 1997

Genomic mismatch scanning (GMS) is a high-throughput, high-resolution identity by descent mapping technique that enriches for genomic DNA fragments that are shared between related individuals. In GMS, DNA heteroduplexes are formed from restriction-digested genomic DNA fragments from two relatives. Mismatch-free DNA heteroduplexes, likely representing DNA shared identical by descent between the two individuals, are relatively purified by depleting the mismatch-containing heteroduplexes using the *Escherichia coli* mismatch repair proteins and exonuclease. Here, we demonstrate using quantitative microsatellite genotyping that, despite the complexity of the human genome, GMS can enrich the majority of restriction fragments that are identical by descent between two related humans. As the entire genome is selected in GMS, an extraordinarily dense set of markers (up to 200,000 markers) may be screened in parallel. The demonstration of the molecular enrichment of identical DNA fragments in the context of the whole human genome establishes conditions for the application of GMS to human genetics. This forms a framework for the further development of GMS as a hybridization-based mapping technique that utilizes DNA microarray technology to map the selected identical by descent DNA fragments. © 1998 Academic Press

INTRODUCTION

Genomic mismatch scanning (GMS) is a hybridization-based linkage method designed to enrich DNA restriction fragments of identical sequence shared between two individuals (Nelson *et al.*, 1993). In GMS, the genomic DNA fragments that are shared identical by descent (IBD) are purified. Restriction-digested DNA from one individual is methylated and mixed with DNA from another individual that is digested with the same restriction enzyme but left unmethylated. The

DNA mixture is denatured and allowed to reanneal to form hemimethylated heterohybrids composed of one strand from each of the two different genomes, as well as homohybrids. The homohybrids, which are either completely methylated or completely unmethylated, are selectively removed by treatment with methylation-sensitive restriction enzymes and exonuclease. Subsequently, mismatch-containing heterohybrids are depleted with a combination of *Escherichia coli* mismatch repair proteins and exonuclease (Nelson *et al.*, 1993). Mismatch-free DNA heterohybrids are less depleted by this procedure and are, therefore, relatively enriched. The genomic locations of the GMS-selected IBD DNA fragments are determined by hybridization onto an array of mapped DNA clones gridded in physical map order. In this process, hundreds of thousands of DNA restriction fragments that span the genome are analyzed in parallel. Thus, GMS promises to be an efficient mapping method for determining identity between genomes.

GMS selection is made possible by the rate of natural sequence polymorphism. Single basepair differences (efficiently recognized by the mismatch repair proteins) are the most abundant polymorphisms and occur with frequency of 1 in 1000 to 1 in 100 bp in noncoding regions (Cooper *et al.*, 1985; Li and Sadler, 1991; Bowcock and Cavalli-Sforza, 1991). Therefore, sufficiently large DNA molecules created by the restriction enzyme (1.5–4.0 kb) and analyzed between genomes by GMS are likely to contain multiple sequence differences if not inherited IBD. Those heterohybrids without mismatches, which are likely to represent DNA fragments that the two individuals have inherited IBD from a common ancestor, are relatively purified in GMS.

Proof of principle of GMS was originally established for the yeast genome (Nelson *et al.*, 1993). Compared to the yeast genome, the human genome is approximately 250-fold larger, has more repetitive elements, and has decreased sequence divergence. Accordingly, we have modified the GMS procedure for human genomic DNA and have adapted techniques to monitor the enrichment of specific IBD alleles during GMS. About 70% of the DNA fragments that are IBD are enriched in the

¹ To whom correspondence should be addressed at UCLA Medical Center, Room A149 RNRC, 710 Westwood Plaza, Los Angeles, CA 90095-1769. Telephone: (310) 794-7981. Fax: (310) 206-9819. E-mail: snelson@ucla.edu.

current GMS protocol. Recently, it was shown that only IBD alleles were detectable by genotyping GMS-selected DNA from two distantly related affecteds with iridogoniodysgenesis (Mirzayans *et al.*, 1997), which establishes that GMS can enrich identical fragments in the context of complex mammalian genomes. This work focuses on the range of selection and enrichment of IBD alleles possible by GMS in an analysis of multiple fragments throughout the genome and forms the framework for developing array-based methods for high-throughput genotyping by GMS.

MATERIALS AND METHODS

GMS selection. High-molecular-weight genomic DNA from two related individuals was digested with 10 U/ μ g *Pst*I at 37°C. One individual's digested DNA was fully methylated with 10 U/ μ g *dam* methylase and 160 μ M *S*-adenosylmethionine in the supplied buffer (NEB) at 37°C for 4 h. Five micrograms of the methylated *Pst*I-digested DNA was mixed with the other person's *Pst*I-digested DNA in a total volume of 100 μ l. The mixture was denatured by addition of 10 N NaOH to 0.3 M NaOH for 15 min at 25°C and subsequently neutralized by adding 3 M Mops acid to 0.4 M Mops acid, which resulted in a pH 8.0 solution. The final volume was adjusted to 400 μ l in 2 M sodium thiocyanate, 10 mM Tris-Cl (pH 8.0), 0.1 mM EDTA, and 8% formamide (Casna *et al.*, 1986). Water-saturated phenol (about 150 μ l) was added until an emulsion just formed, and the mixture was agitated for 24 h at room temperature to perform formamide phenol emulsion reassociation technique (FPERT). The solution was chloroform-extracted twice and ethanol-precipitated. The reannealed DNA solution was adjusted to 0.8 M LiCl and incubated in 40 mg benzoylated naphthylated DEAE cellulose (BNDC), which was also equilibrated in 0.8 M LiCl and 10 mM Tris-1 mM EDTA (pH 8.0) at 25°C for 30 min (Sedat *et al.*, 1967). The sample was passed through a 0.45- μ m cellulose acetate filter (Costar) to remove the BNDC from the supernatant and was ethanol-precipitated. FPERT-reannealed genomic DNA (about 500–1000 ng recovered) was digested with 4 U *Dpn*I and 0.5 U *Mbo*I in 100 mM NaCl, 50 mM Tris-Cl (pH 7.9), 10 mM MgCl₂, 1 mM DTT in a total volume of 60 μ l at 37°C for 30 min. The volume of the sample was doubled by adding an equal volume of 66 mM Tris-Cl (pH 8.0) and 0.66 mM MgCl₂. One unit of exonuclease III was added, and the sample was incubated at 37°C for 15 min (Henikoff, 1984) to create large single-stranded tails on the DNA molecules cleaved by *Dpn*I and *Mbo*I. The DNA mixture was again treated with BNDC to remove the DNA with substantial single-stranded regions (greater than 100 bp). The remaining DNA (about 150 ng), composed primarily of heterohybrids, was then incubated with 5250 ng MutS, 2550 ng MutL, and 78 ng MutH in 50 mM Hepes (pH 8.0), 20 mM KCl, 4 mM MgCl₂, 1 mM DTT, 50 μ g/ml BSA, and 2 mM ATP in a total volume of 250 μ l, at 37°C for 45 min. Then 3.75 U *Exo*III was added, and the sample was incubated at 37°C for 15 min. The mixture again was treated with 0.75 mg equilibrated BNDC, spun through a 0.45- μ m filter, and ethanol-precipitated. The GMS-selected DNA (10–25 ng) was resuspended in 15 μ l of 10 mM Tris-Cl (pH 7.4) and 1 mM EDTA.

Microsatellite analysis. Quantitative genotypings were performed using 1/10 of the heterohybrid DNA or final GMS-selected DNA, fluorescent primers (0.8 μ M each), 200 μ M dNTP, 2.5 units AmpliTaq DNA polymerase (Perkin-Elmer), 10 mM Tris-Cl (pH 9.0), 50 mM KCl, 1.5 mM MgCl₂, 0.1% Triton X-100, 0.01% gelatin in a final volume of 10 μ l. Amplifications were carried out by an initial denaturation at 96°C for 3 min followed by 30 cycles of 93°C for 45 s, 56°C for 45 s, and 72°C for 45 s and a final extension at 72°C for 3 min. Multiplexing with up to four microsatellite primers was typically performed in each amplification. Samples were diluted 10- to 20-fold. One microliter of the diluted sample was mixed with 12 μ l formamide and 1 μ l GeneScan Tamra-500 size standard (Perkin-Elmer), denatured, and loaded onto an Applied Biosystems 310

Genotyper, or 1 to 4 μ l was mixed with 2 μ l formamide, 0.5 μ l loading buffer, and 1 μ l GeneScan Tamra-500 size standard, denatured, and loaded onto an Applied Biosystems 373 Genotyper. On the 310 Genotyper, the samples were electrophoresed with 3% polymer/6.6 M urea solution on the short denatured C program. On the 373 Genotyper, the samples were electrophoresed on a 6% denaturing polyacrylamide gel. Fluorescence quantitation was performed with the Genotyper software (Applied Biosystems). PCR amplification of the samples allowed direct quantitation of the fluorescent peaks for each allele, which accurately reflects the relative amounts of template DNA in a manner analogous to competitive PCR (Cheung and Nelson, 1996). In addition to the markers explicitly mentioned in the text, the following markers were scored on GMS samples: *D1S518*, *D2S1384*, *D3S1766*, *D5S1110*, *D5S408*, *D6S1270*, *D6S1006*, *D7S1824*, *D8S592*, *D8S1128*, *D8S1143*, *D9S302*, *D13S894*, *D15S816*, *D16S539*, *D16S771*, *D17S1290*, *D17S1298*, and *D20S484*.

Measurement of enrichment. Fold enrichment is derived from two ratios: *a*, the ratio of the most abundant allele to the least abundant allele, scored after GMS, and *b*, the ratio of the same alleles at the heterohybrid step. Fold enrichment is the ratio of these two ratios (*a/b*), which adjusts for the starting proportions and corrects for differences in how well specific alleles are amplified in PCRs. However, it is only a relative measure of enrichment. For instance, consider the selection in Fig. 1. The peak area ratio of the 124-bp allele to the 128-bp allele is 1:3, as expected, at the heterohybrid stage. However, the peak area ratio after GMS is 1:12, which represents a relative 4-fold greater depletion of the 124-bp peak compared to the 128-bp peak. Assuming that the 128-bp alleles that are within non-IBD heterohybrids are as equally depleted as the 124-bp alleles, only 2/12 of the 128-bp peak are from remaining non-IBD heterohybrids. Thus, the IBD heterohybrids are relatively enriched up to 10-fold at this locus. However, this level of enrichment reflects only a 3.25-fold relative increase in mass of DNA at the IBD locus at the end of GMS. As there are four heterohybrid products, and three are depleted 10 fold relative to the IBD heterohybrid, the final product will be 1.3/4 of the original DNA, if IBD. If not IBD, all four heterohybrids will be depleted by 10, and the total DNA at that locus will be 0.4/4 of the original. Therefore, the ratio of DNA mass, if IBD, is (1.3/4)/(0.4/4) = 3.25. Alternatively, a fold enrichment of 2 indicates a mass enrichment of DNA at the IBD locus of 1.75, which is above the detection limit for comparative genomic hybridization in which intensity ratios of as little as 1.5 are detectable (Kallioniemi *et al.*, 1992).

RESULTS

In adapting GMS from the yeast to the human genome, we needed to monitor the selection of specific restriction fragments during the steps of GMS to determine the amount of enrichment of the IBD fragments. Due to the larger size and lower yield using human genomic DNA, it was necessary to quantitate specific DNA fragments in small amounts of genomic DNA. Simple sequence repeat polymorphisms provide an informative means with which to follow specific allele abundance at discrete loci during GMS selection (L. McAllister and P. Brown, pers. comm.). Enrichment of IBD alleles was assessed by a fluorescence-based assay for quantifying the relative amount of each allele in a mixture of genomic DNAs (see Materials and Methods). The peak area of fluorescence intensity at a given allele size correlates to the relative abundance of that allele (Pertl *et al.*, 1996; Cheung and Nelson, 1996; McAllister *et al.*, 1996). Figure 1 illustrates the GMS selection procedure and analysis by quantitative genotyping in which a grandfather and grandchild share one allele

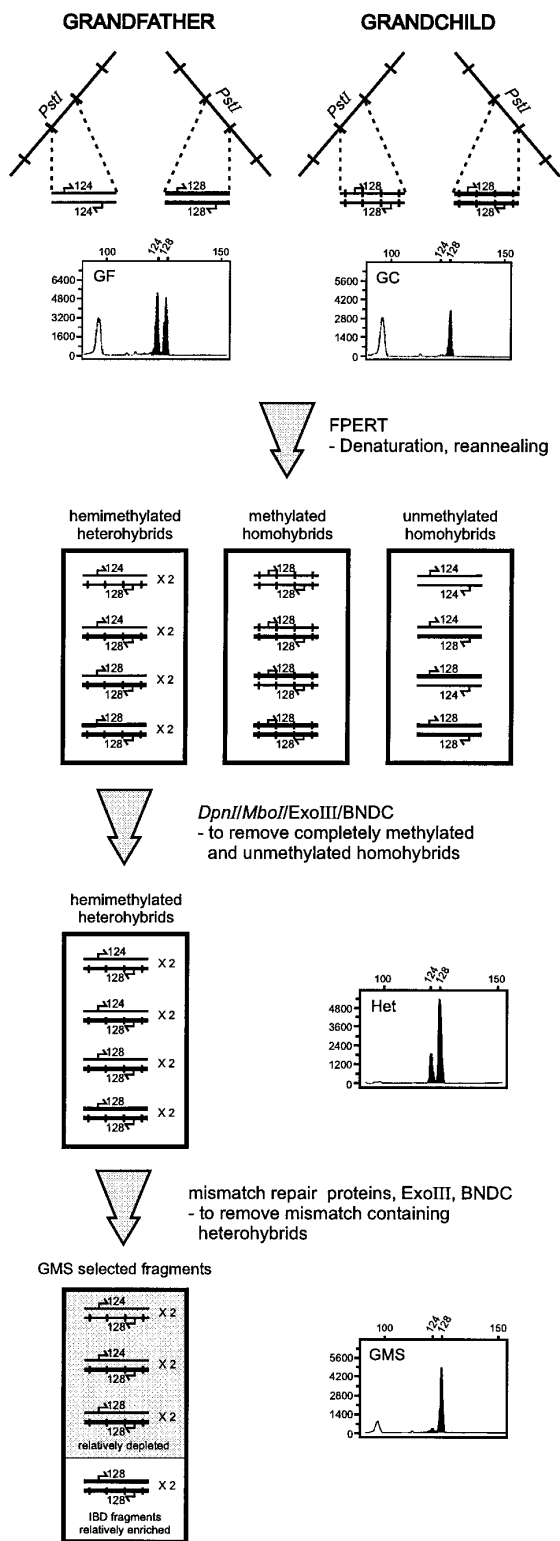


FIG. 1. GMS selection and quantitative genotyping. A flowchart of a GMS selection including electropherograms from a representative GMS experiment using a grandparent–grandchild pair from CEPH (Centre d’Etudes du Polymorphisme Humain) family 1331 at the *D4S2366* locus is shown. Five micrograms of *PstI*-digested/*dam*-methylated genomic DNA from a grandparent (7007) was compared to 5 μ g *PstI*-digested genomic DNA from his grandchild (7023) by GMS. The *PstI* fragment in which locus *D4S2366* is embedded was analyzed in this experiment. (**Top**) The genotyping of the individual

IBD and indicates a fourfold enrichment of the IBD peak (see Materials and Methods).

Fourteen microsatellite DNA markers selected from tri- and tetranucleotide repeat polymorphisms on 11 chromosomes were scored, and allele peaks were quantified on two reciprocal grandparent–grandchild comparisons (Table 1). GMS comparisons were performed to compare a child with her paternal grandfather and paternal grandmother. For each locus, the father transmitted the allele he received from only one of his parents. Therefore, when child and paternal grandparents are compared, there must be one allele that is IBD with one grandparent, and no allele can be IBD with the other at each locus. Nine of the fourteen markers allowed unequivocal determination of the IBD allele from the database genotypes. GMS identified the appropriate allele as IBD in all nine cases (Table 1). At the remaining five markers, only one of the grandparent’s alleles is enriched as expected. There is no enrichment of the most abundant allele in the non-IBD pair. Two of the fourteen markers tested (*D18S535* and *D19S433*) showed low enrichment (less than 2-fold), seven showed medium enrichment (2- to 10-fold), and five indicated a high enrichment (more than 10-fold). Three markers (*D4S2366*, *D7S513*, and *D18S858*) showed extremely strong selection of the IBD allele such that the non-IBD peak was not significantly different from the background tracings (electropherograms not shown).

GMS selection can distinguish IBD and non-IBD alleles even if the microsatellite marker is partially uninformative. For instance, at *D8S1132* the grandmother and grandchild both have the same heterozygous genotype (5, 9), but have only the 9 allele that is IBD is enriched. Thus, GMS coupled with microsatellite scoring can distinguish between identity by state and identity by descent. Moreover, scoring at *D8S1132* also indicates that GMS selection is not dependent on the mismatch created by the microsatellite size difference between alleles, but is instead due to other sequence variations between the unrelated strands. Indeed, most

genomic samples, revealing that the grandfather is a 124/128 heterozygote and the grandchild is a 128/128 homozygote. However, only one of the grandchild’s 128-bp alleles is IBD with the grandfather’s, schematically shown as the thicker line. After renaturation and removal of homohybrids by *DpnI/MboI/ExoIII* digestion followed by BNDC treatment to remove the degraded DNA, genotyping of the mixture (Het) reveals threefold greater allele abundance of the 128-bp allele than the 124-bp allele, as expected. Only one of four of the heterohybrids is the mismatch-free IBD *PstI* fragment that contains only the 128-bp allele. All the combinations of homohybrids and heterohybrids that can be formed from the *PstI* fragments that contain *D4S2366* are illustrated. In the three non-IBD heterohybrids, there is twice as much of the 128-bp strand as of the 124-bp strand. Most importantly, the 124-bp peak derives only from non-IBD heterohybrids. Mismatch-specific nicking followed by exonuclease III digestion and BNDC treatment selectively removes the mismatch-containing heterohybrids (in the shaded box). The final genotyping (GMS) reveals a fourfold decrease in the relative abundance of the 124-bp strand, which is present only in mismatch-containing heterohybrids, compared to the 128-bp strand.

TABLE 1

Quantitative Measurement of Peak Areas of Two Grandparent-Grandchild GMS Comparisons

Markers	GC × pGF			Genotypes			GC × pGM		
	HET	GMS	Fold enrichment	pGF	GC	pGM	HET	GMS	Fold enrichment
<i>D1S1665^a</i>	1:3	1:2	0.7	4,4,	<u>3,4</u>	3,6	2:1:2	9:1:1	<u>4.5</u>
<i>D2S410</i>	1: <u>1</u>	1: <u>7</u>	<u>7</u>	7, <u>9</u>	<u>7,9</u>	1,7	1:2:1	No peak	—
<i>D2S441</i>	1:2:1	1:2:1	1	3,3	<u>2,8</u>	2,6	4:2:1	15:2:1	<u>3.8</u>
<i>D3S2387</i>	1:1:2	1:1:9	<u>4.5</u>	2,6	<u>4,6</u>	3,6	1:1:2	1:1:2	1
<i>D4S2366</i>	1:1:1:1	1:1:1:1	1	2,5	<u>1,6</u>	<u>6,6</u>	1: <u>3</u>	1: <u>1581</u>	>100
<i>D5S1453</i>	1:1:2	1:1:90	45	4,9	2,9	1,9	1:1:2	No peak	—
<i>D7S513</i>	<u>1:2:1</u>	<u>1066:1:1</u>	>100	<u>4,12</u>	<u>4,6</u>	5,15	2:1:1	No peak	—
<i>D8S1132</i>	2:1:1	1:1:1	0.5	5,8	<u>5,9</u>	5,9	1:1	1:10	<u>10</u>
<i>D8S1179^a</i>	1:1	<u>19:1</u>	<u>19</u>	<u>6,8</u>	<u>6,8</u>	3,7	1:1:1:1	No peak	—
<i>D9S301^a</i>	3:1	1:1	0.3	1,1	<u>1,6</u>	1,7	2:1:1	5:1:1	<u>2.5</u>
<i>D12S372</i>	1:2:1	1:5:1	<u>2.5</u>	2,3	<u>3,4</u>	4,4	1:3	1:1	<u>0.3</u>
<i>D18S535^a</i>	3:1:1	5:1:1	<u>1.7</u>	NA	3,3	NA	2:1	2:1	1
<i>D18S858^a</i>	1:3	1:2431	>100	2,5	5,5	5,5	1	1	NI
<i>D19S433</i>	1:1: <u>3</u>	1:1: <u>6</u>	<u>2</u>	2, <u>4</u>	<u>1,4</u>	1,2	1:1:3	1:1:1	0.3

Note. One grandchild was compared by GMS with each paternal grandparent. Aliquots of heterozygotes and final GMS products were quantitatively genotyped with microsatellite markers. Relative peak areas as measured from electropherograms are shown setting the smallest peak to 1. Fold enrichment is derived from two ratios: *a*, the ratio of the most abundant allele to the least abundant allele, scored after GMS, and *b*, the ratio of the same alleles at the heterozygote step. Fold enrichment is the ratio of these two ratios (*a/b*). The precision of relative peak area determination is limited when over 100-fold differences are measured in single electropherograms. Therefore, at these loci, fold enrichment is listed as >100. Genomic DNAs are from CEPH family 1341. IBD alleles and their peak areas are underlined. The allele that is IBD was determined using marker data from a public database (<http://genetics.mfclin.edu/>). In verifying the IBD status, all the parental and grandparental genotypes were examined whenever available in the database. Paternal grandfather (pGF) is individual 7034. Paternal grandmother (pGM) is individual 7055, and the grandchild (GC) is individual 7344. NI, not informative marker. NA, not available in public database. HET, heterozygotes. GMS, GMS selected sample.

^a Markers at which the IBD allele could not be conclusively demonstrated by genotyping.

of the mismatches created by the microsatellite size differences are too large to be bound by the *E. coli* MutS, which binds well up to 3-bp insertions (Learn and Grafstrom, 1989). Thus, most of the mismatch detection in GMS is likely due to single basepair differences between the non-IBD DNA strands.

DISCUSSION

The purpose of this study is to demonstrate the enrichment of IBD DNA by GMS in the human genome. In total, we have performed 25 separate GMS experiments on 14 distinct grandparent-grandchild pairs from CEPH pedigrees, each of which has been scored at up to 33 microsatellites (Table 2). Repeat GMS experiments on the same pairs of relatives always yield enrichment of IBD alleles at specific loci. However, there is some experimental variability in the level of selection as noted in Table 2. Nonetheless, if one marker showed strong selection, then other markers in the rest of the genome were also strongly selected in that experiment. Thus, only a limited number of quantitative genotypings are required to determine that GMS selection occurred in a given sample. Of the 122 genotypings performed, 29% indicated high enrichment, 45% indicated medium enrichment, and 26% showed minimal enrichment of the IBD alleles. Among the 33 markers tested, 7 markers (21%) consistently

showed high enrichment (greater than 10-fold). Four markers ranged from medium (2- to 10-fold) to high enrichment. Seven markers consistently showed medium enrichment. At 9 markers, variable results were obtained with minimal to high enrichment, and 6 markers never showed enrichment. Thus, 18 of 33 loci (55%) were enriched consistently in relative pairs, whereas at 15 of 33 markers, variable or no enrichment was detected despite the existence of regions that are IBD. Although the sample size is small, approximately three-quarters of the *Pst*I fragments monitored are enriched by GMS and approximately one in five are consistently and strongly enriched. Because this analysis tests only *Pst*I fragments containing microsatellites, the true likelihood of an arbitrary *Pst*I fragment being appropriately enriched has not been determined. However, the primary goal of this analysis was to determine whether GMS would enrich for a substantial fraction of alleles proven to be IBD in the context of the whole genome, and this has been accomplished.

The poor selection of specific *Pst*I fragments may be due to several factors. Each DNA molecule must contain at least one sequence mismatch between strands and one GATC site to be appropriately removed by GMS (Su and Modrich, 1986; Smith and Modrich, 1996; Su *et al.*, 1988; Au *et al.*, 1992; Nelson *et al.*, 1993). Restriction fragments that lack these features will not be removed even if they are within non-IBD regions.

TABLE 2

Summary of GMS Enrichment of IBD Allele

Markers	No. of experiments showing <2-fold enrichment	No. of experiments showing 2- to 10-fold enrichment	No. of experiments showing >10-fold enrichment
D1S1665	2	1	2
D1S518	1	4	0
D2S1384	2	0	0
D2S441	0	3	0
D2S410	0	7	0
D3S1766	0	0	1
D3S2387	0	4	0
D4S2366	0	2	4
D5S1453	0	0	3
D5S1110	3	0	0
D5S408	2	0	0
D6S1270	2	0	0
D6S1006	0	7	2
D7S513	0	0	3
D7S1824	0	2	0
D8S1179	4	0	2
D8S592	3	4	0
D8S1128	0	3	0
D8S1132	2	3	2
D9S301	1	3	0
D9S302	0	2	0
D12S372	3	8	2
D13S894	0	0	3
D15S816	0	0	2
D16S539	0	3	0
D16S771	0	0	4
D17S1290	3	0	0
D17S1298	0	0	3
D18S535	1	0	0
D18S858	0	2	2
D19S433	2	3	0
D20S484	2	0	0
Sum	32 (26%)	55 (45%)	35 (29%)

Note. This table is a summary of 122 GMS samples evaluated at random microsatellite sites for fold enrichment of IBD alleles. The results are grouped into three categories, less than 2-fold enrichment (minimal), 2- to 10-fold enrichment (medium), and greater than 10-fold enrichment (high).

This will tend to happen more frequently on smaller restriction fragments. Conversely, large fragments (greater than 20 kb) are degraded during reannealing, and *PstI* fragments with a high density of repetitive sequences may not successfully reanneal with their allelic partners. Thus, both classes are depleted independently of sequence identity. The loci that were variably enriched may reflect polymorphisms that create mismatches that are variably identified by the mismatch repair proteins or polymorphisms in GATC sequences. The spectrum of loci that can be appropriately selected by GMS can likely be altered by choosing different initial restriction enzymes to shift the size of the average fragment.

Overall, the major issue in adapting GMS to human genomic DNA, compared to yeast, is coping with significantly reduced yields at the multiple steps. For a human GMS comparison using 5 μ g of *PstI*-digested

genomic DNA from each of the two relatives, the yield is about 1 μ g (10%) of the reannealed duplex DNA, which is a mixture of 50% heterohybrids and 50% homohybrids, which is much lower than the 50% reannealing of yeast genomes with FPERT. After the homohybrids are removed using methylation-sensitive enzymes, *DpnI* and *MboI*, and exonuclease III, 500–750 ng of heterohybrids remain. Most of the steps of GMS were unchanged from the successful yeast experiments, and the relative amounts of restriction enzyme and exonuclease were unchanged from the yeast experiments. We have found that the use of 0.8 M LiCl₂ in the high-salt buffer of the BNDC slurry improved recovery of double-stranded DNA following the BNDC single-stranded DNA bindings.

Compared to the yeast genome, the human genome has a lower frequency of natural sequence polymorphism; thus unrelated DNA heteroduplexes will have fewer basepair mismatches and possibly be more difficult to discriminate from nonmismatched IBD heteroduplexes. In addition, as the human genome is diploid and substantially more complex than the yeast genome, the relative titration of mismatch repair proteins to distinguish between mismatch-containing and mismatch-free heteroduplexes was difficult to predict. The exact titration of the mismatch repair proteins was empirically optimized to maximize relative enrichment of IBD *PstI* fragments as determined by microsatellite quantitation. Subtle differences in preparative methods for the purification that lead to differences in the specific activity of each mismatch repair protein preparation has required titration with individual lots of protein obtained from Amersham. However, the relative ratios of the mismatch repairs proteins for *in vivo* nicking are similar to those previously published (Au *et al.*, 1992).

Most of the losses during GMS are at the first reannealing step, where only 10–15% yield is achieved, and by nonspecific nicking of IBD duplexes in the final HLS nicking reaction. The yield of the procedure is about 1–2% of the theoretical maximum. Thus, there is a requirement for amplification of the 10–25 ng of GMS-selected genomic DNA to generate sufficient DNA to probe microarrayed genomic fragments. Typical amplification techniques, including inter-*Alu* PCR, require as little as 1 ng of genomic DNA for whole-genome amplification (Nelson *et al.*, 1989). Thus, the low yield is sufficient and does not inhibit the development of GMS as an efficient whole-genome mapping technique. Indeed, smaller starting amounts of genomic DNA may be tenable with appropriate amplification techniques.

In the present study, we used microsatellite genotyping to determine the success of GMS selection on arbitrary *PstI* fragments. However, the genomic locations of the regions that share IBD between a pair of affected individuals will ultimately be determined by hybridization of the GMS-selected DNA fragments onto a DNA array of mapped clones. Hybridization of this type is similar to that of comparative genomic hybridization

and allows all the GMS-selected DNA fragments to be mapped to their physical locations in one hybridization step (Kallioniemi *et al.*, 1992; McAllister *et al.*, 1996; Shalon *et al.*, 1996). It is unknown what level of enrichment is needed to allow detection by hybridization of the GMS-selected DNA onto ordered arrays. Even if the subtle differences detectable by comparative genomic hybridization are not measurable and greater than 10-fold enrichment is required, our results indicate that approximately one-fifth (7/33) of *Pst*I restriction fragments will be enriched to this extent. Based on this estimate and the fact that the average size of a *Pst*I fragment within the human genome is no greater than 3.0 kb, GMS may yield a potential marker every 15 kb (every fifth fragment) throughout the genome. In the approximately 3000-Mb human genome, this indicates that about 200,000 informative *Pst*I restriction fragments may be compared in parallel and successfully depleted by GMS if not IBD. Thus, there is potential for developing an extremely dense, highly informative GMS-based map of the genome. The development of highly parallel genetic analysis systems such as GMS may provide sufficiently robust molecular tools for identifying the genetic basis of complex human diseases (Risch and Merikangas, 1996).

ACKNOWLEDGMENTS

We thank A. J. Lusis and R. Spielman for critically reading the manuscript, and L. McAllister, J. DeRisi, and P. Brown for useful discussions. This work was performed with support from the Joseph Stokes Research Institute to V.G.C. and by grants to S.F.N. from the NCHGR (1R29HG01141), the NARSAD (H940323), and the MDA (H940502).

REFERENCES

- Au, K. G., Welsh, K., and Modrich, P. (1992). Initiation of methyl-directed mismatch repair. *J. Biol. Chem.* **267**: 12142–12148.
- Bowcock, A., and Cavalli-Sforza, L. (1991). The study of variation in the human genome. *Genomics* **11**: 491–498.
- Casna, N., Novack, D., Hsu, M., and Ford, D. (1986). Genomic analysis II, isolation of high molecular weight heteroduplex DNA following differential methylase protection and formamide PERT hybridization. *Nucleic Acids Res.* **14**: 7285–7303.
- Cheung, V. G., and Nelson, S. F. (1996a). Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proc. Natl. Acad. Sci. USA* **93**: 14676–14679.
- Cheung, V. G., and Nelson, S. F. (1996b). Genomic mismatch scanning: An identity-by-descent physical mapping technique in human. *Am. J. Hum. Genet.* **59**(Suppl. 1): A1740.
- Cooper, D. N., Smith, B. A., Cooke, H. J., Niemann, S., and Schmidte, J. (1985). An estimate of unique DNA sequence heterozygosity in the human genome. *Hum. Genet.* **69**: 201–205.
- Henikoff, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28**: 351–360.
- Kallioniemi, A., *et al.* (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**: 818–821.
- Learn, B. A., and Grafstrom, R. H. (1989). Methyl-directed repair of frameshift heteroduplexes in cell extracts from *Escherichia coli*. *J. Bacteriol.* **171**: 6473–6481.
- Li, W. H., and Sadler, L. A. (1991). Low nucleotide diversity in man. *Genetics* **129**: 513–523.
- McAllister, L., Penland, L., DeRisi, J., and Brown, P. O. (1996). Application of genomic mismatch scanning to mammalian genomes. *Am. J. Hum. Genet.* **59**(Suppl. 1): A303.
- Mirzayans, F., Mears, A. J., Guo, S. W., Pearce, W. G., and Walter, M. A. (1997). Identification of the human chromosomal region containing the iridogoniodysgenesis anomaly locus by genomic-mismatch scanning. *Am. J. Hum. Genet.* **61**: 111–119.
- Nelson, D. L., Ledbetter, S. A., Corbo, L., Victoria, M. F., Ramirez-Solis, R., Webster, T. D., Ledbetter, D. H., and Caskey, C. T. (1989). Alu polymerase chain reaction: A method for rapid isolation of human-specific sequences from complex DNA sources. *Proc. Natl. Acad. Sci. USA* **86**: 6686–6690.
- Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P., and Brown, P. O. (1993). Genomic mismatch scanning, a new approach to genetic linkage mapping. *Nat. Genet.* **4**: 11–18.
- Pertl, B., *et al.* (1996). Rapid detection of trisomies 21 and 18 and sexing by quantitative fluorescent multiplex PCR. *Hum. Genet.* **98**: 55–59.
- Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**: 1516–1517.
- Sedat, J. W., Kelly, R. B., and Sinsheimer, R. L. (1967). Fractionation of nucleic acid on benzoylated-naphthoylated DEAE cellulose. *J. Mol. Biol.* **26**: 537–540.
- Shalon, D., Smith, S. J., and Brown, P. O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**: 639–645.
- Smith, J., and Modrich, P. (1996). Mutation detection with MutH, MutL, and MutS mismatch repair proteins. *Proc. Natl. Acad. Sci. USA* **93**: 4374–4379.
- Su, S., Lahue, R. S., Au, K. G., and Modrich, P. (1988). Mismatch specificity of methyl-directed DNA mismatch correction in vitro. *J. Biol. Chem.* **263**: 6829–6835.
- Su, S. S., and Modrich, P. (1986). *E. coli* mutS-encoded protein binds to mismatched DNA base pairs. *Proc. Natl. Acad. Sci. USA* **83**: 5057–5061.