
Kohji Kusano,* Mark E. Berres and William R. Engels*

*Laboratory of Genetics and †Laboratory of Zoology, University of Wisconsin, Madison, Wisconsin 53706

Manuscript received June 17, 1998
Accepted for publication November 10, 1998

ABSTRACT

Several eukaryotic homologs of the Escherichia coli RecQ DNA helicase have been found. These include the human BLM gene, whose mutation results in Bloom syndrome, and the human WRN gene, whose mutation leads to Werner syndrome resembling premature aging. We cloned a Drosophila melanogaster homolog of the RECQ helicase family, \textit{Dmblm} (\textit{Drosophila melanogaster Bloom}), which encodes a putative 1487-amino-acid protein. Phylogenetic and dot plot analyses for the RECQ family, including 10 eukaryotic and 3 prokaryotic genes, indicate \textit{Dmblm} is most closely related to the \textit{Homo sapiens} BLM gene, suggesting functional similarity. Also, we found that \textit{Dmblm} cDNA partially rescued the sensitivity to methyl methanesulfonate of \textit{Saccharomyces cerevisiae} sgsl mutant, demonstrating the presence of a functional similarity between \textit{Dmblm} and SGSI. Our analyses identify four possible subfamilies in the RECQ family: (1) the BLM subgroup (\textit{H. sapiens} Bloom, \textit{D. melanogaster} \textit{Dmblm}, and \textit{Caenorhabditis elegans} \textit{T04A11.6}); (2) the yeast RECQ subgroup (\textit{S. cerevisiae} SGSI and \textit{Schizosaccharomyces pombe} rgh1/rad12); (3) the RECQL/ Q1 subgroup (\textit{H. sapiens} RECQ/ Q1 and \textit{C. elegans} K02F3.1); and (4) the WRN subgroup (\textit{H. sapiens} Werner and \textit{C. elegans} F18C5.2). This result may indicate that metazoans hold at least three RECQ genes, each of which may have a different function, and that multiple RECQ genes diverged with the generation of multicellular organisms. We propose that invertebrates such as nematodes and insects are useful as model systems of human genetic diseases.

DNA/RNA helicase protein families have been found with seven consensus motifs (Gorbalenya et al. 1989). These enzymes play important roles in cellular processes such as genome replication, recombination and repair, transcription, and mRNA translation. Members within each helicase family also share consensus sequences between motifs. The RECQ helicase family includes proteins with extensive amino acid sequence similarity to the \textit{Escherichia coli} helicase, RecQ.

The \textit{E. coli} RecQ gene encodes a DNA helicase (Umez u et al. 1990) involved in homologous recombination (Nakayama et al. 1984). The RecQ helicase appears to interact with RecJ exonuclease because recQ null mutations suppress recJ defects such as decreased crossing over (Kusano et al. 1994), increased sensitivity to DNA-damaging agents (Kusano et al. 1994; Lovett and Sutera 1995), and decreased conjugational recombination (Lovett and Sutera 1995) in a \textit{recB} \textit{C} \textit{sbca} background of \textit{E. coli}. A recQ null mutation increases illegitimate recombination in a wild-type background of \textit{E. coli} (Hanada et al. 1997). The \textit{Saccharomyces cerevisiae} SGSI gene encodes similar helicase motifs (Gangloff et al. 1994; Watt et al. 1995) and possesses a 3′-5′ DNA helicase activity (Lu et al. 1996; Bennett et al. 1998). An sgsl mutation suppresses slow growth in a top3 mutant (Gangloff et al. 1994; Lu et al. 1996) and causes missegregation of chromosomes in meiosis and mitosis (Watt et al. 1995). The SGSI protein physically interacts with topoisomerase III (Gangloff et al. 1994) and topoisomerase II (Watt et al. 1995). The mutations of another RECQ family member, \textit{Schizosaccharomyces pombe} rgh1/rad12, cause increased mitotic recombination with hydroxyurea treatment (Stewart et al. 1997). S. pombe rgh1/rad12 appears to function in a checkpoint dependent DNA damage response during S phase (Murray et al. 1997; Stewart et al. 1997; Davey et al. 1998). Three homologs of RECQ have been found in humans. One is the BLM gene whose defects cause immunodeficiencies and a highly increased rate of cancer, called Bloom syndrome (Ellis et al. 1995). Somatic cells from Bloom patients are susceptible to several DNA-damaging agents (Krepinsky et al. 1979; Ishizaki et al. 1981; Kurihara et al. 1987) and exhibit increased interchanges between homologous chromosomes (German 1993) and increased sister chromatid exchanges (Heartlein et al. 1987; Kurihara et al. 1987). The Bloom gene product has a 3′-5′ DNA helicase activity (Karow et al. 1997). The second homolog is the WRN gene whose defects result in clinical symptoms resembling premature aging, called Werner syndrome (Yu et al. 1996). Cells from Werner patients show various chromosome rearrangements (Salk et al. 1981; Scappaticci et al. 1982). More recently, WRN cells were shown to be hypersensitive to 4-nitroquinoline-1-oxide (Ogburn et al. 1997). The Werner protein has DNA helicase activi-
ity (Gray et al. 1997; Suzuki et al. 1997) with 3'-5' directionality (Shen et al. 1998). The third human RecQ homolog is RecQL/Q1, which has also been demonstrated to possess a 3'-5' DNA helicase activity (Puranam and Blackshear 1994; Seki et al. 1994). The symptoms of RecQL/Q1 deficiency have not been determined.

Phylogenetic analysis of a protein family often reveals the presence of subfamilies that are likely to have functional similarities. For example, an analysis of MutS-related proteins shows that there are several distinct subgroups (Fishel et al. 1993), each corresponding to a different function: mitochondrial mismatch repair, DNA sequence stability, and meiotic crossing over. The family of SNF2-related proteins provides another example. Phylogenetic analysis of this family suggests that there are eight distinct subgroups (Eisen et al. 1995), each corresponding to different functions, including transcription activation, transcription-coupled repair, recombinational repair, and removal of TATA binding protein from DNA. These two phylogenetic analyses indicate that the gene families functionally diverged with the generation of unicellular eukaryotes such as yeast.

In this work, we cloned a cDNA of the Drosophila melanogaster RecQ homolog, Dmblm (Drosophila melanogaster Bloom). It encodes a putative 1487 amino acid protein (Figure 1) and exhibits significant similarity to other RecQ family members. Our phylogenetic analysis of the RecQ family including Dmblm leads us to propose that the currently recognized RecQ family contains four subfamilies and that they functionally diverged with the generation of multicellular organisms. We suggest that Dmblm is included in the same group with Human BLM and Caenorhabditis elegans T04A11.6, each of which appears to be related to S. pombe rqn1/rad12 and S. cerevisiae SGS1. Also, we found that Dmblm partially rescued the sensitivity to methyl methanesulfonate (MMS) of a s. cerevisiae sgs1 mutant.

**Materials and Methods**

**Yeast strains and media:** Yeast strain MRQ966 (from T. Enomoto) is an sgs1::URA3 derivative of strain MR966 (MAT a ura3-52 leu2-3, 112 trp-289 his1-7). The following media were used: YEPD, which contains 1% yeast extract, 2% peptone, and 2% glucose, and SD, which is the selective medium for strains with a plasmid carrying a TRP1 marker, and contains 2% glucose, 0.67% Bacto-yeast nitrogen base, and supplements of uracil and all amino acids except tryptophan.

**Plasmids:** pYCP1305 (from T. Enomoto) contains the entire SGS1 gene of S. cerevisiae, a derivative of pRS314 that includes a centromere element, and an ARS and TRP1 marker (Sikorak and Hieter 1989). Two plasmids were used for the expression of Dmblm cDNA in yeast: (1) The plasmid pAS2-1 (CLONTECH, Palo Alto, CA) contains a GAL4-DNA-binding domain under an ADH1 promoter, 2-μm origin, and a TRP1 marker. (2) The plasmid pGBT9-bd (from A. Shimamoto), a derivative of pGBT9 (CLONTECH), lacks a GAL4-DNA-binding domain but contains an ADH1 promoter, 2-μm origin, and a TRP1 marker. The deletion of the domain was made by replacing the region from the HindIII site (nucleotide 409) to the EcoRI site (nucleotide 878) of pGBT9 (GenBank accession number U07646) with EcoRI oligonucleotides, GGAATTCC (New England Biolabs, Beverly, MA). The plasmid pBSDMBLM1 contains a PCR product that spans the entire region of Dmblm cDNA, nucleotides 1-4725 (GenBank accession number U92-536), at the EcoRV site of Bluescript SK (+). The next two plasmids contain the fragment that spans the region from the Dral site (nucleotide 111), at 15 bp before the putative start codon, to nucleotide 4716 on Dmblm cDNA. The fragment was prepared by partially cutting pBSDMBLM DNA with Dral restriction enzyme, ligating BamHI oligonucleotides CCGGA TCCGG (New England Biolabs) to the resultant Drl and completely cutting with BamHI restriction enzyme. The plasmids pYEDMBLM1 and pYEDMBLM2 possess this fragment at the BamHI site downstream of the GAL4-DNA-binding domain under the ADH1 promoter on pAS2-1, and at the BamHI site downstream of the ADH1 promoter on pGBT9-bd, respectively.

**PCR cloning:** Genomic DNA for PCR was prepared from the Canton-S strain of D. melanogaster. Two degenerate oligonucleotides, sense, ATGCCNACYGGHGGHGGHAA, and antisense, GGRCGGAARTCRTGDCCCCCA, were used for the PCR and cloning of a fragment of the D. melanogaster RecQ homolog. Two gene specific primers based on the nucleotide sequence of the obtained fragment were used for 5' RACE and 3' RACE: GGTGCAGCTAGAGAAATTTGACAGG and GCCGTGACCTTGACGAGGAGTG, respectively. A cDNA library was constructed using total RNA extracted by the RNAzol B method (Chomczynski and Sacchi 1987) from Schneider II cells for the degenerate PCR, 5' RACE and 3' RACE. A CLONTECH Marathon cDNA amplification kit was used for constructing the cDNA library and for 5' and 3' RACE. The 5' and 3' end primers used for amplifying an entire Dmblm cDNA were ATGTGTTTCCATTGGCATATTGCG and CTGACCATTTGGCATATTGCC, respectively. A mixture of Stratagene (La Jolla, CA) Taq DNA polymerase and Strategene Extender was used for the amplification of the 5' RACE, 3' RACES, and the entire cDNA. PCR products were always cloned into Bluescript SK (+).

**Nucleotide sequencing:** The M13-20 primer (GTAAAGACGCCGCAAT) or the T3 primer (AATTAACCTCTCATAAAGGG) was used for two types of sequencing reactions: the dideoxy method and the Dye deoxy termination method (Applied Biosystems International, Foster City, CA). The Promega (Madison, WI) Erase-A-base kit was used for preparing plasmid DNA with nested deletions in the insert region.

**Additional sequences:** Sequences other than D. melanogaster Dmblm (D.m.BLM), the abbreviations for their helicase domain, and their accession numbers are as follows: H. sapiens Bloom (H.s.BLM), U39817 (Ellis et al. 1995); H. sapiens RECOL/ Q1 (H.s.RECOL/ Q1), L36140 (Puranam and Blackshear 1994); H. sapiens Werner (H.s.WRN), L76937 (Yu et al. 1996); C. elegans E03A3.2 (C.e.E03A3.2), Z38112; C. elegans K02F3.1 (C.e.K02F3.1), U00052 (Wilson et al. 1994); C. elegans F18C5.2 (C.e.F18C5.2), U29097 (Wilson et al. 1994); C. elegans T04A11.6 (C.e.T04A11.6), Z83123; S. pombe rqn1/rad12 (S.p.rqn1/rad12), Z54354 (Murray et al. 1997; Stewart et al. 1997); S. cerevisiae SGS1 (S.c.SGS1), L07870; E. coli RecQ (E.coliRecQ), M30198 (Irino et al. 1986); Haemophilus influenzae RecQ (H.i.RecQ), U32756 (Fleischmann et al. 1995); Bacillus subtilis RecQ (B.s.RecQ), L47648 (Roels et al. 1992; Sorokin et al. 1993).

**Intron boundaries:** We noted that the putative amino acid sequence in U00052 (C.e.K02F3.1) did not contain the region around helicase motif VI and that the putative amino acid sequence in U29097 (C.e.F18C5.2) did not contain the region...
around helicase motifs V and VI. We checked each amino acid sequence of the two published nucleotide sequences in three different reading frames and found that motif V and VI of C.e.F18C5.2 were present behind putative introns in U29097 and that motif VI of C.e.K02F3.1 was present in another reading frame in U00052. Therefore, we adjusted the exon-intron boundaries in C.e.K02F3.1 and C.e.F18C5.2 to include motifs V and VI. Thus, the putative introns start at GT and stop at AG. The nucleotide sequence of C.e.F18C5.2 from U29097 contains the region, nucleotide numbers 10192–10497, which was regarded as an intron, and the region, 10498–10580, which was regarded as an exon. The region 10192–10580 was modified as follows: 10192–10239 (intron), 10240–10427 (exon), 10428–10475 (intron), and 10476–10580 (exon). According to these modifications, amino acid positions 302–398 in C.e.F18C5.2, with reference to the coordinates of Figure 2, were used instead of QKEEEVENLT, which was part of the amino acid sequence in U29097. Also, the sequence of C.e.K02F3.1, derived from U00052, contains the regions 34759–35328 (exon), 35329–35532 (intron), 35533–35701 (exon), and 35702 through the end of this clone as an intron. These regions were modified as follows: 34759–35221 (exon), 35222–35532 (intron), 35533–35779 (exon), and 35780 through the end (intron). According to these modifications, amino acid positions 339 through the last site of C.e.K02F3.1 in Figure 2 were used instead of the last 33 amino acid residues in U00052.

**Sequence alignment:** An alignment of the amino acid sequence data in the helicase domains of the above proteins was constructed using clustering algorithms incorporated into the program CLUSTALW version 1.73 (Higgins et al. 1992). Parameters were set as follows: gap penalty is 10.0; gap extension penalty is 0.05; and protein weight matrix is BLOSUM30 (Henikoff and Henikoff 1994). Further visual editing of the algorithmically aligned amino acid sequences was not attempted. We used the protein alignment as a guide to exclude highly variable regions that exhibited many insertion and deletion events from the final alignment to eliminate ambiguous residue positioning (Figure 2). Corresponding codon positions in the nucleotide sequence data were deleted and the resulting in-frame nucleotide alignment was used for all subsequent analyses.

**Phylogenetic analysis:** The skewness of a random sample of all tree-length distributions based on the g statistic (Sokal and Rohlf 1981) as described by Hillis (1991), Hillis and Huelsenbeck (1992), and Huelsenbeck (1991) was estimated using PAUP version 3.1.1 (Swofford 1993). Initial tests for significant variance in nucleotide and codon substitution rates based on maximum likelihood estimates were performed according to Muse and Gaut (1994) and Muse and Weir (1992). PAUP version 3.1.1 (Swofford 1993) was used for an initial heuristic tree search applying the TBR option for branch swapping. The F84 substitution model (Thorne et al. 1992) incorporated into PAML version 1.3 (Yang 1996b) was used to estimate substitution parameters and the maximum likelihood of all tree topologies.

**Other analyses:** Dot plot analysis for protein was performed in the 13 RECQ family proteins. The algorithmically aligned amino acid sequences was not attempted. We used the protein alignment as a guide to exclude highly variable regions that exhibited many insertion and deletion events from the final alignment to eliminate ambiguous residue positioning (Figure 2). Corresponding codon positions in the nucleotide sequence data were deleted and the resulting in-frame nucleotide alignment was used for all subsequent analyses.

**Cloning a Drosophila homolog of RECQ:** To obtain a fragment of a Drosophila RECQ homolog, we carried out PCR on genomic DNA of D. melanogaster using degenerate primers. The primers (see materials and methods) correspond to the amino acid residues MPTGGGK, located within the conserved helicase motif I, and WGHDWRP, located between motifs II and III (Figure 1), which is conserved within the RECQ helicase family (Figure 2). Resulting amplicons were cloned and sequenced. One clone was 426 bp in length and included a putative 63-bp intron whose 5'- and 3'-splice sites were ATgtgatt and tttctttttacagAT, respectively. These are in general agreement with the splice donor and acceptor consensus sequences of Drosophila short introns, AGgt(a/g)agt(a/t) and ttttt(c/t)(c/t)(c/t)(c/t)tnca(g/A/G)T (Mount et al. 1992). The predicted amino acid sequence from the 363-bp region without the intron sequence indicated good homology with the corresponding region of other RECQ genes. Also, PCR on cDNA isolated from Schneider II cells using the same primer pair produced an ~360-bp fragment. This was judged as one related to the clone from genomic DNA based on digestion patterns by HindIII and Mbol restriction enzymes.

Using the RACE cloning method with the two internal primers (materials and methods), designed on the basis of the nucleotide sequence of the 363-bp region, a 2.6-kb 5' RACE clone and a 2.4-kb 3' RACE clone (the largest fragments) were cloned.

**Cytological localization:** Each of the DNA fragments hybridized within the cytological map region 86F1-8. Also, PCR of a P1 clone carrying the region 86F1-4 (BDGP, Hartl et al. 1994) using the above internal primer pair produced the expected PCR fragment size.

**Predicted 1487-amino-acid protein is included in RECQ helicase family:** Sequencing of the two overlapping RACE fragments revealed an open reading frame of 1487 amino acids in length. The putative 5'- and 3'-untranslated regions contain two and six stop codons in the same frame, respectively. The predicted protein sequence is shown in Figure 1. The central region contains seven motifs commonly found in DNA/RNA helicases, including a putative ATP-binding site (motif I) and a DExH box (motif II; Gorbalenya et al. 1989). An alignment of this protein to other members of the RECQ family shows significant similarity among them, especially in the residues of the seven helicase motifs (Figure 2). Therefore, we conclude that this Drosophila gene is a member of the RECQ family.

**Sequence alignment:** The schematic diagram in Figure 3 represents the position of each helicase domain in the 13 RECQ family proteins. The algorithmically aligned protein sequences of the 13 helicase domains formed a consensus length of 419 sites of which 70 sites could not be unambiguously aligned (Figure 2). Thorne et al. have shown that the true alignment between a set of sequences tends to exhibit fewer gaps than an algorithmically generated alignment (Thorne et al. 1991, 1992; Thorne and Kishino 1992). As a result, assignment of any residue to a particular position in an area of introduced gaps would be arbitrary because of
a lack of knowledge concerning prior insertion and deletion events. Therefore, areas with multiple floating gaps were omitted following a rule that usable sites must be within anchored regions of two unambiguously aligned sites. Using these criteria, the final amino acid alignment consisted of 349 aligned residues, of which 61 were constant and 288 were variable. Deletion of the corresponding ambiguously aligned codon positions in the nucleotide sequence data produced an alignment containing a total of 1047 nucleotide sites, of which 164 were constant and 883 were variable.

**Skewness:** On the basis of a statistical measure of skewness called the $g_1$ statistic, Hillis (1991) devised a method for testing whether or not a particular data set contains phylogenetic information. One prediction of the model is that for data sets containing phylogenetic information, the set of all possible topologies (or a subsample thereof) arranged in order of increasing length will have a definite skew to the right. This skewing is attributed to character correlation among sequences with a common phylogenetic history (Hillis 1991). The tree-length distribution analysis of $1 \times 10^6$ random to-

Figure 1.—Predicted protein sequence of the D. melanogaster homolog of RECQ gene, Dmblm (Accession number U92536). Helicase motifs I–VI are underlined by dashes, and their locations are based on previously described helicase domain alignment (Gorbalenya et al. 1989). The repeat near the N-terminal end is underlined. The positions of amino acid substitution in the repeat are double-underlined.
polologies from a possible $8.64 \times 10^6$ unrooted topologies of our nucleotide sequence data was highly skewed ($g_i = -0.5598$, $P < 0.01$).

**Evolutionary rates:** In most molecular sequences, substitution rates across sites are not equally distributed. Zuker and Corbin (1971) have shown that rate differences may be best fitted to a gamma distribution. The shape parameter, $\alpha$, of the gamma distribution, is now used as a measure of substitution heterogeneity across sites (Yang 1996a). The importance of considering substitution heterogeneity across sites in phylogenetic reconstructions is established (Nei 1987, 1991; Yang 1994b). To test the nucleotide sequence data for significant substitution heterogeneity under a particular evolutionary model, we employed a likelihood ratio test (Huelsenbeck and Rannala 1997). Our null hypothesis was that substitution heterogeneity existed across each codon position. Our alternative hypothesis was that substitution heterogeneity existed across each codon position. The likelihood ratio statistic is two times the difference in log likelihood of the model and is approximately $\chi^2$ distributed with one degree of freedom. Substitution heterogeneity was accommodated under the F84 model (Thorne et al. 1992) by applying the discrete gamma model (F84-di; Yang 1994b), which approximates the gamma distribution by allowing different rate categories for each nucleotide position, each with equal probability of occurrence. Our analysis was performed with four rate categories, which provides a good approximation of the gamma distribution (Yang 1994b). It is known that substitution heterogeneity is overestimated if a grossly incorrect tree topology (e.g., a completely multifurcating star tree) is used (Sullivan et al. 1996). Therefore, we used the single most parsimonious tree derived from the nucleotide sequence data (MP1, Figure 4A). When the F84-di model was applied to our data, substitution rates were gamma distributed across the three codon positions with $\alpha$-values of 0.73, 0.61, and 0.65, corresponding to the first, second, and third codon positions, respectively. The likelihood ratio test indicates clearly that accounting for substitution heterogeneity across codon sites improves significantly the likelihood of the maximum parsimony topology (Table 1).

Significant differences in substitution rates across lineages are known to impair tree estimation methods (Felsenstein 1978; Huelsenbeck and Hillis 1993). We initially applied relative rate tests (Wu and Li 1985; Muse and Weir 1992; Muse and Gaut 1994) to determine whether both nucleotides (transition and transversion substitutions) and codons (synonymous and non-synonymous substitutions) evolve in a clocklike fashion among the helicase sequences examined. This test consisted of a pairwise comparison of all 13 helicase domains using the RECO sequence from E. coli as an outgroup to obtain an overall estimation of rate variances. Of 78 individual tests, 32 nucleotide and 24 codon sequence comparisons exhibited a significant deviation ($P < 0.05$) from the expectation that 2 test sequences should show an equal amount of divergence compared to the reference sequence. To test for this effect under the F84 substitution model, we again used a likelihood ratio test. Our null hypothesis was that rate homogeneity existed across all lineages. The alternative was that rate heterogeneity existed across one or more lineages. In this case, the likelihood ratio statistic is two times the difference in log likelihood of the model and is approximately $\chi^2$ distributed with two degrees of freedom. This test, when applied to our data, easily rejects the presence of a molecular clock with high significance (Table 1).

**Phylogenetic analysis:** In maximum likelihood methods, probabilities for possible topologies are computed using all nucleotide sites, each of which is considered separately. The evaluation of the likelihood that a particular topology exhibits is conditional on the given evolutionary model and how well the model can explain the observed data. The topology with the highest likelihood is thus considered the best representation of relationship. Our estimation of the best topology was guided by the subjective value of accounting for parameters that have a significant bearing on tree reconstruction efficiency, such as rate heterogeneity and transition and transversion biases. We used the maximum parsimony method to construct an initial topology for the purposes of relevant parameter estimation. It is known that maximum parsimony often fails when significant rate variation exists across lineages (Felsenstein 1978; Nei 1991). However, maximum parsimony has the advantage of requiring no explicit evolutionary model. Under these assumptions, maximum parsimony recovers a single most parsimonious tree (MP1, Table 1; Figure 4A), with length 4157 and a consistency index of 0.475. MP1 was then used as the starting topology for a maximum likelihood tree search assuming the F84-di model of sequence evolution, which allows for rate heterogeneity across sites, unequal base composition, and different transition and transversion rates (Yang 1994a). This tree search resulted in a best maximum likelihood tree of $-14311.24$ with transition/transversion rate ratio parameters of 0.58, 0.22, and 10.52 and $\alpha$-parameters of 0.71, 0.61, and 0.63 for each codon position respectively (Figure 4B).

**Similarities outside the helicase domain:** A different kind of analysis was needed to determine whether the gene groupings indicated above are also apparent in sequences flanking the helicase domain. In these areas, the similarities were too weak to use the phylogenetic methods employed for the helicase domain. Therefore, we examined dot plots of all 78 pairwise comparisons in the C- and N-terminal regions and ranked each pair according to total length of similarity. The highest-ranking comparisons are shown in Figure 5. For the C-terminal region, 5 of the 6 within-group comparisons had clear similarity (Figure 5, B–E and I) as opposed to only 4 of the 72 between-group comparisons (Figure 5, A and F–H). In the N-terminal regions, only 1 of the 78 comparisons had notable similarity (Figure 5J). This
comparison was also a within-group pair. We conclude that sequence similarities extend beyond the helicase domain for all 6 pairwise comparisons within the proposed groups and that there is relatively little such similarity between groups.

**Domain structure comparison of the RECQ family:**
The 13 RECQ family proteins were arranged in Figure 3 on the basis of the subgrouping postulated above. Recently, the possibility was suggested that RECQ family members can be separated into distinct functional subgroups on the basis of the presence or absence of extensive N-terminal domains (Watt and Hickson 1996). Our analyses provide evidence against this possibility. For example, the N-terminal segment of H. sapiens RECQL/Q1 is short as are the bacterial RECQs. However, in the helicase domain, H. sapiens RECQL/Q1 is more closely related to C. elegans K02F3.1, which has a long N-terminal segment (Figure 3C). Furthermore, the short N-terminal segment of H. sapiens RECQL/Q1 shares limited similarity with the region adjacent to the helicase domain of C. elegans K02F3.1 (Figures 3C and 5J), suggesting that these short segments are important for a common function between H. sapiens RECQL/Q1 and C. elegans K02F3.1. The N-terminal domain of H. sapiens WNR is as long as the N-terminal domains of Dmbm, H. sapiens BLM, C. elegans T04A11.6, S. cerevisiae SGS1, S. pombe rqh1/rad12, and C. elegans K02F3.1; all of which are members of other subgroups. Overall, it is clear from Figures 3 and 4 that the length of the N-terminal regions is not well correlated with the relatedness of the helicase domain sequences.

**Presence of direct sequence repeats in the RECQ family:**
For example, the N-terminal segment of H. sapiens RECQL/Q1 is short as are the bacterial RECQs. However, in the helicase domain, H. sapiens RECQL/Q1 is more closely related to C. elegans K02F3.1, which has a long N-terminal segment (Figure 3C). Furthermore, the short N-terminal segment of H. sapiens RECQL/Q1 shares limited similarity with the region adjacent to the helicase domain of C. elegans K02F3.1 (Figures 3C and 5J), suggesting that these short segments are important for a common function between H. sapiens RECQL/Q1 and C. elegans K02F3.1. The N-terminal domain of H. sapiens WNR is as long as the N-terminal domains of Dmbm, H. sapiens BLM, C. elegans T04A11.6, S. cerevisiae SGS1, S. pombe rqh1/rad12, and C. elegans K02F3.1; all of which are members of other subgroups. Overall, it is clear from Figures 3 and 4 that the length of the N-terminal regions is not well correlated with the relatedness of the helicase domain sequences.
previously (Yu et al. 1996; Figure 3E). There is also a repeat of 20/24 amino acids near the N-terminal end of Dmblm (Figures 1 and 3A), a repeat of 7/9 amino acids in the N-terminal domain of the C. elegans T04-A11.6 gene (Figure 3A), and a repeat of 9/10 amino acids near the N-terminal end of the C. elegans F15C5.2 gene (Figure 3E). These repeat structures might be correlated with the apparent evolutionary flexibility of the length of the N-terminal sequences, as discussed below.

**Rescue of the S. cerevisiae sgs1 mutant by the Dmblm cDNA:** The structural homology between the BLM and yeast RECQ subgroups (Figures 4 and 5) suggests the possibility that Dmblm functions in a manner similar to the SGS1 gene of S. cerevisiae. First we tested sensitivity of the sgs1::URA3 mutant to MMS (Figure 6A). The D. melanogaster Dmblm cDNA placed downstream of the ADH1 promoter in the yeast 2-µm plasmids conferred a 12-fold increase in the survival fraction of the sgs1 mutant cells against 0.02% MMS (Figure 6B), although it did not reach that of the wild-type strain carrying the vector.

**DISCUSSION**

**D. melanogaster Bloom:** We have identified a new D. melanogaster gene and include it in the RECQ helicase family. On the basis of our phylogenetic analysis, dot plot analysis, and comparative domain structure analysis, this new gene, designated Dmblm, is most similar to the BLM gene of H. sapiens, strongly suggesting functional similarity.

**The subgroups in the RECQ family:** Maximum likelihood estimations, assuming the F84-d1 model (Yang 1994b), were applied to the helicase domain present in all 13 RECQs. Dot plot analyses were executed for all domains of all 13 RECQs. On the basis of these results, we suggest that there are four possible subgroups in the RECQ family: (i) the BLM subgroup (H. sapiens Bloom, D. melanogaster Dmblm, and C. elegans TO4A11.6); (ii) the RECQL/Q1 subgroup (H. sapiens RECQL/Q1 and C. elegans K02F3.1); and (iii) the WRN subgroup (H. sapiens WRN and C. elegans K00F3.1); and (iv) the WRN
<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Clade</th>
<th>N-terminal</th>
<th>Helicase</th>
<th>C-terminal</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. BLM</td>
<td>D.m.BLM</td>
<td>733 1110</td>
<td>1487</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H.s.BLM</td>
<td>653 1041</td>
<td>1417</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C.e.T04A11.6</td>
<td>364 372 373 381</td>
<td>486 863</td>
<td>1231</td>
</tr>
<tr>
<td>B. Yeast</td>
<td>S.p.qh1/rad12</td>
<td>515 894</td>
<td>1328</td>
<td></td>
</tr>
<tr>
<td>RECQ</td>
<td>S.c.SGS1</td>
<td>674 1052</td>
<td>1447</td>
<td></td>
</tr>
<tr>
<td>C. RECQL/Q1</td>
<td>H.s.RECQL/Q1</td>
<td>87 458</td>
<td>659</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C.e.K02F3.1</td>
<td>482 853</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. C.e.E03A3.2</td>
<td>C.e.E03A3.2</td>
<td>173 579</td>
<td>809</td>
<td></td>
</tr>
<tr>
<td>E. WRN</td>
<td>H.s.WRN</td>
<td>424 450 451</td>
<td>1432</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C.e.F18C5.2</td>
<td>544 913</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 NEELPETEFE 26</td>
<td>596</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F. Bacterial</td>
<td>B.s.RECQ</td>
<td>12 375</td>
<td>496</td>
<td></td>
</tr>
<tr>
<td>RECQ</td>
<td>H.i.RECQ</td>
<td>24 389</td>
<td>619</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E.c.RecQ</td>
<td>23 387</td>
<td>610</td>
<td></td>
</tr>
</tbody>
</table>
Evolution of RECQ Helicases

Figure 4.—Topology of the 13 helicase domains depicted with branch lengths proportional to the expected substitutions. The four groups were inferred from the helicase domains as well as from the flanking regions (see Figure 5). (A) MP1. Maximum parsimony tree. (B) Tree 2. Maximum likelihood topology of the best tree found by global branch swapping of the initial maximum parsimony tree (MP1) using the F84-ac substitution model.

Discussion:

subgroup (H. sapiens Werner and C. elegans F18C5.2). Our designation of subgroups provides reasonable insights for functional analysis of RECQ family genes.

N-terminal length variation in the RECQ family: The presence of four direct repeats in the N-terminal domains allows us to suggest a correlation with the apparent evolutionary variability of the length of the N-terminal sequences. Gene elongation during evolution seems to have depended largely on domain duplications (Barker et al. 1978). The multiple duplication of small units would partially contribute to elongation of N-terminal ends of an ancestor gene that might be similar to bacterial RECQs. The small duplications shown in Figure 3 in Dmbm, C. elegans T04A11.6, C. elegans F18C5.2, and H. sapiens WRN would likely have occurred recently in evolutionary history. The WRN homologues of two vertebrates, Mus musculus (Imamura et al. 1997) and Xenopus laevis (Yan et al. 1998), have only one such sequence. The length difference of their N-terminal domains might depend on the length of the original repeat unit and the frequency of duplication events. However, the first half of the N-terminal segment of the Werner protein includes a nuclease domain homologous to bacterial RNaseD and the 3'-5' exonuclease domain of DNA polymerase I (Mian 1997; Mushugian et al. 1997). This indicates that other mechanisms such as exon shuffling contributed to the N-terminal length variation.

Features and perspectives in the functional subgroups: The relationships among members of the RECQ family do not follow the distinction between vertebrates and invertebrates, as seen in the three subgroups that contain both human and invertebrate RECQ genes (Figure 4). These features may indicate that gene duplication of ancestral RECQ sequences occurred several times. Over evolutionary time each RECQ gene would diverge and acquire a new function stabilized by natural selection or would be lost, as proposed recently by Fryxell (1996).

Mutations of S. cerevisiae SGS1 increased mitotic recombination in ribosomal DNA (rDNA) repeats (Gangloff et al. 1994) and between direct repeated genes (Watt et al. 1996). We found that an sgs1 mutant is hypersensitive to MMS. Mutations of S. pombe rgh1/rad12 caused the hypersensitivity to UV, hydroxyurea or γ-ray (Murray et al. 1997; Steward et al. 1997; Davey et al. 1998) and increased mitotic recombination with hydroxyurea treatment (Steward et al. 1997). S. pombe rgh1/rad12 appear to function on a checkpoint protein dependent DNA damage response during S phase (Murray et al. 1997; Steward et al. 1997; Davey et al. 1998). Yeast RECQ proteins might inhibit inappropriate homologous recombination associated with postreplication repair against DNA damage. Also, the SGS1 gene was shown to be involved in cellular aging (Sinclair et al. 1997), which was associated with accumulation of extrachromosomal rDNA circles (Sinclair and Guarente 1997).

Figure 3.—(A-F) Schematic representation of proteins of RECQ family members. Each entire protein is exhibited by a box with its total amino acid residue number on the right end. Shaded regions indicate the helicase domains whose sequences are shown in Figure 2 with the initial and final residue numbers. The C-terminal region of C. elegans K02F3.1 is not yet sequenced completely and is shown by a dashed line. Each protein is aligned along the position of the initial residue of the helicase domain to compare the length of N-terminal domains between each protein. Each of the short dotted lines in (C) shows the region sharing sequence similarity to the N-terminal domains of C. elegans K02F3.1 and H. sapiens RECQL/Q1 (see Figure 5J). A pair of triangles indicates a duplication with the indicated amino acid sequence. Substitutions within the duplications are underlined.
TABLE 1

<table>
<thead>
<tr>
<th>Tree</th>
<th>Model</th>
<th>Clock</th>
<th>ln L</th>
<th>Model</th>
<th>Clock</th>
<th>ln L</th>
<th>2(ln L1 – ln L0)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP1</td>
<td>F84</td>
<td>No</td>
<td>-14763.39</td>
<td>F84-dl′</td>
<td>No</td>
<td>-14327.03</td>
<td>436.36</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MP1</td>
<td>F84-dl′</td>
<td>Yes</td>
<td>-14668.76</td>
<td>F84-dl′</td>
<td>No</td>
<td>-14327.03</td>
<td>341.73</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

* The maximum parsimony tree.

Somatic cells from patients with Bloom syndrome show increased sister chromatid exchanges (Heartlein et al. 1987; Kurihara et al. 1987) and increased interchanges between homologous chromosomes (German 1993). BLM cells are shown to be hypersensitive to ethyl methane-sulfonate (Krepinsky et al. 1979). Dmblm cDNA partially rescued the MMS sensitivity of an sgs1 mutant. H. sapiens BLM has a similar rescue activity (Yamagata et al. 1998). These results provide evidence of the functional similarity between the BLM subgroup and the yeast RECQ subgroup. The BLM subfamily members might inhibit extra homologous recombination at/near the site of DNA damage. In D. melanogaster, P-element-induced double strand gaps are repaired by gene conversion accompanied very infrequently with crossing over (Nassif and Engels 1993). Interchromosomal re-

---

Figure 5.—Dot plot analysis between proteins of RECQ family members. The entire amino acid sequences were compared by a PAM250 scoring matrix, a window size of 30 residues, and a minimum score of 35% similarity. Shading indicates the helicase domains aligned in Figure 2. The total of length of the dots (excluding helicase domain) that lie on the diagonal line are shown. Dots with a closed triangle include the sequences similar to the nuclear translocation signal of H. sapiens BLM (Kaneko et al. 1997).
Various chromosome rearrangements (Sal k et al. 1981; Scappaticci et al. 1982), but do not show increased sister chromatid exchange (Gebhart et al. 1988), in contrast to BLM cells. WRN cells exhibited a slower rate of DNA replication (Fujiiwara et al. 1977). More recently WRN cells were shown to be hypersensitive to 4-nitroquinoline-1-oxide (Ogburn et al. 1997). The Werner protein is likely to work with replication protein A (Shen et al. 1998). Yan et al. (1998) suggested the role of Werner protein and replication protein A in DNA replication. The WRN subgroup members might be involved in resumption of DNA replication following postreplication repair. The human WRN gene and the \textit{C. elegans} \textit{F18C5.2} share significant similarity in their helicase and C-terminal domains. This indicates that \textit{C. elegans} \textit{F18C5.2} may be a useful model of Werner syndrome.

In the RECQL/ Q1 subgroup, there is no genetical data. We suggest that a genetical approach using \textit{C. elegans} K02F3.1 may be useful to address the role of RECQL/ Q1 in humans. A biochemical finding suggests that the human RECQL/ Q1 protein possesses DNA helicase activity and translocates in a 3' to 5' direction on the DNA to which it binds (Seki et al. 1994). The manner in which the short N-terminal segment shared by the members of this subgroup affects their DNA helicase activities may be a key for the clarification of their specific function.

In E. coli, homologous recombination is initiated in several ways (Clark and Low 1988; Lloyd and Low 1996). The E. coli RecQ helicase appears to initiate a type of homologous recombination with RecJ exonuclease (Kusano et al. 1994). In a recB \textsuperscript{-} sboA \textsuperscript{-} background, recQ \textsuperscript{+} strains are more resistant to UV-irradiation than recQ \textsuperscript{-} strains (Kusano et al. 1994; Lovett and Sutera 1995). In a recB \textsuperscript{-} sboB C \textsuperscript{-} background, however, a recQ \textsuperscript{+} strain is more sensitive to UV-irradiation than a recQ \textsuperscript{-} strain (data not shown). These two results indicate that E. coli RecQ contains an additional function(s) as well as one epistatic to RecJ function. Also, the E. coli RecQ gene appears to control nonhomologous recombination (Hanada et al. 1997). The functional relationships between bacterial RECQs and eukaryotic RECQs are important in order to understand how RECQ family members have diverged.

\textbf{Meaning of suppression by mutations in the RECQ family members:} Consider two genes A and B. A \textsuperscript{-} B \textsuperscript{-} combination in mitotic cells of mammals is very low compared with that in yeast (Shulman et al. 1995). These reports indicate that mitotic crossing over is strictly inhibited in metazoans. \textit{H. sapiens} BLM and Dmblm might be involved in preventing the loss of heterozygosity induced by DNA damage.

Cells from patients with Werner syndrome exhibit combination in mitotic cells of mammals is very low compared with that in yeast (Shulman et al. 1995). These reports indicate that mitotic crossing over is strictly inhibited in metazoans. \textit{H. sapiens} BLM and Dmblm might be involved in preventing the loss of heterozygosity induced by DNA damage.

Cells from patients with Werner syndrome exhibit combination in mitotic cells of mammals is very low compared with that in yeast (Shulman et al. 1995). These reports indicate that mitotic crossing over is strictly inhibited in metazoans. \textit{H. sapiens} BLM and Dmblm might be involved in preventing the loss of heterozygosity induced by DNA damage.
be explained by the hypothesis that A\textsuperscript{−} activates a pathway substituted for the AB pathway so as to alleviate any severe phenotype caused by B\textsuperscript{−}. In this way, the sgsl\textsuperscript{−} mutation might enable type I topoisomerase to substitute for Top3 in S. cerevisiae cells (Lu et al. 1994). The recQ\textsuperscript{−} mutation might open other homologous recombination pathways for E. coli cells (Kusano et al. 1994). The RECQ family members might therefore operate on DNA helicase is a suppressor of illegitimate recombination in E. coli recQ

1038 K. Kusano, M. E. Berres and W. R. Engels

LITERATURE CITED


Lovell, D. T., and V. A. F. Nair, 1995 Suppression of RecQ exonuclease mutants of Escherichia coli by alterations in DNA Helicases II (uvrD) and IV (hde). Genetics 140: 27±45.


Mount, S. M., C. Burks, G. Hertz, G. D. Stormo, O. White et al., 1038 K. Kusano, M. E. Berres and W. R. Engels


Yang, Z., 1996b PAML: phylogenetic analysis by maximum likelihood, version 1.3. Department of Integrative Biology, University of California, Berkeley, CA.