Dottorato in Biochimica e Biologia Molecolare
XXVIII Ciclo

Dipartimento di Bioscienze
Università degli Studi di Parma

HABITAT E GENOMA: POSSIBILI ADATTAMENTI DI *HELICOBACTER PYLORI* ALLO STOMACO UMANO

Tutor: chiar.mo prof. Riccardo Percudani

Coordinatore: chiar.mo prof. Andrea Mozzarelli

Dottorando: dott. Pietro Cravedi

2013 - 2015

INDICE DEI CONTENUTI

1 - SCOPO DELLA TESI

Helicobacter pylori è un batterio patogeno che infetta e abita lo stomaco umano. La sua diffusione è ubiquitaria nel mondo ed esso è responsabile di varie patologie sia gastriche che extra-gastriche. Vista l'estrema ostilità del suo habitat, *H. pylori* necessita di specifici adattamenti che lo rendono unico nella sua capacità di tollerare l'estrema acidità dello stomaco umano, oltre ad evitare la risposta immunitaria dell'ospite. In questa tesi verranno adottati degli approcci bioinformatici per cercare di individuare quali possano essere gli adattamenti e le caratteristiche del genoma di questo batterio correlati con la sopravvivenza nell'ambiente gastrico e l'adattamento all'ospite umano.

2 – INTRODUZIONE

Sin dagli albori della batteriologia medica, narra J. Robin Warren l'8 dicembre 2005 nella lezione che tiene alla consegna del premio Nobel, era ritenuta conoscenza comune nonché fatto scontato che non esistessero batteri capaci di crescere nello stomaco umano. Eppure, erano noti organismi, come lieviti o funghi, capaci di crescere nel tessuto gastrico necrotico lasciato da ulcere e tumori. Negli anni '70, però, l'invenzione dell'endoscopio flessibile consentì di cominciare ad effettuare facilmente biopsie gastriche e, quindi, ampliò notevolmente la quantità di campioni disponibili per l'analisi. Inoltre, i campioni prelevati con questa nuova tecnica potevano essere fissati rapidamente tramite immersione in formalina e potevano essere sottoposti a varie indagini istologiche. Nei suoi studi su questi campioni bioptici, Warren notò, tramite l'osservazione al microscopio e la colorazione dei tessuti, la presenza di batteri in uno dei campioni che stava esaminando, piccoli bacilli spiraliformi aderenti alla superficie epiteliale e, spesso, organizzati in palizzate. Questi batteri si attaccavano ai microvilli e spesso appiattivano e distruggevano i microvilli (*Figura 2.1*). Le cellule epiteliali perdevano struttura e assumevano un comportamento ameboide. Analizzando ulteriori campioni emerse che questi batteri erano visibili solo ai bordi dei campioni, dove una striscia di mucosa entrava direttamente in contatto con la formalina e veniva fissata rapidamente. Inoltre, Warren osservò anche una forma degenerata di questi batteri in cui assumevano una forma sferica in regioni più lontane dal bordo del taglio. La presenza di queste zone in cui le due forme si mescolano

probabilmente spiegava l'assenza di precedenti informazioni al riguardo, dal momento che, probabilmente, venivano visti come contaminanti. I batteri osservati da Warren, che lui classificò come *Campylobacter pyloridis* erano stati osservati anche da altri in precedenza (Dunn, Cohen, and Blaser 1997) ma, sia per il fatto che non erano stati potuti essere isolati, sia per la radicata convinzione che le stomaco umano fosse sterile, la significatività della loro presenza era stata ignorata ed erano stati dismessi come contaminanti. In seguito, il batterio fu rinominato *Campylobacter pylori* (per rispettare la grammatica latina) e, poi, *Helicobacter pylori* quando i dati dei sequenziamenti degli rRNA lo hanno collocato in un genere a parte (Goodwin CS, Amstrong JA, Chilvers T 1989).

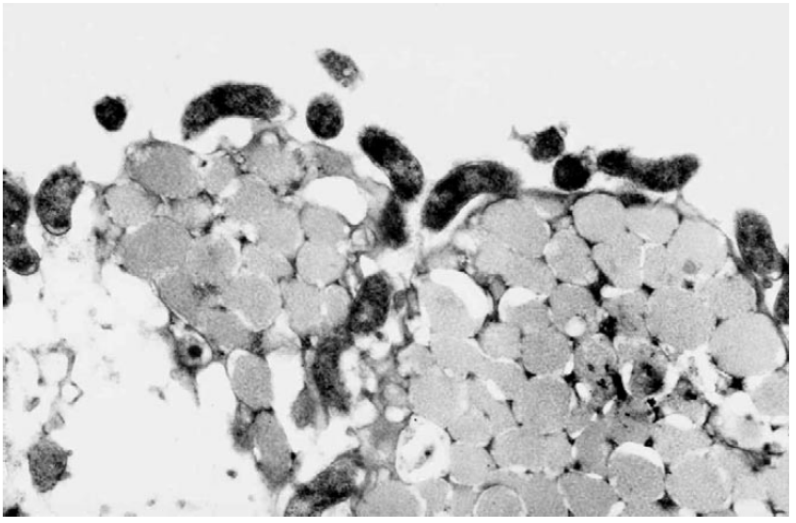

Figura 2.1 *Immagine al microscopio elettronico di due cellule epiteliali gastriche colonizzate da H. pylori (cellule più scure) (Warren 2006)*

A seguito della pubblicazione dei risultati di Marshall e Warren (Dunn et al. 1997), ricercatori da tutto il mondo cominciarono a riportare di aver

trovato il nuovo batterio nei loro campioni di mucosa gastrica. In capo ad alcuni anni diventò evidente che l'infezione di questo batterio era fortemente correlata con l'infiammazione della mucosa gastrica, la gastrite cronica superficiale e la gastrite cronica attiva. Ad oggi, un grande quantità di dati suggerisce che, una volta acquisito, *H. pylori* permanga per il resto della vita del suo ospite, a meno che non venga eradicato tramite terapie antibiotiche (Dunn et al. 1997).

2.1 - EPIDEMIOLOGIA

Helicobacter pylori è diffuso in maniera pressoché capillare in tutto il mondo ed è stato trovato nello stomaco umano in ogni parte del mondo. La prevalenza dell'infezione cambia a seconda dell'area geografica e dello stadio di sviluppo del Paese in questione e dall'intervallo di età dei soggetti considerati. Per i bambini varia dal 4% di prevalenza in Giappone all'82% fra i bambini rifugiati africani. Per gli adulti varia tra il 34% misurato in Svezia e il 60% in Albania, Egitto, Iran, Turchia e Cina. L'infezione sembra essere più diffusa tra gli individui appartenenti a famiglie numerose provenienti dagli strati socioeconomici inferiori della popolazione (Azevedo, Huntington, and Goodman 2009).

2.1.1 – ASSOCIAZIONE CON ALTRE MALATTIE

L'infezione di *Helicobacter pylori* è spesso associata con varie patologie, sia gastriche che extra-gastriche. Una delle associazioni più studiate è quella con l'ulcera peptica. I pazienti infetti da *H. pylori* mostrano un rischio di sviluppare questa patologia che varia dal 3% al 25%. Quasi tutti

i casi di quella forma di ulcera peptica precedentemente ritenuta idiopatica (dal 60 al 95% dei casi totali di ulcera peptica) sono da imputarsi ad *H. pylori* (secondo le conclusioni raggiunte dalla conferenza del NIH del 1990). È stato, inoltre, visto che l'eradicazione del batterio comporta una riduzione significativa della reinsorgenza della patologia. L'infezione da *H. pylori* è inoltre associata a dispepsia e reflusso gastroesofageo e può portare all'insorgenza di linfoma MALT (associato alla mucosa) e adenocarcinoma gastrico. (Kandulski, Selgrad, and Malfertheiner 2008).

H. pylori, infine, può provocare l'insorgenza di patologie extra-gastriche. Tra esse si annoverano la porpora trombocitica idiopatica (per la quale si suppone abbia un ruolo l'attività di CagA) e anemia sideropenica (dovuta alla competizione di *H. pylori* con l'ospite per l'assunzione di ferro). Per contro si è vista una correlazione negativa tra l'infezione da *H. pylori* e l'insorgenza di asma, allergie e altre malattie atopiche (Kandulski et al. 2008).

2.2 – CARATTERISTICHE

2.2.1 – ε-PROTEOBATTERI

Helicobacter pylori appartiene agli ε-proteobatteri, una delle cinque classi in cui è suddiviso il phylum dei Proteobatteri. Si ritiene che gli ε-proteobatteri siano comparsi sulla Terra tra 1,3 e 2 miliardi di anni fa (Campbell et al. 2006). Questa classe è suddivisa in due ordini: i Nautiliales (comprendenti i generi *Nautilia, Caminibacter* e *Lebetimonas*)

e i Campylobacterales (comprendenti le famiglie delle Campylobacteraceae, Helicobacteraceae e Hydrogenimonaceae) (Campbell et al. 2006). Questa classe di batteri abita una vasta gamma di habitat, che varia dall'apparato digerente degli animali a riserve idriche, acque di scarico, depositi di oli fossili e fonti idrotermali (Gupta 2006). Negli habitat che occupano e in cui prosperano, gli ε-proteobatteri rivestono un ruolo ecologico importante per via della loro biomassa, del loro ritmo di crescita e delle loro capacità di adattarsi rapidamente a quelli che, altrimenti, sarebbero ambienti ostili, il che consente loro di colonizzare nicchie ecologiche che altri organismi disdegnano. Essi, inoltre, rivestono l'importante ruolo di fonti primarie di carbonio, azoto e zolfo organicati in molti ecosistemi, in virtù del metabolismo chemolitoautotrofico che gli ε-proteobatteri colonizzanti gli ambienti terrestri e marini possiedono (Campbell et al. 2006).

Molti ε-proteobatteri rivestono un'importanza notevole dal punto di vista clinico. Oltre ad *Helicobacter pylori*, molti Campylobacterales sono patogeni dell'uomo o degli animali. *Helicobacter hepaticus* è considerato essere un fattore di predisposizione allo sviluppo di tumore gastrico nell'uomo e tumore epatico nei roditori. *Campylobacter jejuni* può condurre a degenerazione muscolare causando l'insorgenza della sindrome di Guillain-Barré e, insieme a *Campylobacter coli* è una delle principali cause al mondo di diarrea e altre malattie legate all'ingestione di cibo contaminato (Gupta 2006).

2.2.2 - MORFOLOGIA

Helicobacter pylori è un batterio gram-negativo micro-aerofilo, flagellato. Gli esemplari isolati da campioni di biopsie gastriche hanno una forma a spirale con estremità arrotondate, gli esemplari coltivati su terreno solido assumono prevalentemente una forma a bastoncino mentre a seguito di una coltura prolungata in terreno solido o liquido la forma prevalente è quella coccoide (Figura 2.2). Gli esemplari da biopsie gastriche sono lunghi da 2,5 a 5,0 µm e larghi da 0,5 a 1,0 µm e sono dotati di un numero variabile da 4 a 6 di flagelli unipolari inguainati. Ogni flagello è lungo circa 30 µm e spesso circa 2,5 nm. Gli esemplari di *H. pylori* rinvenuti nei campioni gastrici possono essere visti collegati ai microvilli da estensioni filamentose del glicocalice (Dunn et al. 1997).

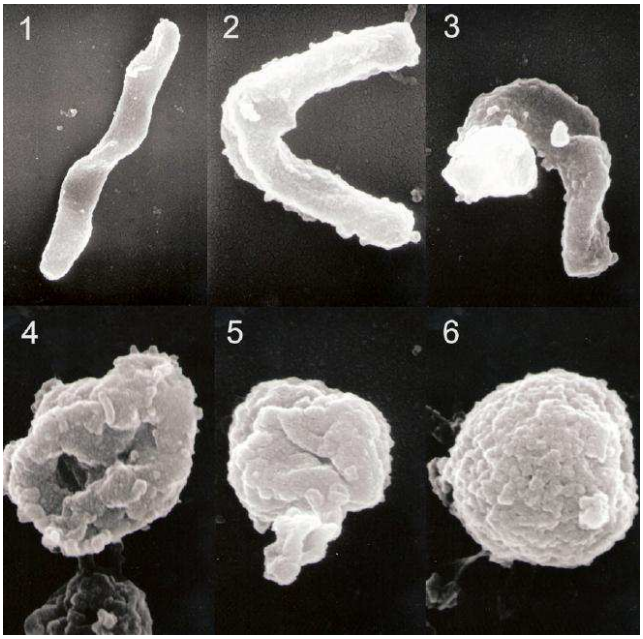

Figura 2.2 *Immagini al microscopio elettronico di H. pylori in transizione dalla forma a bacillo a quella coccoide(Dus et al. 2013)*

2.2.3 - GENOMA

Il genoma di *Helicobacter pylori* (ceppo 26695) è stato pubblicato dal TIGR nel 1997 (Tomb et al. 1997). Il genoma di questo ceppo (da noi considerato come ceppo di riferimento) contiene 1,67 Mb ($1,67*10^6$ paia di basi) circa e ha un contenuto in G+C del 39%. Al 2013 sono presenti in banca dati 43 genomi completi e 198 sequenze genomiche grezze di vari ceppi di *H. pylori*, provenienti da varie parti del mondo (Ahmed et al. 2013). Basandosi su tali dati la dimensione media del genoma di *H. pylori* è stata stimata attorno ad 1,62 (1,51-1,71) Mb, con un contenuto medio di GC del 38,92% (38,40-39,30%) e un numero medio di ORF di 1590 (1429–1749) codificanti per 1532 (1382–1707) proteine. Il genoma di *H. pylori* è caratterizzato da un'elevata diversità tra i vari ceppi rispetto a quanto osservato per altre specie. Questo può essere dovuto alla scarsa presenza di competitori, visto che il microrganismo si trasmette principalmente nell'ambito famigliare e, quindi, in un ospite c'è normalmente un solo ceppo di batterio. Ciò può portare ad un rilassamento della pressione selettiva, in virtù del fatto che ad *H. pylori* sia necessario solo perdurare nella nicchia gastrica senza dover competere con altri microrganismi. Un'altra ragione potrebbe risiedere in una diversità nei tratti umani che potrebbe essere importante per la sopravvivenza del batterio, come l'efficacia del sistema immunitario e l'abbondanza di recettori a cui *H. pylori* aderisce, il che selezionerebbe a favore della divergenza dei vari ceppi (Kersulyte et al. 2000).

Nel genoma di *H. pylori* sono presenti delle zone di plasticità, ovvero delle regioni a livello delle quali è più frequente il trasferimento

orizzontale di geni tra i batteri. Queste regioni contengono geni provenienti da un set fisso che vengono scambiati mediante trasferimento orizzontale tra diversi ceppi quando vengono in contatto in seguito ad un'infezione mista (Fischer et al., 2010, 2014).

Queste caratteristiche di elevata variabilità genomica unitamente al fatto che il tasso di mutazione spontanea in *H. pylori* sia stato visto essere svariati ordini di grandezza superiore a quello medio di *E. coli* e vicino a quello di quei ceppi di *E. coli* difettivi nei sistemi di riparazione del DNA (Kraft and Suerbaum 2005) suggeriscono un quadro in cui ogni individuo infetto abbia, all'atto pratico, il suo proprio ceppo di *Helicobacter pylori* (Suerbaum and Josenhans 2007)

Una caratteristica peculiare di *H. pylori* è lo scarso numero di proteine regolatrici predette. Sono stati, infatti, identificati solo 32 prodotti genici classificabili come aventi un possibile ruolo regolatore. Per dare un termine di paragone, questo numero è circa la metà del numero di geni regolatori identificati in *H. influenzae* (il cui genoma ha dimensioni paragonabili a quelle del genoma di *H. pylori*) e meno di un quarto di quelli predetti per *E. coli* (Scarlato et al. 2001). Di questi, solo 17 loci sono stati predetti avere un ruolo nella regolazione trascrizionale di *H. pylori*, tra cui il regolatore Fur, centrale nella risposta di *H. pylori* allo stress ossidativo (Pelliciari et al. 2015) e nell'assunzione del ferro, e il repressore HspR, coinvolto nella regolazione dell'espressione degli chaperones GroESL e DnaK (Scarlato et al. 2001).

2.2.4 - HABITAT

La nicchia occupata da *Helicobacter pylori* è un ambiente estremamente inospitale. Lo stomaco umano, infatti, è caratterizzato dall'essere un ambiente estremamente acido con un pH che varia tra 1 e 2, grazie alla secrezione di acido cloridrico da parte della mucosa gastrica, a cui si aggiunge la secrezione della proteasi pepsina. Ciò costituisce una forte barriera contro gran parte dei microrganismi che da lì cercano di entrare nell'ospite umano (Smith 2003), e ciò richiede meccanismi specifici anche solo per sopravvivere le poche ore necessarie ad oltrepassare lo stomaco e a raggiungere il distretto preferito. Tali sistemi possono variare dall'infezione con numeri talmente vasti di microrganismi che, anche in mancanza di adattamenti specifici, risulta statisticamente probabile che una carica batterica adeguata riesca a sopravvivere alla barriera gastrica (è questo il caso di *Vibrio cholerae*), a raffinati sistemi volti ad aumentare le probabilità di sopravvivenza del singolo organismo. In ciò è maestro *Helicobacter pylori* che non si limita a sopravvivere il tempo necessario a transitare nello stomaco, bensì fa di esso il suo habitat privilegiato (Foster 2004).

2.2.5 – FATTORI DI VIRULENZA

Data l'estrema ostilità dell'ambiente che *H. pylori* colonizza, il batterio ha dovuto sviluppare, nel corso dell'evoluzione, dei meccanismi che gli consentissero di sopravvivere nello stomaco umano, sia resistendo all'intrinseca ostilità dell'ambiente, sia eludendo l'azione del sistema immunitario umano. Oltre ad un impressionante set di strumenti che

influenzano direttamente le cellule della mucosa gastrica facilitando l'ancoraggio di *H. pylori* al suo ospite, il batterio possiede una gamma di adattamenti metabolici che gli consentono di alterare il microambiente in cui risiede a suo vantaggio.

2.2.5.1 - Adesine

Le adesine sono proteine della superficie della cellula che consentono al batterio di aderire alle cellule dell'ospite, il che rappresenta il primo passo nella colonizzazione e nella patogenesi. Le adesine di *H. pylori* sono coinvolte nella fase precoce e nella fase cronica dell'infezione e appartengono ad una grande famiglia di proteine della membrana esterna, la famiglia Hop. Questa famiglia comprende le principali adesine del batterio: BabA (che lega gli antigeni del gruppo sanguigno), SabA (che lega l'acido sialico della matrice extracellulare), AlpA/B (leganti la laminina), HopZ e OipA (Kalali et al. 2014).

2.2.5.2 - Isola di patogenicità cag

L'isola di patogenicità *cag* (PAI) è un elemento inserzionale di circa 40 Kb contenente 32 geni che codificano per un sistema di secrezione di tipo IV. PAI è un elemento di virulenza noto e ben caratterizzato, presente in circa il 60-70% dei ceppi occidentali di *H. pylori* e nella quasi totalità dei ceppi orientali. Ceppi di *Helicobacter pylori* cag^- vengono trovati prevalentemente nello strato del muco, mentre i ceppi cag^+ risiedono prevalentemente adiacenti o aderenti alle cellule dell'epitelio gastrico. Ciò indica che la presenza o l'assenza di *cag* influenza la topografia della

colonizzazione. Il sistema di secrezione codificato da *cag* consente il trasferimento di molecole batteriche nelle cellule dell'ospite. Una di queste molecole che vengono trasferite è il prodotto di uno dei geni *cag*, CagA, i cui effetti di iperproliferazione dell'epitelio gastrico lo classificano come un'oncoproteina. CagA possiede dei motivi di fosforilazione EPIYA (-A, -B, -C, -D) che si differenziano fra loro in merito agli aminoacidi fiancheggianti e la cui distribuzione varia in base all'area geografica di provenienza del batterio.

Una volta introdotta nelle cellule dell'ospite, CagA viene fosforilata ad opera delle chinasi umane e, a sua volta, attiva ulteriori cascate chinasiche, risultando in sostanziali interferenze nel sistema di trasduzione del segnale. Uno degli effetti di questa interferenza è l'attivazione di circuiti di feedback negativo che servono a regolare la quantità di CagA fosforilata presente nella cellula dell'ospite. CagA fosforilata induce nelle cellule dell'ospite un fenotipo morfologico detto "a colibrì", in cui le cellule si allungano e si disperdono, oltre ad altre possibili aberrazioni morfologiche.

Anche non fosforilata, CagA esercita vari effetti sulle cellule dell'epitelio gastrico che contribuiscono alla patogenesi, portando ad una distruzione dei complessi delle giunzioni apicali. CagA non fosforilata induce un assemblaggio incompleto delle giunzioni strette nei siti di adesione del batterio. Essa, inoltre, disturba la formazione delle giunzioni aderenti (portando ad una perdita della funzione di barriera del tessuto e della polarità cellulare) e induce risposte pro-mitotiche e pro-infiammatorie (Noto and Peek 2012)

2.2.5.3 - Ureasi

L'estremamente basso pH gastrico rappresenta la principale difficoltà che *H. pylori* deve affrontare per sopravvivere nello stomaco. *H. pylori* produce grandi quantità dell'enzima ureasi (Mobley et al. 1988), in grado di produrre ammoniaca e CO_2 a partire da urea. *H. pylori* espone questo enzima sulla sua superficie e lo impiega per generare un gradiente di pH e mantenere attorno a sé un microambiente più ospitale. L'ureasi processa le piccole quantità di urea presenti nello stomaco per produrre ammoniaca che si accumula intorno al batterio. L'ammoniaca, quindi, consuma gli ioni H^+ circostanti causando un innalzamento locale del pH. L'ureasi di *H. pylori*, oltre ad essere espressa in grandi quantità, è anche estremamente efficiente e presenta una K_m notevolmente più bassa (e quindi un'elevata affinità) e una V_{max} notevolmente più elevata (quindi un'efficienza maggiore) rispetto alle ureasi di altre specie batteriche. Essa è, quindi, in grado di legare urea a basse concentrazioni e processarla con grande efficienza producendo, quindi, ammoniaca sufficiente a mantenere il microambiente a pH elevato attorno a sé (Chen et al. 1997).

L'ureasi di *H. pylori* è un enzima dimerico, composto da una subunità grande (UreA) e una subunità piccola (UreB). Sono inoltre presenti tre geni accessori (*ureF,G,H*), essenziali per l'attività dell'enzima che giocano un ruolo importante nell'attivazione dell'apoenzima e nell'inserzione degli ioni Ni^{2+} nel sito attivo dell'enzima nascente durante il suo assemblaggio. La proteina è localizzata nel citoplasma nella fase esponenziale precoce, mentre è, invece, associata alla superficie della

cellula nella fase esponenziale tardiva e deve essere così localizzata per conferire la resistenza all'ambiente acido al batterio. Dal momento che l'ureasi di *H. pylori* non presenta le caratteristiche tipiche delle proteine normalmente indirizzate alla membrana o alla via secretoria (segnali terminali) è stato proposto che il meccanismo di localizzazione sia un'autolisi altruistica, ovvero un meccanismo di autolisi geneticamente programmata attuato da una frazione della popolazione batterica con lo scopo di rilasciare l'ureasi concentrata nel citoplasma e consentire agli altri microrganismi di raccoglierla sulla superficie della membrana (Dunn and Phadnis 2000).

2.2.5.4 - Citotossina vacuolante VacA

Sin dal 1988 è noto che *Helicobacter pylori* secerne una tossina in grado di indurre degenerazione vacuolare nelle cellule epiteliali. Questa proteina, chiamata VacA, presenta un notevole polimorfismo nelle varianti in cui si presenta nei vari ceppi del patogeno. Il suo preciso ruolo nell'infezione non è ancora stato determinato, ma si suppone abbia un ruolo nella colonizzazione e nella persistenza di *H. pylori* nello stomaco umano e che contribuisca al danno all'epitelio gastrico.

Il gene *vacA* è presente in tutti i ceppi di *H. pylori* in svariate varianti alleliche. La proteina VacA da esso codificata dispone di due domini, uno avente la funzione vacuolante e l'altro con una funzione di legame ai recettori della cellula bersaglio. La tossina induce degenerazione vacuolare, ma non porta rapidamente la cellula alla morte. La proteina necessita di un ambiente acido per attivarsi e viene internalizzata dalle

cellule dell'ospite tramite endocitosi mediata da recettore. L'azione di VacA, inoltre, causa un deficit nell'azione endolisosomiale.

Sebbene sia stato proposto un ruolo di VacA primariamente dannoso nei confronti dell'epitelio gastrico, evidenze sperimentali e di considerazioni evolutive (non è ben chiaro perché un patogeno dovrebbe possedere una proteina che abbia l'esclusiva funzione di infliggere all'ospite un danno a cui il batterio non sembra trarre particolare beneficio) portano ad ipotizzare altri ruoli per questa proteina. È stato proposto che l'azione di VacA sia vantaggiosa nella sopravvivenza del batterio, in quanto i danni all'epitelio gastrico potrebbero portare ad un migliore afflusso di nutrienti dalla mucosa allo strato di muco. L'alterazione delle vie endocitotiche, inoltre, introdurrebbe degli ostacoli nei processi di presentazione dell'antigene, il che aiuterebbe il patogeno a mascherarsi dal sistema immunitario dell'ospite (Atherton et al. 2001).

2.2.5.5 - Protezione dalle specie reattive dell'ossigeno

Helicobacter pylori, nel colonizzare lo stomaco umano, si trova ad affrontare un significativo stress ossidativo. Le specie reattive dell'ossigeno (ROS), dannose a livello del DNA e dei lipidi di membrana, vengono normalmente prodotte durante il metabolismo respiratorio. Esse possono, inoltre, essere prodotte da agenti esterni, come agenti chimici e radiazioni. Le ROS, infine, sono prodotte dal burst ossidativo impiegato come difesa dai leucociti polimorfonucleati (PMN). L'infezione da *H. pylori*, oltre ad innescare la risposta infiammatoria e, quindi, l'attivazione dei PMN e l'innesco del burst ossidativo (Ramarao, Gray-

Owen, and Meyer 2000), sembra stimolare la produzione di ioni superossido ($O_2^{\cdot -}$) da parte delle cellule epiteliali gastriche (Bagchi, Bhattacharya, and Stohs 1996). È interessante notare come, a differenza di altri patogeni che cercano di evitare la letale aggressione a base di ROS da parte dei PMN, *H. pylori* sembra avere evoluto un sistema per innescare e sopravvivere al burst ossidativo, probabilmente per trarre vantaggio dal fatto che i ROS che vengono impiegati per combattere il patogeno sono tossici anche per le cellule dell'ospite e inducono danni all'epitelio della mucosa, cosa che costituirebbe un vantaggio per il batterio (Ramarao et al. 2000).

Sono due gli enzimi principali che in *H. pylori* sono responsabili della detossificazione delle ROS: la superossido dismutasi (SOD) e la catalasi (KatA). La prima catalizza la reazione di dismutazione dello ione superossido a acqua ossigenata e ossigeno molecolare ($O_2^{\cdot -} + 2H^+ \rightarrow H_2O_2 + O_2$), mentre la seconda catalizza la degradazione dell'acqua ossigenata ad acqua e ossigeno molecolare ($2H_2O_2 \rightarrow 2H_2O + O_2$). La catalasi di *H. pylori* è principalmente attiva sia nel citoplasma che nel periplasma ed è estremamente stabile anche ad alte concentrazioni di perossido di idrogeno (Wang, Alamuri, and Maier 2006). Per quanto riguarda la superossido dismutasi, nei batteri esistono quattro tipi di SOD, classificate in base al cofattore metallico che legano: Fe-SOD, Zn-SOD, Mn-SOD e Cu-SOD. Di queste quattro, *H. pylori* sembra possedere solo la superossido dismutasi a ferro (Fe-SOD), espressa nel citoplasma (Hazell, Harris, and Trend 2001).

3 – ANALISI IN SILICO DEL METALLOPROTEOMA DI H. PYLORI

3.1 - INTRODUZIONE

3.1.1 - COMPETIZIONE CON L'OSPITE

L'infezione da *Helicobacter pylori* è spesso associata con patologie extragastriche da carenza di metalli. Molti studi clinici hanno riportato l'insorgenza di anemia sideropenica nei pazienti infetti di *H. pylori*. Tale anemia spesso regrediva in seguito alla riuscita eradicazione del batterio (Barabino 2002). Il sequestro del ferro da parte dell'organismo è una delle prime difese dell'organismo che cerca di negare al patogeno il metallo essenziale. Tale sequestro avviene ad opera della ferritina e della lattoferrina che vengono rilasciate nella mucosa e sequestrano il ferro impedendone l'assorbimento da parte del batterio. Ne consegue che *H. pylori* ha dovuto sviluppare dei meccanismi per l'assunzione efficiente del ferro in modo da combattere questi meccanismi (Barabino 2002), dal momento che il ferro è richiesto in vari enzimi, tra cui la catalasi e la superossido dismutasi, importanti nella risposta allo stress da specie reattive dell'ossigeno (Hazell et al. 2001).

Un altro metallo di cui *H. pylori* ha disperato bisogno è il nickel. Esso è un cofattore dell'ureasi, indispensabile per la sopravvivenza all'ambiente acido (Chen et al. 1997). Essendo l'ureasi prodotta in grandi quantità, *H. pylori* ha dovuto sviluppare dei sistemi per accaparrarsi il nickel presente

nella mucosa gastrica. *H. pylori*, inoltre, ha un altro importante enzima nickel-dipendente, la [NiFe]-idrogenasi (Maier et al. 1996). Questo enzima consente al batterio di utilizzare l'idrogeno molecolare come fonte di potere riducente, catalizzandone la scissione della molecola. L'idrogeno molecolare è presente nello stomaco umano in grandi quantità e la capacità di *H. pylori* di sfruttarlo gli conferisce un notevole vantaggio nella colonizzazione dello stomaco umano (Maier 2005).

3.1.2 - OMEOSTASI DEI METALLI

Al fine di competere efficientemente con l'ospite per l'assunzione di ioni metallici, *H. pylori* ha evoluto dei raffinati sistemi regolativi e di trasporto per ottenere efficacemente i micronutrienti metallici necessari alla sua sopravvivenza e persistenza nell'ospite. Inoltre, dal momento che la disponibilità degli ioni metallici nello stomaco umano non è costante, il batterio deve essere in grado di far fronte sia a condizioni di carenza di metalli, sia a situazioni in cui i metalli sono in pericolosa sovrabbondanza

La regolazione del metabolismo di nickel e ferro in *H. pylori* è regolata dai due regolatori trascrizionali NikR e Fur che gestiscono un raffinato circuito di feedback che regola l'omeostasi dei metalli e l'espressione delle metalloproteine in base alla disponibilità di ferro e nickel (Figura 3.1). In *H. pylori*, NikR è un regolatore pleiotropico che controlla operoni coinvolti in varie risposte, tra cui l'assunzione del nickel, la risposta allo stress, i sistemi di detossificazione e l'espressione dell'ureasi. NikR è in grado di legare nickel ed è un sensore della concentrazione intracellulare di nickel, la cui attività dipende dalla disponibilità del metallo. Fra i geni

dell'omeostasi dei metalli controllati da NikR c'è la permeasi NixA (Mobley, Garner, and Bauerfeind 1995) che viene repressa di NikR in abbondanza di nickel intracellulare e de-repressa in carenza di nickel. In presenza di nickel, inoltre, NikR induce l'espressione dell'ureasi (Danielli and Scarlato 2010). Oltre a NixA (indotto da basso pH e represso da alte concentrazioni di nickel), *H. pylori* ha altri sistemi di importazione del nickel, tra cui, probabilmente, il sistema di trasporto ABC costituito da FecCDE/CeuB (de Reuse, Vinella, and Cavazza 2013).

Il regolatore principale dell'omeostasi del ferro, invece, è il repressore Fur. In *H. pylori* Fur, a seconda della concentrazione di ferro intracellulare, varia la sua organizzazione quaternaria a seconda che leghi o meno il metallo. Fur di *H. pylori* (HpFur) agisce da repressore sia su geni repressi in abbondanza di ferro, sia su geni repressi in carenza di ferro. Per esempio, in abbondanza di ferro, Fur reprime la trascrizione di *frpB*, un gene per l'acquisizione del ferro, mentre in carenza di ferro reprime la trascrizione del gene *pfr* codificante per una ferritina che deve essere espressa solo in abbondanza di ferro (Agriesti et al. 2014). Inoltre Fur media la risposta allo stress ossidativo reprimendo in carenza di ferro e di stress ossidativo l'espressione di geni per detossificazione delle specie reattive dell'ossigeno, come *sodB*, codificante per la superossido dismutasi (Pelliciari et al. 2015).

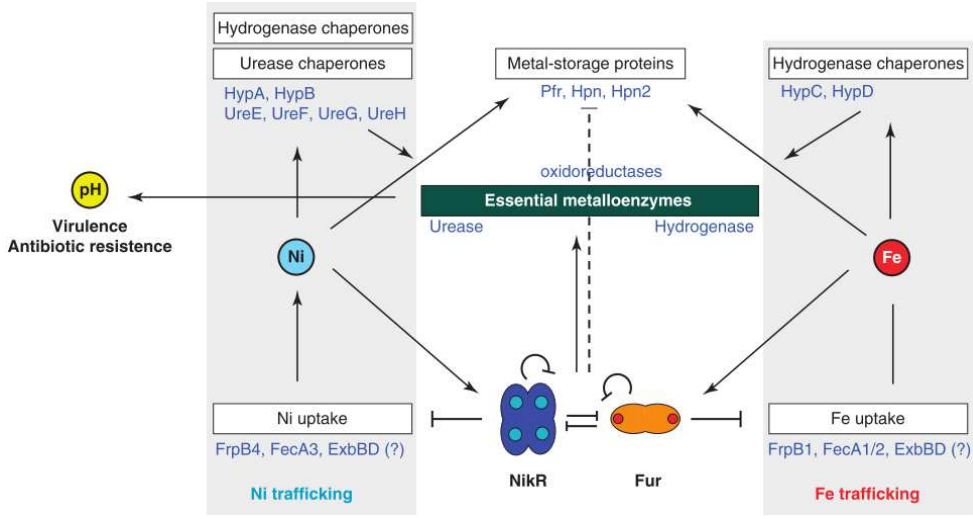

Figura 3.1 *Circuiti regolatori controllati da Fur e NikR (Danielli and Scarlato 2010)*

3.1.3 - PREDIZIONE DI LEGAME AI METALLI

Vista l'importanza rivestita dalle metalloproteine in *H. pylori* come in qualsiasi altro organismo, è evidente l'utilità di sistemi per identificare le metalloproteine su larga scala. La spettroscopia di assorbimento ai raggi X è stata usata come strumento per misurare il contenuto in metalli di transizione in ampi set di proteine (Shi et al. 2011), ma è uno strumento che si presta male all'analisi di proteomi completi, dal momento che richiede che le proteine da analizzare siano clonate, espresse e purificate, operazioni non sempre banali. È, pertanto, indubbia l'utilità di strumenti bioinformatica capaci di processare le sequenze di genomi o proteomi completi e restituire delle predizioni il più possibile accurate su quali proteine leghino quali metalli e in che siti. Ciò consentirebbe di identificare nuove proteine coinvolte nei processi di acquisizione o omeostasi dei metalli o coinvolte nel processo di infezione.

Una strategia si basa sulla conoscenza a priori delle caratteristiche dei siti di legame ai metalli. Utilizzando database come Pfam (Finn et al. 2014) contenenti le descrizioni dei domini di legame ai metalli finora caratterizzati è possibile effettuare delle ricerche nel genoma tradotto di un organismo e identificare quali proteine possiedono questi domini e, quindi, quali metalli legano. Questo approccio può essere raffinato confrontando il database Pfam con altri database (come PDB, SCOP e CATH) in modo da raffinare le annotazioni di domini e ottenere una predizione migliore basata su domini ben identificati. A ciò si può aggiungere la ricerca di motivi di legame ai metalli, ossia di sequenze specifiche di aminoacidi viste in molte proteine coinvolte nel legame ai metalli (Bertini and Cavallaro 2010). Lo svantaggio di questo approccio è che con esso non è possibile identificare nuovi motivi o domini di legame ai metalli e la predizione è subordinata alla completezza e all'accuratezza delle annotazioni presenti in banca dati.

L'altro approccio è di tipo puramente predittivo e consiste nell'utilizzare programmi che, a partire dalla sequenza, identifichino quali residui possano coordinare un atomo di metallo e assegnino ad ogni residuo una predizione che indichi se il residuo è previsto essere libero o impegnato in una coordinazione con uno ione metallico. Esistono vari predittori, i quali si concentrano principalmente sull'identificare le istidine e le cisteine impegnate nel legame ai metalli. Uno di essi è MetalDetector (MD), il quale applica algoritmi di reti neurali per generare le sue predizioni. Esso è in grado di identificare quali istidine e cisteine possano essere coinvolte nel legame ai metalli e fornisce anche una predizione

sulla geometria del sito di legame, indicando quali residui legano quali ioni metallici nel caso in cui la proteina leghi più di un atomo di metallo. MD è, inoltre, in grado di predire la presenza non solo di uno ione metallico da solo, ma anche qualora esso sia incluso in una molecola di eme o in un centro ferro-zolfo (Passerini, Lippi, and Frasconi 2011).

3.2 - MATERIALI E METODI

3.2.1 - OTTENIMENTO DELLE SEQUENZE PROTEICHE

Il proteoma di *H. pylori*, le sequenze RefSeq e il file con le sequenze in formato GenBank su cui è stata effettuata la ricerca per parole chiave sono stati ottenuti dal sito FTP dell'NCBI (ftp://ftp.ncbi.nlm.nih.gov).

3.2.2 - PREDIZIONE DI LEGAME AI METALLI

Per la predizione di legame ai metalli è stato usato il programma MetalDetector (Passerini et al. 2011). La versione scaricabile, operante su una singola sequenza, è stata implementata in uno script in linguaggio Perl. Lo script divide il proteoma di *H. pylori* nelle sequenze proteiche che lo compongono. Successivamente, per ogni sequenza, viene usato PSI-BLAST (Altschul et al. 1997) (con le opzioni *-e 1e-5 -j 2 -a 8*) usando come database le sequenze RefSeq di GenBank per produrre una matrice di sostituzione posizione specifica per ogni proteina. La matrice viene richiesta da MetalDetector per formulare la predizione. Su ogni sequenza viene quindi effettuata la predizione tramite MetalDetector.

3.2.3 - FILTRO DEI RISULTATI

Dal momento che la predizione di MetalDetector si basa sull'analisi della conservazione dei residui secondo quanto indicato nelle matrici di sostituzione, al fine di eliminare il bias introdotto da quelle proteine i cui unici risultati di BLAST erano proteine identiche, è stato applicato un filtro sui risultati che scarti tutti quei risultati per cui MetalDetector non

abbia individuato almeno 2 aminoacidi leganti i metalli che siano conservati almeno al 95% (nelle sequenze omologhe trovate da PSIBlast) e per i quali il rapporto tra la percentuale di conservazione del residuo e la percentuale di conservazione media di tutti i residui calcolata sulla matrice di sostituzione sia maggiore o uguale ad 1,7.

Ai risultati filtrati vengono attribuite le annotazioni attribuite alle relative hits di BLAST.

3.2.4 - RICERCA PER PAROLE CHIAVE

Un script in linguaggio Perl è stato utilizzato per interrogare il file contenente tutte le sequenze codificanti di *H. pylori* in formato GenBank e reperire gli accession numbers delle sequenze la cui descrizione contenga riferimenti a metalli (parole chiave: *metal, zinc, Zn, iron, Fe, magnesium, Mg, nickel, Ni, Calcium, Ca, manganese, Mn, copper, Cu*).

3.2.5 - ANALISI MANUALE

Le sequenze appartenenti a proteine ipotetiche che hanno passato il filtro sono state sottoposte ad un'analisi manuale utilizzando seguenti strumenti per formulare ipotesi sulla loro possibile funzione:

- PSI-BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi)
- HHpred (http://toolkit.tuebingen.mpg.de/hhpred)
- String (http://string-db.org/) per l'analisi di associazione genica
- Microbes Online (http://www.microbesonline.org/) per l'analisi dell'intorno genico

- PSortB (http://www.psort.org/psortb/) per la predizione di localizzazione cellulare
- TMPred (http://www.ch.embnet.org/software/TMPRED_form.html) e TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) per la predizione di tratti transmembrana
- SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP-3.0/) per la predizione del peptide segnale

Sono, inoltre, state consultate le banche dati Pfam (http://pfam.xfam.org/) e Prosite (http://prosite.expasy.org/).

Ove ritenuto interessante, dei modelli 3D della struttura della proteina sono stati realizzati per homology-modelling usando il server Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) e analizzati con PyMol (https://www.pymol.org/).

3.3 - RISULTATI

I risultati completi in formato Excel sono disponibili online sulla cartella Dropbox (Link in tabella Supplementare S 6.1).

3.3.1 - IDENTIFICAZIONE DELLE METALLOPROTEINE

Abbiamo implementato MetalDetector (Passerini et al. 2011) in una procedura automatizzata in Perl per cercare i possibili siti di coordinazione di metalli che includessero cisteine o istidine nella collezione completa delle proteine di *Helicobacter pylori*. La ricerca ha individuato 928 sequenze contenenti almeno una cisteina o istidina a cui MetalDetector (MD) ha assegnato una predizione di partecipazione in un sito di coordinazione (Tabella 3.1). A questi risultati sono state assegnate le annotazioni assegnate alla sequenza stessa e ai suoi omologhi in RefSeq, raggruppando le annotazioni identiche e assegnando ad ogni annotazione un numero che indica quante volte quell'annotazione compare tra le sequenze omologhe. Di queste 928 sequenze, a 549 è stata attribuita una funzione mediante le annotazioni della banca dati mentre le restanti 379 sono annotate come proteine ipotetiche (Figura 3.1).

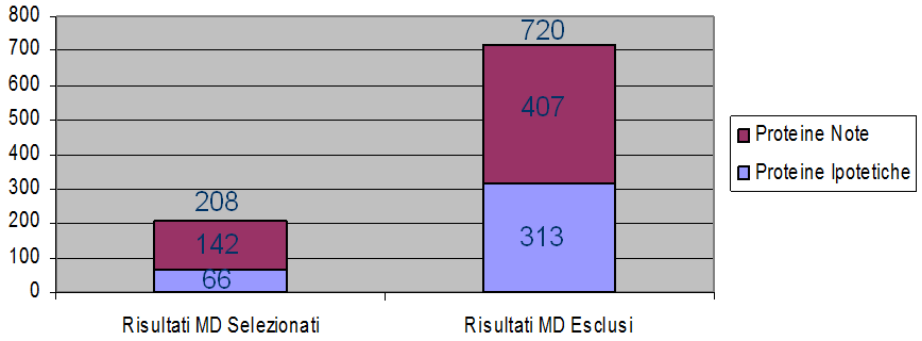

Figura 3.1 *Ripartizione dei risultati di MetalDetector. I risultati selezionati sono quelli che hanno passato il filtro sulla conservazione relativa.*

Posizione	Residuo	Predizione	Sito
92	H		
94	H	M	1
108	H	M	2
195	H	M	1
210	H	M	2
218	H		
262	C		
273	C		
275	H		
296	H		
347	H		

Tabella 3.1 *Esempio di predizione di MetalDetector. Nella colonna "Predizione" M indica che il residuo è predetto coordinare un metallo, diversamente non viene indicato niente. Nella colonna "Sito" il numero indica di che sito di coordinazione fa parte il residuo. Numeri uguali indicano residui facenti parte dello stesso sito.*

Dal momento che la ricerca con MD si basa sull'analisi della conservazione dei residui dedotta da una matrice di punteggi posizione-specifici (PSSM) calcolata da PSI-BLAST, nella predizione viene introdotto

un pregiudizio. La PSSM viene costruita usando PSI-BLAST, quindi analizzando l'allineamento tra la sequenza oggetto dell'analisi e le sequenze ad essa omologhe presenti in banca dati (RefSeq nella nostra ricerca). L'attribuzione dei punteggi è, quindi, influenzata dal risultato della ricerca di omologia e, nello specifico, dalla quantità e dalla varietà di sequenze che vengono individuate. Da ciò ne consegue che, nel caso in cui per la sequenza query venga individuato come unico risultato la sequenza stessa, la matrice riporterà per tutte le posizioni il 100% di conservazione, il che, chiaramente, influenzerà la predizione di MetalDetector che potrebbe attribuire erroneamente una predizione di appartenenza errata ad un sito di coordinazione ad un residuo, interpretando erroneamente l'elevato punteggio di conservazione del residuo. Una situazione simile si ha quando la ricerca di omologia trova più risultati, ma essi sono sequenze con altissima identità con la sequenza di partenza.

Per ovviare al problema abbiamo applicato un filtro basato sulla conservazione relativa dei residui. Avendo a disposizione le PSSM usate nella ricerca (dal momento che avevamo dovuto costruirle per fornirle in input a MD) abbiamo potuto, per ogni sequenza, calcolare la conservazione media dei residui su tutta la matrice, il che è indice della varietà delle sequenze omologhe trovate. Sequenze con omologhi identici o con come unico omologo la sequenza stessa mostrano una conservazione media su tutta la matrice più alta rispetto a sequenze con molti omologhi con percentuale di identità bassa rispetto alla sequenza query. Sempre dalla matrice abbiamo ottenuto i punteggi di

conservazione per i singoli residui a cui MD aveva attribuito una predizione di partecipazione ad un sito di coordinazione. Nel nostro filtro vengono considerati come veri positivi quei residui che MD ha predetto partecipare ad un sito di coordinazione e per i quali si ha un punteggio di conservazione assoluto del 95% (ovvero quei residui sono conservati in almeno il 95% delle sequenze omologhe trovate) e un punteggio di conservazione relativo di 1,7, calcolato come il rapporto tra il punteggio di conservazione assoluto e il punteggio di conservazione medio della PSSM di quella sequenza; questo limite esclude quei risultati per cui la concentrazione assoluta, pur essendo alta, non si discosta sensibilmente dalla conservazione media di tutti i siti, la quale è alta quando la PSSM è costruita con sequenze identiche o quasi. Stanti queste condizioni per considerare la predizione di MD corretta, abbiamo escluso dai risultati grezzi quelle sequenze che non contenevano almeno due residui classificati come veri positivi.

Delle 928 proteine inizialmente trovate, 208 hanno superato il filtro (Tabella Supplementare S3.1). Di esse, 142 hanno un'annotazione funzionale attribuita e 66 sono annotate come proteine ipotetiche (Figura 3.1).

Le 208 proteine che hanno superato il filtro sono state usate per una ricerca di omologia tra le sequenze presenti nella banca dati PDB. Lo scopo di questa ricerca è stato quello di ottenere informazioni in più sulle funzioni delle proteine trovate e vedere quali di esse fossero già state caratterizzate o la cui struttura fosse stata risolta in studi si spettroscopia ai raggi X su larga scala. Per ciascuna sequenza è stata

utilizzata la PSSM già usata per MD per effettuare una ricerca nel database PDB con PSI-BLAST ($E<10^{-4}$, 1 ciclo). Dalle ricerche è emerso che delle 208 proteine individuate da MD che avevano passato il filtro di conservazione relativa, 96 hanno degli omologhi a struttura nota. Per quanto riguarda le proteine ipotetiche, delle 66 trovate da MD, 23 hanno degli omologhi in PDB (Figura 3.2).

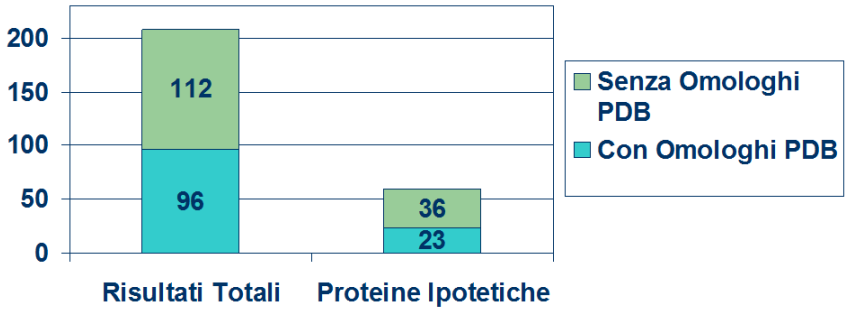

Figura 3.2 *Grafico riportante quanti risultati di MetalDetector hanno omologhi a struttura nota*

3.3.2 - CONFRONTO CON LE ANNOTAZIONI IN BANCA DATI

Per avere una misura di quanto la predizione di MD coincida con le annotazioni presenti in GenBank, abbiamo analizzato tramite una procedura automatica le annotazioni delle proteine di *Helicobacter pylori*, selezionando quelle nella cui annotazione comparivano riferimenti ai metalli di transizione comunemente legati nei centri di coordinazione.

La ricerca (meglio descritta nei Materiali e Metodi) ha identificato 156 sequenze di *Helicobacter pylori* la cui annotazione GenBank facesse riferimento a metalli di transizione (Tabella Supplementare S3.2). Di

queste, 126 hanno un'annotazione funzionale attribuita e 30 sono annotate come proteine ipotetiche (Figura 3.3).

Abbiamo confrontato la lista di proteine così ottenuta con l'elenco di proteine trovate da MetalDetector per individuare in che misura i risultati si sovrappongono. Delle 307 proteine trovate da almeno una delle due ricerche, 57 sono state trovate da entrambe, 151 solo da MD e 99 solo dalla ricerca per parole chiave (KS) [figura]. Limitando il campo solo alle proteine ipotetiche, 80 sono state quelle identificate da almeno una delle due procedure. Di queste 14 sono state trovate da entrambe, 52 solo da MD e 16 solo dalla KS [figura].

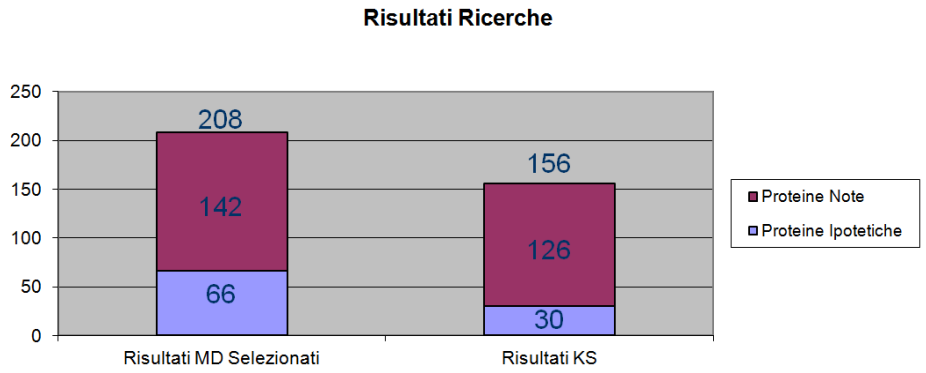

Figura 3.3 *Comparazione tra I risultati della ricerca con MetalDetector (Risultati MD Selezionati) e la ricerca per parole chiave (Risultati KS)*

Figura 3.4 *Sovrapposizione dei risultati tra la ricerca con MetalDetector e la ricerca per parole chiave. A: risultati totali; B: proteine ipotetiche*

3.3.3 - ANALISI DEI RISULTATI

Come successivo passaggio ci siamo concentrati sulle proteine ipotetiche, ovvero quelle proteine la cui annotazione funzionale è assente o ancora incerta. Abbiamo scelto di concentrarci su di esse perché ritenevamo che fra esse potessero trovarsi proteine interessanti coinvolte nell'adattamento e nella sopravvivenza di *H. pylori* nel suo habitat che non fossero ancora state individuate come tali.

Avendo ridotto il numero dei candidati ad una quantità gestibile manualmente e avendo qualche informazione di natura strutturale su di esse, abbiamo proceduto ad un'analisi manuale dei risultati. Tramite l'analisi manuale abbiamo attribuito, ove possibile, una possibile annotazione funzionale alle proteine, basata sulle annotazioni di

omologhi (a struttura nota ove disponibili) e su analisi di associazione genica e dell'intorno genico. L'elenco completo delle proteine ipotetiche sottoposte ad ulteriore analisi è riportato in (Tabella Supplementare S3.3). Nella Tabella 3.2 sono riportati i risultati suddivisi secondo le funzioni biologiche a cui sono stati attribuiti.

Classificazione Funzionale	**Numero di proteine**	**Proteine con omologhi PDB**
Mantenimento e costruzione di membrana e parete cellulare	9	2
Maturazione tRNA	6	3
Metabolismo cofattori	5	3
Trasduzione del segnale	5	3
Nucleasi	4	0
Costruzione del flagello	3	2
Modificazioni post-traduzionali	3	2
Trasporto	2	1
Proteine radical SAM	2	2
Enzimi di restrizione	2	0
Metabolismo degli zuccheri	2	0
Plasticità del DNA	2	0
Ciclo dell'urea	1	1
Competenza	1	0
Detossificazione	2	1
Divisione cellulare	1	1
Proteine fagiche	1	0
Fattori di trascrizione	1	0
Maturazione mRNA	1	0
Proteine ribosomiali	1	0
Riparazione del DNA	1	1
Trasporto di elettroni	1	0
Funzione non assegnabile	11	2

Totale	**67**	**24**

Tabella 3.2*Classificazione funzionale delle proteine ipotetiche trovate da MetalDetector. Per ogni classe funzionale viene indicato il numero di membri di quella classe che hanno omologhi a struttura nota in PDB*

Una delle proteine che abbiamo assegnato alla classe "Trasporto" è HP0129, annotata come proteina ipotetica. Questa proteina ha un paralogo in *H. pylori* 26695 (HP0721) annotata come proteina legante il substrato e parte di un sistema di trasporto ABC per lo zinco. Evidenze di letteratura, però, suggeriscono che, in base alla sua struttura, HP0721 legherebbe un nucleotide nicotinico (Cioci et al. 2011) o l'acido sialico (Bennett and Roberts 2005). Entrambe le proteine hanno un dominio Pfam, la cui funzione, però, è ignota (DUF1104). Abbiamo effettuato una ricerca con HHpred usando come quey sia HP0129 che HP0721, ma nessuna per nessuna delle due proteine sono stati trovati risultati significativi ($E<10^{-4}$) che suggerissero evidenze a favore dell'annotazione come componenti di un sistema ABC (i risultati significativi rappresentano il dominio DUF1104 in vari database), e anche gli intorni genici di entrambe non presentano componenti di sistemi ABC che, spesso, si trovano vicini fra loro sul genoma organizzati in operoni (Tomii and Kanehisa 1998). In seguito a ciò abbiamo concluso che quell'annotazione fosse stata attribuita, probabilmente, in seguito ad un errore. Alla classe "Trasporto" abbiamo assegnato anche HP0311, dal momento che in vari organismi il suo gene si trova adiacente a componenti di sistemi ABC o ad una permeasi ed è quasi sempre affiancata da una proteina etichettata come legante nucleosidi trifosfati (ATP o GTP); essa è, inoltre, quasi sempre vicina alla polisaccaride deacetilasi HP0310 (Shaik et al. 2011)(vedi Capitolo 3), pertanto è

possibile ipotizzare che svolga un qualche ruolo di trasporto o una funzione "ponte" tra HP0310 e un sistema ABC. Altre proteine interessanti sono HP0506 (classe "Mantenimento e costruzione di membrana e parete cellulare"), la quale è stata recentemente caratterizzata come una metallopeptidasi coinvolta nel modellamento del peptidoglicano (Bonis et al. 2010), e HP0518, coinvolta nella metilazione del flagello (Asakura et al. 2010). Fra le proteine appartenenti alla classe "Detossificazione" c'è HP0813. Essa contiene un dominio "Lactamase_B" (Pfam, $E=10^{-18}$) tipico di β-lattamasi e gliossalasi II. In PDB sono presenti omologhi annotati sia come gliossalasi che β-lattamasi, oltre ad almeno due persolfuro diossigenasi (Pettinati et al. 2015; Sattler et al. 2015). Dal momento che gli allineamenti che coprono tutta o quasi la sequenza sono quelli con persolfuro diossigenasi (coinvolte nella detossificazione dei solfuri) e gliossalasi II (coinvolte sia nel metabolismo del piruvato che nella detossificazione del metilgliossale) è nostra opinione che anche HP0813 possa avere un ruolo simile, pertanto la abbiamo classificata come enzima coinvolto in processi di detossificazione. L'alternativa è che HP0813 sia una β-lattamasi, quindi un enzima coinvolto nella resistenza agli antibiotici, ma le β-lattamasi a struttura nota omologhe ad HP0813 sono allineabili solo ad una porzione ridotta di HP0813 mentre le gliossalasi II e le persolfuro diossigenasi si allineano bene all'intera lunghezza della proteina (Figura 3.5).

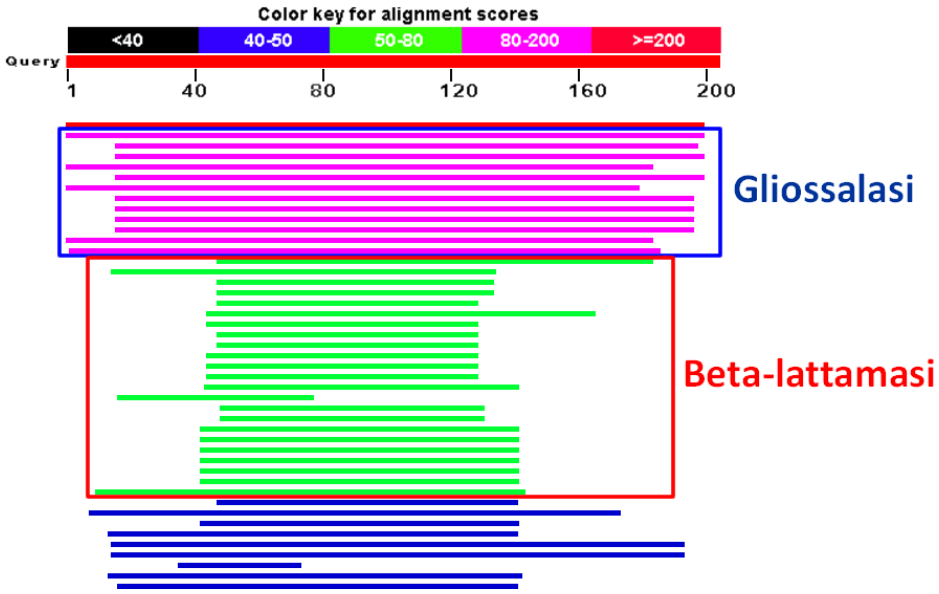

Figura 3.5 *Riepilogo grafico degli allineamenti degli omologhi PDB di HP0813 con la sequenza stessa (In rosso in alto). Il riquadro blu evidenzia le gliossalasi e persolfuro diossigenasi, il riquadro rosso evidenzia le β-lattamasi.*

Usando le sequenze delle proteine ipotetiche individuate dalla nostra procedura per una ricerca nel database dei domini Pfam (http://pfam.xfam.org/) abbiamo visto come 14 delle proteine da noi identificate come metalloproteine non abbiano domini annotati (Tabella 3.3), 6 delle quali non hanno annotazioni funzionali assegnate (né per sé né per i loro omologhi trovati con PSI-BLAST).

Accession number	Locus	Classe funzionale
NP_207029	HP0231	Modificazioni post-traduzionali
NP_207109	HP0311	Trasporto
NP_207416	HP0622	Proteina di membrana
NP_207454	HP0660	Proteina ipotetica
NP_207467	HP0673	Proteina Ipotetica
NP_207574	HP0781	Proteina fagica
NP_207631	HP0838	Proteina Ipotetica
NP_207658	HP0864	Proteina ribosomiale
NP_207695	HP0902	Proteina Ipotetica
NP_207781	HP0990	Proteina Ipotetica
NP_207787	HP0996	Plasticità del DNA
NP_207832	HP1042	Nucleasi
NP_208264	HP1473	Competenza
NP_208310	HP1519	Proteina Ipotetica

Tabella 3.3 *Proteine annotate come proteine ipotetiche che non hanno domini annotate in Pfam e classe funzionale a cui le abbiamo assegnate.*

Infine, controllando le annotazioni dei domini Pfam e analizzando le sequenze con TMHMM abbiamo visto che circa un terzo delle proteine probabilmente includono nella loro struttura dei cluster ferro-zolfo o sono proteine di membrana e, a causa di queste loro caratteristiche, potrebbero presentare delle difficoltà qualora si volesse tentare di esprimerle e purificarle per caratterizzarle (Tabella 3.4).

Locus	Accession Number	Dominio con centro FeS	Eliche transmembrana
HP0117	NP_206917	Radical_SAM	0
HP0138	NP_206938	Fer4_8	0
HP0139	NP_206939	CCG	0
HP0258	NP_207056	no	4
HP0274	NP_207072	CxxCxxCC	0
HP0285	NP_207083	Radical_SAM	0
HP0506	NP_207303	no	1*
HP0568	NP_207363	SPASM	0
HP0595	NP_207390	no	5
HP0622	NP_207416	no	2
HP0650	NP_207444	UDG	0
HP0654	NP_207448	Radical_SAM	0
HP0656	NP_207450	Radical_SAM	0
HP0660	NP_207454	no	1*
HP0734	NP_207528	Radical_SAM	0
HP0764	NP_207557	no	1
HP0861	NP_207655	no	6
HP0864	NP_207658	no	1
HP0980	NP_207771	no	2**
HP1044	NP_207834	no	4
HP1089	NP_207880	PDDEXK_1	0
HP1417m	NP_208208	no	5
HP1580	NP_208371	no	1
Totale		**11**	**10+2**

Tabella 3.4 *Proteine contenenti cluster FeS (secondo l'annotazione dei domini Pfam) o domini transmembrana (secondo TMHMM). Nella colonna "Dominio con centro FeS" viene indicato l'identificativo del dominio Pfam contenente centri FeS presente nella proteina. *: la proteina ha, probabilmente un'ancora segnale all'N-terminale (ha un tratto transmembrana nei primi 30 aminoacidi ma nessun sito di taglio per peptidi segnali secondo SignalP 3.0). Le proteine con un'ancora segnale sono indicate separatamente nel totale. **: le eliche transmembrana non sono state predette da TMHMM, ma sono state individuate solo da TMPred.*

3.4 - CONCLUSIONI

Vista l'importanza che hanno le metalloproteine nel metabolismo di *H. pylori* abbiamo applicato il programma MetalDetector (Passerini et al. 2011) in una procedura per l'identificazione a tappeto delle proteine leganti metalli basandosi sulle sole sequenze aminoacidiche. L'utilità di un simile approccio risiede nel fatto che la predizione tramite MetalDetector è rapida (l'analisi dell'intero proteoma di *H. pylori* non ha richiesto che qualche ora) e non richiede passaggi sperimentali, che sono invece richiesti da altre tecniche per identificare siti di coordinazione su larga scala (Shi et al. 2011). Un altro vantaggio di questo sistema è che l'identificazione di una metallo proteina con esso non richiede che esistano proteine omologhe già annotate come tali, il che consente di identificare metalloproteine non ancora identificate e di evitare identificazioni erronee di proteine i cui omologhi legano metalli ma esse stesse non hanno questa capacità, il che riduce errori dovuti alla disseminazione delle annotazioni in banca dati che si ha quando i programmi di annotazione automatica estendono erroneamente delle annotazioni. Usando questo sistema siamo riusciti, infatti, ad identificare 6 metalloproteine che non sarebbero state identificate né tramite ricerche di omologia né tramite ricerche nei database di domini funzionali. A questo va ad aggiungersi il risultato del confronto con le annotazioni della banca dati che ha rivelato come per le sequenze di *H. pylori* la sovrapposizione tra i nostri risultati e le annotazioni sia scarsa, il che indica che, probabilmente, molte metalloproteine sono state mancate nelle procedure di annotazione automatica.

L'analisi delle proteine annotate come ipotetiche ci ha consentito di proporre per molte di esse una possibile funzione. Fra esse ci sono alcune proteine periplasmatiche tra cui ne abbiamo identificata una che potrebbe essere la componente periplasmatica di legame al substrato di un sistema ABC per il trasporto degli ioni metallici (HP0129). Un gruppo in cui è ricaduta una buona parte delle proteine ipotetiche (9 su 67) è quello delle proteine responsabili del mantenimento delle strutture di superficie del batterio (membrana cellulare e parete cellulare). Queste proteine potrebbero rivelarsi interessanti, dal momento che le modificazioni della parete cellulare e delle proteine della membrana esterna rivestono un ruolo importante nell'evasione della risposta immunitaria dell'ospite.

4 – EVOLUZIONE DEL SELENOPROTEOMA IN H. PYLORI E NEGLI E-PROTEOBATTERI

4.1 - INTRODUZIONE

4.1.1 - RUOLO DEL SELENIO

Il selenio è un microelemento non metallico essenziale per la salute umana. Appropriati livelli di selenio circolante riducono la mortalità degli individui e hanno effetti benefici sul sistema immunitario. Bassi livelli di selenio sono correlati ad una maggiore vulnerabilità a malattie virali, problemi neurologici, ridotta fertilità maschile, danni alla tiroide, ridotta protezione contro vari tipi di cancro e aumento della mortalità in seguito a problemi cardiovascolari (Rayman 2012).

Ci sono scarsi dati riguardo l'utilizzo di selenio da parte di *Helicobacter pylori*. In vari studi condotti non sono state trovate significative differenze nei livelli di selenio circolante tra individui sani e individui infettati da *H. pylori* (Toyonaga et al. 2000; Ustundag et al. 2001). Differenze, invece, sono state individuate nei livelli di selenio nel tessuto gastrico. Le biopsie di individui infetti mostrano livelli di selenio più elevati rispetto alle biopsie di individui sani. I livelli di selenio sono stati visti tornare alla normalità in seguito all'eradicazione di *H. pylori* (Ustundag et al. 2001). Inoltre, il genoma di *H. pylori* (Tomb et al. 1997) mostra la presenza di una proteina annotata come L-seril-tRNA^{Sec} selenio

transferasi. Ciò suggerisce che il selenio possa rivestire un qualche ruolo nel metabolismo del batterio.

4.1.2 - SELENOCISTEINA E SISTEMA SEL

Il ruolo principale del selenio si esplica nella sua incorporazione nelle proteine nella forma dell'aminoacido selenocisteina (Sec), la quale riveste solitamente il ruolo di residuo catalitico in una serie di enzimi redox (Kryukov and Gladyshev 2004; Rayman 2012). Questi enzimi, in certi organismi, hanno una la selenocisteina sostituita con una cisteina (Kryukov and Gladyshev 2004; Kryukov et al. 2003).

La selenocisteina è nota come il 21° aminoacido e ha peculiare caratteristica (condivisa con il 22° aminoacido, la pirrolisina) di essere codificata da un codone di stop. Normalmente, il codone UGA viene interpretato come codone di stop "opale", ma, in presenza di un determinato segnale, esso codifica per l'inserimento della selenocisteina a livello tradizionale. Questo segnale di inserimento della selenocisteina (elemento SECIS) è una struttura secondaria formata dall'mRNA stesso, la cui presenza innesca il reclutamento del $tRNA^{Sec}$ e di un fattore di traduzione simile ad EfTu ma specifico per la selenocisteina (SelB) a livello del ribosoma. Un'altra caratteristica della selenocisteina è l'assenza di una selenocisteinil-tRNA sintetasi. Il $tRNA^{Sec}$, invece, viene dapprima caricato ad opera di una seril-tRNA sintasi con una serina che viene, quindi, modificata in selenocisteina. Nei procarioti, il macchinario molecolare per l'inserimento della selenocisteina richiede quattro componenti (Figura 4.1): l'enzima selenofosfato sintasi (SelD) che

catalizza la formazione di selenofosfato a partire da selenuro, acqua e ATP; il $tRNA^{Sec}$, codificato dal gene *selC*; l'enzima selenocisteina sintasi, anche detto L-seril-$tRNA^{Sec}$ selenio transferasi (SelA), che utilizza il selenofosfato per modificare la serina caricata sul $tRNA^{Sec}$ in selenocisteina; il fattore di traduzione simile ad EfTu specifico per Sec (SelB) (Böck et al. 1991; Heider, Baron, and Böck 1992; Kryukov and Gladyshev 2004; Zinoni et al. 1987; Zinoni, Heider, and Bock 1990).

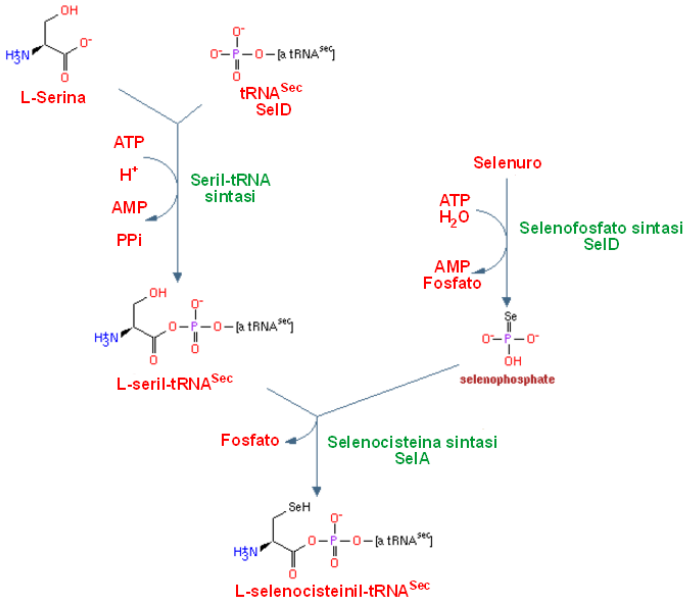

Figura 4.1 *Via biosintetica della selenocisteina (fonte: http://metacyc.org).*

4.1.3 - IDENTIFICAZIONE DELLE SELENOPROTEINE NELLE SEQUENZE GENOMICHE

L'identificazione delle selenoproteine nelle sequenze genomiche è un'operazione non banale e, spesso, non viene eseguita correttamente dai programmi di annotazione automatica. Ciò è dovuto principalmente

al fatto che Sec viene codificata dal codone TGA che, solitamente, è un codone di stop. Per questa ragione, le procedure automatiche normalmente fanno corrispondere a TGA uno stop e questo fa sì che le sequenze codificanti per selenoproteine siano identificate come ORF troncate o che addirittura siano ignorate se la selenocisteina si trova molto vicina all'terminale (Castellano et al. 2008).

L'identificazione delle selenoproteine richiede, quindi, una particolare attenzione e può essere effettuata mediante due strategie: l'identificazione degli elementi SECIS e la ricerca di corrispondenze Cys/Sec nelle sequenze omologhe. Come già menzionato, quando il codone TGA deve codificare per Sec invece che per un codone di stop, esso è seguito da una sequenza di inserzione della selenocisteina (SECIS) che determina una particolare struttura secondaria dell'mRNA. Questi elementi SECIS possono essere identificati tramite modelli di covarianza (Zhang and Gladyshev 2005) e sono un buon indicatore del fatto che il codone TGA codifichi per Sec anziché per uno stop. L'altra strategia si basa sul fatto che tutte le selenoproteine note tranne una (selenoproteina A) hanno degli omologhi in cui la selenocisteina è stata sostituita da una cisteina. È, pertanto, possibile utilizzare tBLASTn usando queste sequenze omologhe per cercare quelle sequenze nucleotidiche in cui si ha corrispondenza tra un codone TGA e una cisteina e l'intorno della corrispondenza è ben conservato (Kryukov and Gladyshev 2004).

4.2 - MATERIALI E METODI

4.2.1 - FILOGENESI DI SPECIE E GENI

L'albero delle specie è basato su un albero radicato di massima verosimiglianza costruito con 454 marker genici specifici per gli ε-proteobatteri (Wang and Wu 2013). L'albero è stato processato in ambiente R usando le funzioni della libreria Ape (Popescu, Huber, and Paradis 2012). Le foglie non necessarie sono state rimosse con la funzione *drop.tip* e l'albero risultante è stato reso ultrametrico con la funzione *chronos*. I pattern di assenza/presenza di geni sono stati graficati a fianco delle foglie con la funzione *table.phylo4d* del pacchetto Adephylo (Jombart, Balloux, and Dray 2010). L'albero filogenetico delle proteine SelA è stato ottenuto usando il metodo di massima somiglianza incluso nel programma RaxML v7.7.8 (Stamatakis 2006) usando il modello PROTCATGTR per le sostituzioni aminoacidiche. L'albero di SelA è stato confrontato con l'albero delle specie usando la funzione *cophyplot* di Ape. L'albero dei SelA procariotici è stato visualizzato con FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

4.2.2 - IDENTIFICAZIONE DELLE PROTEINE SEL E DEI GENI CODIFICANTI PER SELENOPROTEINE

I genomi procariotici completi sono stati scaricati dal sito ftp nell'NCBI; i genomi degli ε-proteobatteri considerati nell'analisi sono stati selezionati dai file con i report genomici in base alla disponibilità della filogenesi su genoma completo (Wang and Wu 2013) e raggruppati usando un

punteggio di somiglianza genomica (GSSa) = 0.95 (Moreno-Hagelsieb et al. 2013). Le proteine del sistema Sel sono state identificate tramite ricerca di omologia usando le proteine annotate di *Campylobacter jejuni*. Le proteine MnmH e YdeF genuine da *Sulfurimonas autotrophica* e *Campylobacter lari*, codificate in operoni Sel, sono state usate come sequenze di riferimento per cercare proteine coinvolte nel metabolismo del selenio. Le ricerche sono state effettuate su entrambi i proteomi completi usando BLASTp ($E<10^{-10}$) e sui genomi tradotti con tBLASTn ($E<10^{-10}$). Per discriminare tra omologhi equivalenti e non equivalenti, paraloghi noti (es.: EfTu) sono stati inclusi nella ricerca.

Le proteine contenenti un codone UGA codificante per Sec sono state identificate tramite ricerca di omologia con un set di 1022 selenoproteine annotate ottenute dal database dbTEU (database of Trace Elements Utilizators: http://gladyshevlab.org/trace_element/). Le selenoproteine annotate sono state usate per interrogare il set di genomi con tBLASTn ($E<10^{-6}$). Gli allineamenti dei risultati migliori sono stati analizzati per identificare il codone (TGA o TGY) corrispondente a Sec nella sequenza query. Le sequenze codificanti delle selenoproteine identificate tramite omologia sono state predette usando il programma *gmhmmp_heuristic.pl* della suite Genmark (Borodovsky et al. 2003) considerando il miglior modello genico (opzione *–b*) ottenuto dopo aver rimpiazzato il codone TGA con NNN. La predizione di Genmark del gene *selD* di *C. jejuni* è stata corretta in virtù della presenza di un raro codone d'inizio CUG (Shaw et al. 2012).

La ricerca di selenoproteine prive di omologia con le selenoproteine note è stata effettuata usando il metodo descritto in Kryukov & Gladyshev, 2004.

4.2.3 - IDENTIFICAZIONE DEI GENI PER I tRNA^{SEC}

La ricerca dei tRNA^{Sec} è stata effettuata col programma tRNAscan-SE usando l'opzione per la massima sensibilità e il modello *infernal* per i tRNA^{Sec} procariotici. Il nuovo modello *infernal* (PSELCinf-c.cm) è stato visto essere più efficace nell'identificazione dei geni rispetto al precedente (PSELC.cm). I risultati finali sono stati ottenuti con un modello di covarianza specifico per i tRNA^{Sec} degli ε-proteobatteri.

4.2.4 - STRUTTURA SECONDARIA DEI tRNA E ANALISI DI COVARIANZA

Il modello di covarianza per tRNA batterici (TRNAinf-bact-c.cm) del pacchetto Infernal (Nawrocki and Eddy 2013) è stato usato per allineare e foldare i tRNA non-Sec nella classica struttura a trifoglio. L'allineamento dei tRNA^{Sec} degli ε-proteobatteri è stato basato sul modello procariotico di covarianza dei tRNA^{Sec} (PSELCinf-c.cm). Diverse strutture secondarie proposte per i tRNA^{Sec} sono state analizzate tramite l'analisi di covarianza confrontando i punteggi ottenuti dai vari tRNA^{Sec} con vari modelli di covarianza che assumono 7/5, 8/5 o 9/4 bp nello stelo aminoacilico e nello stelo T. La struttura secondaria degli RNA codificanti per le selenoproteine è stata analizzata usando il programma Centroid_alifold (Hamada et al. 2009) usando allineamenti di sequenza della regione attorno al codone TGA (-10; +60) delle famiglie di proteine

della formato deidrogenasi (Fdh), SelD e SelW. La predizione degli elementi SECIS è stata effettuata usando il programma bSECISearch (Zhang and Gladyshev 2005) tramite l'interfaccia web (http://genomics.unl.edu/bSECISearch/). L'identificazione degli elementi SECIS nei genomi degli ε-proteobatteri è stata basata sul programma Cmsearch (Nawrocki and Eddy 2013) col modello di covarianza derivato dalla figura [figura] usando un punteggio di cutoff di 16. Le corrispondenze di elementi SECIS col modello di covarianza sono state valutate tramite bit scores (T) e significatività (P) forniti da Cmsearch. L'analisi di covarianza e le immagini delle strutture degli RNA sono state basate sul programma R2R (Weinberg and Breaker 2011).

4.2.5 - SOSTITUZIONI SINONIME E NON SINONIME

Il rapporto tra le sostituzioni non sinonime per sito non sinonimo (dN) e le sostituzioni sinonime per sito sinonimo (dS) nei confronti a coppie è stato calcolato usando il metodo di massima verosimiglianza implementato nel programma Codeml del pacchetto PAML (Yang 2007). Le variazioni nei rapporti dN/dS nelle sequenze codificanti di SelA lungo l'albero degli ε-proteobatteri sono state analizzate con Codeml consentendo al rapporto dN/dS di variare in tra particolari rami dell'albero. I parametri sono stati stimati usando modelli di singoli o multipli cambiamenti di dN/dS (fino a quattro) e modelli di variazioni persistenti o episodiche. Le verosimiglianze dei vari modelli sono state confrontate con il test dei rapporti di verosimiglianza.

4.2.6 - ALLINEAMENTO DI SEQUENZE E ANALISI DI STRUTTURE

Gli allineamenti multipli sono stati costruiti con ClustalW (Larkin et al. 2007). La presenza o assenza del dominio di legame al $tRNA^{Sec}$ nelle proteine SelA è stata valutata tramite confronto con un modello markoviano (Eddy 1996) del dominio. L'allineamento è stato analizzato con GeneDoc e la risorsa online Espript (Robert and Gouet 2014) è stata usata per generare le immagini degli allineamenti. I modelli della struttura di SelA sono stati scaricati da PDB e analizzati con PyMol (http://www.pymol.org). La lista dei residui coinvolti nelle interazioni proteina-proteina è stata scaricata da PDBsum (http://www.ebi.ac.uk/pdbsum). L'analisi della conservazione dei residui mappati sulla struttura di AaSelA è stata effettuata tramite il server Consurf (http://consurf.tau.ac.il/) usando allineamenti multipli delle sequenze di SelA da specie di Helicobacteraceae che possiedono o meno il sistema Sel.

4.3 - RISULTATI

4.3.1 - IDENTIFICAZIONE DELLE SELENOPROTEINE

Una prima ricerca nei genomi degli ε-proteobatteri selezionati (Tabella 4.1) ha identificato 10 proteine contenenti un residuo Sec (U): 6 formato deidrogenasi (Fdh), 2 SelD, 1 selenoproteina W (SelW) e 1 tioredossina (Tr). Utilizzando tBLASTn per cercare sequenze omologhe alle sequenze riportate in dbTEU (Zhang and Gladyshev 2010) ha identificato altri 40 geni codificanti per possibili selenoproteine [tabella], corrispondenti a proteine annotate come troncate al 3' o al 5' in GenBank. Alcuni di questi geni erano, inoltre, stati mancati nelle annotazioni genomiche. Altri geni identificati per omologia con una selenoproteina avevano, invece, una cisteina al posto del residuo Sec. In considerazione della possibile esistenza di selenoproteine ignote, abbiamo effettuato una terza ricerca in cui abbiamo tradotto tutte le possibili ORF nei genomi selezionati codificanti per proteine di almeno 30 aminoacidi assumendo la codifica TGA→U e le abbiamo usate per una ricerca con tBLASTn. I risultati sono stati esaminati in cerca di quegli allineamenti in cui si aveva una corrispondenza U:U oppure U:C rispetto ai casi di corrispondenza U:X (in cui TGA codificava, probabilmente, davvero per un codone di stop). Questa procedura ha identificato tutte le selenoproteine precedentemente trovate e un candidato addizionale: una ORG di *Campylobacter curvus* con somiglianza significativa con le tiolo perossidasi presenti in dbTEU (migliore hit: YP_593855.1, $E=2*10^{-8}$). Questa sequenza contiene Sec in posizione diversa rispetto alle sequenze

di dbTEU e questa è la ragione per cui, probabilmente, era stata mancata dalle ricerche precedenti.

L'analisi degli mRNA col programma bSECISearch (Zhang and Gladyshev 2005) ha identificato degli elementi SECIS ottimali in tutte le sequenze identificate eccetto SelW di *Campylobacter fetus*. Un allineamento guidato dalla struttura degli elementi SECIS trovati è stato usato per costruire un modello di covarianza specifico per gli ε-proteobatteri e questo modello è stato usato per una ricerca nei genomi degli ε-proteobatteri considerati, ottenendo 131 ulteriori candidati, tra cui membri della famiglia DUF466 (Figura 4.2). In queste proteine, la selenocisteina era preceduta da una cisteina e seguita da un codone di stop ed esse hanno degli omologhi in cui Sec è sostituita da Cys, anch'essa preceduta da un'altra Cys e seguita da un codone di stop. Il fatto che in queste proteine Sec sia l'ultimo residuo è la ragione per cui la ricerca con tBLASTn non le trovava, visto che l'allineamento di BLAST non include residui U terminali. Non esiste somiglianza significativa tra le proteine DUF466 e altre selenoproteine (come confermato da ricerche con modelli markoviani; hmmsearch, E>1) e non sono stati riportati altri casi di selenoproteine con una selenocisteina come ultimo residuo al carbossi-terminale.

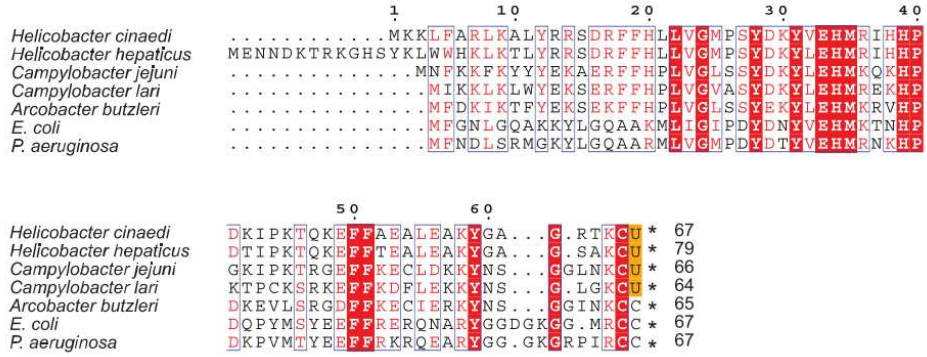

Figura 4.2 *Allineamento delle sequenze proteiche di proteine della famiglia DUF466. La Sec terminale è evidenziata in arancione (Cravedi et al. 2015).*

4.3.2 - IDENTIFICAZIONE DEI COMPONENTI DEL SISTEMA SEL

Il sistema Sel, responsabile dell'utilizzo del selenio sotto forma di selenocisteina (Sec) è composto da 4 geni: *selA* (codificante per la selenocisteina sintasi), *selB* (codificante per un fattore di traduzione simile ad EfTu), *selC* (codificante per il $tRNA^{Sec}$) e *selD* (codificante per la selenofosfato sintasi)(Romero et al. 2005). Allo scopo di individuare gli organismi utilizzatori di selenocisteina negli ε-proteobatteri abbiamo cercato questi geni in un campione di 37 ε-proteobatteri includenti 10 ceppi di *Helicobacter pylori* (Tabella 4.1).

Organismo
Helicobacter pylori 26695
Helicobacter pylori Shi417
Helicobacter pylori PeCan4
Helicobacter pylori India7
Helicobacter pylori P12
Helicobacter pylori B8
Helicobacter pylori G27
Helicobacter pylori ELS37

Helicobacter pylori Gambia94/24
Helicobacter pylori SouthAfrica7
Helicobacter acinonychis str. Sheeba
Helicobacter cetorum MIT 99-5656
Helicobacter cetorum MIT 00-7128
Helicobacter bizzozeronii CIII-1
Helicobacter felis ATCC 49179
Helicobacter mustelae 12198
Helicobacter hepaticus ATCC 51449
Helicobacter cinaedi PAGU611
Wolinella succinogenes DSM 1740
Campylobacter jejuni subsp. jejuni ATCC 700819
Campylobacter lari RM2100
Campylobacter hominis ATCC BAA-381
Campylobacter fetus subsp. fetus 82-40
Campylobacter curvus 525.92
Campylobacter concisus 13826
Sulfurospirillum deleyianum DSM 6946
Sulfurospirillum barnesii SES-3
Arcobacter butzleri RM4018
Arcobacter sp. L
Arcobacter nitrofigilis DSM 7299
Sulfurimonas denitrificans DSM 1251
Sulfurimonas autotrophica DSM 16294
Sulfuricurvum kujiense DSM 16994
Nitratifractor salsuginis DSM 16511
Sulfurovum sp. NBC37-1
Nitratiruptor sp. SB155-2
Nautilia profundicola AmH

Tabella 4.1 *Organismi utilizzati nella ricerca*

I tRNA vengono di solito identificati con grande accuratezza nelle sequenze genomiche tramite l'uso di appositi programmi. I $tRNA^{Sec}$, però, possono sfuggire all'identificazione per via della loro struttura inusuale, con il braccio D e il braccio variabile più lunghi del normale e

varie posizioni la cui sequenza devia dal consenso. Il programma da noi utilizzato, tRNAscan-SE (Lowe and Eddy 1997) dispone di un modello apposito che può essere utilizzato per trovare specificamente i $tRNA^{Sec}$. Usando i parametri di default, tRNAscan-SE identifica i $tRNA^{Sec}$ solo in tre specie (*C. jejuni, C. hominis* e *C. lari*), mentre l'uso del modello di covarianza specifico "Infernal" (Nawrocki and Eddy 2013) consente al programma di identificare 21 $tRNA^{Sec}$ distribuiti in 8 diversi generi. Utilizzando le sequenze così identificate, abbiamo realizzato un modello di covarianza specifico per i $tRNA^{Sec}$ che identifica le sequenze già trovate con elevata significatività, ma non abbiamo trovato nuove sequenze di $tRNA^{Sec}$

Le proteine SelA, SelB e SelD sono state identificate tramite ricerca di omologia con sequenze di riferimento riportate in GenBank. Alla ricerca è stata aggiunta la sequenza di EfTu (omologo a SelB) in modo da poter discriminare tra gli organismi che possiedono solo EfTu e quelli che possiedono sia EfTu che SelB. Abbiamo, inoltre, cercato i geni *mnmH* (codificante per l'enzima tRNA 2-selenouridina sintasi coinvolto nella modificazione della base vacillante in specifici anticodoni (Wolfe et al. 2004)) e *yedF* (codificante per una proteina del metabolismo del selenio a funzione indefinita) che spesso si ritrovano vicino ai geni *sel*. Le ricerche sono state condotte sia nel database di sequenze proteiche, sia in quello di sequenze nucleotidiche tradotto, in modo da individuare proteine annotate male o non annotate. Normalmente, la presenza dei geni *sel* segue uno schema prevedibile in cui il sistema è presente o assente nella sua interezza oppure è presente solo *selD*, probabilmente

richiesto per altri processi che coinvolgono il selenofosfato, il che è corroborato dal fatto che la presenza di *mnmH* e *yedF* implica di solito la presenza di *selD* (Romero et al. 2005; Zhang et al. 2006). Questo schema sembra essere rispettato negli ε-proteobatteri, la cui maggioranza possiede un sistema Sel completo e funzionante (Figura 4.3), probabilmente indizio della presenza del sistema nell'antenato comune della classe. L'unica eccezione sembra costituita da *H. pylori* e *H. acinonychis*, i quali non hanno il sistema Sel né possiedono selenoproteine, eppure conservano *selA*.

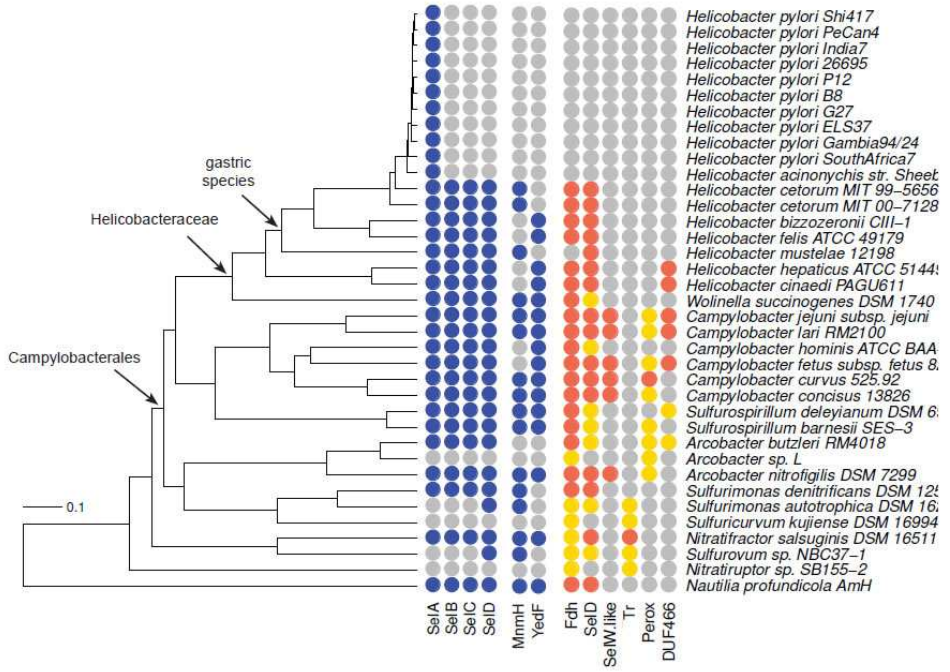

Figura 4.3 *Albero filogenetico degli organismi analizzati e presenza o assenza di geni in quegli organismi. Le prime 4 colonne riportano i geni del sistema Sel; un pallino blu indica la presenza del gene, un pallino grigio la sua assenza. La quarta e la quinta colonna riportano i geni MnmH e YedF; il codice colori è lo stesso del sistema Sel. Le restanti colonne riportano i geni codificanti per alcune selenoproteine note; un pallino grigio indica l'assenza del gene, un pallino rosso la sua presenza e un pallino giallo la sua presenza con la selenocisteina sostituita da una cisteina (Cravedi et al. 2015).*

4.3.3 - EVOLUZIONE DI SELA

Nonostante le ricerche accurate, non siamo stati in grado di identificare geni *sel* oltre a *selA* o geni codificanti per selenoproteine in *H. pylori* e *H. acinonychis*, pertanto abbiamo considerato l'ipotesi che *selA* fosse uno pseudogene in questi due organismi.

L'analisi delle sequenze di *selA* dei vari ceppi di *H. pylori* (*hpselA*) e *H. acinonychis* (*haselA*) ha rivelato l'assenza di mutazioni frameshift o nonsenso comuni negli pseudo geni. Abbiamo, inoltre, analizzato il rapporto tra sostituzioni non sinonime e sinonime (dN/dS) di *hpselA* e *haselA* e abbiamo ottenuto dei valori notevolmente inferiori ad 1 in tutti i confronti a coppie (Tabella Supplementare S4.1). Per *hpselA*, il valore di dN/dS è maggiore della media per *H. pylori*, ma ancora paragonabile a quello di geni metabolici del genoma core (Didelot et al. 2013), il che suggerisce che *hpselA* sia sottoposto ad una debole selezione purificatrice. Tutto ciò suggerisce che *selA* non sia uno pseudogene, ma che conservi una funzione in queste specie prive del sistema Sel.

L'analisi dei rapporti dN/dS lungo i bracci dell'albero degli ε-proteobatteri ha individuato delle variazioni nella pressione selettiva agente sul gene nella filogenesi degli ε-proteobatteri. Il modello migliore (tabella 4.2) prevede un primo aumento del rapporto dopo la separazione tra *Helicobacter mustelae* e le altre specie gastriche e un secondo aumento nel braccio dell'albero che conduce all'antenato comune di *H. pylori* e *H. acinonychis*.

Modello	dN/dS lungo i bracci				Log verosimiglianza	LRT
	#0	#1	#2	#3		
H0 (1 rapporto)	0.11	-	-	-	-23198,31	
H1 (2 rapporti)	0.05	-	-	0.23	-23126,67	3,5E-33[a]
H2 (3 rapporti)	0.02	-	0.08	0.24	-23111,25	1,1E-08[b]
H3 (4 rapporti)	0.02	0.003	0.08	0.24	-23106,83	0,002[c]

Tabella 4.2 *Rapporti dN/dS lungo I bracci dell'albero. La numerazione dei bracci si riferisce alla figura 4.4 (Cravedi et al. 2015).*
a: calcolato rispetto ad H0
b: calcolato rispetto ad H1
c: calcolato rispetto ad H2

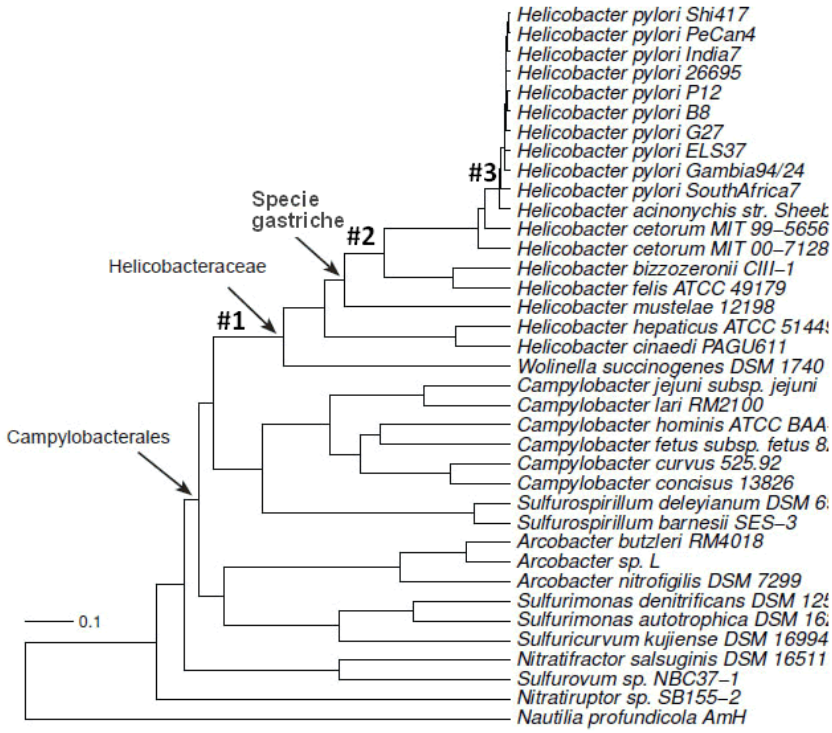

Figura 4.4 *Albero filogenetico degli organismi analizzati con I bracci numerati secondo I riferimenti della Tabella 4.2*

4.3.4 - CARATTERISTICHE DI HPSELA

La struttura tridimensionale di SelA è stata risolta per l'enzima di *Aquifex aeolicus*. La proteina ha una struttura omodecamerica a pentamero di dimeri con forma ad anello pentagonale. La proteina lega e processa 10 molecole di tRNA contemporaneamente. Ogni tRNA è legato cooperativamente da due dimeri adiacenti che formano la tasca di legame al tRNA all'interfaccia tra essi. Il sito attivo dell'enzima, contente PLP come cofattore, è situato all'interfaccia tra le due subunità di ogni dimero (Itoh et al. 2013). Abbiamo, quindi, confrontato HpSelA con la proteina di *A. aeolicus* (AaSelA) per individuare quali caratteristiche fossero state mantenute e quali perse.

Dall'allineamento tra HpSelA, HaSelA, AaSelA e le proteine SelA provenienti da altri *Helicobacter* e da *E. coli* è emerso che HaSelA e HpSelA mancano di un tratto terminale di circa 50 aminoacidi (Figura 4.5). Questo tratto non fa parte della porzione della proteina che lega il PLP. Confrontando un modello per omologia della struttura di HpSelA con la struttura di AaSelA è emerso che questo tratto mancante corrisponde ad un dominio terminale che in AaSelA lega il braccio D e l'ansa T del tRNA (Figura 4.5). La perdita di questo dominio, presente, invece, in tutti gli organismi utilizzatori di Sec, suggerisce che la perdita dei geni *sel* in *H. pylori* e *H. acinonychis* sia stata accompagnata dalla perdita della capacità di legame al tRNA in HpSelA e HaSelA.

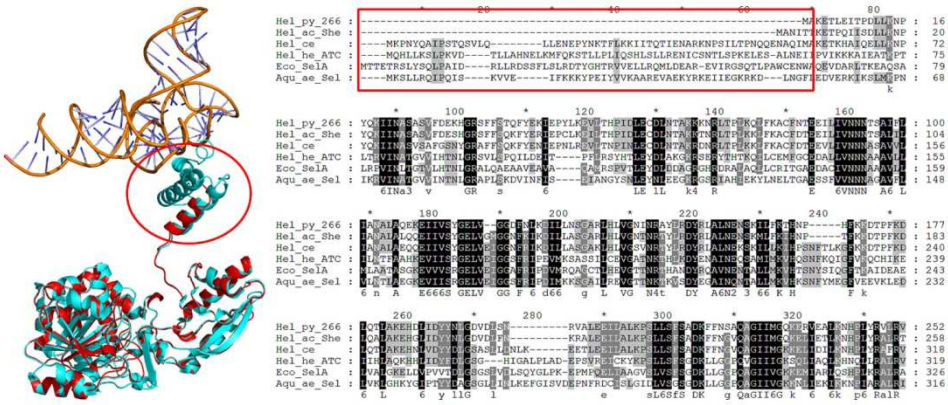

Figura 4.5 *Destra: allineamento delle sequenze di SelA di vari organismi, il riquadro rosso indica il dominio terminale di legame al tRNA, Hel_py_266 = H. pylori 26695, Hel_ac_She = H. acinonychis, Hel_ce = H. cetorum, Hel_he_ATC = H. hepaticus, Eco_SelA = E. coli, Aqu_ae_Sel = A. aeolicus; Sinistra: allineamento della struttura di AaSelA (azzurro) con un modello realizzato con PHYRE2 della struttura di HpSelA (rosso).*

L'analisi della conservazione dei residui e delle variazioni di dN/dS tra i siti indica che i residui coinvolti nell'interazione tra le due subunità di un dimero mostrano profili di conservazione simili in HpSelA e nelle altre *Helicobacteraceae*, suggerendo che l'assemblaggio dei dimeri sia conservato in HpSelA. Al contrario, i residui coinvolti nelle interazioni dimero-dimero risultano poco conservati. In particolare, una regione compresa tra i residui 218 e 224 di AaSelA è quasi completamente deleta in HpSelA e dei residui 191, 192, 199 e 220, la cui mutazione abolisce l'assemblaggio decamerico e fornisce una SelA dimerica inattiva (Itoh et al. 2014), solo Asp199 è conservato in HpSelA (Figura 4.6). Similmente, Arg174, coinvolta nell'interazione dimero-dimero, non è conservata in HpSelA. Di contrasto, i residui cataliticamente attivi mostrano una buona conservazione: la lisina 185 di AaSelA legante il PLP è conservata in HpSelA e lo sono anche le arginine 86, 312 e 315 che legano il

selenofosfato e Arg119 e Asp284, coinvolti nel loro corretto posizionamento.

Vale la pena notare, infine, che SelA dimerica è inattiva, avendo perso la capacità di legare il tRNA, il quale viene legato da due dimeri e processato dai siti catalitici dei dimeri adiacenti (Itoh et al. 2013); la conservazione dei residui catalitici, comunque, suggerisce la conservazione di un qualche tipo di attività. Il legame di un eventuale substrato più piccolo al sito attivo di SelA eliminerebbe la necessità della struttura decamerica ad anello, indispensabile per legare i tRNA.

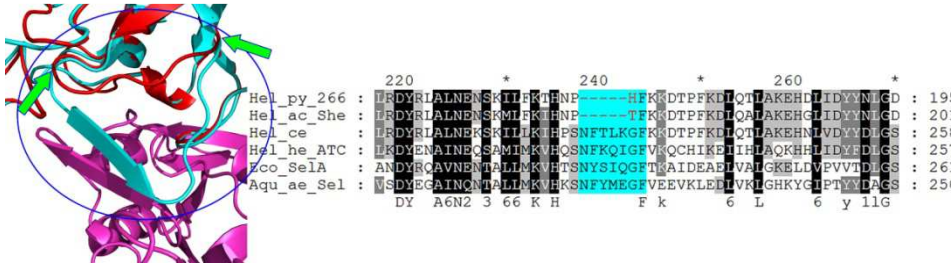

Figura 4.6 *Destra: allineamento della regione di interazione dimero-dimero delle sequenze di SelA di vari organismi, ilombreggiatura azzurra indica il tratto deleto in HpSelA e Ha SelA, Hel_py_266 = H. pylori 26695, Hel_ac_She = H. acinonychis, Hel_ce = H. cetorum, Hel_he_ATC = H. hepaticus, Eco_SelA = E. coli, Aqu_ae_Sel = A. aeolicus; Sinistra: dettaglio allineamento della struttura di AaSelA (azzurro e viola) con un modello realizzato con PHYRE2 della struttura di HpSelA (rosso) relativo alla regione allineata. Il cerchio e le frecce evidenziano il tratto mancante in HpSelA e HaSelA.*

4.4 - CONCLUSIONI

La presenza del gene *selA* in *Helicobacter pylori* e la mancanza di informazioni sulle sue capacità di utilizzare il selenio ci hanno spinto ad investigare più approfonditamente sul ruolo del selenio nel metabolismo del batterio. Queste indagini hanno evidenziato come le selenoproteine e il sistema Sel siano elementi che richiedono una particolare attenzione da parte dei ricercatori. Molte delle selenoproteine e dei geni per i $tRNA^{Sec}$ identificati in questo studio erano stati annotati in modo non corretto o, addirittura, non erano stati annotati affatto. Sebbene le selenoproteine possano essere identificate tramite omologia con proteine esistenti, il fatto che nelle loro sequenze codificanti Sec sia codificata da un codone di stop può facilmente far sì che i programmi di annotazione automatica commettano degli errori nel trovarle quando vengono analizzati i dati dei sequenziamenti. I $tRNA^{Sec}$ sfuggono all'individuazione tramite i modelli per i tRNA canonici e modelli specifici devono essere impiegati per trovarli. I $tRNA^{Sec}$ degli ε-proteobatteri, in particolare, hanno una struttura particolare, diversa da quella classica dei $tRNA^{Sec}$.

L'uso di modelli di covarianza per l'identificazione degli elementi SECIS e degli accorgimenti particolari nell'applicazione di BLAST hanno consentito l'individuazione di una nuova famiglia di selenoproteine, DUF466, che rappresentano in primo esempio di proteine con una Sec come ultimo residuo al C-terminale. Dal momento che Sec è codificata da un codone di stop e l'allineamento di BLAST non include mismatch terminali questa famiglia non era stata identificata nelle precedenti

annotazioni e altre famiglie di proteine che, similmente, hanno una selenocisteina terminale possono sfuggire ai normali metodi di ricerca. Non c'è nessuna funzione proposta per DUF466 ma noi riteniamo che, vista la presenza di una selenocisteina, questa famiglia di proteine possa avere un'attività redox.

I risultati presentati suggeriscono che *H. pylori* non sia un utilizzatore di selenocisteina e che il suo gene *selA* sia un residuo del sistema Sel perso che è stato reclutato per una nuova funzione. Ciò è corroborato dai cambiamenti nella pressione selettiva agente su *selA* nell'antenato comune di *H. pylori* e *H. acinonychis* che sembra essere concomitante con la perdita del sistema Sel. La perdita del dominio terminale e le specifiche mutazioni accumulate nelle sequenze di *H. pylori* e *H. acinonychis* suggeriscono che HpSelA e HaSelA abbiano perso sia la capacità di legare il $tRNA^{Sec}$ sia la capacità di formare la struttura quaternaria decamerica, pur mantenendo un'organizzazione dimerica e la capacità di legare il cofattore PLP. Ciò è in accordo con la nostra ipotesi che HpSelA abbia una funzione enzimatica ma che questa funzione sia differente da quella originaria dell'enzima. Una ricerca in tutti i procarioti ha identificato altri casi di SelA prive del dominio terminale. I membri di questa sottofamiglia di SelA si ritrovano in organismi che non sono utilizzatori di selenocisteina (es.: *Agrobacterium radiobacter*) o che utilizzano Sec ma hanno un paralogo di SelA a lunghezza completa, il che è in accordo con la nostra ipotesi che SelA possa assumere un'altra funzione oltre a quella di selenocisteina sintasi. È possibile che questa seconda funzione sia una caratteristica intrinseca

di tutte le SelA e che la funzione di selenocisteina sintasi la soppianti in quegli organismi utilizzatori di Sec che non dispongono di un SelA paralogo. È altresì possibile che questa seconda funzione possa essere stata evoluta indipendentemente da quegli organismi che hanno perso la capacità di utilizzare Sec. Vista la conservazione dei residui che legano il PLP e il fosfato di $SePO_3$, è possibile ipotizzare che la nuova funzione di SelA coinvolga un'ammina primaria e un gruppo fosfato.

È stato visto che il fattore principale che spinge al mantenimento del sistema Sel è la presenza di una Fdh contenente Sec (Romero et al. 2005). È pertanto possibile che la perdita della Fdh da parte di *H. pylori* sia stato il motivo che ha portato ad un rilassamento della pressione selettiva che manteneva il sistema Sel nel batterio. Dal momento che la perdita del sistema Sel sembra coincidere temporalmente con l'adattamento all'ospite umano, è possibile che questa caratteristica di *H. pylori* faccia parte degli adattamenti che il batterio ha sviluppato per colonizzare lo stomaco dell'uomo.

Parte di questo lavoro è stato oggetto di pubblicazione (Cravedi et al. 2015).

5 - RICERCA DI GENI ORTOLOGHI

5.1 - INTRODUZIONE

5.1.1 - PERDITA E ACQUISIZIONE DI GENI

La pressione selettiva non agisce in modo uguale su tutti i geni di un organismo. In base alla funzione dei geni, al metabolismo dell'organismo e all'habitat che esso abita la pressione selettiva agisce più su alcuni che geni che su altri. Si può avere, inoltre, la perdita di geni non più necessari per l'occupazione di una certa nicchia ecologica e l'acquisizione di altri geni (tramite trasferimento orizzontale, origine *de novo* e duplicazione genica seguita da divergenza funzionale) che migliorano l'adattamento dell'organismo al suo habitat.

La diversa azione della pressione selettiva sui geni in base al loro diverso ruolo in un organismo è un fenomeno noto. Nei batteri la selezione purificatrice (ovvero la pressione selettiva che tende ad eliminare cambiamenti nella sequenza di un gene o una proteina) agisce in maniera diversa a seconda del fatto che il gene sia essenziale o non essenziale. Quantificando la pressione selettiva col rapporto tra sostituzioni non sinonime e sostituzioni sinonime nelle sequenze codificanti (dN/dS), è stato visto che tale rapporto è significativamente più basso per i geni classificati come essenziali, il che indica che la selezione purificatrice agisce con più forza su questi geni (Jordan et al. 2002). Studi su *Saccharomyces cerevisiae* hanno, inoltre, rivelato come la

pressione selettiva non agisca in modo uguale su tutti i geni non essenziali, ma che la forza della selezione purificatrice dipende dal grado di dispensabilità del gene, ovvero sull'impatto che la perdita di quel gene può avere sulla fitness dell'organismo pur non compromettendone la sopravvivenza (Hirsh and Fraser 2001). In merito alle interazioni patogeno-ospite, uno studio di genomica comparativa su *Vibrio cholerae* e le specie vicine ha identificato quali geni fossero sottoposti ad una selezione positiva (ossia una pressione selettiva che favorisce l'accumulo di mutazioni in un gene). Sebbene la maggioranza dei geni di *Vibrio spp.* Fosse sottoposta a selezione purificatrice, un gruppo di geni sono sottoposti a selezione positiva e questi geni ricadono in quelle classi di geni che mediano le interazioni tra ospite e patogeno e la pressione selettiva su questi geni sembrerebbe rappresentare una corsa agli armamenti biologici tra patogeno ed ospite (Rabby et al. 2015).

La perdita o l'acquisto selettivo di geni possono fornire indicazioni sugli adattamenti di un organismo all'ambiente in cui vive. Uno studio su *Burkholderia spp.* ha rivelato come la perdita e l'acquisto di geni siano un fenomeno abbastanza frequente all'interno di un genere batterico e che l'acquisto di nuovi geni sia prevalente sulla perdita, probabilmente per via di eventi di trasferimento genico orizzontale. Inoltre è stato visto come gran parte dei geni che sono stati persi o acquisiti in *Burkholderia spp.* siano geni correlati al metabolismo (Zhu et al. 2011). Uno studio su varie specie di *Streptococcus* ha scoperto come l'acquisizione di geni (principalmente tramite trasferimento orizzontale) sembri essere un fenomeno di adattamento all'ambiente, dal momento che la maggior

parte dei geni acquisiti tramite trasferimento orizzontale ricadono in due categorie: geni che migliorano la sopravvivenza del patogeno nella sua nicchia e geni che aiutano il patogeno nella competizione con altre specie e che questi geni neo-acquisiti sono spesso sottoposti a selezione positiva (Marri, Hao, and Golding 2006). Una cosa simile è stata vista in *Prochlorococcus spp.*, dove i più recenti eventi di perdita e guadagno di geni interessano geni riguardanti caratteristiche della superficie della cellula che probabilmente sono sottoposti a pressione selettiva da parte di predatori e fagi e che giocano un probabile ruolo nella difesa dalle sostanze tossiche e nello sfruttamento dei nutrienti nei diversi habitat occupati (Kettler et al. 2007).

5.1.2 - ALLA RICERCA DEGLI ORTOLOGHI

Il concetto di ortologia è un concetto molto importante nell'ambito delle analisi filogenetiche. La definizione di questo concetto è la seguente: due geni si dicono ortologhi quando sono omologhi (quindi condividono un gene ancestrale comune) e la separazione tra essi è stata dovuta ad un evento di speciazione. Questo concetto è normalmente usato in contrasto con quello di paralogia, ovvero quel rapporto evolutivo tra geni omologhi che si sono separati a seguito di un evento di duplicazione genica. A ciò si aggiungono altri processi, ovvero il trasferimento orizzontale di geni, la fusione e la fissione geniche.

L'identificazione dei geni ortologhi ha un'importanza duplice. innanzitutto, dal momento che i geni ortologhi sono, sostanzialmente, lo stesso gene distribuito in più genomi a seguito della divergenza delle

specie, l'esatta identificazione dei geni ortologhi è necessaria al fine di attuare ricostruzioni della filogenesi delle specie. Tracciando i cambiamenti avvenuti all'interno di un gene quando esso è presente in più genomi è possibile ricavare informazioni riguardo alla filogenesi delle specie in esame. In secondo luogo, spesso (ma non sempre) un gene ortologo mantiene la stessa funzione o una funzione simile durante l'evoluzione, pertanto è possibile estendere le annotazioni funzionali di un gene agli ortologhi degli organismi vicini, sebbene sia spesso necessario da parte del ricercatore adottare la propria discrezione per capire se fidarsi o meno dell'annotazione carpita da un gene ortologo.

La complessità dei meccanismi di divergenza dei geni, rende spesso difficoltosa l'operazione di identificazione precisa degli ortologhi. Eventi di fusione e fissione genica e scambio, perdita o acquisizione di domini funzionali possono introdurre difficoltà nella ricostruzione dei corretti rapporti di ortologia. A ciò si aggiungono eventi di perdita di geni e di incompletezza dei genomi disponibili che possono portare a conclusioni errate sulla presenza o meno dell'ortologo di un gene in una determinata specie. Infine, la combinazione di questi eventi con la presenza di geni paraloghi può dare vita a filogenesi complesse, in cui un determinato gene ha dei geni paraloghi che non sono presenti allo stesso modo in tutti gli organismi considerati, il che porta a ricostruzioni errate dei rapporti filogenetici fra le sequenze (Kuzniar et al. 2008).

5.1.3 - ORTHOLUGE

Uno dei metodi generalmente usati per l'identificazione dei geni ortologhi in un gruppo di organismi è il metodo delle migliori hit reciproche di BLAST (BRH – Best Reciprocal Hits). Questo metodo prevede che due geni *a* e *b* appartenenti a due genomi *A* e *B* rispettivamente siano considerati ortologhi quando effettuando una ricerca con BLAST (Altschul et al. 1990, 1997) in *B* usando *a* come query si ottiene *b* come risultato più significativo ed effettuando la ricerca reciproca (in *A* usando *b* come query) si ottiene *a* come risultato più significativo (Bork et al. 1998; Tatusov et al. 2003).

Questo metodo, però, può essere prono ad errori nel caso in cui un gene sia assente in uno dei genomi analizzati (perché quell'organismo ha perso quel gene o perché c'è stato un errore nel sequenziamento) e quel gene abbia dei paraloghi. In casi del genere può capitare che la procedura del BRH identifichi erroneamente il gene paralogo superstite come gene ortologo. Inoltre, sebbene il rapporto di paralogia sia spesso associato col mantenimento della funzione del gene, non sempre ciò è vero. Possono esserci stati cambiamenti nella pressione selettiva e nella velocità di divergenza in un organismo oppure in un organismo potrebbe essere avvenuta una duplicazione genica antecedente alla separazione delle specie, seguita da un rilassamento della pressione selettiva su uno dei due paraloghi.

Il programma Ortholuge si pone come obiettivo l'identificazione dei "veri ortologhi", ossia di quei geni che divergono esclusivamente in seguito ad

un evento di speciazione e la cui storia evolutiva rifletta la storia evolutiva delle specie. Ortholuge è un miglioramento della procedura BRH che, dati due genomi *A* e *B*, prevede l'introduzione di un terzo genoma *C* che sia un outgroup rispetto ad *A* e *B*, ossia la cui separazione dagli altri due preceda la separazione tra *A* e *B*. Il programma identifica delle terne di geni *a, b* e *c* che siano tutti BRH tra loro imponendo, quindi, che la relazione di BRH tra *a* e *b* abbia un'origine evolutiva. A ciò viene aggiunto un filtro basato sulle distanze evolutive tra *a, b* e *c*. Le distanze vengono calcolate per ogni terna di possibili ortologhi e i rapporti tra esse vengono usati per discriminare tra i veri ortologhi e quei geni che sono hanno avuto una storia evolutiva che non supporta le relazioni filogenetiche tra le specie (Fulton et al. 2006).

Dal momento che, per la nostra ricerca, era necessario seguire l'evoluzione, ovvero la comparsa e la scomparsa dei geni lungo un albero filogenetico, questa procedura, essendo basata su considerazioni di natura evolutiva, ci è sembrata particolarmente adatta al nostro scopo.

5.1.4 - FILOGENESI ACCURATE

L'identificazione dei corretti rapporti filogenetici tra le specie considerate è essenziale per qualunque considerazione sull'evoluzione degli organismi in esame. Al giorno d'oggi, la ricostruzione filogenetica si basa su caratteri molecolari, ossia sul confronto tra geni omologhi appartenenti alle diverse e sull'osservazione delle differenze tra essi tramite l'applicazione di modelli matematici e statistici per ricostruire la storia evolutiva delle specie. Il problema resta su quali e quanti di questi

caratteri omologhi bisogna scegliere per la ricostruzione filogenetica e su quali modelli bisogna applicare.

Attualmente i metodi per la ricostruzione filogenetica cadono in tre categorie: metodi basati sulle distanze, metodi di massima parsimonia e metodi di massima verosimiglianza. I primi prevedono l'uso di matrici di sostituzione che vengono usate per calcolare la distanza evolutiva tra le sequenze a partire dal loro allineamento; le distanze vengono poi usate per collocare le sequenze su un albero filogenetico usando vari algoritmi. I metodi di massima parsimonia stabiliscono la filogenesi selezionando l'albero che richiede il minor numero di variazioni tra i caratteri. I metodi di massima verosimiglianza, infine, utilizzano funzioni che calcolano la probabilità che un albero possa produrre i dati osservati (verosimiglianza) e scelgono l'albero più verosimile (Delsuc, Brinkmann, and Philippe 2005).

Per quanto riguarda la scelta dei caratteri, i caratteri considerati più di frequente nelle ricostruzioni filogenetiche dei batteri sono gli RNA ribosomiali, in particolare l'rRNA 16S, il quale è ubiquitario, facilmente sequenziabile ed è stato raramente soggetto ad eventi di trasferimento orizzontale che ne complicano la filogenesi (Wang and Wu 2013). Basandosi sugli rRNA 16S è stato realizzato l'albero filogenetico della vita che ha portato alla proposizione della classificazione dei viventi in domini tuttora utilizzata (Woese, Kandler, and Wheelis 1990). Un altro gruppo di caratteri che si prestano alla ricostruzione filogenetica sono le proteine ribosomiali. Similmente all'rRNA 16S, le proteine ribosomiali sono altamente conservate, vengono raramente perse o duplicate e sono

raramente soggette a trasferimento orizzontale. Confrontando tutte le proteine ribosomiali dei proteobatteri si è visto che sono un set di caratteri che porta con sé un buon segnale filogenetico (Ramulu et al. 2014). Esistono, infine, metodi per identificare caratteri che non appartengano ad una singola classe di proteine. Uno di questi metodi prevede l'utilizzo di ricerche di omologia su un set rappresentativo di genomi e l'uso di modelli markoviani sui gruppi di sequenze omologhe identificate per individuare dei set di caratteri composti anche da centinaia di geni che siano specifici per ogni phylum batterico, i quali possono essere usati per ricostruire la filogenesi all'interno di quel phylum (Wang and Wu 2013).

5.2 - MATERIALI E METODI

5.2.1 - COSTRUZIONE DELL'ALBERO FILOGENETICO

L'albero filogenetico usato nella presente analisi è lo stesso usato nella sezione 2.

5.2.2 - LINGUAGGI DI PROGRAMMAZIONE E PROGRAMMI PREESISTENTI

Lo script *find_orthologs.sh* che gestisce la procedura di ricerca dei geni omologhi nei genomi selezionati è scritto in linguaggio Bash. Lo script richiama ulteriori script scritti in Perl ed R. La versione di Perl utilizzata è la 5.10.1. La versione di R utilizzata è la 3.2.2. La procedura utilizza il programma Ortholuge (Fulton et al. 2006) ottenibile presso il seguente URL: http://www.pathogenomics.ca/ortholuge/download.html . La procedura utilizza, inoltre, i programmi invocati da Ortholuge, tra cui BLAST (Altschul et al. 1990). Gli script utilizzati in questa sezione sono disponibili online sulla cartella Dropbox e su GitHub (Tabella Supplementare S6.1).

5.2.3 - SCHEMA GENERALE DELLO SCRIPT

Lo script *find_orthologs.sh* è stato sviluppato per applicare Ortholuge (che lavora su terne di genomi in cui uno è outgroup rispetto agli altri due) ad un qualsiasi numero di genomi selezionati su un albero filogenetico. L'albero di partenza deve riportare sulle foglie i taxid degli organismi, ovvero i codici numerici che in GenBank identificano univocamente ogni organismo. L'utente seleziona i genomi su cui

lavorare e un genoma query per i cui geni si vogliono trovare i geni ortologhi corrispondenti nel set di organismi selezionato. Il programma identifica tutte le possibili combinazioni dell'organismo query con altri due organismi che costituiscono una tripletta query-ingroup-outgroup (QIO), recupera i genomi da GenBank e applica Ortholuge ad ogni tripletta, quindi filtra i risultati e li unisce in un'unica tabella finale.

5.2.4 - ANALISI DELL'ALBERO

L'utente fornisce in input un albero filogenetico riportante i taxid degli organismi e indica, inoltre, quale organismo verrà considerato l'organismo query e altri due organismi (organismi delimitanti), in modo da selezionare gli organismi tra essi compresi. È possibile anche indicare un quarto organismo che debba fungere da outgroup per tutti gli organismi selezionati. La procedura utilizza lo script *trim_tree.r* scritto in linguaggio R e utilizzante il pacchetto APE (Popescu et al. 2012) per estrarre un sottoalbero contenente solo gli organismi desiderati. Lo script utilizza la funzione *getMRCA* di APE per risalire al più vicino antenato comune degli organismi delimitanti, quindi utilizza la funzione *getDescendants* (http://blog.phytools.org/2012/01/function-to-get-descendant-node-numbers.html) per ottenere la lista dei nodi da esso discendenti. Da questa lista vengono rimosse le foglie che, sulla rappresentazione grafica dell'albero, non sono comprese tra i due organismi limitanti, quindi vengono aggiunti la query e l'eventuale outgroup. La funzione *drop.tip* infine, viene usata per rimuovere dall'albero tutte le altre foglie.

Una seconda procedura in R (*tree2triplets.r*) identifica le terne QIO di organismi. *tree2triplets.r* analizza il sottoalbero ottenuto con *trim_tree.r* considerando ogni organismo come possibile ingroup. Per ogni coppia query-ingroup, lo script usa *getMRCA* e *getDescendants* per ottenere la lista di tutti gli organismi che condividono lo stesso ancestore comune con la query e l'ingroup e che sono, quindi, ingroup rispetto ad entrambi. La lista complementare di organismi costituisce la lista degli outgroup per quella coppia query-ingroup. *tree2triplets.r* fornisce in output una lista di tutte le possibili terne QIO di organismi (indicati con i loro taxid) così determinate. A queste terne vengono aggiunte le terne query-ingroup-ingroup (QII) che vengono utilizzate per determinare i geni BRH fra ogni coppia di organismi.

5.2.5 - DOWNLOAD DEI GENOMI

Lo script *taxid2an.pl* in linguaggio Perl utilizza la tabella di corrispondenza *prokaryotes.txt* ottenuta dal sito FTP di GenBank (ftp.ncbi.nlm.nih.gov) per convertire i taxid degli organismi negli accession numbers dei loro genomi e proteomi completi. Agli accession numbers viene anteposto un prefisso numerico per numerarli in modo che in seguito possa essere ricostruito l'ordine con cui compaiono nell'albero. Lo script *download_genomes.pl* in linguaggio Perl utilizza la funzione *wget* della Shell Linux per scaricare i file contenti i proteomi (o le sequenze codificanti) e le sequenze genomiche complete degli organismi dal sito FTP di GenBank.

5.2.6 - BLAST E tBLASTn

Lo script *reformat_genomes.pl* in linguaggio Perl viene usato per cambiare le intestazioni delle sequenze in formato fasta in modo che includano solo l'accession number della sequenza (questo migliora la leggibilità della tabella finale) e, nel caso si stia lavorando con sequenze nucleotidiche codificanti, *reformat_genomes.pl* include, oltre all'accession number del cromosoma, anche le relative coordinate.

Se si sta lavorando con sequenze codificanti, lo script *translate.pl* in linguaggio Perl le traduce in sequenze aminoacidiche usando il frame +1. Le analisi successive verranno eseguite usando le sequenze tradotte.

Al fine di ottimizzare il tempo di esecuzione, le ricerche di omologia con BLASTp richieste da Ortholuge vengono eseguite a parte, in modo da evitare di effettuare la stessa ricerca più volte. BLASTp viene lanciato con limite di significatività $E=10^{-4}$ e con le altre impostazioni di default. Vengono, inoltre, effettuate anche delle ricerche con tBLASTn usando come query le sequenze proteiche dell'organismo query e come database di ricerca le sequenze dei cromosomi completi degli altri organismi; il limite di significatività è impostato a 10^{-4} e l'output è in formato tabulare; le altre impostazioni sono a default.

I risultati di BLAST e tBLASTn sono salvati nelle omonime cartelle.

5.2.7 - ORTHOLUGE

Basandosi sul file contenente l'elenco delle terne di organismi, per ogni terna viene lanciato Ortholuge con le impostazioni *–skip-blast* e *–quiet*. Il

programma utilizza una directory temporanea come directory di lavoro di Ortholuge che viene svuotata dopo ogni ricerca. Per ogni ricerca di Ortholuge, il file di output contenente le terne di geni ortologhi con le relative distanze e i rapporti fra esse viene salvato in una directory destinata a contenere i risultati di Ortholuge organizzati in sottocartelle, una per ogni coppia query-ingroup contenete le ricerche effettuate con tutti gli outgroup per quella coppia.

5.2.8 - FILTRO RISULTATI

I risultati di Ortholuge vengono filtrati usando lo script in R *filter_ratios.r*. Lo script calcola per ogni file di output di Ortholuge media e deviazione standard delle distribuzioni dei rapporti delle distanze e determina i cutoff per ciascuno di essi secondo quanto spiegato nella sezione risultati. Per i rapporti R1 (distanza query-ingroup/distanza query-outgroup) e R2 (distanza query-ingroup/distanza ingroup-outgroup) vengono considerati i valori fino al 99° percentile, per R3 (distanza query-ingroup/distanza ingroup-outgroup) i valori tra il 5° e il 95° percentile. I risultati filtrati vengono collocati in una cartella temporanea. I risultati non filtrati vengono copiati in una seconda directory temporanea. In seguito al filtro i risultati delle terne QII vengono esclusi per via del rapporto R2 che assume valore infinito. Durante questo passaggio viene anche prodotto un file riportante la distanza media con relativa deviazione standard tra tutte le sequenze per ogni coppia query-ingroup.

5.2.9 - ANALISI DEI VERI NEGATIVI

L'analisi dei veri negativi è stata effettuata usando come campione le terne *H. pylori – H. acinonychis – H. cetorum, H. pylori – H. hepaticus – W. succinogenes, H. pylori – H. acinonychis – C. jejuni, H. pylori – H. hepaticus – C. jejuni*. Su ciascuna terna è stato applicato Ortholuge, quindi è stato usato lo script *introduce_true_negatives.pl* (in Perl) per sostituire, nei geni dell'ingroup, il miglior risultato di BLAST col secondo migliore risultato ovunque fosse disponibile più di un risultato. È stato usato, quindi, lo script *ortholuge-align.pl* del pacchetto di Ortholuge per calcolare le nuove distanze. I grafici che riportano le distribuzioni dei rapporti sono stati ottenuti con gli script *ortoluge-scatter.pl* e *ortholuge-hist.pl* del pacchetto Ortholuge.

5.2.10 - UNIONE DEI RISULTATI

A partire dai risultati di Ortholuge viene prodotta una tabella (outwalking) per ogni coppia query-ingroup riportante i geni ortologhi tra quei due organismi e le sequenze outgroup che li confermano. Questa operazione viene effettuata tramite lo script in Perl *merge_ortholuge_outwalking.pl*. L'operazione viene effettuata sia per i risultati filtrati che per i risultati in filtrati.

Ai file di outwalking vengono quindi applicati gli script *consensus_count.pl* e *consensus_filter.pl* in linguaggio Perl. Questi script contano, per ogni coppia di geni ortologhi tra un organismo query e un organismo ingroup, in quante terne QIO è stata ritrovata quella coppia; questo valore viene chiamato consenso. Il secondo script procede,

quindi, a scartare tutte le coppie che non sono confermate da almeno *N* outgroup, quindi tiene solo i risultati con un consenso minimo di *N* (con *N* specificato dall'utente). La procedura viene applicata sia agli outwalking filtrati che agli outwalking non filtrati. Dal momento che i risultati delle terne QII sono presenti solo nei risultati non filtrati, negli outwalking non filtrati un singolo outgroup di conferma indica che i due geni sono BRH tra loro senza avere outgroup di conferma (l'unico "outgroup" di conferma è l'ingroup stesso), mentre quelli confermati da almeno un vero outgroup avranno un consenso minimo di 2.

Successivamente, lo script *merge_ortholuge_inwalking.pl* in linguaggio Perl unisce i file di outwalking in un unico file in cui, per ogni gene dell'organismo query, vengono indicati i geni ortologhi negli altri organismi confermati da almeno *no* outgroup e, nel caso dei risultati non filtrati, quali geni sono BRH del gene della query in quegli organismi in cui non è stato trovato un ortologo supportato da almeno *no* outgroup. Questa procedura è applicata sia ai risultati filtrati che a quelli non filtrati, ottenendo, quindi, due tabelle.

L'ultimo passaggio è operato dallo script in Perl *ortholuge_merge.pl* che confronta le due tabelle e ne fornisce una unica in cui gli ortologhi sono classificati come SSD (Supporting Species Divergence) se presenti nella tabella finale dei risultati filtrati, NSD (Not Supporting species Divergence) se presenti solo nella tabella dei risultati non filtrati e BRH (Best Reciprocal Hits) se confermati solo dalle terne QII.

5.2.11 - RICERCHE BLAST E TBLASTN

Facoltativamente, l'utente può scegliere di controllare, in quei casi in cui in un organismo *A* non viene trovato un possibile ortologo del gene *x*, se il gene *x* ha almeno un risultato di BLAST in *A* o, se ciò non avviene, se la ricerca tBLASTn con *x* come query e la sequenza cromosomica di *A* come target dà risultati. Queste ricerche vengono effettuate dagli script *find_best_blast.pl* e *find_best_tblastn.pl* (in Perl) che aggiungono gli eventuali risultati alla tabella catalogandoli come BBH (Best BLAST Hit) o BTH (Best tBLASTn Hit).

5.2.12 - PARAMETRI UTILIZZATI NELLE RICERCHE RIPORTATE NEI RISULTATI

Sono state lanciate più ricerche con *find_orthologs.sh* variando la query ma tenendo fissi gli altri organismi e gli altri parametri. Gli organismi usati come query sono: *Helicobacter pylori 26695* (taxid: 85962), *Helicobacter acinonychis* str. Sheeba (taxid: 382638), *Helicobacter cetorum* MIT 00-7128 (taxid: 182217), *Helicobacter bizzozeronii* CIII-1 (taxid: 1002804), *Helicobacter hepaticus* ATCC 51449 (taxid: 235279). I parametri tenuti fissi tra le varie ricerche sono

- Organismi delimitanti: *Helicobacter pylori* Shi417 (taxid: 1163739) e *Nitratiruptor sp. SB155-2* (taxid: 387092)
- Outgroup: *Nautilia profundicola* AmH (taxid: 598659)
- Consenso minimo *no*: 1
- Limite su R1: media + 1,5 deviazioni standard (limite superiore)

- Limite su R2: media + 1,5 deviazioni standard (limite superiore)
- Limite su R3: media ± 3 deviazioni standard
- Opzioni: *-skip bbh,bth –correct-table NC_018939.1,NZ_CP011330.1*

Abbiamo deciso di non considerare I geni che vengono trovati solo come miglior risultato di BLAST o tBLASTn. L'opzione *–correct-table* serve per quei casi in cui per un organismo ci sono più sequenziamenti, in modo da evitare taxid doppi.

Lo script *filter_results.pl* (in Perl) è stato usato per analizzare la conservazione dei geni. I risultati filtrati in base alla conservazione sono stati importati su foglio di calcolo e uniti e analizzati manualmente.

5.3 - RISULTATI

Ortholuge (Fulton et al. 2006) è un stato utilizzato per identificare i geni ortologhi tra due organismi usando un terzo organismo come outgroup e individuando tutte le terne di geni appartenenti ciascuno ad uno dei tre organismi che siano migliori hit reciproche di BLAST (BRH) fra loro. Successivamente sono state calcolate le distanze evolutive tra i geni di ogni terna. Le distanze sono state usate per calcolare dei rapporti che descrivano quanto le distanze fra le sequenze siano in accordo con la filogenesi delle specie. I rapporti considerati sono i seguenti: R1=distanza ingroup1-ingroup2/distanza ingroup1-outgroup; R2= distanza ingroup1-ingroup2/distanza ingroup2-outgroup; R3=distanza ingroup1-outgroup/distanza ingroup2-outgroup. Per i geni ortologhi R1 e R2 dovrebbero essere <1 (a indicare che le sequenze degli ingroup sono tra loro più vicine rispetto di quanto non lo siano alla sequenza dell'outgroup) e R3 dovrebbe avere un valore vicino ad 1 (a indicare che l'outgroup è circa equidistante a ciascuno degli ingroup).

Abbiamo messo a punto una procedura automatica per applicare Ortholuge ad un set contenente un qualsiasi numero di genomi selezionati su un albero e lo abbiamo applicato allo stesso set di genomi preso in considerazione nella sezione 2. *Sulfurovum* sp. NBC37-1 e *Nitratiruptor* sp. SB155-2 sono stati esclusi dall'analisi poiché la tabella di corrispondenza reperita non riporta gli URL relativi alla posizione dei loro genomi sul sito FTP dell'NCBI.

I risultati completi delle ricerche sono disponibili online sulla cartella Dropbox (Tabella Supplementare S6.1).

5.3.1 - DETERMINAZIONE DELLE SOGLIE PER I RAPPORTI FRA LE DISTANZE

Al fine di discriminare tra geni ortologhi e non ortologhi, è necessario stabilire una soglia per i valori dei rapporti che consenta di selezionare solo i veri ortologhi. Al fine di determinare tali soglie abbiamo analizzato la distribuzione dei valori dei tre rapporti R1, R2 e R3 calcolati sui risultati di Ortholuge per 4 terne di organismi: *H. pylori, H. acinonychis, H. cetorum* (tre organismi vicini fra loro)*; H. pylori, H. acinonychis, Campylobacter jejuni* (ingroup 1 e ingroup2 vicini, outgroup appartenente ad un'altra famiglia)*; H. pylori, H. hepaticus, Wolinella succinogenes* (tre organismi più lontani fra loro rispetto alla prima terna, ma facenti parte della stessa famiglia)*; H. pylori, H. hepaticus, C. jejuni* (ingroup1 e ingroup2 lontani ma appartenenti alla stessa famiglia, outgroup di un'altra famiglia). I grafici riportanti le distribuzioni sono indicati in Figura 5.1. Le distribuzioni di R1 e R2 tendono ad essere simili (tranne nel caso C, dove la distribuzione di R2 è più spostata verso 1 rispetto a quella di R1) e le distribuzioni di R3 tendono ad attestarsi attorno ad 1. Osservando le distribuzioni è evidente come stabilire dei valori soglia fissi per R1, R2, ed R3 non si presti bene alla nostra procedura, visto come le medie e le deviazioni standard delle distribuzioni variano a seconda degli organismi in esame. Per questa ragione abbiamo deciso di implementare un algoritmo che determini le

soglie di R1, R2 e R3 per ogni terna di organismi, calcolate sulla base delle distribuzioni dei rapporti.

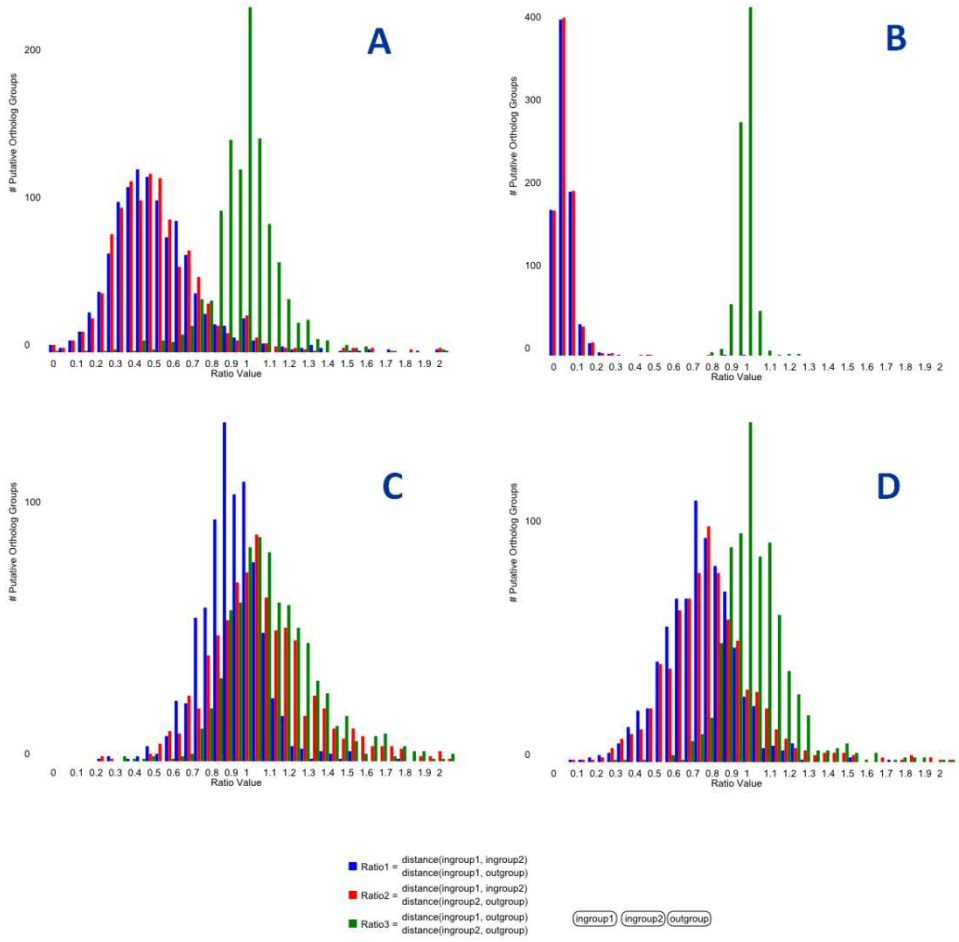

Figura 5.1 *Distribuzione dei rapporti R1 (blu), R2 (rosso) e R3 (verde) per varie terne di organismi. A: Ingroup1 = H. pylori, Ingroup2 = H. acinonychis, Outgroup = H. cetorum; B: Ingroup1 = H. pylori, Ingroup2 = H. acinonychis, Outgroup = C. jejuni; C: Ingroup1 = H. pylori, Ingroup2 = H. hepaticus, Outgroup = W. succinogenes; D: Ingroup1 = H. pylori, Ingroup2 = H. hepaticus, Outgroup = C. jejuni.*

Gli autori che hanno sviluppato Ortholuge (Fulton et al. 2006) hanno riportato una strategia per determinare le soglie dei rapporti che prevede l'introduzione di veri negativi, ossia di risultati che simulino un'errata predizione da parte del programma, e di valutare come le

distribuzioni dei rapporti variano in seguito all'inserimento dei veri negativi. Abbiamo implementato questa strategia sostituendo nelle 4 ricerche già menzionate quanti più possibili geni dell'ingroup2 con dei veri negativi (procedura dettagliata nei Materiali e Metodi). Abbiamo, quindi, osservato come variavano le distribuzioni (Figura 5.2). Quello che si nota di più è il cambiamento della distribuzione di R3 con la comparsa di una secondo massimo della distribuzione a sinistra rispetto al massimo originario. Ciò è in linea con le attese, dal momento che i veri negativi sono i secondi migliori risultati di BLAST nell'ingroup2, quindi ci si aspetta un aumento della distanza tra sequenze ingroup2 e outgroup (al denominatore in R3) mentre le distanze ingroup1-outgroup restano invariate. Per quanto riguarda le distribuzioni di R1 sia ha un aumento delle frequenze dei valori nelle code destre delle distribuzioni e una diminuzione delle frequenze dei valori nel centro della distribuzione. Anche questo è in linea con le attese, dal momento che ci si aspetta che la distanza ingroup1-ingroup2 (al numeratore nel rapporto) sia più alta a seguito della scelta del secondo miglior risultato di BLAST mentre la distanza ingroup1-outgroup resti invariata. Le distribuzioni di R2, infine, mostrano la comparsa di un secondo massimo vicino ad 1 (Figura 5.2 A e B) o, ove il massimo originario fosse già vicino ad 1, un aumento delle frequenze dei valori nel vicino intorno destro di 1. Questo comportamento potrebbe essere spiegato col fatto che, scegliendo il secondo miglior risultato di BLAST nell'ingroup2 si simula l'individuazione di un paralogo. Le distanze ingroup1-ingroup2 e ingroup2-outgroup aumentano entrambe, ma questo aumento sembra essere più importante nella distanza con la sequenza dell'organismo più

evolutivamente prossimo, probabilmente dovuto al fatto che un aumento del numero di sostituzioni per sito abbia un peso maggiore in sequenze che si sono separate da meno tempo (le sostituzioni già accumulate dall'antenato dei due ingroup a partire dalla sua separazione con l'outgroup non si vedono nel confronto tra le due sequenze ingroup). Questa analisi dei veri negativi è stata effettuata solo inserendo i veri negativi tra le sequenze dell'ingroup2, ma ci aspettiamo una situazione speculare nel caso le sequenze vengano inserite tra i risultati dell'ingroup1: uno spostamento verso 1 della distribuzione di R1, un aumento delle frequenze della coda destra della distribuzione di R2 e la comparsa di un massimo a destra del massimo originario nella distribuzione di R3.

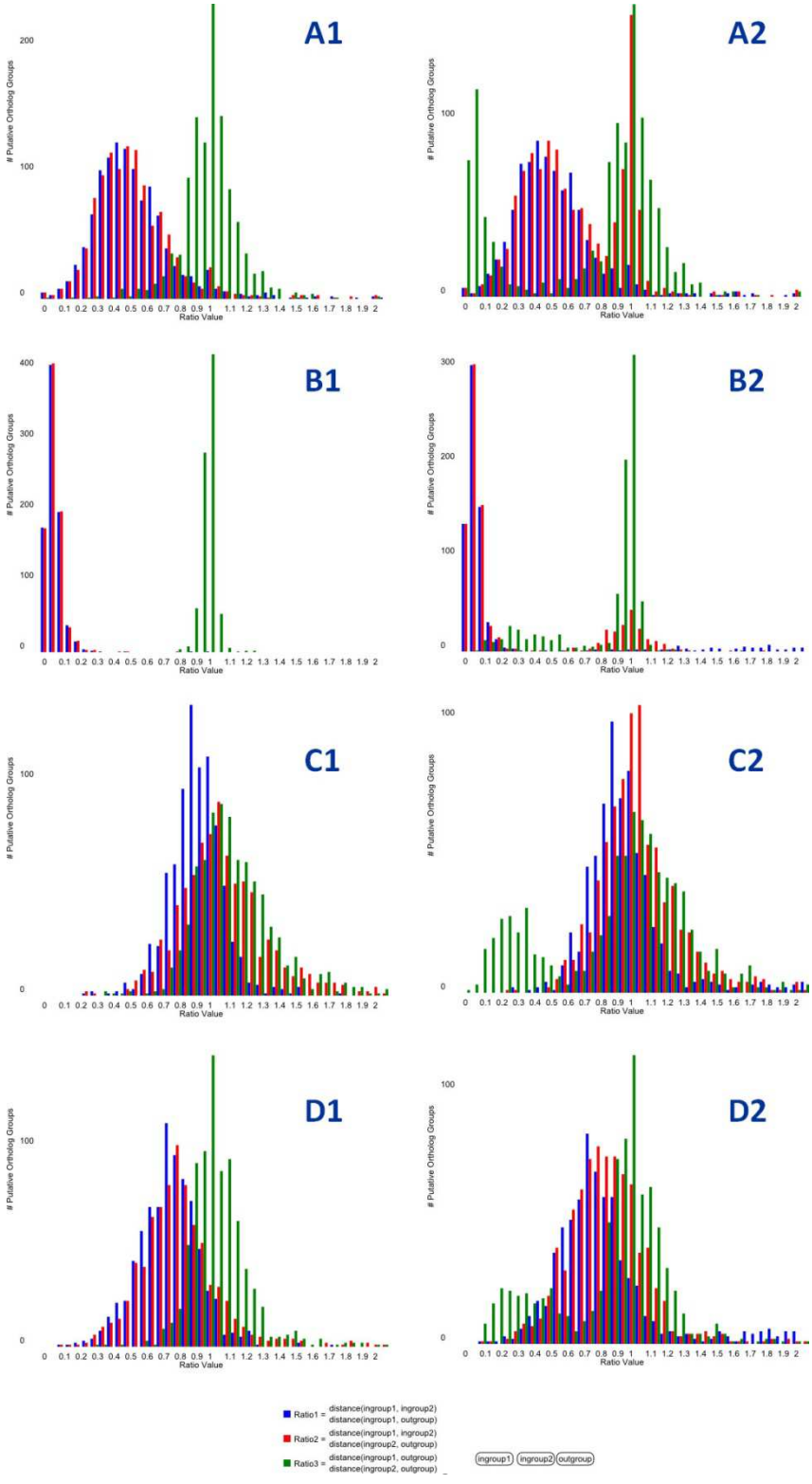
A1
A2
B1
B2
C1
C2
D1
D2
Putative Ortholog Groups
Ratio Value
Ratio1 = distance(ingroup1, ingroup2) / distance(ingroup1, outgroup)
Ratio2 = distance(ingroup1, ingroup2) / distance(ingroup2, outgroup)
Ratio3 = distance(ingroup1, outgroup) / distance(ingroup2, outgroup)
ingroup1
ingroup2
outgroup

Figura 5.2 *Confronto tra le distribuzioni dei rapporti R1 (blu), R2 (rosso) e R3 (verde) prima (A1, B1, C1, D1) e dopo (A2, B2, C2, D2) l'introduzione dei veri negativi per varie terne di organismi. A: Ingroup1 = H. pylori, Ingroup2 = H. acinonychis, Outgroup = H. cetorum; B: Ingroup1 = H. pylori, Ingroup2 = H. acinonychis, Outgroup = C. jejuni; C: Ingroup1 = H. pylori, Ingroup2 = H. hepaticus, Outgroup = W. succinogenes; D: Ingroup1 = H. pylori, Ingroup2 = H. hepaticus, Outgroup = C. jejuni.*

Basandosi sull'analisi dei grafici abbiamo individuato le soglie dei valori di R1, R2 e R3 per discriminare tra veri ortologhi e non. Chiamando M1 e S1 la media e la deviazione standard della distribuzione di R1, M2 e S2 la media e la deviazione standard della distribuzione di R2 e M3 e S3 la media e la deviazione standard della distribuzione di R3 le soglie vengono determinate per ogni terna di organismi come segue:

- R1>M1-1,5*S1
- R2>M2-1,5*S2
- M3-3*S3<R3<M3+3*S3

Medie e deviazioni standard sono calcolate sui primi 99 percentili delle distribuzioni di R1 e R2 e sui 90 percentili centrali delle distribuzioni di R3 per eliminare gli effetti dei valori estremi dei rapporti dovuti a casi di trasferimento orizzontale di geni tra un ingroup e l'outgroup. Tali situazioni danno valori di R1 o R2 estremamente alti e valori di R3 estremamente bassi o estremamente alti a seconda che il trasferimento sia avvenuto tra ingroup1 e outgroup (distanza ingroup1-outgroup prossima a 0) o tra ingroup2 e outgroup (distanza ingroup2-outgroup prossima a 0) rispettivamente.

Queste soglie sui valori dei tre rapporti vengono usate per attribuire un grado di confidenza ai rapporti di ortologia. Nella tabella finale fornita come output dalla procedura, per ogni gene dell'organismo query

vengono indicati quali geni sono ad esso ortologhi negli altri organismi in esame. In base al fatto che abbiano passato o meno il filtro sui rapporti delle distanze, i possibili ortologhi vengono suddivisi in SSD (Supporting Species Divergence) se almeno una ricerca di Ortholuge ha indicato tra i risultati quel gene come possibile ortologo e i rapporti delle distanze per la terna a cui quella coppia apparteneva hanno superato il filtro delle distanze; NSD (Not supporting Species Divergence) se Ortholuge ha trovato almeno una terna di geni con quella coppia come sequenze ingroup, ma nessuno di tali risultati ha superato il filtro sulle distanze; BRH (Best Reciprocal Hits) se nessuna sequenza outgroup è stata trovata essere migliore hit reciproca di BLAST con le due sequenze ma le due sequenze sono migliori hit reciproche fra loro. SSD è il massimo grado di confidenza e indica quei geni il cui rapporto di ortologia è stato confermato da un outgroup e dai rapporti fra le distanze geniche con almeno un outgroup. La ragione per cui nella ricerca vengono incluse anche le sequenze BRH è perché le sequenze specifiche di un gruppo di organismi non vengono individuate da Ortholuge in quegli organismi per cui non ci sono degli outgroup all'interno del gruppo. Ad esempio, considerando gli *Helicobacter* gastrici, per *Helicobacter mustelae* non sono disponibili organismi outgroup all'interno del gruppo degli *Helicobacter* gastrici, quindi le sequenze esclusive degli *Helicobacter* gastrici non verranno identificate in *H. mustelae* per via dell'assenza di sequenze outgroup con cui confrontarle. Per questa ragione le sequenze BRH vengono incluse nella tabella quando Ortholuge non trova risultati per esse e vengono considerate come buone sequenze ortologhe quando in gran parte degli organismi per cui l'outgroup era disponibile

sono stati trovati ortologhi SSD, mentre i BRH si ritrovano solo in quegli organismi per cui l'outgroup non era disponibile.

5.3.2 - RISULTATI DELLE RICERCHE

5.3.2.1 - Validazione della procedura

Abbiamo fatto una prima ricerca utilizzando come query il proteoma di *Helicobacter pylori* 26695 e come bersaglio della ricerca gli tutti gli altri organismi che avevamo analizzato nella sezione 2. I risultati della ricerca sono stati filtrati conservando solo quelle sequenze query che avevano degli ortologhi in tutte i ceppi di *Helicobacter pylori*, in modo da determinare il genoma core della specie per confrontare i nostri risultati con quelli in letteratura. Il filtro è stato applicato sia considerando le sequenze NSD come ortologhe, sia considerandole come non ortologhe. La stessa procedura è stata applicata anche ponendo come limite la presenza di ortologhi in almeno l'80% dei ceppi di *H. pylori* considerati. La ricerca ha trovato 1159 geni presenti in tutti i ceppi considerando le sequenze NSD come ortologhe. Il numero si abbassa a 1132 considerando le sequenze NSD come non ortologhe. Le sequenze presenti in almeno l'80% degli *H. pylori* considerati sono 1272 considerando le sequenze NSD e 1262 non considerando gli NSD. Abbiamo confrontato il risultato con la versione online di Ortholuge (OrtholugeDB; http://www.pathogenomics.sfu.ca/ortholugedb/) che ha individuato 1161 geni presenti in tutti i ceppi. OrtholugeDB, però, individua tutti i risultati come BRH. Il motivo di ciò, probabilmente, risiede in limitazioni tecniche dell'interfaccia della piattaforma online.

OrthologeDB consente solo la selezione di al massimo 9 organismi oltre l'organismo query, pertanto è stato possibile effettuare la ricerca solo all'interno dei 10 ceppi di *H. pylori*. Ciò ha, probabilmente, fatto emergere il già citato problema dell'assenza di organismi outgroup nella ricerca, pertanto Ortholuge ha potuto trovare solo corrispondenze di tipo BRH. I 29 geni trovati in più rispetto alla nostra procedura (considerando come ortologhi solo le sequenze SSD) sono stati, probabilmente, inclusi erroneamente a causa di questa limitazione tecnica.

Un altro strumento online per l'individuazione del genoma core è MicroScope (http://www.genoscope.cns.fr/agc/microscope/compgenomics/pancoreTool.php?) (Miele, Penel, and Duret 2011), il quale stabilisce la composizione del genoma core (inteso come i geni presenti in tutti i ceppi) in 1205 geni, 73 in più rispetto al nostro risultato che considera gli NSD come non ortologhi e 46 in più rispetto al risultato che, invece, li considera come ortologhi. Abbiamo verificato quale fosse la sovrapposizione fra i risultati cercando quali geni fossero stati trovati da entrambi i metodi e quali da uno solo dei due. È emerso che dei geni individuati da MicroScope, 22 indicati come presenti in *H. pylori* 26695 non erano presenti nel proteoma di riferimento ottenuto dal sito dell'NCBI. Per i restanti geni, la sovrapposizione è riassunta in Figura 5.3. Considerando i risultati NSD come ortologhi i risultati comuni ad entrambi i metodi sono 1126, 33 sono trovati solo dalla nostra procedura e 68 solo da MicroScope, altrimenti i risultati comuni sono 1112, quelli

esclusivi della nostra procedura 20 e quelli trovati solo da MicroScope sono 82.

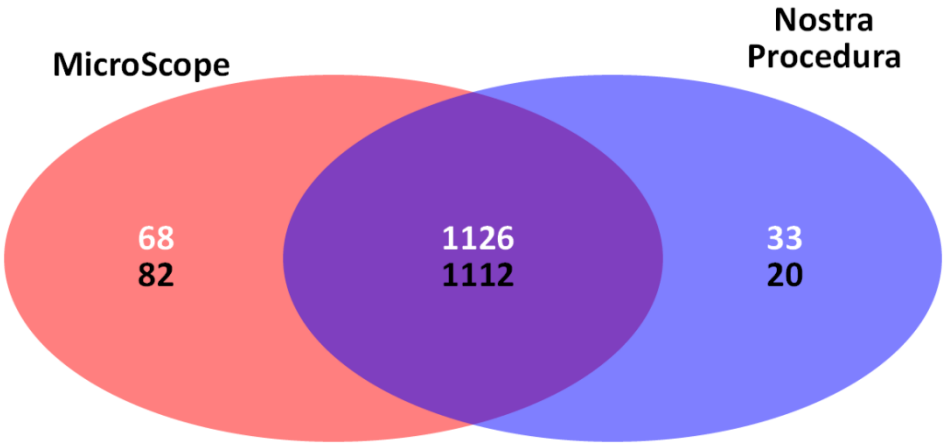

Figura 5.3 *Sovrapposizione dei risultati tra la nostra procedura e la ricerca con MicroScope. In nero sono indicati i risultati del confronto tra la ricerca con MicroScope e l'output della nostra procedura filtrato in modo da tenere solo i geni presenti come SSD o BRH in tutti i ceppi di* H. pylori*. In bianco sono indicati i risultati del confronto tra la ricerca con MicroScope e l'output della nostra procedura filtrato in modo da tenere solo i geni presenti come SSD o BRH in tutti i ceppi di* H. pylori.

La differenza è, probabilmente, giustificabile con la differenza del metodo utilizzato per l'individuazione. MicroScope utilizza un algoritmo che considera come ortologhe quelle sequenze che sono BRH, allineabili per almeno l'80% della loro lunghezza e identiche a livello aminoacidico per almeno l'80%. Tali parametri, sebbene stringenti, non escludono casi di trasferimento orizzontale o filogenesi complesse in cui due sequenze possono essere BRH ma ciò è dovuto alla perdita di un gene paralogo in uno degli organismi.

Fonti di letteratura stimano le dimensioni del genoma core di *H. pylori* intorno a 1200 geni (Lu et al. 2014), con delle stime che questo numero possa scendere intorno a 1100 a seconda dei genomi considerati

(Gressmann et al. 2005). A confronto con questi dati la nostra predizione di un genoma core di 1132 geni cade probabilmente nella parte inferiore dell'intervallo, probabilmente per via della stringenza delle condizioni con cui la procedura attribuisce la predizione di ortologia. Chalker et al., 2001 hanno riportato una lista di geni di *H. pylori* che hanno testato per verificarne l'essenzialità. Di questi geni, 33 sono stati individuati come essenziali. La nostra procedura ha trovato 31 di questi geni come presenti come SSD in tutti i ceppi di *H. pylori* del nostro set (Tabella 5.1). I 2 geni restanti risultano, invece, assenti in un solo ceppo. Questi due geni sono HP1448 (non trovato nel ceppo G27) e HP1159 (non trovato nel ceppo Gambia94/24). Una ricerca con BLAST nel database nr con le sequenze di questi geni del ceppo 26695 non ha trovato omologhi in G27 e Gambia94/24 per HP1448 e HP1159 rispettivamente (E<10), indicando quindi che gli ortologhi nei due ceppi non sono stati individuati dalla procedura non per un difetto dell'algoritmo ma per l'assenza delle sequenze omologhe nel database.

Accession number	**Locus**	**Trovato in tutti gli *H. pylori*?**
NP_206822	HP0020	sì
NP_206874	HP0074	sì
NP_206886	HP0086	sì
NP_207170	HP0372	sì
NP_207306	HP0509	sì
NP_207383	HP0588	sì
NP_207384	HP0589	sì
NP_207385	HP0590	sì
NP_207386	HP0591	sì
NP_207451	HP0657	sì
NP_207452	HP0658	sì
NP_207173	HP0735	sì

NP_207534	HP0740	sì
NP_207594	HP0801	sì
NP_207601	HP0808	sì
NP_207623	HP0830	sì
NP_207625	HP0832	sì
NP_207766	HP0975	sì
NP_207828	HP1038	sì
NP_207890	HP1099	sì
NP_207891	HP1100	sì
NP_207899	HP1108	sì
NP_207900	HP1109	sì
NP_207901	HP1110	sì
NP_207902	HP1111	sì
NP_207931	HP1140	sì
NP_207950	HP1159	Manca in Gambia94/24
NP_207955	HP1164	sì
NP_208023	HP1231	sì
NP_208049	HP1257	sì
NP_208163	HP1372	sì
NP_208176	HP1385	sì
NP_208209	HP1418	sì
NP_208239	HP1448	manca in G27
NP_208344	HP1553	sì
NP_208354	HP1563	sì

Tabella 5.1 *Geni essenziali di* H. pylori *(Chalker et al. 2001). La terza colonna indica se il gene è stato trovato o meno dalla nostra procedura.*

5.3.2.2 - Geni esclusivi degli Helicobacter *gastrici*

Sono state effettuate delle ricerche utilizzando come query i proteomi di *H. acinonychis, H. cetorum MIT 00-7128* e *H. bizzozeronii*. I risultati delle ricerche sono stati analizzati per individuare i geni selettivamente conservati negli *Helicobacter* gastrici. È stato applicato un filtro che trattenesse solo quei geni presenti in almeno l'80% dei ceppi di *H. pylori* e almeno l'80% delle altre specie gastriche, che fossero assenti nelle due

specie non gastriche (*H. hepaticus* e *H. cinaedi*) e che fossero presenti in al più il 20% delle altre specie di ε-proteobatteri in esame. Il filtro è stato applicato sia considerando le sequenze NSD come ortologhe che come non ortologhe. Successivamente i risultati delle ricerche sono stati confrontati manualmente per compilare una lista in cui fosse ridotto il pregiudizio derivante dal basare le ricerche su un genoma query (mentre la ricerca può trovare le sequenze del genoma query assenti negli altri genomi non può fare il contrario, quindi si rende necessario confrontare i risultati delle ricerche con query differenti). Sono state selezionati geni per cui il filtro venisse rispettato in almeno una ricerca, ma non venisse smentito dalle altre. Infine i risultati ottenuti considerando le sequenze NSD come ortologhe (risultati *"full"*) con quelli ottenuti considerandole non ortologhe (risultati *"ssd"*). Per ogni risultato siamo andati ad analizzare gli schemi di conservazione con più attenzione e abbiamo controllato i rapporti delle distanze calcolati da Ortholuge per vedere se ci fossero sequenze NSD i cui rapporti superavano di poco le soglie imposte e che, quindi, potessero essere considerati comunque come SSD.

Complessivamente sono stati trovati 128 geni presenti selettivamente nelle specie gastriche di *Helicobacter.* Di queste, 32 sono stati trovati solo nelle ricerche *full* (15 di essi sono stati confermati) e 19 solo nelle ricerche *ssd* (di cui 15 confermati). Dei restanti 77, in seguito ad un'analisi manuale, 66 sono stati confermati. In totale i geni confermati avere ortologhi nei gastrici sono 96. I risultati sono riportati nella Tabella Supplementare S5.1.

Le sequenze di *H. pylori* sono state usate per effettuare ricerche di omologia con BLAST ($E<10^{-4}$) negli organismi al di fuori degli ε-proteobatteri. Dalle ricerche è emesso che, dei geni confermati, 52 sono esclusivi degli ε-proteobatteri.

I geni di *H. pylori* così individuati sono stati raggruppati in base ai processi cellulari a cui sono stati attribuiti in base alle loro annotazioni. I risultati sono riportati in Tabella 5.2.

Funzione	**Numero di Geni**	**Geni esclusivi degli ε-proteobatteri**
Metabolismo degli zuccheri	6	0
Metabolismo degli aminoacidi	7	0
Metabolismo degli acidi carbossilici	2	0
Metabolismo lipidi	2	1
Altri processi metabolici	3	0
Modellamento della parete	3	1
Motilità	2	1
Competenza	3	1
Virulenza	3	3
Nucleasi	2	1
Trasporto	6	0
Altro	3	2
Proteine di membrana	13	12
Proteine ipotetiche	41	30
Totale	**96**	**52**

Tabella 5.2 *Geni selettivamente conservati dagli* Helicobacter *gastrici raggruppati per funzione. La terza colonna indica quanti geni di ogni classe funzionale sono esclusivi degli ε-proteobatteri.*

Si nota che il grosso dei geni individuati codifica per proteine a funzione ignota. In questo gruppo rientrano anche le proteine di membrana, dal momento che le loro annotazioni riportano semplicemente l'indicazione

di localizzazione sulla membrana esterna ma non forniscono suggerimenti sulla funzione della proteina. La maggioranza di queste proteine a funzione ignota (78%) sembra essere comparsa negli ε-proteobatteri, il che suggerisce che siano proteine di invenzione relativamente recente. Delle proteine a cui è stato possibile attribuire una possibile funzione, 9 sembrano essere coinvolte in interazioni con l'ospite (virulenza + motilità + modellamento della parete) e 19 sembrano avere parte in meccanismi di utilizzo dei nutrienti (metabolismo degli zuccheri e degli aminoacidi, trasporto). Fra i geni legati al metabolismo degli aminoacidi rientra SelA che ha sì degli omologhi in (quasi) tutti gli organismi, ma lo stato di SSD cambia tra *H. felis* e *H. mustelae*: le sequenze tra *H. pylori* e *H. felis* sono SSD fra loro, è così sono fra loro le sequenze da *H. mustelae* in poi, ma i due gruppi sono fra loro NSD, il che riflette il cambiamento nella pressione selettiva evidenziato nel Capitolo 2. Nelle proteine di trasporto sono incluse NixA, necessaria per l'assunzione di nickel e il corretto funzionamento dell'ureasi (Mobley et al. 1995), e una proteina classificata come lattato permeasi, la cui funzione è stata confermata sperimentalmente (Iwatani et al. 2014). Nel genoma di *H. pylori* sono presenti due geni annotati come codificanti per lattato permeasi (Tomb et al. 1997). Una delle due (HP0140) sembra essere ortologa a quella di *E. coli* (BRH da ricerca con BLAST online), l'altra, anch'essa omologa alla lattato permeasi di *E. coli* (HP0141) è quella presente in questa lista. *E. coli* ha anche un trasportatore del glicolato che può trasportare anche lattato (Iwatani et al. 2014), ma questo trasportatore non è BRH con HP1041, il che suggerisce che HP1041 si sia originata in seguito a duplicazione genica di

HP1040 nelle specie gastriche che hanno, quindi, due trasportatori del lattato. È interessante notare, però, che il trasportatore del glicolato di *E. coli* non ha omologhi in *H. pylori* a parte HP0140 e HP0141, il che suggerisce che, possibilmente, gli *Helicobacter* gastrici non hanno un trasportatore del glicolato apposito ma la funzione è svolta dal trasportatore del lattato duplicato. Infine, un membro di questa lista è la proteina associata alla catalasi KapA (HP0874), attualmente in corso di caratterizzazione da parte del nostro gruppo di ricerca.

Il fatto di aver trovato SelA fra queste proteine suggerisce che, probabilmente, anche altri fra i geni che mostrano un pattern di presenza simile (SSD nei gastrici, NSD al di fuori) siano nella stessa situazione. Siamo, pertanto, andati a vedere quali dei geni trovati fossero stati individuati dalla ricerca “ssd” e non dalla ricerca “full”, ossia quei geni che vengono considerati come non presenti come ortologhi solo quando la definizione di ortologia comprende solo i geni SSD. Questi geni sono riportati in Tabella 5.3.

Accession Number	Descrizione	Classe funzionale
NP_207128	ketol-acid reductoisomerase	Metabolismo degli acidi carbossilici
NP_207725*	6-carboxy-5,6,7,8-tetrahydropterin synthase	Metabolismo dei cofattori
NP_207891	phosphogluconate dehydratase	Metabolismo degli zuccheri
NP_207957	glucose-6-phosphate isomerase	Metabolismo degli zuccheri
NP_208071	bifunctional indole-3-glycerol phosphate synthase/phosphoribosylanthranilate isomerase	Metabolismo degli zuccheri
NP_208204*	7-cyano-7-deazaguanine reductase	Metabolismo delle basi azotate
NP_208253**	flagellar motility protein	Motilità
NP_208304	selenocysteine synthase	Sistema Sel
NP_208321**	hypothetical protein HP1531	Proteina Ipotetica

Tabella 5.3 *Geni con un comportamento simile a SelA. Questi geni sono ortologhi SSD fino ad* H. felis *e NSD al di fuori in* H. mustelae *dei gastrici.*
**: assente negli* Helicobacter *non gastrici*
***: presente solo negli* Helicobacter

Per questi geni è possibile ipotizzare che seguano un comportamento simile a SelA, ossia che siano in realtà presenti in tutti gli organismi (tranne le 4 eccezioni segnalate in didascalia) ma che siano stati oggetti di cambiamenti significativi nella pressione selettiva che li ha portati ad avere rapporti fra le distanze anomali.

5.3.2.3 - Possibili geni coinvolti nell'adattamento all'ospite umano

Allo scopo di investigare quali siano i possibili adattamenti di *H. pylori* all'ospite umano abbiamo analizzato i risultati delle ricerche per vedere

quali geni fossero stati selettivamente conservati in *H. pylori* e, eventualmente, in *H. acinonychis*. La ragione per cui *H. acinonychis* è stato incluso come organismo in cui i geni siano opzionalmente presenti è che si suppone che *H. acinonychis* si sia separato da *H. pylori* in seguito ad un salto d'ospite dall'uomo ai grandi felini nelle ultime centinaia di migliaia di anni, probabilmente in seguito ad episodi di predazione (Eppinger et al. 2006).

Abbiamo confrontato i risultati delle ricerche eseguite usando come query *H. pylori* 26695 e *H. acinonychis*. È stato applicato un filtro per tenere solo i geni presenti come SSD in almeno l'80% degli *H. pylori*, al più l'80% degli altri *Helicobacter* e al più l'80% degli altri organismi. Il filtro è stato lasciato indifferente nei confronti di *H. acinonychis*. Questo è stato fatto per tenere in conto che alcuni geni acquisiti da *H. pylori* per l'adattamento all'ospite umano potrebbero essere stati persi da *H. acinonychis* quando ha effettuato il salto d'ospite.

Come risultato abbiamo avuto 99 geni selettivamente presenti in *H. pylori*. Di questi 99, 3 sono stati scartati in seguito ad un'analisi manuale degli schemi di presenza/assenza dei geni a causa di gravi discordanze tra i risultati di *H. pylori* e quelli di *H. acinonychis*. I geni identificati sono riportati nella Tabella Supplementare S5.2. Dei 94 geni confermati dall'analisi manuale, 53 sono presenti anche in *H. acinonychis*.

I risultati sono stati raggruppati in base al possibile ruolo biologico attribuito tramite le annotazioni presenti sulle sequenze in banca dati. I risultati raggruppati per funzione sono riportati in Tabella 5.4.

Funzione	Numero di Geni	Geni presenti in *H. acinonychis*
Processamento acidi nucleici	1	1
Isola di patogenicità CAG	23	0
Enzimi di restrizione	5	4
Metabolismo degli aminoacidi	1	1
Metabolismo dei cofattori	2	1
Metabolismo dei lipidi	1	1
Metabolismo delle basi azotate	4	4
Metabolismo degli zuccheri	1	1
Motilità	1	0
Modellamento della parete	4	4
Trasporto	3	2
Proteine di membrana	1	1
Proteine ipotetiche	46	33
Totale	**94**	**53**

Tabella 5.4 *Geni selettivamente presenti in* H. pylori.

Si nota come *H. acinonychis* abbia perso circa la metà dei geni che erano stati selettivamente acquisiti da *H. pylori*, il che è probabilmente dovuto al cambio d'ospite dall'uomo ai grandi felini.

Anche nei geni selettivamente presenti in *H. pylori*, buona parte è composta da proteine ipotetiche, la cui maggioranza è presente anche in *H. acinonychis*. Un altro ampio gruppo di geni è quello dei geni codificanti per le proteine dell'isola di patogenicità CAG, la cui presenza ci aspettavamo tra i geni specifici di *H. pylori* (Noto and Peek 2012). È interessante notare come le categorie funzionali con più membri (a parte il gruppo delle proteine ipotetiche e dell'isola CAG) siano gli enzimi di restrizione, seguiti da enzimi del metabolismo delle basi azotate e della parete cellulare e proteine coinvolte nel trasporto di sostanze dentro e fuori la cellula. Sono stati identificati e caratterizzati vari fagi che

infettano *H. pylori* (Heintschel von Heinegg, Nalik, and Schmid 1993; Luo et al. 2012; Uchiyama et al. 2012, 2013) ed è noto che gli enzimi di restrizione rappresentino un'importante linea di difesa dei batteri contro questi loro aggressori (Labrie, Samson, and Moineau 2010). Il fatto che il 5% dei geni esclusivi di *H. pylori* codifichi per enzimi di restrizione indica che gli adattamenti specifici del batterio sono dedicati non solo ad una corsa agli armamenti con l'ospite umano, ma anche con i parassiti del batterio stesso. Questi 5 geni, quindi, rappresentano non tanto un adattamento allo stomaco umano ma ai nemici naturali di *H. pylori*. Tra i geni coinvolti nel metabolismo della parete cellulare ci sono tre enzimi della via di sintesi del lipopolisaccaride di membrana e HP0310, annotata come polisaccaride deacetilasi. Essa è stata sperimentalmente caratterizzata come una peptidoglicano deacetilasi atipica, priva di peptide segnale per l'indirizzamento al periplasma e, probabilmente, agente su una porzione del peptidoglicano diversa dalla componente polisaccaridica. È interessante notare che momento le peptidoglicano deacetilasi sono enzimi importanti nell'evasione della risposta immunitaria dell'ospite (Shaik et al. 2011).

Anche in questo caso abbiamo cercato dei geni che seguissero un comportamento analogo a quello di SelA, ovvero la cui predizione SSD/NSD cambiasse tra il gruppo *H. pylori/H. acinonychis* e gli altri organismi, trovando i due geni in Tabella 5.5.

Accession Number	Descrizione	Classe funzionale
NP_207490	methylhydantoinase	Metabolismo delle basi azotate
NP_207557	hypothetical protein HP0764	Proteina ipotetica

Tabella 5.5 *Geni presenti il cui stato di ortologia SSD/NSD cambia tra il gruppo* H. pylori/H. acinonychis *e gli altri organismi.*

Nessuno di questi geni è presente, neanche come NSD, fuori dagli *Helicobacter* gastrici. Inoltre, NP_207490 (HP0696) è assente in *H. cetorum* e NP_207557 (HP0764) è assente in *H. mustelae*.

5.3.2.4 - Geni selettivamente persi da H. pylori *e dagli* Helicobacter *gastrici*

Ci siamo interessati a quei geni che fossero stati persi in seguito all'adattamento di *H. pylori* all'ospite umano. Abbiamo usato la nostra procedura per effettuare una ricerca usando *H. hepaticus* come query e abbiamo filtrato i risultati trattenendo solo quei geni assenti in almeno l'80% dei ceppi di *H. pylori* e presenti in almeno l'80% delle specie di *Helicobacter* gastriche, in almeno il 60% delle altre *Helicobacteraceae* e in almeno il 60% degli altri organismi come ortologhi SSD. Lo stesso filtro è stato applicato alle altre ricerche effettuate in precedenza e i risultati sono stati confrontati manualmente. Così facendo abbiamo ottenuto una lista di 8 geni, riportati in Tabella 5.6.

Accession Number	Annotazione	Gruppo Funzionale
NP_860044	hypothetical protein HH0513	Proteina Ipotetica
NP_859689	nitrate reductase catalytic subunit	Metabolismo dell'azoto
NP_859692	periplasmic nitrate reductase	Metabolismo dell'azoto
NP_859694	periplasmic nitrate reductase component NapD	Metabolismo dell'azoto
NP_859690	quinol dehydrogenase periplasmic component	Produzione di Energia
NP_860272	selenocysteine-specific elongation factor SelB	Sistema Sel
NP_861265	selenophosphate synthase	Sistema Sel

Tabella 5.6 *Geni selettivamente assenti in* H. pylori

Nessuno di questi geni mostra un comportamento come quello di SelA, ovvero venire trovato come SSD fuori dal gruppo *H. acinonychis/H. pylori* e come SSD in *H. acinonychis* e *H. pylori*.

Come ci aspettavamo, in questo elenco di geni rientrano SelB e SelD. Inoltre si nota come la nitrato reduttasi citoplasmatica sia stata persa in *H. pylori*.

Infine abbiamo cercato quei geni che fossero stati persi dagli *Helicobacter* gastrici ma conservati da quelli non gastrici e che, quindi, potessero essere stati persi in seguito all'adattamento all'ambiente gastrico. La ricerca è stata effettuata usando *Helicobacter hepaticus* come query e i risultati sono stati filtrati trattenendo quei geni assenti in almeno l'80% degli *H. pylori* e l'80% delle altre specie gastriche, e presenti in almeno il 60% dei non *Helicobacter*. Abbiamo ottenuto una

lista di 100 geni, riportata in Tabella Supplementare S5.3. I geni, classificati per processo biologico, sono riassunti in Tabella 5.7.

Funzione	Numero di Geni
Metabolismo degli aminoacidi	19
Ciclo dell'urea	8
Metabolismo delle basi azotate	6
Modificazioni degli acidi nucleici	3
Trasporto	3
Chemiotassi	2
Produzione di energia	2
Metabolismo degli acidi carbossilici	2
Metabolismo dei cofattori	2
Metabolismo dei lipidi	2
Regolatori trascrizionali	2
Trasduzione del segnale	2
Metabolismo degli zuccheri	1
Modificazioni delle proteine	1
Trasporto di elettroni	1
Altro	2
Proteine ipotetiche	41
Totale	**99**

Tabella 5.7 *Geni selettivamente persi dagli* Helicobacter *gastrici*

Si nota come gli *Helicobacter* gastrici abbiano perso molti enzimi del ciclo dell'urea e del metabolismo delle basi azotate. Essi sembrano, inoltre, aver perso una ferredossina (classe: trasporto di elettroni). Nel proteoma di *H. mustelae* sono annotate due ferredossine e una di esse è uno dei geni persi dagli *Helicobacter* gastrici.

Anche in questo caso abbiamo visto quali geni avessero un comportamento simile a SelA. Tali geni sono riportati in Tabella 5.8.

Accession number	Descrizione	Classe funzionale	Cambiamento SSD/NSD
NP_859707	S-ribosylhomocysteinase	Metabolismo degli aminoacidi	*H. felis/H. mustelae*
NP_859982*	branched-chain amino acid aminotransferase	Metabolismo degli aminoacidi	*H. felis/H. mustelae*
NP_860022*	glucose-6-phosphate isomerase	Metabolismo degli zuccheri	Gastrici/non gastrici
NP_860381*	dihydroxy-acid dehydratase	Metabolismo degli aminoacidi	Gastrici/non gastrici
NP_860636	hypothetical protein HH1105	Proteina ipotetica	*H. felis/H. mustelae*
NP_860647	hypothetical protein HH1116	Proteina ipotetica	*H. felis/H. mustelae*
NP_860735*	ketol-acid reductoisomerase	Metabolismo degli aminoacidi	*H. felis/H. mustelae*

Tabella 5.8 *Geni il cui stato SSD/NSD cambia tra* Helicobacter *gastrici e non gastrici o tra* H. felis *e* H. mustelae

Si può vedere come alcuni di questi risultati siano stati trovati anche dal filtro che selezionava solo i geni presenti nei gastrici. Ciò non sorprende, dal momento che uno dei filtri che abbiamo applicato considerava come assenti quei casi in cui la relazione di possibile ortologia era NSD, pertanto è logico aspettarsi che alcuni dei risultati per cui si ha un cambiamento della relazione di ortologia da SSD a NSD al limite tra specie gastriche e non gastriche (o ad esso vicino) faccia sì che il gene sia trovato da entrambi i filtri. Questi 4 geni, pertanto, sono da considerarsi presenti in tutti gli organismi ma soggetti a cambiamenti nella pressione selettiva.

5.3.2.5 - Conservazione e perdita dei geni in Helicobacter mustelae

Helicobacter mustelae è il primo degli *Helicobacter* gastrici ad essersi separato dal loro ancestore comune. Ci siamo domandati se l'adattamento all'ambiente gastrico fosse già stato completato in questo organismo, quindi abbiamo analizzato i risultati delle ricerche con cui abbiamo individuato i geni selettivamente conservati e selettivamente persi negli *Helicobacter* gastrici per vedere quanti di essi fossero presenti in *H. mustelae*. I risultati sono riassunti in Tabella 5.9.

	Geni presenti in *H. mustelae*	**Geni assenti in *H. mustelae***	**Totale**
Geni selettivamente conservati da almeno l'80% delle specie gastriche	51	46	95
Geni selettivamente persi da almeno l'80% dalle specie gastriche	47	48	95

Tabella 5.9 *Presenza e assenza dei geni precedentemente individuati in* H. mustelae

Si può notare come circa la metà dei geni selettivamente persi o selettivamente conservati dagli *Helicobacter* gastrici sia assente in *H. mustelae*, il che non è sortprendente, visto che *H. mustelae* è il primo degli *Helicobacter* gastrici ad essersi separato dal loro antenato comune, quindi è logico aspettarsi che abbia dei tratti in comune con gli *Helicobacter* non gastrici, gli organismi ad esso più strettamente

imparentati. Da questi conti sono stati esclusi i 4 geni trovati come duplicati tra l'elenco dei geni selettivamente presenti nelle specie gastriche e quello dei geni selettivamente assenti.

5.4 - CONCLUSIONI

L'adattamento all'ambiente gastrico e all'ospite umano da parte di *H. pylori* hanno richiesto specifici adattamenti. La nostra analisi ha rivelato come, dal punto di vista della perdita o guadagno di geni codificanti per proteine l'adattamento all'ambiente gastrico sembri aver coinvolto 196 geni (97 acquisiti e 99 persi), mentre l'adattamento all'ospite umano sembra averne coinvolti 101 (94 acquisiti e 7 persi). In totale, quasi 200 dei geni codificanti per proteine di *H. pylori* sembrano essere coinvolti nell'adattamento alla sua nicchia ecologica. Questo numero è, probabilmente, sottostimato, dal momento che non include geni non codificanti per proteine o geni presenti in tutti gli organismi ma per cui si hanno avuto cambiamenti nella pressione selettiva. Una parte di questi geni può essere, probabilmente, individuata effettuando considerazioni sulla predizione di ortologia assegnata dalla nostra procedura. Un esempio può essere SelA, le cui sequenze appartenenti agli *Helicobacter* gastrici (tranne *H. mustelae*) sono SSD fra loro, così come lo sono le sequenze appartenenti agli altri ε-proteobatteri, ma i due gruppi non sono SSD fra loro. Questi risultati possono essere interessanti poiché rappresentano geni separati da un evento di cambio della pressione ma comunque presenti in tutti gli organismi. Abbiamo, comunque, visto che casi del genere sono sporadici e sono individuabili tramite una curatura manuale dei risultati che richiede poco tempo.

È interessante notare come *H. mustelae* sembri trovarsi in una situazione intermedia tra le specie gastriche e le specie non gastriche, infatti circa la metà dei geni selettivamente presenti nelle specie gastriche sono assenti

in *H. mustelae* e viceversa. Questo è in linea con il fatto che *H. mustelae* sia un outgroup rispetto a tutti gli altri Helicobacter gastrici, quindi è atteso che alcuni dei geni tipici degli altri gastrici manchino in esso e che alcuni dei geni persi dagli altri gastrici siano ancora presenti in *H. mustelae.* Un'altra cosa interessante è che, mentre nell'adattamento all'ambiente gastrico perdita e guadagno di geni sembrano avere lo stesso peso, nell'adattamento all'ospite umano la forza predominante sembra essere il guadagno dei geni.

Osservando la distribuzione dei geni dell'adattamento nelle varie classi funzionali, si nota come l'adattamento coinvolga spesso enzimi appartenenti a vie metaboliche (escludendo dal conto le proteine ipotetiche e quelle non classificate si ha: 21/54 per i geni selettivamente presenti nelle specie gastriche, 9/48 per i geni selettivamente presenti in *H. pylori*, 3/7 per i geni selettivamente assenti in *H. pylori* e 17/56 per i geni selettivamente assenti nelle specie gastriche) e, in misura minore, sistemi di trasporto (7/54; 3/48; 0/7; 3/56 per gli stessi gruppi rispettivamente). Ciò è atteso, dal momento che è facile immaginare come un nuovo habitat richieda che il batterio abbia a che fare con nuove disponibilità di nutrienti nell'ambiente circostante e debba, quindi, adattare il suo metabolismo a ciò che trova. Di interesse sono anche le 5 nucleasi di restrizione facenti parte dell'adattamento all'ospite umano, dal momento che suggeriscono che lo stanziamento nello stomaco umano abbia comportato l'incontro da parte di *H. pylori* con una popolazione di fagi capaci di infettarlo diversa da quella a cui i suoi antenati che non abitavano l'uomo erano adattati.

6 - CONCLUSIONI

In questa tesi abbiamo analizzato alcuni aspetti di quelli che potrebbero essere gli adattamenti di *Helicobacter pylori* allo stomaco umano.

Vista l'importanza dell'omeostasi dei metalli e delle metalloproteine nel metabolismo e nei meccanismi di sopravvivenza del batterio, abbiamo applicato una procedura bioinformatica per l'identificazione su larga scala delle metalloproteine di *H. pylori* nel proteoma del batterio a partire dalle sole sequenze proteiche. Tramite questa procedura abbiamo identificato come metalloproteine 67 proteine precedentemente annotate come proteine ipotetiche, 6 delle quali non contengono domini con siti di coordinazione di metalli noti e sarebbero, quindi, state mancate dalle normali procedure di analisi bioinformatica.

Abbiamo analizzato il selenoproteoma degli ε-proteobatteri concludendo che *H. pylori* non è un utilizzatore di selenocisteina e che questo tratto sembra essere stato perso nell'adattamento all'ospite umano. Abbiamo anche analizzato l'unico superstite del sistema Sel in *H. pylori* (SelA) scoprendo come esso sia mantenuto come gene funzionante in *H. pylori*, sebbene esso abbia probabilmente perso la sua attività originaria. Le nostre analisi, comunque, suggeriscono che HpSelA mantenga una funzione enzimatica, sebbene diversa da quella dei SelA degli altri organismi.

Infine, abbiamo sviluppato una procedura che applica il programma Ortholuge (Fulton et al. 2006) per identificare i geni ortologhi in un set di

genomi selezionabili su un albero filogenetico. Le predizioni formulate tramite la nostra procedura tengono conto di considerazioni di natura evolutiva basate sulla distanza fra le sequenze e l'utilizzo di sequenze outgroup per confermare la predizione. Tramite la nostra procedura abbiamo individuato una lista di geni che potrebbero essere stati coinvolti nell'adattamento all'ambiente gastrico e all'ospite umano e abbiamo visto come in *H. mustelae* (il primo degli *Helicobacter* gastrici ad essersi separato dal loro antenato comune dopo che questo si era separato dall'antenato comune delle *Helicobacteraceae*) alcuni di questi geni seguono il pattern di conservazione tipico degli *Helicobacter* non gastrici.

7 - BIBLIOGRAFIA

Agriesti, Francesca et al. 2014. "FeON-FeOFF: The Helicobacter Pylori Fur Regulator Commutates Iron-Responsive Transcription by Discriminative Readout of Opposed DNA Grooves." *Nucleic acids research* 42(5):3138–51.

Ahmed, Niyaz, Mun Fai Loke, Narender Kumar, and Jamuna Vadivelu. 2013. "Helicobacter Pylori in 2013: Multiplying Genomes, Emerging Insights." *Helicobacter* 18(March):1–4.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of molecular biology* 215(3):403–10.

Altschul, Stephen F. et al. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25(17):3389–3402.

Asakura, Hiroshi et al. 2010. "Helicobacter Pylori HP0518 Affects Flagellin Glycosylation to Alter Bacterial Motility." *Molecular Microbiology* 78(September):1130–44.

Atherton, John C., Timothy L. Cover, Emanuele Papini, and John L. Telford. 2001. "Vacuolating Cytotoxin - Helicobacter Pylori - NCBI Bookshelf.pdf." in *Helicobacter pylori: Physiology and Genetics.*, edited by H. Mobeley, G. Mendz, and S. Hazell. ASM Press.

Azevedo, Nuno F., Janis Huntington, and Karen J. Goodman. 2009. “The Epidemiology of Helicobacter Pylori and Public Health Implications.” *Helicobacter* 14:1–7.

Bagchi, D., G. Bhattacharya, and S. J. Stohs. 1996. “Production of Reactive Oxygen Species by Gastric Cells in Association with Helicobacter Pylori.” *Free radical research* 24(6):439–50.

Barabino, Arrigo. 2002. “Helicobacter Pylori -Related Iron Deficiency Anemia: A Review.” *Helicobacter* 7(2):71–75.

Bennett, Hayley J. and Ian S. Roberts. 2005. “Identification of a New Sialic Acid-Binding Protein in Helicobacter Pylori.” *FEMS immunology and medical microbiology* 44(2):163–69. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/15866211).

Bertini, Ivano and Gabriele Cavallaro. 2010. “Bioinformatics in Bioinorganic Chemistry.” *Metallomics : integrated biometal science* 2(1):39–51.

Böck, A. et al. 1991. “Selenocysteine: The 21st Amino Acid.” *Molecular microbiology* 5(3):515–20.

Bonis, Mathilde, Chantal Ecobichon, Stephanie Guadagnini, Marie Christine Prévost, and Ivo G. Boneca. 2010. “A M23B Family Metallopeptidase of Helicobacter Pylori Required for Cell Shape, Pole Formation and Virulence.” *Molecular Microbiology* 78(4):809–19.

Bork, P. et al. 1998. “Predicting Function: From Genes to Genomes and Back.” *Journal of molecular biology* 283(4):707–25.

Borodovsky, Mark, Ryan Mills, John Besemer, and Alex Lomsadze. 2003. “Prokaryotic Gene Prediction Using GeneMark and GeneMark.hmm.” *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 4:Unit4.5.

Campbell, Barbara J., Annette Summers Engel, Megan L. Porter, and Ken Takai. 2006. “The Versatile ε-Proteobacteria: Key Players in Sulphidic Habitats.” *Nature Reviews Microbiology* 4(6):458–68.

Castellano, Sergi, Vadim N. Gladyshev, Roderic Guigó, and Marla J. Berry. 2008. “SelenoDB 1.0 : A Database of Selenoprotein Genes, Proteins and SECIS Elements.” *Nucleic acids research* 36(Databse Issue):D332–38.

Chalker, a. F. et al. 2001. “Systematic Identification of Selective Essential Genes in Helicobacter Pylori by Genome Prioritization and Allelic Replacement Mutagenesis.” *Journal of Bacteriology* 183(4):1259–68.

Chen, G., R. L. Fournier, S. Varanasi, and P. A. Mahama-Relue. 1997. “Helicobacter Pylori Survival in Gastric Mucosa by Generation of a pH Gradient.” *Biophysical journal* 73(2):1081–88.

Cioci, Gianluca, Laurent Terradot, Cyril Dian, Christoph Mueller-Dieckmann, and Gordon Leonard. 2011. “Crystal Structure of HP0721, a Novel Secreted Protein from Helicobacter Pylori.”

Proteins 79(5):1678–81. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/21365686).

Cravedi, P., G. Mori, F. Fischer, and R. Percudani. 2015. "Evolution of the Selenoproteome in Helicobacter Pylori and Epsilonproteobacteria." *Genome Biology and Evolution* 7(9):evv177.

Danielli, Alberto and Vincenzo Scarlato. 2010. "Regulatory Circuits in Helicobacter Pylori: Network Motifs and Regulators Involved in Metal-Dependent Responses." *FEMS Microbiology Reviews* 34(5):738–52.

Delsuc, Frédéric, Henner Brinkmann, and Hervé Philippe. 2005. "Phylogenomics and the Reconstruction of the Tree of Life." *Nature Reviews Genetics* 6(5):361–75.

Didelot, Xavier et al. 2013. "Genomic Evolution and Transmission of Helicobacter Pylori in Two South African Families." *Proceedings of the National Academy of Sciences of the United States of America* 110(34):13880–85.

Dunn, Bruce E., Hartley Cohen, and Martin J. Blaser. 1997. "Helicobacter Pylori." *Clinical Microbiology Reviews* 10(4):720–41.

Dunn, Bruce E. and Suhas H. Phadnis. 2000. "Structure , Function and Localization of Helicobacter Pyloni Urease." *Yale Journal OF Biology and Medicine* 71(1998):63–73.

Dus, Irena et al. 2013. "Role of PCR in Helicobacter Pylori Diagnostics and

Research--New Approaches for Study of Coccoid and Spiral Forms of the Bacteria." *Postepy higieny i medycyny doswiadczalnej (Online)* 67:261–68.

Eddy, S. R. 1996. "Hidden Markov Models." *Current opinion in structural biology* 6(3):361–65.

Eppinger, Mark et al. 2006. "Who Ate Whom? Adaptive Helicobacter Genomic Changes That Accompanied a Host Jump from Early Humans to Large Felines." *PLoS genetics* 2(7):e120.

Finn, Robert D. et al. 2014. "Pfam: The Protein Families Database." *Nucleic acids research* 42(Database issue):D222–30.

Foster, John W. 2004. "Escherichia Coli Acid Resistance: Tales of an Amateur Acidophile." *Nature Reviews Microbiology* 2(11):898–907.

Fulton, Debra L. et al. 2006. "Improving the Specificity of High-Throughput Ortholog Prediction." *BMC bioinformatics* 16:1–16.

Goodwin CS, Amstrong JA, Chilvers T, et al. 1989. "Transfer of Campylobacter Pylori and Campylobacter Mustelae to Helicobacter Gen.nov.as Helicobacter Pylori Comb.nov. and Helicobacter Mustelae Comb.nov., Respectively." *Int J Syst Bacteriol* 39:397–405.

Gressmann, Helga et al. 2005. "Gain and Loss of Multiple Genes during the Evolution of Helicobacter Pylori." *PLoS genetics* 1(4):e43.

Gupta, Radhey S. 2006. "Molecular Signatures (unique Proteins and

Conserved Indels) That Are Specific for the Epsilon Proteobacteria (Campylobacterales).” *BMC genomics* 7:167.

Hamada, Michiaki, Hisanori Kiryu, Kengo Sato, Toutai Mituyama, and Kiyoshi Asai. 2009. “Prediction of RNA Secondary Structure Using Generalized Centroid Estimators.” *Bioinformatics (Oxford, England)* 25(4):465–73.

Hazell, Stuart L., Andrew G. Harris, and Mark A. Trend. 2001. “Evasion of the Toxic Effects of Oxygen - Helicobacter Pylori - NCBI Bookshelf.pdf.” in *Helicobacter pylori: Physiology and Genetics.*, edited by G. L. Mendz and S. L. Hazell. Mobley HLT.

Heider, J., C. Baron, and a Böck. 1992. “Coding from a Distance: Dissection of the mRNA Determinants Required for the Incorporation of Selenocysteine into Protein.” *The EMBO journal* 11(1):3759–66.

Heintschel von Heinegg, E., H. P. Nalik, and E. N. Schmid. 1993. “Characterisation of a Helicobacter Pylori Phage (HP1).” *Journal of Medical Microbiology* 38(4):245–49.

Hirsh, AE and HB Fraser. 2001. “Protein Dispensability and Rate of Evolution.” *Nature* 411(6841):1046–49.

Itoh, Yuzuru et al. 2013. “Decameric SelA*tRNA(Sec) Ring Structure Reveals Mechanism of Bacterial Selenocysteine Formation.” *Science (New York, N.Y.)* 340(6128):75–78.

Itoh, Yuzuru, Markus J. Brocker, Shun-ichi Sekine, Dieter Soll, and Shigeyuki Yokoyama. 2014. "Dimer-Dimer Interaction of the Bacterial Selenocysteine Synthase SelA Promotes Functional Active-Site Formation and Catalytic Specificity." *Journal of molecular biology* 426(8):1723–35.

Iwatani, Shun et al. 2014. "Identification of the Genes That Contribute to Lactate Utilization in Helicobacter Pylori." *PloS one* 9(7):e103506.

Jombart, Thibaut, François Balloux, and Stéphane Dray. 2010. "Adephylo: New Tools for Investigating the Phylogenetic Signal in Biological Traits." *Bioinformatics (Oxford, England)* 26(15):1907–9.

Jordan, I. King, Igor B. Rogozin, Yuri I. Wolf, and Eugene V Koonin. 2002. "Essential Genes Are More Evolutionarily Conserved than Are Nonessential Genes in Bacteria." *Genome research* 12(6):962–68.

Kalali, Behnam, Raquel Mejías-Luque, Anahita Javaheri, and Markus Gerhard. 2014. "H. Pylori Virulence Factors: Influence on Immune System and Pathology." *Mediators of inflammation* 2014:426309.

Kandulski, a., M. Selgrad, and P. Malfertheiner. 2008. "Helicobacter Pylori Infection: A Clinical Overview." *Digestive and Liver Disease* 40(8):619–26.

Kersulyte, Dangeruta et al. 2000. "Differences in Genotypes of Helicobacter Pylori from Different Human Populations." *Journal of Bacteriology* 182(11):3210–18.

Kettler, Gregory C. et al. 2007. “Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus.” *PLoS Genetics* 3(12):e231.

Kraft, Christian and Sebastian Suerbaum. 2005. “Mutation and Recombination in Helicobacter Pylori: Mechanisms and Role in Generating Strain Diversity.” *International journal of medical microbiology : IJMM* 295(5):299–305.

Kryukov, Gregory V et al. 2003. “Characterization of Mammalian Selenoproteomes.” *Science (New York, N.Y.)* 300(5624):1439–43.

Kryukov, Gregory V and Vadim N. Gladyshev. 2004. “The Prokaryotic Selenoproteome.” *EMBO reports* 5(5):538–43.

Kuzniar, Arnold, Roeland C. H. J. van Ham, Sándor Pongor, and Jack A. M. Leunissen. 2008. “The Quest for Orthologs: Finding the Corresponding Gene across Genomes.” *Trends in Genetics* 24(11):539–51.

Labrie, Simon J., Julie E. Samson, and Sylvain Moineau. 2010. “Bacteriophage Resistance Mechanisms.” *Nature Reviews Microbiology* 8(5):317–27.

Larkin, M. A. et al. 2007. “Clustal W and Clustal X Version 2.0.” *Bioinformatics (Oxford, England)* 23(21):2947–48.

Lowe, T. M. and S. R. Eddy. 1997. “tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence.” *Nucleic*

acids research 25(5):955–64.

Lu, Wei et al. 2014. “Comparative Analysis of the Full Genome of Helicobacter Pylori Isolate sahul64 Identifies Genes of High Divergence.” *Journal of Bacteriology* 196(5):1073–83.

Luo, C. H., P. Y. Chiou, C. Y. Yang, and N. T. Lin. 2012. “Genome, Integration, and Transduction of a Novel Temperate Phage of Helicobacter Pylori.” *Journal of Virology* 86(16):8781–92.

Maier, R. J. et al. 1996. “Hydrogen Uptake Hydrogenase in Helicobacter Pylori.” *FEMS microbiology letters* 141(1):71–76.

Maier, R. J. 2005. “Use of Molecular Hydrogen as an Energy Substrate by Human Pathogenic Bacteria.” *Biochemical Society transactions* 33(Pt 1):83–85.

Marri, Pradeep Reddy, Weilong Hao, and G. Brian Golding. 2006. “Gene Gain and Gene Loss in Streptococcus: Is It Driven by Habitat?” *Molecular biology and evolution* 23(12):2379–91.

Miele, Vincent, Simon Penel, and Laurent Duret. 2011. “Ultra-Fast Sequence Clustering from Similarity Networks with SiLiX.” *BMC Bioinformatics* 12(1):116.

Mobley, H. L., M. J. Cortesia, L. E. Rosenthal, and B. D. Jones. 1988. “Characterization of Urease from Campylobacter Pylori.” *Journal of clinical microbiology* 26(5):831–36.

Mobley, H. L., R. M. Garner, and P. Bauerfeind. 1995. “Helicobacter Pylori Nickel-Transport Gene nixA: Synthesis of Catalytically Active Urease in Escherichia Coli Independent of Growth Conditions.” *Molecular microbiology* 16(1):97–109.

Moreno-Hagelsieb, Gabriel, Zilin Wang, Stephanie Walsh, and Aisha ElSherbiny. 2013. “Phylogenomic Clustering for Selecting Non-Redundant Genomes for Comparative Genomics.” *Bioinformatics (Oxford, England)* 29(7):947–49.

Nawrocki, Eric P. and Sean R. Eddy. 2013. “Infernal 1.1: 100-Fold Faster RNA Homology Searches.” *Bioinformatics (Oxford, England)* 29(22):2933–35.

Noto, Jennifer M. and Richard M. Peek. 2012. “The Helicobacter Pylori Cag Pathogenicity Island.” *methods Mol Biol* 921(22):41–50.

Passerini, A., M. Lippi, and P. Frasconi. 2011. “MetalDetector v2.0: Predicting the Geometry of Metal Binding Sites from Protein Sequence.” *Nucleic Acids Research* 39(suppl):W288–92.

Pelliciari, Simone, Andrea Vannini, Davide Roncarati, and Alberto Danielli. 2015. “The Allosteric Behavior of Fur Mediates Oxidative Stress Signal Transduction in Helicobacter Pylori.” *Frontiers in Microbiology* 6(August):1–10.

Pettinati, I., J. Brem, M. A. McDonough, and C. J. Schofield. 2015. “Crystal Structure of Human Persulfide Dioxygenase: Structural Basis of Ethylmalonic Encephalopathy.” *Human Molecular Genetics*

24(9):2458–69.

Popescu, Andrei-Alin, Katharina T. Huber, and Emmanuel Paradis. 2012. "Ape 3.0: New Tools for Distance-Based Phylogenetics and Evolutionary Analysis in R." *Bioinformatics (Oxford, England)* 28(11):1536–37.

Rabby, Atai et al. 2015. "Identification of the Positively Selected Genes Governing Host-Pathogen Arm Race in Vibrio Sp. through Comparative Genomics Approach." *Biojournal of Science and Technology* 2.

Ramarao, N., S. D. Gray-Owen, and T. F. Meyer. 2000. "Helicobacter Pylori Induces but Survives the Extracellular Release of Oxygen Radicals from Professional Phagocytes Using Its Catalase Activity." *Molecular microbiology* 38(1):103–13.

Ramulu, H. G. et al. 2014. "Ribosomal Proteins: Toward a next Generation Standard for Prokaryotic Systematics?" *Molecular Phylogenetics and Evolution* 75(1):103–17.

Rayman, Margaret P. 2012. "Selenium and Human Health." *The Lancet* 379(9822):1256–68.

de Reuse, Hilde, Daniel Vinella, and Christine Cavazza. 2013. "Common Themes and Unique Proteins for the Uptake and Trafficking of Nickel, a Metal Essential for the Virulence of Helicobacter Pylori." *Frontiers in Cellular and Infection Microbiology* 3(December):1–6.

Robert, Xavier and Patrice Gouet. 2014. “Deciphering Key Features in Protein Structures with the New ENDscript Server.” *Nucleic acids research* 42(Web Server issue):W320–24.

Romero, Héctor, Yan Zhang, Vadim N. Gladyshev, and Gustavo Salinas. 2005. “Evolution of Selenium Utilization Traits.” *Genome biology* 6(8):R66.

Sattler, Steven A. et al. 2015. “Characterizations of Two Bacterial Persulfide Dioxygenases of the Metallo-β-Lactamase Superfamily.” *Journal of Biological Chemistry* 290(31):18914–23.

Scarlato, V., I. Delany, G. Spohn, and D. Beier. 2001. “Regulation of Transcription in Helicobacter Pylori: Simple Systems or Complex Circuits?” *International journal of medical microbiology : IJMM* 291(2):107–17.

Shaik, Md Munan, Laura Cendron, Riccardo Percudani, and Giuseppe Zanotti. 2011. “The Structure of Helicobacter Pylori HP0310 Reveals an Atypical Peptidoglycan Deacetylase.” *PLoS ONE* 6(4):1–8.

Shaw, Frances L. et al. 2012. “Selenium-Dependent Biogenesis of Formate Dehydrogenase in Campylobacter Jejuni Is Controlled by the fdhTU Accessory Genes.” *Journal of bacteriology* 194(15):3814–23.

Shi, Wuxian et al. 2011. “Characterization of Metalloproteins by High-Throughput X-Ray Absorption Spectroscopy.” *Genome Research* 21(6):898–907.

Smith, James L. 2003. “The Role of Gastric Acid in Preventing Foodborne Disease and How Bacteria Overcome Acid Conditions.”

Stamatakis, Alexandros. 2006. “RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models.” *Bioinformatics (Oxford, England)* 22(21):2688–90.

Suerbaum, Sebastian and Christine Josenhans. 2007. “Helicobacter Pylori Evolution and Phenotypic Diversification in a Changing Host.” *Nature reviews. Microbiology* 5(6):441–52.

Tatusov, Roman L. et al. 2003. “The COG Database: An Updated Version Includes Eukaryotes.” *BMC bioinformatics* 4:41.

Tomb, J. F. et al. 1997. “The Complete Genome Sequence of the Gastric Pathogen Helicobacter Pylori.” *Nature* 388(6642):539–47.

Tomii, Kentaro and Minoru Kanehisa. 1998. “A Comparative Analysis of ABC Transporters in Complete Microbial Genomes.” *Genome Research* 8(10):1048–59.

Toyonaga, A. et al. 2000. “Epidemiological Study on Food Intake and Helicobacter Pylori Infection.” *The Kurume medical journal* 47(3714):25–30. Retrieved December 10, 2015 (http://www.ncbi.nlm.nih.gov/pubmed/10812886).

Uchiyama, J. et al. 2012. “Complete Genome Sequences of Two Helicobacter Pylori Bacteriophages Isolated from Japanese Patients.” *Journal of Virology* 86(20):11400–401.

Uchiyama, J. et al. 2013. “Characterization of Helicobacter Pylori Bacteriophage KHP30.” *Applied and Environmental Microbiology* 79(10):3176–84.

Ustundag, Y., S. Boyacioglu, A. Haberal, B. Demirhan, and B. Bilezikci. 2001. “Plasma and Gastric Tissue Selenium Levels in Patients with Helicobacter Pylori Infection.” *Journal of clinical gastroenterology* 32(5):405–8.

Wang, Ge, Praveen Alamuri, and Robert J. Maier. 2006. “The Diverse Antioxidant Systems of Helicobacter Pylori.” *Molecular Microbiology* 61(4):847–60.

Wang, Z. and M. Wu. 2013. “A Phylum-Level Bacterial Phylogenetic Marker Database.” *Molecular Biology and Evolution* 30(6):1258–62.

Warren, J. Robin. 2006. “Helicobacter: The Ease and Difficulty of a New Discovery (Nobel Lecture).” *ChemMedChem* 672–85.

Weinberg, Zasha and Ronald R. Breaker. 2011. “R2R--Software to Speed the Depiction of Aesthetic Consensus RNA Secondary Structures.” *BMC bioinformatics* 12:3.

Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. “Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya.” *Proceedings of the National Academy of Sciences of the United States of America* 87(12):4576–79.

Wolfe, M. D. et al. 2004. “Functional Diversity of the Rhodanese

Homology Domain: THE ESCHERICHIA COLI ybbB GENE ENCODES A SELENOPHOSPHATE-DEPENDENT tRNA 2-SELENOURIDINE SYNTHASE." *Journal of Biological Chemistry* 279(3):1801–9.

Yang, Ziheng. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular biology and evolution* 24(8):1586–91.

Zhang, Yan and Vadim N. Gladyshev. 2005. "An Algorithm for Identification of Bacterial Selenocysteine Insertion Sequence Elements and Selenoprotein Genes." *Bioinformatics (Oxford, England)* 21(11):2580–89.

Zhang, Yan and Vadim N. Gladyshev. 2010. "dbTEU: A Protein Database of Trace Element Utilization." *Bioinformatics (Oxford, England)* 26(5):700–702.

Zhang, Yan, Hector Romero, Gustavo Salinas, and Vadim N. Gladyshev. 2006. "Dynamic Evolution of Selenocysteine Utilization in Bacteria: A Balance between Selenoprotein Loss and Evolution of Selenocysteine from Redox Active Cysteine Residues." *Genome biology* 7(10):R94.

Zhu et al. 2011. "Characterization and Inference of Gene Gain/Loss Along Burkholderia Evolutionary History." *Evolutionary Bioinformatics* 191.

Zinoni, F., A. Birkmann, W. Leinfelder, and A. Böck. 1987. "Cotranslational Insertion of Selenocysteine into Formate Dehydrogenase from Escherichia Coli Directed by a UGA Codon."

Proceedings of the National Academy of Sciences of the United States of America 84(10):3156–60.

Zinoni, F., J. Heider, and A. Bock. 1990. “Features of the Formate Dehydrogenase mRNA Necessary for Decoding of the UGA Codon as Selenocysteine.” *Proceedings of the National Academy of Sciences of the United States of America* 87(12):4660–64.

APPENDICE 1 – TABELLE SUPPLEMENTARI

Tabella Supplementare S3.1 *Risultati di MetalDetector filtrati in base alla conservazione relative. Per ogni proteina vengono indicati tutti gli aminoacidi predetti facenti parte di un sito di coordinazione. La colonna sito indica a che sito appartiene un determinato residuo; numeri uguali nella stessa proteina indicano residui facenti parte dello stesso sito. Cons. assoluta indica la percentuale di conservazione del residuo negli allineamenti di BLAST con gli omologhi di quella sequenza. Cons. media indica la conservazione media della matrice PSSM calcolata per quella sequenza. Cons. relativa indica la conservazione relativa del residuo (Cons. assoluta/Cons. media).*

Accession Number	Posizione	Residuo	Sito	Cons. Assoluta	Cons. media	Cons. relativa
NP_207460	17	C	2	100	36	2.78
NP_207460	20	C	2	98	36	2.73
NP_207460	23	C	2	99	36	2.76
NP_207460	27	C	2	99	36	2.76
NP_207460	68	C	2	99	36	2.76
NP_207460	71	C	2	100	36	2.78
NP_207460	74	C	3	100	36	2.78
NP_207460	78	C	3	100	36	2.78
NP_207460	189	C	4	90	36	2.51
NP_207460	220	C	1	95	36	2.64
NP_207460	221	C	4	93	36	2.59
NP_207460	262	C	4	86	36	2.39
NP_207460	325	H	4	80	36	2.23
NP_207460	328	C	1	100	36	2.78
NP_207460	329	H	4	96	36	2.67
NP_207460	357	C	1	95	36	2.64
NP_207460	358	C	1	90	36	2.51
NP_207460	396	C	1	60	36	1.67
NP_207460	399	C	1	96	36	2.67
NP_208299	88	C	2	100	41	2.44
NP_208299	92	C	2	100	41	2.44
NP_208299	197	C	2	100	41	2.44
NP_208299	201	C	2	99	41	2.42
NP_208299	249	C	4	100	41	2.44
NP_208299	252	C	4	100	41	2.44

NP_208299	255	C	2	100	41	2.44
NP_208299	259	C	4	100	41	2.44
NP_208299	273	C	1	100	41	2.44
NP_208299	276	C	1	100	41	2.44
NP_208299	279	C	1	100	41	2.44
NP_208299	283	C	3	100	41	2.44
NP_206939	9	C	4	100	52	1.92
NP_206939	41	C	4	100	52	1.92
NP_206939	42	C	4	100	52	1.92
NP_206939	80	C	2	100	52	1.92
NP_206939	135	H	2	99	52	1.90
NP_206939	138	C	2	100	52	1.92
NP_206939	171	C	3	100	52	1.92
NP_206939	172	C	1	100	52	1.92
NP_206939	212	C	1	100	52	1.92
NP_206939	231	H	1	100	52	1.92
NP_206990	60	C	2	97	45	2.17
NP_206990	65	C	2	97	45	2.17
NP_206990	68	C	3	76	45	1.70
NP_206990	80	C	2	98	45	2.19
NP_206990	154	C	4	96	45	2.15
NP_206990	157	C	3	100	45	2.24
NP_206990	160	C	4	100	45	2.24
NP_206990	164	C	4	96	45	2.15
NP_206990	211	C	1	100	45	2.24
NP_206990	217	C	1	100	45	2.24
NP_206990	218	H	1	24	45	0.54
NP_206990	221	C	3	100	45	2.24
NP_206916	32	H	2	100	57	1.76
NP_206916	616	C	2	98	57	1.73
NP_206916	619	C	2	99	57	1.74
NP_206916	636	C	2	100	57	1.76
NP_206916	642	C	1	100	57	1.76

NP_206916	663	C	3	99	57	1.74
NP_206916	666	C	3	98	57	1.73
NP_206916	683	C	1	98	57	1.73
NP_206916	689	C	2	99	57	1.74
NP_206916	702	C	1	96	57	1.69
NP_206916	705	C	1	93	57	1.64
NP_206916	722	C	2	73	57	1.29
NP_206916	725	C	1	96	57	1.69
NP_207900	58	C	2	100	47	2.11
NP_207900	61	C	2	100	47	2.11
NP_207900	64	C	1	100	47	2.11
NP_207900	68	C	1	100	47	2.11
NP_207900	88	C	2	100	47	2.11
NP_207900	91	C	2	100	47	2.11
NP_207900	94	C	1	100	47	2.11
NP_207900	98	C	1	100	47	2.11
NP_207383	19	C	2	100	54	1.86
NP_207383	22	C	2	100	54	1.86
NP_207383	25	C	2	100	54	1.86
NP_207383	29	C	2	100	54	1.86
NP_207383	56	C	1	100	54	1.86
NP_207383	59	C	1	100	54	1.86
NP_207383	62	C	1	100	54	1.86
NP_207383	66	C	2	100	54	1.86
NP_208058	45	C	2	100	41	2.46
NP_208058	48	C	2	100	41	2.46
NP_208058	96	C	2	99	41	2.43
NP_208058	99	C	2	100	41	2.46
NP_208058	105	C	1	100	41	2.46
NP_208058	143	C	1	100	41	2.46
NP_208058	146	C	2	100	41	2.46
NP_208058	149	C	2	100	41	2.46
NP_208058	239	C	1	85	41	2.09

NP_208058	242	C	1	85	41	2.09
NP_208058	246	C	1	85	41	2.09
NP_206849	124	C	2	100	48	2.07
NP_206849	127	C	2	100	48	2.07
NP_206849	146	C	2	100	48	2.07
NP_206849	149	C	2	100	48	2.07
NP_206849	174	C	1	100	48	2.07
NP_206849	177	C	1	100	48	2.07
NP_206849	196	C	1	100	48	2.07
NP_206849	199	C	1	99	48	2.05
NP_208060	82	C	2	100	46	2.16
NP_208060	85	C	2	100	46	2.16
NP_208060	88	C	2	100	46	2.16
NP_208060	92	C	3	100	46	2.16
NP_208060	121	C	1	100	46	2.16
NP_208060	124	C	4	100	46	2.16
NP_208060	127	C	1	100	46	2.16
NP_208060	131	C	1	100	46	2.16
NP_207185	336	C	2	100	49	2.04
NP_207185	339	C	2	100	49	2.04
NP_207185	345	C	2	100	49	2.04
NP_207185	348	C	2	100	49	2.04
NP_207185	363	C	2	100	49	2.04
NP_207185	364	H	2	93	49	1.90
NP_207185	366	C	1	100	49	2.04
NP_207185	376	C	1	100	49	2.04
NP_207185	379	C	1	100	49	2.04
NP_207075	9	C	1	100	59	1.71
NP_207075	12	C	1	100	59	1.71
NP_207075	15	C	1	100	59	1.71
NP_207075	19	C	2	100	59	1.71
NP_207075	38	C	2	100	59	1.71

NP_207075	41	C	2	100	59	1.71
NP_207075	51	C	1	100	59	1.71
NP_207075	55	C	1	100	59	1.71
NP_207557	337	C	1	94	53	1.77
NP_207557	340	C	1	100	53	1.88
NP_207557	343	C	1	100	53	1.88
NP_207557	344	C	2	100	53	1.88
NP_207557	363	C	2	100	53	1.88
NP_207557	373	C	1	100	53	1.88
NP_207557	383	C	1	97	53	1.83
NP_207557	409	C	1	100	53	1.88
NP_208294	3	C	2	100	44	2.25
NP_208294	5	H	2	100	44	2.25
NP_208294	6	C	2	100	44	2.25
NP_208294	28	C	1	92	44	2.07
NP_208294	29	C	1	100	44	2.25
NP_208294	32	C	1	100	44	2.25
NP_208294	93	C	1	100	44	2.25
NP_208294	96	C	1	100	44	2.25
NP_207528	16	C	2	100	50	2.00
NP_207528	50	C	2	100	50	2.00
NP_207528	82	C	3	100	50	2.00
NP_207528	151	C	1	100	50	2.00
NP_207528	155	C	1	100	50	2.00
NP_207528	158	C	1	100	50	2.00
NP_207528	260	H	4	97	50	1.94
NP_206900	4	H	2	96	43	2.23
NP_206900	6	C	2	100	43	2.32
NP_206900	7	C	2	100	43	2.32
NP_206900	87	C	2	100	43	2.32
NP_206900	90	C	1	100	43	2.32
NP_206900	174	C	1	100	43	2.32

NP_206900	176	C	1	100	43	2.32
NP_207083	11	C	1	100	48	2.10
NP_207083	45	C	1	100	48	2.10
NP_207083	74	C	1	95	48	2.00
NP_207083	144	C	1	100	48	2.10
NP_207083	148	C	1	100	48	2.10
NP_207083	151	C	2	100	48	2.10
NP_207198	12	C	1	100	46	2.16
NP_207198	36	H	2	99	46	2.14
NP_207198	70	H	1	100	46	2.16
NP_207198	92	C	1	100	46	2.16
NP_207198	120	H	1	100	46	2.16
NP_207198	186	C	1	100	46	2.16
NP_208148	86	C	1	98	48	2.05
NP_208148	198	C	1	100	48	2.09
NP_208148	201	H	1	100	48	2.09
NP_208148	224	H	1	100	48	2.09
NP_208148	227	C	1	95	48	1.99
NP_208148	288	C	2	100	48	2.09
NP_208252	88	C	2	99	46	2.16
NP_208252	91	C	2	100	46	2.19
NP_208252	92	H	2	100	46	2.19
NP_208252	234	C	1	100	46	2.19
NP_208252	237	C	1	100	46	2.19
NP_208252	238	H	1	100	46	2.19
NP_208331	98	C	1	100	45	2.23
NP_208331	100	H	1	100	45	2.23
NP_208331	103	C	1	100	45	2.23
NP_208331	116	C	1	100	45	2.23
NP_208331	118	C	1	100	45	2.23
NP_208331	119	H	1	100	45	2.23

NP_206946	48	H	2	88	44	2.01
NP_206946	129	C	2	100	44	2.29
NP_206946	132	C	2	100	44	2.29
NP_206946	133	H	3	100	44	2.29
NP_206946	219	C	1	100	44	2.29
NP_206946	222	C	1	100	44	2.29
NP_206946	223	H	1	100	44	2.29
NP_208014	548	C	2	72	38	1.92
NP_208014	551	C	2	59	38	1.57
NP_208014	554	C	2	88	38	2.34
NP_208014	558	C	2	89	38	2.37
NP_208014	610	C	2	99	38	2.64
NP_208014	613	C	2	65	38	1.73
NP_208014	616	C	1	100	38	2.66
NP_208014	620	C	1	100	38	2.66
NP_208014	883	C	1	99	38	2.64
NP_208014	884	C	1	95	38	2.53
NP_208014	924	C	1	98	38	2.61
NP_207067	46	C	1	100	54	1.87
NP_207067	78	C	1	100	54	1.87
NP_207067	148	C	1	100	54	1.87
NP_207067	152	C	1	100	54	1.87
NP_207067	155	C	2	100	54	1.87
NP_208018	13	H	1	100	46	2.19
NP_208018	17	C	1	100	46	2.19
NP_208018	21	C	1	100	46	2.19
NP_208018	24	C	2	100	46	2.19
NP_208018	139	H	2	96	46	2.11
NP_207425	86	C	2	96	44	2.16
NP_207425	89	C	2	67	44	1.51
NP_207425	185	C	2	100	44	2.25

NP_207425	218	C	2	83	44	1.87
NP_207425	256	H	2	86	44	1.93
NP_207425	259	C	1	98	44	2.20
NP_207425	318	C	1	100	44	2.25
NP_207425	321	C	1	100	44	2.25
NP_206938	314	C	2	100	53	1.88
NP_206938	317	C	2	100	53	1.88
NP_206938	320	C	2	100	53	1.88
NP_206938	367	C	1	94	53	1.76
NP_206938	370	C	1	100	53	1.88
NP_206938	374	C	1	100	53	1.88
NP_206992	44	H	2	100	47	2.14
NP_206992	93	H	1	97	47	2.07
NP_206992	120	H	1	95	47	2.03
NP_206992	143	H	1	100	47	2.14
NP_206992	182	H	1	100	47	2.14
NP_207606	52	H	1	100	44	2.29
NP_207606	54	H	2	100	44	2.29
NP_207606	57	H	1	98	44	2.25
NP_207606	121	H	2	55	44	1.26
NP_207606	125	H	1	100	44	2.29
NP_207606	184	H	2	100	44	2.29
NP_207202	36	H	1	100	55	1.81
NP_207202	93	H	1	97	55	1.76
NP_207202	95	H	1	100	55	1.81
NP_207202	97	H	1	100	55	1.81
NP_207202	104	H	1	100	55	1.81
NP_207679	36	H	2	100	57	1.74
NP_207679	204	H	1	100	57	1.74
NP_207679	207	C	1	100	57	1.74
NP_207679	227	H	1	100	57	1.74

NP_207679	237	H	1	100	57	1.74
NP_208364	6	H	2	99	51	1.93
NP_208364	8	H	2	99	51	1.93
NP_208364	63	H	3	96	51	1.87
NP_208364	131	H	1	100	51	1.95
NP_208364	154	H	1	100	51	1.95
NP_208364	155	C	1	89	51	1.74
NP_207385	19	C	1	100	51	1.96
NP_207385	22	C	1	100	51	1.96
NP_207385	67	H	2	89	51	1.75
NP_207385	70	H	2	100	51	1.96
NP_207385	105	H	3	100	51	1.96
NP_207385	108	H	1	89	51	1.75
NP_207385	202	C	4	100	51	1.96
NP_207691	40	C	2	100	55	1.81
NP_207691	43	H	2	87	55	1.58
NP_207691	68	C	2	100	55	1.81
NP_207691	71	C	2	100	55	1.81
NP_207691	322	C	1	92	55	1.67
NP_207691	324	C	1	83	55	1.50
NP_207691	337	C	1	100	55	1.81
NP_207691	344	C	1	100	55	1.81
NP_207426	62	C	1	97	41	2.38
NP_207426	65	C	1	99	41	2.43
NP_207426	69	H	1	95	41	2.33
NP_207426	551	C	1	97	41	2.38
NP_207426	554	C	1	100	41	2.45
NP_206968	167	H	2	99	50	1.96
NP_206968	171	C	2	100	50	1.98
NP_206968	178	C	1	100	50	1.98
NP_206968	194	C	1	100	50	1.98

NP_206968	198	C	1	100	50	1.98
NP_208264	3	C	2	100	52	1.92
NP_208264	6	C	1	100	52	1.92
NP_208264	15	C	1	100	52	1.92
NP_208264	18	C	1	100	52	1.92
NP_208264	100	H	2	96	52	1.84
NP_207561	21	C	1	100	48	2.07
NP_207561	25	C	1	100	48	2.07
NP_207561	28	C	1	100	48	2.07
NP_207561	249	C	1	100	48	2.07
NP_207561	252	C	1	100	48	2.07
NP_207717	61	C	1	100	52	1.92
NP_207717	64	C	1	100	52	1.92
NP_207717	73	C	1	100	52	1.92
NP_207717	76	C	1	100	52	1.92
NP_207717	87	C	1	95	52	1.82
NP_207022	44	C	2	100	57	1.75
NP_207022	300	H	1	100	57	1.75
NP_207022	303	H	1	100	57	1.75
NP_207022	318	C	1	98	57	1.72
NP_207022	359	H	1	100	57	1.75
NP_208221	206	H	2	100	54	1.84
NP_208221	208	H	2	100	54	1.84
NP_208221	450	H	1	93	54	1.72
NP_208221	498	H	1	100	54	1.84
NP_208221	502	H	1	100	54	1.84
NP_208221	524	H	1	98	54	1.81
NP_207072	13	C	1	100	47	2.14
NP_207072	16	C	1	98	47	2.10
NP_207072	20	C	2	91	47	1.95

NP_207072	21	C	2	100	47	2.14
NP_207072	75	C	1	97	47	2.08
NP_207655	23	H	2	100	40	2.51
NP_207655	24	C	1	100	40	2.51
NP_207655	28	C	1	98	40	2.46
NP_207655	144	C	2	98	40	2.46
NP_207454	296	C	1	100	44	2.26
NP_207454	299	C	1	100	44	2.26
NP_207454	310	C	1	100	44	2.26
NP_207454	313	C	1	100	44	2.26
NP_208330	94	H	1	97	55	1.76
NP_208330	108	H	2	99	55	1.79
NP_208330	195	H	1	100	55	1.81
NP_208330	210	H	2	97	55	1.76
NP_207863	12	C	1	100	54	1.84
NP_207863	15	C	1	100	54	1.84
NP_207863	391	C	1	100	54	1.84
NP_207863	393	C	1	100	54	1.84
NP_207535	20	C	1	85	44	1.92
NP_207535	23	C	1	88	44	1.98
NP_207535	57	H	1	97	44	2.19
NP_207535	116	H	2	100	44	2.25
NP_207535	118	H	1	96	44	2.16
NP_207535	120	H	1	95	44	2.14
NP_207663	74	C	1	100	46	2.16
NP_207663	77	C	1	100	46	2.16
NP_207663	91	C	1	100	46	2.16
NP_207663	94	C	1	98	46	2.12
NP_206905	53	H	1	100	54	1.85

NP_206905	57	H	1	100	54	1.85
NP_206905	78	C	1	100	54	1.85
NP_206905	120	C	1	100	54	1.85
NP_206905	126	H	1	92	54	1.70
NP_206814	37	C	1	98	46	2.14
NP_206814	40	H	1	98	46	2.14
NP_206814	58	C	1	99	46	2.17
NP_206814	61	C	1	98	46	2.14
NP_206904	110	H	1	97	42	2.31
NP_206904	210	H	1	97	42	2.31
NP_206904	239	H	1	98	42	2.34
NP_206904	241	H	1	98	42	2.34
NP_207827	197	H	1	100	46	2.18
NP_207827	297	H	1	98	46	2.14
NP_207827	301	H	1	100	46	2.18
NP_207827	308	H	1	98	46	2.14
NP_206944	61	C	1	100	48	2.07
NP_206944	64	C	1	100	48	2.07
NP_206944	65	H	1	100	48	2.07
NP_206944	117	H	1	86	48	1.78
NP_206944	120	H	1	100	48	2.07
NP_207834	154	H	1	99	40	2.50
NP_207834	216	H	1	100	40	2.52
NP_207834	306	H	1	100	40	2.52
NP_207834	308	H	1	99	40	2.50
NP_207459	58	H	1	100	58	1.73
NP_207459	62	C	1	100	58	1.73
NP_207459	66	C	1	100	58	1.73
NP_207459	69	C	1	100	58	1.73

NP_207880	570	H	1	95	38	2.47
NP_207880	767	C	1	98	38	2.55
NP_207880	770	C	1	100	38	2.60
NP_207880	776	C	2	100	38	2.60
NP_207511	62	C	1	96	49	1.97
NP_207511	71	C	1	96	49	1.97
NP_207511	74	C	1	96	49	1.97
NP_207511	77	C	1	100	49	2.05
NP_207750	199	C	1	96	39	2.47
NP_207750	202	C	1	96	39	2.47
NP_207750	223	C	1	96	39	2.47
NP_207750	226	C	1	96	39	2.47
NP_207380	194	C	1	100	56	1.79
NP_207380	201	C	1	99	56	1.77
NP_207380	204	C	1	100	56	1.79
NP_207380	210	C	1	100	56	1.79
NP_206942	190	C	1	100	47	2.14
NP_206942	196	C	1	100	47	2.14
NP_206942	199	C	1	100	47	2.14
NP_206942	205	C	1	98	47	2.10
NP_206942	273	H	2	82	47	1.76
NP_208219	116	C	1	100	51	1.95
NP_208219	123	C	1	100	51	1.95
NP_208219	127	C	1	100	51	1.95
NP_208219	130	C	1	100	51	1.95
NP_207780	321	C	1	100	52	1.94
NP_207780	324	C	1	77	52	1.49
NP_207780	340	C	1	96	52	1.86
NP_207780	343	C	2	96	52	1.86

NP_207951	88	H	1	66	49	1.34
NP_207951	100	H	1	100	49	2.02
NP_207951	104	H	1	100	49	2.02
NP_207951	110	H	1	100	49	2.02
NP_206917	33	C	1	100	39	2.54
NP_206917	37	C	1	100	39	2.54
NP_206917	40	C	1	100	39	2.54
NP_208040	497	H	1	100	45	2.23
NP_208040	507	H	1	100	45	2.23
NP_208040	521	H	1	100	45	2.23
NP_207190	47	H	1	100	46	2.19
NP_207190	66	H	1	91	46	1.99
NP_207190	381	H	1	100	46	2.19
NP_207190	389	H	1	100	46	2.19
NP_207614	169	C	1	98	47	2.10
NP_207614	177	C	1	98	47	2.10
NP_207614	181	C	1	98	47	2.10
NP_208310	14	H	1	100	48	2.07
NP_208310	100	H	1	100	48	2.07
NP_208310	224	H	2	100	48	2.07
NP_207034	42	C	1	100	57	1.77
NP_207034	45	C	1	100	57	1.77
NP_207034	46	H	1	100	57	1.77
NP_207369	16	H	1	95	56	1.70
NP_207369	70	C	1	95	56	1.70
NP_207369	103	H	1	96	56	1.72
NP_207369	138	H	1	91	56	1.63
NP_207743	142	C	1	100	57	1.76

NP_207743	145	C	1	100	57	1.76
NP_207743	163	H	1	96	57	1.69
NP_207743	166	C	1	100	57	1.76
NP_207065	65	H	1	100	36	2.75
NP_207065	67	H	1	99	36	2.72
NP_207065	207	H	2	100	36	2.75
NP_208213	127	C	1	94	56	1.68
NP_208213	895	C	1	100	56	1.79
NP_208213	898	C	1	100	56	1.79
NP_208213	910	C	1	100	56	1.79
NP_207010	79	H	1	100	54	1.84
NP_207010	201	H	1	100	54	1.84
NP_207010	355	H	2	100	54	1.84
NP_208374	126	H	1	97	49	2.00
NP_208374	156	H	1	97	49	2.00
NP_208374	195	H	1	100	49	2.06
NP_208019	29	C	1	100	52	1.93
NP_208019	32	C	1	100	52	1.93
NP_208019	33	H	1	100	52	1.93
NP_207481	340	C	1	100	44	2.28
NP_207481	369	C	1	100	44	2.28
NP_207481	599	C	2	99	44	2.26
NP_207469	309	H	1	100	41	2.42
NP_207469	313	H	1	100	41	2.42
NP_207469	335	H	1	100	41	2.42
NP_208171	103	H	2	100	45	2.21
NP_208171	161	H	1	99	45	2.18
NP_208171	170	H	2	100	45	2.21

NP_208171	173	H	1	94	45	2.07
NP_207306	373	H	1	100	50	1.99
NP_207306	380	H	2	100	50	1.99
NP_207306	417	H	1	99	50	1.97
NP_208240	32	C	1	100	51	1.96
NP_208240	38	C	1	100	51	1.96
NP_208240	65	C	1	100	51	1.96
NP_206929	119	C	2	89	53	1.68
NP_206929	121	C	1	67	53	1.26
NP_206929	126	H	1	90	53	1.69
NP_206929	128	H	2	100	53	1.88
NP_206929	138	H	2	74	53	1.39
NP_206929	139	H	1	100	53	1.88
NP_206929	141	H	1	100	53	1.88
NP_207307	147	H	1	100	52	1.91
NP_207307	205	H	2	100	52	1.91
NP_207307	220	H	1	100	52	1.91
NP_208324	321	C	1	100	52	1.94
NP_208324	324	C	1	77	52	1.49
NP_208324	340	C	1	96	52	1.86
NP_208324	343	C	2	96	52	1.86
NP_207786	297	H	1	95	35	2.71
NP_207786	301	H	1	95	35	2.71
NP_207786	323	H	1	100	35	2.85
NP_207786	334	H	1	74	35	2.11
NP_208073	84	C	1	100	55	1.81
NP_208073	106	H	1	98	55	1.78
NP_208073	133	H	1	99	55	1.79

NP_208064	229	H	2	99	52	1.89
NP_208064	310	H	1	100	52	1.91
NP_208064	336	H	2	96	52	1.84
NP_208170	364	C	1	97	56	1.74
NP_208170	727	H	1	100	56	1.80
NP_208170	729	H	1	100	56	1.80
NP_208197	17	C	1	100	49	2.03
NP_208197	21	C	1	100	49	2.03
NP_208197	24	C	1	100	49	2.03
NP_208197	134	H	1	86	49	1.74
NP_207393	118	H	1	100	44	2.25
NP_207393	137	H	1	99	44	2.23
NP_207393	193	H	1	100	44	2.25
NP_207586	96	C	2	100	56	1.79
NP_207586	138	H	1	100	56	1.79
NP_207586	142	H	1	100	56	1.79
NP_206857	1162	H	2	100	42	2.39
NP_206857	1177	C	1	73	42	1.75
NP_206857	1178	H	1	76	42	1.82
NP_206857	1181	C	1	100	42	2.39
NP_206857	1182	C	1	100	42	2.39
NP_207788	321	C	1	100	51	1.94
NP_207788	324	C	1	77	51	1.50
NP_207788	340	C	1	95	51	1.84
NP_207788	343	C	2	95	51	1.84
NP_208269	94	H	1	100	53	1.89
NP_208269	536	H	1	100	53	1.89
NP_208269	638	H	1	97	53	1.83

NP_207236	321	C	1	100	51	1.94
NP_207236	324	C	1	77	51	1.50
NP_207236	340	C	1	95	51	1.84
NP_207236	343	C	2	95	51	1.84
NP_207394	337	H	1	72	39	1.85
NP_207394	340	C	1	95	39	2.44
NP_207394	368	H	1	97	39	2.49
NP_207394	372	H	1	95	39	2.44
NP_206824	267	C	1	96	45	2.15
NP_206824	368	H	1	100	45	2.23
NP_206824	387	C	1	100	45	2.23
NP_207176	760	H	1	100	38	2.64
NP_207176	857	H	1	98	38	2.59
NP_207176	896	H	1	96	38	2.54
NP_207110	66	C	1	100	47	2.12
NP_207110	68	C	1	100	47	2.12
NP_207110	69	C	1	100	47	2.12
NP_208371	7	C	1	100	50	2.01
NP_208371	128	H	1	100	50	2.01
NP_208371	168	H	2	100	50	2.01
NP_207631	20	C	1	100	56	1.77
NP_207631	103	C	1	100	56	1.77
NP_207631	108	C	1	100	56	1.77
NP_207886	321	C	1	100	51	1.94
NP_207886	324	C	1	77	51	1.50
NP_207886	340	C	1	95	51	1.84
NP_207886	343	C	2	95	51	1.84
NP_208365	40	C	1	100	51	1.97

NP_208365	87	H	1	99	51	1.95
NP_208365	92	H	1	100	51	1.97
NP_207410	400	C	1	100	56	1.79
NP_207410	403	C	1	100	56	1.79
NP_207410	416	C	1	92	56	1.65
NP_207410	421	C	2	100	56	1.79
NP_207448	59	C	1	100	49	2.03
NP_207448	63	C	1	100	49	2.03
NP_207448	66	C	1	100	49	2.03
NP_207427	64	H	2	96	41	2.36
NP_207427	182	H	1	95	41	2.34
NP_207427	196	H	1	100	41	2.46
NP_207450	81	C	1	100	50	2.00
NP_207450	85	C	1	100	50	2.00
NP_207450	88	C	1	100	50	2.00
NP_207192	119	H	1	100	43	2.31
NP_207192	208	H	1	100	43	2.31
NP_207192	210	H	1	100	43	2.31
NP_207568	67	H	2	100	49	2.05
NP_207568	640	C	2	99	49	2.03
NP_207568	641	C	1	100	49	2.05
NP_207568	664	H	1	85	49	1.74
NP_207568	665	H	1	56	49	1.15
NP_207568	668	C	1	91	49	1.86
NP_207081	231	H	1	100	55	1.81
NP_207081	235	H	2	100	55	1.81
NP_207081	248	H	1	100	55	1.81
NP_207696	63	H	2	95	59	1.62

NP_207696	91	H	1	100	59	1.70
NP_207696	124	H	1	100	59	1.70
NP_207696	181	H	1	100	59	1.70
NP_207696	185	H	1	97	59	1.65
NP_207696	209	H	1	99	59	1.69
NP_207087	59	H	3	32	49	0.65
NP_207087	67	C	3	56	49	1.14
NP_207087	75	C	3	50	49	1.02
NP_207087	84	H	4	38	49	0.78
NP_207087	185	H	4	22	49	0.45
NP_207087	233	C	4	23	49	0.47
NP_207087	245	C	4	20	49	0.41
NP_207087	692	H	4	25	49	0.51
NP_207087	693	C	3	24	49	0.49
NP_207087	1158	H	3	30	49	0.61
NP_207087	1159	C	4	29	49	0.59
NP_207087	1385	H	1	26	49	0.53
NP_207087	1667	H	2	65	49	1.33
NP_207087	1775	C	2	100	49	2.04
NP_207087	1782	C	2	100	49	2.04
NP_207087	2191	C	1	64	49	1.31
NP_207087	2206	C	1	45	49	0.92
NP_208296	106	C	1	100	41	2.44
NP_208296	115	C	1	100	41	2.44
NP_208029	268	C	1	100	56	1.78
NP_208029	271	H	1	95	56	1.69
NP_208029	351	H	1	100	56	1.78
NP_207853	193	H	1	100	55	1.80
NP_207853	231	H	1	100	55	1.80
NP_207004	298	C	1	100	57	1.77
NP_207004	301	C	1	97	57	1.72

NP_206975	80	H	1	99	57	1.73
NP_206975	83	H	1	100	57	1.75
NP_208225	105	H	1	99	53	1.86
NP_208225	106	C	1	88	53	1.65
NP_208225	130	H	1	84	53	1.58
NP_208225	202	H	1	100	53	1.88
NP_207501	38	H	1	97	54	1.81
NP_207501	239	H	1	96	54	1.79
NP_207416	72	C	2	100	59	1.71
NP_207416	109	C	1	100	59	1.71
NP_207444	33	C	1	100	55	1.81
NP_207444	36	C	1	100	55	1.81
NP_207444	116	C	1	91	55	1.64
NP_207056	16	H	1	100	41	2.45
NP_207056	20	H	1	100	41	2.45
NP_207470	127	C	1	100	52	1.92
NP_207470	128	H	1	100	52	1.92
NP_207029	159	C	1	100	46	2.17
NP_207029	162	C	1	100	46	2.17
NP_207693	106	C	1	99	53	1.88
NP_207693	107	H	1	98	53	1.86
NP_208147	147	H	1	99	53	1.86
NP_208147	161	H	1	99	53	1.86
NP_208208	381	H	1	78	33	2.35
NP_208208	386	H	2	100	33	3.01

NP_208208	455	H	1	100	33	3.01
NP_207363	243	C	1	100	46	2.17
NP_207363	246	C	1	100	46	2.17
NP_207376	14	H	1	100	58	1.72
NP_207376	207	H	1	90	58	1.55
NP_207376	243	H	2	100	58	1.72
NP_207303	259	H	1	100	43	2.32
NP_207303	296	H	1	87	43	2.01
NP_207303	306	H	1	98	43	2.27
NP_207934	11	C	1	96	44	2.18
NP_207934	14	C	1	97	44	2.20
NP_207733	126	H	1	94	48	1.95
NP_207733	165	H	1	100	48	2.07
NP_207733	202	H	2	96	48	1.99
NP_207733	318	C	1	85	48	1.76
NP_207584	11	C	1	99	55	1.81
NP_207584	14	C	1	99	55	1.81
NP_207574	48	C	1	100	45	2.23
NP_207574	52	C	1	100	45	2.23
NP_207405	685	H	1	64	51	1.26
NP_207405	797	C	1	100	51	1.97
NP_207405	804	C	1	100	51	1.97
NP_207781	3	C	1	100	52	1.92
NP_207781	80	C	1	100	52	1.92
NP_207781	89	C	1	82	52	1.57
NP_208212	236	H	1	97	46	2.12

NP_208212	244	H	1	100	46	2.18
NP_208300	150	H	1	98	50	1.96
NP_208300	246	H	1	100	50	2.00
NP_207199	184	H	1	98	55	1.79
NP_207199	384	H	1	100	55	1.83
NP_207521	98	C	1	100	51	1.98
NP_207521	166	H	1	100	51	1.98
NP_207521	254	H	2	91	51	1.80
NP_208334	160	H	1	59	42	1.42
NP_208334	169	H	1	91	42	2.18
NP_208334	209	H	1	95	42	2.28
NP_208334	219	H	1	98	42	2.35
NP_207748	189	H	2	98	45	2.20
NP_207748	194	H	1	98	45	2.20
NP_207748	221	H	2	87	45	1.95
NP_207816	100	H	1	95	56	1.70
NP_207816	353	H	1	98	56	1.76
NP_207703	86	H	1	100	41	2.41
NP_207703	590	H	1	96	41	2.32
NP_207016	71	H	1	100	43	2.34
NP_207016	134	H	1	100	43	2.34
NP_207016	166	H	1	83	43	1.94
NP_207802	78	H	1	100	42	2.37
NP_207802	82	H	1	100	42	2.37
NP_207802	183	H	2	77	42	1.83
NP_207651	48	C	2	69	56	1.24

NP_207651	59	H	2	94	56	1.68
NP_207651	173	H	1	95	56	1.70
NP_207651	178	H	1	97	56	1.74
NP_207651	181	C	2	86	56	1.54
NP_208024	246	H	1	98	49	2.00
NP_208024	361	H	1	100	49	2.04
NP_208127	89	C	1	67	52	1.29
NP_208127	92	C	1	100	52	1.92
NP_208127	191	C	1	99	52	1.90
NP_207417	190	H	1	100	48	2.10
NP_207417	354	H	1	97	48	2.03
NP_207344	70	C	1	97	46	2.09
NP_207344	181	C	1	100	46	2.15
NP_207810	256	H	1	100	49	2.04
NP_207810	288	H	1	100	49	2.04
NP_207507	33	H	1	95	44	2.14
NP_207507	40	H	1	100	44	2.25
NP_207490	137	H	1	97	37	2.64
NP_207490	162	H	1	96	37	2.61
NP_207725	13	H	1	94	54	1.73
NP_207725	18	C	1	75	54	1.38
NP_207725	23	C	1	94	54	1.73
NP_207725	28	H	1	100	54	1.84
NP_207725	68	H	2	100	54	1.84
NP_207361	69	C	1	97	41	2.38
NP_207361	75	C	1	86	41	2.11
NP_207361	152	H	1	90	41	2.21

NP_207361	209	C	1	97	41	2.38
NP_207116	8	H	2	100	45	2.22
NP_207116	13	H	1	100	45	2.22
NP_207116	136	H	1	89	45	1.98
NP_207116	245	H	2	86	45	1.91
NP_208005	404	H	1	100	57	1.75
NP_208005	442	C	1	100	57	1.75
NP_207902	22	H	1	77	40	1.92
NP_207902	25	C	1	100	40	2.49
NP_207902	28	C	1	100	40	2.49
NP_207902	53	C	1	91	40	2.26
NP_208335	205	H	1	95	39	2.46
NP_208335	215	H	2	98	39	2.53
NP_208335	248	H	1	91	39	2.35
NP_207315	160	H	1	100	42	2.38
NP_207315	176	C	1	100	42	2.38
NP_207909	313	H	1	74	55	1.34
NP_207909	382	H	1	97	55	1.76
NP_207909	504	H	1	95	55	1.72
NP_208167	23	H	1	100	56	1.78
NP_208167	58	H	1	100	56	1.78
NP_207751	98	H	2	100	38	2.60
NP_207751	209	H	1	100	38	2.60
NP_208013	7	H	1	99	57	1.75
NP_208013	30	H	1	97	57	1.71
NP_208290	102	H	1	100	48	2.07

NP_208290	272	C	1	100	48	2.07
NP_207474	216	C	1	98	48	2.05
NP_207474	426	C	1	96	48	2.01
NP_207474	784	C	1	87	48	1.82
NP_207474	787	C	1	91	48	1.91
NP_207162	123	H	1	100	44	2.28
NP_207162	222	H	1	97	44	2.21
NP_207390	39	C	1	100	44	2.26
NP_207390	42	C	1	100	44	2.26
NP_207390	86	H	1	68	44	1.54
NP_207771	17	H	1	100	48	2.07
NP_207771	21	H	1	100	48	2.07
NP_207695	42	H	1	100	55	1.82
NP_207695	82	H	1	100	55	1.82
NP_207933	283	C	1	100	34	2.92
NP_207933	286	C	1	100	34	2.92
NP_208149	129	H	1	100	45	2.23
NP_208149	132	H	1	99	45	2.21
NP_206815	88	C	1	95	47	2.04
NP_206815	91	C	1	95	47	2.04
NP_206815	92	H	2	47	47	1.01
NP_206815	221	C	1	93	47	2.00
NP_207721	85	H	1	99	51	1.95
NP_207721	91	H	1	99	51	1.95
NP_207721	183	H	1	89	51	1.75
NP_207787	91	H	1	100	58	1.73

NP_207787	93	H	1	100	58	1.73
NP_207878	25	H	1	100	44	2.29
NP_207878	28	H	1	100	44	2.29
NP_208115	173	H	1	99	52	1.92
NP_208115	186	H	1	95	52	1.84
NP_207268	368	H	1	100	42	2.38
NP_207268	372	H	1	96	42	2.29
NP_207832	9	H	1	100	53	1.90
NP_207832	317	H	1	100	53	1.90
NP_207321	265	H	1	95	44	2.18
NP_207321	273	H	1	100	44	2.30
NP_207188	60	C	2	100	57	1.75
NP_207188	94	C	1	97	57	1.70
NP_207467	101	C	2	100	53	1.90
NP_207467	266	C	1	98	53	1.86
NP_207860	434	H	1	100	57	1.74
NP_207860	438	H	1	100	57	1.74
NP_208245	147	C	2	100	50	1.99
NP_208245	257	C	1	100	50	1.99
NP_206850	202	H	1	99	47	2.09
NP_206850	323	H	1	100	47	2.11
NP_207203	242	H	1	96	42	2.30
NP_207203	384	C	1	59	42	1.41
NP_207203	386	C	1	98	42	2.35

NP_207658	192	H	1	100	47	2.11
NP_207658	218	H	1	100	47	2.11
NP_207041	25	H	1	100	46	2.18
NP_207041	37	H	1	100	46	2.18
NP_207618	133	C	1	100	55	1.83
NP_207618	136	C	1	100	55	1.83
NP_207719	182	H	1	100	46	2.19
NP_207719	186	H	1	100	46	2.19
NP_207063	31	C	2	100	45	2.22
NP_207063	153	C	1	100	45	2.22
NP_207868	44	H	1	97	44	2.18
NP_207868	79	H	1	99	44	2.23
NP_207795	131	H	1	94	57	1.66
NP_207795	171	H	1	100	57	1.76
NP_207795	173	H	2	100	57	1.76
NP_208192	195	H	1	100	35	2.87
NP_208192	199	H	1	100	35	2.87
NP_208129	24	H	1	97	42	2.30
NP_208129	27	H	1	100	42	2.37
NP_207617	30	C	1	100	57	1.74
NP_207617	33	C	1	100	57	1.74
NP_206955	87	H	1	100	49	2.05
NP_206955	170	H	1	95	49	1.95
NP_206977	172	C	1	95	49	1.94
NP_206977	204	H	1	100	49	2.04

NP_207019	131	C	2	85	48	1.78
NP_207019	161	C	1	88	48	1.84
NP_207019	163	C	2	85	48	1.78
NP_207019	196	C	1	85	48	1.78
NP_207019	199	C	1	90	48	1.88
NP_207019	295	C	1	100	48	2.09
NP_207019	298	C	1	100	48	2.09
NP_207599	158	H	1	100	38	2.61
NP_207599	162	H	1	100	38	2.61
NP_207741	125	H	2	98	40	2.45
NP_207741	149	H	1	100	40	2.50
NP_207973	108	C	1	100	51	1.98
NP_207973	111	C	1	94	51	1.86
NP_207973	135	H	1	86	51	1.70
NP_207973	136	H	1	97	51	1.92
NP_208375	36	H	2	92	53	1.75
NP_208375	111	H	1	100	53	1.90
NP_208375	115	H	1	100	53	1.90
NP_208375	139	H	1	94	53	1.79
NP_208375	302	C	2	86	53	1.64
NP_208283	49	C	1	100	54	1.87
NP_208283	52	C	1	100	54	1.87
NP_208249	28	C	1	98	50	1.96
NP_208249	31	C	1	100	50	2.00
NP_206989	132	H	1	88	44	2.02
NP_206989	167	H	1	100	44	2.30
NP_206989	401	H	1	97	44	2.23

NP_207109	2	C	1	100	53	1.90
NP_207109	5	C	1	100	53	1.90
NP_207109	13	H	1	91	53	1.73

Tabella Supplementare S3.2*Risultati della ricerca per parole chiave. La colonna "Parola Chiave" indica quale delle parole chiave è stata trovata nel record GenBank della sequenza*

Accession Number	Locus	Parola Chiave	Annotazione
NP_206806	HP0004	zinc	carbonic anhydrase (icfA)
NP_206812	HP0010	Mg	molecular chaperone GroEL
NP_206814	HP0012	zinc	DNA primase
NP_206838	HP0036	zinc	hypothetical protein
NP_206848	HP0047	Ni	hydrogenase expression/formation protein HypE
NP_206870	HP0070	nickel	urease accessory protein UreE
NP_206872	HP0072	nickel	urease subunit beta
NP_206873	HP0073	nickel	urease subunit alpha
NP_206875	HP0075	metal	phosphoglucosamine mutase
NP_206902	HP0102	metal	hypothetical protein
NP_206904	HP0104	metal	2',3'-cyclic-nucleotide 2'-phosphodiesterase
NP_206916	HP0116	metal	DNA topoisomerase I
NP_206917	HP0117	Fe	hypothetical protein
NP_206932	HP0132	iron	L-serine deaminase (sdaA)
NP_206938	HP0138	iron	iron-sulfur protein
NP_206953	HP0154	metal	phosphopyruvate hydratase
NP_206956	HP0157	magnesium	shikimate kinase
NP_206958	HP0159	metal	lipopolysaccharide 1,2-glucosyltransferase RfaJ
NP_206975	HP0176	zinc	fructose-bisphosphate aldolase
NP_206990	HP0191	iron	fumarate reductase iron-sulfur subunit
NP_206991	HP0192	iron	fumarate reductase flavoprotein subunit
NP_206992	HP0193	iron	fumarate reductase cytochrome b-556 subunit
NP_206993	HP0194	metal	triosephosphate isomerase
NP_206999	HP0200	zinc	50S ribosomal protein L32
NP_207010	HP0212	metal	succinyl-diaminopimelate desuccinylase
NP_207019	HP0221	Fe	nifU-like protein
NP_207021	HP0223	Mg	DNA repair protein RadA
NP_207045	HP0247	Mg	DEAD/DEAH box helicase
NP_207048	HP0250	nickel	oligopeptide ABC transporter ATP-binding protein (oppD)
NP_207056	HP0258	Zn	hypothetical protein
NP_207067	HP0269	iron	(dimethylallyl)adenosine tRNA methylthiotransferase

NP_207081	HP0283	metal	3-dehydroquinate synthase
NP_207083	HP0285	iron	hypothetical protein
NP_207084	HP0286	Zn	cell division protein (ftsH)
NP_207099	HP0301	nickel	dipeptide ABC transporter ATP-binding protein (dppD)
NP_207100	HP0302	nickel	dipeptide ABC transporter ATP-binding protein (dppF)
NP_207108	HP0310	Zn	hypothetical protein
NP_207127	HP0329	Mg	NAD synthetase
NP_207174	HP0376	iron	ferrochelatase
NP_207185	HP0387	Mg	primosome assembly protein PriA
NP_207187	HP0389	manganese	iron-dependent superoxide dismutase
NP_207192	HP0394	metal	hypothetical protein
NP_207200	HP0402	magnesium	phenylalanyl-tRNA synthetase subunit beta
NP_207201	HP0403	magnesium	phenylalanyl-tRNA synthetase subunit alpha
NP_207215	HP0417	zinc	methionyl-tRNA synthetase
NP_207226	HP0428	metal	phage/colicin/tellurite resistance cluster terY protein
NP_207238	HP0440	metal	DNA topoisomerase I (topA)
NP_207264	HP0466	metal	hypothetical protein
NP_207268	HP0470	Zn	oligoendopeptidase F (pepF)
NP_207280	HP0482	metal	hypothetical protein
NP_207290	HP0493	Mg	phospho-acetylmuramoyl-pentapeptide- transferase
NP_207296	HP0499	calcium	phospholipase A1 precursor (DR-phospholipase A)
NP_207298	HP0501	metal	DNA gyrase subunit B
NP_207304	HP0507	metal	hypothetical protein
NP_207346	HP0551	zinc	50S ribosomal protein L31
NP_207365	HP0570	zinc	leucyl aminopeptidase
NP_207370	HP0575	zinc	hypothetical protein
NP_207380	HP0585	iron	endonuclease III (nth)
NP_207387	HP0592	Mg	type III restriction enzyme R protein (res)
NP_207410	HP0615	zinc	NAD-dependent DNA ligase LigA
NP_207414	HP0620	metal	inorganic pyrophosphatase
NP_207426	HP0632	Ni	quinone-reactive Ni/Fe hydrogenase, large subunit (hydB)
NP_207427	HP0633	Ni	quinone-reactive Ni/Fe hydrogenase, cytochrome b

			subunit (hydC)
NP_207428	HP0634	nickel	quinone-reactive Ni/Fe hydrogenase (hydD)
NP_207434	HP0640	metal	poly(A) polymerase
NP_207448	HP0654	iron	hypothetical protein
NP_207450	HP0656	iron	hypothetical protein
NP_207451	HP0657	Zn	processing protease (ymxG)
NP_207454	HP0660	Zn	hypothetical protein
NP_207456	HP0662	magnesium	ribonuclease III
NP_207459	HP0665	iron	coproporphyrinogen III oxidase
NP_207460	HP0666	Fe	glycerol-3-phosphate dehydrogenase
NP_207462	HP0668	Mg	hypothetical protein
NP_207477	HP0683	Mg	bifunctional acetylglucosamine-1-phosphate uridyltransferase/glucosamine-1-phosphate acetyltransferase
NP_207481	HP0687	iron	iron(II) transport protein (feoB)
NP_207531	HP0737	metal	hypothetical protein
NP_207561	HP0768	iron	molybdenum cofactor biosynthesis protein A
NP_207565	HP0772	metal	acetylmuramoyl-L-alanine amidase
NP_207568	HP0775	metal	penta-phosphate guanosine-3'-pyrophosphohydrolase (spoT)
NP_207584	HP0791	metal	cadmium-transporting ATPase, P-type (cadA)
NP_207585	HP0792	Mg	Fis family transcriptional regulator
NP_207586	HP0793	iron	peptide deformylase
NP_207606	HP0813	Zn	hypothetical protein
NP_207614	HP0821	metal	excinuclease ABC subunit C
NP_207639	HP0846	Mg	type I restriction enzyme R protein (hsdR)
NP_207663	HP0869	nickel	hydrogenase nickel incorporation protein
NP_207673	HP0879	metal	hypothetical protein
NP_207681	HP0888	iron	iron(III) dicitrate ABC transporter ATP-binding protein (fecE)
NP_207696.2	HP0903m	magnesium	acetate kinase
NP_207717	HP0925	metal	recombination protein RecR
NP_207750	HP0958	Zn	hypothetical protein
NP_207771	HP0980	zinc	hypothetical protein
NP_207802	HP1012	Zn	protease (pqqE)

NP_207810	HP1020	zinc	bifunctional 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase/2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
NP_207812	HP1022	metal	hypothetical protein
NP_207817	HP1027	iron	ferric uptake regulation protein
NP_207834	HP1044	metal	hypothetical protein
NP_207842	HP1051	metal	hypothetical protein
NP_207843	HP1052	zinc	UDP-3-O-[3-hydroxymyristoyl] acetylglucosamine deacetylase
NP_207849	HP1058	metal	3-methyl-2-oxobutanoate hydroxymethyltransferase
NP_207863	HP1072	metal	copper-transporting ATPase, P-type (copA)
NP_207864	HP1073	metal	copper ion binding protein (copP)
NP_207866	HP1075	Zn	hypothetical protein
NP_207868	HP1077	nickel	nickel transport protein (nixA)
NP_207869	HP1078	metal	hypothetical protein
NP_207895	HP1104	Zn	cinnamyl-alcohol dehydrogenase ELI3-2 (cad)
NP_207936	HP1145	metal	hypothetical protein
NP_207951	HP1160	metal	metalloprotease
NP_207977	HP1186	zinc	carbonic anhydrase
NP_207996	HP1204	zinc	50S ribosomal protein L33
NP_208012	HP1220	Mn	ABC transporter ATP-binding protein
NP_208014	HP1222	Fe	D-lactate dehydrogenase (dld)
NP_208018	HP1226	iron	coproporphyrinogen III oxidase
NP_208020	HP1228	metal	dinucleoside polyphosphate hydrolase
NP_208021	HP1229	Mg	aspartate kinase
NP_208044	HP1252	nickel	oligopeptide ABC transporter periplasmic oligopeptide-binding protein (oppA)
NP_208051	HP1259	Zn	hypothetical protein
NP_208058	HP1266	iron	NADH dehydrogenase subunit G
NP_208067	HP1275	metal	phosphomannomutase
NP_208079	HP1287	iron	TenA transcriptional regulator
NP_208089	HP1297	zinc	50S ribosomal protein L36
NP_208091	HP1299	Zn	methionine aminopeptidase
NP_208098	HP1306	zinc	30S ribosomal protein S14
NP_208115	HP1323	manganese	ribonuclease HII

NP_208117	HP1325	metal	fumarate hydratase
NP_208124	HP1332	Zn	molecular chaperone DnaJ
NP_208126	HP1334	metal	hypothetical protein
NP_208130	HP1338	nickel	nickel responsive regulator
NP_208136	HP1344	magnesium	magnesium and cobalt transport protein (corA)
NP_208153	HP1361	metal	competence locus E (comE3)
NP_208162	HP1371	Mg	type III restriction enzyme R protein
NP_208164	HP1373	Mg	rod shape-determining protein MreB
NP_208165	HP1374	zinc	ATP-dependent protease ATP-binding subunit ClpX
NP_208170	HP1379	Mg	ATP-dependent protease (lon)
NP_208177	HP1386	metal	ribulose-phosphate 3-epimerase
NP_208190	HP1399	Mn	arginase
NP_208193	HP1402	Mg	type I restriction enzyme R protein (hsdR)
NP_208197	HP1406	iron	biotin synthase
NP_208207	HP1416	metal	lipopolysaccharide 1,2-glucosyltransferase RfaJ
NP_208208	HP1417m	metal	hypothetical protein
NP_208219	HP1428	iron	ribosomal RNA large subunit methyltransferase NO
NP_208251	HP1460	metal	DNA polymerase III subunit alpha
NP_208261	HP1470	metal	DNA polymerase I (polA)
NP_208294	HP1503	metal	cation-transporting ATPase, P-type (copA)
NP_208296	HP1505	Zn	riboflavin biosynthesis protein (ribG)
NP_208312	HP1521	Mg	type III restriction enzyme R protein (res)
NP_208313	HP1523	Mg	ATP-dependent DNA helicase RecG
NP_208316	HP1526	metal	exodeoxyribonuclease III
NP_208331	HP1540	iron	ubiquinol cytochrome c oxidoreductase, Rieske 2Fe-2S subunit (fbcF)
NP_208332	HP1541	Mg	transcription-repair coupling factor (trcF)
NP_208352	HP1561	metal	iron(III) ABC transporter periplasmic iron-binding protein (ceuE)
NP_208353	HP1562	metal	iron(III) ABC transporter periplasmic iron-binding protein

			(ceuE)
NP_208364	HP1573	metal	hypothetical protein
NP_208368	HP1577	metal	ABC transporter permease (yaeE)
NP_208372	HP1581	Mg	methicillin resistance protein (llm)
NP_208375	HP1584	iron	DNA-binding/iron metalloprotein/AP endonuclease

Tabella Supplementare S3.3 *Proteine ipotetiche trovate da MetalDetector e annotazione funzionale che abbiamo loro assegnato. Omologhi PDB: indica se la sequenza ha omologhi in PDB. Dominio Pfam: indica se la proteina ha domini funzionali annotati in Pfam. Nella colonna "Annotazioni" vengono riportate le annotazioni della proteina e dei suoi omologhi; il numero tra parentesi quadre indica quante volte ricorre ogni annotazione.*

Tag	Accession number	Annotazioni	Classe funzionale	Omologhi PDB	Domini Pfam
HP0100	NP_206900	hypothetical protein HP0100[267],DNA polymerase III gamma[3],FIG053235: Diacylglucosamine hydrolase like[2],uncharacterized BCR, COG1636 family protein[3]	nucleasi	no	DUF208
HP1089	NP_207880	hypothetical protein HP1089[91],exonuclease[11],PD-(D/E)XK nuclease superfamily protein[38]	nucleasi	no	PDDEXK_1
HP1573	NP_208364	hypothetical protein HP1573[38],mg-dependent DNase[73],Deoxyribonuclease[19],DNAse of the TatD family[75],glyoxalase II[1],metal-dependent DNA hydrolase of TatD family[48],Uncharacterized deoxyribonuclease yabD[35],sec-independent secretion protein tatD[26],putative TatD related DNase[8]	nucleasi	no	TatD_DNase
HP0049	NP_206850	hypothetical protein HP0049[126],putative agmatine deiminase[4],peptidylarginine deiminase domain-containing protein[162],non-functional type II restriction endonuclease[4]	ciclo dell'urea	sì	PAD_porph
HP0117	NP_206917	hypothetical protein HP0117[112],radical SAM superfamily protein[44],putative radical SAM domain-containing protein[82],Fe-SÌ oxidoreductase[15],MoaA/NifB/PqqE family protein, putative[5],signal recognition particle protein[16],GTPase ObgE[6]	radical SAM	sì	Radical_SAM

HP0568	NP_207363	hypothetical protein HP0568[144],radical SAM domain-containing protein[121],molybdenum cofactor biosynthesis enzyme /related Fe-SÌ[27],pyruvate formate-lyase activating enzyme[1],putative Fe-SÌ oxidoreductase[5],tungsten-containing aldehyde ferredoxin oxidoreductase[4]	radical SAM	sì	SPASM
HP1473	NP_208264	hypothetical protein HP1473[64],amidophosphoribosyltransferase[96],HP1473-transformation associated protein[54],putative[1],late competence protein comFC[37]	competenza	no	no
HP0861	NP_207655	hypothetical protein HP0861[241],integral membrane protein[3],brp/Blh family beta-carotene 15,15'-monooxygenase[1],Heavy-metal-associated domain (terminus) and membrane-bounded[7],putative membrane copper tolerance protein[8],Cbb3-type cytochrome oxidase assembly protein disulfide bond[9],heavy metal transport/detoxification protein[35],Ferric reductase domain protein transmembrane component domain[21]	detossificazione	no	DsbD_2
HP0713	NP_207507	hypothetical protein HP0713[192],filamentation induced by cAMP protein fic[56],toxin-antitoxin system, toxin component, Fic family[24],ORF2[1],LOW QUALITY PROTEIN: fic family protein[18],Oligopeptide ABC transporter, periplasmic oligopeptide-binding[1],uncharacterized conserved protein[4]	divisione cellulare	sì	Fic
HP1401	NP_208192	hypothetical protein HP1401[150],putative metal-dependent hydrolase[78],type I site-specific	enzima di restrizione	no	DUF45

		deoxyribonuclease, HsdR family protein[3],Zinc metalloprotease[25],putative uncharacterized protein[2],peptidase S26A[1],peptidase S26A, signal peptidase I[2]			
HP1499	NP_208290	hypothetical protein HP1499[69],putative methyltransferase[2],putative HKD family nuclease[1],DEAD/DEAH box helicase family protein[7],type III restriction protein res subunit[50],HKD family nuclease fused to DNA/RNA helicase of superfamily II[14],ABC transporter, permease/ATP-binding protein[12],ATP-dependent RNA helicase, DEAD/DEAH box family[59]	enzima di restrizione	no	PLDc_2
HP0781	NP_207574	hypothetical protein HP0781[157],outer membrane HofE domain protein[1],putative periplasmic protein[41],DNA polymerase, bacteriophage-type[1]	proteina fagica	no	no
HP0959	NP_207751	hypothetical protein HP0959[239],dinuclear metal center protein, YbgI family[40],putative NIF3 family protein family protein[6],NGG1p interacting factor 3 protein, NIF3[12],UPF0135 protein Bsu YqfO[6]	fattore di trascrizione	no	NIF3
HP0274	NP_207072	hypothetical protein HP0274[200],flagellin methylase family protein[9],putative oxidoreductase[9],Uncharacterised protein family (UPF0153)[18],LOW QUALITY PROTEIN: putative Fe-SÌ oxidoreductase[7],THAP domain-containing protein 7[1]	flagello	sì	CxxCxxCC
HP0518	NP_207315	hypothetical protein HP0518[155],putative periplasmic protein[34],L,D-transpeptidase catalytic domain	flagello	sì	YkuD

		protein[17],ErfK/YbiS/YcfS/YnhG family protein[51],60 kDa chaperonin (protein Cpn60; groEL protein)[5],LOW QUALITY PROTEIN: Putative exported protein[3]			
HP0958	NP_207750	hypothetical protein HP0958[166],zinc ribbon domain-containing protein[96],Myosin-2 heavy chain, non muscle[1],Zn-ribbon protein, possibly nucleic acid-binding protein[22]	flagello	no	zf-RING_7
HP0156	NP_206955	hypothetical protein HP0156[132],putative periplasmic protein[38]	ipotetica	no	AMIN
HP0205	NP_207004	hypothetical protein HP0205[219],ATP-dependent OLD family endonuclease[30],AAA ATPase domain protein[14],serine acetyltransferase[3],ATP-dependent endonuclease of the OLD family-like protein[10]	ipotetica	no	AAA_21
HP0660	NP_207454	hypothetical protein HP0660[179],putative periplasmic protein[23],SpoOJ regulator[1],TPR repeat-containing protein[3],tetratricopeptide repeat protein[39],predicted acetylglucosaminyl transferase YciM[3],heat shock (predicted periplasmic) protein YciM, precursor[6]	ipotetica	sì	no
HP0673	NP_207467	hypothetical protein HP0673[269]	ipotetica	no	no
HP0838	NP_207631	hypothetical protein HP0838[108],probable lipoprotein Cj0375[34]	ipotetica	no	no
HP0806	NP_207599	hypothetical protein HP0806[229],putative hydrolase[24],zinc metalloprotease[48],PF01863 domain protein[8],putative zinc metallopeptidase protein[9]	ipotetica	no	DUF45
HP0902	NP_207695	hypothetical protein HP0902[184],double-stranded beta-	ipotetica	sì	no

		helix domain-containing protein[7],AraC-like regulator[1],RmlC-like cupin family protein[13],Cupin 2 conserved barrel domain protein[45],Nitric oxide dioxygenase[1]			
HP0990	NP_207781	hypothetical protein HP0990[165]	ipotetica	no	no
HP1519	NP_208310	hypothetical protein HP1519[59],putative cytoplasmic protein[9],macrophage infection , MimD domain protein[12]	ipotetica	no	no
HP1143	NP_207934	hypothetical protein HP1143[257]	ipotetica	no	DUF2130
HP1454	NP_208245	hypothetical protein HP1454[145],putative[1]	ipotetica	no	LPP20
HP0394	NP_207192	hypothetical protein HP0394[137],UDP-2,3-diacylglucosamine hydrolase[48],calcineurin-like phosphoesterase family protein[21],Ser/Thr protein phosphatase[48],pyridine nucleotide-disulfide oxidoreductase YkgC[1]	membrana e parete	no	Metallophos
HP1430	NP_208221	ATP-binding protein[2],ribonuclease J[120],RNA-metabolising Zn-dependent hydrolase[11],metallo-beta-lactamase[7],conserved hypothetical ATP-binding protein[43],peptide ABC transporter periplasmic peptide-binding protein[4],putative metallo-beta-lactamase family protein[125],YkqC[1],hydroxyacylglutathione hydrolase[2]	maturazio ne rRNA	no	Lactamase_B, RMMBL
HP1449	NP_208240	hypothetical protein HP1449[150],yidD[1],DUF37 domain containing protein[1],alpha-hemolysin[7],toxin-antitoxin system toxin component[1]	membrana e parete	no	Haemolytic
HP0318	NP_207116	hypothetical protein HP0318[58],heme oxygenase, HugZ	metabolis	sì	DUF2470,

		family[45],putative heme iron utilization protein[60],pyridoxamine 5'-phosphate oxidase family protein[85],putative fMN-binding split barrel[15],adenylate cyclase[1]	mo dei cofattori		Pyridox_oxidase
HP0654	NP_207448	hypothetical protein HP0654[80],menaquinone biosynthesis protein, SCO4494 family[122],Gene SCO4494, often clustered with other genes in menaquinone via[1],putative Radical-SAM domain-containing protein[15],FO synthase subunit 2 (7,8-didemethyl-8-hydroxy-5-deazariboflavin[4],dehypoxanthinylfutalosine cyclase[8],Thiamine biosynthesis enzyme ThiH or related uncharacterized enzyme[25]	metabolismo dei cofattori	sì	Radical_SAM
HP0656	NP_207450	hypothetical protein HP0656[69],menaquinone biosynthesis protein, SCO4550 family[115],FO synthase subunit 2 1[2],radical SAM domain-containing protein[86],de-hypoxanthine futalosine cyclase[13],FO synthase subunit 2 (7,8-didemethyl-8-hydroxy-5-deazariboflavin[6],thiamine biosynthesis enzyme ThiH and related uncharacterized[26],oxidoreductase[1]	metabolismo dei cofattori	sì	Radical_SAM
HP0933	NP_207725	hypothetical protein HP0933[32],6-carboxy-5,6,7,8-tetrahydropterin synthase[102],queuosine biosynthesis protein QueD[34],6-pyruvoyl tetrahydropterin synthase/QueD family protein[83],beta-phosphoglucomutase[2]	metabolismo dei cofattori	no	PTPS
HP1337	NP_208129	hypothetical protein HP1337[29],nicotinate-nucleotide	metabolis	no	CTP_transf_lik

		adenyltransferase[345],cytidyltransferase-related domain protein[4],metal dependent phosphohydrolase[1]	mo dei cofattori		e
HP0138	NP_206938	iron-sulfur protein[93],putative L-lactate dehydrogenase, Iron-sulfur cluster-binding[11],conserved hypothetical iron-sulfur protein[69],putative oxidoreductase subunit with NAD(P)-binding domain and[15],(4Fe-4S) cluster-containing protein[166],unnamed protein product[3],putative electron transport protein YkgF[5],predicted electron transport protein with ferridoxin-like domai[20],YvfW[1],sn-glycerol-3-phosphate dehydrogenase subunit C[1]	metabolismo degli zuccheri	no	DUF162, DUF3390, Fer4_8
HP0139	NP_206939	hypothetical protein HP0139[188],(SÌ)-2-hydroxy-acid oxidase[17],Fe-SÌ oxidoreductase[5],cysteine-rich domain protein[16],oxidoreductase iron-sulfur subunit[3],putative L-lactate dehydrogenase, Fe-SÌ oxidoreductase subunit YkgE[68],leucine-rich-repeat type III effector protein[2],coB-CoM heterodisulfide reductase[24],unnamed protein product[2],Fe-SÌ oxidoreductase, glycolate/lactate utilization protein[32],adenine phosphoribosyltransferase[3],putative (SÌ)-2-hydroxy-acid oxidase, iron-sulfur subunit[16],YvfV[2],fumarate reductase-related protein[25],Cysteine-rich domain of 2-hydroxy-acid oxidase GlcF homolog[31],putative hydroxyacid oxidoreductase (Fe-SÌ centre)[11]	metabolismo degli zuccheri	no	CCG
HP0764	NP_207557	hypothetical protein HP0764[195],proteobacterial sortase system OmpA family protein[65],ancestral	modificazioni post-	no	CxxCxxCC

		polypeptide[2],flagellin methylase[2]	traduzionali		
HP0013	NP_206815	hypothetical protein HP0013[82],tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase[88],tRNA methyl transferase family protein[48],argininosuccinate synthase[26],thiamine biosynthesis protein:ExsB[31],putative ATP-binding protein[12]	maturazione tRNA	sì	tRNA_Me_trans
HP0285	NP_207083	hypothetical protein HP0285[114],radical SAM methylthiotransferase, MiaB/RimO family protein[82],MiaB-like tRNA modifying enzyme[61],putative 2-methylthioadenine synthase[17],miaB family protein, possibly involved in tRNA or rRNA modification[33],tRNA (2-methylthio-N6-threonylcarbamyl-A37) methylthiotransferase[66],RNA modification enzyme, MiaB family[35],Fe-SÌ oxidoreductase[7],Holliday junction DNA helicase subunit RuvA[1],putative AdoMet-dependent methyltransferase,UPF0004 family[1]	maturazione tRNA	sì	Radical_SAM, UPF0004
HP0734	NP_207528	hypothetical protein HP0734[80],ribosomal protein S12 methylthiotransferase RimO[157],MiaB-like tRNA modifying enzyme YliG[70],putative 2-methylthioadenine synthetase[20],radical SAM domain-containing protein[46],Fe-SÌ oxidoreductase[8],ribosome maturation factor rimO[1]	maturazione tRNA	sì	Radical_SAM, UPF0004
HP0956	NP_207748	hypothetical protein HP0956[52],pseudouridine	maturazio	no	PseudoU_synt

		synthase[5],RNA pseudouridylate synthase family protein[18],Ribosomal large subunit pseudouridine synthase D[87],Pseudouridine synthase (RNA-uridine isomerase) (RNA pseudouridylate[211]	ne tRNA		h_2
HP1142	NP_207933	hypothetical protein HP1142[243],putative ATP-binding cassette[3],methionyl-tRNA formyltransferase[1],trans-Golgi membrane protein p230[12],putative uncharacterized protein[10],cation transport ATPase[5],metallophosphoesterase[2]	maturazione tRNA	no	AAA_13
HP1182	NP_207973	hypothetical protein HP1182[48],tRNA 2-thiocytidine biosynthesis protein TtcA[208],putative PP-loop family ATPase[13],adenine nucleotide alpha hydrolase[1],putative ATPase of the PP-loop superfamily implicated in cell cycle[12],PP-loop domain-containing protein[99],C32 tRNA thiolase[77]	maturazione tRNA	no	ATP_bind_3
HP0595	NP_207390	hypothetical protein HP0595[48],disulfide bond formation protein[151],putative ATP/GTP binding protein[12],disulfide oxidoreductase B (DsbB-like protein); membrane protein[49]	modificazioni post-traduzionali	sì	DsbB
HP0231	NP_207029	hypothetical protein HP0231[70],disulfide interchange protein DsbC[84],isomerase[1],putative disulfide isomerase[9],thiol peroxidase[1],Putative periplasmic protein[11],multi-sensor signal transduction histidine kinase[3],disulfide isomerase/thiol-disulfide oxidase[7],DSBA-like thioredoxin domain protein[1]	modificazioni post-traduzionali	sì	no

HP1042	NP_207832	hypothetical protein HP1042[148],putative phosphoesterase RecJ-like protein[63],3'-to-5' oligoribonuclease B[7],putative DHD superfamily phosphohydrolase[6],DHHA1 domain-containing protein[5],putative DHH family phosphohydrolase involved in recombination YngD[23]	nucleasi	no	no
HP0506	NP_207303	hypothetical protein HP0506[69],putative outer membrane protein[65],peptidase M23 family protein[87],Membrane proteins related to metalloendopeptidases[34],probable periplasmic protein Cj1215[10],cell wall endopeptidase, family M23/M37[35]	membrana e parete	sì	Peptidase_M23
HP0996	NP_207787	hypothetical protein HP0996[132],relaxase[10],relaxase/mobilization nuclease domain protein[76],virD2[2],putative VirD2[14],PREDICTED: ecotropic viral integration site 5[1]	plasticità del DNA	no	no
HP1004	NP_207795	hypothetical protein HP1004[76],relaxase[12],relaxase/mobilization nuclease domain protein[77],putative pZ10b[5],virD2[1],Type IV secretory pathway VirD2 component[1]	plasticità del DNA	no	Relaxase
HP0258	NP_207056	hypothetical protein HP0258[49],RIP metalloprotease RseP[12],peptidase M50 (membrane-associated zinc metallopeptidase), MEROPS[209],integral membrane protein[3],putative peptidase M50 family protein[52],regulator of sigma E protease[12],RseP peptidase. Metallo peptidase. MEROPS family M50B[4]	membrana e parete	sì	Peptidase_M50

HP0980	NP_207771	hypothetical protein HP0980[49],RIP metalloprotease RseP[12],peptidase M50 family protein[26],membrane-associated zinc metalloprotease[174],integral membrane protein[3],regulator of sigma E protease[5],putative transmembane regulator of protease[21],M50 peptidase family metallopeptidase[28],zinc protease, putative[7],Site-2 protease, Metallo peptidase, MEROPS family M50B[19]	membrana e parete	no	Peptidase_M50
HP1037	NP_207827	hypothetical protein HP1037[30],proline dipeptidase[6],metallopeptidase M24 family protein[108],Aminopeptidase YpdF (MP-, MA-, MS-, AP-,NP-specific)[24],prolidase (Xaa-Pro dipeptidase) (pepQ)[36],DNA polymerase III gamma and tau subunits[2],Uncharacterized peptidase yqhT[107],RNA methyltransferase RsmE[3]	membrana e parete	no	Peptidase_M24
HP0622	NP_207416	hypothetical protein HP0622[86],putative[1],putative outer membrane protein[4]	membrana e parete	no	no
HP1417m	NP_208208	hypothetical protein HP1417m[218],sulfatase family protein[40],putative transmembrane sulphatase[44],phosphoribosylformylglycinamidine cyclo-ligase[4],phosphoethanolamine transferase CptA[32],putative membrane-associated, metal-dependent hydrolase[29],UPF0141 membrane protein YijP possibly required for[22]	membrana e parete	no	Sulfatase
HP1580	NP_208371	hypothetical protein HP1580[131],PAP2 superfamily protein[11],phosphatase/haloperoxidase[2],phosphoestera	membrana e parete	no	PAP2

		se pa-phosphatase related protein[3],PAP2 (2 phosphatidic acid phosphatase) family protein[47],membrane-associated phospholipid phosphatase[56],lipoprotein signal peptidase[1]			
HP0813	NP_207606	hypothetical protein HP0813[146],metallo-beta-lactamase[22],putative hydrolase[20],beta-lactamase domain-containing protein[48],probable hydrolase Cj0809c[8],DNA polymerase III subunit epsilon[1],thioredoxin reductase[1],Zn-dependent hydrolase, glyoxylase[25]	resistenza agli antibiotici	sì	Lactamase_B
HP0864	NP_207658	hypothetical protein HP0864[109],putative 50S ribosomal protein L22[75]	proteine ribosomiali	no	no
HP0650	NP_207444	hypothetical protein HP0650[68],family 4 uracil-DNA glycosylase[119],transferase[4],phage spo1 DNA polymerase-like protein[61]	riparazione DNA	sì	UDG
HP0190	NP_206989	hypothetical protein HP0190[49],phospholipase D-family protein[117],putative cardiolipin synthase[12],phosphatidylserine/phosphatidylglycerophosphate/cardiolipin[88],putative protein with phospholipase D/nuclease domain[73],Cardiolipin synthetase Cardiolipin synthase; CL synthase[2]	trasduzione segnale	sì	PLDc_2
HP0218	NP_207016	hypothetical protein HP0218[103],phosphatidylethanolamine-binding family protein[8],phospholipid-binding protein[144],Raf kinase inhibitor-like protein[15],Uncharacterized protein yxkA[32]	trasduzione segnale	no	PBP

HP0404	NP_207202	HIT family protein[3],Hit-like protein involved in cell-cycle regulation[6],ADP hydrolase of the hit protein family[7],protein kinase C inhibitor[5],histidine triad nucleotide-binding protein 1[48],HIT family hydrolase, diadenosine tetraphosphate hydrolase[71],purine nucleoside phosphoramidase[21],Putative 13.2 kDa HIT-like protein in hisE 3'region[31],scavenger mRNA decapping enzyme[4],Bis(5'-nucleosyl)-tetraphosphatase (asymmetrical)[9],possible HIT histidine triad hydrolase[14],PREDICTED: fibroblast growth factor receptor 1-A-like[1],member of HIT (histidine triad) family that contains Ap3A and Ap4A[29],inhibitor of protein kinase C, contains a transferase domain[19]	trasduzione segnale	sì	HIT
HP0741	NP_207535	hypothetical protein HP0741[16],histidine triad family protein[161],HIT family hydrolase, diadenosine tetraphosphate hydrolase[64],scavenger mRNA decapping enzyme[4],AP-4-A phosphorylase[1],putative bis(5'-adenosyl)-triphosphatase, HIT family[35]	trasduzione segnale	sì	HIT
HP1044	NP_207834	hypothetical protein HP1044[113],phosphodiesterase domain-containing protein[7],membrane-bound phosphoesterase; membrane protein[66],Ser/Thr phosphatase family protein[110],Predicted DNA repair exonuclease[2],putative metallophosphoesterase protein[23]	trasduzione segnale	no	Metallophos
HP0129	NP_206929	hypothetical protein HP0129[234],high-affinity zinc	trasporto	sì	DUF1104

		transporter periplasmic component[31]			
HP0311	NP_207109	hypothetical protein HP0311[189]	trasporto	no	no
HP0236	NP_207034	hypothetical protein HP0236[145],cytochrome C oxidase, cbb3-type, subunit III family protein[14],probable periplasmic protein Cj0854c[28]	trasporto elettroni	no	Cytochrome_CBB3

Tabella Supplementare S4.1 *Rapporti dN/dS per I confronti a coppie tra* H. pylori *e* H. acinonychis

	H. acinonychis	***H. pylori* 26695**	***H. pylori* Shi417**	***H. pylori* Gambia94/24**	***H. pylori* P12**	***H. pylori* G27**	***H. pylori* B8**	***H. pylori* ELS37**	***H. pylori* India7**	***H. pylori* PeCan4**
H. acinonychis										
***H. pylori* 26695**	0.261									
***H. pylori* Shi417**	0.289	0.365								
***H. pylori* Gambia94/24**	0.265	0.446	0.369							
***H. pylori* P12**	0.235	0.316	0.276	0.473						
***H. pylori* G27**	0.258	0.327	0.291	0.598	0.371					
***H. pylori* B8**	0.235	0.309	0.284	0.441	0.230	0.325				
***H. pylori* ELS37**	0.249	0.385	0.340	0.521	0.411	0.460	0.348			
***H. pylori* India7**	0.229	0.244	0.294	0.495	0.262	0.298	0.220	0.459		
***H. pylori* PeCan4**	0.257	0.536	0.282	0.377	0.247	0.327	0.324	0.357	0.293	
***H. pylori* SouthAfrica**	0.343	0.295	0.248	0.375	0.288	0.311	0.296	0.342	0.246	0.261

Tabella Supplementare S5.1 *Geni selettivamente conservati negli* Helicobacter *gastrici. "Full" e "Ssd" indicano i rispettivi filtri adottati (vedi testo); una x indica che il gene ha passato quel filtro sui pattern di presenza. Gli accession number sono relativi ad* H. pylori*.*

Accession number	Full	Ssd	Descrizione	Classificazione	Unica degli ε-proteobattteri?	Presente in H. mustelae?
NP_206817	x	x	hypothetical protein HP0015	Ipotetica	sì	no
NP_206839	x	x	NADH-ubiquinone oxidoreductase subunit/ComB6 competence protein	Competenza	sì	no
NP_206984	x	x	hypothetical protein HP0185	Ipotetica	sì	no
NP_207002	x		hypothetical protein HP0203	Ipotetica	sì	no
NP_207040	x	x	hypothetical protein Hac_1376	Ipotetica	sì	sì
NP_207047	x	x	hypothetical protein HP0249	Ipotetica	sì	no
NP_207071	x	x	hypothetical protein HP0273	Ipotetica	sì	no
NP_207087	x		vacuolating cytotoxin (VacA)-like protein	Virulenza	sì	sì
NP_207106	x		hypothetical protein HP0308	Ipotetica	sì	no
NP_207120	x	x	poly E-rich protein	Altro	sì	sì
NP_207122	x	x	hypothetical protein HP0324/HorC	OMP	sì	no
NP_207184	x	x	hypothetical protein Hac_0422	Ipotetica	sì	no
NP_207270	x		hypothetical protein HP0472/HorE	OMP	sì	no
NP_207283	x	x	hypothetical protein HP0486/HofC	OMP	sì	no

NP_207391	x	x	tumor necrosis factor alpha-inducing protein	Virulenza	sì	no
NP_207403	x	x	hypothetical protein HP0608	Ipotetica	sì	sì
NP_207496	x		hypothetical protein HP0702	Ipotetica	sì	sì
NP_207520	x	x	hypothetical protein HP0726/OMP	OMP	sì	sì
NP_207547	x		hypothetical protein HCW_04825	Ipotetica	sì	sì
NP_207555	x	x	hypothetical protein HP0762	Ipotetica	sì	sì
NP_207575	x	x	hypothetical protein HP0782/hof OMP	OMP	sì	no
NP_207576	x	x	hypothetical protein HP0783	Ipotetica	sì	no
NP_207581	x	x	hypothetical protein HP0788/hof OMP	OMP	sì	sì
NP_207590	x	x	neuraminyllactose-binding hemagglutinin (hpaA?)	Parete	sì	no
NP_207610	x	x	hypothetical protein HP0817	Ipotetica	sì	no
NP_207662	x	x	membrane-associated phospholipid phosphatase	Metabolismo lipidi	sì	sì
NP_207668	x	x	KapA	Altro	sì	sì
NP_207706	x	x	hypothetical protein HP0914/hof OMP	OMP	sì	no
NP_207714	x	x	toxin-like outer membrane protein/putative VacA	Virulenza	sì	sì

NP_207805	x		hypothetical protein HCW_03690	Ipotetica	sì	no
NP_207846	x	x	3'-to-5' oligoribonuclease A	Nucleasi	sì	sì
NP_207847	x	x	hypothetical protein HP1056	Ipotetica	sì	sì
NP_207848	x	x	hypothetical protein HP1057	Ipotetica	sì	sì
NP_207857	x	x	hypothetical protein HP1066/OMP	OMP	sì	no
NP_207872	x	x	hypothetical protein HP1081	Ipotetica	sì	no
NP_207876	x	x	hypothetical protein HP1085	Ipotetica	sì	sì
NP_207913	x	x	hypothetical protein HCW_08355	Ipotetica	sì	sì
NP_208009	x		hypothetical protein HCW_03210	Ipotetica	sì	no
NP_208042	x	x	hypothetical protein HBZC1_00360	Ipotetica	sì	no
NP_208119	x	x	hypothetical protein HP1327	Ipotetica	sì	sì
NP_208125	x		hypothetical protein Hac_0263	Ipotetica	sì	no
NP_208186	x	x	hypothetical protein HP1395/HorL	OMP	sì	no
NP_208253		x	flagellar motility protein	Motilità	sì	sì
NP_208258	x	x	hypothetical protein HP1467/OMP	OMP	sì	sì
NP_208260	x	x	outer membrane protein 32	OMP	sì	no
NP_208292	x	x	hypothetical protein	OMP	sì	no

			HP1501/HopK			
NP_208302	x	x	hypothetical protein Hac_0073	Ipotetica	sì	no
NP_208314	x		hypothetical protein Hac_0005	Ipotetica	sì	no
NP_208315	x	x	hypothetical protein HP1525	Ipotetica	sì	no
NP_208317	x	x	hypothetical protein HBZC1_05490	Ipotetica	sì	no
NP_208321		x	hypothetical protein HP1531	Ipotetica	sì	no
NP_208357	x	x	hypothetical protein HP1566	Ipotetica	sì	no

Tabella Supplementare S5.2 *Geni selettivamente conservati in* H. pylori *"Full" e "Ssd" indicano i rispettivi filtri adottati (vedi testo); una x indica che il gene ha passato quel filtro sui pattern di presenza. Gli accession number sono relativi ad* H. pylori. *"Presente in Haci" indica se il gene è presente in* H. acinonychis.

Accession Number	Descrizioni	Classificazione	Full	Ssd	Presente in Haci?
NP_207058	adenine-specific DNA methyltransferase	Modificazioni DNA	x	x	sì
NP_207317	cag pathogenicity island protein cag1	isola cag	x	x	no
NP_207318	cag pathogenicity island protein cag3	isola cag	x	x	no
NP_207319	cag pathogenicity island protein cag4	isola cag	x	x	no
NP_207320	cag pathogenicity island protein cag5	isola cag	x	x	no
NP_207322	cag pathogenicity island protein cag6	isola cag	x	x	no
NP_207323	cag pathogenicity island protein cag7	isola cag		x	no
NP_207324	cag pathogenicity island protein cag8	isola cag	x	x	no
NP_207325	cag pathogenicity island protein cag9	isola cag	x	x	no
NP_207326	cag pathogenicity island protein cag10	isola cag	x	x	no
NP_207327	cag pathogenicity island protein cag11	isola cag	x	x	no
NP_207328	cag pathogenicity island protein cag12	isola cag	x	x	no
NP_207330	cag pathogenicity island protein cag13	isola cag	x	x	no
NP_207333	cag pathogenicity island protein cag16	isola cag	x	x	no
NP_207334	cag pathogenicity island protein cag17	isola cag	x	x	no
NP_207335	cag pathogenicity island protein cag18	isola cag	x	x	no
NP_207336	cag pathogenicity island protein cag19	isola cag	x	x	no
NP_207337	cag pathogenicity island protein cag20	isola cag	x	x	no

NP_207338	cag pathogenicity island protein cag21	isola cag	x	x	no
NP_207339	cag pathogenicity island protein cag22	isola cag	x	x	no
NP_207340	cag pathogenicity island protein cag23	isola cag	x	x	no
NP_207341	cag pathogenicity island protein cag24	isola cag	x	x	no
NP_207342	cag pathogenicity island protein cag25	isola cag	x	x	no
NP_207343	cag pathogenicity island protein cag26	isola cag	x	x	no
YP_008682876	cag pathogenicity island protein	isola cag	x	x	no
NP_207926	ATP synthase F0F1 subunit delta	produzione di energia		x	sì
NP_207260	type I restriction-modification enzyme, SÌ subunit	enzima di restrizione	x		sì
NP_207261	type I restriction enzyme M protein HsdM	enzima di restrizione	x	x	sì
NP_207262	type I restriction enzyme R protein HsdR	enzima di restrizione	x	x	sì
NP_207701	restriction endonuclease	enzima di restrizione	x	x	no
NP_208162	type III restriction enzyme R protein	enzima di restrizione	x	x	sì
NP_206827	hypothetical protein HP0025	Ipotetica	x	x	no
NP_206830	hypothetical protein HP0028	Ipotetica		x	sì
NP_206832	hypothetical protein HP0030	Ipotetica	x	x	sì
NP_206859	hypothetical protein HP0059	Ipotetica	x	x	no
NP_206861	hypothetical protein HP0061	Ipotetica	x	x	sì

NP_206885	hypothetical protein HP0085	Ipotetica	x	x	sì
NP_206947	hypothetical protein HP0148	Ipotetica		x	sì
NP_206988	hypothetical protein HP0189	Ipotetica	x	x	sì
NP_207009	hypothetical protein HP0211	Ipotetica	x	x	no
NP_207027	hypothetical protein HP0229	Ipotetica	x	x	no
NP_207033	hypothetical protein Hac_1383	Ipotetica		x	sì
NP_207059	hypothetical protein HP0261	Ipotetica	x	x	sì
NP_207107	hypothetical protein HP0309	Ipotetica	x	x	sì
NP_207109	hypothetical protein HP0311	Ipotetica	x	x	sì
NP_207135	hypothetical protein HP0337	Ipotetica	x	x	sì
NP_207136	hypothetical protein HP0338	Ipotetica	x	x	sì
NP_207171	hypothetical protein HP0373	Ipotetica	x		sì
NP_207182	hypothetical protein Hac_0471	Ipotetica	x	x	sì
NP_207277	hypothetical protein HP0479	Ipotetica	x	x	sì
NP_207351	hypothetical protein HP0556	Ipotetica	x	x	sì
NP_207374	hypothetical protein Hac_1433	Ipotetica	x	x	sì
NP_207409	hypothetical protein HP0614	Ipotetica	x	x	sì
NP_207423	hypothetical protein Hac_0732	Ipotetica	x	x	sì
NP_207432	hypothetical protein HP0638	Ipotetica	x	x	sì
NP_207461	hypothetical protein HP0667	Ipotetica	x	x	sì
NP_207475	hypothetical protein HP0681	Ipotetica	x	x	sì
NP_207491	hypothetical protein HP0697	Ipotetica		x	sì

NP_207503	hypothetical protein HP0709	Ipotetica		x	sì
NP_207523	hypothetical protein HP0729/ATP /GTP binding protein	Ipotetica	x	x	sì
NP_207524	hypothetical protein HP0730	Ipotetica	x	x	sì
NP_207557	hypothetical protein HP0764	Ipotetica		x	sì
NP_207573	hypothetical protein HP0780	Ipotetica	x	x	no
NP_207673	hypothetical protein HP0879	Ipotetica	x	x	sì
NP_207689	hypothetical protein HP0896	Ipotetica	x	x	no
NP_207727	hypothetical protein HP0935	Ipotetica	x	x	no
NP_207739	hypothetical protein HP0947	Ipotetica		x	sì
NP_207755	hypothetical protein HP0963	Ipotetica	x	x	sì
NP_207871	hypothetical protein HP1080	Ipotetica		x	sì
NP_207968	hypothetical protein HP1177	Ipotetica	x	x	no
NP_208035	hypothetical protein HP1243	Ipotetica	x	x	no
NP_208114	hypothetical protein HP1322	Ipotetica		x	sì
NP_208143	hypothetical protein HP1351	Ipotetica	x	x	no
NP_208173	hypothetical protein HP1382	Ipotetica	x	x	no
NP_208203	hypothetical protein HP1412	Ipotetica	x	x	sì
NP_208228	hypothetical protein HP1437	Ipotetica	x	x	no
NP_208244	hypothetical protein HP1453	Ipotetica	x	x	no
NP_207092	acylamide amidohydrolase	Metabolismo aminoacidi	x	x	sì
NP_207110	ATP-binding protein/cobalamin synthesis protein	Metabolismo	x	x	sì

		cofattori			
NP_207638	hydroxyethylthiazole kinase	Metabolismo cofattori	x	x	no
NP_207665	CDP-diacylglycerol pyrophosphatase	Metabolismo lipidi	x	x	sì
NP_206904	2',3'-cyclic-nucleotide 2'-phosphodiesterase	Metabolismo basi azotate	x	x	sì
NP_207490	methylhydantoinase	Metabolismo basi azotate		x	sì
NP_207648	guanosine 5'-monophosphate oxidoreductase	Metabolismo basi azotate	x	x	sì
NP_207969	purine-nucleoside phosphorylase DeoD	Metabolismo basi azotate		x	sì
NP_207970	phosphopentomutase	Metabolismo zuccheri		x	sì
NP_207983	flagellar motility protein	Motilità	x	x	no
NP_207589	hypothetical protein HP0796/OMP	OMP	x	x	sì
NP_206958	lipopolysaccharide 1,2-glucosyltransferase RfaJ	Parete	x	x	sì
NP_207108	polysaccharide deacetylase	Parete	x	x	sì
NP_208207	lipopolysaccharide 1,2-glucosyltransferase RfaJ	Parete	x	x	sì
YP_008682875	lipopolysaccharide biosynthesis protein	Parete	x	x	sì
NP_208079	TenA	Metabolismo cofattori	x	x	no
NP_207487	short-chain fatty acids transporter	Trasporto	x	x	sì
NP_208353	iron(III) ABC transporter substrate-binding protein	Trasporto	x	x	no

	CeuE				
YP_665124	multidrug ABC transporter permease	Trasporto		x	sì

Tabella Supplementare S5.1 *Geni selettivamente persi dagli* Helicobacter *gastrici. Gli accession number sono relativi ad* H. hepaticus*.*

Accession number	**Descrizioni**	**Gruppo**	**Presente in *H. mustelae***
NP_859543	hypothetical protein HH0012	Ipotetica	no
NP_859548	aspartate oxidase	Metabolismo aminoacidi	no
NP_859552	two-component sensor histidine kinase	Trasduzione segnale	no
NP_859553	two-component response regulator family protein	Trasduzione segnale	no
NP_859572	hypothetical protein HH0041	Ipotetica	no
NP_859579	molybdopterin-guanine dinucleotide biosynthesis protein MobB	Metabolismo cofattori	no
NP_859584	hypothetical protein HH0053	Ipotetica	sì
NP_859631	hypothetical protein HH0100	Ipotetica	no
NP_859655	tRNA pseudouridine synthase B	Maturazione tRNA	sì
NP_859680	ATP phosphoribosyltransferase	Metabolismo basi azotate	sì
NP_859693	hypothetical protein HH0162	Ipotetica	sì
NP_859705	hypothetical protein HH0174	Ipotetica	no
NP_859706	hypothetical protein HH0175	Ipotetica	no
NP_859707	S-ribosylhomocysteinase	Metabolismo aminoacidi	no
NP_859710	acetylglutamate kinase	Ciclo dell'urea	sì
NP_859745	hypothetical protein HH0214	Ipotetica	no
NP_859759	hypothetical protein HH0228	Ipotetica	no
NP_859837	2-acyl-glycerophospho-ethanolamine acyltransferase	Metabolismo lipidi	no

NP_859859	hypothetical protein HH0328	Ipotetica	sì
NP_859909	rRNA methylase (SpoU class)	Maturazione rRNA	sì
NP_859950	gamma-glutamyl kinase	Ciclo dell'urea	no
NP_859980	phosphoribosyl-AMP cyclohydrolase/ phosphoribosyl-ATP pyrophosphohydrolase	Metabolismo basi azotate	sì
NP_859982	branched-chain amino acid aminotransferase	Metabolismo aminoacidi	sì
NP_859986	chemotaxis protein CheR	Chemotassi/Motilità	no
NP_859987	chemotaxis protein methylesterase CheB	Chemotassi/Motilità	no
NP_860014	bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase	Metabolismo basi azotate	sì
NP_860016	glutamine amidotransferase HisH	Metabolismo aminoacidi	sì
NP_860022	glucose-6-phosphate isomerase	metabolismo zuccheri	no
NP_860059	phosphoribosylaminoimidazole carboxylase catalytic subunit	Metabolismo basi azotate	sì
NP_860092	hypothetical protein HH0561	Ipotetica	no
NP_860099	hypothetical protein HH0568	Ipotetica	sì
NP_860136	hypothetical protein HH0605	Ipotetica	no
NP_860143	histidinol-phosphate aminotransferase	Metabolismo aminoacidi	sì
NP_860167	O-acetylhomoserine (thiol)-lyase MetY	Metabolismo aminoacidi	no
NP_860177	ferredoxin	Trasporto elettroni	sì
NP_860199	hypothetical protein HH0668	Ipotetica	sì
NP_860200	acetyl-gamma-glutamyl-phosphate reductase	Ciclo dell'urea	sì
NP_860224	tRNA (uracil-5-)-methyltransferase	Maturazione tRNA	sì

NP_860304	hypothetical protein HH0773	Ipotetica	no
NP_860337	gamma-glutamyl phosphate reductase	Ciclo dell'urea	no
NP_860357	1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino] imidazole-4-carboxamide isomerase	Metabolismo aminoacidi	sì
NP_860381	dihydroxy-acid dehydratase	Metabolismo aminoacidi	no
NP_860382	hypothetical protein HH0851	Ipotetica	no
NP_860383	5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase	Metabolismo aminoacidi	no
NP_860394	acetolactate synthase 3 regulatory subunit	Metabolismo aminoacidi	no
NP_860395	acetolactate synthase 3 catalytic subunit	Metabolismo aminoacidi	no
NP_860427	threonine dehydratase	Metabolismo aminoacidi	no
NP_860451	hypothetical protein HH0920	Ipotetica	no
NP_860473	transcriptional regulator	Regolazione trascrizione	sì
NP_860486	ATP phosphoribosyltransferase	Metabolismo basi azotate	sì
NP_860500	acetylornithine aminotransferase	Ciclo dell'urea	sì
NP_860504	hypothetical protein HH0973	Ipotetica	no
NP_860508	ornithine carbamoyltransferase	Metabolismo aminoacidi	sì
NP_860512	hypothetical protein HH0981	Ipotetica	no
NP_860610	hypothetical protein HH1079	Ipotetica	sì
NP_860633	cytochrome bd quinol oxidase	Produzione di energia	no
NP_860634	cytochrome bd quinol oxidase	Produzione di energia	no

NP_860636	hypothetical protein HH1105	Ipotetica	no
NP_860646	hypothetical protein HH1115	Ipotetica	no
NP_860647	hypothetical protein HH1116	Ipotetica	sì
NP_860648	hypothetical protein HH1117	Ipotetica	sì
NP_860707	cyclase	Altro	sì
NP_860727	isocitrate dehydrogenase	Metabolismo acidi carbossilici	sì
NP_860735	ketol-acid reductoisomerase	Metabolismo aminoacidi	no
NP_860752	argininosuccinate synthase	Ciclo dell'urea	sì
NP_860757	hypothetical protein HH1226	Ipotetica	no
NP_860767	DnaK suppressor protein DskA	Regolazione trascrizione	sì
NP_860815	hypothetical protein HH1284	Ipotetica	no
NP_860817	hypothetical protein HH1286	Ipotetica	sì
NP_860822	bifunctional ornithine acetyltransferase/acetylglutamate synthase	Ciclo dell'urea	sì
NP_860856	imidazoleglycerol-phosphate dehydratase	Metabolismo aminoacidi	sì
NP_860861	acetyltransferase	Altro	sì
NP_860947	methionine sulfoxide reductase B	Metabolismo aminoacidi	sì
NP_861034	hypothetical protein HH1503	Ipotetica	no
NP_861035	hypothetical protein HH1504	Ipotetica	no
NP_861047	alanine racemase	Metabolismo aminoacidi	sì
NP_861091	hypothetical protein HH1560	Ipotetica	sì
NP_861102	malate dehydrogenase	Metabolismo acidi	no

		carbossilici	
NP_861151	hypothetical protein HH1620	Ipotetica	sì
NP_861216	hypothetical protein HH1685	Ipotetica	sì
NP_861227	hypothetical protein HH1696	Ipotetica	no
NP_861240	hypothetical protein HH1709	Ipotetica	sì
NP_861245	hypothetical protein HH1714	Ipotetica	sì
NP_861249	argininosuccinate lyase	Ciclo dell'urea	sì
NP_861252	histidinol dehydrogenase	Metabolismo aminoacidi	sì
NP_861280	aspartate aminotransferase	Metabolismo aminoacidi	no
NP_861289	hypothetical protein HH1758	Ipotetica	sì
NP_861290	ABC transporter	Trasporto	sì
NP_861291	hypothetical protein HH1760	Ipotetica	sì
NP_861356	phosphoribosylformylglycinamidine synthase II	Metabolismo basi azotate	sì
NP_861360	arginyl-tRNA-protein transferase	Modificazioni post-traduzionali	sì
NP_861375	4-methyl-5(beta-hydroxyethyl)-thiazole monophosphate synthesis protein ThiJ	Metabolismo cofattori	sì
NP_861380	acetyl-CoA carboxylase subunit A	Metabolismo lipidi	no
NP_861388	ABC transporter	Trasporto	no
NP_861389	hypothetical protein HH1858	Ipotetica	no
NP_861390	hypothetical protein HH1859	Ipotetica	no
NP_861399	C4-dicarboxylate binding periplasmic protein	Trasporto	no

NP_861400	hypothetical protein HH1869	Ipotetica	no
NP_861401	hypothetical protein HH1870	Ipotetica	no

Tabella Supplementare S6.1 *Link a depositi online contenenti i risultati completi e gli script.*

Link	Contenuto
https://www.dropbox.com/sh/cttrxs2en0fi4kx/AAAwUB_vr9PnijP573o_ZKhIa?dl=0	Risultati Completi
https://www.dropbox.com/sh/1k0g2yo39s02p50/AAAY3sWdvLEDyCs3tt5d2nN6a?dl=0	Script finding_orthologs.sh e procedure correlate su Dropbox
https://github.com/Pietro-Cravedi/find_orthologs	Script finding_orthologs.sh e procedure correlate su GitHub