Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 7514–7519, June 1998
Genetics

Computerized polymorphic marker identification: Experimental validation and a predicted human polymorphism catalog

JOHN W. FONDON III*, GINA M. MELE†, RUTH I. BREZINSCHEK†, DONNA CUMMINGS*, ASHWINI PANDE*, JONATHAN WREN*, KEVIN M. O'BRIEN*, KENNETH C. KUPFER*, MING-HUI WEI‡, MICHAEL LERMAN‡, JOHN D. MINNA†, AND HAROLD R. GARNER*§

*McDermott Center for Human Growth and Development and the Center for Biomedical Inventions, and †Hamon Center for Therapeutic Oncology Research, The University of Texas Southwestern Medical Center, Dallas, TX 75235; and ‡Laboratory of Immunobiology, National Cancer Institute, Frederick Cancer Research and Development Center, Frederick, MD 21702

Communicated by Michael S. Brown, University of Texas Southwestern, Dallas, TX, May 4, 1998 (received for review January 15, 1998)

ABSTRACT **A computational system for the prediction of polymorphic loci directly and efficiently from human genomic sequence was developed and verified. A suite of programs, collectively called POMPOUS (polymorphic marker prediction of ubiquitous simple sequences) detects tandem repeats ranging from dinucleotides up to 250 mers, scores them according to predicted level of polymorphism, and designs appropriate flanking primers for PCR amplification. This approach was validated on an approximately 750-kilobase region of human chromosome 3p21.3, involved in lung and breast carcinoma homozygous deletions. Target DNA from 36 paired B lymphoblastoid and lung cancer lines was amplified and allelotyped for 33 loci predicted by POMPOUS to be variable in repeat size. We found that among those 36 predominately Caucasian individuals 22 of the 33 (67%) predicted loci were polymorphic with an average heterozygosity of 0.42. Allele loss in this region was found in 27/36 (75%) of the tumor lines using these markers. POMPOUS provides the genetic researcher with an additional tool for the rapid and efficient identification of polymorphic markers, and through a World Wide Web site, investigators can use POMPOUS to identify polymorphic markers for their research. A catalog of 13,261 potential polymorphic markers and associated primer sets has been created from the analysis of 141,779,504 base pairs of human genomic sequence in GenBank. This data is available on our Web site (pompous.swmed.edu) and will be updated periodically as GenBank is expanded and algorithm accuracy is improved.**

The positional cloning of many disease loci has been facilitated by high-resolution genetic maps. The precise localization of the DNA sequence responsible for a disease usually requires the development of very high-density physical and genetic maps. The availability of multiple polymorphic genetic markers is crucial to this effort (1). Current widely used methods for the identification of new simple sequence repeat polymorphisms involve PCR-based and subcloning strategies (1, 2). Subcloning, once the primary method of isolating new microsatellite sequences (3), largely has been supplanted by PCR-based methods because of the relatively large amount of work and technical difficulties involved in subcloning strategies, including sublibrary construction and screening with oligonucleotide probes (4, 5). PCR-based strategies, although generally faster and more successful than those using subcloning, still suffer from several disadvantages. The researcher is unable to specify with much precision the size, homogeneity, or locations of the repeat loci identified and is forced to use complex or sophisticated methods to complete the process (6). In addition, many of the repeat loci identified are either redundant, difficult to cleanly amplify, or are otherwise unsuitable for use. These limitations restrict the ability of genetic researchers to create local high-density maps in their area of interest and force many to be dependent on large-scale mapping facilities.

The large quantities of human genomic sequence being generated by the Human Genome Project are rapidly making these approaches obsolete. It is anticipated that the Human Genome Project will be more than 95% complete by the year 2002 (7). The knowledge of the frequency and level of polymorphism of the various types and sizes of microsatellites and variable number tandem repeats allows one to predict, *a priori*, which tandem repeats are likely to be highly polymorphic from a single genomic sequence. For microsatellites, the level of heterozygosity has been observed to be directly proportional to the number of repeated units and inversely proportional to the size of the repeated unit (6, 8–12). Although several software applications for locating some microsatellites or larger tandem repeats currently exist, a comprehensive tool for the identification of, and generation of primer sequences for, those repeats correlated with a high probability of polymorphism has been lacking (13–23). Because of this need, software was written that will take human genomic sequence data as input and will output a list of oligonucleotide sequences that may be used as primers for PCR amplification of those tandem repeat sequences that are predicted to be highly polymorphic based on observations from the literature (Table 1). The 3p21.3 region has been found to be a site of frequent allele loss in lung, breast, and other cancers, as well as being the site of frequent homozygous deletions in lung and breast cancers. As part of a positional cloning effort to identify a putative tumor suppressor gene in this region we constructed a cosmid/P1 contig, which recently has been sequenced by the Washington University Human Genome Center (St. Louis) and the Sanger Centre (Hinxton, Cambridge, U.K.) (31). The closest flanking known polymorphic markers are MFD 93 centromeric and D3S15F2 telomeric. By using standard BLASTN analysis, various sequence-tagged sites and polymorphic markers previously unmapped can be identified, and the GenBank annotation documents some previously identified tandem repeats. We have discovered in this region on the known genetic linkage map that polymorphic markers would be of great use for further studies of allele loss in tumors and associated preneoplastic lesions as well as for placement of the genes. As a test of POMPOUS (polymorphic marker prediction of ubiquitous simple sequences), we applied this software suite to approximately 750 kilobases of genomic sequence from this human chromosome 3p21.3 region and then

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: POMPOUS, polymorphic marker prediction of ubiquitous simple sequences.

[§] To whom reprint requests should be addressed at: McDermott Center for Human Growth and Development and the Center for Biomedical Inventions, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75235-8591. e-mail: garner@swmed.edu.

Table 1. Minimum criteria for polymorphism

Type of repeat	Threshold number* of repeating units	Minimum homogeneity**
Dinucleotide	8†	0.9
Trinucleotide	7‡	0.9
Tetranucleotide	6§	0.9
5-mer–9-mer	6¶	0.9
10-mer–250-mer	4‖	<2 consecutive errors

*The threshold number refers to the minimum number of repeated units required for a locus to be scored as a likely polymorphism.
†Ref. 8.
‡Refs. 8, 11, 24–26.
§Refs. 8, 25–27.
¶Refs. 8, 25.
‖Refs. 12, 28–30.
**The minimum homogeneity refers to the minimum overall identity score for an alignment of the query sequence against the repeat sequence to which it is most similar.

determined the utility of the predicted primers in detecting length polymorphisms and allele loss in a panel of human tumor and normal DNAs.

MATERIALS AND METHODS

Computational Tools. All codes were written in either ANSI C, HP/C, or Perl and run on an HP/Convex Exemplar with eight processors and 512 MB of shared RAM. Database searches for simple repetitive elements in sequences were performed by using a parallel version of BLASTN (23). The accession numbers of the cosmid and P1 clone genomic DNA sequences analyzed are listed in the *Appendix*.

Genomic DNA. Genomic DNA was extracted by standard methods (32) from 36 human B lymphoblastoid cell lines and paired tumor lines (one mesothelioma, one breast carcinoma, 14 small cell lung carcinoma, and 20 nonsmall cell lung carcinoma).

Primers, PCR Conditions, and Allelotyping. Genomic DNA was amplified via PCR using the predicted primers (Table 2). Primers for this study were produced by using a MerMade high-throughput oligonucleotide synthesizer (S. Rayner, S. Brignac, R. Bumeister, Y. Belodludtsev, T. Ward, O. Grant, K.M.O., G. A. Evans, and H.R.G., unpublished work). Reactions were run in a Perkin–Elmer 9700 thermocycler with 1.5 mM $MgCl_2$, 200 μM each dNTP (GIBCO), 2 μCi α^{32}P-dCTP, 10 μM each primer, AmpliTaq 0.5 μg/μl (Perkin–Elmer). A touch-down PCR program with a modified hot start was used: initial denaturation at 94°C for 2.5 min followed by 10 cycles with denaturation at 94°C for 30 sec, annealing starting at 68°C and decreasing by 1°C for each cycle for 30 sec, and extension at 72°C for 30 sec followed by 25 cycles of denaturation at 94°C for 30 sec, annealing at 58°C for 30 sec, and extension at 72°C for 30 sec followed by a final extension at 72°C for 10 min. PCR products were heat-denatured, snap-chilled, and run on a denaturing 6.7% acrylamide gel (acrylamide/bis acrylamide ratio 19:1) with 10 M urea. Results were visualized by autoradiography using BioMax film (Kodak).

RESULTS

Computational Considerations. Current algorithms and software tools for the identification of simple sequence repeats, commonly performed to remove low-complexity elements from a sequence before computational analysis (filtering or masking), are based on one of two primary methods: database analysis or self-comparison of sequences (13–23) or sophisticated hybrids of

Table 2. PCR-allelotyped primers predicted by POMPOUS for region 3p21.3

Oligo name*	Forward primer†	Reverse primer†	Size	Sequence
LUCA1.1	CCTCATCCTCTCTGTTGGG	GCAGAATGACGTGAACCC	266	$(ct)_{15}$
LUCA1.2	TTCCTCCACAGATTCCTCTG	TTGAGAAGGATGTGGATGAC	277	$(ca)_{18}$
LUCA1.3	GTTCATGATGGCAGACTCTG	CCCAGAACAATCACAAGATG	259	$(ca)_{23}$
LUCA2.1	GCTCCTCAGGCAGAGTCC	CCACAGCCATCCACTGGAAGG	371	$(ca)_{14}(ca)_{14}$
LUCA2.2	GGCAGTGTTTGAGCTTACATGGG	CAGGCTCTGGAAACCAAGC	172	$(ca)_{29}$
LUCA4.1	CTGCCTGCCTCACTACTCC	CAAACTGCCCAGCTTCTG	182	$(ca)_{23}$
LUCA4.2	CCAGTGAGTACCAAGATATGGG	GAGAATGTGACCCGGGC	236	$(aaac)_6$
LUCA4.3	ACATGGTGGCATGTGCC	AGACCTTTAGCGTGTCATTTAC	239	$(aaaat)_6$
LUCA4.4	GATCATGGAGGCCTTGAATG	TGTTCATTGCCTGCATGG	407	$(ca)_{13}$
LUCA6.3	AAAGTACAGGGCAGTGTCAG	AGAAGAAGAGTCCAGCAAGC	480	$(aatg)_6$
LUCA7.1	GGAGATCACCTGGACACATC	TAAGTGCTAGGAGGCACTGC	229	$(agcccc)_7$
LUCA7.2	TTGTGGGCACTTCTTGCTAG	ACAACATATGCAAGGCCCAG	523	$(aatg)_6$
LUCA8.1	TAACCCTTGGCTTTGCTG	GGCTACTCTTCAACAAGATGC	274	$(ca)_{30}$
LUCA8.2	GCTGAGAAATCTCAATTGTGGGTG	GGCTGCTGAGCAGTGTCAGAC	125	$(ca)_{11}$
LUCA8.3	AATTCAGAACTGCGCCTTG	GCAGAGTGAGTAAAGGTTCACAG	485	$(aat)_{25}$
LUCA11.1	CACACACATCAGGCCGC	CTGCTCACAACCCCCTACC	252	$(ca)_{22}$
LUCA11.2	AAGACCAGCAAGGTGGC	TACTAGGGCTCCAGAGAGGG	238	$(ca)_{13}$
LUCA11.4	CCTGTGCCCACTCCCTATCTC	CCCAGGCTGGGATGCCTTTGG	177	$(ca)_{12}$
LUCA12.1	TTCTCTGAGGAATGACTCCTGG	AACAGCGCCTTATGATGGAG	308	$(cca)_8$
LUCA13.4	CCGCCTTCCAAGTTTAAGTG	TCAACCTGGGAGGTGGG	199	$(aat)_7$
LUCA17.1	GCTTGATCTTGGCTCACCTC	AGGATCACTTGGGCCTG	374	$(aaat)_8$
LUCA17.2	GTGGGTGGATCATCAGGTC	GTGCCTACAATAGAGAGGGTC	360	$(ttttg)_6$
LUCA17.4	GAACATGGAAGGAGGAGG	CTCGATCTATCTCTTGATCTCAG	220	$(aaat)_8$
LUCA19.1	AGGCAGGCTATATTCAACCC	GAAAAAGCCCCGGAACC	199	$(ta)_{15}$
LUCA19.2	AATTAGGACCTTCAGGAAACC	CCAGCAGATGGAGGTTGC	204	$(aaag)_6$
LUCA20.2	TGTGGTCCCACCTAATCTGG	CACATGCGTGCACATGC	304	$(aaag)_{11}$, $(aagg)_{11}$
LUCA22	GAAGTGGTTCACCCTAAGATC	AAGGGAGCTGGCCTAGGAG	235	$(ct)_{12}$
3938P1.2	AAACCTAGTTGATGCCAGGC	AACTTCTGGGATACATGGGC	413	$(aaac)_6$
3938P1.3	TGCCAGAGCCCCAAAAC	AAATGGGGCCAGGTGTG	416	$(ttttc)_6$
3938P1.4	TAAACCACTAGGCCCGGC	AGGTGTGGTGTTGTGTGCC	246	$(ttttg)_6$
3938P1.5	TGTTCCCTTCCCTATAAACAGAT	GAAAGCAAGGAAGGCACATG	167	$(ca)_{24}$
3938P1.6	CATGGTGGCACACAACTG	CCTCACCTGCCTGATTTTTG	310	$(aaat)_6$
3938P1.7	AGCAGATAGACAGATAAAGAGAGG	TCAGCTTCTTTGTCCCTTTCAC	211	$(10\ mer)_5$

*Oligo name represents the LUCA cosmid the primer maps to (see *Appendix*).
†Sequences are listed in a 5′ to 3′ orientation.

both methods (13). Database comparison methods for masking repeat sequences are preferred for microsatellite repeats because they are easily implemented by using available comparison programs such as BLASTN and are readily optimized to quickly yield meaningful results with a minimum of noise for small repeat sequences (23). These programs typically function by comparing the query sequence with a compiled database of simple sequence repeats and then excluding regions with significant similarity from subsequent genetic database queries. For larger repeating elements, such as minisatellites, self-comparison algorithms are more efficient because repeat databases grow exponentially with repeating element size. Because hypervariable tandem repeats can have repeating units with lengths ranging from two to greater than 50, we chose to use a dual strategy for the large scale, automated analysis of large tracts of genomic sequence, using each type of algorithm where it is most effective. These algorithms are combined with codes that evaluate the likelihood of polymorphism of the identified repeats and then design PCR primers for their amplification to give the researcher a comprehensive tool for length polymorphism identification.

For the strategy we used to detect and evaluate novel microsatellite repeats an exhaustive repeat database was required. A cursory examination of widely used repeat databases found that they are not exhaustive, but contain only those simple sequence repeats that have been previously described and cataloged. For odd repeat unit length, n, the number of nonrepetitive nonredundant sequences is given by:

$$\frac{4^n - \Sigma\ (4^f - \Sigma\ 4^s)}{2n}$$

for all f = factors of n excluding n, and for all s = factors of f excluding f.

The terms in the summation account for repeats that can be represented as repeats with shorter periods, whereas the n in the denominator accounts for all cyclic permutations and the 2 in the denominator accounts for their complements (for even n the calculation is complicated by the occurrence of sequences that are their own complements because simply dividing by 2 is no longer valid). For example, for repeating unit length $n = 9$ there are

$$\frac{4^9 - (4^3 - 4^1) - 4^1}{2(9)} = 14{,}560 \text{ nonredundant sequences,}$$

whereas the simple sequence repeat databases available contained less than 100. Available repeat sequence generation programs are unsuitable for any large n as they generate all 4^n sequences and compare all possible cyclic permutations of each sequence and their complements to all previously selected sequences, compiling a list of unique sequences. This algorithm, though direct, becomes unwieldy quickly [efficiency is $O(4^{2n})$] as it compares every possible sequence to every other sequence in the exponentially growing list of selected sequences, and this effectively limits its utility to small n. To alleviate this computational obstacle to determining the optimal value of n for the database-centered portion of our approach, we developed an algorithm that uses a branch and bound technique (13) to eliminate the computationally expensive cross-comparison used by other codes. It also has the added advantage of lending itself well to parallel processing. We wrote a program based on this algorithm, called SIMPLESEQ, to generate minimal exhaustive simple sequence databases for $n \leq 13$ on an eight-processor HP/Convex Exemplar. These data, once computed, were kept in a series of databases for $n \leq 8, n \leq 9, \ldots, n \leq 13$ for empirically determining the optimal database size for the database-centered portion of our approach.

A Perl script, TANDMIN, was written to take as input the results of a BLASTN of the sequence of interest against this simple repeat database and parse the high scoring hits to determine the repeat unit length, level of homogeneity, type of repeat, and the number of repeating units and evaluate these parameters against the criteria listed in Table 1. These minimum thresholds were drawn from the extensive literature on microsatellite and minisatellite repeat polymorphisms and represent a conservative extrapolation of the data found therein to reliably identify repeat sequence loci with a greater than 50% probability of having a heterozygosity (H) of at least 0.50 (see Table 1 for references). We ran TANDMIN on several cosmid and PAC genomic sequences by using each of the simple sequence repeat databases containing all repeats with length n less than or equal to 13 and found that TANDMIN run times increased dramatically when we went from a database representing all repeats with $n \leq 9$ to one with $n \leq 10$ (from ≈4 sec to several minutes), we therefore selected $n \leq 9$ as the optimal database size for use (Table 1). A large increase was expected because of details of the implementation, primarily because of our minimum required homogeneity threshold of 90% identity. To be scored as repeats, adjacent sequences must be at least 90% identical to database repeat core sequence. For repeat units of length 10 or greater, a single base mismatch in every repeat no longer drops the identity score below 90%.

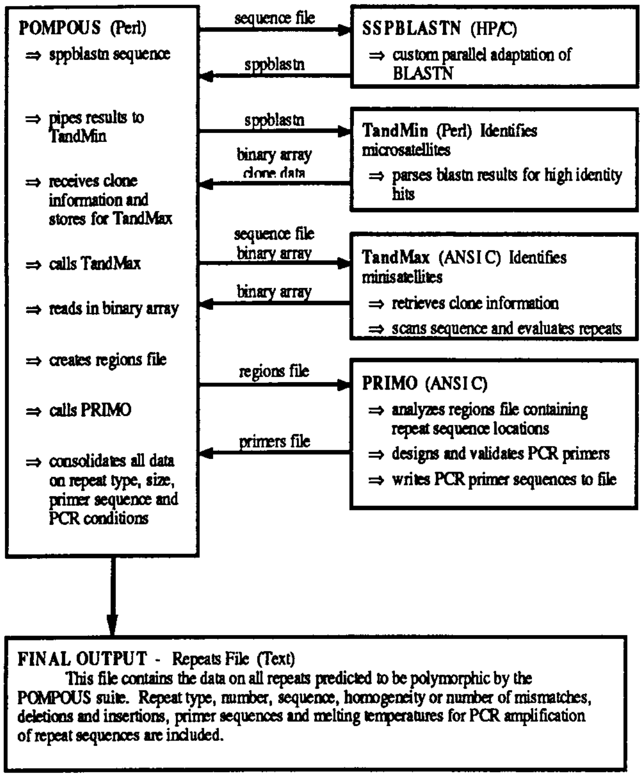
POMPOUS (Perl)
⇒ sppblastn sequence
⇒ pipes results to TandMin
⇒ receives clone information and stores for TandMax
⇒ calls TandMax
⇒ reads in binary array
⇒ creates regions file
⇒ calls PRIMO
⇒ consolidates all data on repeat type, size, primer sequence and PCR conditions
sequence file
sppblastn
sppblastn
binary array clone data
sequence file binary array
binary array
regions file
primers file
SSPBLASTN (HP/C)
⇒ custom parallel adaptation of BLASTN
TandMin (Perl) Identifies microsatellites
⇒ parses blastn results for high identity hits
TandMax (ANSI C) Identifies minisatellites
⇒ retrieves clone information
⇒ scans sequence and evaluates repeats
PRIMO (ANSI C)
⇒ analyzes regions file containing repeat sequence locations
⇒ designs and validates PCR primers
⇒ writes PCR primer sequences to file
FINAL OUTPUT - Repeats File (Text)
This file contains the data on all repeats predicted to be polymorphic by the POMPOUS suite. Repeat type, number, sequence, homogeneity or number of mismatches, deletions and insertions, primer sequences and melting temperatures for PCR amplification of repeat sequences are included.

FIG. 1. Program flow. POMPOUS takes as input a sequence of length up to 250,000 bases in FASTA format and processes it first by a parallel sequence homology search (sppblastn), which then is parsed by TANDMIN for hits with the properties listed in Table 1. The sequence then is scanned for larger tandem repeats by TANDMAX, and the outputs are consolidated. POMPOUS then converts the consolidated data into a format that is usable by PRIMO (regions file), and PRIMO uses this file to predict flanking oligonucleotide primers for subsequent PCR amplification.

For the detection of larger repeats the database approach is clearly not feasible (for $n = 37$ the minimal sequence set is 1.9×10^{22} sequences). To identify these longer repeats a second program, TANDMAX, was written that compares the query sequence to itself at offsets of 10–250 bases, looking for stretches of near identity (no more than two consecutive mismatches or indels) that are longer than the offset. This approach is analogous to graphic matrix analysis wherein the presence of diagonals close to the main diagonal in a binary comparison matrix is indicative of tandem repeats (14, 34). Those sequences satisfying these criteria then are analyzed further to determine the number of

Table 3. Heterozygosity rate for POMPOUS-predicted polymorphic markers

Marker	Repeat type*	Repeat homogeneity	No. of alleles	Rate of heterozygosity† (n = 36 individuals)
LUCA1.1	(ct)$_{15}$	0.97	2	0.42
LUCA1.2	(ca)$_{18}$	0.94	1	<0.03
LUCA1.3	(ca)$_{23}$	1.00	4	0.64
LUCA2.1	(ca)$_{14}$(ca)$_{14}$	0.97	4	0.42
LUCA2.2‡	(ca)$_{29}$	0.98	7	0.69
LUCA4.1	(ca)$_{23}$	1.00	4	0.64
LUCA4.2	(aaac)$_{6}$	1.00	1	<0.03
LUCA4.3	(aaaat)$_{6}$	0.94	1	<0.03
LUCA4.4	(ca)$_{13}$	0.92	1	<0.03
LUCA6.2	(15mer)$_{4}$	0.97	NA	No product
LUCA6.3	(aatg)$_{6}$	0.97	2	0.03
LUCA7.1	(agcccc)$_{7}$	0.95	3	0.08
LUCA7.2	(aatg)$_{6}$	0.96	2	0.03
LUCA7.3	(10 mer)$_{6}$	0.85	NA	No product
LUCA8.1§	(ca)$_{30}$	1.00	7	0.72
LUCA8.2	(ca)$_{11}$	1.00	3	0.53
LUCA8.3	(aat)$_{25}$	0.96	3	0.64
LUCA11.1	(ca)$_{22}$	1.00	4	0.50
LUCA11.2	(ca)$_{13}$	1.00	2	0.03
LUCA11.4	(ca)$_{12}$	1.00	2	0.22
LUCA12.1	(cca)$_{g}$	0.96	1	<0.03
LUCA13.1	(10 mer)$_{4}$	0.73	NA	Nonspecific product
LUCA13.4	(aat)$_{7}$	0.95	2	0.17
LUCA13.5	(ttg)$_{10}$	0.96	NA	Nonspecific product
LUCA17.1	(aaat)$_{g}$	0.91	1	<0.03
LUCA17.2	(ttttg)$_{6}$	1.00	3	0.31
LUCA17.3	(13 mer)$_{4}$	0.85	NA	Nonspecific product
LUCA17.4	(aaat)$_{8}$	1.00	1	<0.03
LUCA19.1	(ta)$_{15}$	0.93	6	0.86
LUCA19.2	(aaag)$_{6}$	0.92	1	<0.03
LUCA20.1	(aaag)$_{12}$, (aaag)$_{12}$, (10 mer)$_{6}$	1.00	NA	Nonspecific product
LUCA20.2	(aaag)$_{11}$, (aagg)$_{11}$	0.94	4	0.83
LUCA22	(ct)$_{12}$	0.96	1	<0.03
P1.1	(aat)$_{13}$	0.97	NA	Nonspecific product
P1.2	(aaac)$_{6}$	0.92	2	0.03
P1.3	(ttttc)$_{6}$	0.90	1	<0.03
P1.4	(ttttg)$_{6}$	0.96	2	0.58
P1.5	(ca)$_{24}$	1.00	5	0.83
P1.6	(aaat)$_{6}$	0.96	1	<0.03
P1.7	(10 mer)$_{5}$	0.82	2	0.03

NA, not available.

*(Repeat sequence)$_{\text{number of repeats}}$.

†Rate of heterozygosity = heterozygous cell lines/36.

‡Previously identified as K1.1CA.

§Previously identified as D3S1568(Z23748).

repeating units. Because large tandem repeats frequently have units deleted in the course of cloning and sequencing, the threshold number of repeats is reduced to four for minisatellites of repeating unit length from 10 to 250 nt. In addition, Jeffreys *et al.* (12) report that minisatellites appear to have a different mechanism of mutation than microsatellites that may not be dependent on copy number (see also ref. 30).

POMPOUS is a Perl script that automates the entire process (Fig. 1). It runs the parallel BLASTN (sppblastn) on the query sequence, sends the output to TANDMIN and then invokes TANDMAX. The outputs of TANDMIN and TANDMAX are integrated, consolidating any redundant or adjacent hits. POMPOUS then takes these results and translates them into the proper input format of the primer prediction software, PRIMO (35). POMPOUS executes PRIMO, which designs multiple flanking oligonucleotides for PCR amplification of all selected repeat regions. Each oligonucleotide is checked for optimal length, melting temperature, GC content, self-complementary, GC clamps, base quality (if available), and complementary to common human repeats (e.g., Alu, LINE1, THE). The final output is a file containing a list of repetitive sequences, statistics such as length and identity, and predicted primers and conditions that can be used for amplification of the repeat loci.

POMPOUS Identifies Polymorphic Loci. To test the effectiveness of POMPOUS in identifying polymorphic repeat loci that are useful in genetic studies, we ran the analysis on an approximately 750-kilobase region of human chromosome 3p21.3 (28). This region is involved in homozygous deletions in small cell lung carcinoma and breast cancer and is an area of intense search for a tumor suppressor gene (31, 36–39). POMPOUS predicted primer sets for 40 tandem repeat loci among the 19 cosmid and P1 clones. For this study we tested all 40 POMPOUS-predicted primer sets. Table 2 lists all primer sets yielding specific products. To determine the frequency of heterozygosity and utility of the POMPOUS predicted primer set we tested it for amplification, specificity, and heterozygosity in a panel of 36 predominately Caucasian patient DNAs, including matched B lymphoblastoid and tumor cell lines. In this way we were able to determine not only the heterozygosity of the selected loci, but also their utility in studies of allele loss. In addition to the usual controls (normal human DNA and a

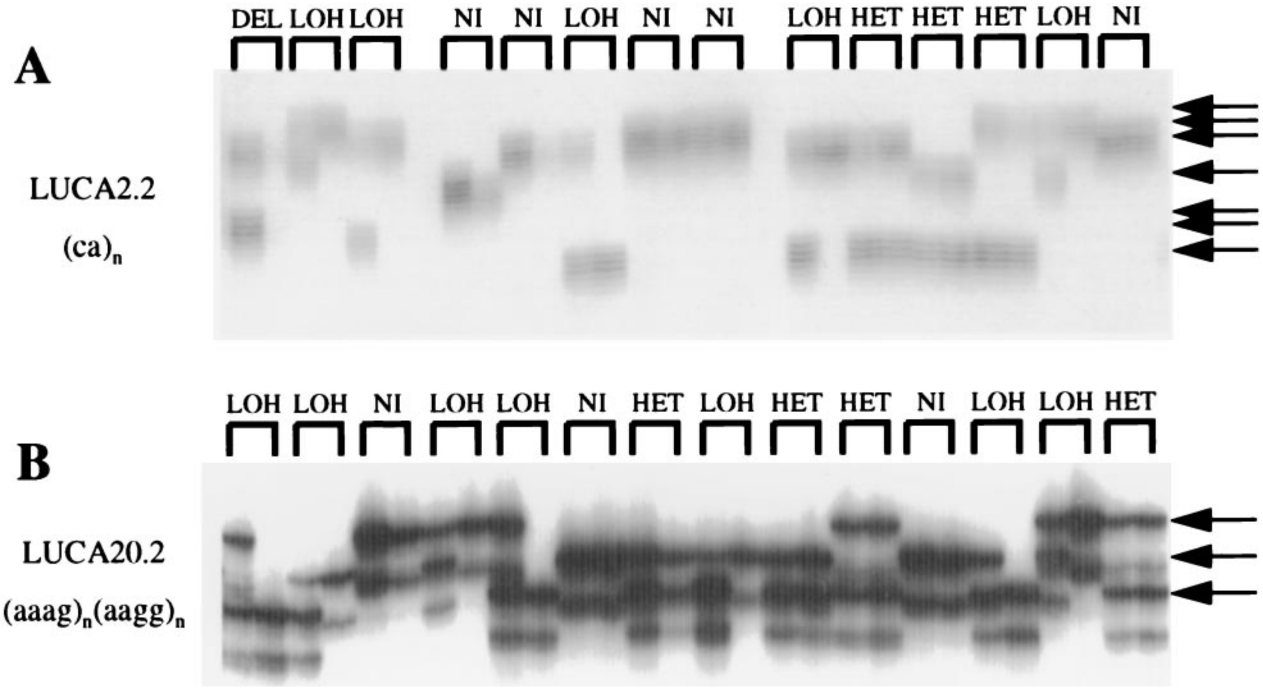
A
DEL LOH LOH NI NI LOH NI NI LOH HET HET HET LOH NI
LUCA2.2
(ca)n
B
LOH LOH NI LOH LOH NI HET LOH HET HET NI LOH LOH HET
LUCA20.2
(aaag)n(aagg)n

FIG. 2. Representative allelotyping results. Polyacrylamide allelotyping gels for paired B lymphoblastoid and tumor DNAs visualized by autoradiography. (*A*) Marker LUCA2.2, (ca)$_n$ repeat. (*B*) LUCA20.2, (aaag)$_n$(aagg)$_n$ repeat. The bracket above each lane contains DNA pairs from the same individual, first normal DNA and second tumor DNA. Arrows denote different alleles for each marker. DEL, H1450 deletion line; HET, retention of heterozygosity; LOH, loss of heterozygosity; NI, noninformative.

water blank), DNAs from two small cell lung cancer lines, NCI-H740 and H1450, were included that carry $\approx$1.5 Mbp and $\approx$.75 Mbp homozygous deletions within the 3p21.3 region, respectively. H1450 has a paired B lymphoblastoid line that also was included in these studies. These homozygous deletion DNAs provide a specificity check, as the predicted primers sets should yield no product in reactions with these templates. The results are summarized in Table 3. Thirty-three reactions yielded no product from the homozygously deleted tumor line DNAs (Fig. 2*A*, DEL) while giving clean product of the approximate expected size in matched B lymphoblastoid lines (Fig. 2). Two previously discovered polymorphic markers [K1.1CA and D3S1568(Z23748)] were identified and predicted by POMPOUS (Table 3). Primer sets yielding product from the homozygously deleted lines (five sets) or giving no product from normal lines (two sets) were excluded from further analysis (Table 3). Of the 33 loci yielding clean amplification patterns 22 were found to be polymorphic with heterozygosities ranging from 0.03 to 0.86 (mean H = 0.42) (Table 3). This gives an estimate of one marker for every 31 kilobases on average. All of the patients were heterozygous for at least one of the 22 informative markers (average of nine heterozygous loci per patient), and as predicted from many previous studies of 3p21.3, allele loss was found to be common, occurring in 75% of the tumor line DNAs.

We then went on to analyze 55,402 entries in GenBank comprising 141,779,504 bp of human genomic sequence selected to include the words "human" or "Sapiens" and exclude the word "mRNA" within the annotation of GenBank gbpri1 and gbpri2 databases. We downloaded these sequences from GenBank on March 16, 1998. POMPOUS identified 13,261 probable polymorphic markers in 4,669 of the GenBank entries. If 67% of the predicted markers are polymorphic as was reported in this study that would mean $\approx$8,885 potential additional markers have been identified, giving an average marker density of 1 marker per 16 kilobases. The actual number of markers undoubtedly will be less than this as many clones in the database contain overlapping or redundant sequences, and POMPOUS analyzes each sequence independently. Additionally, for efficiency reasons, PRIMO is not set to check primers for complementarity to genomic sequences other than common repeats, and this likely contributed to our observed PCR amplification failure rate of 17.5% (7/40). This catalog of putative polymorphic markers, primer sequences, and repeat characteristics, which is now available on our Web site (pompous.swmed.edu), is anticipated to be of value to the post-genomics community of researchers involved in gene identification and localization. Given that this sequence represents $\approx$4% of the human genome, and coding sequences are overrepresented in the accumulated GenBank database, at least 200,000 new polymorphic markers could be identified from the complete sequence now being generated by the Human Genome Project. Repeats with unit sizes from 2 to 229 were identified, for a total of 76 different unit sizes in the range of 2 to 250 checked by this code. As a percentage of the total repetitive DNA analyzed, the repeat unit sizes were: 50.0% di-, 10.2% tri-, 24.4% tetra-, 5.7% penta-, 2.2% 6 mer-9 mer, and 7.5% 10 mer-229 mer. This represents a total of 710,825 bp, or 0.5% of the GenBank human sequence being part of a repetitive unit that is potentially polymorphic.

DISCUSSION

The identification of polymorphic genetic markers from the huge quantities of genomic sequence generated by the Human Genome Project is necessary for the genetic researcher to localize specific sequences associated with phenotypes and disease. Although massively parallel single nucleotide polymorphism genotyping technologies hold great promise for the future of genetic mapping, they currently cannot be predicted from a single genomic sequence, and their identification relies on large-scale resequencing of genomic DNA from multiple individuals. Length polymorphisms therefore are still the preferred markers for genetic linkage studies. We have developed a suite of programs, called POMPOUS, to predict putative length polymorphisms directly from a single copy of genomic DNA sequence. POMPOUS is designed to detect tandem repeats with repeating unit length ranging from 2 to 250, select those that are likely to be highly polymorphic, and design flanking oligonucleotide primers for use in PCR amplification. To test the accuracy of POMPOUS we allelotyped 33 of these predicted loci in B lymphoblastoid cell DNAs from 36 individuals (72 chromosomes) to determine their levels of polymorphism.

We found 22 to be polymorphic with heterozygosity rates ranging from 0.03 to 0.86, with 11 having heterozygosities of at least 0.5. With no consanguinity and modest mutation rates, the probability that any one individual would be homozygous for these 11 loci is 3.4×10^{-6} and for all 22 polymorphic loci is 1.9×10^{-7}. Thus, this panel of polymorphic markers will be useful for tumor allelotyping studies as well as for accurately placing genes in this region on the human recombination map. As a first test of their utility as such we allelotyped the 36 matched tumor cell lines and found very frequent loss of heterozygosity of this region, consistent with the presence of a 3p21.3 tumor suppressor gene.

There was a difference in utility of loci identified by TANDMIN (repeat unit less than 10 bases) and those identified by TANDMAX (repeat unit of length 10 or longer). Longer repeats were more difficult to cleanly amplify, largely because of weak homology to the repeats in the flanking sequences from which primers were designed by PRIMO. Some degenerate highly repeated sequences were identified by POMPOUS (including a minisatellite with a 49-bp core sequence repeated approximately 35 times), but not scored as likely polymorphisms because no more than three consecutive repeats were highly conserved. We currently are developing a more sophisticated algorithm for the identification of longer repeats and modifying the primer selection process to minimize these effects. Additionally, because of the fundamentally different algorithm and criteria used in their detection and selection, longer repeats also had identities well below those of repeats scored by TANDMIN (Table 3). As a more statistically significant quantity of data becomes available, the threshold criteria will be refined to make POMPOUS more efficient at selecting only the most useful loci.

Future studies will allow precise mapping of very small areas of allele loss as have been found for preneoplastic bronchial epithelial lesions (33). Likewise, by comparing the particular alleles at multiple loci in different tumors with normal tissues and individuals, it will be possible to define allelotype sets that are associated with pathogenicity. This information will be useful for future recombination analysis, genetic epidemiology studies, and predisposition diagnosis.

APPENDIX

Database Availability. POMPOUS has been run for all human genomic sequences in GenBank as of March 16, 1998 (GenBank release v105.0). The results are downloadable as a database and will be expanded to include all of GenBank and will be updated monthly as GenBank expands. All results are available for FTP downloading from our Web site (pompous.swmed.edu). In addition, a POMPOUS server will be available at that site for the analysis of any submitted sequences, with the submission and results conducted through the World Wide Web.

Accession/Contig Numbers. Analyses were performed on the following sequences: Luca1, Z74618; Luca2, Z77852; Luca3, Z74023; Luca4, Z74019; Luca5, Z74582; Luca6, Luca06.00286, Luca06.01215, Luca06.01317, Luca06.01351, Luca06.01396 (all Luca6 contigs have accession no. Z84493); Luca7, LUCA7.00598, LUCA7.00720 (all Luca7 contigs have accession no. Z84494); Luca8, Z84495; Luca9, Z75743; Luca10, Z75742; Luca11, LUCA11.00789, LUCA11.00039, LUCA11.01287 (all Luca11 contigs have accession no. Z84492); Luca12, AC002481; Luca13, AC002455; Luca14, U73167; Luca15, U73166; Luca16,U73169; Luca17, AC002077; Luca19, AC000063; Luca20.Contig17, Luca20.Contig18 (St. Louis Genome Center); Luca22, U73168; 3938P1.Contig16 (St. Louis Genome Center). Marker LUCA11.1 is in contig LUCA11.00789, and markers LUCA11.2–11.4 are located in contig LUCA11.00039.

We thank Hewlett Packard, specifically David Valenta, David Field, and Don North, for use of the Exemplar computer on which this work was performed. We thank Dr. Glen Evans for providing space for the Exemplar computer and Dr. Evans and Ron Bumeister for providing some of the oligonucleotide primers used in this study. We also thank Dr. Chris Davies, Peter Schilling, Dr. Adi Gazdar, Tina Chen, Dr. Carmen Behrens, Dr. Ivan Wistuba, Dr. Arvind Virmani, and Victor Stastny for many useful discussions and reagents. This work was supported by the Whitaker Foundation, the P. O'Brien Montgomery Distinguished Chair in Developmental Biology, National Cancer Institute Grants CA 71618 and SPORE P50 CA70907, Texas Advanced Research Projects Grant 003660–080, and the G. Harold and Leila T. Mathers Charitable Foundation.

1. Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., *et al.* (1996) *Nature (London)* **38,** 152–154.
2. Chen, H., Pulido, J. C. & Duyk, G. M. (1995) *Genomics* **25,** 1–8.
3. Malo, D., Vidal, S. M., Hu, J., Skamene, E. & Gros, P. (1993) *Genomics* **16,** 655–663.
4. Cornelis, F., Hashimoto, L., Loveridge, J., MacCarthy, A., Buckle, V., Julier, C. & Bell, J. (1992) *Genomics* **13,** 820–825.
5. Ostrander, E. A., Jong, P. M., Rine, J. & Duyk, G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 3419–3423.
6. Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. & Lathrop, M. (1992) *Nature (London)* **359,** 794–801.
7. Marshall, E. & Pennisi, E. (1996) *Science* **272,** 188–189.
8. Charmley, P., Concannon, P., Hood, L. & Rowen, L. (1995) *Genomics* **29,** 760–765.
9. Iizuka, M., Makino, R., Sekiya, T. & Hayashi, K. (1993) *Genet. Anal. Tech. Appl.* **10,** 2–5.
10. Weber, J. L. (1990) *Genomics* **7,** 524–530.
11. Gastier, J. M., Pulido, J. C., Sunden, S., Brody, T., Buetow, K. H., Murray, J. C., Weber, J. L., Hudson, T. J., Sheffield, V. C. & Duyk, G. M. (1995) *Hum. Mol. Genet.* **4,** 1829–1836.
12. Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985) *Nature (London)* **314,** 67–73.
13. Benson, G. & Waterman, M. S. (1994) *Nucleic Acids Res.* **22,** 4828–4836.
14. Claverie, J.-M. & States, D. J. (1993) *Comput. Chem.* **17,** 191–201.
15. Wooton, J. C. & Federhen, S. (1993) *Comput. Chem.* **17,** 149–163.
16. Brutlag, D. L., Dautricourt, J.-P., Diaz, R., Fier, J., Moxon, B. & Stamm, R. (1993) *Comput. Chem.* **17,** 203–207.
17. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
18. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441.
19. Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E. & Karlin, S. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 2002–2006.
20. Shpaer, E. (1996) *Methods in Molecular Biology Sequence Data Analysis Guidebook*, ed. Swindell, S. (Humana, Totowa, NJ).
21. Hancock, J. M. & Armstrong, J. S. (1994) *Comput. Appl. Biosci.* **10,** 67–70.
22. Staden, R. (1996) *Mol. Biotechnol.* **5,** 233–241.
23. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
24. Watkins, W. S., Bamshad, M. & Jorde, L. B. (1995) *Hum. Mol. Genet.* **4,** 1485–1491.
25. Hammond, H. A., Jin, L., Zhong, Y., Caskey, C. T. & Chakraborty, R. (1994) *Am. J. Hum. Genet.* **55,** 175–189.
26. Sheffield, V. C., Weber, J. L., Buetow, K. H., Murray, J. C., Even, D. A., Wiles, K., Gastier, J. M., Pulido, J. C., Yandava, C., Sunden, S. L., *et al.* (1995) *Hum. Mol. Genet.* **4,** 1837–1843.
27. Liu, Y., Rasool, O., Grander, D., Lindblom, A. & Einhorn, S. (1995) *Hum. Mol. Genet.* **4,** 727–729.
28. Armour, J. A. L. & Jeffreys, A. J. (1992) *Curr. Opin. Genet. Dev.* **2,** 850–856.
29. Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. & White, R. (1987) *Science* **235,** 1616–1622.
30. Andreassen, R., Egeland, T. & Olaisen, B. (1996) *Am. J. Hum. Genet.* **59,** 360–367.
31. Wei, M.-H., Latif, F., Bader, S., Kashuba, V., Chen, J.-Y., Duh, F.-M., Sekido, Y., Lee, C. C., Geil, L., Kuzmin, I., *et al.* (1996) *Cancer Res.* **56,** 1487–1492.
32. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd Ed.
33. Chung, G. T., Sundaresan, V., Hasleton, P., Rudd, R., Taylor, R. & Rabbitts, P. H. (1995) *Oncogene* **11,** 2591–2598.
34. Maizel, J. V., Jr. & Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 7665–7669.
35. Li, P., Kupfer, K. C., Davies, C. J., Burbee, D., Evans, G. A. & Garner, H. R. (1997) *Genomics* **40,** 476–485.
36. Daly, M. C., Xiang, R.-H., Buchhagen, D., Hensel, C. H., Garcia, D. K., Killary, A. M., Minna, J. D. & Naylor, S. L. (1993) *Oncogene* **8,** 1721–1729.
37. Todd, S., Franklin, W. A., Varella-Garcia, M., Kennedy, T., Hilliker, C. E., Jr., Hahner, L., Anderson, M., Wiest, J. S., Drabkin, H. A. & Gemmill, R. M. (1997) *Cancer Res.* **57,** 1344–1352.
38. Kok, K., Naylor, S. L. & Buys, C. H. C. M. (1997) *Advances in Cancer Research*, eds. Klein, G. & Vande Woude, G. F. (Academic, San Diego), Vol. 71, pp. 27–92.
39. Sekido, Y., Ahmadian, M., Wistuba, I. I., Latif, F., Bader, S., Wei, M.-H., Duh, F.-H., Gazdar, A. F., Lerman, M. I. & Minna, J. D. (1998) *Oncogene* **15,** in press.