MINISTÉRIO DA SAUDE

FUNDAÇÃO OSWALDO CRUZ

INSTITUTO OSWALDO CRUZ

Mestrado no Programa de Pós-Graduação Biologia Computacional e Sistemas

Análise do perfil filogenético de enzimas isofuncionais não homólogas entre cepas patogênicas de *Escherichia Coli.*

CHRISTIAN SAGAVE MAZZOCCO DE ALMEIDA

Rio de Janeiro

Agosto de 2020

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

CHRISTIAN SAGAVE MAZZOCCO DE ALEMIDA

Análise do perfil filogenético de enzimas isofuncionais não homólogas entre cepas patogênicas de *Escherichia Coli.*

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Biologia Computacional e Sistemas.

Orientadores: Prof. Dr. Marcos Paulo Catanho

RIO DE JANEIRO

Agosto de 2020

Sagave Mazzocco de Almeida, Christian.

Análise do perfil filogenético de enzimas isofuncionais não homólogas entre cepas patogênicas de Escherichia Coli. / Christian Sagave Mazzocco de Almeida. - Rio de janeiro, 2020.
35 f.

Dissertação (Mestrado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2020.

Orientador: Marcos Paulo Catanho.

Bibliografia: f. 36-45

1. Enzimas análogas. 2. Escherichia Coli. 3. Enzimas isofuncionais. I. Título.

Elaborado pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca de Manguinhos/Icict/Fiocruz com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Igor Falce Dias de Lima - CRB-7/6930.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Programa de Pós-Graduação em Biologia Computacional e Sistemas

CHRISTIAN SAGAVE MAZZOCCO DE ALMEIDA

Análise do perfil filogenético de enzimas isofuncionais não homólogas entre cepas patogênicas de *Escherichia Coli*

ORIENTADOR (ES): Prof. Dr. Marcos Paulo Catanho de Souza

Aprovada em: _____/_____/_____

EXAMINADORES:

Prof. Dr. Rafael Dias Mesquita	**Presidente** (IOC/FIOCRUZ)
Prof. Dr. Diogo Antonio Tschoeke	**Membro**(UFRJ)
Prof. Dra. Kary Ocana	**Membro**(LNCC)
Prof. Dr**.** Fabio Faria da Mota	**Suplente**(IOC/FIOCRUZ)
Prof. Dr**.** Nicolas Carels	**Suplente**(IOC/FIOCRUZ)

Rio de Janeiro, 01 de outubro de 2020

Agradecimentos

Aos meus pais, Luciana e Marcos, pois sem eles nada disso seria possível. Agradeço a vocês pela minha vida, por zelarem por mim e sempre darem todo o apoio que precisei ao longo dessa caminhada que decidi começar

À minha tia, Inha, por ser minha maior incentivadora, você sempre esteve comigo sempre que possível, obrigado por todo amor, afeto e garra. Saiba que só cheguei onde cheguei graças a Deus e a você, mais uma vez obrigado por ser a minha fonte de inspiração na vida como um todo, incluindo a acadêmica.

A minha noiva Emily, por ter me aturado nos dias difíceis durante esta etapa, pela paciência e consolo em meio a perrengues e prazos. Obrigado por ser minha principal fonte de fôlego e perseverança, através de muito carinho e amor.

Ao meu orientador Catanho, primeiramente por ter me aceito como aluno, mesmo em uma situação tão complexa, e por estar sempre ao meu lado, me instruindo e me ensinando a nobre arte da ciência. Não conseguirei expressar o quanto sou grato por isso. Em todos os desafios e conquistas subsequentes lembrarei dos ensinamentos passados durante meu mestrado.

Aos amigos que conheci na FIOCRUZ, Ronald, Pedro, Rafael, Arthur, Lucas, Aline, Gisele, Alessandra, por serem sempre tão companheiros e leais, mesmo quando as coisas ficaram difíceis. Obrigada pelas conversas, pelas ajudas nos scripts, pelas companhias nas disciplinas.

A banca avaliadora, pela disposição em participar da avaliação deste projeto. A Fundação Oswaldo Cruz (FIOCRUZ) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro. Obrigada a todos!

INSTITUTO OSWALDO CRUZ

Análise do perfil filogenético de enzimas isofuncionais não-homólogas entre cepas patogênicas de *Escherichia coli.*

RESUMO

DISSERTAÇÃO DE MESTRADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Christian Sagave Mazzocco de Almeida

Escherichia coli é uma bactéria Gram-negativa e em sua variedade possui cepas não patogênicas e patogênicas. Cepas não patogênicas pertencem naturalmente a nossa microbiota intestinal, desempenhando funções benéficas ao nosso organismo, já inúmeras cepas patogênicas afetam a saúde pública, principalmente em países subdesenvolvidos, por ser uma das principais causadoras de gastroenterites em crianças menores de 5 anos de idade, estimando-se 760.000 mortes por ano em escala global. Sabemos que a compreensão do metabolismo é crucial para entender a expressão fenotípica em todos os organismos vivos. Neste sentido, torna-se fundamental a identificação e caracterização do repertório de enzimas que atuam nestas vias. Entretanto, a classificação comumente usada para enzimas não leva em consideração a sua ancestralidade, que permite diferenciá-las em dois grupos: homólogas, que evoluíram do mesmo ancestral, possuem estruturas tridimensionais semelhantes; e análogas, que possuem histórias evolutivas diferentes, mas desempenham a mesma função. O conhecimento sobre enzimas isofuncionais não-homólogas e as vias bioquímicas em que elas atuam pode dizer muito a respeito da evolução do metabolismo, bem como seu papel nos distintos fenótipos de patogenicidade, não somente em *E. coli*, como também em outras espécies. Neste projeto analisamos os possíveis papeis das enzimas isofuncionais não-homologas na diversidade genética e metabólica e sua relação com fenótipos de patogenicidade em *E. coli*, utilizando métodos computacionais para identificação (AnEnPi), caracterização (Argot 2.5), validação (SUPERFAMILY) e mapeamento metabólico (KEGG) destas enzimas em um conjunto de 52 cepas de *E. coli* com genomas completamente sequenciados e com origem e fenótipo de patogenicidade definidos (NCBI, PATRIC). Dessa forma, detectamos 71 enzimas possivelmente análogas, 45 delas pertencentes ao genoma acessório da espécie, envolvendo um total 66 vias metabólicas. Os padrões de presença/ausência das enzimas isofuncionais não-homólogas e das atividades enzimáticas exercidas por elas foram avaliados frente aos distintos perfis de patogenicidade apresentados pelos organismos em nossa amostra – EHEC (7), EIEC (1), ETEC (1), ExPEC (7), MCR-1 positive (6), STEC (4), UPEC (3), NIA (23). As enzimas e atividades enzimáticas ferredoxian reductase (EC 1.18.1.3), ribokinase (2.7.1.15), manose-6-fofato isomerase (EC 5.3.1.8), amina oxidase (EC 1.4.3.21), chitinase (EC 3.2.1.14), glucosamina-1-fosfato N-acetyltransferase (EC 2.3.1.157) e asparaginase (EC 3.5.1.1)

apresentam padrão de presença (correlação) ou ausência (anticorrelação) associados a grupos patogênicos diferentes, contribuindo para a diversidade genética e metabólica em *E. coli*. Mais investigações são necessárias para estabelecer se tais enzimas de fato contribuem diretamente para a expressão de fenótipos de patogenicidade em *E. coli*.

INSTITUTO OSWALDO CRUZ

Analysis of the phylogenetic profile of non-homologous isofunctional enzymes among pathogenic strains of Escherichia coli.

ABSTRACT

MASTER DISSERTATION IN COMPUTACIONAL BIOLOGY

Christian Sagave Mazzocco de Almeida

Escherichia coli is a Gram-negative bacterium and in its variety has non-pathogenic and pathogenic strains. Non-pathogenic strains naturally belong to our intestinal microbiota, performing beneficial functions to our organism, since numerous pathogenic strains affect public health, especially in underdeveloped countries, as it is one of the main causes of gastroenteritis in children under 5 years of age, estimated 760,000 deaths per year on a global scale. We know that understanding metabolism is crucial to understanding phenotypic expression in all living organisms. In this sense, it is essential to identify and characterize the repertoire of enzymes that act in these pathways. However, the classification commonly used for enzymes does not take into account their ancestry, which allows to differentiate them into two groups: homologues, which evolved from the same ancestor, have similar three-dimensional structures; and analogous, which have different evolutionary histories, but perform the same function. The knowledge about non-homologous isofunctional enzymes and the biochemical pathways in which they act can say a lot about the evolution of metabolism, as well as their role in the different pathogenicity phenotypes, not only in E. coli, but also in other species. In this project we analyze the possible roles of non-homologous isofunctional enzymes in genetic and metabolic diversity and their relationship with E. coli pathogenicity phenotypes, using computational methods for identification (AnEnPi), characterization (Argot 2.5), validation (SUPERFAMILY) and mapping metabolic (KEGG) of these enzymes in a set of 52 E. coli strains with completely sequenced genomes and with defined pathogenic origin and phenotype (NCBI, PATRIC). Thus, we detected 71 possibly analogous enzymes, 45 of which belong to the accessory genome of the species, involving a total of 66 metabolic pathways. The patterns of presence / absence of non-homologous isofunctional enzymes and of the enzymatic activities carried out by them were evaluated against the different pathogenicity profiles presented by the organisms in our sample - EHEC (7), EIEC (1), ETEC (1), ExPEC (7), MCR-1 positive (6), STEC (4), UPEC (3), NIA (23). Enzymes and enzymatic activities ferredoxian reductase (EC 1.18.1.3), ribokinase (2.7.1.15), mannose-6-fofate isomerase (EC 5.3.1.8), amine oxidase (EC 1.4.3.21), chitinase (EC 3.2.1.14) , glucosamine-1-phosphate N-acetyltransferase (EC 2.3.1.157) and asparaginase (EC 3.5.1.1) present a pattern of presence (correlation) or absence (anti-correlation) associated with different pathogenic groups,

contributing to the genetic and metabolic diversity in E coli. Further investigation is needed to establish whether such enzymes actually contribute directly to the expression of pathogenic phenotypes in E. coli.

ÍNDICE

ÍNDICE DE FIGURAS

LISTA DE TABELAS

Lista de abreviaturas

E. coli	*Escherichia coli*
EPEC	*Escherichia coli* Enteropatogênica
DNA	Ácido desoxirribonucleico
ETEC	*Escherichia coli* Enterotoxigênica
EAEC	*Escherichia coli* Enteroagregativa
EIEC	*Escherichia coli* Enteroinvasica
STEC	*Escherichia coli* Produtora de toxina Shiga
EHEC	*Escherichia coli* Enterohemorrágica
DAEC	*Escherichia coli* de adesçao difusa
ExPEC	*Escherichia coli* Extraintestinal
EC number	*Enzyme Commission Number*
AnEnPi	*Analogous Enzyme Pipeline*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
PATRIC	*Pathosystems Resource Integration Center*
GO	*Gene Ontology*
CoA	Coenzima A
NAD	Dinucleotídeo de nicotinamida e adenina
NADp	Dinucleotídeo de nicotinamida e adenina fosfato
OMA	*Ortholohous Matrix*
RefSeq	*Reference Sequence Database*
BLAST	*Basic Local Alignment Tool*

1. INTRODUÇÃO

1.1 *Escherichia Coli*

Escherichia coli é uma bactéria Gram-negativa pertencente à família Enterobactericiae inserida na classe das Gammaproteobacteria, e se apresenta em forma de bastonetes. A *E. coli* é uma das bactérias mais estudadas em todo mundo, possuindo crescimento rápido sob condições ótimas, com a capacidade de se replicar em aproximadamente 20 minutos (1).

Em sua maioria, as linhagens de *E. coli* são consideradas comensais, ou seja, inofensivas e presentes na microbiota intestinal dos humanos e outros animais de sangue quente, desempenhando papel importante em nossos organismos; entretanto, algumas linhagens podem ser patogênicas, até mesmo capazes de causar doenças graves (2).

A infecção por *E. coli* pode causar diarreias severas, sendo relatado aproximadamente 1,7 bilhão de casos no mundo anualmente. Em crianças menores de 5 anos é um problema ainda mais grave de saúde pública, sendo a maior causa de malnutrição, morbidade e mortalidade nessa faixa etária, estimando-se 560.000 mortes por ano, afligindo principalmente países da África, Ásia e América do Sul (3).

Baseado nos primeiros estudos epidemiológicos de *E. coli* em águas recreacionais, foi identificada uma possível relação entre a presença dessa bactéria em águas recreacionais e fontes contaminantes (esgoto etc.), bem como a incidência de doenças gastrointestinais (4).

Um estudo realizado por Hamilton e colaboradores (5), nas águas marinhas de uso recreativo na Califórnia, detectou a presença de cepas potencialmente patogênicas, analisando a composição genômica e frequência dos genes de virulência conhecidos de *E. coli.* Mais de 10% das cepas identificadas neste estudo eram potencialmente *E. coli* Enteropatogênica (EPEC), reforçando a sobrevivência desse patógeno no ambiente (5).

Um outro estudo que corrobora a afirmação sobre a presença de cepas patogênicas de *E. coli* no ambiente é o de Ishii e colaboradores (6). O grupo detectou abundância do gene eaeA, presentes em cepas EPEC e em *Cladophoras*, uma espécie de alga. O gene eaeA é um dos fatores de virulência mais frequentemente detectados no ambiente (6).

1.1.1 Perfis de patogenicidade

Embora a maioria dos membros da espécie da *E. coli* permanecem inofensivos na flora intestinal, alguns clones altamente adaptados podem causar uma ampla gama de doenças em humanos como gastroenterite, cistite, meningite, sepse, dentre outras. As cepas que adquirirem os chamados fatores de virulência, comumente codificados por elementos genéticos móveis, como plasmídeos, são chamadas de *E. coli* patogênicas (7).

A evolução da *E. coli* é extremamente rápida, estudos já mostraram a diferenciação entre cepas de *E. coli* em menos de 200 gerações, e revelaram alterações genômicas após infecção em humanos. O sequenciamento de inúmeros genomas desta espécie também confirmou variação genotípica e fenotípica intraespecífica (8).

Os genomas de alguns patógenos bacterianos podem "adquirir" novos genes através da duplicação gênica, resultando muitas vezes em expansões de famílias inteiras de proteínas já existentes. No entanto, o ganho gênico a partir da transferência horizontal (por conjugação, transdução ou transformação) continua sendo a fonte mais eficiente de inovação e variação em patógenos bacterianos. No caso da *E. coli* K-12 estirpe MG1655, estima-se que cerca de 18% do seu genoma tenha sido adquirido por fenômenos de transferência horizontal (8) (9).

Estudos sugerem que a aquisição de genes por transferência lateral, no meio ambiente, é intermediada por elementos genéticos móveis, (ilhas de patogenicidade, bacteriófagos e plasmídeos), que comumente carreiam genes relacionados a fatores de virulência, e cuja a expressão pode ser o resultado de uma interação entre genes "residentes" e novos genes adquiridos. Entretanto, ainda é difícil determinar *a priori* quais variações genômicas viriam de fato ter algum tipo de influência no metabolismo ou adaptação dessas bactérias (10).

Apesar do ganho de genes ser recorrente, por quaisquer dos mecanismos conhecidos, o genoma bacteriano permanece aproximadamente do mesmo tamanho. O ganho gênico deve ser equilibrado à sua contraparte, a perda gênica, por seleção natural ou deriva gênica, sujeitando o genoma bacteriano à uma dinâmica de "uso ou perda de cada gene". Em *E. coli*, por exemplo, temos os genes Flag-2 e ETT2 que, quando ainda não atingidos por nenhuma pressão seletiva,

ocupavam dezenas de kilobases de DNA, mas que foram reduzidos e passaram a ocupar somente algumas centenas de pares de bases, provavelmente por não contribuírem mais para o *fitness* (11).

Dentre as cepas de *E. coli* diarreicas, podemos destacar os seguintes perfis de patogenicidade conhecidos: *E. coli* Enterotoxigênica (ETEC); *E. coli* Enteropatogênica (EPEC); *E. coli* Enteroagregativa (EAEC); *E. coli* Enteroinvasiva (EIEC); *E. coli* Produtora de toxina Shiga (STEC), na qual está inserida o grupo *E. coli* Enterohemorrágica (EHEC); *E. coli* de adesão difusa (DAEC) (Tabela 1). Entretanto, estudos recentes sugerem que algumas amostras de *E. coli* são capazes de sobreviver, e possivelmente se reproduzir, fora do ambiente intestinal, sendo chamadas de *E. coli* Extraintestinal (ExPEC) (1).

Tabela 1. Perfis de patogenicidade em *E. coli*

Perfil	Descrição
Enterotoxigênica (ETEC)	As cepas ETECs são de grande importância epidemiológica, sendo o perfil mais associado a doenças diarreicas e mortes em países em desenvolvimento (12).
Enteropatogênica (EPEC)	EPEC também é outro perfil patogênico muito importante, associado a diarreia em crianças menores de 5 anos em países em desenvolvimento, causando lesões na mucosa intestinal (13).
Enteroagregativa (EAEC)	Cepas EAEC causam doença diarreica com pouca distinção entre faixa etária e classe social, causando doença em crianças e adultos, em países desenvolvidos e em desenvolvimento. Esse grupo patogênico atua aderindo a células HEp-2 da mucosa intestinal (14).
Enteroinvasiva (EIEC)	O grupo EIEC se distingue dos demais perfis patogênicos de *E.coli*, devido ao grande impacto como principal causador de morbidade e mortalidade em adultos e crianças em países desenvolvidos. É capaz de causar colite inflamatória aguda invasiva, e consequentemente, disenteria, mas na maioria dos casos provoca uma diarreia aquosa (15).
Produtora de toxina Shiga (STEC)	As doenças causadas por cepas pertencentes ao grupo STEC são relevantes pelo seu potencial zoonótico e origem alimentar. Em humanos os sintomas podem abranger desde uma diarreia leve até a síndrome hemolítica urêmica (16).
Enterohemorrágica (EHEC)	EHEC é comummente associada a diarreias sanguinolentas e síndrome hemolítica urêmica. Essas cepas estão inclusas no grupo STEC pois possuem como principal fator de virulência as toxinas Shiga (17).
De adesão difusa (DAEC)	As cepas caracterizadas como DAEC são definidas por um padrão de aderência difusa (DA), em que as bactérias cobrem uniformemente toda a superfície celular. Os sintomas são relativamente leves, causando principalmente diarreia aquosa em crianças (13).

Extraintestinal (ExPEC)	As cepas do perfil patogênico ExPEC são capazes de causar infecções em locais extra-intestinais. Algumas delas se destacam dentro desse grupo: *E. coli* Uropatogênicas (UPEC) e Meningites neonatais associadas a *E. coli* (NMEC). Muitas cepas deste grupo também estão relacionadas com à aquisição de novos genes de resistência a antibióticos (18) (19) (20).

1.1.2 Genômica

A primeira análise genômica de *E. coli* foi feita em 1997, desde então, mais de 5.000 genomas de *E. coli* já foram sequenciados. Por apresentar um rápido crescimento, a *E. coli* é muito utilizada para a estudar a evolução, como por exemplo, Tenaillon *et al.*, que possui um estudo progressivo da evolução da *E.coli*, que em 2016 atingiu 50.000 gerações (21).

Análises comparativas entre genomas sequenciados de *E. coli* têm demonstrado a aquisição e a perda de genes ao longo da evolução nesta espécie (sobretudo através de elementos genéticos móveis, como plasmídeos, ilhas de patogenicidade e bacteriófagos), resultando em diferenças genéticas e fenotípicas, como por exemplo variabilidade metabólica e perfis de patogenicidade, que distinguem as cepas e perfis atualmente conhecidos (22). Tais abordagens genômicas resultaram na caracterização de genes determinantes de virulência entre cepas pertencentes a diferentes grupos de *E. coli*, facilitando avaliação e diagnósticos de riscos, assim como métodos de combate as cepas virulentas de diferentes perfis de patogenicidade de forma mais eficaz (23).

A totalidade de genes contidos em uma espécie é denominado pangenoma, e pode ser dividido em genoma core, composto por um núcleo conservado de genes e proteínas presentes em todos ou quase todos os genomas de uma mesma espécie, constituindo uma “espinha dorsal” gênica contento informações genéticas necessárias para processos celulares; e genoma acessório, o qual possui uma gama de genes e proteínas que forma um *pool* gênico flexível, que não é comum a grande maioria das cepas da espécie, proporcionando uma consistente variedade individual de informações gênicas específicas, capazes de fornecer propriedades adicionais de adaptação em um nicho específico, ou sob condições específicas, dentre outras, tornando provável a distinção dos diferentes grupos patogênicos de *E .coli* através da análise destes genes (24) (25).

1.2 Enzimas

Enzimas foram originalmente descritas como biocatalisadores, em 1883, com a descoberta da diastase, que faz parte de um grupo de enzimas que catalisa a quebra do amino em maltose (26). As enzimas são responsáveis por catalisar reações bioquímicas acelerando a conversão de substratos em produtos (e vice-versa), diminuindo significativamente a energia necessária para ativação de uma reação bioquímica. Sem a catálise enzimática a maioria das reações necessárias à manutenção da vida seriam demasiadamente lentas, inviabilizando a vida na Terra como a conhecemos (27).

As enzimas são classificadas hierarquicamente, de acordo com as reações químicas que catalisam. A classificação mais usual foi proposta em 1992 pelo *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology* <http://www.sbcs.qmul.ac.uk/iubmb>, na qual cada enzima é designada por um número EC (*Enzyme Commission number*) composto de 4 dígitos, no qual o primeiro número corresponde à classe da enzima, os dois números subsequentes têm significados diferentes dependendo da classe da enzima, e o último número descreve a especificidade da reação, definindo o substrato/produto ou cofatores. Por exemplo, o EC 1.1.1.1 (álcool desidrogenase), faz parte da classe das enzimas oxidoredutases (EC 1), que possui como doadores os grupamentos CH-OH (EC 1.1), sendo seus aceptores NAD ou NADP (EC 1.1.1) e tendo como substratos o álcool etílico ou hemiacetais (EC 1.1.1.1) (28) (29) (30).

1.3 Analogia Funcional

Como visto anteriormente, as enzimas são classificadas de acordo com a reação química que catalisam, seus substratos e cofatores. Porém, as enzimas também podem ser classificadas de acordo com sua ancestralidade, sendo as enzimas com origem evolutiva a partir de um ancestral comum, denominadas homólogas. Essas proteínas apresentam comumente estruturas tridimensionais significativamente similares. No entanto, as enzimas que se originaram a partir de eventos evolutivos independentes, não possuindo um ancestral comum portanto, que convergem para o desempenho de uma mesma atividade enzimática, em um processo evolutivo conhecido

como convergência funcional, são denominadas análogas funcionais, isto é, enzimas isofuncionais não-homólogas (31).

Apesar de catalisarem a mesma reação bioquímica, enzimas análogas não possuem semelhanças detectáveis ente duas estruturas primária e terciária (31). Portanto, em outras palavras, o surgimento de caracteres (moleculares, morfológicos, comportamentais, ecológicos e funcionais) análogos representa soluções distintas e independentes (enzimas isofuncionais não-homólogas) para um mesmo problema (catálise enzimática), podendo ocorrer repetidas vezes ao longo da evolução das espécies (32) (33).

HIPÓTESE

Enzimas isofuncionais não-homólogas (análogas funcionais), codificadas por genes não essenciais em *E. coli*, pertencentes ao genoma acessório ou linhagem-específicos, podem contribuir para os fenótipos de patogenicidade conhecidos nesta espécie, demonstrando a importância biomédica da investigação do fenômeno de convergência evolutiva em vias bioquímicas de microrganismos patogênicos.

2. OBJETIVOS

2.1 Geral

Analisar o perfil filogenético (ocorrência e distribuição) de enzimas isofuncionais não-homólogas (análogas funcionais) entre cepas patogênicas de *Escherichia coli*, buscando compreender a contribuição do fenômeno de convergência evolutiva para a diversidade genética e metabólica entre representantes desta espécie com genoma completamente sequenciado e sua relação com os fenótipos de patogenicidade conhecidos nesta espécie.

2.2 Específicos

1. Identificar, computacionalmente, as enzimas isofuncionais não-homólogas codificadas nos genomas completamente sequenciados e anotados de cepas de *E. coli*, patogênicas e não patogênicas disponíveis publicamente;

2. Caracterizar as enzimas isofuncionais não-homólogas preditas segundo sua composição de domínios funcionais e o tipo de enovelamento proteico e sua origem evolutiva;

3. Mapear as enzimas isofuncionais não-homólogas identificadas e caracterizadas nas vias bioquímicas de *E. coli*, com base nos mapas de referência de vias metabólicas disponíveis publicamente;

4. Obter a distribuição dos genes codificadores das enzimas isofuncionais não-homólogas reconhecidas nas diferentes seções do pangenoma de *E. coli*, determinando se tais genes pertencem ao genoma *core* ou acessório;

5. Analisar o perfil filogenético (ocorrência e distribuição) das enzimas isofuncionais não-homólogas entre as cepas patogênicas de *E. coli*, revelando quais enzimas análogas funcionais podem estar relacionadas aos fenótipos de patogenicidade nesta espécie.

3. MATERIAL E MÉTODOS

3.1 Desenho Experimental

Para o desenvolvimento deste trabalho, as etapas realizadas foram ilustradas na **Figura 1**. O desenho experimental pode ser divido em 4 etapas: (1) determinação do *dataset*, (2) identificação das proteínas com atividade enzimática não-homóloga, (3) determinação do pangenoma, genoma *core* e acessório, (4) Mapeamento das enzimas análogas nas vias metabólicas KEGG. Na etapa inicial, definimos um *dataset* reunindo informações dos genomas completamente sequenciados de *E. coli* disponíveis publicamente, assim como, seus metadados. Em seguida, identificamos quais proteínas tinham atividade enzimática dentro desses genomas com a ferramenta Argot2.5 e distinguimos as não-homólogas utilizando o *pipeline* AnEnPi, as atividades enzimáticas foram então validadas de acordo com o tipo e origem evolutiva e composição de domínios funcionais, através dos bancos de dados SUPERFAMILY e Pfam, respectivamente. Posteriormente, determinamos em que sessão do pangenoma cada enzima se encontrava, se elas eram pertencentes ao genoma *core* ou acessório, as atividades enzimáticas pertencentes ao genoma acessório são responsáveis pela diversidade presente nesta espécie, visto que podem ser transferidos horizontalmente e estão presentes apenas em algumas cepas. Por fim, foi feito o mapeamento das enzimas isofuncionais não-homólogas em algumas das principais vias metabólicas de *E. coli* e quando possível correlacionar a presença ou ausência de determinada atividade enzimática ao seu perfil de patogenicidade previamente descrito.

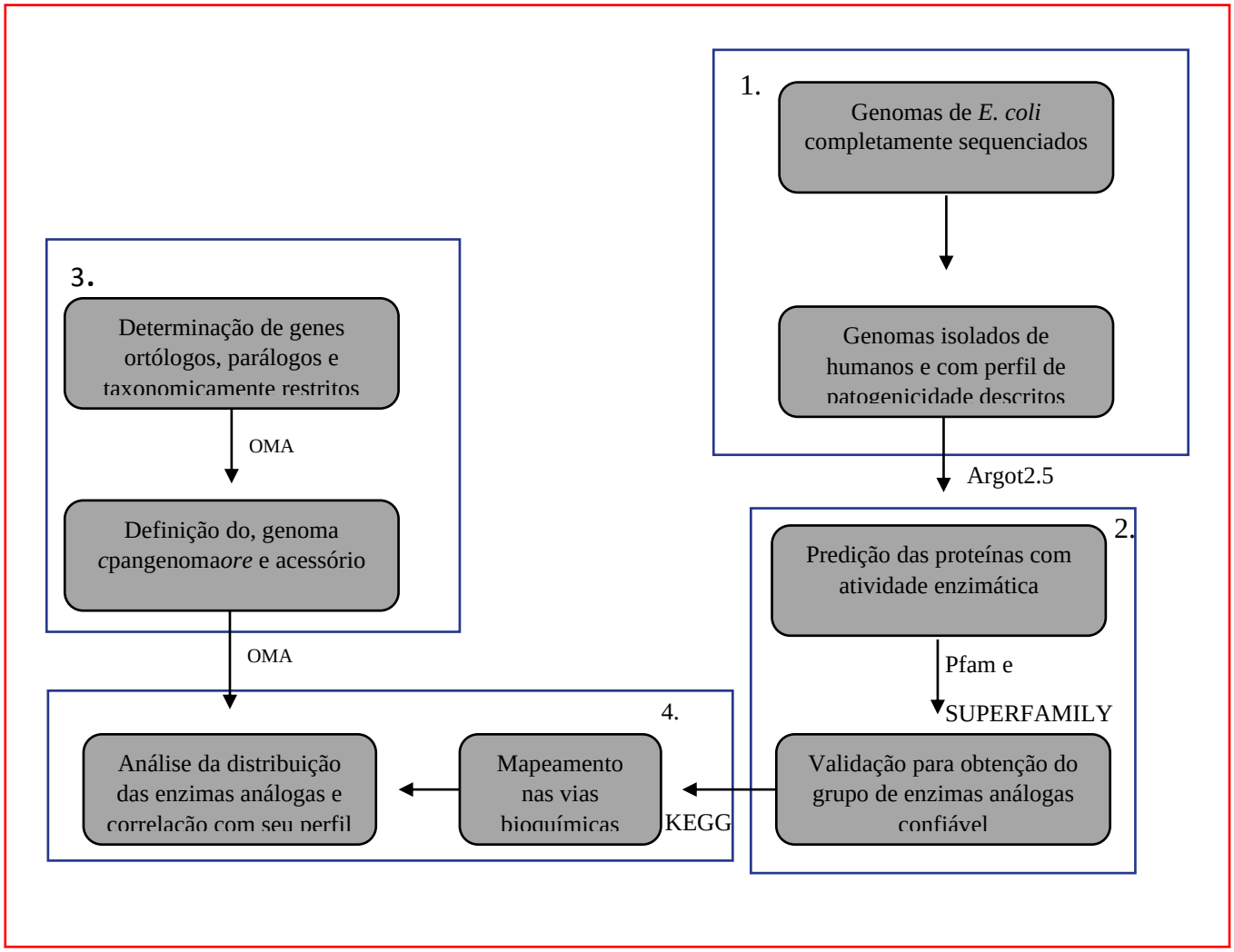

Figura1. Esquema representando as quatro etapas empregadas neste projeto.

Para compor nosso conjunto de dados, selecionamos amostras com genoma completamente sequenciado, disponíveis publicamente no *Reference Sequence Database* (RefSeq) acesso 01-08-2020 <https://www.ncbi.nlm.nih.gov/refseq/> (34). Somente as proteínas preditas nestes genomas foram utilizadas em nossas análises, contidas nos arquivos com terminação ".faa" neste repositório. As informações sobre o perfil de patogenicidade e origem biológica das amostras selecionadas foram obtidos no banco de dados *Pathosystems Resource Integration Center*

(PATRIC) versão 3.3.6 <https://www.patricbrc.org/>, que fornece dados integrados e ferramentas de análise para apoiar a pesquisa biomédica em doenças infecciosas bacterianas (35).

3.2 Determinação do pangenoma de *E. coli*

A etapa fundamental em uma análise de pangenoma é a determinação das famílias de genes ortólogos compartilhadas entre os organismos estudados. Nesse sentido, o método e banco de dados *Orthologous Matrix* (OMA) demonstrou ser capaz de inferir ortólogos em genomas completos, com bastante precisão. O algoritmo de inferência do OMA consiste em três fases principais. Primeiro, para inferir sequências homólogas (sequências de ancestralidade comum), alinhamentos par a par entre todas as sequências são computados e as correspondências significativas são mantidas. Segundo, para inferir pares de ortólogos (o subconjunto de homólogos relacionados a eventos de especiação), os homólogos mutuamente mais próximos são identificados com base nas distâncias evolutivas, levando em consideração a incerteza da inferência baseada em distância e a possibilidade de perdas de genes em linhagens específicas. Terceiro, esses ortólogos são agrupados de duas maneiras diferentes: (a) em grupos de pares de ortólogos, que tendem a ser muito específicos e úteis como genes marcadores para a reconstrução filogenética; (b) em grupos ortólogos hierárquicos, ou seja, grupos de genes definidos para níveis taxonômicos específicos, que correspondem a todos os genes que descenderam de um gene ancestral comum nesse nível taxonômico (36).

A determinação do pangenoma provém da interpretação dos resultados do OMA, a partir da análise da matriz de presença/ausência de grupos ortólogos entre os genomas, ou seja, o perfil filogenético, contido no arquivo de saída do OMA “PhylogeneticProfileOMAGroups”. Esta análise foi realizada utilizando os scripts “parse_pangenome_matrix.pl” e “plot_pancore_matrix.pl” do pacote de programas GET_HOMOLOGUES <https://github.com/eead-csic-compbio/get_homologues> (37) (38), com as configurações fornecidos pelo desenvolvedor, resultando na classificação dos grupos de genes ortólogos em pertencentes ao genoma *core*, *soft-core*, *shell* ou *cloud*, conforme as definições do desenvolvedor do programa:

➔ *Core*: genes presentes em todos os genomas analisados

➔ *Soft-core*: genes presentes em 95% ou mais genomas

➔ *Cloud*: genes presentes em menos que 5% dos genomas

➔ *Shell*: demais genes remanescentes presentes em vários genomas

3.3 Identificação de proteínas com atividade enzimática

A inferência de proteínas com atividade enzimática foi realizada com o uso do servidor Argot versão 2.5 <http://www.medcomp.medicina.unipd.it/Argot2-5/> (39). As sequências em formato FASTA, onde constavam no bando PATRIC patogenicidade a humanos, foram processadas através da busca por similaridade de sequência, através do programa BLAST versão 2.8.1 <https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/> (40), e de uma busca de perfis de proteínas, com o uso do programa HMMER versão 3.0b2 <http://hmmer.org/> (41) nos bancos de dados Swiss-Prot versão 2019_2 <https://www.uniprot.org/downloads> (42) e Pfam versão 31.0 <https://pfam.xfam.org/> (43), respectivamente; essas sequências foram então anotadas com os termos do Gene Ontology (GO) versão 01-07-2019 (44) (45) recuperados do banco de dados UniProtKB-GOA e os termos anotados foram ponderados usando os valores de E-value obtidos do BLAST e HMMER. Os termos GO ponderados foram processados de acordo com suas relações de similaridade semântica descritas pela GO e pela pontuação obtida nesse processo. A execução se deu a partir dos seguintes passos:

1- Configuração das bases de dados para as buscas

```
$ makeblastdb -in uniprot_sprot.fasta -input_type fasta -dbtype prot -out uniprot_sprot.fasta.db
```

makeblastdb – indexa o arquivo dados de entrada em “banco de dados BLAST”;
-in – arquivo de entrada;
-input_type – tipo de arquivo de entrada;
-dbtype – tipo de banco de dados;
-out – nome para os arquivos de dados indexados.

```
$ hmmspress Pfam-A.hmm
```

hmmspress – constrói arquivos de dados compactos binários para o hmmscan.

2- Busca por similaridade

```
$ blastp -num_threads 10 -outfmt '6 qseqid sseqid evalue' -query ../E.coli._protein.faa
-db uniprot_sprot.fasta.db -out E.coli_protein.faa.blast
```

blastp – algoritmo de busca por similaridade local entre proteínas;

-num_threads – número de núcleos de processamento usados na pesquisa;

-outfmt – opções de visualização do alinhamento: “6” (tabular);

-query – arquivo de entrada;

-db – nome fornecido para os arquivos de dados indexados (“banco de dados BLAST”);

-out – nome para o arquivo de saída.

```
$ hmmscan --cpu 10 --tblout E.coli_protein.faa.hmm Pfam-A.hmm E.coli_protein.faa
```

hmmscan – algoritmo de busca por similaridade local entre proteínas e perfis HMM de proteínas;

--cpu – núcleos de processamento utilizados na pesquisa;

--tblout – nome para o arquivo de saída em formato tabular.

Em seguida os arquivos de saída dos programas foram compactados em formato .zip, individualmente, e submetidos ao servidor Argot2.5, utilizando os seguintes parâmetros:

Batch processing	(opção de processamento em lote)
BLAST: E.coli_proteins.faa.blast.zip	(arquivo de entrada BLAST)
HMMER: E.coli_proteins.faa.hmm.zip	(arquivo de entrada HMMER)

Species (name or taxon ID): 562 (Escherichia coli)

E-mail: xxxxxx@xxxxx.xxx.xx

Description: E.coli.proteins

Cut-Off (meaning): >=120 (seguindo recomendações do desenvolvedor)

Semantic similarity metrics: simGIC (seguindo recomendações do desenvolvedor)

Somente as sequências com atividade enzimática predita, ou seja, que possuíam número EC atribuído, foram utilizadas nas análises subsequentes.

3.4 Identificação de enzimas isofuncionais não-homólogas em *E. coli*

Todas as enzimas preditas nos genomas analisados foram comparadas com os grupos de enzimas análogas preditos com o *pipeline* AnEnPi (46), a partir das anotações disponibilizadas na última versão gratuita da base de dados Pathway do KEGG em 2011 <https://www.genome.jp/kegg/pathway.html> (47) (48). As comparações foram feitas com o programa BLAST versão 2.8.1 <https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/> (40), utilizando como limiar de similaridade o *score* de 120, obtido com uso da matriz de substituição BLOSUM 62. Dessa forma, identificamos a que grupos de enzimas análogas formados pelo AnEnPi pertenciam as enzimas do nosso conjunto de dados de *E. coli*. Os parâmetros de busca e suas descrições encontram-se a seguir:

```
$ blastp -num_threads 10 -outfmt "6 qseqid sseqid pident lenght mismatch gapopen qstart qend sstart send evalue bitscore score" -query E.coli_go2ec2seq_proteins.faa -db anenpi.fasta.db -out E.coli_go2ec2seq_proteins.faa.blast
```

blastp – algoritmo de busca por similaridade local entre proteínas;

-num_threads – número de núcleos de processamento usados na pesquisa;

-outfmt – opções de visualização do alinhamento: “6” (tabular);

-query – arquivo de entrada;

-db – nome fornecido para os arquivos de dados indexados (“banco de dados BLAST”);

Ao final deste processo, atividades enzimáticas (e todas as sequências enzimáticas correspondentes) que apresentavam anotação incompleta (cujo número EC não estava definido

até o quarto e último nível de classificação), assim como aquelas atividades enzimáticas que continham uma ou mais sequências cujos descritores continham as palavras *chain* (cadeia), *subunit* (subunidade) e *multimeric* (multimérico) foram retiradas das análises subsequentes.

3.5 Validação das enzimas isofuncionais não-homólogas

Para validarmos as instancias de analo gia preditas, nos baseamos em informações sobre o tipo e origem evolutiva do enovelamento apresentado e a composição de domínios funcionas identificados nas sequências análogas, através de buscas nas bases de dados SUPERFAMILY <https://supfam.org/SUPERFAMILY/> (49) e PFAM (50). Os resultados foram gerados por linha de comando, utilizando a ferramenta InterProScan versão 5.46-81.0 <https://www.ebi.ac.uk/interpro/>, que fornece uma análise funcional de proteínas, classificando-as em famílias e prevendo a presença de domínios e sítios importantes (51):

```
$ ./interproscan.sh --applications Pfam --applications SUPERFAMILY --goterms --pathways --seqtype p --formats TSV,XML,GFF3,HTML,SVG –input E.coli_go2ec2seq_proteins.faa --output-dir E.coli_go2ec2seq_proteins.faa.interpro
```

--applications – argumentos utilizados para executar um ou mais programas específicos;

--goterms – fornece mapeamentos para o Gene Onology;

--pathways – fornece mapeamentos baseados em curadoria manual;

--seqtype – o tipo de sequência de entrada; “p” (proteína);

--formats – formatos de arquivos de saída;

3.6 Mapeamento das formas enzimáticas análogas nas vias metabólicas de *E. coli*

O mapeamento das enzimas análogas funcionais nas vias metabólicas de *E. coli* foi realizado com a ferramenta KEGG Mapper <https://www.genome.jp/kegg/mapper.html> (52); a ferramenta *Search & Color Pathway* <https://www.genome.jp/kegg/tool/map_pathway3.html> foi utilizada para destacar as atividades enzimáticas de interesse, colorindo-as nos mapas de referência.

4 RESULTADOS E DISCUSSÃO

4.1 Seleção dos genomas de *E. coli* e identificação do repertório de enzimas isofuncionais não-homólogas

Para compor nosso conjunto de dados, selecionamos apenas genomas completamente sequenciado, disponíveis publicamente no *RefSeq*, um conjunto abrangente, integrado, não redundante e bem anotado de sequências, incluindo DNA genômico, transcritos e proteínas (34), totalizando 637 genomas de *E. coli*. A completude dos genomas foi um critério aplicado na seleção para que não houvesse dúvidas sobre o conteúdo gênico predito nestes genomas, bem como limitar o número de genomas a serem analisados, adequando esse número à nossa capacidade computacional.

As informações sobre o tipo de isolado e perfil de patogenicidade das amostras selecionadas foram obtidas no banco de dados PATRIC, que fornece dados integrados e ferramentas de análise para apoiar a pesquisa biomédica em doenças bacterianas (35). Uma vez que nossa hipótese sugere que enzimas isofuncionais não-homólogas possam contribuir para os fenótipos de patogenicidade conhecidos em *E. coli*, é crucial dispor de informações sobre a origem e caracterização das amostras utilizadas no sequenciamento dos genomas selecionados. Dessa forma, verificamos que 252 genomas pertencentes ao grupo original de 637 haviam sido isoladas a partir de amostras humanas, dentre as quais 52 genomas apresentavam informações quanto seu perfil de patogenicidade, sendo estes os genomas selecionados para as análises posteriores (Tabela 2).

A Tabela 2 descreve o total de proteínas com atividade enzimática predita, de acordo com a classificação funcional que realizamos, baseada no uso de um vocabulário controlado de ontologia gênica.

Tabela 2. Identificadores, perfis de patogenicidade, conteúdo gênico codificador de proteínas, e proteínas com atividade enzimática predita em 52 genomas de *E.coli* isolados de humanos.

Genoma	Strain ID	Perfil de Patogenicidade	nº de proteínas	nº de atividades enzimáticas	nº de Sequências enzimáticas
GCF_001721125.1	FORC_028	EHEC	5353	942	1874
GCF_001750845.1	FORC_031	EHEC	5033	921	1791
GCF_001890205.1	O157:H7 155	EHEC	5029	933	1817
GCF_001890225.1	O157:H7 272	EHEC	5234	937	1827
GCF_001890245.1	O157:H7 472	EHEC	5228	930	1816
GCF_001890265.1	O157:H7 350	EHEC	5208	936	1816
GCF_001890325.1	O157:H7 319	EHEC	5253	938	1837
GCF_001650275.1	O157 180-PT54	EHEC	5250	935	1846
GCF_001650295.1	O157 644-PT8	EHEC	5240	937	1841
MCR-1positive|GCF_001886535.1	MRSN352231	STEC	5164	944	1827
MCR-1positive|GCF_001886555.1	MRSN346638	STEC	5162	944	1827
MCR-1positive|GCF_001886575.1	MRSN346355	STEC	5164	944	1827
MCR-1positive|GCF_001886755.1	MRSN346595	STEC	5163	944	1827
MCR-1positive|GCF_001890365.1	MRSN346647	STEC	5207	942	1850
GCF_001677515.1	08-00022	STEC	5224	952	1848
GCF_001678925.1	O7 H1827/12	STEC	5103	946	1813
GCF_001682305.2	EC590	STEC	4889	918	1742
GCF_001721225.1	CFSAN004176	STEC	5142	935	1814
GCF_001860505.1	Y5	STEC	5115	947	1825
GCF_001886895.1	FORC_029	STEC	5241	945	1847
GCF_001936315.1	SLK172	STEC	5287	947	1872
GCF_001420935.1	O165:H25 2012C-4227	STEC	5082	931	1779
GCF_001644725.1	O79:H7 2011C-3911	STEC	5158	947	1810
GCF_001644745.1	O55:H7 2013C-4465	STEC	5259	938	1848
GCF_001865295.1	PA20	STEC	5242	936	1822
GCF_001043215.1	NCM3722	STEC	5071	935	1792
GCF_001442495.1	YD786	STEC	5130	946	1825
GCF_001455385.1	CQSW20	STEC	4723	892	1648
GCF_001007915.1	O96:H19 CFSAN029787	EIEC	5168	943	1823
MCR-1positive|GCF_001693635.1	O177:H21	EIEC	5180	944	1834
GCF_001888075.1	O6:H16 FMU073332	ETEC	4992	918	1765
GCF_001469815.1	uk_P46212	ExPEC	5243	939	1885
GCF_001513635.1	JJ2434	ExPEC	5276	936	1889
GCF_001566615.1	SaT040	ExPEC	5191	937	1871
GCF_001566635.1	G749	ExPEC	5255	936	1891
GCF_001566655.1	MVAST0167	ExPEC	4991	916	1802
GCF_001566675.1	ZH193	ExPEC	5245	938	1877
GCF_001577325.1	ZH063	ExPEC	5177	934	1859
GCF_001593565.1	JJ1887	ExPEC	5279	939	1900
GCF_001617565.1	Ecol_732	ExPEC	5291	944	1899
GCF_001618325.1	Ecol_743	ExPEC	5251	946	1884
GCF_001618345.2	Ecol_745	ExPEC	5205	943	1858
GCF_001618365.1	Ecol_448	ExPEC	5253	945	1885
GCF_001663475.1	Eco889	ExPEC	5318	948	1903
GCF_001280325.1	SF-088	UPEC	5234	930	1880
GCF_001280345.1	SF-468	UPEC	5342	932	1901
GCF_001280385.1	SF-166	UPEC	5213	928	1879
GCF_001280405.1	SF-173	UPEC	5270	930	1888
GCF_001485455.1	ST648	UPEC	5200	939	1863
GCF_001693315.1	UPEC 26-1	UPEC	5240	927	1869
GCF_001721525.1	MS6198	UPEC	5473	961	1950
GCF_001874485.1	O25b:H4	UPEC	5324	940	1902

A comparação das sequências enzimáticas anotadas com os distintos grupos de enzimas análogas preditas pelo programa AnEnPi, através de uma busca por similaridade de sequência, resultou na identificação de aproximadamente 1.000 casos de analogia por genoma que, após o

refinamento desses casos eliminando atividades (e respectivas enzimas) com ECs incompletos, bem como aplicando um limiar de similaridade baseado em um BLAST *score* igual ou superior a 120 (mesmo limiar utilizado na predição de análogos pelo AnEnPi), culminou em aproximadamente 500 casos de analogia por genoma, compreendendo cerca de 800 sequências no total. Os números exatos dos casos em cada genoma são demonstrados no Anexo 1.

Uma observação importante é que o agrupamento promovido pelo AnEnPi foi realizado em 2011, data da última versão gratuita do KEGG *Pathway*. Mesmo com a atualização constante deste banco de dados, estudos anteriores conduzidos pelo nosso grupo demonstraram que as atualizações não alteraram de forma significativa os grupos de enzimas análogas formados pelo AnEnPi (46)(53).

4.2 Validação dos casos de analogia por enovelamento estrutural e composição domínios funcionais

As instâncias de analogia preditas foram validadas com base em informações sobre tipo e origem evolutiva do enovelamento proteico (49) e composição de domínios funcionais (50) das sequências envolvidas, usadas para discriminar enzimas homólogas (mesmo enovelamento e composição de domínios) de não-homólogos (distintos enovelamentos e composição de domínios). Após a validação, o número total de atividades enzimáticas e enzimas isofuncionais não-homólogas preditas corresponde respectivamente a 452 atividades enzimáticas e 1307 enzimas.

4.3 Perfil filogenético de atividades enzimáticas e enzimas isofuncionais não-homólogas entre as cepas de *E. coli.*

O repertório de enzimas isofuncionais não-homólogas identificadas e validadas, bem como das atividades enzimáticas desempenhadas por elas, foram analisados em busca de correlações (presença) e anticorrelações (ausência) existentes entre estas enzimas/atividades e os distintos grupos patogênicos de *E. coli*, definidos como presença ou ausência em pelo menos 95% dos organismos pertencentes ao mesmo grupo patogênico. Dessa forma, 71 atividades enzimáticas, representadas por 131 enzimas análogas, foram selecionadas para as análises subsequentes,

conforme apresentadas nas Tabelas 3 e 4 (correlações e anticorrelações envolvendo atividades enzimáticas) e nas Tabelas 5 e 6 (correlações e anticorrelações envolvendo enzimas isofuncionais não-homólogas) a seguir, e complementadas com informações contidas no banco de dados KEGG, para melhor descrição de função exercida e vias metabólicas, apresentadas no Anexo 2.

Tabela 3. Correlações observadas a partir da presença de atividades enzimáticas em pelo menos 95% dos organismos pertencentes ao mesmo grupo patogênico.

EC	Description	Pathotype group	Map numb	General Pathway	pathways	SSF	SUPERFAMILY
3.5.1.5	urease	EHEC	220 230 791 1100 1120	Amino acid metabolism Nucleotide metabolism Xenobiotics biodegradation and metabolism	Arginine biosynthesis Purine metabolism Atrazine degradation Metabolic pathways Microbial metabolism	SSF51278 SSF51338 SSF51556 SSF54111	Urease, beta subunit Urease alpha-subunit Metal-dependent hydrolase Urease, gamma subunit
6.3.1.11	glutamate—putrescine ligase	EHEC EIEC ETEC	330 1100	Amino acid metabolism	Arginine and proline metabolism Metabolic pathways	SSF54368 SSF55931	Glutamine synthetase Glutamine synthetase/guanido kinase, catalytic domain
3.2.1.26	β-fructofuranosidase	EHEC EIEC ExPEC	52 500 1100	Carbohydrate metabolism	Galactose metabolism Starch and sucrose metabolism Metabolic pathways	SSF49899 SSF75005	Concanavalin A-like lectin/glucanase domain Glycosyl hydrolase, five-bladed beta-propellor domain
4.1.1.61	4-hydroxybenzoate	EHEC EIEC ExPEC	627 1120	Xenobiotics biodegradation and metabolism	Aminobenzoate degradation Microbial metabolism in diverse environments	SSF143968 SSF50475	-
3.2.1.122	maltose-6′-phosphate glucosidase	EHEC ETEC	500	Carbohydrate metabolism	Starch and sucrose metabolism	SSF51735 SSF56327	Lactate dehydrogenase/glycoside hydrolase, family 4 NAD(P)-binding domain
1.4.3.21	primary-amine oxidase	EIEC	260 350 360 410 950 960 1100 1110	Amino acid metabolism Metabolism of other amino acids Carbohydrate metabolism Biosynthesis of other secondary metabolites	Glycine, serine and threonine metabolism Tyrosine metabolism Phenylalanine metabolism beta-Alanine metabolism Isoquinoline alkaloid biosynthesis Tropane, piperidine and pyridine alkaloid biosynthesis Metabolic pathways Biosynthesis of secondary metabolites	SSF49998 SSF54416 SSF55383	Copper amine oxidase, catalytic domain Copper amine oxidase Copper amine oxidase-like
3.2.1.24	α-mannosidase	EIEC ETEC ExPEC	511	Glycan biosynthesis and metabolism	Other glycan degradation	SSF74650 SSF88688 SSF88713	Glycoside hydrolase family 38 Glycoside hydrolase families 57/38 Glycoside hydrolase/deacetylase, beta/alpha-barrel
1.1.99.14	glycolate dehydrogenase	EIEC ExPEC UPEC	630 1100 1120	Carbohydrate metabolism	Glyoxylate and dicarboxylate metabolism Metabolic pathways Microbial metabolism	SSF54862 SSF55103 SSF56176	FAD-linked oxidase-like FAD type PCMH-like
1.1.1.11	D-arabinitol 4-dehydrogenase	ETEC	40 51 1100	Carbohydrate metabolism	Pentose and glucuronate interconversions Fructose and mannose metabolism Metabolic pathways	SSF48179 SSF51735	6-phosphogluconate dehydrogenase-like NAD(P)-binding domain superfamily
6.6.1.2	cobaltochelatase	ETEC	860 1100	Metabolism of cofactors and vitamins	Porphyrin and chlorophyll metabolism Metabolic pathways	SSF52540 SSF53300	P-loop containing nucleoside triphosphate hydrolase von Willebrand factor A-like domain superfamily
3.2.1.14	chitinase	ETEC ExPEC	520	Carbohydrate metabolism	Amino sugar and nucleotide sugar metabolism	SSF51055 SSF51445	arbohydrate-binding module superfamily 5/12 Glycoside hydrolase superfamily
3.5.1.47	N-acetyldiaminopimelate deacetylase	ExPEC	300 1100 1110	Amino acid metabolism	Lysine biosynthesis Metabolic pathways Biosynthesis of secondary metabolites	SSF53187 SSF55031	l exopeptidase dimerisation
1.18.1.3	ferredoxin—NAD+ reductase	ExPEC UPEC	71	Lipid metabolism	Fatty acid degradation	SSF51905 SSF55424	FAD/NAD-linked reductase, dimerisation domain FAD/NAD(P)-binding domain superfamily
1.1.1.14	L-iditol 2-dehydrogenase	STEC	40 51 1100	Carbohydrate metabolism	Pentose and glucuronate interconversions Fructose and mannose metabolism Metabolic pathways	SSF51735 SSF50129	NAD(P)-binding domain superfamily GroES-like superfamily

Tabela 4. Anticorrelações observadas a partir da ausência de atividades enzimáticas em pelo menos 95% dos organismos pertencentes ao mesmo grupo patogênico.

EC	Description	Pathotype group	Map numb	General Pathway	pathways	SSF	SUPERFAMILY
1.1.1.251	galactitol-1-phosphate 5-dehydrogenase	EHEC ETEC	52 1100	Carbohydrate metabolism	Galactose metabolism Metabolic pathways	SSF50129 SSF51735	NAD(P)-binding domain superfamily GroES-like superfamily
3.5.2.5	allantoinase	EHEC ETEC	230 1100 1120	Nucleotide metabolism	Purine metabolism Metabolic pathways Microbial metabolism in diverse environments	SSF51338 SSF51556	Metal-dependent hydrolase Metal-dependent hydrolase
4.2.1.39	gluconate dehydratase	EHEC STEC	30 1100 1120	Carbohydrate metabolism	Pentose phosphate pathway Metabolic pathways Microbial metabolism in diverse environments	SSF51604 SSF54826	Enolase-like, C-terminal Enolase-like, N-terminal
4.2.1.90	L-rhamnonate dehydratase	EHEC STEC	51 1120	Carbohydrate metabolism	Fructose and mannose metabolism Microbial metabolism in diverse environments	SSF51604 SSF54826	Enolase-like Enolase-like
3.2.1.31	β-glucuronidase	EHEC UPEC	40 500 531 860 944 983 1100	Carbohydrate metabolism Glycan biosynthesis and metabolism Metabolism of cofactors and vitamins Biosynthesis of other secondary metabolites Xenobiotics biodegradation and metabolism	Pentose and glucuronate interconversions Starch and sucrose metabolism Glycosaminoglycan degradation Porphyrin and chlorophyll metabolism Flavone and flavonol biosynthesis Drug metabolism - other enzymes Metabolic pathways	SSF49303 SSF49785 SSF51445	Galactose-binding-like Glycoside hydrolase superfamily Beta-Galactosidase/glucuronidase
2.1.2.3	phosphoribosylaminoimidazolecarboxamide formyltransferase	EIEC	230 670 1100 1110 1130	Nucleotide metabolism Metabolism of cofactors and vitamins	Purine metabolism One carbon pool by folate Metabolic pathways Biosynthesis of secondary metabolites Biosynthesis of antibiotics	SSF52335 SSF53927	Methylglyoxal synthase-like Cytidine deaminase-like
3.5.4.10	IMP cyclohydrolase	EIEC	230 1100 1110 1130	Nucleotide metabolism	Purine metabolism Metabolic pathways Biosynthesis of secondary metabolites Biosynthesis of antibiotics	SSF52335 SSF53927	Methylglyoxal synthase-like Cytidine deaminase-like
5.3.3.10	5-carboxymethyl-2-hydroxymuconate Δ-isomerase	EIEC	350 1120	Amino acid metabolism	Tyrosine metabolism Microbial metabolism in diverse environments	SSF55331 SSF56529	Tautomerase/MIF superfamily
1.99.1.1	choline dehydrogenase	ETEC	260 110	Amino acid metabolism	Glycine, serine and threonine metabolism Metabolic pathways	SSF51905 SSF54373	
3.5.1.78	glutathionylspermidine amidase	ETEC	480 1100	Metabolism of other amino acids	Metabolic pathways	SSF52440 SSF54001 SSF56059	Papain-like cysteine peptidase Pre-ATP-grasp domain
3.5.3.9	allantoate deiminase	ETEC	230 1120	Nucleotide metabolism	Purine metabolism Microbial metabolism in diverse environments	SSF53187 SSF55031	ial exopeptidase dimerisation d
5.4.99.2	methylmalonyl-CoA mutase	ETEC	280 630 640 720 1100 1120	Amino acid metabolism Carbohydrate metabolism Energy metabolism	Valine, leucine and isoleucine degradation Glyoxylate and dicarboxylate metabolism Propanoate metabolism Carbon fixation pathways in prokaryotes Metabolic pathways Microbial metabolism in diverse environments	SSF51703 SSF52242	Cobalamin (vitamin B12)-dependent enzyme, catalytic Cobalamin-binding domain
6.3.1.8	glutathionylspermidine synthase	ETEC	480 1100	Glutathione metabolism	Glutathione metabolism Metabolic pathways	SSF52440 SSF54001 SSF56059	Papain-like cysteine peptidase Pre-ATP-grasp domain
3.5.4.1	cytosine deaminase	ETEC STEC	240 330 1100	Nucleotide metabolism Amino acid metabolism	Pyrimidine metabolism Arginine and proline metabolism Metabolic pathways	SSF51338 SSF51556	Metal-dependent hydrolase Metal-dependent hydrolase
3.2.1.48	sucrose α-glucosidase	ETEC STEC UPEC	500 1100	Carbohydrate metabolism	Starch and sucrose metabolism Metabolic pathways	SSF49899 SSF75005	Concanavalin A-like lectin/glucanase Glycosyl hydrolase, five-bladed beta-propellor
2.4.1.12	cellulose synthase (UDP-forming)	ExPEC	500 1100	Carbohydrate metabolism	Starch and sucrose metabolism Metabolic pathways	SSF141371 SSF53448	-
1.14.12.19	3-phenylpropanoate dioxygenase	ExPEC UPEC	360 1120	Amino acid metabolism	Phenylalanine metabolism Microbial metabolism in diverse environments	SSF50022 SSF54427 SSF55961	Rieske [2Fe-2S] iron-sulphur domain NTF2-like domain superfamily
4.2.1.104	cyanase	ExPEC UPEC	910	Energy metabolism	Nitrogen metabolism	SSF47413 SSF55234	Lambda repressor-like, DNA-binding domain Cyanate lyase, C-terminal
2.4.1.64	α,α-trehalose phosphorylase	STEC	500	Carbohydrate metabolism	Starch and sucrose metabolism	SSF48208 SSF74650	Six-hairpin glycosidase superfamily Galactose mutarotase-like domain
2.7.7.13	mannose-1-phosphate guanylyltransferase	STEC	51 520 1100 1110	Carbohydrate metabolism	Fructose and mannose metabolism Amino sugar and nucleotide sugar metabolism Metabolic pathways Biosynthesis of secondary metabolites	SSF51182 SSF53448	RmlC-like cupin domain Nucleotide-diphospho-sugar transferases
3.1.4.16	2',3'-cyclic-nucleotide 2'-phosphodiesterase	STEC	230 240	Nucleotide metabolism	Purine metabolism Pyrimidine metabolism	SSF55816 SSF56300	5'-Nucleotidase, C-terminal
5.4.2.8	phosphomannomutase	STEC	51 520 1100 1110	Carbohydrate metabolism	Fructose and mannose metabolism Amino sugar and nucleotide sugar metabolism Metabolic pathways Biosynthesis of secondary metabolites	SSF53738 SSF55957	Alpha-D-phosphohexomutase Alpha-D-phosphohexomutase, C-terminal
4.1.3.39	4-hydroxy-2-oxovalerate aldolase	UPEC	360 362 621 622 1100 1120	Amino acid metabolism Xenobiotics biodegradation and metabolism	Phenylalanine metabolism Benzoate degradation Dioxin degradation Xylene degradation Metabolic pathways Microbial metabolism in diverse environments	SSF51569 SSF89000	-

Tabela 5. Correlações observadas a partir da presença de enzimas isofuncionais não-homólogas em pelo menos 95% dos organismos pertencentes ao mesmo grupo patogênico.

EC	Description	Pathotype group	Map numb	General Pathway	pathways	SSF	SUPERFAMILY
2.3.1.5	arylamine N-acetyltransferase	EHEC	232 633 983 1100 1110 1120	Biosynthesis of other secondary metabolites Xenobiotics biodegradation and metabolism	Caffeine metabolism Nitrotoluene degradation Drug metabolism - other enzymes Metabolic pathways Biosynthesis of secondary metabolites Microbial metabolism in diverse environments	SSF47598	Ribbon-helix-helix
2.7.7.6	DNA-directed RNA polymerase	EIEC ETEC	230 240 1100	Nucleotide metabolism	Purine metabolism Pyrimidine metabolism Metabolic pathways	SSF46785 SSF52540	Winged helix DNA-binding domain P-loop containing nucleoside triphosphate hydrolase
3.5.2.6	β-lactamase	EIEC ETEC	311 1130	Biosynthesis of other secondary metabolites	Penicillin and cephalosporin biosynthesis Biosynthesis of antibiotics	SSF81901	-
3.2.1.86	6-phospho-β-glucosidase	EIEC ETEC ExPEC	10	Carbohydrate metabolism	Glycolysis / Gluconeogenesis	SSF51445	Glycoside hydrolase
2.7.7.7	NA-directed DNA polymerase	ETEC	230 240 1100	Nucleotide metabolism	Purine metabolism Pyrimidine metabolism Metabolic pathways	SSF52141	Uracil-DNA glycosylase-like
3.6.1.3	adenosinetriphosphatase	ETEC	230 340 1100 1110	Nucleotide metabolism Amino acid metabolism	Purine metabolism Histidine metabolism Metabolic pathways Biosynthesis of secondary metabolites	SSF52833	-
2.3.1.5	arylamine N-acetyltransferase	ETEC ExPEC	232 633 983 1100 1110 1120	Biosynthesis of other secondary metabolites Xenobiotics biodegradation and metabolism	Caffeine metabolism Nitrotoluene degradation Drug metabolism - other enzymes Metabolic pathways Biosynthesis of secondary metabolites Microbial metabolism in diverse environments	SSF47598	Ribbon-helix-helix
1.17.1.4	xanthine dehydrogenase	ETEC UPEC	230 1100 1120	Nucleotide metabolism	Purine metabolism Metabolic pathways Microbial metabolism in diverse environments	SSF53187	-
2.7.1.15	ribokinase	ExPEC	30	Carbohydrate metabolism	Pentose phosphate pathway	SSF46785	Winged helix DNA-binding domain
3.2.1.23	β-galactosidase	ExPEC	52 511 531 600 604 1100	Carbohydrate metabolism Glycan biosynthesis and metabolism Lipid metabolism	Galactose metabolism Other glycan degradation Glycosaminoglycan degradation Sphingolipid metabolism Glycosphingolipid biosynthesis - ganglio series Metabolic pathways	SSF51197	-
3.6.1.3	adenosinetriphosphatase	ExPEC	230 340 1100 1110	Nucleotide metabolism Amino acid metabolism	Purine metabolism Histidine metabolism Metabolic pathways Biosynthesis of secondary metabolites	SSF158682	TerB-like
5.4.99.5	chorismate mutase	ExPEC UPEC	400 1100 1120 1130	Amino acid metabolism	Phenylalanine, tyrosine and tryptophan biosynthesis Metabolic pathways Biosynthesis of secondary metabolites Biosynthesis of antibiotics	SSF56322	ADC synthase
2.3.1.157	glucosamine-1-phosphate N-acetyltransferase	STEC	510 1100 1130	Glycan biosynthesis and metabolism	Amino sugar and nucleotide sugar metabolism Metabolic pathways Biosynthesis of antibiotics	SSF117856	-

Tabela 6. Anticorrelações observadas a partir da ausência de enzimas isofuncionais não-homólogas em pelo menos 95% dos organismos pertencentes ao mesmo grupo patogênico.

EC	Description	Pathotype group	Map numb	General Pathway	pathways	SSF	SUPERFAMILY
2.3.1.5	arylamine N-acetyltransferase	EHEC STEC	232 633 983 1100 1110 1120	Biosynthesis of other secondary metabolites Xenobiotics biodegradation and metabolism	Caffeine metabolism Nitrotoluene degradation Drug metabolism - other enzymes Metabolic pathways Biosynthesis of secondary metabolites Microbial metabolism in diverse environments	SSF54001	Papain-like cysteine peptidase superfamily
2.5.1.47	cysteine synthase	EHEC STEC	270 920 1100 1110 1130	Amino acid metabolism Energy metabolism	Cysteine and methionine metabolism Sulfur metabolism Metabolic pathways Biosynthesis of secondary metabolites Biosynthesis of antibiotics	SSF53383	yridoxal phosphate-dependent transferase
3.5.1.1	asparaginase	EHEC STEC	250 460 1100 1110	Amino acid metabolism Metabolism of other amino acids	Alanine, aspartate and glutamate Cyanoamino acid metabolism Metabolic pathways Biosynthesis of secondary metabolites	SSF56235	Nucleophile aminohydrolases, N-terminal
4.2.1.8	mannonate dehydratase	EHEC STEC	40 1100	Carbohydrate metabolism	Pentose and glucuronate interconversions Metabolic pathways	SSF51604 SSF54826	Enolase-like
5.3.1.6	ribose-5-phosphate isomerase	EHEC STEC	30 51 710 1100 1110 1120 1130	Carbohydrate metabolism Energy metabolism	Pentose phosphate pathway Fructose and mannose metabolism Carbon fixation in photosynthetic organisms Metabolic pathways Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Biosynthesis of antibiotics	SSF89623	ugar-phosphate isomerase, RpiB/LacA/LacE
1.2.1.3	aldehyde dehydrogenase (NAD+)	EIEC	10 40 53 71 280 310 330 340 380 410 562 620 625 903 1100 1110 1120 1130	Carbohydrate metabolism Lipid metabolism Amino acid metabolism Metabolism of other amino acids Xenobiotics biodegradation and metabolism Metabolism of terpenoids and polyketides	Glycolysis / Gluconeogenesis Pentose and glucuronate interconversions Ascorbate and aldarate metabolism Fatty acid degradation Valine, leucine and isoleucine degradation Lysine degradation Arginine and proline metabolism Histidine metabolism Tryptophan metabolism beta-Alanine metabolism Glycerolipid metabolism Pyruvate metabolism Chloroalkane and chloroalkene degradation Limonene and pinene degradation Metabolic pathways Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Biosynthesis of antibiotics	SSF54637	HotDog domain
1.6.99.3	NADH dehydrogenase	EIEC	190 1100	Energy metabolism	Oxidative phosphorylation Metabolic pathways	SSF140490 SSF142019 SSF142984 SSF52833	-
2.5.1.17	corrinoid adenosyltransferase	EIEC	860 1100	Metabolism of cofactors and vitamins	Porphyrin and chlorophyll metabolism Metabolic pathways	SSF52540	p containing nucleoside triphosphate hydr
3.1.3.5	5'-nucleotidase	EIEC	230 240 760 1100 1110	Nucleotide metabolism Metabolism of cofactors and vitamins	Purine metabolism Pyrimidine metabolism Nicotinate and nicotinamide metabolism Metabolic pathways Biosynthesis of secondary metabolites	SSF55816 SSF56300	5'-Nucleotidase
5.3.1.13	arabinose-5-phosphate isomerase	EIEC	540 1100	Glycan biosynthesis and metabolism	Lipopolysaccharide biosynthesis Metabolic pathways	SSF54631	-
1.1.1.2	alcohol dehydrogenase (NADP+)	ETEC	10 40 561 930 1100 1110 1120 1130	Carbohydrate metabolism Lipid metabolism Xenobiotics biodegradation and metabolism	Glycolysis / Gluconeogenesis Pentose and glucuronate interconversions Glycerolipid metabolism Caprolactam degradation Metabolic pathways Biosynthesis of secondary metabolites Microbial metabolism in diverse environments Biosynthesis of antibiotics	SSF51430	NADP-dependent oxidoreductase domain
1.2.1.8	betaine-aldehyde dehydrogenase	ETEC	260 1100	Amino acid metabolism	Glycine, serine and threonine metabolism Metabolic pathways	SSF51905 SSF54373	FAD/NAD(P)-binding domain superfamily
1.3.1.1	dihydropyrimidine dehydrogenase (NAD+)	ETEC	240 410 770 1100	Nucleotide metabolism Metabolism of other amino acids Metabolism of cofactors and vitamins	Pyrimidine metabolism beta-Alanine metabolism Pantothenate and CoA biosynthesis Metabolic pathways	SSF51395 SSF54862	-
1.3.5.2	dihydroorotate dehydrogenase (quinone)	ETEC	240 1100	Nucleotide metabolism	Pyrimidine metabolism Metabolic pathways	SSF54862	-
3.1.3.27	phosphatidylglycerophosphatase	ETEC	564 1100	Lipid metabolism	Glycerophospholipid metabolism Metabolic pathways	SSF56784	HAD-like superfamily
3.1.1.23	acylglycerol lipase	ETEC STEC ExPEC	561 1100	Lipid metabolism	Glycerolipid metabolism Metabolic pathways	SSF53335	osyl-L-methionine-dependent methyltrans
2.7.1.29	glycerone kinase	ExPEC	561 680 1100 1120	Lipid metabolism Energy metabolism	Glycerolipid metabolism Methane metabolism Metabolic pathways Microbial metabolism in diverse environments	SSF82549	-
2.5.1.3	thiamine phosphate synthase	ExPEC UPEC	730 1100	Metabolism of cofactors and vitamins	Thiamine metabolism Metabolic pathways	SSF53613	Ribokinase-like
2.7.7.23	UDP-N-acetylglucosamine diphosphorylase	STEC	520 1100 1130	Carbohydrate metabolism	Amino sugar and nucleotide sugar metabolism Metabolic pathways Biosynthesis of antibiotics	SSF117856	-
3.1.3.6	3'-nucleotidase	STEC	230 240	Nucleotide metabolism	Purine metabolism Pyrimidine metabolism	SSF55816 SSF56300	5'-Nucleotidase, C-terminal domain
5.3.1.8	mannose-6-phosphate isomerase	STEC	51 520 1100 1110	Carbohydrate metabolism	Fructose and mannose metabolism Amino sugar and nucleotide sugar metabolism Metabolic pathways Biosynthesis of secondary metabolites	SSF53448	Nucleotide-diphospho-sugar transferases
1.2.1.10	acetaldehyde dehydrogenase (acetylating)	UPEC	360 362 620 621 622 650 1100 1120	Amino acid metabolism Xenobiotics biodegradation and metabolism Carbohydrate metabolism	Phenylalanine metabolism Benzoate degradation Pyruvate metabolism Dioxin degradation Xylene degradation Butanoate metabolism Metabolic pathways Microbial metabolism in diverse environments	SSF51735 SSF55347	NAD(P)-binding domain superfamily

4.4 Distribuição dos genes codificadores de enzimas isofuncionais não-homólogas nas diferentes seções do pangenoma de *E. coli*.

Os genes pertencentes ao genoma *core* são compartilhados entre todas as cepas dentro de uma espécie, e estão associados a aspectos básicos da biologia e aos principais traços fenotípicos do organismo; estes genes são compreendidos pelos subgrupos *core* e *soft-core,* obtidos através da interpretação dos resultados do OMA. Os genes pertencentes ao genoma acessório compreendem os subgrupos *cloud* e *shell,* e contribuem para a diversidade observada nas amostras desta espécie, por exemplo codificando atividades enzimáticas de vias bioquímicas suplementares, realizando funções não essências para o crescimento, mas que conferem vantagens seletivas, como a adaptação a diferentes nichos, resistência a antibióticos ou colonização de um novo hospedeiro. Há ainda genes taxonomicamente restritos, ou seja, genes exclusivos de uma determinada linhagem ou cepa. Os critérios aplicados para a determinação do pangenoma de *E. coli*, definindo operacionalmente os distintos grupos foi a seguinte: *core:* genes presentes em todos os genomas; *soft-core:* genes compartilhados por ≥ 95% dos genomas; *cloud:* genes encontrados em ≤ 5% das cepas; *shell:* demais genes, linhagem-específicos (54).

Uma parte considerável do genoma de *E. coli* foi supostamente obtida através de ganho gênico por transferência horizontal, um dos principais meios de diversificação nesta espécie e inúmeras outras, geralmente conferindo vantagens adaptativas para o hospedeiro (55). A distribuição das enzimas análogas nas diferentes seções do pangenoma de *E. coli* mostrou que, entre as 131 enzimas isofuncionais não-homólogas, representantes de 71 atividades enzimáticas, 81 destas enzimas, totalizando 45 atividades, pertencem ao genoma acessório, as quais selecionamos para as análises posteriores (Tabela 6).

Tabela 7. Enzimas análogas pertencentes ao genoma acessório

legenda:				
	verde escuro: indica presença de determinada atividade enzimática			
	vermelho escuro: indica ausência de determinada atividade enzimá			
	verde claro: Indica presença de determinada forma análoga indica			
	vermelho claro: Indica ausência de determinada forma análoga inc			
EC	Pathotype group	SSF	OMA	Pangenome
1.1.1.11	ETEC	SSF48179 SSF51735	1542	shell
1.1.1.14	STEC	SSF51735 SSF50129	2820 2855	shell
1.1.1.2	ETEC	SSF51430	4094	core
1.17.1.4	ETEC UPEC	SSF53187	1036	core
1.18.1.3	ExPEC UPEC	SSF51905 SSF55424	2181	shell
1.2.1.10	UPEC	SSF51735 SSF55347	2610	core
1.2.1.3	EIEC	SSF54637	598	shell
1.2.1.8	ETEC	SSF51905 SSF54373	872	core
1.3.1.1	ETEC	SSF51395 SSF54862	1964	core
1.3.5.2	ETEC	SSF54862	1964	core
1.4.3.21	EIEC	SSF49998 SSF54416 SSF55383	425	shell
1.6.99.3	EIEC	SSF140490 SSF142019 SSF142984 SSF52833	1672	core
1.99.1.1	ETEC	SSF51905 SSF54373	-	-
2.1.2.3	EIEC	SSF52335 SSF53927	1068	core
2.3.1.157	STEC	SSF117856	7052	shell
2.3.1.5	EHEC	SSF47598		shell
2.3.1.5	EHEC ETEC ExPEC	SSF47598	-	-
2.3.1.5	EHEC STEC	SSF54001	3917	shell
2.4.1.64	STEC	SSF48208 SSF74650	459	shell
2.5.1.17	EIEC	SSF52540	5727	core
2.5.1.3	ExPEC UPEC	SSF53613	4317	core
2.5.1.47	EHEC STEC	SSF53383	2740	shell
2.7.1.15	ExPEC	SSF46785	2042 2159	shell
2.7.1.29	ExPEC	SSF82549	2768	core
2.7.7.13	STEC	SSF51182 SSF53448	1375	shell
2.7.7.23	STEC	SSF117856	7052	shell
2.7.7.6	EIEC ETEC	SSF46785 SSF52540	854	shell
2.7.7.7	ETEC	SSF52141	123	shell
3.1.1.23	ETEC STEC ExPEC	SSF53335	833	shell
3.1.3.27	ETEC	SSF56784	5388	core
3.1.3.5	EIEC	SSF55816 SSF56300	962	core
3.1.3.6	STEC	SSF55816 SSF56300	673	core
3.1.4.16	STEC	SSF55816 SSF56300	673	core

3.2.1.122	EHEC ETEC	SSF51735 SSF56327	1674	shell
3.2.1.14	ETEC ExPEC	SSF51055 SSF51445	233	shell
3.2.1.23	ExPEC	SSF51197	6673	shell
3.2.1.24	EIEC ETEC ExPEC	SSF74650 SSF88688 SSF88713	254	shell
3.2.1.26	EHEC EIEC ExPEC	SSF49899 SSF75005	1248	shell
3.2.1.31	EHEC UPEC	SSF49303 SSF49785 SSF51445	755	shell
3.2.1.48	ETEC STEC UPEC	SSF49899 SSF75005	1248	shell
3.2.1.86	EIEC ETEC ExPEC	SSF51445	1374	shell
3.5.1.1	EHEC STEC	SSF56235	7982 3372	shell
3.5.1.47	ExPEC	SSF53187 SSF55031	2319	shell
3.5.1.78	ETEC	SSF52440 SSF54001 SSF56059	719	core
3.5.2.5	EHEC ETEC	SSF51338 SSF51556	1566	shell
3.5.2.6	EIEC ETEC	SSF81901	2521	shell
3.5.3.9	ETEC	SSF53187 SSF55031	1996	core
3.5.4.1	ETEC STEC	SSF51338 SSF51556	1815	core
3.5.4.10	EIEC	SSF52335 SSF53927	1068	core
3.6.1.3	ETEC	SSF52833	-	-
3.6.1.3	ExPEC	SSF158682	5373	core
4.1.1.61	EHEC EIEC ExPEC	SSF143968 SSF50475	1403	shell
4.1.3.39	UPEC	SSF51569 SSF89000	3025	shell
4.2.1.104	ExPEC UPEC	SSF47413 SSF55234	6824	shell
4.2.1.39	EHEC STEC	SSF51604 SSF54826	2012	shell
4.2.1.8	EHEC STEC	SSF51604 SSF54826	2012	shell
4.2.1.90	EHEC STEC	SSF51604 SSF54826	2061	shell
5.3.1.13	EIEC	SSF54631	3385	core
5.3.1.6	EHEC STEC	SSF89623	8200	shell
5.3.1.8	STEC	SSF53448	1340	shell
5.3.3.10	EIEC	SSF55331 SSF56529	7512	shell
5.4.99.2	ETEC	SSF51703 SSF52242	2743	shell
5.4.99.5	ExPEC UPEC	SSF56322	1776	shell
6.3.1.11	EHEC EIEC ETEC	SSF54368 SSF55931	1404	shell
6.3.1.8	ETEC	SSF52440 SSF54001 SSF56059	719	core
6.6.1.2	ETEC	SSF52540 SSF53300	2077	shell

4.5 Mapeamento das atividades enzimáticas e das enzimas isofuncionais não-homólogas em vias bioquímicas de referência de *E. coli*.

O mapeamento das enzimas isofuncionais não-homólogas mostrou a participação destas enzimas em vias bioquímicas relevantes, totalizando 32 vias nas quais observamos correlações, e 34 vias nas quais observamos anticorrelações. Algumas vias essenciais e a atuação destas atividades e respectivas enzimas análogas serão discutidas a seguir, através dos mapas metabólicos de referência do KEGG. Em todos os mapas a cor rosa indica uma atividade enzimática ou uma enzima isofuncional não-homóloga correlacionada ou anticorrelacionada a um ou mais perfis de patogenicidade. A cor verde destaca as atividades enzimáticas presentes no metabolismo de *E.coli*. As demais atividades representam enzimas que podem catalisar reações na via metabólica analisada, porém em organismos diferentes de *E.coli*.

Na via da interconversão de pentose e glucuronato encontramos uma atividade enzimática e uma enzima isofuncional não-homóloga anticorrelacionadas ao perfil EHEC: a β-guluronidase (EC 3.2.1.31), codificada pelo gene uidA em bactérias, ausente nos demais perfis de patogenicidade. Em humanos a deficiência dessa enzima resulta na doença metabólica síndrome de Sly. O gene uidA é frequentemente utilizado como alvo de marcação na medição de coliformes fecais no ambiente. E uma, dentre as três enzimas isofuncionais não-homólogas com atividade enzimática mannonate dehydratase (4.2.1.8), codificada pelo gene uxuA, ausente nos demais grupos patogênicos. (Figura 2).

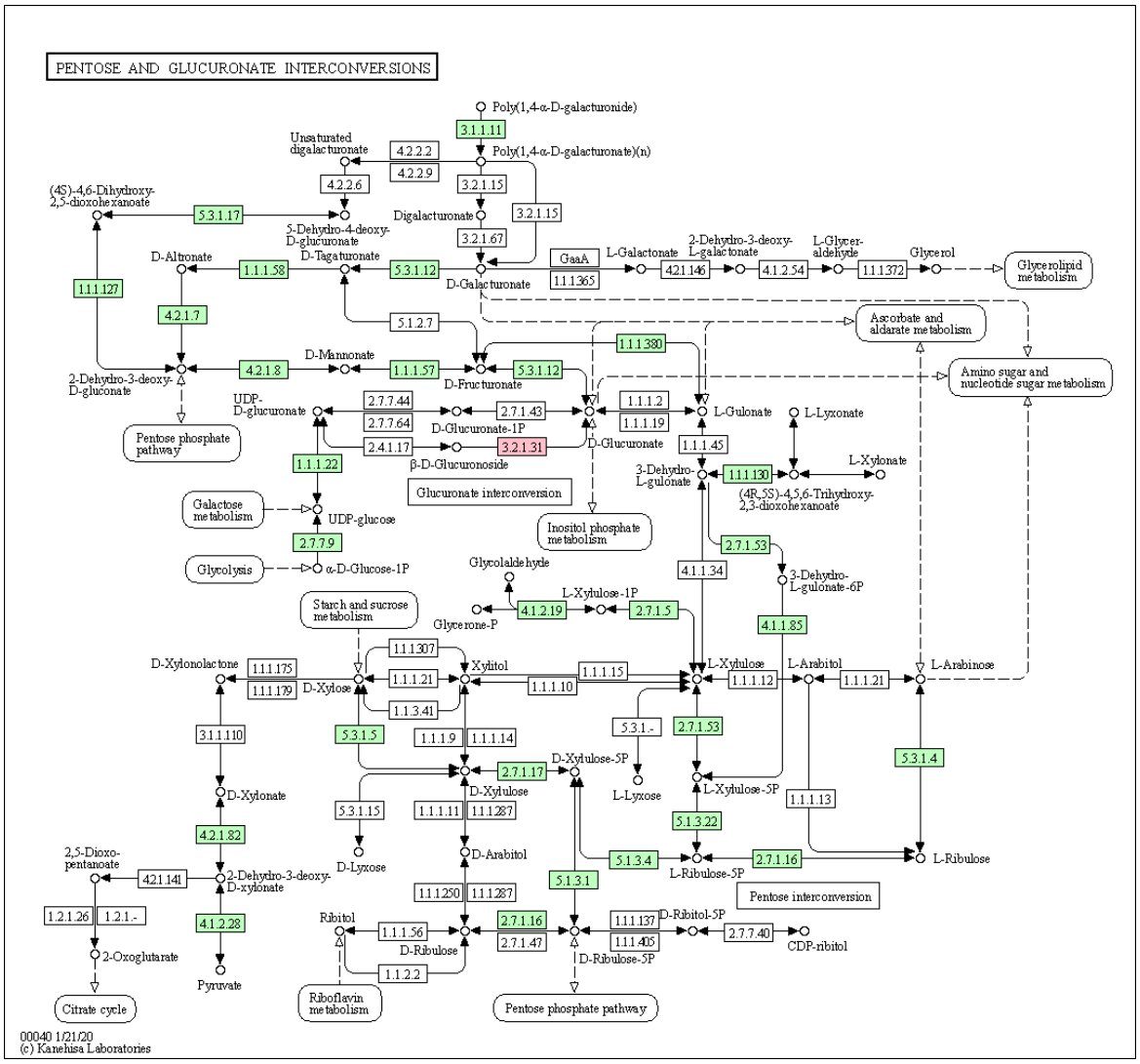

Figura 2. Via da interconversão de pentose e gluconato (map 00040) de *E.coli.*

Na via de degradação de ácidos graxos (Figura 3), via muito conservada do ponto de vista evolutivo e que ocupa uma posição central no metaboloma microbiano, identificamos a enzima ferredoxina NAD+ reductase (EC 1.18.1.3) exclusivamente encontradas nos perfis de patogenicidade extraintestinais ExPEC e UPEC, em plantas, essa enzima possui papel central na fotossíntese. Pouco se sabe a respeito da atuação dela nas bactérias, mas trata-se de cadeias anaeróbicas de transporte de elétrons, sendo estruturas bem conservadas (56).

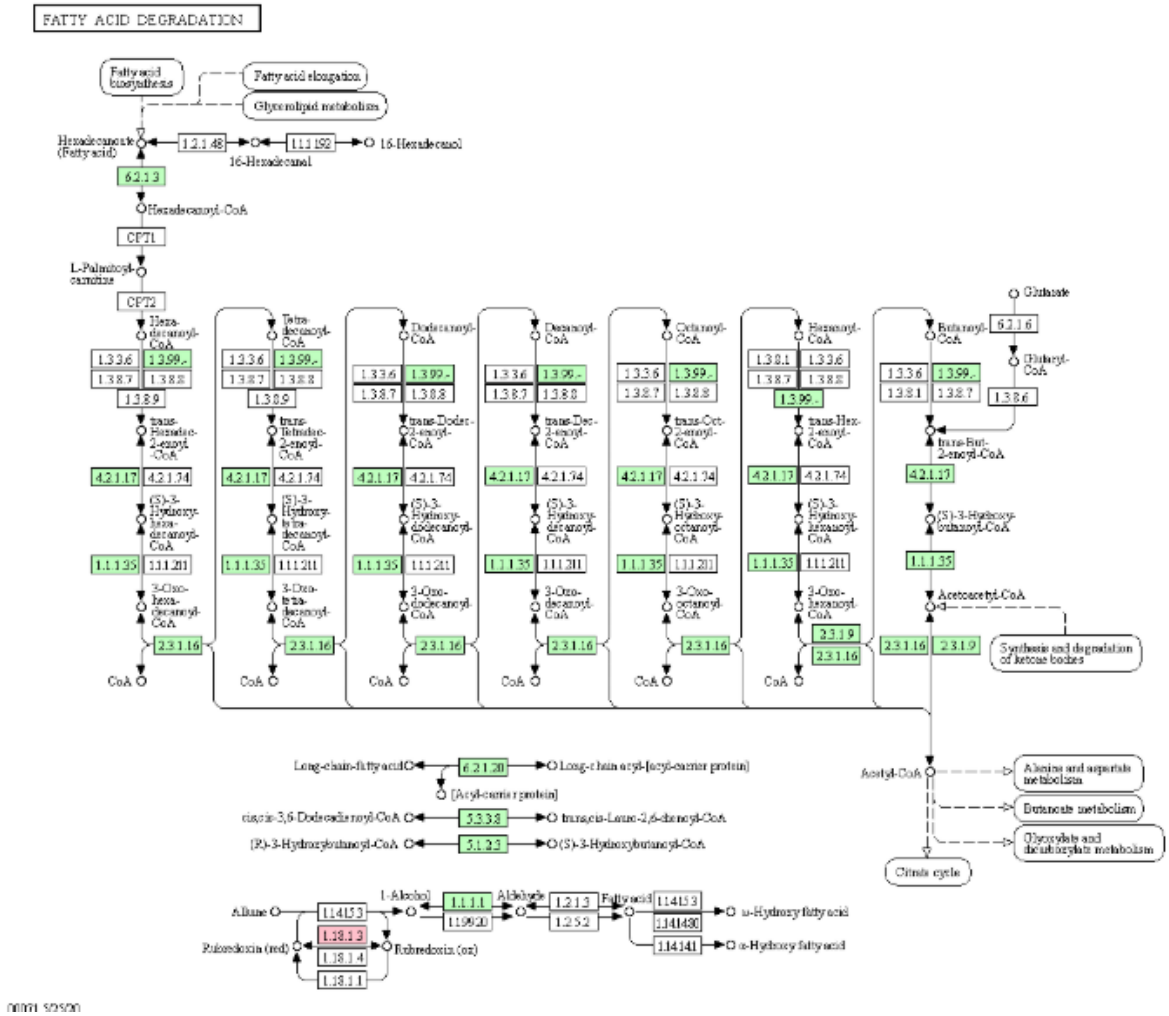

Figura 3. Via de degradação de ácidos graxos (map 00071) de *E.coli*.

Foi identificada a ausência de uma forma análoga da enzima ribose-5-fosfato isomerase (EC 5.3.1.6), exclusivamente nos perfis de patogenicidade EHEC e STEC. Esta enzima faz parte do ciclo da pentose fosfato (Figura 4), presente na via não oxidativa, cujo resultado é a produção de intermediários utilizados na via glicolítica (57).

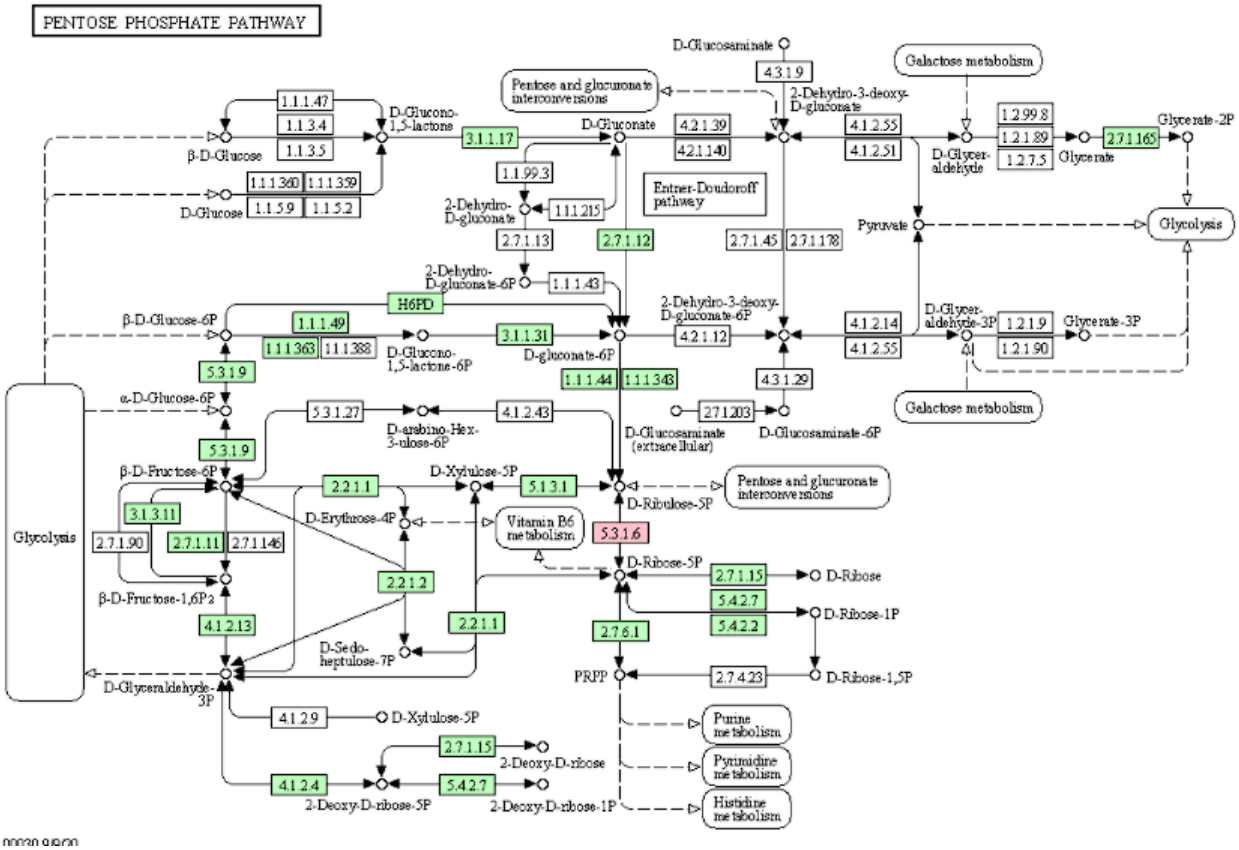

Figura 4. Via pentose fosfato (map 00030) de *E.coli*.

Na via de degradação das purinas (Figura 5), identificamos a ausência da enzima alantoinase (EC 3.5.2.5), exclusivamente nos perfis de patogenicidade EHEC e ETEC, codificada pelo gene allB, que ao estar presente juntamente com o gene HyuA, sugere fortemente que as ureidas cíclicas podem ser utilizadas como fonte de nutrientes em *E.coli* (58).

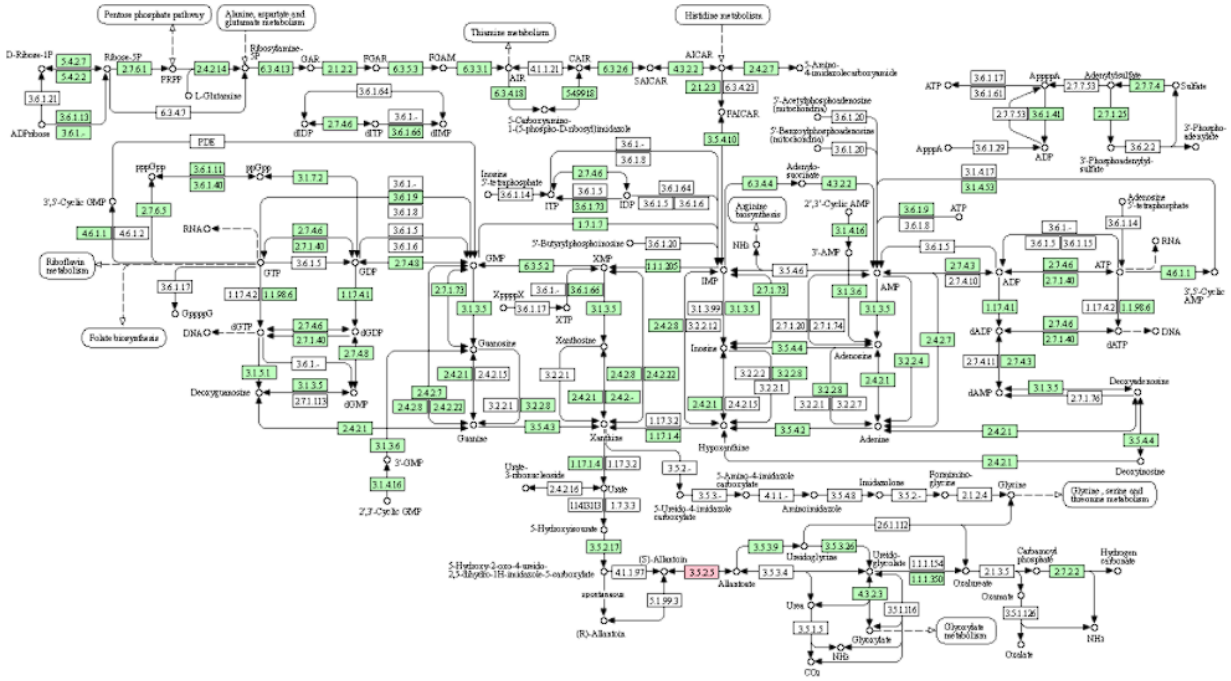

Figura 5. Via do metabolismo das purinas (map 00230) de *E. coli*.

No metabolismo da glicina, serina e treonina (Figura 6), identificamos a presença da enzima amina oxidase (EC 1.4.3.21) exclusivamente nas cepas com perfil de patogenicidade EIEC. Esta enzima catalisa a desaminação oxidativa de aminas aromáticas em aldeídos por meio de um mecanismo bem estabelecido, sendo codificada pelo gene tynA. Acredita-se ser usado em condições ambientais rigorosas em que ECAO pode, devido à sua produção de peróxido de hidrogênio, fornecendo uma vantagem de crescimento sobre outras bactérias que são incapazes de controlar altos níveis deste oxidante (59).

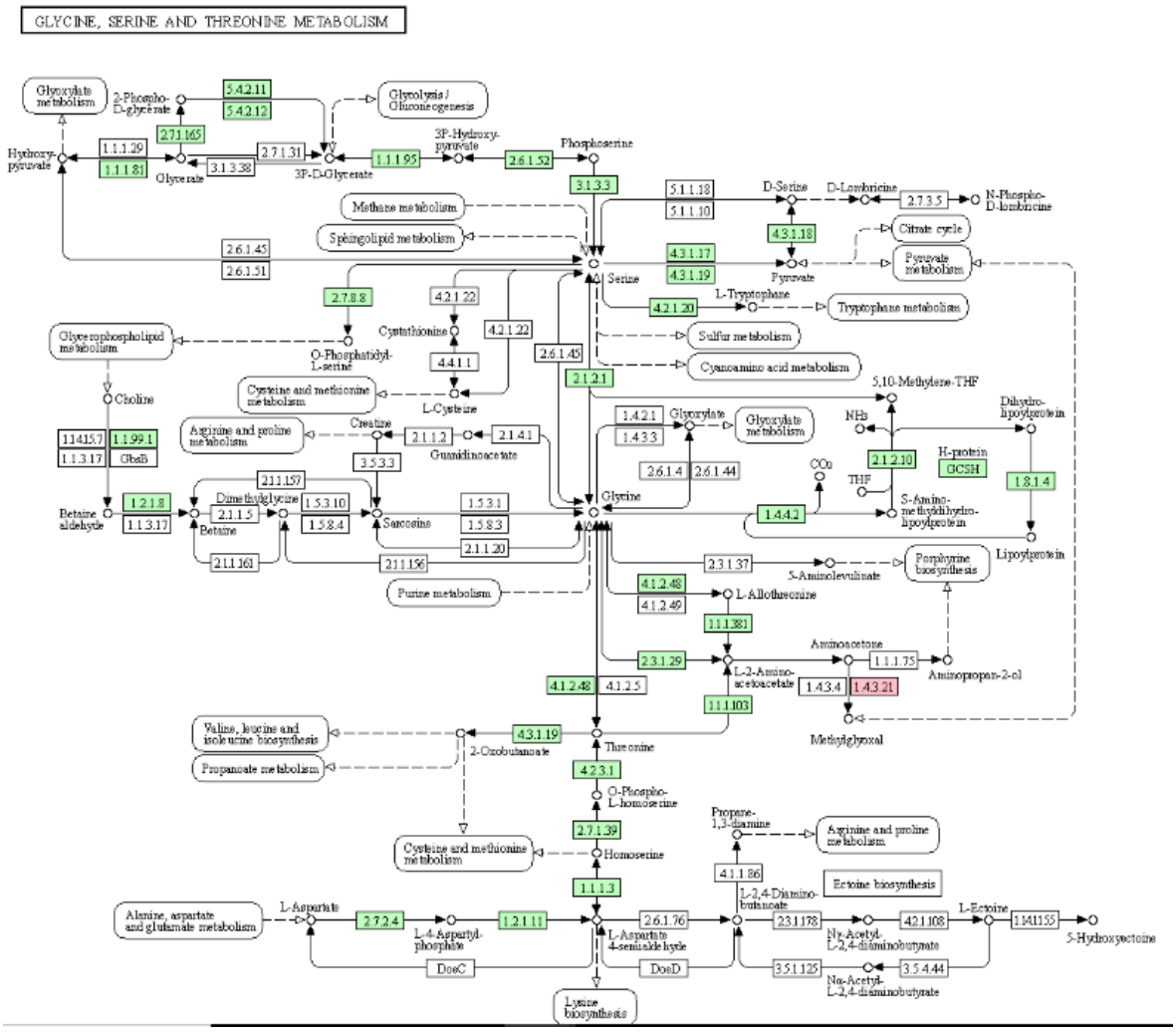

Figura 6. Via de metabolismo da glicina, serina e treonina (map 00260) de *E.coli*.

Na via de metabolismo de amino açúcar e nucleotídeos (Figura 7), identificamos a presença da enzima chitinase (EC 3.2.1.14), somente nos perfis de patogenicidade ETEC e ExPEC. Catalisando a transformação da chitina em chitobiose e N-acetyl-glucosamine, essa enzima desempenha funções em uma variedade de processos, como parasitismo, nutrição, mecanismo de defesa e morfogênese (60).

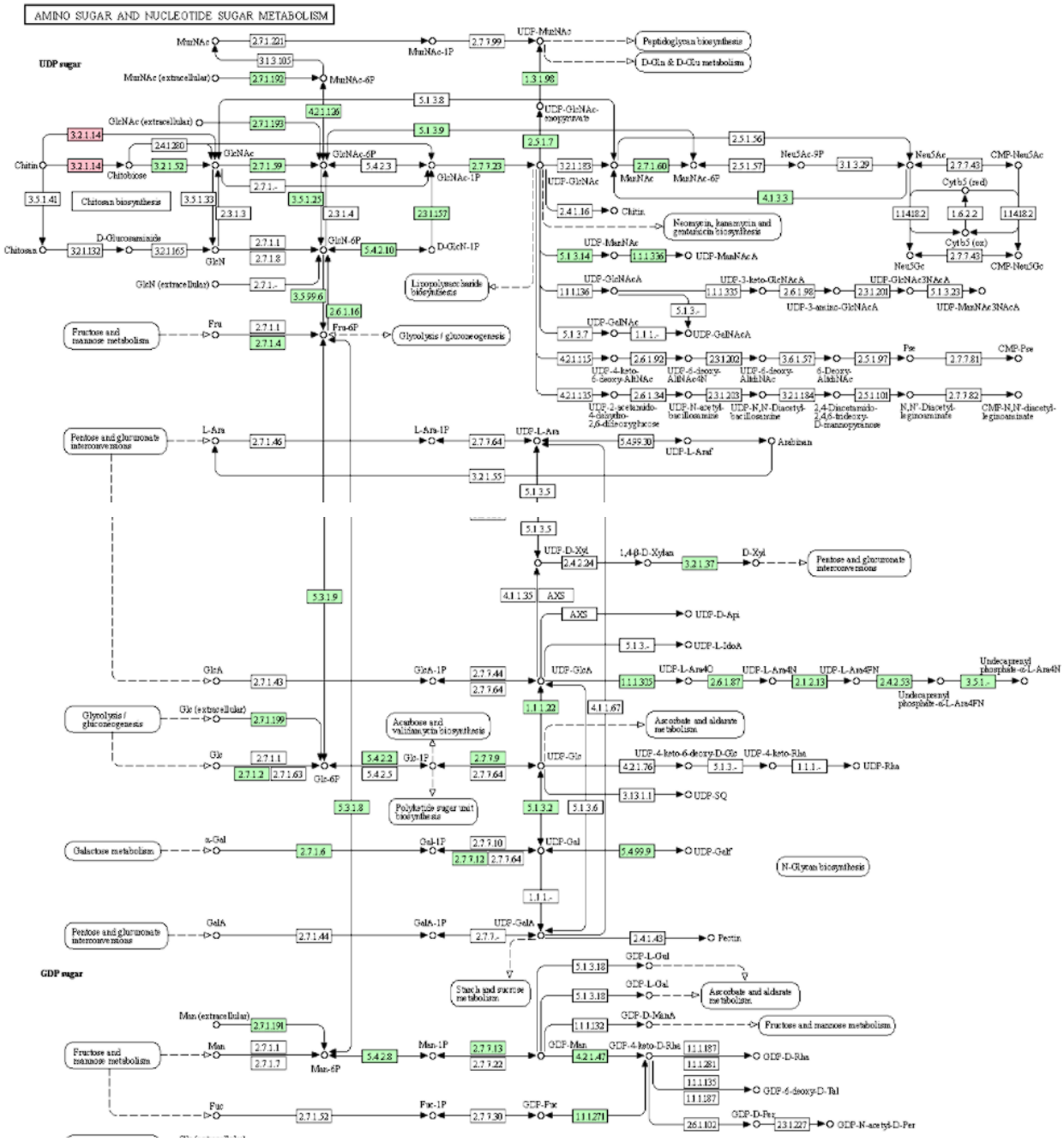

Figura 7. Via de metabolismo de amino açúcar e nucleotídeos (map 00520) de *E.coli*.

Na via de metabolismo da fenilalanina (Figura 8), mais especificamente na degradação de trans-cinamato, foi identificada a ausência de duas enzimas, a 3-phenylpropanoate dioxygenase (EC 1.14.12.19) nos perfis de patogenicidade extraintestinais ExPEC e UPEC e a 3-(3-hydroxy-

phenyl) propionate hydroxylase (EC 1.14.13.127). Estudos realizados por outros grupos já mostraram a relação da 3-phenylpropanoate dioxygenase com a resposta ao stress oxidativo (61).

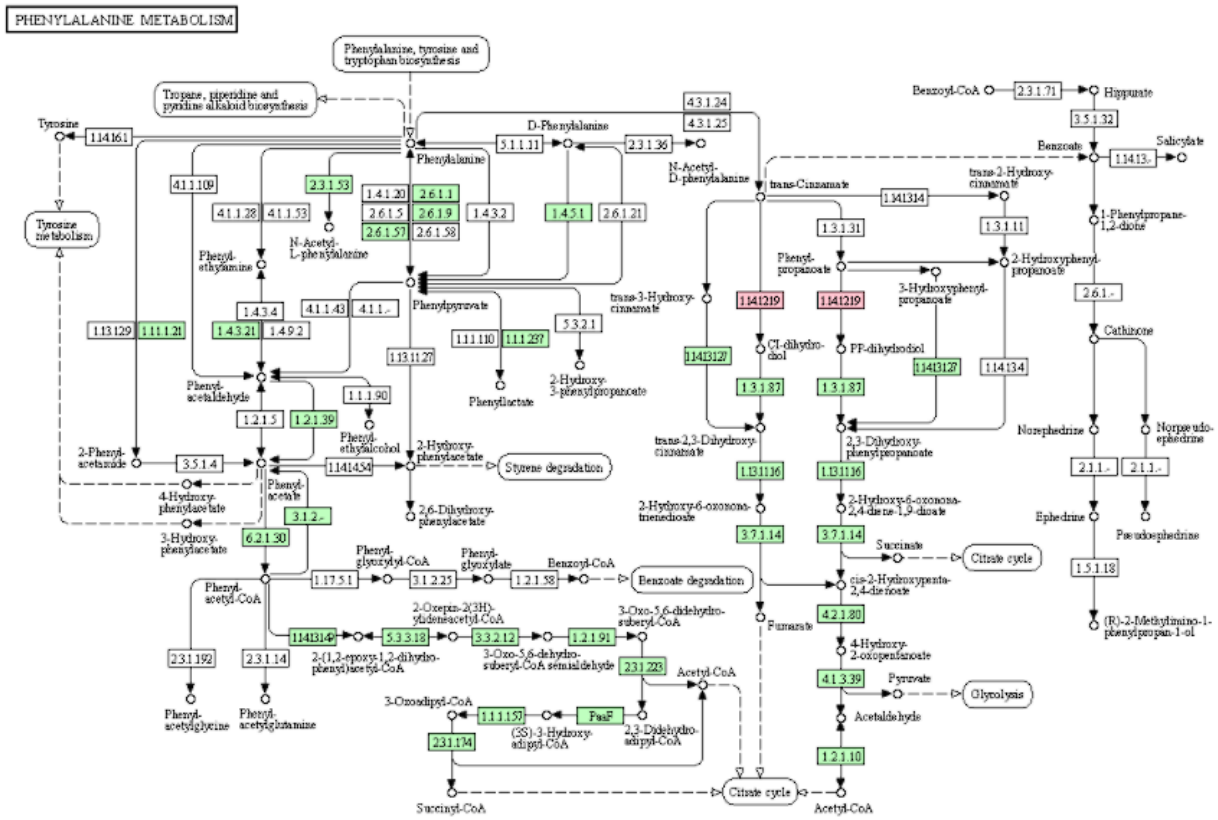

Figura 8. Via de metabolismo da fenilalanina (map 00360) de *E. coli.*

5 CONCLUSÃO

Foram analisados 52 genomas de *E. coli* completamente sequenciados, nos quais investigamos o repertório de enzimas análogas presentes em cada um deles e se essas enzimas contribuíam para a diversidade da gama de cepas potencialmente patogênicas deste organismo e se podiam estar associadas aos fenótipos de patogenicidade conhecidos nesta espécie.

Mais de 1.000 proteínas com atividade enzimática foram identificadas em cada genoma, que após uma etapa de validação, contavam aproximadamente 500 casos de analogia por genoma. Com as etapas de validação subsequentes, quanto ao tipo e origem evolutiva no enovelamento proteico e composição de dóminos funcionais, foram avaliadas os tipos de variações encontradas na comparação destes genomas e definimos quatro possíveis variações relevantes: (i) correlação da presença da atividade enzimática em determinado perfil de patogenicidade; (ii) anticorrelação da ausência da atividade enzimática em determinado perfil de patogenicidade; (iii) correlação do número de formas enzimáticas em determinado perfil de patogenicidade; (iv) anticorrelação do número de formas enzimáticas em determinado perfil de patogenicidade. Com isso, destacamos 71 atividades enzimáticas onde pudemos detectar alguma relevância para uma ou mais variações destacadas acima.

Dessas 71 atividades enzimáticas, 45 delas estão presentes no genoma acessório, a análise dessa sessão do pangenoma é interessante por estar envolvida em um dos principais meios de diversificação da espécie *E. coli*, a transferência horizontal de genes. Portanto, essas enzimas foram continuadas em nossas análises. O mapeamento dessas enzimas culminou em vias de grande relevância para o sistema como um todo, foram encontradas vias fundamentais, como a via de degradação de ácidos graxos, via pentose fosfato, via de degradação das purinas, dentre outras.

A presença da enzima ferroxina NAD+ reductase (EC 1.18.1.3) nos perfis de patogenicidade de *E.coli* responsáveis por doenças em ambiente extraintestinal sugere uma forte ligação entre esta enzima e o perfil de patogenicidade associado, atua na via de degradação de ácidos graxos e pode ser um potencial alvo terapêutico, visto que, ocupa posição central no metaboloma microbiano (56). A presença da enzima amina ocidase (EC 1.4.3.21) pode conferir

vantagens de crescimento de *E. coli* EIEC sobre outras bactérias, por possuir um mecanismo de desaminação oxidativa, tornando-a capaz de controlar altos níveis deste antioxidante (59).

Alguns dos casos identificados também geram um interessante ponto para estudos posteriores, como por exemplo, a ausência da enzima 3-phenylpropanoate dioxygenase (EC 1.14.12.19) nos perfis extraintestinas de *E. coli*, porém a reação catalisada pela enzima também pode ser realizada de forma muito similar, inclusive utilizando mesmo substrato, pela 3-(3-hydroxy-phenyl) propionate hydroxylase (EC 1.14.13.127), ambas relacionadas a resposta ao stress oxidativo (61). Pelo fato de as duas enzimas catalisarem reações muito similares, qual etapa evolucionária pode estar relacionada as duas enzimas, e quais suas semelhanças, são possíveis pontos a serem abordados futuramente.

As enzimas isofuncionais não-homólogas identificas em *E. coli*, se revelaram como potenciais alvos para estudos posteriores, seja para potenciais alvos terapêuticos, a importância dessas enzimas nas vias metabólicas, dentre outros. Incluindo a relação mais aprofundada de como essas enzimas podem interferir em cada perfil de patogenicidade em que estão associadas e sobre quais condições.

6 REFERÊNCIAS BIBLIOGRÁFICAS

1. Jang J, Hur HG, Sadowsky MJ, Byappanahalli MN, Yan T, Ishii S. Environmental Escherichia coli: ecology and public health implications-a review. *J Appl Microbiol*. 2017;123(3):570-581. doi:10.1111/jam.13468.

2. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in understanding enteric pathogenic Escherichia coli. *Clin Microbiol Rev*. 2013;26(4):822-880. doi:10.1128/CMR.00022-13.

3. Chowdhury F, Kuchta A, Khan AI, Faruque AS, Calderwood SB, Ryan ET, Qadri F. The increased severity in patients presenting to hospital with diarrhea in Dhaka, Bangladesh since the emergence of the hybrid strain of Vibrio cholerae O1 is not unique to cholera patients. Int J Infect Dis. 2015 Nov;40:9- 14. doi: 10.1016/j.ijid.2015.09.007. Epub 2015 Sep 25.

4. Edberg SC, Rice EW, Karlin RJ, Allen MJ. Escherichia coli: the best biological drinking water indicator for public health protection. *Symp Ser Soc Appl Microbiol*. 2000;(29):106S-116S. doi:10.1111/j.1365-2672.2000.tb05338.x.

5. Hamilton MJ, Hadi AZ, Griffith JF, Ishii S, Sadowsky MJ. Large scale analysis of virulence genes in Escherichia coli strains isolated from Avalon Bay, CA. *Water Res*. 2010;44(18):5463-5473. doi:10.1016/j.watres.2010.06.058.

6. Ishii K, Hamamoto H, Sekimizu K. Establishment of a bacterial infection model using the European honeybee, Apis mellifera L. *PLoS One*. 2014;9(2):e89917. Published 2014 Feb 24. doi:10.1371/journal.pone.0089917.

7. Kim NH, Cho TJ, Rhee MS. Current Interventions for Controlling Pathogenic Escherichia coli. *Adv Appl Microbiol*. 2017;100:1-47. doi:10.1016/bs.aambs.2017.02.001.

8. Pallen MJ, Wren BW. Bacterial pathogenomics. *Nature*. 2007;449(7164):835-842. doi:10.1038/nature06248.

9. Jain R, Rivera MC, Moore JE, Lake JA. Horizontal gene transfer in microbial genome evolution. *Theor Popul Biol*. 2002;61(4):489-495. doi:10.1006/tpbi.2002.1596.

10. Escobar-Páramo P, Sabbagh A, Darlu P, et al. Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the Escherichia coli case study. *Mol Phylogenet Evol*. 2004;30(1):243-250. doi:10.1016/s1055-7903(03)00181-7.

11. Ren CP, Beatson SA, Parkhill J, Pallen MJ. The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from Escherichia coli. *J Bacteriol*. 2005;187(4):1430-1440. doi:10.1128/JB.187.4.1430-1440.2005.

12. Fleckenstein JM, Rasko DA. Overcoming Enterotoxigenic Escherichia coli Pathogen Diversity: Translational Molecular Approaches to Inform Vaccine Design. *Methods Mol Biol*. 2016;1403:363-383. doi:10.1007/978-1-4939-3387-7_19.

13. Farfán-García AE, Ariza-Rojas SC, Vargas-Cárdenas FA, Vargas-Remolina LV. Mecanismos de virulencia de Escherichia coli enteropatógena [Virulence mechanisms of enteropathogenic Escherichia coli]. *Rev Chilena Infectol*. 2016;33(4):438-450. doi:10.4067/S0716-10182016000400009.

14. Hebbelstrup Jensen B, Olsen KE, Struve C, Krogfelt KA, Petersen AM. Epidemiology and clinical manifestations of enteroaggregative Escherichia coli. *Clin Microbiol Rev*. 2014;27(3):614-630. doi:10.1128/CMR.00112-13.

15. Mohammadzadeh M, Goudarzi H, Dabiri H, Fallah F. Molecular detection of lactose fermenting enteroinvasive Escherichia coli from patients with diarrhea in Tehran-Iran. *Iran J Microbiol*. 2015;7(4):198-202.

16. Frank E, Bonke R, Drees N, Heurich M, Märtlbauer E, Gareis M. Shiga toxin-producing Escherichia coli (STEC) shedding in a wild roe deer population. *Vet Microbiol*. 2019;239:108479. doi:10.1016/j.vetmic.2019.108479.

17. Nataro JP, Kaper JB. Diarrheagenic Escherichia coli [published correction appears in Clin Microbiol Rev 1998 Apr;11(2):403]. *Clin Microbiol Rev*. 1998;11(1):142-201.

18. Yamamoto S, Tsukamoto T, Terai A, Kurazono H, Takeda Y, Yoshida O. Genetic evidence supporting the fecal-perineal-urethral hypothesis in cystitis caused by Escherichia coli. *J Urol*. 1997;157(3):1127-1129.

19. Liu YY, Wang Y, Walsh TR, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect Dis*. 2016;16(2):161-168. doi:10.1016/S1473-3099(15)00424-7.

20. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. Global Extraintestinal Pathogenic Escherichia coli (ExPEC) Lineages. *Clin Microbiol Rev*. 2019;32(3):e00135-18. Published 2019 Jun 12. doi:10.1128/CMR.00135-18.

21. Desroches M, Royer G, Roche D, et al. The Odyssey of the Ancestral Escherich Strain through Culture Collections: an Example of Allopatric Diversification. *mSphere*. 2018;3(1):e00553-17. Published 2018 Jan 31. doi:10.1128/mSphere.00553-17.

22. Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. Parallel evolution of virulence in pathogenic Escherichia coli. *Nature*. 2000;406(6791):64-67. doi:10.1038/35017546.

23. Dobrindt U. (Patho-)Genomics of Escherichia coli. *Int J Med Microbiol*. 2005;295(6-7):357-371. doi:10.1016/j.ijmm.2005.07.009.

24. Dobrindt U, Hentschel U, Kaper JB, Hacker J. Genome plasticity in pathogenic and nonpathogenic enterobacteria. *Curr Top Microbiol Immunol*. 2002;264(1):157-175.

25. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. *Trends Genet*. 2009;25(3):107-110. doi:10.1016/j.tig.2008.12.004.

26. A. Payen and J.-F. Persoz (1833) "Mémoire sur la diastase, les principaux produits de ses réactions et leurs applications aux arts industriels" (Memoir on diastase, the principal products of its reactions, and their applications to the industrial arts), *Annales de chimie et de physique*, 2nd series, vol. 53, pages 73–92

27. Keller MA, Piedrafita G, Ralser M. The widespread role of non-enzymatic reactions in cellular metabolism. *Curr Opin Biotechnol*. 2015;34:153-161. doi:10.1016/j.copbio.2014.12.020.

28. Espadaler J, Eswar N, Querol E, et al. Prediction of enzyme function by combining sequence similarity and protein interactions. *BMC Bioinformatics*. 2008;9:249. Published 2008 May 27. doi:10.1186/1471-2105-9-249.

29. Ann Benore M. What is in a name? (or a number?): The updated enzyme classifications. *Biochem Mol Biol Educ*. 2019;47(4):481-483. doi:10.1002/bmb.21251.

30. Martínez Cuesta S, Rahman SA, Furnham N, Thornton JM. The Classification and Evolution of Enzyme Function. *Biophys J*. 2015;109(6):1082-1086. doi:10.1016/j.bpj.2015.04.020.

31. Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970 Jun;19(2):99-113. PMID:5449325.

32. Storz JF. Causes of molecular convergence and parallelism in protein evolution. *Nat Rev Genet*. 2016;17(4):239-250. doi:10.1038/nrg.2016.11.

33. Piergiorge RM, de Miranda AB, Guimarães AC, Catanho M. Functional Analogy in Human Metabolism: Enzymes with Different Biological Roles or Functional Redundancy?. *Genome Biol Evol*. 2017;9(6):1624-1636. doi:10.1093/gbe/evx119.

34. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Res.* 2016 Jan 4;44(D1):D733-45 .

35. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, Gerdes S, Henry CS, Kenyon RW, Machi D, Mao C, Nordberg EK, Olsen GJ, Murphy-Olson DE, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Vonstein V, Warren A, Xia F, Yoo H, Stevens RL. **Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center**. *Nucleic Acids Res*. 2017 Jan 4;45(D1):D535-D542. doi: 10.1093/nar/gkw1017. PMID: 27899627. PMCID: PMC5210524.

36. Dalquen DA, Altenhoff AM, Gonnet GH, Dessimoz C. The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PLoS One. 2013;8(2):e56925. doi:10.1371/journal.pone.0056925.

37. Contreras-Moreira,B and Vinuesa,P (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl.Environ.Microbiol. 79:7696-7701.

38. Vinuesa P and Contreras-Moreira B (2015) Robust Identification of Orthologues and Paralogues for Microbial Pan-Genomics Using GET_HOMOLOGUES: A Case Study of pIncA/C Plasmids. In Bacterial Pangenomics, Methods in Molecular Biology Volume 1231, 203-232, edited by A Mengoni, M Galardini and M Fondi.

39. Lavezzo, E., Falda, M., Fontana, P., Bianco, L., & Toppo, S. (2016). *Enhancing protein function prediction with taxonomic constraints – The Argot2.5 web server. Methods, 93, 15–23.* doi:10.1016/j.ymeth.2015.08.021.

40. Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. (1994) "Issues in searching molecular sequence databases." Nature Genet. 6:119-129. PubMed.

41. Profile Hidden Markov Models. S. R. Eddy. *Bioinformatics*, 14:755-763, 1998.

42. Anne Morgat, Thierry Lombardot, Elisabeth Coudert, Kristian Axelsen, Teresa Batista Neto, Sebastien Gehant, Parit Bansal, Jerven Bolleman, Elisabeth Gasteiger, Edouard de Castro, Delphine Baratin, Monica Pozzato, Ioannis Xenarios, Sylvain Poux, Nicole Redaschi, Alan Bridge, The UniProt Consortium, Enzyme annotation in UniProtKB using Rhea, *Bioinformatics*, Volume 36, Issue 6, 15 March 2020, Pages 1896–1901, https://doi.org/10.1093/bioinformatics/btz817.

43. El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. *Nucleic Acids Res*. 2019;47(D1):D427-D432. doi:10.1093/nar/gky995.

44. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;47(D1):D330-D338. doi:10.1093/nar/gky1055.

45. Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.

46. Guimarães ACR, Otto TD, Alves-Ferreira M, Miranda AB, Degrave WM. In silico reconstruction of the amino acid metabolic pathways of Trypanosoma cruzi. Genet Mol Res [Internet]. 2008 Sep 23 [cited 2017 Nov 6];7(3):872–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18949706.

47. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi:10.1093/nar/28.1.27.

48. anehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M.; New approach for understanding genome variations in KEGG. Nucleic Acids Res. 47, D590-D595 (2019).

49. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res [Internet]. 2016 Jan 4 [cited 2017 Nov 6];44(D1):D279–85. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1344.

50. Alex L Mitchell, Teresa K Attwood, Patricia C Babbitt, Matthias Blum, Peer Bork, Alan Bridge, Shoshana D Brown, Hsin-Yu Chang, Sara El-Gebali, Matthew I Fraser, Julian

Gough, David R Haft, Hongzhan Huang, Ivica Letunic, Rodrigo Lopez, Aurélien Luciani, Fabio Madeira, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Gift Nuka, Christine Orengo, Arun P Pandurangan, Typhaine Paysan-Lafosse, Sebastien Pesseat, Simon C Potter, Matloob A Qureshi, Neil D Rawlings, Nicole Redaschi, Lorna J Richardson, Catherine Rivoire, Gustavo A Salazar, Amaia Sangrador-Vegas, Christian J A Sigrist, Ian Sillitoe, Granger G Sutton, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Siew-Yit Yong and Robert D Finn **InterPro in 2019: improving coverage, classification and access to protein sequence annotations**. *Nucleic Acids Research, Jan 2019, (doi: 10.1093/nar/gky1100).*

51. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res [Internet]. 2016 Jan 4 [cited 2017 Nov 6];44(D1):D457–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26476454.

52. Otto TD, Guimaraes ACR, Degrave WM, de Miranda AB. AnEnPi: Identification and annotation of analogous enzymes. BMC Bioinformatics [Internet]. 2008; 9(1):544.

53. Jimenez-Diaz L., Caballero A., Segura A. (2017) Pathways for the Degradation of Fatty Acids in Bacteria. In: Rojo F. (eds) Aerobic Utilization of Hydrocarbons, Oils and Lipids. Handbook of Hydrocarbon and Lipid Microbiology.

54. Kaas RS, Friis C, Ussery DW, Aarestrup FM. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. *BMC Genomics*. 2012;13:577. Published 2012 Oct 31. doi:10.1186/1471-2164-13-577.

55. Tenaillon O, Skurnik D, Picard B, Denamur E: The population genetics of commensal Escherichia coli. Nat Rev Microbiol 2010, 8:207–17.

56. Salerno C, D'Eufemia P, Finocchiaro R, et al. Effect of D-ribose on purine synthesis and neurological symptoms in a patient with adenylosuccinase deficiency. *Biochim Biophys Acta*. 1999;1453(1):135-140. doi:10.1016/s0925-4439(98)00093-3.

57. Pathan N, Faust SN, Levin M. Pathophysiology of meningococcal meningitis and septicaemia. Archives of Diseases of Childhood, 2003 vol. 88:p 601-607; (Oelschlaeger TA, Dobrindt U, Hacker J. Virulence factors of uropathogens. Current Opinion Urology 12:33-38, 2002.

58. Elovaara H, Huusko T, Maksimow M, et al. Primary Amine Oxidase of Escherichia coli Is a Metabolic Enzyme that Can Use a Human Leukocyte Molecule as a Substrate. *PLoS One*. 2015;10(11):e0142367. Published 2015 Nov 10. doi:10.1371/journal.pone.0142367.

59. Ubhayasekera W: Structure and function of chitinases from glycoside hydrolase family 19. Polym Int 2011, 60:890–896.

60. DOMINIQUE MENGIN LECREULX* Arw JEAN vA HEIJENOORT. Copurification of Glucosamine-1-Phosphate Acetyltransferase and N-Acetylglucosamine-1-Phosphate Uridyltransferase Activities of Escherichia coli: Characterization of the glmU Gene Product as a Bifunctional Enzyme Catalyzing Two Subsequent Steps in the Pathway for UDP-N-Acetylglucosamine Synthesis. JOURNAL OF BACTERIOLOGY, Sept. 1994, p. 5788-5795.

61. LOMBARD, V. et al. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research, v. 42, n. 1, p. D490-D495, 2014.