J Hum Genet (2002) 47:567–575

ORIGINAL ARTICLE

Gale Brightwell · Rachel Wycherley · Gemma Potts
Andrew Waghorn

A high-density SNP map for the FRAX region of the X chromosome

Received: June 26, 2002 / Accepted: July 22, 2002

Abstract Single-nucleotide polymorphisms (SNPs) are the most common type of genetic variation within the human genome, occurring approximately once every kilobase. However, for association studies, SNPs are not as informative as microsatellite markers and a large number of SNPs and substantial population sizes are required for linkage and mapping studies. A SNP map was generated for the FRAX region of the X chromosome, approximately 0.8 Mb proximal and 1.8 Mb distal to the FRAXA repeat, at a density of at least 1 SNP every 100 kb. SNPs were identified in a population of 28 women with a FRAXA expansion (including three women with a FRAXE expansion) on a background of different DXS548, CA1 and CA2 haplotypes, and a normal X chromosome with a different microsatellite haplotype. Fifty-four polymorphisms were identified in a total of 52 257 bp distributed over 2.6 Mb. This represented about 1 SNP every 1024 bp, which was consistent with a nondesert region (1 : 1000 bp). Because the SNPs identified in this study have haplotype and frequency data from an affected population, they should provide a useful resource for researchers to investigate the genetic mechanisms behind instability and expansion of both FRAXA and FRAXE triplet repeats.

Key words Fragile X · FRAXA · FRAXE · Single-nucleotide polymorphism · SNP map · dHPLC

Introduction

The FRAX region on Xq27–q28 contains two loci that are subject to dynamic mutation, FRAXA and FRAXE. Expansion and subsequent methylation of a CGG trinucleotide repeat (FRAXA) located in the 5′ untranslated region of the *FMR1* gene gives rise to the fragile X syndrome, an inherited disorder associated with mental retardation (Fu et al. 1991; Verkerk et al. 1991). Distal to *FMR1* (approximately 0.77 Mb) is *FMR2*, which contains a GCC triplet repeat (FRAXE) in the 5′ untranslated region. Full mutations (over 200 repeats) in FRAXE are rare (Sutherland and Baker 1992) but appear to be associated with nonsyndromic mental impairment (Knight et al. 1994).

Instability at FRAXA is thought to be due at least in part to loss of an AGG interspersion from the CGG repeat, resulting in slippage during replication (Eichler et al. 1994). However, haplotype studies using microsatellites suggest that other factors may also be important determinants of risk for FRAXA expansion (Murray et al. 1997). Eichler et al. (1996) proposed three independent mechanisms for repeat expansion: (1) loss of 3′ stabilizing AGG interspersion followed by rapid expansion to a pre- or full mutation, (2) slow expansion from common to pre- and full mutation without loss of an AGG interspersion, and (3) generalized instability that affects other polymorphic microsatellite markers. Instability at FRAXE usually appears to increase with increasing repeat number. In contrast to FRAXA, the repeat does not contain any interspersions, suggesting a simple relationship between size of repeat and risk of expansion. However, work by Ennis et al. (2001) suggested that, although triplet repeat size was a significant predisposing factor for FRAXE expansion, there were other genetic determinants involved. For example, they described a significant association between unusual FRAXA and unusual FRAXE alleles. FRAXA repeat sizes of more than 50 were positively associated with FRAXE repeats of less than 11. Although these associations occurred on similar haplotype backgrounds, founder effects could not be confirmed without looking at comparative data from different ethnic backgrounds.

With the near completion of the human genome sequencing project, there has been much interest in the use of single-nucleotide polymorphisms (SNPs) for studying genetic factors associated with complex disease traits. SNPs

G. Brightwell (✉) · R. Wycherley · G. Potts · A. Waghorn
Wessex Regional Genetics Laboratory, Salisbury District Hospital, Salisbury, Wiltshire SP2 8BJ, UK
Tel. +44-1722-425047; Fax +44-1722-338095
e-mail: galebrightwell@hotmail.com

are the most common type of genetic variation within the human genome, occurring approximately once every kilobase (Wang et al. 1998). They have several advantages over microsatellites: SNPs have low mutation rates and are biallelic and hence, can more easily accommodate analysis with statistical computer packages. However, for association studies, SNPs are not as informative as microsatellite markers. Therefore, a large number of SNPs and substantial population sizes are required. Taillon-Miller and Kwok (2000) developed a high-density SNP map for Xq25–q28 with an average distance between SNPs of about 100 kb. However, this map did not include the FRAX region at Xq27.3–q28. Mathews et al. (2001) previously carried out a SNP analysis within the *FMR1* gene and concluded that long contiguous regions must be studied to accurately understand the phylogeny and evolutionary mechanisms behind fragile X.

The present study aimed to generate a SNP map for the FRAXA and FRAXE repeat region of the X chromosome at a similar density to that of Taillon-Miller and Kwok (2000). To look for possible *cis*-acting factors and to investigate the phylogeny and evolutionary mechanisms behind fragile X, we have extended the region of the X chromosome analyzed to approximately 1 Mb proximal and 2 Mb distal to the FRAXA repeat, which also included the genes *FMR2* and *IDS*.

Materials and methods

DNA samples

Individual genomic DNA samples were obtained from the Wessex Regional Genetics Laboratory, Salisbury District Hospital. Each individual DNA was genotyped for FRAXA and FRAXE repeat size (Murray et al. 1996) and the microsatellites DXS548, FRAXAC1, FRAXAC2 (Macpherson et al. 1994; Oudet et al. 1993; Jacobs et al. 1993). Primate (*Pan troglodytes*, PTR9 and *Pan paniscus*, PPA2) DNA samples were obtained from Mariano Rocchi, DAPEG-Sezione di Genetica, Bari, Italy. This study was approved by the United Kingdom's National Health Service Regional and Multi-Centre Ethical Committees.

Target sequences

Target sequences for SNP detection were identified using the Golden Path Working Draft Genome Browser August 2001 freeze (http://genome.ucsc.edu/index.html), which mapped the *FMR1* gene sequence to a contig containing 13 overlapping GenBank sequences (AC016897.4, AL589669.10, AL137841.9, AL13742.9, AL096861.9, AL592439.4, AL450484.1, AL009048.1, AC007538.5, AL450486.1, AC016925.15, L29074.1, and AC006054.2) and the *FMR2* gene sequence to a contig containing 9 overlapping GenBank sequences (U40455.1, AC079462.2, AC006399.6, AC002368.1, AC006516.10, AC0015552.12, AC006522.5, AC005731.2, and AC002523.1). There still remains a gap between the FRAXA and FRAXE contigs of unknown size, which has been arbitrarily given a size of 200 kb. This gave a region of approximately 2.8 Mb in total. Sequences for polymerase chain reaction (PCR) product amplification for denaturing high-pressure liquid chromotography (dHPLC, WAVE) analysis were identified at approximately every 100 kb, 0.8 Mb proximal and 1.8 Mb distal to (avoiding repetitive elements for which sequence data was available) the FRAXA triplet repeat. Before the availability of computer software for repetitive element screening, all primer sequences were blasted using the National Centre for Biotechnology Information (NCBI) BLAST facility on the World Wide Web (WWW). PCR target sequences for direct sequencing, and chemical cleavage mismatch (CCM) were concentrated within 250 kb of the FRAXA repeat. If no SNP was detected, a new target sequence was identified within 20 kb of the original sequence until a SNP was confirmed.

Primers and PCR conditions

All primers were obtained from Interactiva (Ulm, Germany). All amplimers were designed using Primer3 (Rozen and Skaletsky 1998). PCR reactions for dHPLC heteroduplex analysis, CCM, and sequencing were carried out using Amplitaq Gold Kit reagents (Applied Biosystems, Warrington, UK) as per manufacturer's guidelines. DNA template final concentrations were 5 ng and primer concentrations were 0.1 μM. All PCR reactions were carried out using a MJ Research DNA Engine Tetrad Thermo Cycler (Waltham, MA, USA) as follows: 95°C for 15 min, 35 cycles of 95°C for 30 s, appropriate annealing temperature for 30 s, and 72°C for 30 s followed by 72°C for 10 min and a 4°C soak.

SNP detection

PCR amplification for CCM analysis was carried out using biotinylated primers and fluorescent R6G 2′-deoxyuridine 5′-triphosphate (dUTP) from Applied Biosystems (Warrington, UK). Heteroduplexes were prepared by heating PCR products at 95°C for 5 min followed by incubation overnight at 65°C. PCR fragments were then purified using Streptavidin MagneSphere Paramagnetic Particles from Promega (Southampton, UK), following the manufacturer's instructions. CCM was carried out as described in Gogos et al. (1990), except the cleaved products were analyzed using an ABI 377 and GeneScan software according to the manufacturer's guidelines. dHPLC was carried out using the WAVE system from Transgenomics (Crewe, UK) as per the manufacturer's recommendations. PCR products before analysis were denatured to allow heteroduplex species to form. Appropriate temperature for analysis was predicted using WaveMaker software. All sequencing reactions were carried out using the ABI Prism BigDye Terminator Cycle Sequence Ready Reaction Kit Applied Biosystems and an ABI 377 according to the manufacturer's guidelines.

Results

Polymorphisms identified

PCR primers were designed to amplify approximately 600-bp target sequences, and individual SNPs were detected either by direct sequencing, CCM, or dHPLC heteroduplex analysis (Table 1). Fifty-four polymorphisms (51 SNPs and 3 insertion/deletions) were identified as follows: 31 (57.5%) A/G (and C/T) transversions; 10 (18.5%) A/C (and G/T), 7 (13%) G/C, and 3 (5.5%) A/T transitions, and 3 (5.5%) deletions, ranging from 1 bp to 31 bp. Each polymorphism was confirmed by sequencing, and its Golden Path August 2001 freeze locations and SNP allele status are shown in Table 1. These findings were similar to those observed by Taillon-Miller (2000), who estimated that the mutational spectrum in humans and orangutans was A/G (and C/T, 63%), A/C (and G/T, 17%), C/G (8%), insertions/deletions (8%), and A/T (4%). However, we found 5.5% fewer A/G (C/T) transversions and 2.5% fewer deletion/insertions with a corresponding increase in C/G (5.5%), A/C (T/G)(1.5%), and A/T (1.5%) transitions within the Xq27.3–q28 region. The ancestral allele status for each SNP was determined by sequencing the appropriate DNA from either *Pan troglodytes* or *Pan paniscus* (Table 1).

Forty-one of the polymorphisms identified were novel and, of the 13 found in the NCBI SNP database, only one, Rs544682, had any details on allele frequency (Table 1) or study population. Approximately one third of the polymorphisms were found to be located in repetitive elements such as long interspersed element (LINEs) or short interspersed element (SINEs). To limit the possibility that the SNP identified in a repetitive element was not due to the amplification of two or more of these elements from different locations in the genome, we searched the primer sequences of each of the amplimers used to amplify the SNP target sequences against the GenBank sequence database using the NCBI standard nucleotide–nucleotide BLAST [blastn] WWW facility. If the primer pairs showed no significant homology with the consensus repetitive element or flanking regions, then the SNP was accepted as real. Two of the deletions, WEX3 (T) and WEX27 (31 bp), were located in repetitive elements. However, WEX26 (6-bp deletion) was in intron 5 of *FMR2* and was identified in three different women.

FRAXA and FRAXE triplet repeats are located in a non-SNP desert region of the X chromosome

Fifty-four polymorphisms (51 SNPs) were identified in a total of 52257 bp distributed over 2.6 Mb. This represented about one polymorphism every 968 bp and one SNP every 1024.6 bp, slightly more common than other estimates of SNP frequency by Taillon-Miller and Kwok (2000) in flanking regions Xq25 (1:1400 bp) and Xq28 (1:2600 bp), but consistent with a nondesert region (1:1000 bp).

Because the average incidence of SNPs in the FRAX region was one SNP per 1024 bp and the average size of PCR product screened was 600 bp, the probability of identifying one SNP in any 100-kb interval on the first attempt was 600/1024.6 and on the second attempt about 1200/1024.6. For this reason, any interval in which the SNP frequency was greater than 1:1500 nucleotides was designated as a region of genomic DNA where it was relatively difficult to find an SNP (Fig. 1 and Table 2). For example, intervals 1 (1:2702), 2 (1:1940), 16 (1:2638), and 17 (1:3990) were located between 650 and 850 kb upstream of *FMR1* and 100 kb proximal and distal to the start of *FMR2*. However, SNPs in this region were still fairly common and thus it did not correspond to an SNP desert ($1 < 10000$, Miller et al. 2001). For those intervals in which more than one SNP was identified, the average number of nucleotides per SNP was determined (Table 2). The SNPs identified in this study had an SNP incidence of between 1:338 (interval 25) and 1: 3990 (interval 17), consistent with the genes *FMR1* and *FMR2* residing in a non-SNP desert region of the X chromosome.

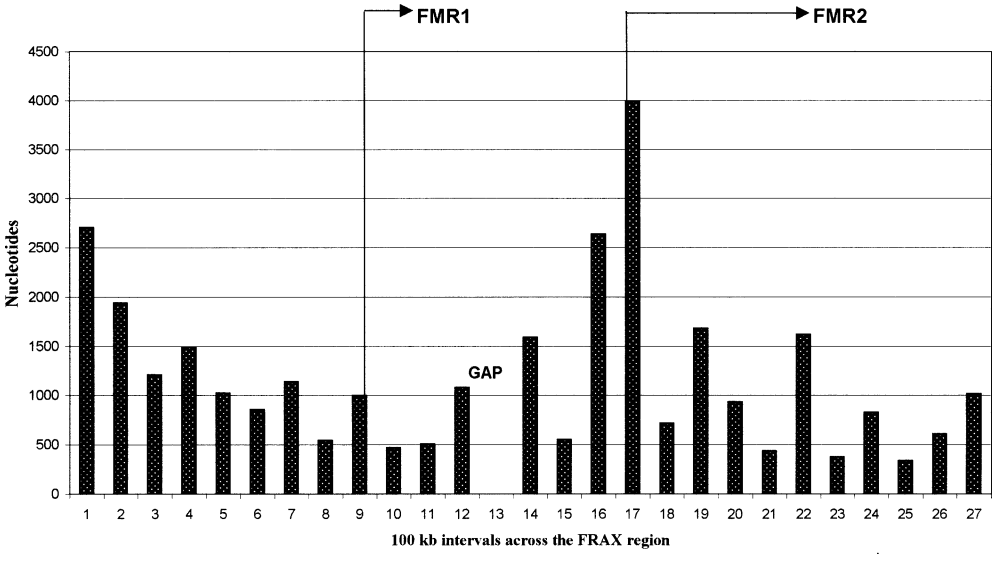


Fig. 1. Polymorphism incidence across the FRAX region of the X chromosome. The FRAX region of the X chromosome (Golden Path August Freeze 2001 locations 149.7 Mb to 152.4 Mb) was separated into 100-kb intervals (1–27). The incidence of polymorphisms per 100-kb interval is shown by the height of the columns. The *two arrows* indicate the length of the coding sequences of *FMR1* and *FMR2*. Interval 13 represents the arbitrary 200-kb gap between the two contigs containing *FMR1* and *FMR2*

Table 1. Polymorphisms identified in the FRAX region of the X chromosome

Golden Path location[a]	Polymorphism	Primate allele	Name	Identified by	5′ to 3′ Flanking sequence[b]	NCBI Database	Comments
149785093	T/C	T[c]	WEX62	dHPLC	AAAATAGGCCTGACTGTTGCACGTTTGT [T/C] TGTGCTGGGTTCATTGGGATGAGGCCCCTC	Novel	
149854302	A/T	T[c]	WEX47	dHPLC	TTTTTTACTTCCAGACATTAATGGTCTTA [A/T] ACATATTCATCCATTCTGAACGTAACAAAA	Rs2218611	No data available on dbSNP database
149942870	C/T	C[c]	WEX53	dHPLC	TATTCAAAATCCAGCAGACCATGTGCTG [C/T] TGCAAAATCCAGATGTTTCTGACTTTATTAT	Novel	
150055334	C/A	C[c]	WEX46	dHPLC	GCAACTAGGTAGTTTAGAGAGCAGATAG [C/A] ATAATGCTGTACCTATGGAAAAGATTCAGTA	Novel	
150155252	C/T	ND	WEX54	dHPLC	AGATTCATTTCTGCCATCACAGAGCTTTT [C/T] ACCAAATAATCTGTTAAATTTATATTCCAATG	Rs555559	No data available on dbSNP database
150155571	T/G	ND	WEX55	dHPLC	TTTTCAATTAAAATCAGTGATGATTATTT [T/G] AATATACACTTCTCTTTAGAGAGATTTCTCC	Rs609033	No data available on dbSNP database
150251763	T/C	T[c]	WEX33	dHPLC	CTTTGTCAAAAATCAGTTGGCTTCACATT [T/C] ATGAATTTATTTCTGGGTTCTTTATTATTGTTC	Novel	Present in L1, LINE
150251799	A/G	A[c]	WEX34	dHPLC	AGAGAATGTCCTTACCTCAATATGTGTTC [A/G] TGGCACCTTTGTCAAAAATCAGTTGGCTTCAC	Novel	Present in L1, LINE
150251825	C/G	C[c]	WEX32	dHPLC	CCAGTTATTCCAGCACCATTTATTGAAGA [C/G] AATGTCCTTACCTCAATATGTGTTC	Novel	Present in L1, LINE
150279808	A/G	G[c]	WEX25	dHPLC	GACTTCAATTCTTTAAAATTAGCAATGAT [A/G] TTTCACCATAGGTTTATAAATCAAGTAAGTAT	Rs2742911	No data available on dbSNP database
150369029	A/C	T[c]	WEX28	dHPLC	AGGGAGGAGTATGGGAATAGTTTACATA [A/C] AGGGATTAAAAATAGTTCACATACATACTCC	Novel	
150438312	G/A	A[d]	WEX43	dHPLC	CTAGTGTCATCAACTGTGACAGCTGATG [G/A] CACTGACCATGAAGGCCATCAAATTTTCTCT	Novel	
150438448	C/A	A[c]	WEX44	dHPLC	AGAATAGTTTCAGTTTCTCAGTTTAAATT [C/A] TGTGTTTCCATCATGGTCTATACTGTTAGGGT	Rs1868140	No data available on dbSNP database
150543464	A/G	ND	WEX8	CCM	CCATGCCACTGCATCAACAGTGCTCCAGT [A/G] TAATAGCAATAGATTACAATGGGAGAGCTGA	Novel	Present in L1MC2, L1, LINE
150543643	DEL T	+ T[c]	WEX3	CCM	CCCCAACTCCTAACTGATTAACATAAAAAG [T] CAAACTAAAACCTAATTACCTCGTTCCTATTA	Novel	Present in L1MC2, L1, LINE
150551214	C/A	C[c] C[d]	WEX1	CCM	ACATTATCTGTGTTAAATTATCAAGGATCT [C/A] TATCGAACATATTGCAGCTTGTGTCTAGAAGA	Novel	
150551332	T/C	T[c]	WEX4	CCM	TGTAAACAAGGACCCTGTAGGGACTGATA [T/C] GACAATGTGCTGAAAATTGAGGAGCAAAGTTA	Novel	
150552016	C/G	C[c] C[d]	WEX5	CCM	TTCATACCCCTTATCACAGCTGCAACTACT [C/G] ATTTACTTGTCTGACAATTTGATTTATGTCCAC	Rs1805420	Ancesteral allele C
150552318	G/A	A[d]	WEX6	CCM	GGGTTGCAAGGAGGTGCATCGGCCCCTGT [G/A] GACAGGACGCATGACTGCTACACACGTGTTCA	Novel	
150599327	T/G	T[c]	WEX20	SEQ	TGCATACACAGTGGTGGATCCAGAAGGG [T/G] AGCATCTGGGGCTGTCAGTCAATTATGCTCTC	Novel	Present in LTR16B, ERVL, LTR
150659826	G/C	ND	WEX16	SEQ	ATGGGGAAAGCATTCCCTATTTGATAAAT [G/C] GTGCTGGGAAAACTGGCTAGCCATATGTAGAA	Novel	Present in L1, LINE
150660613	T/C	C[c]	WEX17	SEQ	TGAATTCTTTGTTTGAAAGATAATATGTCT [T/C] TGTTTCTCCAGGATTAGTTTCTGGTGCCTTATT	Novel	Present in L1, LINE

Table 1. Continued

Golden Path location[a]	Polymorphism	Primate allele	Name	Identified by	5′ to 3′ Flanking sequence[b]	NCBI Database	Comments
150660751	A/G	A[c]	WEX18	SEQ	AGAGTTAGGTATTCATTGTACTGTTCACT [A/G] TCTGGGCTTGTTTGTACCTGTCTTTCTTCGGAC	Novel	Present in L1, LINE
150696756	T/C	T[c]	WEX51	dHPLC	CTAGTTCTGAGTTACATAGGAAACTTCTC [T/C] CAGAAAGGATTGAATTATGGGAATTTTTTGC	Novel	
150698054	G/A	G[c]	WEX52	dHPLC	CTGAAAAAAGCACTTTTAAACATTAACAC [G/A] TTCATCTTCCTATTTGTCTTTATTAATGAGA	Novel	
150764280	G/A	A[c]	WEX58	dHPLC	ATCCGTCCTCCCTCAGCCGTGTTGGACCT [G/A] CTGTGGTCTTATTTTCCCTCTTCACTTGAACA	Novel	
150764389	A/G	A[c]	WEX63	dHPLC	GATGGCAATCAACGTCCAAGAGTTCATAC [A/G] TTTACCAAAAAAGGGGACGTTTTTGGTTTAGA	Novel	
150764518	C/G	C[c]	WEX64	dHPLC	AGCTACTTATGTCTCCATTTTTTAAAGAAC [C/G] AATTGCAAGTCTAATGAGTACTAGTCCTAGTT	Novel	
150808016	G/A	T[c]	WEX10	SEQ	GATGACTCATTTGCATACTGCTGACTTCAG [G/A] TAGATTGAATATTTTCCTTGCACCAAGTATGGC	Novel	
151013423	A/C	C[c]	WEX31	dHPLC	CAGTGTCATACTTTCCATTGTCAGATAGTT [A/C] TTCTATAGGTCCGTTTATGTTTGATTCTAATTTC	Novel	Present in Tigger6a, MER2, DNA
151113269	T/C	T[c]	WEX19	dHPLC	AATCCTGTGGCTATTTTCCAACTTTAAACAC [T/C] AACTACAGTACTTTCCTCACTTTTTGGTTCATTTA	Novel	
151151916	G/A	A[c]	WEX50	dHPLC	ATTGTACAAAAATGTGAGAAAATGTTATTG [G/A] GAACTGCTTTAAAGAGCTTAGGACAATGATGGC	Rs2536592	No data available on dbSNP database
151251275	G/A	G[c]	WEX48	dHPLC	GCCTAGTGCAAGGCACACAGTAGATACTCA [G/A] TATTTTTTTTTATTTAATAATGCAGAAGATGAGA	Novel	MIR, SINE
151324005	G/C	C[c]	WEX61	dHPLC	CCCTCATCATTATAAACTAC [G/C] AATCTGCAT TTGGCCATGGACTGGAGGTTATATAAAGGGAC	Novel	Present in exon 2 of *FMR3*
151368030	T/A	T[c]	WEX56	dHPLC	AGGCTCTGCTAACCAGAAATTAAAAAAAA [T/A] TTTTTTTCATTAAGCATGTGATGGGTTATAGAG	Rs993421	Present in intron 1 of *FMR2* between run of As and Ts
151460356	G/A	A[c]	WEX21	dHPLC	TCATATATATATAAATATATACGACTGTCAA [G/A] CACATACGTAGAAGTACAAGGGCAATGCGCAAC	Novel	Present in intron 1 of *FMR2* in a REP3, ERV, LTR
151562024	C/T	C[c]	WEX38	dHPLC	AATTAACTTCCCATCAATAACTGTGTTCAAC [C/T] TAAAATAATTATAATAATAAATATTATTTTATTG	Novel	Present in intron 3 of *FMR2* in AT-rich repeat of low complexity
151638995	T/C	C[c]	WEX35	dHPLC	TTTCAAGGAGTTTCATGATGGGGAATCCTGA [T/C] ATTCTACTTCAAGACCCTTCTGCTTATTCCAGTG	Novel	Present in intron 4 of *FMR2*
15165592025	DEL 6BP	+ [GTGACT][c]	WEX26	dHPLC	TAGAGAAAAACGAATCTGGGGAGAGGA [GTGACT] GTCCAGGCATTCAGAAATACTCAACCTATTAGGG	Novel	Present in intron 5 of *FMR2*
151666232	C/G	C[c]	WEX41	dHPLC	TACTGTCCTCCCTTGAGCTGACAGATCACAC [C/G] TGGAACAGGATACCATGAGTTACAAGAGCCCCA	Rs1265396	Present in intron 5 of *FMR2* in L3, CR1, LINE; no data available
151666883	T/G	T[d]	WEX36	dHPLC	CCAGGACAGGTCAGTTCTCTTCCTTCCTGCA [T/G] TTTTGTTTGTCTTATTTTAATTATACCAATCTTCC	Novel	Present in intron 5 of *FMR2*
151734195	G/A	G[c]	WEX23	dHPLC	TTAATTTATTTTGGCTGACTTCCTTTATAGTT [G/A] GTACCAAATACAGAAATGGAGAGTACATAGATT	Rs1265416	Present in intron 9 of *FMR2*. No data available on dbSNP database
151782800	G/A	A[d]	WEX60	dHPLC	GCTCCAGCGAATCGGAGAGCAGCTCTGAGTC [G/A] GATTCAGACACTGAAAGTAGCACCACTGACAGCG	Novel	Present in exon 10 of *FMR2*
151833292323	DEL 31BP	+ [31 bp]	WEX27	dHPLC	GACTTTTAAAAATGCAAATGT [CAGGCTTCCACCT CAGAGATTGATTCCATAG] GTGTGTGTCAAAGGTC	Novel	Present in MER5B, Mer1-type, DNA repeat

Table 1. Continued

Golden Path location[a]	Polymorphism	Primate allele	Name	Identified by	5′ to 3′ Flanking sequence[b]	NCBI Database	Comments
151856666	G/T	T[c]	WEX45	dHPLC	GAGAGTTTCCCACAACTACAAAGGATATATT [G/T] CAGCAGAATGAAGAAAAATGATGGAAGAAGGAA	Novel	
151935815	A/G	A[c]	WEX22	dHPLC	AAAAGCTAAG CCTCAAAGAGCACATATTGT [A/G] TGATTCTGCTTATATGAAATGTCCAGAAAAGACA	Novel	Present in L1MB8, L1, LINE
151936625	A/T	ND	WEX49	dHPLC	GAATAGCTGGGACTGCAGGAACACACCACC [A/T] TGCCCAGGAAAATGTTTTAATTTTGTAGAGACAG	Novel	Present in AluJo, Alu, SINE
151936628	C/T	ND	WEX37	dHPLC	ATAGCTGGGACTGCAGGAACACACCACCATG [C/T] CCAGGAAAATGTTTTAATTTTGTAGAGACAGAGT	Novel	Present in AluJo, Alu, SINE
152038635	G/C	C[c]	WEX29	dHPLC	ATGAGGAAGCTCAAAATTTTGCTGTCTGCTAA [G/C] TTACATTTTTTGCTTTTCCAAAATATTAACTCTTTG	Novel	
152063259	G/A	ND	WEX39	dHPLC	ATGCTGGCAAGACTGTAGAGAATTAGGGAAC [G/A] CTTTTACATTTTTTGGTGGGAATGTAAATTAGTTCA	Novel	Present in L1PA10, L1, LINE
152102784	G/A	G[c]	WEX59	dHPLC	GTGTTTGTGTGTGTGTGCGTGTGCCCATGTGC [G/A] TATGTGCAGCTGTGTGCGTGCAAATGGTGTGGGT	Rs2056833	No data available on dbSNP database; present in simple (TG)n repeat
152102864	C/T	T[c]	WEX57	dHPLC	GATTCCCAGTGGGGAGGCTCAGAATATTTCTA [C/T] CACAGAAAGAGGGAGCAGAGCTCAGGGGTTATTC	Rs741733	No data available on dbSNP database
152225748	G/A	G[c]	WEX30	dHPLC	GCCTGAGTGCTGGTTTCACTTTGTGGCACCAA [G/A] CATTTATTCATAGAGGATTTTTATAGCCAACAACC	Novel	Present in LTR10B, ERV1, LTR
152343017	A/C	C[d]	WEX40	dHPLC	GAGTAAGCCCTGAGCACCACTGTCTAAAGAA [A/C] TTTATGGCCCTACAATGCTGAGATGTGGGTTCTACC	Rs544682	G — 0.130, T — 0.870 Present in *IDS2* pseudogene mRNA

NCBI database, dbSNP at http://www.ncbi.nlm.gov/SNP/index.html
Del, deletion polymorphism; +, does not have deletion polymorphism; ND, not determined; NCBI, National Center for Biotechnology Information
[a]Nucleotide numbering corresponds to location of polymorphism on the Golden Path August 2001 freeze map of the X chromosome
[b]Variation is shown 5′ to 3′ in brackets
[c]Allele status determined from *Pan troglodytes*
[d]Allele status determined from *Pan paniscus*

Table 2. Interval allocation and polymorphism incidence across the FRAX region of the X chromosome

100-kb Interval[a]	Golden Path location (Mb)	Distance from FRAXA repeat (100 kb)	Number of nucleotides screened	Number of polymorphisms identified	Incidence of polymorphisms[b]
1	149.7–149.8	−753 to −853	2702	1	2702
2	149.8–149.9	−653 to −753	1940	1	1940
3	149.9–150.0	−553 to −653	1207	1	1207
4	150.0–150.1	−453 to −553	1489	1	1489
5	150.1–150.2	−353 to −453	2044	2	1022
6	150.2–150.3	−253 to −353	3408	4	852
7	150.3–150.4	−153 to −253	1136	1	1136
8	150.4–150.5	−53 to −153	1082	2	541
9	150.5–150.6	+53 to −53	6979	7	997
10	150.6–150.7	+53 to +153	2350	5	470
11	150.7–150.8	+153 to +253	1515	3	505
12	150.8–150.9	+253 to +353	1080	1	1080
Gap					
14	151.0–151.1	+453 to +553	1590	1	1590
15	151.1–151.2	+553 to +653	1100	2	550
16	151.2–151.3	+653 to +753	2638	1	2638
17	151.3–151.4	+753 to +853	3990	1	3990
18	151.4–151.5	+853 to +953	1430	2	715
19	151.5–151.6	+953 to +1053	1678	1	1678
20	151.6–151.7	+1053 to +1153	3724	4	931
21	151.7–151.8	+1153 to +1253	874	2	437
22	151.8–151.9	+1253 to +1353	3234	2	1617
23	151.9–152.0	+1353 to +1453	1119	3	373
24	152.0–152.1	+1453 to +1553	1650	2	825
25	152.1–152.2	+1553 to +1653	676	2	338
26	152.2–152.3	+1653 to +1753	608	1	608
27	152.3–152.4	+1753 to +1853	1014	1	1014
Total	2.6		52257	54	

−, Distance proximal to FMR1; +, distance distal to FMR1

[a] Golden Path August Freeze 2001 locations, 149.7 Mb to 152.4 Mb, separated into 100-kb intervals (1–27)

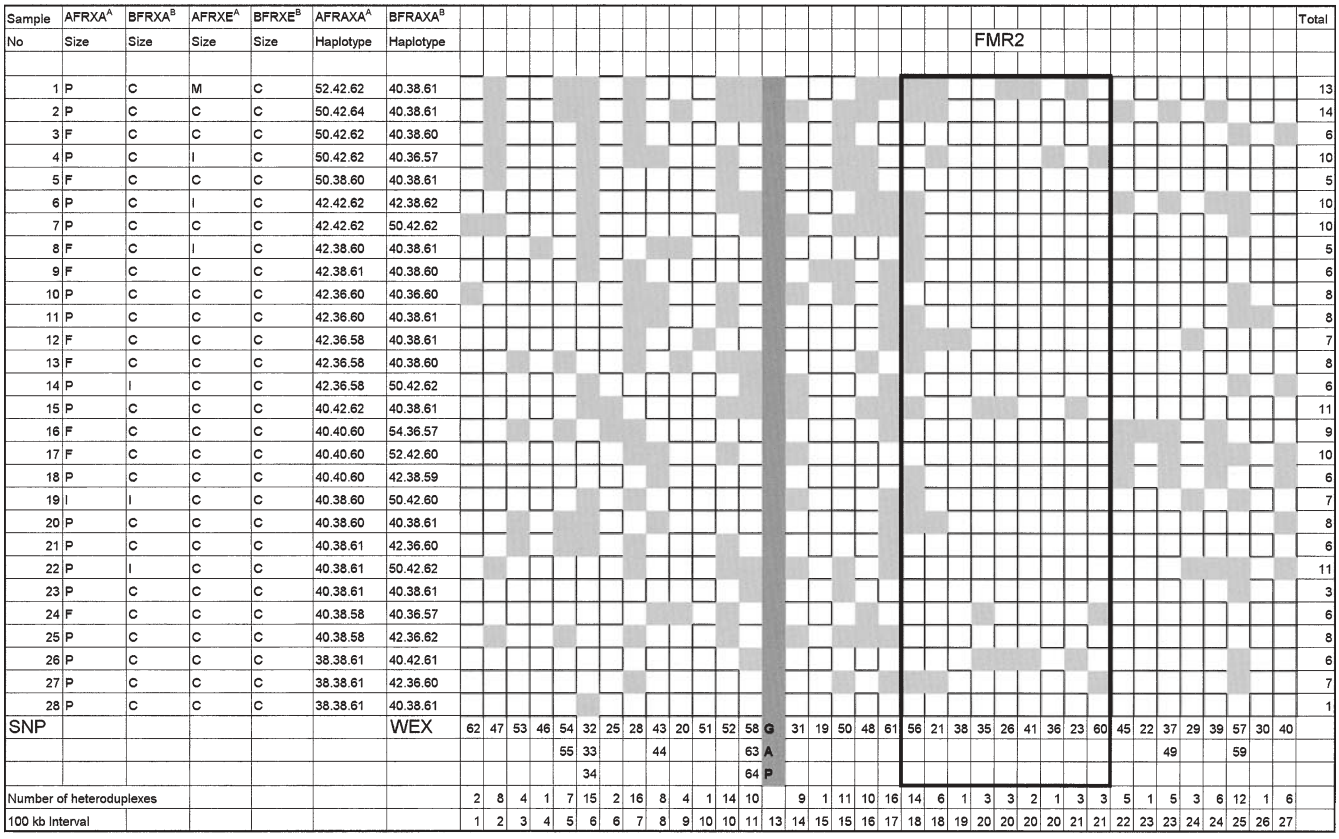
[b] Base pairs per polymorphism

SNP frequency in a population of women with FRAX expansions

To increase the probability of identifying SNPs associated with FRAX mutations, we generated a panel of 28 female genomic DNA samples in which each woman had either an intermediate (I, 41–60 repeats), pre- (P, 61–200 repeats), or full (F, over 200 repeats) FRAXA mutation on a background of different DXS548, CA1, and CA2 haplotypes, together with a normal X chromosome (C, 11–40 repeats) with a different microsatellite haplotype (Fig. 2). Three of these women (sample numbers 4, 6, and 8) also had an expansion at FRAXE and one woman (sample number 1) had a minimal FRAXE size of nine repeats. A normal male DNA sample was also included to act as a homoduplex control for heteroduplex analysis.

SNPs were identified in female DNA samples on the basis of heteroduplex formation generated by the PCR amplification of both copies of the X chromosome. Heteroduplexes were detected by either sequencing, CCM, or dHPLC. Individual SNPs were then confirmed by direct sequencing of the heteroduplex PCR product. Figure 2 shows the frequency of heteroduplexes in the population of 28 women with FRAX expansions for each PCR product containing a SNP analyzed by dHPLC. Within our panel, a SNP with an allele sample frequency of 50:50 would theoretically result in 14 heteroduplexes. Assuming Hardy-Weinburg, of the 56 chromosomes, 28 would be allele A and 28 allele B, giving 14 × AB heterozygotes, and 7 × AA plus 7 × BB homozygotes. A frequency of heteroduplexes greater than 14 may be indicative of an association with expansion, whether due to founder effect or an association with a *cis* element affecting triplet repeat instability. If only one heteroduplex was observed, then the sample frequency for that allele would be 1 in 56, i.e., 2:98. As the number of heteroduplexes increases, the chance of a homoduplex for both alleles increases, which would not be resolved by dHPLC. For this reason, it is difficult to determine the exact sample frequency for each SNP allele, and Fig. 2 represents an underestimate of the allele frequency for each SNP. Similarly, those PCR products identified as having more than one SNP will be an overestimation of allele frequency.

Of the 43 SNPs identified by dHPLC, 29 were unique to one PCR product, and six PCR products contained multiple SNPs (Fig. 2). The number of heteroduplexes identified in a individual PCR product ranged from 1 to 16. There appeared to be no relationship between the incidence of polymorphisms in any 100-kb interval and the frequency of heteroduplexes. For example, intervals 1, 2, 16, and 17 had a relatively low polymorphism incidence (greater than 1:1500), but the number of heterozygous women identified for each individual polymorphism ranged from 2 to 16.



Sample No	AFRXA[A] Size	BFRXA[B] Size	AFRXE[A] Size	BFRXE[B] Size	AFRAXA[A] Haplotype	BFRAXA[B] Haplotype	Total
1	P	C	M	C	52.42.62	40.38.61	13
2	P	C	C	C	50.42.64	40.38.61	14
3	F	C	C	C	50.42.62	40.38.60	6
4	P	C	I	C	50.42.62	40.36.57	10
5	F	C	C	C	50.38.60	40.38.61	5
6	P	C	I	C	42.42.62	42.38.62	10
7	P	C	C	C	42.42.62	50.42.62	10
8	F	C	I	C	42.38.60	40.38.61	5
9	F	C	C	C	42.38.61	40.38.60	6
10	P	C	C	C	42.36.60	40.36.60	8
11	P	C	C	C	42.36.60	40.38.61	8
12	F	C	C	C	42.36.58	40.38.61	7
13	F	C	C	C	42.36.58	40.38.60	8
14	P	I	C	C	42.36.58	50.42.62	6
15	P	C	C	C	40.42.62	40.38.61	11
16	F	C	C	C	40.40.60	54.36.57	9
17	F	C	C	C	40.40.60	52.42.60	10
18	P	C	C	C	40.40.60	42.38.59	6
19	I	I	C	C	40.38.60	50.42.60	7
20	P	C	C	C	40.38.60	40.38.61	8
21	P	C	C	C	40.38.61	42.36.60	6
22	P	I	C	C	40.38.61	50.42.62	11
23	P	C	C	C	40.38.61	40.38.61	3
24	F	C	C	C	40.38.58	40.36.57	6
25	P	C	C	C	40.38.58	42.36.62	8
26	P	C	C	C	38.38.61	40.42.61	6
27	P	C	C	C	38.38.61	42.36.60	7
28	P	C	C	C	38.38.61	40.38.61	1

SNP	WEX	62	47	53	46	54	32	25	28	43	20	51	52	58	G	31	19	50	48	61	56	21	38	35	26	41	36	23	60	45	22	37	29	39	57	30	40
						55	33			44				63	A																	49			59		
							34							64	P																						
Number of heteroduplexes		2	8	4	1	7	15	2	16	8	4	1	14	10		9	1	11	10	16	14	6	1	3	3	2	1	3	3	5	1	5	3	6	12	1	6
100 kb Interval		1	2	3	4	5	6	6	7	8	9	10	10	11	13	14	15	15	16	17	18	18	19	20	20	20	20	21	21	22	23	23	24	24	25	26	27

Fig. 2. Haplotype status and single-nucleotide polymorphism (*SNP*) frequency in a population of females with FRAX expansions determined by denaturing high-pressure liquid chromatography *Light grey squares* represent a heteroduplex product. The *dark grey column* indicates the location of the gap between the two contigs containing *FMR1* and *FMR2*. The *boxed region* shows the location of the *FMR2* coding sequences. The total number of heteroduplexes for each female sample is shown in the *last column* and the total number of heteroduplexes for each SNP is shown in the *penultimate row*. AFRXA, FRAXA repeat size on the expanded X chromosome[A]; BFRXA, FRAXA repeat size on the normal X chromosome[B]; AFRXE, FRAXE repeat size on the expanded X chromosome[A]; BFRXE, FRAXE repeat size on the normal X chromosome[B]; AFRAXA, DXS548.CA1.CA2 microsatellite haplotype for chromosome[A]; BFRAXA, DXS548.CA1.CA2 microsatellite haplotype for chromosome[B]. FRAXA repeat size is indicated by: *M*, minimal, <10 repeats; *C*, common, 11–40 repeats; *I*, intermediate, 41–60 repeats; *P*, premutation, 61–200 repeats; *F*, full mutation, 200 repeats

Similarly, for those 100-kb intervals in which it was relatively easy to find polymorphisms (less than 1:1000), the number of heteroduplexes ranged from 1 to 16. However, the distribution of heteroduplex frequency was not consistent across the FRAX region (Fig. 2). For 100-kb intervals 1 to 17, the number of heteroduplexes for each PCR product ranged from 1 to 16 and were fairly randomly distributed among the female samples. In comparison, for intervals 18 to 24, the number of heteroduplexes only ranged from 1 to 6 and, of the 13 polymorphisms identified in this region, 8 appeared to be found in a subset of individuals. For example, heteroduplexes for SNPs WEX35, 26, 41, and 23 were identified in individuals 1, 15, and 26 and for SNPs WEX45, 37/49, and 39, heteroduplexes were found in individuals 2, 6, 16, 17, and 18. Although it was relatively easy to find polymorhisms in the FRAX region between 100-kb intervals 18 and 24, it would appear that the frequency and distribution of the alleles in this region for our FRAX population was restricted compared with intervals 1 to 17. This observation does not appear to be haplotype related, because the heterozygous individuals have a range of different DXS548, CA1, and CA2 haplotypes. The gap between the two sequence contigs containing either *FMR1* or *FMR2* has been given an arbitrary value of 200 kb by Golden Path; this makes any estimate of the distance between FRAXA and FRAXE inaccurate. However, using Golden Path's data, *FMR2* was located between 100-kb intervals 17 to 21, and the boundary between the region of low incidence (intervals 18 to 24) and random incidence of heteroduplexes (intervals 1 to 17) occurs in intron 1 of *FMR2*. The region of low incidence of heteroduplexes extends over a region of approximately 700 kb, which includes the remaining coding sequences of *FMR2* plus another 300 kb distal to *FMR2.*

Discussion

Evidence of *cis*-acting factors affecting triplet repeat stability has been demonstrated by Cleary et al. (2002), who looked at the effect of the location of an SV40 origin of replication on (CTG)n/(CAG)n triplet repeat stability in

primate cells. Cleary et al. (2002) showed that the repeat, depending on the position of the SV40 origin proximal or distal to the CAG/CTG repeat, either remained stable or underwent an expansion, a deletion, or both. In addition, *cis*-acting control elements have been identified over 100 kb away from the associated gene. For example, Pfeifer et al. (1999) found several campomelic dysplasia translocation and inversion cases mapping to >130 kb proximal SOX9. No evidence of other genes or transcripts was found in this region, suggesting that chromosomal rearrangement had removed one or more *cis*-regulatory elements from an extended SOX9 control region. The genetic mechanisms behind triplet repeat expansions and the relationship between FRAXA and FRAXE mutations are still not fully understood, and more work is needed to investigate the effects of *cis*- or *trans*-acting elements on repeat stability.

We have started to genotype a large population of men from the Wessex region of the United Kingdom, with a range of FRAXA and FRAXE (1 to >200) repeat sizes, using four SNPs identified in this study (WEX1, WEX10, WEX17, and WEX28) and two from the HGP SNP database (ATL1 and FMRb) (Brightwell 2002). Each SNP correlated with a distinct haplogroup (A, B, C, D, and E), previously identified by the microsatellite DXS548, FRAXAC1, and FRAXAC2 repeat patterns (described in Ennis et al. 2001). In our predominantly Caucasian population, these haplogroups have been shown to be associated with FRAXA repeat instability (Ennis et al. 2001). For example, the majority of individuals in haplogroup C with FRAXA mutations undergo a rapid expansion to a pre- or full-sized (over 200 repeats) mutation with the loss of a 3′ stabilizing AGG interspersion from the CGG triplet repeat. In contrast, individuals in group D usually undergo expansion slowly without the associated loss of an AGG interspersion. It is important to study populations from a number of different ethnic backgrounds to dissect founder effects from molecular causes of repeat instability. Because the markers identified in this study are from an affected population, they should provide a useful resource for researchers to investigate the genetic mechanisms behind instability and expansion of both FRAXA and FRAXE triplet repeats.

Acknowledgments This work was funded by the Wellcome Trust. We thank Patricia Jacobs, Sarah Ennis, and James Macpherson for comments and suggestions on this manuscript.

References

Brightwell G, Wycherley R, Waghorn A (2002) SNP genotyping using a simple and rapid single-tube modification of ARMS illustrated by analysis of 6 SNPs in a population of males with FRAXA repeat expansions. Mol Cell Probes Aug 16(4): 297

Cleary JD, Nichol K, Wang Y-H, Pearson CE (2002) Evidence of *cis*-acting factors in replication-mediated trinucleotide repeat instability in primate cells. Nat Genet 31:37–45

Eichler EE, Holden JJA, Popovich BW, Reiss AL, Snow K, Thibodeau SN, Richards CS, Ward PA, Nelson DL (1994) Length of uninterrupted CGG repeats determines instability in the FMR1 gene. Nat Genet 8:88–94

Eichler EE, Macpherson JN, Murray A, Jacobs PA, Chakravarti A, Nelson DL (1996) Haplotype and interspersion analysis of the FMR1 CGG repeat identifies two different mutational pathways for the origin of the fragile X syndrome. Hum Mol Genet 5:319–330

Ennis S, Murray A, Morton NE (2001) Haplotypic determinants of instability in the FRAX region: concatenated mutation or founder effect? Hum Mutat 18:61–69

Fu YH, Kuhl DPA, Pizzuti A (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell 67:1047–1058

Gogos JA, Karayiorgou M, Aburatani H, Kafatos FC (1990) Detection of single base mismatches of thiamine and cytosine residues by potassium permanganate and hydroxylamine in the presence of tetralkylammonium salts. Nucleic Acids Res 18:6807–6814

Jacobs PA, Bullman H, Macpherson J, Youings S, Rooney V, Watson A, Dennis NR (1993) Population studies of the fragile (X): a molecular approach. J Med Genet 30:454–459

Knight SJL, Oelckel MA, Hirst MC, Flanery AV, Moncla A, Davies KE (1994) Triplet repeat expansion at the FRAXE locus and X-linked mild mental handicap. Am J Hum Genet 55:81–86

Macpherson JN, Bullman H, Youings SA, Jacobs PA (1994) Insert size and flanking haplotype in fragile X and normal populations: possible multiple origins for the fragile X mutation. Hum Mol Genet 3:399–405

Mathews DJ, Kashuk C, Brightwell G, Eichler EE, Chakravarti A (2001) Sequence variation within the fragile X locus. Genome Res 11:1382–1391

Miller RD, Taillon-Miller P, Kwok P-Y (2001) Regions of low single-nucleotide polymorphism incidence in human and orang-utan Xq: deserts and recent coalescences. Genomics 71:78–88

Murray A, Youings S, Dennis N, Latsky L, Lineham P, McKechnie N, Macpherson J, Pound M, Jacobs P (1996) Population screening at the FRAXA and FRAXE loci: molecular analyses of boys with learning difficulties and their mothers. Hum Mol Genet 5:727–735

Murray A, Macpherson JN, Pound MC, Sharrock A, Youings SA, Dennis NR, Mckechnie N, Linehan P, Morton NE, Jacobs P (1997) The role of size, sequence and haplotype in the stability of FRAXA and FRAXE alleles during transmission. Hum Mol Genet 6:173–184

Oudet C, Mornet E, Serre JL, Thomas F, Lentes-Zengerling S, Kretz C, Deluchat C, Tejada I, Boue J, Boue A, Mandel JL (1993) Linkage disequilibrium between the fragile X mutation and two closely linked CA repeats suggest that fragile X chromosomes are derived from a small number of founder chromosomes. Am J Hum Genet 52:297–304

Pfeifer D, Dewer KR, Devon K, Lander ES, Birren B, Korniszewski I, Back E, Scherer G (1999) Campomelic dysplasia translocation breakpoints are scattered over 1 Mb proximal to SOX9: evidence for an extended control region. Am J Hum Genet 65:111–124

Rozen S, Skaletsky HJ (1998) Primer3 (http://www.genome.wi.mit.edu/genome software/other/primer3.html)

Sutherland GR, Baker E (1992) Characterisation of a new rare fragile site easily confused with the fragile X. Hum Mol Genet 1:111–113

Taillon-Miller P, Kwok P-Y (2000) A high-density single nucleotide polymorphism map of Xq25–28. Genomics 65:195–202

Verkerk AJMH, Pieretti M, Sutcliffe JS (1991) Identification of a gene (*FMR-1*) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65:905–914

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES (1998) Large-scale identification, mapping and genotyping of single nucleotide polymorphisms in the human genome. Science 280:1077–1082