Title

Bioinformatic Characterization of Genes and Proteins Involved in Blood Clotting in Lampreys

Journal

Journal of Molecular Evolution, 81(3-4)

ISSN

0022-2844

Author

Doolittle, Russell F

Publication Date

2015-10-01

DOI

10.1007/s00239-015-9701-0

Peer reviewed

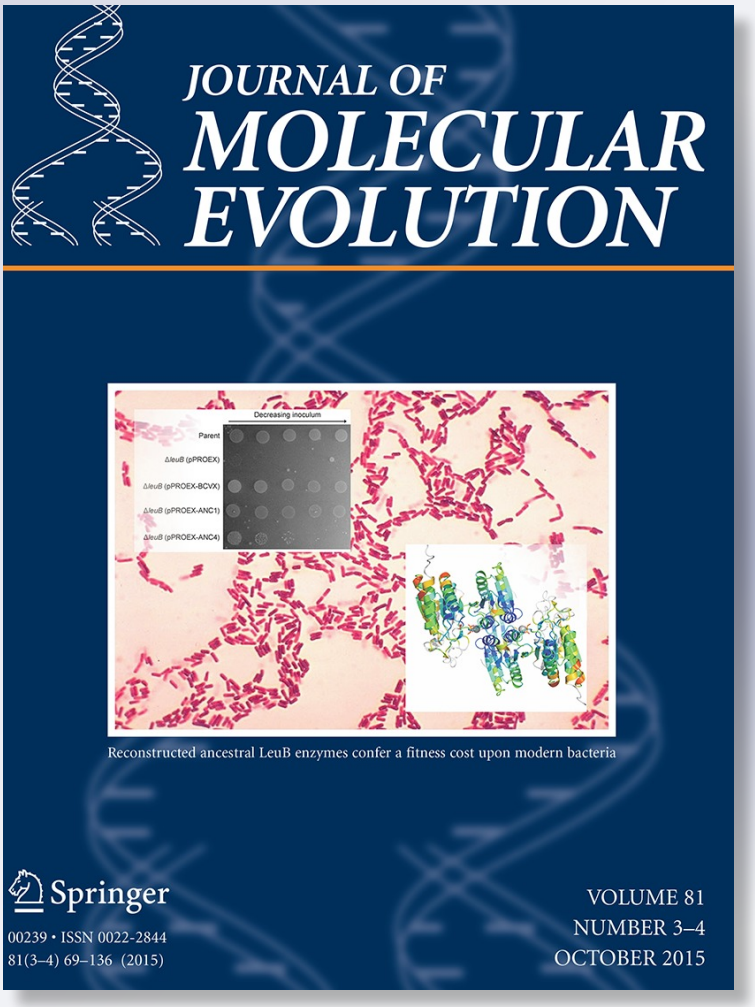
Bioinformatic Characterization of Genes and Proteins Involved in Blood Clotting in Lampreys

Russell F. Doolittle

Journal of Molecular Evolution

ISSN 0022-2844
Volume 81
Combined 3-4

J Mol Evol (2015) 81:121-130
DOI 10.1007/s00239-015-9701-0



Springer

J Mol Evol (2015) 81:121–130
DOI 10.1007/s00239-015-9701-0

ORIGINAL ARTICLE

Bioinformatic Characterization of Genes and Proteins Involved in Blood Clotting in Lampreys

Russell F. Doolittle[1]

Received: 2 September 2015 / Accepted: 24 September 2015 / Published online: 5 October 2015

Abstract Lampreys and hagfish are the earliest diverging of extant vertebrates and are obvious targets for investigating the origins of complex biochemical systems found in mammals. Currently, the simplest approach for such inquiries is to search for the presence of relevant genes in whole genome sequence (WGS) assemblies. Unhappily, in the past a high-quality complete genome sequence has not been available for either lampreys or hagfish, precluding the possibility of proving gene absence. Recently, improved but still incomplete genome assemblies for two species of lamprey have been posted, and, taken together with an extensive collection of short sequences in the NCBI trace archive, they have made it possible to make reliable counts for specific gene families. Particularly, a multi-source tactic has been used to study the lamprey blood clotting system with regard to the presence and absence of genes known to occur in higher vertebrates. As was suggested in earlier studies, lampreys lack genes for coagulation factors VIII and IX, both of which are critical for the "intrinsic" clotting system and responsible for hemophilia in humans. On the other hand, they have three each of genes for factors VII and X, participants in the "extrinsic" clotting system. The strategy of using raw trace sequence "reads" together with partial WGS assemblies for lampreys can be used in studies on the early evolution of other biochemical systems in vertebrates.

Keywords Lampreys · Blood clotting · Trace databases · Whole genome sequence assembly problems

Introduction

Lampreys and hagfish are the only two extant genera of jawless fish (Agnatha). As such, they are central to our understanding of the early evolution of vertebrates, offering the possibility of finding simpler versions of complex physiological systems observed in mammals. For example, it was long ago determined that lampreys have single-chain hemoglobins and not the tetrameric kind found in most vertebrates (Wald and Riggs 1951). Similarly, early biochemical studies suggested that lampreys have a simpler blood clotting scheme than do higher vertebrates and one that is limited to the so-called "extrinsic" clotting system (Doolittle and Surgenor 1962). The same underlying question has been asked about numerous other biochemical systems: namely, can the early evolutionary stages of complex systems found in higher vertebrates be better understood by examining the situation in lampreys or hagfish?

Over the years, we and others have been attempting to trace the appearance of the various genes involved in vertebrate blood coagulation. Currently, the most direct

In the article describing the whole genome assembly for the Japanese lamprey (Mehta et al. 2013), the organism is referred to as *Lethenteron japonicum*. However, the downloaded version of the database describes each entry as *Lethenteron camtschaticum*. In the present article, the first-named version is used. Moreover, because of the frequent back and forth comparing the sea lamprey and Japanese lamprey, occasionally the shortened terms PM (for *Petromyzon marinus*) and LJ (for *Lethenteron japonicum*) are used.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-015-9701-0) contains supplementary material, which is available to authorized users.

✉ Russell F. Doolittle
rdoolittle@ucsd.edu

[1] Departments of Chemistry & Biochemistry and Molecular Biology, University of California, San Diego, La Jolla, CA 92093-0314, USA

method of attempting such evolutionary reconstructions is the examination of whole genome DNA sequences (WGS) of representative organisms. In this regard, high-quality genomes are available for several jawed fish, frogs, lizard, birds, and numerous mammals, and it has been possible to detail many of the events that have made mammalian blood clotting so complex (Davidson et al. 2003a, b; Jiang and Doolittle 2003; Ponczec et al. 2008). Unhappily, whole genome sequences for lamprey and hagfish have been late in coming and are still imperfect, especially for cases where a consideration of gene absence is critical.

Several years ago, we attempted to circumvent the problem of not having a lamprey WGS by exhaustively studying individual DNA sequences stored in the NCBI Trace Archive, which, in the case of the sea lamprey (*Petromyzon marinus*), is an especially rich trove of raw sequence data. As it happened, the concurrent appearance of a draft assembly based on the same Trace data made it possible to make judgments about the presence and likely absence of the various clotting factors found in mammals, and in the end we cautiously proposed that lampreys have a reduced set of such genes (Doolittle et al. 2008).

In particular, the data suggested that lampreys lack genes for factors VIII and IX, two genes that are essential for the "intrinsic" clotting system in higher vertebrates and defects in which are responsible for hemophilia in humans. It was already known that genes for these two proteins are present in jawed fishes like the pufferfish (Davidson et al. 2003a; Jiang and Doolittle 2003) and zebrafish (Hanumanthaiah et al. 2002), and the proposal was that in the interval between the divergence of the jawless and jawed fish more or less simultaneous duplications of genes for factors X and V gave rise to those for factors IX and VIII. Crucially, factors VIII and IX interact with each other in the same way that factors V and X do, the first-named of each pair (VIII and V) serving as a large molecular weight co-factor for the second-named vitamin K-dependent protease (factors IX and X).

The study also revealed that lampreys have two genes for factor X and three for factor VII, vitamin K-dependent factors at the heart of the "extrinsic" clotting system. The main shortcoming of the work was that it was not possible to link together all the various trace sequences for every one of the clotting genes, as would be required for proving absence.

In the interval since that report, two different assemblies for lamprey genomes have appeared. The first (Smith et al. 2013) was a new assembly for *P. marinus* but one that was based on sequences that also appear in the Trace Archive and were used in the 2007 draft assembly. Although the new assembly significantly increased the lengths of scaffolds relative to the contigs generated in the earlier draft, it was actually slightly less complete as measured by base pairs provided (Table 1). The other recently reported assembly was for the closely related Japanese lamprey (*Lethenteron japonicum*) (Mehta et al. 2013). It has only a slightly higher coverage, but the much longer scaffolds provide vital overlaps for joining segments found in the other databases (Table 1). The lamprey genome has been estimated to contain between 1.6 and 2.2 billion bp (Gregory 2005), suggesting that the average coverage in both of the new assemblies may be as low as 60 % (Table 1). This may be misleading, however, because gene-rich regions of the genome were likely easier to sequence than gene-poor, high-repeat regions.

In spite of their limitations, in the work reported here, these assemblies have been used in combination with the abundant short sequences stored in the Trace Archive collection to reconstruct fully most of the genes and proteins involved in the lamprey coagulation pathway. All the evidence supports the earlier conclusions about the absences of factors VIII and IX and the presence of additional factors VII and X in lampreys. The biggest challenge was that most of the gene products of interest are the result of gene duplications that took place about the same time as the appearance of vertebrates, exacerbating the problem of distinguishing orthologs from highly similar paralogs. The cases of coagulation factors V and VIII are even more problematic because of their being members of the ferroxidase family of proteins, which includes ceruloplasmin and hephaestin proteins, themselves composed of three major domains that are the result of (tandem) duplications.

Methods

Various lamprey databases were downloaded on to an in-house computer. These included current versions of the Trace Archive for *P. marinus*, which, in addition to the vast numbers of random shotgun reads, also contains numerous sequences from non-genomic sources, including extensive cDNA and EST entries. The Trace Archive also contains many mate pairs that can be used to establish neighborliness. As noted in our earlier report, we had also downloaded a 2007 draft assembly based on the same Trace data from ftp://genome.wustl.edu/pub/petromyzon_marinus. More recently, a 2012 (and subsequently updated to 2015 versions) top level assembly was downloaded from http://www.ensembl.org, as well as the 2013 release of the *L. japonicum* from http://lampreygenome.imcb.a-star.edu.sg/.

BLAST software (Altschul et al. 1997) was downloaded from the NCBI website; tblastn was used for searching amino acid sequences against raw DNA data. Phylogenetic reconstructions were made by a distance-matrix method (Feng and Doolittle 1996) as well as with a parsimony procedure (Doolittle and Feng 1990). Trees were drawn on

Table 1 Sources of sequence data used for re-constructing lamprey clotting genes

Database[a]	DNA seqs[b]	Base pairs (bp)[c]	bps/seq[d]	N50[e]	UNSEQ[f] (%)
2007 PM trace archive	18,787,613	14,640,144,063	780	–	–
2014 PM trace archive	19,213,524	14,973,428,366	779	–	–
2008 PM draft assembly[g]	108,246	1,027.242,766	9490	–	19
2012 PM assembly[h]	25,005	885,534,757	36,076	174,000	27
2013 LJ scaffolds[i]	6011	887,131,771	147,855	1,051,965	17
2013 LJ contigs[i]	80,114	143,530,947	1794	9240	<1

[a] Sources are for *P. marinus* (PM) or *L. japonicum* (LJ)

[b] Number of fragments (scaffolds, contigs, or raw sequences)

[c] Total number DNA base pairs

[d] Average length of sequences in data base

[e] The N50 is the size of for which one-half of scaffolds or contigs are larger

[f] UNSEQ = percent unknown bp (NNNN) in database

[g] Obtained from ftp://genome.wustl.edu/pub/petromyzon_marinus

[h] Smith et al. (2013) (updates to assembly included up to version 7.0 80)

[i] Mehta et al. (2013)

the PHYLODENDRON website http://iubio.bio.indiana.edu/treeapp/treeprint-form.html.

In one case, cDNA prepared from sea lamprey liver was subjected to PCR to obtain an overlap between two key segments of the lamprey factor V gene. Primers were based on sequences found in the Trace database. I am grateful to Russell Darst for conducting these experiments in the laboratory of Dr. Lorraine Pillus.

Searching Strategy

As described in our earlier report (Doolittle et al. 2008), the exploration began by searching relatively short (exon-sized) sequences from human clotting factors against all entries in the 2007 Trace Archive. Identical (or near identical) sequences were extracted from the individual reads and clustered in "hit-groups." As an example, the preliminary search for sequences corresponding to clotting factors V and VIII yielded 257 hits, which after accounting for redundancy reduced to about 50 "hit-groups."

The challenge from that point on was to arrange the hit-groups in the proper order in which they occur in their parent proteins, either with the aid of other data or by homology with factors from other vertebrates. In the cases of some of the clotting factors, longer non-genomic sequences (cDNA and/or ESTs) are available, including some especially useful entries added to the Trace Archive after our first report. The 2008 draft lamprey assembly cited above had been used to link about half of the remaining hit-groups to contigs, and the posting of the 2012 assembly with its longer scaffolds gave rise to a few more. In a few cases, connections in contigs and scaffolds revealed segments that had been missed in the initial searching. At this point, all the collected matches for the sea lamprey were re-searched against the 2015 Trace Archive, as well as against the 2013 assembly for the Japanese lamprey. The latter has very long scaffolds and provided additional linkages between contigs and scaffolds from the sea lamprey assembly. Exon–intron boundaries were determined on the basis of the GT-AG rule for the start- and endpoints of introns.

Results

The Ferroxidase Gene Family

The initial evidence for there being only four genes in the hephaestin-ceruloplasmin-factor V–VIII family was based on there never being more than four different sequences for any set of aligned segments, and only one of those being more similar to factors V and VIII (Doolittle et al. 2008). The weakness of the proposal was that fewer than half of the segments (based on reads from the trace archive) could be linked together by cDNA sequences or the draft assembly. Two important developments have taken place since that first effort. First, some long non-genomic DNA sequences have been added to the Trace Archive that link together many of the trace reads for ferroxidase family sequences and demonstrate unequivocally that none of these have intervening B domains and cannot be either factors V or VIII. Second, an assembly for the Japanese lamprey with its very long scaffolds showed that these same three constructs are in complete agreement and located on separate scaffolds (Fig. 1). The corresponding exons for these three genes are very similar in length but the introns separating them vary greatly (Figs. 1, 2a–c). The same is true for a comparison of the lamprey

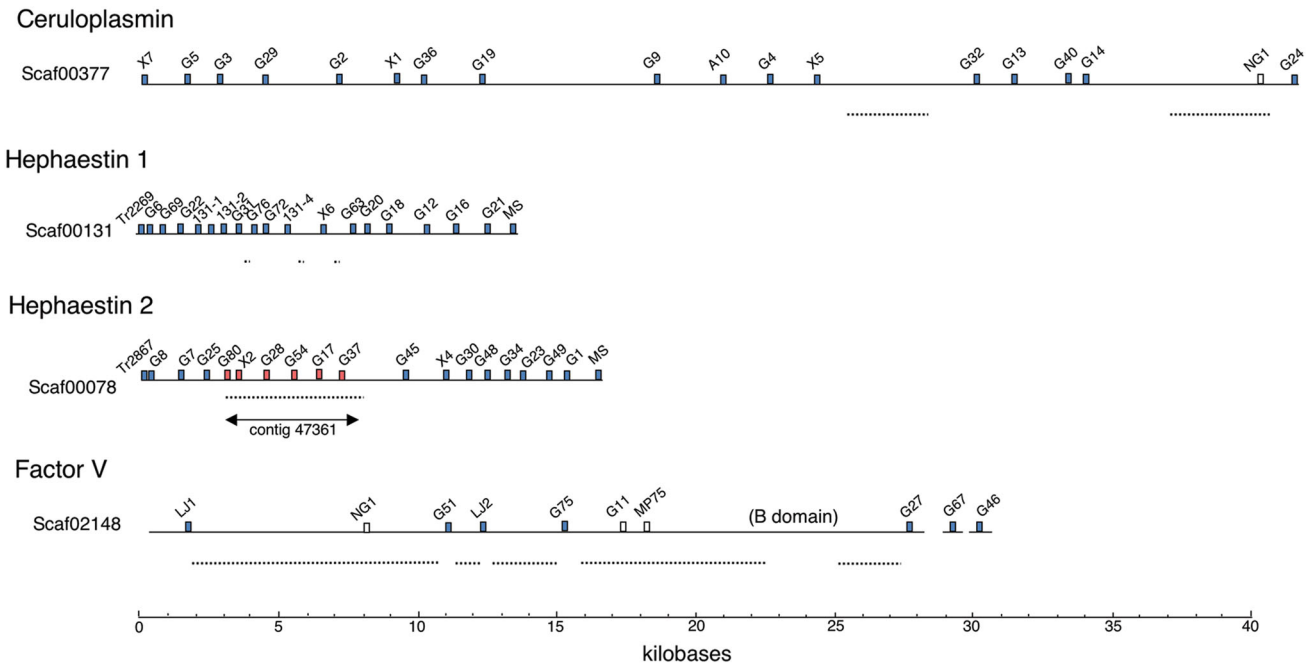


Fig. 1 Gene structures of four ferroxidase family members (hephaestin-1, hephaestin-2, ceruloplasmin, and factor V) in *L. japonicum*. Locations of "hit-groups" from *P. marinus* are shown as small *boxes*: *blue boxes* mean sequences available for both species; *orange boxes* occur in unsequenced regions of scaffolds (*dotted lines*) and are positioned by independent data from *P. marinus* assemblies (see Fig. 2). *Open boxes* (*no color*) indicate approximate locations; MS denotes region encoding membrane-spanner (Color figure online)

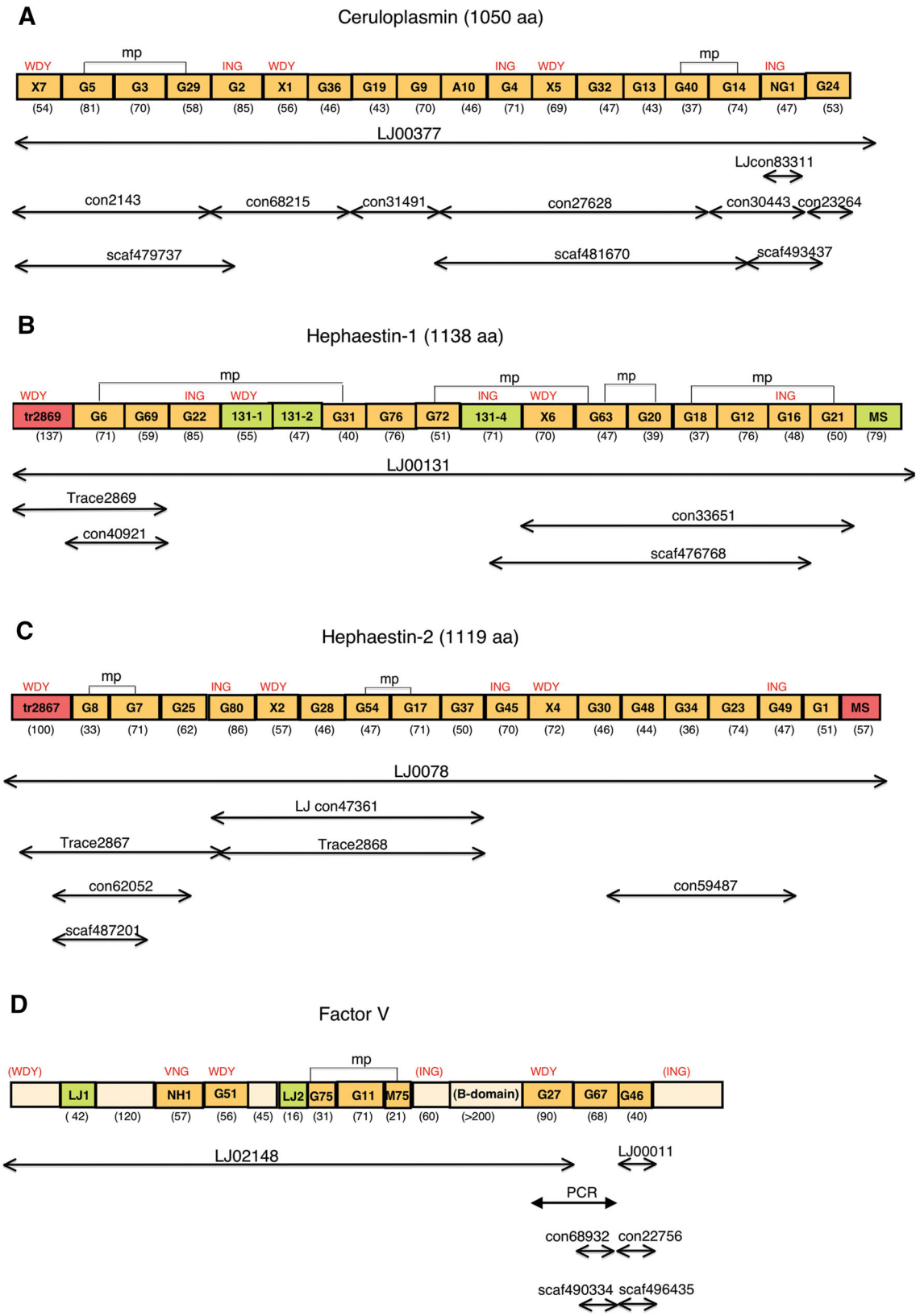
ceruloplasmin gene with its human counterpart (Daimon et al. 1995), the exon lengths corresponding closely but the introns exhibiting significant differences. Two of the constructs also have membrane-spanning segments at their carboxyl ends and must be hephaestins. The third does not have such a segment, as would be expected for ceruloplasmin.

The fourth member of the ferroxidase family has only been partially reconstructed, but all of its parts are more similar to factors V or VIII from other vertebrates than they are to hephaestin or ceruloplasmin. Unhappily, the gene occurs mainly on a scaffold (LJ scaf02148) with long interrupting regions of unsequenced DNA (Fig. 1). Nonetheless, the gene was pieced together further by a variety of means. For one, PCR of sea lamprey (PM) cDNA was used to show that hit-groups G27 and G67 are adjacent to each other. Second, various other hit-groups were found to match segments in the same region of the scaffold (LJ scaf02148). The few remaining unplaced hit-groups (from the original 50) were positioned in the most propitious gaps. Additionally, a previously unidentified mate pair for hit-group G45 was found and logically placed nearby (MP45 in Fig. 1). On a negative note, although there are numerous discoidin sequences in the various lamprey databases, it was not possible to link any specific pair to the putative factor V gene.

As implied above, it was possible to assign all 50 of the original hit-groups for the ferroxidase family by the use of

Fig. 2 Reconstructed blocks of "hit-groups" for four ferroxidase proteins from lamprey: **a** ceruloplasmin-1, **b** hephaestin-1, **c** hephaestin-2, and **d** factor V. *Double-arrows* denote overlap sequences from various sequence data sources that support the arrangements. LJ and LJcon denote scaffolds and contigs from *L. japonicum* assembly; the prefix con denotes contigs from the 2008 draft assembly of *P. marinus*, and the prefix scaf is for scaffolds from the 2012 *P. marinus* assembly. *Orange boxes* are "hit groups" from Trace Archive; *green boxes* are based only on sequences in *L. japonicum* scaffolds. *Pink boxes* are based on non-genomic sequences in trace archive, and *blank boxes* in factor V denote missing regions based on homology expectations. The motifs "WDY" and "ING" (or "VNG") found in all A domains are shown atop the group in which they occur. *MP* mate pair (Color figure online)

overlaps from the various databases (Fig. 2a–d). The matches for the putative factor V gene were scattered and occurred in parts of all three A domains. A concatenated sequence of five coding sections amounting to 481 amino acids was prepared and aligned with the corresponding segments from 12 other ferroxidase proteins, including the three other lamprey proteins (hephaestins −1 and −2 and ceruloplasmin), five human counterparts, pufferfish hephaestin, and zebrafish ceruloplasmin. A phylogenetic tree generated from the alignment is in complete accord with there being a single encoded protein of the factor V–VIII kind in lamprey genomes (Fig. 3). (The amino acid sequences of the four lamprey ferroxidase proteins are provided in the Supplementary Material.)



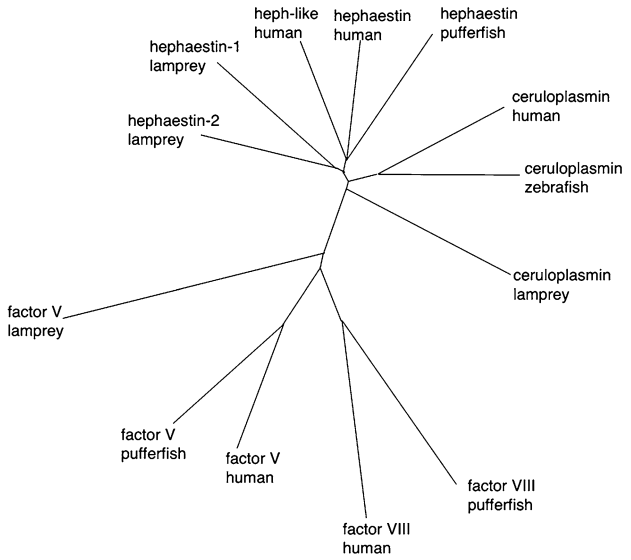


Fig. 3 Phylogenetic tree (unrooted) of ferroxidase family proteins (coagulation factors V and VIII, hephaestins, and ceruloplasmin) from human (5 entries), lamprey (4 entries), pufferfish (*Takifugu rubripes*) (3 entries), and zebrafish (*Danio rerum*) (1 entry). Each of the 13 entries was a concatenated sequence of five segments corresponding to the available partial sequences for lamprey factor V. The five segments represented all three A domains

Vitamin K-Dependent Factors

Previously, we had reported the existence of seven vitamin K-dependent proteases in the sea lamprey, *P. marinus*, based on Trace Archive data (which included some EST and non-genomic sequences). All told, 215 hits were gathered into 38 hit-groups, independent of eight others for GLA domains (see below). The seven proteins identified were prothrombin, protein C, two factors X, and three factors VII. One of the factors X (factor XB) was found to lack an activation cleavage site, casting doubt on its role as a prothrombinase. At the time, it had not been possible to provide sequences or link together traces for all three of the factors VII.

Since that report, an article has appeared reporting the cloning of messages from the liver of the Japanese lamprey (*L. japonicum*) for prothrombin, protein C, a factor VII, and two factors X, one of which, like the putative sea lamprey protein, lacks an activation cleavage site (Kimura et al. 2009).

As a reminder, the vitamin K-dependent clotting factors contain amino-terminal sections with several gamma-carboxylated glutamic acid residues, casually referred to as "GLA domains." In contrast to the 16 GLA domains found in most vertebrates, we had only found eight in sea lampreys (Doolittle et al. 2008), but now a search of the genome for the Japanese lamprey has revealed 12 GLA domains, two of which are 96 % identical. All 12 have been linked with their parent proteins (Table 2). Four of these are not associated with serine proteases and correspond to protein S, growth-arrest protein (homologous to protein S), and two GLA-transmembrane proteins. The other eight occur in prothrombin, protein C, three factors VII, and three factors X. The third factor X was a surprise and found to lie adjacent to what had been termed the factor XA gene (LJ scaffold 00078). The two putative proteins are 92 % identical and clearly the result of a fairly recent duplication, not much different from the time the sea (PM) and Japanese (LJ) lampreys diverged. There was no evidence for a comparably recent duplication in *P. Marinus*, the single factor XA being located in the middle of a vary large scaffold in the 2012 assembly.

As was the case for the ferroxidase family proteins, it was possible to assemble the sequences of the vitamin K-dependent factors by taking advantage of sequences being found in one or the other databases that were not present in others. The approach was particular helpful in characterizing the three factor VII proteins. As an example, in the case of factor VII C, scaffold 00627 of the Japanese lamprey contains the GLA domain and almost all of the protease domain, but the regions of the gene between them, which include the two EGF domains, remain unsequenced (NNNN). By good fortune, the 2007 draft assembly for *P. marinus* had a contig (35757) that contained the GLA domain and the first EGF section, and a scaffold from the 2012 PM assembly contained both EGF domains (but no GLA domain) (Fig. 4).

Similarly, the scaffold containing the factor VIIA gene (scaffold 00301) also had large unsequenced regions, one of which fell in the protease region of the gene (Fig. 4). In this case, matters were greatly facilitated by the cDNA for this protein having been determined (Kimura et al. 2009). With the aid of that sequence, it was possible to use various traces, contigs, and scaffolds to reconstitute most of the protein sequence for the sea lamprey, *P. marinus*.

The putative sequences for the eight vitamin K-dependent factors of interest are included in the Supplementary Material. A phylogenetic tree of various vitamin K-dependent proteases that included factor IX sequences from various species was wholly consistent with its absence in lampreys (Fig. 5).

Some Other Clotting Proteins

Fibrinogen

Although the sequences for the various chains of sea lamprey fibrinogen have been long known, the Japanese lamprey allows the full gene structures to be revealed. Unlike the situation in mammals where the α, β, and γ genes are clustered together in a 50-kilobase region that is coordinately regulated with regard to gene expression, in

Table 2 Vitamin K-dependent proteins identified in *L. japonicum* assembly

Protein	Protein length (aa)	Scaffold	Scaffold length (bp)	Start[a]	End	Gene length (bp)	GLA-motif[b]
Prothrombin	605	Scaf 00194	1,091,192	904,400	894,937	9,463	CLEETCSH
Protein C	413	Scaf 00651	225,932	127,753	118,776	8,977	CVEETCTM
Factor XA1	478	Scaf 00078	2,371,380	260,519	246,330	14,189	CMEERCSF
Factor XA2	478	Scaf 00078	2,371,380	270,333	263,087	7,246	CMEERCSF
Factor XB	471	Scaf 00705	203,506	183,730	118,882	64,848	CNEERCSI
Factor VIIA	484	Scaf 00301	662,907	528,771	509,235	19,536	CREETCNF
Factor VIIB	396	Scaf 00131	1,685,660	644,438	632,313	12,110	CREETCSF
Factor VIIC	393	Scaf 00627	242,945	73,192	59,516	13,715	CMEEHCSL
Protein S	ca 650	Scaf 00174	1,243,958	444,836	425,538	19,298	CVEEFCNK
Growth arrest	ca 645	Scaf 00520	307,154	227,176	>238,937	>11,761	CVEEVCSK
GLA-TM1	ca 215	Scaf 00100	2,104,269	>964,076	<962,198	>1,878	CVEERCNY
GLA-TM2	ca 240	Scaf 00023	4,607.062	1,442,828	<1,440,770	>2,058	CNEELCSY

[a] Start denotes position of first amino acid codon. If "start" nucleotide number is larger than "end" nucleotide number, reverse strand coding

[b] As a rule, the various GLA domains in lamprey proteins can be distinguished from each other on the basis of a short motif. The one exception is that the motifs for the two factors XA, whose genes are adjacent in *L. japonicum*, are identical. The corresponding motifs are very similar in the *P. marinus* proteins, in only three cases there being a single difference between the two species

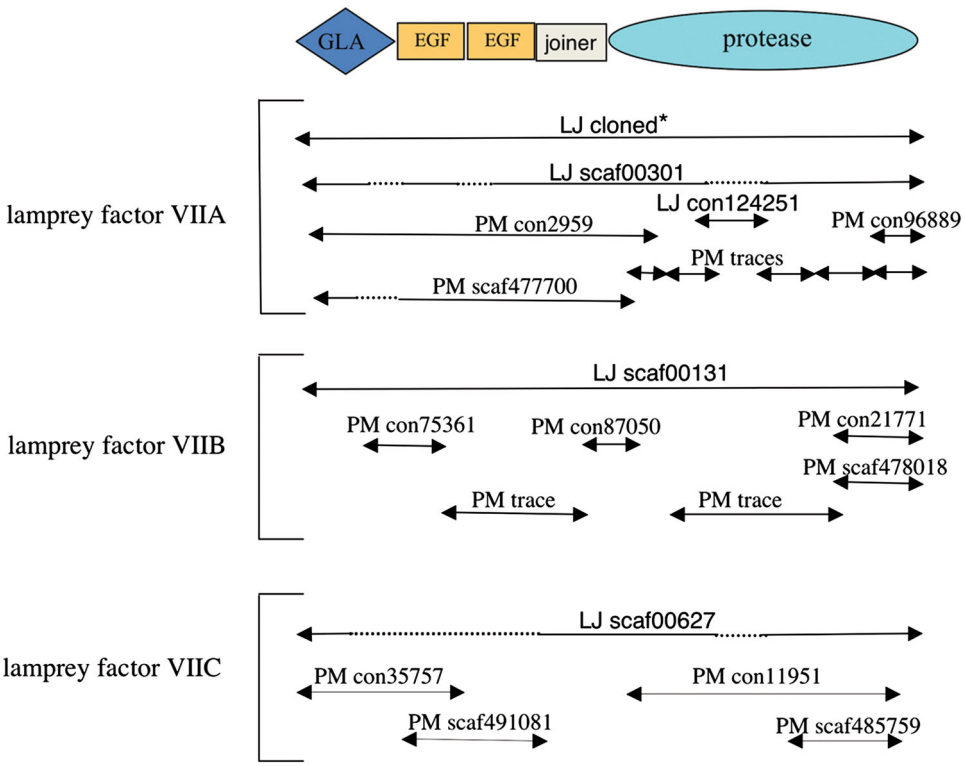


Fig. 4 Schematic depiction of three different factors VII from lamprey showing sources of sequence data. *LJ* data from *L. japonicum*; *PM* data from *P. marinus*. The cartoon across the *top* shows the domainal arrangement of the proteins: *GLA* γ-carboxy-glutamic acid containing domain; *EGF* epidermal growth factor domain; joiner, joining region; *SP* serine protease section. *Dotted lines* denote unsequenced regions of scaffolds

the lamprey the α, β, and γ genes are on separate scaffolds, the great lengths of which show that in the lamprey these genes cannot be within a megabase of each other. Not surprisingly, the α (Wang et al. 1989) and α_2 (Pan and Doolittle 1992) genes are adjacent to each other on the same scaffold (LJ scaffold 00123).

Tissue Factor

In our earlier report, we had been unable to identify a gene in the sea lamprey for tissue factor, a notoriously fast changing protein, even though lamprey tissue factor had been long ago characterized biochemically. We have now

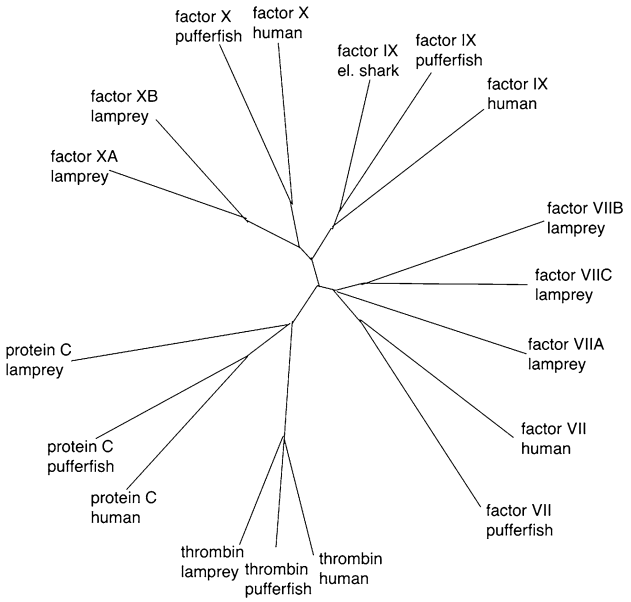


Fig. 5 Phylogenetic tree (unrooted) of 18 vitamin K-dependent clotting factors (serine protease domains only) from lamprey (*P. marinus* and/or *L. japonicum*), elephant shark (*Callorhinchus milii*) (only factor IX), pufferfish (*Takifugu rubripes*), and human

re-examined the various databases and found appropriate matches covering the whole protein in the 2007 draft assembly, two scaffolds amounting about half the protein in the 2012 assembly, and a virtually complete sequence in the Japanese lamprey genome. (The amino acid sequences are provided in the Supplementary Material.)

Species Differences

The use of the LJ assembly in coordination with sundry sea lamprey data afforded an opportunity for numerous direct amino acid sequence comparisons for many of the clotting factors (Table 3). On the average, these proteins are about 95 % identical in the two species, in line with the suggested divergence time of 10–30 million years based on similar data for other proteins (Kuraku and Kuratani 2006).

Discussion

The primary aim of this report is to update and solidify the proposal that the lamprey genome contains a smaller number of standard blood clotting factors than jawed vertebrates. Beyond that, a strategy is described for using a combination of various sources of lamprey sequence data that can overcome the limitations and incompleteness of currently available lamprey genome assemblies.

Particularly, we had cautiously proposed that the lamprey has a reduced set of clotting factors, corresponding to

Table 3 Percent identities for 12 orthologous proteins in *P. marinus* and *L. Japonicum*

Protein	LJ length	PM length	Percent identity[a]
Prothrombin	605	582	93.1
Protein C	413	414	95.6
Factor XA	478	467	95.7
Factor XB	471	472	96.0
Factor VIIA	463	402	94.8
Factor VIIB	396	299	96.3
Factor VIIC[b]	234	222	92.8
Hephaestin-2	1101	1119	94.2
Ceruloplasmin	1050	1050	96.4
Fibrinogen γ	330	408	97.3
Fibrinogen β	448	480	96.2
Fibrinogen α[c]	219	219	90.9

In some cases, sections of DNA encoding protein from one of the species is still missing, as indicated by a significantly smaller number of residues. All sequences are available as Supplemental Material

[a] Average percent identity for 12 sequences = 94.9; weighted average (for lengths) = 94.5

[b] Serine protease portion only

[c] Residues 1–219 only

what would have been in place before the duplication of two different kinds of protein, a vitamin K-dependent protease, for one, and a (factor V–VIII) ferroxidase family protein, for the other. Even with the newly available data, it has not yet been possible to reconstruct the full sequence for the putative pre-duplication factor V/VIII gene, but the fact that there is one, and only one, is now certain. The full reconstruction of the closely related ferroxidase proteins hephaestins −1 and −2 and ceruloplasmin has been completed.

Several kinds of evidence speak for the absence of the vitamin K-dependent protease factor IX. First, all of the 12 GLA domains in the Japanese lamprey have been identified. Of these, the only possible candidates for a factor IX would be one of the two "extra" factors VII. These sequences are very different from factor IX, however, and contain features typically found in factor VII and X, including an additional disulfide bond near the amino-terminus of the A-chain in the activated protease. It might also be mentioned again that the sequence similarities between known factors IX and X are only slightly less than the resemblances observed between fish and human sequences for those two factors, implying that the duplication event occurred not long before the appearance of jawed fishes (Fig. 5). It is always more difficult to prove the absence of a gene than its presence in a sequence database, but the case is strong that a gene corresponding to factor IX is not present in the lamprey genome.

What makes the question of when the coagulation factors VIII and IX first made their appearance interesting is that it offers a unique example of two interacting proteins in a cascade, one a protease and the other a large MW cofactor, having their genes duplicated at (or about) the same time and as a consequence expanding an existing pathway in a kind of double-jump. Fuller descriptions of how the clotting system may have gotten started and how it continued to evolve during vertebrate evolution have been presented (Doolittle 2009, 2012).

It would have been better, of course, if all the clotting factors considered here could have been identified (or not) directly in a high-quality, fully completed genome assembly for the lamprey. Neither the 2012 assembly for the sea lamprey (Smith et al. 2013) nor the 2013 assembly for the Japanese lamprey (Mehta et al. 2013) fully meet the criterion, however. It should be emphasized that the missing data from the 2012 sea lamprey assembly are not because of programmed loss of DNA during development (Smith et al. 2009). The main reason for the deficiency is because a large number of trace reads were omitted during the assembly process, mostly because of being considered repetitive DNA.

To emphasize the point, more than a third of the ($\sim$50) hit-groups for the ferroxidase family of proteins were not found in the sea lamprey assembly even though they are present in the Trace Archive. Many appear on the long list of "unplaced reads" provided as part of Supplementary Material (Smith et al. 2013). That some key genes reside in parts of the genome that are extensively populated with repeats is unfortunate but something that needs to be dealt with, difficulties with the assembly process aside.

The strategy of using numerous short sequences in the *Petromyzon marinus* Trace Archive together with the two partially assembled genomes should be applicable for studies already under way for other vertebrate systems, including, for example, the evolution of complement systems (Kimura et al. 2009), the extracellular matrix (Hynes 2012; Adams et al. 2015), bone formation (Zhang et al. 2006), and numerous other systems that have been expanded by gene duplications during the course of vertebrate evolution. In a sense, the use of this approach allows one to make use of information discarded during the assembly process. Clearly, different assemblies of the same sequence data can end up with different regions being excluded. Used together and with "reads" from the Trace database, these assemblies can be used to encompass entire genes, introns included.

The strategy described here, which depends heavily on manual curation, has the advantage that individual researchers bring a familiarity to focused projects that can greatly aid interpretation of the data. In this regard, it was disconcerting to read in the report covering the 2012 assembly (Smith et al. 2013) that the sea lamprey had *lost clotting genes*, in contrast to our earlier (and current) findings that several clotting genes had never been there to lose (Doolittle 2009, 2012). It was also dispiriting to read of the discovery that sea lamprey codon usage is greatly biased in favor of G and C at the third position, an observation reported three decades ago (Strong et al. 1985; Bohonus et al. 1986; Pontes et al. 1988, *inter alia*) and rediscovered in 2011 (Qiu et al. 2011).

About Whole Genome Duplications

Certainly there is uniform agreement that the vertebrate blood coagulation pathway is the result of a series of gene duplications; it is the timing of those duplications that is an issue. At one point, it was suggested that the gene duplications responsible for clotting factor genes may be linked to whole genome duplication events (Davidson et al. 2003b) in accord with the proposal that two rounds of whole genome duplication occurred during the early evolution of vertebrates (Ohno 1970). The 2R hypothesis, as it has come to be known, has been much debated ever since, and the two recent lamprey genome assemblies seem to have yielded two different interpretations of the matter (Mehta et al. 2013; Smith and Keinath 2015).

We have no wish to become embroiled in the debate about the timing of the two rounds of duplication and whether or not one or both preceded the divergence of Agnatha except to note that our findings are in complete accord with an early proposal of one round of whole genome duplication occurring before the advent of cyclostomes and another after their appearance (Escriva et al. 2002). Certainly in the wake of this event, the lineage leading to jawed vertebrates experienced a large scale block duplication—even if less than genome wide—that simultaneously gave rise to genes that were destined to encode factors VIII and IX, quite apart from several independent duplications that gave rise to additional genes for factors VII and X on the lineage leading to lampreys.

Acknowledgments I thank Kyle Mcnamara, Russell Darst, Akash Sonkar, and Willie Chao for their assistance during the course of this study.

Funding This work was supported in part by a Grant from the Academic Senate of U. C. San Diego.

References

Adams JC, Chiquet-Ehrismann R, Tucker RP (2015) The evolution of tenascins and fibronectin. Cell Adhes Migr 9:22–33

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new

generation of protein database search programs. Nucl Acids Res 25:3389–3402

Bohonus VL, Doolittle RF, Pontes M, Strong DD (1986) Complementary DNA sequence of lamprey fibrinogen β chain. Biochemistry 25:6512–6516

Daimon M, Yamatani K, Igarashi M, Fukase N, Kawanami T, Kato T, Tominaga M, Sasaki H (1995) Fine structure of the human ceruloplasmin gene. Biochem Biophys Res Commun 208:1028–1035

Davidson CJ, Hirt RP, Lal K, Elgar G, Tuddenham EGD, McVey JH (2003a) Molecular evolution of the vertebrate blood coagulation network. J Thromb Haemost 1:1487–1494

Davidson CJ, Tuddenham EG, McVey JH (2003b) 450 million years of hemostasis. Thromb Haemost 89:420–428

Doolittle RF (2009) Step-by-step evolution of vertebrate blood coagulation. Cold Spring Harbor Symp Quant Biol 74:35–40

Doolittle RF (2012) The evolution of vertebrate blood coagulation. University Science Books, Mill Valley

Doolittle RF, Feng D-F (1990) Nearest neighbor procedure for relating progressively aligned amino acid sequences. Methods Enzymol 183:659–669

Doolittle RF, Surgenor DM (1962) Blood coagulation in fish. Am J Physiol 203:964–970

Doolittle RF, Jiang Y, Nand J (2008) Genomic evidence for a simpler clotting scheme in jawless vertebrates. J Mol Evol 66:185–196

Escriva H, Manzon L, Youson J, Laudet V (2002) Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol Biol Evol 19:1440–1450

Feng D-F, Doolittle RF (1996) Progressive alignment and phylogenetic tree construction of protein sequences. Methods Enzymol 266:368–372

Gregory TR (2005) Animal genome size database. http://www.genomesize.com

Hanumanthaiah R, Day K, Jagadeeswaran P (2002) Comprehensive analysis of blood coagulation pathways in Teleostei: evolution of coagulation factor genes and identification in zebrafish factor VIIi. Blood Cells Mol Dis 29:57–68

Hynes RO (2012) The evolution of metazoan extracellular matrix. J Cell Biol 196:671–679

Jiang Y, Doolittle RF (2003) The evolution of vertebrate blood coagulation as viewed from a comparison of pufferfish and sea squirt genomes. Proc Natl Acad Sci USA 100:7527–7532

Kimura A, Ikeo K, Nonaka M (2009) Evolutionary origin of the blood complement and coagulation systems inferred from liver EST analysis of lampreys. Dev Comp Immunol 33:77–87

Kuraku S, Kuratani S (2006) Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. Zool Sci 23:1053–1064

Mehta TK, Ravi V, Yamasaki S, Lee AP, Lian MM, Tay B-H, Tohari S, Yanai S, Tay A, Brenner S, Venkatesh B (2013) Evidence for at least six Hox clusters in the Japanese lamprey (*Lethenteron japonicum*). Proc Natl Acad Sci USA 110:16044–16049

Ohno S (1970) Evolution by gene duplication. Springer, New York

Pan Y, Doolittle RF (1992) cDNA sequence of a second fibrinogen α-chain in lamprey: an archetypal version alignable with full-length β and γ chains. Proc Natl Acad Sci USA 89:2066–2070

Ponczec MB, Gailani D, Doolittle RF (2008) Evolution of the contact phase of blood coagulation. J Thromb Haemost 6:1–8

Pontes M, Xu X, Graham D, Riley M, Doolittle RF (1988) cDNA sequences of two apolipoproteins from lamprey. Biochemistry 26:1611–1617

Qiu H, Hildebrand F, Kuraku S, Meyer A (2011) Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case. BMC Genom 12:325

Smith JJ, Keinath MC (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. Genome Res. doi:10.1101/gr.184135.11

Smith JJ, Antonacci F, Eichler EE, Amemiya CT (2009) Programmed loss of millions of base pairs from a vertebrate genome. Proc Natl Acad Sci USA 106:11212–11217

Smith JJ et al (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat Genetics 45:415–421

Strong DD, Moore M, Cottrell BA, Bohonus VL, Pontes M, Evans B, Riley M, Doolittle RF (1985) Lamprey fibrinogen γ chain: cloning, cDNA sequencing and general characterization. Biochemistry 24:92–101

Wald G, Riggs A (1951) The hemoglobin of the sea lamprey, *Petromyzon marinus*. J Gen Physiol 35:45–53

Wang Y-Z, Patterson J. Gray JE, Yu C, Cottrell BA. Shimizu A, Graham D, Riley M, Doolittle RF (1989) Complete sequence of the lamprey fibrinogen α chain. Biochemistry 28: 9801-9806

Zhang G, Miyamoto MM, Cohn MJ (2006) Lamprey type II collagen and Sox9 reveal an ancient origin of the vertebrate collagenous skeleton. Proc Natl Acad Sci USA 103:3180–3185