# Overlapping Genomic Sequences: A Treasure Trove of Single-Nucleotide Polymorphisms

**Patricia Taillon-Miller,[1] Zhijie Gu,[1] Qun Li,[1] LaDeana Hillier,[2] and Pui-Yan Kwok[1,3]**

[1]Division of Dermatology and [2]Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63110 USA

An efficient strategy to develop a dense set of single-nucleotide polymorphism (SNP) markers is to take advantage of the human genome sequencing effort currently under way. Our approach is based on the fact that bacterial artificial chromosomes (BACs) and P1-based artificial chromosomes (PACs) used in long-range sequencing projects come from diploid libraries. If the overlapping clones sequenced are from different lineages, one is comparing the sequences from 2 homologous chromosomes in the overlapping region. We have analyzed in detail every SNP identified while sequencing three sets of overlapping clones found on chromosome 5p15.2, 7q21–7q22, and 13q12–13q13. In the 200.6 kb of DNA sequence analyzed in these overlaps, 153 SNPs were identified. Computer analysis for repetitive elements and suitability for STS development yielded 44 STSs containing 68 SNPs for further study. All 68 SNPs were confirmed to be present in at least one of the three (Caucasian, African-American, Hispanic) populations studied. Furthermore, 42 of the SNPs tested (62%) were informative in at least one population, 32 (47%) were informative in two or more populations, and 23 (34%) were informative in all three populations. These results clearly indicate that developing SNP markers from overlapping genomic sequence is highly efficient and cost effective, requiring only the two simple steps of developing STSs around the known SNPs and characterizing them in the appropriate populations.

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. AC003015 (for GS113423), AC002380 (GS330J10), AC000066 (RG293F11), AC003086 (RG104F04), AC002525 (257C22A), and U73331 (96A18A).]

There is increasing agreement that association studies using a set of single-nucleotide polymorphism (SNP) markers across the genome with markers evenly distributed at ~100-kb intervals would provide the necessary power to detect small genetic effects for a given complex disease trait (Collins et al. 1997; Kruglyak 1997). To develop 30,000 or more SNP markers is a priority of the consortium of National Institutes led by the National Human Genome Research Institute (Marshall 1997). Although the frequency of SNPs is approximately 1 in 1000 bp between any two chromosomes (Cooper et al. 1985; Kwok et al. 1996), there are currently no efficient ways to find and map them from scratch.

In general, development of SNP markers requires five different steps: obtain DNA sequence, develop STSs from the DNA sequence, screen STSs for SNPs, characterize SNPs, and map SNPs to specific chromosomal locations. To date, much effort has been devoted to devising more efficient ways to screen STSs for SNPs and to characterize them. Being largely ignored is the fact that the most costly aspects of developing SNP markers are the obtaining of DNA sequence for STS development initially and mapping the SNPs at the end of the process. We and others have developed various strategies to improve the efficiency of this process by utilizing existing resources to our advantage. For example, we have screened mapped STSs for SNPs, thereby reducing the development of SNPs to two steps (screening for and characterizing SNPs), abrogating the need for genomic DNA sequencing, STS development, or mapping (Kwok et al. 1996). Because the mapped STSs were developed for YAC library screening in physical mapping, they amplified short DNA fragments and screening several of them was required before a SNP could be found. Moreover, as the physical mapping effort is winding down over the next 2 years, this resource will not be available for further SNP marker development.

Fortunately, the human genome sequencing effort that is currently under way provides a better

[3]Corresponding author.
E-MAIL kwok@im.wustl.edu; FAX (314) 362-8159.

 8:748–754 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

way to develop a dense set of SNP markers in the genome. Like screening mapped STSs for SNPs, our approach by passes the need for DNA sequence acquisition and mapping for SNPs. In addition, it eliminates the polymorphism screening step altogether, leaving only the development of STSs around the SNPs found during the course of genome sequencing and their characterization. The strategy is based on the fact that bacterial artificial chromosomes (BACs) and P1-based artificial chromosomes (PACs), the substrates of choice for long-range sequencing, come from diploid libraries. With an average insert size of ~120 kb, one can expect a significant overlap between clones selected for sequencing at ~100 kb intervals. If the clones are from different libraries (presumably from different individuals), one is comparing the sequences from two lineages in the overlapping region. If the clones are from the same library, there is still a 50% chance that the overlapping clones are derived from different lineages (paternal or maternal). This probability could increase to close to 100% if libraries made from mixtures of individuals are used. Although one is sampling just two copies of the same region for polymorphisms in this approach, the chance of identifying a polymorphism in the region is the same as the heterozygosity of the polymorphism in the population. Therefore, for the more informative markers (defined as having heterozygosity of >30%) there is a >30% chance of identifying them when sequences of overlapping clones are examined. Even with a minimal 10% overlap when clone libraries with 10-fold genomic redundancy are used to provide sequencing clones, a 6-kb overlap at each end of the clone insert will result. Given the general observation that one informative marker is found in 1.5–2.0-kb in the human genome (Kwok et al. 1996), at least one such polymorphism should be found in each 4.5–6-kb overlap. In practice, the overlaps are much larger than 6 kb because of the stringent requirements of physical mapping by fingerprinting in selecting clones for large-scale genome sequencing (Marra et al. 1997), and as described here, one can almost certainly find informative SNPs in them.

To test the validity of this approach, we have analyzed in detail every SNP identified while sequencing three sets of overlapping clones found on chromosome 5p15.2, 7q21–7q22, and 13q12–13q13. We report here that this approach of SNP marker development is highly efficient and cost effective, requiring only the two simple steps of developing STSs around the known SNPs and characterizing them in the appropriate populations.

## RESULTS

In the course of sequencing overlapping BAC and PAC clones from human genomic libraries at the Genome Sequencing Center (GSC) at Washington

Table 1. Results of Analyzing 68 SNPs Found in 3 Overlapping Clones

| Chromosome location | Length of overlap (bp) | Polymorphisms found | SNPs[a] | SNPs analyzed[b] | Informative SNPs[c] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | African–American | Caucasian | Hispanic | populations 3 | populations >2 | populations >1 |
| 5p15.2 | 81,830 | 20 | 18 | 10 | 6 | 8 | 6 | 6 | 6 | 8 |
| | | | | 8 STSs 3892 bp | 60% | 80% | 60% | 60% | 60% | 80% |
| 7q21–7q22 | 59,048 | 97 | 83 | 20 | 14 | 15 | 17 | 12 | 16 | 18 |
| | | | | 13 STSs 6360 bp | 70% | 75% | 85% | 60% | 80% | 90% |
| 13q12–q13 | 59,739 | 66 | 52 | 38 | 11 | 8 | 11 | 5 | 10 | 16 |
| | | | | 23 STSs 7720 bp | 29% | 21% | 29% | 13% | 26% | 42% |
| Totals | 200,617 | 183 | 153 | 68 | 31 | 31 | 34 | 23 | 32 | 42 |
| | | | | 44 STSs 18172 bp | 46% | 46% | 50% | 34% | 47% | 62% |

[a]SNPs found among all of the polymorphisms identified in the overlap.
[b]SNPs analyzed by determining their frequencies in the population pools, after discarding from consideration polymorphisms found in repetitive regions. The number of STSs and base pairs scanned in the process of analyzing the SNPs are also listed.
[c]Number of SNPs (and the proportion among the SNPs analyzed) with allele frequencies of >20% for the minor allele are categorized according to the individual populations studied and the number of populations in which these informative SNPs are found.

University in St. Louis, MO, 183 polymorphisms were found in 200.6 kb in three overlapping regions, for an average of 1 polymorphism every 1.1 kb (see Table 1). In the 81.8-kb overlap on chromosome 5, 20 polymorphisms were identified. Of these, 17 polymorphisms were single-base substitution polymorphisms (85%), 1 was an unique insertion/deletion polymorphism (5%), and 2 were short tandem repeat polymorphisms (STRP) (10%). In the 59.0-kb overlap on chromosome 7, 97 polymorphisms were found, with 83 being substitution polymorphisms (86%), 11 insertion/deletion polymorphisms in a run of a single base such as a poly(A) (11%) and 3 STRPs (3%). In the 59.7-kb overlap on chromosome 13, 66 polymorphisms were found, with 49 substitution polymorphisms (74%), 3 unique insertion/deletions (5%), 12 insertion/deletions in a run of a single base (18%), and 2 STRPs (3%). Overall, there were 153 SNPs (substitution and unique insertion/deletion polymorphisms) at a frequency of 1 per 1.3 kb. In contrast, there were only 7 STRPs (1 per 28.7 kb).

The 153 SNPs were evaluated further for their usefulness as genetic markers. Those found in common repeat regions masked by the GSC during the sequence annotation such as Alu and L1 were discarded. In all, 55 SNPs were eliminated by computer analysis 5/19 (26%) in the chromosome 5 overlap, 43/83 (52%) in the chromosome 7 overlap, and 7/52 (13%) in the chromosome 13 overlap. The oligonucleotide selection program (osp) (Hillier and Green 1991) was used to design primers to amplify each of the 98 remaining SNPs. Thirty SNPs were in regions of DNA in which no suitable amplimers could be found, including SNPs in repeat regions other than *Alu*s and L1s and a small number of PCR failures. In all, 44 STSs were developed to amplify the remaining 68 SNPs (spanning 18,172 bp of DNA sequence). Among them, 16 STSs contained 2 SNPs and 8 STSs contained 3 or more SNPs.

The 68 SNPs were confirmed by sequence analysis with the homozygous CHM1 DNA from a homozygous complete hydatidiform mole and pooled DNA samples from 30 individuals each from the Caucasian, Hispanic, and African-American populations (Kwok et al. 1994; Taillon-Miller et al. 1997). The pooled DNA sequencing approach for allele frequency estimation is highly reproducible and has been found to give estimates of within 5% of the true allele frequency as found by genotyping all the individuals in the population pool (Kwok et al. 1994). Allele frequency estimates of the 68 SNPs revealed that the minor allele in 23 SNPs (34% of the analyzed SNPs) had a frequency >20% in all three populations (>32% heterozygosity, assuming Hardy–Weinberg equilibrium), 32 (47%) had a frequency >20% in at least two populations, and 42 (62%) had a frequency >20% in at least one population.

In addition, 18 new SNPs were discovered during the course of analyzing the 44 STSs in 18.2 kb of DNA sequence contained in the STSs (found at a rate of 1 per 1.0 kb). Among these, 9 (50%) were found to be informative in one or more populations and 3 (17%, 1 per 2.0 kb) were informative in all three populations.

The 26 SNPs with frequencies >20% in all 3 populations tested are presented in Table 2. Information about the remaining SNPs with their estimated allele frequencies in each of the three populations can be found on our public database (currently under construction) accessible through the internet (http://www.ibc.wustl.edu/SNP).

## DISCUSSION

The overall results of this study confirmed that the chance of finding informative SNPs by use of overlapping regions of clones sequenced as part of the human genome sequencing project was in line with our expectations. We had expected to find one informative SNP per 4.5–6.0 kb, and we found one informative SNP per 4.8 kb (3.9 kb if those markers discovered during the sequence analysis of STSs for characterization of the SNPs were included). Although it is safe to assume that the ethnic origins of the donors of the BAC and PAC libraries are Caucasian, many of the SNPs found were also polymorphic in the African-American and Hispanic populations. These results point to the fact that the more informative SNPs are more ancient and are therefore informative in most populations (Kimura 1983).

On closer examination, however, it is clear that there is a large variation in the frequency of finding SNPs among the different overlaps. For example, whereas SNPs are found at a rate of 1 per 4.5 kb in the 5q15.2 overlap, the rate is 1 per 0.7 kb in the 7q21–7q22 overlap and 1 per 1.1 kb in the 13q12–13q13. After computer screening, the rate of analyzable SNPs ranges from 1 in 8.1 kb for 5q12.2, to 1 in 3.0 kb for 7q21–7q22, to 1 in 1.6 kb for 13q12–13q13.

Furthermore, the chromosome 7q21–7q22 overlap was unique in that it had an extremely large number of repeat elements within which the bulk of the SNPs were found, leaving only 20 of the 83 SNPs (24%) suitable for sequence analysis. Among the analyzable SNPs, >80% were informative in at least

Table 2. SNPs with Frequencies >0.20 in Three Populations

| STS Name | Chromosome Location | STS Size (bp) | STS Primer 1 Sequence | STS Primer 2 Sequence | Sequence Context of Polymorphism | Allele 1 | Allele 2 | African-American Frequency Allele 1 | Hispanic Frequency Allele 1 | Caucasian Frequency Allele 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 5p0002-1 | 5p15.2 | 819 | CTTCCCTAAAC AAAACCTAC | ACAGTGAAAGG AAATGAGTC | GCACTTAGTCTTTGTATTAG[TCAG/*] AGTCCTCCAGAAAGAATCAA | TCAG | * | 0.41 | 0.48 | 0.31 |
| 5p0003-2 | 5p15.2 | 719 | GTATTGATAGA CTTGCTTCC | AGCCTACACATT TTCTTCTG | AACTATTATTACTCAAACATA[T/C]CT GGAACTGATTTCAATAAT | C | T | 0.55 | 0.62 | 0.38 |
| 5p0003-4 | 5p15.2 | 719 | GTATTGATAGA CTTGCTTCC | AGCCTACACATT TTCTTCTG | CCTAAACTTTTCTAGCTTTG[A/T]TTG TAACAAAAAAACTCTAC | A | T | 0.68 | 0.55 | 0.65 |
| 5p0004-1 | 5p15.2 | 307 | TTCTTCTGTGC TTGACAAAG | TGCATCTTAATC CACTCAAC | AACTATGTCACAAAGAATCT[A/G]TG GTCAGCGAAATCATGAGT | G | A | 0.79 | 0.63 | 0.56 |
| 5p0004-2 | 5p15.2 | 307 | TTCTTCTGTGC TTGACAAAG | TGCATCTTAATC CACTCAAC | AGAAAGAGAAAGATAGTATTC[C/T]CT GTTGTCCTACCAGGAACA | C | T | 0.61 | 0.51 | 0.47 |
| 5p0008-1 | 5p15.2 | 137 | CTTACTTGATC TGCATGG | AGTAGGTAAGG AAGAGAG | TGCCCTTCTTGCAAGTACTA[C/G]GA CACTCCAATCCCATTTCT | C | G | 0.63 | 0.62 | 0.68 |
| 5p0009-1 | 5p15.2 | 755 | GGAGACAAAA CAGATGAG | TCACATCCTGA GAGATTG | TTTTTAAATTAAAAGTGATTA[T/C]ACT GACCAACCACATAACAA | T | C | 0.46 | 0.62 | 0.34 |
| 7q0001-1 | 7q21-22 | 627 | TGAACGATTC GCAGATTG | TACTTCATGCAC CACTTC | GTTTAAACTATTTGTTCTGG[A/C]AGA AGGAGGGAATGTGAGAC | A | C | 0.43 | 0.48 | 0.34 |
| 7q0002-1 | 7q21-22 | 170 | AAAACATGATT ACAGTGGG | TCTCTTCTTTTC TTCTCTC | AATGGAGCAGAAACTGGAAA[T/C]GT GGGTTTAAGAAAAGGCTG | T | C | 0.26 | 0.22 | 0.37 |
| 7q0003-1 | 7q21-22 | 396 | ACTTCTACCCT CCTTAAC | CCAGTGTTCAA AGGTATC | AACAGAGGATACTGAGGTGA[G/T]AT AATTATAAGGAAACAGAT | G | T | 0.72 | 0.37 | 0.59 |
| 7q0004-1 | 7q21-22 | 373 | TTTCAAGAAAG ATCATGGG | GAAAAGAAAAAT GGCTTGAG | ATCCTGCTTGAGCAATATTC[C/G]AG TCTTGATGAATGTAGTTC | C | G | 0.72 | 0.71 | 0.70 |
| 7q0004-2 | 7q21-22 | 373 | TTTCAAGAAAG ATCATGGG | GAAAAGAAAAAT GGCTTGAG | CAGAGAATTTGCATTTAGCT[A/C]TAT GACTTGCTGAGCAACAA | A | C | 0.80 | 0.75 | 0.57 |
| 7q0006-1 | 7q21-22 | 918 | TATAATCCCAG CTACTCAG | CTACGTTTTCAT AGACAATTC | TCCAGCCTGGGAGACAGAGG[A/G]A AACTCTGTCTCAAAAAGAA | A | G | 0.45 | 0.28 | 0.34 |
| 7q0007-1 | 7q21-22 | 539 | TAGCATGAACT TCAGAAGG | CGATGAATTAAC AGAGCC | CTCTAAGCCCAGTAACACTA[C/G]GG CTTTGGGAGAGAAACAGC | C | G | 0.43 | 0.44 | 0.36 |

Table 2. (*Continued*)

| STS Name | Chromosome Location | STS Size (bp) | STS Primer 1 Sequence | STS Primer 2 Sequence | Sequence Context of Polymorphism | Allele 1 | Allele 2 | African-American Frequency Allele 1 | Hispanic Frequency Allele 1 | Caucasian Frequency Allele 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 7q0007-2 | 7q21-22 | 539 | TAGCATGAACT TCAGAAGG | CGATGAATTAAC AGAGCC | TGTTCTGCTCAAGTATTCTA[C/G]ATT ATTATGTTAATCTGTTT | C | G | 0.54 | 0.35 | 0.43 |
| 7q0008-1 | 7q21-22 | 274 | ATAATGGGCAT TATTTGCTG | ACTTTCTTAGCT GTGGAAC | ACAAAAAAAAGGATGGTGCT[C/T]CA TGCAAGCTGTATATTGAT | C | T | 0.80 | 0.75 | 0.75 |
| 7q0011-1 | 7q21-22 | 161 | ATTCAAAGGGT GAAATAAGG | TAGAGAGAAAG AGAGTGAG | TAAGGTAGGCAATTTTAATA[A/G]GC CTCAGAATTTTAATTGTA | A | G | 0.46 | 0.29 | 0.38 |
| 7q0011-2 | 7q21-22 | 161 | ATTCAAAGGGT GAAATAAGG | TAGAGAGAAAG AGAGTGAG | TTAAATTTTTTTAATGTTCT[A/G]AGCT TTCTCTTCTTTCTTTC | A | G | 0.56 | 0.3 | 0.38 |
| 7q0012-1 | 7q21-22 | 611 | GATCCTTTGAA TTATTCTGTG | ATCATCTTAGCA AAGTGCC | TAAATTAAAAATTTGTCTCA[C/A]ACG CTTGCAGTGAGCCAAGA | C | A | 0.67 | 0.33 | 0.46 |
| 13q0003-2 | 13q12-q13 | 170 | TGGAATTGAG ATACAGTGG | TGCATACACAC AATTAGGTC | CCCTGAAACTCGATGCAATG[C/G]TT GACTTATGAAGAGACCT | C | G | 0.77 | 0.48 | 0.58 |
| 13q0004-2 | 13q12-q13 | 241 | CAAGGTCATG TATCTAAGG | TCTAAGAAAGC AGTAGCAC | CCCTCGTAAGTTCCTTAGGC[A/G]TT TTTATTCCTTTGTTCAAC | A | G | 0.76 | 0.53 | 0.54 |
| 13q0006-1 | 13q12-q13 | 352 | CAGCACATTTC AATTCAGC | GAGTCAGAAAC CCTTATATG | CAATTCAGCTGTGTTTCA[C/T]GTGCT CAGTAGCACATGTAC | C | T | 0.23 | 0.71 | 0.23 |
| 13q0013-3 | 13q12-q13 | 858 | GAAACTGAAG TACAAAGGG | ACCTGGCTAAG AAATGAATC | TAGGATGAAATACAATAAAC[C/T]GA GCATAAAACTTTTTGCAC | C | T | 0.57 | 0.72 | 0.46 |
| 13q0020-3 | 13q12-q13 | 783 | CAGGAACAAT ACTAGCAAC | GCAGAAACAGA ATATCCTTG | TCCTCCTTCCCGTTTCTGTTT[TC/*]T GTGTCTCACACAGACATAT | TC | * | 0.41 | 0.49 | 0.48 |
| 13q0021-2 | 13q12-q13 | 614 | GAATCACAAA GGTTTGAGG | CGAGAACTCTA ATTCAAGG | GCAGCACCTTACTTCAGCTG[T/C]CC AGCTCTAAGCCAGAGTTC | T | C | 0.42 | 0.68 | 0.31 |
| 13q0022-1 | 13q12-q13 | 261 | ATATTCTTTCC TCTGCCTG | TCACTATGTATG GTGAGTC | TTTCCCAGAAGATTAAAATA[T/C]GCT TTGCAGCTTA TTATCTC | T | C | 0.53 | 0.69 | 0.45 |

[a]GenBank accession nos. are AC003015 (for GS113H23) and AC002380 (for GS330J10) on 5p15.2; AC000066 (for RG293F11) and AC003086 (for RG104F04) on chromosome 7q21–7q22; AC002525 (for 257C22A) and U73331 (for 96A18A) on chromosome 13q12–13q13.

one population in both the chromosome 5 and 7 overlaps. In contrast, 38 of the 58 SNPs (73%) in the chromosome 13q12–13q13 overlap were analyzable but only 16 of 38 SNPs (42%) were informative.

Despite these regional differences, however, the overall results show that even in the worst case scenario, one only has to analyze two to three SNPs to find an informative SNP marker. Given that almost all overlaps produced by large-scale genome sequencing projects are >20 kb, there will be more than enough analyzable SNPs to choose from.

Our approach has many advantages not found in other current methods. First, all long-range sequencing groups produce high quality sequence data, and because every base is sequenced at least twice from each clone (so-called double-stranding), the error rate is therefore much lower than the polymorphism rate. Second, because the polymorphism data are generated by examining existing data, the amount of sequencing required is minimal and the cost of the project is shifted from identification of polymorphisms toward estimates of the usefulness of the polymorphism by population sequencing. Third, because they are derived from long-range sequence data, the markers are precisely mapped, not just assigned to an interval of a clone-based contig by STS content mapping. Fourth, the physical distance between markers is known precisely. Fifth, because they are detected when only two chromosomes are examined, each SNP identified in this way has a higher chance of being informative. Sixth, this approach scales easily because the basic methods used are simple and robust, making it possible to keep up with the expanding sequencing efforts around the world. Consequently, the genetic map could be completed along with the sequencing of the genome. Seventh, the markers would be intrinsically distributed more evenly than those based on genes.

By use of this approach, a high-density genetic map with precisely placed SNP markers that are evenly placed in the genome can be assembled with minimal effort and will be available for use to study complex genetic traits as soon as the genome sequencing is completed in year 2005 (Collins and Galas 1993).

## METHODS

### DNA Sequences

Three overlap regions containing SNPs were identified by the GSC for this study. On chromosome 5p15.2, an 81,830-bp overlap between BAC clones GS113H23 (GenBank accession no. AC003015) and GS330J10 (accession no. AC002380); on chromosome 7q21–7q22, a 59,048-bp overlap between BACs RG293F11 (accession no. AC000066) and RG104F04 (accession no. AC003086); and the BRCA2 gene region on chromosome 13q12–13q13, a 59,739-bp overlap between PAC clones 257C22A (accession no. AC002525 ) and 96A18A (accession no. U73331).

### Primary Analysis of SNPs

The GSC provided us with the ability to access the database remotely and the primary assembly data was viewed by use of the XGAP program (Bonfield et al. 1995). Up to four aligned sequencing traces could be opened and viewed simultaneously for close inspection. In the XGAP program, one can set the level of discrepancies at each nucleotide position over which it is declared an ambiguous base. When the limit is set at 80%, all differences between the consensus sequences from the two overlapping clones are designated as ambiguous and flagged. Given the fact that the base-calling and assembly programs used (PHRED, Ewing et al. 1998; PHRAP, P. Green, pers. comm.) take into account the sequence data quality, one can easily tell the sequence variations caused by poor data quality from the real polymorphisms. These polymorphisms were unmistakable because all subclones from one PAC exhibited one nucleotide but all subclones of the second PAC possessed another nucleotide. Because the primary data were available for quality check, variations caused by base-calling errors were easily eliminated. The sequence context of each polymorphism was recorded at this step. Simple sequence repeats and long runs of poly(A)s or poly(T)s were eliminated from further consideration.

Annotation and masking of common repetitive elements (such as Alu and L1 repeats) were done automatically by the GSC and these regions were removed from further consideration. PCR assays were designed for all the remaining SNPs by use of the oligonucleotide selection program (Hillier and Green 1991).

[Note: At the GSC, the sequence of only one of the two overlapping clones is fully finished and deposited to GenBank. Typically, the shotgun sequence data in the overlapping region found in the second clone sequenced are assembled into contigs with a few gaps in between. The prefinished data are archived and are not deposited in a public database. However, all of the sequencing traces for both clones are freely accessible. Any researcher interested in the overlapping sequences is encouraged to contact the GSC for access to these data.]

### Determining Frequencies of SNPs in Population Pools

All PCR assays were amplified against the complete hydatidiform mole 1 (CHM1) DNA, a completely homozygous DNA described in detail previously and the Caucasian, African-American, and Hispanic population pools (30 anonymous individuals each) (Taillon-Miller et al. 1997). PCR reaction conditions and preparation of DNA template for sequencing have been described previously (Kwok et al 1994). DNA sequencing was done with the dichloro-rhodamine dye terminators analyzed on the 377 DNA Sequencer (PE Applied Biosystems, Foster City, CA) according to the manufacturer's instructions. The frequency of each allele in a population pool was determined by comparing the DNA sequencing trace of a PCR

product amplified from a pooled DNA sample with that of a PCR product amplified from the DNA sample of the CHM1 (Kwok et al 1994). The CHM1 DNA serves as a homozygous control, its normalized peak height equal to a frequency of 100%.

## ACKNOWLEDGMENTS

We thank M. Boyce-Jacino for pooled DNA samples; and Ellen Piernot, Jenna Putzel, and Jenica Lee for technical assistance. This work is supported by the National Institute of Health and National Science Foundation.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

## REFERENCES

Bonfield, J.K., K.F. Smith, and R. Staden. 1995. A new DNA sequence assembly program. *Nucleic Acids Res.* 23: 4992–4999.

Collins, F.S. and D. Galas. 1993. A new five-year plan for the U.S. Human Genome Program. *Science* 262: 43–46.

Collins, F.S., M.S. Guyer, and A. Chakravarti. 1997. Variations on a theme: Cataloging human DNA sequence variation. *Science* 278: 1580–1581.

Cooper, D.N., B.A. Smith, H.J. Cooke, S. Niemann, and J. Schmidtke. 1985. An estimate of unique DNA sequence heterozygosity in the human genome. *Hum. Genet.* 69: 201–205.

Ewing, B.G., L. Hillier, M.C. Wendl, and P. Green. 1998. Basecalling of automated sequencer traces using PHRED. I. Accuracy assessment. *Genome Res.* 8: 175–185.

Hillier, L. and P. Green. 1991. OSP: A computer program for choosing PCR and DNA sequencing primers. *PCR Methods Applic.* 1: 124–128.

Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, UK.

Kruglyak, L. 1997. The use of a genetic map of biallelic markers in linkage studies. *Nature Genet.* 17: 21–24.

Kwok, P.-Y., C. Carlson, T. Yager, W. Ankener, and D.A. Nickerson. 1994. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* 23: 138–144.

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

Marra M.A., T.A. Kucaba, N.L. Dietrich, E.D. Green, B. Brownstein, R.K. Wilson, K.M. McDonald, L.W. Hillier, J.D. McPherson, R.H. Waterston. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res.* 7: 1072–1084.

Marshall, E. 1997. Snipping away at genome patenting. *Science* 277: 1752–1753.

Taillon-Miller, P., I. Bauer-Sardiña, H. Zakeri, L. Hillier, D.G. Mutch, and P-Y. Kwok. 1997. The homozygous complete hydatidiform mole: A unique resource for genome studies. *Genomics* 46: 307–310.

*Received February 26, 1998; accepted in revised form April 28, 1998.*