ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

GENOMICS

Genomics 82 (2003) 10–19

www.elsevier.com/locate/ygeno

An exhaustive DNA micro-satellite map of the human genome using high performance computing

Jack R. Collins,[a,1] Robert M. Stephens,[a,1] Bert Gold,[b,1] Bill Long,[c] Michael Dean,[b,*] and Stanley K. Burt[a]

[a] *Advanced Biomedical Computing Center, NCI-Frederick, Frederick, MD, USA*
[b] *Laboratory of Genomic Diversity, NCI-Frederick, Frederick, MD, USA*
[c] *Cray Inc., Mendota Heights, MN, USA*

Received 31 July 2002; accepted 27 February 2003

Abstract

The current pace of the generation of sequence data requires the development of software tools that can rapidly provide full annotation of the data. We have developed a new method for rapid sequence comparison using the exact match algorithm without repeat masking. As a demonstration, we have identified all perfect simple tandem repeats (STR) within the draft sequence of the human genome. The STR elements (chromosome, position, length and repeat subunit) have been placed into a relational database. Repeat flanking sequence is also publicly accessible at http://grid.abcc.ncifcrf.gov. To illustrate the utility of this complete set of STR elements, we documented the increased density of potentially polymorphic markers throughout the genome. The new STR markers may be useful in disease association studies because so many STR elements manifest multiallelic polymorphism. Also, because triplet repeat expansions are important for human disease etiology, we identified trinucleotide repeats that exist within exons of known genes. This resulted in a list that includes all 14 genes known to undergo polynucleotide expansion, and 48 additional candidates. Several of these are non-polyglutamine triplet repeats. Other examinations of the STR database demonstrated repeats spanning splice junctions and identified SNPs within repeat elements.

Introduction

Availability of raw draft and finished sequence data necessitates increasingly complex computational methods for data annotation. As the data acquisition rate increases, there is a considerable lag between data production and data mining, including the annotation of sequence features and comparative genomics. In an effort to decrease this lag period, we have developed a rapid method for sequence comparison, initially using special hardware in the Cray SV1 computer architecture. Later experiments permitted us to reproduce our supercomputer results using conventional processors. As sequence comparison forms the basis of most genome annotation, this method should be generally useful in achieving the goal of decreasing annotation lag time.

Polymorphism annotation is a requirement for investigations of human predisposition and susceptibility for disease. STR repeat annotation is central to successful association mapping. Microsatellite and short tandem repeat (STR) maps have been highly valued since their discovery early in the genome project, because these features are highly abundant and polymorphic [1]. Microsatellites usually consist of a 2–6 nucleotide long core motif, while minisatellites may be much longer [2,3]. Satellite markers have been used to simplify complex biological problems such as estimation of the spatio-temporal relationships between chromosome segments [4,5], evolutionary relationships between species [6], to provide a basis for DNA fingerprinting [7,8], and allow for population genetic studies. Microsatellite polymorphisms have also been implicated in the causes of disease states such as fragile sites in chromosomes [5,9] and certain neurodegenerative disorders attributed to repeat sequence instability [10,11].

* Corresponding author. Bldg. 560, Rm. 21-18, Laboratory of Genomic Diversity, NCI-Frederick, Frederick, MD 21702. Fax: +1-301-846-1909.
E-mail address: dean@ncifcrf.gov (M. Dean).
[1] These authors contributed equally to this study.

0888-7543/03/$ – see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0888-7543(03)00076-4

Many computational procedures have been advanced to detect STRs in DNA sequences over the past 10 years. Among these are alignment matrices [12–14], alignment matrices coupled with data compression algorithms [15,16], heuristics [17–19], k-tuples [20] and direct measures [21]. Until recently however, only a k-tuple triplet repeat finder was able to deal with sequences up to millions of nucleotides [22]. To be generally useful, computational repeat finder procedures should require no prior knowledge of the repeat patterns being searched, should be exhaustive, and should be very fast. The commonly used methods are not exhaustive and because of practical computational constraints are not generally feasible for sequences of the scale of eukaryotic genomes.

In our implementation of the exact match algorithm, we initially took advantage of the unique hardware instructions and rapid vector register comparison of the Cray SV1 to detect exact tandem repeats for sequences over 3 billion bases in length (manuscript in preparation). In the current analysis, we have focused on a subset of exact matches that are at least nine repeats units long, containing between two and sixteen nucleotides in the motif. We have chosen these stringent criteria in order to maximize the potential for polymorphism among the newly delineated markers [2]. In addition, among triplet repeats within genes, we assert that patterns longer than nine within coding sequences will likely have significance for human triplet repeat expansion diseases [10,11]. In this communication, we provide new, powerful reagents likely to display polymorphism for human gene mapping and, in addition, we suggest a series of potential new triplet repeat disease loci.

Materials and methods

High performance STR finder

Preliminary investigations were conducted using unique vector register instructions available on the Cray SV1 (Cray, Inc, Mendota Heights, MN). The initial exact match STR finding procedure uses these features to compress the sequence data into two bits per nucleotide. Since the Cray vector register contained 64 elements of 32 nucleotides each, we compared 64 different starting positions per vector iteration. This level of fine-grain parallelism led to the tremendous speed advantage in this implementation for finding exact patterns. A detailed description of the algorithm is being submitted elsewhere. Since the tasks for each pattern size are independent, up to 15 processors were used quite effectively. Using 15 processors and the 64-element vector registers effectively resulted in a parallel performance of approximately 960 over a single scalar processor of the same speed. Our tandem repeat program was run interactively on 8 processors of the NCI SV1 in a full production environment. The elapsed time for running the repeat finder on the entire set of 24 chromosomes in the human genome took approximately 5 minutes. Once the repeated patterns were recorded, along with position and length, a program that filters or clusters them for further analysis processes the patterns. The nature of the filtering or clustering depends on the specific application. In the case of the STR elements, repeats that were in any way subsets of already identified repeats were removed from the list.

C language implementation

Subsequent to several successful whole genome tandem repeat extractions using the high performance capability described above, one of us (Collins) wrote a C-implementation of an exact match algorithm. In brief, this program reads sequences in FASTA format, and begins by comparing an initial two nucleotides to each successive string, in order to count whether a minimum number of repeats, span, and mismatch criteria are met. Whether the initial sequence is accepted as meeting STR criterion or not, an increment moves the comparison to the next character in the string, and so on. The C-code implementation of the algorithm is rapid and efficient: Chromosome 22 can be exhaustively searched for tandem repeats of size 2–16 in about a minute, using a minimum repeat criterion of 6 on an 800 MHz Powerbook. Executables are available on the tandem repeats web page: http://ncisgi.ncifcrf.gov/~collinsj/Tandem_Repeats/downloads/

Genome data analysis

For the data analysis, the UCSC (genome.ucsc.edu) datasets from December 2000, April 2001 and August 2001 were all analyzed. We have also analyzed the NCBI NT contig dataset as of October 2001. In all cases the fasta format files were downloaded from the source and then analyzed with the processing method described above, producing a separate output file for each input sequence (chromosome). The derived database consists of all repeat elements with the following criteria (repeat length:minimum number of elements) 2:6, 3:6, 4:6, 5:5, 6:4, 7:4, 8:3, 9:3, 10:3, 11:3, 12:2, 13:2, 14:2, 15:2, 16:2. The resulting data was then placed into a MySQL database containing the repeats along with additional information pertaining to other genome features including genes, STS and SNP marker positions. This additional genome information was derived from the downloaded UCSC genome database and stored in a slightly different schema to allow for more detailed SQL queries of the positional information. This data can be accessed for the December 2000, April 2001 and August 2001 datasets at http://grid.abcc.ncifcrf.gov.

Map density analysis

A program that determines minimum distance between the closest marker of any particular class and each element in a second input file was written in C. This program was used to determine the average minimum distance between each

marker type and each exon for each chromosome. A second program, also in C, that simply scans one input file for gaps and then records the number of elements in a second file that fall within that gap was also used to determine whether STR elements can be found within gaps in the STS or SNP map. Data were calculated from these programs using either the August release of data for STS or the April set of data for SNPs (SNPs were not yet in the August freeze). For the STR data, either the full data set or the subset of elements repeated 9 or more times was used (see text).

Electronic validation protocol

In order to assess the polymorphism status of these STR elements we have devised a simple protocol that searches the entire GenBank database using BLAST with each STR being scanned. Our protocol does an initial BLAST using the 5′ flanking unique region (after repeat masking) to generate a list of sequences with homology to that region in the database. Relatively high stringency scores are used to ensure that only very good alignments will be identified. A second BLAST run then uses the full query sequence (both flanking sequences and the repeat region) to identify sequences in the subset that span the repeat region. An even more stringent cutoff score is used in the second step to ensure that identified sequences contain at least some of each unique region in addition to the repeat sequence. The resulting sequences are then aligned and analyzed for inserts and deletions within the repeat. Subsequent evaluation procedures only consider deletions and insertions as polymorphic and thus exclude simple SNPs as polymorphic. Also, as a result of blast gap penalty issues, elements with significant insertions or deletions do not register using this method. We are currently enhancing the protocol to detect these situations.

Denaturing HPLC and sequencing

Primers were designed to putative coding triplet repeats and used to amplify fragments from unrelated humans. Samples from a heterogeneous set of individuals (8 Caucasian CEPH parents, 4 Asian and 4 African Americans) were run on denaturing HPLC using a Varian Helix column and detected by UV absorbance. Putative polymorphic samples were sequenced using ABI Big Dye chemistry and run on an ABI 310 sequencer.

Results

Whole genome statistics

A comparison of the total number of microsatellites described by our exhaustive search, Tandem Repeat Finder, and those within the Cooperative Human Linkage Consortium (CHLC) and Research Genetics databases is presented in Table 1. As can be seen in the Table, we found a total of 128,021 di-nucleotide repeats with a span of 9 repeats or greater. This compares with 52,865 found using tandem repeat finder (TRF, reference [22], perfect match subset); 5,757 well-characterized dinucleotide repeats in the CHLC database, and 6,570 such repeats in the Research Genetics catalog. Both CHLC and Research Genetics showed far fewer triplet repeats, yet our exhaustive effort reveals 8,741 of these. By downloading the CHLC (http://lpgws.nci.nih.gov/cgi-bin/ABI/ABIIntegratedMapsChr?v8c8.description.table) and Research Genetics data (ftp://ftp.resgen.com/pub/mappairs/human/Human.txt) and summing across all chromosomes, there are 1,453 tetranucleotide repeats denoted as useful for mapping in the CHLC database and 2,360 of these in the Research Genetics catalog. This is in contrast to the 23,676 tetranucleotide repeats resulting from our search and analysis. As can be seen from Table 1, there are a paucity of penta- and hexanucleotide repeats in the CHLC and Research Genetics databases while results from our computations yield 4,313 and 233, respectively. It is notable that we found approximately two and one-half times as many tetranucleotide repeats of repeat pattern length nine or more, as tri-nucleotide repeats. This trend is found in the TRF, CHLC, and Research Genetics data as well. Graphical analysis (not shown) demonstrates that di- and trinucleotide repeats show a steady, gradual extinction, as repeat span lengthens. However, tetranucleotide repeats are unique in that they show a maximum at repeat pattern number of 10, with an approximately equal number at repeat pattern length of 11. An analysis of all the repeat motifs 2 through 16 reveals that only the tetra-nucleotide

Table 1
STR database comparison

Tandem repeat length	ABCC	TRF	CHLC	Research genetics
2 (di)	128,021	52,865	5,757	6,570
3 (tri)	8,741	5,898	400	431
4 (tetra)	23,676	8,835	1,453	2,360
5 (penta)	4,313	1,193	8	6
6 (hexa)	233	73	0	2
7 (hepta)	10	6		
8 (octa)	14	2		
9 (ennea)	6	2		
10 (deca)	4	0		
11 (hendeca)	3	0		
12 (dodeca)	6	2		
13 (triskaideca)	7	0		
14 (tetrakaideca)	6	1		
15 (pentakaideca)	5	0		
16 (hexakaideca)	6	3		

The exhaustive database and TRF STRs compared repeats that had at least 9 units repeated tandemly. ABCC refers to our current exhaustive repeats database and TRF to the results from Tandem Repeats Finder [22] retaining only perfect repeats. CHLC (see text) and Research Genetics (see text) STRs consist of those polymorphic sequences, listed in their databases, summed over each chromosome. No repeats in the Research Genetics or CHLC/ABI Prism database were excluded, even if precise mapping coordinates are unknown.

repeats display this behavior. This observation is independent of any particular base-sequence subset of the repeats for each length (data not shown).

To demonstrate the utility of the newly characterized STRs described by our exhaustive search, we chose two approaches: (1) we demonstrated the application of the complete STR map in increasing the density of the physical map of the human genome, and (2) we queried the database for examples of trinucleotide repeats contained within coding regions of known genes.

Map marker density

One of the largest immediate contributions to genomic research made by this exhaustive collection of STRs with known chromosomal location and considerable potential for polymorphism is increased density of physical markers. In order to assess this, we determined the predicted increase in map density using our exhaustive dataset of STRs. Based on our STR data and the SNP and STS data from the UCSC browser, we calculated the minimum distance between each marker type and the known RefSeq exons within the genome.

As expected from the number of each type of marker and a uniform distribution of those markers along the genome, the SNPs showed the shortest average distance to an exon followed by the STR elements and then the STS markers. The average minimum distances for all chromosomes (excluding Y) were 2537 bp for the SNPs, 9365 bp for the STRs and 12500 for the STSs. If the STS markers used in the search are limited to those placed on either the Marshfield or Genethon maps, the average STS to gene distance increased to about 200 kbp. Chromosome Y displays large distances between STRs presumably due to the high levels of unknown sequences and low gene content. Our analysis reveals that on a global scale the STR marker set can be used to extend the density of markers over the existing STS set. This relative enhancement in density would be further increased if only the genetically mapped subset of STS markers was used (data not shown).

When we reduce the set of STR markers to include only those repeated at least 9 times, the average minimum distance between the STR markers and each exon is closer to that observed for the STS to exon distances; however, the maximum value for a minimum distance between STRs was still approximately one half that observed for the STS to exon distance maximum. Thus, even using a restricted set of STR elements more likely to be polymorphic, the map density near critical regions is likely increased by our work. Of course, the overall average distances between all marker element types and exons will decrease as the known gene set expands.

To further characterize the utility of the complete STR map in increasing known marker density, we wanted to determine whether STR markers could be used to bridge gaps in the known marker map. Casual observation of the localization of these genetic markers on chromosomal maps reveals that there are particular locations with considerable gaps even in the SNP marker set. These gaps are often large enough to contain entire genetic loci (250 kb). Thus, genetic analysis could be facilitated if the density in these areas could also be increased. In order to determine whether the STR markers would be helpful in these cases, we wrote a second computer program that searched through the STS and SNP data sets for each chromosome to identify gaps of over 250 kb and then looked for STR elements within those gaps. Over the entire genome, there were a total of 146 gaps of this size in the SNP marker set (a total of 231 Mb). Within these gaps, we have identified 4839 repeats that lie within them. For the less dense STS marker set (all known STS markers), we identified a total of 1,218 such gaps for a total of 625 Mb. Within these gaps, 63,517 STR elements have been identified. This analysis was done using the complete set of identified STR markers.

As a ‘proof of principle’ of the increased marker density from our STR collection, we have analyzed the marker density near the *BRCA1* locus on chromosome 17. The results from this analysis using the UCSC genome browser (mirrored at the ABCC at http://ucsc.abcc.ncifrf.gov) with the added track of the ABCC repeats are shown in Figure 1. In this case, we used the subset of STS markers with known genetic map identification and we only show STR elements which are repeated perfectly at least 9 times. Only one known polymorphic marker appears near the *BRCA1* gene on the current genetic map. However, our work elucidates 23 potentially polymorphic markers in this region that may find an important use in gene deletion analysis, or association studies. The presence of the additional potential STR markers in this region has been verified in both more recent releases of the public genome assembly and by examination of this region using Celera’s sequence assembly.

As a practical demonstration of the utility of the newly characterized STRs described by our exhaustive search, we characterized three STR markers in the vicinity of the *TP73* gene at 1p36.1. Two out of three of these showed sufficient polymorphism for association studies.

Repeats within coding regions

As a second measure of the utility of the complete STR map of the human genome, we also scanned the genome for those repeats that are contained entirely within the exons of known genes. Figure 2 Panel A shows a graph of the number of nucleotides that occur within repeats that are composed of 9 or more repeat units and are also within exons as defined in RefSeq. The analysis differentiates between internal exons, in those genes containing three or more exons, and exons that may contain either 3′ or 5′ UTRs (denoted in Figure 2 as 1st/last). As can be seen from the graph, there are no repeats longer than penta-nucleotides that occur within exons and meet the length criteria outlined above. Tetra-nucleotide repeats occur in several of the chro-

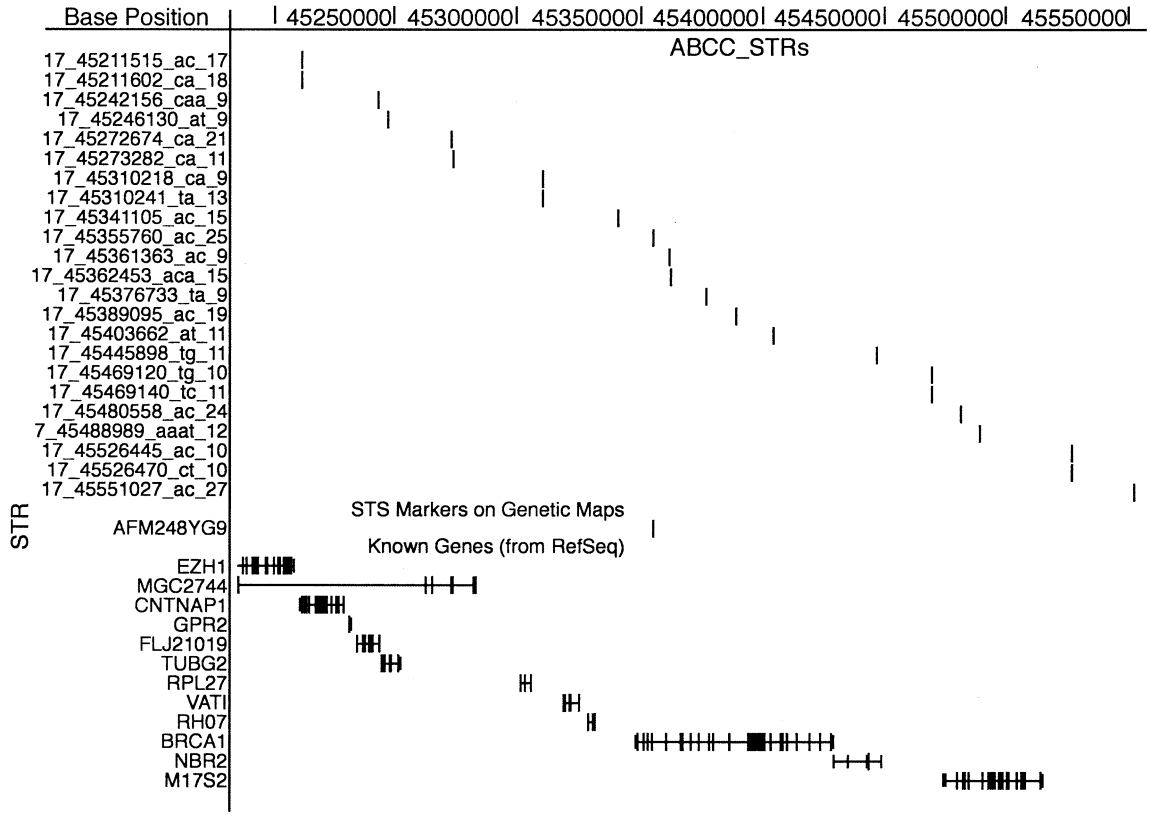

Fig. 1. STR marker comparison in the BRCA1 region. Only known polymorphic STS elements (map position indicated in either Marshfield or Genethon maps) are shown in this comparison with the GRID database described in this paper. The upper tracks delineate 23 putative polymorphic markers in a region where only Afm248YG9 was heretofore known.

mosomes while the only penta-nucleotide repeat is found in an internal exon on chromosome 11 in the *KCNJ5* gene, a cardiac potassium channel gene. As expected, the occurrence of repeats within internal exons is lower than 1st/last exons (Figure 2). One important statistic that is derived from our studies is the ratio of the repeats per base on the whole genome to repeats per base in exons. This ratio is approximately 30:1 for the entire draft human genome sequence (data available at the supplementary website). Consistent with earlier investigations, our data suggests that, given the known set of genes, simple tandem repeats may be selected against in regions containing exons. On the basis of this evolutionary reasoning, the di-, tri-, tetra-, and penta-nucleotide repeats within genes that are nine units or longer are of broad interest for further analysis. Figure 2 Panel B demonstrates the variability of the repeat sequence distribution by chromosome. Although the distribution of the number of nucleotides in repeats to the number of nucleotides per chromosome meets the test of a Gaussian Distribution, both human chromosomes X and Y are outliers. Chromosome X contains between 2 and 3 standard deviations more repeats per nucleotide than expected, while Chromosome Y contains more than 3 standard deviations of the 'normal' number of repeats per chromosome. This is consistent with much literature on the repetitive nature of human Chromosome Y. Supplementary annotation concerning our global characterization of human repeats is available at the Tandem Repeats Web page: http://ncisgi.ncifcrf.gov/~collinsj/Tandem_Repeats/

As part of our effort to mine this new database for new, potentially etiologic triplet repeat expansions, we developed the list of sixty-three triplet repeats shown in Table 2. The Fredrich's Ataxia expansion was not included in the list because it is intronic, even though it is present in the GRID database. In addition, the SCA8 and SCA12 polymorphisms were too short to pass our length criterion. After identifying the forty-nine novel triplet repeat sequences, we searched the LocusLink and OMIM databases for their function and to inquire whether polymorphism had previously been described. As can be seen from Table 2, nineteen of the newly identified intragenic triplet repeats were found to have been described as polymorphic or undergo expansion through articles cited in the OMIM database. We chose twelve of the newly identified putative triplet repeat expansions for experimental analysis. PCR primers flanking each of these putative, novel triplet repeat expansions were designed and amplification was attempted. In six out of the twelve cases, amplification was sufficiently robust for polymorphism analysis using Denaturing High Performance Liquid Chromatography (DHPLC). However, among the sixteen ethnically diverse individuals no polymorphism was observed in

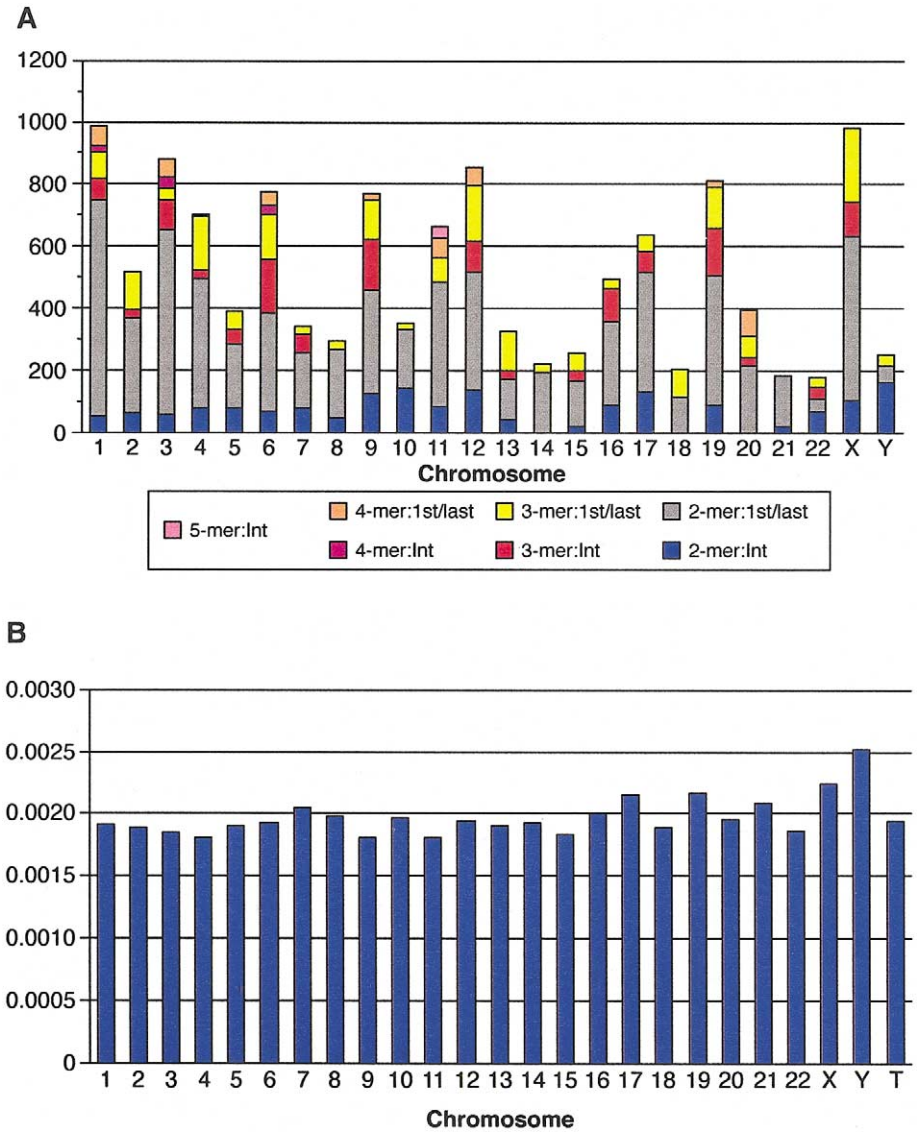

Fig. 2. Comparison by Chromosome of Short Tandem Repeat Content. Panel A. Number of nucleotides within STRs repeated nine or more times that occur within exons. The number of nucleotides contained within the set of STRs repeated 9 or more times and within exons of known genes is shown for each chromosome in the human genome. Of the entire set of repeat lengths 2–16, only di-, tri-, tetra-, and penta- nucleotide motifs had non-zero contributions. The exons are separated into distinct sets: Internal and 1st/last. Internal exons refer to those genes that are composed of 3 or more exons with internal referring to those that do not contain the 3′ or 5′ UTR regions. The set of 1st/last corresponds to those genes that have only 1 or 2 exons or the first and last exons of genes with 3 or more exons. Panel B. Total Repeat Content per Base: The ordinate represents the quotient of the number of nucleotides per chromosome present in the exact match repeats described in this report divided by the total number of nucleotides in each chromosome as reported in the August 2001 UCSC assembly. Chromosomes are displayed by number on the abscissa, with T standing for Total.

5 out of six of these amplifications. In the remaining amplification, that for *SMARCA2*, a mitosis control protein, a variety of triplet repeat polymorphism was observed, varying up to 18 repeats. Among the other newly identified putative triplet repeat expansions, *IGSF3* did not show polymorphism in our ethnic diversity panel; yet this newly identified expansion is also probably polymorphic, based on our observation that the BAC AL355794 contains six CCT's while the contig NT_004754 and contig NT_457126 contain 9 CCT's. Six reads in the Celera database all contain 6 CCT's (data not shown).

Electronic validation of polymorphism for mapping

As a means of identifying potential polymorphism within the identified triplet repeats, we developed a means to screen for virtual polymorphism within the existing database on sequence information. Since GenBank contains sequences derived from both messenger RNA (ESTs) and genomic fragments and represents data obtained from many different individuals and species, it provides a rich source of polymorphism information.

Our electronic validation procedure was tested using the

Table 2
Sixty-two identified and putative etiologic triplet repeat genes

Gene	Repeat sequence (span)	Chromosome location	Status	Disease or function	In silico validation
AR	**GGC (17)**	**X**	**Known**	**Kennedy disease**	**Yes**
ARMET	GGA (11)	3	Known	Arginine-rich protein	Yes
ASCL1	TGC (12)	12	Known	Neuroendocrine tumors	Yes
ATBF1	TTG (10)	16	New	AFP enhancer binding	No (4)
ATP7A	TTG (11)	X	Known	Cu transport ATPase	No (6)
CACNA1A	**TGC (13)**	**19**	**Known**	**SCA**	**No[a] (25)**
CACNA1A	**GGT (9)**	**19**	**Known**	**Migraine, SCA**	**No[a] (25)**
CAPN6	GCA (12)	X	New	Calpain 6	No (5)
CAPNS1	CCG (10)	19	New	Calpain, small subunit 1	No (5)
CBX4	GGT (10)	17	New	Protooncogene	No (3)
CD28	AAC (10)	2	New	Antigen CD28	No (3)
DIAPH1	CTC (10)	5	Known	Deafness	No (11)
DMPK	**CTG (11)**	**19**	**Known**	**Myotonic dystrophy**	**No (22)**
DRIL1	GTG (10)	19	New	*Drosophila* dead ringer homolog	No (4)
DRPLA	**TGC (15)**	**12**	**Known**	**DRPLA**	**Yes**
E2F4	CAG (13)	16	Known	Transcription factor	Yes
FMR1	**GCG (10)**	**X**	**Known**	**Fragile X MR**	**Yes**
FOXE1	CGC (10)	9	New	Thyroid dysfunction	Yes
FOXF2	GGC (9)	6	New	Lung transcription factor	No (0)
FRAXE	**GCG (15)**	**X**	**Known**	**Fragile XE MR**	**Yes**
FRDA	GAA (7)	9	Known	Friedrich ataxia	No
GDF11	CGC (10)	12	New	Differentiation factor	No (4)
GSPT1	CCG (10)	16	Known	G to S phase transition 1	Yes
HD	**GCA (18)**	**4**	**Known**	**HD**	**Yes**
HLXB9	CGC (9)	7	New	Homeobox	No (0)
HRC	TCA (13)	19	Known	Sarcoplasmic reticulum	No (2)
IGSF3	CTC (9)	1	New	Ig-like	No
ITGAV	TTG (10)	2	New	α V integrin	No (0)
KCNN3	**GCA (18)**	**1**	**Known**	**SCA**	**Yes**
KIAA0040	CTT (10)	1	New	KIAA0040 gene product	Yes
KIAA0476	CAG (9)	X	New	KIAA0476 protein	Yes
KIF3B	CCG (12)	20	Known	Kinesin family member 3B	No (1)
MAB21L1	CTG (19)	13	Known	*Caenorhabditis elegans* homolog-like 1	No (5)
MAP3K4	TGC (10)	6	New	MAP kinase	Yes
MLLT3	CTG (10)	9	New	Myeloid leukemia	No (3)
MN1	TCA (10)	22	Known	Meningioma	No (5)
NCOR2	TGC (11)	12	Known	Neuro retinoid	Yes
NOTCH4	GCA (9)	6	Known	Protooncogene	Yes
NR4A3	CCA (9)	9	New	Chondrosarcoma	No (8)
PCQAP	CAG (13)	22	New	Positive C-factor-2	Yes
PIM1	GCA (10)	6	New	Oncogene PIM1	Yes
POLG	GCA (13)	15	Known	Mitochondrial DNA polymerase	Yes
POU4F1	CCA (9)	13	New	Ceroid lipofuscinosis	Yes
POU4F2	GGC (11)	4	Known	POU transcription factor	Yes
PRDM12	GGC (11)	9	New	Unknown	No (4)
PRKCSH	GGA (10)	19	Known	PKC substrate	Yes
PTPNS1	CCA (10)	20	New	Protein tryosine phosphatase nonreceptor	No (7)
RAI1	CAG (13)	17	New	Retinoic acid induced	Yes
RPF1	CAG (10)	7	New	Retina POU-factor 1	Yes
RPL14	CTG (10)	3	New	Ribosomal protein L14	Yes
SALL1	GCT (10)	16	New	SAL-like	No (4)
SCA1	**CAG (13)**	**6**	**Known**	**SCA1**	**Yes**
SCA12	**CAG (7)**	**5**	**Known**	**Spinocerebrellar ataxia**	**No**
SCA2	**TGC (13)**	**12**	**Known**	**SCA**	**Yes**
SCA7	**TGC (10)**	**3**	**Known**	**SCA**	**Yes**
SCA8	**CTG (8)**	**13**	**Known**	**Spinocerebellar ataxia**	**Yes**
SMARCA2	GCA (13)	9	New	Mitosis control	No (4)
TBP	GCA (19)	6	Known	TATA binding	Yes
TIG1	GCA (12)	22	Known	Transcription mediator	Yes
TNRC1	GCA (14)	12	Known	Candidate	Yes
UBE2B	CGG (10)	5	New	Ubiquitin-conjugating enzyme	Yes
ZIC2	CCA (9)	13	Known	Holoprosencephaly	Yes

Each triplet repeat expansion within exons of known genes gathered from the GRID database is listed. The span of each is shown in column 2 and the chromosomal location in column 3. Each triplet repeat expansion was looked up in both the LocusLink and the OMIM database to assign a known or unknown polymorphism status. Functional assignments were assigned based on the literature cited in OMIM. The 14 known disease-causing triplet repeat expansions are demarcated in bold.

subset of repeat elements that are entirely contained within exons. The results of this analysis are shown in Table 2. Using the virtual polymorphism detection protocol, we identified potential polymorphism in the *FOXE1, KIAA0040, MAP3K4, PIM1, POU4F1, RAI1, RPF1, RPL14,* and *UBE2B* genes. We conclude that we have identified polymorphism in ten (including *SMARCA2*) among the twenty-eight new candidate triplet repeat expansions.

The ability of the electronic screening protocol to detect variation within the existing sequence dataset confirms the possibility that many of these STR markers may be polymorphic. In further studies using the protocol, we also examined the set of STRs that span either the splice donor or acceptor of the same set of exons. The results of this analysis revealed that virtual polymorphism could be detected in more than 50% of the samples analyzed by this protocol. The results of the electronic validation support our contention that many of the computationally detected STRs are potentially polymorphic and therefore useful in genome mapping and genetic studies. We are currently refining the protocol for application to the entire set of STR markers. The resulting data will be accessible at the web site as well.

Discussion

In this report, we describe a method for the rapid annotation of all perfect tandem repeats within the human genome. The method is thorough and yet rapid enough to enable analysis of the increasing amounts of refined sequence data that are being generated from the human and other genomes. The incorporation of identified STR elements into a relational database will enable several important potential applications of this data to the study of genomics and evolutionary biology. We focus our analysis here on two such applications—increasing map marker density along the genetic map and identifying repeat elements that might disrupt or otherwise contribute to gene dysfunction. Both of these applications have the potential of significantly enhancing our knowledge of human disease etiology.

As seen in Table 1, our method identifies many additional STR elements beyond those detected by any of the existing methods. We verified the thoroughness of our STR map by comparing our results with those derived from two other methods: 1) a modified regular-expression based repeat finder developed by Temnykh et al. [23] and 2) a C implementation of our perfect tandem repeat finding algorithm. The results of this comparison showed that an identical set of repeats was obtained using all three methods over the entire human genome.

The overall composition of the observed STR elements in the human genome reflects expected values. For all di-tri- and tetranucleotide repeat elements, there is a relative decrease in those elements containing the CpG dinucleotide (data available at website). Also, as expected the relative density of STR elements in the human genome is significantly higher than would be expected from a random distribution. In fact, the overall frequency of STR elements in the human genome is significantly higher than would be expected from a random distribution of nucleotides throughout the genome, even accounting for base compositional bias (Table 1). This discrepancy suggests the existence of a mechanism for the generation and/or amplification and maintenance of such repeats and may implicate a role for the STR elements in regulation. The summaries of the distributions and contents of the STR elements for the entire genome can be accessed as auxiliary information at our website (http://grid.abcc.ncifcrf.gov). We were surprised to find 67 repeats between 7 and 16 base pairs in length which were repeated 9 or more times (Table 1).

In our evaluation of the marker density relative to existing marker information, we asked whether the addition of the complete set of STR markers would increase the likelihood that there would be a new marker near a given gene. We saw that the average distance from an STS marker to an exon was dramatically decreased and also that where there were deserts in the SNP or STS data set, the STR data frequently filled those gaps. In addition, a typical cancer gene, *BRCA1* has only one STS marker on the genetic linkage map, whereas our set includes 22 additional STR elements in this region. Of these, at least one displays polymorphism (17_45272674_ca_21) electronically. Several additional STRs were present only once in the database and their polymorphism status could not be determined. In addition to increasing the power of genetic linkage studies these new STRs could also allow more detailed analysis of chromosome delections in the germ line and/or in tumor cells. Our exhaustive STR map has particular utility in linkage and association studies where marker statistical power and informativeness are paramount. Both the transmission disequilibrium tests (TDT, Sib-TDT) and allele sharing tests (e.g.-haplotype relative risk, affected-sib-pair methods) benefit through the use of multi-allelic markers [24]. A dense map of highly informative markers will increase the ability to go from either linkage or initial association to identifying the disease mutation or haplotype.

Incorporation of the complete STR map into a relational database allows querying genomic information in novel ways. As an example, we queried the dataset of known annotated genes for those containing repeat elements. As shown in Figure 2 there are a considerable number of di- and trinucleotide repeats in coding regions. Because trinucleotide repeats in coding regions are known to be associated with triplet expansion disorders, we investigated these further. Twenty-eight of these newly identified triplet repeat-containing genes are candidates for undescribed trinucleotide expansion disorders and include genes that potentially play a role in cancer. We have experimentally demonstrated that one of these, *SMARCA2,* is polymorphic and nine others (*FOXE1, KIAA0040, MAP3K4, PIM1,*

POU4F1, *RAI1*, *RPF1*, *RPL14*, and *UBE2B*) are polymorphic on the basis of disparate sequence deposits into the public databases. We are currently experimentally evaluating the polymorphic status for many of the identified triplet repeats where amplification failed.

In addition to the triplet repeats within genes, we have also begun to analyze other interesting STR-based features we have identified within the genome. These include repetitive deserts, repeats within repeats, repeats flanking genes, SNPs within repeats and repeats spanning intron/exon junctions, and the development of a saturation repeat map of such a density that it would allow repeats to serve as markers in regions where other methods fail.

The method described here initially made use of the special architecture of the Cray computer to analyze whole genome sequences. Searching for perfect tandem repeats is a particularly straightforward demonstration of the utility of high performance computing vector platforms to bioinformatics. We identified all perfect tandem repeats in the human genome between 2 and 16 bases in length and incorporated the repeats along with other genomic annotation into an open resource for querying. Because we found analysis of the entire human genome sequence feasible on a high performance platform, we were encouraged to develop methods to carry out this same analysis using conventional computers. The database that resulted from this computational work has increased the marker density throughout the genome.

In considering other important bioinformatics problems, such as delineating non-tandem repeats and providing whole genome comparisons, supercomputers may be a useful platform for high throughput genome annotation. In addition, we speculate that the special hardware features of vector machines may be amenable to a number of other bioinformatics applications. These include genome assembly, EST clustering, unique primer design for PCR, identification of segmental duplications, proteomics and comparative genomics.

Acknowledgments

The authors wish to thank Dr. David Haussler and Mr. Jim Kent for early access to certain features of the UC Santa Cruz browser, which have enhanced our own database mining capabilities. We also wish to thank Bernard Gerrard and Amanda Robert for timely and outstanding technical assistance searching for triplet repeat polymorphism and in demonstrating the utility of identified repeats as microsatellite markers. Ms. Beva Langdon assisted in the editing of some portions of the manuscript. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This project has been funded in whole or in part with Federal Funds from the National Cancer Institute, National Institutes of Health under contract No. N01-C0-12400.

References

[1] J.L. Weber, P.E. May, Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction, Am. J. Hum. Genet. 44 (1989) 388–396.
[2] J.A.L. Armour, Minisatellites and mutation processes in tandemly repetitive DNA, in: D.B. Goldstein, C. Schlotterer (Eds.), Microsatellites Evolution and Applications, Oxford University Press, Oxford, 1999, pp. 24–33.
[3] H. Zischler, C. Kammerbauer, R. Studer, K.H. Grzeschik, J.T. Epplen, Dissecting (CAC)5/(GTG)5 multilocus fingerprints from man into individual locus-specific, hypervariable components, Genomics 13 (1992) 983–990.
[4] P.S. Subramanian, D.L. Nelson, A.C. Chinault, Large domains of apparent delayed replication timing associated with triplet repeat expansion at FRAXA and FRAXE, Am. J. Hum. Genet. 59 (1996) 407–416.
[5] Y.H. Wang, R. Gellibolian, M. Shimizu, R.D. Wells, J. Griffith, Long CCG triplet repeat blocks exclude nucleosomes: a possible mechanism for the nature of fragile sites in chromosomes, J. Mol. Biol. 263 (1996) 511–516.
[6] W. Amos, A comparative approach to the study of microsatellite evolution, in: D.B. Goldstein, C. Schlotterer (Eds.), Microsatellites Evolution and Applications, Oxford University Press, Oxford, 1999, pp. 66–79.
[7] D.G. Monckton, A.J. Jeffreys, DNA profiling, Curr. Opin. Biotechnol. 4 (1993) 660–664.
[8] S. Cosso, R. Reynolds, Validation of the AmpliFLP D1S80 PCR Amplification Kit for forensic casework analysis according to TWGDAM guidelines, J. Forensic. Sci. 40 (1995) 424–434.
[9] J.P. Jakupciak, R.D. Wells, Genetic instabilities in (CTG.CAG) repeats occur by recombination, J. Biol. Chem. 274 (1999) 23468–23479.
[10] D.C. Rubinstein, Trinucleotide expansion mutations cause diseases which do not conform to classical Mendelian expectations, in: D.B. Goldstein, C. Schlotterer (Eds.), Microsatellites Evolution and Applications, Oxford University Press, Oxford, 1999, pp. 80–97.
[11] C.J. Cummings, H.Y. Zoghbi, Fourteen and counting: unraveling trinucleotide repeat diseases, Hum. Mol. Genet. 9 (2000) 909–916.
[12] S.K. Kannan, E.W. Myers, An algorithm for locating nonoverlapping regions of maximum alignment score, SIAMS J. Comput. 25 (1996) 648–662.
[13] G. Benson, A space efficient algorithm for finding the best nonoverlapping alignment score, Theor. Comput. Sci. 145 (1995) 357–369.
[14] J.P. Schmidtz, All highest scoring paths in weighted grid graphs and their application to finding all approximate repeats in strings, SIAM J. Comput. 27 (1998) 972–992.
[15] E. Rivals, O. Delgrange, J.P. Delahaye, M. Dauchet, M.O. Delorme, A. Henaut, E. Ollivier, Detection of significant patterns by compression algorithms: the case of approximate tandem repeats in DNA sequences, Comput. Appl. Biosci. 13 (1997) 131–136.
[16] A. Milosavljevic, J. Jurka, Discovering simple DNA sequences by the algorithmic significance method, Comput. Appl. Biosci. 9 (1993) 407–411.
[17] S. Karlin, M. Morris, G. Ghandour, M.Y. Leung, Efficient algorithms for molecular sequence analysis, Proc. Natl. Acad. Sci. USA 85 (1988) 841–845.
[18] G. Benson, M.S. Waterman, A method for fast database search for all k-nucleotide repeats, Nucleic Acids Res. 22 (1994) 4828–4836.
[19] M. Sagot, E. Myers, Identifying satellites in nucleic acid sequences, in: S. Istrail, P. Pevzner, M. Waterman (Eds.), Proceedings of the

Second Annual International Conference on Computational Molecular Biology, ACM Press, New York, 1998, pp. 234–242.
[20] G. Benson, An algorithm for finding repeats of unspecified pattern size, in: S. Istrail, P. Pevzner, M. Waterman (Eds.), Proceedings of the Second Annual International Conference on Computational Molecular Biology, ACM Press, New York, 1998, pp. 20–29.
[21] G. Landau, J. Schmidt, An algorithm for approximate tandem repeats, in: A. Apostiolico, M. Crochemore, Z. Galil, U. Manber (Eds.), Proceedings of the 4th Annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science, Vol. 648, Springer-Verlag, Berlin, 1993, pp. 120–133.
[22] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res. 27 (1999) 573–580.
[23] S. Temnykh, G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, S. McCouch, Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential, Genome Res. 11 (2001) 1441–1452.
[24] M. Pericak-Vance, Linkage disequilibrium and allelic association in approaches to gene mapping in complex human diseases, in: M. Pericak-Vance, J.J. Haines (Eds.), Approaches to Gene Mapping in Complex Diseases, Wiley-Liss, New York, 1998, pp. 323–333.