Genome complexity reduction for SNP genotyping analysis

Barbara Jordan*, Alain Charest*, John F. Dowd†, Justin P. Blumenstiel*, Ru-fang Yeh*, Asiah Osman*, David E. Housman*‡, and John E. Landers†

*Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and †Polygenyx, Inc., Worcester, MA 01605

Contributed by David E. Housman, December 29, 2001

Efficient single nucleotide polymorphism (SNP) genotyping methods are necessary to accomplish many current gene discovery goals. A crucial element in large-scale SNP genotyping is the number of individual biochemical reactions that must be performed. An efficient method that can be used to simultaneously amplify a set of genetic loci across a genome with high reliability can provide a valuable tool for large-scale SNP genotyping studies. In this paper we describe and characterize a method that addresses this goal. We have developed a strategy for reducing genome complexity by using degenerate oligonucleotide primer (DOP)-PCR and applied this strategy to SNP genotyping in three complex eukaryotic genomes; human, mouse, and *Arabidopsis thaliana*. Using a single DOP-PCR primer, SNP loci spread throughout a genome can be amplified and accurately genotyped directly from a DOP-PCR product mixture. DOP-PCRs are extremely reproducible. The DOP-PCR method is transferable to many species of interest. Finally, we describe an *in silico* approach that can effectively predict the SNP loci amplified in a given DOP-PCR, permitting the design of an efficient set of reactions for large-scale, genome-wide SNP studies.

Single nucleotide polymorphisms (SNPs) are the most abundant resource of genetic variation among individuals of a species. They are the focus of large-scale genotyping projects in both humans and model organisms and the projected demands for human genome-wide association studies range into the tens and hundreds of millions of genotypes (1, 2). Large-scale mutagenesis projects and genetic modifier screens in model organisms such as *Mus musculus* and *Arabidopsis thaliana* are also driving the need for low-cost, high-efficiency genotyping. SNPs have the inherent potential for highly automated genotyping necessary for these studies and a broad range of SNP-based genotyping methodologies have been developed to date (for a recent review, see ref. 3).

However, SNP genotyping protocols developed to date require individual amplification of SNP-containing loci by PCR or some other biochemical reaction. The cost and labor required to carry out individual enzymatic reactions have led to a desire to combine, or multiplex, biochemical reactions. Multiplexing can improve efficiency, but predicting compatibility of reactions is increasingly challenging as the number of loci increases (4). An alternative strategy to address this problem is to develop a biochemical reaction that is preoptimized to achieve a similar goal to that of multiplexing.

One approach that has been applied to this problem has been to use PCR primers complementary to interspersed repetitive sequences, such as the human alu sequence, to amplify subsets of unique genomic sequences with a single PCR primer. Although this approach has great utility in many contexts, it is limited in the subset of genomic sequences that can be amplified by the position of interspersed repetitive sequences in the genome. To achieve a broader representation of possible sequences that can be amplified, we have adapted a PCR amplification method originally designed for whole-genome amplification, degenerate oligonucleotide primed (DOP)-PCR, for use in SNP genotyping. Amplification by DOP-PCR utilizes a partially degenerate oligonucleotide where approximately one third of the positions in the center of the oligonucleotide are degenerate (5, 6). When the 3′ end of such an oligonucleotide consists of six unique nucleotides, effective amplification of a high proportion of a eukaryotic genome is observed. We have systematically altered the length of the unique region of the oligonucleotide primer in the DOP-PCR protocol. By carrying out analyses of DOP-PCR products resulting from primers with 3′ unique nucleotide sequences between 6 and 10 nt in length, we have accomplished a genome complexity reduction that effectively supports SNP genotyping for the three species that we have studied, human, mouse, and *A. thaliana*. We have carried out a systematic comparison in humans of an *in silico* methodology for prediction of DOP-PCR products with experimental results, indicating that an optimized set of DOP-PCR products can be generated for species for which complete genomic sequence is available. Because DOP-PCR-based genome complexity reduction does not depend on the presence of an interspersed repetitive element, we anticipate that the conclusions reached in the studies reported here will generalize to any eukaryotic species.

Materials and Methods

DOP-PCRs. The DOP-PCR mix was the same for all species and all primers: 1 ng/μl genomic DNA, 0.2 mM dNTPs, 1.25 units of AmpliTaq DNA polymerase (Applied Biosystems), 3.0 μM degenerate primer, and 1× enzyme buffer with $MgCl_2$ at 1.5 mM final concentration. DOP-PCR cycling profiles for human, mouse, and *Arabidopsis* are presented in supporting information, which is published on the PNAS web site, www.pnas.org. The 5′ end sequence, CTCGAGNNNNNN, was the same for all 35 degenerate primers; the specific 3′ end sequences are given in Table 3, which is published as supporting information on the PNAS web site. Radioactive body-labeled DOP-PCR products were prepared as described above except that the reactions were performed with [α-^{33}P]dCTP as described in *Supporting Methods*, which is published as supporting information on the PNAS web site. The resulting reactions were separated on a 6% denaturing PAGE, dried, and exposed to x-ray films.

DOP-PCR Library Statistics. Forty-one different DOP-PCR product mixtures from human, mouse, and *Arabidopsis* were shotgun cloned in a plasmid vector as described in *Supporting Methods*. For each library, 96 or more colonies were analyzed. Sequencing data were analyzed manually or with the Phred/Phrap/Consed system (http://genome.washington.edu/index.html), and within each library the total number of unique sequences was counted

Abbreviations: SNP, single nucleotide polymorphism; IRS, interspersed repetitive sequences; ASO, allele-specific oligonucleotide; DOP, degenerate oligonucleotide primer.

Data deposition: The SNPs reported in this paper have been deposited in the NCBI dbSNP database (accession nos. ss3173385–ss3173615 and ss3176793–ss3176830).

‡To whom reprint requests should be addressed. E-mail: dhousman@mit.edu.

www.pnas.org/cgi/doi/10.1073/pnas.261710699

and the frequency of each unique sequence determined. The average length of the cloned human and *Arabidopsis* DNA fragments was estimated by agarose gel analysis. The average clone length for each mouse library was estimated from the lengths of the contigs containing overlapping reads from identical clones. Library size was estimated by solving for N in the equation: $xL/NL = 1 - (1 - 1/N)^m$, where n is the total number of unique clones in the library, assuming each clone is equally represented, m is the number of clones sampled from the library, x is the number of unique clones observed, and L is the average length in bp of observed clone inserts. For the human library with three data points, we used nonlinear regression to obtain the best fit to the above equation. In all species NL estimated the library complexities in base pairs. We estimated the percent coverage as $100^*(x/N)$.

SNP Genotyping. DOP-PCR product locus-specific primer pairs, sequences of which are in Table 4, which is published as supporting information on the PNAS web site, were used to PCR amplify a unique product in *A. thaliana* Columbia and Landsberg erecta; *Mus musculus* C57BL/6J and *Mus spretus* SPRET/Ei; or CEPH individuals 12-01, 104-01, 884-01, and 1331-01 (Coriell Cell Repositories), using a Touchdown cycling profile (see *Supporting Methods*). Gel-purified DNA fragments were cycle sequenced and putative SNPs were found by comparing sequences from two to four samples by visual inspection of PAGE bands, with POLYPHRED software (http://droog.mbt.washington.edu/PolyPhred.html), or with the LASERGENE SEQMAN II program (DNASTAR, Madison, WI). For each allele of a putative SNP, a 17-mer oligonucleotide, centered on the putative polymorphism (see Table 4) was synthesized, end-labeled with [γ-^{33}P]ATP by using polynucleotide kinase (NEB, Beverly, MA), and hybridized at 52°C in a tetramethyl ammonium chloride (TMAC)-based buffer (7) to Hybond N + membranes (Amersham Pharmacia Biotech) supporting DOP-PCR products. All validated SNP data were submitted to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/); ID's are in Tables 4–6.

Genetic and Physical Mapping of DOP-PCR SNPs. Mouse DOP-PCR products were mapped on the interspecific backcross BSS panel from The Jackson Laboratory (8) by DOP-PCR product Hybond N + membrane stamping and allele-specific oligonucleotide (ASO) hybridization as described above. The data has been deposited at The Jackson Laboratory web site database (www.jax.org). Thirty-seven *Arabidopsis* SNPs were genotyped directly from DOP-PCR product mixtures from each of 29 recombinant inbred (RI) lines (9), as described above. The data were submitted to the appropriate database (http://nasc.nott.ac.uk/new_ri_map.html). RI map ID's are provided in Table 6, which is published as supporting information on the PNAS web site. Physical mapping was accomplished by BLASTing the sequences obtained for each unique human and *Arabidopsis* clone against the NCBI human genome contig database build 22 (blastcl3 -d/hs_genome/contig), or the *Arabidopsis* sequences in GenBank on 05/07/01 (blastcl3 -d nr -l Arabidopsis_thaliana.n.gil). BLASTN searches using the network-client BLAST program, BLASTCL3, were performed on REDHAT LINUX 6.2. Human chromosome positions were obtained in base pairs by finding the start position of the BLAST hit contig (if localized) in the NCBI MAPVIEWER and adding the BLAST result query start position to the contig start position. Note that these are approximate and will continue to shift until the genome sequence is complete. We found the *Arabidopsis* chromosome positions by retrieving the BLAST hit clone (BAC, PAC, etc.) start position on the *Arabidopsis* Genome Initiative sequence map, from TAIR Clone Search (http://www.arabidopsis.org). We made quantile–quantile plots and histograms with the R statistics software.

Electronic DOP-PCR and Human SNP Identification. The sequence obtained for 172 unique clones was BLASTed against the NCBI dbSNP (blastcl3 -d/snp/snpch1.fas. . ./snp/snpchY.fas) and each potential hit was verified to see whether the SNP was within the bounds of the locus-specific primer pair. The sequence of the human genome was obtained from (http://genome.ucsc.edu/) and a script to search the genome for exact matches to a given 8-mer sequence was designed to identify whether a second match lies in opposite orientation and less than 2 kb away, thus predicting a DOP-PCR product. The sequence of each eDOP-PCR product was analyzed with REPEATMASKER as above, and BLASTed against dbSNP on 09/24/01. DOP-PCRs with primers e1, e2, and e3 were performed on human genomic DNA as described above and in *Supporting Methods*. Each DOP-PCR was diluted 1:10,000. Of this dilution, 0.4 μl was used in each 20-μl Touchdown PCR as described above with each pair of specific primers for 324 electronically predicted products (95 for each of DOP-PCR primers e1, e2, and e3; 39 for primer 8BJ).

Results

DOP-PCR Genotyping. We have developed a stepwise methodology by which SNPs are directly genotyped from several unique fragments that are amplified by a genome complexity reducing PCR. Figure 1*A* depicts the steps of DOP-PCR genome-wide SNP loci amplification and parallel genotyping. In the amplification step, a partially degenerate primer binds to many cognate sites throughout the genome and amplifies a product wherever two sites, in opposite orientation, lie close to each other. This reaction results in the amplification of a mixture of PCR products, many of which contain SNPs. These SNPs can then be genotyped directly from the DOP-PCR product mixture. One DOP-PCR allows detection of one subset of the SNPs from a genome.

DOP-PCR Design and Characteristics. Our initial design of the DOP-PCR primers for whole genome amplification included a C/G-rich 5′ anchor (CCGACTCGAG), 6 "N" where n = A, C, G, or T, resulting in a 4,096-fold degenerate region, and a 6-nt specific sequence at the 3′ end of the primer. We reasoned that the complexity of the product would be inversely related to the length of this specific 3′ end sequence. As such, we tested DOP-PCR primers with lengths of specific 3′ end sequence ranging from 6 to 10 nt (18–22 nt overall) for their ability to amplify human genomic DNA (Fig. 2*A*). The oligonucleotide primer containing 6 nt of specific sequence at the 3′ end resulted in a number of DOP-PCR products so high as to produce a smear on a denaturing PAGE. However, the primers with 8 and 10 nt of specific sequence at the 3′ end led to a discrete number of products. We estimated the complexity of the product mixtures, amplified with primers having 3′ specific sequences of 8 to 10 nt, at a few hundred unique products.

A second parameter of the complexity of a given PCR product is the average size of each amplified fragment. Fragment size is affected by the DOP-PCR cycling parameters. DOP-PCR uses a two-part cycling program: 5 cycles with a very low annealing temperature, then 35 cycles at a higher annealing temperature. The initial low temperature cycles are necessary to allow primer binding, whereas the later cycles improve specificity and yield (5). We observed that lowering annealing temperatures and/or shortening extension times could shift the size range of the amplified products downward. We optimized the DOP-PCR cycling times and temperatures for each species so that fragments in the size range of ≈200 to 1,500 bp were amplified. Using the average fragment size and the estimated number of unique fragments amplified in each DOP-PCR, we estimated the complexities of the DOP-PCRs by using the 8 and 10 3′ end nucleotide primer that we tested to be in the range of 50,000 to 500,000 unique base pairs. We next examined the applicability of

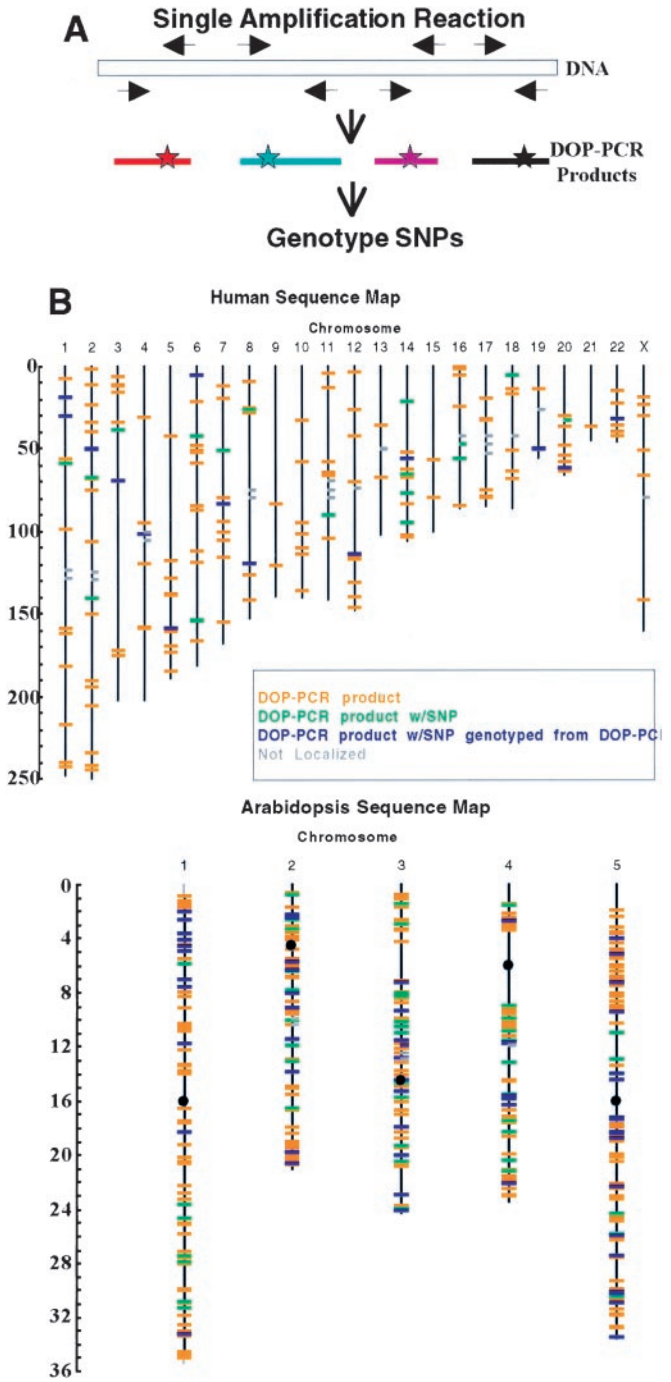

Fig. 1. (*A*) Simultaneous amplification of SNP-based markers by DOP-PCR. Schematic representation of a DOP-PCR reaction amplifies hundreds of products, distributed throughout the genome. Many of the products contain SNPs (depicted by stars), which were genotyped directly from the DOP-PCR products. The horizontal arrows represent degenerate primer binding sites. (*B*) Sequence-based mapping of human and *Arabidopsis* DOP-PCR products. The sequences of DOP-PCR products were compared with the genomic sequence of human and *Arabidopsis* to find their physical map location. The scales are in million bp. The largest gaps on *Arabidopsis* chromosomes 1–5 are 2.5, 1.2, 2.8, 4.5, and 2.8 million bp, respectively.

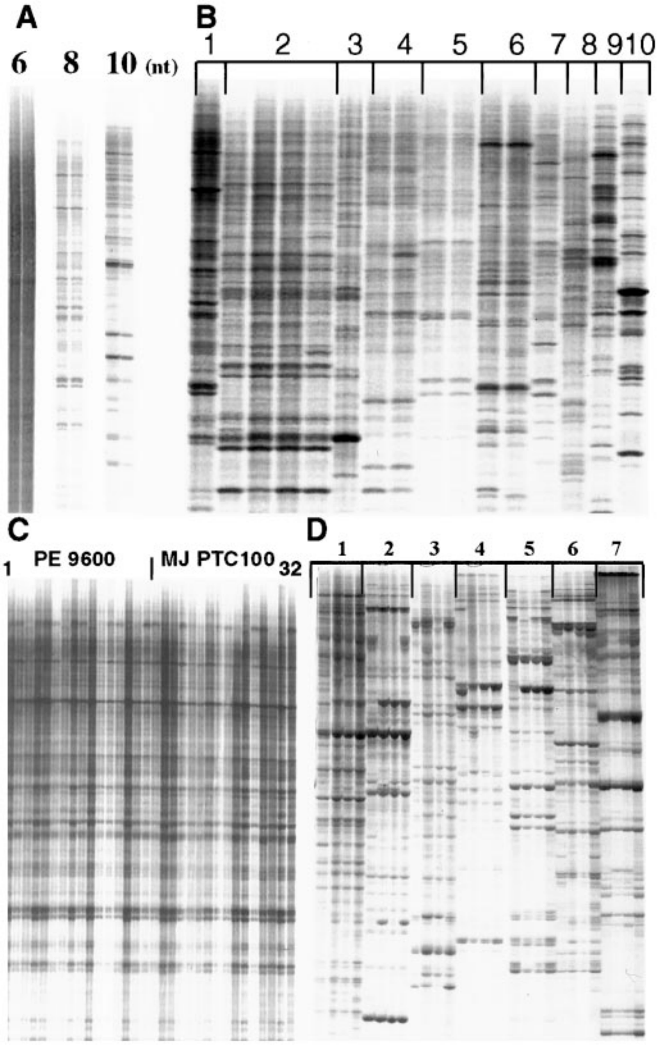

Fig. 2. DOP-PCR characteristics. (*A*–*D*) 6% denaturing polyacrylamide gel electrophoresis (PAGE) of [α-^{33}P]dCTP body-labeled DOP-PCR products. The size range shown for all panels is from approximately 250 bp up to the base of the loading wells. (*A*) Changing the length of the DOP-PCR primer 3′ end results in different complexities of amplified product sets. We amplified two human genomic DNA samples each with DOP-PCR primers of 3′ end length of 6, 8, and 10 nt. (*B*) DOP-PCR is applicable to any species. We used the same 3′ end 8-nt primer as in *A* and conditions (human cycling profile) to amplify a DOP-PCR product set from each of 16 genomic DNA samples representing ten different species: 1, *Homo sapiens*; 2, *Mus musculus*, C57BL/6J, 129/SvJ, AKR/J, BALB/cJ; 3, *Mus spretus*, SPRET/Ei; 4, *Gorilla gorilla*; 5, *Pan troglodytes* (common chimpanzee); 6, *Pongo pygmaeus* (orangutan); 7, *Pan paniscus* (Bonobo chimpanzee); 8, *Danio rerio* (zebrafish); 9, *A. thaliana*, Columbia; 10, *Saccharomyces cerevisiae*. (*C*) DOP-PCR is highly reproducible among genomic DNA samples. Thirty-two unrelated human genomic DNA samples from the CEPH reference panel (http://www.cephb.fr/) were amplified with a DOP-PCR primer. We amplified 16 of the samples in a Perkin–Elmer 9600 thermal cycler and the other 16 samples in an MJ Research PTC100 (Cambridge, MA), using the same primer and cycling conditions. (*D*) Changing the 3′ end specific sequence of the DOP-PCR primer resulted in different amplified product sets. We amplified four *Arabidopsis* recombinant inbred (RI) line genomic DNA samples with seven DOP-PCR primers that each had a different 3′ end sequence.

DOP-PCR amplification to a broad range of species (Fig. 2*B*). We expected that DOP-PCR primers with 3′ end specific sequences of 8 nt or more would amplify a series of specific targets in any complex genomic DNA. The product complexity for each species was reduced to a set of discrete fragments that could be resolved by PAGE. Each of the ten species gave a distinct PAGE banding pattern, whereas different individuals of the same species gave identical PAGE banding patterns.

A critical factor for the use of a reduced complexity genome-wide

Table 1. Summary of 41 DOP-PCR libraries

	Primer	*m*	*n*	*x*	*N*	*L*	bp	%G:C	%IRS	%LCS	Total%
Human	8	352	352	271	602	400	240,800	50.47	14.15	0.61	14.76
Arabidopsis	8–10	1,110	1,015	493	79	400	31,600	42.85	4.94	0.26	5.20
Mouse	8.5	656	549	307	211	504	106,344	49.02	23.08	2.69	25.76
	9	2,488	2,257	1,302	291	434	126,294	47.43	20.06	1.69	21.75
	9.5	1,031	1,005	683	429	409	175,461	48.49	19.38	1.74	21.12
	10	867	801	441	215	434	93,310	48.77	14.90	1.70	16.59
	11	2,769	2,300	1,175	145	368	53,360	48.75	16.77	1.97	18.73
	12	1,428	1,350	694	154	316	48,664	49.02	14.03	1.31	15.34
	8.5–12	9,239	8,262	4,602	216	391	87,029	48.50	17.44	1.78	19.22

Each library is described by the DOP-PCR primer 3′ end length. One library was made from a human DOP-PCR done with a primer having an 8-mer on the 3′ end. For *Arabidopsis* the results of eight different libraries, made with primers having 8–10-mers on the 3′ end, were summed (*m*, *x*) or averaged. For 32 mouse libraries, the libraries were grouped by primer 3′ end length and the results summed or averaged within each group. An 8.5-mer is a 9-mer with a degenerate base at the 3′-most position; i.e. A or T = W, C or G = S. The total number of clones that were sampled = *m*; total number of unique clones observed = *x*; estimated number of unique clones per library = *N*; approximate average clone insert length (bp) = *L*; %GC content of sequenced clones; %IRS + %LCS = Total%.

PCR amplification for SNP genotyping is the reproducibility of the PCR. Fig. 2*C* shows the results obtained with a given DOP-PCR on 32 human genomic DNA samples from two different thermocyclers. Similar results were obtained when using DNA from the other species (data not shown). Detailed comparison of the products revealed that an identical banding pattern is seen in each individual. We then tested the effect of slight variations in the PCR cycling parameters. Changing the denaturing step from 30 to 60 seconds or changing the primer annealing temperature by 1°C was generally tolerated; however, variations beyond these ranges affected the results (data not shown). The reactions were also scaleable; volumes of 10 μl to 100 μl were amplified successfully (data not shown).

We wished to determine the potential for complete genome coverage by DOP-PCR. We therefore determined the extent to which different DOP-PCR primers amplify different product mixtures. We tested several degenerate primers, each with a different 3′ end specific sequence, to determine the uniqueness or similarity of amplified product mixtures. Fig. 2*D* shows the PAGE banding pattern for each of seven DOP-PCR primers with *Arabidopsis* DNA. Each pattern is unique. Sequence sampling from libraries made from a series of such products confirmed that the sequences represented in the PCR product mixture for a given DOP-PCR primer in a given species were distinct (see next sections).

Libraries of Cloned DOP-PCR Products. DOP-PCR product complexity is the most important variable for direct SNP genotyping. If the complexity is too low, then few SNP-containing fragments are amplified. As the complexity approaches that of the whole genome, the efficiency and accuracy of SNP genotyping becomes problematic. To quantitatively determine the effect of DOP-PCR primer length on complexity, we analyzed plasmid libraries of DOP-PCR from three species. One human, eight *Arabidopsis*, and 32 mouse DOP-PCRs libraries were analyzed for sequence content and complexity. DNA sequences were compared with each other within a given library, so that the number of unique fragments, *x*, observed in a library of *m* sequenced clones could be counted for each DOP-PCR (Table 1). The number of times each unique cloned fragment was observed in a library of size *m* was also recorded (Fig. 3). These data were used to estimate the total number of unique clones, *N*, which is a measure of the library complexity (see *Materials and Methods*).

In the human DOP-PCR library of 352 sequenced clones, 271 or 77% contained unique fragments. Only three fragments were present more than four times, and the most frequently represented fragment accounted for only 3.4% of the clones (Fig. 3). By plotting the number of observed unique clones as a function of the number of clones sampled and assuming a randomly cloned library, our data follows the theoretically expected exponential curve (data not shown). A sampling size of 352 clones falls on the linear phase of the curve and corresponds to only about 45% coverage of the estimated library size of 602 unique clones.

Among the *Arabidopsis* DOP-PCR libraries, 31–58% (*x*/*m**100) of the sequenced clones in any one library contained unique fragments and 44% of the observed clones were unique when all eight libraries were combined (Table 1). In two of the eight *Arabidopsis* libraries there were one or two clones observed >20 times each. These high frequency clones were amplified from chloroplast DNA or contained a long stretch of interspersed repetitive sequences (IRS), or were very short and were subject to preferential cloning. The *Arabidopsis* DOP-PCR libraries were on average much less complex than the mammalian libraries; $N_{Arabidopsis} = 76$ vs. $N_{Mammals} = 219$. This was not surprising because the *Arabidopsis* genome size is about 1/30th of the size of a mammalian genome. However, the *Arabidopsis* library complexities were more than 1/30th of the mammalian complexities, because the DOP-PCR cycling parameters were adjusted to increase the fraction of the *Arabidopsis* genome amplified, relative to the fraction of the mammalian genomes amplified. Sampling coverage of the *Arabidopsis* libraries was estimated at 81%—i.e., sequencing 1,110 clones yielded an estimated 81% of the unique clones present in the eight DOP-PCR libraries.

In the mouse, we chose 32 different primers and a set of cycling parameters that, on average, amplified product mixtures of lower

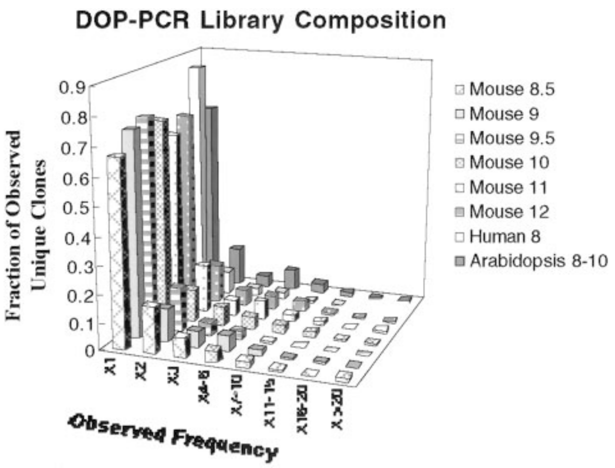

Fig. 3. Characteristics of DOP-PCR libraries. Clone frequency distribution in libraries of cloned DOP-PCR products. The libraries are grouped by primer 3′ end length and/or species.

complexity than the human DOP-PCR. The 32 primers and resulting DOP-PCR libraries could be grouped by primer 3′ end length. Six 3′ end lengths, 8.5, 9, 9.5, 10, 11, and 12 nt, were tested. Decreased library complexity was observed with increased primer length (Table 1). Among all 32 mouse DOP-PCR libraries, the number of sequenced clones, *m*, averaged 289 with a range of 129 to 371, and the percentage of unique clones observed within a library averaged 50% with a range of 28% to 79%. Sixteen of 32 DOP-PCR libraries contained at least one clone that was observed >20 times. The sampling coverage of the mouse libraries was 70% overall.

%GC and IRS Content of DOP-PCR Products. The %GC content of the sequence from the unique cloned PCR fragments is presented in Table 1. Each DOP-PCR library contained sequences with a %GC content somewhat higher than the overall content of the genome from which it was derived. The human and mouse DOP-PCR libraries contained about 50% GC, whereas the overall mammalian genome composition is estimated at between 41% (10) and 43% (11). The *Arabidopsis* libraries contained about 43% GC, whereas the *Arabidopsis* genome contains about 36% GC (12, 13). The unique cloned PCR fragment sequences from each DOP-PCR library were analyzed with the repetitive element filter software REPEATMASKER (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker), to identify IRS and low complexity sequences (LCS). The *Arabidopsis* libraries contained about 5.2% IRS and LCS; the human library contained about 14.8% IRS and LCS, and the mouse libraries contained on average 19.2% IRS and LCS. These percentages were significantly lower than whole-genome estimates of at least 45% for human (10), roughly 50% for mouse (14), and 10% for *Arabidopsis* (15). We chose 3′ specific sequences that avoided known repetitive elements in each genome that we believe led to the low percentage of IRS and LCS in our DOP-PCR products.

SNP Identification and Genotyping in DOP-PCR Products. We tested the feasibility of genotyping SNPs directly from the mixture of DNA sequences amplified in a DOP-PCR. First, we collected samples of SNPs from fragments amplified in each of nine DOP-PCRs from human, mouse, and *Arabidopsis*. From this sequence data 98, 38, and 279 SNPs were identified of 172 human, 21 mouse, and 304 *Arabidopsis* DOP-PCR products. We then validated these putative SNPs by performing ASO hybridization on the locus-specific PCR product containing each SNP from each individual, ecotype, or strain (see *Materials and Methods*). Using ASO hybridization in a tetramethyl ammonium chloride (TMAC)-based buffer (see *Materials and Methods*), we validated 54 human, 38 mouse, and 209 *Arabidopsis* SNPs. Detailed information and public database ID's for each of the 269 validated SNPs found in single-copy sequence are provided in Tables 4–6.

We then tested whether each validated SNP could be genotyped directly from the DOP-PCR product mixtures. Of the 269 validated SNPs from all species, 48% were successfully genotyped directly from the DOP-PCR product mixture under a single set of ASO hybridization conditions. Fifty-five percent of the PCR products from all species contained at least one SNP that was successfully genotyped directly on a DOP-PCR product mixture by ASO hybridization. Lastly, we ascertained the reliability of genotyping SNPs directly from the DOP-PCR. As such, we successfully genotyped each of 37 unique *Arabidopsis* SNPs on a panel of 31 independent samples (data not shown), thus demonstrating the robustness of direct SNP genotyping on DOP-PCR products.

Genetic Mapping of Mouse and *Arabidopsis* SNP Loci. Sixteen mouse and 62 *Arabidopsis* SNP-containing DOP-PCR product loci were genetically assigned a chromosomal location. The 16 mouse DOP-PCR products mapped to 11 different chromosomes (data not shown). The *Arabidopsis* SNPs were mapped by genotyping the Lister and Dean (9) recombinant inbred (RI) lines. The public database ID's and map locations are provided for each SNP in Tables 5 and 6.

Table 2. Description of electronically predicted DOP-PCR product sets

	e1	e2	e3
eDOP-PCR Primer			
8-mer sequence	GATACAGC	ATCATCCC	TTGAGGTG
No. predicted products	503	1,008	1,731
dbSNPs in products	140	364	612
% GC of predicted products	40.67	42.48	42.22
% IRS of predicted products	30.69	31.47	33.64
Product generated on			
Genomic DNA	90	94	91
Uncycled DOP-PCR	0	0	1
Cycled DOP-PCR	71	76	67
% on cycled DOP-PCR	78.9%	80.9%	73.6%

Experimental results show that on average 78% of the predicted products are actually amplified in DOP-PCR reactions with human genomic DNA.

Physical Mapping of DOP-PCR Products. We assigned a physical map location for each DOP-PCR product based on its presence in the genomic sequence for either human or *Arabidopsis*. DOP-PCR products were well distributed over the human and *Arabidopsis* genomes (Fig. 1*B*). Of the 271 unique human DOP-PCR products that we analyzed, we were able to place 200 on the working draft sequence by searching the NCBI human genome contig database for exact matches to our sequences. We located all 480 unique *Arabidopsis* DOP-PCR products on the *Arabidopsis* Genome Initiative sequence map (15). Four hundred and five *Arabidopsis* products had a single exact match hit, the locations of which are shown in Fig. 1*B*.

We wished to characterize the distribution of the DOP-PCR products in the human and *Arabidopsis* genomes. If the DOP-PCR markers were randomly spaced, then the intermarker distances would fit an exponential distribution, and a quantile–quantile plot of the intermarker distances vs. an exponential distribution with the sample mean would produce a straight line. A χ^2 test produced *P* values of 0.3 for the human and 0.1 for the *Arabidopsis* data. As such, neither dataset deviated significantly from the expected exponential distribution, suggesting that these DOP-PCR products are randomly distributed over the genomes.

Electronic DOP-PCR and SNP Identification. We next experimented with developing a sequencing-independent, *in silico* method to identify loci that are amplified in a given human DOP-PCR. We have developed a script to simulate a DOP-PCR electronically. By scanning the genome for the occurrence of a given arbitrary sequence (i.e., DOP-PCR primer) present twice in a head-to-head fashion within a distance that can be amplified (<2 kb), we hypothesized that we could predict which loci would be amplified in a given DOP-PCR. To investigate this approach, we compiled predicted DOP-PCR product sets for 1,000 different 8-mer primers having a GC content of 50%. The median number of predicted products for each primer was 1494, but ranged from zero to over 10,000. We determined the GC and IRS content of a sample of the 1,000 predicted product sets to be 35.9% and 42.6%, respectively. We chose three of the 1,000 8-mer electronic DOP-PCR primers for further analysis based on their predicted product set sizes (500–2000 products) and the %GC and %IRS content of their predicted product sets, which were close to the average percentages for all 1,000 predicted product sets (Table

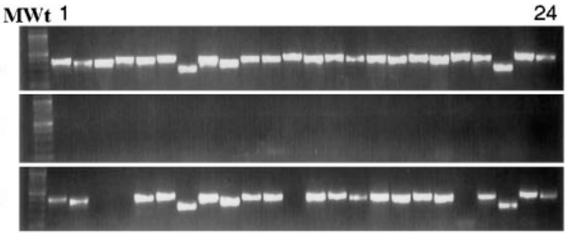

Fig. 4. Electronically predicted products are amplified in DOP-PCRs with genomic DNA. Specific primers designed to amplify predicted products from three electronic DOP-PCRs (e1–e3) were tested for their ability to amplify genuine products from genomic DNA (row A-positive control), a 1:10,000 dilution of an uncycled, mock DOP-PCR (row B-negative control for amplifying any remaining genomic DNA), and a 1:10,000 dilution of cycled DOP-PCR (row C). Results for 24 predicted DOP-PCR products (from primer e1) are shown, and similar results were obtained with 300 other primer pairs for predicted products (data not shown). MWt, DNA standard.

2). We assessed the randomness of the predicted products' distribution over the human genome by using the quantile–quantile plot method described above. χ^2 tests produced *P* values of 0.83 for set e1, 0.14 for set e2, and 0.03 for set e3.

We next determined whether the products predicted by electronic DOP-PCR were indeed amplified in DOP-PCRs. We performed DOP-PCRs with the above three primers on human genomic DNA (Fig. 4). Approximately 78% of the predicted products were amplified in the three DOP-PCRs we tested. To demonstrate that the amplification was also specific, we tested the specific primer pairs for each DOP-PCR to see whether they could amplify products from a different DOP-PCR. We found only four weakly "cross-reacting" primer pairs of the 275 tested, showing that each DOP-PCR amplified a specific set of products (data not shown).

Discussion

Genome-wide association studies and large-scale linkage analysis present significant logistical challenges. Kruglyak and Nickerson (16) and Kwok (17) have suggested that a SNP genotyping strategy based on whole-genome amplification may be critical to meeting this challenge. We present here the analysis of such an approach.

For human SNP genotyping, the optimal embodiment of the strategy we present will begin with an *in silico* analysis designed to optimize the number of SNP loci genotypes while minimizing the number of PCR products required. A critical strategic parameter in this analysis will be the sensitivity of the genotyping method used. In the current study, we have used ASO hybridizations as a standard genotyping method. However, the strategy presented here is compatible with many available high-throughput SNP genotyping platforms including high-density variation–detection DNA chips (4), single base extension (SBE) reactions (18), arrayed primer extension (APEX) (19), and bead-based ASO hybridizations (20) that have the sensitivity to independently genotype hundreds of SNP loci in a complex PCR product mixture. The use of any such high-sensitivity SNP genotyping method would allow the completion of a genome-wide SNP scan with several hundred DOP PCR products.

Complexity reduction by DOP-PCR in conjunction with *in silico* genomic analysis could clearly facilitate array based resequencing projects designed for initial SNP discovery. For examples, methods currently applied to a single human chromosome (21) could be extended by an integrated strategy in which *in silico* genomic analysis is used to design DOP-PCR primers and resequencing microarrays to efficiently cover an entire genome.

The range of applications for complexity reduction-based, genome-wide genotyping extends well beyond genotyping for human SNP association studies. Complexity reduction would reduce the time and cost involved in any genotyping project that could be accomplished with randomly distributed SNPs. These projects include linkage mapping in model organisms, marker-assisted selection programs to improve commercially valuable plant and animal species, and forensic and agricultural DNA fingerprinting. Additionally, because no prior sequence information is required to use DOP-PCR to amplify multiple loci from a genome, DOP-PCR is a method by which a set of SNPs can be collected from a species with minimal prior genome characterization.

The U.S. Department of Energy Joint Genome Institute sequenced the mouse DOP-PCR products. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48, and by Los Alamos National Laboratory under contract No. W-7405-ENG-36. We thank Frederick Ausubel for use of his greenhouse facilities; Damien Burke for computation and system administration; Chris Burge for statistical advice; Beth Nickerson for the ape DNA; and Peter Reddien, James Weber, David Page, and Linda Engle for helpful comments on the manuscript. This work was supported by Grant HG00299-21 from the National Institutes of Health (to D.E.H.), the Virginia and D. K. Ludwig Fund for Cancer Research fellowship (to B.J.), the Terry Fox Postdoctoral Fellowship from the National Cancer Institute of Canada and a Medical Research Council of Canada Postdoctoral Fellowship (to A.C.), and Grant 5 F32 CA73118 from the National Cancer Institute (to J.E.L.).

1. Risch, N. & Merikangas, K. (1996) *Science* **273,** 1516–1517.
2. Kruglyak, L. (1999) *Nat. Genet.* **22,** 139–144.
3. Kwok, P. Y. (2001) *Annu. Rev. Genomics Hum. Genet.* **2,** 235–258.
4. Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., *et al.* (1998) *Science* **280,** 1077–1082.
5. Telenius, H., Carter, N. P., Bebb, C. E., Nordenskjold, M., Ponder, B. A. & Tunnacliffe, A. (1992) *Genomics* **13,** 718–725.
6. Cheung, V. G. & Nelson, S. F. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14676–14679.
7. Handelin, B. & Shuber, A. P. (1994) in *Current Protocols*, eds. Dracopoli, N. C., Haines, J. L., Korf, B. R., Moir, D. T., Morton, C. C., Seidman, C. E., Seidman, J. G. & Smith, D. R. (Current Protocols, New York), pp. 9.4.1–9.4.8.
8. Rowe, L. B., Nadeau, J. H., Turner, R., Frankel, W. N., Letts, V. A., Eppig, J. T., Ko, M. S., Thurston, S. J. & Birkenmeier, E. H. (1994) *Mamm. Genome* **5,** 253–274.
9. Lister, C. & Dean, C. (1993) *Plant J.* **4,** 745–750.
10. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature (London)* **409,** 860–921.
11. Bernardi, G. (2000) *Gene* **259,** 31–43.
12. Carels, N. & Bernardi, G. (2000) *FEBS Lett.* **472,** 302–306.
13. Nekrutenko, A. & Li, W. H. (2000) *Genome Res.* **10,** 1986–1995.
14. Silver, L. M. (1995) *Mouse Genetics: Concepts and Applications* (Oxford Univ. Press, New York).
15. Initiative, A. G. (2000) *Nature (London)* **408,** 796–815.
16. Kruglyak, L. & Nickerson, D. A. (2001) *Nat. Genet.* **27,** 234–236.
17. Kwok, P. Y. (2001) *Science* **294,** 1719–1723.
18. Hirschhorn, J. N., Sklar, P., Lindblad-Toh, K., Lim, Y. M., Ruiz-Gutierrez, M., Bolk, S., Langhorst, B., Schaffner, S., Winchester, E. & Lander, E. S. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 12164–12169.
19. Kurg, A., Tonisson, N., Georgiou, I., Shumaker, J. & Metspalu, A. (2000) *Genet. Test.* **4,** 1–7.
20. Chen, J., Iannone, M. A., Li, M. S., Taylor, J. D., Rivers, P., Nelsen, A. J., Slentz-Kesler, K. A., Roses, A. & Weiner, M. P. (2000) *Genome Res.* **10,** 549–557.
21. Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., *et al.* (2001) *Science* **294,** 1719–1723.