DOI: 10.1534/genetics.104.029744

Note

Genes Encoding Vitamin-K Epoxide Reductase Are Present in Drosophila and Trypanosomatid Protists

Hugh M. Robertson[1]

Department of Entomology, University of Illinois, Urbana, Illinois 61801

Manuscript received April 6, 2004
Accepted for publication June 7, 2004

ABSTRACT

Vitamin-K epoxide reductase is encoded by the VKORC1 gene in mammals and other vertebrates, which also have a paralog, VKORC1L1. Single homologs are present in basal deuterostome and insect genomes, including Drosophila, and three trypanosomatid protists. VKOR is therefore an ancient gene/protein that can be studied in the Drosophila model system.

VITAMIN K is an important nutrient for animals. It is a cofactor in carboxylation of glutamine residues of proteins by vitamin-K-dependent gamma-glutamyl carboxylase. While this reaction is probably important for many protein modifications, it is particularly important in vertebrates for activation of several blood-clotting proteins and hence for blood coagulation. The resultant vitamin-K 2,3-epoxide is recycled to the reduced active form by vitamin-K epoxide reductase (VKOR). Inhibition of VKOR by the drug warfarin can be beneficial at low concentrations to humans with blood clots but causes lethal bleeding in rats at high concentrations (see Sadler 2004). The gene encoding one subunit of this enzyme complex, VKORC1, was recently identified in mammals by Rost *et al.* (2004) and Li *et al.* (2004). Both articles note that there is a homologous protein encoded by a gene in the African malaria mosquito *Anopheles gambiae* genome, but apparently not in *Drosophila melanogaster*. Rost *et al.* (2004) noted that this was unusual because *D. melanogaster* encodes a homolog of vitamin-K-dependent gamma-glutamyl carboxylase (Li *et al.* 2000), an enzyme activity also known from molluscs, which they proposed might therefore predate VKOR activity (despite the existence of the Anopheles insect VKORC1 homolog). This observation is also unusual in that these two, admittedly highly divergent, flies encode mostly the same repertoire of conserved genes (Zdobnov *et al.* 2002). I therefore examined the Drosophila genome sequences, using TBLASTN searches with the vertebrate and Anopheles proteins, and the VKORC1 homolog is readily apparent, but not annotated in Release 3.1 (Misra *et al.* 2003). It is now partly annotated in the Third Party Annotation (TPA) database at NCBI through the work of Hild *et al.* (2003), who undertook an independent annotation of the *D. melanogaster* genome, although the annotation (HDC06808) fuses this gene and a downstream gene. I also undertook a phylogenetic analysis of this gene/protein in animals, revealing its presence in all available insect genomes, as well as in the sea urchin *Strongylocentrotus purpuratus* and urochordate Ciona genomes. Finally, I root these phylogenetic analyses using three trypanosomatid homologs that are the only confident available outgroup.

TBLASTN searches of the Release 3.1 assembly of the *D. melanogaster* genome revealed a compact three-coding-exon gene of 570 bp encoding a VKORC1 homolog at cytological position 53C15 on chromosome arm 2R in the 4-kb region between Actin-related protein 53D (Arp53D or CG5409) and CG15920, which I propose to name *vitamin-K epoxide reductase* (*vkor*). This gene structure with two introns in the coding region (phases 2 and 1) is present throughout the animals, the only modifications being loss of the second intron in *A. gambiae* and loss of both introns in *Ciona savignyi*. The gene annotation in Hild *et al.* (2003; HDC06808) splices from within the third exon to a downstream exon just before CG15920—this downstream exon is likely to represent yet another unannotated gene, because there is an expressed sequence tag (EST) from it. The homolog was also identified in the draft genome sequence of *D. pseudoobscura*, although raw trace information was required to extend the assembled contig sequence that contains the gene. The homolog in the honey bee *Apis mellifera* draft genome sequence v1.1 was readily identified in

[1]*Address for correspondence:* Department of Entomology, University of Illinois, 505 S. Goodwin Ave., Urbana, IL 61801.
E-mail: hughrobe@uiuc.edu

Contig5445, as was the silk moth *Bombyx mori* homolog in three short contigs of the Japanese assembly (Mita *et al.* 2004). The homolog in the sea urchin *S. purpuratus* was assembled manually from the raw sequence traces (as is true for much of this genome, two quite divergent haplotypes are present, so the manually assembled sequence might represent a chimera of the two alleles), while the urochordate sea squirt *C. intestinalis* and *C. savignyi* genes were identified in the assembled genome sequences available at NCBI (Dehal *et al.* 2002), where again each gene has two haplotypes.

Among the vertebrates, the available human, rat, and mouse orthologs of VKORC1 were obtained, along with the chicken, Xenopus, Takifugu, and zebrafish orthologs from genomic and/or cDNA sequences. These are all relatively straightforward except for the VKORC1 ortholog in the chicken. This protein is rather divergent and the gene is not present in the draft genome assembly. Instead it was built from a combination of two ESTs and a single available whole-genome shotgun trace containing most of the middle exon. As Rost *et al.* (2004) and Li *et al.* (2004) note, the human genome contains two independently generated retropseudogenes of VKORC1 with 85 and 83% encoded amino acid identity on chromosomes X and 1, respectively. The chromosome 1 copy is truncated at the 5′ end. The other vertebrate genomes do not contain VKORC1 pseudogenes.

Rost *et al.* (2004) also noted that there is a paralog of VKORC1 called VKORC1-like1 (VKORC1L1) in these vertebrate genomes and that this gene/protein is actually more conserved in sequence across the vertebrates than is VKORC1. The human VKORC1L1 gene on chromosome 7 is somewhat unusual in that there is a copy of the first exon encoding five different amino acids ±835 kbp upstream from the first exon that Rost *et al.* (2004) identified and also an identical duplication of the second exon ±48 kbp downstream of the end of the gene. There is also a retropseudogene fragment 2.1 Mbp upstream of VKORC1L1 with 78% encoded amino acid identity. While the VKORC1L1 orthologs in the same set of other vertebrates were easily identified in genomic and/or cDNA sequences, the mouse (Waterston *et al.* 2002) and rat (Gibbs *et al.* 2004) genomes contain several pseudogenes. In mouse there is a very young retropseudogene on chromosome 15 that could encode an identical protein (the coding sequence differs by just one synonymous base change). The mouse genome contains three additional older retropseudogenes whose coding capacity is interrupted by frameshifts and encoded stop codons on chromosomes 1, 7, and 17 encoding 89, 79, and 66% amino acid identity to VKORC1L1, respectively. There is also a regular pseudogene on the X chromosome that retains the first intron (the second intron appears to have been lost in a deletion that also removed five exonic bases—alternatively this is a retropseudogene derived from an incompletely spliced mRNA) and encodes 77% amino acid identity to VKORC1L1. In the rat genome there are three retropseudogenes on chromosomes X, 8, and 17, each encoding 88% amino acid identity to VKORC1L1. These mouse and rat retropseudogenes all appear to have formed independently in these two rodent lineages as they do not cluster with each other in phylogenetic analyses, nor are they microsyntenous with each other, unlike the VKORC1 and VKORC1L1 genes.

No homolog could be identified in the two available Caenorhabditis nematode genomes, and in agreement with Li *et al.* (2000) they also do not appear to have a homolog of vitamin-K-dependent carboxylase, so this entire pathway of protein modification appears to have been lost from at least these two nematodes. Searches of all other publicly available eukaryotic genome sequences at NCBI yielded three homologs encoded by single-coding-exon genes in the three available kinetoplastid genomes of *Trypanosoma cruzi*, *T. brucei*, and *Leishmania major* (among the 13 available protist genome sequences at NCBI; see Ghedin *et al.* 2004). These matches are clearly significant; for example, the matches with human VKORC1 are nearly full length and range from $E = 1e\text{-}12$ to 5e-17. PSI-BLASTP searches initiated with these or the animal proteins yielded possible distantly related proteins in the plant *Arabidopsis thaliana* genome and in many bacterial genomes that share the four completely conserved cysteines of VKORC1, but few other residues and no fungal homologs were identified. The three kinetoplastid proteins were therefore used to root phylogenetic analysis of the animal proteins.

Phylogenetic analysis of these 25 short proteins of 150–180 amino acids cannot be expected to reconstruct known animal phylogeny accurately, and indeed the rapidly diverging VKORC1 lineage in the vertebrates caused considerable difficulty (in particular the divergent chicken VKORC1 sequence branched basally within the animals), as did the rapidly evolving Drosophila proteins. Nevertheless the key features of the evolution of this protein can be deduced from the partially constrained tree in Figure 1. The vertebrate VKORC1 lineage was constrained to be monophyletic and follow accepted phylogenetic relationships (as shown naturally by the more conservative VKORC1L1 lineage); the sea urchin and urochordate lineages were swapped in position in the tree to reflect accepted phylogeny of these basal deuterostomes; and the Drosophila and Anopheles dipteran proteins, along with the other insect proteins, were constrained to be monophyletic. None of the constrained relationships were supported by bootstrapping; indeed, the only strong support is for clear intralineage relationships like those of the Drosophila, Ciona, fish, mammalian, and kinetoplastid proteins, as well as all branches in the vertebrate VKORC1L1 lineage. There is weak support for the paralogous relationship of the VKORC1 and VKORC1L1 genes in vertebrates.

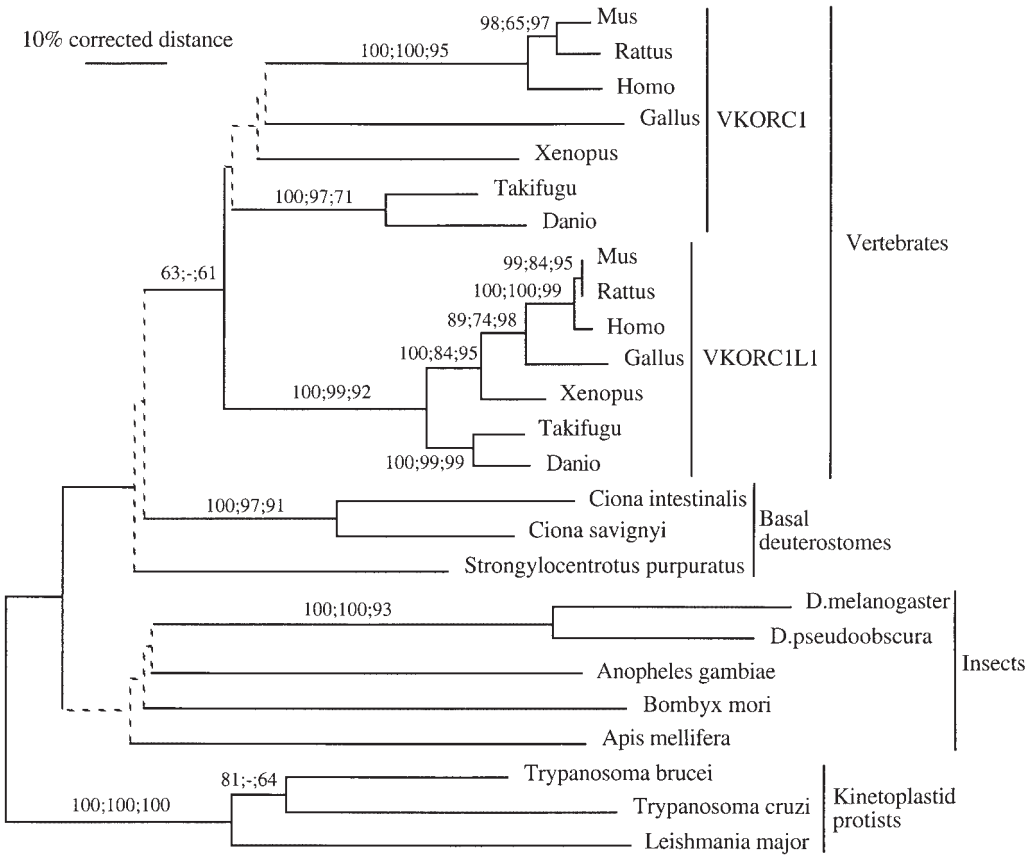

FIGURE 1.—Phylogenetic tree relating the VKOR homologs in animals, rooted with the three kinetoplastid proteins. The branches that were constrained are indicated by dashed lines (see text for details). The scale bar is 10% divergence in corrected distances. Numbers on branches indicate the percentage of bootstrap support from 1000 replications of uncorrected distance and parsimony analysis, followed by percentage of representation in 1000 quartet puzzling steps. Phylogenetic analysis of these 25 proteins, after alignment with CLUSTALX using default settings (CHENNA *et al.* 2003; the alignment is available as supplemental data at http://www.genetics.org/supplemental/ or from the author), was conducted with corrected distances and minimum evolution methods in PAUP* v4.0b10 (SWOFFORD 1998). Distances were corrected using the maximum-likelihood model in TREE-PUZZLE v5.0 (SCHMIDT *et al.* 2002), using default settings and the BLOSUM62 amino acid exchange matrix.

From the tree in Figure 1 it is evident that, as expected, the protist proteins are highly divergent from the animal proteins. Nevertheless their homology seems clear, including conservation of the four absolutely conserved cysteines in the animal proteins, as well as another ±30 residues highly conserved in the animal proteins. While it is remotely possible that this gene was acquired by a trypanosomatid ancestor from an ancestral animal lineage and then diverged rapidly, the tree suggests that it is indeed an ancient protein in this protist lineage, in which case it must have been lost from most of the other protists whose genome sequences are available, as well as from the major fungal, plant, and nematode lineages for which genome sequences are available. Searches for the carboxylase gene in protist, plant, and fungal genome sequences were all negative, so this VKORC1 homolog in kinetoplastids might be serving a quite different function in them. Convincing homologs of the carboxylase are encoded by all the animal genomes in Figure 1, as well as by several bacterial genomes, including the spirochaete *Leptospira interrogans* (REN *et al.* 2003), the sphingobacterium *Cytophaga hutchinsonii* (NCBI accession ZP_00119016.1), and many unknown environmental sequences (VENTER *et al.* 2004), so it might still be an older enzyme, although not necessarily using vitamin K as a cofactor.

Within the animals it is clear that both protostome and deuterostome lineages contained a single gene, and, as is the case for so many other vertebrate genes, the split into two genes appears to have occurred between the urochordate and vertebrate lineages, that is, in a primitive vertebrate (DEHAL *et al.* 2002). As noted by ROST *et al.* (2004), the paralogous VKORC1L1 protein is actually the more highly conserved of the two in amino acid sequence (for example, the mouse and rat proteins are identical, whereas VKORC1 differs by 8% in these two rodents), although its function is unknown. If this duplication follows the neofunctionalization model of gene duplication (OHNO 1970), then the rapid divergence of VKORC1 in vertebrates might suggest that its function related to vitamin K recycling might be the derived function, in which case the unknown VKORC1L1 function might better reflect the role of this protein in the other animals and trypanosomatids. Alternatively, if this duplication follows the subfunctionalization model of LYNCH and FORCE (2000), then both proteins might still be involved in vitamin K recycling; however, for some reason VKORC1 has been free to diverge more rapidly in

vertebrates than has VKORC1L1. VKORC1 appears to be twice as highly and somewhat more widely expressed than VKORC1L1 (~400 *vs.* 200 ESTs in the human Unigene database at NCBI). What is clear is that VKOR proteins are much older than Rost *et al.* (2004) concluded, perhaps as old as carboxylation of glutamates by the carboxylation enzyme.

Recognition of the Drosophila *vkor* gene means that it is possible to study this gene/protein in this model system, although failure to annotate it in the first three releases of the Drosophila annotation means that it is not represented in most current microarray or proteomic projects so no information is available about it from those sources. It is not even on the Heidelberg microarray (Hild *et al.* 2003) because they designed their amplicon HFA05627 for this fused gene annotation to the downstream exon that is not part of *vkor*. The gene is not highly expressed in *D. melanogaster* with no ESTs among ±374,000 in dbEST at NCBI, although there are five ESTs among ±135,000 for *A. gambiae*, one among ±116,000 for *B. mori*, and none among ±23,000 for *A. mellifera*. The nearest available transposon insert is P{EPgy2}EY00984 ±4 kb downstream in the next annotated gene, CG15920, so it is possible that a local hopping insertion might be obtained by remobilization of this insert. The *vkor* genes in *D. melanogaster* and *D. pseudoobscura*, despite their rapid divergence shown in the phylogenetic tree (only 57% amino acid identity with N- and C-terminal and internal length differences), are under strong purifying selection as revealed by a nonsynonymous substitution frequency (K_a) of 0.31 ± 0.04 compared to a synonymous substitution frequency (K_s) of 2.14 ± 0.74, yielding a K_a/K_s ratio of 0.15. Thus these *vkor* genes and their encoded VKOR enzymes are important for Drosophila.

I thank the Baylor College of Medicine Human Genome Sequence Center (GSC) for access to the unpublished genome sequences and raw traces for *Drosophila pseudoobscura*, *Apis mellifera*, and *Strongylocentrotus purpuratus*; the Massachusetts Institute of Technology GSC for access to the unpublished *Ciona savignyi* genome assembly; the Washington University GSC for help with finding the chicken VKORC1 gene in unassembled raw traces; and the various other genome projects, especially the protist genome projects, for access to their unpublished assemblies via the National Center for Biotechnology Information (NCBI). Where possible, new gene constructs and protein translations have been deposited in the TPA database at NCBI (accession nos. BK005184–BK005187 and BK005471). Alignments are available as supplementary information on the Genetics website at http://www.genetics.org/supplemental/.

Note added in proof: The *Drosophila melanogaster vkor* gene is now represented on the second generation microarray chip from Affymetrix. Preliminary results of expression assays using these chips with whole adult male mRNA indicates that *vkor* is expressed at low levels (K. Hughes, personal communication), consistent with the absence of ESTs from *D. melanogaster*. It is unclear whether this indicates generally low expression in all cells, or whether *vkor* might be expressed at higher levels in certain cells/tissues, and other stages of development remain unexamined.

Goodstadt, L., and C. P. Ponting (2004, Vitamin K epoxide reductase: homology, active site and catalytic mechanism. Trends Biochem. Sci. **29:** 289–292) have also noted that *D. melanogaster* has a *vkor* gene and examined the distant bacterial homologs in detail, including proposing an active site and catalytic mechanism for the enzyme.

LITERATURE CITED

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson *et al.*, 2003 Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. **31:** 3497–3500.

Dehal, P., Y. Satou, R. K. Campbell, J. Chapman, B. Degnan *et al.*, 2002 The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science **298:** 2157–2167.

Ghedin, E., F. Bringaud, J. Peterson, P. Myler, M. Berriman *et al.*, 2004 Gene synteny and evolution of genome architecture in trypanosomatids. Mol. Biochem. Parasitol. **134:** 183–191.

Gibbs, R. A., G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren *et al.*, 2004 Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428:** 493–521.

Hild, M., B. Beckmann, S. A. Haas, B. Koch, V. Solovyev *et al.*, 2003 An integrated gene annotation and transcriptional profiling approach towards the full gene content of the Drosophila genome. Genome Biol. **5:** R3.

Li, T., C. T. Yang, D. Jin and D. W. Stafford, 2000 Identification of a Drosophila vitamin K-dependent gamma-glutamyl carboxylase. J. Biol. Chem. **275:** 18291–18296.

Li, T., C. Y. Chang, D. Y. Jin, P. J. Lin, A. Khvorova *et al.*, 2004 Identification of the gene for vitamin K epoxide reductase. Nature **427:** 541–544.

Lynch, M., and A. Force, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459–473.

Misra, S., M. A. Crosby, C. J. Mungall, B. B. Matthews, K. S. Campbell *et al.*, 2003 Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol. **3:** RESEARCH0083.

Mita, K., M. Kasahara, S. Sasaki, Y. Nagayasu, T. Yamada *et al.*, 2004 The genome sequence of silkworm, *Bombyx mori*. DNA Res. **11:** 27–35.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg, Germany.

Ren, S. X., G. Fu, X. G. Jiang, R. Zeng, Y. G. Miao *et al.*, 2003 Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. Nature **422:** 888–893.

Rost, S., A. Fregin, V. Ivaskevicius, E. Conzelmann, K. Hortnagel *et al.*, 2004 Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature **427:** 537–541.

Sadler, J. E., 2004 Medicine: K is for koagulation. Nature **427:** 493–494.

Schmidt, H. A., K. Strimmer, M. Vingron and A. von Haeseler, 2002 TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18:** 502–504.

Swofford, D. L., 1998 *PAUP* 4: Phylogenetic Analysis Using Parsimony and Other Methods*. Sinauer Associates, Sunderland, MA.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch *et al.*, 2004 Environmental genome shotgun sequencing of the Sargasso sea. Science **304:** 66–74.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520–562.

Zdobnov, E. M., C. von Mering, I. Letunic, D. Torrents, M. Suyama *et al.*, 2002 Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science **298:** 149–159.

Communicating editor: S. Henikoff