A putative nucleic acid-binding domain in Bloom's and Werner's syndrome helicases

Bloom's syndrome and Werner's syndrome are autosomal recessive disorders associated with a marked predisposition to a variety of cancers; at the cellular level, the hallmark of both syndromes is chromosomal instability[1–3]. The genes carrying mutations causing Werner's and Bloom's syndromes have been positionally cloned and sequenced, and both have been shown to encode proteins highly similar to helicases of the RecQ family[4,5]. The Werner's and Bloom's syndrome proteins (WRNp and BLMp, respectively) consist of over 1400 amino acids each, but except for the helicase domain no other motifs indicative of possible functions had originally been found. We have recently identified a domain in WNRp with highly significant similarity to RNase D and that is also related to the 3′–5′ proofreading exonuclease domain of bacterial DNA polymerase I (Ref. 6). This domain in WRNp is predicted to possess exonuclease activity[6]; it is missing in BLMp and other RecQ helicases (Fig. 1). Additional functions of BLMp are suggested by experiments on its apparent yeast ortholog, Sgs1, which is required for chromosome segregation and that interacts with topoisomerases I and II (Refs 7, 8). Furthermore, it has been shown that RecQ DNA helicase from *Escherichia coli* is a suppressor of illegitimate recombination, which may be directly relevant to the role of both WRNp and BLMp in the maintenance of the human genome stability[9].

All RecQ-like proteins, including WRNp and BLMp, contain, compared with other DEAH/DEAD helicases, a C-terminal extension that can be divided into two domains. While the first, proximal domain is apparently restricted to RecQ-like helicases, we show here that the distal domain is also present in RNase D and its eukaryotic homologs as well as in a

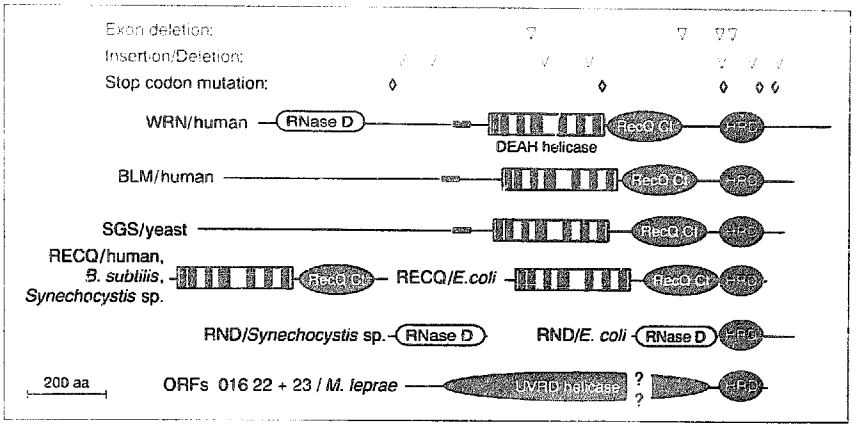

Figure 1

Modular structure of proteins containing the HRDC domain. The HRDC domain is conserved in RecQ helicases and RNase D homologs. RecQ Ct denotes a conserved domain in RecQ-like helicases that is not present in other homologs DEAH helicases and thus carries a distinct function. The location of the seven conserved motifs of DEAH helicases[16] are indicated (dark red); a blue bar preceding some helicase domains is a significant negative charge run as revealed using the program SAPS[17]. Different classes of mutations in WRNp and BLMp are indicated at the top. *Mycobacterium leprae* UvrD, a superfamily I helicase only distantly related to RecQ[16] is encoded by two neighboring ORFs, which might be due to a frameshift.

Predicted 2D	LL-hHHHHHHHHhHHHHHHhh-LLLLL------hHHHHHHHHLLL--hhhh---LLLLhhHHhhHHHHHHHHHHH--LLLL	from	to	Acc. No.
RecQ				
WRN/HUMAN	QETQIV GK VE QK ANKM VP-PAI TNK D KM TTVEN R DG SEGK AMLAP- E KH CQTNSVQ	1150	1229	L76937
BLM/HUMAN	EEMVKK GE TE CFS GKVFGVH YFN NTV K ES SSDPEV Q DG TEDK EKYGAE S QK SEWTSPA	1212	1292	P54132
SGS1_YEAST	LNNLRM ER RE SLN GNRMVPP VGN PDS K AI MNDSA T GT -EDK RRRFKY A AD SKKRSSE	1272	1351	P35187
YABC_SCHPO	IDVMTR KD KL ESN AIDSR-VSS TDS S AKK PRNVKE E HG SNEK VNLGPK Q QK IDEKEQN	1116	1195	Q09811
RECQ_ECOLI	GNYDRK AK RK EKS ADESNVP-PYV NDA E AEQ PITASE S NG GMRK ERFGKP A RA VDGDDEE	527	607	P15043
HI0728/HAEIN	ANYDKD AR RF EKQ ADKENIP-PYI NDA QE AQY PTSNIE Q NG GSIK ERFGQP A QE KAILANA	535	615	U32756
F18C5.2/CAEEL	PEKIDQ SR DD KVG ANMHEVA-PFQ SNT DC ANL PISASN EM DG SAQQ SRYGKR D VQ SKETGIA	736	816	U29097
T04A11/CAEEL	GDVFTR QD TH ITA AESSGLSGPYS SREG EQ AAL PRTNSD R DS TQIK TRYGRL E AT WKQVDGM	33818	34101	Z83123
RNase D				
RND_ECOLI	RTRQLA QL AD RLR ARERDLA-VNF REE S ARY GSLGE DS GLSGSEI FHGKT- A EK QTLPEDA	210	289	P09155
RND_HAEIN	NPLELS RV AQ RQN AIERDLA-LSY KSE K AKNN PRNTSE E GL ENEV VRGKE- Q SQ RRISSND	240	318	P44442
U1764u/MYCLE	DRRALA RE WA RDH AQRRDIA-PRR PDT D AIAD PKTIDE A PV GGANQRRSAAM A ET RQSQDLP	245	325	U15181
ORFQ/BACNO	RKIHQN EA VQ RET AEKYDLP-RRK SDE E ALAQ PQKFDD A LP EHYF AEEN-- E EQ RTQTPPA	143	220	U17138
PMSC_HUMAN	NTQQLT QL FA RDK ARREDES-YGY PNH K AEE PKEPQG A CN VPPL RQQINE L QQ REMPLLK	503	583	Q01780
C14A4.4/CAEEL	NTRQDY TH FK RDV ARAEDES-PHF PNH S SET PRDVGG A CN LPYF KQRTGD K VE RDVKLEF	489	569	Z49909
YASB_SCHPO	GREALM RA HD RDS ARKEDES-VRY PNR A AAS PVEAAD S SKQLTPI RMYVED K QE EKLYNEQ	442	522	Q10146
UNC733/YEAST	PPEREV RE YQ RDL ARRDDES-PRF PNQ A AY PTDVIG S TNGVTEH RQNAKL N RD LRNIKNT	435	515	U43491
UvrD				
016_23/MYCLE	ADVDEE LQ KA ALS AKEQNVP-AYV TDN A AEL PADEAA A PG SVRK EQYGSD Q RC AVAVRTQ	161	241	U00016

Figure 2

Multiple alignment of HRDC domains. The first column gives the protein or SWISS-PROT database names followed by the species (BACNO, *Bacteroides nodosus*; CAEEL, *Caenorhabditis elegans*; HAEIN, *Haemophilus influenzae*; MYCLE, *Mycobacterium leprae*; SCHPO, *Schizosaccharomyces pombe*). Residues invariant in more than 60% of the sequences are red; conserved hydrophobic positions are shown in green. Secondary structure predictions using the PhD program[18] are shown above the alignment (H/h denotes helix positions with 82%/72% expected accuracy, L is a loop position predicted with 90% confidence). Database searches with the region of WRNp that is not present in several other RecQ-like proteins readily delineated the HRD domain; for example, using the recently developed program PSIBLAST (Position-Specific Iterative BLAST, an extension of the BLAST2 algorithm[19]), which produces position-specific weight matrices from the BLAST2 output and employs them to iterate the database search (A. J. Schaffer *et al.*, pers. commun.). Screening of the non-redundant protein sequence database at the National Center for Biotechnology Information (NIH, Bethesda, MD, USA) with the HRDC domain from BLMp using PSIBLAST retrieved the HRDC sequences from WRNp and from *Escherichia coli* RecQ with $P \sim 10^{-5}$ and $P \sim 10^{-3}$, respectively. Motif and profile searches further confirmed the statistical significance of the findings (for details of the search strategy see Ref. 21). For example, motif searches using MoST[22] with HRDC domains of helicases identified regions in RNases D with probabilities of chance matches as low as $P = 3.08 \times 10^{-6}$. The alignment was constructed using the MACAW program[20] and was statistically significant ($P < 1C^{-4}$).

Published by Elsevier Science Ltd. 0968–0004/97 PII: S0968-0004(97)01128-6

Mycobacterium leprae UvrD helicase (Fig. 1; hereinafter HRDC domain, after Helicase and RNase D C-terminal). The conservation of the HRDC domain is supported, at a statistically significant level, by database screening and multiple alignment analysis (see legend to Fig. 2). In all three distinct groups of proteins that contain the HRDC domain, it is located C-terminal of either a nuclease or a helicase domain (Fig. 1). The HRDC domain appears not to be essential for either activity as it is missing from RecQ-like helicases from several species, from RNase D from *Synechocystis* sp. and from the *Drosophila* developmental protein Egalitarian[10], which contains an RNase D domain. The presence of this domain only in proteins from purple bacteria and eukaryotes (with the only exception of *M. leprae* UvrD) suggests that WRN and BLM genes may have originated from the mitochondrial genome.

Inspection of the multiple alignment of the HRDC domains shows primarily conservation of bulky, hydrophobic residues, without a single polar residue being invariant in all of the sequences; the predicted all-α secondary structure (Fig. 2) also strongly suggests that HRDC does not harbor an enzymatic activity. The most obvious common denominator among all the proteins containing HRDC is their binding to RNA or DNA, pointing to a possible role in nucleic acid-binding. Given the existence of helicases and RNases D without this domain, HRDC seems not to be the principal determinant of DNA or RNA binding though.

Multiple mutations have already been mapped to WRNp[4,11–14] and a considerable fraction is within the HRDC domain supporting its functional relevance. Although most of the mutations mapped to HRDC so far result in truncations, there is a considerable chance that the N-terminal domains remain functional and only HRDC is destroyed. Furthermore, HRDC belongs to one of three potential regions of mutational susceptibility[11] with identical mutations in various ethnic groups. Fewer mutations have been mapped to BLMp and they remain to be detected in the HRDC domain. However, whereas mutations in the helicase domain of the BLMp homolog in yeast, SGS1, have no apparent effect *in vivo*, a truncated version missing the HRDC domain was non-functional[15]. Elucidation of the role of HRDC domain is expected to be important for deciphering the molecular underpinning of the pathogenesis of Werner's and Bloom's syndromes, and more generally, for understanding the role of helicases in genome stability.

References

1 Ellis, N. A. (1996) *Nature* 381, 110–111
2 Epstein, C. J. and Motulsky, A. G. (1996) *BioEssays* 18, 1025–1027
3 Lombard, D. B. and Guarente, L. (1996) *Trends Genet.* 12, 283–286
4 Yu, C. E. *et al.* (1996) *Science* 272, 258–262
5 Ellis, N. A. *et al.* (1995) *Cell* 83, 655–666
6 Mushegian, A. R. *et al.* (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 5831–5836
7 Gangloff, S. *et al.* (1994) *Mol. Cell. Biol.* 14, 8391–8398
8 Watt, P. M. *et al.* (1995) *Cell* 81, 253–260
9 Hanada, K. *et al.* (1997) *Proc. Natl. Acad. Sci. U. S. A.* 94, 3860–3865
10 Mach, J. M. and Lehmann, R. (1997) *Genes Dev.* 11, 423–435
11 Oshima, J. *et al.* (1996) *Hum. Mol. Genet.* 5, 1909–1913
12 Yu, C. E. *et al.* (1997) *Am. J. Hum. Genet.* 60, 330–341
13 Ye, L. *et al.* (1997) *Am. J. Med. Genet.* 68, 494–498
14 Goto, M. *et al.* (1997) *Hum. Genet.* 99, 191–193
15 Lu, J. *et al.* (1996) *Nature* 383, 678–679
16 Gorbalenya, A. E. and Koonin, E. V. (1993) *Curr. Opin. Struct. Biol.* 3, 419–429
17 Brendel, V. *et al.* (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 2002–2006
18 Rost, B. *et al.* (1994) *CABIOS* 10, 53–60
19 Altschul, S. F. and Gish, W. (1996) *Methods Enzymol.* 266, 460–480
20 Schuler, G. D. *et al.* (1991) *Protein Struct. Funct. Genet.* 9, 180–190
21 Bork, P. and Gibson, T. J. (1996) *Methods Enzymol.* 266, 162–182
22 Tatusov, R. L. *et al.* (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 12091–12095

VLADIMIR MOROZOV

LION, Bioscience AG, 69120 Heidelberg, Germany.

ARCADY R. MUSHEGIAN

Sequana Therapeutics, Inc., 11099 North Torrey Pines Rd, La Jolla, CA 92037, USA.

EUGENE V. KOONIN

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

PEER BORK

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg and Max-Delbrück-Center, 13189 Berlin-Buch, Germany.

A eukaryotic XPB/ERCC3-like helicase in *Mycobacterium leprae*?

XPB/ERCC3 and XPD/ERCC2 are two helicases with distinct DNA-dependent ATPase activities. Mutations in these polypeptides result in DNA-repair deficiency, giving rise to genetic disorders in humans such as xeroderma pigmentosum, Cockayne's syndrome and trichothiodystrophy[1,2]. Unlike most helicases (see Ref. 3 for a review on helicases), these two proteins, as well as their counterparts in *Saccharomyces cerevisiae*, were only found in a multiprotein complex called TFIIH, which is involved in transcription and nucleotide excision repair (NER)[4,5].

Whereas various proteins related to XPD/ERCC2 have been described in prokaryotes[6], homologues of XPB/ERCC3 have only been identified in eukaryotes. XPB/ERCC3 and related proteins are organized in two main domains connected by a variable linker region (Fig. 1). A basic N-terminal module with a putative nuclear localization sequence (NLS-module) is followed by a highly conserved 20 kDa N-terminal domain. The second domain contains the helicase signature motifs followed by a 12 kDa module in which at least one known XPB/CS mutation is located[1,7].

Characteristic patterns from eukaryotic XPB/ERCC3 homologues were used to search the nucleotide database dynamically translated in all six open-reading frames (ORFs). Interestingly, the pattern PCGAGK(S/T) from the ATP-binding site (motif I, Fig. 2), present in all eukaryotic XPB/ERCC3-like proteins, was also found in *Mycobacterium leprae*. A careful analysis of the surrounding region revealed the presence of all the other helicase motifs within the same reading frame and of the N-terminal domain in an ORF that has been frame-shifted by −1 position.

Re-sequencing of the gene confirmed the existence of an extra guanine nucleotide. The corrected sequence codes for a putative 60 kDa protein, which when run through BLAST, detects the eukaryotic XPB/ERCC3 homologues with significant P-values ranging from 2.1×10^{-62} to 1.8×10^{-48}, while other helicases score around 10^{-3} (Ref. 8).

The ATG start codon is located a few nucleotides downstream from a canonical ribosome-binding site[9], suggesting that this putative protein is effectively translated in *M. leprae*. However, both the NLS- and the C-terminal 12 kDa modules are absent and the linker region is extremely short. Sequence conservation extends over the helicase catalytic core and the N-terminal domain, which is specific for this helicase subfamily (31% identity and 55% similarity in a 540-residue overlap).

 0968-0004/97/$17.00 PII: S0968-0004(97)01124-9