Biochemical and Biophysical Research Communications 455 (2014) 98–106

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Surveying genetic variants and molecular phylogeny of cerebral cavernous malformation gene, CCM3/PDCD10

Abhishek Kumar [a,*], Anita Bhandari [b], Chandan Goswami [c]

[a] Department of Genetics & Molecular Biology in Botany, Institute of Botany, Christian-Albrechts-University at Kiel, Kiel, Germany
[b] Molecular Physiology, Christian-Albrechts-University at Kiel, Kiel, Germany
[c] National Institute of Science Education and Research, Bhubaneswar, Orissa, India

ARTICLE INFO

Article history:
Received 15 October 2014
Available online 27 October 2014

Keywords:
Cerebral cavernous malformations
Endothelium
CCM3/PDCD10
Synteny
Phylogenetic analysis
Genetic variants

ABSTRACT

The three cerebral cavernous malformations (CCMs) genes namely CCM1/KRIT1, CCM2/MGC4607 and CCM3/PDCD10 have been identified for which mutations cause cerebral cavernous malformations. However, the protein products of these genes involved in forming CCM signaling, are still poorly understood imposing an urgent need to understand these genes and their signaling processes in details. So far involvement of CCM3/PDCD10 in the cavernous angioma has been characterized from biochemical and biophysical analyses. However, there is no comprehensive study illustrating the phylogenetic history and comprehensive genetic variants of CCM3/PDCD10. Herein, we explored the phylogenetic history and genetic variants of CCM3/PDCD10 gene. Synteny analyses revealed that CCM3/PDCD10 gene shared same genomic loci from *Drosophila* to human and the gene structure of CCM3/PDCD10 is conserved from human to *Branchiostoma floridae* for about 500 MYs with some changes in sea urchin and in insects. The conserved CCM3/PDCD10 is characterized by presence of indels in the N-terminal dimerization domain. We identified 951 CCM3/PDCD10 variants by analysis of 1092 human genomes with top three variation classes belongs to 84% SNPs, 6.9% insertions and 6.2% deletions. We identified 22 missense mutations in the human CCM3/PDCD10 protein and out of which three mutations are deleterious. We also identified four stop-codon gaining mutations at the positions E34*, E68*, E97* and E140*, respectively. This study is the first comprehensive analysis of the CCM3/PDCD10 gene based on phylogenetic origin and genetic variants. This study corroborates that the evolution of CCM proteins with tubular organization evolvements by endothelial cells.

1. Introduction

Cerebral cavernous malformations (CCMs; OMIM 116860) are hamartomatous vascular malformations that can occur as a sporadic or a familial autosomal dominant disorder [1,2]. CCMs are characterized by enlarged vascular cavities without intervening brain parenchyma with an estimated prevalence of 0.1–0.5% in general population. Single or multiple malformations may develop, which can lead to cerebral hemorrhage (30–40%), seizures (40–70%), headache (10–30%) and focal neurological symptoms (35–50%). The onset age is variable with higher incidence between 10 and 40 years. CCM may occur sporadically or with an autosomal dominant inheritance pattern with variable expression and incomplete penetrance. Almost one fourth of CCM carriers remain symptom-free throughout their lives. In the human, so far three CCM genes have been identified which are located in three different CCM loci: CCM1/KRIT1 in chromosome 7q21.2, CCM2/MGC4607 in 7p13, and CCM3/PDCD10 in 3q25.2-q27 respectively [3,4]. So far, most of the mutations in these genes reported lead to either introduce a premature termination codon or large deletions, again strongly suggesting that most of these mutations are "loss-of-function" mutations [5,6]. Human CCM3/PDCD10 gene encodes for a protein, known as programmed cell death 10 (PDCD10) and hence called as CCM3/PDCD10. Recent studies revealed the presence of CCM3 along with CCM1–CCM2 in the CCM protein complex and this suggest for a common signaling mechanism/pathway. However, this common-signaling pathway involved in CCM-signaling has not been established yet. CCM3 is known to bind GCKIII family of sterile 20-like serine/threonine kinases STK24, STK25 and MST4 [5,6]. These signaling are critical for vascular development and for cell survival. CCM3 has also been shown to bind paxillin, membrane protein VEGF-R2 and HEG1

* Corresponding author.
E-mail addresses: abhishek.abhishekkumar@gmail.com, akumar@bot.uni-kiel.de (A. Kumar), anita.bhandari@gmail.com (A. Bhandari), chandan@niser.ac.in (C. Goswami).

http://dx.doi.org/10.1016/j.bbrc.2014.10.105
0006-291X/

[5,6]. This signaling network is known to regulate cell adhesion, angiogenesis and vascular integrity [5,6]. After a decade of research on vascular malformations, it appears CCM3/PDCD10 is one of the hallmark genes of the CCM complex or the signaling network.

The crystal structures of CCM3/PDCD10 become available recently [7,8]. The α-helical structure of human CCM3/PDCD10 contains a N-terminal dimerization domain (four α-helices) and a C-terminal focal adhesion targeting (FAT)-homology domain (four α-helices), separated by an α-helix served as a linker between two these domains [7,8].

However, there is no report yet known on molecular evolution and a through study of genetic variants of human CCM3/PDCD10. Hence, investigations from these aspects of CCM3/PDCD10 are urgent requirements. In order to gain insight of structure–function relationship in details, in the current study, we analyzed the phylogenetic history of CCM3/PDCD10 genes by focusing gene structures, synteny and sequence-structural properties from selected eukaryotic genomes. Furthermore, we have identified 951 CCM3/PDCD10 variants from 1092 human genomes and 84% of these variants are SNPs. These genetic variants are critical indicators of cerebral cavernous malformations due to CCM3/PDCD10in human population. This studies in general shade lights on the molecular evolution of CCM proteins, which play important roles in endothelial cells and in angiogenesis.

2. Materials and methods

We extracted genomic DNA and protein sequences from different vertebrate genomes via Ensembl release 73 (September 2013) [9] using BLAST suite for CCM3/PDCD10 are provided in Table S1. We scanned chromosomal locus for CCM3/PDCD10 gene for each species using Ensembl genome browser [9] and the map viewer from NCBI (website: http://www.ncbi.nlm.nih.gov/mapview/). We predicted gene structures of CCM3/PDCD10 using AUGUSTUS suite [10] and we combined with gene structure prediction within the Ensembl [9], which ensured accuracy.

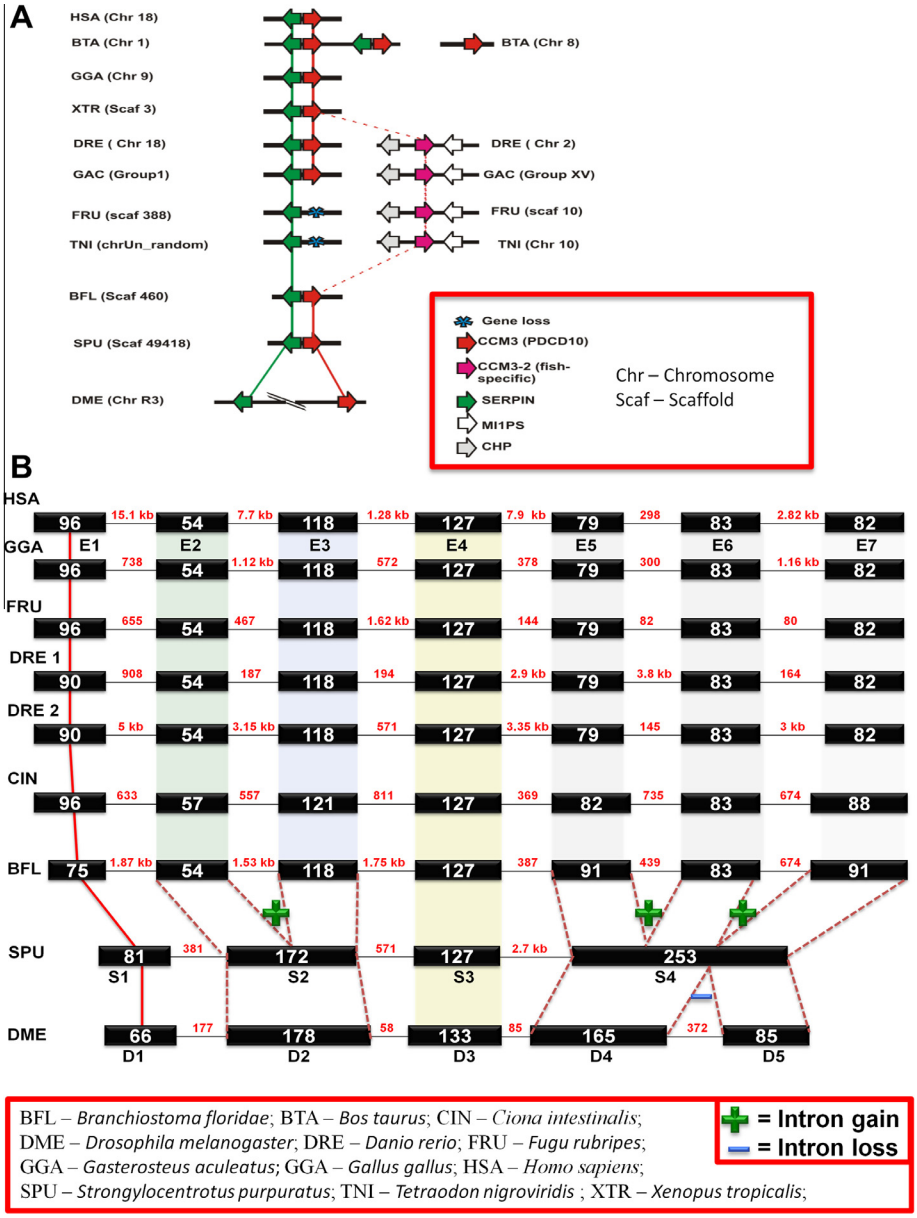

Fig. 1. Genomic and gene organizations of CCM3/PDCD10 genes. (A) Syntenic comparisons of CCM3/PDCD10 locus in different eukaryotes. (B) Gene structural changes of CCM3/PDCD10 gene from selected animals.

We created protein alignment of CCM3/PDCD10 using the CLUSTALW 1.83 [11,12] and edited and visualized in GENEDOC [13] as shown in Fig. S1. We computed genetic variants of human CCM3/PDCD10 using 1092 human genomes from 14 different populations available in 1000 genomes project [14]. Sorting Intolerant From Tolerant (SIFT) is a software tool, which predicts whether an amino acid substitution affects protein function and it helps in prioritizing substitutions for further study [15]. The SIFT value $\leqslant$0.05 indicates the deleterious effects of missense variants on protein function [15]. Phenotyping V2 (PolyPhen-V2) is a tool that predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations [16]. PolyPhen-V2 score (close to 1) indicates the damaging effects of missense variants on protein function. We used these two methods to predict the impact of these variants on function of CCM3/PDCD10. Missense variants were analyzed using variant modeler (VarMod) tool [17]. We constructed the phylogenetic tree of CCM3/PDCD10 proteins aided by Neighbor Joining (NJ) method [18] with 1000 bootstraps using MEGA5 [19].

3. Results

3.1. CCM3/PDCD10 gene is maintained from insects to human on the same locus

The CCM3/PDCD10 gene is located in reverse orientation with respect to a neuroserpin (red color) in the human chromosome 18 (Fig. 1A). This genomic organization is maintained in all known mammalian genome. Interestingly, bovine has two copies of CCM3/PDCD10-neuroserpin, cluster tandemly duplicated on the chromosome 1 and additionally a single copy is found in the chromosome 8. Chicken and *Xenopus tropicalis* has conserved loci of CCM3/PDCD10-neuroserpin in the chromosome 9 and in the scaffold 3, respectively. Noteworthy, selected fishes have two copies of CCM3/PDCD10 gene, localized in two different genomic locations. For example, *Danio rerio* and *Gasterosteus aculeatus* have two CCM3/PDCD10 genes localized on chromosomes 1 and 2, and groups 1 and XV, respectively. In the first location, it is maintained with neuroserpin, while in other location second copy of CCM3/PDCD10 gene is flanked by CHP and M1PS genes. It is notable that *Fugu* and *Tetraodon* have lost CCM3/PDCD10 gene in first location and retained second copy of CCM3/PDCD10 gene. However, CCM3/PDCD10 clustered is maintained in *Branchiostoma floridae* and in sea urchin. Insects have CCM3/PDCD10 on the same chromosome but serpin gene is localized 60 Mb upstream region such as in the chromosome R3 in *Drosophila melanogaster*. This suggested that CCM3/PDCD10 and serpin gene are localized on the same syntenic organization from *Drosophila* to human for about 700 MY.

3.2. Variations in CCM3/PDCD10 gene structure

Human CCM3/PDCD10 gene has seven exons E1–E7 with exon size ranged from 54–127 bp while intron length is >1 kb except for intron joining exons E5–E5 (Fig. 1B). This gene structure is maintained in CCM3/PDCD10 genes of chicken, *Fugu* and also in both copies of CCM3/PDCD10 from *D. rerio*. Interestingly, CCM3/PDCD10 from *Ciona* and *Branchiostoma* also possesses this gene structures. However, variations in gene structures have been observed in CCM3/PDCD10 gene from sea urchin, *Strongylocentrotus purpuratus* and *D. melanogaster*. Sea urchin possesses exon E1 and E4 which are conserved and dubbed as S1 and S3, where exons E2–E3 are clubbed together forming a new exon named as S2.

Similarly, exons E5–E7 are fused to form exon S4 in CCM3/PDCD10 gene from sea urchin. Likewise, *D. melanogaster* has conserved exons E1 and E4 as D1 and D3, respectively. But, it retained sea urchin exon S2 as exon D2. Whereas, exon S4 is divided into two exons, called as D4 and D5. The exon D4 has size of 165 bp and the size of exon D5 is 85 bp, which is comparable to human exon E7. This "intron gain" events are found from sea urchin to *Branchiostoma floridae* as the exon S2 is divided into exons E2–E3. Similarly, exon S4 is divided by two intron gains into three exons E5–E6–E7. There is a "loss of intron" too as exons D4–D5 (from *Drosophila*) joined to form exon S4 in sea urchin. Taken together, CCM3/PDCD10 gene structure is conserved from human to *B. floridae* for about 500 MYs. Whereas changes of gene structures are observed in sea urchin, separated about 550 MY ago (MYA) and changes are also seen in insect CCM3, which separated about 700 MYA.

3.3. Conserved CCM3/PDCD10 is characterized by indels

This protein is highly conserved in vertebrates so that human CCM3/PDCD10 shares 94.3% and 92.9% identities with *two* copies of CCM3/PDCD10 present in *D. rerio* (Table 1). Additionally, this protein is conserved in different invertebrate species such as

Table 1
Sequence identity scores for CCM3/PDCD10 across different eukaryotes. Above 90% of sequence identities are marked in bold.

	CEL	SPU	DME	DPS	AAE	AGA	NVI	AME	BFL	GAC1	BTA3	GAC2	TNI	FRU	DRE1	DRE2	HSA	BTA1	BTA2
CBR	**96.7**	25.5	35.3	35.8	38.6	38.6	37.3	39.5	31.8	32.4	31.0	33.3	33.8	33.8	32.9	33.8	33.8	33.3	38.6
CEL		25.0	35.3	35.8	38.6	38.6	37.3	38.6	31.4	32.4	31.5	33.3	33.8	33.8	32.9	33.8	34.3	33.8	39.2
SPU			36.9	36.9	40.7	40.7	33.6	37.9	49.0	41.9	41.9	42.9	42.9	42.9	45.2	43.8	44.8	44.8	46.7
DME				98.6	72.5	72.0	66.7	69.5	45.8	44.3	44.8	46.2	45.8	45.8	44.3	45.8	47.2	47.6	50.3
DPS					72.5	72.0	66.7	69.5	45.8	44.3	44.8	46.2	45.8	45.8	44.3	45.8	47.2	47.6	50.3
AAE						**97.2**	69.2	73.5	49.1	46.2	45.8	47.2	47.2	47.2	46.2	47.2	49.1	48.6	53.3
AGA							69.7	73.0	49.5	47.2	46.7	48.1	48.1	48.1	47.2	48.1	50.0	49.5	53.9
NVI								83.8	44.4	45.3	42.9	44.8	44.8	44.8	43.9	44.3	46.2	45.8	47.3
AME									47.2	46.7	46.2	47.6	47.6	47.6	47.2	47.6	49.5	49.1	50.9
BFL										57.4	59.8	62.2	61.7	61.7	64.1	62.2	64.1	64.1	67.5
GAC1											71.9	79.5	79.5	79.5	80.5	79.0	79.5	79.0	79.6
BTA3												85.7	85.2	85.2	85.7	85.2	**90.5**	**91.0**	**91.4**
GAC2													**99.5**	**99.5**	**92.9**	**94.8**	**93.8**	**94.3**	**94.4**
TNI														**100.0**	**92.4**	**94.3**	**93.3**	**93.8**	**93.8**
FRU															**92.4**	**94.3**	**93.3**	**93.8**	**93.8**
DRE1																**92.4**	**94.3**	**94.3**	**95.7**
DRE2																	**92.9**	**93.3**	**96.3**
HSA																		**99.5**	**99.4**
BTA1																			**100.0**

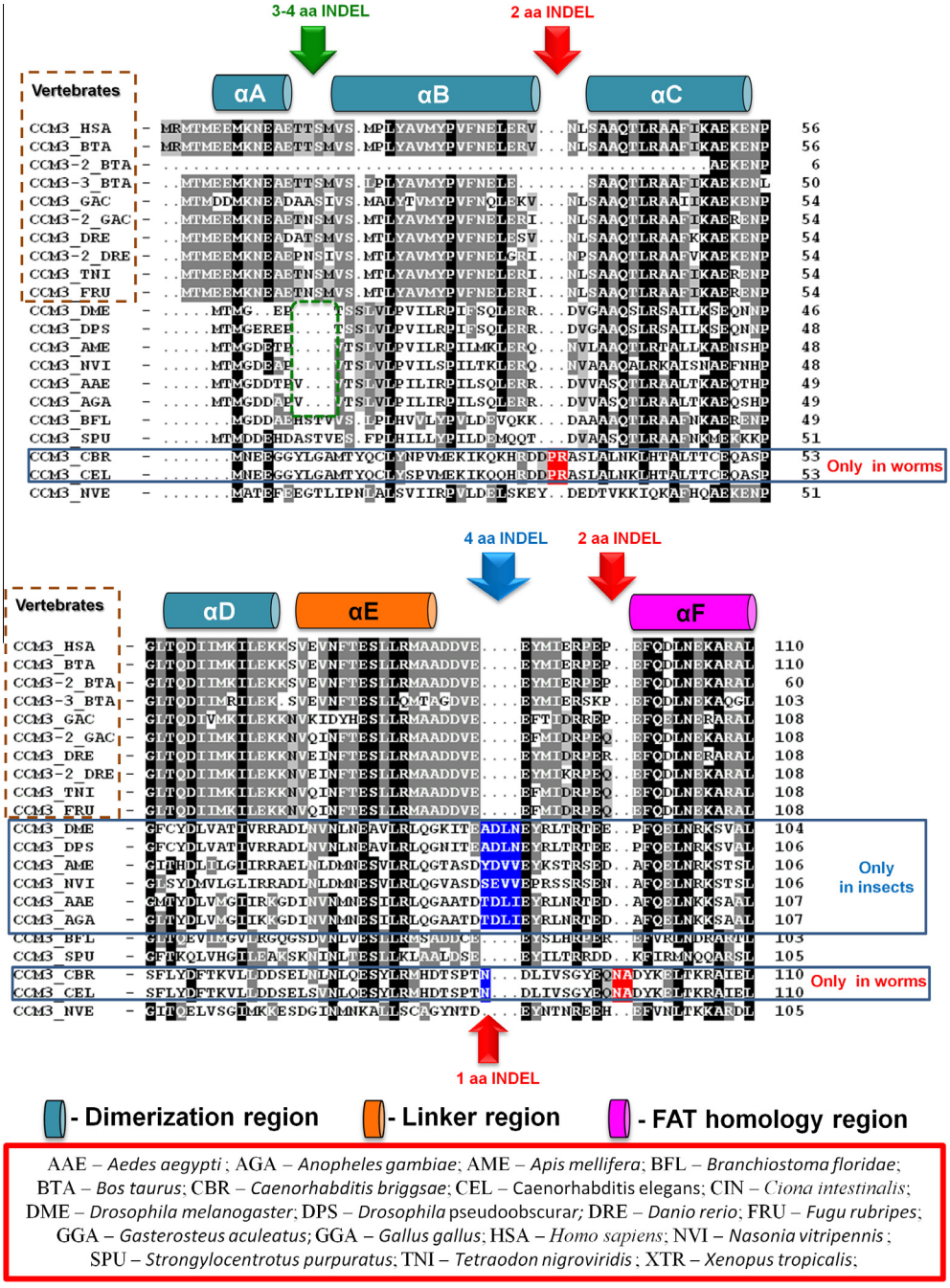

Fig. 2. Alignment of N-terminal region of CCM3/PDCD10 demonstrates that insertions/deletions (indels) are characteristic features of CCM3/PDCD10 proteins.

human CCM3/PDCD10 shares 64.1%, 50% and 33.8% identities with CCM3/PDCD10 orthologs in *B. floridae*, *Anopheles gambiae* and *Caenorhabditis briggsae*, respectively. This suggests that this protein is significantly conserved. Interestingly, CCM3/PDCD10 from different animal lineages are characterized by insertions/deletions (Indels) as shown in Fig 2. Insects have a deletion of 3–4 amino acids at the position 13 within the loop joining helices αA and αB and an insertion of four amino acids at the position 87, within the loop joining helices αE and αF (numbering according to human CCM3/PDCD10 sequence). Similarly, worms (Caenorhabditis *elegans* and Caenorhabditis *briggsae*) possess three insertions:(a) two amino acids P–R at the position 35 (within the loop joining the helices αB–αC), (b) one amino acid, a glutamine (N) at the position 87 (within the loop joining helices αE and αF) and (c) two amino acids N–A at the position 97 (just before helix αF), respectively. All these indels are in the N-terminal region. Only few are found in the C-terminal end or the FAT-homology-domain of PDCD10 protein. CCM3/PDCD10 proteins of *B. floridae* and *S. purpuratus* have a four amino acid indel at the position 154 between helices αG and αH (Fig. S1) and at the same position, insects have a single amino acid indel. Overall, indels are devising method for diversities in CCM3/PDCD10 gene across metazoa, largely in the N-terminal segment, which involved in dimerization.

3.4. Surveying genetic variants of human CCM3/PDCD10

We deduced 951 genetic variants of CCM3/PDCD10 from analysis of 1092 human genomes originated from 14 different global populations (Table 2). Four major genetic variants are intron variants (77.8%), downstream gene variants (7.7%), upstream gene

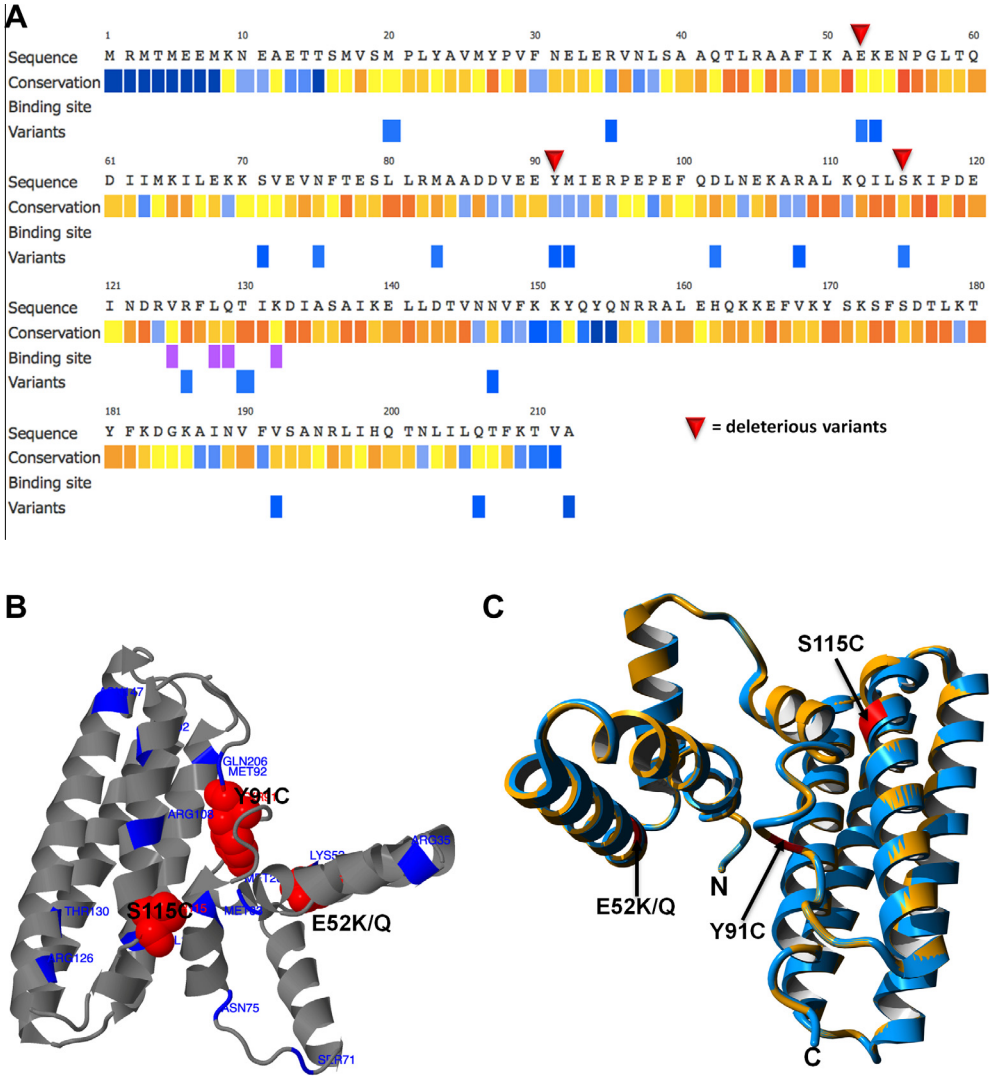

Fig. 3. Summary of missense variants of CCM3/PDCD10 protein. (A) Location of missense variants on the human CCM3/PDCD10 protein sequence. (B) Mapping missense variants on the human CCM3/PDCD10 structure. (C) Structural comparison of deleterious variations.

Table 2
Summary of human genetic variants of human CCM3/PDCD10 deduced from 1000 genome data.

Variant types	Deletion	Insertion	Sequence_alteration	SNP	Somatic_SNV	Somatic_deletion	Somatic_substitution	Total
Intron variant	43	60	1	636	0	0	0	740
Downstream gene variant	8	3	0	62	0	0	0	73
Upstream gene variant	1	2	0	53	0	0	0	56
Missense variant	0	0	0	13	8	0	1	22
3 prime UTR variant	3	0	0	13	0	0	0	16
Synonymous variant	0	0	0	8	4	0	0	12
Coding sequence variant	3	1	0	5	0	0	0	9
Splice region variant	1	0	1	4	1	0	0	7
5 prime UTR variant	0	0	0	7	0	0	0	7
Stop gained	0	0	0	0	0	4	0	4
Splice acceptor variant	0	0	1	0	2	0	0	3
Splice donor variant	0	0	1	0	0	0	0	1
Frameshift variant	0	0	0	0	0	1	0	1
Total	59	66	4	801	15	5	1	951

variant (5.9%), and missense variants (2.3%). Similarly, major constituents of variant classes were SNPs (84%), insertions (6.9%), deletions (6.2%) and somatic_SNVs (1.6%). Our analysis identified twenty-two missense mutations, which include three deleterious mutations (Table 3). Locations of these mutants are localized in the full length of CCM3/PDCD10 proteins with top secondary structures being the loop joining helices αE–αF (four variants), the helices αC, αF and αG (three each) as summarized in Table 3. These mutations are mapping onto protein sequence (Fig. 3A) and structure (Fig. 3B and C). Three deleterious mutations are shown as red

Table 3
List of missense variants of human CCM3/PDCD10 deduced from 1000 genome data. Deleterious mutations are underlined.

Protein mutant	Structural location	Variant ID	Chromosomal location	Alleles	gmaf	Class	Source	Status		SIFT value	POLYPHEN value
MT3IS	At the N-terminal end	COSM353187	3:167437936–167437937	TC/AA	–	SS	COSMIC	–		–	–
E7D	In the helix αA	COSM1484824	3:167437925	C/G	–	sSNV	COSMIC	–		0.66	0
M20V	In the helix αB	rs138275885	3:167437888	T/C	–	SNP	dbSNP	MO:Frequency ESP		0.26	0.002
R35P	At the end of the helix αB	COSM94947	3:167422676	C/G	–	sSNV	COSMIC	–		0.05	0.158
R35Q	At the end of the helix αB	rs201796692	3:167422676	C/T	–	SNP	dbSNP	–		0.45	0.002
E52K	In the helix αC	COSM345452	3:167414911	C/T	–	sSNV	COSMIC	–		0.01	0.998
E52Q	In the helix αC	COSM419725	3:167414911	C/G	–	sSNV	COSMIC	–		0.01	0.999
K53I	At the end of the helix αC	rs11541686	3:167414907	T/A	–	SNP	dbSNP	–		0.08	0.349
S71R	At the end of the helix αE	TMP_ESP_3_167414854	3:167414854	T/G	–	SNP	ESP	ESP		0.34	0.074
N75K	In the loop joining helices αE-αF	COSM582305	3:167414840	G/C	–	sSNV	COSMIC	–		0.07	0.587
M83T	In the loop joining helices αE-αF	rs11541685	3:167414817	A/G	–	SNP	dbSNP	–		0.07	0.207
Y91C	In the loop joining helices αE-αF	rs200148887	3:167413507	T/C	0.001 (C)	SNP	dbSNP	–		0	0.86
M92V	In the loop joining helices αE-αF	TMP_ESP_3_167413505	3:167413505	T/C	–	SNP	ESP	ESP		0.34	0
D102A	At the end of the helix αF	rs1129087	3:167413474	T/G	–	SNP	dbSNP	–		0.09	0.021
R108L	At the end of the helix αF	COSM729220	3:167413456	C/A	–	sSNV	COSMIC	–		0.32	0.479
S115C	At the end of the helix αF	COSM582306	3:167413436	T/A	–	sSNV	COSMIC	–		0	0.995
R126M	In the helix αG	rs182501365	3:167414023:167413402	C/A	0.001 (A)	SNP	dbSNP	–		0.11	0.419
T130I	In the helix αG	rs74635000	3:167413903:167413390	G/A	0.001 (A)	SNP	dbSNP	MO:1000 Genomes		0.13	0.331
N147D	In the helix αG	rs201857174	3:167405383:167405438	T/C	0.001 (C)	SNP	dbSNP	–		0.46	0.001
V192I	In the helix αI	rs151267430	3:167402161	C/T	–	SNP	dbSNP	ESP		0.47	0
Q206E	In the helix αI	COSM729222	3:167402119	G/C	–	sSNV	COSMIC	–		0.78	0.003
A212S	At the C-terminal end	TMP_ESP_3_167402101	3:167402101	C/A	–	SNP	ESP	ESP		0.53	0.001

SS, somatic_substitution; sSNV, somatic_SNV; MO, Multiple_observations.

arrows in Fig. 3A as E52Q (in the helix αC), Y91C (in the loop joining helices αE–αF), and S115C (at the end of the helix αF) as marked in red surface areas in Fig. 3B. The comparison of known structure and with these three mutations are shown in Fig. 3C. However, there is no significant difference can be visualized using homology modeling. Four stop-gained mutations were also identified (Table 4) at positions E34 (at the end of the helix αB), E68 (at the end of the helix αD), E97 (in the loop joining helices αD-αE) and E140 (in the helix αG), where glutamate codon G-A-[AG] changed at the first position, leading into stop codons, T-A-[AG].

4. Discussion

The common vascular dysplasia, cerebral cavernous malformation affects both systemic and central nervous system blood vessels. Understanding of the CCM complexes and associated functional networks are still in their infancies as individual components or genes are still not fully characterized. Here we have compiled and analyzed CCM3/PDCD10 gene from various animal genomes using sequence, genomic structure, gene structure pattern, and genetic variants.

CCM3/PDCD10 gene structure is conserved for about 500 MYs from *B. floridae* to human (Fig. 1A). There are three intron gains and single intron loss events when compared to sea urchin to *B. floridae* and insects to sea urchin, respectively (Fig. 1B). These "intron gains" and "intron loss" have been occurred between 700–550 MY. In contrast, reversely flanking neuroserpin orthologs depict different fates of exon–intron patterns, which are only conserved only invertebrates for about 450MYA. This serpin is member of the group V3 in the six group-wise vertebrate serpin (V1–V6) classification system, based on gene structure patterns [20]. But gene structure of neuroserpin is conserved as an "8-exon" architecture only in vertebrates and although neuroserpin ortholog is detected from sea anemone to human by tracing amino acid motifs, indels and syntentic organization [20]. None of the invertebrate serpins share gene structures [21,22] with this serpin [20]. These two physically-linked genes depicts that gene structural changes occurs at the different time points and are independent of close proximities. Genome changes normally accompanied by influx of changes of several types including gene structural genes. Spliceosomal introns and their splicing machineries are hallmarks of eukaryotic genomes and complexities in the gene regulation. However, the conundrums about creation of introns are associated with the discovery of introns about three and half decades ago [23]. Introns are either created or lost throughout eukaryotic evolution [24–29]. So far there are seven different mechanisms have been proposed for intron gain/invasions [27,28], which included intron transposition with partial recombination, transposon insertion, tandem genomic duplication using duplicated splice sites, double-strand break repair (DSBR), group II intron insertion, intron transfer (vii) intronization. These mechanisms are working on intron creation and losses in general and these can explain origin of novel introns and losses in CCM3/PDCD10 gene.

Indels are devising method for diversities in CCM3/PDCD10 gene across different animal genomes, largely in the N-terminal dimerization region (Fig. 2). Short indels are equally sensitive as primary sequences. These are significant factors in the mutational input for the evolutionary process. Indels are rare genomic characters, which are phylogenetic markers during evolution. There are several examples of indels being used as phylogenetic markers in eukaryotic genomes such as in case of vertebrate serpins [20,30–34], melanocortin receptors [35], bem46 gene [36] and also used for phylogenetic positioning of trichomonads [37]. Highly complex mutational mechanisms are needed to deal with exonic indels in comparison to single base substitutions.

Table 4
List of genetic variants of human CCM3/PDCD10 that caused gain of stop codon deduced from 1000 genome data.

Mutants	Structural location	Variant ID	Chromosomal location	Alleles	Class	Source
E34*	At the end of the helix αB	COSM729219	3:1674226803:167422680	C/A	Somatic_SNV	COSMIC
E68*	At the end of the helix αD	COSM336719	3:1674148633:167414863	C/A	Somatic_SNV	COSMIC
E97*	In the loop joining helices αD-αE	COSM370867	3:1674134903:167413490	C/A	Somatic_SNV	COSMIC
E140*	In the helix αG	COSM1219904	3:1674054593:167405459	C/A	Somatic_SNV	COSMIC

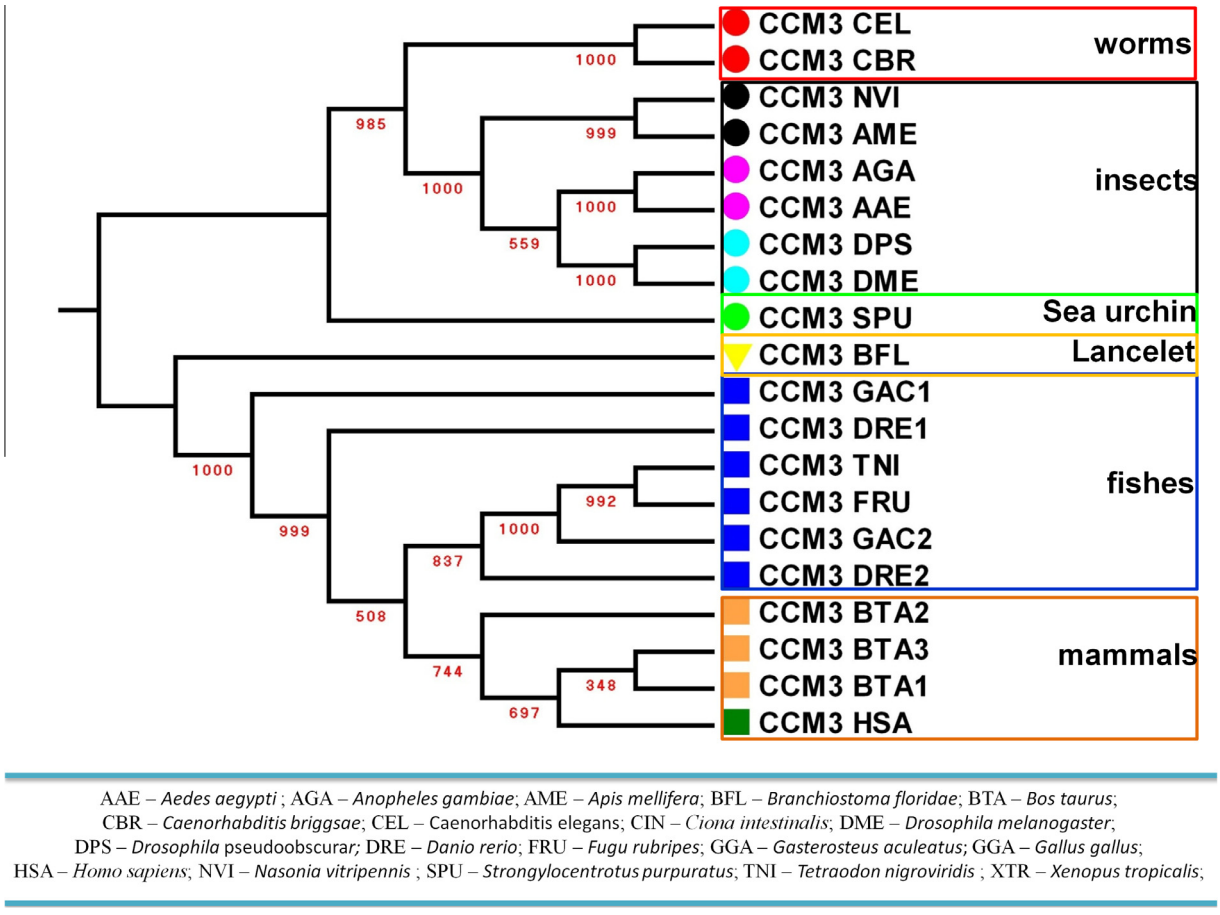

Fig. 4. Phylogenetic history of CCM3/PDCD10 in selected vertebrates (square) and invertebrates (circles) demonstrates presence of this gene for about 700 MY. This is a bootstrap consensus tree based on 1000 replicates. Following color-coding was followed in this tree:insects (cyan), worm (red), sea urchin (light green circle), human (green square), cow (orange square) and fishes (blue square), amphioxus (yellow triangle; BLF). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We identified 951 CCM3/PDCD10 variants by analysis of 1092 human genomes with SNPs (84%), being top variation class and followed by insertions (6.9%) and deletions (6.2%). The human CCM3/PDCD10 protein possesses 22 missense mutations and three were detected as deleterious for CCM3/PDCD10 protein by SIFT and POLYPHEN V2 tools (Table 3 & Fig 3). We also identified four terminal codon gaining mutations at the positions E34*, E68*, E97* and E140*, respectively (Table 4). These terminal codon-gaining mutations are medically important as the "loss-of-function" mutational sites of CCM3/PDCD10 are important for understanding the pathogenicity associated with the development of cerebral cavernous malformation. Even a single loss-of-function mutations are valued as such as in the case of a recent studies, where Q200* was identified by direct sequencing approach [38]. Hence, the four loss-of-function mutations identified in this work are of great medical importance. In addition, the other 951 genetic variants/mutations identified in this study may also have clinical significance (Tables 2–4 and S2). Our study shade light into angiogenesis and function of endothelial cell functions. This is due to the fact that CCM proteins are associated with the origin of endothelial cells and development of tubes (due to differentiation) by these cells. Although the true endothelial cells are found in the vertebrates with cardiovascular organizations only, however invertebrates too contain a transitional circulatory systems [39], such as an open circulatory systems in arthropods (insects and crustaceans) and non-cephalopod molluscs (clams, snails and slugs) [40]. Invertebrate vessels are associated by extracellular matrix. The blood vessels of some invertebrates (including cephalopods, annelids, and amphioxus) have cells clinging to the luminal surface, internal to the basement membrane [40]. These cells are called as 'endothelial cells' in *B. floridae* [41,42] and *D. melanogaster* [43]. But, these have incomplete lining as these lack intercellular junctions, which are prominent features of vertebrate endothelial cells. Sometimes, these cells appear attached to the underlying basal lamina. A more

appropriate term for this cell type is an amoebocyte. These amoebocytic cells may serve as an evolutionary precursor of the endothelial cell [39,40].

Majority of organs are composed of epithelial/endothelial cells, which form tubular organizations, and the junctional complexes maintain the intercellular or auto-cellular connections that seal the cells into selectively permeable tubes [44]. Recently, it is shown that CCM3 and GCKIII are required to avoid terminal cell tube dilation in the *Drosophila* tracheal system [44], in the similar manner as it is expected for CCM networking in the human [5,6]. By combining conserved sequence and syntenic organizations and presence of CCM3/PDCD10 gene from insects to human as shown in the phylogenetic tree (Fig. 4) with CCM3/GCKIII signaling pathway, it appears that at least a common pathway exists from insects to human for about 700 MY. These findings also support the existence of cardiovascular physiological similarities between vertebrates and insects as reported earlier for *D. melanogaster* [40]. We corroborates that CCM proteins are essential for endothelial cell-based tube formation and dilation. This hints that these networks have been evolved from *Drosophila* to human for about 700 MY, at least in primitive forms as junctional complexes. These complexes are required to be studies extensively in human, amphioxus, and insects, so to elaborate on the conserved signaling pathways that are functional from insects to human. It is also important to investigate further the roles of junctional proteins in the context of development as well as maintenance of endothelial cells. This will provide insight into the CCM complexes, and signaling, required for normal vasculogenesis including tube formation and regulation of tube dilation.

Taken together, our study provides important insights into chromosomal organization, exon–intron patterns, indels and genetic variants of CCM3/PDCD10 gene. CCM3/PDCD10 is conserved on same genomic loci for 700 MYA. Gene structural variations are observed in sea urchin and flies. The N-terminal segment of CCM3/PDCD10 is demarked by indels. CCM3/PDCD10 has 951 variants in 1092 human genomes originated from 14 populations. There are 22 missense variants in CCM3/PDCD10 among which three are deleterious in nature. Four terminal codon-gaining mutations were identified in this gene. Upon comparing with roles and corresponding signaling pathways in human to insects, this gene appears to support endothelial developments in invertebrates, a process which is conserved for about 700 MYs.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2014.10.105.

References

[1] P. Labauge, C. Denier, F. Bergametti, et al., Genetics of cavernous angiomas, Lancet Neurol. 6 (2007) 237–244.
[2] E. Tournier-Lasserve, Contribution of molecular genetics in cavernous angiomas, Neurochirurgie 53 (2007) 136–140.
[3] H.D. Craig, M. Gunel, O. Cepeda, et al., Multilocus linkage identifies two new loci for a mendelian form of stroke, cerebral cavernous malformation, at 7p15–13 and 3q25.2–27, Hum. Mol. Genet. 7 (1998) 1851–1858.
[4] J. Dubovsky, J.M. Zabramski, J. Kurth, et al., A gene responsible for cavernous malformations of the brain maps to chromosome 7q, Hum. Mol. Genet. 4 (1995) 453–458.
[5] K.M. Draheim, O.S. Fisher, T.J. Boggon, et al., Cerebral cavernous malformation proteins at a glance, J. Cell Sci. 127 (2014) 701–707.
[6] O.S. Fisher, T.J. Boggon, Signaling pathways and the cerebral cavernous malformations proteins: lessons from structural biology, Cell. Mol. Life Sci.: CMLS 71 (2014) 1881–1892.
[7] X. Li, R. Zhang, H. Zhang, et al., Crystal structure of CCM3, a cerebral cavernous malformation protein critical for vascular integrity, J. Biol. Chem. 285 (2010) 24099–24107.
[8] J. Ding, X. Wang, D.-F. Li, et al., Crystal structure of human programmed cell death 10 complexed with inositol-(1,3,4,5)-tetrakisphosphate: a novel adaptor protein involved in human cerebral cavernous malformation, Biochem. Biophys. Res. Commun. 399 (2010) 587–592.
[9] P. Flicek, I. Ahmed, M.R. Amode, et al., Ensembl 2013, Nucleic Acids Res. 41 (2013) D48–D55.
[10] M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, Nucleic Acids Res. 33 (2005) W465–W467.
[11] D.G. Higgins, J.D. Thompson, T.J. Gibson, Using CLUSTAL for multiple sequence alignments, Methods Enzymol. 266 (1996) 383–402.
[12] J.D. Thompson, T.J. Gibson, F. Plewniak, et al., The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876–4882.
[13] K.B. Nicholas, H.B. Nicholas Jr., D.W.I. Deerfield, GeneDoc: analysis and visualization of genetic variation, EMBNEW.NEWS 4 (14) (1997).
[14] G.R. Abecasis, A. Auton, L.D. Brooks, et al., An integrated map of genetic variation from 1092 human genomes, Nature 491 (2012) 56–65.
[15] P.C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucleic Acids Res. 31 (2003) 3812–3814.
[16] I.A. Adzhubei, S. Schmidt, L. Peshkin, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[17] M. Pappalardo, M.N. Wass, VarMod: modelling the functional effects of non-synonymous variants, Nucleic Acids Res. 42 (2014) W331–336.
[18] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.
[19] K. Tamura, D. Peterson, N. Peterson, et al., MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, Mol. Biol. Evol. 28 (2011) 2731–2739.
[20] A. Kumar, H. Ragg, Ancestry and evolution of a secretory pathway serpin, BMC Evol. Biol. 8 (2008) 250.
[21] A. Kumar, A. Bhandari, Urochordate serpins are classified into six groups encoded by exon-intron structures, microsynteny and bayesian phylogenetic analyses, J. Genomics 2 (2014).
[22] H. Ragg, A. Kumar, K. Koster, et al., Multiple gains of spliceosomal introns in a superfamily of vertebrate protease inhibitor genes, BMC Evol. Biol. 9 (2009) 208.
[23] S.W. Roy, W. Gilbert, The evolution of spliceosomal introns: patterns, puzzles and progress, Nat. Rev. Genet. 7 (2006) 211–221.
[24] B. Verhelst, Y. Van de Peer, P. Rouzé, The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution, Genome Biol. Evol. 5 (2013) 2393–2401.
[25] Y.C.G. Lee, H.-H. Chang, The evolution and functional significance of nested gene structures in *Drosophila melanogaster*, Genome Biol. Evol. 5 (2013) 1978–1985.
[26] T. Zhu, D.K. Niu, Mechanisms of intron loss and gain in the fission yeast *Schizosaccharomyces*, PLoS One 8 (2013) e61683.
[27] P. Yenerall, L. Zhou, Identifying the mechanisms of intron gain: progress and trends, Biol. Direct 7 (2012) 29.
[28] P. Yenerall, B. Krupa, L. Zhou, Mechanisms of intron gain and loss in *Drosophila*, BMC Evol. Biol. 11 (2011) 364.
[29] S.W. Roy, M. Irimia, Mystery of intron gain: new data and new models, Trends Genet. 25 (2009) 67–73.
[30] A. Kumar, Delving into vertebrate serpins for understanding their evolution, Nature Precedings (2009) 1–5.
[31] A. Kumar, A. Bhandari, S.J. Sarde, et al., Sequence, phylogenetic and variant analyses of antithrombin III, Biochem. Biophys. Res. Commun. 440 (2013) 714–724.
[32] A. Kumar, A. Bhandari, S.J. Sarde, et al., Molecular phylogeny of C1 inhibitor depicts two immunoglobulin-like domains fusion in fishes and ray-finned fishes specific intron insertion after separation from zebrafish, Biochem. Biophys. Res. Commun. 450 (2014) 219–226.
[33] A. Kumar, A. Bhandari, S.J. Sarde, et al., Genetic variants and evolutionary analyses of heparin cofactor II, Immunobiology 219 (2014) 713–728.
[34] A. Kumar, S.J. Sarde, A. Bhandari, Revising angiotensinogen from phylogenetic and genetic variants perspectives, Biochem. Biophys. Res. Commun. 446 (2014) 504–518.
[35] A. Kumar, A. Bhandari, R. Sinha, et al., Spliceosomal intron insertions in genome compacted ray-finned fishes as evident from phylogeny of MC receptors, also supported by a few other GPCRs, PLoS One 6 (2011). e22046–e22046.
[36] A. Kumar, K. Kollath-Leiss, F. Kempken, Characterization of bud emergence 46 (BEM46) protein: sequence, structural, phylogenetic and subcellular localization analyses, Biochem. Biophys. Res. Commun. 438 (2013) 526–532.
[37] E. Bapteste, H. Philippe, The potential value of indels as phylogenetic markers: position of trichomonads as a case study, Mol. Biol. Evol. 19 (2002) 972–977.
[38] W. Schröder, J. Najm, S. Spiegler, et al., Predictive genetic testing of at-risk relatives requires analysis of all CCM genes after identification of an unclassified CCM1 variant in an individual affected with cerebral cavernous malformations, Neurosurg. Rev. 37 (2014) 161–165.
[39] R. Munoz-Chapuli, R. Carmona, J.A. Guadix, et al., The origin of the endothelial cells: an evo-devo approach for the invertebrate/vertebrate transition of the circulatory system, Evol. Dev. 7 (2005) 351–358.
[40] R. Monahan-Earley, A.M. Dvorak, W.C. Aird, Evolutionary origins of the blood vascular system and endothelium, J. Thromb. Haemost. 11 (Suppl 1) (2013) 46–66.

[41] P.C. Moller, C.W. Philpott, The circulatory system of amphioxus (*Branchiostoma floridae*). I. Morphology of the major vessels of the pharyngeal area, J. Morphol. 139 (1973) 389–406.
[42] P.C. Moller, C.W. Philpott, The circulatory system of amphioxus (*Branchiostoma floridae*). II. Uptake of exogenous proteins by endothelial cells, Z. Zellforsch. Mikrosk. Anat. 143 (1973) 135–141.
[43] M.A. Choma, M.J. Suter, B.J. Vakoc, et al., Physiological homology between *Drosophila melanogaster* and vertebrate cardiovascular systems, Dis. Model Mech. 4 (2011) 411–420.
[44] Y. Song, M. Eng, A.S. Ghabrial, Focal defects in single-celled tubes mutant for cerebral cavernous malformation 3, GCKIII, or NSF2, Dev. Cell 25 (2013) 507–519.