BMC Genomics

BioMed Central

Research article

Open Access

STRP Screening Sets for the human genome at 5 cM density

Nader Ghebranious[4], David Vaske[2], Adong Yu[1], Chengfeng Zhao[1], Gabor Marth[3] and James L Weber*[1]

Address: [1]Center for Medical Genetics, Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA, [2]Pioneer Hi-Bred International, Johnston, IA USA, [3]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD USA and [4]Molecular Diagnostic Genotyping Laboratory, Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA

Email: Nader Ghebranious - ghebranious.nader@marshfieldclinic.org; David Vaske - dave.vaske@pioneer.com; Adong Yu - yua@cmg.mfldclin.edu; Chengfeng Zhao - zhaoc@cmg.mfldclin.edu; Gabor Marth - marth@ncbi.nlm.nih.gov; James L Weber* - weberj@cmg.mfldclin.edu

* Corresponding author

Published: 24 February 2003

BMC Genomics 2003, **4**:6

Received: 10 December 2002
Accepted: 24 February 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/6

Abstract

Background: Short tandem repeat polymorphisms (STRPs) are powerful tools for gene mapping and other applications. A STRP genome scan of 10 cM is usually adequate for mapping single gene disorders. However mapping studies involving genetically complex disorders and especially association (linkage disequilibrium) often require higher STRP density.

Results: We report the development of two separate 10 cM human STRP Screening Sets (Sets 12 and 52) which span all chromosomes. When combined, the two Sets contain a total of 782 STRPs, with average STRP spacing of 4.8 cM, average heterozygosity of 0.72, and total sex-average coverage of 3535 cM. The current Sets are comprised almost entirely of STRPs based on tri- and tetranucleotide repeats. We also report correction of primer sequences for many STRPs used in previous Screening Sets. Detailed information for the new Screening Sets is available from our web site: http://research.marshfieldclinic.org/genetics.

Conclusion: Our new human STRP Screening Sets will improve the quality and cost effectiveness of genotyping for gene mapping and other applications.

Background

Since their discovery in 1988, multiallelic short tandem repeat polymorphisms (STRPs) (also called microsatellites or simple sequence length polymorphisms (SSLPs)) have been the polymorphisms of choice for linkage mapping and many other genetic studies.

Although there are hundreds of thousands of reasonably informative STRPs in the human genome [1,2], only a small fraction are optimal for genotyping and genome scans. Optimal properties of an STRP include: high heterozygosity, strong and specific PCR amplification, capability to be amplified simultaneously with other STRPs (multiplexed), sharp bands on gels, easy and accurate scoring of allele sizes, relatively low mutation rate, and appropriate position along the genetic map.

We have performed human genome polymorphism scans in our lab since 1989 [3]. Our first human Screening Set of STRPs developed in 1992 had an average STRP spacing of ~20 cM, no sex chromosome STRPs, and consisted almost entirely of dinucleotide repeat STRPs identified at Marshfield. Each subsequent Screening Set from our lab improved on the previous version by adding STRPs, by using more accurate genetic maps to make STRP spacing more uniform and to eliminate large gaps, and especially

by replacing relatively low quality STRPs with superior ones. Typing better STRPs leads to higher data quality through fewer missing genotypes and fewer incorrect allele calls. Typing optimal STRPs also leads to lower genotyping costs by providing more information, by reducing the need for duplicate genotyping, by permitting the use of shorter gels (with lower resolving power but shorter run times), and by increasing the efficiency of allele calling.

We have replaced nearly all of the dinucleotide repeat STRPs in our Screening Sets with tri- and tetranucleotide STRPs. Although dinucleotide STRPs are abundant and meet many of the criteria for optimal STRPs, they are also in our hands more difficult to score accurately because of substantial strand slippage during PCR [4]. We also find that dinucleotide STRPs are more difficult to PCR multiplex than tri- or tetranucleotide STRPs.

Similarly, we have eliminated nearly all of the STRPs with frequent (> 2%) "non-integer" alleles. Non-integer alleles are defined as have length differences from the most frequent alleles which are other than integer multiples of the repeat length. For example, an allele of 221 bp (PCR product length) would be a non-integer allele for a tetranucleotide STRP with frequent alleles of 230, 226, 222, and 214 bp. Non-integer alleles are not typing artifacts as they have been observed in many labs and have been confirmed by sequencing of individual alleles [5]http://www.cstl.nist.gov/biotech/strbase. Non-integer alleles probably exist somewhere in the human population for all or nearly all STRPs, but a significant fraction of STRPs do not have frequent non-integer alleles.

We have also excluded or repaired STRPs with weak or null alleles. In at least most cases, weak and null alleles appear to be due to substitution polymorphisms within the primer annealing sites [6]. They can be repaired by sliding the offending PCR primer to a nearby position along the chromosome.

For most applications of genome polymorphism scans, higher STRP densities are preferable. This is particularly important for gene mapping by association. While analysts have predicted that very high polymorphism densities will be required for association mapping in mixed or outbred populations [see for example reference [7]], promising results have been obtained using genome scans of 600–1200 STRPs in isolated populations where levels of linkage disequilibrium are particularly high [8–10]. In this manuscript we describe the development of two new 10 cM human STRP Screening Sets (Sets 12 and 52) which when combined provide average STRP spacing of 4.8 cM.

Results

Building new human Screening Sets

Over about the last decade we have produced at Marshfield twelve separate, but related 10 cM Screening Sets of STRPs for the human genome (see Table 1 and http://research.marshfieldclinic.org/genetics). For each of these Sets, the lowest quality STRPs in the previous Set were replaced with superior ones. Of the most recent collections, Sets 6, 7, 10, and 11 were major overhauls, with 21–52% of the STRPs replaced (Table 1). Sets 5 and 8 were described in the literature [11,12]. Beginning particularly with Set 6, many of the dinucleotide STRPs were replaced with tri- and tetranucleotide STRPs from the Cooperative Human Linkage Center (CHLC) [13]. CHLC STRPs still comprise 81% and 55% of our current Sets (Sets 12 and 52, respectively). Starting in about 2001, the availability of the human genomic draft sequence greatly expanded the number of STRPs from which to choose. Sets 12 and 52 contain 15% and 44%, respectively, newly derived STRPs from the genomic sequence.

Table 1: History of Marshfield 10 cM STRP Screening Sets

Set	Year	Number of STRPs	Number of Dinucleotide STRPs (Fraction)	Number of STRPs Shared with Previous Set (Fraction)
1	1992	231	211 (0.91)	
2	1993	366	347 (0.95)	131 (0.36)
3	1993	319	226 (0.71)	176 (0.55)
4	1994	347	243 (0.70)	274 (0.79)
5	1994	363	191 (0.53)	265 (0.73)
6	1995	391	55 (0.14)	186 (0.48)
7	1995	390	47 (0.12)	297 (0.76)
8	1996	387	43 (0.11)	377 (0.97)
9	1997	387	44 (0.11)	378 (0.98)
10	1999	405	49 (0.12)	313 (0.77)
11	2001	410	2 (0.01)	324 (0.79)
12	2002	408	3 (0.01)	405 (0.99)

We began the construction of Sets 12 and 52 with specific goals. For Set 12 (and the very similar Set 11), we intended to replace all or nearly all dinucleotide repeat polymorphisms and to eliminate other problematic STRPs such as those with frequent non-integer alleles. For Set 52 (and the preceding Set 51), we aimed to identify one or two high quality tri- or tetranucleotide STRPs at approximately the midpoint between each pair of STRPs within Set 12. Set 52 is therefore the second in a new series of 10 cM Screening Sets. For both Sets, we needed to identify new, high quality STRPs within specific, relatively small (~1 mb) chromosomal segments.

Altogether, we screened 2262 STRPs for possible inclusion within the Screening Sets. Of these, 1103 were STRPs developed within the CHLC or at Utah [14]. The remaining 1159 were developed from human genomic sequences. Most of these (961) were identified by searching for tri- or tetranucleotide STRs with ≥ 7 or 8 uninterrupted repeats within the sequence assembly available from the University of California – Santa Cruz web site, December 2000 version http://genome.cse.ucsc.edu. Others (198) were identified by examining a collection of ~1 gb of overlapping BAC sequences [15] for the presence of variable tri- and tetranucleotide STRs. We focused efforts on AAT and AGAT repeats because these sequences are known to be abundant and to yield useful polymorphisms [13,16].

New PCR primers selected from the sequences flanking the tandem repeats were tested by amplification with ten individual DNA samples and one DNA pool using incorporation of a nucleotide tagged with a fluorescent dye (see Methods). PCR primers labelled with a fluorescent dye at the 5' end were then synthesized for those STRPs which displayed ≥ 4 alleles in the first screen. These were combined with existing CHLC and Utah STRPs, and were used to screen 12 individuals and one pool. All donors of DNA samples used in these first two screens had Northern European ancestry. STRPs which passed these first two hurdles were then used in genome scans within the Mammalian Genotyping Service (see Marshfield web site) using hundreds of DNA samples from various geographical locations.

Only 11% of the 2262 genomic STR sequences that were screened were included within Sets 12 and 52. The great majority of excluded STRPs were rejected because of limited numbers of alleles (low informativeness). About 9% were rejected because of the presence of frequent non-integer alleles. We found that use of candidate genomic sequences with larger numbers of uninterrupted tandem repeats and use of overlapping BAC sequences with alleles which differed by *two or more* repeats led to higher rates of STRP inclusion into the Screening Sets. Information on all of the STRPs that we found to be polymorphic can be obtained from the comprehensive list of indel polymorphisms on the Marshfield web site.

We also improved the amplification efficiency of Screening Set STRPs. Most of the human STRPs developed in the early and mid 90s were based on relatively crude, single-pass sequencing of genomic DNA subclones. Comparison of the PCR primer sequences for Set 10 STRPs with the new public genomic sequences revealed that a surprisingly high 25% of the STRPs had mismatches in at least one of the primers (an example is shown in Figure 1). Nearly all of the mismatches were near the middle or 5' ends of the primers. New primers designed using the public genomic sequences were then tested side by side with old primers. At 55°C annealing temperature and no PCR multiplexing, few differences were observed between the old and new primer pairs, but under more stringent conditions (60°C annealing temperatures), 79 STRPs were found to amplify better with the new primer pairs (two examples are shown in Figure 2).

STRPs alleles are usually identified and labelled as the length of the PCR product as measured on denaturing polyacrylamide gels. Only in a handful of cases have the full spectrum of STRP alleles been sequenced. Therefore, STRP alleles are referenced to allele sizes for standard DNA templates (we use the parents of CEPH family 1331 available from the NIGMS Human Genetic Cell Repository). Allele sizes will also, of course, often change if the PCR primer sequences for a polymorphism are altered. To avoid null and weak alleles, to prevent the formation of doublet bands during PCR [17,18], and to achieve optimal PCR product length, we have modified original primer sequences for a substantial fraction of our Screening Set STRPs. We have used several different letters following the STRP name to indicate changes in PCR primers (see Marshfield web site). As two examples for STRPs on chromosome 1 in Set 12: GATA26G09N indicates that one of the original primers for GATA26G09 was changed to correct a sequencing error without change in allele sizes, and GGAA3A07Z indicates that one of the primers for GGAA3A07 was shifted along the chromosome resulting in different allele sizes. Current PCR primer sequences for all Screening Set STRPs are listed on the Marshfield web site along with allele sizes for individuals 133101 and 133102.

Genetic Map Positions

Initially, new STRPs were selected and incorporated into our Screening Sets based on physical distances obtained from the December 2000 UC-Santa Cruz draft sequence assembly. However, we soon found that the draft assembly contained many errors [see for example reference [19]] and resulted therefore in many STRPs being in the wrong map positions. To correct these mistakes, we utilized the

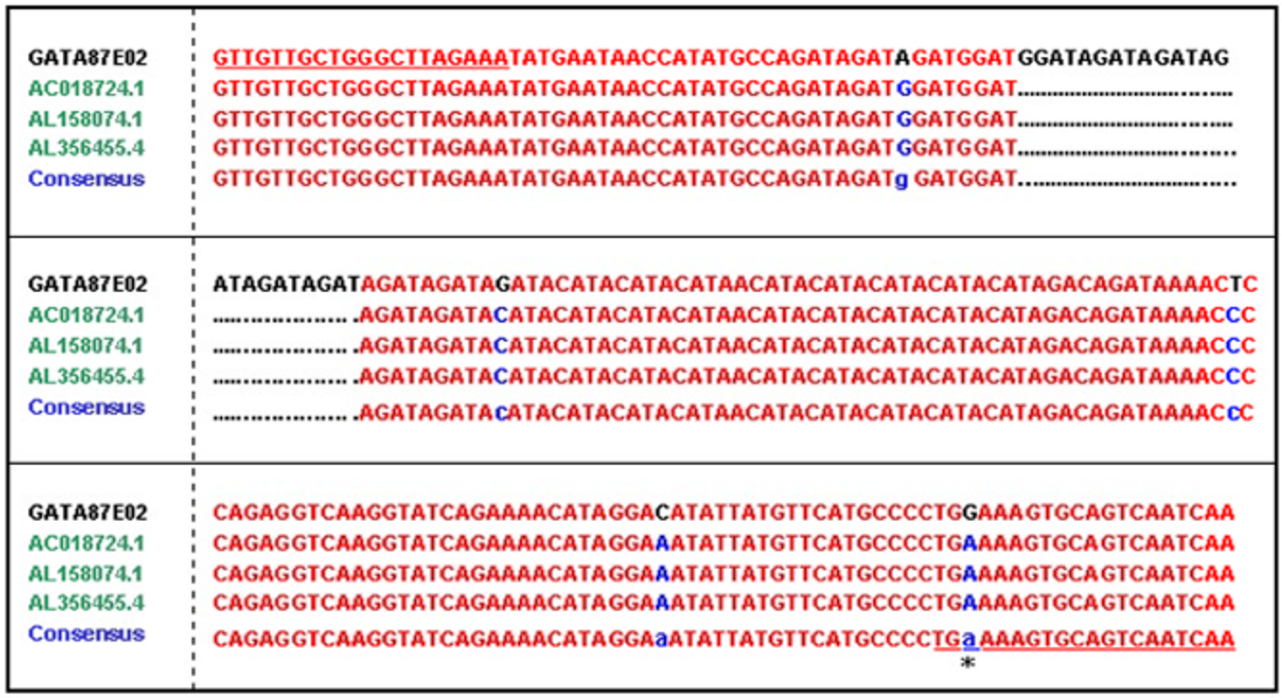

Figure 1
Correction of PCR Primer Sequences using Genomic Sequence Assemblies. The original single pass sequence for GATA87E02 is aligned with the sequences from several BACs containing overlapping genomic DNA. The original reverse PCR primer mismatched the BAC sequences near its 3' end. Note that because the great majority of the public human genomic sequence was generated from BAC libraries prepared from just a few donors, it is possible that two or even all three of the BAC sequences shown in the figure came from the same chromosome.

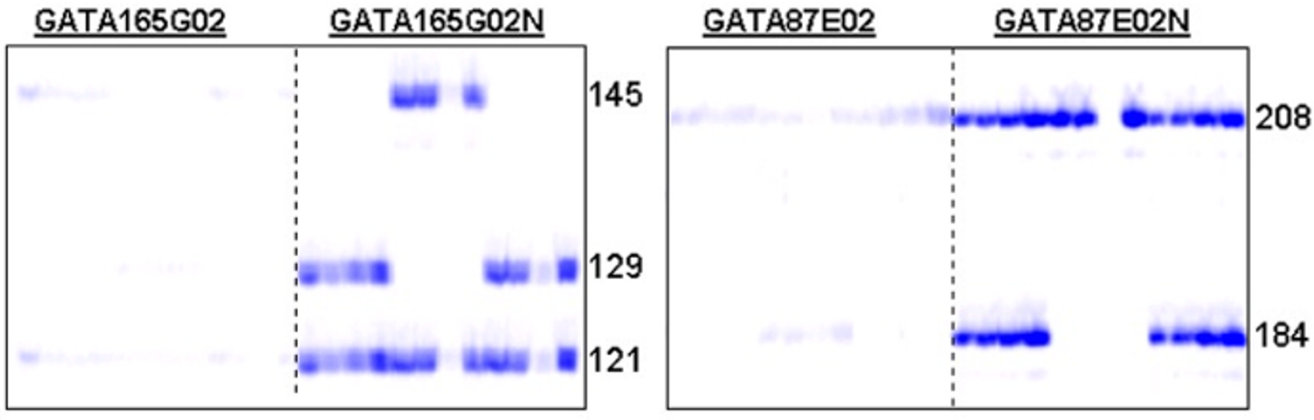

Figure 2
Comparison of PCR Amplification using Original and Corrected PCR Primer Sequences. Shown are electrophoretic separations of DNA fragments from unrelated individuals amplified at 60°C annealing temperature. Fragments obtained with the corrected PCR primer pairs are indicated by the N suffixes after the STRP names.

most recent (June 2002) sequence assembly in addition to linkage analysis using three large Sets of families. In all cases except one (4ptel04), the linkage results matched the June 2002 assembly in terms of STRP order (we assumed the linkage results were correct for 4ptel04). Our confidence in STRP order is therefore high.

With one exception on chromosome 6p (see below) genetic map positions for the Screening Set STRPs were taken from the most recent Marshfield map [20] or by interpolation using the Marshfield map and the genetic and physical map positions described in the previous paragraph. Although the new Iceland genetic map [19] is higher resolution than the Marshfield map, a large fraction (62%) of the Screening Set 12 and 52 STRPs were not typed in the Iceland families. We did, however, check STRP order for all STRPs that were typed in the Iceland families and found no disagreements with the Marshfield map, except for two close (~1 mb apart), adjacent STRPs on chromosome 6p, ATA50C05 and ATC4D09, where the linkage results, the Iceland map and the June 2002 sequence assembly all disagreed with the Marshfield map.

Characterization of Sets 12 and 52

Numbers of STRPs, heterozygosity values, and sex-average genetic map properties for Screening Sets 12, 52, and 12 plus 52 combined, broken down by chromosome, are displayed in Table 2. Of the 39 total X chromosome STRPs in the combined Sets, 3 (GATA2A12, GGAT3F08, and GATA42G01) are in the pter pseudoautosomal region, and 1 (SDF1) is in the qter pseudoautosomal region. The 9 Y chromosome STRPs are all male-specific. Also, two small, tightly-spaced clusters of STRPs are included in Set 12 (six STRPs near the centromere of chromosome 11 and three STRPs on the short arm of chromosome 1) for the purpose of gauging linkage disequilibrium.

Table 2: General Properties of Sets 12 and 52 by Chromosome.

	Numbers of STRPs			Average Heterozygosity			Total Distance Covered (cM)			Average Spacing (cM)		
Chr.	**Set 12**	**Set 52**	**Both Sets**	**Set 12**	**Set 52**	**Both Sets**	**Set 12**	**Set 52**	**Both Sets**	**Set 12**	**Set 52**	**Both Sets**
1	31	36	67	0.76	0.68	0.72	274.6	261.1	282.9	9.2	7.5	4.3
2	28	28	56	0.78	0.66	0.72	266.2	215.7	266.2	9.9	8.0	4.8
3	25	24	49	0.75	0.73	0.74	222.5	207.1	222.5	9.3	9.0	4.6
4	23	16	39	0.76	0.69	0.73	208.0	198.6	208.0	9.5	13.2	5.5
5	20	25	45	0.78	0.66	0.71	196.6	190.3	197.5	10.3	7.9	4.5
6	23	18	41	0.75	0.69	0.72	192.4	188.0	192.4	8.7	11.1	4.8
7	21	20	41	0.76	0.68	0.72	178.6	157.3	178.6	8.9	8.3	4.5
8	19	12	31	0.74	0.67	0.72	159.7	141.5	159.7	8.9	12.9	5.3
9	18	17	35	0.74	0.67	0.71	158.2	151.8	158.2	9.3	9.5	4.7
10	20	18	38	0.75	0.66	0.71	163.1	152.3	163.1	8.6	9.0	4.4
11	20	12	32	0.78	0.65	0.73	145.5	127.7	145.5	7.7	11.6	4.7
12	17	19	36	0.79	0.72	0.75	161.5	151.4	165.8	10.1	8.4	4.7
13	12	12	24	0.75	0.69	0.72	98.9	94.3	102.0	9.0	8.6	4.4
14	14	14	28	0.75	0.69	0.72	122.5	109.4	122.5	9.4	8.4	4.5
15	13	9	22	0.76	0.71	0.74	114.8	81.5	114.8	9.6	10.2	5.5
16	15	13	28	0.75	0.67	0.71	127.7	114.5	127.7	9.1	9.5	4.7
17	13	13	26	0.76	0.65	0.71	118.5	103.6	118.5	9.9	8.6	4.7
18	13	14	27	0.78	0.69	0.74	113.2	121.0	121.0	9.4	9.3	4.7
19	10	8	18	0.77	0.65	0.72	91.2	86.6	91.2	10.1	12.4	5.4
20	11	12	23	0.78	0.69	0.73	98.4	85.1	98.4	9.8	7.7	4.5
21	6	5	11	0.82	0.62	0.73	44.8	49.6	54.8	9.0	12.4	5.5
22	8	9	17	0.75	0.71	0.73	59.4	44.6	59.4	8.5	5.6	3.7
X	22	17	39	0.70	0.62	0.66	184.0	149.1	184.0	8.8	9.3	4.8
Y	6	3	9	0.66	0.41	0.58						
	Total STRPs			**Average Heterozygosity**			**Total Coverage (cM)**			**Average Spacing (cM)**		
	408	**374**	**782**	**0.76**	**0.67**	**0.72**	**3500**	**3182**	**3535**	**9.3**	**9.5**	**4.8**

Genetic distances for the autosomes are sex-average, and for the X chromosome are female (except for pseudoautosomal regions).

Table 3: Breakdown of Screening Set STRPs by Repeat Length

	Total STRPs	Dinucleotide	Trinucleotide	Tetranucleotide	Pentanucleotide
Set 12	408	3	82	318	5
Set 52	374	0	98	267	9
Both Sets	782	3	180	585	14

Table 4: Breakdown of Screening Set STRPs by Repeat Sequence.

	Total STRPs	AAAT	AAGG	AAT	AATG	AGAT	Other
Set 12	408	13	27	78	4	265	21
Set 52	374	22	11	85	6	218	32
Both Sets	782	35	38	163	10	483	53

Repeat sequences are listed in their alphabetically minimal forms.

Set 12 STRPs with overall average heterozygosity of 76% are more informative than Set 52 STRPs with overall average heterozygosity of 67%. At least part of this difference may simply be a reflection of the populations used to deduce these values (see Methods). As shown in Table 2, X and especially Y chromosome STRPs had lower average informativeness than autosomal STRPs.

The average, sex-average STRP spacing of the combined Sets was 4.8 cM. The maximum gaps are 18.4, 37.5, and 15.5 cM for Set 12, Set 52 and Sets 12 and 52 combined, respectively. There were 13 gaps ≥ 15 cM in Set 12, 51 such gaps in Set 52, and 29 gaps ≥ 10 cM in Sets 12 and 52 combined. Set 12 STRPs were generally closer to telomeres than Set 52 STRPs, resulting in greater total chromosomal coverage.

A summary of repeat length in the Screening Set STRPs is presented in Table 3. Only 3 dinucleotide STRPs remain in Set 12. Fourteen pentanucleotide STRPs were also included in the combined Sets.

Breakdown of the Screening Set STRPs by repeat type is shown in Table 4. STRPs with AGAT and AAT repeats together accounted for 83% of the STRPs in the combined Sets. Note that because of permutation and the complementary strand there are several names for each repeat type. As just one example, AGAT repeats can also be presented as GATA, ATAG, TAGA, ATCT, TATC, CTAT, and TCTA repeats. Following the suggestion of Jin et al. [21] we have chosen the alphabetically minimal name.

We found that AAT repeats in particular, have a relatively low level of non-integer alleles. For example, within Set 10, 11.1% of GGAA and 10.2% of AGAT STRPs had frequent non-integer alleles, compared to only 1.8% of AAT STRPs. Because of high rates of non-integer alleles, STRPs with purines on one strand and pyrimidines on the other (eg AAGG) were avoided even though they are reasonably abundant and often especially informative [13].

Association of Screening Set STRPs with interspersed repeat elements (IREs) is shown in Table 5. STRPs were considered to be associated with IREs if the IRE fell in the 50 bp flanking the STR on either side (total of 100 bp of flanking sequence). Although total numbers for some of the STR types are relatively small, it appears that each type of STR has its own particular signature of IRE association. For example, AAAT STRs are very often (86%) associated with Alu elements, consistent with the hypothesis that most of these repeats evolved from the polyA tail of Alus [22]. An unexpectedly large fraction of AGAT STRs (16%) were found to be associated with LTRs. The results in Table 5 may generally provide clues about the evolution of STRs.

Discussion

Development of human STRP Screening Sets has paralleled advances in construction of genetic and physical maps. Except in regions with long inversion polymorphisms [23], it should soon be possible to specify STRP *order* within Screening Sets with near certainty. However, because of individual and even possibly population differences in recombination rates [24–26], it may never be possible to specify genetic *distances* between STRPs with high precision.

Table 5: Association of Interspersed Repetitive Elements with Selected STR Types.

	Number of STRPs Associated with Indicated Repeats (Fraction STRPs)						
Repeat	**Total STRPs**	**AAAT**	**AAGG**	**AAT**	**AATG**	**AGAT**	**Other**
ALU	189 (0.24)	31 (0.89)	9 (0.24)	64 (0.39)	2 (0.20)	63 (0.13)	20 (0.38)
L1	96 (0.12)	2 (0.06)	5 (0.13)	42 (0.26)	0 (0.00)	40 (0.08)	7 (0.13)
LTR	119 (0.15)	4 (0.11)	3 (0.08)	14 (0.09)	0 (0.00)	97 (0.20)	1 (0.02)
L2	16 (0.02)	1 (0.03)	1 (0.03)	4 (0.02)	4 (0.40)	4 (0.01)	2 (0.04)
MER	27 (0.03)	0 (0.00)	0 (0.00)	5 (0.03)	0 (0.00)	20 (0.04)	2 (0.04)
MIR	22 (0.03)	1 (0.03)	0 (0.00)	8 (0.05)	0 (0.00)	8 (0.02)	5 (0.10)
Other	6 (0.01)	0 (0.00)	0 (0.00)	2 (0.01)	0 (0.00)	3 (0.01)	1 (0.02)
Not Associated	352 (0.45)	3 (0.09)	23 (0.61)	44 (0.27)	5 (0.50)	259 (0.54)	18 (0.35)
TOTAL TESTED	781	35	38	163	10	483	52

STRPs were screened using Repeat Masker for IREs that are within 50 bp in either direction of the short tandem repeats (excluding the tandem repeats). Sums of the numbers in the columns do not match the totals because some sequences had two different interspersed repeats within the 100 bp.

Screening Set 10 STRPs have been typed in the ~1000 members of the Human Diversity Panel [27]. We hope to also type the new Set 12 and 52 STRPs through this Panel in the relatively near future. It is beneficial to have a global perspective on informativeness and allele frequencies for each Screening Set STRP. Although the Screening Set STRPs were initially screened using European DNA samples, we have found that in almost all cases, they are highly or modestly informative in other human populations. Consistent with previous results [28–30], average heterozygosities for the Screening Set STRPs in Sub-Saharan Africans are the highest [27]. Isolated European populations such as Sardinian villagers, and Old Order Amish in the U.S. have only slightly diminished heterozygosities. Although some Screening Set STRPs have reduced informativeness in East Asian populations such as Han Chinese, average heterozygosities in East Asians are also only slightly diminished compared to Europeans. So far, only Native American and some Oceanic populations have average heterozygosities for Screening Set STRPs that are substantially reduced. With enough effort, it probably would be possible to develop Screening Set STRPs with higher average heterozygosities for populations such as Native Americans, but the practicality of such an undertaking is uncertain.

Similarly, it would also be helpful to carry out extensive sequencing of at least the frequent alleles for each Screening Set STRP. This would eliminate the need to approximate allele sizes. However, this would also be a large and expensive project, and may have to wait until sequencing costs drop so that many human genomes from around the world can be sequenced.

Although nearly all Screening Set STRPs are at least modestly polymorphic in all human populations examined to date, this does not guarantee that they will be free of frequent non-integer alleles or weak or null alleles in some populations. For example, we have observed apparent null alleles for some STRPs in Chinese that were not present in Europeans (eg GATA29A01 on chromosome 6 and GGAA20G04 on chromosome 2). We have also observed non-integer alleles in Sub-Saharan Africans that have not been seen at appreciable frequency in other populations (eg GATA104 on chromosome 7 and GATA11A06 on chromosome 18).

Despite having much higher mutation rates than diallelic polymorphisms, there is abundant evidence that highly informative STRPs of the type found within Screening Sets are generally powerful markers for detection of linkage disequilibrium [eg [31,32]]. It is unclear, however, whether dinucleotide or tetranucleotide STRPs are superior in this regard. Experimental evidence seems to favour higher average mutations rates for tetranucleotide STRPs [33], while theoretical results favour higher average rates for dinucleotide STRPs [34]. Analysis of STRPs typed in CEPH reference families for construction of human genetic maps revealed that the fraction of dinucleotide/dinucleotide STRP pairs < 200 kb apart with linkage disequilibrium at $p < 0.01$ was 18.7%, whereas the fraction for dinucleotide/tetranucleotide pairs was 7.9% and for dinucleotide/trinucleotide pairs was 22.1% (Broman K, Weber J unpublished results).

Many of the Set 12 and 52 STRPs are superior to the thirteen STRPs used routinely in forensic DNA testing in the U.S. http://www.cstl.nist.gov/biotech/strbase/fbicore.htm. Several of the thirteen forensic STRPs have fre-

quent non-integer alleles. Several are not especially informative. Five of the thirteen forensic STRPs are currently included within Set 12. This occurred by chance rather than design. If genome polymorphism scans for either research or clinical purposes become widespread, then overlap between our Screening Sets and forensic Sets will have to be carefully considered.

Although our newest Screening Sets are substantial improvements over previous versions, they are still not perfect. Some STRPs have lower informativeness than desired, and some large gaps in coverage remain. The Set 12 STRPs are generally superior to Set 52 polymorphisms because Set 52 is new. There has not yet been a chance to make many replacements.

We will continue to make improvements in our human STRP Screening Sets and to post upgrades on the Marshfield web site. But are there limits to the quality of STRP Sets? The answer is undoubtedly yes. There are only approximately 65,000 modestly to highly informative tri- and tetranucleotide STRPs in the human gene pool [1,2]. Within some ~1 mb regions of the genome, we have already exhausted all likely tri- and tetranucleotide STRP candidates. Only a small fraction (11%) of the new STRPs we screened from the genomic sequence were selected for the new Sets. It is quite conceivable, that over the next decade or two we will characterize all human STRPs that have reasonable informativeness. Resequencing different human genomes will undoubtedly contribute much to this effort.

Quite a few investigators have speculated that diallelic polymorphisms such as SNPs or diallelic indels will supplant STRPs in human Screening Sets. Our position continues to be that this question will likely be ultimately determined by typing costs [4]. STRPs provide *much* more information than diallelic polymorphisms, so diallelic typing costs would need to drop well below those for STRPs. This might happen, but it hasn't yet, and it's not clear that it ever will. There may also be advantages to including both high and low mutation rate polymorphisms within Screening Sets (ie STRPs and diallelics) [35]. In any case, we believe that our STRP Screening Sets will continue to be highly valuable and widely used for many years.

Conclusions

The development of Screening Sets 12 and 52 will improve gene mapping in general, and specifically genome scans where a relatively high STRP density is required. Complete information on all of our Screening Sets is freely available from the Marshfield web site http://research.marshfieldclinic.org/genetics along with lists of over 200,000 candidate and confirmed human indel polymorphisms, both multi- and diallelic. We plan to continue to improve our human STRP Screening Sets until we have exhausted all available STRPs at specific chromosomal sites.

Materials and Methods

Identification of candidate polymorphisms

Two different approaches were used to search for new polymorphisms. One approach was to use overlapping BAC genomic sequences to select polymorphisms that varied by more than two repeats [15]. The other approach was to browse the genome for STRs using the December 12, 2000 version of the genomic sequence at University of California – Santa Cruz http://genome.ucsc.edu/[36].

Once a sequence containing the desired polymorphism was selected (usually 400–700 bp in length), it was run through the Repeat Masker program http://ftp.genome.washington.edu/cgi-bin/RepeatMasker in order to avoid selecting PCR primers within Alu, L1, or other repeats. The Primer 3 program http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi was used to select PCR primers. Candidate sequences which did not permit the placement of at least one PCR primer within unique sequence (ie outside of a repeat identified by Repeat Masker) were not tested further. In cases where one PCR primer was located within a repeat, the primer from within the unique sequence was tagged with a fluorescent dye.

Sequence Alignments

All of the 406 single read STRP sequences from Set 10 were Blasted against genomic sequences from the public labs. For nearly all STRPs, we identified 1 to 3 BACs that showed high homology (Blast criteria were score (bits) > 200, expect (E) value < e-50, and ratio of matched bases to STRP sequence length >85%). Two different multiple alignment programs were then used to align the single read and the genomic sequences: "multalin" http://protein.toulouse.inra.fr/multalin/multalin.html and "clustalw" http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html.

Screening of candidate polymorphisms

For initial screening of the PCR primers, we incorporated a dye-labelled nucleotide with a two-step PCR protocol. Briefly, the first step contained 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.001% gelatin, 250 μM each dNTP, 4.7 μM of the forward and reverse primers, 0.15 units of Taq polymerase (Roche) in a total 5 μl reaction volume. The second reaction had the same components and volume as in the first step, except that the forward primer was present at 6.2 μM and R6G dUTP (Applied Biosystems) at 0.5 μM with no reverse primer. About 0.5 μl of step 1 PCR product was used as a DNA template for step 2 PCR. Each PCR step initiated with a 95°C soak for 4

min, followed by 30 and 25 cycles for steps 1 and 2, respectively, consisting of 95°C for 40 sec, 55°C for 75 sec, 72°C for 40 sec, and a final extension of 7 min at 72°C. An equal volume of loading solution composed of EDTA (10 mM) and Orange G dye (13.6 mM) (Sigma) dissolved in formamide was added to the reaction following PCR, and 0.6 μl of the product was fractionated on denaturing acrylamide gels (6.0% acrylamide, 7.7 M urea, 89 mM Tris, 89 mM borate, 2.5 mM EDTA, pH 8.3).

For use of fluorescent-labelled primers, 45 ng of template DNA is dried in the wells of 96 well polypropylene plates. PCR amplifications were carried out in a 4 μl volume containing 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.001% gelatin, 100 μM each dNTP, 0.075 μM of fluorescent-labelled forward and unlabeled reverse primer, and a 0.12 units of Taq polymerase. PCR amplification was carried out for 27 cycles with the same times and temperatures as listed above.

Genetic map positions for new STRPs

Genetic distances for the new STRPs were obtained by typing the STRPs in several projects with large numbers of European families. The CRIMAP program was used to order the STRPs and to deduce genetic distances. In order to fit new STRPs into the Marshfield map [20], approximate genetic values were obtained by extrapolations using the new sex-average genetic distances and the Marshfield map genetic distances for two flanking, older STRPs. In rare instances, when no neighbouring STRPs with known Marshfield map distances were available, the genetic distances were extrapolated from physical distances from the UC-Santa Cruz sequence assembly, June 2002 version. For the X-chromosome analysis, female genetic distances were used in place of sex-average genetic distances.

Heterozygosity values were determined by typing STRPs in two different population groups. For Cooperative Human Linkage Center (CHLC) and Utah STRPs in Set 12, heterozygosity values were deduced by typing the STRPs through several populations of different ethnic groups (African, Asian and European), whereas for newly developed STRPs in Set 12 and all the STRPs within Set 52 (newly developed, CHLC, and Utah STRPs), a European population was used. Heterozygosity estimates of the Set 10 (and many Set 12) STRPs are also available from genotyping of the Human Diversity Panel [see Marshfield web site and reference [27]].

Authors' contributions

NG led the building of Screening Sets 11, 12, 51 and 52 and drafted the manuscript. DV led the building of Set 10. AY worked out conditions for initial screening of new STRPs. CZ carried out ePCR and STR computer searches. GM identified candidate polymorphisms from overlapping BAC sequences. JLW conceived the study and coordinated all efforts. All authors read and approved the final manuscript.

Acknowledgements

We thank Jan Wood, Heather Pagenkopf, Jocelyn Schroeder, Thao Le, Robert Kuntz, Vani Natarajan, Jessica Kayhart, Jennifer Kislow, Matt Williamson and Kate Buehler for expert laboratory assistance. This work was supported through NHLBI Contract HV48141 for the Mammalian Genotyping Service.

References

1. Zhao C, Heil J and Weber JL **A genome-wide portrait of short tandem repeats.** *Am J Hum Genet* 1999, **65(supplement):**A102
2. Tóth G, Gáspári Z and Jurka J **Microsatellites in different eukaryotic genomes: survey and analysis.** *Genome Res* 2000, **10:**967-981
3. Wijmenga C, Frants RR, Brouwer OF, Moerer P, Weber JL and Padberg GW **Location of the fascioscapulohumeral muscular dystrophy gene on chromosome 4.** *Lancet* 1990, **336:**651-653
4. Weber JL and Broman KW **Genotyping for human whole-genome scans: past, present, and future.** *Adv Genet* 2001, **42:**77-96
5. Brinkmann B, Klintschar M, Neuhuber F, Hühne J and Rolf B **Mutation in human microsatellites: Influence of the structure and length of the tandem repeat.** *Am J Hum Genet* 1998, **62:**1408-1415
6. Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC and Sutherland GR **Incidence and origin of "null" alleles in the (AC)n microsatellite markers.** *Am J Hum Genet* 1993, **52:**922-927
7. Kruglyak L **Prospects for whole-genome linkage disequilibrium mapping of common disease genes.** *Nat Genet* 1999, **22:**139-144
8. Simonic I, Gericke GS, Ott J and Weber JL **Identification of genetic STRPs associated with Gilles de la Tourette Syndrome in an Afrikaner population.** *Am J Hum Genet* 1998, **63:**839-846
9. Ober C, Abney M and McPeek MS **The genetic dissection of complex traits in a founder population.** *Am J Hum Genet* 2001, **69:**1068-1079
10. Ophoff RA, Escamilla MA, Service SK, Spesny M, Meshi DB, Poon W, Molina J, Fournier E, Gallegos A and Mathews C **Genomewide linkage disequilibrium mapping of severe bipolar disorder in a population isolate.** *Am J Hum Genet* 2002, **71:**565-574
11. Dubovsky J., Sheffield VC, Duyk GM and Weber JL **Sets of short tandem repeat polymorphisms for efficient linkage Screening of the human genome.** *Hum Mol Genet* 1995, **4:**449-452
12. Yuan B, Vaske D, Weber JL, Beck J and Sheffield VC **Improved Set of short tandem repeat polymorphisms for screening the human genome.** *Am J Hum Genet* 1997, **60:**459-460
13. Sheffield VC, Weber JL, Buetow KH, Murray JC, Even DA, Wiles K, Gastier JM, Pulido JC, Yandava C and Sunden SL **A collection of tri- and tetranucleotide repeat STRPs used to generate high quality, high resolution human genome-wide linkage maps.** *Hum Mol Genet* 1995, **4:**1837-1844
14. Utah Marker Development Group **A collection of ordered tetranucleotide repeat markers from the human genome.** *Am J Hum Genet* 1995, **57:**619-628
15. Weber JL, David D, Heil J, Fan Y, Zhao C and Marth G **Human** *Am J Hum Genet* 2002, **71:**854-862
16. Gastier JM, Pulido JC, Brody T, Sheffield VC, Weber JL, Buetow KH, Murray JC, Hudson TJ and Duyk GM **Survey of trinucleotide repeats in the human genome: assessment of their utility as genetic markers.** *Hum Mol Genet* 1995, **4:**1829-1836
17. Brownstein MJ, Carpten JD and Smith JR **Modulation of non-templated nucleotide addition by *Taq* DNA polymerase: primer modifications that facilitate genotyping.** *Biotechniques* 1996, **20:**1004-1010
18. Magnuson VL, Ally DS, Nyland SJ, Karanjawala ZE, Rayman JB, Knapp JI, Lowe AL, Ghosh S and Collins FS **Substrate nucleotide-determined non-templated addition of adenine by *Taq* polymerase: implications for PCR-based genotyping and cloning.** *Biotechniques* 1996, **21:**700-709
19. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B and Masson G **A**

high-resolution recombination map of the human genome. *Nat Genet* 2002, **31:**241-247
20. Broman KW, Murray JC, Sheffield VC, White RL and Weber JL **Comprehensive human genetic maps: individual and sex-specific variation in recombination.** *Am J Hum Genet* 1998, **63:**861-869
21. Jin L, Zhong Y and Chakraborty R **The exact numbers of possible microsatellite motifs.** *Am J Hum Genet* 1994, **55:**582-583
22. Beckmann JS and Weber JL **Survey of human and rat microsatellites.** *Genomics* 1992, **12:**627-631
23. Giglio S, Calvari V, Gregato G, Gimelli G, Camanini S, Giorda R, Ragusa A, Guerneri S, Selicorni A and Stumm M **Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation.** *Am J Hum Genet* 2002, **71:**276-285
24. Weber JL **The Iceland Map.** *Nature Genet* 2002, **31:**225-226
25. Cullen M, Perfetto SP, Klitz W, Nelson G and Carrington M **High-resolution patterns of meiotic recombination across the human major histocompatibility complex.** *Am J Hum Genet* 2002, **71:**759-776
26. Lynn A, Koehler KE, Judis L, Chan ER, Cherry JP, Schwartz S, Seftel A, Hunt PA and Hassold TJ **Covariation of synaptonemal complex length and mammalian meiotic exchange rates.** *Science* 2002, **296:**2222-2225
27. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW **Genetic structure of human populations.** *Science* 2002, **298:**2381-2385
28. Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL **High resolution of human evolutionary trees with polymorphic microsatellites.** *Nature* 1994, **368:**455-457
29. Deka R, Jin L, Shriver MD, Yu LM, DeCroo S, Hundrieser J, Bunker CH, Ferrell RE and Chakraborty R **Population genetics of dinucleotide (dC-dA)n (dG-dT)n polymorphisms in world populations.** *Am J Hum Genet* 1995, **56:**461-474
30. Calafell F, Shuster A, Speed WC, Kidd JR and Kidd KK **Short tandem repeat evolution in humans.** *Eur J Hum Genet* 1998, **6:**38-49
31. Huttley GA, Smith MW, Carrington M and O'Brien SJ **A scan for linkage disequilibrium across the human genome.** *Genetics* 1999, **152:**1711-1722
32. Varilo T, Paunio T, Parker A, Perola M, Meyer J, Terwilliger JD and Peltonen L **The interval of linkage disequilibrium detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories.** *Hum Mol Genet* 2003, **12:**51-59
33. Weber JL and Wong C **Mutation in short tandem repeat polymorphisms.** *Hum Mol Genet* 1993, **2:**1123-1128
34. Chakraborty R, Kimmel M, Stivers DN, Davison LJ and Deka R **Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci.** *Proc Natl Acad Sci USA* 1997, **94:**1041-1046
35. de Kniff P **Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome.** *Am J Hum Genet* 2000, **67:**1055-1061
36. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006