Thaís Cabrera Galvão Rojas

"DETECÇÃO DE GENES SOB SELEÇÃO POSITIVA EM LINHAGENS DE *ESCHERICHIA COLI* PATOGÊNICAS PARA AVES (APEC) E PARA HUMANOS"

Campinas, 2012

1

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

SECRETARIA DE PÓS-GRADUAÇÃO I. B.

THAÍS CABRERA GALVÃO ROJAS

"DETECÇÃO DE GENES SOB SELEÇÃO POSITIVA EM LINHAGENS DE *ESCHERICHIA COLI* PATOGÊNICAS PARA AVES (APEC) E PARA HUMANOS"

Este exemplar corresponde à redação final da tese defendida pelo(a) candidato (a) THAÍS CABRERA GALVÃO ROJAS

e aprovada pela Comissão Julgadora.

Tese apresentada ao Instituto de Biologia para obtenção do Título de Doutor em Genética e Biologia Molecular, na área de Microbiologia.

Orientador: Prof. Dr. Wanderley Dias da Silveira

Campinas, 2012

FICHA CATALOGRÁFICA ELABORADA POR
ROBERTA CRISTINA DAL' EVEDOVE TARTAROTTI – CRB8/7430
BIBLIOTECA DO INSTITUTO DE BIOLOGIA - UNICAMP

R638d

Rojas, Thaís Cabrera Galvão, 1980-
Detecção de genes sob seleção positiva em linhagens de *Escherichia coli* patogênicas para aves (APEC) e para humanos / Thaís Cabrera Galvão Rojas. – Campinas, SP: [s.n.], 2012.

Orientador: Wanderley Dias da Silveira.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. *Escherichia coli* patogênica aviária. 2. Patogenicidade. 3. Genes. 4. Seleção positiva. I. Silveira, Wanderley Dias da, 1956-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em Inglês: Detection of genes under positive selection in Avian Pathogenic Escherichia coli (APEC) and humans pathogenic strains
Palavras-chave em Inglês:
Avian pathogenic *Escherichia coli*
Pathogenicity
Genes
Positive selection
Área de concentração: Microbiologia
Titulação: Doutora em Genética e Biologia Molecular
Banca examinadora:
Wanderley Dias da Silveira [Orientador]
Marcelo Brocchi
Fabiana Fantinatti Garboggini
Renata Katsuko Takayama Kobayashi
Rodrigo Tavanello Hernandes
Data da defesa: 15-10-2012
Programa de Pós Graduação: Genética e Biologia Molecular

Campinas, 15 de outubro de 2012

BANCA EXAMINADORA

Prof. Dr. Wanderley Dias da Silveira
(Orientador)

Assinatura

Prof. Dr. Marcelo Brocchi

Assinatura

Profª. Drª. Fabiana Fantinatti-Garboggini

Assinatura

Profª. Drª. Sueli dos Santos Freitas

Assinatura

Prof. Dr. Rodrigo Tavanelli Hernandes

Assinatura

Profª. Drª. Clarice Weis Arns

Assinatura

Prof. Dr. Tomomasa Yano

Assinatura

Profª. Drª. Eliana Guedes Stehling

Assinatura

Agradecimentos

Agradeço ao meu orientador Prof. Dr. Wanderley Dias da Silveira, pela oportunidade, confiança, compreensão e ensinamentos.

Aos meus amigos companheiros do Laboratório de Biologia Molecular Bacteriana: Cristiane Mobilon, Fernanda de Pace, Telma Piacese, Murillo Lino Bution, Monique Ribeiro Tiba, Lívia Pilatti Mendes da Silva, Flávio Silveira, Renato Pariz Maluta, Ricardo Destéfano, Izildinha Coli e em especial à Jacqueline Boldrin de Paiva e à Janaína Luíza Leite Garbim, pela grande amizade e ajuda.

Ao Lucas Pedersen Parizzi e ao Prof. Dr. Gonçalo Amarante Guimarães Pereira pela montagem dos genomas.

Ao Pesquisador Dr. Francisco Pereira Lobo e ao Laboratório Multiusuário de Bioinformática da Embrapa Informática Agropecuária, pela colaboração nas análises de seleção positiva.

A todos meus amigos, pelo apoio e companheirismo.

A minha enorme família, e aos meus três pais que sempre me amaram muito e possibilitaram que eu chegasse até aqui.

A minha filha amada, por ter transformado minha vida numa festa e me mostrar o que é a felicidade.

Ao meu amado esposo, que sempre permanece firme ao meu lado, pelo apoio incondicional e pelas muitas horas de trabalho.

Ao programa de Pós-Graduação em Genética e Biologia Molecular– IB – UNICAMP.

Ao CNPq (142718/2010-6) pela bolsa de estudo.

Dedico este trabalho ao meu amado Luciano, que nunca me deixou desistir e sempre me guiou pelo caminho certo.

Resumo

A bactéria *Escherichia coli* coloniza o trato intestinal de aves e humanos, de maneira comensal sem causar processos infecciosos. No entanto alguns clones adquiriram fatores de virulência específicos, permitindo o desenvolvimento de diferentes doenças como infecção do trato urinário, diarreia e meningite em humanos e colibacilose em aves. As linhagens que causam doença em aves são tipicamente denominadas APEC (*Avian Pathogenic Escherichia coli*). Neste trabalho foram sequenciados e anotados os genomas de quatro linhagens APECs (SCI-07, SEPT362, S17 e O8) que, juntamente com mais nove genomas referentes a linhagens de *Escherichia coli* patogênicas para aves e patogênicas para humanos foram utilizados para a busca de genes sob seleção positiva. Os genes homólogos foram agrupados, e posteriormente submetidos ao alinhamento de códons e das sequencias proteicas correspondentes. Uma árvore filogenética foi gerada para cada grupo de proteínas homólogas. Testes estatísticos determinaram qual entre os modelos de seleção neutra ou seleção positiva melhor explicou os dados existentes (alinhamentos de códons e árvores filogenéticas). Essas análises detectaram duzentos e cinquenta e quatro grupos de genes homólogos com evidência de seleção positiva. Para cada grupo foi realizado um teste de recombinação para verificar se o aumento na variação das sequencias não era devido à conversão gênica, resultando em cento e dezesseis grupos de genes homólogos sob seleção positiva. A proteína correspondente a um gene de cada grupo de genes homólogos foi identificada, por meio da ferramenta Blast. Diversos fatores de virulência, já conhecidos, e proteínas regulatórias puderem ser detectados. Os genes sob seleção positiva, também foram submetidos à anotação considerando o termo GO (Gene Ontology), apenas da categoria processo

biológico. Dos cento e dezesseis genes apenas cinquenta e sete puderam ser identificados por meio dessa metodologia. O resultado da classificação dos genes dentro da classe GO, considerando o terceiro nível hierárquico, mostrou que a maioria dos genes anotados (31) tinha relação com o metabolismo primário. As proteínas cuja identificação, por meio do blast, não foi possível (proteínas hipotéticas) foram submetidas à análise de predição de localização subcelular e de peptídeo sinal. Essas análises revelaram que três proteínas desconhecidas (*hypothetical protein* ECIAI39_1028, *hypothetical protein* Z0639 e *hypothetical protein* EC042_3791) são potenciais alvos para estudos que visam à busca de novos fatores de virulência de *Escherichia coli* patogênicas.

Palavras- chave: *Escherichia coli*, APEC, genes homólogos e seleção positiva.

Abstract

The bacterium *Escherichia coli* colonizes the intestinal tract of birds and humans, in a commensal relationship without causing infection. However, some clones have acquired specific virulence factors allowing the development of various diseases such as urinary tract infection, diarrhea and meningitis in humans and colibacillosis in poultry. The strains that cause disease in birds are typically named APEC (Avian Pathogenic *Escherichia coli*). In this study we sequenced and annotated the genomes of four APECs strains (SCI-07, SEPT362, S17 and O8). These genomes and nine others avian pathogenic *Escherichia coli* and humans pathogenic strains genomes were used for studying genes under positive selection. The homologous genes were grouped and then subjected to codons and corresponding protein sequences alignment. A phylogenetic tree was generated for each group of homologous proteins. Statistical tests determined which among neutral or positive selection models best explains the existing data (codon alignments and phylogenetic trees). These analyzes detected two hundred fifty-four groups of homologous genes with positive selection evidence. For each group a recombination test was conducted to verify if the variation increase in the sequences was not due to gene conversion, resulting in one hundred and sixteen groups of homologous genes under positive selection. The protein corresponding to a gene of each group of homologous genes under positive selection was identified through Blast tool. Genes under positive selection were annotated considering the GO term (Gene Ontology), just for the biological process category. Only fifty-seven genes could be identified using this methodology. The gene classification within the GO classes, considering only the third hierarchical level showed that most of the annotated genes (31)

were related with the primary metabolism. Proteins which blast identification was not possible (hypothetical proteins) were subjected to subcellular localization and signal peptide prediction analyzes. These analyzes revealed that three unknown proteins (hypothetical protein ECIAI39_1028, hypothetical protein Z0639 e hypothetical protein EC042_3791) are potential targets for studies, in order to search for new virulence factors of pathogenic *Escherichia coli*.

Key words: *Escherichia coli*, APEC, homologous genes and positive selection.

Sumário

Resumo.....6

Abstract.....8

1. Introdução.....12

1.1 Colibacilose em frangos.....12

1.2 *Escherichia coli* patogênica para humanos.....16

1.3 Bioinformática.....20

1.4 Seleção positiva.....21

2. Objetivos.....28

2.1 Objetivo geral.....28

2.2 Objetivos específicos.....28

3. Material e Métodos.....29

3.1 Linhagens sequenciadas.....29

3.2 Sequenciamentos dos genomas e anotação.....29

3.3 Linhagens utilizadas nas análises de seleção positiva.....30

3.4 Detecção dos grupos de genes homólogos sob seleção positiva.....31

3.4.1 Grupo de homólogos.....31

3.4.2 Alinhamentos.....32

3.4.3 Análise de seleção positiva.....32

3.4.4 Recombinação.....33

3.4.5 Identificação das proteínas.....34

3.4.6 Processamento das análises.....34

3.5 Classificação dos genes baseada em termos de ontologia gênica (GO).....36

3.6 Análise das proteínas hipotéticas sob seleção positiva.....36

4. Resultados.....37

5. Discussão.....56

6. Conclusões.....73

7. Referências bibliográficas.....74

1. Introdução

1.1 Colibacilose em frangos

Atualmente, o Brasil é o maior exportador de carne de frango e o terceiro maior produtor do mundo (BRAZILIANCHIKEN, 2012), sendo, por isso, a indústria avícola importante para as economias interna e exportadora, gerando empregos, alimentos de alto valor proteico, com relativo baixo custo e gerando divisas em moedas fortes (dólares e euros). Apesar de ser um importante segmento econômico, a indústria avícola está sujeita a fatores que podem afetar significativamente seus rendimentos e custos. Destacam-se as infecções causadas por micro-organismos patogênicos (vírus, bactérias e fungos). Entre estes micro-organismos, as bactérias, principalmente *Escherichia coli*, são agentes significativos de morbidade e mortalidade (GROSS, 1991).

Escherichia coli é um bacilo Gram-negativo, anaeróbio facultativo, pertencente à família Enterobacteriaceae que, tipicamente, coloniza o trato gastrointestinal de aves (SOKJA & CARNAGHAM, 1961) e mamíferos (DRASAR & HILL, 1974; SOKJA, 1965). Linhagens não patogênicas desta bactéria, geralmente, permanecem confinadas ao lúmen intestinal de maneira comensal entretanto, em hospedeiros debilitados ou imunossuprimidos, ou quando as barreiras gastrointestinais são violadas, podem causar infecções. Além disso, existem linhagens originárias de clones patogênicos altamente adaptados que são capazes de causar doenças em seres saudáveis (NATARO & KAPER, 1998).

As linhagens patogênicas de *E. coli* relacionadas a doenças em aves são coletivamente

denominadas APEC (*Avian Pathogenic Escherichia coli*), podendo ser classificadas de acordo com os principais sorogrupos O associados às doenças, especialmente os sorogrupos O1, O2, O8, O15, O18, O35, O78, O88, O109 (EWERS *et al.,* 2004; KNÖBL *et al.,* 2001; MCPEAKE *et al.,* 2005; NGELEKA *et al.,* 2002).

Linhagens patogênicas para aves estão associada a infecções extra-intestinais, sendo o termo colibacilose empregado para englobar o grande número de doenças existentes, dentre as quais: septicemia, celulite, pericardite, perihepatite, onfalite, peritonite, infecções do trato respiratório e da gema do ovo, e síndrome da cabeça inchada. Acredita-se que, exceto para as infecções da gema do ovo, todas estas infecções requerem fatores predisponentes que incluem condições ambientais e infecções primárias, principalmente virais, o que faz com que se acredite que linhagens APEC atuem como agentes patogênicos secundários (DHO-MOULIN & FAIRBROTHER, 1999).

Um importante fator para o estabelecimento, colonização e posterior infecção por *E. coli* é a adesão dos bacilos aos tecidos epiteliais do hospedeiro, a qual pode ser facilitada por adesinas (MOON, 1990). Em linhagens APEC, a fímbria tipo I é amplamente encontrada estando relacionada à adesão ao trato respiratório superior. Ela seria necessária para a colonização inicial do epitélio respiratório, enquanto outros fatores como motilidade e a produção de colicina V seriam responsáveis pela manutenção da infecção e pelo desenvolvimento de lesão na traqueia (WOOLEY *et al*.,1998).

O papel da adesina fimbrial P ainda não foi completamente esclarecido como fator de virulência em APEC (NAKAZATO *et al*., 2009). Ela é codificada pelo operon *pap*, um grupo de

11 genes, o qual está localizado no cromossomo bacteriano (CAMPOS *et al.,* 2005; LATHAM & STAMM, 1994). Análises feitas por Kariyawasam e colaboradores (2006) demostraram que esse operon foi encontrado na linhagem APEC- O1, causadora de colicepticemia, em uma região do genoma portadora de genes de patogenicidade horizontalmente adquiridas, ou seja, ilha de patogenicidade (PAI). Assim, os autores sugerem que a presença do operon *pap* em uma ilha de patogenicidade pode demonstrar que a fimbria P, possivelmente, estaria envolvida na transformação de linhagens não virulentas em linhagens virulentas.

A fímbria *curli* é uma estrutura filamentosa, fina e enrolada, encontrada na superfície de *E. coli* e *Salmonella spp.* (COLLINSON *et al*., 1993; OLSÉN *et al*., 1989). Essa estrutura auxiliaria a ligação da bactéria à matriz extracelular e a outras proteína do hospedeiro, como fibronectina, por exemplo (OLSÉN *et al*., 1989). As propriedades de aderência da fimbria *curli* sugerem que ela exerça um papel na ligação e colonização da bactéria durante o estágio inicial da infecção (DHO-MOULIN & FAIRBROTHER, 1999).

Existe uma serino-protease autotransportadora, secretada por APEC, que causa a aglutinação dos eritrócitos de aves. A expressão desta proteína é regulada pela temperatura, sendo ativada a 26 ºC e inativada a 42 ºC, sendo, por isso, denominada Tsh (*temperature sensitive haemagglutinin*) (PROVENCE & CURTISS III, 1994). O gene *tsh* está localizado em plasmídeos de alto peso molecular, principalmente no plasmídeo ColV (JOHNSON *et al*., 2006; STEHLING *et al*., 2003). Por esse gene ser, frequentemente, encontrado em APEC, Ewers e colaboradores (2004) propuseram a sua utilização como marcador molecular de APEC.

Há poucos relatos que estabelecem que a produção de toxinas por APEC possa estar relacionada ao processo de patogenicidade (DHO-MOULIN & FAIRBROTHER, 1999). No

entanto, algumas linhagens produzem enterotoxinas LT (*labile temperature*) e ST (*stable temperature*) (SMITH & GYLES, 1969). Também com efeito tóxico, a proteína autotransportadora Vat (*vacuolating autotransporter toxin*) que foi localizada dentro de uma ilha de patogenicidade apresenta 75% de homologia com Tsh e induz a formação de vacúolos intracelulares causando um efeito citotóxico semelhante ao causado pela toxina VacA de *Helicobacter pylori* (PARREIRA & GYLES, 2003).

A resistência das bactérias aos efeitos bactericidas do sistema complemento do soro do hospedeiro é mediada por estruturas da superfície da bactéria tais como, cápsula, lipopolissacarídeos, produtos do plasmídeo ColV e outras proteínas de membrana (GROOS, 1994; WOOLEY *et al*., 1998). O fator Iss (*increased serum survival*), codificado pelo gene *iss* está fortemente associado à patogenicidade de APEC, estando, significativamente, mais presente nestas linhagens do que em linhagens de *E. coli* isoladas de aves saudáveis (PFAFF-MCDONOUGH *et al*., 2000; RODRIGUEZ-SIEK *et al*., 2005). O gene *iss,* à semelhança do gene *tsh,* está localizado em uma ilha de patogenicidade (PAI) do plasmídeo ColV/BM (JOHNSON *et al.*, 2006).

Devido à baixa concentração de ferro livre nos fluídos animais, várias bactérias patogênicas possuem um sistema de aquisição de ferro, para sobreviver no hospedeiro. Esse sistema compete com o ferro e transferrinas do hospedeiro (ANDRADE *et al*., 2000; DHO-MOULIN & FAIRBROTHER, 1999). Assim, Dho & Lafont (1984) observaram uma correlação positiva entre a baixa concentração de ferro e a capacidade de crescimento de APEC e à letalidade a frangos de um dia de idade, infectados com linhagem patogênica destas bactérias. Em *E. coli*, sistemas de aquisição de ferro são codificados por genes localizados no plasmídeo ColV (JONHSON *et al.,*

2006; SABRI *et al*., 2006) e em ilha de patogenicidade presente no cromossomo (KARIYAWASAM *et al.*, 2006).

Alguns estudos demonstram que linhagens APEC e linhagens de *E. coli* patogênicas extra-intestinais de origem humana (ExPEC) compartilham características de patogenicidade. Em um trabalho realizado por Rodriguez-Siek e colaboradores (2005) demonstrou-se que linhagens de colibacilose aviária e de *E. coli* isoladas de infecções urinárias de humanos (UTI) apresentaram sobreposição de características como: sorogrupos, grupos filogenéticos e genótipos de virulência incluindo sequências relacionadas a plasmídeos, à adesão bacteriana, sistemas de captação de ferro e toxinas. Johnson e colaboradores (2003) isolaram linhagens do produto retal da cloaca de galinhas que apresentaram perfil de virulência e antígenos O semelhantes aos encontrados em linhagens de *E. coli* humanas associadas a diferentes infecções extra-intestinais. Esses estudos sugerem que linhagens APEC e ExPEC humanas podem encontrar desafios similares no momento do estabelecimento da infecção e que compartilham conteúdos de genes de virulência e capacidade patogênica semelhantes. Essas similaridades podem indicar que APEC poderia servir como causa de infecção ou como reservatório de genes de virulência para linhagens ExPEC humanas, constituindo um possível risco zoonótico.

1.2 *Escherichia coli* patogênica para humanos

A bactéria *Escherichia coli* possui uma relação comensal com seres humanos. Esta bactéria coloniza a camada de muco do trato intestinal de pessoas saudáveis as quais podem ter mais de um bilhão de células de *E. coli* em seu intestino (SAVAGEAU, 1983). No entanto, alguns clones adquiriram fatores de virulência específicos, que possibilitaram a adaptação a novos nichos,

permitindo que os mesmos causem diferentes doenças, como por exemplo: infecção do trato urinário, pneumonia, septicemia e meningite, além de diarreia (DONNENBERG, 2002; KAPER *et al*., 2004). Esses fatores de virulência que, anteriormente, eram elementos móveis, acabaram fixados nos genomas, originando diferentes patótipos de *E. coli* capazes de causar doenças em organismo saudáveis (KAPER *et al*., 2004).

Essa bactéria é o principal agente de morbidade e mortalidade no mundo, sendo que mais de dois milhões de crianças morrem todo ano devido à diarreia (KOSEK *et al*., 20003) e outras doenças extraintestinais (RUSSO & JOHNSON, 2003).

Linhagens de *E. coli* enterohemorrágicas (EHEC) podem colonizar o íleo e a parte distal do intestino humano podendo causar surtos severos de gastroenterites. A contaminação por esse patovar normalmente se dá pela ingestão de comida e água contaminadas, sendo o gado o principal reservatório para essas linhagens (CROXEN & FINLAY, 2010).

Quase todas as EHEC O157:H7 carregam o plasmídeo pO157 que codifica diversos genes de virulência, inclusive a shiga toxina, o fator de virulência mais importante (NATARO & KAPER, 1998). Nessas linhagens não existe maquinário de secreção para essa toxina, sua liberação é efetuada após a lise da célula bacteriana mediada pelo fago Lambda. Em infeções causadas por essas linhagens o tratamento com antibiótico não é indicado, já que a destruição das células bacterianas resulta na maior liberação da toxina (CROXEN & FINLAY, 2010; TOSHIMA *et al*., 2007), piorando o quadro clínico do paciente.

Os mecanismos utilizados no processo de adesão inicial das linhagens EHEC ainda não estão completamente elucidados, mas acredita-se que ela ocorra pela ligação do pili tipo IV às células do intestino (XICOHTENCATL-CORTES *et al*., 2009). Em seguida, a ação das intiminas resulta

na formação de um pedestal na superfície das células do hospedeiro e vários fatores de virulência são injetados (CROXEN & FINLAY, 2010).

Em linhagens de *E. coli* uropatogênicas (UPEC), a primeira etapa necessária para o estabelecimento da infecção é o deslocamento das bactérias a partir do trato intestinal até o trato urinário. Diversos fatores de virulência contribuem para esse processo, como a expressão de diversos pili, cápsula de polissacarídeo, sistemas de aquisição de ferro, secreção de toxinas (ex: Vat- *vacuolating autotransporter toxin*), entre outros (CROXEN & FINLAY, 2010). A ligação da bactéria às células uroteliais é mediada pela fímbria tipo I cuja ligação inicia uma via de sinalização para o processo de invasão e apoptose (THUMBIKAT *et al*., 2009). Dentro da célula do hospedeiro a bactéria rapidamente se multiplica, formando um complexo semelhante ao biofilme, causando danos ao tecido (ANDERSON *et al*., 2003). A morte de células teciduais do hospedeiro também pode ocorre pela liberação da toxina HlyA (*haemolysin A*) (WILES *et al*., 2008).

A contaminação do recém-nascido com *E. coli* causadora de meningite neonatal (NMEC), ocorre durante o parto. Após a colonização do intestino do recém-nascido, a bactéria alcança a corrente sanguínea e atinge o sistema nervoso central (CROXEN & FINLAY, 2010). A habilidade da bactéria sobreviver no sangue é conferida pela existência de uma cápsula (K1) antifagocítica. Adicionalmente, a proteína OmpA permite a resistência dessa bactéria ao sistema complemento do soro (WOOSTER *et al*., 2006).

A barreira hemato encefálica é formada pelas células endoteliais microvasculares cerebrais. A ligação da bactéria a essas células é mediada pela fímbria tipo I e pela proteína OmpA que se liga a receptores específicos (KHAN *et al*., 2007). O processo de invasão é mediado por essas

mesmas duas proteínas, além das proteínas IbeA, IbeB e IbeC (*invasion protein*) e CNF1 (*cytotoxic nectrotizing factor 1*) (KIM, 2008). Uma vez no sistema nervoso central a infecção resultará em edema, inflamação e dano encefálico (CROXEN & FINLAY, 2010).

Os mecanismos utilizados durante o estabelecimento da infecção bacteriana variam para cada linhagem, mas de maneira geral, podem incluir os processos de: adesão (podendo existir injeção de fatores), invasão (apenas para alguns patótipos), ativação de fatores de sobrevivência intracelular (ex: hemolisinas, toxinas), danos teciduais e doença. No entanto, dentro de uma mesma espécie existem diferenças nas características relacionadas à resistência a antibióticos, habilidade de utilizar um determinado açúcar como fonte de carbono, temperatura ótima de crescimento, presença de fatores de virulência, proteínas de adesão, presença de plasmídeos, tipo do sorogrupo, presença de sistemas de secreção, etc. Tais diferenças possibilitam que determinada linhagem possa causar certo tipo de doença apenas em um determinado órgão de um determinado organismo.

Apesar dessas especificações existentes para os diferentes patótipos, quando dentro de uma mesma espécie co-existem linhagens patogênicas e comensais, o risco de linhagens comensais adquirirem fatores de virulência e de linhagens patogênicas transferirem fatores de virulência é iminente. Assim, a comparação de genomas dentro de uma mesma espécie permite a realização de estudos sobre a modelação dos fenótipos, uma vez que as semelhanças permitem acessar alterações pontuais e o elevado número de genes homólogos possibilita o estudo da história e do processo evolucionário (PETERSEN *et al*., 2007).

1.3 Bioinformática

O sequenciamento de genomas é uma tecnologia relativamente recente na biologia molecular e microbiologia. O primeiro genoma bacteriano a ser sequenciado foi o de *Haemophilus influenzae*, com apenas 1,8 Mpb (FLEISCHMANN *et al.*, 1995). Em seguida, foram sequenciados os genomas da *Escherichia coli* (BLATTNER *et al*., 1997), Archaea *Archaeoglobus fulgidus* (KLENK *et al*., 1997) e do *Mycobacterium tuberculosis* (COLE *et al*., 1998). Genomas eucarióticos como o do *Plasmodium falciparum* (GARDNER *et al*., 2002a, GARDNER *et al*., 2002b, HALL *et al*., 2002, HYMAN *et al*., 2002), o genoma humano (LANDER *et al*., 2001) e do chimpanzé (CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM, 2005), também foram sequenciados. Estes estudos foram determinantes para a popularização do sequenciamento de genomas (HALL, 2007), sendo que em 2010, já existiam mais de 1500 genomas procarióticos sequenciados (NCBI, 2012a).

No ano de 2000, o Brasil iniciou a pesquisa genômica com o sequenciamento do fitopatógeno *Xylella fastidiosa* (SIMPSON *et al*., 2000). Desde então, o Brasil já sequenciou mais de uma centena de genomas (NCBI, 2012b), sendo a sua grande maioria referente a genomas de procariotos. No ano de 2012, foi publicado, pela 1º vez, o genoma de uma linhagem brasileira de *Escherichia coli* patogênica para aves (APEC) pelo grupo de pesquisa do Prof. Dr. Wanderley Dias da Silveira (ROJAS *et al*., 2012).

Com o surgimento dos sequenciadores automáticos de DNA, na década de 90, ocorreu um aumento bastante significativo na quantidade de dados que precisariam ser armazenados e analisados, resultando por uma demanda maior de recursos computacionais na rotina de estudos relacionados à genômica, surgindo assim a Bioinformática.

A Bioinformática, também conhecida como biologia computacional, pode ser definida como a aplicação de técnicas computacionais para estudar dados biológicos (HUGHEY *et al*., 2003), aplicando técnicas quantitativas e analíticas à modelação de sistemas biológicos (GIBAS & JAMBECK, 2001). O sequenciamento genômico permite a identificação e caracterização de genes e suas respectivas proteínas, o que possibilita à análise de sequências de DNA, expressão gênica e proteica, regulação da expressão gênica, predição de estrutura proteica, genômica comparativa (WEISS, 2010), relações evolutivas entre diferentes níveis taxonômicos, a identificação de elementos móveis e de genes adquiridos por transferência lateral (NIERMAN *et al*., 2000). Com esse novo campo do conhecimento é possível, agora, realizar estudos genômicos comparativos, apenas *in silico.*

Neste trabalho, ferramentas de bioinformática foram utilizadas para a detecção de grupos de genes homólogos sob seleção positiva de linhagens de *Escherichia coli* patogênicas para aves (APEC) e humanos (EHEC, NMEC e UPEC).

1.4 Seleção positiva

A evolução não é um processo totalmente determinista ou totalmente estocástico, ela sempre é influenciada por fatores como o tamanho da população, deriva genética, distribuição de espécies ou frequência alélica. No entanto, as mudanças genéticas, ou seja, as mutações ocorrem sempre de forma randômica e a pressão de seleção irá determinar se essa mutação será mantida ou não na população. Mutações não sinônimas que resultam em mudança de fenótipo, alterando a interação de um organismo com seu ambiente, estarão sob pressão de seleção. Já as mutações sinônimas são neutras e poderão ser fixadas na população, sendo que o aumento da frequência

deste alelo está diretamente relacionado ao tamanho da população. Uma mudança adaptativa que esteja sob pressão negativa pode até ser fixada em uma população, por deriva genética, se não possuir efeito deletério. No entanto, uma mutação que resulte em uma mudança adaptativa aumentando assim o sucesso reprodutivo, ou seja, sob pressão de seleção positiva, irá aumentar a sua frequência e ser fixada na população após poucas gerações (SALEMI & VANDAMME, 2003).

Na história evolutiva de diversos micro-organismos, a seleção positiva e a recombinação homóloga aturam como alguns dos fatores determinantes para a adaptação a novos nichos ecológicos. Ambos os fatores contribuem para a variabilidade genética entre organismos que acaba determinando a diversidade populacional e a adaptação dos micro-organismos patogênicos a diferentes grupos de hospedeiros (LEWIS-ROGERS *et al*., 2008; PETERSEN *et al*., 2007). A comparação de dados genômicos de diferentes espécies pode ser utilizada para a identificação de genes afetados pela seleção natural (PETERSEN *et al*., 2007).

O processo de co-evolução entre patógeno e hospedeiro atua favorecendo o surgimento de novos alelos relacionados à resistência e à virulência (CHISHOLM *et al*., 2006). O conhecimento dos genes que apresentam uma taxa de mutação maior tem aplicação prática, ajudando a desenvolver vacinas e drogas cujos alvos sejam regiões conservadas do genoma do patógeno (POLLEY *et al*., 2003), dessa forma, evita-se utilizar proteínas não conservadas ou então domínios não conservados dentro das proteínas imunogênicas durante o desenvolvimento dessas vacinas e drogas.

As mutações nas sequências de DNA de diferentes organismos são utilizadas, por diferentes métodos filogenéticos, para se calcular a relação entre genes homólogos. As análises filogenéticas

estimam a relação entre esses genes inferindo a história comum entre eles. Genes homólogos que divergiram depois de evento de especiação são denominados ortólogos, enquanto os genes homólogos que divergiram devido a um evento de duplicação são chamados de parálogos, e estes, normalmente, estão relacionados à neo funcionalização ou divisão da função da proteína ancestral (SALEMI & VANDAMME, 2003).

A redundância do código genético garante que as sequências de nucleotídeos evoluam mais rapidamente do que as sequências de proteínas. Uma mudança na última base de um códon resulta na alteração do amino ácido correspondente em apenas 30% dos casos (SALEMI & VANDAMME, 2003)**.** Assim, existe uma frequência maior de mutações sinônimas quando comparadas às mutações não sinônimas (YANG & BIELAWSKI, 2000). Isto ocorre porque mutações não sinônimas podem alterar a eficiência funcional da proteína codificada em comparação com o alelo selvagem, o que resulta na exclusão do alelo mutante da população, devido à ação da seleção purificadora ou negativa (AGUILETA *et al*., 2010). No entanto, o alinhamento de códons pode mostrar que em alguns casos podem existir posições em que a frequência de mutações não sinônimas é mais elevada do que a frequência desse tipo de mutação no restante do alinhamento em questão, indicando uma pressão de seleção positiva para a variação ao invés da conservação dessa proteína (AGUILETA *et al*., 2009; 2010). Os modelos de substituição de códons demonstram como a seleção natural atua sobre um determinado fenótipo (proteína), através de mudanças no genótipo (DNA). Por isso, a análise de seleção positiva é realizada diretamente nas sequências codificadoras (SCHERER, 2002).

A taxa de substituições não sinônimas/sinônimas ($\omega = d_N/ d_S$) é uma medida que indica que tipo de pressão de seleção está atuando em uma determinada proteína, sendo que $\omega = 1$ indica

evolução neutra, ω <1 indica seleção purificadora e ω >1 indica seleção positiva (YANG & NIELSEN, 2002).

A análise de diversos genes por meio da taxa de substituições sinônimas e não sinônimas revelou a ação da seleção positiva sobre eles (YANG & BIELAWSKI, 2000). A maioria destes genes está envolvida na relação parasita-hospedeiro, tanto na defesa ou imunidade contra os ataques de parasitas (vírus, bactérias, fungos, nematoides e insetos), quanto naqueles genes envolvidos no processo de fuga dos fatores de defesa do hospedeiro e de invasão do parasita (YANG, 2005). Como exemplo, podemos citar genes envolvidos no complexo de histocompatibilidade (HUGHES & NEI, 1988; YANG & SWANSON, 2002), proteína CD45 dos linfoides (FILIP & MUNDY, 2004), genes de planta do tipo R (LEHMANN, 2002), genes de quitinases em plantas, os quais conferem resistência à planta ao ataque de fungos na parede celular (BISHOP *et al*., 2000), proteínas virais de superfície e capsídeo (HAYDON *et al*., 2001; SHACKELTON *et al*., 2005), antígenos da membrana de plasmódio (POLLEY & CONWAY, 2001), entre outros. Os exemplos acima elucidam a corrida evolucionária existente entre parasita e hospedeiro, descrita na teoria da Rainha vermelha (VAN VALEN, 1973). Dessa forma, o conhecimento de genes sob seleção positiva pode ajudar o melhor entendimento da interação gênica entre parasita e hospedeiro.

Inúmeros estudos demonstraram a importância da seleção positiva e da recombinação na evolução molecular de bactérias patogênicas como *Escherichia coli* (PETERSEN *et al*., 2007), *Listeria monocytogenes* (ORSI *et al*., 2008), *Salmonella spp* (SOYER *et al*., 2009), *Streptococcus ssp* (LEFÉBURE & STANHOPE, 2007) e *Campylobacter spp*. (LEFÉBURE & STANHOPE, 2009). Essas análises revelaram que muitos dos genes que estão sob pressão de seleção positiva

estão, geralmente, envolvidos na interação entre patógeno e hospedeiro.

Atualmente, pelo que sabemos, apenas 2 trabalhos analisaram a seleção positiva em diferentes linhagens de *E. coli.* Chen, S. e colaboradores (2006) identificaram genes sob seleção positiva na linhagem UPEC UTI89, utilizando a linhagem K12 (MG1655) (BLATTNER *et al*., 1997) e as linhagens patogênicas para humanos CFTO73 (UPEC) (WELCH *et al*., 2002), EDL933 e SAKAI (enterohemorrágicas) (HAYASHI *et al*., 2001; PERNA *et al*., 2001), além de linhagens de *Shigella* 2547T e 301 (JIN *et al*, 2002; WEI *et al.*, 2003) como genomas de comparação. Na linhagem UTI89, os genes: *fepE*, que regula o comprimento da cadeia do antígeno lipossolúvel O, a adesina *yfaL*, o gene *cedA*, que codifica uma proteína regulatória da divisão celular, *top*B, uma topoisomerase, e 3 genes relacionados ao processo de aquisição de ferro no hospedeiro, *fhuA*, *entD* e *entF* foram considerados como estando sob seleção positiva.

No estudo realizado por Petersen e colaboradores (2007), as análises de seleção positiva foram feitas por meio da comparação do genoma de 4 linhagens diferentes de *E. coli* (K12, CFT073, Sakai O157:H7 e EDL933) e duas de *Shigella flexneri.* (2457T e 301), diferindo do trabalho anteriormente discutido, por não focar em uma linhagem específica. Vinte e três genes mostraram evidencia de seleção positiva. Muitas das proteínas codificadas por esses genes estão diretamente relacionadas à interação patógeno-hospedeiro, ou até na interação entre diferentes bactérias, sugerindo a existência de um processo de co-evolução. Proteínas pertencentes a outras funções, como repressor transcricional, função enzimática e transferência horizontal de DNA foram detectadas, indicando que a seleção também deve estar atuando na resposta adaptativa da interação com fagos e outros micro-organismos e em respostas às alterações do ambiente (PETERSEN *et al*., 2007).

A seleção positiva também foi analisada em diferentes espécies de *Streptoccocus*. Nesses estudos, a maioria dos genes positivamente selecionados estava relacionada à colonização, sobrevivência e propensão de causar doença em um determinado hospedeiro, mostrando, assim, relação com a adaptação específica da bactéria a um determinado nicho (LEFÉBURE & STANHOPE, 2007).

Mesmo linhagens diferentes, de um mesmo gênero, podem apresentar diferenças em determinados grupos gênicos, principalmente aqueles responsáveis por processo fisiológicos ou de virulência (DODRINDT & HACKER, 2001). Este subconjunto de genes deve ser responsável pela sobrevivência do organismo em um determinado nicho ecológico, sendo que esta adaptação pode ser resultante do ganho de genes, por transferência horizontal (HUGHES & FRIEDMAN 2005; LAWRENCE, 1997; OCHMAN *et al*., 2000; SPRINGAEL & TOP, 2004), perda de genes (FOSTER *et al.*, 2005) ou alteração dos genes já existentes (FELDGARDEN *et al*., 2003). O estudo da presença de vários genomas intimamente relacionados de diferentes linhagens, do mesmo patógeno que exploram nichos diferentes, deve fornecer um melhor entendimento sobre os processos de aquisição de novos genes (MARRI *et al*., 2006). Estudos sobre a evolução do genoma de *Escherichia coli*, através da seleção positiva, já foram realizados por Chen, S. e colaboradores (2006) e Petersen e colaboradores. (2007), como mostrado acima. No entanto, nenhum deles incluiu genomas de linhagens APEC nas análises. Com o presente estudo, pretendemos avaliar a evolução do genoma em linhagens patogênicas de *E. coli* para aves (APEC) e humanos, determinando quais proteínas estão fornecendo um maior sucesso às diferentes bactérias, resultando na seleção positiva de seus genes. Os resultados obtidos podem ajudar a elucidar como ocorre a interação dessas linhagens com o hospedeiro, sendo que as

proteínas resultantes poderão ser alvo de estudos futuros, uma vez que elas estão diretamente relacionadas ao sucesso adaptativo das linhagens em questão.

2. Objetivos

2.1 Objetivo geral

Detectar quais grupos de genes homólogos referentes aos genomas de diferentes linhagens APEC, UPEC, NMEC, EHEC e de uma linhagem laboratorial de *Escherichia coli* estão sofrendo pressão de seleção positiva.

2.2 Objetivos específicos

- Efetuar as montagens e anotações dos genomas das linhagens APECs sequenciadas para este estudo;
- Estabelecer os grupos de genes homólogos para os genomas analisados;
- Gerar uma árvore filogenética para cada grupo de proteínas homólogas;
- Determinar quais grupos de genes homólogos apresentam evidências da ação da seleção positiva sob eles, utilizando o método da máxima verossimilhança, para aceitar ou rejeitar modelos de substituições de códons que analisam diferentes taxas de mutações não-sinônimas/sinônimas em sequências codificantes de DNA;
- Verificar quais grupos de genes homólogos com evidência de seleção positiva não apresentam eventos de conversão gênica;
- Identificar qual é a proteína codificada por cada grupo de genes homólogos sob seleção positiva.

3. Material e Métodos

3.1 Linhagens sequenciadas

Foram sequenciadas quatro diferentes linhagens APEC para esse trabalho (Tabela 1).

Tabela 1: Linhagens sequenciadas.

Linhagem	Doença	Órgão/Região de isolamento	Sorogrupo
SEPT362	Septicemia	Fígado de galinha poedeira	OR:H10
SCI-07	Síndrome da cabeça inchada	Edema gelatinoso do tecido infra-orbital	ONT:H31
O8	Onfalite	Gema presente no abdômen de pinto de 1 dia	O38:H10
S17	Septicemia	Fígado de ave de corte	Não tipável (NT)

3.2 Sequenciamentos dos genomas e anotação

O DNA total das quatro linhagens SEPT362, S17, O8 e SCI-07, foram sequenciados, por empresa terceirizada, utilizando-se a tecnologia 454 Life Sciences (MARGULIES *et al*., 2005) para obtenção de sequências de aproximadamente 400 pares de base (pb). A montagem foi realizada com o programa GS De novo Assembler versão 2.5.3 com parâmetros padrão para genoma de bactéria. Os *contigs* resultantes foram ordenados através de alinhamento com genoma de referência de *E. coli* APEC O1 (*Genbank Acession Number* CP000468.1) com auxilio do aplicativo PROmer (KURTZ *et al.* 2004) versão 3.0 e mapeamento "um para um" da opção *show-tiling*.

A cobertura dos *reads* foi calculada assumindo como referência um genoma de 5 Mbp, assim como o genoma de *Escherichia coli* APEC-01 (JOHNSON *et al*., 2007).

Esse trabalho foi realizado pelo grupo de pesquisa do Prof. Dr. Gonçalo Amarante Guimarães Pereira, do Departamento de Genética, Evolução e Bioagentes, no Laboratório de Genômica e

Expressão, da UNICAMP.

A anotação dos genomas foi realizada por um serviço automatizado de anotação - RAST (*Rapid Annotation using Subsystem Technology*), o qual fornece a anotação do genoma de bactérias e arqueas. Essa ferramenta fornece uma anotação de alta qualidade, levando em consideração toda a filogenia do organismo em questão. O site do RAST pode ser acessado em http://rast.nmpdr.org/rast.cgi (AZIZ *et al*., 2008).

3.3 Linhagens utilizadas nas análises de seleção positiva

No presente estudo sobre a evolução dos genomas de diferentes linhagens de *E. coli* por meio da detecção de genes com evidências de estarem sob seleção positiva, foram analisados os genomas de diferentes patótipos, incluindo linhagens APEC, UPEC, EHEC, NMEC. Além de uma linhagem laboratorial não patogênica.

Os genomas das linhagens que não foram sequenciadas durante a realização deste estudo foram obtidos do GeneBank (NCBI- Centro Nacional de Informação Biotecnológica, USA) (Tabela 2).

Tabela 2: Linhagens que foram utilizadas nos testes de seleção positiva.

Nome	**Isolado/Linhagem**	**Número de acesso (NCBI)**
Escherichia coli K-12 str MG1655	Laboratorial	NC_000913.2 (RILEY e*t al*., 2005)
Escherichia coli O157:H7 str Sakai pO157 pOSAK1	Diarreica	NC_002695.1 (BERGHOLZ *et al*., 2007) NC_002128.1 NC_002127.1
Escherichia coli IAI39	Linhagem uropatogênica	NC_011750.1 (TOUCHON *et al*., 2009)
Escherichia coli 536	Linhagem uropatogênica	NC_008253.1 (HOCHHUT *et al*., 2006)
Escherichia coli CFT073	Linhagem uropatogênica	NC_004431.1 (WELCH *et al*, 2002)
Escherichia coli UTI89 pUTI89	Linhagem uropatogênica	NC_007946.1 (CHEN, S. *et al*., 2006) NC_007941.1
Escherichia coli HE3034	Meningite neonatal	CP001969.1 (MORIEL *et al*., 2010);
Escherichia coli S88 pECOS88	Meningite neonatal	NC_011742.1 (TOUCHON *et al*., 2009) NC_011747.1
Escherichia coli APEC 01 pAPEC-O1-ColBM pAPEC-O1-R	Patogênica para aves	NC_008563.1 (JONHSON *et al*., 2007) NC_009837.1 NC_009838.1
Escherichia coli O8	Patogênica para aves	x
Escherichia coli S17	Patogênica para aves	x
Escherichia coli SEPT362	Patogênica para aves	x
Escherichia coli SCI-07	Patogênica para aves	AJFG01000000

3.4 Detecção dos grupos de genes homólogos sob seleção positiva

3.4.1 Grupo de homólogos

Inicialmente, os genomas dos plasmídeos sequenciados independentemente foram concatenados aos genomas cromossômico da linhagem em questão. O genoma de cada linhagem foi utilizado como arquivo de entrada para o software OrthoMCL (CHEN F. *et al*., 2006) para determinar os grupos de genes homólogos.

Em seguida cada gene foi analisado individualmente e filtrado de acordo com os seguintes critérios: múltiplo de 3, códon de início válido (ATG ou GTG), códon de término válido (TAA, TAG ou TGA), sem nucleotídeos ambíguos.

Para obter um conjunto de genes homólogos, que melhor represente os genes ortólogos somente conjuntos com 1 gene por linhagem foram analisados, excluindo-se assim genes com mais de uma cópia, os quais provavelmente são parálogos.

3.4.2 Alinhamentos

Todos os grupos de genes homólogos que continham pelo menos 4 sequências de linhagens diferentes (número mínimo para a realização do teste de seleção positiva) e as sequências proteicas correspondentes foram alinhados pelo software MUSCLE (EDGAR, 2004). As sequências proteicas foram utilizadas como guia para o alinhamento das sequências de nucleotídeos, para obter assim um alinhamento de códon. Os trechos considerados ruins desses alinhamentos, ou seja, qualquer coluna do alinhamento múltiplo que fosse uma lacuna em mais de 20% das sequências foram removidas pelo software trimAll (CAPELLA-GUTIÉRREZ *et al*., 2009).

3.4.3 Análise de seleção positiva

O programa PROML, do pacote PHYLIP (RETIEF, 2000), foi utilizado para gerar árvores de máxima verossimilhança para as sequências proteicas homólogas correspondentes a cada grupo de genes homólogos. Essas árvores e os alinhamentos de códons foram utilizados pelo software CODELM do pacote PAML 4.3 (YANG, 2007) para testar qual modelo de probabilidade de ω, razão entre as taxas de mutações não sinônimas e sinônimas (d_N/ d_S), melhor explicou a distribuição dos dados.

Modelos que verificam a razão ω entre os diferentes sítios de um alinhamento foram

utilizados. Para tal, os modelos M1 x M2 e M7 x M8 foram testados. Os modelos M1 e M2 possuem uma distribuição discreta. M1 assume que os dados estão sob seleção neutra e que assim ω pode apenas assumir os valores 0 e 1. Já o modelo M2 não possui uma razão pré-determinada para ω, permitindo que ele varie entre os valores 0, 1 e > 1, possibilitando a detecção da seleção positiva.

Os modelos M7 e M8 assumem uma distribuição de probabilidade beta para ω. O modelo M7 permite que ω varie entre 0 a 1, assim esse modelo assume que os dados estão sob seleção neutra. Já o modelo M8 permite que ω assuma valores entre 0 até >1, permitindo a seleção positiva.

Para testar os modelos que melhor se adaptam a cada um dos grupos de homólogos, o teste de hipótese χ^2 (Qui quadrado), com 2 graus de liberdade (modelo nulo e alternativo), foi utilizado para os dois diferentes conjuntos de modelos. Apenas os grupos de homólogos com p-value <0.05 para os dois conjuntos de modelos foram considerados como grupos sob seleção positiva.

3.4.4 Recombinação

Os testes para detectar a seleção positiva levam em consideração o polimorfismo entre os genes homólogos de diferentes linhagens. No entanto, o polimorfismo pode ser resultado de recombinação, assim o software GENECONV foi utilizado para a detecção de algum evento de conversão gênica nos grupos de genes homólogos com evidência de seleção positiva (SAWYER, 1989). Os parâmetros utilizados pelo software não foram alterados com exceção da inclusão de um parâmetro que indica que o mesmo deve utilizar apenas polimorfismo de códons na verificação de conversão gênica. O p-value é o indicador de recombinação no grupo, p-value < 0,05 indica recombinação, assim apenas os grupos com p-value > 0,05 foram considerados sem

recombinação e realmente sob seleção positiva.

3.4.5 Identificação das proteínas

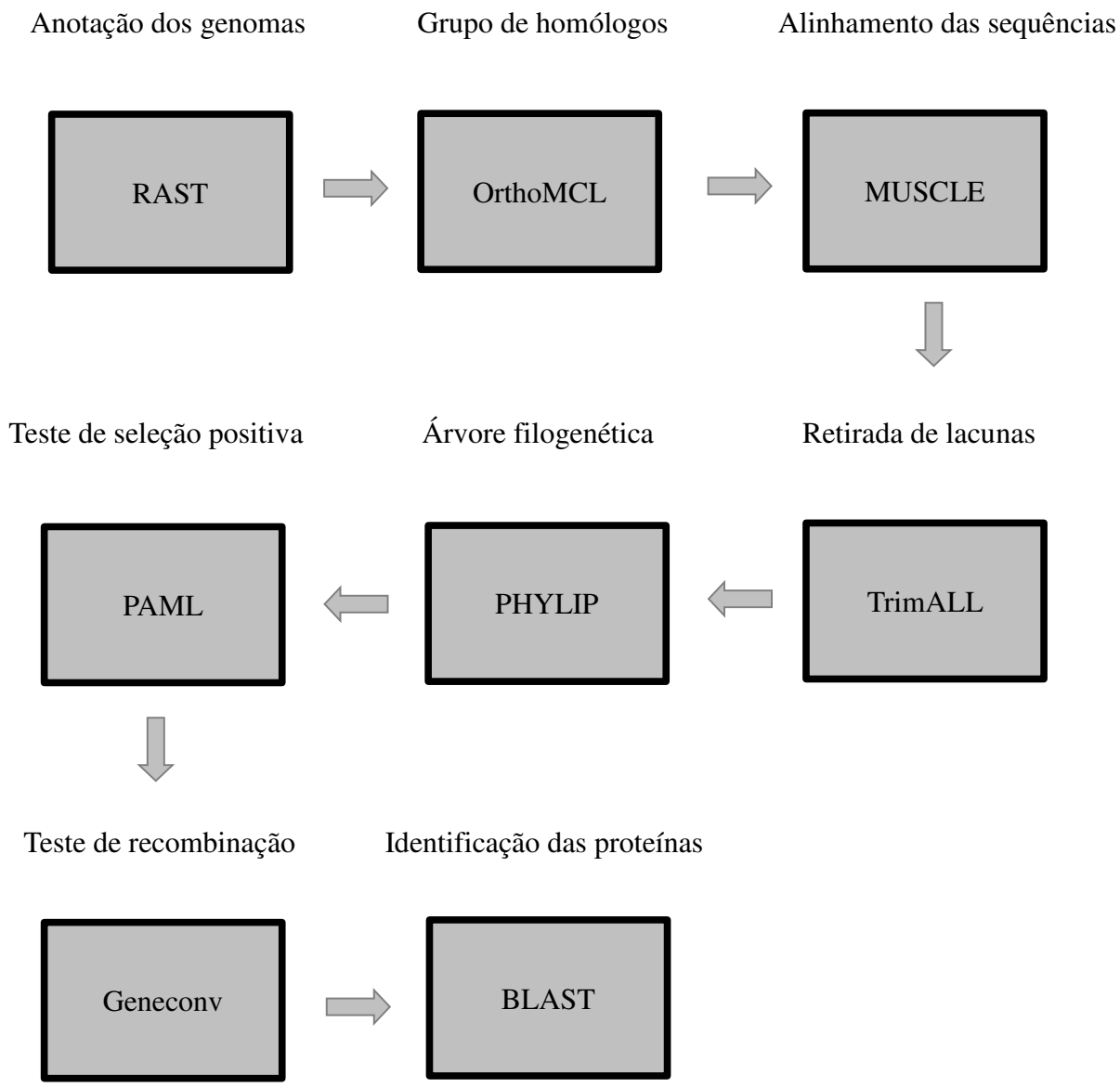
Após a determinação de quais grupos de genes homólogos estavam sob seleção positiva, a sequência de um gene de cada grupo foi utilizada para identificar a proteína correspondente. Tal processo foi realizado pelo programa blastx do software BLAST, localizado na página eletrônica do Centro Nacional de Informação sobre Biotecnologia (NCBI) (http://www.ncbi.nlm.nih.gov/).

3.4.6 Processamento das análises

As análises para detectar grupos de genes homólogos sob seleção positiva foram executadas em parceria com o Laboratório Multiusuário de Bioinformática da EMBRAPA INFORMÁTICA AGROPECUÁRIA, de forma automatizada por um software ainda em implementação, sob supervisão do pesquisador Dr. Francisco Pereira Lobo. As análises de recombinação e identificação das proteínas foram executadas separadamente.

Fluxograma das análises realizadas para a detecção de genes sob seleção positiva

O fluxograma abaixo ilustra os programas e análises realizadas para a detecção de grupos de genes homólogos sob seleção positiva.



3.5 Classificação dos genes baseada em termos de ontologia gênica (GO)

Os genes sob seleção positiva, foram anotados de acordo com a classe referente à ontologia do gene (GO). A anotação dos genes de acordo com a ontologia faz parte do projeto *Gene Ontology Consortium* (**http://www.geneontology.org/**), que tem como objetivo a utilização de termos controlados para a descrição dos genes e seus produtos, para todos os organismos. A descrição dos genes é feita de forma hierárquica, de maneira que classes maiores incluem classes mais específicas. O agrupamento dos genes foi realizado considerando o terceiro nível hierárquico da categoria processos biológico. Essa análise foi realizada por meio do programa Blast2Go (CONESA *et al*., 2005).

3.6 Análise das proteínas hipotéticas sob seleção positiva.

A identificação das proteínas realizada pelo programa blastx, identificou proteínas como sendo hipotéticas, pois suas funções e características são ainda desconhecidas. Para estudar essa proteínas, uma proteína de cada grupo de homólogos identificado como *hypothetical protein* foi analisada pelo programa SOSUIgranN do pacote SOSUI (*Classification and Secondary Structure Prediction of Membrane Proteins*) (IMAI *et al*., 2008) para determinar a provável localização subcelular dessa proteína. Posteriormente, o programa PHOBIUS (KÄLL *et al*., 2004) foi utilizado para a predição de peptídeo sinal nessas mesmas proteínas. A presença de peptídeo sinal indica quais proteínas deverão ser exportada para o espaço periplasmático ou meio extracelular.

4. Resultados

Os softwares empregados na montagem dos genomas (GS De novo Assembler versão 2.5.3 e PROmer 3) não permitiram realizar a montagem completa do genoma de nenhuma das quatro linhagens sequenciadas. No entanto a alta cobertura obtida permitiu que praticamente todo o genoma estivesse representado pelos *contigs* obtidos (Tabela 3).

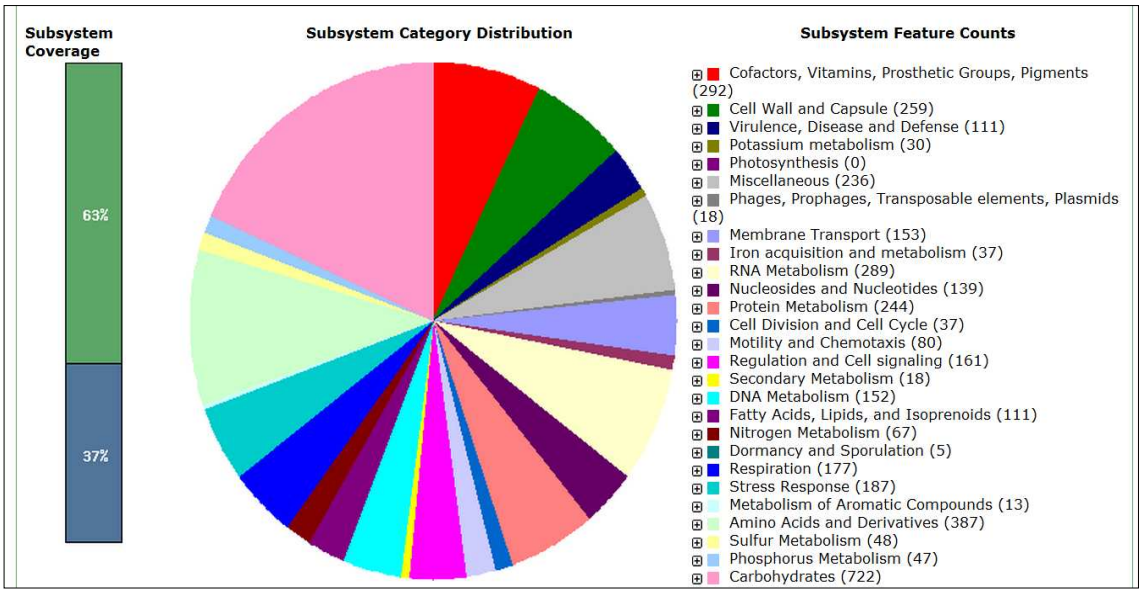
Tabela 3: Resultado dos sequenciamentos das linhagens SCI-07, O8, S17 e SEPT362.

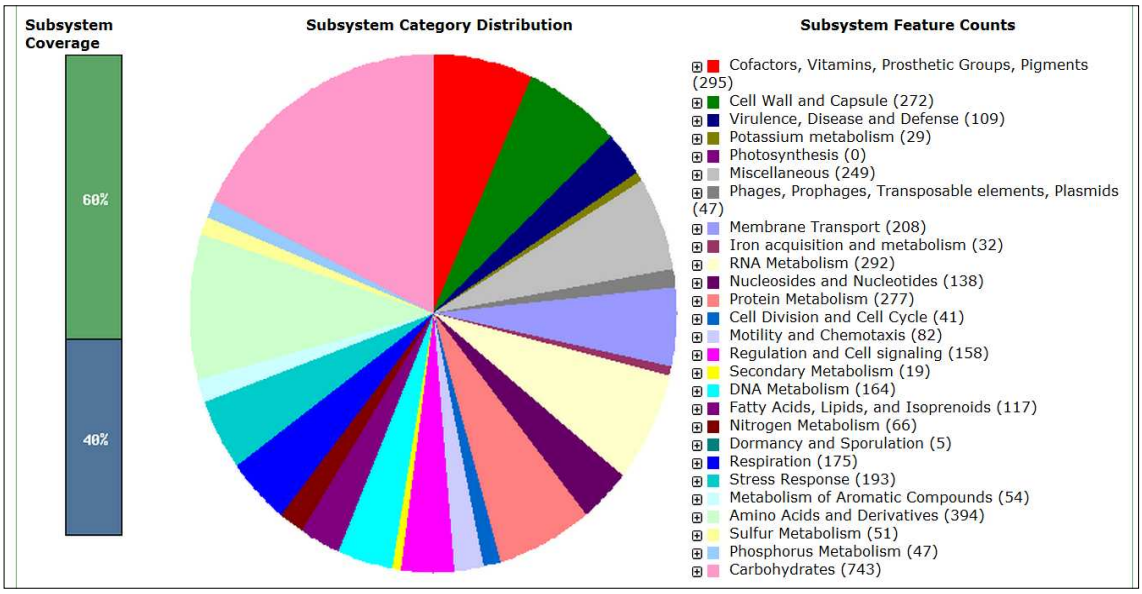
Linhagem	**Número de *reads***	**Tamanho médio do *read* (pb)**	**Cobertura (assumindo genoma de 5 Mpb)**	**Tamanho da montagem (pb)**	**Número de *contigs***
SCI-07	382.008	424	32X	4.972.327	68
O8	301.852	445	26X	5.084.722	160
S17	342.817	369	25X	4.591.470	187
SEPT362	305381	345	21X	5.279.952	173

Os genes anotados pelo RAST (AZIZ *et al*., 2008) foram agrupados em subsistemas, os quais são divididos em grupos de acordo com o papel funcional de cada gene que juntos implementam um processo biológico específico ou um complexo estrutural que normalmente constituem uma via metabólica ou um grupo de proteínas (Figura 1). A anotação do genoma de cada linhagem está descrita na Tabela 4.

Tabela 4: Descrição do número de sequências codificantes e tRNAs para cada linhagem.

Linhagem	**Nº de sequências codificantes**	**Nº de tRNAs**	**Nº de subsistemas**
SCI-07	4881	59	599
SEPT362	5239	65	604
O8	4982	70	610
S17	4442	64	588





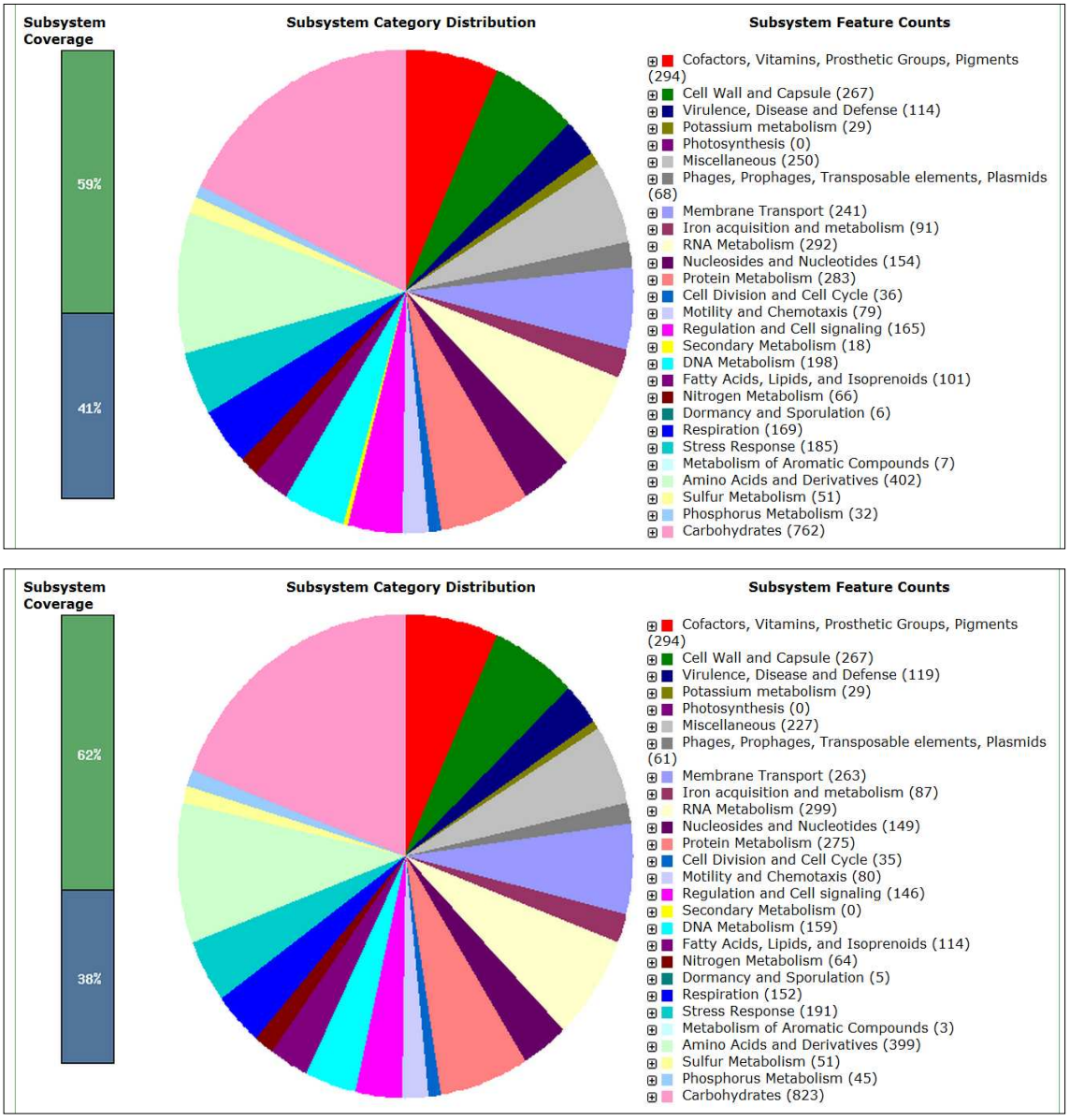


Figura 1: Gráficos da distribuição do genoma nos diferentes subsistemas das linhagens S17, O8, SEPT362 e SCI-7, respectivamente. A coluna *Subsytem feature counts* descreve os diferentes subsistemas e o número de genes agrupados em cada classe. Na coluna *Subsystem coverage* está descrita a porcentagem dos genes que foram agrupados (verde) e dos genes não foram agrupados

em nenhum subsistema (azul). Os genes não agrupados ou possuem função desconhecida ou estão anotados com hypothetical protein.

O genoma de cada uma das treze linhagens submetidos ao programa OrthoMCL, geraram 6597 diferentes grupos de genes homólogos, sendo que após a seleção dos grupos que continham apenas sequências múltiplo de 3, códon de início válido (ATG ou GTG), códon de término válido (TAA, TAG ou TGA), sem nucleotídeos ambíguos, apenas no máximo uma sequência de cada linhagem e no mínimo 4 sequências, restaram 4962 grupos de genes homólogos para as análises posteriores.

Devido ao grande número de grupos de genes homólogos encontrados os resultados das análises serão apresentados somente para um grupo ORTHOMCL 4046.

Os genes homólogos e as sequências proteicas codificadas por eles foram alinhados pelo programa MUSCLE (Figuras 2 e 3). Para cada grupo de proteínas homólogas alinhadas, foi gerada uma árvore filogenética pelo método de Máxima verossimilhança e um teste de confiança denominado Bootstrap (com 1000 repetições) foi feito para as árvores resultantes de todos os grupos de proteínas homólogas. Este teste verifica a probabilidade de um nó realmente retratar a monofilia de um determinado clado. Uma árvore consenso foi gerada para cada grupo de homólogo. Segue abaixo o exemplo da árvore do grupo ORTHOMCL 4046 (Figura 4).

```
                1        10        20        30        40        50
                |        |         |         |         |         |
A62546/1-1005   ATGAAAAAATGGCTCCCTGCTTTTTTATTT---TTATCCCTGTCAGGCTGTAATGATG
A14614/1-1005   ATGAAAAAATGGTTCCCTGCTTTTTTATTT---TTATCCCTGTCAGGCTGTAATGATG
A29192/1-1008   ATGAAAAAATGGTTCCCAGCTTTGTTATTTTCCTTGTGTGTGTCTGGTGAGTCCTCTG
A17692/1-1008   ATGAAAAAATGGTTCCCAGCTTTGTTATTTTCCTTGTGTGTGTCTGGTGAGTCCTCTG
A51024/1-1008   ATGAAAAAATGGTTCCCAGCTTTGTTATTTTCCTTGTGTGTGTCTGGTGAGTCCTCTG

A62546/1-1005   CTGGCTGCAAACCAGAGTACAATGTTTTACTCGTTTAATGATAACATTTATCGTCCTC
A14614/1-1005   CTGGCTGCAAACCAGAGTACAATATTTTACTCGTTTAATGATAACATTTATCATCCTC
A29192/1-1008   TGG---------AATAATATTGTCTTTTACTCCCTTGGAGACGTTAACTCTTATCAGG
A17692/1-1008   TGG---------AATAATATTGTCTTTTACTCCCTTGGAAACGTTAACTCTTATCAGG
A51024/1-1008   TGG---------AATAATATTGTCTTTTACTCCCTTGGAAACGTTAACTCTTATCAGG

A62546/1-1005   CTTAGTGTTAAAGTAACCGATATTGTTCAATTCATAGTGGATATAAACTCCGCATCAA
A14614/1-1005   CTTAGTGTTAAAGTAACCGATATTGTTCAATTCATAGTGGATATAAACTCCGCATCAA
A29192/1-1008   GGGAATGTGGTGATTACTCAAAGGCCACAATTTATAACTTCGTGGCGCCCGGGC----
A17692/1-1008   GGGAATGTGGTGATTACTCAAAGGCCACAATTTATAACTTCGTGGCGCCCGGGC----
A51024/1-1008   GGGAATGTGGTGATTACTCAAAGGCCACAATTTATAACTTCGTGGCGCCCGGGC----

A62546/1-1005   ACGGCAACTTTAAGCTATGTGGCCTGCAATGGATTTACCTGGACTCATGGTCTT----
A14614/1-1005   ACGGCAACTTTAAGCTATGTGGCCTGCAATGGATTTACCTGGACTCATGGTCTT----
A29192/1-1008   ATTGCTACGGTAACCTGGAATCAGTGTAATGGTCTTGGGTTCGCTGATGGCTCCTGGG
A17692/1-1008   ATTGCTACGGTAACCTGGAATCAGTGTAATGGTCCTGGGTTTGCTGATGGTTCCTGGG
A51024/1-1008   ATTGCTACGGTAACCTGGAATCAGTGTAATGGTCCTGGGTTTGCTGATGGTTTCTGGG

A62546/1-1005   TACTGGTCTGAGTATTTTGCATGGCTGGTTGTTCCTAAACATGTT---TCCTATAATG
A14614/1-1005   TACTGGTCTGAGTATTTTGCATGGCTGGTTGTTCCTAAACATGTT---TCCTATAATG
A29192/1-1008   TACTACAGGGAGTATATTGCGTGGGTAGTATTCCCCAAAAAGGTTATGACCCAAAATG
A17692/1-1008   TACTACAGGGAGTATATTGCGTGGGTAGTATTCCCCAAAAAGGTTATGACCAAAAATG
A51024/1-1008   TACTACAGGGAGTATATTGCGTGGGTAGTATTCCCCAAAAAGGTTATGACCAAAAATG

A62546/1-1005   TATAATATATATCTTGAACTTCAGTCCAGAGGAAGTTTTTCACTTGAT---GCAGAAG
A14614/1-1005   TATAATATATATCTTGAACTTCAGTCCAAAGGAGGTTTTTCACTTGAT---GCAGAAG
A29192/1-1008   TATCCCTTATTTATTGAGGTTCATAATAAAGGTAACTGGAGTGAGGAGAATAGTGGTG
A17692/1-1008   TATCCCTTATTTATTGAGGTTCATAATAAAGGTAGCTGGAGTGAGGAGAATACTGGTG
A51024/1-1008   TATCCCTTATTTATTGAGGTTCATAATAAAGGTAGCTGGAGTGAGGAGAATACTGGTG

A62546/1-1005   AATGATAATTACTATCTTACCAAGGGATTTGCATGGGATGAA---GTAAACTCATCTG
A14614/1-1005   AATGATAATTACTATCTTACCAAGGGATTTGCATGGGATGAA---GTAAACTCATCTG
A29192/1-1008   AATGACAGCTATTTTTTTCTCAAGGGGTATAAGTGGGATGAGCGGGCCTTTGATGCAG
A17692/1-1008   AATGACAGCTATTTTTTTCTCAAGGGGTATAAGTGGGATGAGCGGGCCTTTGATGCAG
A51024/1-1008   AATGACAGCTATTTTTTTCTCAAGGGGTATAAGTGGGATGAGCGGGCCTTTGATGCAG

A62546/1-1005   CGGGTATGTTTCGATATCGGAGAAAAAAGAAGTCTGGCATGGTCATTTGGTGGTGTTA
A14614/1-1005   CGGGTATGTTTCGATATCGGAGAAAAAAGAAGTCTGGCATGGTCATTTGGTGGTGTTA
A29192/1-1008   AATTTGTGTCAGAAACCAGGAGAAACAACTCGTCTGACTGAGAAATTTGACGATATTA
A17692/1-1008   AATTTGTGTCAGAAACCAGGAGAAACAACCCGTCTGACTGAGAAATTTGACGATATTA
A51024/1-1008   AATTTGTGTCAGAAACCAGGAGAAACAACCCGTCTGACTGAGAAATTTGACGATATTA
```

```
A62546/1-1005   CTGAACGCCAGATTGCCTGTTGACCTTCCTAAGGGGGATTATACGTTTCCAGTTAAGT
A14614/1-1005   CTGAACGCCAGATTGCCTGTTGACCTTCCTAAGGGGGATTATACGTTTCCAGTTAAGT
A29192/1-1008   TTTAAAGTCGCCTTACCTGCAGATCTTCCTTTAGGGGATTATTCTGTTAAAATTCCAT
A17692/1-1008   TTTAAAGTCGCCCTACCTGCAGATCTTCCTTTAGGGGATTATTCTGTTACAATTCCAT
A51024/1-1008   TTTAAAGTCGCCCTACCTGCAGATCTTCCTTTAGGGGATTATTCTGTTACAATTCCAT

A62546/1-1005   TTACGTGGCATTCAGCGTAATAATTATGATTATATTGGTGGACGCTACAAAATCCCTT
A14614/1-1005   TTACGTGGCATTCAGCGTAATAATTATGATTATATTGGTGGACGCTACAAAATCCCTT
A29192/1-1008   ACTTCCGGCATGCAGCGTCATTTCGCGAGTTACTTGGGGGCCCGTTTTAAAATCCCAT
A17692/1-1008   ACTTCCGGCATGCAGCGTCATTTCGCGAGTTACTTGGGAGCCCGTTTTAAAATCCCAT
A51024/1-1008   ACTTCCGGCATGCAGCGTCATTTCGCGAGTTACTTGGGAGCCCGTTTTAAAATCCCAT

A62546/1-1005   TCGTTAATGAAAACATTTCCTTTTAATGGTACATTGAATTTCTCAATTAAAAATACCG
A14614/1-1005   TCGTTAATGAAAACATTTCCTTTTAATGGTACATTGAATTTCTCAATTAAAAATACCG
A29192/1-1008   AATGTGGCCAAAACCCTCCCAAGAGAGAATGAAATGTTATTCTTATTTAAGAATATCG
A17692/1-1008   AATGTGGCCAAAACTCTCCCAAGAGAGAATGAAATGTTATTCTTATTTAAGAATATCG
A51024/1-1008   AATGTGGCCAAAACTCTCCCAAGAGAGAATGAAATGTTATTCTTATTTAAGAATATCG

A62546/1-1005   GGATGCCGTCCTTCTGCACAGTCTCTGGAAATAAATCATGGTGATCTGTCGATTAATA
A14614/1-1005   GGATGCCGTCCTTCTGCACAGTCTCTGGAAATAAATCATGGTGATCTGTCGATTAATA
A29192/1-1008   GGATGCCGTCCTTCTGCACAGTCTCTGGAAATAAAGCATGGTGATCTGTCTATTAATA
A17692/1-1008   GGATGCCGTCCTTCTGCACAGTCTCTGGAAATAAAGCATGGTGATCTGTCTATTAATA
A51024/1-1008   GGATGCCGTCCTTCTGCACAGTCTCTGGAAATAAAGCATGGTGATCTGTCTATTAATA

A62546/1-1005   GCTAATAATCATTATGCGGCTCAGACTCTTTCTGTGTCTTGCGATGTGCCTACAAATA
A14614/1-1005   GCTAATAATCATTATGCGGCTCAGACTCTTTCTGTGTCTTGCGATGTGCCTACAAATA
A29192/1-1008   GCTAATAATCATTATGCGGCTCAGACTCTTTCTGTGTCTTGCGATGTGCCTGCAAATA
A17692/1-1008   GCTAATAATCATTATGCGGCTCAGACTCTTTCTGTGTCTTGCGATGTGCCTGCAAATA
A51024/1-1008   GCTAATAATCATTATGCGGCTCAGACTCTTTCTGTGTCTTGCGATGTGCCTGCAAATA

A62546/1-1005   CGTTTTTTCCTGTTAAGCAATACAACTCCGGCATACAGCCATGGTCAGCAATTTTCGG
A14614/1-1005   CGTTTTTTCCTGTTAAGCAATACAACTCCGGCATACAGCCATGGTCAGCAATTTTCGG
A29192/1-1008   CGTTTTATGCTGTTAAGAAATACAACTCCGACATACAGCCATGGTAAGAAATTTTCGG
A17692/1-1008   CGTTTTATGCTGTTAAGAAATACAACTCCGACATACAGCCATGGTAAGAAATTTTCGG
A51024/1-1008   CGTTTTATGCTGTTAAGAAATACAACTCCGACATACAGCCATGGTAAGAAATTTTCGG

A62546/1-1005   GGTCTGGGTCATGGCTGGGACTCCATTGTTTCGATTAATGGCGTGGACACAGGAGAGA
A14614/1-1005   GGTCTGGGTCATGGCTGGGACTCCATTGTTTCGATTAATGGCGTGGACACAGGAGAGA
A29192/1-1008   GGTCTGGGGCATGGCTGGGACTCCATTGTTTCAGTTAACGGGGTGGACACAGGAGAGA
A17692/1-1008   GGTCTGGGGCATGGCTGGGACTCCATTGTTTCGGTTAACGGGGTGGACACAGGAGAGA
A51024/1-1008   GGTCTGGGGCATGGCTGGGACTCCATTGTTTCGGTTAACGGGGTGGACACAGGAGAGA

A62546/1-1005   ACGATGAGATGGTACAGAGCAGGTACACAAAACCTGACCATCGGCAGTCGCCTCTATG
A14614/1-1005   ACGATGAGATGGTACAGAGCAGGTACACAAAACCTGACCATCGGCAGTCGCCTCTATG
A29192/1-1008   ACGATGAGATGGTACAAAGCAGGTACACAAAGCCTGACCATCGGCAGTCGCCTCTATG
A17692/1-1008   ACGATGAGATGGTACAAAGCAGGTACACAAAACCTGACCATCGGCAGTCGCCTCTATG
A51024/1-1008   ACGATGAGATGGTACAAAGCAGGTACACAAAACCTGACCATCGGCAGTCGCCTCTATG
```

```
A62546/1-1005   GAATCTTCAAAGATACAACCAGGAGTACTATCTGGTTCAGCAACGCTGCTCATGATAT
A14614/1-1005   GAATCTTCAAAGATACAACCAGGAGTACTATCTGGTTCAGCAACGCTGCTCATGATAT
A29192/1-1008   GAATCTTCAAAGATACAACCAGGAGTACTATCTGGTTCAGCAACGCTGCTCATGATAT
A17692/1-1008   GAATCTTCAAAGATACAACCAGGAGTACTATCTGGTTCAGCAACGCTGCTCATGATAT
A51024/1-1008   GAATCTTCAAAGATACAACCAGGAGTACTATCTGGTTCAGCAACGCTGCTCATGATAT

A62546/1-1005   CCA
A14614/1-1005   CCA
A29192/1-1008   CCA
A17692/1-1008   CCA
A51024/1-1008   CCA
```

Figura 2: Alinhamento de códons do grupo de genes homólogos ORTHOMCL 4046, obtidos pelo programa MUSCLE.

```
          1        10        20        30        40        50        60
          |        |         |         |         |         |         |
A62546    MKKWLPAFLF-LSLSGCNDALAANQSTMFYSFNDNIYRPQLSVKVTDIVQFIVDINSASS
A14614    MKKWFPAFLF-LSLSGCNDALAANQSTIFYSFNDNIYHPQLSVKVTDIVQFIVDINSASS
A17692    MKKWFPALLFSLCVSGESSAW---NNIVFYSLGNVNSYQGGNVVITQRPQFITSWRPG--
A51024    MKKWFPALLFSLCVSGESSAW---NNIVFYSLGNVNSYQGGNVVITQRPQFITSWRPG--
A29192    MKKWFPALLFSLCVSGESSAW---NNIVFYSLGDVNSYQGGNVVITQRPQFITSWRPG--

A62546    TATLSYVACNGFTWTHGL--YWSEYFAWLVVPKHV-SYNGYNIYLELQSRGSFSLD-AED
A14614    TATLSYVACNGFTWTHGL--YWSEYFAWLVVPKHV-SYNGYNIYLELQSKGGFSLD-AED
A17692    IATVTWNQCNGPGFADGSWAYYREYIAWVVFPKKVMTKNGYPLFIEVHNKGSWSEENTGD
A51024    IATVTWNQCNGPGFADGFWAYYREYIAWVVFPKKVMTKNGYPLFIEVHNKGSWSEENTGD
A29192    IATVTWNQCNGLGFADGSWAYYREYIAWVVFPKKVMTQNGYPLFIEVHNKGNWSEENSGD

A62546    NDNYYLTKGFAWDE-VNSSGRVCFDIGEKRSLAWSFGGVTLNARLPVDLPKGDYTFPVKF
A14614    NDNYYLTKGFAWDE-VNSSGRVCFDIGEKRSLAWSFGGVTLNARLPVDLPKGDYTFPVKF
A17692    NDSYFFLKGYKWDERAFDAGNLCQKPGETTRLTEKFDDIIFKVALPADLPLGDYSVTIPY
A51024    NDSYFFLKGYKWDERAFDAGNLCQKPGETTRLTEKFDDIIFKVALPADLPLGDYSVTIPY
A29192    NDSYFFLKGYKWDERAFDAGNLCQKPGETTRLTEKFDDIIFKVALPADLPLGDYSVKIPY

A62546    LRGIQRNNYDYIGGRYKIPSSLMKTFPFNGTLNFSIKNTGGCRPSAQSLEINHGDLSINS
A14614    LRGIQRNNYDYIGGRYKIPSSLMKTFPFNGTLNFSIKNTGGCRPSAQSLEINHGDLSINS
A17692    TSGMQRHFASYLGARFKIPYNVAKTLPRENEMLFLFKNIGGCRPSAQSLEIKHGDLSINS
A51024    TSGMQRHFASYLGARFKIPYNVAKTLPRENEMLFLFKNIGGCRPSAQSLEIKHGDLSINS
A29192    TSGMQRHFASYLGARFKIPYNVAKTLPRENEMLFLFKNIGGCRPSAQSLEIKHGDLSINS
```

```
A62546   ANNHYAAQTLSVSCDVPTNIRFFLLSNTTPAYSHGQQFSVGLGHGWDSIVSINGVDTGET
A14614   ANNHYAAQTLSVSCDVPTNIRFFLLSNTTPAYSHGQQFSVGLGHGWDSIVSINGVDTGET
A17692   ANNHYAAQTLSVSCDVPANIRFMLLRNTTPTYSHGKKFSVGLGHGWDSIVSVNGVDTGET
A51024   ANNHYAAQTLSVSCDVPANIRFMLLRNTTPTYSHGKKFSVGLGHGWDSIVSVNGVDTGET
A29192   ANNHYAAQTLSVSCDVPANIRFMLLRNTTPTYSHGKKFSVGLGHGWDSIVSVNGVDTGET

A62546   TMRWYRAGTQNLTIGSRLYGESSKIQPGVLSGSATLLMILP
A14614   TMRWYRAGTQNLTIGSRLYGESSKIQPGVLSGSATLLMILP
A17692   TMRWYKAGTQNLTIGSRLYGESSKIQPGVLSGSATLLMILP
A51024   TMRWYKAGTQNLTIGSRLYGESSKIQPGVLSGSATLLMILP
A29192   TMRWYKAGTQSLTIGSRLYGESSKIQPGVLSGSATLLMILP
```

Figura 3: Alinhamento das proteínas homóloga do grupo ORTHOMCL 4046 obtido pelo programa MUSCLE.

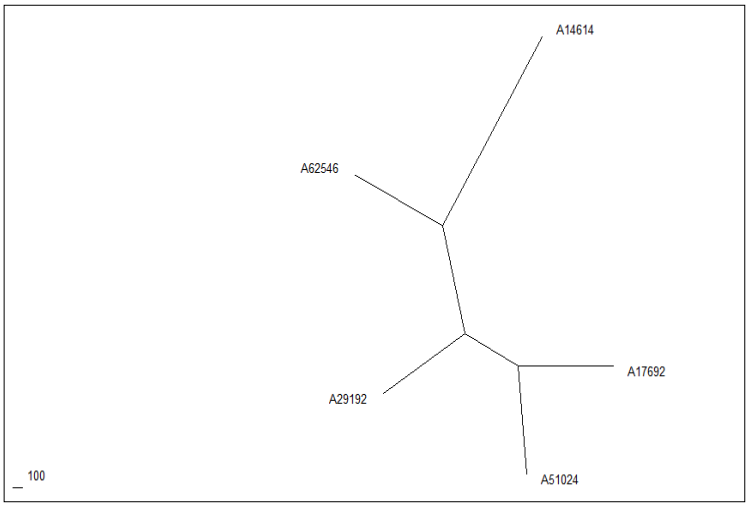


Figura 4: Árvore consenso de Máxima verossimilhança para o grupo de homólogos ORTHOMCL 4046, obtida pelo programa PHYLIP.

Os testes estatísticos utilizados para determinar a ocorrência de seleção positiva empregaram o cálculo da razão ω para cada sítio (códon) das sequências alinhadas nos ramos da árvore filogenética correspondente. Os testes de seleção positiva realizados pelo programa PAML detectaram 254 grupos de homólogos com evidência de seleção positiva. Após o teste de recombinação, realizado pelo programa Geneconv, 138 grupos apresentaram eventos de conversão gênica, restando assim 116 grupos de genes homólogos sob seleção positiva.

O resultado dos testes para o grupo ORTHOMCL 4046 está descrito abaixo (Tabela 5)

A identificação de uma proteína correspondente a cada grupo de genes homólogos sob seleção positiva está descrita na Tabela 6.

Tabela 5: Resultados das análises efetuadas para o grupo de homólogos ORTHOMCL 4046.

Grupo de homólogo:	ORTHOMCL 4046
Valor do modelo M1 (seleção neutra):	-2.342.112.264
Valor do modelo M2 (seleção positiva):	-2.330.437.366
Teste da razão de verossimilhança (M1xM2):	233.497.959.999.995
P-value M1xM2 (teste de hipótese χ^2):	0,0000085046
Resultado para os modelos M1xM2:	Seleção positiva
Valor do modelo M7 (seleção neutra):	-2.342.552.808
Valor do modelo M8 (seleção positiva):	-2.330.440.975
Teste da razão de verossimilhança (M7xM8):	242.236.659.999.999
P-value M7xM8 (teste de hipótese χ^2):	0,0000054941
Resultado para os modelos M7xM8:	Seleção positiva
P-value para o teste de recombinação:	0,7855
Resultado do BLAST:	p pilus adhesin PapG protein [Escherichia coli UTI89]

Tabela 6: Identificação das proteínas referentes a cada grupo de genes homólogos sob seleção positiva.

Grupo de homólogos	Gene ID	Símbolo do gene	Referencia da sequência NCBI	Proteínas
ORTHOMCL 1053	959776	yjgP	NP_290894.1	yjgP gene product
ORTHOMCL 1106	6147080	EcSMS35_3825	YP_001745791.1	LysR family transcriptional regulator
ORTHOMCL 1134	3781797	srlB	YP_409167.1	PTS system glucitol/sorbitol-specific transporter subunit IIA
ORTHOMCL 1154	961306	rpoH	NP_290009.1	rpoH gene product
ORTHOMCL 1272	958094	ybiT	NP_286584.1	ybiT gene product
ORTHOMCL 1404	959923	rpsF	NP_290830.1	rpsF gene product
ORTHOMCL 142	7153522	hslJ	YP_002407698.1	heat-inducible protein
ORTHOMCL 1445	4190150	ECP_3336	YP_671217.1	regulatory protein CsrD
ORTHOMCL 1568	7063996	ydhS	YP_002329274.1	ydhS gene product
ORTHOMCL 1582	1036152	ydhM	NP_753937.1	transcriptional regulator YdhM
ORTHOMCL 159	4495112	yafV	YP_851436.1	yafV gene product
ORTHOMCL 1608	12693969	yafE	ZP_07594298.1	Methyltransferase type
ORTHOMCL 1609	956973	yafD	NP_285897.1	yafD gene product
ORTHOMCL 1718	958280	ompC	NP_288795.1	ompC gene product
ORTHOMCL 177	6145043	EcSMS35_1347	YP_001743405.1	copper resistance protein D
ORTHOMCL 1952	960924	rfaC	NP_290204.1	rfaC gene product
ORTHOMCL 1992	916932	ECs2315	NP_310342.1	sensor protein RstB
ORTHOMCL 1993	4187334	ECP_1552	YP_669460.1	DNA-binding transcriptional regulator RstA
ORTHOMCL 210	956819	yegT	NP_288600.1	yegT gene product
ORTHOMCL 2116	5594876	EcHS_A3589	YP_001460187.1	PilN family protein
ORTHOMCL 2130	1026805	cpxA	NP_709717.1	cpxA gene product
ORTHOMCL 2211	4188168	ECP_0458	YP_668388.1	phosphate regulon transcriptional regulatory protein PhoB
ORTHOMCL 2512	958586	tdcB	NP_289689.1	tdcB gene product
ORTHOMCL 2589	7152322	ybgF	YP_002406743.1	tol-pal system protein YbgF
ORTHOMCL 2653	959940	yjfK	NP_290813.1	yjfK gene product
ORTHOMCL 2681	948415	uspD	NP_418358.1	stress-induced protein
ORTHOMCL 2696	946533	plaP	NP_416518.2	putrescine importer, low affinity
ORTHOMCL 2720	3993304	argB	YP_543493.1	acetylglutamate kinase
ORTHOMCL 273	948548	lamB	NP_418460.1	maltose outer membrane porin (maltoporin)
ORTHOMCL 2788	961919	fliZ	NP_288382.1	fliZ gene product
ORTHOMCL 2816	7132741	insO	YP_002394341.1	transposase ORF B (fragment), IS911
ORTHOMCL 2840	5587940	EcE24377A_2023	YP_001463097.1	IS605 family transposase
ORTHOMCL 2888	5595048	phnP	YP_001460876.1	carbon-phosphorus lyase

				complex accessory protein
ORTHOMCL 2957	4493769	hipA	YP_852651.1	hipA gene product
ORTHOMCL 3013	7151781	allB	YP_002406506.1	allantoinase
ORTHOMCL 3058	4494089	bcsA	YP_859139.1	bcsA gene product
ORTHOMCL 306	12682576	yjdA	YP_006141411.1	yjdA gene product
ORTHOMCL 3241	7158629	nhoA	YP_002412452.1	N-hydroxyarylamine O-acetyltransferase
ORTHOMCL 3344	945757	treA	NP_415715.1	periplasmic trehalase
ORTHOMCL 3353	7149616	gltA	YP_002406703.1	type II citrate synthase
ORTHOMCL 3359	12682671	yjeF	YP_006141503.1	yjeF gene product
ORTHOMCL 3407	3990845	UTI89_C4957	YP_543891.1	DNA-binding transcriptional regulator DsdC
ORTHOMCL 3444	7061829	rpiB	YP_002331872.1	rpiB gene product
ORTHOMCL 3487	3992321	UTI89_C0389	YP_539421.1	structural protein
ORTHOMCL 3542	1036698	yadC	NP_752116.1	fimbrial-like adhesin protein
ORTHOMCL 3602	12693287	ECOK1_4718	YP_006103772.1	HipA-like N-terminal domain family
ORTHOMCL 3618	4190002	ECP_3422	YP_671303.1	general secretion pathway protein M
ORTHOMCL 3682	12728409	ECABU_c30960	YP_006107099.1	VGR-related
ORTHOMCL 3685	12687669	UMNK88_4410	YP_006136091.1	UDP-galactose LPS alpha1,2-galactosyltransferase
ORTHOMCL 3691	960256	yiiG	NP_290525.1	yiiG gene
ORTHOMCL 3708	4495712	yieL	YP_859315.1	yieL gene product
ORTHOMCL 3711	945738	ycgG	NP_415686.2	predicted cyclic-di-GMP phosphodiesterase
ORTHOMCL 3715	959719	yjiJ	NP_290950.1	yjiJ gene product
ORTHOMCL 3730	8480954	ECO26_4015	YP_003230940.1	DNA-binding protein
ORTHOMCL 3731	1038989	c3225	NP_755107.1	hypothetical protein c3225
ORTHOMCL 3790	4189441	ECP_3027	YP_670911.1	capsule polysaccharide export protein KpsS
ORTHOMCL 3796	7151459	irp	YP_002407084.1	salicyl-AMP ligase YbtE
ORTHOMCL 380	1035560	pspA	NP_753678.1	phage shock protein PspA
ORTHOMCL 3839	7133315	ECS88_3097	YP_002392728.1	hypothetical protein ECS88_3097 [ICMF-related]
ORTHOMCL 3840	4495821	aec31	YP_851414.1	aec31 gene product [impA]
ORTHOMCL 3893	7150038	ECIAI39_2691	YP_002408630.1	hypothetical protein ECIAI39_2691
ORTHOMCL 3898	3992544	UTI89_C2165	YP_541169.1	outer membrane protein **N**
ORTHOMCL 3903	7131573	ECS88_1339	YP_002391111.1	hypothetical protein ECS88_1339
ORTHOMCL 3913	4191388	ECP_3517	YP_671394.1	minor fimbrial subunit
ORTHOMCL 3923	4495672	hlyD	YP_852503.1	hlyD gene product
ORTHOMCL 3928	6144833	EcSMS35_4387	YP_001746332.1	hypothetical protein EcSMS35_4387
ORTHOMCL 3942	5586718	EcE24377A_0230	YP_001461389.1	type VI secretion-associated protein [VC_A0118]
ORTHOMCL 3965	7149688	ECIAI39_1028	YP_002407049.1	hypothetical protein ECIAI39_1028
ORTHOMCL 3993	12689137	ECOK1_0569	YP_006099808.1	tail fiber assembly protein
ORTHOMCL 4046	3992036	papG	YP_543821.1	p pilus adhesin PapG protein

ORTHOMCL 4067	1037359	c5349	NP_757197.1	carbamate kinase
ORTHOMCL 4068	3992062	argF	YP_543790.1	ornithine carbamoyltransferase
ORTHOMCL 4085	7154568	ECIAI39_3173	YP_002409100.1	hypothetical protein ECIAI39_3173
ORTHOMCL 4099	1037920	c2608	NP_754496.1	hypothetical protein c2608
ORTHOMCL 4116	957474	Z0639	NP_286235.1	hypothetical protein Z0639
ORTHOMCL 4177	12884549	EC042_3791	YP_006098077.1	hypothetical protein EC042_3791
ORTHOMCL 4178	7150243	ECIAI39_3992	YP_002409884.1	hypothetical protein ECIAI39_3992
ORTHOMCL 4189	7119472	trbC	YP_002401080.1	conjugal transfer pilus assembly protein TrbC
ORTHOMCL 422	958375	ygfZ	NP_289467.1	ygfZ gene product
ORTHOMCL 4223	1037251	c5390	NP_757238.1	hypothetical protein c5390
ORTHOMCL 4256	7393823	LF82_272	YP_002556436.1	tail component of prophageCP-933K
ORTHOMCL 4274	7130464	neuE	YP_002392935.1	polysialic acid biosynthesis protein
ORTHOMCL 4276	3990582	UTI89_C3047	YP_542034.1	hypothetical protein UTI89_C3047
ORTHOMCL 4326	7152795	hsdS	YP_002410668.1	Type I restriction enzyme EcoAI specificity protein (S protein) (S.EcoAI)
ORTHOMCL 4335	1039581	iucB	NP_755501.1	IucB protein
ORTHOMCL 4361	3991957	UTI89_C3390	YP_542368.1	hypothetical protein UTI89_C3390
ORTHOMCL 4363	7150558	ECIAI39_2267	YP_002408230.1	hypothetical protein ECIAI39_2267
ORTHOMCL 4383	8481823	mhpT	YP_003227473.1	3-hydroxyphenylpropionic transporter MhpT
ORTHOMCL 4423	7144277	trbJ	YP_002405984.1	conjugal transfer protein TrbJ
ORTHOMCL 4429	958665	Z4585	NP_289795.1	cryptic C4-dicarboxylate transporter DcuD
ORTHOMCL 4430	3853559	inv	YP_444138.1	putative invertase
ORTHOMCL 4475	3996438	istA	YP_538670.1	transposase for insertion sequence element IS21
ORTHOMCL 4525	3993876	UTI89_C3199	YP_542179.1	hypothetical protein UTI89_C3199
ORTHOMCL 4531	4493527	APECO1_4190	YP_853478.1	hypothetical protein
ORTHOMCL 454	958861	ycaD	NP_286775.1	ycaD gene product
ORTHOMCL 4555	1037008	c1220	NP_753137.1	phospho-2-dehydro-3-deoxyheptonate aldolase
ORTHOMCL 4579	4493270	APECO1_1035	YP_853054.1	hypothetical protein
ORTHOMCL 4676	961700	Z2759	NP_288163.1	hypothetical protein
ORTHOMCL 4718	7151723	hyfD	YP_002408571.1	hydrogenase 4 subunit D
ORTHOMCL 472	945702	icd	NP_415654.1	e14 prophage isocitrate dehydrogenase, specific for NADP+
ORTHOMCL 4751	7063467	E2348C_0565	YP_002328136.1	TerC family integral membrane protein
ORTHOMCL 4753	3990200	UTI89_C2862	YP_541853.1	hypothetical protein

				UTI89_C2862
ORTHOMCL 4850	3989632	UTI89_C4074	YP_543034.1	hypothetical protein UTI89_C4074
ORTHOMCL 4867	4494219	APECO1_4020	YP_853646.1	integrase-like protein
ORTHOMCL 4871	7150040	ECIAI39_2693	YP_002408632.1	putative replication protein P of bacteriophage
ORTHOMCL 4949	1036444	c0608	NP_752542.1	hypothetical protein c0608
ORTHOMCL 4959	5616590	APECO1_O1CoBM82	YP_001481237.1	hypothetical protein APECO1_O1CoBM82
ORTHOMCL 5029	945245	djlB	NP_415179.1	predicted co-chaperone
ORTHOMCL 506	3993485	pabC	YP_540235.1	4-amino-4-deoxychorismate lyase
ORTHOMCL 5092	12657552	ECO55CA74_00110	YP_006157211.1	non-LEE-encoded type III secreted effector
ORTHOMCL 5105	3990815	UTI89_C2441	YP_541440.1	hypothetical protein UTI89_C2441
ORTHOMCL 685	4491754	fucR	YP_854021.1	fucR gene product
ORTHOMCL 72	12879578	ETEC_3215	YP_006116755.1	putative GTP-binding protein
ORTHOMCL 721	7063565	agaB	YP_002330892.1	agaB gene product
ORTHOMCL 758	4494298	rlpA	YP_851762.1	rlpA gene product
ORTHOMCL 899	1036882	c4208	NP_756072.1	fimbrial chaperone

Os genes sob seleção positiva foram anotados de acordo com a Ontologia do Gene (GO), considerando apenas a categoria processo biológico. Os resultados dessa anotação, para o nível hierárquico 3, estão descritos na Figura 5.

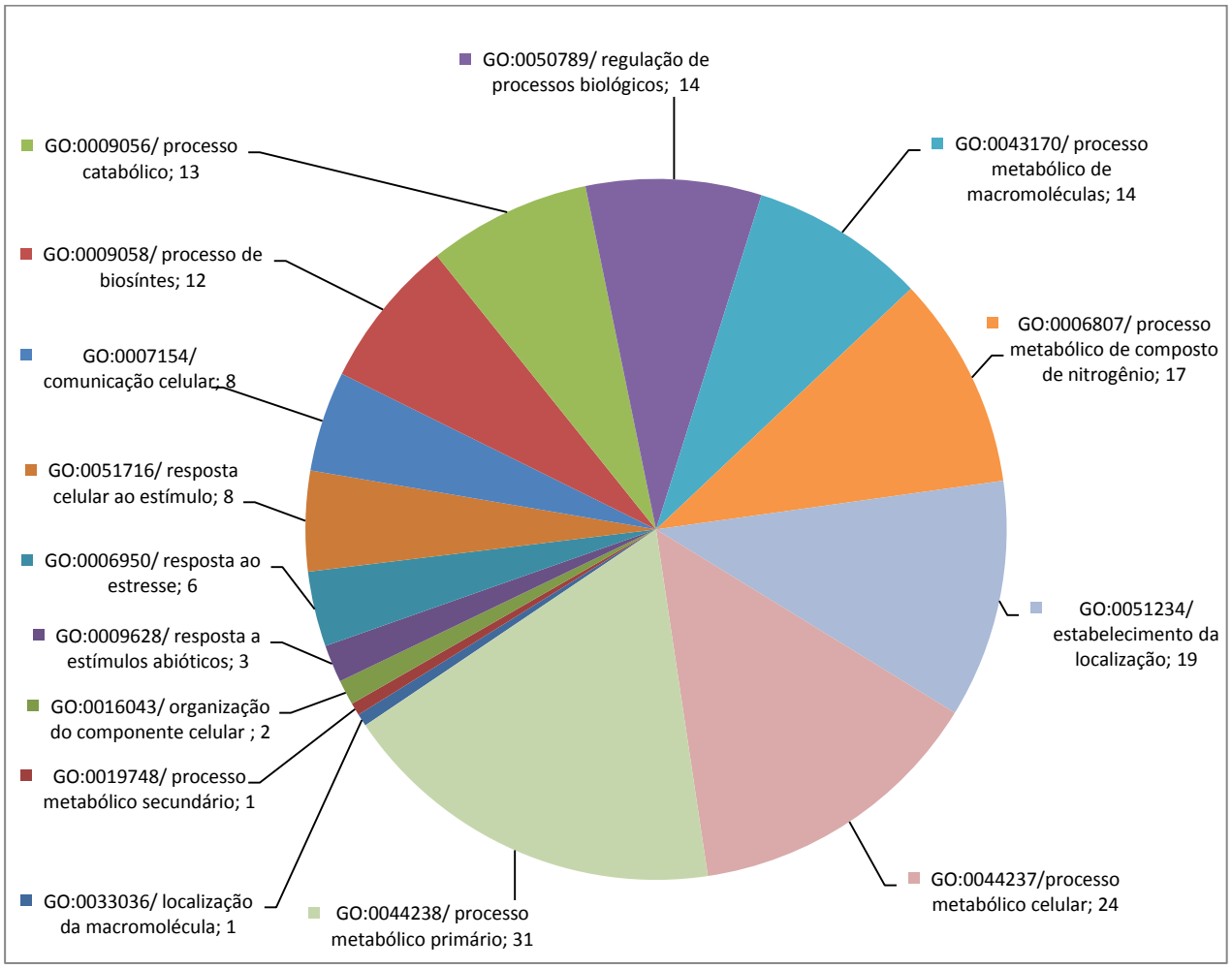


Figura 5: Distribuição dos genes nas diferentes classes GO anotados pelo software Blast2go. Os valores das legendas incluem GO id, termo GO e número de genes pertencentes a cada classe.

A anotação dos genes de acordo com o processo biológico ao qual ele participa permitiu agrupar 57 dos 116 genes sob seleção positiva, em pelo menos uma classe. Vários genes foram agrupados em mais de uma classe indicando que esses genes estão envolvidos em mais de um processo biológico. A relação dos genes que foram agrupados em alguma classe GO, segue na Tabela 7.

Tabela 7: Descrição dos genes sob seleção positiva anotados pela ontologia do gene, para as classes referentes á processo biológico, considerando o nível hierárquico 3.

Grupo de homólogos	**Referencia da sequência NCBI**	**Proteína**	**GO id**
ORTHOMCL 1053	NP_290894.1	yjgP gene product	GO:0051234
ORTHOMCL 1106	YP_001745791.1	LysR family transcriptional regulator	GO:0050789
ORTHOMCL 1134	YP_409167.1	PTS system glucitol/sorbitol-specific transporter subunit IIA	GO:0051716/GO:0050789 GO:0051234
ORTHOMCL 1154	NP_290009.1	rpoH gene product	GO:0050789/GO:0006950 GO:0016043/GO:0009628
ORTHOMCL 1272	NP_286584.1	ybiT gene product	GO:0044238/GO:0044237 GO:0006950/GO:0051234 GO:0009056/GO:0006807 GO:0009628
ORTHOMCL 1404	NP_290830.1	rpsF gene product	GO:0044238/GO:0043170 GO:0044237/GO:0009058
ORTHOMCL 1445	YP_671217.1	regulatory protein CsrD	GO:0044238/GO:0044237 GO:0009058/GO:0006807
ORTHOMCL 1582	NP_753937.1	transcriptional regulator YdhM	GO:0050789
ORTHOMCL 1718	NP_288795.1	ompC gene product	GO:0051234
ORTHOMCL 1952	NP_290204.1	rfaC gene product	GO:0044238/GO:0009058
ORTHOMCL 1992	NP_310342.1	sensor protein RstB	GO:0051716/GO:0050789 GO:0044238/GO:0007154 GO:0043170/GO:0044237
ORTHOMCL 1993	YP_669460.1	DNA-binding transcriptional regulator RstA	GO:0051716/GO:0050789 GO:0007154
ORTHOMCL 210	NP_288600.1	yegT gene product	GO:0051234
ORTHOMCL 2130	NP_709717.1	cpxA gene product	GO:0051716/GO:0050789 GO:0044238/GO:0007154 GO:0043170/GO:0044237
ORTHOMCL 2211	YP_668388.1	phosphate regulon transcriptional regulatory protein PhoB	GO:0051716/GO:0050789 GO:0007154
ORTHOMCL 2512	NP_289689.1	tdcB gene product	GO:0009056
ORTHOMCL 2681	NP_418358.1	stress-induced protein	GO:0006950
ORTHOMCL 2696	NP_416518.2	putrescine importer, low affinity	GO:0051234
ORTHOMCL 2720	YP_543493.1	acetylglutamate kinase	GO:0009058
ORTHOMCL 2788	NP_288382.1	fliZ gene product	GO:0044238/GO:0043170 GO:0044237/GO:0006807
ORTHOMCL 2816	YP_002394341.1	transposase ORF B (fragment), IS911	GO:0044238/GO:0043170 GO:0044237/GO:0006807
ORTHOMCL 2840	YP_001463097.1	IS605 family transposase	GO:0044238/GO:0043170 GO:0044237/GO:0006807
ORTHOMCL 3013	YP_002406506.1	allantoinase	GO:0044238/GO:0044237 GO:0009056/GO:0006807
ORTHOMCL 3058	YP_859139.1	bcsA gene product	GO:0044238/GO:0044237 GO:0006807
ORTHOMCL 306	YP_006141411.1	yjdA gene product	GO:0044238/GO:0044237

			GO:0009056/GO:0006807
ORTHOMCL 3344	NP_415715.1	periplasmic trehalase	GO:0044238/GO:0006950 GO:0009056/GO:0009628
ORTHOMCL 3353	YP_002406703.1	type II citrate synthase	GO:0044238/GO:0044237 GO:0009056
ORTHOMCL 3407	YP_543891.1	DNA-binding transcriptional regulator DsdC	GO:0050789
ORTHOMCL 3444	YP_002331872.1	rpiB gene product	GO:0044238/GO:0044237 GO:0009056/GO:0019748 GO:0006807
ORTHOMCL 3618	YP_671303.1	general secretion pathway protein M	GO:0051234
ORTHOMCL 3685	YP_006136091.1	UDP-galactose LPS alpha1,2-galactosyltransferase WaaW	GO:0044238/GO:0009058
ORTHOMCL 3708	YP_859315.1	yieL gene product	GO:0044238/GO:0009056
ORTHOMCL 3790	YP_670911.1	capsule polysaccharide export protein KpsS	GO:0044238/GO:0051234 GO:0009058
ORTHOMCL 3796	YP_002407084.1	salicyl-AMP ligase YbtE	GO:0050789/GO:0044238 GO:0044237/GO:0051234 GO:0009056/GO:0006807
ORTHOMCL 3898	YP_541169.1	outer membrane protein N	GO:0051234
ORTHOMCL 3923	YP_852503.1	hlyD gene product	GO:0051234
ORTHOMCL 4067	NP_757197.1	carbamate kinase	GO:0009058
ORTHOMCL 4068	YP_543790.1	ornithine carbamoyltransferase	GO:0051716/GO:0050789 GO:0044238/GO:0007154 GO:0043170/GO:0044237 GO:0009056/GO:0009058
ORTHOMCL 422	NP_289467.1	ygfZ gene product	GO:0044238/GO:0044237 GO:0009056/GO:0006807
ORTHOMCL 4326	YP_002410668.1	Type I restriction enzyme EcoAI specificity protein (S protein) (S.EcoAI)	GO:0044238/GO:0043170 GO:0044237/GO:0006807
ORTHOMCL 4335	NP_755501.1	IucB protein	GO:0009058
ORTHOMCL 4383	YP_003227473.1	3-hydroxyphenylpropionic transporter MhpT	GO:0051234
ORTHOMCL 4429	NP_289795.1	cryptic C4-dicarboxylate transporter DcuD	GO:0051234
ORTHOMCL 4430	YP_444138.1	putative invertase	GO:0044238/GO:0043170 GO:0044237/GO:0006807
ORTHOMCL 4475	YP_538670.1	transposase for insertion sequence element IS21	GO:0044238/GO:0043170 GO:0044237/GO:0006807
ORTHOMCL 4525	YP_542179.1	hypothetical protein UTI89_C3199	GO:0033036
ORTHOMCL 454	NP_286775.1	ycaD gene product	GO:0051234
ORTHOMCL 4555	NP_753137.1	phospho-2-dehydro-3-deoxyheptonate aldolase	GO:0009058
ORTHOMCL 472	NP_415654.1	e14 prophage isocitrate dehydrogenase, specific for NADP+	GO:0044238/GO:0044237 GO:0009056
ORTHOMCL 4867	YP_853646.1	integrase-like protein	GO:0044238/GO:0043170 GO:0044237/GO:0006807
ORTHOMCL 4871	YP_002408632.1	putative replication protein P of bacteriophage	GO:0044238/GO:0043170 GO:0044237/GO:0009058

			GO:0006807
ORTHOMCL 5029	NP_415179.1	predicted co-chaperone	GO:0006950
ORTHOMCL 506	YP_540235.1	4-amino-4-deoxychorismate lyase	GO:0009058
ORTHOMCL 5105	YP_541440.1	hypothetical protein UTI89_C2441	GO:0051716/GO:0050789 GO:0007154/GO:0051234
ORTHOMCL 685	YP_854021.1	fucR gene product	GO:0050789/GO:0044238
ORTHOMCL 721	YP_002330892.1	agaB gene product	GO:0051716/GO:0050789 GO:0007154/GO:0051234
ORTHOMCL 899	NP_756072.1	fimbrial chaperone	GO:0016043

A identificação das proteínas correspondentes aos genes sob seleção positiva resultou em 23 proteínas anotadas como *hypothetical protein.* A análise dessas proteínas (Tabela 8) realizada pelo programa SOSUIgranN identificou 14 proteínas com localização subcelular predita como citoplasmática, 1 extracelular (ORTHOMCL 3965) e 2 localizadas na membrana externa (ORTHOMCL 4116 e ORTHOMCL 4177). O programa não foi capaz de predizer a localização subcelular de 6 proteínas.

O programa PHOBIUS identificou a presença de peptídeo sinal em apenas 1 proteína (ORTHOMCL 4116) (Figura 6) indicando que, provavelmente, essa proteína deve ser exportada e atuar em outro compartimento, como o espaço periplasmático ou o ambiente extracelular.

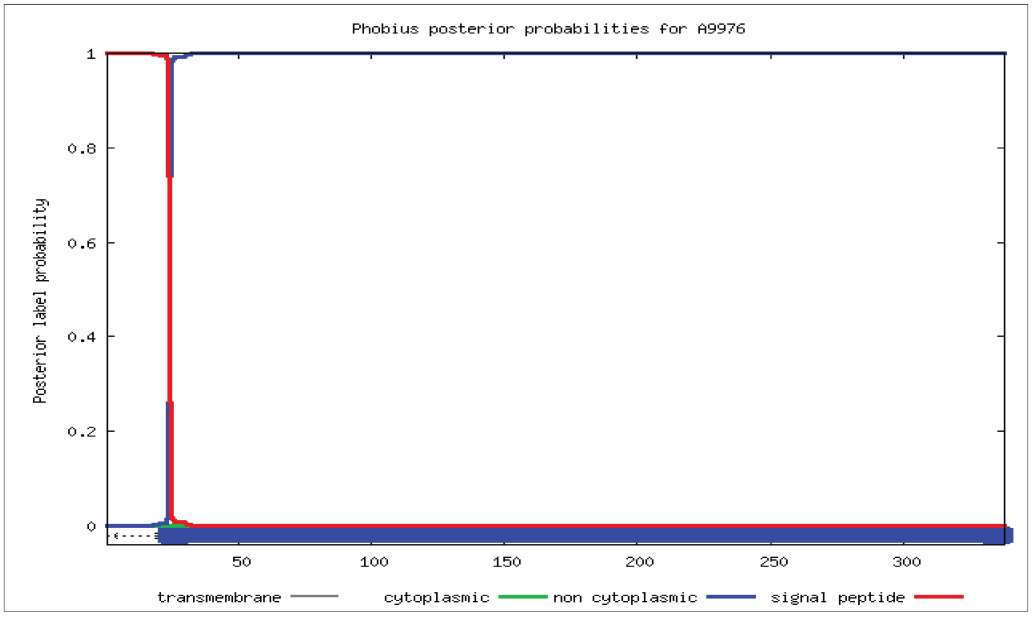


Figura 6: Identificação de peptídeo sinal para uma proteína do grupo de homólogo ORTHOMCL 4116, análise realizada pelo programa PHOBIUS.

Tabela 8: Resultado da predição da localização subcelular das proteínas sob seleção positiva anotadas com *hypothetical protein* realizada pelo programa SOSUIgranN.

Grupo de homólogos	**Referência da proteína-NCBI**	**Localização celular**
ORTHOMCL 3731	NP_755107.1	Citoplasmática
ORTHOMCL 3893	YP_002408630.1	Citoplasmática
ORTHOMCL 3903	YP_002391111.1	Desconhecida
ORTHOMCL 3928	YP_001746332.1	Citoplasmática
ORTHOMCL 3965	YP_002407049.1	Extracelular
ORTHOMCL 4085	YP_002409100.1	Citoplasmática
ORTHOMCL 4099	NP_754496.1	Desconhecida
ORTHOMCL 4116	NP_286235.1	Membrana Externa
ORTHOMCL 4177	YP_006098077.1	Membrana Externa
ORTHOMCL 4178	YP_002409884.1	Desconhecida
ORTHOMCL 4223	NP_757238.1	Citoplasmática
ORTHOMCL 4276	YP_542034.1	Citoplasmática
ORTHOMCL 4361	YP_542368.1	Desconhecida
ORTHOMCL 4363	YP_002408230.1	Citoplasmática
ORTHOMCL 4525	YP_542179.1	Citoplasmática
ORTHOMCL 4531	YP_853478.1	Citoplasmática
ORTHOMCL 4579	YP_853054.1	Citoplasmática
ORTHOMCL 4676	NP_288163.1	Citoplasmática
ORTHOMCL 4753	YP_541853.1	Desconhecida
ORTHOMCL 4850	YP_543034.1	Desconhecida
ORTHOMCL 4949	NP_752542.1	Citoplasmática
ORTHOMCL 4959	YP_001481237.1	Citoplasmática
ORTHOMCL 5105	YP_541440.1	Citoplasmática

5. Discussão

A aquisição de fatores de virulências específicos por alguns clones de *E. coli* permitiu que estas bactérias pudessem ocupar diferentes nichos, resultando em diferentes patótipos responsáveis por causar diversas infecções em humanos (DONNENBERG 2002; KAPER *et al*., 2004) e aves. Além disso, a coexistência de diferentes linhagens patogênicas e comensais, dentro de uma mesma espécie, representa risco iminente de linhagens comensais adquirirem fatores de virulência e de linhagens patogênicas incluírem novos fatores de virulência.

O estudo evolucionário de diferentes linhagens de *E. coli*, por meio da seleção positiva de genes, permite a identificação dos fatores que estão proporcionando uma vantagem adaptativa para essas linhagens. De maneira geral, já foi demonstrado que grande parte dos genes sob seleção positiva apresenta relação íntima com o ambiente e, no caso de linhagens patogênicas, com o hospedeiro (BISHOP *et al*., 2000; FILIP & MUNDY, 2004; HAYDON *et al*., 2001; HUGHES & NEI, 1988; LEHMANN, 2002; POLLEY & CONWAY, 2001; SHACKELTON *et al*., 2005; YANG & SWANSON, 2002). O conhecimento das proteínas cujos genes estão sofrendo pressão de seleção positiva em diferentes linhagens patogênicas pode ajudar no entendimento do processo de patogênese e no desenvolvimento de vacinas e drogas que possam atuar num amplo espectro de linhagens (POLLEY *et al*., 2003).

O estudo de genes homólogos de diversas linhagens permite a identificação das proteínas comuns aos processos de infecção. A inclusão de linhagens APECs em estudos relacionados à seleção positiva é inédita. Os genomas sequenciados para esse trabalho, bem como os resultados apresentados podem ajudar a entender os fatores de virulência de uma espécie bacteriana

amplamente distribuída em diversos ambientes e com diferentes relações ecológicas.

O método de pirossequenciamento utilizado para se obter o sequenciamento dos genomas das quatro linhagens APECs (SEPT362, SCI-07 O8 e S17) iniciou-se pela quebra do DNA total (cromossômico e plasmideal) em milhares de fragmentos, cujas extremidades foram sequenciadas. A montagem do genoma é realizada de forma automatizada, por algoritmos complexos, de maneira que a sobreposição das bases sequenciadas (extremidades dos *reads*) permita a montagem de fragmentos maiores (*contigs*). Para obter uma montagem adequada, o genoma total do organismo deve estar representado diversas vezes pelo conjunto de *reads*, o que é definido pela cobertura obtida. As altas coberturas obtidas, de 21x a 32x, permitiram que praticamente todos os genomas estivessem bem representados pelos conjuntos de *contigs*.

As análises comparativas entre os genomas incluíram apenas a comparação de genes, sendo que nenhuma comparação sintênica foi empregada. Assim, decidiu-se não alocar recursos, humanos e financeiros, no fechamento das lacunas existentes entre dois *contigs* consecutivos, resultando em uma montagem denominada *draft.* No entanto, tal montagem é altamente representativa, de maneira que o resultado do sequenciamento do genoma da linhagem SCI-07 resultou em um artigo cientifico publicado no periódico *Journal of Bacteriology* (ROJAS *et al.,* 2012).

Para o estudo de detecção dos genes sob seleção positiva, os testes realizados com os grupos de genes homólogos permitiram a identificação de 116 grupos sob seleção positiva. A abordagem adotada para o trabalho detectou grupos de genes homólogos sob seleção darwiniana, sem que nenhuma análise linhagem específica fosse efetuada. Tal metodologia foi escolhida porque esses resultados permitem a identificação de genes com mais mutações não sinônimas do que mutações

sinônimas, ou seja, proteínas sofrendo pressão de seleção positiva independentemente do nicho específico de cada linhagem, evidenciando assim, fatores relacionados à adaptação das diferentes estirpes de *E. coli*.

Diversas proteínas relacionadas ao processo de patogenicidade foram identificadas, como por exemplo, proteínas fimbriais e flagelares (PilN*; structural protein; fimbrial-like adhesin protein; minor fimbrial subunit; tail fiber assembly protein; fimbrial chaperone*), proteínas relacionadas ao sistemas de secreção tipo VI (*type VI secretion-associated protein* VC_A0118 *family; aec31 gene product-* ImpA*;* VGR-*related protein; hypothetical protein ECS88_3097 (*IcmF-*related*), porinas (OmpC*; maltose outer membrane porin (maltoporin); outer membrane protein N*), proteínas relacionadas à aquisição de ferro (IucB, UspD; yieL), proteínas relacionadas à manutenção dos plasmídeos (*conjugal transfer pilus assembly protein* TrbC*; conjugal transfer protein* TrbJ), proteínas de adesão (PapG *protein; fimbrial-like adhesin protein*), proteínas de capsula e biogênese de lipídios (KpsS; rlpA; yjgP; rfaC), proteínas relacionadas ao processo de formação de células persistentes (HipA *family,* HipA-*like N-terminal domain family*), além de proteínas regulatórias, proteínas relacionas ao estresse, metabolismo e diversas proteínas cujas funções ainda não são conhecida (*hypothetical protein*).

A comparação dos genes sob seleção positiva desse trabalho, com os outros dois trabalhos que analisaram a seleção positiva em *E. coli* (CHEN, S. *et al.*, 2006; PETERSEN *et al*., 2007) mostrou que em comum para os 3 estudos existe apenas a proteína OmpC, e que os genes *rfac* e *lamB* estavam sob seleção positiva nesse trabalho e no estudo realizado por Petersen e colaboradores (2007). Os demais genes detectadas pelos 2 estudos citados: *fhuA, eaeH, nmpC, ubiF, ompF, ompA, ycgV, yciD, rzpR, tra8_2, yddK, pqqL, nohA, purR, yeeU, yeeV, mdtC,*

innsA_6, wecD, e tra8_3 (PETERSEN *et al*., 2007) e *cedA, yhgA, yidQ, UTI89_C4862, fdnG, ygcR, cycA, argL, yjjN, fruA, agaL, yicM, yjiL, yjbN, topB, xseA, recC, fepE, cutE, ompF, yfaL, amiA, fhuA, entD, entF, yoji, ycdT, yegO* (CHEN, S. *et al.*, 2006), não foram encontrados no presente trabalho, tal fato deve ser resultado da diferente abordagem utilizada na análise realizada por Chen, S. e colaboradores (2006) que verificaram quais genes estavam sob seleção positiva apenas na linhagens UTI89, e ao pequeno número de linhagens analisadas por Petersen e colaboradores (2007).

A inclusão de um número maior de linhagens nesse trabalho (13) em comparação com o número de linhagens analisadas nos outros estudos (7) (CHEN, S. *et al.*, 2006) e (6) (PETERSEN *et al*., 2007), dos genomas dos plasmídeos e de linhagens APEC, permitiu a obtenção de mais grupos de genes homólogos e a inclusão de um número maior de genes nos diferentes grupos, resultando em um valor diferente para ω, uma vez que a taxa de d_N/d_S sofre alterações de acordo com o tamanho da amostra. A análise de genes sob seleção positiva com um número alto de amostras (13) é algo novo e talvez não tenha sido realizado anteriormente devido à necessidade de se possuir um processo automatizado, com computadores com alta capacidade de processamento e à dificuldade da análise de uma grande quantidade de informações geradas durante todo o fluxo do trabalho.

A anotação dos genes de acordo com a ontologia (GO) considerando o processo biológico ao qual ele participa, permitiu incluir 57 genes em pelo menos uma das 15 classes: resposta celular ao estímulo, localização da macromolécula, regulação de processos biológicos, processo metabólico primário, comunicação celular, processo metabólico de macromoléculas, processo metabólico celular, resposta ao estresse, estabelecimento da localização, processo catabólico,

processo metabólico secundário, processo de biossíntese, processo metabólico de composto de nitrogênio, organização do componente celular e resposta a estímulos abióticos.

Diversas proteínas foram incluídas em mais de uma classe, o que indica que ela participa de mais de um processo biológico, podendo atuar nesse processo por meio do seu produto gênico ou regulação. Como exemplo podemos citar a proteínas CpxA, que atua nos processos: resposta celular ao estímulo, regulação de processos biológicos, processo metabólico primário, comunicação celular, processo metabólico de macromoléculas e processo metabólico celular.

A classe mais enriquecida, ou seja, aquela que englobou mais genes foi a classe processo metabólico primário, com 31 genes, indicando que os a maioria dos genes anotados de acordo com a GO estão relacionados a processos metabólicos primários, como processos anabólicos e catabólicos. Tal achado era esperado uma vez que a bactéria necessita ajustar seu metabolismo para a sobrevivência dentro do hospedeiro. Genes, cujo produto, apresenta função enzimática, também foram detectados nos estudos realizados por Petersen e colaboradores (2007), indicando que a seleção também deve estar atuando na resposta adaptativa e em respostas às alterações do ambiente.

A classe processo metabólico celular, englobou 24 genes, responsáveis pela transformação de substâncias químicas apenas em células específicas, indicando assim que mudanças específicas do metabolismo bacteriano acabam proporcionando uma vantagem adaptativa para as bactérias dentro do hospedeiro. A terceira classe, em número de genes, estabelecimento da localização, incluiu 19 genes, que atuam no movimento dirigido, tanto da célula como de composto para uma determinada localização. O movimento dirigido é importante para processos de secreção de moléculas efetoras, as quais podem atuar como fatores de virulência. Diversas porinas fazem

parte dessa classe, como OmpC e HlyD, evidenciando a importância dessa classe de proteínas para a sobrevivência da bactéria dentro do hospedeiro.

Com 17 genes a classe processo metabólico de composto de nitrogênio, também está bem representada, elucidando a importância do nitrogênio na síntese e decomposição de DNA, RNA e proteínas, para a sobrevivência da bactéria (TORTORA *et al*., 2006).

As classes regulação de processos biológicos, processo metabólico de macromoléculas, processo catabólico e processo de biossíntese englobaram de 14 a 12 genes, mostrando novamente que grande parte dos genes sob seleção positiva atua no metabolismo, seja na regulação, biossíntese ou catabolismo, ilustrando novamente que alguns genes relacionados aos processos metabólicos ou no ajuste do metabolismo da bactéria proporcionam uma vantagem adaptativa.

Algumas proteínas identificadas, cujos genes estão sob seleção positiva, foram discutidas individualmente devido à relevância dessas, como fatores de virulência, fatores regulatórios, porinas, adesinas, fatores relacionados à produção de biofilme, etc.

A proteína OmpC, comum em todos os estudo, é uma porina trimérica com um domínio que forma uma estrutura em forma de barril. Esta porina fica localizada na membrana externa da parede celular e permite a passagem de pequenas moléculas como íons e solutos hidrofílicos (NIKAIDO, 1994, 1996; SCHIRMER, 1998). OmpC é uma das proteínas mais abundantes em *E. coli* (NIKAIDO, 1996) e já foi demonstrado que está significativamente mais expressa durante o processo de infecção do trato urinário por *E. coli* (NIKAIDO, 1996). O alto nível de expressão desta proteína deve colaborar para a sua imunogenicidade. Além disso, a alta expressão de OmpC *in vivo* pode ser uma vantagem em ambientes onde pequenas moléculas de toxinas são

abundantes, uma vez que o poro formado por essa proteína também serve como caminho de saída para toxinas (NIKAIDO, 1996). Essa proteína é regulada pelo sistema de resposta ao estresse do envelope CpxA/CpxR (BATCHELOR *et al.*, 2005), sendo que o gene da proteína CpxA também mostrou evidências de seleção positiva no presente estudo. Outras características importantes dessa proteína é que a porina formada pela OmpC também funciona como receptor para vários fagos (YU *et al*., 2000) e que ela é importante para o mecanismo de resistência à tetraciclina e ao ácido nalidíxico em *E. coli,* de maneira que o aumento dos níveis de sua expressão está diretamente relacionado ao aumento da resistência a esses antibióticos (LIN *et al*., 2008; ZHANG *et al.*, 2008).

A proteína ompN, identificada aqui como *outer membrane protein N*, apresenta características bioquímicas e funcionais muito similares à OmpC. No entanto, essa porina não apresentou altos níveis de expressão em laboratório (PRILIPOV *et al*., 1998) e, até então, ainda não se sabe quais fatores ou condições estão relacionadas ao aumento de sua expressão.

Outra proteína de membrana sob seleção positiva é a porina maltoporina (*maltose outer membrane porin*) cuja função é facilitar a difusão de maltodextrinas (WANG *et al*., 1997). O nome do gene que codifica essa porina é *lamB*, e faz referência ao fago lambda por atuar como receptora para o mesmo (RANDALL-HAZELBAUER *et al*., 1973). Da mesma maneira que a porina OmpC, os níveis de expressão dessa maltoporina também estão relacionados à resistência à tetraciclina. No entanto, nesse caso, o mecanismo funciona de forma oposta, sendo que a diminuição da expressão do gene *lamB* aumenta a resistência à tetraciclina (ZHANG *et al.*, 2008).

Também formadora de poro a proteína HlyD faz parte da família MFP (*membrane fusion*

protein), cuja função é ligar a membrana interna à externa (PIMENTA *et al*., 1999). Essa família de proteínas está envolvida no transporte de diversos fatores desde pequenas moléculas até grande proteínas, como a hemolisina HlyA. Essa hemolisina é uma citoxina secretada por *E. coli* que forma poros na membrana da célula alvo, causando oscilações no níveis de Ca^{2+} intracelular (KOSCHINSKI *et al*., 2006) e lise das células de eritrócitos (BHAKDI *et al*., 1988).

Localizado também na membrana da bactéria, o sistema de secreção tipo VI (SST6), descrito em 2006 por dois grupos distintos (MOUGOUS, 2006; PUKATZKI, 2006), ainda não teve sua função e a descrição dos fatores que são secretados por esse sistema, totalmente elucidados, embora se acredite que o SST6 possa estar envolvido no processo de patogenicidade de uma das linhagens aqui estudadas (SEPT362) (DE PACE *et al*., 2010, 2011). Além disso, sabe-se que as proteínas Hcp (*Haemolysin- Coregulated Protein*) e VgrG (*Val-Gly Repeats*) são secretadas, que a estabilização na membrana é feita pela proteína IcmF (*intracellular multiplication F*) e que ClpV (*caseinolytic protease V*) é a ATPse do sistema (FILLOUX *et al*., 2008).

Hcp é uma proteína que forma hexâmetros e acredita-se que essas estruturas formam nanotubos na membrana bacteriana que permitem a passagem de proteínas, incluindo VgrG para o meio extra celular. (BALLISTER *et al*., 2008). Já a proteína VgrG pode ser dividida em 3 diferentes domínios: na região N-terminal da proteína estão localizados os dois primeiros domínios que tem homologia com proteínas que constituem a cauda de um bacteriófago e na região C-terminal, existente em apenas parte dos VgrGs, está localizado um domínio que apresenta homologia com adesinas, proteases ou proteínas que se ligam à actina (LEIMAN *et al*., 2009; PUKATZKI, 2006).

Diversos estudos demonstraram a importância desse sistema na interação patógeno

hospedeiro. Em *Burkholderia mallei*, a proteína Hcp foi produzida durante ensaio i*n vivo* em um modelo de infecção animal e o SST6 foi requerido para sua proliferação em macrófagos (SCHELL *et al*., 2007). Em *Pseudomonas aeruginosa*, um patógeno humano, foram detectados anticorpos contra Hcp em pacientes com fibrose cística, o que mostra que o sistema estava ativo durante a infecção sendo reconhecido pelo sistema imunológico do hospedeiro (MOUGOUS *et al*., 2006). Na linhagem APEC SEPT362, de Pace e colaboradores (2010) mostraram que o mutante Δhcp apresentou diminuição na adesão e invasão de células epiteliais *in vitro* e nenhum dos mutantes, Δhcp e ΔclpV, foram capazes de se replicarem dentro de macrófagos, o que diminuiu a capacidade desses mutantes causarem infecções.

Sabe-se, também, que o SST6 varia muito na composição dos genes, no número de conjuntos existentes em diferentes espécies e linhagens (BOYER *et al*., 2009) e na sequência de DNA dos genes que codificam para a mesma proteína.

No presente estudo foram detectadas 4 proteínas do SST6 cujos genes estão sob seleção positiva: *type VI secretion-associated protein VC_A0118 family; aec31 gene product- impA; VGR-related protein; hypothetical protein ECS88_3097*. A proteína *VGR-related protein* possui em sua sequencia os domínios conservados homólogos às proteínas de bacteriófago, como descrito acima, e na sua região C-terminal um domínio conservados, já encontrado em diversas proteínas semelhantes, mas com função ainda desconhecida. O VgrG está intimamente em contato com o hospedeiro/ambiente, uma vez que se não for ela a proteína efetora, ela tem a função de inserir a proteína efetora na célula alvo e, dessa maneira, deve sofrer pressão de seleção constantemente, uma vez que o ambiente em que ela se encontra não é tão constante quanto o interior da célula bacteriana, explicando assim a alta taxa de mutações não sinônimas

encontrada.

A proteína *hypothetical protein ECS88_3097* apresenta homologia com a proteína IcmF. Localizada na membrana interna da parede das bactérias gram negativas essa proteína atua na ancoragem do aparato e possui vários domínios transmembrânicos. Diversos estudos já identificaram sua relação próxima com DoutU, outra proteína do SST6, sugerindo que essas proteínas trabalham juntas (SEXTON *et al*., 2004; VANRHEENEN *et al.*, 2004). VC_A0118 e ImpA também apresentaram evidencias de seleção positiva. Sabe-se que seus genes estão sempre agrupados com os demais genes do SST6, mas não se tem conhecimento da forma como elas atuam no aparato, se através da regulação ou formando a estrutura do sistema.

Bactérias patogênicas possuem estruturas adesivas que se ligam à célula do hospedeiro para iniciar uma infecção (LI *et al*., 2010). *E. coli* uropatogênicas usam a fimbria P para aderirem a receptores da bexiga (ROBERTS *et al*., 1994). Em APEC, a fímbria P não é requerida no estágio inicial da infecção, mas deve ser importante em estágios posteriores, uma vez que está presente em bactérias que se ligam a órgãos internos, mas não à traqueia (DZIVA & STEVENS, 2008).

A fimbria P é codificada por um conjunto de genes cromossômicos denominado conjunto pap. Essa fimbria é formada por 6 proteínas diferentes que se conectam uma à outra variando no número de cópia de cada uma (MU & BULLITT, 2006). A proteína PapG, sob seleção positiva, está localizada no final do pili e se liga a receptores específicos (LUND *et al*., 1987). A intima relação da proteína PapG com seu hospedeiro mostra que é necessária a constante adaptação dessa proteína a seu meio ambiente.

Ainda relacionado ao contato direto da bactéria com o ambiente, as proteínas da membrana externa sofrem constante pressão de seleção. Os produtos dos genes *rlpA, yjgP* e *rfaC,* sob

seleção positiva, possuem relação com o processo de biossíntese do polissacarídeo da membrana externa da bactéria. Além disso, a proteína KpsS, responsável pela biossíntese do polissacarídeo presente na cápsula da bactéria, também foi detectada como estando sob seleção positiva. Sabe-se que a cápsula aumenta a resistência à fagocitose e a capacidade de resistir aos efeitos bactericidas do soro (RUSSO *et al*., 1993) sendo importante para a capacidade da bactéria colonizar a bexiga, o rim e a urina (BAHRANI-MOUGEOT *et al*., 2002).

Localizado no envelope celular de *E. coli*, o sistema Tol/Pal atua na manutenção da integridade do envelope celular e no transporte de componentes pelo espaço periplasmático. Além disso, seu poro serve de entrada para bacteriófagos filamentosos e para colicina A (CASCALES *et al.*, 2002; WALBURGER *et al*., 2002).

Esse complexo é formado por proteínas codificadas por dois operons, sendo que o gene *ygbF*, sob seleção positiva, está presente nesses dois operons. Esse gene é uma proteína periplasmática que interage diretamente com a proteína TolA, localizada na membrana externa.

A pressão de seleção positiva sob *ygbF* pode estar relacionada ao aumento de sua eficiência na interação com TolA. A manutenção do envelope celular é indispensável para a sobrevivência da bactéria em ambientes com diversos fatores de estresse, como dentro do hospedeiro. Além disso, esse sistema pode atuar também como mecanismo de defesa da bactéria, uma vez que a colicina A interage diretamente com TolA.

A competência natural, a aquisição de DNA exógenos como fonte de carbono, a formação de células persistentes e a adesão célula a célula são processos que ocorrem nas bactérias quando elas estão organizadas em biofilme, caracterizando assim, um mecanismo de sobrevivência e de virulência, ou seja, mecanismos relacionados à adaptação da bactéria ao ambiente.

A proteína PilN, sob seleção positiva, é homóloga à proteína HofN da *E. coli* K-12 substr. MG1655. Embora sua função seja ainda desconhecida, esse gene faz parte do operon relacionado à aquisição de DNA exógeno e sua utilização como fonte de carbono em ambientes com restrições de nutrientes. Este processo está relacionado à competência natural de *E. coli*, um importante mecanismo para a transferência horizontal de genes. Além disso, os genes relacionados a esse operon, também conferiram uma vantagem competitiva para *E. coli* incubadas por longo período durante a fase estacionária (PALCHEVSKIY & FINKEL, 2006). A competência natural é um processo que também ocorre frequentemente quando as bactérias estão organizadas em biofilme.

A formação de células persistentes é mediada por um complexo toxina – antitoxina. A proteína HipA (*high persistence*) é a toxina do complexo toxina - antitoxina hipAB e está sob seleção positiva. Esse complexo está relacionado à formação das células persistentes, processo no qual uma pequena fração das células bacterianas de uma população (aproximadamente de 10^{-5} a 10^{-6} do total) entram num período de dormência, diminuindo as taxas metabólicas, tornando-as resistentes a antibióticos, independentemente da classe (JAYARAMAN *et al.*, 2008). As células persistentes estão relacionadas a processos de recidiva de infecções bacterianas devido à capacidade dessas bactérias restabelecerem uma nova população após o período de estresse (KUSSELL *et al*, 2005).

Outro mecanismo relacionado à adaptação da bactéria ao ambiente/hospedeiro é a capacidade de aquisição de ferro pelas bactérias dentro do hospedeiro, pois as bactérias dependem do íon férrico para crescerem e se multiplicar. O ferro está envolvido em diferentes processos como síntese de DNA, co-fator enzimático, síntese de ATP, entre outros. Entretanto, apesar da

abundância desse elemento na natureza, o ferro não se encontra livre nos sistemas biológico (LORENZO *et al*., 1986). A proteína IucB, responsável pela síntese da aerobactina, um sideróforo responsável pela aquisição de ferro dos fluídos do hospedeiro, está sob seleção positiva. Um estudo realizado por Xiong e colaboradores (2012) mostrou que o mutante ΔiucB de uma linhagem APEC, apresentou uma redução da contagem das bactérias nos órgãos de frango com três semanas de vida, mostrando que a colonização pelo mutante foi significativamente mais baixa do que a da bactéria selvagem. O ferro é um fator limitante para a sobrevivência das bactérias, existindo diferentes sistemas de aquisição de ferro, e muitas vezes a presença de um acaba suprindo a ausência de outro, mas, de maneira geral, acredita-se que genes de aquisição de ferro são fatores importantes para a virulência de linhagens APEC e causadoras de infecção no trato urinário (DHO & LAFONT, 1984; SNYDER *et al*., 2004).

A proteína UspD (*stress-induced protein*) possui sua expressão elevada em condições de baixa oferta de nutrientes e em outras condições de estresse. Essa proteína também atua na proteção da célula contra danos no DNA, como aqueles causados pela exposição à luz ultravioleta (GUSTAVSSON *et al*., 2001). Um estudo realizado por Nachin e colaboradores (2005) demonstrou que mutantes para o gene *uspD* foram mais suscetíveis ao estresse oxidativo. Além disso, a ausência deste gene aumentou a sensibilidade ao antibiótico estreptonigrina, cuja toxidade depende da concentração de ferro intracelular, demonstrando que a deleção do gene *uspD* causou um aumento do influxo de ferro nas células. Este experimento concluiu que a proteína UspD atua no metabolismo do ferro, um importante fator de virulência.

Além das proteínas IucB e UspD, outra proteína relacionada ao metabolismo do ferro também foi detectada: a enterobactina codificada pelo gene *yieL*, cuja função exata ainda não é

conhecida.

Ainda ligada à capacidade de sobrevivência da bactéria em diferentes micro-ambientes, a ativação de um sistema de resposta às alterações ambientais é necessária. Essa resposta pode ser mediada por um sistema regulatório de dois componentes, composto por uma proteína sensória do tipo histidina quinase ligada à membrana interna e por uma proteína de ligação ao DNA, responsável pela expressão ou repressão dos genes controlados por este regulão (WEST & STOCK, 2001).

Sabe-se que o fósforo é encontrado em diversas moléculas biológicas como lipídios de membrana, ácidos nucleicos e açúcares, sendo assim imprescindível para a vida bacteriana à semelhança de outros seres vivos (WANNER, 1960). A baixa concentração desse elemento, no meio ambiente, ativa o sistema de sinalização e regulação, composto pelo PhoB e PhoR em bactérias.

A proteína PhoB (*phosphate regulon transcriptional regulatory protein PhoB*), que é a proteína de ligação ao DNA do sistema regulatório de sensibilidade ao fósforo, apresentou evidências de seleção positiva em sua sequencia codificante. A ativação do PhoB regula a expressão de 31 genes pertencentes a diversos processos relacionados à sobrevivência da bactéria em condições de estresse, inclusive de fatores de virulência (HSIEH & WANNER, 2010). O operon *pst* (*phosphate-specific transport*), responsável pelo transporte de diversas categorias de fosfato inorgânico, também é regulado pelo PhoB. O gene desse operon, responsável pela síntese da proteína *PTS system glucitol/sorbitol-specific transporter subunit IIA,* também apresentou evidências de seleção positiva.

Na linhagem APEC χ 7122, o nível da ativação de PhoB estava diretamente relacionado ao

nível de atenuação da infeção dessa linhagem (BERTRAND *et al*., 2010). No entanto, um estudo anterior realizado por Dozois e colaboradores (2003) demonstrou que PhoB estava expresso durante a infecção em um frango por esta mesma linhagem, sugerindo assim que a regulação fina de PhoB é necessária para a ativação de fatores de virulência.

A ativação de PhoB pode alterar a expressão de genes ligados a diversos fatores de virulência e estresse, principalmente adesinas. PhoB é importante não apenas para controlar os níveis de fosfato, mas também por regular a expressão de inúmeros genes relacionados a condição de estresse, sendo que a expressão inapropriada ou a repressão desse sistema tem consequências direta na virulência da bactéria (CRÉPIN *et al*., 2011).

Outro sistema regulatório de dois componentes que estão presentes em *E. coli* é composto pela proteínas PhoQ e PhoP, que respondem aos níveis extracelular de Mg^{2+} (MINAGAWA. *et al*., 2003). Sob controle desse sistema está o operon *rstAB* o qual apresentou evidencias de seleção positiva (*sensor protein RstB e DNA-binding transcriptional regulator RstA*).

O aumento nos níveis de expressão do operon *rstAB* reprime a expressão do genes *csgD*, que codifica a principal proteína regulatória do processo de formação da fimbria *curli* (OGASAWARA *et al*., 2007). A expressão da fimbria *curli* está sujeita a ação de outros reguladores, como repressor transcricional FliZ, o qual, também sofre pressão de seleção positiva. Sabe-se que a fimbria *curli* é importante para os processos de adesão, invasão e formação de biofilme (BROMBACHER *et al*., 2003; BROWN *et al*., 2001), sendo um fator expresso em condições de estresse, como a baixa oferta de nutrientes, por exemplo.

Diversos fatores relacionados ao contato diretos da bactéria com o ambiente/hospedeiro e proteínas que fazem parte do processo de adaptação da bactéria às condições de estresses foram

detectados, corroborando os dados já obtidos para diversos estudos de genes sob seleção positiva para diferentes organismos (BISHOP *et al*., 2000; FILIP & MUNDY, 2004; HAYDON *et al*., 2001; HUGHES & NEI, 1988; LEHMANN, 2002; POLLEY & CONWAY, 2001; SHACKELTON *et al*., 2005; YANG & SWANSON, 2002) e com o processo de co-evolução existente entre hospedeiro e patógeno (VAN VALEN, 1973), o qual acaba selecionando positivamente diversos genes que codificam para as proteínas envolvidas no processo de patogênese

A investigação e discussão da função das principais proteínas, cujos genes estão sob seleção positiva, foi descrita anteriormente. O entendimento dos demais produtos gênicos ainda precisa ser feito para as demais proteínas. No entanto, a falta de informações e o grande número de fatores desconhecidos requerem uma análise mais minuciosa.

As análises realizada nas proteínas sob seleção positiva, identificadas como *hypothetical protein* representam o início do estudo da possibilidade de atuarem como fatores de virulência. Muitos desses fatores são extracelulares, podendo ser secretados para o meio extra celular ou estarem ligados à membrana externa. A presença de peptídeo sinal numa determinada proteína indica que ela deve exercer seu papel biológico em outro compartimento, e assim, deve ser exportada. Diversas proteínas já identificadas como fatores de virulência, estão localizadas na membrana externa, como flagelo, fímbrias, adesinas, porinas, proteínas autotransportadoras, proteínas de sistema de secreção, etc.

A combinação dos resultados das análises de predição de peptídeo sinal e localização subcelular identificaram uma proteína (*hypothetical protein* Z0639) que poderia atuar como um possível fator de virulência, pois ela está sofrendo pressão de seleção positiva, possui peptídeo

sinal e teve sua localização subcelular predita na membrana extracelular. Adicionalmente, as análises revelaram mais uma proteína localizada na membrana externa (*hypothetical protein* EC042_3791) e outra com localização extracelular (*hypothetical protein* ECIAI39_1028), as quais poderão ser alvo de trabalhos futuros, enriquecendo o estudo sobre os processos de patogênese e a combinação de fatores de virulência para as diferentes linhagens de *Escherichia coli*.

6. Conclusões

As análises empregadas nas sequencias codificantes dos genomas analisados permitiram a detecção de cento e dezesseis grupos de genes homólogos sob seleção positiva. A identificação das proteínas correspondentes encontrou diversos fatores de virulência de linhagens patogênicas de *Escherichia coli* já amplamente descritos, confirmando a premissa de que fatores de virulência, normalmente, proporcionam uma vantagem adaptativa aos patógenos, já que a expressão dos mesmos aumenta as chances do patógeno colonizar, se multiplicar e, portanto, sobreviver dentro do hospedeiro.

A anotação dos genes de acordo com a ontologia (GO) permitiu verificar que a classe mais enriquecida está relacionada ao metabolismo primário, indicando que muitos dos genes sob seleção positiva atuam no metabolismo. Dessa maneira, verifica-se que o ajuste metabólico bacteriano de maneira adequada é determinante para a sobrevivência da bactéria dentro do hospedeiro.

Diversas proteínas ainda não conhecidas também apresentaram evidências de seleção positiva. O estudo dessas proteínas pode identificar fatores de virulências ainda não descritos, aumentando, assim, o conhecimento sobre a interação patógeno-hospedeiro. Nesse trabalho foram identificadas três proteínas desconhecidas (*hypothetical protein* ECIAI39_1028, *hypothetical protein* Z0639 e *hypothetical protein* EC042_3791) com características comuns a fatores de virulência e que, por isso, deverão ser alvo de estudos futuros.

7. Referências bibliográficas

AGUILETA, G. et al. **Finding candidate genes under positive selection in Non-model species: examples of genes involved in host specialization in pathogens**. Molecular ecol., 19(2), 292–306, 2010.

AGUILETA, G. et al. **Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists**. Infect Genet Evol. Jul;9(4):656-70, 2009.

ANDERSON, G. G. et al. **Intracellular bacterial biofilmlike pods in urinary tract infections**. Scien. 301, 105–107, 2003.

ANDRADE, A. et al. **The iron uptake mechanisms of enteroinvasive *Escherichia coli***. Braz. J. Microbiol. v. 31, p. 200-2005, 2000.

AZIZ, R. K. et al. **The RAST Server: Rapid Annotations using Subsystems Technology** BMC Genom., 2008.

BAHRANI-MOUGEOT, F. K. et al. **Type 1 fimbriae and extracellular polysaccharides are preeminent uropathogenic *Escherichia coli* virulence determinants in the murine urinary tract.** Mol. Microbiol. Aug;45(4):1079-93, 2002.

BALLISTER, E. R. et al. **In vitro self-assembly of tailorable nanotubes from a simple protein building block**. Proc. Natl. Acad. Sci. U S A. Mar 11;105(10):3733-8, 2008.

BATCHELOR, E. et al. **The *Escherichia coli* CpxA-CpxR envelope stress response system regulates expression of the porins ompF and ompC**. J. Bacteriol. Aug;187(16):5723-31, 2005.

BERGHOLZ, T. M. et al. **Transcriptional response of *Escherichia coli* O157:H7 to growth transitions in glucose minimal medium**. BMC Microbiol. Oct 29;7:97, 2007.

BERTRAND, N. et al. **Increased Pho regulon activation correlates with decreased virulence of an avian pathogenic *Escherichia coli* O78 strain**. Infect. Immun. 78, 5324–5331, 2010.

BHAKDI, S. et al. **The hemolysin of *Escherichia coli***. Eur J. Epidemiol. 4:135–143, 1988.

BISHOP, J. G.; DEAN, A. M.; MITCHELL-OLDS, T. **Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution.** Proc. Natl. Acad. Sci. U S A, May 9;97(10):5322-7, 2000.

BLATTNER, F. R. et al. **The complete genome sequence of *Escherichia coli* K-12.** Scien. Sep 5;277(5331):1453-62, 1997.

BOYER, F. et al. **Dissecting the bacterial type VI secretion system by a genome wide in silico analysis: what can be learned from available microbial genomin resources?** BMC genom. v.10, p.104, 2009.

BRAZILIANCHICKEN**.** Disponível em:<.http://www.brazilianchicken.com.br/industria-avicola/o-frango-pelo-mundo.php >. Acesso em: 20, Ago. 2012.

BROMBACHER, E. et al. **The curli biosynthesis regulator CsgD co-ordinates the expression of both positive andnegative determinants for biofilm formation in *Escherichia coli***. Microbiol. 149:2847–2857, 2003.

BROWN, P. K., et al. **MlrA, a novel regulator of curli (AgF) and extracellular matrix synthesis by *Escherichia coli* and Salmonella enterica serovar Typhimurium**. Mol. Microbiol. 41:349–363, 2001.

CAMPOS, T. A. et al. **Adhesion properties, fimbrial expression and PCR detection of adhesin-related genes of avian *Escherichia coli* strains.** Vet. Microbiol., v. 106, p. 275-285, 2005.

CAPELLA-GUTIÉRREZ, S.; SILLA-MARTÍNEZ, J. M.; GABALDÓN, T. **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses**. Bioinform.1;25(15):1972-3, 2009.

CASCALES, E. et al. Pal **lipoprotein of *Escherichia coli* plays a major role in outer membrane integrity.** J. Bacteriol. Feb;184(3):754-9, 2002.

CHEN, F. et al. **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups**. Nucleic Acids Res. 34: D363-8, 2006.

CHEN, S. L. et. al. **Identification of genes subject to positive selection in uropathogenic of *Escherichia coli*: a comparative genome approach**. Proc. Natl. Acad. Sci. U.S. 103, 5977-5982, 2006.

CHIMPANZEE SEQUENCING AND ANALYSIS CONSORTIUM. **Initial sequence of the chimpanzee genome and comparison with the human genome.** Natur. Sep1;437(7055):69-87, 2005.

CHISHOLM, S. T. et al. **Host-microbe interactions: shaping the evolution of the plant immune response**. Cell, Feb 24;124(4):803-14, 2006.

COLE, S. T. et al. **Deciphering the biology of Mycobacterium turbeculosis from the complete genome sequence**. Natur., Nov. 12;396(6707):190, 1998.

COLLINSON, S. K. et al. **Thin, aggregative fimbrae mediate binding of *Salmonella enteritidis* to finectin.** J. Bacteriol., (1):12-8, 1993.

CONESA, A. et al. **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. Bioinformatics. Sep 15;21(18):3674-6, 2005

CRÉPIN, S. et al. **The Pho regulon and the pathogenesis of *Escherichia coli***. Vet Microbiol. Nov 21;153(1-2):82-8, 2011.

CROXEN, M. A.; FINLAY, B. B. **Molecular mechanisms of *Escherichia coli* pathogenicity**. Nat. Rev. Microbiol. Jan;8(1):26-38, 2010.

DE PACE, F. et al. **Characterization of IcmF of the type VI secretion system in an avian pathogenic *Escherichia coli* (APEC) strain**. Microbiol. Oct;157(Pt 10):2954-62, 2011.

DE PACE, F. et al. **The type VI secretion system plays a role in type 1 fimbria expression and pathogenesis of an avian pathogenic *Escherichia coli* strain**. Infect. Immun. Dec;78(12):4990-8, 2010.

DHO, M.; LAFONT, J. P. **Adhesive properties and iron uptake abilities in *E. coli* lethal and non-lethal for chicks**. Avian Dis., v. 28, p.1016-1025, 1984.

DHO-MOULIN, M.; FAIRBROTHER, J. M. **Avian pathogenic *Escherichia coli* (APEC).** Vet. Res., v. 30, p 299- 316, 1999.

DODRINDT, U.; HACKER, J. **Whole genome plasticity in pathogenic bacteria.** Curr. Opin. Microbiol., (5):550-7, 2001.

DONNENBERG, M. S. ***Escherichia coli*** **Virulence mechanisms of a versatile pathogen**. Baltimore: Academic press, Elsevier Science, 2002.

DOZOIS, C. M. et al. **pap- and pil related DNA sequences and other virulence determinants associated with *Escherichia coli* isolated from septicemic chickens and turkeys.** Infect. Immun., v. 60, p. 2648-2656, 1992.

DRASAR, B. S.; HILL, M. J. **Human intestinal flora.** Academic Press, Ltda., London, United Kingdom, p. 36-43, 1974.

DZIVA, F.; STEVENS, M. P. **Colibacillosis in poultry: unravelling the molecular basis of virulence of avian pathogenic *Escherichia coli* in their natural hosts.** Avian Pathol. 37, 355-366, 2008.

EDGAR, R. C. **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** Nucleic Acids Res., 32(5), 1792-97, 2004.

EMERY, D. A. et al. **Virulence factors of *Escherichia coli* associated with colisepticemia in chicken and turkeys**. Avian Dis., v. 36, p. 504-511, 1992.

EWERS, C. et al. **Molecular epidemiology of avian pathogenic *Escherichia coli* (APEC) isolated from olisepticemia in poultry.** Vet. Microbiol., v. 104, p. 91-101, 2004.

FELDGARDEN, M;, BYRD, N.; COHAN, F. M. **Gradual evolution in bacteria: evidence from *Bacillus systematics***. Microbiol. 149:3565–3573, 2003.

FILIP, L. C., MUNDY, N. I. **Rapid evolution by positive Darwinian selection in the extracellular domain of the abundant lymphocyte protein CD45 in primates.** Mol. Biol. Evol. Aug;21(8):1504-11, 2004.

FILLOUX, A.; HACHANI, A.; BLEVES, S. **The bacterial type VI secretion machine: yet another player for protein transport across membranes.** Microbiol. v.154, p.1570-1583, 2008.

FLEISCHMANN, R. D. et al. **Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd.** Scien.**,** v. 269, n. 5223 p. 496-512. 1995.

FOSTER, J. et al. **The Wolbachia genome of *Brugia malayi*: endosymbiont evolutionwithin a human pathogenic nematode.** PLoS Biol. 3:e121, 2005.

GARDNER, M. J. et al. **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** Natur., v. 419, p. 498-511, 2002a.

GARDNER, M. J. et al. **Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14**. Natur., v. 419, p. 531-534, 2002b.

GIBAS, C.; JAMBECK, P. **Developing Bioinformatics Computer Skills**, First Edition, O'Reilly ; Associates, Inc., 2001.

GROSS, W. B. **Diseases due to *Escherichia coli* in poultry** In: Gyles C.L. (Ed.), *Escherichia coli* in Domestic Animals and Humans. CAB International, Wallingford, UK, p.237- 259, 1994.

GROSS, W. B. **The Development of "air sac disease "**. Avian Dis. 5: 431-439, 1991.

GUSTAVSSON, N.; DIEZ, A.; NYSTRÖM, T. **The universal stress protein paralogues of *Escherichia coli* are co-ordinately regulated and co-operate in the defence against DNA damage.** Mol Microbiol. Jan;43(1):107-17, 2002.

HALL, N. **Advanced sequencing technologies and their wider impact in microbiology.** J. Exp. Biol**.,** v. 210, p. 1518-1525, 2007.

HALL, N. et al. **Sequence of *Plasmodium falciparum* chromosomes 1, 3-9 and 13.** Natur., v. 419, p. 527-531, 2002.

HAYASHI, T. et al. **Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12.** DNA Res. 8(1):11-22, 2001.

HAYDON, D. T. et al. **Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates.** Genet., Jan;157(1):7-15, 2001.

HOCHHUT, B. et al. **Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic *Escherichia coli* strain 536.** Mol. Microbiol. Aug;61(3):584-95, 2006.

HOLUB, E. B. **The arms race is ancient history in Arabidopsis, the wildflower.** Nat. Rev. Genet. Jul;2(7):516-27, 2001.

HSIEH, Y. J.; WANNER, B .L. **Global regulation by the seven-component Pi signaling system**. Curr. Opin. Microbiol. 13, 198–203, 2010.

HUGHES, A. L.; FRIEDMAN, R. **Gene duplication and the properties of biological networks.** J. Mo.l Evol., 61(6):758-64, 2005.

HUGHES, A. L.; NEI, M. **Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection**. Natur., 8;335(6186):167-70, 1988.

HUGHEY, R.; KARPLUS, K. **Bioinformatics:a new field in engineering education.** J. Engineering Educ., p.101–104, 2003.

HYMAN, R. W. et al. **Sequence of *Plasmodium falciparum* chromosome 12**. Natur., v. 419, p. 534-537, 2002.

IMAI, K. et al. **SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in gram-negative bacteria.** Bioinformat. Jul 14;2(9):417-21, 2008.

JAYARAMAN, R. **Bacterial persistence: some new insights into an old phenomenon**. J. Biosci. Dec;33(5):795-805, 2008

JIN, Q. et al. **Genome sequence of *Shigella flexneri 2a*: insights into pathogenicity trough comparison with genomes of *Escherichia coli* K12 and O157.** Nucleic. Acids. Res., 15;30(20):4432-4441, 2002.

JOHNSON, J. R. et al. **Isolation and molecular characterization of nalidixic acid resistant extraintestinal pathogenic *Escherichia coli* from retail chicken products**. Antimicrob. Agents Chemot., v. 47, p. 2161-2168, 2003.

JOHNSON, T. J. et al. **DNA sequence of a ColV plasmid and prevalence of selected plasmid-encoded virulence genes among avian *Escherichia coli* strains.** J. Bacteriol., v. 188, p. 745-758, 2006.

JOHNSON, T. J. et al. **The Genome Sequence of Avian Pathogenic *Escherichia coli* Strain O1:K1:H7 Shares Strong Similarities with Human Extraintestinal Pathogenic *E. coli* Genomes.** J. of Bacteriol., p. 3228–3236, 2007.

KÄLL, L.; KROGH, A.; SONNHAMMER, E. L. **A combined transmembrane topology and signal**

peptide prediction method. J. Mol. Biol. May 14;338(5):1027-36, 2004.

KAPER, J. B., NATARO, J. P.; MOBLEY, H. L. **Pathogenic *Escherichia coli*.** Nat. Rev. Microbiol. 2004 Feb;2(2):123-40.

KARIYAWASAM, S.; JOHNSON, T. J.; NOLAN, L. K. **The pap operon of avian pathogenic *Escherichia coli* strain O1:K1 is located on a novel pathogenicity island.** Infect. Immun., v. 74, p. 744-749, 2006.

KHAN, N. A. et al. **FimH mediated *Escherichia coli K1* invasion of human brain microvascular endothelial cells**. Cell. Microbiol. 9, 169–178, 2007.

KIM, K. S. **Mechanisms of microbial traversal of the blood–brain barrier.** Nature Rev. Microbiol. 6, 625–634, 2008.

KLENK, H. P. et al. **The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus.** Natur., v. 390, p. 364-370, 1997.

KNÖBL, T. et al. **Virulence properties of *Escherichia coli* isolated from ostriches with respiratory disease**. Vet. Microb., 83, 71-80, 2001.

KOSCHINSKI, A. et al. **Why *Escherichia coli* alpha-hemolysin induces calcium oscillations in mammalian cells—the pore is on its own**. FASEB J 20:973–975, 2006.

KOSEK, M.; BERN, C.; GUERRANT, R. L. **The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000**. Bull World Heal. Organ. 81(3):197-204, 2003.

KURTZ, S. et al. **Versatile and open software for comparing large genomes**, Genome Biol., 5:R12, 2004.

KUSSELL, E. et al. **Bacterial persistence: a model of survival in changing environments.** Genetics. Apr;169(4):1807-14, 2005.

LANDER, E. S. et al. **Initial sequencing and analysis of the human genome.** Natur., v. 409, p. 860-921, 2001.

LATHAM, R. H.; STAMM, W. E. **Role of fimbriated *Escherichia coli* in urinary tract infections in adult women: Correlation with localization studies**. J. Infect. Dis., v. 149, p.835-840, 1994.

LAWRENCE, J. G. **Selfish operons and speciation by gene transfer.** Trends Microbiol. 5:355–359, 1997.

LEFÉBURE, T.; STANHOPE, M. J. **Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition.** Genome Biol., 8(5):R71, 2007.

LEFÉBURE, T.; STANHOPE, M. J. **Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*.** Genome Res.,19(7):1224-32, 2009.

LEHMANN P. **Structure and evolution of plant disease resistance genes**. J. Appl. Genet. 43(4):403-14, 2002.

LEIMAN, P.G. et al. **Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin.** Proc. Natl. Acad. Sci. USA. v.106, p.4154-4159, 2009.

LEWIS-ROGERS, N.; McCLELLAN, D. A.; CRANDALL, K. A. **The evolution of foot-and-mouth disease virus: impacts of recombination and selection**. Infect. Genet. Evol., 8(6):786-98, 2008.

LI, Q. et al. **The differential affinity of the usher for chaperone-subunit complexes is required for assembly of complete pili**. Mol. Microbiol. Apr;76(1):159-72, 2010.

LIN, X. M. et al. **Proteomic analysis of nalidixic acid resistance in *Escherichia coli*: identification and functional characterization of OM proteins**. J. Proteome Res. 7(6):2399-405, 2008.

LINGGOOD, M. A. et al. **Incidence of the aerobactin iron uptake system among *Escherichia coli isolates* from infections of farm animals**. J. Gen. Microbiol., v. 133, p. 835-842, 1987.

LORENZO, V. D.; PAW, H. B.; NEILANDS, J. B. **Aerobactin biosynthesis and transport genes of plasmid Co1V-K30 in *Escherichia coli K-12***. J. Bacteriol. 165, 570–578, 1986.

LUND, B. et al. **The PapG protein is the a-D-galactopyranosyl-(1-4)-b-Dgalactopyranose- binding adhesin of uropathogenic *Escherichia coli*.** Proc. Natl. Acad. Sci. USA 84: 5898–5902, 1987.

MARGULIES, M. et al. **Genome sequencing in microfabricated high-density picolitre reactors**. Natur. 437, 376-380, 2005.

MARRI, P. R.; HAO, W.; GOLDING, G. B. **Gene gain and gene loss in *Streptococcus*: is it driven by habitat?** Mol. Biol. Evol., (12):2379-2391, 2006.

MCPEAKE, S. J. W.; SMUTH, J. A.; BALL, H. J. **Characterisation of avian pathogenic *Escherichia coli* (APEC) associated with colisepticaemia compared to faecal insolates from healthy birds**. Vet. Microb., 110, 245-253, 2005.

MINAGAWA, S. et al. **Identification and molecular characterization of the Mg2 stimulon of *Escherichia coli*.** J. Bacteriol. 185: 3696–3702, 2003.

MOON, H. W. **Colonization factor antigens of Enterotoxigenic *Escherichia coli* in animals.** Curr. Top. Microbiol. Immunol., v. 151, p.148-165, 1990.

MORIEL, D. G. et al. **Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*.** Proc. Natl. Acad. Sci. U S A. May 18;107(20):9072-7, 2010.

MOUGOUS, J. D. et al. **A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus.** Scien. Jun 9;312(5779):1526-30, 2006.

MU, X.Q.; BULLITT, E. **Structure and assembly of P-pili: a protruding hinge region used for assembly of a bacterial adhesion filament.** Proc. Natl. Acad. Sci. USA 103: 9861–9866, 2006.

NACHIN, L.; NANNMARK, U.; NYSTRÖM, T. **Differential roles of the universal stress proteins of *Escherichia coli* in oxidative stress resistance, adhesion, and motility.** J. Bacteriol. Sep;187(18):6265-72, 2005.

NAKAZATO, G. et al. **Virulence factors of avian pathogenic *Escherichia coli* (APEC).** Pesq. Vet. Bras., v. 29 , n. 7, p.479-486, 2009.

NATARO, J. P.; KAPER, J. B. **Diarrheagenic *Escherichia coli***. Clin. Microbiol. Rev., v. 11, p.142-201, 1998.

NCBI a, **National Center for Biotechnology Information.** Disponível

em:<http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_growth.html >. Acesso em: 05, Jan. 2012.

NCBI b, **National Center for Biotechnology Information.** Disponível em:<http://www.ncbi.nlm.nih.gov/bioproject/browse/ >. Acesso em: 06, Ago. 2012.

NGELEKA, M. et al. **Pathotypes of avian *Escherichia coli* as related to tsh-, pap-, pil-, and iuc-DNA sequences, and antibiotic sensitivity of isolated from internal tissues and the cloacae of broilers.** Av. Dis.,46,143-152, 2002.

NIERMAN, C. W. et al. **Genome data: what do we learn?** Structural Biol., v.10, p. 343-348, 2000.

NIKAIDO, H. Outer membrane. **In: *Escherichia coli* and *Salmonella* cellular and molecular biology** (eds. F.C. Neidhardt et al.) ASM Press, Washington, DC., 1996. p. 29–47.

NIKAIDO, H. **Porins and specific diffusion channels in bacterial outer membranes**. J. Biol. Chem. 269: 3905–3908, 1994.

OCHMAN. H.; LAWRENCE, J. G.; GROISMAN, E. A. **Lateral gene transfer and the nature of bacterial innovation.** Natur. 405:299–304, 2000.

OGASAWARA, H. et al. **Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade**. J Bacteriol. Jul;189(13):4791-9, 2007.

OLSÉN, A.; JONSSON, A.; NORMARK, S. **Fibronectin binding mediated by a novel class of surface organelles on *Escherichia coli*.** Natur., v. 338, p. 652-655, 1989.

ORSI, R. H.; SUN, Q.; WIEDMANN, M. **Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes.*** BMC Evol. Biol. 12;8:233, 2008.

PALCHEVSKIY, V.; FINKEL, S. E. ***Escherichia coli* competence gene homologs are essential for competitive fitness and the use of DNA as a nutrient.** J. Bacteriol. Jun;188(11):3902-10, 2006.

PARREIRA, V. R.; GYLES, C. L. **A novel pathogenicity island integrated adjacent to the thrW tRNA gene of avian pathogenic *Escherichia coli* encodes a vacuolating autotransporter toxin.** Infect. Immun., Sep;71(9):5087-96, 2003.

PERNA, N. T. et al. **Genome sequence of enterohaemorrhagic *Escherichia coli O157:H7***. Natur.**,** 409 (6819): 529-33, 2001.

PETERSEN, L. et al. **Genes under positive selection in *Escherichia coli.*** Genome Res.,17(9):1336-1343, 2007.

PFAFF-MCDONOUGH, S. J. et al. **Complement resistance-related traits among *Escherichia coli* isolates from apparently healthy birds and birds with colibacillosis**. Avian Dis., v. 44, p.23-33, 2000.

PIMENTA, A. et al. **Antibody analysis of the localisation, expression and stability of HlyD, the MFP component of the *E. coli* haemolysin translocator**. Mol. Gen. Genet. 261:122– 132, 1999.

POLLEY, S. D.; CHOKEJINDACHAI, W.; CONWAY, D. J. **Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen**. Genetics. Oct;165(2):555-61, 2003.

POLLEY, S. D.; CONWAY, D. J. **Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene**. Genet., Aug;158(4):1505-12, 2001.

PRILIPOV, A. et al. **Identification and characterization of two quiescent porin genes, nmpC and ompN, in *Escherichia coli BE*.** J. Bacteriol. 180(13):3388-92, 1998.

PROVENCE, D. L.; CURTISS III, R. **Isolation and characterization of a gene involved in hemagglutination by an avian pathogenic *Escherichia coli* strains.** Infect. Immun., v. 62, p. 1369-1380, 1994.

PUKATZKI, S. et al. **Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the Dictyostelium host model system**. Proc. Natl. Acad. Sci. USA. v.103, p.1528–1533, 2006.

RANDALL-HAZELBAUER, L.; SCHWARTZ, M. **Isolation of the bacteriophage lambda receptor from *Escherichia coli***. J. Bacteriol. 116(3):1436-46, 1973.

RETIEF, J. D. **Phylogenetic analysis using PHYLIP**. Met. Molecul. Biol., n. 132: 243-258, 2000.

RILEY M. et al. ***Escherichia coli K-12*: a cooperatively developed annotation snapshot--2005.** Nucleic Acids Res. Jan 5;34(1):1-9, 2006.

ROBERTS, J. A. et al. **The Gal(alpha 1-4)Gal-specific tip adhesin of *Escherichia coli* P-fimbriae is needed for pyelonephritis to occur in the normal urinary tract**. Proc. Natl. Acad. Sci. U S A. Dec 6;91(25):11889-93, 1994.

RODRIGUEZ-SIEK, K. E. et al. **Characterizing the APEC pathotype.** Vet. Res., v. 36, p. 241- 256, 2005.

ROJAS, T. C. G. et al. **Draft genome of a Brazilian avian-pathogenic *Escherichia coli* strain an in silico characterization of virulence-related genes**. J. Bacteriol., 194(11)3023.doi:10.1128/JB.00394-12, 2012.

RUSSO, T. A. et al. **TnphoA-mediated disruption of K54 capsular polysaccharide genes in *Escherichia coli* confers serum sensitivity**. Infect Immun. 61**:** 3578–3582, 1993.

RUSSO, T. A.; JOHNSON, J. R. **Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem.** Microbes Infect. 5: 449–456, 2003.

SABRI, M.; LÉVEILLÉ, S.; DOZOIS, C. M. **A SitABCD homologue from an avian pathogenic *Escherichia coli* strain mediates transport of iron and manganese and resistance to hydrogen peroxide.** Microbiol., 152(Pt 3):745-58, 2006.

SALEMI, M.; VANDAMME, A. M. In: **Basic concepts of molecular evolution:** The phylogenetic Handbook: a pratical approach to DNA and protein phylogeny. Cambridge: Cambridge University Press, 2003. 1-21.

SAUVONNET, N.; GOUNON, P.; PUGSLEY, A. P. **PpdD type IV pilin of *Escherichia coli* K-12 can Be assembled into pili in Pseudomonas aeruginosa**. J. Bacteriol. Feb;182(3):848-54, 2000.

SAVAGEAU, M. A. ***Escherichia coli* habitats, cell types, and molecular mechanisms of gene control.** Am. Nat. 122: 732–744, 1983.

SAWYER, S. A. **Statistical tests for detecting gene conversion.** Mole. Biol. and Evol., 6, 526-538, 1989.

SCHELL, M. A. et al . **Type VI secretion is a major virulence determinant in *Burkholderia mallei***. Mol. Microbiol. v.64, p.1466–1485, 2007.

SCHERER, N. M. **Evolução molecular darwiniana nas proteínas relacionadas à patogênese (prs) em plantas**. 2002. Dissertação apresentada ao Programa de Pós Graduação em Genética e Biologia Molecular da Universidade Federal do Rio Grande do Sul para a obtenção do Grau de Mestre em Genética e Biologia Molecular. (Orientador: Francisco M. Salzano).

SCHIRMER, T. **General and specific porins from bacterial outer membranes**. J. Struct. Biol. 121: 101–109, 1998.

SEXTON, J. A. et al. ***Legionella pneumophila*** **DotU and IcmF are required for stability of the Dot/Icm complex.** Infect. Immun. v.72, p. 5983–5992, 2004.

SHACKELTON, L. A. et al. **High rate of viral evolution associated with the emergence of carnivore parvovirus.** Proc. Natl. Acad. Sci. U S A, Jan 11;102(2):379-84, 2005.

SILVEIRA, W. D. et al. **Biological characteristics and pathogenicity of avian *Escherichia coli* strains**. Vet. Microb., 85,47-53, 2002.

SIMPSON, A. J. et al., **The genome sequence of the plant pathogen *Xylella fastidiosa*. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis.** Natur., 13;406(6792):151-9, Jul., 2000.

SMITH, H. W.; GYLES, C. L. ***Escherichia coli*** **enterotoxin**. Vet. Res., v. 85, n. 24, p. 694-725, 1969.

SNYDER, J. A. et al. **Transcriptome of uropathogenic *Escherichia coli* during urinary tract infection.** Infect. Immun. Nov;72(11):6373-81, 2004.

SOKJA, W. J. ***Escherichia coli*** **in domestic animals and poultry**. Common wealth Agricultural Bureaux Farmham Royal, Bucks, England, 157-168, 1965.

SOKJA, W. J., CARNAGHAN, R. B. A. ***Escherichia coli*** **infection in poultry**. Res, Vet. Sch., 2, 340-352, 1961.

SOYER, Y et al., **Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected *Salmonella* serotypes.** BMC Evol. Biol., 14;9:264, 2009.

SPRINGAEL, D.; TOP, E. M. **Horizontal gene transfer and microbial adaptation to xenobiotics: new types of mobile genetic elementes and lessons from ecological studies**. Trends Microbiol., 12(2):53-58,

2004.

STEHLING, E. G. et al. **Characterization of a plasmid-encoded adhesion of an avian pathogenic *Escherichia coli* (APEC) strains isolated from a case of swollen head syndrome (SHS)**. Vet. Microb., 95, 111-120, 2003.

THUMBIKAT, P. et al. **Bacteria-induced uroplakin signaling mediates bladder response to infection**. PLoS Pathog. 5, e1000415, 2009.

TIFFIN, P.; MOELLER, D. A. **Molecular evolution of plant immune system genes.** Trends Genet. Dec; 22(12):662-70, 2006.

TORTORA, G. D.; FUNKE, B. R.; CASE, C. L. **Microbiology: an introduction**. Pearson Education, Inc., publishing as Benjamin Cummings, San Francisco, 2006, 579.

TOSHIMA, H. et al. **Enhancement of Shiga toxin production in enterohemorrhagic *Escherichia coli* serotype O157:H7** by DNase colicins. Appl. Environ. Microbiol. 73, 7582–7588, 2007.

TOUCHON, M. et al. **Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths**. PloS Genet. Jan;5(1):e1000344, 2009.

VAN VALEN, L. **A new evolutionary law. Evolutionary Theory**. 1: 1-30, 1973.

VANRHEENEN, S.; DUMÉNIL, G.; ISBERG, R.R. **IcmF and DotU Are Required for Optimal Effector Translocation and Trafficking of the *Legionella pneumophila* Vacuole.** Infect. Immun. v.72, p.5972-5982, 2004.

WALBURGER, A. LAZDUNSKI, C. CORDA, Y. **The Tol/Pal system function requires an interaction between the C-terminal domain of TolA and the N-terminal domain of TolB**. Mol Microbiol. May;44(3):695-708, 2002.

WANG, Y. F. et al. **Channel specificity: structural basis for sugar discrimination and differential flux rates in maltoporin.** J. Mol. Biol. 12;272(1):56-63, 1997.

WANNER, B. L. **Phosphorus assimilation and control of the phosphate regulon**. In: Neidhardt, R.C.I., Ingraham, J.L., Lin, E.E.C., Low, K.B., Magasanik, B., Reznikoff, W.S., Riley, M., Schaechter, M., Umbrager, H.E. (Eds.), *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, vol. 1.

American Society for Microbiology, Washington, DC, 1996, pp. 1357–1381.

WEI, J. et al. **Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T**. Infect Immun. May;71(5):2775-86, 2003.

WEISS, A. M. **Estratégias de finalização da montagem do genoma da bactéria diazotrófica endofítica *Herbaspirillum seropedae* SmR1**. 2010. Dissetação de Mestrado em Ciências- Bioquimica, do setor de Ciências biológicas da Universidade Federal do Paraná. (Orientador: Leonardo Magalhães Cruz).

WELCH, R. A. et al. **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*.** Proc. Natl. Acad. Sci. U S A, Dec 24;99(26):17020-17024, 2002.

WEST, A. H.; STOCK, A. M. **Histidine kinases and response regulator proteins in two-component signaling systems**. Trends Biochem. Sci. 26, 369–376, 2001.

WILES, T. J. et al. **Inactivation of host Akt/protein kinase B signaling by bacterial pore-forming toxins.** Mol. Biol. Cell. 19, 1427–1438, 2008.

WOOLEY, R. E., et al. **Colonization of the chicken trachea by an avirulent avian *Escherichia coli* transformed with plasmid pHK11**. Avian Dis., v. 42, p.194-198, 1998.

WOOSTER, D. G. et al. **Logarithmic phase *Escherichia coli* K1 efficiently avoids serum killing by promoting C4bpmediated C3b and C4b degradation**. Immunol. 117, 482–493, 2006.

XICOHTENCATL-CORTES, J. et al. **The type 4 pili of enterohemorrhagic *Escherichia coli* O157:H7 are multipurpose structures with pathogenic attributes.** J. Bacteriol. 191, 411–421, 2009.

XIONG, L. et al. **Construction of iucB and iucBiutA mutants of avian pathogenic *Escherichia coli* and evaluation of their pathogenicity.** Vet Microbiol. 2012.

YANG Z. **The power of phylogenetic comparison in revealing protein function**. Proc Natl Acad. Sci. U S A. Mar 1;102(9):3179-80, 2005.

YANG, Z. **PAML 4: phylogenetic analysis by maximum likelihood.** Mol. Biol. and Evol., v. 24, n. 8, p.1586-1591, May, 2007.

YANG, Z.; BIELAWSKI, J. P. **Statistical methods for detecting molecular adaptation**. Trends Eco.l Evol., Dec 1;15(12):496-503, 2000.

YANG, Z.; NIELSEN, R. **Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages**. Mol. Biol. and evol., 19(6), 908–17, 2002.

YANG, Z.; SWANSON, W. J. **Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes**. Mol. Biol. Evol. Jan;19(1):49-57, 2002.

YU, S. L. et al. **Characterization of the distal tail fiber locus and determination of the receptor for phage AR1, which specifically infects *Escherichia coli* O157:H7**. J. Bacteriol. Nov;182(21):5962-8 2000.

ZHANG, D. F. et al. **Functional characterisation of altered outer membrane proteins for tetracycline resistance in *Escherichia coli*.** Int. J. Antimicrob. Agents. Oct;32(4):315-9, 2008.