ELSEVIER

Mutation Research Genomics 406 (1999) 115–120

Mutation Research Genomics

www.elsevier.com/locate/mrgenomics
Community address: www.elsevier.com/locate/mutres

Polymorphisms in the human DNA repair gene XPF

Fan Fan *, Cheng-pin Liu, Simon Tavaré, Norman Arnheim

Program in Molecular Biology, University of Southern California, Los Angeles, CA 90089-1340, USA

Received 25 January 1999; accepted 25 May 1999

Abstract

DNA sequence polymorphisms were sought in the coding region and at the exon–intron boundaries of the human XPF gene, which plays a role in nucleotide excision repair. Based on a survey of 38 individuals, we found six single nucleotide polymorphisms, one in the 5′ non-coding region of the XPF gene, and five in the 2751 bp coding region. At each site, the frequency of the rarer allele varies from about 0.01 to over 0.38. Except for the 5′ non-coding and one coding sequence polymorphism, the rarer alleles for the remaining four polymorphisms were found only in heterozygotes. Of the five polymorphisms in the coding region, one is silent, one results in a conserved amino acid difference, and the remaining three result in non-conserved amino acid differences. Because of its biological function in nucleotide excision repair, functionally significant XPF gene polymorphisms are candidates for influencing cancer susceptibility and overall genetic stability. Nucleotide sequence diversity estimates for XPF are similar to the lipoprotein lipase and beta-globin genes.

Keywords: Human XPF gene; Nucleotide excision repair; Single nucleotide polymorphism, SNP; Nucleotide diversity; Mutation detection

1. Introduction

The nucleotide excision repair (NER) pathway can recognize and repair a broad range of DNA lesions, including those generated by environmental DNA-damaging agents. The NER pathway includes steps that recognize the DNA lesion, melt the helix, excise part of the damaged DNA strand, synthesize new DNA using the undamaged strand as a template, and finally, ligation [1,2]. Defects in the NER pathway, as observed in xeroderma pigmentosum patients, leads to UV sensitivity, development of cancer and decreased survival.

Many proteins are involved in the NER pathway. The XPF gene, which was initially identified from xeroderma pigmentosum complementation group F (XPF) individuals, encodes a key protein in the NER pathway [1,3,4]. The XPF polypeptide forms a complex with the ERCC1 protein and carries out an incision at the 5′ end of the recognized DNA lesion, while the XPG protein carries out an incision at the 3′ end of the lesion. The XPF family of genes is evolutionarily conserved. Extensive homology exists between human XPF, *Drosophila* Mei-9, *Saccharomyces cerevisiae* RAD1, and *S. pombe* Rad16 [3,4], all of which have similar functions in NER.

Because of its importance in DNA repair, we searched for polymorphisms in the XPF gene among

* Corresponding author. Department of Genetics, Beckman Center B203, Stanford University School of Medicine, Stanford, CA 94305, USA. Tel.: +1-650-725-8091; fax: +1-650-723-1399; E-mail: fanfan@cmgm.stanford.edu

1383-5734/99/$ - see front matter
PII: S1383-5726(99)00008-4

38 independent DNA samples. The significance of any polymorphic changes could be further evaluated for possible effects on the efficiency of NER in vitro and on cancer susceptibility and genetic instability.

2. Material and methods

2.1. DNA sample preparation and PCR

DNA from a total of 38 individuals, mostly Caucasian, were analyzed. A frozen semen sample from each individual was lysed and used as PCR template [5]. Each 50 μl amplification reaction contained 10 mM Tris–HCl (pH 9.0), 50 mM KCl, 2.5 mM $MgCl_2$, 0.1% Triton X-100, 40 μM of each dNTP, 0.2 μM of each primer, and 2 units of Taq polymerase. PCR was carried out for 30 or 35 cycles of: denaturation at 95°C for 30 s, annealing at 60°C for 30 s and extension at 72°C for 1 min. Primers (Table 1) were designed to produce PCR product of about 150–200 bp that covered the coding sequences and the intron–exon boundaries. Polymorphisms in exon 11 were screened by direct sequencing of PCR products. Polymorphisms in exon 1 to exon 10 were found using single strand conformation polymorphism (SSCP) analysis followed by sequencing of PCR products from samples with aberrant SSCP patterns.

2.2. SSCP analysis

PCR product was mixed with SSCP loading buffer (90% formamide, 20 mM EDTA, 0.05% xylene cyanol and 0.05% bromophenol blue), denatured and analyzed on 12% acrylamide gels (29:1) in 192 mM glycine, 25 mM Tris–HCl. The gels were silver stained to visualize the bands. Samples of known sequences were used as a positive control to confirm the sensitivity of the electrophoresis conditions for detecting a single nucleotide polymorphism.

2.3. Sequencing

PCR product was purified and sequenced using the fmol DNA cycle sequencing system (Promega) with ^{33}P labeled primer under standard conditions. A total of 30 cycles of denaturation at 95°C for 30 s, annealing at 60°C for 30 s, and extension at 72°C for 45 s were carried out. Homozygotes for the rarer allele were not detected in several cases in our study (see below). To verify the accuracy of these sequences, PCR products from the heterozygous individuals were cloned and several independent clones from each individual were sequenced using an ABI377 automated sequencing system.

3. Results

The XPF coding sequence is 2751 bp in length and is made up of 11 exons [4]. We screened for polymorphic variants in the entire coding region as well as at the intron–exon boundaries in 38 individual samples. We found three single nucleotide polymorphisms (SNPs) where the rarer allele had a frequency greater than 0.1, and another three with frequencies of about 0.01 (Table 2).

One polymorphism, g2603 T/A, occurs in the 5′ region of the XPF gene (the position 2603 is determined according to the published XPF genomic sequence with GenBank accession number L76568), http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=1905925&form=6&db=n&Dopt=g. Heterozygotes as well as both homozygotes were detected with an allele frequency of 0.378 for the rarer A allele. This polymorphism has been reported with a frequency of 0.46 for the A allele when 12 DNA samples were screened for SNPs by direct sequencing [6]. This difference changes the T/C at the last position in one of the ten core repeat TTCGGC(T/C) sequences in the region upstream of the starting ATG codon as identified by Brookman et al. [4].

The remaining two SNPs with frequency greater than 0.1 for the rarer allele were found in the coding region of the XPF gene. All the polymorphisms in the coding region are numbered according to the cDNA sequence with the A position at the starting ATG codon as +1 [4]. Polymorphism 1244 G/A in exon 8 results in an arginine or a glutamine at amino acid 415 (R415Q). The glutamine allele was found in seven heterozygotes among 33 individuals (allele frequency of 0.106). Polymorphism 2505 T/C is

Table 1
Primers

Exon	Primer[a]	Position[b]	Primer sequence
1	19	g2021	TGA GTT CGG CCT ACT CTC CA
	42L	g2179	CGC ACA CTA CTA GCC CGT CA
	43	g2159	GTG CTG GAA CTG CTC GAC AC
	20L	g2323	ACA CTC TCG GAG TCC CCT CA
2	21	g3898	TGA CCT ATT AAA AAC TGC CCT G
	44L	g4090	CAG TCA AGA AG TCA ACC ACA AG
	9	g4044	TGA AGT TTA CAC ACA AGG TGG T
	22L	g4202	TCT CTG GAG AAA AAT AAA ATG GA
3	23	g8418	AGA AAA ATG TGA TGA ATG AAT GG
	45L	g8595	ATC AAA GGC AAC AGC ATT GTC
	47	g8563	AAA ACA AAC GTG GTT TTA TTA AAG
	46L	g8742	AGT GCA ACT AAG TTA CAG TAG A
4	25	g9884	GTT AAA CCA AGA CCA GAA ATA C
	48L	g10025	GTA TAG CAA GCA TGG TAG GTG T
	49	g10001	GTT GTA GAA ATC CAT GTT TCT AT
	26L	g10196	ACT TGC ATT CTT TAA AAT TAC AAC
5	27	g12556	AAT CCT TTT GAA AGT ATG ATT TGT A
	12L	g12714	ACT GCA GCA AAG TTC GTA ATA
	50	g12697	AGT TCA GGA TTT GAA GAT ATT AC
	28L	g12867	ATT CCC CCT AAA ACC TAA CAC
6	63	g13976	AGG AAG ACA GGA TGA CAG CCA G
	64L	g14290	TTT TCA CAT GGC CAA AGA AGA C
7	31	g16046	GTT TTA AAA GCC TTT GGA AGA C
	32L	g16276	GAG CTA ATT AAA TCA GGA CTC A
	13	g16151	CCA AAG TGG GAG GCA CTG AC
8	33	g17011	GCA CAG GGA AAC TAG GAG GA
	14L	g17232	TGA GAT TTC CTA ATT CTC TTT GAA C
	15	g17183	AAG GAT AGC AAA GCT GAA GA
	52L	g17385	GAC GAT ATC CTT CCT CGA CAT
	17	g17369	CTG GAA GAG GAA GGA GAT GT
	16L	g17513	AGA AGC GGA TGG ATG ATA GT
	53	g17506	ACC CCT CAC TAT CAT CCA TC
	34L	g17715	TGG TAT AGT TTC TGG GTT AAA TC
9	35	g19647	TAA AAA TGC TGT TTT TCC AAC CT
	36L	g19795	GGG GCT TTC ATA CTG ATT TTC T
10	59	g26647	CAG GGA AAA AGC AAG CAT G GT
	60L	g26774	AAT GGA GGT GCC TGA CAA AGT
11	3	g29548	GGA ACA GAA TGG TAC ACA GCA A
	6L	g30283	GCA TGG GAT AAG AAA ACA GCC
	65	g29440	CAT CCA TCA GAG TTA ACA ACA G
	41L	g29684	TTG CGC TCC ACG CAC ATT TC
	4L	g29902	TCC GTA GTC TGG GGA AGT GA
	39	g29870	GAC ATT AGT TCC AAA CTC ACT C

[a] Primers ending with an ''L'' correspond to the antisense strand.
[b] Primer positions are numbered according to published XPF/ERCC4 genomic sequence (GenBank L76568, http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid = 1905925&form = 6&db = n&Dopt = g). ''g'' indicates genomic sequence numbers.

silent and does not change the amino acid at position 835 (S835S). The rarer C allele has an frequency of 0.344. This polymorphism was also reported by Sijbers et al. [3] and Shen et al. [6] at a frequency of 0.38 in Shen's report. In our study both homozygotes and heterozygotes for the rarer allele were detected.

Table 2
SNPs identified in the XPF gene

cDNA bp	Exon no.	Nt change	AA change	Genotypes			Total[e]	# Chrom	$p(N)$	$q(n)$	Reference
				NN	Nn	nn					
g2063[a]	5′ non-coding	T/a	core repeat	TT 15	a/T 16	aa 6	37	74	0.622	0.378	this study, [6]
1244	8	G/a	R415Q[b]	GG 26	a/G 7	aa 0	33	66	0.894	0.106	this study
1727	8	G/c	R576T[b]	GG 36	c/G 1	cc 0	37	74	0.986	0.014	this study
2117	11	T/c	I706T[c]	TT 35	c/T 1	cc 0	36	72	0.986	0.014	this study
2505	11	T/c	S835S[d]	TT 14	c/T 14	cc 4	32	64	0.656	0.344	this study, [6]
2624	11	A/g	E875G[b]	AA 31	g/A 1	gg 0	32	64	0.984	0.016	this study

[a] g2063 indicates that this position is determined according to the published genomic sequence with GenBank accession number L76568, http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid = 1905925&form = 6&db = n&Dopt = g.
[b] Indicates non-conserved amino acid substitutions.
[c] Indicates conserved amino acid substitutions.
[d] Indicates silent amino acid substitutions.
[e] Total number of sequences counted differ for each SNP because the data were not available on all 38 individuals.

For each of the next three sequence variants, only a single heterozygote was detected. One in exon 8 was ascertained by SSCP screening followed by sequencing, and the other two in exon 11 by direct sequencing. Sequence variant 1727 G/C results in a non-conservative amino acid difference; the rarer allele has a threonine instead of an arginine at amino acid 576 (R576T). Sequence variant 2117 T/C results in a conservative isoleucine to threonine difference at amino acid 706 (I706T). Sequence variant 2624 A/G results in a non-conservative amino acid difference, the rarer allele has a glycine instead of glutamic acid at position 875 (E875G). The frequency for the rarer alleles is about 0.015. We confirmed these variants by directly sequencing PCR products that were obtained by reamplification from the original templates of the heterozygous individuals. This minimized the possibility that the variants arose in the earliest PCR cycles due to base misincorporation. We also cloned PCR product from each heterozygous individual and sequenced several independent clones. In every case we detected each of the two alleles previously deduced from direct sequencing of PCR product. Some cloned fragments also contained nucleotide substitutions at positions other than the polymorphic site that are not observed in direct sequencing of PCR products. However, the nucleotide changes and the position of these changes were random, and were presumably due to Taq polymerase misincorporation.

Of the five polymorphic variants, we found in the coding region of XPF, two occur in exon 8 and three in exon 11. Clustering is also observed for 14 disease causing mutations found in XPF patients; seven are in exon 8 and two are in exon 11 [3,7]. Exon 8 (598 bp) and exon 11 (731 bp) are the two largest exons and represent about half of the total length of the XPF coding sequence. More data are needed to decide whether this distribution is simply due to the larger size of these two exons.

4. Discussion

In our survey of the XPF gene, we observed six SNPs, one in the 5′ non-coding and five in the coding region. Among the five in the coding region, one is silent, one results in a conserved amino acid difference, and the other three result in nonconserved amino acid differences. The frequencies for the three least common coding sequence variants predicts that, in each case, the frequency of homozygotes for the rare allele is expected to be around 2×10^{-4}, which could be the reason why we did not find any homozygotes in our samples. None of these rare alleles has been observed in XPF patients [3,7].

The XPF family of proteins, including human XPF, *Drosophila* MEI-9, *S. cerevisiae* RAD1, and *S. pombe* Rad16, have extended regions with sequence similarity and are functionally conserved [4].

The 3′ end of exon 8 and all of exon 11 encode the XPF region homologous to Mei-9, Rad16 and RAD1. Based on the species alignment of XPF homologous proteins, SNP 2624A/G is most interesting. This position can be a glutamic acid (XPF), glutamine (MEI-9 and Rad16), or lysine (RAD1), all relatively bulky amino acids. The nonconserved substitution of the relatively bulky and negatively charged glutamic acid by the small neutral amino acid glycine (E875G) in the human XPF gene may influence the DNA repair function of the protein. Functionally significant SNPs in the XPF gene may contribute to individual differences in the fine details of DNA repair. Since mutations in different NER genes that interfere with the NER pathway result in increased susceptibility to cancer as presented in XP patients [8], subtle differences due to polymorphisms in the XPF and/or other NER genes may alter cancer susceptibility. Genetic instability of simple repeated sequences might also be influenced by XPF polymorphisms. XPF/ERCC1, which makes incisions at the 5′ end of DNA loops [3,9,10], may contribute to the repair of large trinucleotide repeat containing loops that are generated due to replication slippage and are too long to be repaired by the postreplicative DNA mismatch repair system [11]. Polymorphisms in enzymes involved in large loop repair could be responsible for the observed variation in stability of similar-sized trinucleotide repeat disease alleles among different individuals [12].

It has been estimated that the number of SNPs ranges from 0.5 to 10 per 1000 bp when any two human chromosomes are compared [13,14]. In the 2751 bp XPF coding region, we found five SNPs among the 38 samples we surveyed, which corresponds to about one variable site in every 550 nucleotides. These calculations are based on data obtained using SSCP screening before sequencing (for exons 1 to 10) as well as sequencing without a screening step (for exon 11). It is well known that SSCP does not pick up all sequence variations. Although, theoretically, direct sequencing of PCR product could find all the variation, technical imperfections, such as uneven band intensity and background bands due to artifactual polymerase stops can cause sequence ambiguity and lower the sensitivity of this method, especially in heterozygotes. In our study, only the base changes that are unambiguously detected and confirmed in multiple runs of sequencing both strands are considered valid. We could be underestimating the number of SNPs in the XPF region.

Shen et al. found seven SNPs in their screen for XPF SNPs using 12 individuals [6]. Two, g2063 T/A and 2505 T/C, were also observed in this study. Four intronic SNPs were not seen in our study because we focused only on intron sequences immediately adjacent to the splice sites. One of their coding region SNPs, 1135 C/T (P379S) in exon 7 with a frequency of 0.08 for the rarer allele, was not observed in our study. It is possible that the samples we screened did not contain this SNP. Alternatively, the different success rates of the two mutation detection strategies may be the explanation since our study screened exon 7 using SSCP.

To estimate the sequence variability in the XPF coding region, we calculated two widely used measures of nucleotide diversity, π (defined to be the average pairwise number of segregating sites per nucleotide; Ref. [15]) and θ_W (Watterson's segregating sites estimator; Ref. [16]). In Table 3, we give the diversity statistics π and θ_W for exon 8 (where SSCP screening followed by sequencing was used), exon 11 (where only sequencing was used), and for the whole XPF gene. The number of sequences being compared differs among these regions (see Table 3) because data on all 38 individuals were not available for all SNPs. To compare nucleotide diversity values across loci, we also calculated the estimates of the variability of π and θ_W, which can be approximated

Table 3
Nucleotide diversity in XPF gene

Region	Length (bp)	# Sequences	π	sd(π)	θ_W	sd(θ_W)
Exon 8	598	66	3.31×10^{-4}	4.64×10^{-4}	7.02×10^{-4}	2.53×10^{-3}
Exon 11	731	60	7.24×10^{-4}	6.80×10^{-4}	8.80×10^{-4}	2.62×10^{-3}
Whole gene	2751	52	2.58×10^{-4}	2.19×10^{-4}	4.02×10^{-4}	9.61×10^{-4}

by the standard deviation of π and θ_W under the infinitely-many-sites mutation model; (Ref. [15], Eq. 10.9; Ref. [16]). These values are given in Table 3 under the column headed sd(π) and sd(θ_W). The nucleotide diversity data on XPF exon 11 can be compared to data on the lipoprotein lipase (LPL; [17]) and β-globin genes [18] that were also obtained by direct sequencing. The estimated nucleotide diversity π in the LPL coding region was $5.0 \times 10^{-4} \pm 5.0 \times 10^{-4}$ based on sequencing DNA from 76 individuals. In sequencing, a β-globin gene region in 24 unrelated Melanesians, Fullerton et al. found a single polymorphic site in 444 bp of exon sequence, corresponding to a nucleotide diversity of $\pi = 9.1 \times 10^{-4} \pm 9.5 \times 10^{-4}$ and $\theta_W = 5.1 \times 10^{-4} \pm 2.3 \times 10^{-3}$. The nucleotide diversity estimates in exon 11 and the whole XPF gene shown in Table 3 are consistent with the diversity estimates in both the LPL and β-globin genes.

In the 38 randomly selected samples we have screened, at least 13 are heterozygous for one of the five polymorphic variants in the coding region, and five are heterozygous for two polymorphic variants. Therefore, even for the functionally important XPF gene, five polymorphic sites are found in the 2751 bp coding region, and almost half of the randomly selected individuals screened are heterozygous for at least one polymorphic mutation, even when the heterozygosity for several of the SNPs is not high.

Acknowledgements

We are grateful to Dr. Uta Francke for critical reading of the manuscript and for the use of ABI sequencing facilities. We thank Ms. Kim Bogard for excellent technical assistance. This work is supported by NIH grant GM36745 (N.A.).

References

[1] A. Sancar, DNA excision repair, Annu. Rev. Biochem. 65 (1996) 43–81.
[2] R.D. Wood, Nucleotide excision repair in mammalian cells, J. Biol. Chem. 272 (1997) 23465–23468.
[3] A.M. Sijbers, W.L. de Laat, R.R. Ariza et al., Xeroderma pigmentosum group F caused by a defect in a structure-specific DNA repair endonuclease, Cell 86 (1996) 811–822.
[4] K.W. Brookman, J.E. Lamerdin, M.P. Thelen et al., ERCC4 (XPF) encodes a human nucleotide excision repair protein with eukaryotic recombination homologs, Mol. Cell. Biol. 16 (1996) 6553–6562.
[5] E.P. Leeflang, R. Hubert, L. Zhang et al., Single sperm typing, in: E. Board (Ed.), Current Protocol in Human Genetics, Wiley.
[6] M.R. Shen, I.M. Jones, H. Mohrenweiser, Nonconservative amino acid substitution variants exist at polymorphic frequency in DNA repair genes in healthy humans, Cancer Res. 58 (1998) 604–608.
[7] Y. Matsumura, C. Nishigori, T. Yagi et al., Characterization of molecular defects in xeroderma pigmentosum group F in relation to its clinically mild symptoms, Hum. Mol. Genet. 7 (1998) 969–974.
[8] J.E. Cleaver, K.H. Kraemer, Xeroderma pigmentosum and Cockayne syndrome, in: C.R. Scriver, A.L. Beaudet, W.S. Sly, D. Valle (Eds.), The metabolic and molecular bases of inherited disease, McGraw-Hill, New York, 1995, pp. 4393–4419.
[9] W.L. de Laat, E. Appeldoorn, N.G.J. Jaspers et al., DNA structural elements required for ERCC1–XPF endonuclease activity, J. Biol. Chem. 273 (1998) 7835–7842.
[10] T. Matsunaga, C.H. Park, T. Bessho et al., Replication protein A confers structure-specific endonuclease activities to the XPF–ERCC1 and XPG subunits of human DNA repair excision nuclease, J. Biol. Chem. 271 (1996) 11047–11050.
[11] P. Modrich, R. Lahue, Mismatch repair in replication fidelity, genetic recombination, and cancer biology, Annu. Rev. Biochem. 65 (1996) 101–133.
[12] E.P. Leeflang, S. Tavaré, P. Marjoram et al., Analysis of germline mutation spectra at the Huntington disease locus supports a mitotic mutation mechanism, Hum. Mol. Genet. 8 (1999) 173–183.
[13] A.J. Jeffreys, DNA sequence variants in the Gγ-, Aγ-, δ-, and β-globin genes of man, Cell 18 (1979) 1–10.
[14] D.N. Cooper, B.A. Smith, H.J. Cooke et al., An estimate of unique DNA sequence heterozygosity in the human genome, Hum. Genet. 69 (1985) 201–205.
[15] M. Nei, Molecular Evolutionary Genetics, Columbia Univ. Press, New York, 1987.
[16] G.A. Watterson, On the number of segregating sites in genetical models without recombination, Theor. Popul. Biol. 7 (1975) 256–276.
[17] D.A. Nickerson, S.L. Taylor, K.M. Weiss et al., DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene, Nature Genet. 19 (1998) 233–240.
[18] S.M. Fullerton, R.M. Harding, A.J. Boyce et al., Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus, Proc. Natl. Acad. Sci. U.S.A. 91 (1994) 1805–1809.