Università degli studi di Modena e Reggio Emilia

Dipartimento di Scienze della Vita

Corso di laurea Magistrale
Biologia sperimentale ed applicata

Analisi di trascrittomi tessuto-specifici di *Pomacea canaliculata* tramite sequenziamento Nanopore: Ipervariabilità ed approccio AI

Relatore
Prof. Nicola Franchi

Tesi di laurea di
Filippo Bertolasi

Anno Accademico 2023/2024

Indice

Riassunto

Il sistema del complemento è un componente fondamentale della difesa immunitaria sia per vertebrati che per invertebrati. Indagare circa le sue origini ed evoluzione è un passaggio chiave per comprendere come sono cambiate le sue proteine in modo da adattarsi a nuove condizioni ambientali, cicli vitali o, nel caso dei vertebrati, all'interazione con l'immunità adattativa. Le proteine responsabili della sua regolazione vengono definite Regolatori del Complemento, sono sempre più studiate perché è dimostrato il loro coinvolgimento in numerosi processi fisiologici e patologici. In *Pomacea canaliculata* queste indagini possono essere utili anche per comprendere un eventuale coinvolgimento del sistema del complemento nei processi definiti "non canonici" come la rigenerazione.

Questo studio ha dimostrato che l'espressione dei regolatori del complemento in *P. canaliculata* è tessuto-specifica, almeno per quanto riguarda emociti e rene posteriore. È stato inoltre dimostrato quanto l'ipervariabilità di questi sia determinata da fenomeni di splicing alternativo. È stato poi con successo applicato il modello AI ProstT5 permettendo la creazione di due dataset contenenti sequenze caratterizzate da un simile significato biologico a delle già note proteine regolatrici del complemento di *P. canaliculata*. L'analisi dei dati ottenuti ha individuato che la maggior parte delle sequenze non è identificabile e non presenta domini noti ai database SMART, gettando le fondamenta per future analisi circa potenziali proteine regolatrici del complemento con struttura diversa dalle convenzionali presentanti CCP.

1. Introduzione

Tutti i metazoi hanno evoluto dei meccanismi cellulari ed umorali per difendersi dal "non-self". Storicamente, il sistema immunitario è stato suddiviso in due: immunità innata ed immunità adattativa. L'immunità innata fornisce una difesa aspecifica e rappresenta la prima forma di difesa che andrà ad agire contro l'organismo patogeno e/o qualsiasi altra molecola riconosciuta come non appartenente all'ospite, al "self". L'immunità adattativa invece rappresenta una seconda linea di difesa, specifica per i singoli organismi/molecole antigeniche (Alberts et al., 2009).

1.1 Immunità innata

L'immunità innata viene detta anche naturale, il nome deriva dal fatto che fornisce protezione da agenti patogeni senza necessità di una prima esposizione ad essi. In altre parole, quando il sistema immunitario innato incontra un antigene non-self, reagirà istantaneamente per isolarlo o rimuoverlo dall'ospite (Kimbrell, 2001). Questa funzione è molto importante dato che la risposta immunitaria adattativa e specifica richiede anche una settimana per diventare efficace. In generale ci sono tre linee di difesa immunitaria innata che possono agire: la prima è rappresentata dalle barriere fisiche e chimiche che impediscono il semplice ingresso di microrganismi all'interno del corpo. Queste barriere comprendono lo strato di tegumento esterno, le giunzioni strette tra cellule epiteliali, il pH acido nello stomaco e i componenti degli strati di muco che inibiscono la colonizzazione ed uccidono i batteri estranei. La flora simbionte ha un ruolo anche nel proteggere le superfici corporee dagli invasori, competendo per la stessa nicchia ecologica e limitando così la colonizzazione. La seconda linea di difese innate comprende le risposte intrinseche della cellula, tramite cui una singola cellula può riconoscere una cellula bersaglio e rispondere di conseguenza con misure atte alla fagocitosi o all'attività citotossica. La maggior parte delle cellule che hanno internalizzato un batterio tramite fagocitosi indotta da patogeni, per esempio, dirigerà immediatamente il fagosoma a fondersi con un lisosoma, esponendo il microrganismo invasore agli enzimi digestivi. Un altro meccanismo di difesa intrinseco è la capacità di cellule ospiti di degradare l'RNA a filamento doppio, che è un intermediario comune nella replicazione virale; le cellule infette degraderanno anche qualsiasi RNA a singolo filamento che ha innescato il meccanismo. La terza linea di difese immunitarie dipende da una serie specializzata di proteine e cellule ad attività immunitaria (fagociti e cellule citotossiche) che riconoscono le caratteristiche conservate dei patogeni e si attivano velocemente per aiutare

a distruggere gli invasori. Tra queste troviamo le cellule specializzate degli invertebrati e neutrofili, macrofagi e cellule natural-killer dei Vertebrati oltra a sistemi molecolari complessi come il sistema del complemento (Alberts et al., 2009).

1.1.1 Immunità cellulo-mediata

Il riconoscimento di pattern molecolari conservati dei patogeni causa l'attivazione, nei Vertebrati, dei leucociti: granulociti neutrofili, macrofagi, monociti, cellule dendritiche. Il loro ruolo principale è quello di immunità cellulo-mediata che include: fagocitosi, lisi della cellula bersaglio, ed immunomodulazione. Per l'attivazione di queste cellule vengono riconosciuti i PAMP (Pathogen Associated Molecular Patterns), si tratta di pattern molecolari, costituiti ad esempio da zuccheri sulla superficie del microrganismo, che durante l'evoluzione si sono conservati e quindi permettono all'organismo di individuare la presenza di un patogeno e generare una risposta aspecifica, ma selettiva (Alberts et al., 2009).

1.1.2 Immunità umorale

La componente umorale della risposta immunitaria innata comprende: il sistema del complemento, il CAS (Contact Activation System) e le lectine. Il CAS è una cascata di proteasi plasmatiche che legano la superficie batterica e promuovono coagulazione e infiammazione, molti componenti di questo sistema inoltre promuovono il taglio proteolitico di C3 e C5, questo suggerisce un certo livello di collaborazione tra i due sistemi (Shishido, 2012). Molte lectine fanno parte di famiglie di PRR (Pattern Recognition Receptors) che riconoscono specifici zuccheri sulla superficie delle cellule e quindi diversi patogeni e ligandi intrinseci come cellule apoptotiche e componenti della matrice extracellulare (Bassi, 2009. Hirschfield, 2003). Queste proteine vengono riconosciute dai macrofagi ed altre cellule dell'immunità innata facilitando la clearance dei patogeni e dei detriti cellulari tramite fagocitosi (Shishido, 2012).

1.2 Sistema del complemento

Il sistema del complemento, pur essendo parte dell'immunità umorale innata, ha un ruolo chiave sia nella risposta immunitaria innata che in quella adattativa (Kohl et al., 2006). É formato da più di 50 glicoproteine solubili o di membrana che sono coinvolte in numerose interazioni proteina-proteina, causando l'assemblaggio ed attivazione di complessi

enzimatici e la generazione di frammenti bioattivi che iniziano svariate risposte cellulari mediante il legame di recettori del complemento e regolatori. (Ricklin et al., 2010). Il sistema del complemento contribuisce a numerose funzioni come fagocitosi, lisi cellulare, infiammazione, solubilizzazione degli immunocomplessi, rimozione delle cellule apoptotiche e stimolazione della risposta immune umorale (Shmidt, 2000).

Nei vertebrati sono conosciute 3 vie di attivazione: la classica, la via lectinica e la via alternativa. Tutte comprendono al loro interno una molecola capace di riconoscere e legare la superficie microbica ed una serina proteasi, che quando attivata, esegue il taglio della proteina C3 in C3a e C3b. C3 rappresenta l'elemento fondamentale del sistema del complemento dato che ha la capacità di esercitare un ruolo nell'opsonizzazione mediante legame di C3b sulla superficie microbica, ma può anche reclutare gli immunociti sul sito d'infezione tramite rilascio di C3a ed eventualmente portare alla lisi cellulare tramite formazione del Complesso di Attacco alla Membrana (MAC) (Song et al., 2000).

Nella via alternativa C3b richiede l'interazione con il fattore B (Bf), una proteasi, per formare la C3 convertasi che promuove la formazione di C3a e C3b. Il legame di C3b alla convertasi attivata causa un cambiamento nell'affinità per il substrato del complesso enzimatico che esegue, nei soli Vertebrati, il taglio di C5 in C5a e C5b che avrà un ruolo iniziale nell'assemblaggio del MAC. La via lectinica inizia con il riconoscimento di zuccheri su superfici cellulari da parte di lectine come le MBLs (Mannose-Binding Lectins), che portano all'attivazione di una serina proteasi associata alle MBLs (MASPs) la quale eseguirà il taglio di C3 in C3a e C3b (Wallis, 2007). La via classica, a differenza delle precedenti, inizia con lo specifico riconoscimento microbico da parte degli anticorpi che interagiranno con C1q che è associato a due serin-proteasi (C1r e C1s), questa via è esclusiva dei Vertebrati (Walport, 2001).

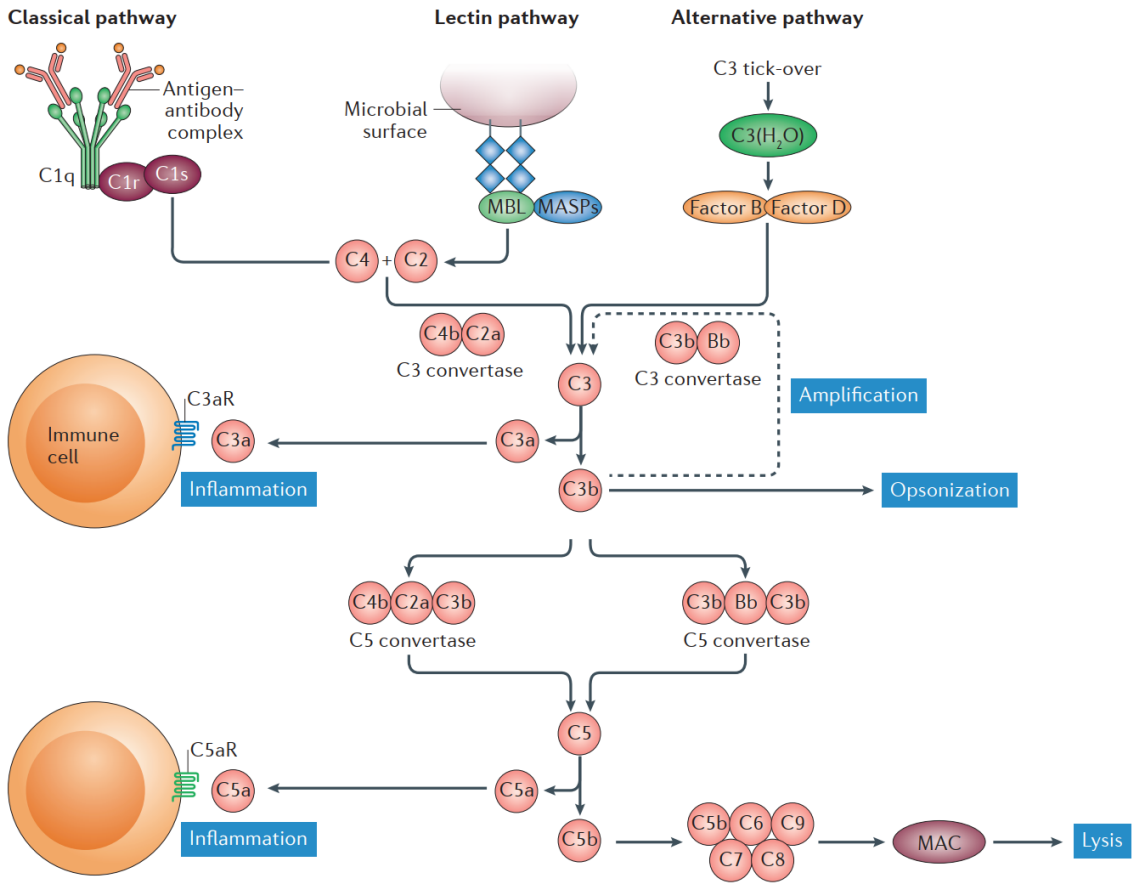


Figura 1 *Sistema del complemento. Modificata da: Trouw, 2017.*

L'attivazione del Sistema del Complemento può avvenire anche in maniera non canonica indipendente dalla convertasi. Proteasi del sistema cinina-callicreina, o delle cascate coagulativa e fibrinolitica, come plasmina e trombina, possono portare al clivaggio di C3, C5 o del fattore B, generando frammenti che possono assemblare una C3 convertasi funzionante o partecipare a risposte effettrici downstream (Irmscher et al., 2018. Amara et al., 2010) . Queste vie convertasi-indipendenti sono coinvolte in meccanismi di amplificazione dell'attivazione del complemento o agiscono separatamente dalle normali vie. È necessario specificare che la loro rilevanza fisiologica è dipendente dal contesto e dovrebbe essere considerata in relazione alle specifiche patologie (Nilsson et al., 2021).

1.2.1 Funzioni non immunitarie del complemento

Con l'avanzare della ricerca sono state evidenziate molte funzioni definite anche "non canoniche" per il sistema del Complemento tra le quali: controllo nella morfogenesi dei tessuti, riparazione delle ferite e pruning sinaptico. Come costante, la presenza di molecole

associate al complemento garantisce la prevenzione della risposta autoimmune, ad esempio, in spermatozoi e durante lo stadio di blastocisti. In particolare, negli spermatozoi, alcune molecole del Sistema del Complemento aiutano l'interazione oocita-spermatozoo e facilitano la fecondazione (Harris et al., 2006. Rooney et al., 1996), allo stadio di blastocisti contribuiscono allo sviluppo della massa cellulare interna e del trofoblasto (Taylor et al., 1996), durante l'organogenesi (più specificatamente durante la neurulazione) facilitano la chiusura del tubo neurale e contribuiscono alla migrazione delle cellule delle creste neurali (Danny et al., 2016), inoltre partecipano al pruning sinaptico ossia il rimodellamento di alcune connessioni sinaptiche per la maturazione dei circuiti neurali (Stevens et al., 2007) mentre nella formazione della placenta sembrano avere un compito di prevenzione contro eventuali patogeni (Stevens et al., 2007. Faulk et al., 1980.).

È Stato inoltre visto che C1q ed il Fattore H hanno ruoli nella proliferazione di cellule tumorali, immunosoppressione in ambito di microambiente tumorale, angiogenesi e progressione neoplastica. L'attività del complemento a livello intracellulare è inoltre stata collegata alla regolazione dell'autofagia e della secrezione dell'insulina a livello delle isole pancreatiche (King et al., 2023).

A livello di organismo adulto le proteine del complemento sono coinvolte in processi vitali delle cellule andando ad influenzarne proliferazione, differenziazione e collocazione (Hawksworth et al., 2018).

Sono sempre più numerose le prove che sottolineano come le proteine del complemento siano profondamente coinvolte nella regolazione di processi biologici che sono molto distanti dall'immunosorveglianza e dalla difesa dell'organismo (Mastellos et al., 2013. Stephan et al., 2012).

1.2.2 Sistema del Complemento negli invertebrati

La prima indagine circa il sistema del complemento a livello di invertebrati aveva portato all'ipotesi che potesse esservi un complemento primitivo o archeo-complemento auto-attivato dal "tick-over" di C3 o dal taglio proteolitico da parte di proteasi di patogeni (Lachmann, 1979). Questo è stato poi confermato da studi seguenti che hanno portato ad individuare una sequenza parziale C3-simile (Smith et al., 1996), seguita dal rispettivo cDNA completo (Al-Sharif et al., 1998) la cui espressione è stata localizzata in cellule correlate all'immunità: i celomociti di *Strongylocentrotus purpuratus* (Clow et al., 2000).

Dalle prime individuazioni di analoghi di C3 in invertebrati numerosi studi hanno poi portato ad identificarne la presenza nei principali phyla (Peng et al., 2016).

A seguito venne individuato anche un omologo del Fattore B (Smith et al., 1996) ricostruendone la relativa sequenza di cDNA completa, espressa anch'essa in celomociti (Terwilliger et al., 2004). Col progredire degli studi l'idea di archeo-complemento si è rivelata corretta ma il concetto è stato modificato ed espanso, aggiungendo una proposta di ulteriore via la cui attivazione è dipendente da lectine. Questo inoltre presenta un feedback d'amplificazione del segnale tipico della via alternativa corredato dalle proteine chiave della via come le proteasi che portano all'attivazione del Fattore B e aumentano l'attivazione di C3 (Smith et al., 1999, 2001). Risultano assenti invece le proteine C4 e C5 perché frutto di una duplicazione genica avvenuta nel clade dei Vertebrati (Franchi, non pubblicato).

1.2.3 Regolatori del complemento

C3 nella sua forma attiva (C3b) espone il sito tioestere che è in grado di legare covalentemente gruppi idrossilici ed amminici, potenzialmente presenti sulle cellule self, ed attivare la cascata del complemento. Perciò, a livello di vertebrati, le proteine regolatrici del complemento sono presenti sulla superficie cellulare e nel liquido extracellulare per proteggere l'organismo da un attacco autoimmune da parte della cascata del Complemento (Merle et al., 2015a). Tutte le proteine codificate dal cluster genico dei regolatori del complemento (RCA) in genomi di mammifero presentano una notevole somiglianza in struttura e composizione. Sono principalmente costituite da domini denominati Short Consensus Repeats (SCRs) o domini CCP (Complement Control Proteins) i quali presentano una struttura tridimensionale fortemente conservata tra loro (Krych-Goldberg and Atkinson, 2001). Sono varie le funzioni attribuite a differenti domini CCP di ogni proteina regolatrice, spaziano da attività inibitorie a legame con superfici cellulari o ad altre componenti del complemento. Questa organizzazione modulare dei geni permette di codificare per proteine ibride con diverse serie di domini CCP assolvendo a differenti funzioni (Pouw et al., 2015).

Informazioni circa i regolatori del complemento a livello degli invertebrati sono scarse. La ricerca di queste proteine nei non mammiferi ed invertebrati è particolarmente complessa a causa della modularità dei geni che porta ad una semplice riorganizzazione dei domini CCP e a causa della similarità nella struttura proteica che portano a difficoltà nel predire parentele con sequenze di mammifero. Questo potrebbe essere dovuto alla variabilità della sequenza insorta nel corso dell'evoluzione originatasi sia tramite: mutazioni puntiformi, duplicazioni

geniche, delezioni e/o riorganizzazione di domini ed esoni, in aggiunta a varianti di splicing portando a geni che codificano per proteine contenenti domini CCP multipli. Un'ulteriore problematica, che vale anche per tutte le altre proteine, è rappresentata dal fatto che non è possibile individuare similarità funzionali tra le proteine regolatrici del complemento tramite discendenza evolutiva o similarità di sequenza. I domini CCP potrebbero variare di numero all'interno e tra le diverse specie, e possono includere differenze dovute a splicing alternativo (Smith et al., 2023).

Ad oggi, nei Vertebrati, sono state individuate una serie di proteine con comprovata azione di regolazione del complemento: Decay Accelerating Factor (DAF), Membrane Cofactor Protein (MCP), Fattore H (FH) e C4b-Binding Protein (C4BP). Queste hanno una forte influenza sulla regolazione della localizzazione e deposizione del complemento (Medzhitov and Janeway, 2002). Data però la probabile ipervariabilità di questi trascritti, anche nei Vertebrati, vengono costantemente scoperte sempre più proteine con azione regolatrice del complemento caratterizzate dal pattern di domini CCP, tra cui CUB and SUSHI Multiple Domains 1 (CSMD1) (Escudero-Esparza et al., 2013; Håvik et al., 2011; Kraus et al., 2006). Per questa proteina, in particolare, è stato recentemente dimostrato un ruolo diretto nella gestione dei processi di eliminazione sinaptica durante lo sviluppo, un dato significativo verso la conclamata funzione del complemento in ambiti non immunitari (Baum et al., 2024).

1.3 *Pomacea canaliculata*

Pomacea canaliculata (Pc) è un mollusco gasteropode d'acqua dolce interessante da più punti di vista, soprattutto per la sua grande resistenza a condizioni di stress e spiccate capacità rigenerative nell'organismo adulto (Bever et al., 1988. Accorsi et al., 2017). L'elevata fecondità e capacità di adattamento di *Pc* ad un'ampia varietà di ambienti la rende una delle 100 specie più invasive secondo l'ISSG (Invasive Species Specialist Group) (View 100 of the world's worst invasive alien species, 2000). *Pc* è anche un ospite intermedio del nematode parassita umano *Angiostrongylus cantonensis* perciò è molto studiata per la sua eradicazione (Yang et al., 2013). Il sistema immunitario di *Pc* è composto da una componente cellulare, gli emociti, ed una componente umorale (Smith et al., 2016). Attualmente esistono informazioni circa la morfologia (Accorsi et al., 2013. Cueto, 2015) e la proliferazione (Accorsi, 2014) degli emociti ma sono pochi i dati molecolari.

2. Scopo della tesi

Numerosi studi hanno provato l'importanza del sistema del complemento ed i suoi regolatori in processi fisiologici immunitari e non, puntualizzando anche quanto alterazioni dei regolatori possano portare a condizioni patologiche. Con il seguente lavoro l'obbiettivo è quello di valutare l'ipotesi che l'ipervariabilità dei regolatori del complemento sia dovuta a splicing alternativo. Disponendo poi delle strutture 3D delle proteine note aventi domini CCP e con sussidio di un'intelligenza artificiale (AI) cercare di raggruppare le proteine dedotte dai trascrittomi degli emociti e del rene posteriore di *P. canaliculata* secondo la loro struttura tridimensionale costituendo così un dataset da poter indagare per individuare nuovi regolatori potenzialmente composti anche da domini non convenzionali.

3. Risultati

3.1 Ipervariablità da splicing alternativi

Sono state individuate all'interno dei trascrittomi di *P. canaliculata* ottenuti con sequenziamento Nanopore 11 sequenze putative per Regolatori del Complemento contenenti CCP, 3 di queste appartenenti al trascrittoma di emociti e le restanti 8 del rene posteriore (Figura 2) dimostrando che i Regolatori in *Pomacea*, almeno per i tessuti presi in esame, risultano essere tessuto-specifici.

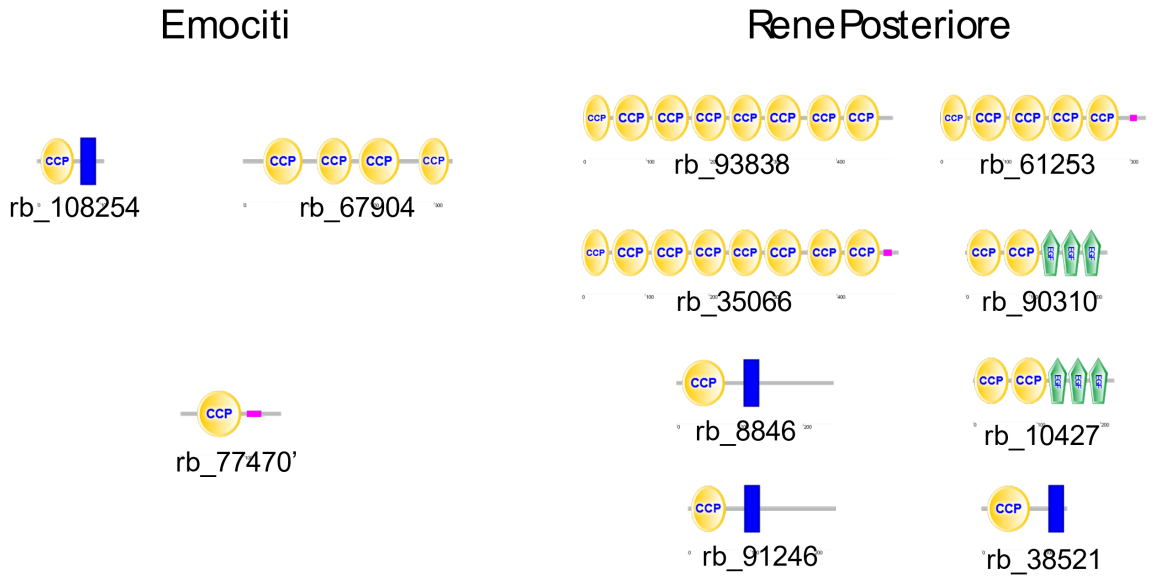


Figura 2 Sequenze putative per Regolatori del Complemento individuate dall'analisi BLAST sui rispettivi trascrittomi.

Allineando queste sequenze con il genoma di *P. canaliculata* depositato in GeneBank sono riuscito ad individuare due geni specifici all'interno delle regioni *SZHN201 Linkage group LG10 ASM307304v1* e *SZHN201 Linkage group LG2 ASM307304v1*. I trascritti di interesse dal rene posteriore risultano generati da splicing alternativo da due geni: PcPKRCA1 e PcPKRCA2. In particolare, sono state individuate 5 varianti di splicing dal gene PcPKRCA1 e 3 varianti di splicing dal gene PcPKRCA2 (Figura 3). L'analisi dei singoli domini CCP delle diverse varianti di splicing conferma questa constatazione (Allegato 1).

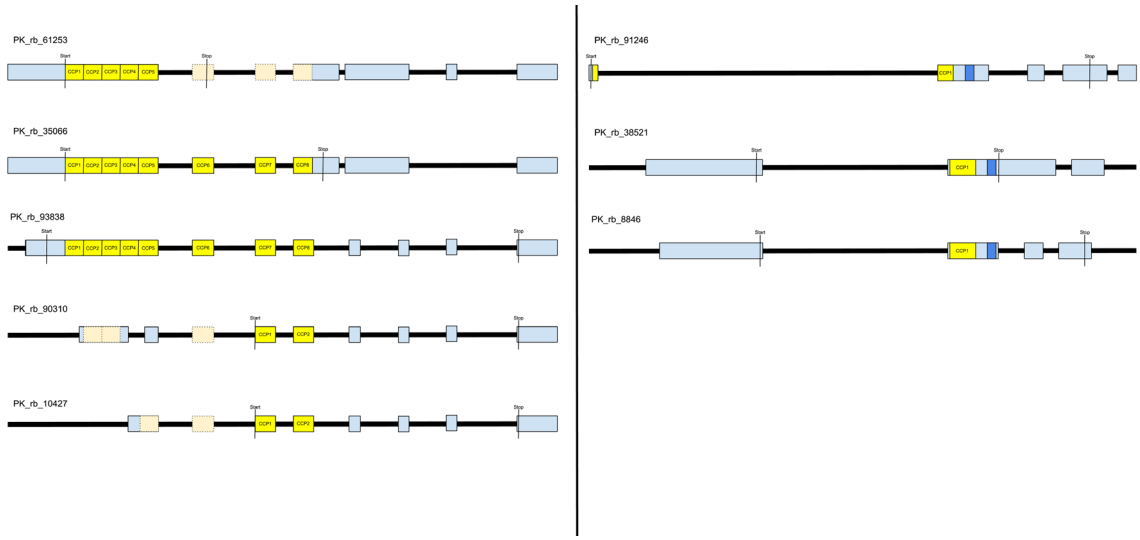


Figura 3 Rappresentazione grafica dei geni e splicing.

Nel caso del trascrittoma di emociti non è stato possibile indagare in modo specifico la presenza di varianti di splicing poiché è stata individuata solo una regione genomica compatibile con uno dei tre trascritti: *SZHN2017 linkage group LG2, ASM307304v1* (Figura 4). Gli altri due trascritti non risultano come splicing alternativo di questa regione genomica.

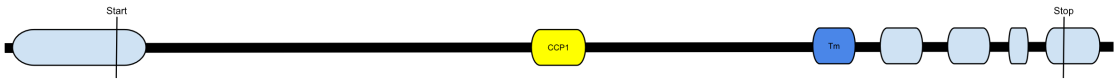


Figura 4 Unica sequenza individuata dal confronto con i database nel trascrittoma di Emociti

3.2 Analisi 3D da algoritmo AI

L'output del modello ProstT5 è stato rappresentato, dopo accurata riduzione dimensionale con metodo Umap, in un grafico bidimensionale a dispersione dove ogni punto è assegnato ad un cluster e la maggior vicinanza tra punti indica una maggiore similarità in termini di significato biologico (Figura 5). Ho quindi identificato il cluster di appartenenza delle sequenze con CCP note individuandole nei cluster 407 e 362, questi risultano vicini tra loro garantendoci un maggior grado di sicurezza nel proseguire con l'analisi.

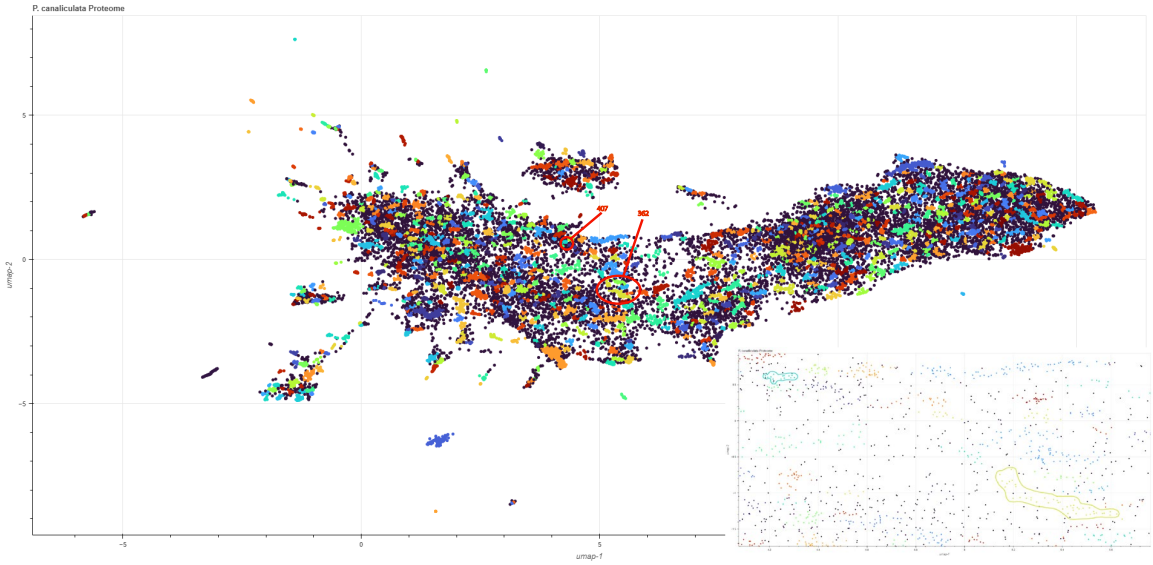


Figura 5 rappresentazione 2D dei dati ricavati tramite approccio AI. Evidenziati i cluster 407 e 362 contenenti le sequenze CCP note.

E' stata quindi eseguita un analisi Blast per verificare l'identità delle sequenze contenute nei due cluster, seguita poi da un'analisi SMART per determinarne l'organizzazione in termini di domini proteici. Questo ci ha portato a costituire un dataset con 14 sequenze appartenenti al cluster 407 e 36 appartenenti al cluster 362 (Allegato 2). Osservando i dati è stato individuato che, la maggior parte delle sequenze non restituisce alcun dominio noto dall'analisi SMART e per una buona parte di queste non è stata possibile l'identificazione con Blast perché non sono presenti sequenze compatibili all'interno dei database (Tabella 1).

Tabella 1 **Analisi del dataset**

	Cluster 407	Cluster 362
Sequenze identificate	14	36
"Uncharacterized"	4 (28.57%)	16 (44.44%)
No domini noti con SMART	7 (50%)	27 (75%)

Avendo ottenuto un riscontro positivo dal primo approccio con il modello AI, sono state inserite nel trascrittoma delle sequenze umane notoriamente appartenenti ai Regolatori del Complemento (CFAH, CD55, TNF) che potessero funzionare da controllo interno. Questo ci ha permesso di costituire un secondo dataset estremamente più abbondante (Allegato 3). Analizzando i dati abbiamo individuato che le sequenze note con CCP sono localizzate nei cluster 231 e 232, estremamente vicine tra di loro, mentre quelle umane in 207, 308 e 374.

4 Discussione

I regolatori del complemento presentano la caratteristica conservata di essere costituiti da ripetizioni di domini CCP e la diversa disposizione di questi ultimi determina alterazioni funzionali (Pouw et al., 2015). È nota la difficolta nell'analisi di sequenze che codificano per regolatori del complemento soprattutto per il fatto che la tecnologia di sequenziamento più utilizzata ad oggi è basata sulla piattaforma Illumina che genera delle reads molto brevi che necessitano di essere poi assemblate sul genoma per formare cDNA completi. Questa metodica risulta inefficace per valutare sequenze ipervariabili come appunto le varianti di splicing (Smith et al.,2023). Il nostro approccio è stato quindi quello di impiegare la tecnologia di sequenziamento Nanopore (*MinION mk1*) che permette di generare reads molto più lunghe che possano essere assemblate facilmente senza l'ausilio di sequenze genomiche di riferimento. In questo modo emergono chiaramente i reali trascritti e quindi tutte le possibili varianti dei diversi geni. Con questo approccio sono stati costituiti due trascrittomi: il primo da rene posteriore ed il secondo da emociti derivati dallo stesso individuo. Abbiamo effettuato un'analisi Blast utilizzando come chiave di ricerca delle sequenze da *P. canaliculata* contenenti domini CCP e precedentemente individuate nei database pubblici, questo ci ha portato a identificare una serie di sequenze con un ottimo grado di similarità. Abbiamo proseguito analizzando le sequenze restituite dall'analisi per individuare quali potessero codificare per proteine contenenti domini CCP. Per questo sono stati valutati due approcci differenti: il primo metodo si serviva del software InterPro Scan, uno strumento bioinformatico capace di identificare domini proteici, famiglie e caratteristiche funzionali partendo da sequenze nucleotidiche o amminoacidiche tramite confronto della sequenza indagata con vari database di famiglie proteiche e modelli di domini. Questo metodo si è rilevato estremamente esoso in termini di tempo e risorse computazionali e non ha restituito dei buoni risultati. Siamo quindi passati al secondo approccio tramite SMART, un database e strumento bioinformatico in grado di identificare domini proteici e motivi modulari maggiormente incentrato verso la comprensione funzionale delle proteine prese in analisi, questo ha restituito 11 sequenze con caratteristiche tipiche dei Regolatori del Complemento (Figura 2) e si è rivelato il metodo migliore.

Numerosi studi hanno evidenziato la presenza di fenomeni di splicing alternativo in alcune delle principali proteine regolatrici del complemento a livello di vertebrati (Post et al., 1991, Russel et al., 1992, Mannes et al., 2020), ma sono poche ad oggi le informazioni circa questo meccanismo a livello degli invertebrati. Ottenute quindi le sequenze abbiamo mappato

queste sul genoma pubblicato in NCBI di *Pomacea canaliculata* descrivendo la composizione in introni ed esoni dei singoli geni. Quello che è emerso immediatamente è che, almeno per i tessuti presi in esame in questo studio, i Regolatori del Complemento di *P. canalicuta* risultano essere tessuto-specifici: 3 trascritti unici negli emociti ed 8 nel rene posteriore. Per quanto riguarda il rene posteriore le sequenze individuate sono risultate appartenere a due geni distinti: PcPKRCA1 e PcPKRCA2. Analizzando le sequenze e creando delle rappresentazioni grafiche di questi trascritti abbiamo evidenziato come questi siano verosimilmente originati da splicing alternativo (Figura 2). A conferma di questo abbiamo generato un albero che raggruppa le sequenze dei diversi CCP sulla base della loro similarità evidenziando come gli stessi CCP si possano trovare in alcune delle varianti trascrizionali sottolineandone la provenienza dal medesimo esone (Allegato 1).

Fino ad oggi, per condurre ricerche di sequenze specifiche su invertebrati, si è sempre utilizzato l'approccio dell'omologia o dissimilarità di sequenza con trascritti dai Vertebrati. Per quanto questo metodo si sia rivelato esaustivo in molti settori rimane sempre vincolato alla ricerca di qualcosa che deve essersi conservato nel corso dell'evoluzione appunto fino ai Vertebrati ed escludendo la possibilità di trovare qualcosa di nuovo. Oggi si apre una nuova frontiera di calcolo computazionale con algoritmi che possono andare molto oltre il semplice confronto di sequenze lineari. Questi nuovi approcci si basano sulla capacità di *machine learning* ed Intelligenza Artificiale (AI).

L'intelligenza artificiale è diventata uno strumento essenziale per l'elaborazione di grandi quantità di dati. Grazie al machine learning, le AI sono capaci di analizzare in modo molto rapido dataset complessi portando all'individuazione di pattern ed informazioni che difficilmente potrebbero essere colte in altro modo (Russell, 2021). L'adattamento di queste tecniche in ambito proteomico tramite sostituzione delle parole con amminoacidi e frasi con sequenze proteiche ha aperto la strada a nuovi e potenti strumenti bioinformatici per la modellizzazione di sequenze proteiche detti *protein Language Models* (pLMs) (Heinzinger et al., 2019; Chen et al., 2023). Il principale, AlphaFold2, ha reso disponibili le previsioni delle strutture 3D di più di 200 milioni di proteine al pubblico. Questo grande database ha consentito di allenare un metodo denominato ProstT5 che, attribuendo un vettore numerico di dimensione 512/1024/2048 ad ogni amminoacido, riesce ad attribuirvi un significato biologico e ricostruire la struttura 3D della sequenza indagata (Heinzinger et al., 2023). Sono stati quindi uniti ed analizzati i due trascrittomi con il pLM ProstT5. Sono stati ottenuti una serie di vettori ciascuno ascrivibile ad una diversa proteina. Sono poi stati rappresentati

graficamente sfruttando la tecnica Umap (*Uniform Manifold Approximation and Projection for Dimension Reduction*) che ha permesso di ridurre l'informazione in un piano bidimensionale dato che il vettore costituisce un dato di tante dimensioni quanti sono i valori di cui è formato. È stato ottenuto quindi un grafico a dispersione dove i punti rappresentano le singole proteine e la loro vicinanza indica un simile significato biologico (Figura 5). Questo approccio porta ad una perdita sostanziosa di informazioni data la riduzione di dimensionalità, ma rappresenta comunque il metodo più idoneo in quanto non è possibile impiegare la PCA (Principal Component Analysis) visto che disponiamo di vettori con molti caratteri. Il metodo ci ha consentito di individuare due cluster dove ricadono le sequenze note contenenti CCP. Il contenuto è stato analizzato prima con Blast per stabilire l'identità delle sequenze e poi con SMART ottenendo un risultato estremamente interessante poiché vengono raggruppate insieme principalmente proteine con risaputa struttura da Regolatore del Complemento, ma anche proteine con predizione di similarità ignota o proteine in cui non si riescono a riconoscere domini proteici. Non va dimenticato che il riconoscimento di un dominio specifico è possibile solo nel caso in cui quel dominio sia stato registrato e gli sia stata data una sequenza consenso nei diversi database. L'assenza di dominio, quindi, può significare una porzione di proteina di collegamento, senza funzione specifica o la presenza di una proteina la cui funzione è effettivamente ignota. È quest'ultima ipotesi che rende interessante il nostro dato allargando la possibile indagine non solo a proteine con domini CCP, ma anche ad altre sequenze. Chiaramente solo studi futuri potranno far luce se queste proteine ignote e se queste abbiano effettivamente un ruolo nella regolazione del Complemento, ma risulta quanto mai interessante riuscire a poter ipotizzare una funzione per gruppi di proteine che non hanno alcuna collocazione funzionale nei database.
In seguito, a scopo di controllo, è stata eseguita nuovamente l'analisi utilizzando lo stesso trascrittoma ma con l'aggiunta di sequenze umane che fanno riferimento a regolatori del complemento noti (CFAH, CD55, TNF) in modo da valutare come venisse influenzata la clusterizzazione. Analizzando l'output è stato visto che le sequenze di *pomacea* hanno mantenuto la clusterizzazione indipendente da quelle umane, questo è con molta probabilità dovuto alla marcata differenza delle proteine tra invertebrati e vertebrati. Inoltre, andando ad analizzare il contenuto dei singoli cluster si riscontrano parecchie somiglianze strutturali tra le sequenze comprese nei singoli raggruppamenti a rafforzare l'efficacia del metodo d'indagine. Questa seconda analisi ha in ogni caso permesso di ottenere un altro abbondante dataset, contenente molte sequenze "*Uncharacterized"* e con previsione dei domini ignota (Allegato 3), che si presta ad ulteriori e più approfondite future analisi.

5 Conclusioni

Lo studio ha individuato 11 sequenze che potrebbero essere dei regolatori del complemento in *P. canaliculata. Questi*, almeno per i tessuti presi in analisi, hanno un'espressione tessuto-specifica e che la loro ipervariabilità è definita da fenomeni di splicing alternativo a partire da, almeno, 3 geni distinti.

Lo studio con AI, tramite applicazione del pLM ProstT5, ha portato alla generazione di due dataset molto popolati. Questi sono un importante risultato nell'ambito dello studio di regolatori del complemento in invertebrati data la scarsità di informazioni e dati a riguardo. Si tratta quindi di uno studio preliminare che ha gettato le basi per l'indagine dei regolatori del complemento in *P. canaliculata*.

6 Materiali e metodi

6.1 Animali

Le specie di *P.canaliculata* sono state acquistate da T.A.F. Trans Aquarium Fish S.r.l. (Torino, Italia) e mantenute in acquari riempiti d'acqua areata a 23.1°C. Organismi adulti ed attivi nella riproduzione (diametro della conchiglia 35-55 mm) sono stati mantenuti in un acquario separato. Gli animali sono stati nutriti *ad libitum* con verdure e frutta. Il 10-20% del contenuto dell'acquario è stato regolarmente sostituito. Il benessere degli animali è stato accertato dalla continua valutazione della riproduzione in acquario. Per i prelievi di campioni gli animali sono stati lasciati in una vasca separata per 24 ore prima di proseguire al prelievo.

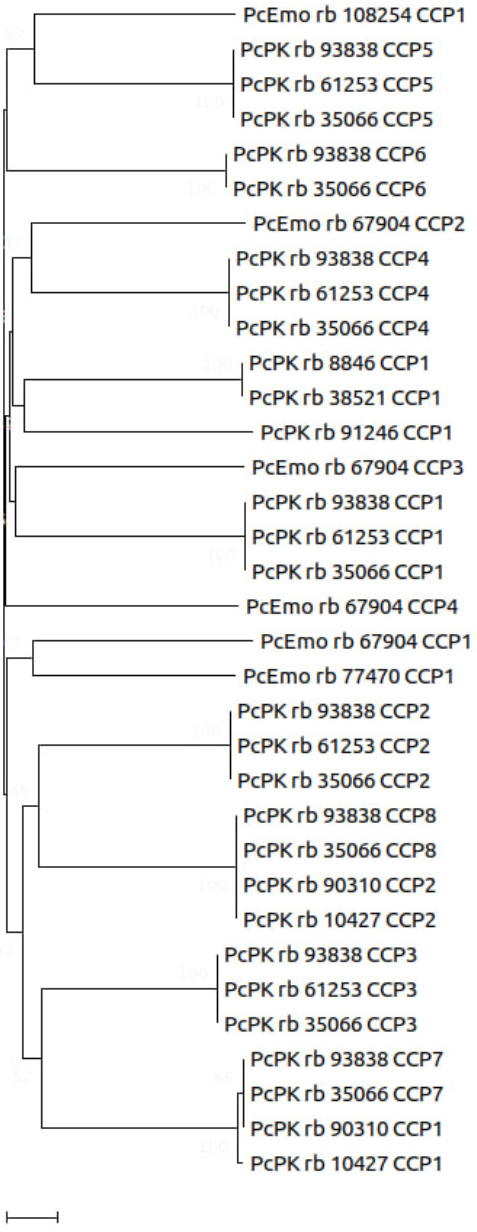
6.2 Software e Strumenti bioinformatici

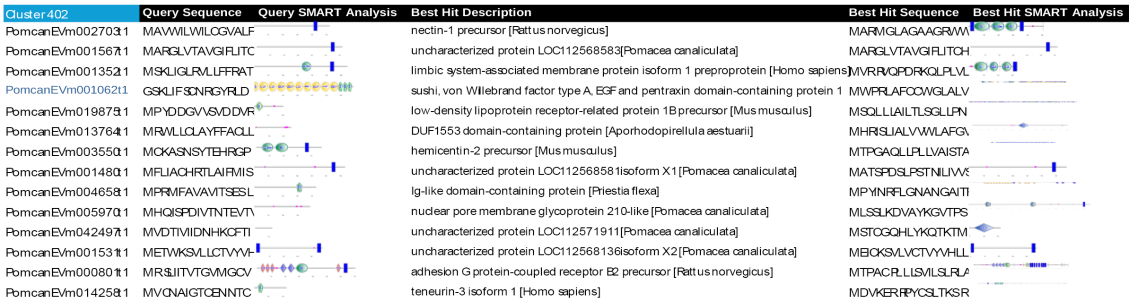
Blast	Allineamento sequenze.	https://blast.ncbi.nlm.nih.gov/Blast.cgi
InterPro Scan	Caratterizzazione funzionale di sequenze proteiche e nucleotidiche.	https://www.ebi.ac.uk/interpro/about/interproscan/
SMART	Identificazione ed annotazione di domini proteici	https://smart.embl.de/
MEGA	Suite di strumenti per l'analisi di sequenze proteiche e nucleotidiche	https://www.megasoftware.net/

ProstT5	pLM capace di tradurre tra sequenza amminoacidica e struttura proteica	https://github.com/mheinzinger/ProstT5
Umap	Tecnica di riduzione dimensionale dei dati	https://umap-learn.readthedocs.io/en/latest/
Dorado	Basecaller ufficiale di Nanopore	https://github.com/nanoporetech/dorado
Pychopper	Strumento capace di identificare, orientare e tagliare le reads cDNA.	https://github.com/epi2me-labs/pychopper
RNAbloom	Assemblatore trascrittomico	https://github.com/bcgsc/RNA-Bloom

Allegati

Allegato 1 Rappresentazione grafica della similarità di sequenza dei domini CCP individuati





Cluster 402	Query Sequence	Query SMART Analysis	Best Hit Description	Best Hit Sequence	Best Hit SMART Analysis
PomcanEVm002703t1	MAVWILWILCGVALF		nectin-1 precursor [Rattus norvegicus]	MARMGLAGAAGRVW	
PomcanEVm001567t1	MARGLVTAVGIFLITC		uncharacterized protein LOC112568583[Pomacea canaliculata]	MARGLVTAVGIFLITCH	
PomcanEVm001352t1	MSKLIGLRVLLFFRAT		limbic system-associated membrane protein isoform 1 preproprotein [Homo sapiens]	MVRRVQPDRKQLPLVL	
PomcanEVm001062t1	GSKLIFSCNRGYRLD		sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein 1	MWPRLAFCCWGLALV	
PomcanEVm019875t1	MPYDDGVVSVDDVR		low-density lipoprotein receptor-related protein 1B precursor [Mus musculus]	MSQLLLAILTLSGLLPN	
PomcanEVm013764t1	MRWLLCLAYFFACLL		DUF1553 domain-containing protein [Aporhodopirellula aestuarii]	MHRSLIALVVWLAFGV	
PomcanEVm003550t1	MCKASNSYTEHRGP		hemicentin-2 precursor [Mus musculus]	MTPGAQLLPLLVAISTA	
PomcanEVm001480t1	MFLIACHRTLAIFMIS		uncharacterized protein LOC112568581isoform X1 [Pomacea canaliculata]	MATSPDSLPSTNILIVVS	
PomcanEVm004658t1	MPRMFAVAVITSESL		Ig-like domain-containing protein [Priestia flexa]	MPYYNRFLGNANGAITI	
PomcanEVm005970t1	MHQISPDIVTNTEVT		nuclear pore membrane glycoprotein 210-like [Pomacea canaliculata]	MLSSLKDVAYKGVTPS	
PomcanEVm042497t1	MVDTIVIIDNHKCFTI		uncharacterized protein LOC112571911[Pomacea canaliculata]	MSTCGQHLYKQTKTM	
PomcanEVm001531t1	METWKSVLLCTVYVH		uncharacterized protein LOC112568136isoform X2 [Pomacea canaliculata]	MEICKSVLVCTVYVHLL	
PomcanEVm000801t1	MRSLIITVTGVMGCV		adhesion G protein-coupled receptor B2 precursor [Rattus norvegicus]	MTPACRLLLSVLSLRLA	
PomcanEVm014258t1	MVCNAIGTCENNTC		teneurin-3 isoform 1 [Homo sapiens]	MDVKERRPYCSLTKSR	

Cluster 362	Query Sequence	Query SMART Analysis	Best Hit Description	Best Hit Sequence	Best Hit SMART Analysis
PomcanEVm007152t1	MTPRSRRRTSTTLSSS		heat shock-related 70 kDa protein 2 [Rattus norvegicus]	MSARGPAIGIDLGTTYS	
PomcanEVm007237t1	MRSPEYLRVGMKLIF		GTP-binding protein 1 [Mus musculus]	MAAERSRSPVDSPVPA	
PomcanEVm019887t1	MFESRNIPGLYIAVQA		actin-related protein 3C isoform a precursor [Homo sapiens]	MFESRNVPGLYIAVQA	
PomcanEVm047790t1	MNRFLGKAVVAVQRI		E3 ubiquitin-protein ligase HECW2-like [Pomacea canaliculata]	MVKAKMSRSNSDYSSI	
PomcanEVm007492t1	MSRRALRCVADVTIH		spermatogenesisassociated 6-like protein [Rattus norvegicus]	MPLEVVVELQIRAISCP	
PomcanEVm007909t1	MTAKLYKIPLFVAKIL		benzoate transporter [Chloroflexus aurantiacus]	MNRPIPRALLSTLAVILL	
PomcanEVm004014t1	MLRVWARHVPYRKS		hypothetical protein C0Q70_17910[Pomacea canaliculata]	MDWESSSCCEYTETKV	
PomcanEVm005258t1	MRYNPTAPDVGVFD		uncharacterized protein PB18E9.04c-like isoform X2 [Pomacea canaliculata]	MEIVWKGGSLRPAVYNI	
PomcanEVm010954t1	MRFSVVLITLLFAVA		hypothetical protein I4U23_029159[Adineta vaga]	MRLSLVVLLCFTIVLVN	
PomcanEVm018021t1	MSASRPPYTPETLFTS		hypothetical protein SARC_06627[Sphaeroforma arctica JP610]	MKVVAALIVLVFGSCIS	
PomcanEVm008182t1	MARASECALVFVVVV		fibronectin type III domain-containing protein [Aquisphaera giovannonii]	MEDLEPRLVMSAGVLT	
PomcanEVm016768t1	MPEHQEAKIYTAEKIL		E3 SUMO-protein ligase CBX4 [Mus musculus]	MELPAVGEHVFAVESE	
PomcanEVm004949t1	MPGILSSLLNGLVVM		uncharacterized protein LOC112559808isoform X1 [Pomacea canaliculata]	MPGILSSLLKGLVVMRF	
PomcanEVm004950t1	MPGILSSLLNGLVVM		uncharacterized protein LOC112559808isoform X1 [Pomacea canaliculata]	MPGILSSLLKGLVVMRF	
PomcanEVm004152t1	MMKISVFLLLVFTEVV		uncharacterized protein LOC112559049isoform X1 [Pomacea canaliculata]	MMKISVFLLLVFTEVLC	
PomcanEVm024213t1	MKRSPEEEPREPPE		hypothetical protein QJQ45_000725[Haematococcus lacustris]	MSPESVQMHVDALWT	
PomcanEVm004468t1	MVLYLGLPDKKRTGI		hypothetical protein C0Q70_06994[Pomacea canaliculata]	MQGILASVSESVAIAVF	
PomcanEVm005393t1	DCRNVWQNCHNLIT		uncharacterized protein LOC112568838isoform X1 [Pomacea canaliculata]	MRGLIITVTGVVLLLTVC	
PomcanEVm007369t1	KTTRRLNLTYPPRPF		uncharacterized protein LOC112568756isoform X2 [Pomacea canaliculata]	MLTVPKVTRADNGTVF	
PomcanEVm011452t1	MAKRKKGAGGGGAL		uncharacterized protein LOC112574796[Pomacea canaliculata]	MVKRKKGAGGGGEDS	
PomcanEVm008358t1	MPKECAAPDCQEKG		THAP domain-containing protein 6 [Rattus norvegicus]	MVKCGSAIGCASRCLP	
PomcanEVm005007t1	MSPVRHPGLTSALVC		uncharacterized protein LOC112569258[Pomacea canaliculata]	MSPVCLPGLTSALVCLL	
PomcanEVm004616t1	MEANYFMLFTLGLFL		uncharacterized protein LOC112571572[Pomacea canaliculata]	MEANYFMLFTLGLFLL	
PomcanEVm011292t1	MQFFKRSDPPQLDLY		leucine-rich repeats and immunoglobulin-like domains protein 2 isoform X1 [Pomac	MKTYVTLTALILGFMVW	
PomcanEVm013203t1	MIAMKAEKCISERLP		uncharacterized protein LOC112556833isoform X1 [Pomacea canaliculata]	MAAKATVCCPDLASAS	
PomcanEVm005707t1	TYISPQSVCLPAPWR		uncharacterized protein LOC112568447[Pomacea canaliculata]	MARDPKEASKQRKFR	
PomcanEVm017325t1	MEISFYRIFLFLLITAV		pituitary tumor-transforming gene 1 protein-interacting protein precursor [Mus mus	MAPANLGLTPHWVML	
PomcanEVm009378t1	MGCNFQCDNSLTM		uncharacterized protein LOC112568712[Pomacea canaliculata]	MGCNFQCDSSLKMNF	
PomcanEVm008614t1	MPFGIKPICVSCKVTS		GATA zinc finger domain-containing protein 1 [Homo sapiens]	MPLGLKPTCSVCKTTS	
PomcanEVm005857t1	MAWCLLLVLLLLVST		uncharacterized protein LOC112571063isoform X3 [Pomacea canaliculata]	MAWCLLLVLLLLISTAA	
PomcanEVm008388t1	METLRNMLLRRPYKI		uncharacterized protein LOC112575090isoform X1 [Pomacea canaliculata]	MGGNKVGVSSAWEITL	
PomcanEVm005414t1	MGNKRTSNACIGFIF		uncharacterized protein LOC112568641[Pomacea canaliculata]	MKLRPQTLAKPPDDQN	
PomcanEVm005456t1	MSHKSSGETCADLE		hypothetical protein C0Q70_12045[Pomacea canaliculata]	MTHVEFAGYKDGDTVI	
PomcanEVm005750t1	MRVREETLLLLWVTV		uncharacterized protein LOC112573282isoform X1 [Pomacea canaliculata]	MRVREETLLLLWVTVC	
PomcanEVm004383t1	MAQRTWQGFQTLTL		mucin-5B precursor [Homo sapiens]	MGAPSACRTLVLALAA	
PomcanEVm009121t1	MGRVCVVFGCGFLN		hypothetical protein C0Q70_20392[Pomacea canaliculata]	MPFVWSPKSIKFEMPS	

Allegato 3 Dataset analisi AI con sequenze umane di controllo

231

	Query Sequence	Query SMART Analysis	Best Hit Description	Best Hit Sequence	Best Hit SMART Analysis
PomcanEVm003674t1	MNENTSTGSITGTEC		uncharacterized protein LOC112574294 [Pomacea canaliculata]	SRQHNSYLKDNLDI	
PomcanEVm004126t1	MENDTNAVWNIPG		uncharacterized protein LOC112558992 [Pomacea canaliculata] >ref|XP_025(	MENDTNAVWNIPG	
PomcanEVm011850t1	MVKNSPWICCMLLL		uncharacterized protein LOC112559493 [Pomacea canaliculata] >ref|XP_025(	MVKNSPWICCMLLL	
PomcanEVm014616t1	MVLSLRSPDVCLCG		hypothetical protein BaRGS_00003034 [Batillaria attramentaria]	LQTDDFVSSAQVSA	
PomcanEVm007152t1	MTPRSRRRTSTTLSS		hsp70-like protein [Diaporthaceae sp. PMI_573]	PKIQKLLSDFFDGKK	
PomcanEVm011469t1	MSSVIFVTIDWLIKVI		uncharacterized protein LOC112559256 isoform X1 [Pomacea canaliculata]	MSSVIFVTIDWLIKLL	
PomcanEVm006347t1	MEQGTTFKIVVMSN		-	-	
PomcanEVm019887t1	MFESFNIPGLYIAVQ		prolyl oligopeptidase (S09 family) [Schistosoma mansoni]	MFETFNIPGLYIAVQ	
PomcanEVm007400t1	METLANLTPTDIENC		unnamed protein product [Rotaria sordida] >emb|CAF3688405.1| unnamed p	MEKLANLTQNDINN	
PomcanEVm047790t1	MNRFLGKAVVAVQI		hypothetical protein C0Q70_10123 [Pomacea canaliculata]	MNRFLGKAVVAVQR	
PomcanEVm019480t1	MLSFDLHKLTCILMI		uncharacterized protein LOC112568809 [Pomacea canaliculata]	MLSFDLHKLTCILML	
PomcanEVm007492t1	MSRRALRCVADVTI		spermatogenesis-associated protein 6-like isoform X3 [Pomacea canaliculata]	MSRRALRCVADVTI	
PomcanEVm004919t1	MGDAATCCTVETPC		uncharacterized protein LOC112576411 [Pomacea canaliculata] >ref|XP_025(	MGDAATCCTVETPC	
PomcanEVm039645t1	MGQWGEWSACSTS		thrombospondin type-1 domain-containing protein [Chitinophagales bacterium	VGPWSPWSACDTS	
PomcanEVm006636t1	MNSFEHSTRELCLL/		uncharacterized protein LOC112569225 isoform X3 [Pomacea canaliculata]	MNSLGHSARELCLL	
PomcanEVm014478t1	MMTKLLFLLLTVVCI		unnamed protein product [Rotaria sp. Silwood1] >emb|CAF3540664.1| unnar	KVFFIFFATVFLPICT-	
PomcanEVm040049t1	SLSILFLRHFIPRSRT		unnamed protein product [Adineta ricciae]	TISILFFGRIIPSYSPR	
PomcanEVm004014t1	MLRVWARHVPYRK(		hypothetical protein C0Q70_17910 [Pomacea canaliculata]	LNYTLVFWRRNDSL	
PomcanEVm021805t1	MKEKSFFCVPFFTIK		unnamed protein product [Rotaria sp. Silwood2]	KNKPILCIPYFTIKHY	
PomcanEVm009909t1	MADIQNFLKELKSQ		hypothetical protein I4U23_026529 [Adineta vaga]	SRSPIRRSSPVPNEN	
PomcanEVm017524t1	MELDEEDERKEDLL/		MYCBP-associated protein-like isoform X1 [Pomacea canaliculata]	MELDEEDERKEDLLA	
PomcanEVm014394t1	MVNEEVGTAEIDINT		unnamed protein product [Moneuplotes crassus]	FSNDEVGSTQLDINT	
PomcanEVm023698t1	MRKFVLTVVIFLSIC/		unnamed protein product [Adineta steineri]	KNYILIVVIYTIALFSTV	
PomcanEVm031199t1	GAYYNLPRDDLKTIL		unnamed protein product [Rotaria sp. Silwood1] >emb|CAF1653488.1| unnar	LKSILPLIISQRSNCY	
PomcanEVm015491t1	MVLFINVYVGQRVE		hypothetical protein C0Q70_20847 [Pomacea canaliculata]	MVLFINVYVGQRVEV	
PomcanEVm017214t1	MEGVAAQSDSDPEI		uncharacterized protein LOC112573858 [Pomacea canaliculata]	MEGVAAQSDSDPEH	
PomcanEVm005258t1	MRYNPTAPDVGVFI		hypothetical protein C0Q70_05358 [Pomacea canaliculata]	MRYNPTTPDVGVFD	
PomcanEVm007657t1	MVVVCSVSEREKIRR		uncharacterized protein LOC112569182 [Pomacea canaliculata] >ref|XP_025:	LCLGVLFLWLMPLA	
PomcanEVm010954t1	MRFSLVVLITLLFAV/		hypothetical protein I4U23_029159 [Adineta vaga]	MRLSLVVL--	
PomcanEVm017024t1	MPPLRQIPDSIKALS		LEM domain-containing protein 1-like [Pomacea canaliculata]	MPPLRQIPDSIKALS	
PomcanEVm018021t1	MSASRPPYTPETLFT		hypothetical protein SARC_06627 [Sphaeroforma arctica JP610] >gb|KNC810:	PPYTHSVIFQAPCFIS	
PomcanEVm008182t1	MARASECALVFVVV		uncharacterized protein LOC112567540 [Pomacea canaliculata]	MARASECALVFVVV	
PomcanEVm016768t1	MPEHQEAKIYTAEKI		unnamed protein product [Rotaria socialis] >emb|CAF3247914.1| unnamed p	MPDDKEAPVYTAEKI	
PomcanEVm008607t1	MLEDEEDTTACEVC		PHD and RING finger domain-containing protein 1-like [Pomacea canaliculata]	MLEDEEDTTACEVC	
PomcanEVm004949t1	MPGILSSLLNGLWM		uncharacterized protein LOC112559808 isoform X1 [Pomacea canaliculata]	MPGILSSLLKGLWM	
PomcanEVm004546t1	MTEGVDTDDILNVS		uncharacterized protein LOC112574827 [Pomacea canaliculata]	MTEVVDTDDILNGN	
PomcanEVm011618t1	MVATRVRLIDALVL/		uncharacterized protein LOC112559070 isoform X3 [Pomacea canaliculata]	MMATRVRLIDALVLA	
PomcanEVm023468t1	MGTFIIQFNKITGNP		-	-	
PomcanEVm004950t1	MPGILSSLLNGLWM		uncharacterized protein LOC112559808 isoform X1 [Pomacea canaliculata]	MPGILSSLLKGLWM	
PomcanEVm006842t1	MIRTENGFQLSLFW		uncharacterized protein LOC112567779 [Pomacea canaliculata] >ref|XP_025:	MRSGSFSCDSATGY	
PomcanEVm022933t1	MAKTLFLPYFSVLTT		transmembrane protein 9-like [Pomacea canaliculata]	QASFEDVRCKCVCP	
PomcanEVm005462t1	MTEYCTVFFTWEID:		uncharacterized protein LOC112564904 isoform X1 [Pomacea canaliculata] >g	MTEYCTVFFTWEIDS	
PomcanEVm006020t1	MTWREGHRIVDMC		zinc finger protein 62 homolog [Pomacea canaliculata]	MTWREGHRIVDMC	
PomcanEVm004152t1	MMKISVFLLLVFTEM		uncharacterized protein LOC112559049 isoform X1 [Pomacea canaliculata] >r	MMKISVFLLLVFTEIV	
PomcanEVm024213t1	MKRSPEEEVPREPPE		hypothetical protein QJQ45_000725 [Haematococcus lacustris]	LSKEDDVRKYVNTYR	
PomcanEVm032766t1	MMFVGDMCNKKRL		uncharacterized protein LOC112553726 [Pomacea canaliculata] >ref|XP_025(	MMFVGDMCNKKRL	
PomcanEVm020166t1	IEYYEEVIRDIHYYHN		LOW QUALITY PROTEIN: calsyntenin-1-like [Pomacea canaliculata]	IEYYEEVIRDIHYYHN	
PomcanEVm010999t1	MFIMKITVLRGIYFIV		hypothetical protein C0Q70_08092 [Pomacea canaliculata]	MCHEASSNMSRCC	
PomcanEVm004468t1	MVLYLGLPDKKRTG		hypothetical protein C0Q70_06994 [Pomacea canaliculata]	ALGADFQIEKLHASK	
PomcanEVm012455t1	MKQIGRSQMLLVLV		delta-like protein A [Pomacea canaliculata]	MNQIGRSQMLLVLV	
PomcanEVm005393t1	DCRNVWQNCHNLI		uncharacterized protein LOC112568838 isoform X1 [Pomacea canaliculata]	FFVRRRRAAQTKTSE	
PomcanEVm006351t1	MNCSLQATSSGHEI		uncharacterized protein LOC112567783 [Pomacea canaliculata]	FTCANPGKDPLDIYF	
PomcanEVm008013t1	MGDKVRAALVTLM		uncharacterized protein LOC112575946 isoform X2 [Pomacea canaliculata]	MGDKVRAALVTLMF	
PomcanEVm007369t1	KTTRRLNLTYPPPKP		uncharacterized protein LOC112568756 isoform X2 [Pomacea canaliculata]	KTTLRLNLTYPPPKPP	
PomcanEVm011452t1	MAKRKKGAGGGGA		uncharacterized protein LOC112574796 [Pomacea canaliculata]	RVKVYLHAVTNDEA	
PomcanEVm008358t1	MPKECAAPDCQEK(		uncharacterized protein LOC112554666 isoform X1 [Pomacea canaliculata]	MPKECAAPDCQEKG	
PomcanEVm005795t1	MAHWSQWFNLDD		uncharacterized protein LOC112572815 [Pomacea canaliculata]	MAHWSQWFNLDDP	
PomcanEVm005007t1	MSPVRHPGLTSALV		uncharacterized protein LOC112569258 [Pomacea canaliculata] >ref|XP_025:	MSPVCLPGLTSALV	
PomcanEVm004616t1	MEANYFIVLFTLGLF		LOW QUALITY PROTEIN: uncharacterized protein LOC112571572 [Pomacea ca	MEANYFIVLFTLGLFL	
PomcanEVm008828t1	MKPGLLKMAGVRD		hypothetical protein C0Q70_12470 [Pomacea canaliculata]	SLTFHIEGNKTSFTFP	
PomcanEVm011292t1	MQFFKRSDPPCLDL		leucine-rich repeats and immunoglobulin-like domains protein 2 isoform X1 [P	FSRSDPPCLDLYTTT	
PomcanEVm017685t1	LKPPTRWSARCTIQ:		hypothetical protein C0Q70_10661 [Pomacea canaliculata]	REGKAVHQVVVPKA	
PomcanEVm019110t1	MEDDVVTAARRRIM		uncharacterized protein LOC112558065 [Pomacea canaliculata]	MEDDVVTAARRRIM	
PomcanEVm013203t1	MIAMKAEKCISEIRLI		uncharacterized protein LOC112556833 isoform X1 [Pomacea canaliculata	AKATVQCPDLASAS	
PomcanEVm007115t1	MWGKDDGQSNGTI		uncharacterized protein LOC112568994 isoform X3 [Pomacea canaliculata]	MWGKDDGQSNGTF	
PomcanEVm006758t1	MRHVVLVIAIVVCC\		uncharacterized protein LOC112569162 isoform X2 [Pomacea canaliculata]	MRHVVLVIAIVVCCV	
PomcanEVm024271t1	MVPSVFLMLMLAV\		uncharacterized protein LOC112570348 [Pomacea canaliculata]	VVVPTHGRPSCPPG	
PomcanEVm013885t1	MTLSTSSLLTLVMLA		uncharacterized protein LOC112560378 [Pomacea canaliculata] >ref|XP_025(	VMLAIVTSPSLAAEQ	
PomcanEVm004501t1	MTDNNEGNLAGRR		uncharacterized protein LOC112574294 [Pomacea canaliculata]	SRQHNSYLKDNLDI	
PomcanEVm015067t1	MPSTTSLCASATAV		tumor necrosis factor receptor superfamily member 13B-like [Pomacea canali	MPSTTSLCASATAVV	

PomcanEVm017567t1	MVTKYFIDSNNNNK	uncharacterized protein LOC112568770 [Pomacea canaliculata]	IVTKYFIDSNNNNKE
PomcanEVm012801t1	MSGETSLELLIPDSS	uncharacterized protein LOC112567443 [Pomacea canaliculata]	MSGETSLKFLIPESS
PomcanEVm019471t1	MVPSVFLVLMLAVV	uncharacterized protein LOC112570348 [Pomacea canaliculata]	VVVPTHGRPSCPPG
PomcanEVm004276t1	MKISVFLLLVFTVMV	uncharacterized protein LOC112559049 isoform X3 [Pomacea canaliculata] >r	MKI-
PomcanEVm008585t1	MKSNLLVHLSLPLFF	uncharacterized protein LOC112568269 [Pomacea canaliculata] >ref|XP_025:	MKSNLLVHLSLPLFF
PomcanEVm022529t1	MSSSCACCSKQTSA	uncharacterized protein LOC112561109 isoform X2 [Pomacea canaliculata]	MFSSCACCSKQTSA
PomcanEVm023372t1	MLVQAVASQLLEAV	zinc finger C2HC domain-containing protein 1A-like isoform X1 [Pomacea cana	RFGVKAGMSSPSLLL
PomcanEVm005707t1	TYISPQSVCLPAPWF	uncharacterized protein LOC112568447 [Pomacea canaliculata]	EISATRRRATCSH-----
PomcanEVm007462t1	MKMSCSTAHQGAS	uncharacterized protein LOC112568957 [Pomacea canaliculata]	MKMSCSTAHQGASL
PomcanEVm017325t1	MEISFYRIFILFLLITA	LOW QUALITY PROTEIN: pituitary tumor-transforming gene 1 protein-interactin	MSNVTDVASTTEVLT
PomcanEVm009378t1	MGCNFQCDNSLTM	uncharacterized protein LOC112568712 [Pomacea canaliculata] >ref|XP_025:	MGCNFQCDSSLKM
PomcanEVm005857t1	MAWCLLLVLLLLVS	uncharacterized protein LOC112571063 isoform X3 [Pomacea canaliculata]	MAWCLLLVLLLLLST
PomcanEVm013965t1	MTCYIESSRLHCTST	uncharacterized protein LOC112567786 [Pomacea canaliculata] >ref|XP_025:	MTCYIESSRLHCTST
PomcanEVm008388t1	METLRNMLLRRPYK	uncharacterized protein LOC112575090 isoform X1 [Pomacea canaliculata]	METLRNMLLRRPYKI
PomcanEVm005414t1	MGNKRTSNACIGFIF	uncharacterized protein LOC112568641 [Pomacea canaliculata] >gb|PVD273	DDQNTVSTLSMFTR
PomcanEVm005456t1	MSIHKSSGETCADLE	hypothetical protein C0Q70_12045 [Pomacea canaliculata]	ADLEFHGFRDGETV
PomcanEVm023124t1	MSLGLTSVAMVALL	hypothetical protein C0Q70_14184 [Pomacea canaliculata]	MSLGLTSVVMVALLT
PomcanEVm005750t1	MRVREETLLLLWVT	uncharacterized protein LOC112573282 isoform X1 [Pomacea canaliculata] >ş	MRVREETLLLLWVTV
PomcanEVm004383t1	MAQRTWQQFQTLT	uncharacterized protein LOC112555430 [Pomacea canaliculata] >ref|XP_025(	MAQRTRAATWQRF
PomcanEVm009121t1	MGRVCVVFGCGFL	hypothetical protein C0Q70_20392 [Pomacea canaliculata]	MPFVWSPKSIKFEM
PomcanEVm005317t1	MTTTEEMKKFIAASL	uncharacterized protein LOC112558245 [Pomacea canaliculata]	MTTTEEMKKFIAASD

232

	Query Sequence	Best Hit Description	Best Hit Sequence
PomcanEVm012876t1	MTTVEVFLLTAFVM	uncharacterized protein LOC112567706 isoform X2 [Pomacea canaliculata]	MTTVEVFLLTAFVMT
PomcanEVm018504t1	MSCSFNHLAHLMC	uncharacterized protein LOC112569223 isoform X1 [Pomacea canaliculata] >r	MSCTTEVEEKSQLFC
PomcanEVm006247t1	MWPSPRRLASPTTG	predicted protein, partial [Nematostella vectensis]	GGLVCSMNRSLTR
PomcanEVm017937t1	MSGETFVKFLIPESS	uncharacterized protein LOC112567443 [Pomacea canaliculata]	MSGETSLKFLIPESS
PomcanEVm012429t1	MPLNLPIVELVLDRC	synaptojanin-2-binding protein-like isoform X1 [Pomacea canaliculata]	MPLNLPIVELVLDRG
PomcanEVm053562t1	MVLMVPFMIHKIGEI	40S ribosomal protein S3 [Aplysia californica]	LMIHSGEPLNDYVDT
PomcanEVm028252t1	MKADPPPVISWEGV	uncharacterized protein LOC112568675 isoform X2 [Pomacea canaliculata]	EGNPPPVISWEGQG
PomcanEVm012884t1	MKTDLVIILSLSFLIS	uncharacterized protein LOC112567701 [Pomacea canaliculata]	MKTDLVIILSLSFLIS
PomcanEVm009527t1	YKKPLSRYSAGRCSF	hypothetical protein C0Q70_03286 [Pomacea canaliculata]	VAYPVPGILYGWVQ
PomcanEVm007909t1	MTAKLYKIPLFVAKIL	mucin-5AC-like isoform X2 [Pomacea canaliculata]	MTAKLYTIPLFVAKIL-
PomcanEVm041492t1	MIKPRLVFAGHTHQ	hypothetical protein I4U23_009426 [Adineta vaga]	MIKPRLIFSGHTHRY
PomcanEVm014764t1	MSTNDHNDPDISST	unnamed protein product [Rotaria sp. Silwood2] >emb|CAF2528129.1| unnar	MSSNKQNNQEISTS
PomcanEVm028451t1	MTMIRSNLIFETLHV	emp24/gp25L/p24 family protein [Dictyostelium discoideum]	LTDPKGNTIFERLHF
PomcanEVm015814t1	MSGEMILKLQIPESS	uncharacterized protein LOC112567439 isoform X2 [Pomacea canaliculata]	MSGETSLKFLIPESS
PomcanEVm005389t1	MAVRLPFLVVTVTL	hypothetical protein C0Q70_04880 [Pomacea canaliculata]	MALRQPFLVVTVTL
PomcanEVm028833t1	MAVGECSTLPSVLD	uncharacterized protein LOC112557320 [Pomacea canaliculata]	MAVGECSTLPSVLD
PomcanEVm016953t1	MTCEIKSKHHPVQIK	uncharacterized protein LOC112568155 [Pomacea canaliculata]	LDRRGLRRQTEGFN
PomcanEVm004866t1	MAFNFRHQSSAKM	hypothetical protein C0Q70_04882 [Pomacea canaliculata]	MYGRNQGARATLTS
PomcanEVm008463t1	MIRKHTFDCVLLCV	hypothetical protein C0Q70_07607 [Pomacea canaliculata]	TTRVQRKCEDPDVV
PomcanEVm026073t1	MTRSSKRREFRIEYK	uncharacterized protein LOC112564392 [Pomacea canaliculata] >ref|XP_025(	ERREFRIEYKNPGQK
PomcanEVm009960t1	MLVVRPVLLITLIFVI	unnamed protein product [Rotaria sp. Silwood1] >emb|CAF4684318.1| unnar	NPPESTSLERAHWC
PomcanEVm004469t1	MAAPAYVHSENAC	protein draper-like [Pomacea canaliculata]	ECVCVEGWTGAYC
PomcanEVm012981t1	MREQMVLDEKIIGLE	unnamed protein product [Adineta ricciae]	MREQMVLDEKIVGV
PomcanEVm013006t1	MCSPGYYSAEGACF	platelet endothelial aggregation receptor 1-like [Pomacea canaliculata]	MCSPGYYSAEGACF
PomcanEVm021264t1	MSGETSVKFLIPESS	uncharacterized protein LOC112567443 [Pomacea canaliculata]	MSGETSLKFLIPESS
PomcanEVm016062t1	MNFVGIKLNEANPL	hypothetical protein M9434_003998 [Picochlorum sp. BPE23]	VNFFGVVLKPDAPTA
PomcanEVm013711t1	MSVRTQSREPLLRNI	uncharacterized protein LOC112568424 [Pomacea canaliculata]	VISFLQLAVLISFASG
PomcanEVm009221t1	MHIMCSASSTLLTS	uncharacterized protein LOC112569050 [Pomacea canaliculata]	MSGFFCLVLVTMVT
PomcanEVm035510t1	MQRQLTTLTRISCV	PLAC8 family-domain-containing protein [Jimgerdemannia flammicorona]	LGDGSCFSQGAIYC
PomcanEVm024833t1	MDTDQPVQRHTAV	unnamed protein product [Rotaria sp. Silwood2] >emb|CAF2538398.1| unnar	FSEDFLNRTFHFHNP
PomcanEVm015607t1	MSGILFYVNNRNESI	uncharacterized protein LOC112567463 [Pomacea canaliculata]	GQKHDANVCKLIVT
PomcanEVm012140t1	MIIIYFITIHCLLKAHC	uncharacterized protein LOC112567789 [Pomacea canaliculata]	VQAGQGPDLYFPSF
PomcanEVm016722t1	MRCTTTSTTPFHLD	uncharacterized protein LOC112569064 [Pomacea canaliculata] >ref|XP_025:	MRCASTRTAPFHVD
PomcanEVm012785t1	MNCISLEFDGLRSS	hypothetical protein C0Q70_12062 [Pomacea canaliculata]	VEFSGLGSGENWTEI
PomcanEVm011206t1	MSLYILLLVWTLGA	hypothetical protein C0Q70_13127 [Pomacea canaliculata]	MSLYVLLLVWTLGAV
PomcanEVm011598t1	VENCAELVFGNDTD	uncharacterized protein LOC112568328 isoform X1 [Pomacea canaliculata]	VENCAELVFGNDTD
PomcanEVm007128t1	MVTCVACGVCLMII	uncharacterized protein LOC112568313 [Pomacea canaliculata] >ref|XP_025:	MVTCVACGVCLMMI
PomcanEVm011187t1	MEQTGLLHGHYFVI	uncharacterized protein LOC112569216 [Pomacea canaliculata]	EMTKSSSQESSQMT
PomcanEVm060094t1	MQVSDLKVKRITPRI	histone H2A.V-like [Lagopus leucura]	DLKVKRITPRHLQLA
PomcanEVm011397t1	MLSSLAVYRALKEW	uncharacterized protein LOC112568183 isoform X1 [Pomacea canaliculata]	APRLPGNTDHPSLIK
PomcanEVm013981t1	MMFSIWLRHTAISR	uncharacterized protein LOC112569144 [Pomacea canaliculata] >ref|XP_025:	EEVSLTCTFGENINE
PomcanEVm017692t1	MQRMELFYRGDISR	uncharacterized protein LOC112569144 [Pomacea canaliculata] >ref|XP_025:	MHFSSDAHKL-
PomcanEVm012422t1	MFTSFCLFTFVLFTL	hypothetical protein C0Q70_12137 [Pomacea canaliculata]	TEDDSPRPGVVVAV
PomcanEVm013622t1	MLRNHDDSVMQQC	uncharacterized protein LOC112569996 [Pomacea canaliculata]	SRASPSTSYRKK------
PomcanEVm007606t1	VVFERVKCAQQEKC	uncharacterized protein LOC112569625 isoform X2 [Pomacea canaliculata]	LVFITVAYTHLENCA
PomcanEVm011476t1	MMSTQGQISELYLTI	uncharacterized protein LOC112567729 isoform X1 [Pomacea canaliculata]	MMSAQGQISELYLTI
PomcanEVm016658t1	KHPLKQNVSTPSSA	furin-like protease kpc-1 isoform X5 [Pomacea canaliculata]	KHPLKQNVSTPSSAA

PomcanEVm007776t1	MFLSSAKKRWSQVI	hypothetical protein C0Q70_06059 [Pomacea canaliculata]	IWRIHTRAMLRTAEL
PomcanEVm006007t1	MTSRLLFWCTFVAV	uncharacterized protein LOC112574344 isoform X1 [Pomacea canaliculata]	MTSRLLFWCTFVAVL
PomcanEVm014346t1	MTCEIKKEKCPFEIQ	uncharacterized protein LOC112569680 [Pomacea canaliculata]	MSCEIKNEKCPFEIE
PomcanEVm014235t1	MRCTTTRTTPFHLD	uncharacterized protein LOC112569064 [Pomacea canaliculata] >ref|XP_025	MRCASTRTAPFHVD
PomcanEVm014348t1	MSFSTATTQFLERR	uncharacterized protein LOC112569075 [Pomacea canaliculata]	MSLSTDKTQFLKMR
PomcanEVm008955t1	MSTTRPSPEVSLTN	LOW QUALITY PROTEIN: uncharacterized protein LOC112567035 [Pomacea ca	MSTTRPSPEVSLTNY
PomcanEVm005626t1	SGKEKKYLDRCIGIIC	uncharacterized protein LOC112568766 isoform X1 [Pomacea canaliculata]	GEEKKYIGECIGIICN
PomcanEVm018370t1	MSGETSLELLIPDSS	uncharacterized protein LOC112567439 isoform X2 [Pomacea canaliculata]	MSGETSLKFLIPESS
PomcanEVm011503t1	MMKIALLIVLTATVT	hypothetical protein C0Q70_06060 [Pomacea canaliculata]	MMKIALLIVLTAAVTT
PomcanEVm012464t1	MFIFMYLALLLAFG	uncharacterized protein LOC112568201 [Pomacea canaliculata]	VFAIDDKCNTEYETT
PomcanEVm014238t1	MATYGICATDKTCA	uncharacterized protein LOC112569050 [Pomacea canaliculata]	MVTYGICATEKTCAH
PomcanEVm007921t1	MDTTPSTQGKTSSG	uncharacterized protein LOC112569182 [Pomacea canaliculata] >ref|XP_025	LGVLFLWLMPLANC
PomcanEVm004292t1	MCLHRLSHNKLKTL	uncharacterized protein LOC112567763 [Pomacea canaliculata]	MWSPYLTNIVRTVLT
PomcanEVm005714t1	MAAILPFISVAMTLV	hypothetical protein C0Q70_04879 [Pomacea canaliculata]	MAATVPFISVAMTFW
PomcanEVm014158t1	MSCTTRHQATHQL	uncharacterized protein LOC112569223 isoform X1 [Pomacea canaliculata] >r	MSCTTEVEEKSQLFC
PomcanEVm016944t1	MLFCTTHLSTGVIID	uncharacterized protein LOC112569520 [Pomacea canaliculata]	SEGVIIDNCDNEIRIV
PomcanEVm001416t1	MELEKLYFWDLFLF	uncharacterized protein LOC112568578 [Pomacea canaliculata]	KPTPLIRGLYTYNMR
PomcanEVm010717t1	MTGHFYSCDDSRTT	unnamed protein product [Adineta steineri]	MAGHFYSCHET-
PomcanEVm010990t1	MSCQLKNTKGPFEI	hypothetical protein C0Q70_12375 [Pomacea canaliculata]	MSCQLKNTKSPFEIQ
PomcanEVm008564t1	MNPEELFKRFFGFTI	uncharacterized protein LOC112553847 isoform X1 [Pomacea canaliculata]	MNPEELFKRFFGFTK
PomcanEVm010835t1	MEHVHRAVLRVFM	uncharacterized protein LOC112572452 isoform X2 [Pomacea canaliculata] >r	MEHVHRAVLRVFMG
PomcanEVm009842t1	MAATTAYLLVLVTW	uncharacterized protein LOC112568212 [Pomacea canaliculata]	MATTTAYLLVFVTWC
PomcanEVm006395t1	MHERVLDRCLTNVI	uncharacterized protein LOC112558823 [Pomacea canaliculata]	MIGTDQGFEHAILTSP
PomcanEVm012206t1	MAAGVCRHSLLSRL	uncharacterized protein LOC112569555 isoform X1 [Pomacea canaliculata]	MAAGVCRHSLLSRL
PomcanEVm018311t1	MSCSFNHLAHLMC	uncharacterized protein LOC112569223 isoform X1 [Pomacea canaliculata] >r	MSCTTEVEEKSQLFC
PomcanEVm011721t1	MRNFSNYFQLLLLT	uncharacterized protein LOC112568273 isoform X3 [Pomacea canaliculata]	MRNFSNYFQLLLLTT
PomcanEVm009652t1	MTLLVSRLAFVVVIII	uncharacterized protein LOC112569165 [Pomacea canaliculata]	MTLLVSRLAFVVVIIIA
PomcanEVm008722t1	MAYTSRASFNLRIFC	uncharacterized protein LOC112572874 [Pomacea canaliculata]	IVAIVCEVF----

207

	Query Sequence	Best Hit Description	Best Hit Sequence
PomcanEVm007646t1	MPVKEFNDNYINVG	uncharacterized protein LOC112555791 [Pomacea canaliculata] >gb|PVD190	MPVKEFNDNYINVG
PomcanEVm045415t1	ISAKIALSTCFCISTT	hypothetical protein C0Q70_06414 [Pomacea canaliculata]	FILENDSLFYVGVDG
PomcanEVm034938t1	MLLLISLNVLVAKSS	hypothetical protein [Philodina roseola]	ISHMKKPQIRAVLLN
PomcanEVm036462t1	RSFLRCNVASSLVIA	hypothetical protein I4U23_020465 [Adineta vaga]	MGSDSSKSKKRATQ
PomcanEVm053774t1	SFFGTVVKPCLKAV	non-canonical purine NTP pyrophosphatase, partial [Escherichia coli]	ILREARGAHGFGYDP
PomcanEVm009016t1	MSAQPIVSVNSDSH	uncharacterized protein LOC112560396 [Pomacea canaliculata] >gb|PVD350	MSAQPIVSVNSDSH
PomcanEVm006685t1	MEHPMSQQVFEDA	uncharacterized protein LOC112555076 [Pomacea canaliculata] >gb|PVD187	MEHPMSQQVFEDA
PomcanEVm015891t1	MPLSRPVFYRLLIFL	unnamed protein product [Rotaria magnacalcarata]	KRKLFHACFI-
PomcanEVm012091t1	MTTRNEKAPRYFPV	hypothetical protein SeMB42_g04412 [Synchytrium endobioticum] >gb|TPX46	THAICNSCDKSLVR
PomcanEVm008466t1	MTGTVHFDVPDTAL	tumor necrosis factor ligand superfamily member 11-like [Pomacea canalic	MPPSRIPPGADLLTC
PomcanEVm077461t1	MEKFNIEKDIAAYIKI	hypothetical protein INT44_005811 [Umbelopsis vinacea]	LEKYNIEKDIAAHIKR
PomcanEVm010351t1	MTIYKTPGGEKKTEK	integral membrane protein 2C-like [Pomacea canaliculata]	LCDRLTGVKDEVEC
PomcanEVm049432t1	MQEKDDTLQRVRLL	myosin heavy chain, embryonic smooth muscle isoform-like [Pomacea canalic	MQEQDDTLQRVRRL
PomcanEVm011325t1	MVGVSLNNANKSD	unnamed protein product [Rotaria sp. Silwood1] >emb|CAF1616688.1| unnar	MVGVSLNTPNKADD
PomcanEVm061939t1	DEIRFGKEQFGPAVI	hypothetical protein SteCoe_35774 [Stentor coeruleus]	DELRFGPNRIPKNEP
PomcanEVm022269t1	MAMLSNHELIRRVP	cyclic nucleotide-binding domain-containing protein [Burkholderiales bacteriu	MAMLSNHELIRRVPI
PomcanEVm008678t1	MQRSTSGHVSWPT	uncharacterized protein LOC112555273 [Pomacea canaliculata] >ref|XP_025	MQTSTSGHVSWPTN
PomcanEVm014712t1	MLCIQQIFTEPNILS	alpha-N-acetylgalactosamine-specific lectin-like isoform X1 [Pomacea canalic	MATSRSPRFTFITRVA
PomcanEVm010699t1	MECVKRSIHVALVFI	tumor necrosis factor ligand superfamily member 10-like [Pomacea canalic	QERGFKTTYKDNHTL
PomcanEVm016587t1	MLGVVDQAKGLYRI	microtubule-actin cross-linking factor 1-like isoform X6 [Pomacea canaliculata	MLGVVDQAKGLYRN
PomcanEVm015528t1	WDDRPNSVLLASGI	RING finger protein 150-like [Pomacea canaliculata] >gb|PVD35007.1| hypoth	MNPCTSIFVFSVIVLL
PomcanEVm009195t1	MGFSTLFGSNTITTA	tumor necrosis factor ligand superfamily member 11-like [Pomacea canalic	MGFSTLFGSKTITTD
PomcanEVm005429t1	MAATERQWSIREDC	LOW QUALITY PROTEIN: baculoviral IAP repeat-containing protein 6-like [Poma	MAATERQWSIREDG
PomcanEVm009897t1	MSESRVSSLASGSIG	hypothetical protein C0Q70_11300 [Pomacea canaliculata]	MSDSRVSSLASGSIG
PomcanEVm020932t1	MLETTEYTLQQVPSI	hypothetical protein Ctob_013751 [Chrysochromulina tobinii]	NWMGVKLEEDSFG
PomcanEVm006118t1	MSVSLQDLGVREKC	uncharacterized protein LOC112555073 isoform X7 [Pomacea canaliculata]	MSVSLQDLGVREKG
PomcanEVm009258t1	MKTRLSSESSMDST	tumor necrosis factor-like [Pomacea canaliculata]	MKTRLSSESSMDSTT
PomcanEVm012947t1	MPSRHLTGILRGRGI	hypothetical protein C0Q70_14854 [Pomacea canaliculata]	VGFTVRFSADPVTNV

NP_000585.2

308

	Query Sequence	Best Hit Description	Best Hit Sequence
PomcanEVm001120t1	MMDFNSKVFFFLLIF	unnamed protein product [Rotaria sp. Silwood2] >emb|CAF4070867.1| unnar	MNFIVKTSFEILLVIVL
PomcanEVm009674t1	MAMLSMDSTVLLL	unnamed protein product [Rotaria sp. Silwood1] >emb|CAF1605507.1| unnar	EDGFEAVGCCLCCA
PomcanEVm007060t1	MACARNWWILLVV	uncharacterized protein LOC112568895 isoform X3 [Pomacea canaliculata]	DVVTSCPEYFIAGQV
PomcanEVm018791t1	MQDRRLNLVTDRIE	uncharacterized protein LOC112571157 [Pomacea canaliculata]	MQDRRLNLATDRIEE

CD55

374

	Query Sequence	Best Hit Description	Best Hit Sequence
PomcanEVm002703t1	MAVWILWILCGVAL	uncharacterized protein LOC112568828 isoform X1 [Pomacea canaliculata]	LWRLWILCGVALLAI
PomcanEVm000688t1	MYRLIASCLICYLLIC	uncharacterized protein LOC112569288 isoform X3 [Pomacea canaliculata]	MYRLIASCLLCYLLIP
PomcanEVm010228t1	MCVLVLLTAAQEAC	uncharacterized protein LOC112568119 [Pomacea canaliculata]	MCVLVLLTAAQEAG
PomcanEVm037242t1	MTLRKGGGVGRPM	cubilin-like isoform X2 [Pomacea canaliculata]	MTLRKGGGMGRPM
PomcanEVm001480t1	MFLIACHRTLAIFMI	uncharacterized protein LOC112568581 isoform X1 [Pomacea canaliculata]	MFLIACHRTLAISLQ
PomcanEVm008328t1	MKFCESYFRICQLQS	uncharacterized protein LOC112560063 [Pomacea canaliculata]	MGLEYKPDGQKVYC
PomcanEVm001993t1	MRLLSCVIIVVAAAL	tolloid-like protein 2 [Pomacea canaliculata]	MESCINCSCDFLAL

CFAH_HUMAN

Bibliografia

A.Accorsi, E. Ross, E. Ottaviani, A. S. Alvarado, Eur. J. Histochem. 2017, 61, 1.

Accorsi, A. Soluble Factors in the Immune-Neuroendocrine System of Invertebrate Models. Ph.D. Thesis, University of Modena and Reggio Emilia, Modena, Italy, 2015.

Accorsi, Alice, Enzo Ottaviani, and Davide Malagoli. "Effects of repeated hemolymph withdrawals on the hemocyte populations and hematopoiesis in Pomacea canaliculata." *Fish & Shellfish Immunology* 38.1 (2014): 56-64.

Accorsi, Alice, et al. "Comparative analysis of circulating hemocytes of the freshwater snail Pomacea canaliculata." *Fish & shellfish immunology* 34.5 (2013): 1260-1268.

Accorsi, Alice, et al. "Complete Regeneration of a Camera-type Eye in the Research Organism Pomacea canaliculata." *The FASEB Journal* 32 (2018): 232-4.

Akira, Shizuo, Satoshi Uematsu, and Osamu Takeuchi. "Pathogen recognition and innate immunity." *Cell* 124.4 (2006): 783-801.

Alberts, Bruce, et al. *Biologia molecolare della cellula*. Bologna: Zanichelli, 2009.

Al-Sharif, W.Z., Sunyer, J.O., Lambris, J.D., Smith, L.C., 1998. Sea urchin coelomocytes specifically express a homologue of the complement component C3. J. Immunol. 160, 2983–299.

Amara, U. et al. Molecular intercommunication between the complement andcoagulation systems. J. Immunol. 185, 5628–5636 (2010).

B. Chen *et al.*, “xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein.” bioRxiv, p. 2023.07.05.547496, Jul. 06, 2023.

Bassi, N., et al. "Pentraxins, anti-pentraxin antibodies, and atherosclerosis." *Clinical Reviews in Allergy & Immunology* 37.1 (2009): 36-43.

Baum, Matthew L., et al. "CSMD1 regulates brain complement activity and circuit development." *Brain, Behavior, and Immunity* 119 (2024): 317-332.

Bergamini, Giulia, et al. "A New Protocol of Computer-Assisted Image Analysis Highlights the Presence of Hemocytes in the Regenerating Cephalic Tentacles of Adult Pomacea canaliculata." *International journal of molecular sciences* 22.9 (2021): 5023.

Beutler, Bruce. "Innate immunity: an overview." *Molecular immunology* 40.12 (2004): 845-859.

Bever, Michele Miller, and Richard B. Borgens. "Eye regeneration in the mystery snail." *Journal of Experimental Zoology* 245.1 (1988): 33-42.

Brehélin, Michel, and Patricia Roch. "Specificity, learning and memory in the innate immune response." *Invertebrate Survival Journal* 5.2 (2008): 103-109.

Clow, L.A., Gross, P.S., Shih, C.-S., Smith, L.C., 2000. Expression of SpC3, the sea urchin complement component, in response to lipopolysaccharide. Immunogenetics 51, 1021–1033.

Cueto, J. A., Rodriguez, C., Vega, I. A., & Castro-Vazquez, A. (2015). Immune defenses of the invasive apple snail Pomacea canaliculata (Caenogastropoda, Ampullariidae): phagocytic hemocytes in the circulation and the kidney. *PLoS One*, *10*(4), e0123964.

Denny, Kerina J., et al. "C5a receptor signaling prevents folate deficiency–induced neural tube defects in mice." *The Journal of Immunology* 190.7 (2013): 3493-3499.

Escudero-Esparza, A., Kalchishkova, N., Kurbasic, E., Jiang, W.G., Blom, A.M., 2013. The novel complement inhibitor human CUB and Sushi multiple domains 1 (CSMD1) protein promotes factor I-mediated degradation of C4b and C3b and inhibits the membrane attack complex assembly. FASEB J. 27, 5083–5093.

Faulk, W. P., et al. "Immunological studies of human placentae: complement components in immature and mature chorionic villi." *Clinical and experimental immunology* 40.2 (1980): 299.Akira, Shizuo, Satoshi Uematsu, and Osamu Takeuchi. "Pathogen recognition and innate immunity." *Cell* 124.4 (2006): 783-801.

Ghosh, Julie, et al. "Invertebrate immune diversity." *Developmental & Comparative Immunology* 35.9 (2011): 959-974.

Harris, Claire L., Masashi Mizuno, and B. Paul Morgan. "Complement and complement regulators in the male reproductive system." *Molecular immunology* 43.1-2 (2006): 57-67.Accorsi, A. Soluble Factors in the Immune-Neuroendocrine System of Invertebrate Models. Ph.D. Thesis, University of Modena and Reggio Emilia, Modena, Italy, 2015.

Håvik, B., Le Hellard, S., Rietschel, M., Lybæk, H., Djurovic, S., Mattheisen, M., Mühleisen, T.W., Degenhardt, F., Priebe, L., Maier, W., Breuer, R., Schulze, T.G., Agartz, I., Melle, I., Hansen, T., Bramham, C.R., Nothen, ¨ M.M., Stevens, B., Werge, T., Andreassen, O.A., Cichon, S., Steen, V.M., 2011. The complement controlrelated genes CSMD1 and CSMD2 associate to schizophrenia. Biol. Psychiatry 70, 35–42.

Hawksworth, Owen A., et al. "Complement in stem cells and development." *Seminars in Immunology*. Vol. 37. Academic Press, 2018.

Heinzinger, Michael, et al. "Prostt5: Bilingual language model for protein sequence and structure. bioRxiv." (2023): 2023-07.

Hirschfield, Gideon M., et al. "Human C-reactive protein does not protect against acute lipopolysaccharide challenge in mice." *The Journal of Immunology* 171.11 (2003): 6046-6051.

Irmscher, S. et al. Kallikrein cleaves C3 and activates complement. J. Innate Immun. 10, 94–105 (2018)

Kimbrell, Deborah A., and Bruce Beutler. "The evolution and genetics of innate immunity." *Nature Reviews Genetics* 2.4 (2001): 256-267.

King, B. C. & Blom, A. M. Intracellular complement: evidence, definitions, controversies, and solutions. Immunol. Rev. 313, 104–119 (2023). Comprehensive overview of the mechanisms and functions of intracellular complement, which also highlights experimental and conceptual challenges.

Klein, Jan, and Vaclav Horejsi. *Immunology*. Wiley-Blackwell, 1999.

Köhl, Jörg. "The role of complement in danger sensing and transmission." *Immunologic research* 34.2 (2006): 157-176.

Kolev, Martin, Gaelle Le Friec, and Claudia Kemper. "Complement—tapping into new sites and effector systems." *Nature Reviews Immunology* 14.12 (2014): 811-820.

Kraus, D.M., Elliott, G.S., Chute, H., Horan, T., Pfenninger, K.H., Sanford, S.D., Foster, S., Scully, S., Welcher, A.A., Holers, V.M., 2006. CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. J. Immunol. 176, 4419–4430.

Krych-Goldberg, M., Atkinson, J., 2001. Structure-function relationships of complement receptor type 1. Immunol. Rev. 180, 112–122.

Lachmann, P.J., 1979. An evolutionary view of the complement system. Behring Institute Mitteilungen 63, 25–37.

M. Heinzinger *et al.*, “Modeling aspects of the language of life through transfer-learning protein sequences,” *BMC Bioinformatics*, vol. 20, no. 1, p. 723, Dec. 2019.

Mannes, Marco, et al. "Tuning the functionality by splicing: factor H and its alternative splice variant FHL-1 share a gene but not all functions." *Frontiers in Immunology* 11 (2020): 596415.

Mastellos, D. C., Deangelis, R. A. & Lambris, J. D. Complement-triggered pathways orchestrate regenerative responses throughout phylogenesis. Semin. Immunol. 25, 29–38 (2013).

Medzhitov, R., Janeway Jr, C.A., 2002. Decoding the patterns of self and nonself by the innate immune system. Science 296, 298–300.

Merle, N.S., Church, S.E., Fremeaux-Bacchi, V., Roumenina, L.T., 2015a. Complement system part I - molecular mechanisms of activation and regulation. Front. Immunol. 6, 262.

Merle, Nicolas S., et al. "Complement system part I–molecular mechanisms of activation and regulation." *Frontiers in immunology* 6 (2015): 262.

Nilsson, P. H. et al. A conformational change of complement C5 is required for thrombin-mediated cleavage, revealed by a novel ex vivo human whole blood model preserving full thrombin activity. J. Immunol. 207, 1641–1651 (2021)

Nonaka, Masaru, et al. "Opsonic complement component C3 in the solitary ascidian, Halocynthia roretzi." *The Journal of Immunology* 162.1 (1999): 387-391.

Ochsenbein, Adrian F., et al. "Control of early viral and bacterial distribution and disease by natural antibodies." *Science* 286.5447 (1999): 2156-2159.

Peng, M., Niu, D., Wang, F., Chen, Z., Li, J., 2016. Complement C3 gene: expression characterization and innate immune response in razor clam Sinonovacula constricta. Fish Shellfish Immunol. 55, 223–232.

Peng, Maoxiao, et al. "Complement C3 gene: expression characterization and innate immune response in razor clam Sinonovacula constricta." *Fish & shellfish immunology* 55 (2016): 223-232.

Post, Theodore W., et al. "Membrane cofactor protein of the complement system: alternative splicing of serine/threonine/proline-rich exons and cytoplasmic tails produces multiple isoforms that correlate with protein phenotype." *The Journal of experimental medicine* 174.1 (1991): 93-102.

Pouw, R.B., Vredevoogd, D.W., Kuijpers, T.W., Wouters, D., 2015. Of mice and men: the factor H protein family and complement regulation. Mol. Immunol. 67, 12–20.

Ricklin, D., Hajishengallis, G., Yang, K. & Lambris, J. D. Complement: a key system for immune surveillance and homeostasis. Nat. Immunol. 11, 785–797 (2010).

Ricklin, Daniel, et al. "Complement: a key system for immune surveillance and homeostasis." *Nature immunology* 11.9 (2010): 785-797.

Rooney, Isabelle A., John E. Heuser, and John P. Atkinson. "GPI-anchored complement regulatory proteins in seminal plasma. An analysis of their physical condition and the mechanisms of their binding to exogenous cells." *The Journal of clinical investigation* 97.7 (1996): 1675-1686.

Russell, S. (2021). *Artificial Intelligence: A Modern Approach*. Pearson.

Russell, Sarah M., et al. "Tissue-specific and allelic expression of the complement regulator CD46 is controlled by alternative splicing." *European journal of immunology* 22.6 (1992): 1513-1518.

Schmidt, Bela Z., and Harvey R. Colten. "Complement: a critical test of its biological importance." *Immunological reviews* 178 (2000): 166-176.

Shishido, Stephanie N., et al. "Humoral innate immune response and disease." *Clinical immunology* 144.2 (2012): 142-158.

Smith, L. Courtney, et al. "The echinoid complement system inferred from genome sequence searches." *Developmental & Comparative Immunology* 140 (2023): 104584.

Smith, L.C., Azumi, K., Nonaka, M., 1999. Complement systems in invertebrates. The ancient alternative and lectin pathways. Immunopharmacology 42, 107–120.

Smith, L.C., Chang, L., Britten, R.J., Davidson, E.H., 1996. Sea urchin genes expressed in activated coelomocytes are identified by expressed sequence tags. Complement homologues and other putative immune response genes suggest immune system homology within the deuterostomes. J. Immunol. 156, 593–602.

Smith, L.C., Clow, L.A., Terwilliger, D.P., 2001. The ancestral complement system in sea urchins. Immunol. Rev. 180, 16–34.

Smith, Valerie J., Alice Accorsi, and Davide Malagoli. "Hematopoiesis and hemocytes in pancrustacean and molluscan models." *The evolution of the immune system*. Academic Press, 2016. 1-28.

Song, W. C., Sarrias, M. R. & Lambris, J. D. Complement and innate immunity. *Immunopharmacology* 49, 187–198 (2000).

Stephan, A. H., Barres, B. A. & Stevens, B. The complement system: an unexpected role in synaptic pruning during development and disease. Annu. Rev. Neurosci. 35, 369–389 (2012).

Stevens, Beth, et al. "The classical complement cascade mediates CNS synapse elimination." *Cell* 131.6 (2007): 1164-1178.Accorsi, Alice, et al. "Complete Regeneration of a Camera-type Eye in the Research Organism Pomacea canaliculata." *The FASEB Journal* 32 (2018): 232-4.

Stevens, Beth, et al. "The classical complement cascade mediates CNS synapse elimination." *Cell* 131.6 (2007): 1164-1178.

Taylor, Clare T., and Peter M. Johnson. "Complement-binding proteins are strongly expressed by human preimplantation blastocysts and cumulus cells as well as gametes." *Mol Hum Reprod* 2.1 (1996): 52-59.Accorsi, Alice, Enzo Ottaviani, and Davide Malagoli. "Effects of repeated hemolymph withdrawals on the hemocyte populations and hematopoiesis in Pomacea canaliculata." *Fish & Shellfish Immunology* 38.1 (2014): 56-64.

Terwilliger, D.P., Clow, L.A., Gross, P.S., Smith, L.C., 2004. Constitutive expression and alternative splicing of the exons encoding SCRs in Sp152, the sea urchin homologue of complement factor B. Implications on the evolution of the Bf/C2 gene family. Immunogenetics 56, 531–543.

Trouw, Leendert A., Matthew C. Pickering, and Anna M. Blom. "The complement system as a potential therapeutic target in rheumatic disease." *Nature Reviews Rheumatology* 13.9 (2017): 538-547.

V. J. Smith, A. Accorsi, D. Malagoli, in The Evolution of the Immune System: Conservation and Diversification, 1st ed. (Ed: D. Malagoli), Academic Press/Elsevier, Cambridge, Massachusetts, USA 2016, Ch. 1.

View 100 of the world's worst invasive alien species, De- cember 2000. Updated and reprinted version: November 2004. https://www.issg.org/worst100 species.html.

Wallis, Russell. "Interactions between mannose-binding lectin and MASPs during complement activation by the lectin pathway." *Immunobiology* 212.4-5 (2007): 289-299.

Walport, Mark J. "Complement first of two parts." *N Engl J Med.* 344 (2001): 1058-1066.

Yang, Ting-Bao, Zhong-Dao Wu, and Zhao-Rong Lun. "The apple snail Pomacea canaliculata, a novel vector of the rat lungworm, Angiostrongylus cantonensis: its introduction, spread, and control in China." *Hawai'i Journal of Medicine & Public Health* 72.6 Suppl 2 (2013): 23.