UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Dissertação de Mestrado

Estudo da diversidade genética da *Escherichia coli* BH100 via genômica estrutural e comparativa

DISCENTE: Gustavo Santos de Oliveira
ORIENTADOR: Prof. Dr. Vasco Ariston de Carvalho Azevedo
CO-ORIENTADORES: Prof. Dr. Edmar Chartone de Souza
Profa. Dra. Andrea Maria Amaral Nascimento

BELO HORIZONTE
Abril – 2018

Gustavo Santos de Oliveira

Estudo da diversidade genética da *Escherichia coli* BH100 via genômica estrutural e comparativa

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo programa Interunidades de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo
Co-orientadores: Prof. Dr. Edmar Chartone de Souza
Profa. Dra. Andrea Maria Amaral Nascimento

BELO HORIZONTE
Abril – 2018

043 Oliveira, Gustavo Santos de.
Estudo da diversidade genética da Escherichia coli BH100 via genômica estrutural e comparativa [manuscrito] / Gustavo Santos de Oliveira. – 2018.

204 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo. Co-orientadores: Prof. Dr. Edmar Chartone de Souza, Profa. Dra. Andrea Maria Amaral Nascimento.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Bioinformática. 2. Genoma Microbiano. 3. Escherichia coli Uropatogênica. 4. Fatores R. I. Azevedo, Vasco Ariston de Carvalho. II. Souza, Edmar Chartone de. III. Nascimento, Andrea Maria Amaral. IV. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. V. Título.

CDU: 573:004

Ficha catalográfica elaborada pela Biblioteca do Instituto de Ciências Biológicas da UFMG

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG
Avenida Presidente Antônio Carlos, 6627 – Pampulha
31270-901 - Belo Horizonte – MG
Endereço eletrônico: bioinfo@icb.ufmg.br 55 31 3409-2554

"Estudo da diversidade genética da Escherichia coli BH100 via genômica estrutural e comparativa"

Gustavo Santos de Oliveira

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Prof. Edmar Chartone Souza - Coorientador
UFMG

Profa Andrea Maria Amaral Nascimento - Coorientadora
UFMG

Profa Glória Regina Franco
UFMG

Prof. Francisco Pereira Lobo
UFMG

Prof. Anderson Miyoshi
UFMG

Belo Horizonte, 09 de maio de 2018.

Em memória daquele quem me ensinou o valor do trabalho e a sempre fazer o melhor sem esperar reconhecimento.

Sobre meu modo de ser, pensar e agir....

"*Dizem que sou louco por pensar assim*
Se eu sou muito louco por eu ser feliz
Mas louco é quem me diz
E não é feliz, não é feliz"
(Arnaldo Baptista e Rita Lee, Balada do Louco)

Para mim, daqui a 20 anos....

"*For what is a man, what has he got*
If not himself, then he has naught
To say the things he truly feels
And not the words of one who kneels
The record shows I took the blows
…"
(Claude Francois, Gilles Thibaut, Jacques Revaux e Marcelo Drumond Nova, My Way)

Sobre este trabalho...

"*Não vamos colocar meta. Vamos deixar a meta aberta. Quando atingirmos a meta, vamos dobrar a meta*"
(Dilma Rousseff)

AGRADECIMENTOS

Agradeço à minha mãe, pelo apoio incondicional e por estar sempre estar comigo em todos os momentos. Obrigado por todo suporte, carinho e estímulo para que eu siga em frente principalmente nos momentos nebulosos.

O meu muito obrigado à Isadora Froes Torrent, quem teve toda paciência do mundo para tolerar minhas frustrações e quem sempre esteve ao meu lado, me apoiando e me incentivando, em todos os momentos. Agradeço também aos seus pais, dona Consuelo e seu Evandro, que abriram as portas de suas casas, onde eu pude morar nos primeiros meses dessa jornada de pós-graduando, e onde sempre fui tratado como um filho.

Meus agradecimentos ao Prof. Vasco Ariston de Carvalho Azevedo e a todo o seu grupo de pesquisa pela oportunidade.

Agradeço a todos os colegas que leram e de alguma forma contribuíram para revisão desse manuscrito, em particular à Mariana Parise e ao Felipe Pereira, pelos comentários pertinentes à parte de Bioinformática e à Bruna Silverio, pela extensiva e detalhada revisão gramatical.

Agradeço à Profa. Andrea Maria Amaral Nascimento, pelas riquíssimas discussões e prestatividade para comigo.

Agradeço ao Prof. Edmar Chartone de Souza, homem cuja história tenho grande respeito e admiração. Sinto-me muito honrado de ter trabalhado com a bactéria, objeto de estudo de toda uma vida acadêmica. Sua prestatividade, descontração e as diversas histórias compartilhadas comigo ficarão para sempre em minha memória.

Agradeço aos amigos da Bioinfo, em especial, Fenícia, Rodrigo, Anderson, Thaís, Núbia e Alessandra, que tive o prazer de fazer ao longo desses 2 anos. Pessoas incríveis, de histórias riquíssimas, cuja convivência me proporcionou e continua proporcionando momentos de grandes risadas e companheirismo.

Agradeço na pessoa do Prof. Rommel Ramos à toda equipe de Biologia Computacional da Universidade Federal do Pará, onde pude observar uma outra vertente de pesquisas com Bioinformática. Em especial, agradeço aos amigos Thiago Souza e Glen Jasper, pessoas de muita seriedade profissional, com quem compartilhei diversos momentos alegres em Belém e cujas conversas serviram de inspiração para diversas análises que realizei ao longo deste trabalho.

No tocante à minha formação intelectual, agradeço em especial aos amigos do LGCM: Marcus Viana, para mim o melhor bioinformata do laboratório, pessoa de riquíssimo conhecimento e cujas discussões de altíssimo nível, sem sombra de dúvidas, contribuíram para meu amadurecimento cientifico; Vinícius Rodovalho, amigo já de longa data, uma das mentes mais brilhantes que já tive o prazer de conhecer, muito grato pelos seus conselhos e nossas discussões que sempre vão dos fatos mais técnicos a grandes embates filosóficos; Raquel Hurtado: pessoa quem me faz pensar em espanhol para falar em português, muito obrigado pelo interesse e preocupação que sempre demostrou para comigo e meu trabalho, pelas sugestões de literatura e pelas grandes discussões cientificas que tivemos ao longo desses anos; Rafael Gantois, pela sinceridade, companheirismo e paciência para com minhas frustrações.

Agradeço aos professores, com quem tive o privilégio de conviver ao longo de suas disciplinas. Profissionais inequiparáveis, que aturaram minhas chatices em sala de aula e que contribuíram, e muito, para meu crescimento profissional. Para mim os professores excepcionais são aqueles que ensinam o que sequer percebem o que estão ensinando, e nesse contexto, agradeço em especial, à Prof. Gloria Regina Franco, cuja paixão por ciência e pela sala de aula, servirão sempre de inspiração para que eu atinja meus objetivos profissionais.

Agradeço também à minha família, amigos e todos aqueles que zelam e torcem pelo meu sucesso. Em especial ao meu padrasto Warley, minha irmã Gabriella, meus avós e meus grandes amigos Rafael Ferreira, Reginaldo Melo e Jéssica Silveiro.

Agradeço a todos aqueles profissionais de antes do início desse mestrado, cujo um pouco da personalidade e ensinamentos carrego comigo, em especial aos meus ex-orientadores, Prof. Carlos Ueira Vieira e Prof. Ana Maria Bonetti, quem muito bem iniciaram e alinharam minha carreira cientifica; ao Prof. Anderson Santos, quem me fez descobrir meu interesse e capacidade para programação, além de sempre ter incitado meu desenvolvimento; a todos os professores da Universidade Federal de Uberlândia, que sempre me incentivaram e a todos os professores de antes da graduação que ajudaram a modelar meu senso analítico e que sempre me estimularam a pensar alto.

Agradeço à secretaria do programa de Bioinformática por todo apoio e prontidão.

Agradeço à toda comunidade da Universidade Federal de Minas Gerais

Agradeço às agências de fomento, em especial à Fapemig

E agradeço a todos que de alguma forma contribuíram direta ou indiretamente para realização deste trabalho.

Sumário

Capítulo 1 INTRODUÇÃO 1
A *E. coli* BH100: do isolamento ao sequenciamento 1
1.1 A *Escherichia coli* 3
1.1.1 Histórico 3
1.1.2 Patótipos 4
1.1.3 Importância 6
1.2 A Multirresistência Bacteriana 9
1.3 A DIVERSIDADE GENÉTICA BACTERIANA 10
1.3.1 Mecanismos associados à geração de variabilidade 11
1.3.1.1 Mutação de ponto 12
1.3.1.2 Elementos móveis 13
1.3.1.3 Transferência horizontal 15
1.3.1.3.1 Conjugação e mobilização plasmidial 15
1.3.1.3.2 Transformação 17
1.3.1.3.3 Transdução 18
1.4 A Era Genômica 19
1.4.1 Sequenciamento de nova geração 20
1.4.2 Pré-Processamento 24
1.4.3 Montagem 25
1.4.4 *Scaffolding* 28
1.4.5 Anotação 29
1.4.6 Curadoria 30
1.5 Adventos da genômica comparativa 30
JUSTIFICATIVA 32
OBJETIVOS 33
Objetivos específicos 33
Capítulo 2 A OBTENÇÃO DO GENOMA COMPLETO DA *E. COLI* BH100 E SUAS VARIANTES 34
2.1 MATERIAL E MÉTODOS 34
2.1.1 Linhagens bacterianas 34
2.1.2 Extração do DNA genômico 35
2.1.3 Sequenciamento 36

2.1.4 Desenvolvimento e aquisição de *scripts* 36
2.1.5 Montagem do genoma 37
2.1.5.1 Pré-processamento 37
2.1.5.2 Montadores 38
2.1.5.3 *Scaffolding* 38
2.1.5.4 Refinamento 40
2.1.5.4.1 Remontagem de reads não mapeadas 40
2.1.5.4.2 Avaliação de regiões de baixa cobertura 40
2.1.5.4.3 Avaliação de regiões de mapeamento espúrio 41
2.1.5.4.4 Circularização 42
2.1.6 Anotação e curadoria 42
2.2 RESULTADOS E DISCUSSÃO 44
2.2.1 Obtenção dos replicons da *E. coli* BH100 44
2.2.2 Qualidade do sequenciamento 44
2.2.3 Pré-tratamentos 45
2.2.4 Montagem 46
2.2.4.1 A montagem do genoma da *E. coli* BH100 subtr. MG2014 46
2.2.4.1.1 Cromossomo 46
2.2.4.1.2 O plasmídeo pAp 47
2.2.4.1.3 O plasmídeo pBH100-1 47
2.2.4.2 A montagem do genoma da *E. coli* BH100 subtr. MG2017 50
2.2.4.2.1 Cromossomo 50
2.2.4.2.2 O plasmídeo pApR 50
2.2.4.2.3 O plasmídeo pBH100-1 53
2.2.4.3 A montagem do genoma *E. coli* BH100L subtr. MG2017 53
2.2.4.4 A montagem do genoma da *E. coli* BH100N subtr. MG2017 53
2.2.4.5 A montagem dos demais plasmídeos 54
2.2.4.6 A montagem dos cromossomos das demais linhagens das subcepas MG2014 54
2.2.5 Anotação e curadoria 54
2.2.5.1 Replicons cromossomicos 54
2.2.5.2 Replicons plasmidiais 58
Capítulo 3 GENÔMICA COMPARATIVA 60
3.1 MATERIAL E MÉTODOS 60

3.1.1 Desenvolvimento e aquisição de *scripts* 60
3.1.2 Obtenção dos genomas RefSeq 61
3.1.3 Sintenia 61
3.1.4 Identificação de elementos móveis 61
3.1.5 Similaridade cromossômica e plasmidial 62
3.1.6 Ilhas genômicas e de patogenicidade 62
3.1.7 Investigação funcional 62
3.1.8 Predição funcional de pseudogenes 63
3.1.9 Predições complementares 63
3.2 RESULTADOS E DISCUSSÃO 64
3.2.1 Cromossomos das linhagens BH100, BH100L e BH100N 64
3.2.1.1 Análise de sintenia 64
3.2.1.2 Análise de elementos móveis 65
3.2.1.3 Quadro comparativo de genes anotados no Patric 68
3.2.1.4 Análise de pseudogenes 78
3.2.2 A E. coli BH100 num contexto patogênico 83
3.2.2.1 Contextualização filogenética 84
3.2.2.2 Análise de ilhas genômicas 86
3.2.2.3 Genes de virulência 91
3.2.3 O contexto filogenético dos plasmídeos da *E. coli* BH100 98
3.2.3.1 Árvore do plasmídeo pAp 98
3.2.3.2 Árvore do plasmídeo pBH100 100
3.2.4 Os plasmídeos do grupo pBH100 102
3.2.4.1 Genes de resistência 103
3.2.4.1.1 Resistência ao mercúrio 103
3.2.4.1.2 Resistência à estreptomicina 104
3.2.4.1.3 Resistência ao cloranfenicol 104
3.2.4.1.4 Resistência à canamicina 105
3.2.4.1.5 Resistência à tetraciclina 105
3.2.4.1.6 Outros marcadores de resistência 106
3.2.4.2 Outros elementos de destaque 106
3.2.4.2.1 Operon tra/trb 106
3.2.4.2.2 Grupo PSH 107

3.2.4.2.3 Grupo Rep-Her 107
3.2.4.3 Elementos móveis 108
3.2.4.4 Diferenças entre os plasmídeos pBH100 110
3.2.4.5 Comparativo funcional 112
3.2.4.6 Identificação das origens de replicação (*oriV*) e origem de transferência (oriT) 113
3.2.4.7 Inferência do número de cópias 113
3.2.5 Os plasmídeos do grupo pAp 114
3.2.5.1 Genes de resistência e toxicidade 115
3.2.5.1.1 A resistência à β-lactâmicos 115
3.2.5.1.2 O operon de colicina 115
3.2.5.2 Outros genes de destaque 116
3.2.5.2.1 Genes associados à segregação 116
3.2.5.2.2 Gene de controle da replicação 117
3.2.5.2.3 Genes de mobilização 117
3.2.5.3 Diferenças entre os plasmídeos 117
3.2.5.4 Um possível mecanismo de mobilização 123
3.2.5.5 Identificação das origens de replicação (*oriV*) e origem de transferência (*oriT*) 124
3.2.5.6 Predição do número de cópias 124
Capítulo 4 MODELO DE SURGIMENTO DE MUTANTES DE ALTA FREQUÊNCIA RESISTENTES À ESTREPTOMICINA 127
CONCLUSÃO 134
PERSPECTIVAS 137
REFERÊNCIAS 138
ANEXOS 163
APÊNDICES 169

Lista de Figuras

Figura 1.1 Perfil de resistência completa, intermediária e susceptibilidade a diferentes antimicrobianos computados para algumas linhagens de *E. coli* 15

Figura 1.2 Mecanismo de aquisição de cassetes gênicos por integrons**Error! Bookmark not defined.**

Figura 1.3 Etapas gerais de um projeto genômico..... 20

Figura 1.4 Representação esquemática de um sequenciamento na plataforma Illumina..... 22

Figura 1.5 Representação esquemática de sequenciamento IonTorrent 23

Figura 1.6 Exemplo de algoritmo de montagem utilizando a estratégia de *Overlap Layout Consensus* (OLC)..... 26

Figura 1.7 Exemplo de algoritmo de montagem utilizando grafos de Brujin..... 27

Figura 1.8 Representação esquemática do problema das repetições 28

Figura 1.9 Gráficos de sintenia (A) entre duas linhagens de *E. coli* uropatogênicas e (B) entre dois biovares distintos de *C. pseudotuberculosis* (MEX31 – equi, 1002 – ovis)..... 29

Figura 2.1 Representação da estratégia de *contig walking* para fechamento de *gaps* 41

Figura 2.2 Exemplo de região de mapeamento espúrio..... 42

Figura 2.3 Representação dos plasmídeos pAp e pBH100-x obtidos no processo de montagem de *reads* oriundas da *E. coli* BH100 subtr. MG2014..... 47

Figura 2.4 Mapeamento de *reads fragment* (A) e *reads fragment* e *mate-pair* (B) na região do flanco esquerdo de pAP (destacado em roxo) no replicon pBH100-x..... 48

Figura 2.5 Mapeamento de *reads fragment* (A) e *reads fragment* e *mate-pair* (B) na região do flanco direito de pAp (destacado em roxo) no replicon pBH100-x após rearranjo da sequência. 48

Figura 2.6 Representação esquemática da construção do plasmídeo pBH100-1 49

Figura 2.7 Representação esquemática da construção do plasmídeo pApR 51

Figura 2.8 Mapeamento de *reads pair-end* Illumina no flanco esquerdo (A) e direito (B) de pAp (destacado em roxo) 52

Figura 2.9 Mapeamento de *reads pair-end* no plasmídeo pAp da *E. coli* BH100 subtr. MG2014 52

Figura 2.10 Exemplos de vieses de anotação..... 56

Figura 3.1 Representação dos blocos contíguos produzidos pelo Mauve, indicando o grau de sintenia entre os cromossomos da *E. coli* BH100 e suas variantes 65

Figura 3.2 Esquematização de duplicação e inversão de uma região presente no cromossomo da *E. coli* BH100N subtr. MG2017 65

Figura 3.3 Predição de prófagos ao longo dos cromossomos da *E. coli* BH100 subtr. MG2014 (1), *E. coli* BH100 subtr. MG2017 (2), *E. coli* BH100L subtr. MG2017 (3) e *E. coli* BH100N subtr. MG2017 (4) 66

Figura 3.4 Composição de ISs nos cromossomos da *E. coli* BH100 e suas variantes..... 67

Figura 3.5 Distribuição de Elementos de Inserção (IS) ao longo dos cromossomos da *E. coli* BH100 e suas variantes. 68

Figura 3.6 Relação de genes presentes/ausentes segundo anotação do PATRIC..... 69

Figura 3.7 Exemplo de incoerências no processo de anotação do PATRIC..... 70

Figura 3.8 Novo diagrama de Venn após identificação de CDSs falsamente preditas como ausentes nas diversas linhagens desse estudo..... 71

Figura 3.9 Termos GO (A) e GSO (B) associados a processos biológicos vinculados a CDSs ausentes no cromossomo da *E. coli* BH100 subtr. MG2014 e presentes em pelo menos outros dois replicons cromossomais. 74

Figura 3.10 Termos GO (A) e GSO (B) associados a processos biológicos vinculados a CDSs ausentes no cromossomo da *E. coli* BH100 subtr. MG2017 e presentes em pelo menos outros dois replicons cromossomais. 74

Figura 3.11 Termos GO (A) e GSO (B) associados a processos biológicos vinculados a CDSs ausentes no cromossomo da *E. coli* BH100L subtr. MG2017 e presentes em pelo menos outros dois replicons cromossomais 75

Figura 3.12 Termos G.O (A) e G.S.O (B) associados a processos biológicos vinculados a CDSs ausentes no cromossomo da *E. coli* BH100N subtr. MG2017 e presentes em pelo menos outros dois replicons cromossomais 76

Figura 3.13 Diagrama de Venn indicando o número de *features* anotadas como pseudogenes, em comum a cada uma das linhagens da *E. coli* BH100 e suas variantes. 79

Figura 3.14 Novo diagrama de Venn, após refinamento das anotações de pseudogenes e aplicação dos filtros para *Recoding* e duplicidade gênica..... 81

Figura 3.15 Região do gene codificador do transportador de C4-dicarboxilato 82

Figura 3.16 Contexto filogenético da *E. coli* BH100 85

Figura 3.17 Ilhas de patogenicidade (PI) e genômicas (GI) preditas pelo IslandPath-DIMOB e compiladas no IslandViewer4 da *E. coli* BH100 subtr. MG2017 em comparação com os demais cromossomos das variantes da *E. coli* BH100, além das linhagens uropatogênicas *E. coli* 536 e *E. coli* CFT073, e da comensal *E. coli* K-12 subtr. MG1655. 87

Figura 3.18 Ilhas de patogenicidade (PI), genômicas (GI) e de resistência (RI), identificadas pelo IslandPath-DIMOB e compilado no IslandViewer4, da *E. coli* 536, em comparação aos cromossomos da *E. coli* BH100 e suas variantes. 88

Figura 3.19 Ilhas de patogenicidade (PI) e genômicas (GI), identificadas pelo IslandPath-DIMOB e compilado no IslandViewer4, da *E. coli* CFT073, em comparação aos cromossomos da *E. coli* BH100 e suas variantes. 89

Figura 3.20 Diagrama evidenciando os diferentes grupos de plasmídeos identificados nesse trabalho..... 98

Figura 3.21 Árvore filogenética do plasmídeo pAp 99

Figura 3.22 Contexto filogenético dos plasmídeos do grupo pBH100..... 101

Figura 3.23 Representação esquemática do plasmídeo pBH100alpha. 102

Figura 3.24 Mapeamento de *reads fragment* no plasmídeo pBH100alpha_MG2014..... 104

Figura 3.25 Atividade dos genes do operon *tet* responsáveis pela resistência à tetraciclina. 105

Figura 3.26 Composição de Elementos de Inserção (ISs) nos diferentes replicons do grupo pBH100. 109

Figura 3.27 Blocos contínuos construídos pelo Mauve revelam a sintenia entre os plasmídeos do grupo pBH100..... 110

Figura 3.28 Posição dos diferentes ISs, ao longo dos replicons do grupo pBH100..... 111

Figura 3.29 Nuvem de palavras de termos GO associados a processos biológicos e vinculados ao grupo pBH100..... 112

Figura 3.30 Representação esquemática das CDSs anotadas nos plasmídeos do grupo pAp115

Figura 3.31 Representação de blocos contíguos, criados pelo Mauve evidenciando a sintenia dos diferentes plasmídeos do grupo pAp. 118

Figura 3.32 Mecanismo de surgimento do plasmídeo pApR. Nesse caso, plasmídeos do grupo pBH100 servem como fonte do Tn21 (em vermelho), capaz de se inserir no plasmídeo pAp. 118

Figura 3.33 Mapeamento de *reads fragment* no plasmídeo pAp_MG2014..... 119

Figura 3.34 Mapeamento de *reads fragment* no plasmídeo pAp_MG2014..... 120

Figura 3.35 Anotações diferenciais de genes *mob* 121

Figura 3.36 Alinhamento global das sequências protéicas, referentes ao gene *bla,* encontradas nos diferentes replicons plasmidiais da *E. coli* BH100..... 122

Figura 3.37 Diferentes mecanismos de mobilização não envolvendo cointegração e condução 123

Figura 4.1 Representação esquemática do Tn21 127

Figura 4.2 Alinhamento múltiplo da região a montante do gene *hns* das 4 linhagens da BH100 e outras 6 *E. coli* 129

Figura 4.3 Níveis de resistência à estreptomicina de 70 mutantes de alta frequência isolados em ágar nutriente contendo 25 $\mu g.Ml^{-1}$ de estreptomicina..... 132

Figura 4.4 Modelo para explicação de diferentes níveis de resistência à estreptomicina como um balanço na atividade das proteínas HHA e H-NS, e da dosagem dos genes *aadA1* 133

Lista de Tabelas

Tabela 2.1 Estatísticas de montagem dos genomas da Escherichia coli BH100 e suas variantes. 44

Tabela 2.2 Estatísticas de anotação referentes aos cromossomos de 4 linhagens completas da E. coli BH100 e suas variantes. 55

Tabela 2.3 Distribuição das CDSs anotadas como pseudogenes..... 55

Tabela 2.4 Frequência de tRNAs anotados automaticamente pelo RAST..... 57

Tabela 2.6 Estatísticas de anotação dos diferentes replicons plasmidiais presentes na E. coli BH100 e suas variantes, e na E. coli K12T. 58

Tabela 3.1 Distribuição das CDSs ausentes nos cromossomos de cada linhagem e presentes em pelo menos dois outros replicons, segundo as categorias descritas no texto..... 72

Tabela 3.2 Nova distribuição de CDSs anotadas como pseudogenes após remoção de features referentes a transposases e a elementos de fagos..... 78

Tabela 3.3 Inferência de oriV e oriT em pBH100 a partir de BLASTn contra uma base de dados contendo sequências referentes às origens de replicação e transferência. 113

Tabela 3.4 Estimativa do número de cópias de cada plasmídeo do grupo pBH100, a partir da proporção entre a profundidade de cobertura no plasmídeo em relação ao cromossomo 114

Tabela 3.5 Predição da origem de replicação (ori) nos plasmídeos do grupo pAp 124

Tabela 3.6 Predição do número de cópias dos plasmídeos do grupo pAp a partir da proporção da profundidade da cobertura média observada nos plasmídeos em relação ao cromossomo 125

Lista de Quadros

Quadro 1.1 Características gerais das estratégias de montagem via grafos de Brujin (DBJ) e *Overlap Layout Consensus* (OLC) 27

Quadro 2.1 Caracteristicas da *E. coli* BH100 utilizadas nesse trabalho e suas marcadores de resistência. 34

Quadro 3.1 Relação de genes pertencentes a uma região exclusiva da *E. coli* BH100 subtr. MG2017. 77

Quadro 3.2 Relação de potenciais pseudogenes exclusivos aos diferentes cromossomos em análise..... 83

Quadro 3.3 Associação das diferentes ilhas anotadas na *E. coli* 536, *E. coli* CFT073 e *E. coli* BH100 subtr. MG2017..... 90

Quadro 3.4 Relação de genes preditos pelo VirulenceFinder no cromossomo da *E. coli* BH100 subtr. MG2017, suas eventuais presenças em ilhas genômicas e o número de parálogos. 92

Quadro 3.5 Genes associados à virulência e *fitness* de UPECs revisados em Subashchandrabose e Mobley (2015) e identificados em diferentes linhagens de *E. coli* (Continua)..... 93

Quadro 3.6 Relação de genes preditos como não funcionais nos membros de pBH100..... 112

LISTA DE ABREVIATURAS

µg	Micrograma
µL	Microlitro
CDS	*Coding Sequence*
DBJ	de Brujin
DNA	*Deoxyribonucleic Acid*
IS	*Insersion Sequence*
In	Integron
ITU	Infecção do Trato Urinário
LGCM	Laboratório de Genética Celular e Molecular
LGM	Laboratório de Genética de Microrganismos
NCBI	*National Center for Biotechnology Information*
NGS	*Next Generation Sequencing*
NT	*Nucleotide Database*
mL	Mililitro
OLC	*Overlap Layout Consensus*
ori	Origem
PATRIC	*Pathosystems Resource Integration Center*
PCR	*Polymerase Chain Reaction*
qPCR	*Real Time PCR*
RAST	*Rapid Annotations Using Subsystems Technology*
SNP	*Single Nucleotide Polymorphism*
Tn	Transposon

RESUMO

A presença de variabilidade genética na natureza é a grande força motriz de processos evolutivos. Em procariotos, diversos mecanismos foram desenvolvidos de modo a incrementar essa variabilidade e, com isso, conquistar os mais diversos nichos ecológicos. Em um contexto clínico, maior atenção é dada à capacidade desses microrganismos em colonizar tecidos humanos e, causar doenças. Dentre as mais diversas infecções bacterianas existentes, destaca-se a infecção do trato urinário (ITU), tendo na *Escherichia coli* o seu principal agente etiológico. O estudo, cada vez mais detalhado, dos mecanismos associados à geração de variabilidade e desenvolvimento de patogênese, somente foi possível graças aos adventos da Biologia molecular e, mais recentemente, do campo da Bioinformática. Neste trabalho, objetivou-se aplicar ferramentas e desenvolver *scripts* de bioinformática, com o propósito de montar, anotar e caracterizar o genoma da *E. coli* BH100, isolada em 1973, em Belo Horizonte, da urina de uma paciente com suspeita de ITU. Essa bactéria apresenta resistência aos antibióticos ampicilina, tetraciclina, canamicina, cloranfenicol e, a baixas concentrações de estreptomicina, e ao íon mercúrio, codificadas por um plasmídeo mobilizável e outro autoconjugativo. A partir do uso das tecnologias de sequenciamento de nova geração foram montados os genomas de variantes da *E.* coli BH100, obtidos a partir da cura desses plasmídeos, de modo a observar potenciais efeitos na geração de variabilidade provocada por esses elementos. A montagem do genoma completo revelou a presença de um cromossomo de aproximadamente 5,2 Mb, do plasmídeo menor, mobilizável, contendo aproximadamente 15 kb e do maior plasmídeo, autoconjugativo, contendo aproximadamente 107 kb. A constatação de um número relativamente elevado de sequências de inserção (IS), da família dos IS3, dispostos em posições distintas nos cromossomos das bactérias analisadas, mostrou-se como um dos principais agentes de geração de variabilidade. As análises comparativas indicaram que a *E. coli* BH100 talvez seja uma autêntica uropatogênica (UPEC), causadora de pielonefrite. A anotação funcional revelou a presença de todos os marcadores de resistência, previamente observados, em transposons. Dentre esses elementos destaca-se a presença do Tn21, idêntico ao encontrado no plasmídeo NR1, e um potencial novo transposon, carreador do gene de resistência à canamicina, flanqueado por IS5. Por fim, propôs-se neste trabalho um mecanismo capaz de explicar o surgimento de variantes com resistência instável à estreptomicina, provocado por um número variável de cópias do gene *aadA1,* presente no Tn21.

ABSTRACT

Genetic variability can be seen as the driving force to evolution. In prokaryotes, many mechanisms emerged such as microorganisms could enhance variability, allowing their spread throughout different ecological niches. In a clinical context, major attention is given to the capacity of some microorganisms in colonize human tissues and developing diseases. From many different bacterial infections currently known, the urinary tract infection (UTI), which is caused mainly by *Escherichia coli*, can be highlighted as a major concern for public health. The comprehension of mechanisms associated with the emergency of variability and pathogenicity only was possible thanks to the advancements of Molecular Biology and more recently to Bioinformatics. In the present work, we aimed to use and develop bioinformatic tools in order to assembly, annotate and fully characterize the complete genome of *E. coli* BH100, which was isolated in 1973 in Belo Horizonte, Brazil, from urine of a patient suffering from UTI. This strain is resistant to ampicillin, tetracycline, kanamycin, chloramphenicol, inorganic mercury and in some cases to streptomycin. This multiresistance is due the presence of a mobilizable plasmid carrying a β-lactamase gene (*bla*) and a self-transmissible R plasmid carrying the genetic resistance marks of all other elements. Thanks to the use of Next Generation Sequencing (NGS) we were able to assemble the genome of this strain as well as other variants built from curation of the original plasmids, such as we could interrogate the effects of these elements to the emergence of variability. The complete genome of *E. coli* BH100 resulted in a chromosome of ~ 5.2 Mb, the smaller mobilizable plasmid of ~15 kb, and a self-transmissible plasmid of ~107 kb. This strain has shown a considerable number of insertion sequences from the family IS3 spread differently along the chromosome of BH100 and its variants. The comparative analysis indicate that this strain might be an authentic uropathogenic *E. coli* (UPEC) causing pyelonephritis. The functional annotation confirmed the presence of all resistance marks in transposons (Tn). Special attention was given to the presence of Tn21, identical to the one found in plasmid NR1, and to a potential new transposon carrying a gene for kanamycin resistance flanked by IS5 elements. In the end, we propose a new mechanism capable to explain the emergence of unstable and diversified resistance to streptomycin within the population of *E. coli* BH100. In this model, we suggest streptomycin resistance shows up due an increase in the copy number of the gene *aadA1* present in Tn21

ESTRUTURA DO TEXTO

Este trabalho foi dividido em 4 capítulos, de modo a tornar mais cômoda a compreensão dos diferentes achados deste trabalho. Ao final do último capítulo podem ser vistas as seções de Conclusão e Perspectivas, onde são destacados os principais achados e planos de atividades futuras que possam ser desenvolvidas com a *E. coli* BH100.

No primeiro capítulo será apresentado o histórico dos trabalhos conduzidos com a *E. coli* BH100, bem como os detalhes técnicos associados à geração deste trabalho. Em seguida, são apresentados o estado da arte de mecanismos associados à geração de variabilidade em procariotos e práticas de bioinformática aplicadas em projetos genômicos. Em seguida, destacam-se a justificativa para o desenvolvimento do presente trabalho e os objetivos almejados.

No capítulo dois, discorre-se acerca das problemáticas da montagem do genoma da *E. coli* BH100 e as estratégias utilizadas para montagem do genoma completo dessa bactéria.

No terceiro capítulo, o leitor encontrará os resultados e discussões pertinentes aos cromossomos sequenciados, montados e anotados, bem como aos plasmídeos presentes na *E. coli* BH100. Serão abordadas as diferentes estratégias comparativas, aplicadas com a finalidade de descrever cada replicon, bem como discussões envolvendo os processos de anotação e curadoria de genomas.

Já no capítulo quatro, apresenta-se um mecanismo capaz de explicar o surgimento de variantes resistentes, mas instáveis, à estreptomicina em diversas populações de *E. coli* BH100.

Por fim, o leitor poderá observar, na seção de Anexos, figuras complementares geradas por diferentes ferramentas de bioinformática e, na sessão Apêndices, tabelas suplementares que dão suporte aos resultados apresentados ao longo do texto.

CAPÍTULO 1 INTRODUÇÃO

A *E. coli* BH100: do isolamento ao sequenciamento

A *Escherichia coli* BH100, não colicinogênica e fermentadora de lactose (*lac*$^+$), foi isolada no município de Belo Horizonte (MG) em 1973 de amostras de urina de mulheres hospitalizadas, no ambulatório do IPSEMG (CHARTONE-SOUZA, 1975). Essa bactéria carrega consigo um plasmídeo de aproximadamente 60 MDa, que foi denominado pBH100, e outro de 10 MDa, denominado pAp. Investigações conduzidas pelo grupo do pesquisador Edmar Chartone-Souza (Informação verbal)[1] revelaram que o maior plasmídeo apresenta propriedades conjugativas, além de possuir marcadores de resistência ao mercúrio inorgânico, cloranfenicol, canamicina, tetraciclina e a baixos níveis (< 5,0 $\mu g.mL^{-1}$) de estreptomicina. O mesmo grupo, também, determinou que o plasmídeo menor codifica resistência à β-lactâmicos e é mobilizável pelo plasmídeo de 60 MDa.

Essa linhagem tem sido utilizada desde a década de 80 em aulas práticas de genética e microbiologia, para alunos de graduação e pós-graduação, e serviu de objeto de estudo para geração de diversos trabalhos acadêmicos (Informação verbal)[2]. Muitas das observações fenotípicas acerca dessa linhagem advêm de anos de observação de experimentos de obtenção de mutantes resistentes a antibióticos e conjugativos realizados no Laboratório de Genética de Microrganismos (LGM) ou em aulas práticas, e servem de base para as várias elucubrações apresentadas neste trabalho.

Os diversos trabalhos realizados com a BH100 focaram na caracterização dos elementos plasmidiais. GUIMARÃES (1986) determinou que a *E. coli* BH100 apresenta resistência de curto espectro ao mercúrio, *i.e* apenas ao mercúrio inorgânico, também destacando a liberação de fatores, possivelmente ricos em grupos tiólicos, no meio de cultivo, capazes de inativar a ação do íon. NASCIMENTO (1987), ao clonar os genes do operon *mer*, presentes no plasmídeo pBH100, chama a atenção para proximidade desse operon com a origem de replicação desse plasmídeo e ao gene capaz de conferir resistência ao cloranfenicol. DE PAULA (1989) identifica mutantes de alta frequência e com resistência variável à estreptomicina, possivelmente associados a um efeito sinérgico entre elementos do plasmídeo pBH100 e do cromos-

[1] Chartone-souza, E. (Instituto de Ciências Biológicas, UFMG). Comunicação pessoal, Maio 2017.
[2] Chartone-Souza, E.; Nascimento, A.M.A. (Instituto de Ciênicas Biológicas, UFMG). Comunicação pessoal. Maio 2017.

somo bacteriano. DE LACERDA (1995) demonstrou que a enzima β-lactamase, sintetizada pelo plasmídeo pAp, é do tipo TEM-1 e é liberada para o meio extracelular.

A maior parte dos trabalhos realizados com a *E. coli* BH100 datam do final do século XX. Poucos estudos, no entanto, propuseram caracterizar o genoma dessa linhagem de maneira holística. Tamanho estudo foi estimulado graças ao advento das pesquisas genômicas e o crescente desenvolvimento das tecnologias de sequenciamento. Nesse contexto, propôs-se a caracterização funcional e estrutural do genoma completo dessa linhagem.

Segundo Chartone-Souza e Nascimento (Informação verbal)[3] foram obtidas, no final da década de 80, variantes da *E. coli* BH100 construídos a partir da cura dos plasmídeos. Para tanto, conseguiram obter uma variante dessa bactéria, contendo apenas o plasmídeo de 60 MDa, outra variante contendo apenas os plasmídeos de 10 MDa, e uma variante ausente de ambos os plasmídeos. A *E. coli* BH100 tal como suas variantes, sempre foram armazenadas à temperatura ambiente ou refrigerador e foram repicadas, sempre que necessário, ao longo dos anos, para fins de pesquisa ou de aula prática. Em 1990, foram sessadas as atividades com a variante ausente dos dois plasmídeos, sendo a mesma não mais repicada desde então (Informação verbal)[4]. Em 2014 foram cedidas ao Laboratório de Genética Celular e Molecular (LGCM) amostras da *E. coli* BH100 e suas variantes para isolamento e sequenciamento do genoma na plataforma IonTorrent, além da montagem e anotação. À época, não foram feitos nenhum tipo de ensaio fenotípico, de modo a validar a presença dos marcadores de resistência em cada variante, de modo que as linhagens foram, assim que recebidas, cultivadas e seus genomas extraídos para sequenciamento. Os pesquisadores, então responsáveis pelo projeto, não foram capazes de identificar *in silico* os plasmídeos de cada linhagem, bem como, corretamente, montar o genoma dessa bactéria e de suas variantes, a partir dos dados desses sequenciamentos.

No início de 2017, cedeu-se novamente ao LGCM a *E. coli* BH100 e suas variantes, para extração e sequenciamento dos plasmídeos. Foram entregues a linhagem selvagem da BH100, a variante contendo apenas o plasmídeo de 60 MDa e a variante ausente dos plasmídeos. À época, o Laboratório de Genética de Microrganismos (LGM) não mais possuía a variante contendo apenas o plasmídeo de 10 MDa. Dessa forma, fora entregue uma linhagem transconjugante da *E. coli* K-12, contendo o mesmo plasmídeo. Com isso, procedeu-se com os ensaios fenotípicos, de modo a validar os marcadores de resistência esperados para cada

[3] Chartone-Souza, E.; Nascimento, A.M.A. (Instituto de Ciênicas Biológicas, UFMG). Comunicação pessoal. Maio 2017.
[4] Nascimento, A.M.A. (Instituto de Ciências Biológicas, UFMG). Comunicação pessoal. 2017

variante. A extração dos plasmídeos não foi bem-sucedida, de forma que, alguns meses depois, optou-se por trabalhar com a remontagem dos genomas sequenciados na plataforma IonTorrent.

Em agosto de 2017 surgiu a oportunidade de sequenciar, na plataforma Illumina HiSeq, as linhagens da BH100, obtidas em 2017, na esperança de conseguir obter as sequências plasmidiais. Na mesma época, no entanto, também conseguiu-se resolver a montagem do cromossomo e plasmídeos da *E. coli* BH100, obtida em 2014, bem como do plasmídeo de 60 MDa, presente exclusivamente em uma das variantes da BH100. De posse dos novos dados advindos de Illumina, julgou-se pertinente a montagem dos genomas.obtidos em 2017, e a realização de análises comparativas entre as variantes obtidas em 2014 e 2017.

O *International Code of Nomenclature of Bacteria* (LAPAGE et al, 1992) recomenda a utilização do termo clone para se referir a uma população de células derivadas em cultura de uma única célula parental e o termo linhagem[5] para se referir aos descendentes de um isolado único. Uma vez que essas bactérias não são armazenadas congeladas, e levando-se em conta o uso diferencial de cada variante, ao longo dos anos, hipotetizou-se que variações genéticas expressivas poderiam ser observadas entre esses genomas. Com efeito, de modo a estruturar os diferentes achados desse trabalho, optou-se por denominar as variantes da BH100 como linhagens derivadas da *E. coli* BH100. De igual forma, utilizou-se o termo sublinhagem para distinguir os genomas de uma mesma linhagem para se referir aos genomas sequenciados em 2014 e em 2017. Tal subdivisão tornar-se-á clara quando o leitor observar as diferenças estruturais (Capítulo 2) e funcionais (Capítulo 3), encontradas em cada genoma sequenciado.

1.1 A *Escherichia coli*

1.1.1 HISTÓRICO

A *Escherichia coli* pertence ao reino Eubacteria, domínio Bacteria, filo Proteobacteria, classe Gammaproteobacteria, ordem Enterobacteriales, família Enterobacteriaceae e gênero *Escherichia*. Ela foi isolada em 1885 pelo bacteriologista alemão Theodor Escherich de fezes humanas (ESCHERICH, 1885; KHETRAPAL; MEHERSHAHI; CHEN, 2017). A *E. coli* é

[5]Em inglês o termo é referido como *strain.* Em português, no entanto, parece não haver consenso na utilização de um termo universal. Certos pesquisadores utilizam os termos cepa, linhagem ou estirpe como sinônimos da palavra inglesa. Outros pesquisadores, no entanto, consideram o termo "cepa" como informal, outros como de uso exclusivo do meio clinico. Neste trabalho, utilizarei os três termos como sinônimos do vocábulo *strain*.

uma bactéria Gram-negativa, apresenta formato de bastão, é anaeróbia facultativa, não esporulante e comumente presente no intestino de mamíferos.

Seu genoma consiste de um cromossomo circular, de tamanho variando de 4,0 a 6,0 Mb, possuindo de 4.000 a 6.000 genes, e, em certos casos, também associado a plasmídeos ("Genomes - Genome - NCBI", [s.d.]). Atualmente, contam na base de dados do NCBI 9.899 genomas de *E. coli*, compreendendo uma diversidade de variantes dessa espécie. A mundialmente conhecida *E. coli* K-12, isolada em 1922 de fezes de pacientes com difteria, tem sido objeto de manipulação genética e geração de diversos variantes (BACHMANN, 1972; KOLI et al., 2011). A primeira linhagem sequenciada e amplamente caracterizada foi uma sublinhagem derivada dessa estirpe e denominada *E. coli* K-12 subtr. MG1655 (BLATTNER et al., 1997). Outra linhagem de destaque é a *E. coli* B, que deu origem à *E. coli* BL21, amplamente utilizada em experimentos de clonagem (DAEGELEN et al., 2009).

1.1.2 PATÓTIPOS

Diferentemente das linhagens comensais descritas acima, algumas linhagens de *E. coli* têm a capacidade de provocar doenças. Baseado nos fatores de virulência, hospedeiro e modo de ação, as diferentes linhagens patogênicas podem ser classificadas nos seguintes patótipos:

a) *Intestinal Pathogenic E. coli* (IPEC): nesse grupo incluem-se as variantes patogênicas obrigatórias, associadas ao trato gastrointestinal. Podem ser classificadas de acordo com seu modo de virulência nos seguintes tipos:

— *Enteropathogenic E. coli* (EPEC): responsável por causar diarreia em crianças e animais, esse grupo caracteriza-se por causar lesões na mucosa intestinal, levando ao comprometimento das microvilosidades. Um dos principais genes de virulência produzidos por linhagens desse patótipo, codifica para a proteína intimina, e serve para identificação molecular do grupo (OCHOA; CONTRERAS, 2011);

— *Enterohemorrhagic E. coli* (EHEC): responsável por casos de colite hemorrágica e síndrome hemolítica-urêmica, esse grupo caracteriza-se pela síntese de uma toxina antigenicamente e estruturalmente similar à toxina Shiga, capaz de inibir a síntese proteica nas células hospedeiras (NGUYEN; SPERANDIO, 2012);

— *Enterotoxigenic E. coli* (ETEC): responsável pela diarreia do viajante e casos de diarreia em bovinos, esse grupo caracteriza-se pela síntese de fatores de colonização e de enterotoxinas. Os primeiros possibilitam a permanência dessas linhagens no intes-

tino delgado, onde a ação das toxinas termoestáveis e termolábeis interferem com a regulação de íons cloreto pelas células intestinais (QADRI et al., 2005);

— *Enteroaggregative E. coli* (EAEC): responsáveis por casos de diarreia persistente, em humanos. Esse patótipo pode, ainda, ser distinguido pela subclasse das *Diffusely Adherent E. coli* (DAEC), responsável por casos de diarreia em crianças. Em ambos os casos, as bactérias desse grupo induzem a formação de um biofilme capaz de interferir com as funções normais da mucosa intestinal (OKEKE; NATARO, 2001);

— *Enteroinvasive E. coli* (EIEC): responsável por casos de diarreia e desinteria, as linhagens desse grupo caracterizam-se pela capacidade de penetrar nas células epiteliais do intestino (PASQUA et al., 2017).

b) *Extraintestinal E. coli* (ExPEC): nesse grupo incluem as variantes tidas como patogênicas facultativas e, geralmente, associadas à microbiota normal de indivíduos, onde coabitam como bactérias comensais (KÖHLER; DOBRINDT, 2011). A patogenicidade emerge, quando fora do trato gastrointestinal, podendo apresentar características únicas frente ao nicho de contaminação. Esse grupo é subdividido, de acordo com o local de infecção:

— *Uropathogenic E. coli* (UPEC): responsáveis por cerca de 80% dos casos de infecção do trato urinário em humanos e animais, essas linhagens têm sido amplamente estudadas a respeito dos fatores de virulência, que incluem adesinas, toxinas, polissacarídeos de membrana, flagelos, sistemas de aquisição de ferro, entre outros (TERLIZZI; GRIBAUDO; MAFFEI, 2017);

— *Neonatal Meningitis E. coli* (NMEC): responsáveis por casos de meningite e sepse, essas linhagens detêm a capacidade de colonizar o fluido cérebro-espinhal nas primeiras horas de vida de recém-nascidos (KU; BOGGESS; COHEN-WOLKOWIEZ, 2015);

— *Sepsis-causing Pathogenic E. coli* (SEPEC): responsáveis por casos de contaminação generalizada no organismo (BOZCAL, 2016);

— *Avian Pathogenic E. coli* (APEC): responsáveis por infecções sistêmicas em aves (DHO-MOULIN; FAIRBROTHER, 1999);

— *Mammary Pathogenic E. coli* (MPEC): associadas a infecções em glândulas mamárias de bovinos (SHPIGEL; ELAZAR; ROSENSHINE, 2008);

— *Endometrial Pathogenic E. coli* (EnPEC): associados a infecções no tecido endometrial (SHELDON et al., 2010).

As diferentes estratégias utilizadas na patogenicidade dependem da síntese de fatores de virulência que auxiliam as células bacterianas nos processos de invasão, indução da doença e evasão dos mecanismos de defesa do organismo. PETERSON (1996) enumera os seguintes tipos: a) fatores de aderência: associados à capacidade da bactéria em se aderir aos tecidos do hospedeiro; b) fatores de invasão: elementos presentes na superfície bacteriana e associados à capacidade de invadir as células do hospedeiro; c) cápsulas: envoltórios da parece celular capazes de conferir proteção contra o sistema imune; d) endotoxinas: componentes lipopolissacaridicos de bactérias Gram-negativas responsáveis por causar febre, inflamação e outros efeitos tóxicos; e) exotoxinas: estruturas proteicas, enzimáticas ou não, secretadas pelo microrganismo e com efeitos citotóxicos, neurotóxicos ou enterotóxicos; f) sideróforos: esses elementos são responsáveis por competir com o hospedeiro pela biodisponibilidade de íons Fe^{2+}/Fe^{3+} do meio.

Esses fatores de virulência podem ser encontrados ao longo do cromossomo em transposons, fagos temperados, ilhas de patogenicidade e plasmídeos (WALDOR; DAVIS, 2002). Estes últimos, tem grande relevância nos mecanismos de patogenicidade de diversas linhagens pertencentes aos diferentes patótipos, descritos acima (JOHNSON; NOLAN, 2009). Em ExPEC, por exemplo, plasmídeos do grupo colV têm sido associados ao favorecimento de crescimento bacteriano em meio ácido (RAJA et al., 1991) e em urina (SKYBERG et al., 2006), além de elevar os níveis de virulência em modelos animais para meningite neonatal (AGUERO et al., 1989).

1.1.3 IMPORTÂNCIA

A *E. coli* tem sido extensivamente utilizada como um organismo modelo em estudos microbiológicos e genéticos (IDALIA; BERNARDO, 2017). De fato, apenas nessa espécie, diversos estudos possibilitaram o embasamento da biologia molecular, incluindo a elucidação da replicação do DNA, o desvendamento do código genético e as origens da síntese proteica. Essa bactéria é bem adaptada às condições laboratoriais e capaz de se desenvolver em meios de cultivo relativamente simples. Populações de *E. coli* podem ser geradas com relativa facilidade a partir de um único clone e servem de base para estudos de mutantes resistentes a antibióticos ou mesmo detentores de um fenótipo metabólico particular.

Essa bactéria também tem sido utilizada em laboratórios de biologia molecular como veículo de expressão e clonagem gênica, o que tem possibilitado importantes avanços biotecno-

lógicos. Uma das maiores contribuições na utilização dessa bactéria talvez esteja relacionada à síntese recombinante de insulina humana para tratamento da diabetes (BAESHEN et al., 2014) ou mesmo para geração de antígenos vacinais (IHSSEN et al., 2010).

A *E. coli* também serve de modelo para estudo de resistência a antimicrobianos. Essa bactéria apresenta um perfil de resistência e susceptibilidade a essas moléculas bastante diversificado (**Figura 1.1**). Esse fato encontra-se relacionado à aquisição de resistência por algumas linhagens e ao desenvolvimento de novos tipos de antibióticos capazes de impedir a propagação de espécimes patogênicas, pela indústria farmacêutica.

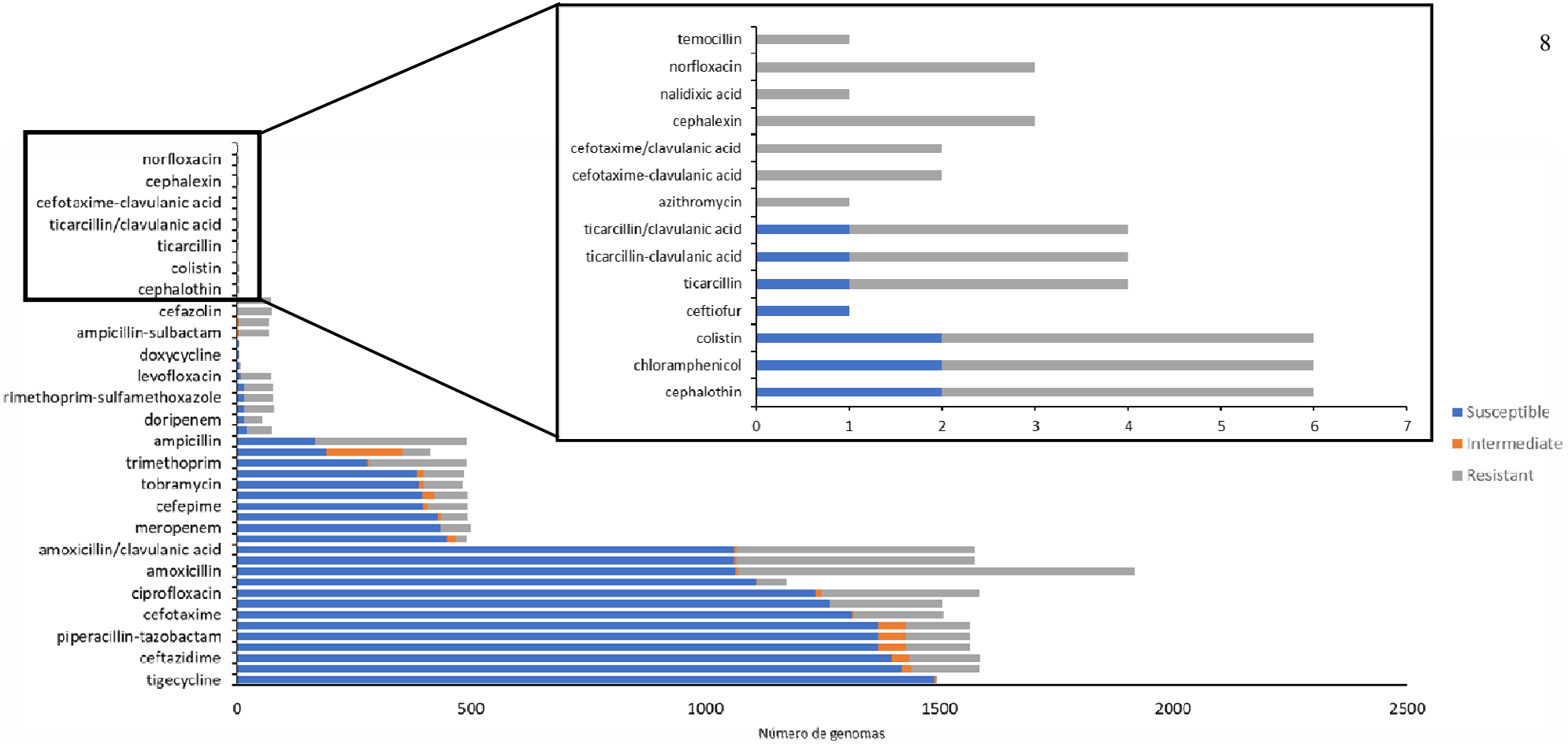


Figura 1.1 Perfil de resistência completa, intermediária e susceptibilidade a diferentes antimicrobianos computados para algumas linhagens de *E. coli*. Adaptado de: PATRIC. <https://www.patricbrc.org/view/Taxonomy/562#view_tab=overview>.Acesso:Fev.2018.

1.2 A Multirresistência Bacteriana

O uso de antibióticos tem sido essencial no tratamento de infecções bacterianas. Infelizmente, a eficiência de qualquer antimicrobiano é reduzida frente a linhagens resistentes. Essa resistência pode surgir devido a um ou mais mecanismos de geração de variabilidade discutidos em 1.1.1. Por exemplo, quinolonas são capazes de inibir a enzima DNA girase e com isso impedir a replicação do DNA. A resistência contra estes antibióticos pode ser adquirida mediante mutações de ponto que alteram a estrutura da enzima e com isso se tornam menos susceptíveis a interagir com essas moléculas (HOOPER; JACOBY, 2015). De forma geral, a resistência a antimicrobianos pode emergir a partir de alterações no alvo da droga, redução da permeabilidade da membrana ao antibiótico, inativação da droga por modificação enzimática e/ou expulsão da droga por bombas de efluxo (MUNITA; ARIAS, 2016).

Poderia se pensar que o surgimento de multirresistência bacteriana dependeria de eventos mutacionais em diversos genes. Apesar de possível, esse cenário encontra-se mais vinculado à presença dos chamados plasmídeos R, os quais carregam consigo genes de resistência a diversos antibióticos. Esses elementos são responsáveis pela rápida disseminação de resistência à antimicrobianos e tem dificultado ações terapêuticas em casos de infecções por microrganismos detentores desses plasmídeos (STONE, 1975).

Um estudo de caso associado aos efeitos clínicos de espécimes detentoras desses plasmídeos pode ser ilustrado pelo uso de sulfonamidas para tratar infecções provocadas por *Shigella dysenteriae*. Nesse exemplo, após cerca de uma década de tratamento com esse antimicrobiano, observou-se que mais de 90% dos casos associados à disenteria no Japão estavam relacionados a variantes resistentes de *S. dysenteriae* (NIYOGI, 2007). Nessa época, também eram administrados cloranfenicol e tetraciclina. A primeira cepa dessa bactéria resistente a todos os antibióticos foi identificada na década de 50, e cerca de 15 anos depois, a maioria das linhagens de *Shigella* já eram resistentes a estes, e também à ampicilina.

Devido ao fato de plasmídeos poderem se transferir entre diversas linhagens bacterianas, eles são frequentemente alvo de transposons. De fato, estudos nos plasmídeos R evidenciaram que o relativo alto número de genes de resistência estaria associado a esses elemento (CLEWELL, 2014; DATTA et al., 1979; HEDGES; JACOB, 1974; SHAW et al., 1979). Transposons que carregam genes de resistência podem inclusive se inserir em outros transposons e com isso formar verdadeiras estruturas móveis codificantes de diversos genes de resistência (NIKAIDO, 2009).

Também vale destacar a ação de integrons, que funcionam como instrumentos de captação de cassetes gênicos. Esses elementos podem também ser incorporados a transposons e com isso podem assegurar a dispersão de diversos genes de resistência de uma única vez. Com efeito, a combinação de transposons, integrons e plasmídeos fornecem um forte mecanismo para rápido espalhamento da resistência a múltiplos antibióticos entre bactérias.

Outro aspecto relevante está associado à correlação de outros genes não necessariamente vinculados à resistência, mas que acabam sendo carreados em conjunto com genes que confiram algum tipo de vantagem adaptativa. Esse fato merece destaque uma vez que sequências presentes em plasmídeos e fagos podem se incorporar nos cromossomos. Dessa forma, caso esses genes assegurem algum tipo de vantagem adaptativa à bactéria, eles podem ser permanentemente incorporados. Essas regiões constituem as chamadas ilhas genômicas e são caracterizadas segundo suas funções (JUHAS et al., 2009). Por exemplo, regiões que contenham genes que confiram patogenicidade são denominadas ilhas de patogenicidade; ilhas que possuam genes associados a algum tipo de adaptação simbiótica são denominadas ilhas de simbiose, etc. A comparação de diversas linhagens bacterianas, muitas vezes distantemente relacionadas, revela a relativa conservação dessas ilhas e corrobora com as implicações da transferência horizontal como força motriz da evolução bacteriana (JUHAS et al., 2009; SCHMIDT; HENSEL, 2004).

1.3 A DIVERSIDADE GENÉTICA BACTERIANA

"*Everything is everywhere but the environment selects*" (BASS BECKING, 1934 *apud* DE WIT; BOUVIER, 2006). Essa frase sumariza a pluralidade de microrganismos, sujeitos às mais diversas condições ambientais. Seus tamanhos diminutos, somados a um alto número de indivíduos e baixas taxas de extinção, contribuem, sumariamente, para dispersão e conquista de nichos ecológicos (MARTINY et al., 2006).

A variabilidade genética está intrinsicamente associada à capacidade adaptativa de microrganismos. Esse fato aliado às leis de seleção natural possibilitaram a coexistência de microrganismos adaptados a um determinado ambiente. A relação homeostática observada nessas populações é fruto de ações cooperativas e destrutivas essenciais para sobrevida em situações de alterações ambientais.

Populações bacterianas, em geral, apresentam maior diversidade genética, quando comparadas a eucariotos. A diversidade alélica média por *locus* em *Escherichia coli*, por exemplo, varia de 0,343 a 0,542 e é relativamente alta, quando comparada com espécies eucarióticas

(SELANDER; CAUGANT; WHITTAM, 1987). Em humanos a estimativa da diversidade é de 0,063 (NEEL, 1984), e, em geral, inferior a 0,15 para maioria dos demais eucariotos (NEVO; BEILES; BEN-SHLOMO, 1984). Já entre espécies bacterianas, observam-se populações mais ou menos variáveis. O genoma de populações de *E. coli* mostra-se mais variável, quando comparado com outras populações bacterianas, ainda que seja menos variável, quando comparado com *Haemophilus influenzae* (MUSSER et al., 1986) ou *Neisseria meningitidis* (CAUGANT et al., 1986).

Não raro, sob condições apropriadas de seleção, novos genes podem evoluir e se espalhar rapidamente. A compreensão dos mecanismos de geração de variabilidade em microrganismos é de suma importância para o entendimento do surgimento e propagação de resistência a antibióticos e patogênese. Os diversos mecanismos genéticos que contribuem para geração de variabilidade são explorados a seguir.

1.3.1 MECANISMOS ASSOCIADOS À GERAÇÃO DE VARIABILIDADE

Uma única célula bacteriana pode dar origem a uma população de células idênticas em um curto espaço de tempo, via fissão binária. Ocasionalmente, mudanças genéticas espontâneas ocorrem em uma dessas células e são herdadas pelas futuras gerações. Se a alteração for prejudicial, a mesma provavelmente não prevalecerá, mas se for benéfica ou neutra poderá ser selecionada e fará parte de toda a população.

Mutações espontâneas em bactérias podem ser de dois tipos: mutações de ponto ou estruturais. O primeiro tipo é responsável por alterações em, pelo menos, um único nucleotídeo, já o segundo caso, refere-se a eventos recombinatórios associados à inserção, deleção ou inversão do DNA. Esses eventos ocorrem geralmente a uma baixa frequência, mas quando ocorrem podem rapidamente se espalhar pela população (DENAMUR; MATIC, 2006; WANG et al., 2001).

Outra forma de geração de variabilidade, em bactérias, está associada à transferência horizontal de genes. Nesse caso, o material genético de uma célula ou vírus é transferido para uma célula receptora e pode contribuir para geração do *fitness* necessário para sobrevivência em uma determinada condição ambiental (RAZ; TANNENBAUM, 2010).

1.3.1.1 Mutação de ponto

Nas mutações de ponto, as alterações nucleotídicas podem ocorrer via substituição de base ou i*ndels.* No primeiro caso, destacam-se as transversões, caracterizadas pela mudança de pirimidinas para purinas e vice-versa; e as transições, definidas como alterações intrapurinicas ou intrapirimidinicas. Já os *indels* se referem a eventos de inserção ou deleção de nucleotídeos. Esses eventos podem levar a uma série de consequências para a célula. Em regiões codificantes elas podem ser responsáveis por mudanças de um aminoácido em uma cadeia polipeptídica e, com isso, levar à inativação da proteína, redução/aumento de atividade (ISHIDA, 2010) ou até mesmo geração de uma nova função (SUHL; WARREN, 2015). Já em regiões não codificantes, essas alterações podem interferir em sítios associados ao controle da expressão gênica (BOLDRIN et al., 2018; HOFFMANN et al., 2017).

Essas alterações podem ocorrer espontaneamente devido a erros na replicação do DNA ou podem ser induzidas por mutagênicos presentes no ambiente. Bactérias possuem diversos mecanismos capazes de corrigir danos ao DNA, muito embora, alterações desses sistemas podem, por vezes, ser selecionadas e acelerar a geração de variabilidade. Tal evento faz-se útil, por exemplo, no surgimento de resistência a antibióticos (MARTINEZ; BAQUERO, 2000; SUZUKI et al., 2018).

A seleção de genótipos que favoreçam o aumento nas taxas mutacionais é denominada seleção de ordem secundária e é explicada pelo modelo *hitchhiking*. Nesse modelo, antecipa-se que em populações assexuadas, alelos favoráveis permaneçam associados à estrutura genética que os deram origem. A vantagem seletiva desses alelos proporciona aumento da sua frequência juntamente com seu *background*. Genótipos mutadores podem gerar alelos favoráveis mais frequentemente que não-mutadores, aumentando sua frequência caso a vantagem conferida por alelos benéficos seja superior ao custo de ser um mutador (TENAILLON et al., 2001).

Em *E. coli* estima-se que as taxas de mutação sejam da ordem de 10^{-10} mutações por nucleotídeo por geração ou 10^{-3} mutações por genoma por geração (LEE et al., 2012). Já a frequência de mutadores em populações dessa espécie é da ordem de 10^{-5} (BOE et al., 2000). Os mecanismos relacionados ao surgimento de um genótipo mutador estão associados ao comprometimento dos sistemas vinculados à fidedignidade do processo de cópia do DNA. Como exemplos de comprometimento desses sistemas pode-se citar: a redução de especificidade da polimerase para incorporar nucleotídeos corretos à fita nascente de DNA (OSHEROFF et al., 1999); atenuação da capacidade de *proofreding* da DNA polimerase

(COX, 1976; RAYNER et al., 2016); alterações no sistema de reparo de *mismatches* vinculadas à metilação do DNA (MARINUS, 2012); e alterações na expressão ou funcionalidade de enzimas vinculadas à reversão das reações de alquilação de guanina (LOECHLER; GREEN; ESSIGMANN, 1984).

1.3.1.2 Elementos móveis

Os elementos móveis são os grandes responsáveis por alterações na estrutura do DNA. Um importante representante desse grupo são as Sequências de Inserção (IS). ISs podem ser definidas como fragmentos curtos de DNA que possuem sequências repetidas e invertidas em suas extremidades, e que codificam proteínas associadas à sua translocação (MAHILLON; CHANDLER, 1998). Em geral, essas sequências podem ser inseridas em qualquer região do genoma em uma frequência relativamente baixa (MALOY, 2001).

A inserção de ISs em uma região codificante pode levar à inativação do gene (OOKA et al., 2009). Alguns ISs também podem agir sob regiões promotoras e com isso levar à ativação de genes silenciosos (SAEDLER et al., 1974). A translocação desses elementos está intrinsicamente associada à formação de deleções e inversões no genoma. Outra importante propriedade desses elementos é a capacidade de dois ISs idênticos transpor regiões gênicas, por eles flanqueadas, para outros sítios no genoma. Essa estrutura é denominada transposon (Tn) composto e é de suma importância na propagação de genes associados à resistência a antimicrobianos. Esse é o caso, por exemplo, do Tn10 que confere resistência à tetraciclina e apresenta duas unidades de IS10 invertidas entre si na sua extremidade (CHALMERS et al., 2000), do Tn5 que possui duas unidade de IS50, flanqueando um gene de resistência à canamicina (REZNIKOFF, 2008), e do Tn9 que codifica resistência ao cloranfenicol e constituido por duas unidades de IS1 (ALTON; VAPNEK, 1979).

Outros elementos de destaque são os transposons não-compostos. Nesse caso, não há a presença de ISs, mas a presença de estruturas repetidas e invertidas nas extremidades e, em geral, a presença de genes associados à transposição. Um exemplo clássico de transposon não-composto é o Tn21, que carrega consigo o operon *mer*, responsável por codificar resistência ao mercúrio inorgânico (LIEBERT; HALL; SUMMERS, 1999a).

De maneira geral, esses elementos móveis operam segundo um mecanismo de transposição conservativa (*cut-and-paste*) onde são removidos do sítio original e reinseridos em outro local. No entanto, a duplicação desses elementos pode ser observada caso esse evento ocorra durante a fase de replicação do DNA. Nesse caso, um elemento móvel já duplicado pode ser

excisado e reinserido em uma região ainda não duplicada, proporcionando um aumento no seu número de cópias. Esses elementos também podem ser classificados quanto à natureza de sua transposição. Os elementos ditos autônomos são capazes de se deslocar por si próprios, enquanto os não-autônomos dependem de outros elementos por, geralmente, não possuírem genes codificadores de transposases (PRAY, 2008).

Também vale destacar aqui, os integrons (In). Apesar de não serem elementos móveis, propriamente ditos, os mesmos são responsáveis pela captação e eventual acúmulo de cassetes gênicos, podendo ser mobilizados, caso inseridos em um transposon (MAZEL, 2006). Esses elementos podem ser classificados em quatro classes distintas, dependendo do tipo de sítio de inserção presente no integron (DENG et al., 2015). Novamente, o Tn21 pode ser utilizado para exemplificar a importância desses elementos. Nesse caso, verifica-se em sua composição um integron de classe 1, denominado In2, e que apresenta os genes de resistência à estreptomicina e à sulfonamida em sua estrutura (LIEBERT; HALL; SUMMERS, 1999b).

Os integrons são capazes de integrar cassetes gênicos graças à ação da enzima integrase. Nesse caso, essa enzima é responsável por reconhecer os sítios *attC* dos cassetes gênicos e inseri-los nos sítios *attI*, presentes na estrutura do integron (**Figura 1.2**), via recombinação sítio-dirigida. Os cassetes gênicos inseridos ficam sujeitos à expressão mediada por promotores próprios do integron. Esse sistema proporciona o acúmulo de genes e é bastante difundido entre as linhagens multirresistentes a antibióticos (GILLINGS, 2014; MAZEL et al., 2015).

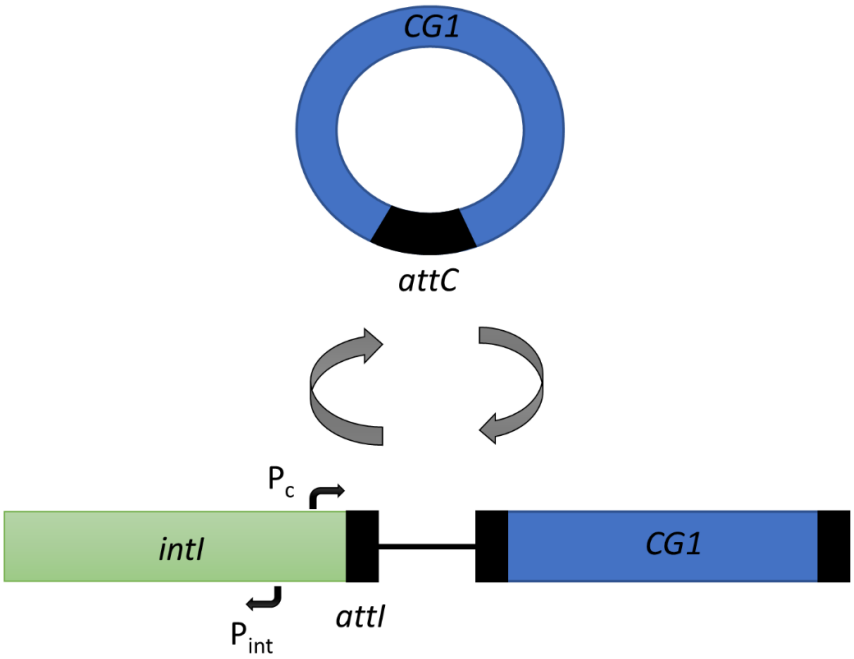


Figura 1.2 Mecanismo de aquisição de cassetes gênicos por integrons. A integrasse é codificada pelo gene *intI* regulada pelo promotor Pint e age recombinando o sítio *attC*, presente no cassete gênico (CG), no sítio *attI* do integron. O cassete fica então sujeito à expressão via promotor Pc.

1.3.1.3 Transferência horizontal

1.3.1.3.1 Conjugação e mobilização plasmidial

Plasmídeos são pequenas[6] moléculas de DNA fita-dupla, autoreplicativos e mantidos na célula juntamente com o cromossomo. Em geral, essas sequências são circulares, ainda que plasmídeos lineares já tenham sido identificados (AHSAN; KABIR, 2013; HINNEBUSCH; TILLY, 1993). Esses elementos, em geral, não são essenciais para os processos básicos de crescimento celular, mas são de extrema valia para adaptação e conquista de diversos ambientes, por carregarem genes capazes de conferir vantagens adaptativas. Por exemplo, plasmídeos podem codificar bacteriocinas capazes de matar outras células (JEZIOROWSKI; GORDON, 2007), codificar genes que permitam o surgimento de vias metabólicas associadas à utilização de compostos exóticos, como fontes de energia (BORONIN et al., 1984; CHAKRABARTY; CHOU; GUNSALUS, 1973; PAUWELS et al., 1999), além de conterem genes associados à resistência a antibióticos (COX; SCHILDBACH, 2017) e patogenicidade (OLOOMI; JAVADI; BOUZARI, 2015).

[6] Em relação ao DNA cromossomal

Uma importante propriedade desses replicons é a habilidade de se transferir para outras células, podendo ser classificados segundo um dos seguintes mecanismos de transferência: 1) plasmídeos autoconjugativos (GRIFFITHS, 1999) e 2) plasmídeos mobilizáveis (RAMSAY; FIRTH, 2017).

No primeiro caso, os plasmídeos contêm todos os genes necessários para promover sua transferência. Cerca de 20 – 30 genes associados aos operons *tra* e *trb* codificam produtos capazes de induzir o pareamento de células, a formação de poros e o carreamento do plasmídeo para a célula receptora. Um dos plasmídeos autoconjugativos mais bem estudados é o plasmídeo F de *E. coli* (GRIFFITHS, 1999). Esse é um plasmídeo de aproximadamente 100 kb, com cerca de um terço de sua estrutura dedicada a genes associados à transferência. Células que contenham esse replicon são denominadas F^+, já as células receptoras ausentes desse replicon são denominadas F^-. Os genes presentes nesse plasmídeo são responsáveis pela formação do chamado pili sexual. Essa é uma estrutura tubular que se projeta para o exterior celular e que ao entrar em contato com uma célula F^- é capaz de aderir à sua superfície. Em seguida, essa estrutura se despolimeriza, aproximando as duas células e estimulando a formação de um poro capaz de conectar o interior de ambas. Uma das fitas do plasmídeo F é então clivada em uma região denominada origem de transferência (*oriT*) e carreada para a célula receptora através desse poro. Ao fim da transferência, a fita complementar é sintetizada em ambas as células e a célula receptora passa a ser F^+.

Já os plasmídeos mobilizáveis não possuem a maquinaria necessária para se autotransferirem, muito embora, possam ser carreados para outras células mediante a presença de um plasmídeo autoconjugativo. Esses plasmídeos podem manifestar diferentes mecanismos de mobilização, a saber: 1) Cointegração; 2) Condução; 3) dependente de proteína Mob*;* 4) independente de Mob.

No caso de plasmídeos mobilizáveis serem transferidos via cointegração, verifica-se a recombinação desses replicons com o plasmídeo autoconjugativo (SCHARBAAI-VÁZQUEZ; CANDELAS; TORRES-BAUZÁ, 2007). Nesse caso, ambos os plasmídeos formam uma estrutura circular única, fruto de recombinação homóloga entre duas regiões idênticas presentes em cada replicon. Uma das formas de gerar tal evento é através da transposição de um IS para o outro plasmídeo seguido de um evento de recombinação homóloga (MACHIDA et al., 1982).

Outro mecanismo denominado condução está associado à integração do plasmídeo ao cromossomo e à transferência de regiões deste para a célula receptora. Células que possuam essa capacidade são denominadas Hfr e estão associadas à integração de plasmídeos autocon-

jugativos ao cromossomo e à sua conseguinte transferência para a célula receptora. Quanto ao mecanismo de mobilização, verifica-se que o próprio plasmídeo não-conjugativo pode se integrar no cromossomo e se transferir para a célula receptora de forma análoga a observada em células Hfr, mas utilizando a maquinaria do plasmídeo autoconjugativo. Esse mecanismo de mobilização é também conhecido como transferência *Hfr-like* (CLARK; ADELBERG, 1962).

Alguns plasmídeos mobilizáveis são capazes de sintetizar uma relaxase denominada Mob. Diversos mecanismos de mobilização dependentes e independentes dessa enzima foram recentemente revisados (RAMSAY; FIRTH, 2017). Em todos os casos o plasmídeo mobilizável aproveita-se da maquinaria do plasmídeo autoconjugativo para promover sua transferência. De forma geral, Mob é capaz de reconhecer a *oriT* desses plasmídeos estimulando a clivagem de uma das fitas e o deslocamento do replicon pelo poro formado pelas proteínas do plasmídeo autoconjugativo. No entanto, certos plasmídeos não-conjugativos podem ser transferidos por mecanismos independentes de Mob, graças à similaridade de suas origens de transferência com a do plasmídeo autoconjugativo. Nesse caso, as enzimas associadas à clivagem das fitas do plasmídeo autoconjugativo podem reconhecer essas regiões e carrear o plasmídeo mobilizável pelo poro.

1.3.1.3.2 Transformação

A transformação consiste na aquisição de DNA do meio extracelular. Esse fenômeno foi descrito por GRIFFITH (1928) ao observar que a injeção de lisados de uma linhagem virulenta de *Streptococcus. pneumoniae* juntamente com uma cepa avirulenta levava à morte de camundongos devido a aquisição de virulência pela linhagem não-virulenta. Mais tarde, (AVERY; MACLEOD; MCCARTY, 1944) descobriram que a substância transformante tratava-se do DNA.

Bactérias podem captar DNA do ambiente apenas se estiverem em um estado denominado competente. Bactérias naturalmente transformáveis como ocorre com *Pasteurella* e *Haemophilus*, por exemplo, são capazes de expressar proteínas, sob determinadas condições de cultivo, associadas à captação de DNA exógeno (MELL; REDFIELD, 2014). Já bactérias naturalmente não-competentes, como *E. coli*, podem tornar-se competentes num contexto laboratorial, mediante tratamento com $CaCl_2$, por exemplo (SAMBROOK; RUSSELL, 2001).

1.3.1.3.3 Transdução

Vírus são moléculas de DNA ou RNA envolvidas em um capsídeo proteico. Esses microrganismos são encontrados no meio extracelular, mas podem se replicar apenas no interior de células vivas. Os vírus que infectam bactérias são denominados bacteriófagos. Estes estão presentes em grande número no ambiente e são relevantes no carreamento de genes entre bactérias (CLOKIE et al., 2011). Somente na base de dados do *National Center for Biotechnology Information* (NCBI) estão presentes sequências de fagos, cujos tamanhos variam de 1.417 pb referente ao *Campylobacter* phage C10 (GenBank: MG065651.1) até 497.513 pb referentes ao *Bacillus* phage G (GenBank: JN638751.1).

Os bacteriófagos ditos virulentos replicam-se no meio intracelular via ciclo lítico. Nesse caso, eles infectam células bacterianas, utilizam de sua maquinaria para se replicar e então rompem a célula para liberar centenas de novos fagos no ambiente. A interação de fagos com a parede celular bacteriana é altamente específica, de modo que esses vírus são capazes de discriminar linhagens da mesma espécie (CHIBANI-CHENNOUFI et al., 2004; FLORES et al., 2011; PARMAR et al., 2017). Dentro da célula os fagos estão sujeitos a um estrito mecanismo de desenvolvimento. Nos fagos líticos T7, por exemplo, capazes de infectar *E. coli*, os primeiros genes a serem expressos são os associados à replicação do DNA, silenciamento de genes bacterianos e expressão exclusiva dos genes do fago. Em seguida, são expressos genes associados à síntese do capsídeo – proteínas da cabeça e cauda. Ao final do ciclo, o DNA é empacotado na estrutura viral e a célula é lisada (KUTTER; GUTTMAN, 2001).

Já os fagos denominados temperados apresentam duas diferentes estratégias de replicação: lítica e lisogênica. No ciclo lisogênico, em particular, ocorre a incorporação do DNA viral no cromossomo bacteriano, passando a ser denominado prófago. Nessa situação, esses fagos expressam genes capazes de silenciar a expressão de genes associados ao ciclo lítico. Esse é o caso do bacteriófago lambda, capaz de infectar *E. coli*. Nesse caso, quando no ciclo lisogênico, o DNA viral é capaz de se integrar no cromossomo utilizando enzimas próprias de recombinação. Esse estado, entretanto, não é permanente. Quando, por exemplo, algum dos repressores de genes líticos é inativado, o prófago é estimulado a sair do cromossomo e dar início ao ciclo lítico (GUTTMAN, 2001).

Fagos detém a capacidade de transferir genes de uma bactéria para outra segundo um processo denominado transdução. Nesse caso, dois mecanismos podem estar presentes. Na transdução generalizada, durante o ciclo lítico e na fase de empacotamento do DNA, eventualmente alguns fragmentos aleatórios do material genético bacteriano podem ser empacotados

ao invés do DNA viral (WADDELL et al., 2009). Dessa forma, uma pequena fração dos fagos liberados contém em seu interior sequências bacterianas que podem ser conduzidas para outras células. Esses fragmentos de DNA bacteriano podem estar associados a ilhas de patogenicidade (SCHMIDT; HENSEL, 2004) e até mesmo plasmídeos completos (DEICHELBOHRER et al., 1985).

Outro mecanismo refere-se à transdução especializada, e ocorre por meio de fagos temperados (TUFARIELLO et al., 2002). Nesse caso, ao ser estimulado o ciclo lítico, regiões cromossomicas próximas à sequência do prófago podem ser excisadas em conjunto. Essas regiões podem carregar genes não sujeitos à ação repressora dos fatores de inibição dos genes líticos e com isso podem ser eficientemente expressos nas novas células sujeitas à infecção desses fagos. Um exemplo desse mecanismo ocorre nos fagos b presentes em *Corynebacterium diphtheriae*, e que carregam consigo o gene da toxina diftérica. Nesse caso, células infectadas por esses fagos, mas não sujeitos a um ciclo lisogênico, são incapazes de causar difteria. Já as células que apresentam o fago inserido no cromossomo o são (TODAR, [s.d.]).

1.4 A Era Genômica

A possibilidade de estudar genomas a partir dos benefícios advindos das plataformas de sequenciamento de nova geração (NGS) não seria possível sem o desenvolvimento de ferramentas de bioinformática. O desenvolvimento de algoritmos capazes de lidar com o grande número de dados gerados e o surgimento de pesquisadores qualificados para analisá-los foi de suma importância para o estabelecimento dessas novas tecnologias. Seu impacto tem possibilitado a diferentes grupos de pesquisa responder questões vinculadas às áreas de medicina, agricultura, virologia, microbiologia, ciências forenses, entre outras.

O surgimento e aprimoramento da bioinformática permitiu o desenvolvimento de softwares capazes de interpretar os sinais advindos de sequenciamentos NGS em um processo denominado *base calling* (LEDERGERBER; DESSIMOZ, 2011). Também foi graças ao advento desse campo que estratégias de montagem de genomas, além de procedimentos de anotação funcional, mapeamento e visualização se tornaram possíveis (**Figura 1.1**).

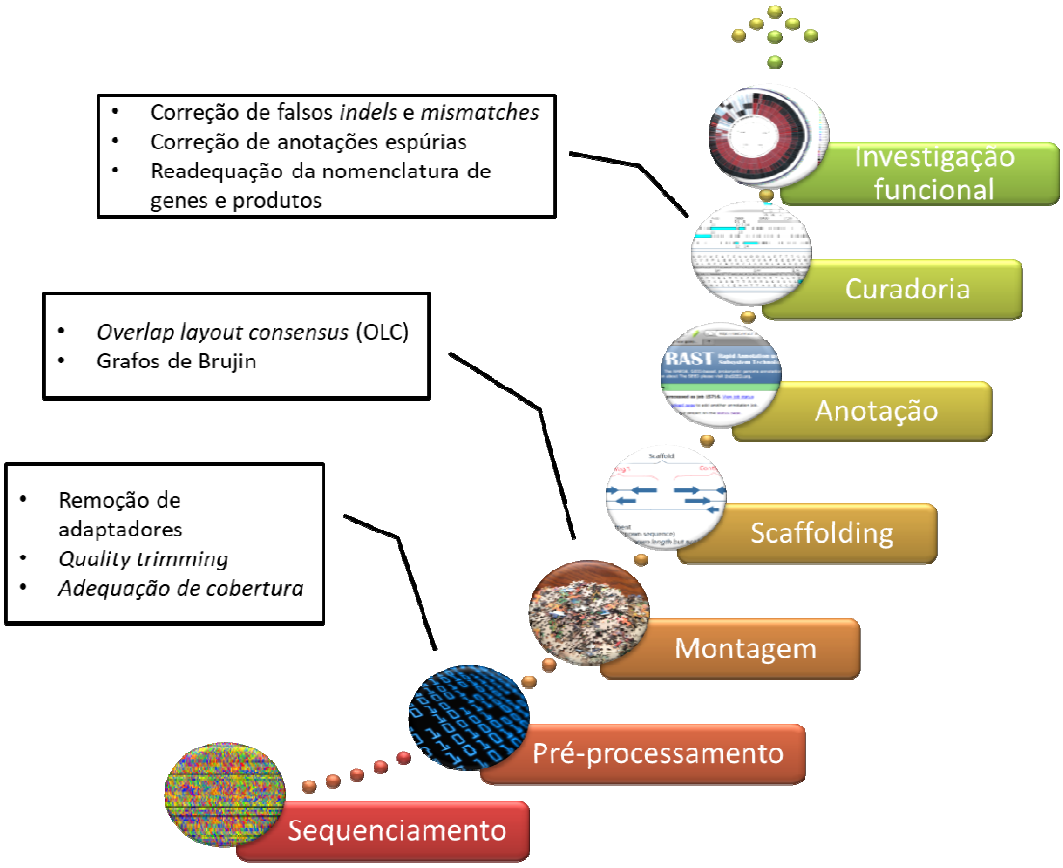


Figura 1.1 Etapas gerais de um projeto genômico. Após sequenciamento em uma ou mais plataformas, as *reads* geradas são pré-processadas, observando se os valores de qualidade e cobertura são adequados. Na etapa de montagem objetiva-se a construção de *contigs* que mais tarde virão a ser unidos em *scaffolds* e, eventualmente, no genoma completo. Dessa etapa em diante, dedica-se à caracterização funcional da sequência, além de etapas de curadoria visando a correção de erros de montagem e anotação.

1.4.1 SEQUENCIAMENTO DE NOVA GERAÇÃO

Os sequenciamentos de nova geração referem-se às tecnologias de sequenciamento de DNA que surgiram após o método de Sanger. Essas novas tecnologias permitiram a análise massiva e paralela de um número muito maior de amostras a um menor custo e tempo, quando comparado com este método. As tecnologias ditas de segunda geração (Ion Torrent e Illumina) caracterizam-se pela necessidade de amplificar o material genético, via reação em cadeia da polimerase (PCR), antes do sequenciamento em si. Já as tecnologias de terceira geração (PacBio, NanoPore) caracterizam-se pelo sequenciamento de uma única molécula de DNA. Detalhes associados às diferentes tecnologias de sequenciamento, incluindo a construção de bibliotecas e aspectos inerentes de cada plataforma, já foram extensivamente revisados (BUERMANS; DEN DUNNEN, 2014; CAO et al., 2017; KULSKI, 2016; LEVY; MYERS, 2016)

No tocante às plataformas de sequenciamento de segunda geração utilizadas neste trabalho, IonTorrent e Illumina, destacam-se a presença de fragmentos aleatórios do genoma, sem a necessidade de cloná-los em uma célula hospedeira, tal como ocorria nos sequenciamentos do tipo Sanger. Nesse caso, sequências adaptadoras são conectadas às extremidades dos fragmentos e ligadas de forma isolada a superfícies sólidas (Illumina) ou *beads* (IonTorrent), para amplificação via PCR. Na reação de sequenciamento em si, a incorporação de bases é monitorada pela detecção de luminescência (Illumina) ou alterações na concentração de íons H^+ (IonTorrent). No primeiro caso, os nucleotídeos modificados, inseridos no fragmento pela polimerase, são submetidos a reações capazes de gerar fluorescência específica, possibilitando a identificação da base inserida pelos fotodetectores (**Figura 1.2**). No segundo caso, os nucleotídeos são adicionados à reação segundo uma ordem pré-estabelecida e induzem mudanças no pH, quando incorporados aos diversos fragmentos, capazes de serem percebidas por detectores eletroquímicos (**Figura 1.3**).

Em relação aos tipos de bibliotecas de sequenciamento existentes relacionadas aos tipos de *reads* geradas, destacam-se as estratégias *paired-end* e *mate-pair* (ECSEQ, 2017). No primeiro caso, em Illumina, o fragmento anexado à superfície sólida é sequenciado em ambas as extremidades, gerando as chamadas *reads forward* e *reads reverse*, cujo somatório dos tamanhos é, em geral, inferior ao tamanho dos fragmentos sequenciados. Uma vez que no processo de construção da biblioteca executa-se uma etapa de seleção de tamanhos dos fragmentos, essa estratégia permite a estimativa da distância entre as *reads* obtidas. A informação de pareamento e de distância entre as *reads* facilita os processos de montagem, por contribuir na resolução de regiões repetitivas. Nas bibliotecas *mate-pair* também se obtém *reads forward e reverse.* No entanto, utilizam-se diferentes estratégias com o propósito de obter fragmentos cujo conteúdo a ser sequenciado contém informações de maior distanciamento entre as *reads*, quando comparado aos fragmentos gerados pela estratégia *paired-end*. Essa última estratégia é de grande-valia para resolução de regiões repetitivas, cujo tamanho seja superior às distâncias geralmente obtidas em sequenciamentos *paired-end* (TREANGEN; SALZBERG, 2011).

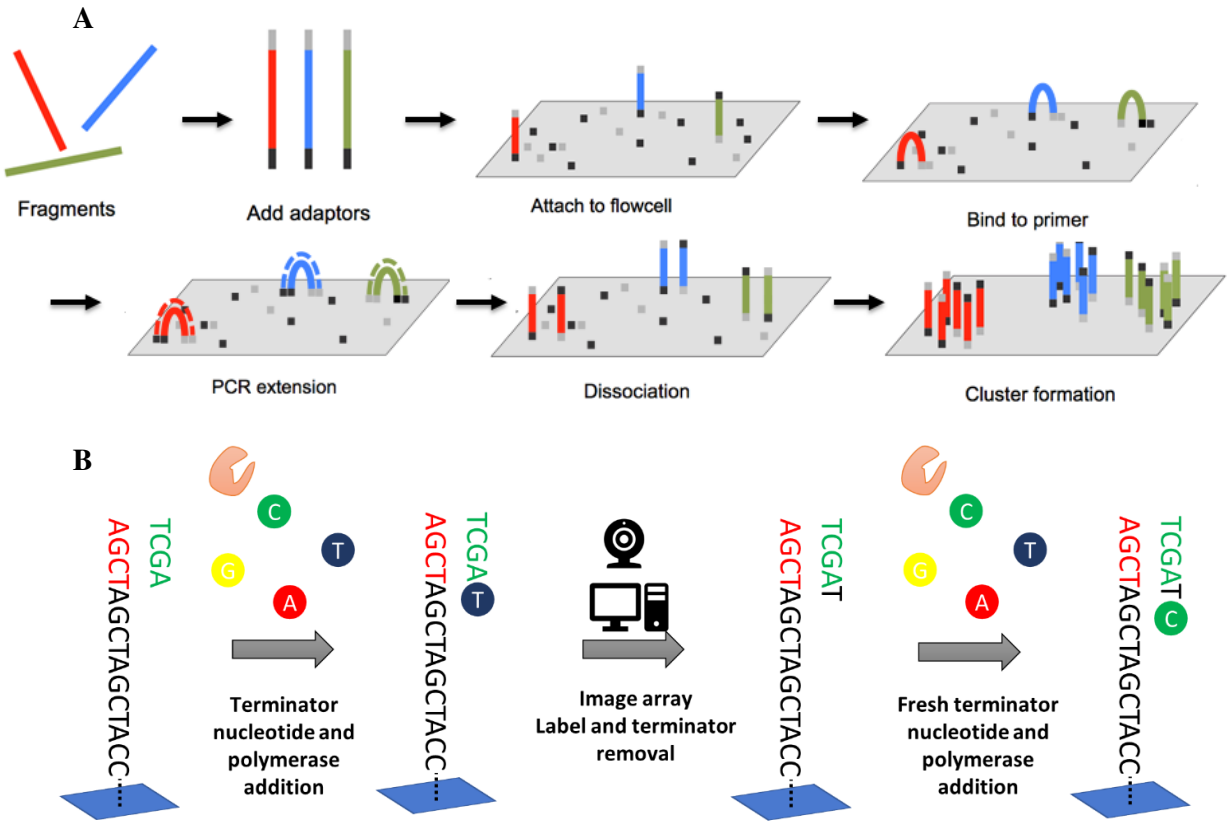


Figura 1.2 Representação esquemática de um sequenciamento na plataforma Illumina. A) A construção da biblioteca envolve a ligação de adaptadores aos fragmentos genômicos, ligação a uma superfície sólida e amplificação dos mesmos via *Bridge Amplification*. Adaptado de (LU et al., 2016). B) Após formação dos *clusters*, os nucleotídeos modificados são adicionados juntamente com DNA polimerases. Após reação de alteração da base adicionada, fluorescência é emitida pelo *cluster* e identificada por fotodetectores específicos. Adaptado de (WILANTHO et al., 2012).

Idealmente, *reads* obtidas de sequenciamento de nova geração deveriam se distribuir uniformemente no genoma. Esse fato, no entanto, está longe de ser uma realidade. A construção de bibliotecas afeta consideravelmente no surgimento de vieses de sequenciamento. Os níveis de cobertura parecem ser menos enviesados em sequenciamentos PacBio, seguido de Illumina, ainda que todas as plataformas incluam alguma espécie de viés associados a regiões de alto ou baixo conteúdo de guanina e citosina, ou mesmo em regiões homopoliméricas (ROSS et al., 2013).

Lidar com vieses de sequenciamento é de suma importância, uma vez que podem afetar os processos de montagem e a descoberta de variantes. Preferencialmente, a melhor opção para lidar com esses efeitos seria utilizar diferentes plataformas de sequenciamento, cujos vieses fossem diferentes entre si (ROSS et al., 2013). Assim sendo, potenciais erros vinculados às *reads*, provenientes de uma determinada plataforma, seriam diluídos por *reads* advindas de outra. Muito embora essa seja uma estratégia útil para contornar esses vieses, genomas

relativamente pequenos podem ser, em geral, montados e estudados com razoável grau de precisão a partir de *reads* provenientes de sequenciamentos advindos de uma única plataforma.

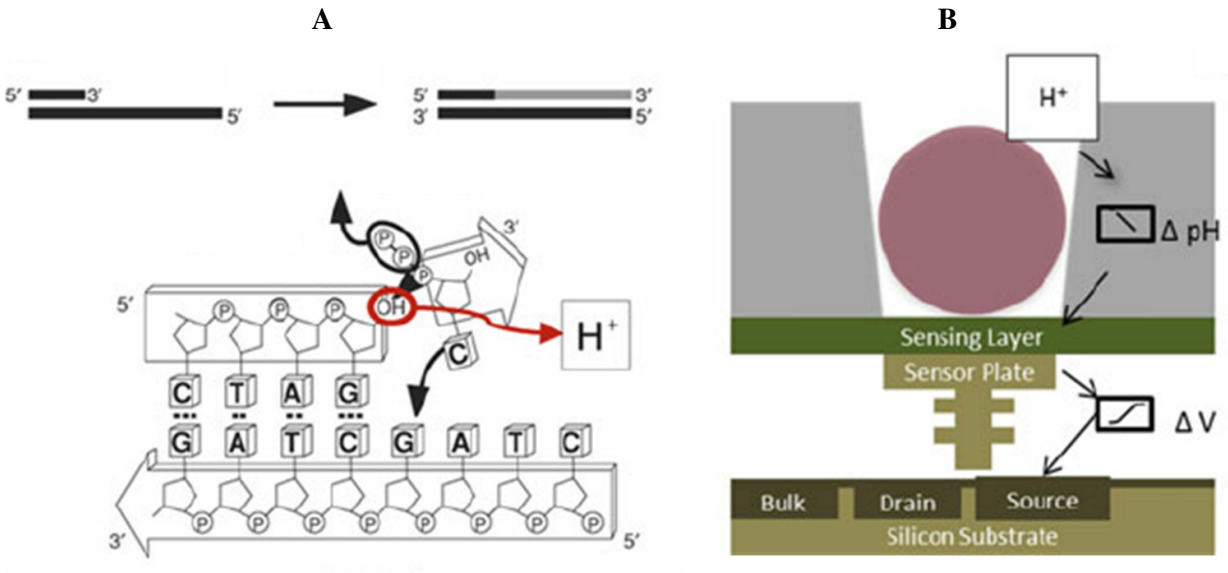


Figura 1.3 Representação esquemática de sequenciamento IonTorrent. A) Reação de sequenciamento evidenciando a liberação de íons H^+, quando a base correta é adicionada. B) Detector eletroquímico capaz de associar a variação no potencial hidrogeniônico a um potencial elétrico capaz de ser interpretado por softwares específicos, destinados à correta predição do tipo e do número de nucleotídeos adicionados (*base calling*). Adaptado de: (KOHN et al., 2013)

Erros de sequenciamento que contribuam para presença de *mismatches* e, principalmente, *indels* no genoma montado podem induzir falsas conclusões acerca da composição genética (FANG et al., 2014; MU et al., 2016). Nesse caso, por exemplo, nucleotídeos erroneamente assinalados podem induzir à predição de uma sequência polipeptídica não condizente com a realidade. Falsos *mismatches* podem ser responsáveis pela predição de pseudogenes decorrentes da presença de códons de parada prematuros, ou, ainda, modificações na estrutura polipeptídica devido alterações de aminoácidos. Falsos *indels* são ainda mais prejudiciais às análises de bioinformática, pois induzem errôneas predições da fase de leitura. Assim, sequências codificantes (CDSs) que naturalmente podem ser funcionais são preditas como potencialmente não-funcionais e vice-versa (MEADER et al., 2010; PHILLIPPY; SCHATZ; POP, 2008; SALZBERG; YORKE, 2005)

Muito embora erros associados à presença de falsos *mismatches* e *indels* possam ser observados nos genomas bacterianos completos, esses fatos são relativamente raros e sua frequência é altamente dependente da plataforma de sequenciamento utilizada. ROSS *et al* (2013) estimam em *E. coli* taxas de erros em relação à cobertura e vinculadas a *mismatches*,

deleções e inserções, respectivamente, de 0,36%, 0,00097% e 0,00051%, para Illumina, e 0,18%, 0,53% e 0,54%, para IonTorrent. Essas observações estão intrinsicamente associadas a artefatos vinculados a essas tecnologias. Por exemplo, as maiores taxas de erros do tipo *indel*, vinculadas ao sequenciamento Torrent, estão diretamente associadas às regiões homopoliméricas.

As taxas de erros observadas, no entanto, podem variar dependendo de uma série de fatores, tais como conteúdo GC; construção da biblioteca; erros associados à DNA polimerase, na etapa de amplificação do material genético; estratégia de montagem e parâmetros utilizados; cobertura do sequenciamento; presença de variantes naturais na população de células sequenciadas, entre outros (ALLHOFF et al., 2013; ROBASKY; LEWIS; CHURCH, 2014). Para todos os efeitos, releva-se a importância das etapas de curadoria, responsáveis por identificar e corrigir a presença de falsos *mismatches* e, principalmente, falsos *indels* no genoma montado.

1.4.2 Pré-Processamento

Na etapa de pré-processamento dedica-se à avaliação da qualidade das *reads* geradas, atentando-se para a correção e/ou remoção de *reads* errôneas. Em geral, observa-se a distribuição da qualidade associada a cada base das *reads*, que serve de parâmetro para julgar a qualidade do sequenciamento (TRIVEDI et al., 2014). Essa qualidade é assinalada pelos valores de phred. Nesse sistema, representa-se em uma escala logarítmica a probabilidade de determinado nucleotídeo da *read* ter sido erroneamente assinalado (STAWINSKI et al., 2016).

Essa informação pode ser aplicada para filtragem de *reads*. Nesse caso, o usuário pode optar por removê-las por completo, caso não satisfaçam um determinado critério de qualidade, ou mesmo descartar as porções de baixa qualidade das mesmas (BOUCHOT et al., 2014). O último caso é denominado “trimagem”, que também pode ser realizada a partir da remoção de um tamanho fixo de uma ou ambas as extremidades da *read* (BOLGER; LOHSE; USADEL, 2014; STAWINSKI et al., 2016). Outro aspecto relevante está relacionado à remoção de sequencias de adaptadores, completos ou não, das *reads* geradas e cuja permanência pode afetar consideravelmente as análises subsequentes (SCHUBERT; LINDGREEN; ORLANDO, 2016).

1.4.3 MONTAGEM

Na etapa de montagem objetiva-se a construção de *contigs*, a partir das *reads* de sequenciamento. Nesse caso, a obtenção desses fragmentos pode ocorrer a partir de duas abordagens. Uma delas consiste no mapeamento de *reads* a um genoma de referência e a conseguinte geração de uma sequência consenso. Essa abordagem é denominada montagem por referência e resulta em boas montagens, quando o organismo de estudo é relativamente próximo do organismo utilizado como modelo (LISCHER; SHIMIZU, 2017). Apesar dessa abordagem ser de mais fácil implementação, ela muitas vezes induz, no genoma montado, vieses do genoma de referência, fruto de erros de montagem, ou mesmo a não inserção de regiões exclusivas do organismo de estudo (SCHNEEBERGER et al., 2011). A segunda abordagem consiste na geração de *contigs* a partir da sobreposição de *reads* em um processo denominado montagem *de novo* ou *ab initio*. Existem atualmente diversos softwares dedicados à resolução desse problema e que empregam diferentes algoritmos destinados a identificar, corrigir e eliminar erros de sequenciamentos. Esses softwares diferem entre si, principalmente, na acurácia, eficiência e qualidade da montagem, além da capacidade de resolver regiões repetitivas e discriminar erros de sequenciamento de variantes reais na amostra (ZHANG et al., 2011).

Em uma recente avaliação dos softwares de montagem mais comumente utilizados, Brandam *et al* (2013) recomendam a observação de resultados de mais de um montador e não confiar cegamente em uma única métrica de qualidade ou acurácia de montagem. Os autores sugerem, ainda, que a escolha do montador ideal deva ser referente àquele que melhor se sobressaia no campo de interesse e que melhor lide com as características inerentes de cada sequenciamento, tais como cobertura e vieses.

Apesar do elevado número de softwares de montagem disponíveis, a maioria se enquadra em uma das seguintes estratégias (LI et al., 2012): 1) *Overlap Layout Consensus* (OLC); 2) Grafos de Brujin (DBJ). Nas estratégias baseadas em OLC as *reads* são alinhadas entre si e a qualidade do alinhamento, calculada. Um grafo de sobreposição é então construído, onde os vértices correspondem às *reads* e as arestas indicam a qualidade dos alinhamentos. A resolução do *contig* é dada pela determinação do consenso desses alinhamentos, indicado pelo caminhamento Hamiltoniano (*i.e.,* onde cada vértice deve ser visitado apenas uma única vez) ótimo no grafo de sobreposição (**Figura 1.4**).

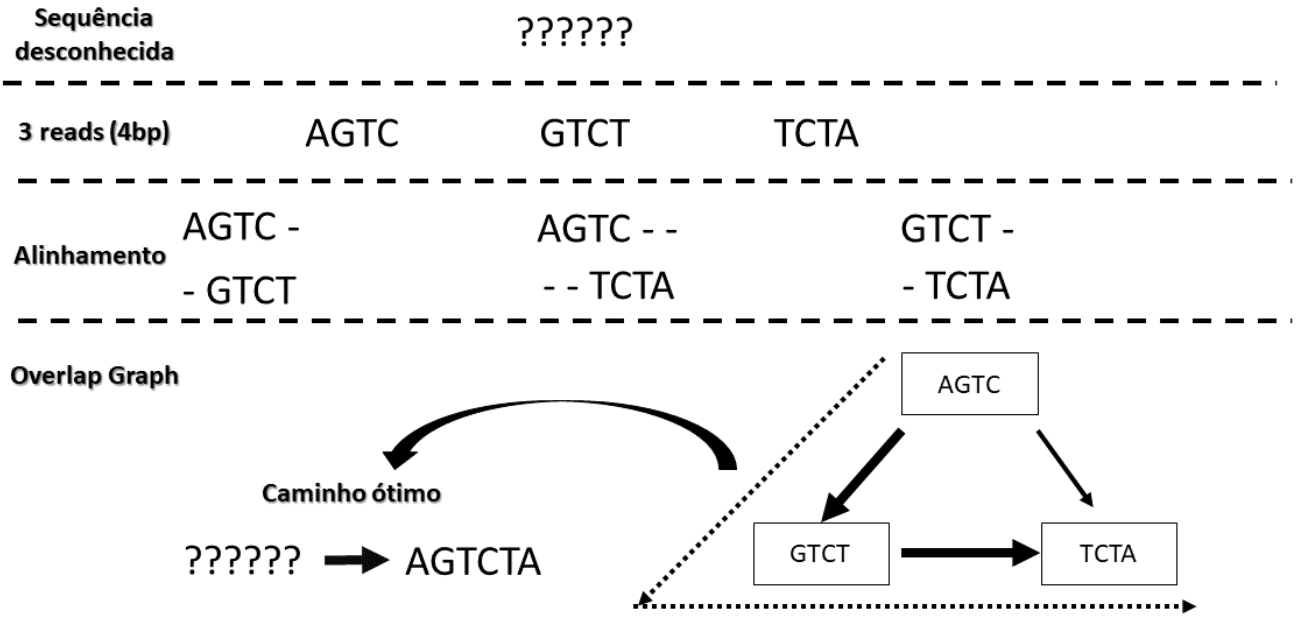


Figura 1.4 Exemplo de algoritmo de montagem utilizando a estratégia de *Overlap Layout Consensus* (OLC). Nesse caso, as *reads* referentes à sequência desconhecida são alinhadas entre si para construção de um grafo de sobreposição, onde os vértices, representando as *reads*, são conectados por arestas direcionadas, cujos pesos refletem o *score* dos alinhamentos. A resolução da sequência desconhecida se dá através do descobrimento do caminho ótimo no grafo.

Já os algoritmos baseados nos grafos de Brujin iniciam o processo de montagem pela segmentação das *reads* em fragmentos de tamanho k denominadas *k-mers*. Na construção do grafo todos os pares de *k-mers* cujas extremidades opostas tiverem k – 1 bases sobrepostas são conectadas por arestas direcionadas. A resolução do *contig* é obtida pelo caminhamento Euclidiano nesse grafo (*i.e.,* onde cada aresta só pode ser visitada uma única vez). A **Figura 1.5** ilustra um exemplo de montagem utilizando essa estratégia e indica a importância da escolha do parâmetro k.

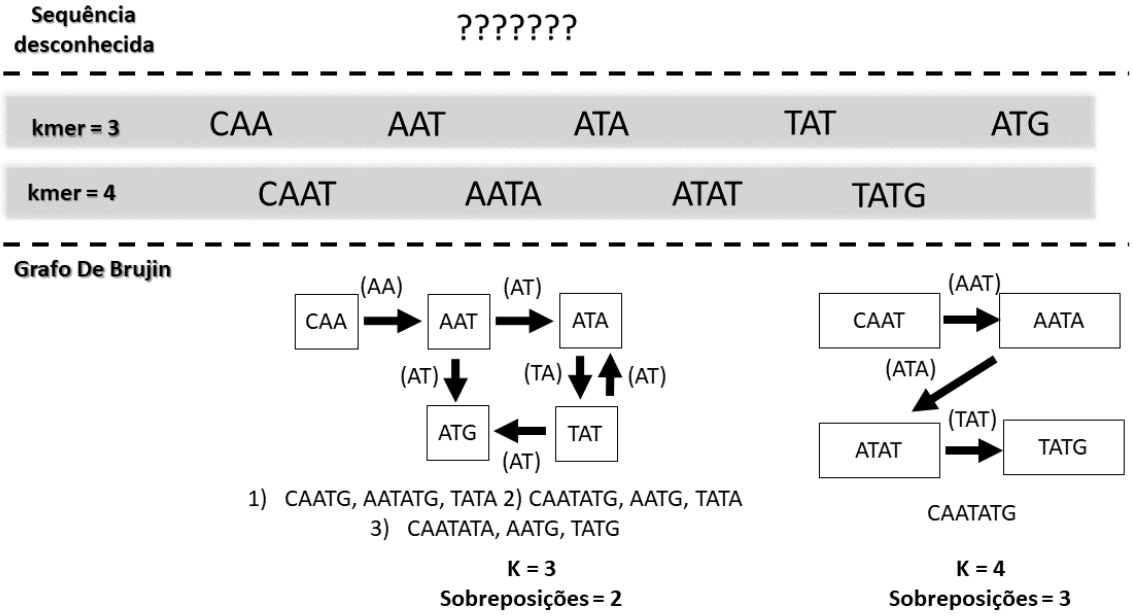


Figura 1.5 Exemplo de algoritmo de montagem utilizando grafos de Brujin. Nesse exemplo a sequência desconhecida foi fragmentada em *kmers* de tamanho 3 e 4. No primeiro caso, cada nó do grafo é conectado por arestas direcionadas se houver sobreposição de 2 pb nas extremidades direita e esquerda dos *kmers*. Analogamente, no segundo caso, os vértices do grafo são conectados se a sobreposição for de 3 pb. Notar que o tamanho do kmer pode afetar a complexidade do grafo, o que pode dificultar a determinação da sequência.

Cada uma dessas estratégias apresenta, de forma geral, pontos positivos e negativos que devem ser considerados de acordo com os tipos de dados trabalhados e dos recursos computacionais disponíveis (**Quadro 1.1**).

Quadro 1.1 Características gerais das estratégias de montagem via grafos de Brujin (DBJ) e *Overlap Layout Consensus* (OLC)

Grafos de Brujin (DBJ)	***Overlap Layout Consensus* (OLC)**
Mais sensíveis a repetições e erros de sequenciamento	Menos sensível a repetições e erros de sequenciamento
Convergência do grafo em regiões de repetições de tamanho k	Maior demanda por recursos computacionais
Erros nas *reads* induzem a formação de vértices espúrios	Não há escalonamento para um volume muito grande de *reads*
Parâmetros relevantes: tamanho do *k-mer*, *cutoff* de cobertura	Parâmetros relevantes: comprimento mínimo do alinhamento; identidade do alinhamento

Fonte: (SEEMANN, 2010)

1.4.4 *SCAFFOLDING*

Devido à natureza repetitiva de algumas regiões do genoma, a utilização dos softwares de montagem dificilmente leva à construção de uma sequência única, abrangendo todas as *reads* obtidas no sequenciamento. Esse fato ocorre porque essas regiões repetitivas elevaram o grau de complexidade dos grafos de montagem, dificultando a correta predição das diferentes regiões únicas em relação às regiões repetitivas (**Figura 1.6**). Por conseguinte, os montadores acabam por gerar diferentes fragmentos do genoma completo (TREANGEN; SALZBERG, 2011),

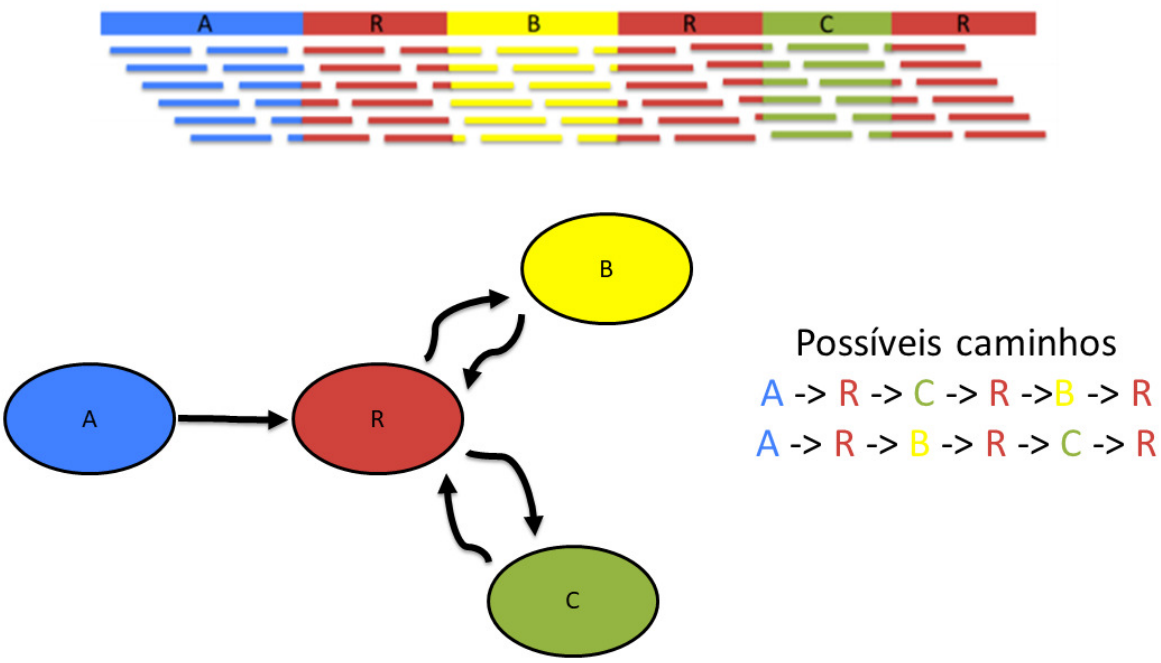


Figura 1.6 Representação esquemática do problema das repetições. A região genômica, contendo sítios únicos (A, B e C) e repetitivos (R) é fragmentada em diferentes *reads* para montagem. Quando os grafos de montagem são construídos e simplificados, o montador concluí que mais de uma possibilidade de caminhamento ótimo é possível. Por conseguinte, o mesmo opta por apresentar cada fragmento, separadamente (*contigs*).

A geração desses *contigs* pode variar dependendo da complexidade do organismo em estudo, da qualidade do sequenciamento, das estratégias de montagem, e dos softwares e parâmetros utilizados. No tocante à complexidade do organismo, de forma geral aqueles que apresentam um menor tamanho e que possuam um menor número de regiões repetitivas, tendem a gerar um menor número *de contigs* (KOREN et al., 2013; TREANGEN; SALZBERG, 2011; ZIMIN et al., 2017). Um exemplo desse fenômeno pode ser ilustrado com os genomas

de *E. coli* comparados com genomas de *Corynebacterium pseudotuberculosis*. O efeito da complexidade pode ser visto ao observar o maior nível de sintenia nos genomas de *C. pseudotuberculosis*, quando comparados com *E. coli* e que se deve à relativa ausência de regiões repetitivas do tipo elementos móveis (**Figura 1.7**). Nesse caso, não coincidentemente em *E. coli*, observa-se um maior número de *contigs* gerados em relação à *C. pseudotuberculosis* (> 100 em *E. coli,* < 15 em *C. pseudotuberculosis*).

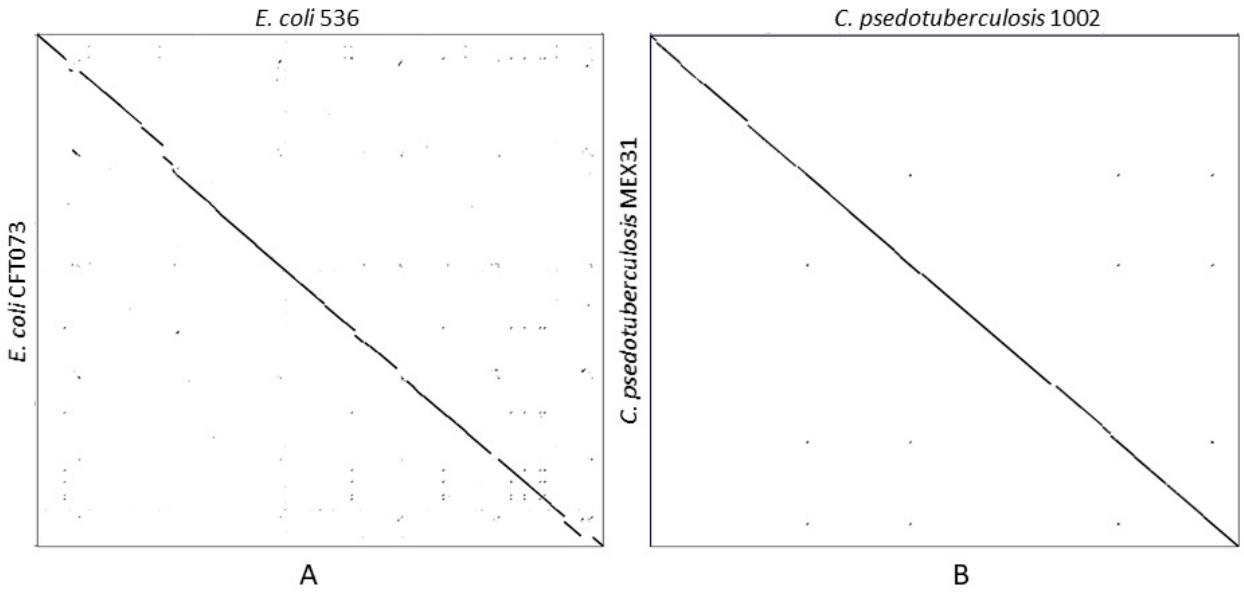


Figura 1.7 Gráficos de sintenia (A) entre duas linhagens de *E. coli* uropatogênicas e (B) entre dois biovares distintos de *C. pseudotuberculosis* (MEX31 – equi, 1002 – ovis).

A obtenção do genoma completo depende, portanto, do correto posicionamento dos *contigs* e da resolução das regiões repetitivas. Para tanto, faz-se o uso da chamada montagem híbrida, onde através de um genoma de referência pode-se inferir a correta disposição das regiões repetitivas (TAMAZIAN et al., 2016). Nesse caso, busca-se a melhor orientação dos diferentes *contigs* e a resolução dos *gaps* entre eles, mediante montagem local.

1.4.5 ANOTAÇÃO

Nessa etapa objetiva-se inferir em uma escala genômica aspectos fenotípicos do organismo em estudo através da anotação estrutural e funcional do genoma. No primeiro caso, busca-se a identificação de CDSs, peptídeos sinais, elementos repetitivos, tRNAs, rRNAs, sequências de prófagos, entre outras. Já no segundo caso, objetiva-se a compreender a função desses elementos.

A etapa de anotação em geral consiste na utilização de softwares de anotação automática. Alguns destes o fazem mediante a transferência de informações de uma espécie filogeneticamente próxima para o genoma de estudo. Tal abordagem, no entanto, pode induzir vieses uma vez que regiões exclusivas do genoma de interesse podem não ser caracterizadas (RICHARDSON; WATSON, 2013).

De modo geral, as ferramentas de anotação contam com alguma estratégia de predição de CDSs seguida do alinhamento local com as sequencias proteicas preditas em um genoma de referência ou com bases de dados especificas. As anotações dos melhores *hits* são então transferidas para as CDSs preditas, quando possível. Os diferentes softwares existentes realizam esses procedimentos somados a algumas variações associadas à correção de erros e adição de novas informações (RICHARDSON; WATSON, 2013)

1.4.6 CURADORIA

A fase de curadoria talvez seja a mais onerosa. Softwares de anotação automática podem introduzir e propagar anotações espúrias. Disso posto, propõe-se, nessa etapa, avaliar a anotação inicial de modo a encontrar discrepâncias na sequência e corrigir, por exemplo, anotações incorretas, *frameshifts* e falsos pseudogenes. Também, é nesse momento onde adequa-se a nomenclatura de genes e produtos às normas dos bancos de dados de submissão, tal como o NCBI (NCBI, 2017). Para maiores detalhes sobre esse tópico consultar Richard e Watson (2013).

1.5 Adventos da genômica comparativa

Toda informação hereditária e funcional de um organismo está contida no seu genoma. A rápida expansão no número de sequências nucleotídicas depositadas nos bancos públicos nos últimos anos é prova do impacto que as tecnologias NGS têm provocado em estudos biológicos. O termo genômica foi primeiro utilizado por Thomas Roderick em 1986 para se referir à análise estrutural e funcional de genes. O mesmo autor também fala em genômica comparativa como instrumento de avaliação dos relacionamentos evolutivos entre diferentes indivíduos (KUSKA, 1998).

Análises da estrutura organizacional detalhada de um genoma somente são possívis após sua montagem e anotação. Graças a isso, diversas conclusões têm sido elaboradas, possibilitando a compreensão de diversos aspectos fisiológicos em fungos (GALAGAN et al.,

2005), algas (GROSSMAN, 2007) e humanos (HANDSAKER et al., 2015), por exemplo. Em bactérias rotineiramente utiliza-se da genômica comparativa, como forma de compreender as razões de algumas linhagens serem patogênicas e outras não (PRENTICE, 2004). Também, faz-se uso dessa estratégia para elucidação de mecanismos de virulência (VIANA et al., 2017) ou mesmo de disseminação de resistência a antimicrobianos (HOLT et al., 2012).

Os trabalhos com genômica têm se tornado mais fáceis, rápidos e baratos frente às melhorias nas tecnologias de sequenciamento e à disposição de um número, cada vez mais crescente, de genomas sequenciados capazes de servir de referência para outros. Graças às tecnologias NGS, atualmente tem-se desenvolvido o campo da genômica personalizada (GAWAD; KOH; QUAKE, 2016; REHM, 2017). Nessa área, objetiva-se a caracterização de indivíduos numa população e a identificação de variantes tais como *Single Nucleotide Polymorphisms* (SNPs), *indels*, alterações no número de cópias e variações estruturais.

A genômica comparativa tem sido de grande valia para descoberta de alvos vacinais mais específicos (GUIMARÃES et al., 2015; SEIB; ZHAO; RAPPUOLI, 2012), tem possibilitado a compreensão dos mecanismos associados à adaptabilidade (ZHANG et al., 2014) e, principalmente, tem servido de embasamento para predições fenotípicas e o conseguinte direcionamento de análises experimentais.

JUSTIFICATIVA

A caracterização genômica da *E. coli* BH100 justifica-se como forma de elucidar, sob um contexto genético, as mais diversas observações fenotípicas contatadas desde a data de seu isolamento. Além disso, por estar estocada à temperatura ambiente desde 1973 essa linhagem faz-se importante objeto de estudo para a compreensão de eventuais mudanças na estrutura genômica dessa bactéria.

A descrição genética da *E. coli* BH100 também permite somar informações ao vasto campo de estudo envolvendo bactérias multirresistentes. Tal fato faz-se importante, uma vez que a investigação extensiva e holística desses microrganismos contribui para o desenvolvimento de tecnologias e políticas públicas que visem conter a rápida expansão desses microrganismos.

Somente as infecções do trato urinário contabilizam elevados custos à saúde pública. Estima-se que cerca de 150 milhões de pessoas desenvolvam casos de ITU por ano (FLORES-MIRELES et al., 2015), sendo a *E. coli* o principal agente etiológico (FOXMAN, 2014). Nessa linha, os estudos com a *E. coli* BH100 também visam somar informações genéticas em um contexto de patogenicidade. Isso pois essa bactéria, por ter sido isolada de urina, é presumivelmente uma uropatogênica. Com efeito, estudos envolvendo a *E. coli* BH100 podem contribuir com dados que ajudem na compreensão dos mecanismos e disseminação da uropatogênese.

Do ponto de vista da Bioinformática, a compreensão completa envolvendo a caracterização de regiões codificantes, regiões móveis e de ilhas de patogenicidade somente faz-se possível mediante a determinação do genoma completo. Ainda que estudos possam ser realizados com genomas não completos (*drafts*), análises comparativas envolvendo a *E. coli* BH100 seriam mais significativas se realizadas com as sequências completas dos cromossomos e plasmídeos, haja vista a necessidade de caracterizar funcional e quantitativamente os elementos móveis presentes nessa bactéria. Disso posto, justifica-se aqui o esforço, *in silico*, de se obter o genoma completo dessa linhagem.

OBJETIVOS

O presente trabalho possui três principais objetivos: 1) descrever o genoma completo da *E. coli* BH100, como também de suas variantes e contrastar sob um contexto genômico os achados fenotípicos descritos nessa linhagem; 2) predizer potencial influencia dos plasmídeos nas taxas mutacionais da bactéria; 3) identificar a origem genética das alterações estruturais observadas nas variantes da *E. coli* BH100.

Objetivos específicos

- Realizar montagem e anotação funcional dos diferentes *replicons* da *E. coli* BH100 e suas variantes;
- Comparar marcadores cromossomicos da *E. coli* BH100 com linhagens de *E. coli* uropatogênicas;
- Identificar, sob um contexto filogenético, a similaridade dos cromossomos da *E. coli* BH100 e suas variantes com outras linhagens de *E. coli*.
- Identificar os tipos e posição dos elementos móveis presentes nos cromossomos e plasmídeos da *E. coli* BH100;
- Identificar sob um contexto filogenético a similaridade do plasmídeo de 60 MDa e 10 MDa com outros plasmídeos de *E. coli*;
- Descrever os principais marcadores genéticos, com ênfase à resistência a drogas, anotados no plasmídeo de 60 MDa;
- Descrever os principais marcadores genéticos, com ênfase à resistência a drogas, anotados no plasmídeo de 10 MDa;
- Propor possível mecanismo de mobilização do plasmídeo de 10 MDa;
- Comparar os diferentes plasmídeos identificados neste trabalho, observando semelhanças e diferenças na sintenia e conteúdo gênicos;
- Descrever mecanismos que expliquem os desvios fenotípicos esperados em experimentos de conjugação envolvendo plasmídeos da *E. coli* BH100;
- Propor mecanismo para explicar o surgimento de mutantes de alta frequência resistentes à estreptomicina na *E. coli* BH100.

CAPÍTULO 2 A OBTENÇÃO DO GENOMA COMPLETO DA *E. COLI* BH100 E SUAS VARIANTES

2.1 MATERIAL E MÉTODOS

2.1.1 Linhagens bacterianas

Em 2014 foram cedidas pelo Laboratório de Genética de Microrganismos (LGM) alíquotas da *Escherichia coli* BH100, juntamente com variantes[7], obtidos a partir da cura dos plasmídeos, aqui denominados *E. coli* BH100L, *E. coli* BH100S e *E. coli* BH100N referentes às linhagens contendo apenas o plasmídeo de 60 MDa, apenas o plasmídeo de 10 MDa e nenhum dos plasmídeos, respectivamente. Em 2017, foram cedidas pelo mesmo grupo a *E. coli* BH100, *E. coli* BH100L, *E. coli* BH100N e uma *E. coli* K-12 transconjugante, contendo o plasmídeo de 10 MDa da *E. coli* BH100[8] (**Quadro 2.1**)

Quadro 2.1 Caracteristicas da *E. coli* BH100 utilizadas nesse trabalho e suas marcadores de resistência.

Linhagem	Marcadores de resistência	Plasmídeos (MDa)
E. coli BH100 subtr. MG2014	HgKnCmTcSm[a] Ap	pBH100 (60) pAp (10)
E. coli BH100L subtr. MG2014	HgKnCmTcSm[b]	pBH100alpha (60)
E. coli BH100S subtr. MG2014	Ap	pAp (10)
E. coli BH100N subtr. MG2014[c]	-	-
E. coli BH100 subtr. MG2017	HgKnCmTcSm[a] Ap	pBH100 (60) pAp (10)
E. coli BH100L subtr. MG2017	HgKnCmTcSm[b]	pBH100 (60)
E. coli BH100N subtr. MG2017[c]	-	-
E. coli K12T subtr. MG2017	Ap	pAp (10)

a) resistência a baixos níveis laboratoriais (< 20,0 $\mu g.mL^{-1}$) de estreptomicina, mas com potencial de gerar mutantes resistentes. b) resistência a baixos níveis laboratoriais (< 20,0 $\mu g.mL^{-1}$) de estreptomicina. c) Linhagens estocadas a temperatura ambiente e em desuso desde 1990. Hg: Mercúrio inorgânico; Kn: Canamicina; Cm: Cloranfenicol; Tc: Tetraciclina; Sm: Estreptomicina; Ap: Ampicilina

[7] Como explicado no início do capítulo 1, as variantes da *E. coli* BH100 serão, ao longo deste trabalho, tratadas como linhagens derivadas da bactéria selvagem, e, portanto, os termos "variante" e "linhagem" serão tratados como sinônimos.

[8] As designações de cada variante ocorreram como segue: BH100L – associado ao termo em inglês, *Large*, para lembrar que essa linhagem contém apenas o maior plasmídeo; BH100S – associado ao termo em inglês, *Small*, por conter apenas o menor plasmídeo; BH100N – associado ao termo em inglês, *No,* para lembrar que essa linhagem não contém plasmídeos; K12T – para se referir a uma *E. coli* K-12 transconjugante.

2.1.2 Extração do DNA genômico

As subcepas MG2014 foram replicadas em 30 mL de meio BHI (HIMEDIA) a 37 ºC por 48h em tubos Falcon (50 mL) e centrifugados a 2.500 g por 15 min. O precipitado foi ressuspenso em 600,0 µL de Solução I (50% sacarose (SYNTH), 10 mM Tris-Cl (PROMEGA), 50 mM EDTA pH 8,0 (SYNTH)). À mistura foi adicionada *beads* VK01 e esta, em seguida, sujeita a um ciclo de 15 s a 5.000 rpm no sistema Precellys (BERTIN TECHNOLOGIES). Em seguida, completou-se o volume do tubo com solução de fenol (25 fenol: 24 clorofórmio: 1 álcool isoamílico (SIGMA), e o mesmo submetido à agitação em vortex e centrifugado a 12.000 g por 7 min. O sobrenadante foi coletado e os processos de adição de solução de fenol e centrifugação repetidos. Ao término da segunda centrifugação, ao sobrenadante foi adicionado 1,0 mL de clorofórmio (SYNTH), e a mistura agitada em vortex e centrifugada a 12.000 g por 7 min. A fase aquosa foi transferida para novo tubo contendo 2,5 volumes de etanol (SYNTH), 0,1 volume de acetato de sódio 3M (SYNTH) e 0,01 volume de glicogênio 20,0 $mg.mL^{-1}$ (USB). A mistura foi levada ao freezer -20 ºC *overnight* para precipitação e centrifugada a 12.000 g por 15 min. O precipitado foi lavado com etanol 70% e centrifugado a 12.000 g por 20 min. Para remoção de resquícios de etanol, os tubos foram deixados para secar em estufa a 37 ºC. O DNA foi ressuspenso em 40,0 µL de água MiliQ estéril.

As subcepas MG2017 foram crescidas em 5,0 mL de meio LB (10% Peptona (HYSOY), 5% extrato de levedura (KASVI), 5% NaCl (SYNTH)) a 37 ºC por 18h, em tubos Falcon (50 mL). Os meios de crescimento das linhagens BH100 e BH100L foram ajustados para 15,0 $µg.mL^{-1}$ de cloranfenicol (SIGMA), já para a linhagem K12T o meio foi ajustado para 50,0 $µg.mL^{-1}$ de ampicilina (SIGMA). Em seguida as culturas foram centrifugadas a 12.000 g por 5 min e o sobrenadante descartado. Ao precipitado foi adicionado 150,0 µL de Solução I juntamente com 10,0 µL de lisozima 10 $mg.mL^{-1}$ (SIGMA). A mistura foi incubada a 37 ºC por 20 min. Logo após, 50,0 µL de SDS 1,5% (SIGMA) foram adicionados e a mistura incubada em gelo por 10 min. Ao final desse período, foram adicionados 100,0 µL de Solução de Fenol e a mistura foi homogeneizada por inversão e centrifugada a 12.000 g por 5 min. O sobrenadante foi transferido para microtubos e a adição da Solução de Fenol, homogeneização e centrifugação repetidos. Após nova coleta do sobrenadante, foram adicionados à mistura 300,0 µL de isopropanol (SYNTH) gelado, homogeneizados por inversão e centrifugados a 12.000 g por 10 min. Descartou-se o sobrenadante e ao precipitado foram adicionados 500,0 µL de etanol 70%, sendo os tubos centrifugados à 7.500 g por 10 min. Após descarte do etanol e afim de eliminar seu excesso, os tubos foram deixados entreabertos em estufa a 37 ºC por 20

min. O precipitado contendo DNA foi ressuspenso em 20,0 µL de solução contendo RNAse 50 $\mu g.mL^{-1}$ (INVITROGEN).

Para ambas subcepas, a fim de averiguar a qualidade do DNA extraído, o mesmo foi submetido à eletroforese em gel de agarose 0,8% (NUTRISCIENCE) por 30 min à 100V e corado com 0,5 µL de Brometo de etídio 10,0 $mg.mL^{-1}$ (AMRESCO). A quantificação foi realizada utilizando espectrofotômetro do tipo Nanodrop (THERMOFISCHER).

2.1.3 Sequenciamento

Todas subcepas MG2014 foram sequenciadas na plataforma IonTorrent PGM™ e a construção das bibliotecas de fragmentos de 200 pb realizada a partir de 1,0 µg de DNA, utilizando Ion Xpress™ Plus Fragment Library Kit (THERMO FISCHER), conforme recomendações do fabricante.

A linhagem BH100 subtr. MG2014 também foi sequenciada utilizando bibliotecas *mate-pair* de 3 kb e 6 kb. Para tanto, a biblioteca foi construída a partir de 10,0 µg de DNA, utilizando 5500 SOLiD® Mate-Paired Library Kit, conforme recomendações técnicas da ThermoFischer.

Já as subcepas MG2017 foram sequenciadas na plataforma Illumina HiSeq, a partir de 0,1 µg de DNA, utilizando bibliotecas *pair-end* (2 x 150 bp) e fragmentos de aproximadamente 500 pb, conforme recomendações do fabricante.

2.1.4 Desenvolvimento e aquisição de *scripts*

Para obtenção dos diferentes resultados desta sessão, uma série de *scripts in-house* foram desenvolvidos, de modo a colaborar no tratamento de dados das diversas etapas executadas para obtenção do genoma completo da *E. coli* BH100 e suas variantes. Os *scripts* a seguir foram elaborados na linguagem Perl e testados em um ambiente Linux:

a) **xmlclip.pl**: esse *script* foi desenvolvido com a finalidade de remover resquícios de adaptadores anotados em um arquivo *xml*, criado a partir do programa **sff_extract**;
b) **fastqsplitter.pl**: desenvolvido com a finalidade de separar as *reads mate-pair forward* e *reverse* de um arquivo *fastq*, em arquivos distintos;

c) **fastqpairselector.pl**: esse *script* foi desenvolvido de modo a selecionar, a partir de um arquivo de *reads* filtradas *forward,* os respectivos pares de *reads reverse* presentes em outro arquivo *fastq*, e vice-versa.

Outros scripts utilizados no decorrer desta seção foram desenvolvidos por outros autores:

a) **fast_sample**: esse *script* permite o redimensionamento de arquivos *fastq*, além da filtragem de *reads*, segundo parâmetros delineados pelo usuário[9].
b) **moveDnaa.py**: *script* utilizado para reposicionar a sequência nucleotídica do genoma de interesse, de acordo com uma referência[10].
c) **sff_extract**: responsável por converter um arquivo *sff* em um arquivo *fastq*[11].
d) **ffn2pseudo.fasta36.pl:** desenvolvido com o propósito de transferir anotação de um genoma referência para um genoma de interesse[12].

2.1.5 Montagem do genoma

2.1.5.1 PRÉ-PROCESSAMENTO

Para obtenção do arquivo *fastq* utilizou-se a ferramenta **sff_extract** nos arquivos *sff* provenientes do sequenciamento IonTorrent. Para remoção de resquícios de adaptadores no arquivo *fastq* utilizou-se o *script in-house* **xmlclip.pl**. O arquivo de *reads* do sequenciamento *mate-pair* foi dividido em dois arquivos distintos contendo as *reads forward* e *reverse*, utilizando o *script in-house* **fastqsplitter.pl**.

Para observação da qualidade das *reads,* utilizou-se o programa FASTQC[13]. A partir dos valores de qualidade, procedeu-se como segue para os dados provenientes da linhagem BH100 subtr. MG2014: para as *reads mate-pair forward* utilizou-se o *script* **fast_sample**, para filtrar por *reads* com tamanhos de no mínimo de 50 pb e máximo de 250 pb; a fim de manter as correspondentes *reads reverse*, utilizou-se o *script in-house* **fastqpairselector.pl**; o novo arquivo de *reads mate-pair reverse* foi filtrado para *reads* de no mínimo 50 pb e no máximo 230 pb e os correspondentes pares no arquivo de *reads forward* foram selecionados.

---

[9] Aquacen. (2016). Fast_sample. Disponível em: https://github.com/aquacen/fast_sample
[10] Mariano, D. (2014). moveDNAa.py. Disponível em: https://github.com/dcbmariano/scripts/blob/master/moveDNAA.py
[11] Blanca, J. (2015). sff_extract. Disponível em: https://github.com/JoseBlanca/seq_crumbs
[12] Soares, S. (2014). ffn2pseudo.fasta36.pl. Disponível em: https://github.com/dcbmariano/scripts/blob/master/ffn2pseudo.fasta36.pl
[13] Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Disponível em: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Já as *reads* provenientes da biblioteca *fragment* foram filtradas para tamanhos mínimos de 30 pb e máximos de 250 pb. Em seguida, a fim de testar o melhor nível de cobertura para montagem, utilizou-se o programa **seqtk**[14], para amostrar 40% das *reads mate-pair* e 25%, 30%, 35%, 40%, 45% e 50% das *reads fragment*. Combinações desses conjuntos foram testados na montagem.

As demais *reads*, referentes às linhagens MG2017, não foram submetidas a nenhum processo de filtragem, devido à qualidade do sequenciamento e da estratégia de montagem utilizada.

2.1.5.2 MONTADORES

Os *contigs* da *E. coli* BH100 subtr. MG2014 foram construídos a partir das *reads* advindas da biblioteca *fragment* e *mate-pair* de IonTorrent, a partir montadores do tipo OLC Newbler (v. 2.9) e Mira (v. 4.9.1)[15] e, do tipo DBJ, Spades (v. 3.9.1) (BANKEVICH et al., 2012) e utilizando diferentes combinações de amostragens de *reads*. A avaliação da qualidade de cada montagem foi observada no QUAST (GUREVICH et al., 2013). Os *contigs* das demais subcepas MG2014 foram construídos utilizando-se apenas o Spades.

Já a montagem dos *contigs* da *E. coli* BH100 subtr. MG2017 foi realizada utilizando-se Spades e Edena (v. 3.131028) (HERNANDEZ et al., 2008). O melhor valor de *k-mer* foi estimado utilizando-se o KmerGenie (v. 1.7023) (CHIKHI; MEDVEDEV, 2014). Os *contigs* das demais subcepas MG2017 foram construídos utilizando-se apenas o software Spades, ajustando o parâmetro de *k-mer* para 121.

Na tentativa de montagem dos plasmídeos, utilizou-se o PlasmidSpades (ANTIPOV et al., 2016), sendo o parâmetro referente aos *k-mers* das MG2017 ajustado para 121 e das MG2014, submetidas às condições padrão da ferramenta. Os *contigs* gerados por essa abordagem foram investigados via BLASTn (ALTSCHUL et al., 1990), para observação de possíveis sequências plasmidiais.

2.1.5.3 *SCAFFOLDING*

[14] Li, H. (2012). Seqtk. Disponível em: https://github.com/lh3/seqtk
[15]Chevreux, B (2014). Mira. Disponível em: https://sourceforge.net/projects/mira-assembler/

A etapa de ordenação dos *scaffolds* criados no Newbler, a partir do uso de *reads fragment* e *mate-pair* da *E. coli* BH100 subtr. MG2014 foi realizada no software Contiguator (v. 2.7.3) (GALARDINI et al., 2011), utilizando-se o genoma da *E. coli* 536 (RefSeq: NC_008253.1) como referência, escolhido a partir de sua similaridade com os maiores *scaffolds*. O Contiguator constrói um *superscaffold* a partir do posicionamento dos diferentes fragmentos, de acordo com sua similaridade ao genoma de referência e insere entre os mesmos 100 N's, denominados *gaps*[16]. As sequências que não apresentam similaridade ou que poderiam ser posicionadas em mais de um local são adicionadas a um arquivo de *contigs* excluídos. A partir desse arquivo foram extraídos os *scaffolds* referentes ao plasmídeo, identificados via BLASTn, e submetidos à etapa de fechamento de gaps e refinamento.

As demais linhagens das subcepas MG2014 foram tratadas de forma similar, mas utilizando-se o genoma completo da *E. coli* BH100 subtr. MG2014 como referência.

Para os *contigs* da *E. coli* BH100 subtr. MG2017, utilizou-se o software SSPACE (v. 2.0) (BOETZER et al., 2011), que utiliza das informações de pareamento das *reads* a fim de posicionar os *contigs* de forma independente de referência. Essa ferramenta foi utilizada no seu módulo de extensão, que permite o preenchimento de *gaps* com as próprias *reads* do sequenciamento. A fim de posicionar os *supercontigs* montados por essa abordagem, os mesmos foram submetidos ao Contiguator utilizando-se como referência a *E. coli* 536. A partir do arquivo de *contigs* excluídos, realizou-se outro procedimento de *scaffolding* utilizando o plasmídeo NR1 (RefSeq: NC_009133.1), selecionado como referência a partir da constatação de sua semelhança com alguns dos *contigs* excluídos.

Para fechamentos dos *gaps* criados a partir do Contiguator, utilizou-se a ferramenta GFinisher (v. 1.4) (GUIZELINI et al., 2016) que busca encontrar sobreposições entre os flancos de cada *gap* com os diferentes *contigs*, construídos a partir de diferentes estratégias de montagem. Aqueles *gaps* não fechados por essa ferramenta foram finalizados de forma manual utilizando-se o CLC Genome Workbench (v. 7.0) (https://www.qiagenbioinformatics.com/), observando o consenso do mapeamento das *reads* com os seguintes genomas de referência, selecionados a partir da similaridade dos flancos do *gap* via BLASTn: *E. coli* 536, *E. coli* Ecol_542 (RefSeq: NZ_CP018970.1) e *E. coli* K-15KW01 (RefSeq: NZ_CP016358.1). O fechamento de *gaps* dos plasmídeos foi realizado da mesma forma, mas utilizando-se o plasmídeo NR1 como referência.

[16]O termo *gap* será utilizado no contexto de montagem para se referir a uma região desconhecida presente entre dois *contigs/scaffolds* de um mesmo replicon.

2.1.5.4 REFINAMENTO

Após fechamento de *gaps* procedeu-se às seguintes etapas de refinamento: 1) remontagem de *reads* não mapeadas; 2) avaliação de regiões de baixa cobertura; 3) avaliação de regiões de mapeamento espúrio; 4) circularização.

2.1.5.4.1 Remontagem de reads não mapeadas

Durante a etapa de fechamento de *gaps* algumas regiões podem ser erroneamente excluídas do genoma final, além do fato de *contigs* excluídos na etapa de *Scaffolding* poderem não estar representados. A fim de corrigir esse problema, *reads* de cada linhagem foram mapeadas nos respectivos genomas montados, utilizando-se a ferramenta de mapeamento do CLC Genome Workbench. A partir de um arquivo contendo *reads* não mapeadas, utilizou-se o Spades para remontagem de *contigs*. Para remoção de contaminantes, utilizou-se BLASTn contra base de dados do NCBI para remoção de *contigs* não pertencentes à Enterobacteriaceae. A partir de BLASTn desses *contigs* com os *superscaffolds* originais e/ou com os genomas montados, procurou-se por regiões de similaridade local que pudessem indicar as regiões de inserção dessas sequências. *Gaps* flanqueando as regiões inseridas foram criados e fechados manualmente, como descrito anteriormente.

2.1.5.4.2 Avaliação de regiões de baixa cobertura

Regiões de baixa cobertura podem ser artefatos da construção de biblioteca ou indicar montagens incorretas. A fim de analisar essas regiões, e buscando corrigir falhas de montagem, utilizou-se o CLC Genome Workbench com intuito de observar regiões contendo até oito *reads* mapeadas. Tais regiões foram desconstruídas manualmente e remontadas localmente, utilizando a estratégia aqui denominada *contig walking*. Nessa abordagem, utilizou-se do fato de alguns montadores deixarem resquícios de regiões repetitivas nas extremidades dos *contigs*. Dessa forma, foram utilizados *contigs* construídos no Spades para extensão e remontagem manual da região, a partir da sobreposição de pelo menos 120 pb. Esse processo foi repetido até o fechamento do gap construído. A **Figura 2.1** ilustra o procedimento de *contig walking.*

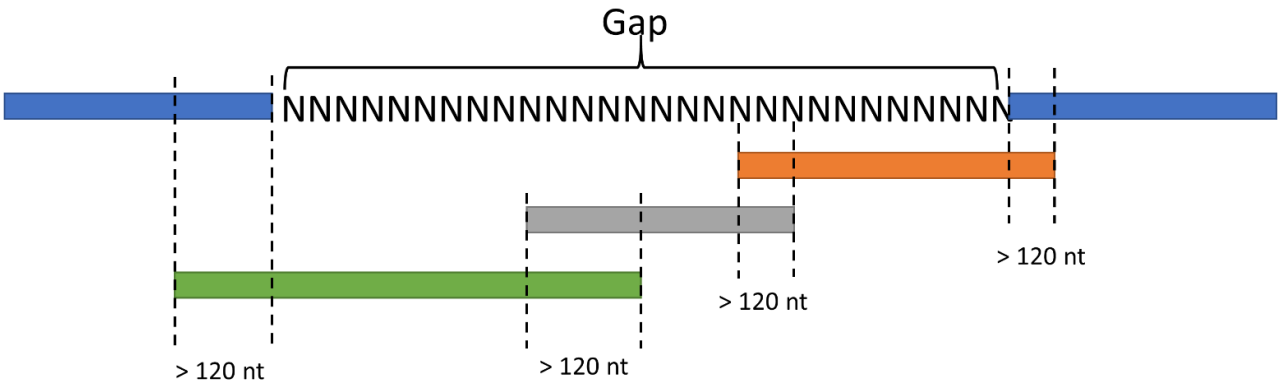


Figura 2.1 Representação da estratégia de *contig walking* para fechamento de *gaps*. Nessa abordagem, os flancos esquerdo e direito do gap (representados em azul) são unidos por *contigs* que apresentam, no mínimo, uma sobreposição de 120 pb.

2.1.5.4.3 Avaliação de regiões de mapeamento espúrio

Certas regiões podem possuir bons valores de profundidade de cobertura, mas que refletem uma baixa qualidade de mapeamento. Essas regiões compreendem *reads* no flanco esquerdo que apresentam alguma sobreposição com *reads* do flanco direito, mas com nenhuma mapeando completamente na região (**Figura 2.2**). Tal fato também sugere erros de montagem. Com efeito, esses sítios foram identificados a partir de um escaneamento manual do mapeamento, e as regiões desconstruídas e remontadas localmente, utilizando a estratégia de *contig walking*.

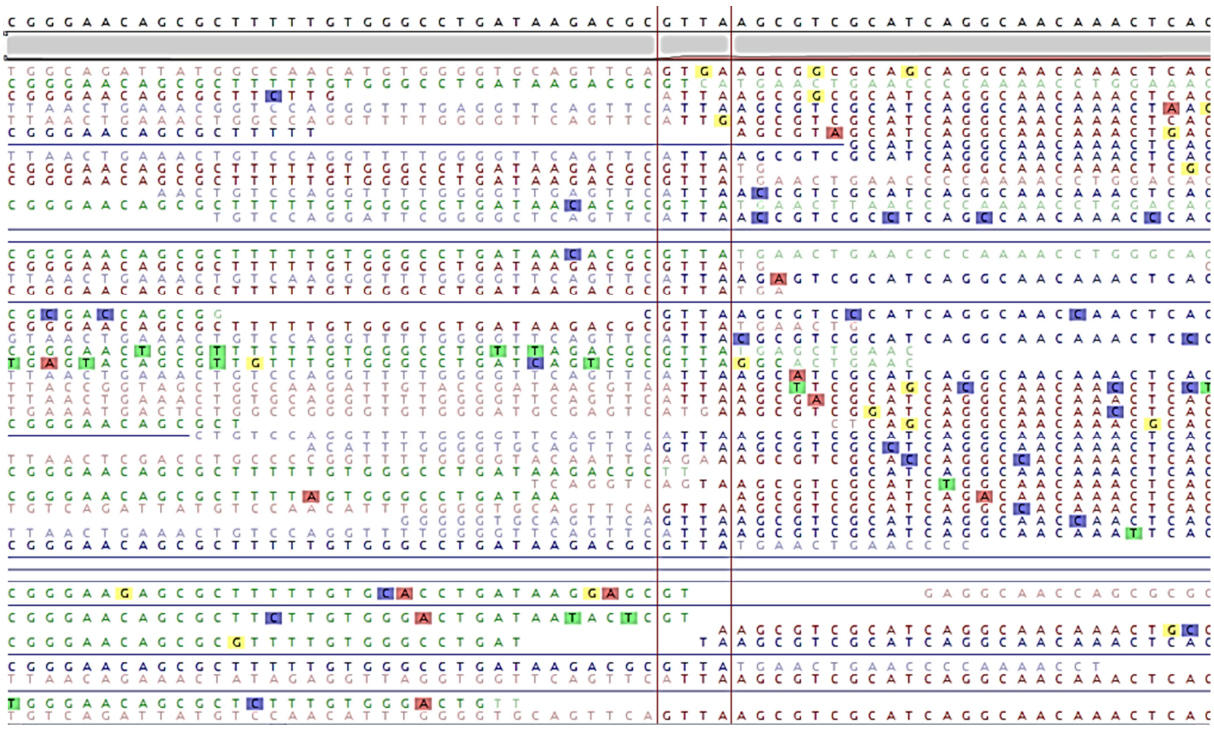


Figura 2.2 Exemplo de região de mapeamento espúrio. Nessa região, verifica-se que as *reads* provenientes do flanco esquerdo e do flanco direito apresentam sobreposição do *motif* GTTA. As regiões em tons mais claros referem-se à porção das *reads* não mapeadas na referência.

2.1.5.4.4 Circularização

Mediante o fato de a maioria dos replicons bacterianos serem circulares, procedeu-se a uma etapa de reorganização da sequência, a fim de mostrar a conexão entre as extremidades previamente montadas. Para tanto, segmentos de 1.000 pb foram excisados do início das sequências e adicionados ao final, mediante a inserção de um *gap* de 100 N's. O fechamento desse *gap* foi realizado utilizando *contigs* oriundos de outras montagens, via FGAP (PIRO et al., 2014) ou manualmente, a partir de comparações com uma referência.

2.1.6 Anotação e curadoria

Após finalização da montagem, as sequências foram anotadas, utilizando-se o *script* de transferência de anotação *in-house* **ffn2pseudo.pl**. Inicialmente, para a linhagem BH100 subtr. MG2014, transferiu-se a anotação contida na *E. coli* 536. As sequências também foram submetidas à ferramenta de anotação automática RAST (OVERBEEK et al., 2014). Uma vez anotados, os genomas foram visualizados e curados no Artemis (v. 16.0) (RUTHERFORD et al., 2000). As CDS anotadas pelo RAST e que se sobrepunham às CDS transferidas da *E. coli*

536 foram eliminadas, de modo a favorecer as anotações advindas do NCBI. Potenciais *frameshits* foram corrigidos, conforme já descrito (OLIVEIRA, 2014).

A fim de preparar o arquivo *gbk* para submissão ao banco de dados, faz-se necessária a criação de *features* do tipo gene, além da remoção e inserção de *qualifiers* obrigatórios, pelo NCBI. Por exemplo, deve-se remover os *qualifiers* referentes à cor e EC *number*, e adicionar locus_tag, protein_id e sequência de tradução. Com o intuito de automatizar e dar agilidade a esse procedimento, foi desenvolvido neste trabalho o programa GBKFinisher. Esse programa, além de preparar o arquivo *gbk* para posterior construção do arquivo *sqn*, também permite a classificação dos potenciais pseudogenes em *frameshifts*, códon de parada prematuro na CDS e fragmentos gênicos. O programa também permite a identificação de regiões contendo sobreposição de CDSs acima de um *treshold* especificado pelo usuário, além de colorir de forma diferencial CDSs contendo possíveis problemas de anotação, tornando mais cômodo o processo de refinamento da curadoria manual.

De posse do arquivo *gbk* finalizado, gerou-se os arquivos *tbl* e *fasta*, via Artemis. O arquivo *sbt* também foi gerado, segundo modelo do NCBI. De posse desses arquivos, utilizou-se o script **tbl2asn**[17] para geração do arquivo *sqn*. O arquivo de submissão *sqn* foi validado no **Sequin**[18]. Erros adicionais de anotação foram corrigidos no *gbk* original e o procedimento para geração do arquivo sqn, repetido.

Uma vez submetido e processado pelo NCBI, o genoma da *E. coli* BH100 subtr. MG2014 foi adicionado à base RefSeq, onde o genoma é reanotado, utilizando a ferramenta PGAP (TATUSOVA et al., 2016). De posse do genoma anotado no RefSeq, utilizou-se essa entrada como referência de anotação para as demais linhagens de *E. coli* deste trabalho, e todo o procedimento acima foi repetido.

[17]NCBI (2017). tbl2asn. Disponível em:ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/tbl2asn/
[18] NCBI (2017). Sequin. Disponível em: ftp://ftp.ncbi.nih.gov/sequin/

2.2 RESULTADOS E DISCUSSÃO

2.2.1 Obtenção dos replicons da *E. coli* BH100

Os diferentes replicons montados nesse projeto e suas respectivas estatísticas de montagem podem ser vistos na **Tabela 2.1**.

Tabela 2.1 Estatísticas de montagem dos genomas da *Escherichia coli* BH100 e suas variantes.

Genoma	Tecnologia	Contigs	N50	Replicon	Situação	Tamanho (bp)	Cobertura
BH100 subtr. MG2014	MP/F Torrent	11[a]	3.094.598	Cromossomo	C	5.072.848	151x
				pBH100-1	C	107.274	157x
				pAp	C	14.241	956x
BH100L subtr. MG2014	F Torrent	123	77.933	Cromossomo	D	-	-
				pBH100alpha	C	102.304	221x
BH100N subtr. MG2014	F Torrent	101	88.085	Cromossomo	D	-	-
BH100S subtr. MG2014	F Torrent	-	-	Cromossomo	-	-	-
BH100 subtr. MG2017	PE Illumina Hi-Seq	130	117.787	Cromossomo	C	5.131.212	439x
				pBH100-1	C	105.801	1.464x
				pApR	C	33.924	1.050x
BH100L subtr. MG2017	PE Illumina Hi-Seq	173	87.230	Cromossomo	C	5.033.217	368x
				pBH100alpha	C	103.103	1.408x
BH100N subtr. MG2017	PE Illumina Hi-Seq	113	141.477	Cromossomo	C	5.116.036	368x
K12T subtr. MG2017	PE Illumina Hi-Seq	-	-	Cromossomo	-	-	-
				pAp	C	14.247	3.811x

[a] Valor referente ao número de *scaffolds* criados pelo Newbler. MP: *Mate-pair*; F: *Fragment*; PE: *Pair-end*; C: Completo; D: Draft.

2.2.2 Qualidade do sequenciamento

A qualidade dos sequenciamentos foi realizada observando o histograma de qualidade das *reads* criados no programa FASTQC. As *reads fragment*, para as linhagens sequenciadas em 2014, apresentavam queda de qualidade após a posição 220 (em anexo: *E. coli* BH100 subtr. MG2014 – **Figura S1A**; *E. coli* BH100S subtr. MG2014 – **Figura** ; *E. coli* BH100L

subtr. MG2014 – **Figura S3**; *E. coli* BH100N subtr. MG2015 – **Figura S4**). Já a maior parte das *reads mate-pair* 3 kb (em anexo: **Figura S1B)** e 6kb (em anexo: **Figura S1C)** apresentam valor de phred superior a 25.

Já as *reads* advindas dos sequenciamentos Illumina resultaram em melhores valores de qualidade e um maior número de *reads.* A redução da qualidade observada nas *reads reverse* é esperada e é devido artefatos intrínsecos do método de sequenciamento (em anexo: *E. coli* BH100 subtr. MG2017 – **Figura** ; *E. coli* BH100L subtr. MG2017 – **Figura** ; *E. coli* BH100N subtr. MG2017 – **Figura** ; *E. coli* K12T subtr. MG2017 – **Figura**).

2.2.3 Pré-tratamentos

Todas as *reads* provenientes de sequenciamento IonTorrent foram compiladas em arquivos binários do tipo *sff.* Para obtenção do arquivo de *reads* no formato *fastq*, comumente utilizado para montagem, utilizou-se o *script* **sff_extract**. Esse *script* fornece, além do arquivo *fastq*, um arquivo *xml*, contendo informações acerca da posição de resquícios de adaptadores nas *reads.* Disso posto, elaborou-se o *script* **xmlclip.pl** com a finalidade de remover essas regiões de cada *read* no arquivo *fastq* e especificada no arquivo *xml*. No tocante às *reads mate-pair*, o script **sff_extract** dispunha das *reads forward* e *reverse* alternadas entre si, em um mesmo arquivo *fastq*. Ainda que esse arquivo pudesse ser utilizado por certos montadores, optou-se pela separação de cada uma dessas *reads* em arquivos distintos, para facilitação dos procedimentos de filtragem. Para tanto, desenvolveu-se o *script* **fastqsplitter.pl**, de modo a separar essas *reads.*

Uma vez que dentre todos os sequenciamentos realizados via IonTorrent, a *E. coli* BH100 subtr. MG2014 era a única que possuía bibliotecas *mate-pair*, um maior esforço fora alocado para resolução da montagem desse genoma. Dito isso, foram aplicados diversos procedimentos de filtragem de modo a obter um conjunto de *reads* ótimo para montagem. Dentre esses processos destaca-se a filtragem pelo tamanho das *reads*. Da análise da distribuição da qualidade das bases e da frequência de *reads* de diferentes tamanhos, utilizou-se o *script* **fast_sample** de modo a filtrar as *reads fragment* para tamanhos superiores a 30 pb e inferiores a 250 pb; as *reads mate-pair* 3 kb e 6 kb *forward* para tamanhos superiores a 50 pb e inferiores a 250 pb; e as *reads mate pair* 3 kb e 6 kb *reverse* para tamanhos superiores a 50 pb e inferiores a 230 pb. De modo que as *reads forward ou reverse* selecionadas no arquivo *fastq* tivessem o seu respectivo par também filtrado, desenvolveu-se o *script* **fastqpairselector.pl.**

A observação da cobertura aproximada do genoma também orientou a execução de filtragens das *reads* obtidas com IonTorrent do genoma da *E. coli* BH100 subtr. MG2014. O Newbler é um dos softwares de montagem que geralmente resulta em melhores métricas (menor número de *contigs* e maior N50) com dados advindos dessa plataforma. Isso ocorre, pois, esse software foi desenvolvido para lidar com *reads* da plataforma 454 cujos vieses e procedimentos de *base calling* são similares aos de IonTorrent. No entanto, esse software executa a montagem seguindo a estratégia OLC. Ao considerar as *reads* advindas das bibliotecas *fragment*, *mate-pair* 3 kb e *mate-pair* 6 kb tem-se aproximadamente 12,5 milhões de *reads* cuja cobertura média no genoma resultaria em ~ 500x. Tamanha cobertura além de poder acentuar erros de montagem também resultaria em um longo tempo de espera para montagem em um software do tipo OLC. Para contornar esses problemas, foram ensaiadas diferentes combinações de amostragens de cada conjunto de *reads* de modo que a cobertura total não fosse superior à 100x. A combinação que resultou em um menor número de *contigs* refere-se ao uso de 25% das *reads fragment*, 40% das *reads mate-pair* 3kb e 33% das *reads* mate-pair 6kb

2.2.4 Montagem

2.2.4.1 A MONTAGEM DO GENOMA DA *E. COLI* BH100 SUBTR. MG2014

2.2.4.1.1 Cromossomo

Ao total foram geradas 4.269.641 *reads fragment*, 3.596.096 *reads* pareadas *mate-pair* 3 kb e 4.518.000 *reads* pareadas *mate-pair* 6 kb. O conjunto de fragmentos utilizados para etapa de *scaffolding* do genoma completo da *E. coli* BH100 subtr. MG2014 se deu pela montagem de *scaffolds* no Newbler utilizando 25% das *reads* da biblioteca *fragment*, 40% das *reads* da biblioteca *mate-pair* 3 kb e 33% das *reads* da biblioteca *mate-pair* 6 kb. Essa combinação foi utilizada com o propósito de manter a cobertura abaixo de 100x e com isso possibilitar a montagem em um montador do tipo OLC, como Newbler e Mira. Dentre as montagens realizadas no Mira, Spades e Newbler, o resultado do último foi selecionado por ter resultado em um menor número de *scaffolds*[19].

Utilizando a *E. coli* 536 como referência, foi possível ordenar os *scaffolds* criados pelo Newbler. Após etapa de fechamento de *gaps* o cromossomo fechado da *E. coli* BH100 subtr.

[19]Como foram utilizadas *reads mate-pair*, o Newbler v. 2.9 é capaz de unir contigs em *scaffolds*.

MG2014 foi sujeito a 54 procedimentos de refinamento, a fim de corrigir regiões de baixa cobertura e/ou de mapeamentos espúrios.

2.2.4.1.2 O plasmídeo pAp

A montagem do plasmídeo de 10 MDa ocorreu utilizando a ferramenta PlasmidSpades. A saída desse programa resultou em uma sequência circular única de 14.241 pb, que foi aqui denominada de pAp.

2.2.4.1.3 O plasmídeo pBH100-1

Dos 11 *scaffolds* obtidos via Newbler a partir das *reads fragment* e *mate-pair* de IonTorrent, foi possível montar dois *scaffolds* referentes a plasmídeos da *E. coli* BH100 subtr. MG2014. O menor desses *scaffolds* era referente a um fragmento de 2.560 pb do plasmídeo pAp. O outro era um misto do plasmídeo de 60 MDa, mas contendo um fragmento de 11.681 pb do plasmídeo pAp. Após etapa de fechamento de *gaps* e refinamento desse possível replicon, obteve-se uma sequência circularizável de 112.410 pb denominada pBH100-x (**Figura 2.3**).

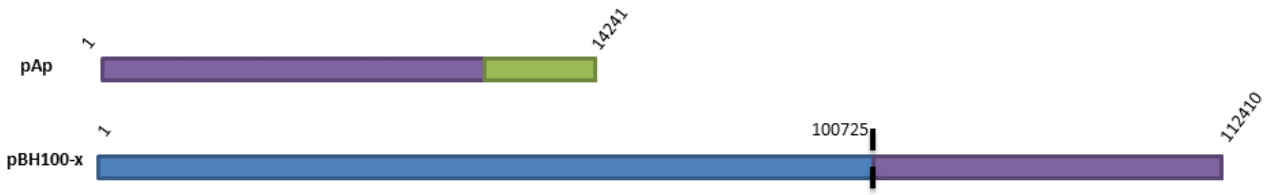


Figura 2.3 Representação dos plasmídeos pAp e pBH100-x obtidos no processo de montagem de *reads* oriundas da *E. coli* BH100 subtr. MG2014. Em pBH100-x destaca-se a porção do plasmídeo pAp integrada na posição 100.725.

A fim de determinar se a recombinação observada era fruto biológico ou um artefato de montagem, observou-se o mapeamento de *reads* na junção do que seria referente ao plasmídeo de 60 MDa e pAp. O mapeamento na região do flanco esquerdo do plasmídeo pAp, com *reads* provenientes das bibliotecas *fragment* e *mate-pair*, sugere que a junção seja real e fruto de uma recombinação biológica (**Figura 2.4**). A fim de verificar se a região do flanco direito de pAp também corroboraria para esse fato, realizou-se um rearranjo da sequência, de forma que a região correspondente ao plasmídeo pAp permanecesse no início de pBH100-x e seu flanco direito mapeado ao restante do replicon (**Figura 2.5**).



Figura 2.4 Mapeamento de *reads fragment* (A) e *reads fragment* e *mate-pair* (B) na região do flanco esquerdo de pAP (destacado em roxo) no replicon pBH100-x.

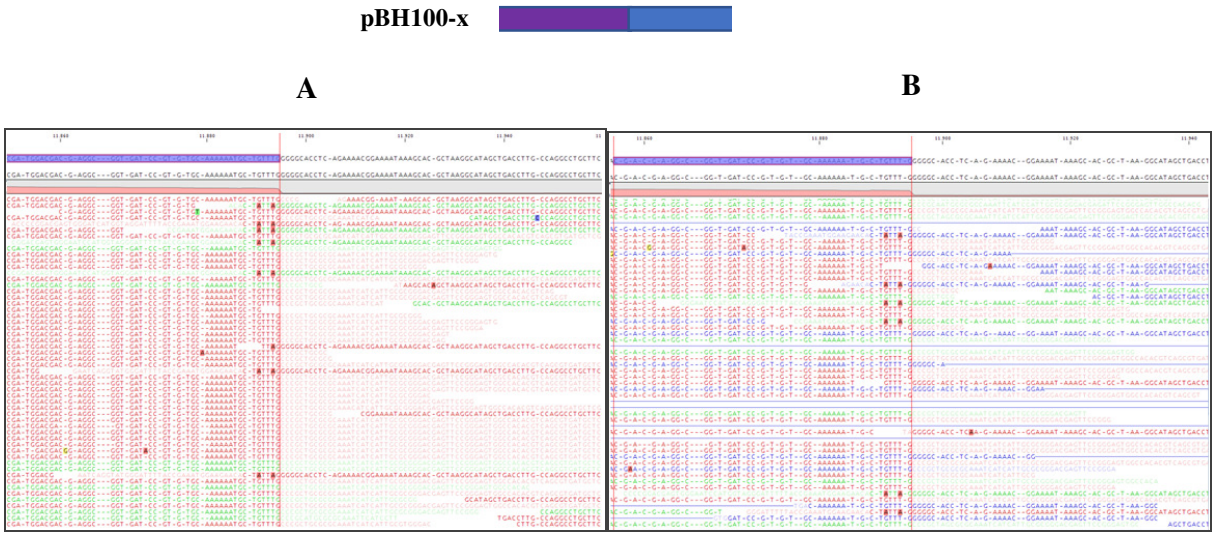


Figura 2.5 Mapeamento de *reads fragment* (A) e *reads fragment* e *mate-pair* (B) na região do flanco direito de pAp (destacado em roxo) no replicon pBH100-x após rearranjo da sequência.

Das observações do mapeamento de *reads fragment* não foram verificadas *reads* que justificassem essa região (**Figura 2.5A**). No entanto, o mapeamento de *reads mate-pair* corroboram para presença dessa junção (**Figura 2.5B**). Essas observações podem sugerir dois episódios: 1) as *reads mate-pair* que mapeiam no flanco direito de pAp são *reads* quimeras, artefatos da construção da biblioteca; 2) como para construção da biblioteca *mate-pair* foram utilizados 10 µg de DNA, em contraste com 1 µg para biblioteca *fragment*, é possível que um

evento recombinatório menos frequente entre os plasmídeos tenha sido amostrado apenas nas bibliotecas *mate-pair*. Disso posto, optou-se por não considerar o pBH100-x nas análises subsequentes.

A hipótese de que a junção entre os plasmídeos fosse um artefato de montagem foi descartada devido sustentação com mapeamento de *reads* na região do flanco esquerdo de pAp (**Figura 2.4**). Para tanto, assumiu-se que parte do plasmídeo pAp deveria estar presente no replicon. Para identificação e montagem da sequência completa do mesmo, e utilizando-se como base o replicon pBH100-x, removeu-se a região correspondente ao pAp e rearranjou-se 1.000 pb da posição inicial para o término da sequência, inserindo-se um *gap* artificial entre os fragmentos (**Figura 2.6**). A partir da estratégia de *contig walking*, para resolução desse *gap*, foi possível construir uma sequência circularizável contendo 4.950 pb do plasmídeo pAp e possuindo o total de 107.276 pb, aqui denominada pBH100-1.

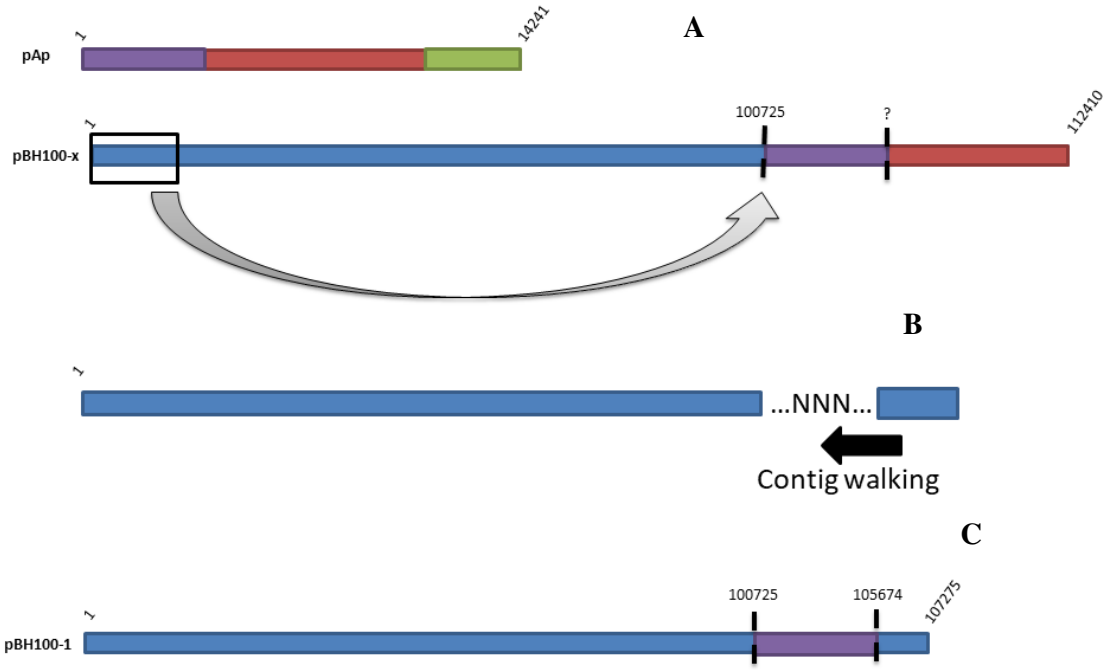


Figura 2.6 Representação esquemática da construção do plasmídeo pBH100-1. A) Fragmento inicial de 1.000 pb do replicon pBH100-x fora removido e inserido na região referente ao flanco esquerdo do plasmídeo pAp, que também fora removida. B) Aplicação da estratégia de *contig walking* para resolução do *gap*. C) Plasmídeo pBH100-1 contendo fragmento de 4.950 pb do plasmídeo pAp.

2.2.4.2 A MONTAGEM DO GENOMA DA *E. COLI* BH100 SUBTR. MG2017

2.2.4.2.1 Cromossomo

Ao total foram gerados 16.388.308 *reads paired-end* que serviram para montar 3 replicons da *E. coli* BH100 subtr. MG2017. A partir da montagem gerada pelo Spades e da informação de pareamento das *reads*, os *contigs* foram submetidos ao SSPACE para geração de 98 *scaffolds*. Estes foram utilizados no Contiguator para formação de um *superscaffold* utilizando a *E. coli* 536 como referência. Após etapa de fechamento de *gaps* esse replicon foi submetido a execução de 5 procedimentos de refinamento.

2.2.4.2.2 O plasmídeo pApR

Nenhum plasmídeo foi montado utilizando o PlasmidSpades. No entanto, após BLASTn do plasmídeo pAp com os *scaffolds* excluídos da etapa de construção do *superscaffold* cromossomico, verificou-se sua integração ao *Scaffold* 66. Esse *scaffold* também apresentava regiões, cujo reverso complementar era idêntico ao *Scaffold* 33 (**Figura 2.7A**). Dessa forma, hipotetizou-se que o plasmídeo pAp possuía um fragmento adicional na linhagem BH100 subtr. MG2017.

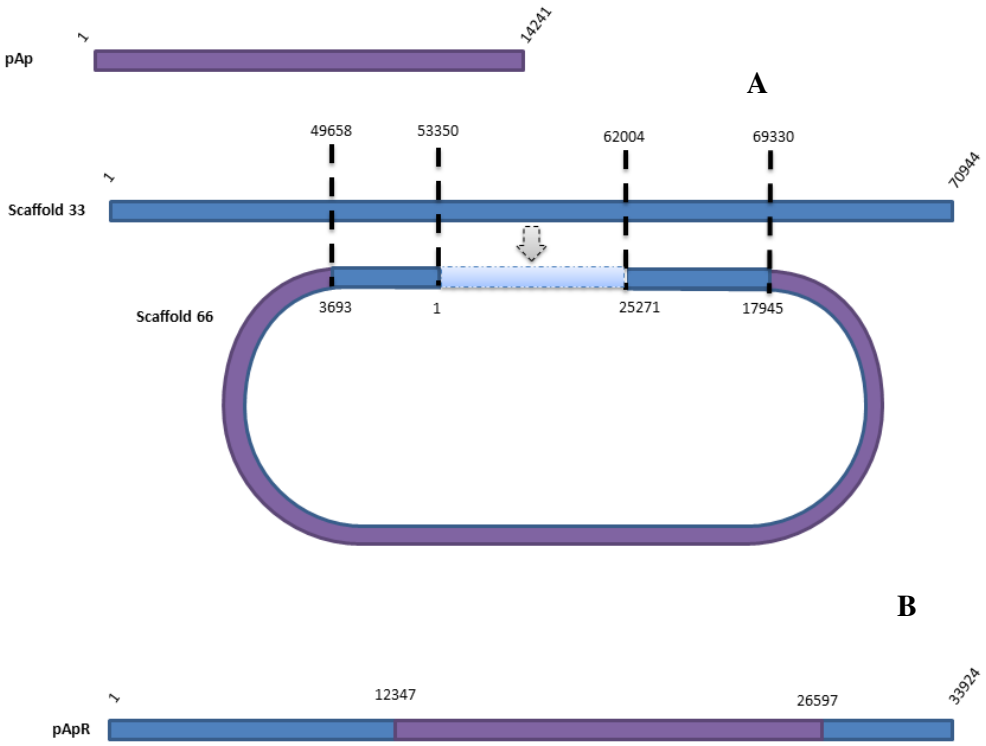


Figura 2.7 Representação esquemática da construção do plasmídeo pApR. A) O plasmídeo pAp encontrava-se integrado ao *Scaffold 66.* As regiões desse *scaffold* não pertencentes ao plasmídeo pAp eram identicas a uma região do *Scaffold* 33. Dessa forma inferiu-se que a região 49.658 – 69.330 do *Scaffold* 33 estaria presente em uma variante do plasmídeo pAp. B) O plasmídeo pApR contendo o plasmídeo pAp e região de 14.253 pb do *Scaffold* 33.

A fim de validar essa proposta observou-se o mapeamento das *reads* no flanco esquerdo (**Figura 2.8A**) e direito (**Figura 2.8B**) do plasmídeo pAp integrado ao *Scaffold* 66. Esses mapeamentos demonstraram que há justificativa para a presença do plasmídeo pAp intercalada a outra região. Com isso, o plasmídeo recombinado, aqui denominado pApR, foi manualmente construído adicionando-se o reverso complementar de um fragmento de 8.655 pb do *Scaffold* 33 ao *Scaffold* 66 (**Figura 2.7B**).

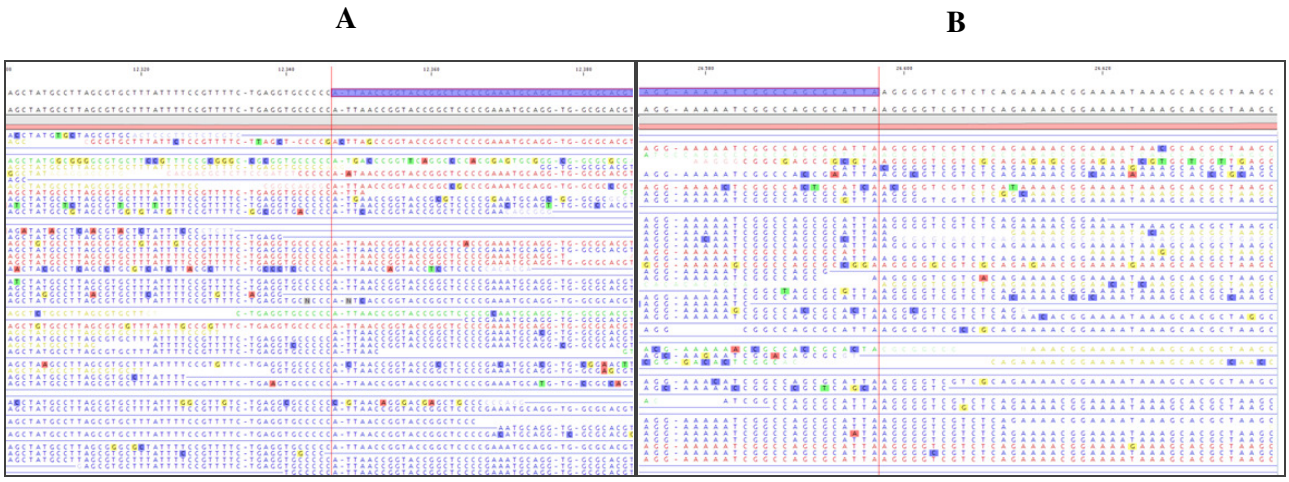


Figura 2.8 Mapeamento de *reads pair-end* Illumina no flanco esquerdo (A) e direito (B) de pAp (destacado em roxo)

A fim de averiguar se o plasmídeo pAp também existia na *E. coli* BH100 subtr. MG2017, as *reads* dessa linhagem foram mapeadas ao plasmídeo pAp, observando-se sua circularização. A **Figura 2.9** revela a presença de um mapeamento espúrio na região, sugerindo que essa variante plasmidial não esteja presente ou seja relativamente rara nessa linhagem. De fato, das 778 *reads* mapeadas nessa região apenas 11 corroboram essa junção.

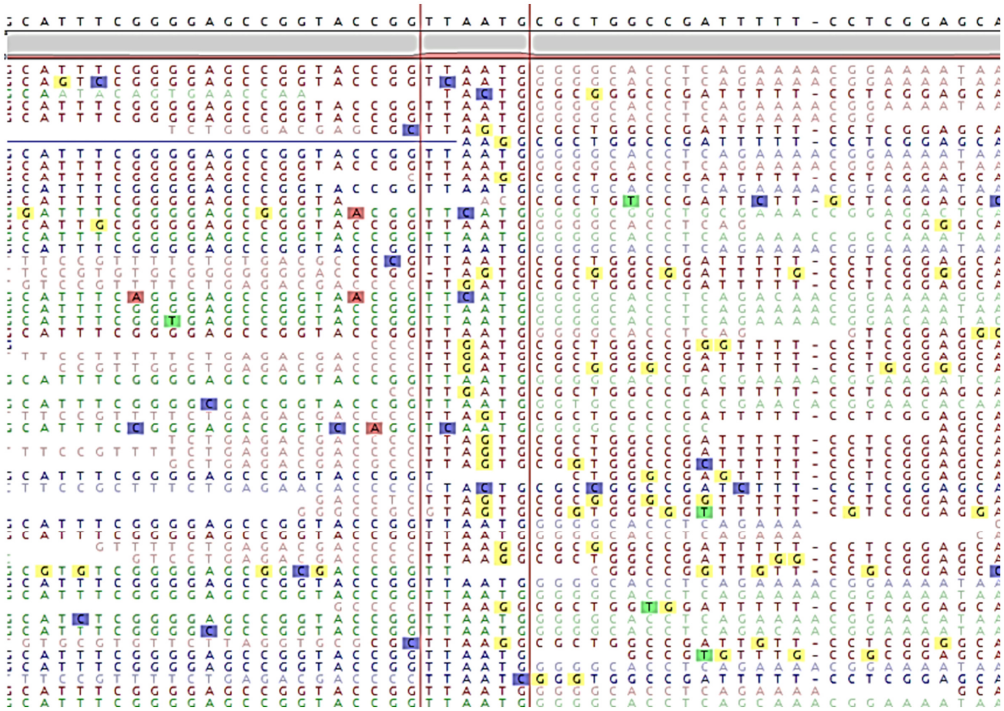


Figura 2.9 Mapeamento de *reads pair-end* no plasmídeo pAp da *E. coli* BH100 subtr. MG2014. Os tons mais claros indicam porções das *reads* não alinhadas na referência. Das 778 *reads* alinhadas na região apenas 11 corroboram para a presença do plasmídeo pAp.

2.2.4.2.3 O plasmídeo pBH100-1

A partir dos *scaffolds* excluídos do *superscaffold* referente ao cromossomo, foi construído o plasmídeo pBH100-1, a partir da criação de um *superscaffold* utilizando o plasmídeo NR1 como referência. Após etapa de fechamento de *gaps*, esse replicon não careceu de nenhuma etapa de refinamento. Mediante seu alto grau de identidade ao plasmídeo pBH100-1 obtido na *E. coli* BH100 subtr. MG2014, utilizou-se a mesma identificação para se referir ao replicon obtido na *E. coli* BH100 subtr. MG2017.

2.2.4.3 A MONTAGEM DO GENOMA *E. COLI* BH100L SUBTR. MG2017

Ao total foram gerados 13.371.138 *reads paired-end* referente a dois replicons da *E. coli* BH100L subtr. MG2017. Os *contigs* gerados no Spades foram alinhados ao cromossomo da *E. coli* BH100 subtr. MG2017 para criação do *superscaffold*. Após etapa de fechamento de *gaps* o cromossomo foi submetido a 13 procedimentos de refinamento.

Os *contigs* excluídos do processo de *Scaffolding* foram alinhados ao plasmídeo pBH100-1 da *E. coli* BH100 subtr. MG2017 para formação do *superscaffold*. Após etapa de fechamento de *gaps*, o replicon foi submetido a 5 procedimentos de refinamento. Uma vez que esse replicon possui similaridade ao pBH100-1 mas não contém fragmentos de pAp, o mesmo foi aqui denominado pBH100alpha[20].

2.2.4.4 A MONTAGEM DO GENOMA DA *E. COLI* BH100N SUBTR. MG2017

Ao total foram geradas 12.550.562 *reads paired-end* referentes ao genoma da *E. coli* BH100N subtr. MG2017. Os 113 *contigs* gerados pelo Spades foram utilizados para construção do *superscaffold* no Contiguator, utilizando o cromossomo da *E. coli* BH100 subtr. MG2017 como referência. Da análise dos *contigs* excluídos do *superscaffold*, bem como da montagem de *reads* via PlasmidSpades não foram detectados, conforme o esperado, replicons pertencentes a plasmídeos. Após etapa de fechamento de gaps do replicon cromossomal, 22 procedimentos de refinamento foram executados.

[20]Esse nome foi dado pois acredito que esse tenha sido o plasmídeo original do isolamento da *E. coli* BH100 por Chartone-Souza.

2.2.4.5 A MONTAGEM DOS DEMAIS PLASMÍDEOS

O plasmídeo pAp, presente na linhagem transconjugante *E. coli* K12T foi montado via PlasmidSpades, a partir de 359.590 *reads* de um total de 14.312.172, e não careceu de refinamentos subsequentes. O cromossomo da E. coli K12T subtr. MG2017 não foi montado.

Já na amostra referente à *E. coli* BH100L subtr. MG2014, foram geradas 2.247.497 *reads fragment* referentes ao cromossomo e a um plasmídeo. A partir dos arquivos excluídos do *superscaffold* construído utilizando o cromossomo da *E. coli* 536 como referência, criou-se o *superscaffold* de um replicon utilizando o plasmídeo pBH100-1 da *E. coli* BH100 subtr. MG2014, como referência. Após etapa de fechamento de *gaps*, esse replicon foi submetido a dois procedimentos de refinamento. Por não conter fragmentos do plasmídeo pAp e por ser similar ao plasmídeo identificado na *E. coli* BH100L subtr. MG2017, esse replicon foi também denominado pBH100alpha. O cromossomo dessa linhagem não foi submetido à etapa de fechamento de *gaps* e refinamento.

2.2.4.6 A MONTAGEM DOS CROMOSSOMOS DAS DEMAIS LINHAGENS DAS SUBCEPAS MG2014

Os *contigs* oriundos da *E. coli* BH100S subtr. MG2014 não foram submetidos ao processo de fechamento de *gaps* e refinamento, uma vez que nenhum dos *contigs* montados eram referentes a uma sequência plasmidial. Além disso, as *reads* desse projeto não mapeavam no plasmídeo pAp, que era esperado ocorrer nessa variante. Esse fato pode ser devido a falhas durante a entrega da linhagem ou cura do plasmídeo durante o longo tempo de cultivo, sem a presença de ampicilina como agente seletivo. Com efeito, optou-se por não dar prosseguimento com as análises dessa estirpe. Já o genoma da *E. coli* BH100N subtr. MG2014 não foi submetido aos processos de fechamento de *gaps* e refinamento, permanecendo na condição de *draft*.

2.2.5 Anotação e curadoria

2.2.5.1 REPLICONS CROMOSSOMICOS

Optou-se por utilizar apenas os replicons completos nas análises subsequentes. Um resumo das anotações cromossomicas, conjuntamente anotadas via transferência automática e RAST, pode ser observada na **Tabela 2.2**.

Tabela 2.2 Estatísticas de anotação referentes aos cromossomos de 4 linhagens completas da *E. coli* BH100 e suas variantes.

Cromossomo	CDSs	Pseudogenes	Genes totais	tRNA	rRNA	%GC	Tamanho (bp)
BH100 subtr. MG2014	4.971	68	5.039	84	22	50,38	5.072.943
BH100 subtr. MG2017	5.080	27	5.107	82	22	50,41	5.131.212
BH100L subtr. MG2017	4.827	85	4.912	81	22	50,53	5.033.214
BH100N subtr. MG2017	4.981	79	5.060	93	22	50,44	5.116.034

Durante a etapa de curadoria, objetivou-se a correção de CDSs vizinhas cujos produtos provavelmente pertenciam à mesma proteína. Através de Blastp das sequências proteicas dessas regiões com a base de dados do Uniprot e a partir da observação do consenso das *reads* mapeadas nesses sítios, procurou-se corrigir as CDSs a partir da inserção e/ou deleção de bases, justificadas pela análise do mapeamento. Dessa forma, após etapa de transferência de anotação, foram anotadas 168 CDSs parciais e/ou fragmentadas referentes ao cromossomo da *E. coli* BH100 subtr. MG2014, 145 referentes ao cromossomo da *E. coli* BH100 subtr. MG2017, 157 referentes ao cromossomo da *E. coli* BH100L subtr. MG2017 e 143 referentes ao cromossomo da *E. coli* BH100N subtr. MG2017. Ao final da etapa de curadoria as CDSs que possuíssem códons de parada no quadro de leitura, *frameshits* ou que se mostrassem parciais quando comparadas à base de dados do Uniprot, foram anotadas como pseudogenes. A distribuição das CDSs indicadas como pseudo pode ser visto na **Tabela 2.3**.

Tabela 2.3 Distribuição das CDSs anotadas como pseudogenes.

Cromossomo	*Frameshift*	Códon de parada prematuro	CDS parcial	Total
BH100 subtr. MG2014	62	6	0	68
BH100 subtr. MG2017	17	3	7	27
BH100L subtr. MG2017	77	6	2	85
BH100N subtr. MG2017	68	11	0	79

O número total de anotações indicadas como pseudogenes na *E. coli* BH100 subtr. MG2017 é consideravelmente inferior às demais. Esse fato, no entanto, se deve a anotação de transposases de elementos móveis, em particular ao elemento IS3. Nesse caso, como pode ser observado na **Tabela 2.2**, um maior número de CDSs foi anotada nessa linhagem, o que corresponde a não identificação da junção de duas ou mais CDSs como sendo pertencentes a uma única *feature*. Isso ocorre, devido ao procedimento de anotação do RAST que não realiza anotação observando para mudanças no quadro de leitura. De acordo com os critérios estabelecidos de anotação nesse trabalho, as anotações provenientes do RAST que se sobrepunham à anotação das *features* transferidas eram eliminadas, de modo a conservar a anotação da refe-

rência advinda do NCBI. O *script* de transferência, no entanto, apresenta falhas ao transferir anotações de regiões repetitivas, muito similares entre si. Um exemplo desse fato pode ser visto na **Figura 2.10A**, onde diversas CDSs de regiões repetitivas correspondentes à transposase de IS3 são anotadas em um mesmo locus. Com efeito, essas anotações duplicadas são removidas e os loci repetitivos permanecem com a anotação do RAST (**Figura 2.10B**).

A



B



Figura 2.10 Exemplos de vieses de anotação. A) Acúmulo de anotações de regiões repetitivas em um mesmo locus, fruto de erros no script ffn2pseudo.fasta36.pl. B) Anotação de duas CDSs (em amarelo) pelo RAST quando na verdade referem-se a um produto proteico único após comparação com outras sequências depositadas no Uniprot.

Em relação aos rRNAs nota-se que em todas as linhagens foram preditos 7 operons de rRNA, que incluem 5S rRNA, 16S rRNA e 23S rRNA. Os 7 operons estão de acordo com o total achado em outras linhagens de Enterobacteriaceae (VĚTROVSKÝ; BALDRIAN, 2013). A presença de um 5S rRNA adicional está vinculada ao operon *rrnD* (Escherichia coli K-12 substr. MG1655 Compound: 5S ribosomal RNA (ECOCYC, [s.d.]). Ao todo esses operons contribuem para síntese diferencial de ribossomos sob diferentes condições de crescimento, contribuindo para geração de *fitness* adaptativo em *E. coli* (MAEDA; SHIMADA; ISHIHAMA, 2015).

A **Tabela 2.4** indica o número de cópias de cada tRNA por replicon. Nela vale destacar a presença do tRNA carreador de selenocistína, capaz de proporcionar a codificação de CDSs contendo o códon de parada TGA (BRÖCKER et al., 2014). Também se salienta o maior número de tRNAs contendo o anticódon TTT, na *E. coli* BH100N subtr. MG2017. Esse fato deve-se a uma duplicação da região contendo 5 desses tRNAs.

Tabela 2.4 Frequência de tRNAs anotados automaticamente pelo RAST. (Continua)

Anticodon(a.a)	E. coli BH100 subtr. MG2014	E. coli BH100 subtr. MG2017	E. coli BH100L subtr. MG2017	E. coli BH100N subtr. MG2017
ACG(Arg)	4	4	4	5
CAA(Leu)	1	1	1	1
CAG(Leu)	4	4	4	4
CAT(Met)	6	7	7	7
CCA(Trp)	1	1	1	1
CCC(Gly)	1	1	1	1
CCG(Arg)	1	1	1	1
CCT(Arg)	1	1	1	1
CGA(Ser)	1	1	1	1
CGG(Pro)	1	1	1	1
CGT(Thr)	1	1	1	1
CTG(Gln)	1	1	1	1
GAA(Phe)	2	2	1	2
GAC(Val)	2	2	2	2
GAG(Leu)	1	1	1	1
GAT(Ile)	3	1	1	3
GCA(Cys)	1	1	1	1
GCC(Gly)	4	4	4	4
GCT(Ser)	1	1	1	1
GGA(Ser)	2	2	2	2
GGC(Ala)	2	2	2	2
GGG(Pro)	1	1	1	1
GGT(Thr)	2	2	2	2
GTA(Tyr)	3	3	3	3
GTC(Asp)	3	3	3	3

GTG(His)	1	1	1	1
GTT(Asn)	4	4	4	4
TAA(Leu)	1	1	1	1
TAC(Val)	5	5	5	7
TAG(Leu)	1	1	1	1
TCA(SeC(p))	1	1	1	1
TCC(Gly)	1	1	1	1
TCT(Arg)	2	2	2	2
TGA(Ser)	1	1	1	1
TGC(Ala)	3	1	1	3
TGG(Pro)	1	1	1	1
TGT(Thr)	1	1	1	1
TTC(Glu)	4	5	5	4
TTG(Gln)	2	2	2	2
TTT(Lys)	6	6	6	11

2.2.5.2 REPLICONS PLASMIDIAIS

Um quadro comparativo, contendo as estatísticas de anotação dos diferentes replicons referentes a plasmídeos, pode ser visualizado na **Tabela 2.5**.

Tabela 2.5 Estatísticas de anotação dos diferentes replicons plasmidiais presentes na *E. coli* BH100 e suas variantes, e na E. coli K12T.

Plasmídeo	CDSs	Pseudo	%GC	Tamanho (pb)	Massa Estimada (MDa)
pAp_MG2014	17	0	53,31	14.241	9,4
pAp_MG2017	16	0	53,33	14.247	9,4
pApR_MG2017	36	0	57,39	33.924	22,4
pBH100alpha_MG2014	133	3	52,01	102.316	67,5
pBH100alpha_MG2017	136	0	52,02	103.103	68,0
pBH100-1_MG2014	133	0	51,91	107.274	70,8
pBH100-1_MG2017	140	0	52,02	105.801	69,8

Verificou-se a presença de potenciais pseudogenes apenas no plasmídeo pBH100alpha da *E. coli* BH100L subtr. MG2014. Todos os pseudogenes preditos se resumiam a *frameshfits*. O primeiro deles está associado à TniB, uma proteína de ligação ao ATP e associada ao integron de classe 1 presente no Tn21 (LIEBERT; HALL; SUMMERS, 1999a). O segundo está associado a uma proteína hipotética e o terceiro está associado ao gene *merA* responsável pela síntese da mercúrio redutase.

A fim de saber se os tamanhos dos plasmídeos estavam de acordo com as estimativas elaboradas em trabalhos anteriores, utilizou-se a relação mencionada em Nascimento (1987),

onde cada 1 kb corresponderia à aproximadamente 0,66 MDa, para estimar as massas moleculares de cada um dos replicons plasmidiais obtidos neste trabalho (**Tabela 2.5**). Para os plasmídeos pAp nota-se um valor próximo ao estimado em trabalhos anteriores para o menor plasmídeo presente na *E. coli* BH100 (10 MDa). Já os plasmídeos pBH100, apresentam massa molecular cerca de 10 unidades acima do predito (60 MDa). Entretanto, deve-se levar em consideração que essas estimativas foram realizadas em géis de agarose através da comparação entre o deslocamento das bandas plasmidiais da BH100 em relação às bandas plasmidiais de replicons com massa conhecida. Se ainda considerar-se o fato de que fragmentos maiores de DNA tendem a se difundir menos em géis de agarose, pode-se antecipar que diferenças de 10 unidades na massa molecular resultaria em, inexpressivas, diferenças migracionais das bandas. Com efeito, em relação aos tamanhos, os replicons plasmidiais encontrados neste trabalho parecem corroborar as observações dos plasmídeos da BH100 visualizados em géis de agarose, ao longo dos anos.

CAPÍTULO 3 GENÔMICA COMPARATIVA

3.1 MATERIAL E MÉTODOS

3.1.1 Desenvolvimento e aquisição de *scripts*

Os seguintes *scripts in-house* foram desenvolvidos com a finalidade de ajustar os dados advindos de outros programas para formatos específicos de modo a servirem de *input* para outros programas ou até mesmo para facilitar a interpretação. Todos os seguintes *scripts* foram desenvolvidos em Perl e testados em um ambiente Linux:

a) **BPGAfiles2tab.pl**: esse *script* foi desenvolvido com o propósito de transformar os cabeçalhos dos arquivos *multifasta* oriundos do programa BPGA e referentes aos grupos de CDSs componentes do *core*, acessórios e únicos em um arquivo tabular para melhor identificação e comparação.

b) **intersect2faa.pl**: esse *script* foi elaborado de forma a escrever arquivos *multifasta* exclusivos para cada intersecção de genes acessórios determinados pelo BPGA, a partir do arquivo *multifasta* contendo todas as proteínas classificadas como acessórias.

c) **IStable2gff.pl**: esse *script* foi desenvolvido de modo a escrever em um arquivo de anotação *gff* as posições preditas pelo ISFinder, para visualização no Artemis.

d) **multigbk2single.pl**: esse *script* foi desenvolvido com a finalidade de obter diferentes arquivos *gbk* a partir de um arquivo *multigbk*.

e) **patricmultifasta_splitter.pl**: esse script foi desenvolvido para separar as proteínas inerentes de cada organismo presentes em um único arquivo *multifasta*, em arquivos *multifasta* próprios de cada organismo.

f) **tBLASTnParser.pl**: esse *script* foi desenvolvido com a finalidade de obter informações de um arquivo de alinhamento oriundo do tBLASTn. Seu propósito consiste em identificar em um arquivo tabular todas as instâncias das proteínas utilizadas como *query* presentes nos organismos utilizados como *subject*, segundo critérios de identidade, cobertura e *e-value*.

Além dos scripts acima também foi utilizado o script **gbk2faa.pl**[21], capaz de ler um arquivo multigbk e retornar em diferentes arquivos a sequência de proteínas de cada genoma.

3.1.2 Obtenção dos genomas RefSeq

Para fins de análises comparativas, foram extraídos do RefSeq a sequência do cromossomo completo de 414 linhagens de *Escherichia coli* (Apêndice B), além dos genomas de *Salmonella enterica* enterica (RefSeq: NC_003197) e *Escherichia fergusonii* ATCC 35469 (RefSeq: NC_011740), utilizadas como grupo externo nas análises filogenéticas.

3.1.3 Sintenia

A posição relativa de fragmentos genômicos entre os diferentes replicons foi estudada utilizando-se o Mauve (v. 2.4) (DARLING et al., 2004) via ProgressiveMauve a partir de seus parâmetros padrão.

3.1.4 Identificação de elementos móveis

A anotação de ISs foi realizada utilizando-se a plataforma ISFinder (SIGUIER et al., 2006), que funciona como um repositório de diversas sequências de ISs previamente anotados. Os arquivos *fasta* foram utilizados para submissão nessa plataforma. Para identificação de transposons foram extraídos do NCBI 1.380 sequências que serviram de banco de dados para procura via BLASTn. Esse banco foi criado assumindo como critério a presença do termo "transposon" no título das entradas presentes no repositório *Nucleotide* (NT) e filtrados para bactérias.

A identificação e anotação de sequências oriundas de prófagos foi realizada no PHASTER (ARNDT et al., 2016). A sequência de nucleotídeos no formato fasta de cada cromossomo foi submetida ao programa e as coordenadas das regiões correspondentes a prófagos completos e incompletos extraídas para futura comparação.

[21] Menzel, P. (2017). gbk2faa.pl. Disponível em: https://github.com/bioinformatics-centre/kaiju/blob/master/util/gbk2faa.pl

3.1.5 Similaridade cromossômica e plasmidial

Arquivos *fasta* contendo os cromossomos das diferentes linhagens montadas da BH100, outras 414 sequências cromossômicas de *E. coli* e 2 cromossomos referentes à *Escherichia. fergussini* e *Salmonella enterica*, utilizadas como grupo externo, foram submetidas ao Gegenees (v. 2.2) (AGREN et al., 2012). Os fragmentos gerados de cada um dos genomas foram comparados todos contra todos via BLASTn. A partir do *heatmap* contendo os *scores* de similaridade, a árvore foi construída a partir do arquivo *nexus* no SplitsTree4 (HUSON; BRYANT, 2006), utilizando-se o método de agrupamento Neighbor-joining (NJ) (SAITOU; NEI, 1987), e a figura refinada para visualização no FigTree (v. 1.4.3)[22]..

A observação da homologia dos plasmídeos da BH100 foi obtida via BLASTn com a base de dados NT. A fim de observar os plasmídeos sob um contexto filogenético, em relação aos já identificados em *Escherichia coli*, foram coletados do GenBank 1.354 sequências referentes a esses replicons. As árvores foram construídas a partir do método NJ, já implementado na ferramenta online do BLAST, e editadas para visualização no FigTree.

3.1.6 Ilhas genômicas e de patogenicidade

A predição de ilhas genômicas, incluindo as de patogenicidade, na *E. coli* BH100 subtr. MG2017, foi realizada mediante a anotação no IslandViewer (BERTELLI et al., 2017). De modo a comparar essas predições com outras uropatogênicas, realizou-se o mesmo procedimento para as *E. coli* 536 e *E. coli* CFT073 (RefSeq: NC_004431.1). Os estudos das ilhas foram realizadas observando para as predições realizadas via IslandPath-DIMOB (HSIAO et al., 2005) e comparadas com as demais linhagens no BRIGG (v. 0.95) (ALIKHAN et al., 2011)

3.1.7 Investigação funcional

Estudos mais detalhados acerca da funcionalidade das diferentes CDSs, anotadas ao longo dos cromossomos e plasmídeos, foram realizados utilizando a ferramenta GoFeat (ARAUJO et al., 2018). As sequências de aminoácidos foram utilizadas nessa ferramenta para caracterização funcional, utilizando e-value de 10^{-10}. O GoFeat é capaz de comparar sequências proteicas com o repositório do Uniprot e importar diversas informações de diferentes

[22] Rambaut, Andrew (2007). FigTree. Disponível em: http://tree.bio.ed.ac.uk/software/figtree/

bases de dados de anotação, tais como Gene Ontology (G.O), KEEG, PFAM e InterproScan. Com isso, um maior conjunto de informações acerca de uma CDSs, em específico, pode ser obtida.

A comparação do conteúdo gênico foi realizada nas plataformas PATRIC (WATTAM et al., 2014) e BPGA (CHAUDHARI; GUPTA; DUTTA, 2016). A primeira foi também utilizada para anotação funcional e a segunda, primariamente, para identificação de CDSs pertencentes ao genoma central (*core*), acessório (pelo menos 2 genomas) e *singletons* (exclusivo de um único genoma). Nesse último caso, o agrupamento das diferentes sequências de aminoácidos com no mínimo 70% de similaridade foi obtido pela ferramenta Usearch (EDGAR, 2010).

3.1.8 Predição funcional de pseudogenes

É sabido que alguns genes podem sofrer processos de controle da expressão gênica em nível traducional. Um desses casos ocorre quando há mudança no quadro de leitura pelo ribossomo. Tal fenômeno pode ocorrer devido a presença de *motfis* específicos na sequência do mRNA, induzindo alterações na estrutura secundária e levando à mudança em sua leitura. A fim de identificar possíveis assinaturas e revelar possíveis genes regulados dessa maneira, as CDSs anotadas como pseudogenes, em decorrência de mudanças no quadro de leitura foram investigadas no software FSFinder (MOON et al., 2004).

Potenciais pseudogenes anotados dessa forma, devido à presença de códons de parada prematuros, também foram investigados. Nesse caso, foram observados os pseudogenes que continham o códon de parada TGA, já que o mesmo pode codificar para o aminoácido selenocisteina ou triptofano (CLARK et al., 2013; LU et al., 1995).

3.1.9 Predições complementares

Para determinação do sorotipo de cada uma das linhagens da BH100 utilizou-se a ferramenta SerotypeFinder (v. 1.1) (JOENSEN et al., 2015). A identificação de fatores de virulência nos plasmídeos e cromossomos se deu pela ferramenta VirulenceFinder (v 1.5) (JOENSEN et al., 2014). A confirmação dos marcadores de resistência identificados nos plasmídeos, bem como a procura por potenciais marcadores de resistência nos cromossomos, ocorreu mediante o uso da ferramenta ResFinder (v. 3.0) (ZANKARI et al., 2012). Por fim, a predição dos grupos de incompatibilidade de cada um dos plasmídeos investigados ocorreu via PlasmidFinder (v. 1.3) (CARATTOLI et al., 2014).

3.2 RESULTADOS E DISCUSSÃO

3.2.1 Cromossomos das linhagens BH100, BH100L e BH100N

As análises do SerotypeFinder revelaram que todas as variantes da BH100, salvo a *E. coli* BH100N subtr. MG2017, são H31:O6. A BH100N parece ter perdido o antígeno O, sendo apenas H31. Da investigação no ResFinder, não foram vinculados genes de resistência a nenhum dos cromossomos, salvo na *E. coli* BH100L subtr. MG2017, onde constatou-se a presença do gene *aadA1*, capaz de conferir resistência à estreptomicina.

3.2.1.1 ANÁLISE DE SINTENIA

A estrutura cromossômica da *E. coli* BH100 subtr. MG2014, *E. coli* BH100 subtr. MG2017, *E. coli* BH100L subtr. MG2017 e *E. coli* BH100N subtr. MG2017 foi inicialmente comparada, utilizando o software Mauve (**Figura 3.1**). Foi evidenciado alto grau de similaridade nucleotídica entre essas linhagens. No entanto, regiões aparentemente exclusivas de determinados genomas também puderam ser destacadas, sugerindo rearranjos na estrutura cromossômica.

De fato, ao observar o conteúdo de duas regiões em destaque na **Figura 3.1**, referente à *E. coli* BH100L subtr. MG2017, foi constatada a presença do Tn21, provavelmente adquirido do plasmídeo pBH100-1. Tal fato explica a identificação do gene de resistência à estreptomicina no cromossomo dessa linhagem. Outras alterações referiam-se basicamente à inserção de genes carreados por elementos móveis.



Figura 3.1 Representação dos blocos contíguos produzidos pelo Mauve, indicando o grau de sintenia entre os cromossomos da *E. coli* BH100 e suas variantes. Em destaque, as regiões correspondentes ao Tn21, presente no cromossomo da *E. coli* BH100L subtr. MG2017.

Outra região de destaque é referente a uma região circundada por elementos de inserção do tipo IS3 e que se encontra duplicada e invertida na *E. coli* BH100N subtr. MG2017 (**Figura 3.2**). Essa duplicação é sustentada pela análise da cobertura de *k-mers*, realizada pelo Spades no *contig* desse fragmento. Nessa análise, encontra-se um valor de 160,7 de cobertura de *k-mers* nessa região, em contraste com uma cobertura de ~ 70,0 associada a outros *contigs* da montagem. Esse fato, portanto, corrobora a existência de duplicidade real da região.

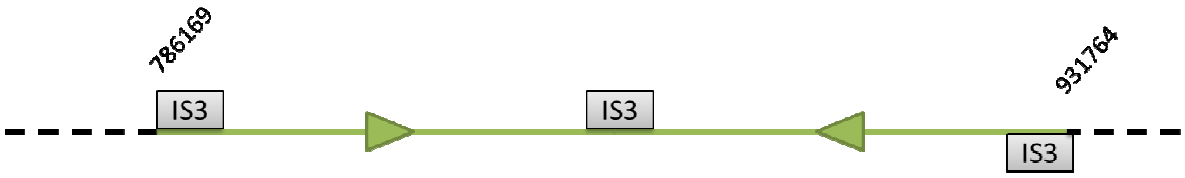


Figura 3.2 Esquematização de duplicação e inversão de uma região presente no cromossomo da *E. coli* BH100N subtr. MG2017. A região compreendida entre os dois primeiros ISs é vista duplicada e invertida entre os dois últimos. Esse evento sugere a presença de um evento recombinatório, via transposição do IS3 mais externo, e um evento de recombinação homóloga do IS3 interno.

3.2.1.2 Análise de elementos móveis

Diferentes elementos móveis foram investigados nos cromossomos das linhagens aqui trabalhadas, uma vez que podem ser objeto de variabilidade genotípica e fenotípica.

A identificação de prófagos, foi realizada na plataforma Phaster. Por meio dela, foram preditas 6 regiões classificadas como completas, referentes a prófagos, e outras regiões de caráter duvidoso ou incompletas (**Figura 3.3**). Pode-se notar que as regiões preditas, como de prófagos completos, são bastante conservadas ao longo dos diferentes cromossomos, sendo a principal diferença relacionada à perda do prófago Entero_lambda (GenBank: NC_001416) na *E. coli* BH100L subtr. MG2017, o que pode estar associado à inserção do Tn21 nessa região.

As demais regiões referem-se aos prófagos: Entero_I2_2 (GenBank: NC_001332); Salmon_Sen34 (GenBank: NC_028699); Entero_P4 (GenBank: NC_001609) e Entero_mEp460 (GenBank: NC_019716).

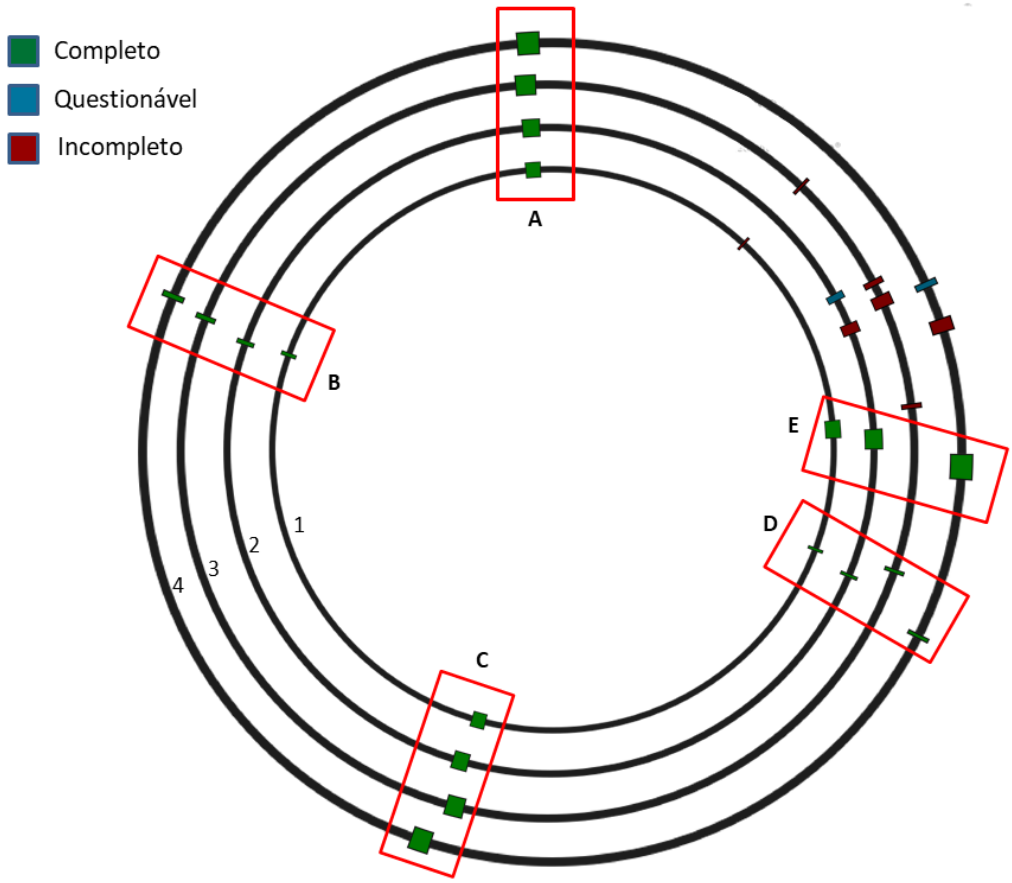


Figura 3.3 Predição de prófagos ao longo dos cromossomos da *E. coli* BH100 subtr. MG2014 (1), *E. coli* BH100 subtr. MG2017 (2), *E. coli* BH100L subtr. MG2017 (3) e *E. coli* BH100N subtr. MG2017 (4). Cada um dos grupos em destaque corresponde a um prófago distinto: Entero_mEp460 (A); Entero_P4 (B); Salmon_Sen34 (C); Entero_I2_2 (D); Entero_lambda (E).

Uma das principais razões para a dificuldade da montagem dos genomas de *E. coli* desse trabalho foi devido a presença de elementos móveis do tipo IS. A plataforma ISFinder,

juntamente com o *in-house script* **ISFinder2gff.pl** foram utilizados para identificação e anotação desses elementos móveis. A **Figura 3.4** resume a composição dos diferentes ISs presentes em cada cromossomo.

Aparentemente, não existem ISs da família dos IS2 nas *E. coli* BH100. Os elementos IS1 estão presentes em quantidades similares nos diferentes cromossomos e podem estar associados com os plasmídeos, que também os contêm. A presença do IS5, apenas na *E. coli* BH100 subtr. MG2014, pode ser devido aos plasmídeos do grupo pBH100, que possuem dois ISs dessa família intercalando o gene de resistência à canamicina. A verificação desse IS no cromossomo pode sugerir que em alguns indivíduos a marca de resistência a esse antibiótico possa se transferir para o cromossomo.

Em comparação com as linhagens de *E. coli* K-12*, E. coli* B e *E. coli* C constata-se um maior número de ISs da família IS3 (HU; DEONIER, 1981) presentes nos cromossomos da *E. coli* BH00. Esse fato pode ter implicações na geração de variabilidade da expressão gênica, uma vez que ISs dessa família podem agir como regiões promotoras mobilizáveis e, com isso, afetar os níveis de expressão de genes próximos (CHARLIER; PIETTE; GLANSDORFF, 1982).

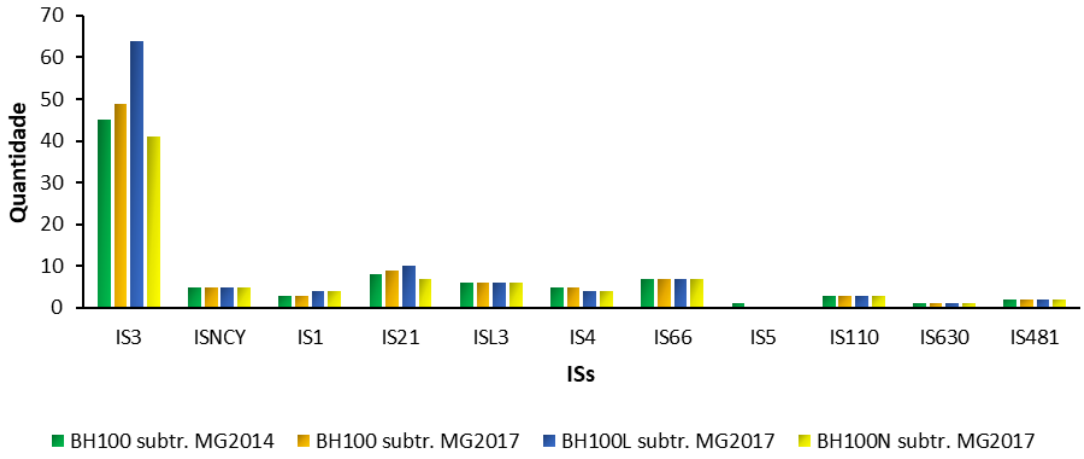


Figura 3.4 Composição de ISs nos cromossomos da *E. coli* BH100 e suas variantes.

Fora questionado se os diferentes elementos móveis eram conservados ao longo do cromossomo das diferentes linhagens desse estudo. A **Figura 3.5** revela que esses elementos se encontram espalhados, de forma diferencial, nessas linhagens. Tal fato revela a capacidade dessa bactéria em adaptar-se geneticamente, uma vez que esses elementos são responsáveis por afetar o conteúdo gênico e transcricional de uma espécie.

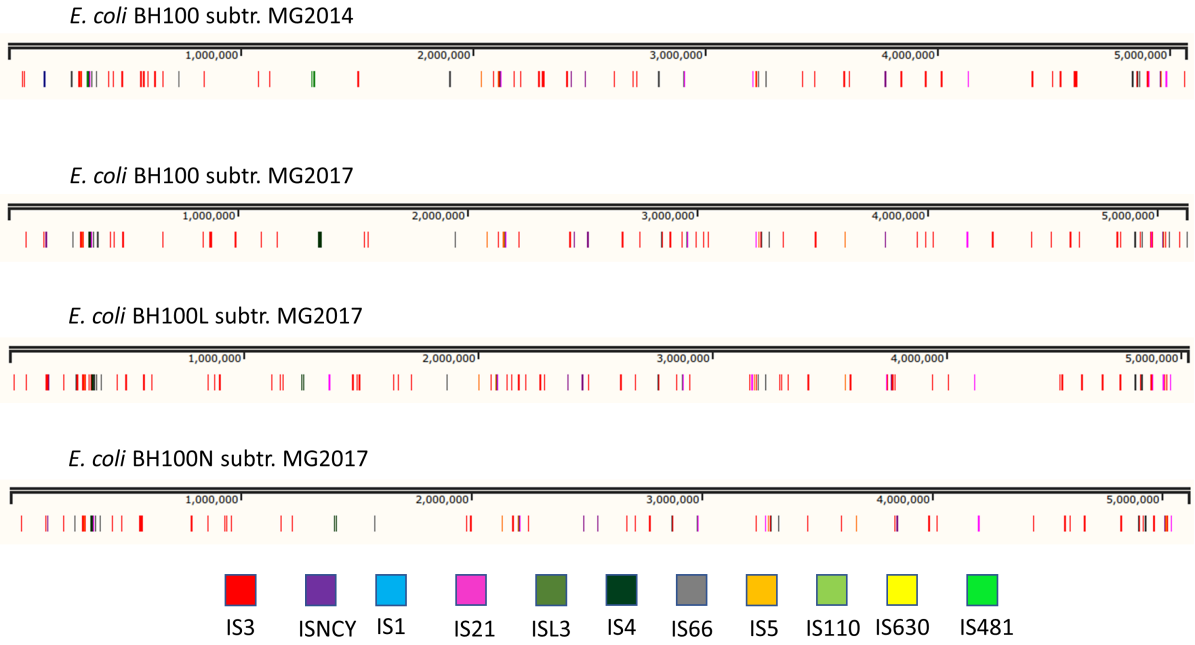


Figura 3.5 Distribuição de Elementos de Inserção (IS) ao longo dos cromossomos da *E. coli* BH100 e suas variantes.

3.2.1.3 QUADRO COMPARATIVO DE GENES ANOTADOS NO PATRIC

Os diferentes cromossomos foram anotados, utilizando-se a plataforma PATRIC e as diferenças funcionais, entre cada linhagem, foram investigadas. A partir da observação da região diferencial, entre as linhagens, foram coletadas as sequências proteicas correspondentes, para caracterização (**Figura 3.6**).

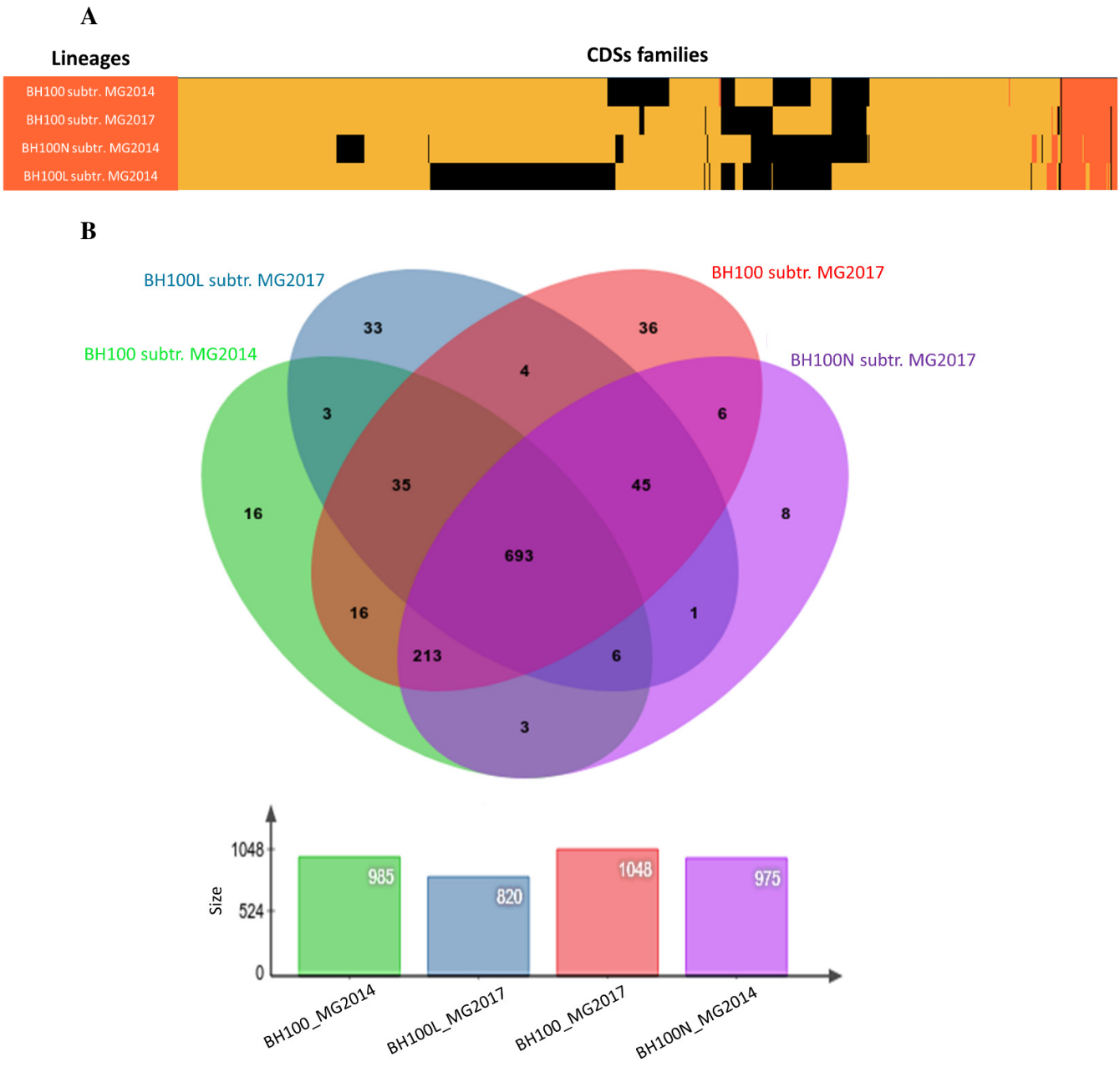


Figura 3.6 Relação de genes presentes/ausentes segundo anotação do PATRIC. A – *Heatmap* referente à região correspondente a 1.118 CDSs obtidas após algoritmo de clusterização padrão do PATRIC seguido por método de clusterização Euclidiana. A cor amarela indica CDSs presentes uma única vez no cromossomo; a cor laranja indica mais de uma CDS codificante para a mesma proteína e a cor preta indica a ausência da CDS. B – Diagrama de Venn das 1.118 CDSs pertencentes ao grupo dos *core,* acessórios e únicos dos cromossomos em estudo.

Antes de analisar o conteúdo gênico de cada grupo destacado na **Figura 3.6**, questionou-se se as CDSs identificadas como únicas ou acessórias eram artefatos de anotação ou curadoria. Nesse caso, devido fatores intrínsecos ao processo de anotação pelo PATRIC somado à ausência/presença de nucleotídeos que interfiram no quadro de leitura, CDSs poderi-

am não ser anotadas em alguns replicons. De igual forma, CDSs poderiam ser erroneamente identificadas, quando deveriam ser classificadas como pseudogenes. Um exemplo de desvios no processo de anotação do PATRIC é evidenciado na **Figura 3.7**. Nela, pode-se observar a ausência da CDS "B" na anotação do cromossomo da *E. coli* BH100 subtr. MG2014 e a ausência da CDS "C" no cromossomo na *E. coli* BH100 subtr. MG2017. Tal fato é inerente à metodologia de anotação do PATRIC, uma vez que a sequência nucleotídica dessa região é idêntica em ambos os cromossomos.

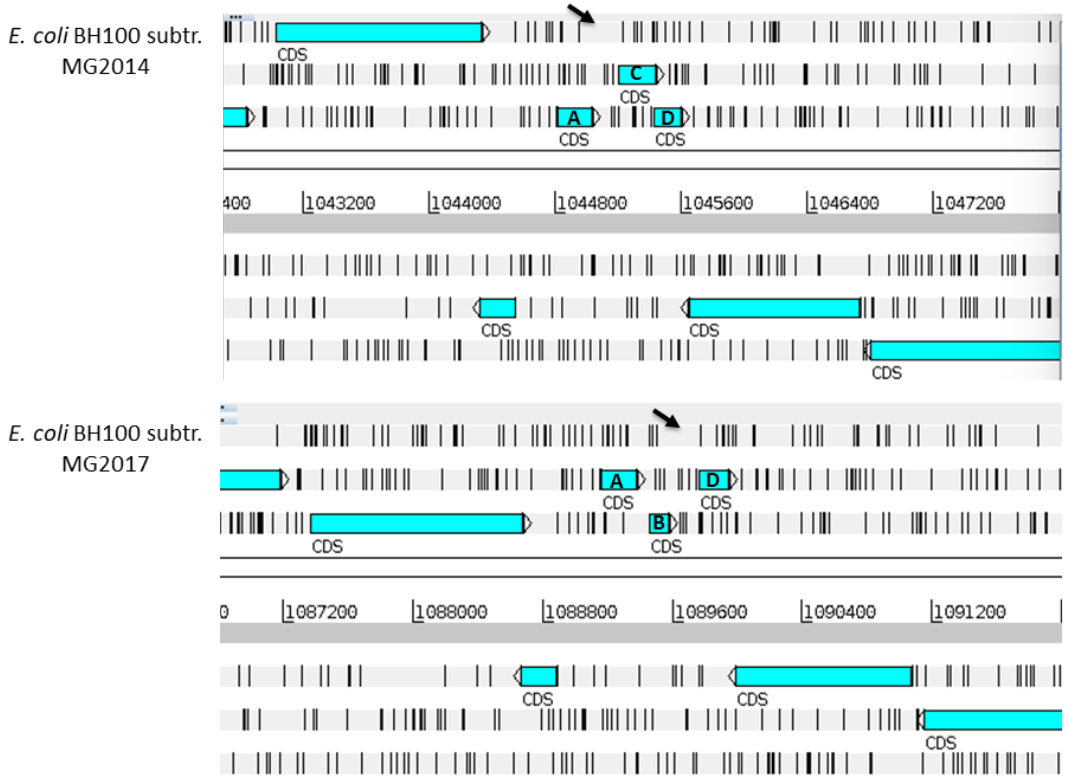


Figura 3.7 Exemplo de incoerências no processo de anotação do PATRIC. A região compreendida entre as CDSs A e D em ambas as linhagens são idênticas entre si. As flechas indicam local em que seria previsto a anotação da CDS **B** na *E. coli* BH100 subtr. MG2014 e da CDS **C** na *E. coli* BH100 subtr. MG2017.

De modo a assegurar maior acurácia na comparação do conteúdo gênico de cada linhagem, as CDSs previamente pertencentes ao grupo dos acessórios ou únicos foram investigadas. Para tanto, utilizou-se a ferramenta tBLASTn, de modo a identificar CDSs não anotadas nos demais replicons cromossomicos. Com isso, foi possível corrigir CDSs falso-negativas e reestruturar cada grupo (*core*, acessório e *singletons*) (**Figura 3.8**).

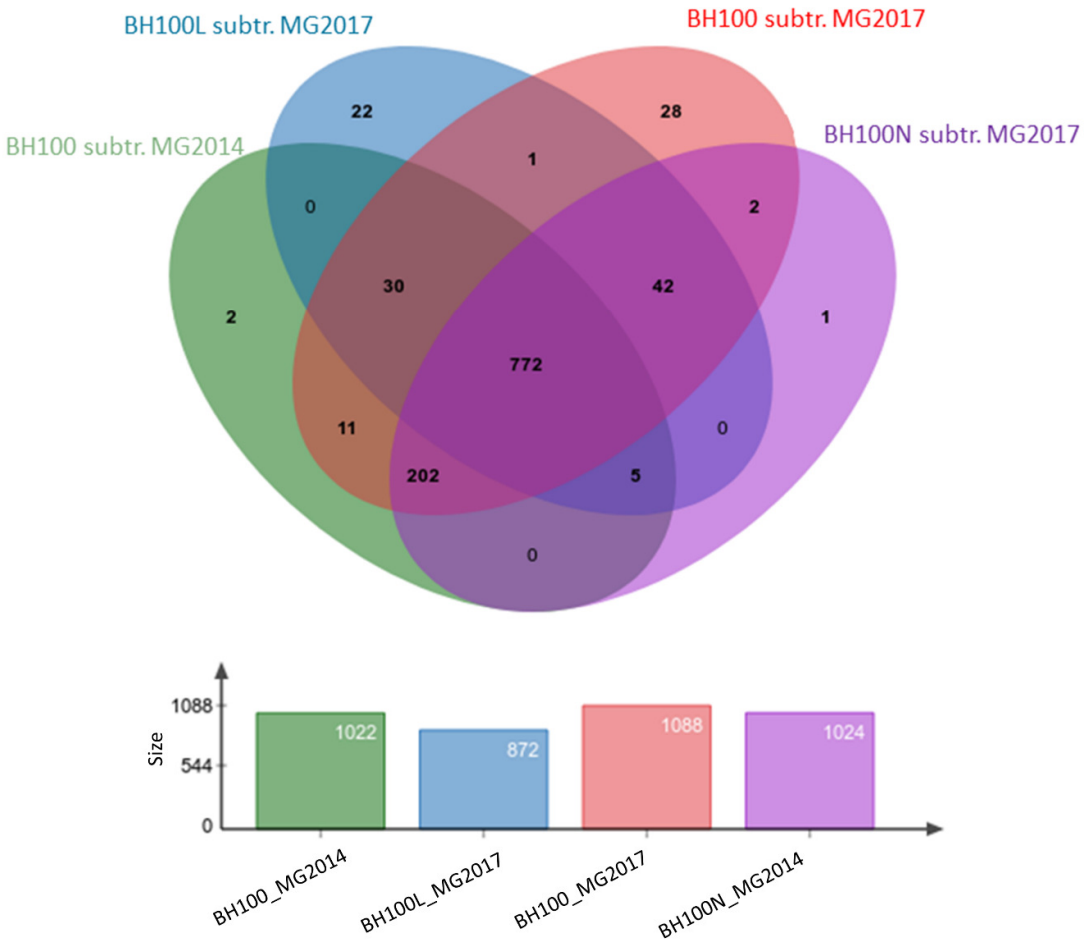


Figura 3.8 Novo diagrama de Venn após identificação de CDSs falsamente preditas como ausentes nas diversas linhagens desse estudo.

Pode-se notar diferenças consideráveis em termos de conteúdo gênico ao longo dos diferentes cromossomos estudados nesse trabalho. Essa diferença deve-se, primariamente, ao efeito dos elementos móveis, em especial aos IS da família IS3. A fim de compreender os possíveis efeitos da ausência de certos genes nas diferentes linhagens, foi coletado o conjunto de proteínas não presente em determinada linhagem, mas presente em, pelo menos, outros dois replicons cromossomicos para caracterização funcional no GoFeat.

Diante do fato de nem todas as CDSs estudadas apresentarem termos GO associados a processos biológicos, úteis para identificação funcional em um contexto metabólico da proteína em estudo, as demais CDSs foram categorizadas em diferentes grupos: 1) proteínas anotadas como hipotéticas ou sem anotação pelo Uniprot e sem anotação de domínios proteicos conhecidos e associação com termos GO; 2) proteínas anotadas como hipotéticas pelo Uniprot mas que apresentassem informação de domínios proteicos e/ou termos GO não associados a processos biológicos; 3) proteínas anotadas como não hipotéticas pelo Uniprot e não contendo

termos GO associados; 4) proteínas anotadas como não hipotéticas pelo Uniprot e não contendo, apenas, termos GO associados à processos biológicos.

No primeiro caso, as proteínas identificadas como hipotéticas podem se referir a produtos exclusivos das demais linhagens em comparação, mas não presentes na linhagem investigada, ou podem se referir a artefatos intrínsecos do processo de predição de CDSs, ou, ainda, a proteínas reais ainda não caracterizadas em nenhum organismo, onde a proteína já fora predita/observada. De toda forma, as proteínas que satisfizeram esse critério não foram consideradas para efeitos comparativos. No segundo caso, optou-se por investigar as bases do Interpro e Pfam, de modo a buscar pistas que pudessem contextualizar essas proteínas em um determinado processo biológico. Já no terceiro e quarto casos, optou-se pela investigação nas bases do Interpro e Pfam, além da literatura, baseado na nomenclatura anotada no Uniprot, por uma possível inferência do processo biológico associado à proteína. A **Tabela 3.1** indica o número de sequências não pertencentes a uma determinada linhagem e distribuídas em cada uma das categorias descritas anteriormente.

Tabela 3.1 Distribuição das CDSs ausentes nos cromossomos de cada linhagem e presentes em pelo menos dois outros replicons, segundo as categorias descritas no texto.

Linhagem	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Processos Biológicos associados
E. coli BH100 subtr. MG2014	5	3	5	11	21
E. coli BH100 subtr. MG2017	2	0	1	1	5
E. coli BH100L subtr. MG2017[a]	27	13	15	46	61
E. coli BH100N subtr. MG2017	1	0	4	17	20

Categoria 1: proteínas anotadas como hipotéticas ou sem anotação pelo Uniprot, sem anotação de domínios proteicos conhecidos e sem associação com termos GO; Categoria 2: proteínas anotadas como hipotéticas pelo Uniprot, mas que apresentam informação de domínios proteicos e/ou termos GO não associados a processos biológicos; Categoria 3: proteínas anotadas como não hipotéticas pelo Uniprot e não contendo termos GO associados; Categoria 4: proteínas anotadas como não-hipotéticas pelo Uniprot e não contendo, apenas, termos GO associados à processos biológicos.

[a] Além das CDSs distribuídas nas diferentes categorias aqui descritas também estão ausentes nessa linhagem 52 genes associados ao profago lambda.

Todas regiões codificantes pertencentes às categorias 2, 3 e 4 foram classificadas de modo a sugerir uma terminologia semelhante as dos termos GO para processos biológicos, e foram aqui denominados termos GSO (*Gene Side Ontology*). Com isso, foi possível investigar de forma sistemática o perfil funcional e possíveis implicações metabólicas que a ausência dessas CDSs pode provocar.

Em relação ao conjunto das 45 CDSs não presentes na *E. coli* BH100 subtr. MG2014, destacam-se genes associados à transcrição e ao transporte de compostos pela membrana. Em geral, os genes ausentes nessa linhagem participam de processos metabólicos que confeririam certa vantagem adaptativa, sob determinadas condições. Exemplo disso são os genes responsáveis pelo uso de etanolamina, como fonte de carbono e nitrogênio, além daqueles associados à formação de metabolossomos, estrutura similar aos carboxissomos, provavelmente úteis no acúmulo de enzimas metabólicas e na eliminação de compostos tóxicos (BRINSMADE; PALDON; ESCALANTE-SEMERENA, 2005; KOFOID et al., 1999). Também se revela a ausência de CDSs relacionadas à neutralização de peptídeos antimicrobianos, como é o caso de uma proteína de membrana predita como hipotética, mas, possivelmente, capaz de contribuir com a presença de cargas positivas na membrana, devido a um domínio de transferência de lisina ao fosfatidilglicerol. A **Figura 3.9** revela os termos GO e GSO mais associados às CDSs ausentes nessa linhagem.

O cromossomo da *E. coli* BH100 subtr. MG2017 é o que menos possui perdas de CDSs acessórias, quando comparado às outras linhagens. Os genes ausentes nessa linhagem, mas presentes em pelo menos dois outros cromossomos das outras, resumem-se a um gene associado ao metabolismo de trealose, e a uma proteína de membrana associada ao transporte de aminoácidos para o meio intracelular (**Figura 3.10**). Destaca-se também a ausência do gene da subunidade β da galactosidase, que pode estar associada a não metabolização de galactose e/ou à formação de biofilme (CHAI et al., 2012).

Já o cromossomo da *E. coli* BH100L subtr. MG2017 apresenta a maior perda de CDSs acessórias quando comparado com os demais cromossomos. Esse fato deve-se, basicamente, à inserção do Tn21 em dois locais distintos no cromossomo e à perda do prófago Entero_lambda. Dentre os genes perdidos, destacam-se aqueles ligados ao controle da expressão gênica, transporte de nutrientes e metabolismo em geral (carboidratos, lipídios, nucleotídeos e proteínas) (**Figura 3.11**). Também, vale destacar a ausência de genes associados à transdução de sinal, o que pode sugerir dificuldades dessa linhagem em adaptar-se sob determinadas condições ambientais, quando comparada com as demais. Como exemplo desses genes, pode-se citar o gene *phoQ*, responsável por estimular a atividade de *phoP*, associado a diversos mecanismos de virulência, quando em baixas concentrações de Mg^{2+} (BEIER; GROSS, 2006; TIWARI et al., 2014).

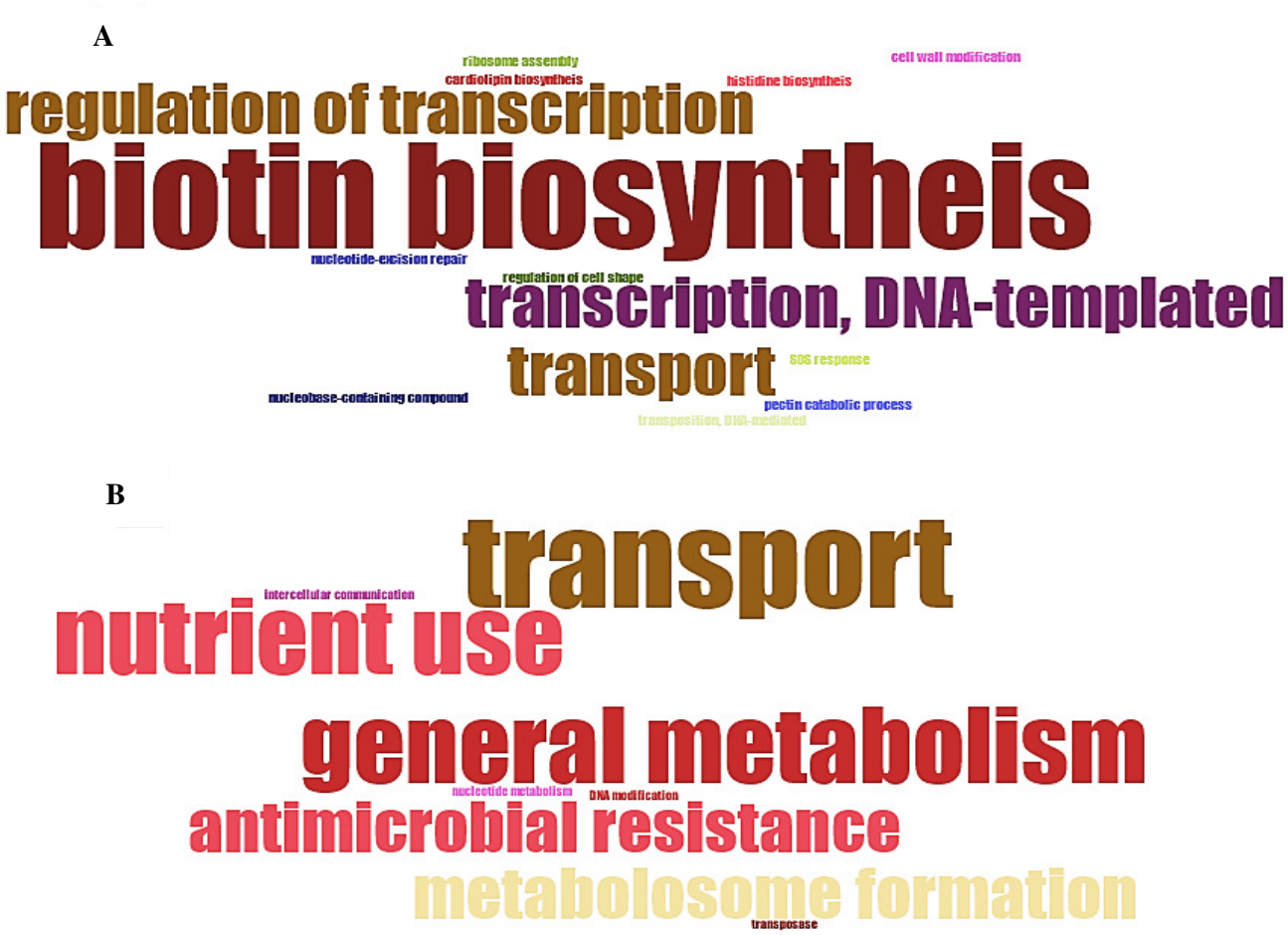


Figura 3.9 Termos GO (A) e GSO (B) associados a processos biológicos vinculados a CDSs ausentes no cromossomo da *E. coli* BH100 subtr. MG2014 e presentes em pelo menos outros dois replicons cromossomais.

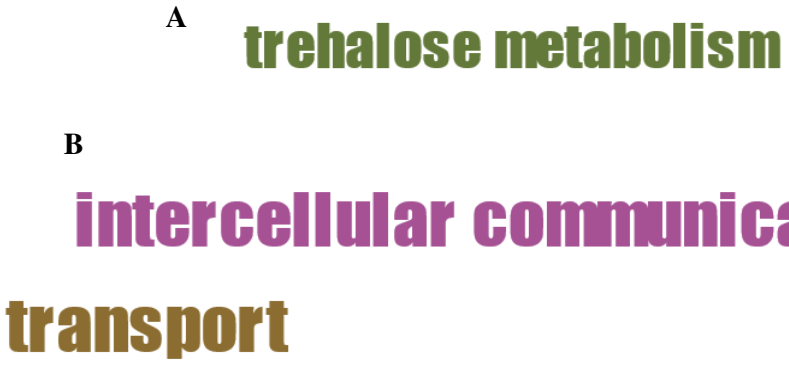
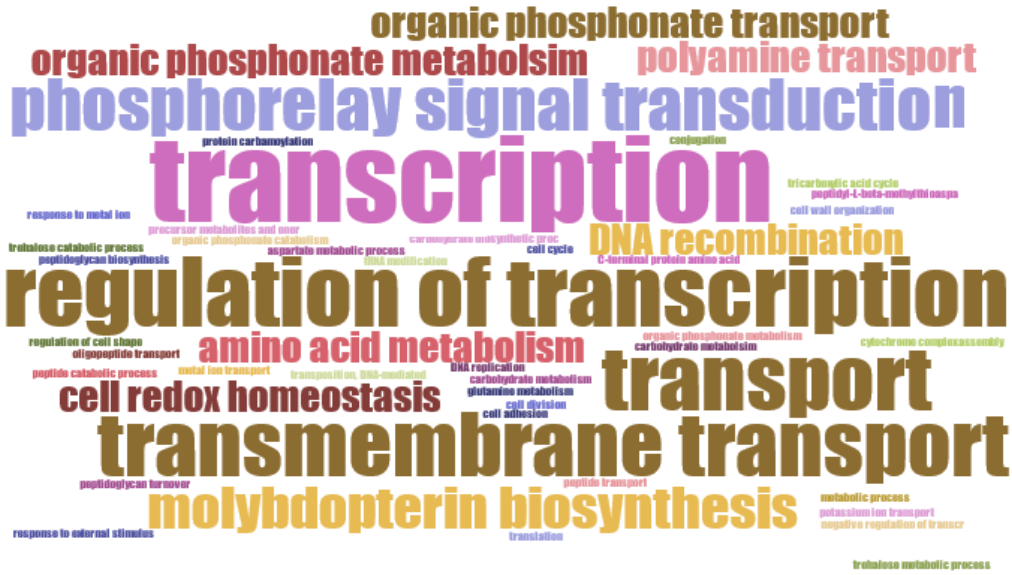


Figura 3.10 Termos GO (A) e GSO (B) associados a processos biológicos vinculados a CDSs ausentes no cromossomo da *E. coli* BH100 subtr. MG2017 e presentes em pelo menos outros dois replicons cromossomais.

A



B

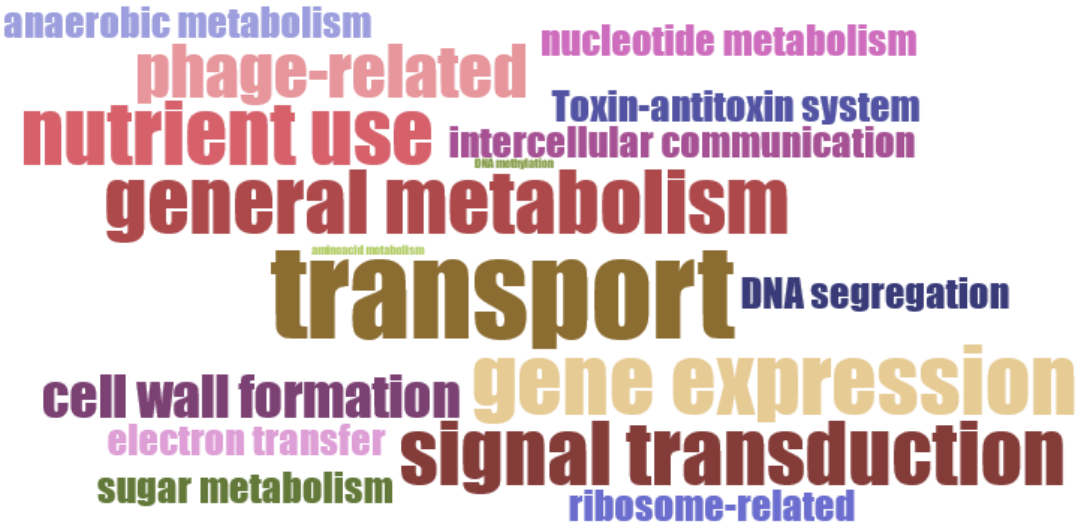


Figura 3.11 Termos GO (A) e GSO (B) associados a processos biológicos vinculados a CDSs ausentes no cromossomo da *E. coli* BH100L subtr. MG2017 e presentes em pelo menos outros dois replicons cromossomais

Por fim, as 42 CDSs acessórias, ausentes na *E. coli* BH100N subtr. MG2017, estão, em sua maioria, associadas à transcrição, transporte e metabolismo em geral (**Figura 3.12**). Evidencia-se também genes associados à construção da parede celular e a genes associados ao operon *phn*, vinculado ao uso de compostos fosfonados, como fonte de carbono e fósforo.

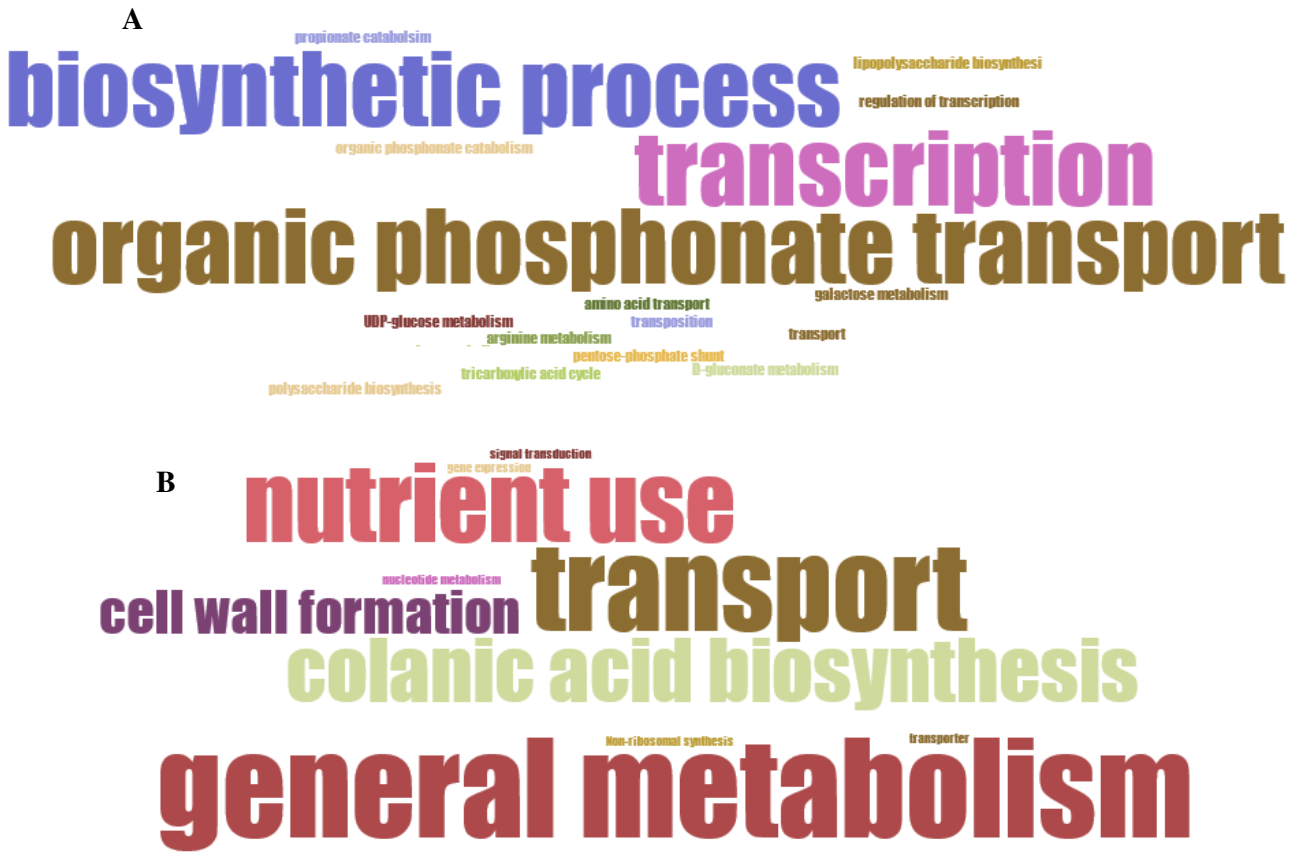


Figura 3.12 Termos G.O (A) e G.S.O (B) associados a processos biológicos vinculados a CDSs ausentes no cromossomo da *E. coli* BH100N subtr. MG2017 e presentes em pelo menos outros dois replicons cromossomais

Já em relação aos genes exclusivos de cada linhagem, destacam-se os genes codificadores da transposase de IS5 e de uma proteína hipotética na *E. coli* BH100 subtr. MG2014. Na *E. coli* BH100N subtr. MG2017, verifica-se a presença exclusiva de um gene hipotético. Já, em relação à *E. coli* BH100L subtr. MG2017, todos os genes, salvo um predito como hipotético, pertencem ao Tn21, inserido no cromossomo. Por fim, dos genes exclusivos da *E. coli* BH100 subtr. MG2017 destacam-se um gene hipotético e a presença de uma região contendo genes associados ao operon *mdt*, associado à síntese de bombas de efluxo de drogas (NISHINO; NIKAIDO; YAMAGUCHI, 2007) e *phn*, além de outros associados à regulação transcricional e metabolismo, de uma forma geral (**Quadro 3.1**). Essa região, aparentemente exclusiva da *E. coli* BH100 subtr. MG2017 é verificada em outras estirpes de *E. coli*, mas encontra-se fragmentada nas demais linhagens da BH100, fruto da inserção de elementos móveis da família dos IS3.

Quadro 3.1 Relação de genes pertencentes a uma região exclusiva da *E. coli* BH100 subtr. MG2017.

Locus_tag (GenBank)	Nomenclatura Uniprot	Termos GO
BH100B_04581	Uncharacterized protein	-
BH100B_04582	ABC transporter family protein	GO:0005524 - ATP binding|GO:0016887 - ATPase activity
BH100B_04583	Oligopeptide ABC transporter ATP-binding protein	GO:0005524 - ATP binding|GO:0016887 - ATPase activity
BH100B_04584	ABC transporter permease	GO:0005886 - plasma membrane|GO:0006810 - transport|GO:0016021 - integral component of membrane
BH100B_04585	ABC transporter permease	GO:0005886 - plasma membrane|GO:0006810 - transport|GO:0016021 - integral component of membrane
BH100B_04586	ABC transporter substrate-binding protein	GO:0043190 - ATP-binding cassette (ABC) transporter complex|GO:0055085 - transmembrane transport
BH100B_04587	Formate dehydrogenase subunit alpha	GO:0008863 - formate dehydrogenase (NAD+) activity|GO:0009055 - electron transfer activity|GO:0015942 - formate metabolic process|GO:0030151 - molybdenum ion binding|GO:0051539 - 4 iron
BH100B_04588	Uncharacterized protein	-
BH100B_04589	Multidrug resistance outer membrane protein mdtP	GO:0005215 - transporter activity|GO:0006810 - transport|GO:0008289 - lipid binding|GO:0009279 - cell outer membrane|GO:0016021 - integral component of membrane
BH100B_04590	Multidrug transporter	GO:0005886 - plasma membrane|GO:0006810 - transport|GO:0016021 - integral component of membrane
BH100B_04591	Multidrug resistance protein mdtN	GO:0015238 - drug transmembrane transporter activity
BH100B_04592	Uncharacterized protein	GO:0016021 - integral component of membrane
BH100B_04593	Alkyl sulfatase	GO:0046983 - protein dimerization activity
BH100B_04594	D-allose kinase	GO:0005524 - ATP binding|GO:0008787 - allose kinase activity|GO:0019316 - D-allose catabolic process
BH100B_04595	Allulose-6-phosphate 3-epimerase	GO:0005975 - carbohydrate metabolic process|GO:0016857 - racemase and epimerase activity
BH100B_04596	Branched-chain amino acid transport system / permease component family protein	GO:0005215 - transporter activity|GO:0005886 - plasma membrane|GO:0006810 - transport|GO:0016021 - integral component of membrane
BH100B_04597	Allose ABC transporter ATP-binding protein	GO:0005524 - ATP binding|GO:0016887 - ATPase activity
BH100B_04598	D-allose-binding periplasmic protein	-

BH100B_04599	DNA-binding transcriptional repressor	GO:0003677 - DNA binding|GO:0003700 - DNA binding transcription factor activity|GO:0006351 - transcription
BH100B_04600	Ribose-5-phosphate isomerase B	GO:0004751 - ribose-5-phosphate isomerase activity|GO:0005975 - carbohydrate metabolic process|GO:0006098 - pentose-phosphate shunt
BH100B_04601	Uncharacterized protein	-
BH100B_04602	Protein phnP	GO:0008081 - phosphoric diester hydrolase activity|GO:0019700 - organic phosphonate catabolic process
BH100B_04603	Acyltransferase	GO:0008080 - N-acetyltransferase activity
BH100B_04604	Ribose 1	-
BH100B_04605	Phosphonate metabolism protein PhnM	GO:0016810 - hydrolase activity
BH100B_04606	Phosphonate transport ATP-binding protein PhnL	GO:0005524 - ATP binding|GO:0016887 - ATPase activity

3.2.1.4 ANÁLISE DE PSEUDOGENES

De modo a identificar possíveis alterações mutacionais nos cromossomos investigados, pretendeu-se avaliar o perfil de pseudogenes de cada replicon. Para tanto, a fim de melhor analisar o perfil de potenciais pseudogenes nas diferentes linhagens, foram removidas as CDSs correspondentes a transposases e a elementos de fagos. Com isso, foram observados 29 pseudogenes para a *E. coli* BH100 subtr. MG2014, 21 para a *E. coli* BH100 subtr. MG2017, 24 para a *E. coli* BH100L subtr. MG2017 e 36 para a *E. coli* BH100N subtr. MG2017. A distribuição dos diferentes tipos de pseudogenes encontrados seguindo essas condições pode ser observada na **Tabela 3.2**.

Tabela 3.2 Nova distribuição de CDSs anotadas como pseudogenes após remoção de *features* referentes a transposases e a elementos de fagos.

Cromossomo	*Frameshift*	Codon de parada prematuro	CDS parcial	Total
BH100 subtr. MG2014	23	6	0	29
BH100 subtr. MG2017	16	2	3	21
BH100L subtr. MG2017	17	5	2	24
BH100N subtr. MG2017	25	11	0	36

A análise dos possíveis pseudogenes, em particular, aqueles associados a mudanças no quadro de leitura e presença de códons de parada prematuros, pode indicar possíveis diferenças nas taxas de mutação dessas linhagens. A fim de avaliar esse possível efeito, utilizou-se o software BPGA para identificar as sequências proteicas advindas dos potenciais pseudogenes, pertencentes a todos os cromossomos (*core*), a mais de um, mas não a todos (acessórios) e os

exclusivos de cada replicon. Para tanto, utilizou-se as recomendações padrão do software, ajustando a similaridade proteica para 70%. A distribuição desses grupos pode ser vista na **Figura 3.13**.

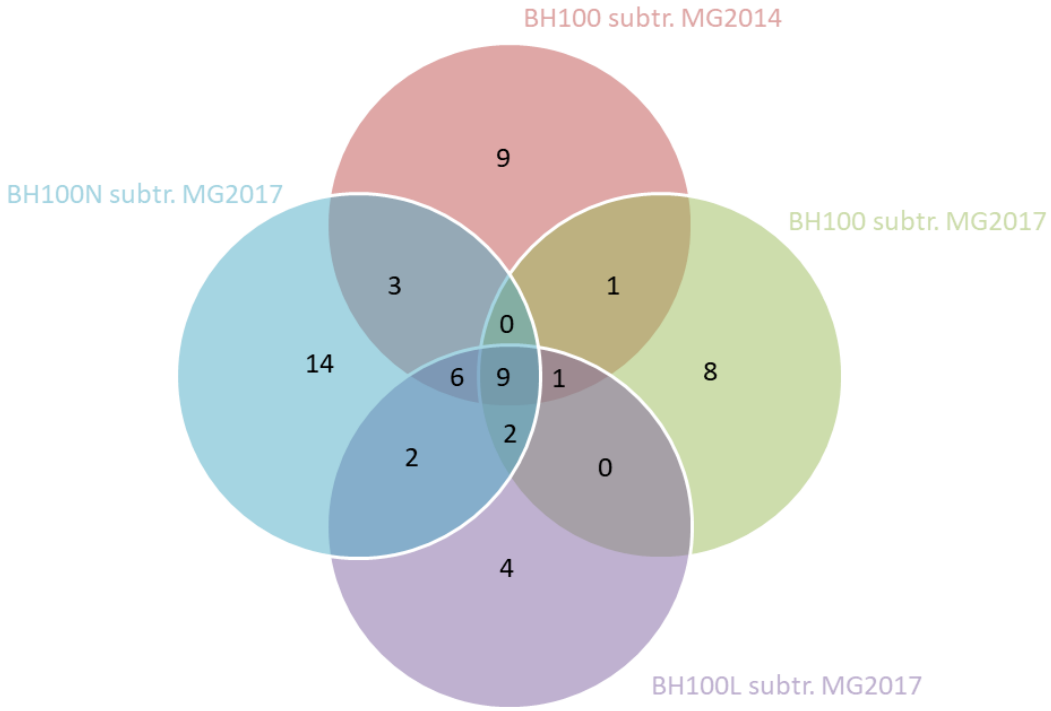


Figura 3.13 Diagrama de Venn indicando o número de *features* anotadas como pseudogenes, em comum a cada uma das linhagens da *E. coli* BH100 e suas variantes.

Nota-se da **Figura 3.13** uma distribuição assimétrica em relação à presença/ausência dos potenciais pseudogenes, ao longo de cada grupo. A fim de confirmar essa distribuição, adotou-se a seguinte metodologia para fins de refinamento: 1) identificação, via tBLASTn, das CDSs não identificadas como pseudogenes, mas homólogas às sequências pertencentes aos grupos acessórios. 2) identificação das CDSs não identificadas como pseudogenes nos demais cromossomos, mas preditas como pseudogenes exclusivos em um determinado replicon. 3) avaliação dessas CDSs, identificando a) falhas na anotação de *frameshifts*, b) potencial redução na extensão da CDS, corrigindo para códons de parada prematuro, c) potencial extensão/redução de CDSs identificadas como parciais, d) reavaliação do mapeamento das *reads* buscando por nucleotídeos passiveis de serem inseridos/deletados e e) identificação de transposases que possam estar associadas à fragmentação de CDSs.

Utilizando dos critérios acima, foi possível identificar 6 anotações imprecisas na *E. coli* BH100 subtr. MG2014, 18 na *E. coli* BH100 subtr. MG2017, 7 na *E. coli* BH100L subtr.

MG2017 e 6 na *E. coli* BH100N subtr. MG2017. Além disso, também foram identificadas duas sequências de transposases da família dos IS3, inseridas entre 2 CDSs anotadas na *E. coli* BH100L subtr. MG2017, resultando na identificação de CDSs parciais. Já as CDSs parciais, outrora preditas na *E. coli* BH100 subtr. MG2017, foram erroneamente assinaladas, uma vez que a redução de seus comprimentos resultaria em CDSs idênticas às anotadas nos demais cromossomos, bem como compatíveis às presentes na base do Uniprot.

A fim de observar tão somente as sequências possivelmente não funcionais no genoma, dois filtros adicionais foram utilizados: 1) a presença de cópias anotadas como funcionais; 2) a ausência de mecanismos descritos de recodificação. No primeiro caso, propôs-se remover as sequências, preditas como pseudogenes, mas que apresentavam cópias aparentemente funcionais em outro locus. Esse foi o caso dos genes codificantes para DNA polimerase II e das proteínas RapA, PapG e PapC da *E. coli* BH100 subtr. MG2014, da subunidade do transportador de efluxo de multidrogas RND da *E. coli* BH100 subtr. MG2017, da adesina Ag43 da *E. coli* BH100L subtr. MG2017 e da fosfoporina PhoE da *E. coli* BH100N subtr. MG2017. Já no segundo caso, observou-se mecanismos de recodificação, onde o mRNA gerado poderia levar à presença de uma proteína funcional, ainda que sua sequência codificante fosse predita como pseudogene. Em procariotos, destacam-se dois mecanismos desse tipo: *ribosomal shifting* e leitura de TGA.

No *ribosomal shifting*, o ribossomo é capaz de alterar o quadro de leitura, gerando proteínas menores ou maiores do que as preditas. Nesse mecanismo, *motifs* específicos na sequência do mRNA podem induzir a esse feito. O FSFinder é um software que se propõe a identificar potenciais sítios de indução na alteração do quadro de leitura. Dentre uma das formas de predição, destaca-se o cenário induzido pelo *motif* NNNWWWH, seguido de uma sequência espaçadora de 4 – 11 pb e de um *stem loop* ou *pseudoknot*. Prevê-se que essa estrutura favoreça a leitura do *frame* -1. No entanto, essa ferramenta não indicou nenhum caso de pseudogene, predito via *frameshift*, como sendo capaz de sofrer recodificação.

Em relação ao segundo mecanismo de recodificação, é sabido que em *E. coli* o tRNA carreador da selenocisteína apresenta em seu anticódon sequência capaz de parear com o tradicional códon de parada TGA. Com efeito, CDSs, preditas como pseudogenes, por conterem esse códon de parada no seu interior, poderiam dar origem a um mRNA funcional. Esse é o caso verificado para a formato desidrogenase H, anotada exclusivamente na *E. coli* BH100 subtr. MG2017 e da transglicosilase lítica F da *E. coli* BH100N subtr. MG2017.

Dessa forma, um novo diagrama de Venn foi construído, de forma a contabilizar todos os filtros adotados, sendo identificados 24 potenciais pseudogenes comuns a todas às linhagens

(**Figura 3.14**). O gene de uma desidrogenase comum às linhagens *E. coli* BH100 subtr. MG2014, *E. coli* BH100N subtr. MG2017 e *E. coli* BH100L subtr. MG2017 não foi encontrado na *E. coli* BH100 subtr. MG2017, por estar, aparentemente, presente em um elemento móvel, não encontrado nessa linhagem. Dois potenciais pseudogenes são encontrados em todas as linhagens, salvo na *E. coli* BH100 subtr. MG2014. Um desses genes é referente ao *prfB* (*peptide chain release fator* 2), que é classicamente conhecido por sofrer recodificação, favorecendo à leitura do *frame* +1 pelo ribossomo (DINMAN, 2006). O segundo gene é referente a uma proteína ligante de ATP, associada ao transportador do tipo ABC e, embora esteja anotada como gene na *E. coli* BH100 subtr. MG2014, seu ponto de predição de *frameshift* nos demais replicons encontra-se em uma região homopolimérica. Por essa linhagem ter sido sequenciada pela plataforma Ion Torrent, a predição dessa região, como sendo uma autêntica CDS, pode ser falso-positiva. Já a sequência compartilhada apenas entre os replicons da *E. coli* BH100 e referente ao transportador de C4-dicarboxilato parece corresponder a um gene funcional na *E. coli* BH100N subtr. MG2017, mas não se encontra presente na E. coli BH100L subtr. MG2017, por, possivelmente, ter sido substituída por um elemento móvel do tipo IS3 (**Figura 3.15**).

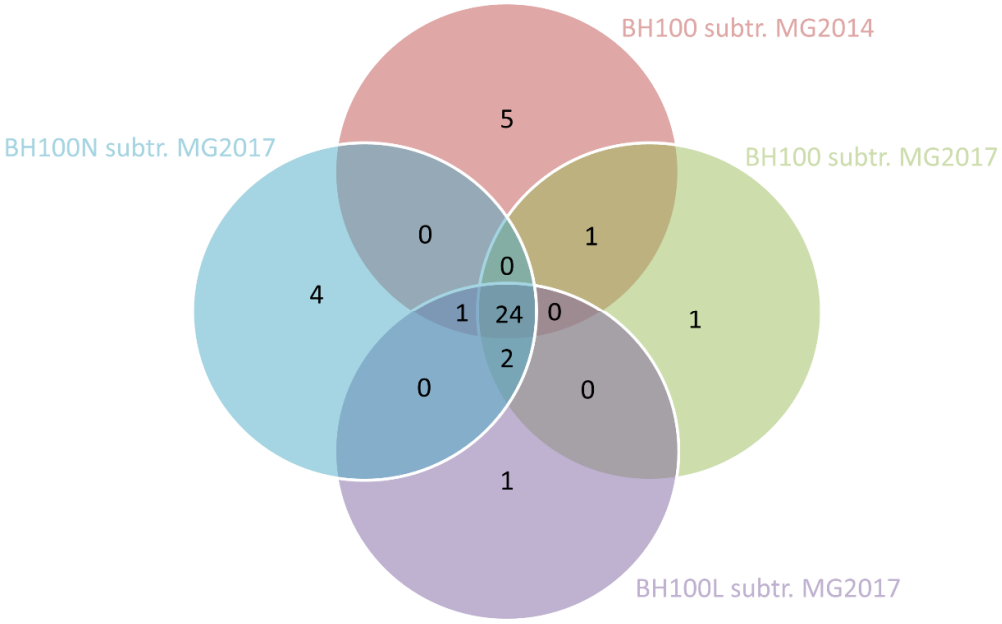


Figura 3.14 Novo diagrama de Venn, após refinamento das anotações de pseudogenes e aplicação dos filtros para *Recoding* e duplicidade gênica

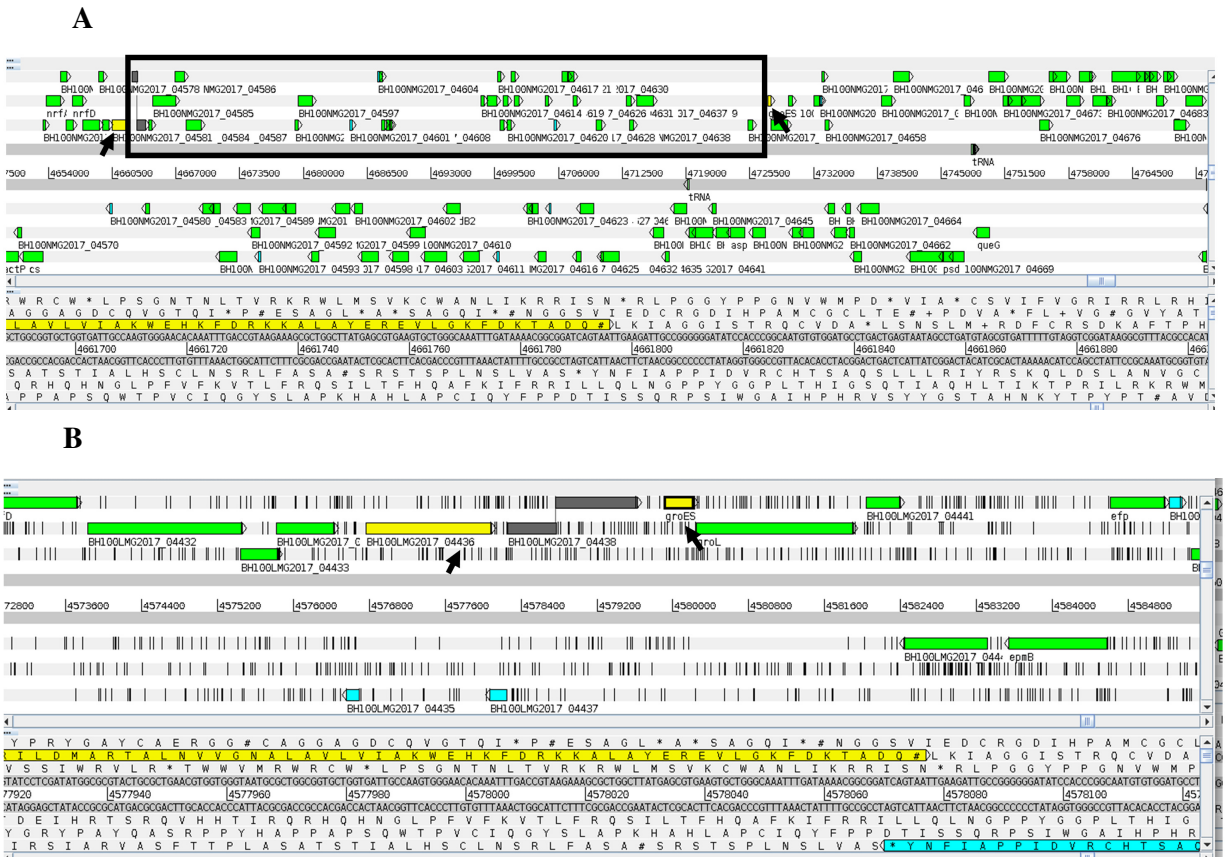


Figura 3.15 Região do gene codificador do transportador de C4-dicarboxilato. Essa região na *E. coli* BH100N subtr. MG2017 (A) parece ter sido substituída por um elemento móvel da família dos IS3, na E. coli BH100L subtr. MG2017 (B). As setas apontam para os mesmos genes no flanco esquerdo e direito, destacados em amarelo em A e B.

Os potenciais pseudogenes, associados exclusivamente a cada uma das linhagens, bem como a causa de sua predição como não funcionais, podem ser observados no **Quadro 3.2**. Em relação às sequencias oriundas da *E. coli* BH100 subtr. MG2014, vale destacar que as regiões preditas de mudanças no quadro de leitura ocorrem em regiões relativamente ricas em conteúdo GC. Esse fato pode sugerir vieses nas *reads* provenientes de Illumina, o que levaria a falsas predições de CDSs funcionais nas demais linhagens. Se esse fosse o caso, seria esperado um aumento no número de potencias pseudogenes comuns a todas as linhagens.

Por fim, levando em conta todos os filtros aplicados e as discussões aqui levantadas, não parece haver diferenças relevantes em termos de potenciais pseudogenes que corroborem a possibilidade de alterações significativas das taxas mutacionais dessas linhagens. Com efeito, ao se levar em conta que todas essas linhagens foram utilizadas de forma constante e diferencial ao longo dos anos e, ainda, considerando as diferentes estratégias de sequenciamento das subcepas MG2014 e MG2017 seria razoável esperar maiores diferenças nos potenciais pseudogenes oriundos de *indels* e substituições para surgimento de códons de parada prematuros.

Dito isso, os resultados aqui apresentados parecem não sugerir que os plasmídeos interfiram, de alguma forma, com mecanismo associado à fidedignidade da replicação do DNA, conforme previamente sugerido (DE PAULA, 1989).

Quadro 3.2 Relação de potenciais pseudogenes exclusivos aos diferentes cromossomos em análise.

Linhagem	Locus_tag (GenBank)	Produto	Razão
BH100 subtr. MG2014	BH100MG2014_03037	*hypothetical protein*	Deleção de G
	BH100MG2014_03288	*hypothetical protein*	Deleção de G
	BH100MG2014_03752	*3-dehydroquinate synthase*	Deleção de G
	BH100MG2014_03805	*glycogen debranching enzyme*	Deleção de C
	BH100MG2014_04136	*Phosphate transport system permease protein PstA*	Deleção de G
BH100 subtr. MG2017	BH100B_05099*	*Nicotinamide-nucleotide adenylyltransferase*	Substituição G:T
BH100L subtr. MG2017	BH100LMG2017_02850	*two-component sensor histidine kinase BarA*	Deleção de 7 pb
BH100N subtr. MG2017	BH100NMG2017_00300	*colicin V synthesis protein*	Deleção de A
	BH100NMG2017_00930	*molybdopterin molybdotransferase*	Inserção de G
	BH100NMG2017_02964	*RNA polymerase sigma factor RpoS*	Inserção de A
	BH100NMG2017_04406	*HTH-type transcriptional repressor CytR*	Deleção de 14 pb

* Presença de códon de parada prematuro.

3.2.2 A E. coli BH100 num contexto patogênico

A *E. coli* BH100, por ter sido isolada de urina, é tida como uma potencial linhagem uropatogênica. Levando-se em conta que essa linhagem tem sido sujeita a um nicho laboratorial desde a data de seu isolamento até os dias atuais (Informação verbal)[23], parece razoável supor que genes associados à patogenicidade possam ter se perdido, por ausência de seleção positiva. Disso posto, procedeu-se às seguintes estratégias *in silico*, adotadas com a finalidade de inferir o potencial patogênico[24] da *E. coli* BH100: contextualização filogenética; análise de ilhas genômicas; presença de genes de virulência.

[23] Nascimento, A.M.A. (Instituto de Ciência Biológicas, UFMG). Comunicação pessoal, 2017.
[24] Por potencial patogênico, me refiro à capacidade da *E. coli* BH100 atual ou de sua ancestral (datada do isolamento) de ser patogênica.

3.2.2.1 CONTEXTUALIZAÇÃO FILOGENÉTICA

O contexto filogenético dos cromossomos estudados nesse trabalho sustenta a hipótese da *E. coli* BH100 ser de fato uma uropatogênica. Utilizando o software Gegenees, foi possível construir uma matriz de distâncias, que serviu de base para construção de uma árvore filogenética, associada à similaridade/distância genômica dos cromossomos de 416 linhagens de *E. coli* em relação à *S. enterica* e *E. fergusonii*, utilizadas como grupo externo. Devido à natureza das análises do Gegenees, pode-se dizer que as linhagens mais próximas entre si, na árvore, revelam genomas mais similares em relação ao conteúdo nucleotídico, bem como ao nível de sintenia. Esse fato, portanto, indica que a *E. coli* BH100 seja similar a outras linhagens componentes de um grupo predominantemente constituído de comensais e ExPECs (**Figura 3.16**).

Em particular, nota-se a presença de um subgrupo imediatamente mais próximo da BH100 e formado pelas linhagens ECONIH2 (RefSeq: NZ_CP0114667.1), isolada de ambiente hospitalar no *National Institute of Health* (NIH) e vinculada à resistência contra carbapenos (CONLAN et al., 2014; HARDIMAN et al., 2016); 536 (RefSeq: NC_008253.1), isolada de paciente com infecção do trato urinário (MIDDENDORF et al., 2004); e K-15KW01 (RefSeq: NZ_CP016358.1), uma linhagem uropatogênica carreadora de gene codificante da β-lactamase do grupo CTX-M (ZURFLUH; TASARA; STEPHAN, 2016). Pode-se identificar ainda, uma linhagem mais distante, mas inserida em um maior subgrupo, referente à linhagem 317 (RefSeq: NZ_CP023357.1), isolada de cães com infecção do trato urinário (BioProject: PRJNA402083). Essas observações, com efeito, sugerem que a *E.* coli BH100, não somente seja similar a outras linhagens de *E. coli* patogênicas, como também seja uma autêntica UPEC.

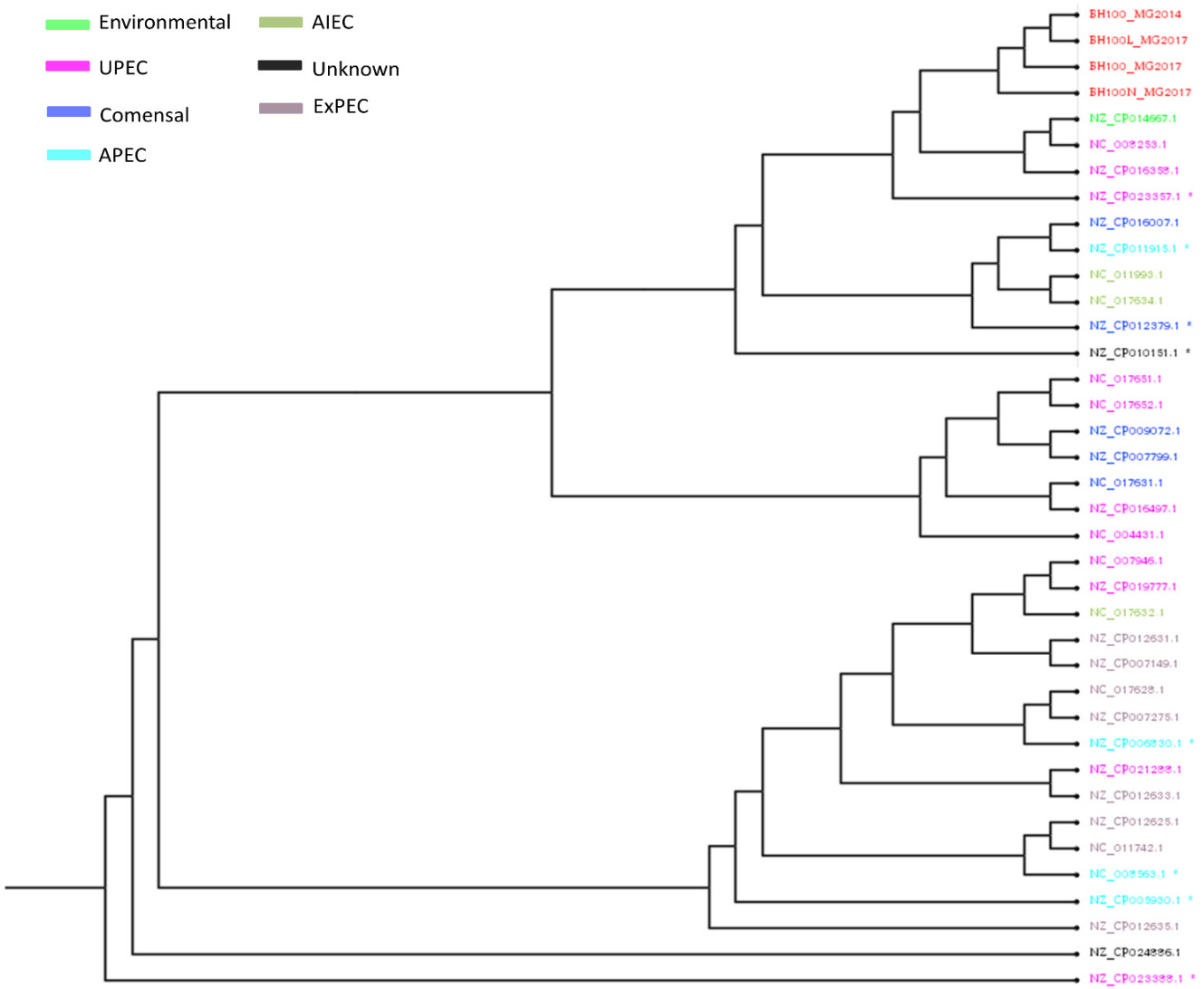


Figura 3.16 Contexto filogenético da *E. coli* BH100. As linhagens marcadas com * são de origem animal. As linhagens classificadas como ExPEC referem-se à NMEC e outras linhagens isoladas de amostras sanguíneas.

O grupo filogenético em que a BH100 está incluída (**Figura 3.16**), não somente indica a presença de linhagens UPEC, mas também revela a presença de linhagens não-patogênicas junto a linhagens patogênicas, incluindo ExPEC e exemplares de AIEC. Essas observações, entretanto, são esperadas. A presença de linhagens não-patogênicas encontra-se vinculada ao surgimento de linhagens ExPEC, fruto da evolução de linhagens comensais (MULVEY, 2002). Essas linhagens podem alcançar nichos ambientais ou animais e apresentar considerável similaridade de fatores de virulência com UPEC, podendo inclusive servir de fonte de contaminação (ROCHA et al., 2017; SINGER, 2015). Já a presença de linhagens AIEC, patógenas do trato gastro-intestinal, encontra-se associada a linhagens UPECs, pelo compartilhamento de diversos fatores de virulência associados à aderência e invasão celular (CONTE et al., 2016).

3.2.2.2 ANÁLISE DE ILHAS GENÔMICAS

A observação de ilhas genômicas pode fornecer pistas que indiquem os potenciais fenótipos, associados aos microrganismos que as carregam. Isso ocorre mediante o fato de genes associados a essas ilhas poderem conferir vantagens adaptativas frente à resistência a antibióticos, conquista de nichos, melhora do *fitness* e/ou patogenicidade. Diante disso, avaliou-se as ilhas genômicas anotadas pelo IslandPath-DIMOB e compiladas no IslandViewer4 da *E. coli* BH100 subtr. MG2017, utilizada por ter sido a linhagem com menos perda de genes, quando comparada às demais. Essa análise resultou na predição de 10 ilhas genômicas e três ilhas de patogenicidade (**Figura 3.17**).

De forma geral, parece haver conservação das ilhas nos diferentes replicons cromossomicos da *E. coli* BH100. As diferenças notáveis referem-se às ilhas GI3 e GI7, em relação ao cromossomo da *E. coli* BH100L subtr. MG2017, devido a perdas de regiões, em função da inserção do Tn21. Também, pode-se notar que algumas dessas ilhas estão presentes em regiões de prófagos. Esse é o caso de GI3, presente no prófago Entero_lambda; GI4, inserido no prófago Salmon_Sen34; e GI10 e GI11 dispostos no profago Entero_mEp460. Essas observações destacam a importância de bacteriófagos e transposons, como veículos de genes capazes de somar/subtrair a composição genética de um microrganismo.

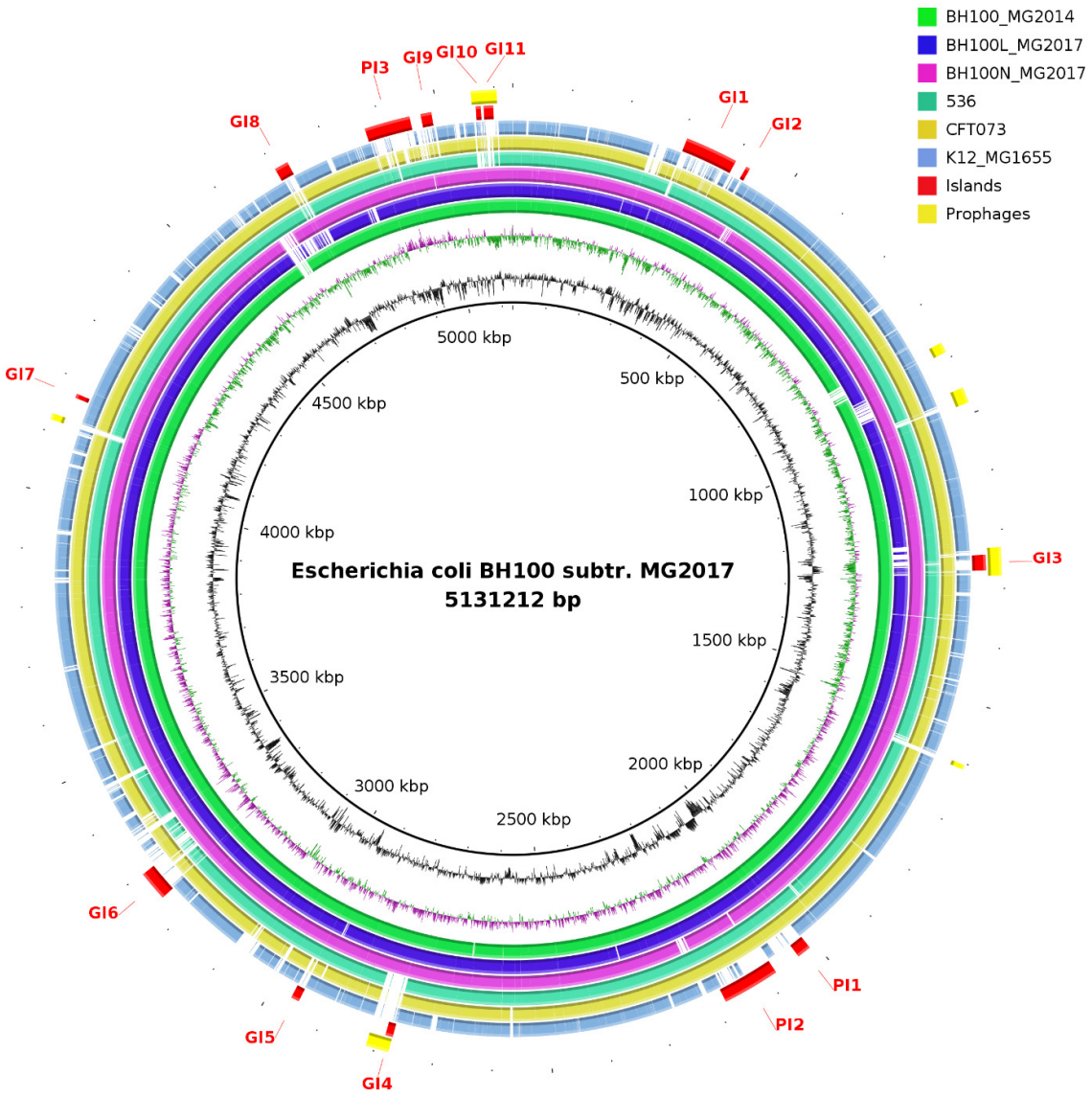


Figura 3.17 Ilhas de patogenicidade (PI) e genômicas (GI) preditas pelo IslandPath-DIMOB e compiladas no IslandViewer4 da *E. coli* BH100 subtr. MG2017 em comparação com os demais cromossomos das variantes da *E. coli* BH100, além das linhagens uropatogênicas *E. coli* 536 e *E. coli* CFT073, e da comensal *E. coli* K-12 subtr. MG1655.

O estudo de ilhas genômicas e, em especial, das ilhas de patogenicidade tem servido de grande valia na compreensão de mecanismos de patogenicidade, associados às bactérias uropatogênicas. Nesse cenário, destacam-se a *E. coli* CFT073, associada a casos de pielonefrite e sepse, e *E. coli* 536, associada a casos de pielonefrite, como modelos de uropatogênese, amplamente estudados em relação às suas ilhas genômicas (VEJBORG et al., 2011). Desse modo, objetivou-se investigar a presença de ilhas identificadas nesses modelos nas variantes da *E. coli* BH100. Para tanto, as ilhas preditas pelo IslandPath-DIMOB e compiladas no IslandViewer4 da *E. coli* 536 (**Figura 3.18**) e *E. coli* CFT073 (**Figura 3.19**) foram investigadas. Esses genomas, também, foram submetidos ao PHASTER para identificação de regiões de prófagos, de modo a identificar ilhas genômicas, even-

tualmente adquiridas por esses elementos. Essas análises resultaram na identificação de 10 ilhas genômicas, três ilhas de patogenicidade e uma ilha de resistência a antimicrobianos (RI) na *E. coli* 536. Dessas ilhas, observou-se a presença de GI3 e GI4, inseridas no prófago Entero_lambda e a ilha GI9, carreada pelo prófago Pseudo_phi3. Já para a *E. coli* CFT073 foram identificadas oito ilhas genômicas e sete ilhas de patogenicidade. Dessas ilhas, destacam-se a presença de PI1, potencialmente vinculado ao prófago Klebsi_vB_KpnP_KpV475; GI3, inserido no prófago Salmon_Fels_2; GI5, presente no prófago Entero_lambda; GI6 e PI3, também inseridos em outro prófago adjacente do tipo Entero_lambda, e, por fim, GI7 inserido no prófago Entero_HK629.

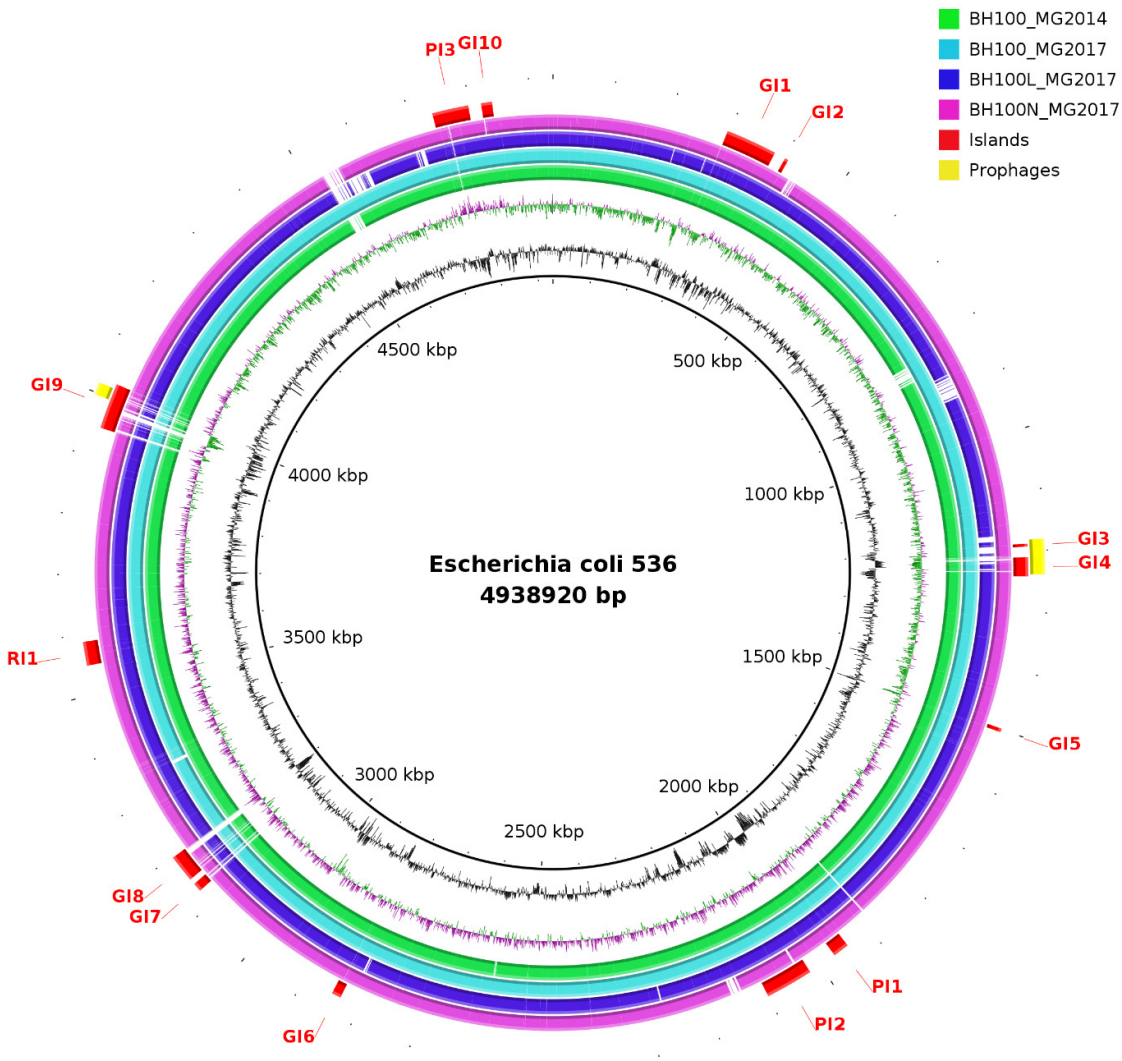


Figura 3.18 Ilhas de patogenicidade (PI), genômicas (GI) e de resistência (RI), identificadas pelo IslandPath-DIMOB e compilado no IslandViewer4, da *E. coli* 536, em comparação aos cromossomos da *E. coli* BH100 e suas variantes.

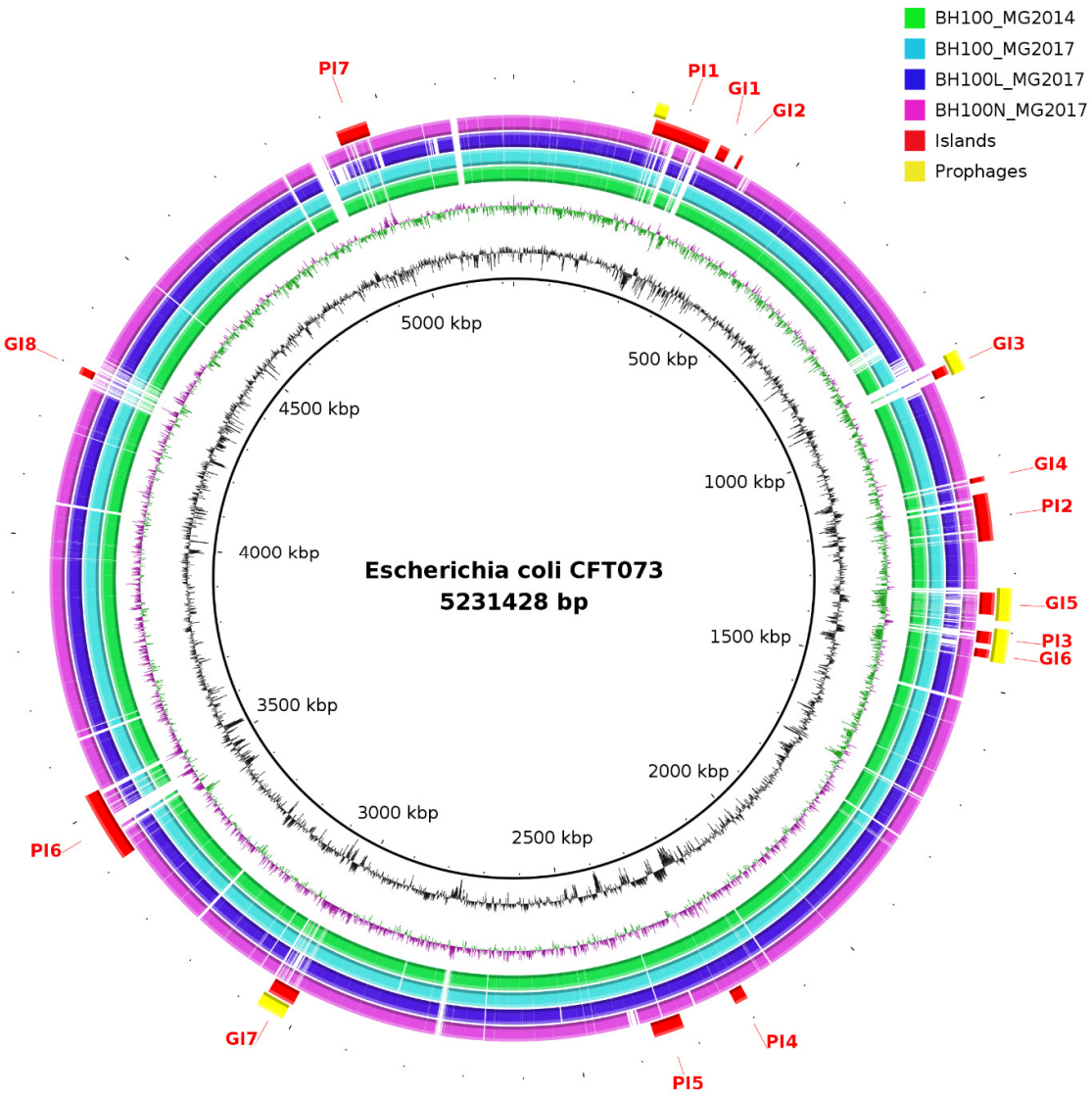


Figura 3.19 Ilhas de patogenicidade (PI) e genômicas (GI), identificadas pelo IslandPath-DIMOB e compilado no IslandViewer4, da *E. coli* CFT073, em comparação aos cromossomos da *E. coli* BH100 e suas variantes.

A comparação dos cromossomos das linhagens derivadas da *E. coli* BH100 com as ilhas identificadas na *E. coli* CFT073 (**Figura 3.19**) revela uma menor conservação, quando comparadas àquelas presentes na *E. coli* 536 (**Figura 3.18**). Esse fato, contudo, era esperado, por esta ser uma linhagem filogeneticamente mais próxima à BH100 (**Figura 3.16**). A associação das ilhas anotadas na *E. coli* BH100 subtr. MG2017 com as ilhas presentes nas *E. coli* 536 e CFT073 se deu via BLASTn, utilizando valores de cobertura acima de 80%, identidade maior que 80 % e *e-value* inferior à 10^{-15}, escolhidos empiricamente. Essa alta estringência foi adotada de modo a identificar as ilhas conservadas entre os genomas com maior acurácia. O **Quadro 3.3** revela a associação entre as diferentes ilhas anotadas em cada replicon.

Quadro 3.3 Associação das diferentes ilhas anotadas na *E. coli* 536, *E. coli* CFT073 e *E. coli* BH100 subtr. MG2017.

***E. coli* 536**	***E. coli* CFT073**	***E. coli* BH100 subtr. MG2017**
GI1	GI1	GI1
GI2	GI2	GI2
PI1	PI4	PI1
PI2	PI2	PI2
GI6	-	GI5
PI3	-	PI3
GI10	-	GI9

GI: Ilhas genômicas (*Genomic islands*); PI: Ilhas de patogenicidade (*Pathogenicity islands*)

A conservação das ilhas PI1 e PI2, e mesmo GI1 e/ou GI2, da *E. coli* BH100 subtr. MG2017 nas *E. coli* 536 e CFT073 podem sugerir o fenótipo uropatogênico à *E. coli* BH100. Já a ilha PI3, bem como GI5 e/ou GI9, podem estar associadas a mecanismos de patogênese, mais similar à *E. coli* 536.

As **Figura 3.18** e **Figura 3.19** revelam a presença de ilhas aparentemente fragmentadas nos replicons cromossomicos da BH100. Uma das possíveis razões para esse fenômeno está relacionado ao fato de que essas ilhas são muitas vezes compostas por elementos móveis (JUHAS et al., 2009), capazes de se dissociar da mesma e se inserir em outros sítios do cromossomo. Além disso, a presença dessas ilhas pode levar a algum ganho adaptativo por estar associada a um maior número de cópias de genes relevantes para um determinado fenótipo. Com isso em mente, extraiu-se as sequências protéicas codificadas por essas ilhas e, via tBLASTn, almejou-se identificar todas as regiões codificantes para esses genes, presentes nos replicons cromossômicos da BH100. Para tanto, o gene pertencente à ilha foi considerado presente no cromossomo da BH100, se o alinhamento satisfizesse os seguintes critérios: cobertura acima de 70%; identidade acima de 40% e *e-value* inferior a 10^{-15}. Esses parâmetros foram determinados de forma empírica, de modo a abrigar todos os possíveis parálogos presentes. Os resultados dessa análise revelam a diversidade no número de possíveis parálogos ao longo dos cromossomos das variantes da *E. coli* BH100 (Apêndice F). Vale destacar, no entanto, que apesar de satisfazerem os critérios de similaridade utilizados, não se pode atestar aumento no número de cópias, já que não é imediata a transferência de função. O indicativo de um maior número de cópias de alguns dos genes dessas ilhas, pode, entretanto, ser sugestivo de alguma adapatação fenotípica nessas linhagens e podem ser alvo de investigações futuras.

As ilhas de patogenicidade assinaladas pelo IslandViewer4 ocorrem mediante a identificação de genes associados a algum dos bancos de dados vinculados a genes envolvidos com patogênese (BERTELLI et al., 2017). Na *E. coli* BH100 subtr.MG2017, por exemplo, as ilhas de patogenicida-

de PI1 e PI2 foram assim denominadas por conter genes previamente identificados apenas em linhagens patogênicas, mas ausentes em linhagens não-patogênicas (HO SUI et al., 2009). Já a ilha PI3, foi assim identificada por conter fatores de virulência compilados na base Victors[25]. Esse fato demonstra que a classificação de ilhas de patogenicidade depende da base de dados utilizada. Com efeito, ilhas genômicas, carreando genes de virulência, podem não ter sido classificadas como de patogenicidade pelo fato desses genes não estarem indexados na base de dados investigada.

Outro aspecto relevante está associado à predição da ilha genômica em si. Atualmente, existem diferentes estratégias vinculadas à predição das ilhas, cada qual associada a taxas específicas de falsos-positivos e falsos-negativos. De fato, os diferentes softwares de predição de ilhas resumem uma série de vantagens e desvantagens, que ao fim, levam o usuário à tarefa de determinar o que for mais conveniente para os tipos de dados e condições de análise.

Os problemas vinculados à caracterização da ilha genômica e de imprecisões, na predição das mesmas, pode ser de certo modo circundado pela análise sistemática dos genes associados a um determinado efeito fenotípico. Nesse caso, por exemplo, a identificação de genes associados à patogenicidade em ilhas genômicas não anotadas como de patogenicidade, pode levar ao pesquisador à devida caracterização manual da ilha. Já em relação às imprecisões nas predições das ilhas, pode-se observar o consenso na predição de outras ferramentas. Em ambos os casos, o pesquisador seria capaz de julgar o quão factível seria (des)considerar determinada ilha genômica na análise.

3.2.2.3 GENES DE VIRULÊNCIA

Até aqui, todos os indícios têm apontado a *E. coli* BH100 como uma linhagem pertencente ao patótipo UPEC. As análises do programa VirulenceFinder revelaram a presença de sete genes de virulência na *E. coli* BH100 subtr. MG2017 (**Quadro 3.4**). Esses achados corroboram a discussão do tópico 3.2.2.2, referente à não assinatura de uma ilha genômica, como sendo de patogenicidade, por conta de vieses nas bases de dados. A presença de fatores de virulência em GI1 e GI3 poderiam, portanto, levar à classificação dessas ilhas como de patogenicidade[26].

[25] He, Yonggun (2011). Victors: Virulence Factors. Disponível em: www.phidias.us/victors/index.php
[26] Por conveniência de notação, continuarei utilizando a nomenclatura de GIs para essas ilhas.

Quadro 3.4 Relação de genes preditos pelo VirulenceFinder no cromossomo da *E. coli* BH100 subtr. MG2017, suas eventuais presenças em ilhas genômicas e o número de parálogos.

Gene	GenBank	Ilha*	Número de parálogos**
iss	WP_001298464.1	GI3	1[a]
gad	WP_000358849.1/ WP_000372166.1	-	2
mcmA	WP_024186735.1	GI1	1
sfaS	WP_000767892.1	GI1	2
iroN	WP_001222188.1	GI1	2[b]
vat	WP_001034003.1	GI1	1
cnf1	WP_000528123.1	PI3	1

* GI: Ilhas genômicas (*Genomic islands*); PI: Ilhas de patogenicidade (*Pathogenicity islands*)
**Cópias no cromossomo. Valores validos para todos os demais replicons cromossomais (BH100 subtr. MG2014, BH100L subtr. MG2017 e BH100N subtr. MG2017), salvo quando especificado.
[a]Gene ausente na *E. coli* BH100L subtr. MG2017
[b]Na *E. coli* BH100L subtr. MG2017 existe apenas uma cópia do gene

O gene *iss* é responsável por codificar uma proteína de membrana capaz de conferir resistência à célula contra o sistema do complemento presente no soro (LYNNE et al., 2007). A ilha genômica intitulada GI3 e presente no fago Entero_lambda, não está presente na *E. coli* BH100L subtr. MG2017, o que explica a ausência do gene *iss* nessa variante. Esse gene está presente próximo à GI4 da *E. coli* 536 e na porção terminal de PI3 da *E. coli* CFT073, ambos associados ao prófago Entero_Lambda, o que sugere uma relativa similaridade entre essas ilhas com GI3_EcolBH100[27]. De fato, as ilhas GI4_Ecol536, PI3_EcolCFT073 e GI3_EcolBH100 compartilham regiões com alto valor de identidade (> 90%), mas baixa cobertura (30 - 40 %). De todo modo, o gene está presente em todas as linhagens aqui analisadas, apesar de ser menos frequente em UPEC (JOHNSON; WANNEMUEHLER; NOLAN, 2008).

O gene *gad* codifica a enzima glutamato descarboxilase, responsável pela sobrevivência de *E. coli* em meios de baixo pH (LIN et al., 1995). A duplicidade desse gene já havia sido relatada em outras linhagens de *E. coli* (SMITH et al., 1992). No entanto, vale destacar que apesar de importante para sobrevivência celular nesses meios, esse gene, não necessariamente, se correlaciona bem com outros fatores de virulência (GRANT; WEAGANT; FENG, 2001).

Na ilha genômica GI1, comum à *E. coli* 536 e CFT073, observa-se a presença de diversos marcadores de virulência. O gene *mcmA* está associado à síntese da microcina H37, capaz de eliminar bactérias susceptíveis (LAVIÑA; GAGGERO; MORENO, 1990). Esse gene está altamente

[27] Essa notação será utilizada para me referir a cada uma das ilhas de cada replicon cromossomal. Nesse caso, por exemplo GI3_EcolBH100 refere-se à ilha genômica 3 presente na *E. coli* BH100.

correlacionado à presença de outros fatores de virulência presentes em UPEC e parece estar associada a casos de bacteremia (BUDIČ et al., 2011). O gene *sfaS* codifica para adesina S capaz de permitir a associação da bactéria a receptores contendo ácido siálico, presentes na membrana de células eucarióticas (SCHMOLL et al., 1989). O gene *iroN*, associado à captação de catecolatos (SORSA et al., 2003), contribui para geração do *fitness* necessário para geração da patogenicidade (SUBASHCHANDRABOSE et al., 2013). Esse gene possui um parálogo em uma região, não situada em uma ilha genômica, em todas as linhagens avaliadas, salvo na *E. coli* BH100L subtr. MG2017. O gene *vat*, codificante de uma exotoxina contendo um domínio serina protease (WELCH, 2016), colabora para o *fitness* necessário para geração de patogenicidade (NICHOLS et al., 2016) e está fortemente associado à UPECs (PARHAM et al., 2005).

Por fim, o gene *cnf1* está associado ao escape do sistema imune, por interferir nos processos de fagocitose e estimular estresse oxidativo em leucócitos (DAVIS; RASMUSSEN; O'BRIEN, 2005). Esse gene está presente na ilha de patogenicidade PI3, ausente na *E. coli* CFT073. Apesar de PI3_EcolBH100 e PI3_Ecol536 serem equivalentes, o gene *cnf1* não está presente na última, conforme já previamente observado (BINGEN-BIDOIS et al., 2002).

A análise mais abrangente dos fatores de virulência presentes na *E. coli* BH100 depende da observação de outros genes intrínsecos do cromossomo, *i.e*, não presentes em ilhas genômicas, ou ainda de genes não vinculados à base de dados do VirulenceFinder e que podem estar associados a essas ilhas. Para tanto, utilizando-se como base o trabalho de SUBASHCHANDRABOSE e MOBLEY (2015), onde propuseram a revisão dos fatores de virulência e de *fitness* associados à UPEC, objetivou-se identificar esses genes no cromossomo da *E. coli* BH100. Para isso, as sequências proteicas destacadas como relevantes, pelos autores, para uropatogênese foram obtidas do Uniprot. A identificação das potenciais CDSs associadas a esses genes e seus possíveis parálogos foi possível via tBLASTn. As CDSs e seus parálogos foram considerados presentes no genoma quando o alinhamento resultasse em identidade > 40%, cobertura > 70% e *e-value* < 10^{-15}. O **Quadro 3.5** compila todos os genes encontrados nas *E. coli* desse trabalho, nas uropatogênicas CFT073 e 536, e na não-patogênica K12 subtr. MG1655.

Quadro 3.5 Genes associados à virulência e *fitness* de UPECs revisados em Subashchandrabose e Mobley (2015) e identificados em diferentes linhagens de *E. coli* (Continua).

Locus_tag	Gene	Mecanismo	BH100 subtr. MG2017	BH100 subtr. MG2014	BH100L subtr. MG2017	BH100N subtr. MG2017	CFT073	536	K-12 subtr. MG1655
BH100B_RS01690	*iroN*[a]	Aquisição de ferro	xx	xx	x	xx	xx	xx	x

BH100B_RS01925	*vat*[a]	Toxina	x	x	x	x	xx	x	A
BH100B_RS03895	*fur*	Aquisição de ferro	x	x	x	x	x	x	x
BH100B_RS04565	*mntR*	Aquisição de manganês	x	x	A	x	x	x	x
BH100B_RS05690	*csgG*	Biofilme	x	x	x	x	x	x	x
BH100B_RS05695	*csgF*	Biofilme	x	x	x	x	x	x	x
BH100B_RS05700	*csgD*	Biofilme	x	x	x	x	x	x	x
BH100B_RS05705	*csgD*	Biofilme	x	x	x	x	x	x	x
BH100B_RS05720	*csgB*	Biofilme	x	x	x	x	x	x	x
BH100B_RS05725	*csgA*	Biofilme	x	x	x	x	x	x	x
BH100B_RS05730	*csgC*	Biofilme	x	x	x	x	x	x	x
BH100B_RS06190	*phoQ*	Regulador transcricional global	x	x	A	x	x	x	x
BH100B_RS06195	*phoP*	Regulador transcricional global	x	x	A	x	x	x	x
BH100B_RS06525	*sitD*[b]	Aquisição de ferro	x	x	A	x	x	x	A
BH100B_RS06530	*sitC*[b]	Aquisição de ferro	x	x	A	x	x	x	A
BH100B_RS06535	*sitB*[b]	Aquisição de ferro	x	x	A	x	x	x	A
BH100B_RS06540	*sitA*[b]	Aquisição de ferro	x	x	A	x	x	x	A
BH100B_RS07040	*hns*	Regulador transcricional global	xxxx	xxxx	xxx	xxxx	xxx	xxx	xx
BH100B_RS07120	*tonB*	Aquisição de ferro	x	x	x	x	x	x	A
BH100B_RS09960	*znuA*	Aquisição de zinco	x	x	x	x	x	x	x

BH100B_RS09965	*znuB*	Aquisição de zinco	x	x	x	x	x	x	x
BH100B_RS09970	*znuC*	Aquisição de zinco	x	x	x	x	x	x	x
BH100B_RS10270	*fliY*	Flagelo	x	x	x	x	x		x
BH100B_RS10275	*fliZ*	Flagelo	x	x	x	x	x	x	x
BH100B_RS10280	*fliA*	Flagelo	x	x	x	x	x	x	x
BH100B_RS10285	*fliC*	Flagelo	x	x	x	x	x	x	A
BH100B_RS10295	*fliD*	Flagelo	x	x	x	x	x	x	x
BH100B_RS10300	*fliS*	Flagelo	x	x	x	x	x	x	x
BH100B_RS10305	*fliT*	Flagelo	x	x	x	x	x	x	x
BH100B_RS14355	*rpoE*	Regulador transcricional global	x	x	x	x	x	x	x
BH100B_RS16470	*papX*^{c}	Fimbria P	xxxx	xxxx	xxxx	xxxx	xxx	xxxx	A
BH100B_RS16480	*papG*^{c}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS16485	*papF*^{c}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS16490	*papE*^{c}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS16495	*papK*^{c}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS16500	*papJ*^{c}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS16505	*papD*^{c}	Fimbria P	xxx	xxx	xxx	xxx	xxxx	xxx	xx
BH100B_RS16510	*papC*^{c}	Fimbria P	xxx	xxx	xxx	xxx	xxx	xxx	x
BH100B_RS16515	*papH*^{c}	Fimbria P	xx	xx	xx	xx	xx	xx	A
BH100B_RS16520	*papA*^{c}	Fimbria P	xx	xx	xx	xx	xx	xx	A
BH100B_RS16525	*papB*^{c}	Fimbria P	xxx	xxx	xxx	xxx	xxx	xx	A
BH100B_RS16530	*papI*^{c}	Fimbria P	xxx	xxx	xxx	xxx	xxx	xx	A
BH100B_RS16965	*kpsF*	Capsula	xxx	xxx	xxx	xxx	xxx	xxx	xx
BH100B_RS16970	*kpsE*	Capsula	x	x	x	x	x	x	A
BH100B_RS16975	*kpsD*	Capsula	x	x	x	x	x	x	A
BH100B_RS16980	*kpsB*	Capsula	xx	xx	xx	xx	xx	xx	x
BH100B_RS16985	*kpsC*	Capsula	x	x	x	x	x	A	A
BH100B_RS16990	*kpsS*	Capsula	x	x	x	x	x	A	A
BH100B_RS17310	*exbD*	Aquisição de	x	x	x	x	x	x	x

		ferro							
BH100B_RS17315	*exbB*	Aquisição de ferro	x	x	x	x	x	x	x
BH100B_RS17430	*qseB*	Quorum sensing	xxx	xxx	xx	xxx	xxx	xxx	xxx
BH100B_RS17435	*qseC*	Quorum sensing	x	x	x	x	x	x	x
BH100B_RS17600	*zupT*	Aquisição de ferro	x	x	x	x	x	x	x
BH100B_RS20040	*chuA*	Aquisição de ferro	xx	xx	xx	xx	xx	xx	A
BH100B_RS24140	*hfq*	Regulador transcricional global	x	x	x	x	x	x	X
BH100B_RS24720	*papX*^{d}	Fimbria P	xxxx	xxxx	xxxx	xxxx	xxx	xxxx	A
BH100B_RS24730	*papG*^{d}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS24735	*papF*^{d}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS24740	*papE*^{d}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS24745	*papK*^{d}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS24750	*papJ*^{d}	Fimbria P	xx	xx	xx	xx	xx	x	A
BH100B_RS24755	*papD*^{d}	Fimbria P	xxx	xxx	xxx	xxx	xxxx	xxx	xx
BH100B_RS24760	*papC*^{d}	Fimbria P	xxx	xxx	xxx	xxx	xxx	xxx	X
BH100B_RS24765	*papH*^{d}	Fimbria P	xx	xx	xx	xx	xx	xx	A
BH100B_RS24770	*papA*^{d}	Fimbria P	xx	xx	xx	xx	xx	xx	A
BH100B_RS24775	*papB*^{d}	Fimbria P	xxx	xxx	xxx	xxx	xxx	xx	A
BH100B_RS24785	*papI*^{d}	Fimbria P	xxx	xxx	xxx	xxx	xxx	xx	A
BH100B_RS24905	*cnf1*^{d}	Toxina	x	x	x	x	A	A	A
BH100B_RS24920	*hlyD*^{d}	Toxina	x	x	x	x	x	xx	A
BH100B_RS24925	*hlyB*^{d}	Toxina	x	x	x	x	x	xx	A
BH100B_RS24930	*hlyA*^{d}	Toxina	x	x	x	x	x	xx	A
BH100B_RS24935	*hlyC*^{d}	Toxina	x	x	x	x	x	xx	A
BH100B_RS25370	*fimB*	Fimbria F	xxx	xxx	xxx	xxx	xxxxx	xxx	xx
BH100B_RS25375	*fimE*	Fimbria F	xxx	xxx	xxx	xxx	xxxxx	xxx	xx

BH100B_RS25385	*fimA*	Fimbria F	xxx	xxx	xxx	xxx	xxx	xx	xx
BH100B_RS25390	*fimI*	Fimbria F	xx	xx	xx	xx	xx	xx	x
BH100B_RS25395	*fimC*	Fimbria F	xxxx	xxxx	xxxx	xxxx	xxxx	xxx	xxx
BH100B_RS25400	*fimD*	Fimbria F	xxxx	xxxx	xxxx	xxxx	xxxx	xxx	xxx
BH100B_RS25405	*fimF*	Fimbria F	xx	xx	xx	xx	xx	x	xx
BH100B_RS25410	*fimG*	Fimbria F	xxx	xxx	xxx	xxx	xxx	xx	xx
BH100B_RS25415	*fimH*	Fimbria F	xx	xx	xx	xx	xx	x	xx

x: Quantidade de parálogos presentes; A: Ausência; a) ilha GI1_BH100MG2017; b) fago Entero_lambda; c) ilha GI6_BH100MG2017; d) ilha PI3_BH100MG2017.

O **Quadro 3.5** revela a presença de alguns genes já identificados pelo programa Virulence-Finder (*iroN*, *vat* e *cnf1*) e que foram discutidos no trabalho de Subashchandrabose e Mobley (2015). Esse quadro também revela a presença de genes exclusivos de UPECs e ausentes na comensal *E.* coli K-12 subtr. MG1655. Esses genes referem-se basicamente à sintese da fimbria P, capsula e toxinas relevantes no processo de infecção do trato urinário. Interessante, também, foi observar a presença do operon *pap* na ilha GI6_BH100MG2017, caracterizando-a portanto como uma ilha de patogenicidade, e do operon *sit* no prófago Entero_lambda.

A ausência do operon *sit* e principalmente dos genes *phoP* e *phoQ* devem reduzir o *fitness* da *E. coli* BH100L subtr. MG2017 frente uma potencial contaminação do trato urinário. Isso pois, o operon *sit* é contribui para aquisição de íons Fe^{2+} do meio. Já os genes *phoP* e *phoQ* atuam como reguladores da expressão de diversos fatores de virulência. A ausência desses genes, portanto, devem afetar, substancialmente, o desenvolvimento de uropatogenicidade dessa linhagem. De toda forma, ensaios laboratorias adicionais seriam interessantes, para melhor compreensão da importância desses genes na patogenicidade da *E. coli* BH100.

A presença de genes parálogos em cada uma das linhagens observadas podem sugerir uma maior contribuição dos efeitos de cada gene na patogenicidade, caso ainda se verifique a conservação funcional desses elementos. Sabe-se que os mecanismos associados à uropatogênese estão intimamente relacionados às alterações nos níveis de transcrito dos genes associados à virulência e *fitness* (SUBASHCHANDRABOSE; MOBLEY, 2015). Disso posto, a análise da duplicidade de genes pode fornecer pistas relevantes, a uma resolução genômica, acerca da importância de determinados genes para patogenicidade. De forma geral, o perfil de parálogos de cada gene parece ser semelhante em todas as linhagens uropatogênicas observadas. As diferenças encontradas podem estar associadas aos mecanismos de patogenicidade, particulares de cada linhagem. De toda forma, ensaios adicionais que permitam classificar esses parálogos como verdadeiras isoformas do gene em questão são necessárias, já que a mera similaridade proteica não implica em funções análogas.

3.2.3 O contexto filogenético dos plasmídeos da *E. coli* BH100

De modo a simplificar a análise e discussão dos plasmídeos aqui identificados, optou-se pela sua classificação, a partir das métricas de montagem e anotação, nos grupos pAp e pBH100. No primeiro foram alocados os plasmídeos pAp da *E. coli* BH100 subtr. MG2014, pAp da *E. coli* K12T subtr. MG2017 e pApR da *E. coli* BH100 subtr. MG2017. No segundo grupo foram inseridos todos os plasmídeos pBH100-1 e pBH100alpha (**Figura 3.20**).



Figura 3.20 Diagrama evidenciando os diferentes grupos de plasmídeos identificados nesse trabalho. A intersecção entre os grupos representa o compartilhamento de sequencias de cada um desses grupos.

Objetivando-se observar as relações filogenéticas dos plasmídeos originalmente presentes na *E. coli* BH100, foram selecionados os plasmídeos pAp da *E. coli* BH100 subtr. MG2014 e o plasmídeo pBH100alpha da *E. coli* BH100 subtr. MG2014 como representantes de cada grupo. Essa escolha se deve ao fato de ambos replicons terem sido sequenciados de uma subcepa, provavelmente, mais antiga e por não apresentarem fragmentos de outro grupo, *i.e* o plasmídeo pAp da *E. coli* BH100 subtr. MG2014 não apresenta fragmentos observados em comum ao grupo pBH100 e o plasmídeo pBH100alpha da *E. coli* BH100L subtr. MG2014 não apresenta fragmentos em comum ao grupo pAp, salvo pApR. Dessa forma pressupõe-se que ambos replicons seriam representativos dos plasmídeos originalmente presentes do isolamento da *E. coli* BH100.

3.2.3.1 Árvore do plasmídeo pAp

A **Figura 3.21** releva o contexto filogenético em que se inserem os plasmídeos do grupo pAp.

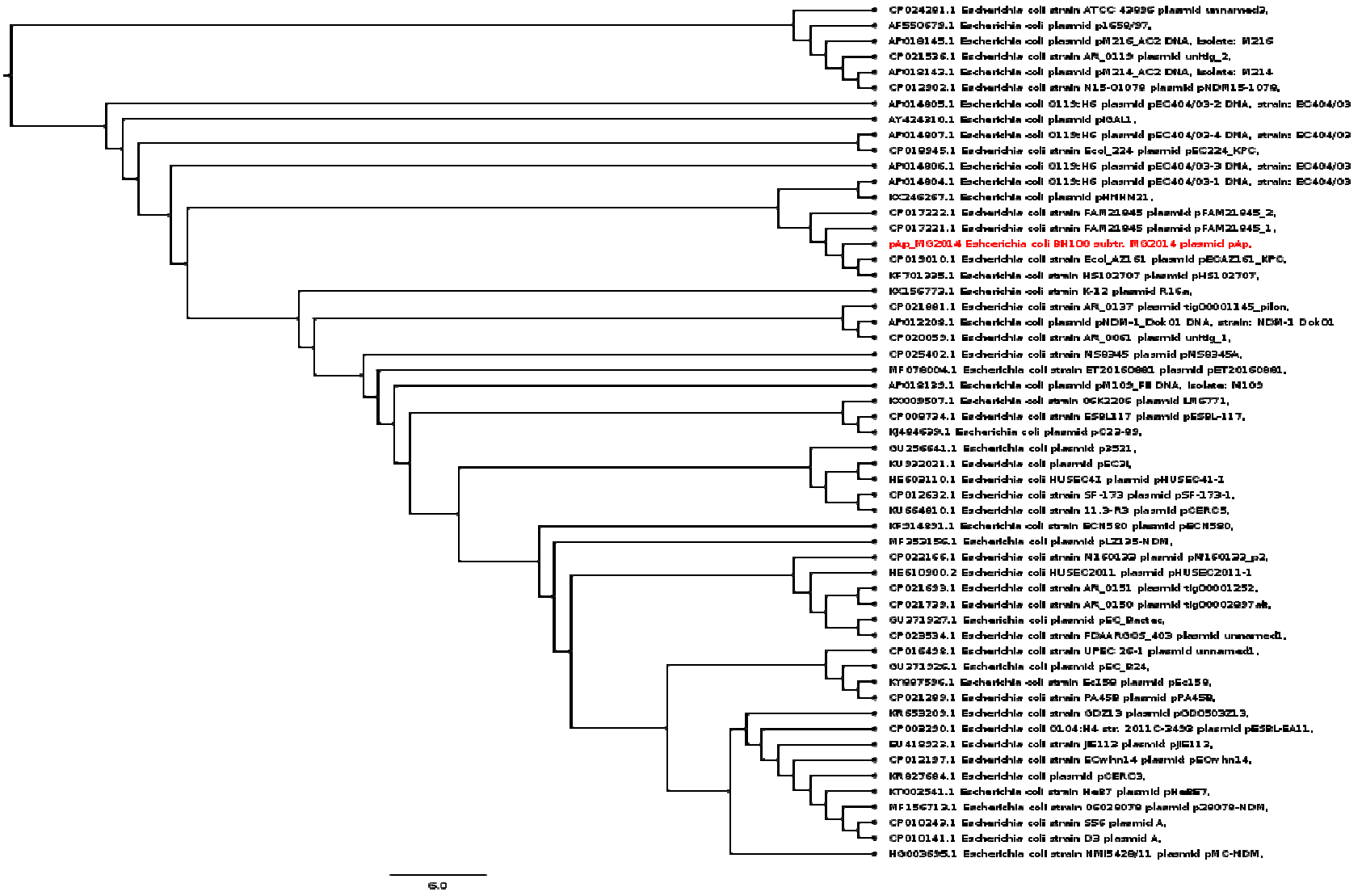


Figura 3.21 Árvore filogenética do plasmídeo pAp

A partir da procura na base de dados do Genbank via BLASTn foi possível identificar a sequência do plasmídeo pECAZ161_KPC (GenBank: CP019010.1) como o mais próximo representante do plasmídeo pAp no gênero *Escherichia*. Outras sequências plasmidias pertencentes predominantemente à *Kleibsiella* também foram observadas. O plasmídeo pECAZ161_KPC encontra-se associado a uma linhagem e *E. coli* resistente a carbapenos. Embora não tenham sido encontrados trabalhos publicados com essa linhagem a análise da anotação funcional desse plasmídeo revela a presença de uma beta-lactamase associada à hidrólise de antibióticos dessa classe.

A similaridade dos plasmídeos pAp com sequências advindas de *Kleibsiella* podem indicar que esse plasmídeo tenha sido adquirido de uma linhagem desse gênero. De fato, a transferência horizontal interespecífica entre representantes do gênero *Kleibsiella* e *Escherichia coli* já foram reportadas (GOREN et al., 2010; WACHINO et al., 2006).

3.2.3.2 ÁRVORE DO PLASMÍDEO PBH100

Novamente através da procura na base de dados do GenBank via BLASTn, observou-se que o plasmídeo pBH100alpha é homólogo ao plasmídeo NR1 (GenBank: DQ364638.1). A observação desse plasmídeo num contexto de outros plasmídeos de *E. coli* pode ser visto na **Figura 3.22**.

O plasmídeo NR1, também conhecido como R100, é um plasmídeo que contém marcadores de resistência a diversos antimicrobianos e naturalmente presente em amostras de *Shigellla flexneri* 2b isoladas de fezes humanas (NAKAYA; NAKAMURA; MURATA, 1960; ROWND; NAKAYA; NAKAMURA, 1966). Esse plasmídeo confere resistência à 20,0 $\mu g.mL^{-1}$ de estreptomicina e tetraciclina, e a concentrações superiores a 100,0 $\mu g.mL^{-1}$ de cloranfenicol e espectinomicina (WILLETTS; FINNEGAN, 1970).

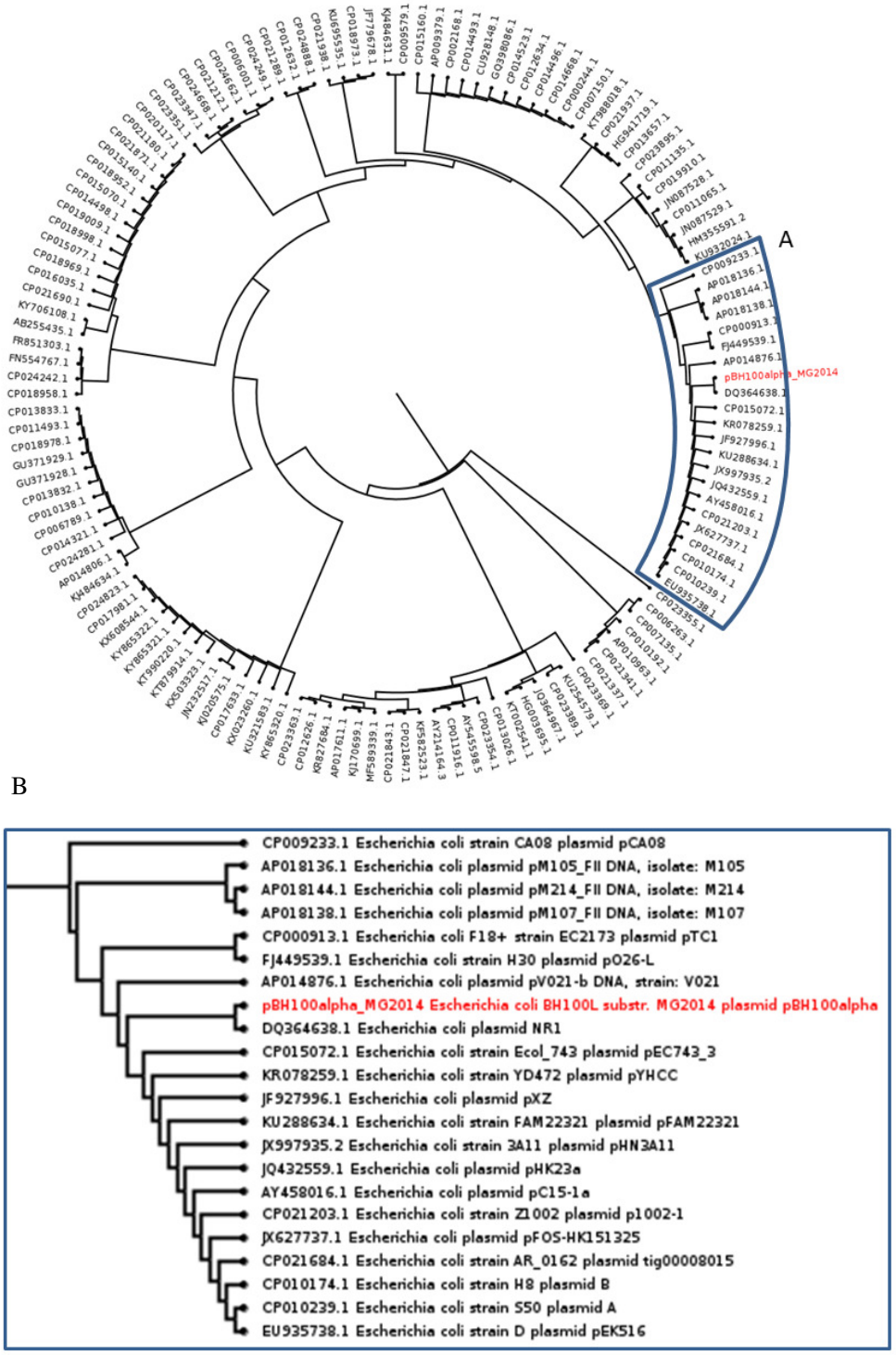


Figura 3.22 Contexto filogenético dos plasmídeos do grupo pBH100 (A). Em (B) destaca-se o clado, onde os plasmídeos pBH100 estão localizados.

3.2.4 Os plasmídeos do grupo pBH100

A análise do PlasmidFinder identificou os plasmídeos desse grupo como pertencentes ao grupo de incompatibilidade IncFII. As principais características anotadas nos plasmídeos do grupo pBH100 podem ser visualizadas na **Figura 3.23**. Nessa figura destacam-se os principais genes presentes nos replicons desse grupo. Outros genes não estão representados, mas estão associados de forma geral a processos metabólicos, conforme descritos abaixo. Nessa figura, justifica-se a distinção dos subgrupos pBH100alpha e pBH100-1, que diferem entre si, basicamente, pela presença do transposon Tn3 em pBH100-1. Outras menores diferenças serão discutidas mais adiante.

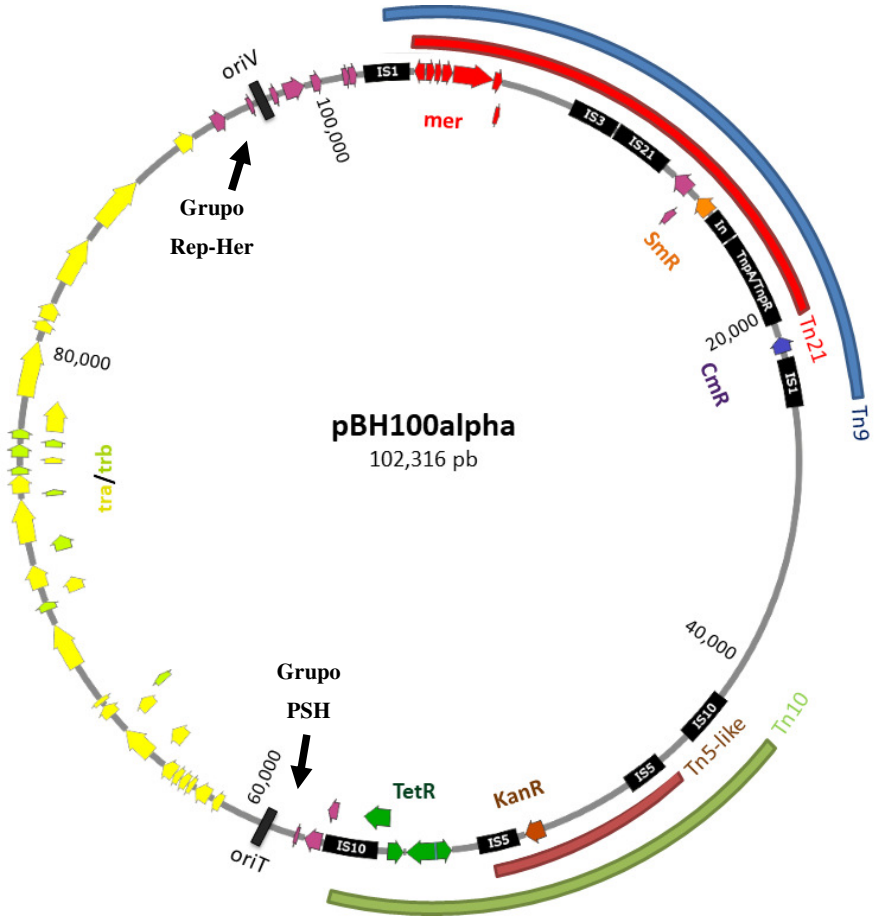


Figura 3.23 Representação esquemática do plasmídeo pBH100alpha.

3.2.4.1 GENES DE RESISTÊNCIA

Os plasmídeos do grupo pBH100 apresentam diversos marcadores de resistência que corroboram as observações fenotípicas constatadas em populações dessa linhagem, bem como outros marcadores ainda não verificados.

3.2.4.1.1 Resistência ao mercúrio

A resistência ao íon mercúrio é conferida pelo operon *mer*. A estrutura desse operon apresenta a seguinte sintenia gênica: *merR*, *merT*, *merP*, *merC*, *merA*, *merD* e *merE* e está de acordo com o verificado no Tn21 (NASCIMENTO; CHARTONE-SOUZA, 2003).

Esse operon mantém a mesma estrutura em todos os plasmídeos identificados do grupo pBH100, salvo no plasmídeo pBH100-1 da *E. coli* BH100 subtr. MG2014, onde o gene codificador da mercúrio redutase (*merA*) possui deleção de uma guanina na posição 100.315 que leva a separação da região codificante em dois quadros de leitura. Essa deleção é justificada pelo mapeamento com as *reads* provenientes da plataforma IonTorrent e que não corroboram a presença do nucleotídeo na região (**Figura 3.24**). Esse fato pode indicar algum viés nas *reads* obtidas na biblioteca *fragment*, capaz de induzir indels ou a inativação real do gene na *E. coli* BH1000 subtr. MG2014. No caso da não funcionalidade do gene, vale salientar que a resistência ao mercúrio ainda pode ser passível de verificação, mas a menores concentrações, uma vez que essa bactéria é capaz de secretar compostos no meio de cultivo, capazes de neutralizar o efeito do íon.

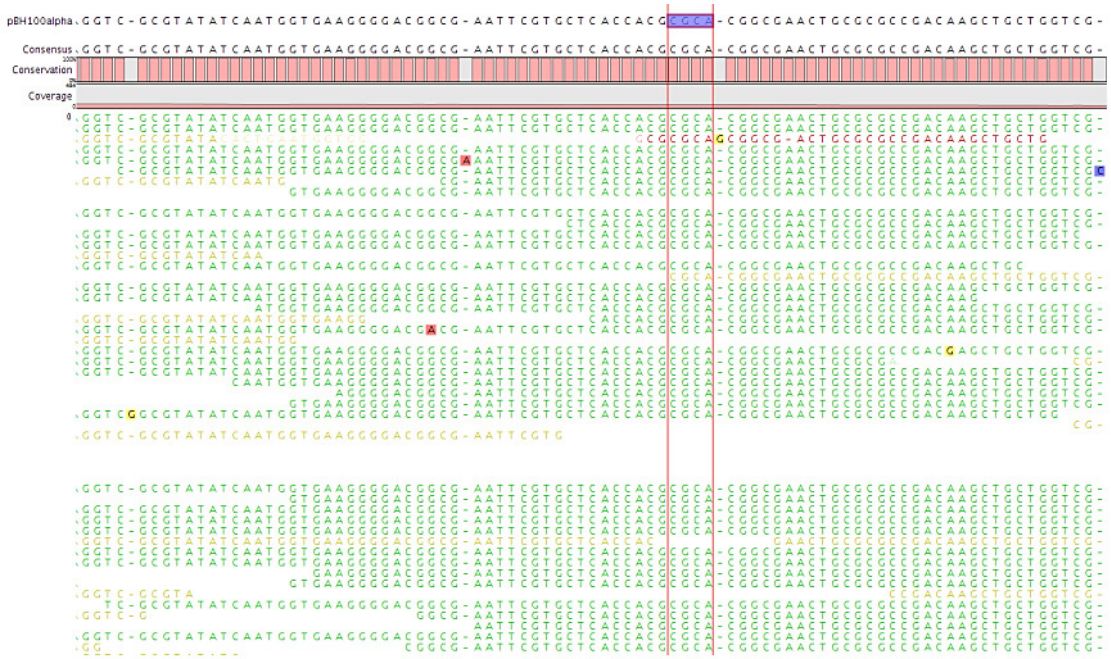


Figura 3.24 Mapeamento de *reads fragment* no plasmídeo pBH100alpha_MG2014. Em destaque, região do gene *merA*, indicando a não sustentação para adição de uma guanina na posição 100.315 (imediatamente a montante da região em destaque), presente nos demais replicons do grupo pBH100.

3.2.4.1.2 Resistência à estreptomicina

O gene *aadA1*, codificante da aminoglicosídeo nucleotidiltransferase (3")-Ia, está localizado a jusante do operon *mer* e é responsável por conferir resistência à estreptomicina e espectinomicina, via transferência de adenosil monofosfato (AMP) para um grupo hidroxila, presente no aminoglicosídeo (RAMIREZ; TOLMASKY, 2010). Esse gene é comum em diversos cassetes gênicos, presentes em transposons, integrons e plasmídeos.

3.2.4.1.3 Resistência ao cloranfenicol

A resistência ao cloranfenicol é devida à presença do gene *cat*, codificante da cloranfenicol acetiltransferase, situado logo após ao gene *aadA1*. Essa enzima é capaz de inativar a atividade do antibiótico a partir da transferência de grupos acetil de acetil-CoA, levando à formação de um composto bi-acetilado, não funcional (SHAW, 1967).

3.2.4.1.4 Resistência à canamicina

Em seguida, responsável por conferir resistência à canamicina, está o gene *aphA1*, codificante da aminoglicosídeo O-fosfotransferase APH(3')-Ia. Essa enzima age através da adição de um grupo fosfato na molécula do aminoglicosídeo e também é capaz de conferir resistência a outros aminoglicosídeos da classe da canamicina, tais como neomicina, lividomicina e ribostamicina (RAMIREZ; TOLMASKY, 2010).

3.2.4.1.5 Resistência à tetraciclina

A jusante do gene *cat* observa-se a presença do operon *tet,* codificante dos genes *tetR*, *tetB*, *tetC e tetD,* e responsável por conferir resistência à tetraciclina. O primeiro gene é responsável pela síntese de um regulador transcricional, capaz de controlar a expressão de *tetB*, que codifica um transportador de membrana capaz de remover o antibiótico do meio intracelular (**Figura 3.25**). Nesse caso, a tetraciclina, quando no citosol, interage com íons bivalentes, possibilitando sua interação com TetR. Essa interação reduz a afinidade dessa proteína com as regiões operadoras de *tetB* e *tetR*, garantindo a ativação desses genes.

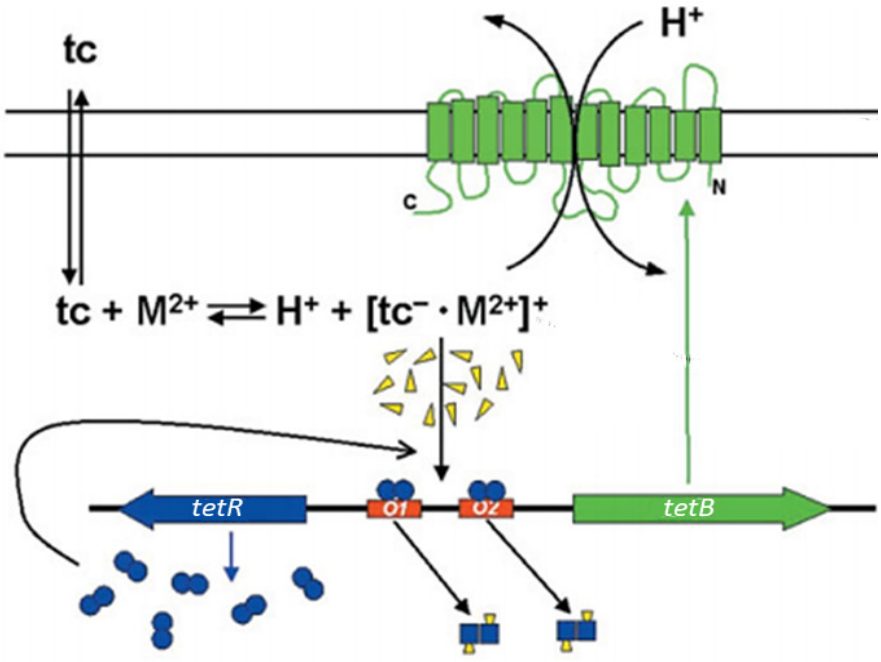


Figura 3.25 Atividade dos genes do operon *tet* responsáveis pela resistência à tetraciclina. Esta, ao entrar no meio intracelular, é conjugada com íons bivalentes, o que a torna capaz de se associar com a proteína repressora TetR. Essa interação induz mudança conformacional em TetR, estimulando sua dissociação das regiões operadoras do gene *tetB* e do próprio *tetR*. Dessa forma, induz-se à expressão de TetB, que age como uma bomba de efluxo da tetraciclina. Adaptado de (HILLEN; BERENS, 2002).

Já os genes *tetC* e *tetD* não parecem essenciais à resistência (BRAUS; ARGAST; BECK, 1984). É sabido que o gene *tetC* age como regulador transcricional negativo de *tetD* (PEPE et al., 1997) e este seria um regulador transcricional genérico, pertencente à família de reguladores positivos AraC/XylS (GRIFFITH; BECKER; E. WOLF, 2005).

3.2.4.1.6 Outros marcadores de resistência

Outros dois marcadores de resistência foram observados entre o operon *mer* e o gene *aadA1*. A potencial resistência a sulfonamidas ocorre graças ao gene *sul1*, que sintetiza a enzima diidropteroato sintase, não susceptível a antibióticos dessa classe. Essa enzima é essencial para biossíntese de folato em bactérias e tem sido associada a isolados clínicos de bactérias gram-negativas, cujo gene, geralmente presente em plasmídeos, é adquirido via transferência horizontal (SKÖLD, 2000). O outro gene, disposto imediatamente a montante de *sul1*, é o gene *qacEΔ1*, responsável por codificar a uma proteína resistente a compostos quartenários de amônio. É sabido que compostos contendo quatro grupos amônio em sua composição estão associados a diversos bactericidas e desinfetantes encontrados no mercado (MCDONNELL, 2009). Os plasmídeos do grupo pBH100, portanto, parecem conferir resistência, por meio desse gene, a antissépticos e corantes intercalantes (KAZAMA et al., 1999). O gene *qacEΔ1* é uma variante semi-funcional do gene *qacE* e capaz de sintetizar um transportador de membrana do tipo bomba de efluxo capaz de ejetar esses compostos para o meio extracelular (PAULSEN et al., 1993).

3.2.4.2 OUTROS ELEMENTOS DE DESTAQUE

3.2.4.2.1 Operon tra/trb

O operon *tra* é responsável por conferir a capacidade auto-conjugativa dos plasmídeos do grupo pBH100. Esse operon é composto por diversos genes associados à síntese de proteínas de membrana e acessórias, relacionadas ao sistema conjugativo. Em pBH100, esses genes apresentam a seguinte disposição: *traM*, *traJ*, *traY*, *traA*, *traL*, *traE*, *traK*, *traB*, *traP*, *traV*, *traR*, *traC*, *traW*, *traU*, *traN*, *traF*, *traQ*, *traH*, *traG*, *traS*, *traT*, *traD*, *traI*, *traX*.

Também, é identificado nesse grupo o operon *trb* intercalado aos genes do operon *tra*, presentes entre *traP* e *traH* e disposto na seguinte ordem: *trbD*, *trbI*, *trbC*, *trbE*, *trbA*, *trbB*, *trbJ*, *trbF*.

3.2.4.2.2 Grupo PSH

O grupo aqui denominado PSH, consiste nos genes *psiB*, *psiA* e *hok* dispostos a jusante do Tn10. Os genes *psiA* e *psiB* estão envolvidos na inibição do sistema SOS (LENGELER; DREWS; SCHLEGEL, 1999). Esse é um mecanismo essencial durante a conjugação, uma vez que a presença de DNA fita-simples na célula receptora poderia induzir esse sistema, levando à degradação do DNA transferido. Aparentemente, *psiA* parece não ser essencial à inibição desse sistema, mas *psiB* é relatado como um inibidor da proteína RecA, principal componente do sistema SOS (ALTHORPE et al., 1999; BAILONE et al., 1988).

Já o gene *hok* codifica um peptídeo de 52 aminoácidos, associado à toxicidade celular em bactérias gram-negativas. Esse peptídeo provoca morte celular, por ser capaz de interferir com a membrana lipídica, levando à sua despolarização (PECOTA et al., 2003). Esse gene está associado ao sistema hok/sok descrito no plasmídeo NR1 e é responsável por assegurar a presença do plasmídeo após evento de segregação em cada célula filha. Nesse sistema, descrito por THISTED e GERDES (1992), ambas as sequências desses genes são complementares entre si, mas o mRNA do gene *hok* apresenta maior estabilidade, quando comparado ao mRNA de sok (GERDES; THISTED; MARTINUSSEN, 1990). Nesse caso, após divisão celular, as células filhas herdam, primariamente, mRNA do gene *hok,* e suas sobrevidas passam a depender da herança do gene *sok*, comumente associado a um promotor forte e presente no plasmídeo. Com isso, essas células são capazes de produzir quantidades suficientes de RNA, que agem inibindo a tradução do mRNA do gene *hok.* Esse sistema, portanto, assegura que células filhas que não herdem o plasmídeo conjugativo sejam mortas pelo peptídeo hok.

3.2.4.2.3 Grupo Rep-Her

No grupo, aqui denominado Rep-Her, verifica-se a presença de genes associados à herdabilidade e replicação do plasmídeo, localizados entre o operon *tra* e o operon *mer*.

O primeiro gene em destaque desse grupo na **Figura 3.23** é referente ao *finO*, responsável por codificar uma proteína de 186 aminoácidos, capaz de inibir a tradução do gene *traJ,* o regulador positivo do operon *tra* (VAN BIESEN; FROST, 1992). O produto do gene *finO* é capaz de estabilizar o mRNA sintetizado por *traJ*, reprimindo sua tradução através de um anti-RNA, codificado pelo gene *finP* (DEMPSEY, 1989), e inibindo o sistema conjugativo. A presença desse gene,

portanto, sugere que os plasmídeos do grupo pBH100 necessitem de um mecanismo adicional para ativação do operon *tra*. Trabalhos realizados com o plasmídeo NR1 e outros plasmídeos do tipo F sugerem que a ausência do promotor de *finP* (DEMPSEY, 1989), ou a inativação de *finO*, via inserção de um IS (YOSHIOKA; OHTSUBO; OHTSUBO, 1987), podem ser responsáveis pelo silenciamento da repressão do sistema conjugativo.

Em seguida, é apresentado o gene *rmoA*, que apresenta homologia às proteínas Hha, envolvidas como moduladores da expressão gênica de diversos fatores de virulência (MADRID; NIETO; JUÁREZ, 2002). Esse gene, também, está associado à regulação das taxas de transferência do plasmídeo NR1, sob diferentes condições de osmolaridade (NIETO et al., 1998), sugerindo que a conjugação dos plasmídeos do grupo pBH100 sejam influenciadas por fatores ambientais.

A jusante do gene *rmoA* verificam-se os genes responsáveis pela replicação do plasmídeo: *repA2*, *repA1* e *repA4*. Esses genes estão associados, respectivamente, à regulação transcricional de *repA1* (LIU; CHURCHWARD; CARO, 1983); à síntese da proteína replicativa, que age como fator de iniciação capaz de se associar à origem de replicação (WU et al., 1992); e à estabilidade de sua herdabilidade (JIANG et al., 1993).

Logo após, destacam-se os genes *pemI* e *pemK*, associados à morte de células filhas que não tenham, porventura, herdado o plasmídeo. Em específico, *pemK* codifica uma endonuclease capaz de interferir na função de diversos mRNAs (ZHANG et al., 2004) e *pemI* agiria como regulador da expressão de *pemK* (TSUCHIMOTO; OHTSUBO; OHTSUBO, 1988).

3.2.4.3 ELEMENTOS MÓVEIS

Dentre os elementos móveis observados em todos os indivíduos desse grupo, destacam-se a presença de ISs e transposons. O Tn21 está presente em todos os indivíduos desse grupo e é responsável pela mobilização do operon *mer* e do gene de resistência à estreptomicina. Esse transposon contém o integron do grupo 1, In2, apresentando o gene *aadA1*, como elemento mais proximal do promotor desse integron. Outros dois IS são encontrados a jusante do gene *aadA1*, a saber o IS1353, da família do IS3 e o IS1326, da família do IS21 (LIEBERT; HALL; SUMMERS, 1999a). Esse transposon parece estar inserido no Tn9, carreador do gene *cat* (ALTON; VAPNEK, 1979).

Outros elementos presentes são o Tn10 e um possível transposon, aqui denominado Tn5-like. O Tn10 possui em sua composição o operon que confere resistência à tetraciclina (PEPE et al., 1997). Dentro desse transposon, também foi verificada a presença do gene *aphA1*, circundado pe-

los elementos de inserção do tipo IS5. Uma vez que o transposon Tn5 é descrito como carreador de um gene de resistência à canamicina, mas circundado por elementos móveis do tipo IS50 (STEINIGER-WHITE; RAYMENT; REZNIKOFF, 2004), e mediante a não caracterização de um transposon composto, por sequências de inserção do tipo IS5, optou-se por denominar a sequência flanqueada por esses elementos, como sendo um potencial transposon, denominado Tn5-like. De fato, estudos anteriores corroboram a hipótese de o gene de resistência à canamicina estar presente em um transposon (informação verbal)[28].

Dos elementos de inserção, preditos no ISFinder, foram removidos os IS91 e Tn3, uma vez que suas regiões eram referentes ao gene da integrase do In2 e dos genes *tnpA* e *tnpR*, do Tn21, respectivamente. A localização dos demais elementos de inserção foram obtidas através do script **ISFInder2gff.pl**. De forma geral, os elementos de inserção, salvo IS1, apresentam a mesma composição em todos os replicons do grupo (**Figura 3.26**).

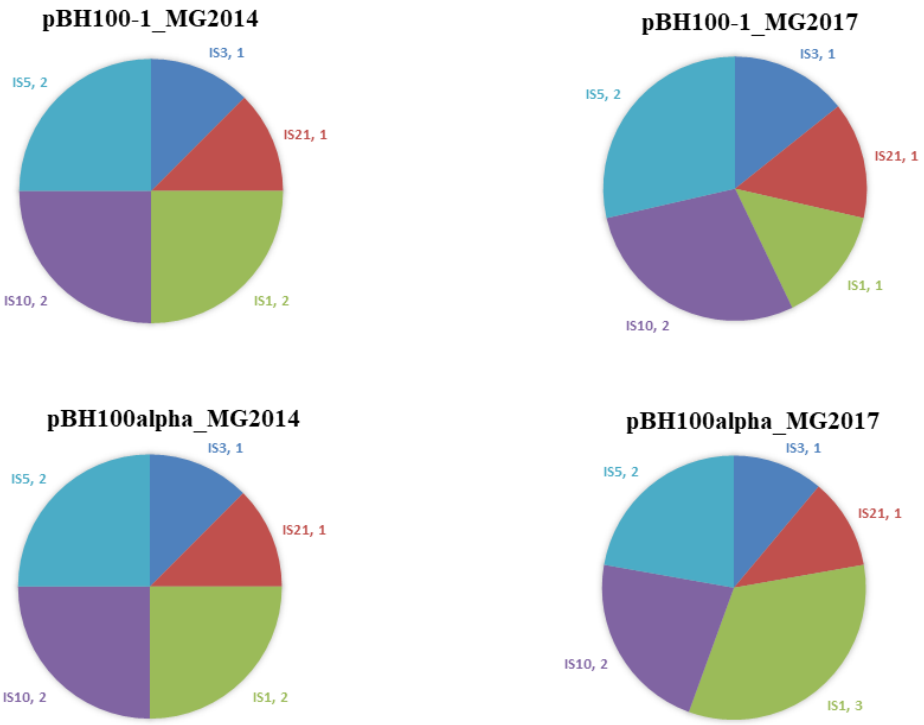


Figura 3.26 Composição de Elementos de Inserção (ISs) nos diferentes replicons do grupo pBH100.

[28]Nascimento, A.M.A. (Instituto de Ciências Biológicas, UFMG). Comunicação pessoal, 2017.

3.2.4.4 DIFERENÇAS ENTRE OS PLASMÍDEOS PBH100

Os gráficos de sintenia dos plasmídeos do grupo pBH100 sugerem alta similaridade entre os diferentes elementos do grupo (**Figura 3.27**). De fato, as maiores diferenças entre esses replicons se deve ao posicionamento de elementos móveis ao longo de suas sequências. Os plasmídeos do subgrupo pBH100alpha não apresentam o transposon Tn3 na sua composição (**Figura 3.28**). Esse elemento, característico dos plasmídeos do grupo pAp, é visto apenas nos plasmídeos pBH100-1, o que sugere que tenham sido adquiridos do plasmídeo pAp. A presença desse transposon já havia sido predita, mediante as observações de resistência fenotípica à ampicilina mesmo na ausência de plasmídeos do grupo pAp (DE LACERDA, 1995; GUIMARÃES, 1986). Além disso, esse transposon parece ter se estabilizado nos plasmídeos desse subgrupo, fato que pode explicar a dificuldade na obtenção de transconjugantes não resistentes à ampicilina (informação verbal)[29].



Figura 3.27 Blocos contínuos construídos pelo Mauve revelam a sintenia entre os plasmídeos do grupo pBH100.

As diferenças estruturais observadas entre os replicons desse grupo devem-se à posição dos elementos móveis, conforme ilustrado na **Figura 3.28**. Nela, pode-se observar a presença do IS1 inserido no transposon Tn10, entre IS5 e IS10. Esse elemento foi inserido no gene *arsR*, responsável por codificar uma proteína regulatória do operon *ars*, que confere resistência ao arsênio. No entanto, esse operon não está presente nos plasmídeos do grupo pBH100. RASTORGUEV,

[29]Nascimento, A.M.A. (Instituto de Ciências Biológicas, UFMG). Comunicação pessoal, 2017.

ZAVILGELSKY e TCHURIKOV (1998) ao avaliarem esse gene no plasmídeo R64, também ausente do operon *ars*, sugerem que o produto desse gene possua ação similar aos genes *ard*, responsáveis por prevenir a ação de enzimas de restrição do tipo I.

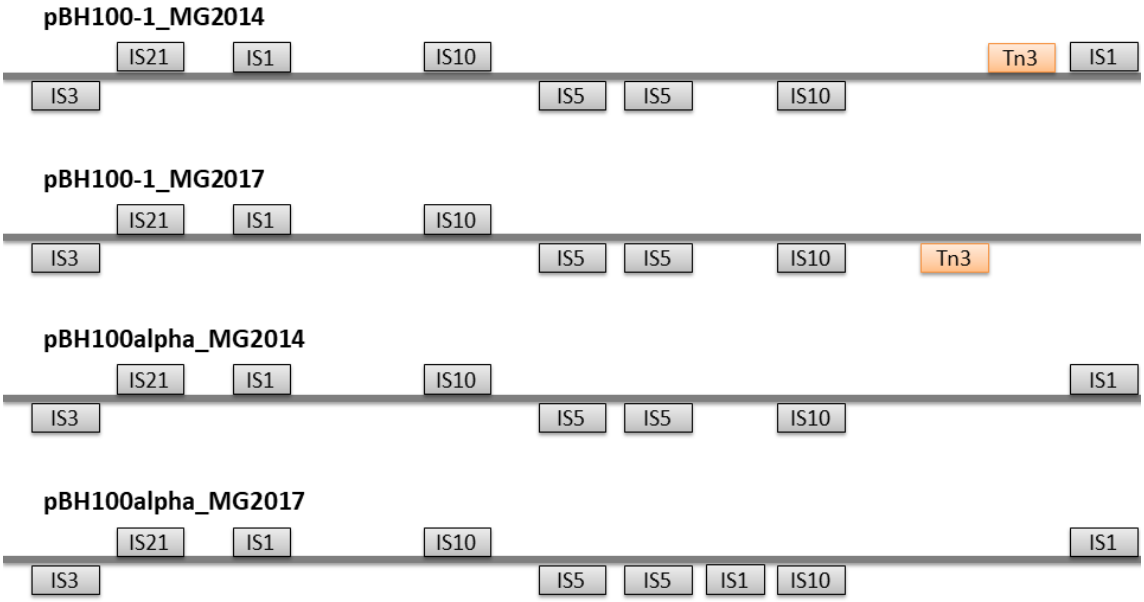


Figura 3.28 Posição dos diferentes ISs, ao longo dos replicons do grupo pBH100. Anotações na parte superior da linha em cinza indicam ISs presentes na fita *sense*. Já anotações dispostas na porção inferior indicam a presença de ISs na fita complementar.

Outra observação é a ausência do IS1 na porção terminal de pBH100-1 da *E. coli* BH100 subtr. MG2017. Esse elemento de inserção parece não estar associado à fragmentação de nenhum gene nos demais replicons desse grupo, além de não estar próximo dos promotores dos genes a montante (gene hipotético) ou a jusante (*merR*). Com efeito, sua presença/ausência nesse sítio, possivelmente não deve interferir com a função de outros genes nesses replicons. Entretanto, esse elemento móvel é predito como sendo um componente do transposon Tn9 (ALTON; VAPNEK, 1979), e portanto é possível que esse transposon seja não-mobilizável em pBH100-1_MG2017.

A terceira diferença estrutural refere-se ao Tn3, presente no subgrupo pBH100-1. Nesse caso, o Tn3 foi inserido mais a jusante do gene *pemK* no pBH100-1 da *E. coli* BH100 subtr. MG2014, quando comparado com pBH100-1 da *E. coli* BH100 subtr. MG2017, inserido, invertido, após o gene *repA4*.

3.2.4.5 COMPARATIVO FUNCIONAL

A análise comparativa funcional das CDSs, anotadas nos diferentes replicons desse grupo, revelou grande semelhança no seu conteúdo gênico. As diferenças observadas da **Tabela 2.5** devem-se, primariamente, a erros sistemáticos do processo de anotação e curadoria. Nesse caso, por exemplo, algumas CDSs completamente sobrepostas foram removidas em alguns replicons, mas não em outros. A análise manual dos genes classificados como únicos ou acessórios revelou que a maior parte dos mesmos era comum a todos os replicons.

As diferenças reais consistem em três genes anotados como pseudogenes em pBH100alpha da *E. coli* BH100 subtr. MG2014 e de duas falsas CDSs anotadas pelo RAST, mas que se encontravam fragmentas pela presença de um IS. O **Quadro 3.6** resume as principais diferenças em termos de conteúdo gênico entre os replicons do grupo pBH100.

Quadro 3.6 Relação de genes preditos como não funcionais nos membros de pBH100

Gene	Ausente em:	Motivo
Gene hipotético	pBH100-1_MG2014	*frameshifted*
tniB	pBH100-1_MG2014	*frameshifted*
merA	pBH100-1_MG2014	*frameshifted*
arsR	pBH100alpha_MG2017	Inserção de IS

A **Figura 3.29** indica os termos GO associados a processos biológicos presentes no grupo pBH100. Nessa figura pode-se notar a presença de genes associados ao metabolismo geral e que se referem a alguns dos demais genes presentes nos plasmídeos desse grupo, mas não representados na **Figura 3.23**.

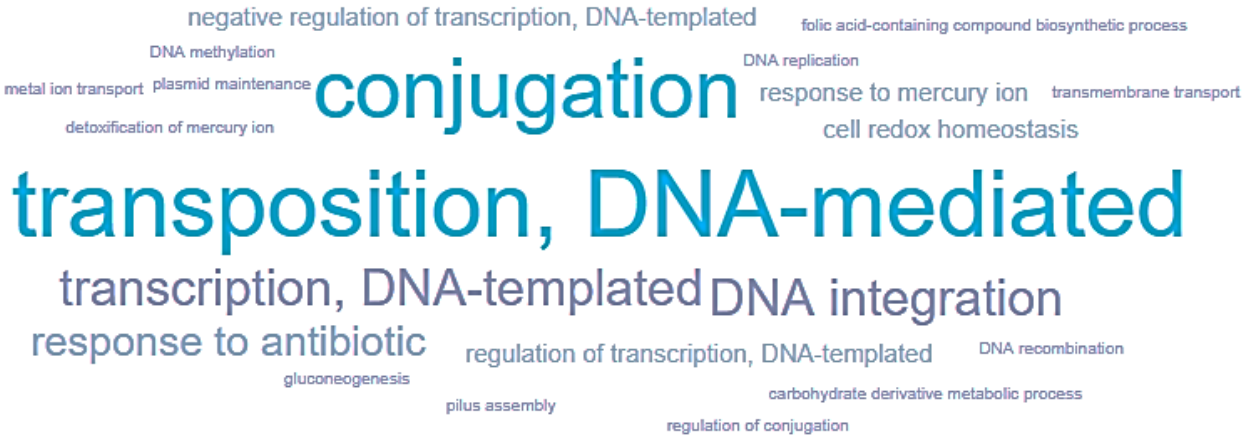


Figura 3.29 Nuvem de palavras de termos GO associados a processos biológicos e vinculados ao grupo pBH100.

3.2.4.6 IDENTIFICAÇÃO DAS ORIGENS DE REPLICAÇÃO (*ORIV*) E ORIGEM DE TRANSFERÊNCIA (ORIT)

A identificação das origens de replicação e transferência dos plasmídeos do grupo pBH100 permite a elaboração de estudos futuros a fim de torná-los verdadeiras ferramentas de engenharia genética. Para tanto, objetivou-se aqui estimar *in silico* a provável região responsável pelo início da replicação e da transferência desses plasmídeos.

A partir da comparação desses plasmídeos com uma base de dados contendo origens de transferência encontradas em *E. coli,* pôde-se identificar a *oriT* dos plasmídeos do grupo pBH100 como sendo homóloga a do plasmídeo NR1. Essa origem de transferência é encontrada a montante do operon *tra,* tal como a origem de transferência do plasmídeo NR1 (WOMBLE; ROWND, 1988). De igual forma, usando uma base de dados de sequências contendo origens de replicação de *E. coli*, foi possível a identificação da provável origem de replicação, próxima ao gene *repA1*, conforme descrito na **Tabela 3.3**.

Tabela 3.3 Inferência de oriV e oriT em pBH100 a partir de BLASTn contra uma base de dados contendo sequências referentes às origens de replicação e transferência.

Região	Inicio	Fim	Identidade	Tamanho no GenBank	Acesso
oriV	93.031	94.142	1.109/1.112	1.112	V00325.1
oriT	53.668	53.830	156/163	450	X00783.1

3.2.4.7 INFERÊNCIA DO NÚMERO DE CÓPIAS

Diversas metodologias para se estimar o número médio de cópias de plasmídeos por célula já foram desenvolvidas e dependem de quantificação de DNA (FRIEHS, 2004). A fim de inferir *in silico* o número de cópias desses plasmídeos, a partir dos dados de NGS, buscou-se avaliar de forma proporcional a cobertura de *reads* nos plasmídeos em relação ao cromossomo. Assumindo um sequenciamento não enviesado, pode-se supor que o número de *reads* mapeadas seja proporcional ao número de vezes que essa região esteja amostrada. De fato, a inferência do número de cópias de regiões repetitivas no genoma, a partir da comparação da cobertura dessas regiões com regiões não repetitivas já vem sendo aplicada (CAMPBELL et al., 2008; MAGI et al., 2012; SWAMINATHAN; VARALA; HUDSON, 2007). Desse modo, observando a proporção da cobertura média do mapeamento de *reads* nos *contigs* presentes uma única vez no cromossomo e nos *contigs* presentes de forma única nos plasmídeos, inferiu-se seu número de cópias por célula. Para tanto, calculou-se a média aritmética da profundidade de cobertura observadas em cinco posições escolhidas arbitrariamente e que satisfizessem o critério de unicidade descrito acima. A **Tabela**

3.4 indica a estimativa do número de cópias em cada uma das linhagens carreadoras de plasmídeos do grupo pBH100.

Tabela 3.4 Estimativa do número de cópias de cada plasmídeo do grupo pBH100, a partir da proporção entre a profundidade de cobertura no plasmídeo em relação ao cromossomo

Linhagem	Plasmídeo	Cobertura média do cromossomo	Cobertura média do plasmídeo	Estimativa do número de cópias
E. coli BH100 subtr. MG2014	pBH100-1	162,8 ± 8,8	152,8 ± 35,5	1 – 2
E. coli BH100 subtr. MG2017	pBH100-1	449,4 ± 14,8	1610,6 ± 122,8	3 – 4
E. coli BH100L subtr, MG2014	pBH100alpha	76,8 ± 4,9	265,4 ± 10,0	3 – 4
E. coli BH100L subtr. MG2017	pBH100alpha	383,4 ± 25,9	1.487,8 ± 92,4	3 – 4

Interessantemente, o número de cópias de pBH100-1 estimados na linhagem *E. coli* BH100 subtr. MG2014 é cerca de metade do estimado nas demais linhagens. Esse fato pode ser devido a condições intrínsecas da metodologia de cultivo ou de extração de DNA dessa linhagem.

As estimativas apontadas na **Tabela 3.4** sugerem que os plasmídeos do grupo pBH100 sejam *low*-copy. Esses dados estão de acordo com as seguintes evidências: 1) os plasmídeos do grupo pBH100 não são facilmente extraídos em grandes quantidades a partir de protocolos convencionais de Miniprep (Informação verbal)[30]; 2) é fato conhecido que o plasmídeo NR1 apresenta replicação estringente (DONGT et al., 1989), e a homologia verificada de sua *oriV* com as sequências desse grupo corroboram os dados observados; 3) o tamanho aumentado desses plasmídeos também sugerem, em geral, um menor número de cópias por célula.

3.2.5 Os plasmídeos do grupo pAp

As análises do PlasmidFinder sugerem que os plasmídeos desse grupo pertençam ao grupo de incompatibilidade ColRNAI. O conteúdo gênico dos principais marcadores anotadas nos plasmídeos desse grupo está indicado na **Figura 3.30**. Genes não representados nessa figura codificam para proteínas hipotéticas preditas pelo RAST.

[30]Nascimento, A.M.A.; Chartone-Souza, E. (Instituto de Ciências Biológicas, UFMG). Comunicação pessoal, 2017.

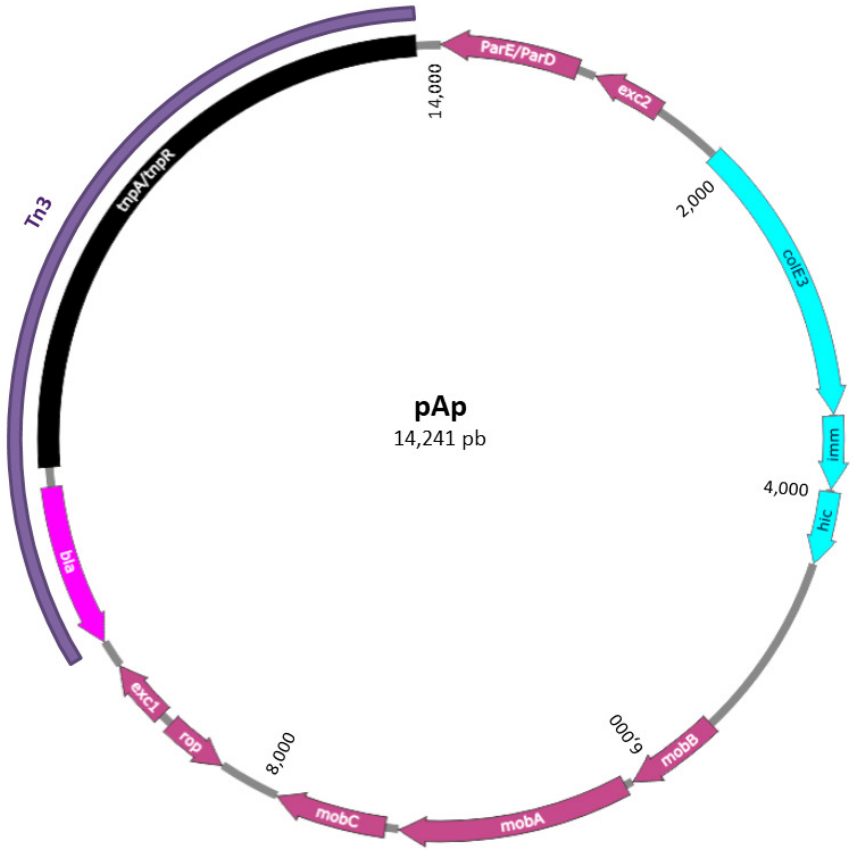


Figura 3.30 Representação esquemática das CDSs anotadas nos plasmídeos do grupo pAp

3.2.5.1 GENES DE RESISTÊNCIA E TOXICIDADE

3.2.5.1.1 A resistência à β-lactâmicos

Conforme já abordado os plasmídeos do grupo pAp carregam consigo o Tn3 que apresenta o gene *bla* codificante de uma β-lactamase. Essa enzima é responsável por clivar antibióticos que contenham um anel β-lactâmico impedindo sua ação inibitória contra a síntese de peptidoglicano. A β-lactamase codificada pelo gene *bla* desse transposon foi identificada como do grupo TEM-1A, corroborando os achados bioquímicos descritos anteriormente na *E. coli* BH100 (DE LACERDA, 1995).

3.2.5.1.2 O operon de colicina

Surpreendentemente, foi observado um operon contendo os genes *colE3*, *imm* e *hic,* responsável pela síntese e liberação de colicina E3, muito embora e *E. coli* BH100 tenha sido descrita

como não-colicinogênica (CHARTONE-SOUZA, 1975). Em específico, o gene *colE3* é responsável por codificar a colicina do tipo E3; o gene *imm*, responsável por assegurar a imunidade da célula produtora de colicina, inibindo sua atividade no citosol (JAKES; MODEL, 1979); e o gene *hic,* descrito como importante regulador da atividade da colicina, garantindo sua liberação para o meio extracelular (WATSON et al., 1984).

A não verificação de síntese de colicinas pela *E. coli* BH100 pode ser devido a não ativação do promotor desse operon. De fato, a síntese de muitas colicinas estão condicionadas à ativação do sistema SOS (GILLOR; VRIEZEN; RILEY, 2008), o que pode não ter sido verificado nas condições ensaiadas. De Lacerda (1995), por exemplo, ao observar redução na taxa de crescimento de *E. coli* C600, quando em cultivo misto com a *E. coli* BH100, sugere a síntese de colicinas como possível explicação para o fenômeno. De toda forma, experimentos adicionais são necessários a fim de observar a funcionalidade desse operon.

3.2.5.2 OUTROS GENES DE DESTAQUE

3.2.5.2.1 Genes associados à segregação

Três genes relacionam-se à presença desses plasmídeos nas células-filhas após divisão celular. O primeiro deles codifica para uma toxina vinculada à família das proteínas do sistema toxina-antitoxina ParDE. Nesse sistema a toxina sintetizada por *parE* codificaria para toxina capaz de inibir a proliferação celular de células que não tenham herdado o plasmídeo e *parD* codificaria para a respectiva anti-toxina. A inibição da divisão celular ocorreria graças às diferentes taxas de *turnover* dessas proteínas (JENSEN; GERDES, 1995) e a capacidade da toxina ParE em inibir a DNA girase (JIANG et al., 2002).

Os demais genes são denominado *exc1* e *exc2* e codificam pra *entry exclusion protein 1* e *entry exclusion protein 2*, respectivamente. Esses produtos estariam em tese associados à capacidade de tornar a célula um recipiente não favorável à entrada de outros plasmídeos (GARCILLÁN-BARCIA; DE LA CRUZ, 2008). No entanto, YAMADA, YAMADA e NAKAZAWA (1995) não observaram evidências para esse efeito em relação a esses genes presentes no plasmídeo colE1, sugerindo que a função desses genes ainda carece de maiores observações.

3.2.5.2.2 Gene de controle da replicação

O gene *rop*, localizado a jusante de *exc1* codifica para a proteína Rop, responsável por estabilizar a interação entre RNAI e RNAII, no chamado *kissing complex*, e contribuindo para o controle do número de cópias do replicon (CESARENI; MUESING; POLISKY, 1982; SOM; TOMIZAWA, 1983).

3.2.5.2.3 Genes de mobilização

Esse conjunto de genes está associado à transferência dos plasmídeos desse grupo, via mecanismo de mobilização. Nesse caso, o gene *mobA* seria responsável pelo reconhecimento da *oriT* e levaria à mobilização do plasmídeo (SMILLIE et al., 2010). Já, o gene *mobC* deve atuar como uma proteína acessória do processo (NÚÑEZ; DE LA CRUZ, 2001).

3.2.5.3 DIFERENÇAS ENTRE OS PLASMÍDEOS

Em termos de sintenia os membros do grupo pAp são muito similares entre si, salvo o plasmídeo pApR (**Figura 3.31**). Este plasmídeo contém o transposon Tn21, identificado nos plasmídeos do grupo pBH100. Esse fato sugere que na *E. coli* BH100 subtr. MG2017 esse transposon tenha sido transferido para o plasmídeo pAp, conforme ilustrado na **Figura 3.32**. A porção desse plasmídeo, não referente ao Tn21 é idêntica ao plasmídeo pAp obtido na *E. coli* K12T.



Figura 3.31 Representação de blocos contíguos, criados pelo Mauve evidenciando a sintenia dos diferentes plasmídeos do grupo pAp.

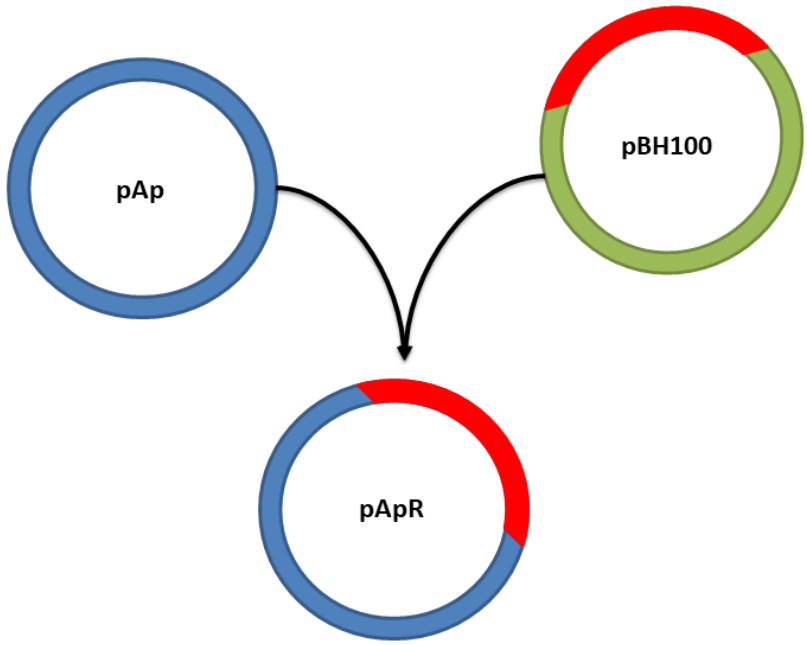


Figura 3.32 Mecanismo de surgimento do plasmídeo pApR. Nesse caso, plasmídeos do grupo pBH100 servem como fonte do Tn21 (em vermelho), capaz de se inserir no plasmídeo pAp.

As principais diferenças em termos de composição se resumem a 6 nucleotídeos presentes/ausentes em pAp da *E. coli* BH100 subtr. MG2014, conforme descritos a seguir:

A. Ausência de uma citosina na posição 3.994: essa deleção trata-se de um artefato intrínseco do processo de curadoria. Nesse caso, verifica-se a presença do nucleotídeo no mapeamento das *reads* nessa posição (**Figura 3.33**). Nessa região, foi anotada uma CDS hipotética, apenas na

E. coli BH100 subtr. MG2014. Dessa forma, a inserção do nucleotídeo deve tornar essa possível CDS num pseudogene.

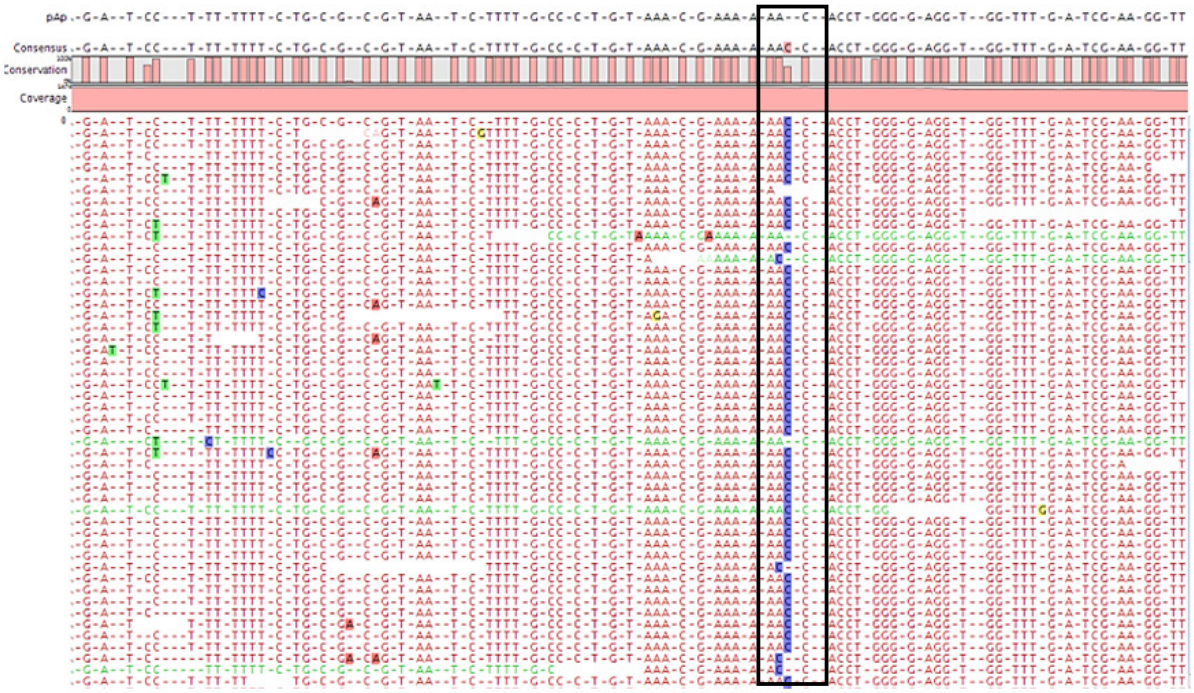


Figura 3.33 Mapeamento de *reads fragment* no plasmídeo pAp_MG2014. Em destaque, a posição 3.994, situada após região homopolimérica, revela a presença de citosina, não adicionada ao replicon plasmidial, por inacurácia no processo de montagem e curadoria.

B. Ausência de uma guanina na posição 4.909: Essa deleção parece se tratar de uma possível variante na *E. coli* BH100 subtr. MG2014, uma vez que as *reads* mapeadas nessa região corroboram a presença e a ausência da base (**Figura 3.34**)



Figura 3.34 Mapeamento de *reads fragment* no plasmídeo pAp_MG2014. Em destaque, a posição 4.909, cujas *reads* corroboram tanto a presença quanto a ausência de guanina.

C. Ausência de uma citosina na posição 5.054: a presença dessa base nos demais plasmídeos levou à anotação de uma CDS hipotética menor, quando comparada à mesma anotação na *E. coli* BH100 subtr. MG2014.

D. Ausência de uma guanina na posição 5.887: a ausência desse nucleotídeo leva à anotação de duas CDSs referentes ao gene *mob* na *E. coli* BH100 subtr. MG2014. No plasmídeo contido na *E. coli* K12T, a presença desse nucleotídeo levou anotação de uma única CDSs codificante para uma proteína de 654 aminoácidos. Já, em pApR, o nucleotídeo encontra-se presente, mas o gene é fragmentado pela inserção do Tn21, e, por conseguinte, anotado como duas CDSs distintas pelo RAST (**Figura 3.35**). Comparando a sequência proteica da maior CDS anotada em pAp_MG2017 e as CDSs fragmentadas nos demais plasmídeos desse grupo, via BLASTp com a base de dados do Uniprot, observou-se que ambos os produtos, também já foram anotados em outras bactérias. Esse fato pode ser devido a inconsistências na anotação desse gene em diversas linhagens depositadas no banco de dados.



Figura 3.35 Anotações diferenciais de genes *mob* (em amarelo). A ausência de uma guanina na posição 5.887 do plasmídeo pAp_MG2014 leva à anotação de duas CDSs, também anotadas em outros organismos presentes na base do Uniprot.

E. Ausência de uma guanina na posição 10.255: a inserção de uma guanina nos demais plasmídeos desse grupo leva a diferente anotação da porção 5' do gene *bla*. Nesse caso, não se verificam os primeiros 49 resíduos de aminoácidos da β-lactamase nos demais plasmídeos, sendo essa região substituída por outros 18 resíduos presentes em outro quadro de leitura. A comparação das sequências protéicas preditas nesses plasmídeos, via BLASTp com a base de dados do Uniprot, apresenta maior suporte à anotação contendo 286 aminoácidos do que aquela contendo 316. Esse fato pode sugerir a não funcionalidade da β-lactamase, sintetizada pelo gene *bla* em pAp_MG2014, ou a presença de uma possível variante estrutural dessa enzima. A **Figura 3.36** revela o alinhamento global da sequência polipeptídica dessa enzima, presente nos grupos pAp e subgrupo pBH100-1.

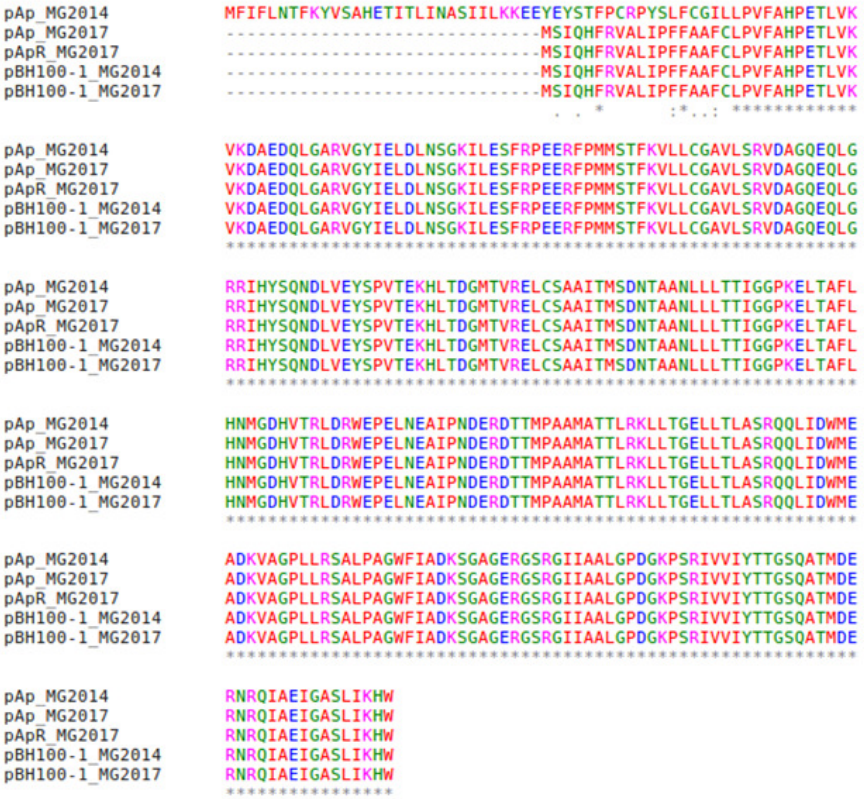


Figura 3.36 Alinhamento global das sequências protéicas, referentes ao gene *bla,* encontradas nos diferentes replicons plasmidiais da *E. coli* BH100.

F. Ausência de uma guanina na posição 11.203: essa região carece de anotação de CDSs, em todos os replicons desse grupo. A mudança da base, portanto, pode não ter efeito fisiológico.

As diferenças apresentadas acima podem se refletir em efeitos biológicos. Em particular, as alterações descritas nos itens D e E podem afetar, respectivamente, o mecanismo de mobilização e a funcionalidade da β-lactamase. No entanto, vale destacar que as diferenças aqui apresentadas também podem ser devido a vieses intrínsecos das tecnologias de sequenciamento utilizadas.

Já em relação às características funcionais anotadas nesses plasmídeos, identificou-se que as diferenças funcionais observadas da **Tabela 2.5** devem-se a incoerências no processo de anotação e curadoria. De fato, a análise cuidadosa dos genes classificados como acessórios e únicos pelo BPGA indicam sua presença nos demais replicons do grupo pAp. As alterações nucleotídicas, discutidas acima, entretanto, revelam a presença de potenciais pseudogenes, surgidos por mudanças no quadro de leitura e não anotados nos replicons.

3.2.5.4 UM POSSÍVEL MECANISMO DE MOBILIZAÇÃO

Ramsay e Firth (2017) revisaram diversos mecanismos de transferência de plasmídeos não-conjugativos não envolvendo cointegração ou condução. Devido a presença do gene *mob* nos plasmídeos do grupo pAp e comparando aos mecanismos propostos da **Figura 3.37**, prevê-se que a mobilização dos plasmídeos desse grupo ocorra segundo descrito em **ia**. Nesse caso, os plasmídeos do grupo pBH100 codificariam para o canal transportador T4SS e a proteína acessória T4CP sintetizados pelo operon *tra*.

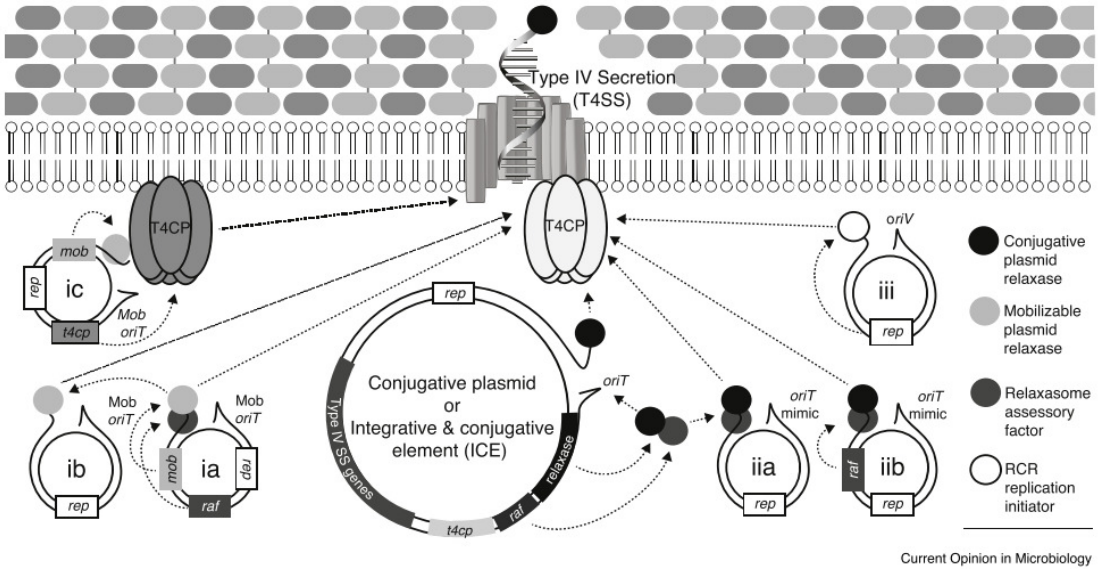


Figura 3.37 Diferentes mecanismos de mobilização não envolvendo cointegração e condução. Nesse caso a mobilização poderia ocorrer através da atividade do gene *mob* sob uma oriT distinta do plasmídeo conjugativo, e podendo em certos casos contar com a presença de uma proteína acessória (ia); através de uma oriT específica para ação de Mob, não sintetizada pelo plasmídeo (ib); através da presença de oriT, *mob* e gene de síntese de T4CP distinta do plasmídeo conjugativo (ic); através da presença de uma oriT similar à oriT do plasmídeo conjugativo (iia); através da presença de uma oriT similar à oriT do plasmídeo conjugativo, mas contendo proteínas acessórias (iib); através da síntese uma relaxase envolvida com processos replicativos e capaz de se associar à T4CP (iii). Fonte: (RAMSAY; FIRTH, 2017)

A presença do gene *mob* nos plasmídeos do grupo pAp corroboram com o modelo de mobilização proposto em **ia**. Entretanto, a partir da observação dos mecanismos **iia** e **iib**, questionou-se a possibilidade da origem de transferência desses plasmídeos ser homóloga as dos plasmídeos conjugativos do grupo pBH100, o que sugeriria outro mecanismo de mobilização. Para tanto, comparou-se a oriT predita em pBH100 com os plasmídeos do grupo pAp via BLASTn, e como resultado, não foi verificada similaridade entre as sequências. Com efeito, sugere-se aqui que a aborda-

gem descrita em **ia** por Ramsay e Firth (2017) seja a mais provável para descrever o mecanismo de mobilização dos plasmídeos do grupo pAp.

Entretanto, a sequência admitida como sendo um erro de montagem e aqui denominada pBH100-x, observada utilizando *reads mate-pair* na *E. coli* BH100 subtr. MG2014 pode corroborar um potencial mecanismo de mobilização do plasmídeo pAp via cointegração aos plasmídeos do grupo pBH100. O mapeamento de *reads fragment* e até mesmo *paired-end* Illumina da *E. coli* BH100 subtr. MG2017 não corroboram para a montagem de pBH100-x. No entanto, sua verificação mediante o uso das *reads mate-pair* pode ser devido uma maior amostragem das sequências, já que 10x mais DNA foram utilizadas para construção dessa biblioteca.

3.2.5.5 IDENTIFICAÇÃO DAS ORIGENS DE REPLICAÇÃO (*oriV*) E ORIGEM DE TRANSFERÊNCIA (*oriT*)

A identificação da origem de replicação (ori) dos plasmídeos do grupo pAp se deu através de BLASTn contra uma base de dados contendo origens de replicação anotadas em *E. coli*. Dessa forma, observou-se homologia de uma região à origem de replicação descrita no plasmídeo colE1 (**Tabela 3.5**). Essa região está presente entre os genes de síntese de colicina e dos genes de mobilização tal como descritos no plasmídeo colE1 (DOUGAN et al., 1978).

Em relação à origem de transferência não foi encontrada homologia entre as sequências desse grupo com sequências referentes a oriT presentes na base de dados. No entanto, a identificação da origem de replicação homóloga à ori de colE1 sugere que oriT esteja a jusante desse sítio. De fato, oriT nos plasmídeos colE1 é encontrada logo após a origem de replicação (DOUGAN et al., 1978; VARSAKI et al., 2009). Além disso, é altamente provável que a origem de transferência esteja situada a jusante dos genes *mob* (Informação verbal)[31].

Tabela 3.5 Predição da origem de replicação (ori) nos plasmídeos do grupo pAp

Região	Inicio	Fim	Identidade	Tamanho no GenBank	Acesso
ori	3935	4481	459/566	602	V00270.1

3.2.5.6 PREDIÇÃO DO NÚMERO DE CÓPIAS

A fim de investigar o número de cópias dos plasmídeos do grupo pAp por célula, seguiu-se a mesma abordagem descrita para identificação do número de cópias dos plasmídeos do grupo

[31]Ramsay, J.P. (School of Biomedical Sciences, Curtin University). Comunicação pessoal, 2017

pBH100. As análises da profundidade de cobertura e estimativa do número de cópias são demonstradas na **Tabela 3.6**.

Tabela 3.6 Predição do número de cópias dos plasmídeos do grupo pAp a partir da proporção da profundidade da cobertura média observada nos plasmídeos em relação ao cromossomo

Linhagem	Plasmídeo	Cobertura média no cromossomo	Cobertura média no plasmídeo	Estimativa do número de cópias
E. coli BH100 subtr. MG2014	pAp	162,8 ± 8,8	1.151,6 ± 38,5	7 - 8
E. coli BH100 subtr. MG2017	pApR	449,4 ± 14,8	798,8 ± 35,5	1 - 2
E. coli K12T subtr. MG2017	pAp	422,8 ± 34,8	4037,0 ± 194,7	9 - 10

A classificação *low-copy* ou *high-copy* do plasmídeo pAp devido sua homologia com a ori de colE1 não é trivial, uma vez que alterações na sequência dessa origem podem alterar consideravelmente o número de cópias do plasmídeo. No entanto, as estimativas do número de cópias por célula obtidos pela análise da profundidade da cobertura sugerem um baixo número de cópias desses replicons. Com efeito, tomando a homologia encontrada com oriV do plasmídeo colE1, originalmente possuindo de 15 a 20 cópias por célula (SAMBROOK; RUSSELL, 2001), somado às estimativas do número de cópias aqui calculadas, sugere-se que os plasmídeos desse grupo sejam *low-copy*.

Baseado na possível origem de replicação desses plasmídeos e a presença do gene *rop* pode-se supor que esses replicons sigam o modelo de replicação do tipo teta. Nesse sistema, o número de cópias sofre extensivo controle dependendo das condições do metabolismo celular (LINCHAO; BREMER, 1986). Nesse caso, por exemplo, pode haver controle da expressão dos transcritos RNAI/RNAII, ou ainda mutações que interfiram no pareamento desses RNAs (SAMBROOK; RUSSELL, 2001) e que são capazes de afetar o início da replicação. Além disso mutações que alterem a estrutura da proteína Rop, capazes de comprometer o *kissing complex*, também podem levar a alterações significativas no número de cópias (CASTAGNOLI et al., 1989).

As diferentes estimativas no número de cópias observada nos plasmídeos pAp da *E. coli* BH100 subtr. MG2014 e *E. coli* K12T subtr. MG2017 podem estar relacionadas a diferenças na capacidade metabólica natural dessas linhagens, também provavelmente influenciada pela presença do plasmídeo pBH100-1. Já na *E. coli* BH100 subtr. MG2017 observa-se uma estimativa muito inferior para o número de cópias de pApR. Esse fato pode ser devido ao aumento no tamanho do plasmídeo (LOPES et al., 1996; SMITH; BIDOCHKA, 1998) que pode contribuir para o surgi-

mento de estruturas secundárias capazes de interferir com a iniciação da replicação. Além disso, a redução pode estar associada a um aumento na pressão metabólica provocada pelos genes presentes no Tn21.

CAPÍTULO 4 MODELO DE SURGIMENTO DE MUTANTES DE ALTA FREQUÊNCIA RESISTENTES À ESTREPTOMICINA

Conforme descrito por Chartone-Souza, 1975, a *E. coli* BH100 não apresenta resistência clínica à estreptomicina. A presença do gene *aadA1* nos plasmídeos do grupo pBH100 já havia sido observada (Informação verbal)[32] e seria responsável pela resistência mínima a estreptomicina verificada nessa bactéria (Informação verbal)[33]. Neste trabalho constatou-se que o gene *aadA1* está localizado no integron In2 inserido no Tn21 (**Figura 4.1**).

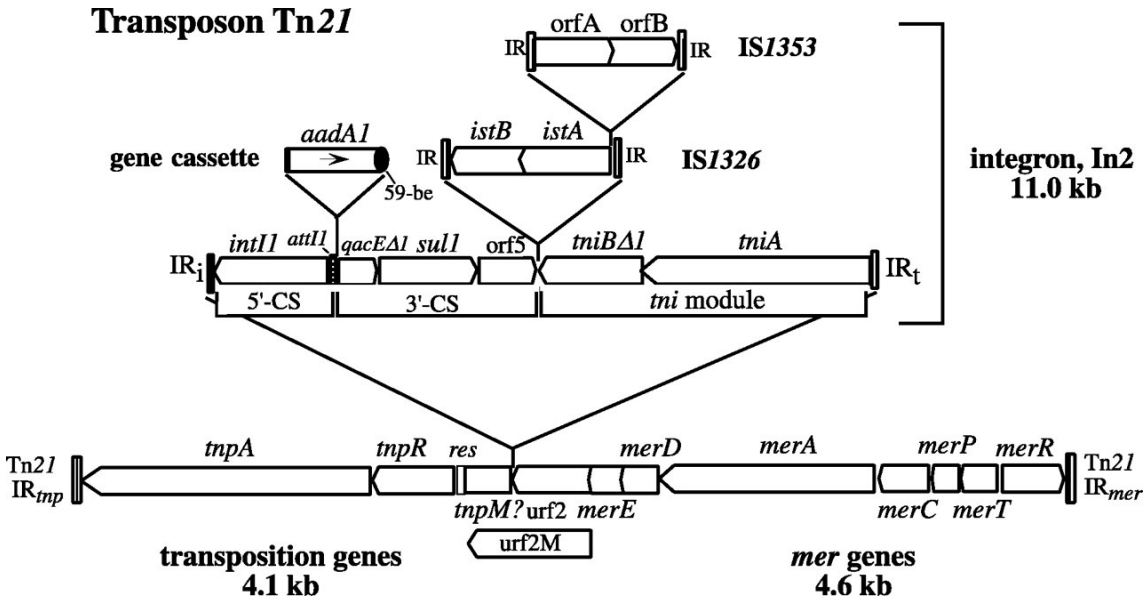


Figura 4.1 Representação esquemática do Tn21. Fonte: (LIEBERT; HALL; SUMMERS, 1999c)

De Paula (1988) determinou que a *E. coli* BH100 cresce a 1,0 µg.mL^{-1} e 15,0 µg.mL^{-1} de estreptomicina em ágar nutriente e meio Müller-Hinton, respectivamente. MIKI, EASTON e ROWND (1978) utilizam 6,0 µg.mL^{-1} desse antibiótico para seleção de indivíduos contendo os plasmídeos NR1 em ágar nutriente. COLLIS e HALL (1995) no entanto, determinaram IC50 de 360,0 µg.mL^{-1} para resistência à estreptomicina, em meio Luria-Bertani (LB), devido a presença do gene *aadA1* inserido no In2. A partir dessas observações, indagou-se as razões que levariam a *E. coli* BH100 não apresentar maiores níveis de resistência a esse antibiótico.

[32] Nascimento, A.M.A. (Instituto de Ciência Biológicas, UFMG). Comunicação pessoal, 2017.
[33] Chartone-Souza, E. (Instituto de Ciências Biológicas, UFMG). Comunicação pessoal, 2017.

Collis e Hall (1995) investigaram o efeito de diversas variantes dos promotores P_c, além da posição dos genes do cassete em integrons, na expressão gênica. Esses autores constataram a presença de uma variante fraca do promotor P_c em In2, responsável por uma taxa de expressão 20x inferior quando comparada à variante forte desse promotor. No entanto, também destacam a presença do promotor P_2 nesse integron, capaz de elevar os níveis de transcritos de *aadA1*. De fato, a presença de P_c e P_2 no integorn In2 colaboram para expressão de *aadA1* a uma taxa 3x menor do que a variante forte de P_c.

A sequência nucleotídica do Tn21, presente em pBH100, mostrou-se idêntica à descrita no plasmídeo NR1, e que carrega consigo o In2. Esse fato indica que em pBH100, o gene *aadA1* deve ser precedido pelos promotores P_2 e P_c, e que, portanto, deveria apresentar um maior nível de resistência à estreptomicina, conforme indicado nos experimentos conduzidos por Collis e Hall (1995). Diante desses fatos, e sabendo que o gene *aadA1* é o mais proximal desses promotores, sugere-se aqui, que na *E. coli* BH100 esse gene apresente menores níveis de expressão, provavelmente devido à ação de algum fator regulador do promotor P_2.

Dentre os possíveis reguladores, destaca-se a proteína H-NS, relatada como um ativador do promotor P_2 (CAGLE; SHEARER; SUMMERS, 2011). Essa proteína é caracterizada como um repressor genérico da expressão de diferentes genes e está envolvida em diversos processos biológicos (GRAINGER, 2016). Com efeito, presume-se aqui, que uma expressão reduzida do gene *hns* possa estar associada à potencial redução da expressão do gene *aadA1* por P_2.

A expressão do gene *hns* é regulada pela própria H-NS, além da proteína FIS (FALCONI et al., 1996). Diversos *motifs* associados à interação desses reguladores, contidos na região promotora e em outros sítios mais à montante, já foram descritos como importantes para regulação da expressão desse gene (GIANGROSSI; GUALERZI; PON, 2001). Disso posto, propôs-se analisar as sequências nucleotídicas a montante da CDS de *hns*, buscando por alterações que pudessem interferir na associação dos fatores regulatórios desse gene.

Foram extraídos 1.000 pb localizados a montante da CDS de *hns* presentes nos quatro cromossomos completos obtidos nesse trabalho, e de outros seis cromossomos obtidos do GenBank (*E. coli* 536 – CP000247.1; *E. coli* K-12 subtr. DH10B – CP000948.1; *E. coli* K-12 subtr. MG1655 – U00096.3; *E. coli* CFT073 – AE014075.1; *E. coli* O157:H7 str. Sakai – BA000007.2; *E. coli* J53 – AICK01000017). Observando para o alinhamento múltiplo dessas regiões pode-se notar um alto grau de identidade entre as sequências. Analisando as diferenças pontuais no alinhamento em uma região associada a sítios de interação de FIS e H-NS, foi possível constatar algumas substituições em relação às linhagens desse estudo (**Figura 4.2**). Essas modificações po-

dem contribuir para alteração da afinidade desses fatores transcricionais, induzindo a diferentes taxas de expressão do gene *hns*.



Figura 4.2 Alinhamento múltiplo da região a montante do gene *hns* das 4 linhagens da BH100 e outras 6 *E. coli*. Essa região é capaz de interagir com as proteínas H-NS e FIS e regular a expressão do gene *hns*. Em destaque as diferenças nucleotídicas observadas entre essas cepas.

Outra observação de destaque refere-se à presença de elementos de inserção localizados logo a montante do gene *hns* na *E. coli* BH100 subtr. MG2014 e *E. coli* BH100 subtr. MG2017. Esse fato, de igual forma, pode estar associado a alterações nos níveis desses transcritos, já que esses elementos podem contribuir para regulação da expressão gênica.

Além da proteína H-NS, também já foi relatada a influência de proteínas associadas à partição plasmidial no silenciamento gênico (KIM; WANG, 1999; RODIONOV; LOBOCKA; YARMOLINSKY, 1999). Nesse caso, presume-se que o silenciamento ocorreria mediante a inacessibilidade do promotor P_2 aos fatores transcricionais.

Disso posto, sugere-se aqui que a possível redução na atividade do promotor P_2 na *E. coli* BH100 seja devido a uma menor expressão do gene *hns*, à ação dos genes associados à partição do

plasmídeo pBH100, e/ou ainda à ação de algum outro gene repressor de P_2 vinculado ao plasmídeo. Dessa forma, a baixíssima resistência à estreptomicina verificada em BH100 poderia ser alterada frente a mecanismos que contrabalanceassem a potencial baixa atividade desse promotor.

De Paula (1988) chama a atenção para o surgimento de mutantes[34] de alta frequência com resistência variável a estreptomicina na *E. coli* BH100. A autora executa uma série de experimentos a fim de caracterizar a natureza desse fenômeno. Por exemplo, ao transferir o plasmídeo pBH100 de um desses mutantes para uma variante da *E. coli* K-12, e não observar a transferência da marca de resistência à estreptomicina, concluiu que a mesma deveria ser de origem cromossomica. No entanto, também observou que as linhagens da *E. coli* BH100 curadas do plasmídeo pBH100 não davam origem a mutantes de alta frequência resistentes a esse antibiótico, e concluiu que o plasmídeo era essencial para o surgimento da mutação cromossomal. A autora reforçou essa conclusão ao observar o fato de que, quando o plasmídeo pBH100 era reinserido na linhagem curada, era possível verificar o ressurgimento dos mutantes de alta frequência. De Paula, por fim, sugere que o plasmídeo pBH100 contribua de alguma forma para alterações das taxas mutacionais da bactéria, o que induziria à ativação de um ou mais gene(s) cromossômico(s) responsável(is) por conferir resistência à estreptomicina.

Entretanto, baseado nos achados desse trabalho, sugere-se, aqui, uma releitura das observações experimentais obtidas por De Paula (1988). Propõe-se assim, que o marcador de resistência à estreptomicina nos mutantes de alta frequência seja de origem plasmidial e não cromossomico. Nesse caso, supôs-se em um primeiro momento, que alterações nas taxas de expressão de fatores transcricionais, capazes de interferir com a atividade do promotor P_2, seriam responsáveis pelo surgimento desses mutantes. As observações deste trabalho também não corroboram com a hipótese de alterações nas taxas mutacionais. Considerando o uso diferencial das linhagens, aqui trabalhadas, ao longo dos anos e a presença/ausência dos plasmídeos, observa-se que o conteúdo nucleotídico é bastante similar entre as diferentes variantes da *E. coli* BH100, sendo a principal diferença de ordem estrutural e provocada por elementos móveis. Além disso, nenhum dos genes anotados e caracterizados nos plasmídeos do grupo pBH100 parece estar relacionado a algum mecanismo conhecido, que contribuísse para alterações das taxas mutacionais. Portanto, essa hipótese parece não ser adequada para explicar o surgimento diferencial de mutantes resistentes à estreptomicina.

[34] Esse foi o termo adotado por de Paula para se referir aos indivíduos resistentes à estreptomicina. A nomenclatura será aqui preservada por conveniência, ainda que não se tenha conhecimento se o surgimento dessas variantes seja fruto de mutações.

Com efeito, a presença do plasmídeo pBH100 em linhagens transconjugantes deveria manifestar maior resistência à estreptomicina, uma vez que diferentes taxas de expressão de fatores transcricionais do promotor P_2 deveriam ser observadas. A presença de níveis de resistência similares à *E. coli* BH100, obtida em uma variante da *E. coli* K-12, carreadora do plasmídeo pBH100, pode ser devido a dois motivos: 1) Essa linhagem também manifesta fatores de transcrição, cuja expressão mimetiza os efeitos observados na *E. coli* BH100; 2) Genes presentes no próprio plasmídeo pBH100 contribuiriam para supressão do gene *aadA1*. Considerando a primeira hipótese, seria esperado, em algum outro transconjugante, que a expressão do gene *aadA1* levaria a maiores níveis de resistência à estreptomicina. Já no segundo caso, independente da linhagem receptora utilizada, a expressão desse gene sempre seria reduzida, levando a baixos níveis de resistência. A validação dessas hipóteses dependeria, portanto, da avaliação da resistência a esse antibiótico em diferentes transconjugantes, obtidos a partir de diferentes espécies/linhagens receptoras.

Apesar das elucubrações aqui propostas carecerem de ensaios transcriptômicos, acredita-se que o próprio plasmídeo pBH100 seja responsável pela potencial baixa expressão do gene *aadA1* pelo promotor P_2. Isso pois, após anos de trabalho envolvendo esse plasmídeo ainda não foi observada uma linhagem transconjugante com resistência à estreptomicina advinda do plasmídeo pBH100. Além disso, o próprio plasmídeo NR1 confere baixos níveis de resistência a esse antibiótico, mas a clonagem do transposon Tn21 desse plasmídeo em outros replicons parece conferir maiores níveis de resistência à droga. Essas observações corroboram para a presença de algum gene presente no plasmídeo pBH100 capaz de silenciar o promotor P_2. Aqui, sugere-se que esse elemento seja o gene *rpsO* codificante da proteína HHA. Essa proposta surgiu frente os trabalhos de ALBA et al (2011), onde revelaram que essa proteína é capaz de interagir fisicamente com o repressor transcricional H-NS, regulando sua atividade e possibilitando a expressão de genes adquiridos via transferência horizontal. A presença do gene codificante de HHA, próximo à origem de replicação dos plasmídeos NR1 e pBH100 seria então responsável por reduzir a atividade de H-NS e com isso impedir a ativação do promotor P_2. Se essa hipótese for verdadeira dois modelos poderiam explicar o surgimento de mutantes de alta frequência com resistência variável à estreptomicina.

No primeiro modelo, mutações no plasmídeo pBH100 seriam responsáveis por amenizar o efeito de HHA. Nesse caso, a expressão diferencial desses genes, ou mesmo a interferência na afinidade de ligação à H-NS explicaria a observação de níveis distintos de resistência à estreptomicina dos mutantes de alta frequência (**Figura 4.3**).

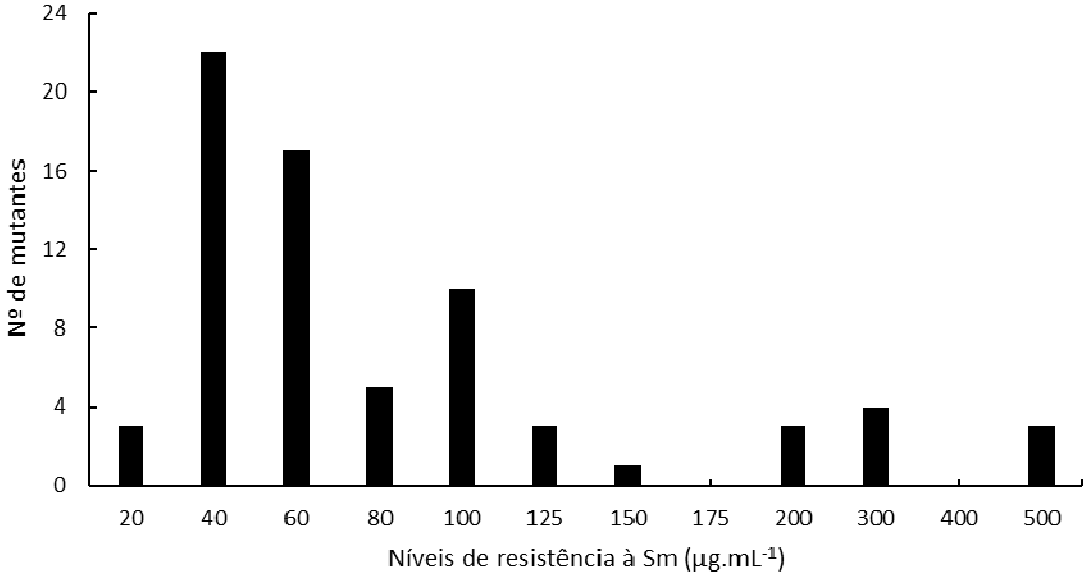


Figura 4.3 Níveis de resistência à estreptomicina de 70 mutantes de alta frequência isolados em ágar nutriente contendo 25 µg.Ml^{-1} de estreptomicina. Adaptado de (DE PAULA, 1989).

Entretanto esse modelo pouco se adequa às observações de instabilidade da resistência à estreptomicina, verificada nesses mutantes de alta frequência. De Paula (1988) ao fazer um estudo populacional desses mutantes constatou que as colônias desses indivíduos poderiam dar origem a colônias com níveis de resistência inferiores ou superiores aos isolados parentais, após passagens sucessivas em meios contendo diferentes concentrações do antibiótico.

Um segundo modelo foi criado de modo a adequar esses efeitos e foi aqui denominado IS-TIP (*Independent Segregation of Tn21 Inserted in pAp*). Nesse modelo, sugere-se que a resistência à estreptomicina, observada nos mutantes de alta frequência, seja devido a um maior número de cópias do gene *aadA1*. De fato, o aumento na dosagem gênica tende a elevar os níveis de resistência a antimicrobianos (UHLIN; NORDSTRÖM, 1977). Em particular, acredita-se que esse efeito seja devido à transferência do Tn21 para o plasmídeo pAp. Essa hipótese surgiu mediante a constatação da presença desse transposon nesse plasmídeo (aqui denominado pApR). Nesse caso, a baixa nos níveis transcricionais do gene *aadA1* seria compensada pelo aumento no número de cópias do Tn21, carreados pelos plasmídeos do grupo pAp. Esse modelo também explicaria os diferentes níveis de resistência observados nos mutantes de alta frequência, além de sua instabilidade, devido à segregação independente dos plasmídeos pAp. Nesse caso, em um evento de divisão celular os plasmídeos pApR, se disporiam de forma independente em cada célula filha. Com efeito, seriam gerados indivíduos na população contendo diferente número de cópias desse plasmídeo, e levando ao surgimento de indivíduos com diferentes níveis de resistência à estreptomicina. A ins-

tabilidade, de igual forma, seria explicada pela perda ou ganho desses plasmídeos após eventos de divisão celular.

Vale destacar que na *E. coli* BH100 subtr. MG2017 um menor número de cópias foi estimado para o plasmídeo pApR. Esse fato também pode indicar que mutações no plasmídeo pAp sejam necessárias para induzir aumento no número de cópias. Além disso, por se tratar de um plasmídeo com origem de replicação supostamente similar à colE1, presume-se que fatores externos possam influenciar sua quantidade. Com efeito, na presença de estreptomicina, indivíduos que estimulem um maior número de cópias desses plasmídeos podem ser selecionados.

Dessa forma, sugere-se aqui que a transferência do Tn21 ao plasmídeo pAp seja a provável explicação para o surgimento dos mutantes de alta frequência com resistência variável e instável à estreptomicina. No entanto, vale salientar que efeitos sinérgicos associados a mutações plasmidiais e/ou cromossomais também possam estar relacionados, podendo contribuir para geração do *fitness* necessário para o surgimento desses variantes. Por fim, o modelo aqui proposto assume que um desbalanço na expressão do gene HHA, de H-NS e/ou no número de cópias do gene *aadA1* associado ao promotor P_2 expliquem o surgimento de indivíduos com resistência variável nas populações de BH100 (**Figura 4.4**).

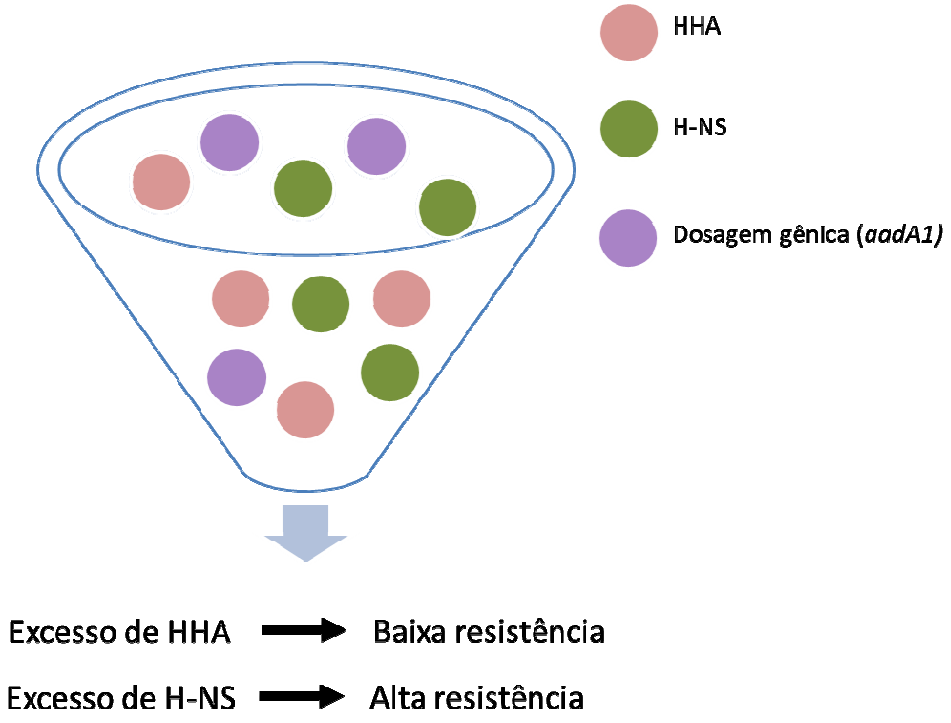


Figura 4.4 Modelo para explicação de diferentes níveis de resistência à estreptomicina como um balanço na atividade das proteínas HHA e H-NS, e da dosagem dos genes *aadA1*. Quando HHA sobrepõe-se à atividade de H-NS verifica-se a diminuição na expressão do gene *aadA1*. Quando H-NS sobrepõe-se à HHA verifica-se aumento na expressão do gene *aadA1*. Nesse modelo, o aumento no número de cópias do gene *aadA1* tende a compensar os efeitos da inibição de HHA em H-NS.

CONCLUSÃO

O sequenciamento, montagem, anotação funcional e as diversas análises comparativas realizadas permitiram corroborar observações experimentais previamente realizadas com a *E. coli* BH100*,* somar novas informações acerca do seu genoma, bem como prever fenótipos ainda não observados.

No capítulo 2, propôs-se discutir a problemática envolvendo a montagem do genoma da *E. coli* BH100 e suas variantes. Os diversos procedimentos de refinamento realizados não permitem afirmar, que o genoma montado não contenha erros, mas possibilita a identificação de regiões que certamente os contêm. Esse fato possibilitou uma melhor e mais acurada montagem capaz de servir de base para anotações mais acuradas e consequentemente análises comparativas mais significativas.

A anotação funcional ainda é tida como um problema em aberto no campo da Bioinformática. Diferentes regras e exceções fazem-se presentes, na tentativa de buscar a uniformização e a reprodutibilidade dessa etapa. Diversas ferramentas de anotação de genomas de procariotos estão disponíveis atualmente, cada qual embasada em estratégias especificas e assumindo determinadas abordagens para lidar com exceções. A comparação do conteúdo gênico entre diferentes organismos, anotados a partir de diferentes softwares, pode induzir a erros significativos na interpretação dos resultados. Neste trabalho, foram utilizadas em conjunto as estratégias de transferência automática de anotações, a partir de uma referência RefSeq, e via anotação na plataforma RAST. Ao longo do trabalho, constatou-se diversos casos de imprecisão quanto a anotação funcional, utilizando-se essas abordagens. O refinamento dos resultados comparativos dependeu da avaliação cuidadosa dos genes anotados, segundo essa estratégia, de modo a tecer conclusões mais acuradas. Exemplo disso foi a análise do *core* de pseudogenes das variantes da *E. coli* BH100. Tal análise foi executada com a finalidade de compreender as diferenças e semelhanças dos potenciais genes não funcionais de cada linhagem e revelou diversas incoerências fruto dos processos de anotação e curadoria realizados.

Já no capítulo 3, propôs-se uma extensiva caracterização do cromossomo e plasmídeos da *E. coli* BH100 e suas variantes. Da análise cromossomica, destacou-se o elevado número de elementos de inserção da família dos IS3, distribuídos de maneira diferencial ao longo dos cromossomos analisados. Tal fato indica a potencial plasticidade do genoma dessa bactéria, uma vez que esses elementos podem interferir nas taxas de expressão gênica, bem como eliminar ou inserir regiões gênicas. Outro aspecto interessante na análise cromossomica foi referente à patogenicidade. Os resultados das análises das ilhas de patogenicidade, filogenômica e da investigação de fatores de

virulência indicam que a *E. coli* BH100 seja uma autentica uropatogênica. Mais do que isso, a semelhança com as *E. coli* 536 e CFT073 ainda sugerem que a mesma seja capaz de provocar infecções renais levando a casos de pielonefrite.

Em relação aos plasmídeos, dois grupos distintos foram observados. No grupo dos plasmídeos pAp estão presentes os plasmídeos que carregam o gene de resistência a β-lactâmicos, o que está de acordo com ensaios experimentais. Esse gene foi verificado no Tn3, também presente em alguns representantes dos plasmídeos do grupo pBH100. Os plasmídeos intitulados pAp_MG2014 e pAp_MG2017 apresentam pequenas modificações de ponto que podem ser devido a mutações espontâneas originadas em suas respectivas células receptoras ou ainda artefatos associados a vieses das plataformas de sequenciamento. Esses plasmídeos parecem estar presentes em um baixo número de cópias dentro da célula, presumivelmente sujeitos a um rigoroso processo de controle, haja vista a homologia de suas *oriV* com a de colE1. Surpreendente, também foi constatada a presença do operon de síntese de colicina E3 nos plasmídeos pAp. Tal fato, não era esperado, uma vez que a *E. coli* BH100 foi classificada como não colicinogênica. Ensaios adicionais, portanto, seriam necessários para verificar a atividade desses genes. Outro aspecto dos plasmídeos desse grupo refere-se à capacidade de mobilização. Aqui, propõe-se que esse mecanismo ocorra mediante a ação da proteína Mob, sendo o processo como um todo dependente da maquinaria de formação do poro conectivo entre as células doadora e receptora, estimulados pelo plasmídeo autoconjugativo

No tocante aos plasmídeos do grupo pBH100, foram constatados todas os marcadores de resistência experimentalmente verificados. Constatou-se a presença desses elementos em transposons, o que explica os diversos desvios fenotípicos observados em linhagens transconjugantes, contendo esses plasmídeos. Interessante foi observar o empacotamento desses elementos móveis. Por exemplo no Tn9, carreador de um gene de resistência ao cloranfenicol, verifica-se a inserção do Tn21, carreador do operon *mer* e do integron In2, cujos cassetes compõem genes de resistência à estreptomicina, sulfonamida e a compostos de amônia quaternários, comumente presentes em sanitizantes. Outro caso é o do Tn10, que carrega consigo o operon *tet*, responsável por conferir resistência à tetraciclina e que apresenta em sua composição um potencial novo transposon, carreador do gene que confere resistência à canamicina e flanqueado por elementos de inserção do tipo IS5. A sobreposição desses elementos corrobora com as observações de geração de variabilidade e disseminação de multirresistência bacteriana. Os plasmídeos desse grupo diferem entre si basicamente na sua composição de elementos móveis. O caso mais notável refere-se aos plasmídeos

intitulados pBH100-1 e pBH100alpha, cuja diferença principal consiste na presença do Tn3, do plasmídeo pAp, no primeiro, e sua ausência no segundo.

A presença de uma variante do plasmídeo pAp, aqui denominada pApR, permitiu a elaboração de um modelo capaz de explicar a geração de indivíduos com resistência variável e instável à estreptomicina. Esse plasmídeo, identificado na *E. coli* BH100 subtr. MG2017, contém o Tn21, que carrega consigo o gene *aadA1,* capaz de conferir resistência à estreptomicina. Aqui, propõe-se que esse gene possua naturalmente expressão reduzida na *E. coli* BH100, como consequência do gene *rmoA*, codificante da proteína HHA, capaz de interagir com o regulador transcricional H-NS. Nesse caso, a repressão da atividade de H-NS, um regulador positivo do promotor P_2, presente no In2, seria responsável pelos baixos índices de resistência à estreptomicina na *E. coli* BH100 e nos transconjugantes contendo o plasmídeo pBH100. O surgimento de variantes resistentes a esse antibiótico seria, portanto, devido a um aumento no número de copias do Tn21, mediado por sua presença no plasmídeo pAp. Esse modelo foi denominado ISTIP (*Independnet Segregation of Tn21 inserted in pAp*) e explica o surgimento de indivíduos na população contendo resistência variável e instável à estreptomicina.

Por fim, os trabalhos conduzidos com essa bactéria sustentam o fato de como microrganismos podem gerar variabilidade para adaptação a diferentes nichos ecológicos. Em especial, tal variabilidade faz-se um problema grave do ponto de vista clínico, devido ao aumento no *fitness* e principalmente no surgimento de variantes resistentes a antimicrobianos. Plasmídeos de multirresistência, presentes na microbiota normal, por exemplo, podem mais tarde surgir em variantes patogênicas. Uma das contramedidas possíveis para amenizar tal efeito é através da redução da pressão seletiva. Tal ação pode e deve ser estimulada por médicos mediante a prescrição de antibióticos apenas quando houver certeza de sua necessidade, ou sempre que possível optar pela escolha de medicamentos de espectro mais curto. O uso excessivo de antibióticos, aliados à habilidade desses microrganismos em desenvolver e disseminar resistência, representa um sério problema de saúde pública e que merece a devida atenção.

PERSPECTIVAS

Os diversos achados apresentados neste trabalho fazem da *E. coli* BH100 um modelo interessante para investigações envolvendo o surgimento de variabilidade genética e fenotípica em populações bacterianas, estocadas em laboratório. Diversas propostas aqui elaboradas podem ser validadas ou invalidas mediante ensaios microbiológicos, transcriptômicos e/ou proteômicos. Por exemplo, a investigação da síntese de colicina E3 pela BH100 pode ser investigada, mediante a observação da taxa de expressão dos transcritos desse gene, via PCR em tempo real (qPCR). Outro uso dessa técnica seria a validação da hipótese ISTIP, mediante a checagem do número de cópias de cada plasmídeo e da presença do Tn21 inserido no plasmídeo pAp.

Ensaios transcriptômicos, também seriam relevantes para compreensão sistêmica da expressão de genes associados aos fatores de virulência presentes na *E. coli* BH100. Essa abordagem, correlacionada ao uso dessa bactéria em modelos de infecção animal, poderia servir de grande valia para compreensão dos mecanismos de geração da infecção do trato urinário, bem como dos efeitos que elementos móveis possam provocar na geração e desenvolvimento dessa enfermidade.

REFERÊNCIAS

AGREN, J. et al. Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. **PloS one**, v. 7, n. 6, p. e39107, 2012.

AGUERO, M. E. et al. ColV increases the virulence of Escherichia coli K1 strains in animal models of neonatal meningitis and urinary infection. **Medical microbiology and immunology**, v. 178, n. 4, p. 211–6, 1989.

AHSAN, S.; KABIR, M. S. Linear plasmids and their replication. **Stamford Journal of Microbiology**, v. 2, n. 1, p. 1–5, 4 jun. 2013.

ALBA, C. F. DE et al. Essential residues in the H-NS binding site of Hha, a co-regulator of horizontally acquired genes in Enterobacteria. **FEBS Letters**, v. 585, n. 12, p. 1765–1770, 23 jun. 2011.

ALIKHAN, N.-F. et al. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. **BMC genomics**, v. 12, p. 402, 8 ago. 2011.

ALLHOFF, M. et al. Discovering motifs that induce sequencing errors. **BMC bioinformatics**, v. 14 Suppl 5, n. Suppl 5, p. S1, 2013.

ALTHORPE, N. J. et al. Transient transcriptional activation of the IncI1 plasmid anti-restriction gene (ardA) and SOS inhibition gene (psiB) early in conjugating recipient bacteria. **Molecular Microbiology**, v. 31, n. 1, p. 133–142, 1 jan. 1999.

ALTON, N. K.; VAPNEK, D. Nucleotide sequence analysis of the chloramphenicol resistance transposon Tn9. **Nature**, v. 282, n. 5741, p. 864–9, 1979.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403–410, 5 out. 1990.

ANTIPOV, D. et al. plasmidSPAdes: assembling plasmids from whole genome sequencing data. **Bioinformatics**, v. 32, n. 22, p. btw493, 27 jul. 2016.

ARAUJO, F. A. et al. GO FEAT: a rapid web-based functional annotation tool for genomic and transcriptomic data. **Scientific Reports**, v. 8, n. 1, p. 1794, 29 dez. 2018.

ARNDT, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. **Nucleic**

acids research, v. 44, n. W1, p. W16-21, 8 jul. 2016.

AVERY, O. T.; MACLEOD, C. M.; MCCARTY, M. Studies On The Chemical Nature Of The Substance Inducing Transformation Of Pneumococcal Types : Induction Of Transformation By A Desoxyribonucleic Acid Fraction Isolated From Pneumococcus Type III. **The Journal of experimental medicine**, v. 79, n. 2, p. 137–58, 1 fev. 1944.

BACHMANN, B. J. Pedigrees of some mutant strains of Escherichia coli K-12. **Bacteriological reviews**, v. 36, n. 4, p. 525–57, dez. 1972.

BAESHEN, N. A. et al. Cell factories for insulin production. **Microbial cell factories**, v. 13, p. 141, 2 out. 2014.

BAILONE, A. et al. PsiB polypeptide prevents activation of RecA protein in Escherichia coli. **MGG Molecular & General Genetics**, v. 214, n. 3, p. 389–395, nov. 1988.

BANKEVICH, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. **Journal of computational biology : a journal of computational molecular cell biology**, v. 19, n. 5, p. 455–77, maio 2012.

BEIER, D.; GROSS, R. Regulation of bacterial virulence by two-component systems. **Current Opinion in Microbiology**, v. 9, n. 2, p. 143–152, 1 abr. 2006.

BERTELLI, C. et al. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. **Nucleic Acids Research**, v. 45, n. W1, p. W30–W35, 3 jul. 2017.

BINGEN-BIDOIS, M. et al. Phylogenetic analysis and prevalence of urosepsis strains of Escherichia coli bearing pathogenicity island-like domains. **Infection and immunity**, v. 70, n. 6, p. 3216–26, jun. 2002.

BLATTNER, F. R. et al. The Complete Genome Sequence of Escherichia coli K-12. **Science**, v. 277, n. 5331, p. 1453–1462, 1997.

BOE, L. et al. The frequency of mutators in populations of Escherichia coli. **Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis**, v. 448, n. 1, p. 47–55, 14 mar. 2000.

BOETZER, M. et al. Scaffolding pre-assembled contigs using SSPACE. **Bioinformatics**, v. 27, n. 4, p. 578–579, 15 fev. 2011.

BOLDRIN, F. et al. Promoter mutagenesis for fine-tuning expression of essential genes in

Mycobacterium tuberculosis. **Microbial Biotechnology**, v. 11, n. 1, p. 238–247, 1 jan. 2018.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 1 ago. 2014.

BORONIN, A. M. et al. Plasmids specifying Îμ-caprolactam degradation in *Pseudomonas* strains. **FEMS Microbiology Letters**, v. 22, n. 3, p. 167–170, 1 maio 1984.

BOUCHOT, J.-L. et al. Advances in Machine Learning for Processing and Comparison of Metagenomic Data. In: **Computational Systems Biology**. [s.l.] Elsevier, 2014. p. 295–329.

BOZCAL, E. Distribution and virulence properties of extra-intestinal pathogenic Escherichia coli in Turkey. **Microbiologia Medica**, v. 31, n. 4, 22 dez. 2016.

BRADNAM, K. R. et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. **GigaScience**, v. 2, n. 1, p. 10, 22 jul. 2013.

BRAUS, G.; ARGAST, M.; BECK, C. F. Identification of Additional Genes on Transposon TnlO: tetC and tetD. **JOURNAL OF BACTERIOLOGY**, v. 160, n. 2, p. 504–509, 1984.

BRINSMADE, S. R.; PALDON, T.; ESCALANTE-SEMERENA, J. C. Minimal functions and physiological conditions required for growth of salmonella enterica on ethanolamine in the absence of the metabolosome. **Journal of bacteriology**, v. 187, n. 23, p. 8039–46, 1 dez. 2005.

BRÖCKER, M. J. et al. Recoding the Genetic Code with Selenocysteine. **Angewandte Chemie (International ed. in English)**, v. 53, n. 1, p. 319, 2014.

BUDIČ, M. et al. Escherichia coli bacteriocins: antimicrobial efficacy and prevalence among isolates from patients with bacteraemia. **PloS one**, v. 6, n. 12, p. e28769, 2011.

BUERMANS, H. P. J.; DEN DUNNEN, J. T. Next generation sequencing technology: Advances and applications. **Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease**, v. 1842, n. 10, p. 1932–1941, 1 out. 2014.

CAGLE, C. A.; SHEARER, J. E. S.; SUMMERS, A. O. Regulation of the integrase and cassette promoters of the class 1 integron by nucleoid-associated proteins. **Microbiology**, v. 157, n. 10, p. 2841–2853, 1 out. 2011.

CAMPBELL, P. J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. **Nature Genetics**, v. 40, n. 6, p. 722–729, 27 jun. 2008.

CAO, Y. et al. A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies. **Frontiers in microbiology**, v. 8, p. 1829, 2017.

CARATTOLI, A. et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. **Antimicrobial agents and chemotherapy**, v. 58, n. 7, p. 3895–903, 28 jul. 2014.

CASTAGNOLI, L. et al. Genetic and structural analysis of the ColE1 Rop (Rom) protein. **The EMBO journal**, v. 8, n. 2, p. 621–9, fev. 1989.

CAUGANT, D. A. et al. Intercontinental spread of a genetically distinctive complex of clones of Neisseria meningitidis causing epidemic disease. **Proceedings of the National Academy of Sciences**, v. 83, n. 13, 1986.

CESARENI, G.; MUESING, M. A.; POLISKY, B. Control of ColE1 DNA replication: the rop gene product negatively affects transcription from the replication primer promoter. **Proceedings of the National Academy of Sciences of the United States of America**, v. 79, n. 20, p. 6313–7, out. 1982.

CHAI, Y. et al. Galactose metabolism plays a crucial role in biofilm formation by Bacillus subtilis. **mBio**, v. 3, n. 4, p. e00184-12, 31 ago. 2012.

CHAKRABARTY, A. M.; CHOU, G.; GUNSALUS, I. C. Genetic regulation of octane dissimilation plasmid in Pseudomonas. **Proceedings of the National Academy of Sciences of the United States of America**, v. 70, n. 4, p. 1137–40, abr. 1973.

CHALMERS, R. et al. Complete nucleotide sequence of Tn10. **Journal of bacteriology**, v. 182, n. 10, p. 2970–2, 15 maio 2000.

CHARLIER, D.; PIETTE, J.; GLANSDORFF, N. IS3 can function as a mobile promoter in E. coli. **Nucleic acids research**, v. 10, n. 19, p. 5935–48, 11 out. 1982.

CHARTONE-SOUZA, E. **Resistência a drogas e propriedade colicinogênica em Escherichia coli**. [s.l.] Universidade Federal de Minas Gerais, 1975.

CHAUDHARI, N. M.; GUPTA, V. K.; DUTTA, C. BPGA- an ultra-fast pan-genome analysis pipeline. **Scientific Reports**, v. 6, n. 1, p. 24373, 13 jul. 2016.

CHIBANI-CHENNOUFI, S. et al. Phage-host interaction: an ecological perspective. **Journal of bacteriology**, v. 186, n. 12, p. 3677–86, jun. 2004.

CHIKHI, R.; MEDVEDEV, P. Informed and automated k-mer size selection for genome assembly. **Bioinformatics**, v. 30, n. 1, p. 31–37, 1 jan. 2014.

CLARK, A. J.; ADELBERG, E. A. Bacterial Conjugation. **Annual Review of Microbiology**, v. 16, n. 1, p. 289–319, 28 out. 1962.

CLARK, D. P. et al. Protein Synthesis. In: **Molecular Biology**. [s.l.] Elsevier, 2013. p. 369–416.

CLEWELL, D. B. Antibiotic Resistance Plasmids in Bacteria. In: **eLS**. Chichester, UK: John Wiley & Sons, Ltd, 2014.

CLOKIE, M. R. et al. Phages in nature. **Bacteriophage**, v. 1, n. 1, p. 31–45, jan. 2011.

COLLIS, C. M.; HALL, R. M. Expression of antibiotic resistance genes in the integrated cassettes of integrons. **Antimicrobial agents and chemotherapy**, v. 39, n. 1, p. 155–62, jan. 1995.

CONLAN, S. et al. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. **Science translational medicine**, v. 6, n. 254, p. 254ra126, 17 set. 2014.

CONTE, M. P. et al. The Adherent/Invasive Escherichia coli Strain LF82 Invades and Persists in Human Prostate Cell Line RWPE-1, Activating a Strong Inflammatory Response. **Infection and immunity**, v. 84, n. 11, p. 3105–3113, nov. 2016.

COX, E. C. Bacterial Mutator Genes and the Control of Spontaneous Mutation. **Annual Review of Genetics**, v. 10, n. 1, p. 135–156, dez. 1976.

COX, K. E. L.; SCHILDBACH, J. F. Sequence of the R1 plasmid and comparison to F and R100. **Plasmid**, v. 91, p. 53–60, maio 2017.

DAEGELEN, P. et al. Tracing Ancestors and Relatives of Escherichia coli B, and the Derivation of B Strains REL606 and BL21(DE3). **Journal of Molecular Biology**, v. 394, n. 4, p. 634–643, 11 dez. 2009.

DARLING, A. C. E. et al. Mauve: multiple alignment of conserved genomic sequence with rearrangements. **Genome research**, v. 14, n. 7, p. 1394–403, 1 jul. 2004.

DATTA, N. et al. Plasmids and transposons and their stability and mutability in bacteria isolated during an outbreak of hospital infection. **Plasmid**, v. 2, n. 2, p. 182–196, 1 abr. 1979.

DAVIS, J. M.; RASMUSSEN, S. B.; O'BRIEN, A. D. Cytotoxic necrotizing factor type 1 production by uropathogenic Escherichia coli modulates polymorphonuclear leukocyte function. **Infection and immunity**, v. 73, n. 9, p. 5301–10, set. 2005.

DE LACERDA, P. V. **Detecção de B-lactamase liberada para o meio sólido por Escherichia coli BH100 selvagem e outras linhagens hospedeiras do plasmídio pAp**. [s.l.] Universidade Federal de Minas Gerais, 1995.

DE PAULA, M. O. **Associação entre mutação cromossômica para resistência à estreptomicina e plasmídio R em Escherichia coli BH100**. [s.l.] Universidade Federal de Minas Gerais, 1989.

DE WIT, R.; BOUVIER, T. "Everything is everywhere, but, the environment selects"; what did Baas Becking and Beijerinck really say? **Environmental Microbiology**, v. 8, n. 4, p. 755–758, 1 abr. 2006.

DEICHELBOHRER, I. et al. Plasmid transduction by Bacillus subtilis bacteriophage SPP1: effects of DNA homology between plasmid and bacteriophage. **Journal of bacteriology**, v. 162, n. 3, p. 1238–43, jun. 1985.

DEMPSEY, W. B. Derepression of conjugal transfer of the antibiotic resistance plasmid R100 by antisense RNA. **Journal of bacteriology**, v. 171, n. 5, p. 2886–8, maio 1989.

DENAMUR, E.; MATIC, I. Evolution of mutation rates in bacteria. **Molecular Microbiology**, v. 60, n. 4, p. 820–827, 1 maio 2006.

DENG, Y. et al. Resistance integrons: class 1, 2 and 3 integrons. **Annals of clinical microbiology and antimicrobials**, v. 14, p. 45, 20 out. 2015.

DHO-MOULIN, M.; FAIRBROTHER, J. M. Avian pathogenic Escherichia coli (APEC). **Veterinary research**, v. 30, n. 2–3, p. 299–316, 1999.

DINMAN, J. D. Programmed Ribosomal Frameshifting Goes Beyond Viruses: Organisms from all three kingdoms use frameshifting to regulate gene expression, perhaps signaling a paradigm shift. **Microbe (Washington, D.C.)**, v. 1, n. 11, p. 521–527, nov. 2006.

DONGT, X. et al. DNA Bending Near the Replication Origin of IncFII Plasmid NR1. **JOURNAL OF BACTERIOLOGY**, v. 171, n. 2, p. 703–707, 1989.

DOUGAN, G. et al. A Functional Map of Plasmid ColE1. **Molec. gen. Genet**, v. 158, p. 325–327, 1978.

ECOCYC. **Escherichia coli K-12 substr. MG1655 Compound: 5S ribosomal RNA**. Disponível em: <https://biocyc.org/ECOLI/NEW-IMAGE?type=POLYPEPTIDE&object=RRFA-RRNA>. Acesso em: 1 jan. 2018.

ECSEQ. **What is mate pair sequencing for?** Disponível em: <https://www.ecseq.com/support/ngs/what-is-mate-pair-sequencing-useful-for>. Acesso em: 28 fev. 2018.

EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. **Bioinformatics**, v. 26, n. 19, p. 2460–2461, 1 out. 2010.

ESCHERICH, T. Die darmbakterien des neugeborenen und säuglings. **Fortschr Med**, v. 3, p. 515–522, 1885.

FALCONI, M. et al. Antagonistic involvement of FIS and H-NS proteins in the transcriptional control of hns expression. **Molecular Microbiology**, v. 19, n. 5, p. 965–975, 1 mar. 1996.

FANG, H. et al. Reducing INDEL calling errors in whole genome and exome sequencing data. **Genome Medicine**, v. 6, n. 10, p. 89, 28 dez. 2014.

FLORES-MIRELES, A. L. et al. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. **Nature Reviews Microbiology**, v. 13, n. 5, p. 269–284, 8 maio 2015.

FLORES, C. O. et al. Statistical structure of host-phage interactions. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 28, p. E288-97, 12 jul. 2011.

FOXMAN, B. Urinary Tract Infection Syndromes. **Infectious Disease Clinics of North America**, v. 28, n. 1, p. 1–13, mar. 2014.

FRIEHS, K. Plasmid Copy Number and Plasmid Stability. In: [s.l.] Springer, Berlin, Heidelberg, 2004. p. 47–82.

GALAGAN, J. E. et al. Genomics of the fungal kingdom: Insights into eukaryotic biology. **Genome Research**, v. 15, n. 12, p. 1620–1631, 1 dez. 2005.

GALARDINI, M. et al. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. **Source Code for Biology and Medicine**, v. 6, n. 1, p. 11, 21 jun. 2011.

GARCILLÁN-BARCIA, M. P.; DE LA CRUZ, F. Why is entry exclusion an essential feature of conjugative plasmids? **Plasmid**, v. 60, n. 1, p. 1–18, jul. 2008.

GAWAD, C.; KOH, W.; QUAKE, S. R. Single-cell genome sequencing: current state of the science. **Nature Reviews Genetics**, v. 17, n. 3, p. 175–188, 25 mar. 2016.

Genomes - Genome - NCBI. Disponível em: <https://www.ncbi.nlm.nih.gov/genome/genomes/167>. Acesso em: 25 fev. 2018.

GERDES, K.; THISTED, T.; MARTINUSSEN, J. Mechanism of post-segregational killing by the hoklsok system of plasmid R1: sok antisense RNA regulates formation of a hok mRNA species correlated with killing of plasmid-free cells. **Molecular Microbiology**, v. 4, n. 11, p. 1807–1818, 1 nov. 1990.

GIANGROSSI, M.; GUALERZI, C. O.; PON, C. L. Mutagenesis of the downstream region of the Escherichia coli hns promoter. **Biochimie**, v. 83, n. 2, p. 251–259, 1 fev. 2001.

GILLINGS, M. R. Integrons: past, present, and future. **Microbiology and molecular biology reviews : MMBR**, v. 78, n. 2, p. 257–77, jun. 2014.

GILLOR, O.; VRIEZEN, J. A. C.; RILEY, M. A. The role of SOS boxes in enteric bacteriocin regulation. **Microbiology (Reading, England)**, v. 154, n. Pt 6, p. 1783–92, jun. 2008.

GOREN, M. G. et al. Transfer of carbapenem-resistant plasmid from Klebsiella pneumoniae ST258 to Escherichia coli in patient. **Emerging infectious diseases**, v. 16, n. 6, p. 1014–7, jun. 2010.

GRAINGER, D. C. Structure and function of bacterial H-NS protein. **Biochemical Society transactions**, v. 44, n. 6, p. 1561–1569, 15 dez. 2016.

GRANT, M. A.; WEAGANT, S. D.; FENG, P. Glutamate decarboxylase genes as a prescreening marker for detection of pathogenic Escherichia coli groups. **Applied and environmental microbiology**, v. 67, n. 7, p. 3110–4, jul. 2001.

GRIFFITH, F. The Significance of Pneumococcal Types. **The Journal of hygiene**, v. 27, n. 2, p. 113–59, jan. 1928.

GRIFFITH, K. L.; BECKER, S. M.; E. WOLF, R. Characterization of TetD as a transcriptional activator of a subset of genes of the Escherichia coli SoxS/MarA/Rob regulon. **Molecular Microbiology**, v. 56, n. 4, p. 1103–1117, 24 mar. 2005.

GRIFFITHS, A. J. F. **Modern genetic analysis**. [s.l.] W.H. Freeman, 1999.

GROSSMAN, A. R. In the Grip of Algal Genomics. In: **Transgenic Microalgae as Green Cell**

Factories. New York, NY: Springer New York, 2007. v. 616p. 54–76.

GUIMARÃES, L. et al. Genome informatics and vaccine targets in Corynebacterium urealyticum using two whole genomes, comparative genomics, and reverse vaccinology. **BMC Genomics**, v. 16, n. Suppl 5, p. S7, 26 maio 2015.

GUIMARÃES, L. M. V. **Resistência cromossômica e plasmidial ao mercúrio em Escherichia coli**. [s.l.] Universidade Federal de Minas Gerais, 1986.

GUIZELINI, D. et al. GFinisher: a new strategy to refine and finish bacterial genome assemblies. **Scientific Reports**, v. 6, n. 1, p. 34963, 10 dez. 2016.

GUREVICH, A. et al. QUAST: quality assessment tool for genome assemblies. **Bioinformatics (Oxford, England)**, v. 29, n. 8, p. 1072–5, 15 abr. 2013.

GUTTMAN, B. S. Lysogeny. In: **Encyclopedia of Genetics**. [s.l.] Elsevier, 2001. p. 1130–1131.

HANDSAKER, R. E. et al. Large multiallelic copy number variations in humans. **Nature genetics**, v. 47, n. 3, p. 296–303, 26 mar. 2015.

HARDIMAN, C. A. et al. Horizontal Transfer of Carbapenemase-Encoding Plasmids and Comparison with Hospital Epidemiology Data. **Antimicrobial agents and chemotherapy**, v. 60, n. 8, p. 4910–9, 1 ago. 2016.

HEDGES, R. W.; JACOB, A. E. Transposition of ampicillin resistance from RP4 to other replicons. **MGG Molecular & General Genetics**, v. 132, n. 1, p. 31–40, 1974.

HERNANDEZ, D. et al. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. **Genome research**, v. 18, n. 5, p. 802–9, maio 2008.

HILLEN, W.; BERENS, C. Tetracyclin-gesteuerte Genregulation: Vom bakteriellen Ursprung zum eukaryotischen Werkzeug. **BIOspektrum ·**, v. 4, p. 355–358, 2002.

HINNEBUSCH, J.; TILLY, K. Linear plasmids and chromosomes in bacteria. **Molecular microbiology**, v. 10, n. 5, p. 917–22, dez. 1993.

HO SUI, S. J. et al. The Association of Virulence Factors with Genomic Islands. **PLoS ONE**, v. 4, n. 12, p. e8094, 1 dez. 2009.

HOFFMANN, T. et al. Synthesis of the compatible solute proline by *Bacillus subtilis* : point

mutations rendering the osmotically controlled *proHJ* promoter hyperactive. **Environmental Microbiology**, v. 19, n. 9, p. 3700–3720, 1 set. 2017.

HOLT, K. E. et al. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. **Nature Genetics**, v. 44, n. 9, p. 1056–1059, 5 set. 2012.

HOOPER, D. C.; JACOBY, G. A. Mechanisms of drug resistance: quinolone resistance. **Annals of the New York Academy of Sciences**, v. 1354, n. 1, p. 12–31, set. 2015.

HSIAO, W. W. L. et al. Evidence of a large novel gene pool associated with prokaryotic genomic islands. **PLoS genetics**, v. 1, n. 5, p. e62, nov. 2005.

HU, M.; DEONIER, R. C. Comparison of IS1, IS2 and IS3 copy number in Escherichia coli strains K-12, B and C. **Gene**, v. 16, n. 1–3, p. 161–170, 1 dez. 1981.

HUSON, D. H.; BRYANT, D. Application of Phylogenetic Networks in Evolutionary Studies. **Molecular Biology and Evolution**, v. 23, n. 2, p. 254–267, 1 fev. 2006.

IDALIA, V.-M. N.; BERNARDO, F. Escherichia coli as a Model Organism and Its Application in Biotechnology. In: **Escherichia coli - Recent Advances on Physiology, Pathogenesis and Biotechnological Applications**. [s.l.] InTech, 2017.

IHSSEN, J. et al. Production of glycoprotein vaccines in Escherichia coli. **Microbial cell factories**, v. 9, p. 61, 11 ago. 2010.

ISHIDA, T. Effects of Point Mutation on Enzymatic Activity: Correlation between Protein Electronic Structure and Motion in Chorismate Mutase Reaction. **Journal of the American Chemical Society**, v. 132, n. 20, p. 7104–7118, 26 maio 2010.

JAKES, K. S.; MODEL, P. Mechanism of Export of Colicin El and Colicin E3. **JouRNAL OF BACTZRIOLOGY**, v. 138, n. 3, p. 770–778, 1979.

JENSEN, R. B.; GERDES, K. Programmed cell death in bacteria: proteic plasmid stabilization systems. **Molecular Microbiology**, v. 17, n. 2, p. 205–210, 1 jul. 1995.

JEZIOROWSKI, A.; GORDON, D. M. Evolution of microcin V and colicin Ia plasmids in Escherichia coli. **Journal of bacteriology**, v. 189, n. 19, p. 7045–52, 1 out. 2007.

JIANG, T. et al. Insertion and deletion mutations in the repA4 region of the IncFII plasmid NR1 cause unstable inheritance. **Journal of bacteriology**, v. 175, n. 17, p. 5350–8, 1 set. 1993.

JIANG, Y. et al. ParE toxin encoded by the broad-host-range plasmid RK2 is an inhibitor of Escherichia coli gyrase. **Molecular microbiology**, v. 44, n. 4, p. 971–9, maio 2002.

JOENSEN, K. G. et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. **Journal of clinical microbiology**, v. 52, n. 5, p. 1501–10, maio 2014.

JOENSEN, K. G. et al. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. **Journal of clinical microbiology**, v. 53, n. 8, p. 2410–26, 1 ago. 2015.

JOHNSON, T. J.; NOLAN, L. K. Pathogenomics of the virulence plasmids of Escherichia coli. **Microbiology and molecular biology reviews : MMBR**, v. 73, n. 4, p. 750–74, dez. 2009.

JOHNSON, T. J.; WANNEMUEHLER, Y. M.; NOLAN, L. K. Evolution of the iss gene in Escherichia coli. **Applied and environmental microbiology**, v. 74, n. 8, p. 2360–9, 15 abr. 2008.

JUHAS, M. et al. Genomic islands: tools of bacterial horizontal gene transfer and evolution. **FEMS microbiology reviews**, v. 33, n. 2, p. 376–93, mar. 2009.

KAZAMA, H. et al. Characterization of the antiseptic-resistance gene *qacE* Î” *1* isolated from clinical and environmental isolates of *Vibrio parahaemolyticus* and *Vibrio cholerae* non-O1. **FEMS Microbiology Letters**, v. 174, n. 2, p. 379–384, 1 maio 1999.

KHETRAPAL, V.; MEHERSHAHI, K. S.; CHEN, S. L. Complete Genome Sequence of the OriginalEscherichia coliIsolate, Strain NCTC86. **Genome announcements**, v. 5, n. 16, 20 abr. 2017.

KIM, S. K.; WANG, J. C. Gene silencing via protein-mediated subcellular localization of DNA. **Proceedings of the National Academy of Sciences of the United States of America**, v. 96, n. 15, p. 8557–61, 20 jul. 1999.

KOFOID, E. et al. The 17-gene ethanolamine (eut) operon of Salmonella typhimurium encodes five homologues of carboxysome shell proteins. **Journal of bacteriology**, v. 181, n. 17, p. 5317–29, set. 1999.

KÖHLER, C.-D.; DOBRINDT, U. What defines extraintestinal pathogenic Escherichia coli? **International Journal of Medical Microbiology**, v. 301, n. 8, p. 642–647, dez. 2011.

KOHN, A. B. et al. Single-Cell Semiconductor Sequencing. In: **Methods in molecular biology (Clifton, N.J.)**. [s.l: s.n.]. v. 1048p. 247–284.

KOLI, P. et al. Conversion of commensal Escherichia coli K-12 to an invasive form via expression of a mutant histone-like protein. **mBio**, v. 2, n. 5, p. e00182-11, 1 nov. 2011.

KOREN, S. et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. **Genome Biology**, v. 14, n. 9, p. R101, 13 set. 2013.

KU, L. C.; BOGGESS, K. A.; COHEN-WOLKOWIEZ, M. Bacterial meningitis in infants. **Clinics in perinatology**, v. 42, n. 1, p. 29–45, vii–viii, mar. 2015.

KULSKI, J. K. Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. In: **Next Generation Sequencing - Advances, Applications and Challenges**. [s.l.] InTech, 2016.

KUSKA, B. Beer, Bethesda, and Biology: How "Genomics" Came Into Being. **JNCI: Journal of the National Cancer Institute**, v. 90, n. 2, p. 93–93, 21 jan. 1998.

KUTTER, E.; GUTTMAN, B. T Phages. In: **Encyclopedia of Genetics**. [s.l.] Elsevier, 2001. p. 1921–1930.

LAPAGE, S. et al. **International Code of Nomenclature of Bacteria**. [s.l.] ASM Press, 1992.

LAVIÑA, M.; GAGGERO, C.; MORENO, F. Microcin H47, a chromosome-encoded microcin antibiotic of Escherichia coli. **Journal of bacteriology**, v. 172, n. 11, p. 6585–8, 1 nov. 1990.

LEDERGERBER, C.; DESSIMOZ, C. Base-calling for next-generation sequencing platforms. **Briefings in bioinformatics**, v. 12, n. 5, p. 489–97, set. 2011.

LEE, H. et al. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. **Proceedings of the National Academy of Sciences**, v. 109, n. 41, p. E2774–E2783, 9 out. 2012.

LENGELER, J. W.; DREWS, G. GERHART; SCHLEGEL, H. G. **Biology of the prokaryotes**. [s.l.] Thieme, 1999.

LEVY, S. E.; MYERS, R. M. Advancements in Next-Generation Sequencing. **Annual Review of Genomics and Human Genetics**, v. 17, n. 1, p. 95–115, 31 ago. 2016.

LI, Z. et al. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. **Briefings in Functional Genomics**, v. 11, n. 1, p. 25–37, 1 jan. 2012.

LIEBERT, C. A.; HALL, R. M.; SUMMERS, A. O. Transposon Tn21, flagship of the floating genome. **Microbiology and molecular biology reviews : MMBR**, v. 63, n. 3, p. 507–22, set. 1999a.

LIEBERT, C. A.; HALL, R. M.; SUMMERS, A. O. Transposon Tn21, flagship of the floating genome. **Microbiology and molecular biology reviews : MMBR**, v. 63, n. 3, p. 507–22, 1 set. 1999b.

LIEBERT, C. A.; HALL, R. M.; SUMMERS, A. O. Transposon Tn21, flagship of the floating genome. **Microbiology and molecular biology reviews : MMBR**, v. 63, n. 3, p. 507–22, 1 set. 1999c.

LIN-CHAO, S.; BREMER, H. Effect of the bacterial growth rate on replication control of plasmid pBR322 in Escherichia coli. **Molecular & general genetics : MGG**, v. 203, n. 1, p. 143–9, abr. 1986.

LIN, J. et al. Comparative analysis of extreme acid survival in Salmonella typhimurium, Shigella flexneri, and Escherichia coli. **Journal of bacteriology**, v. 177, n. 14, p. 4097–104, jul. 1995.

LISCHER, H. E. L.; SHIMIZU, K. K. Reference-guided de novo assembly approach improves genome reconstruction for related species. **BMC bioinformatics**, v. 18, n. 1, p. 474, 10 nov. 2017.

LIU, C. P.; CHURCHWARD, G.; CARO, L. The repA2 gene of the plasmid R100.1 encodes a repressor of plasmid replication. **Plasmid**, v. 10, n. 2, p. 148–55, set. 1983.

LOECHLER, E. L.; GREEN, C. L.; ESSIGMANN, J. M. In vivo mutagenesis by O6-methylguanine built into a unique site in a viral genome. **Proceedings of the National Academy of Sciences of the United States of America**, v. 81, n. 20, p. 6271–5, out. 1984.

LOPES, T. S. et al. Factors affecting the mitotic stability of high-copy-number integration into the ribosomal DNA of Saccharomyces cerevisiae. **Yeast**, v. 12, n. 5, p. 467–477, 1 abr. 1996.

LU, Y. et al. Next Generation Sequencing in Aquatic Models. In: **Next Generation Sequencing - Advances, Applications and Challenges**. [s.l.] InTech, 2016.

LU, K. V et al. Mistranslation of a TGA termination codon as tryptophan in recombinant platelet-derived growth factor expressed in Escherichia coli. **The Biochemical journal**, v. 309 (Pt 2), n. Pt 2, p. 411–7, 15 jul. 1995.

LYNNE, A. M. et al. Characterization of a series of transconjugant mutants of an avian pathogenic Escherichia coli isolate for resistance to serum complement. **Avian diseases**, v. 51, n. 3, p. 771–6, set. 2007.

MACHIDA, Y. et al. Factors determining frequency of plasmid cointegration mediated by insertion sequence IS1. **Proceedings of the National Academy of Sciences of the United States of America**, v. 79, n. 2, p. 277–81, jan. 1982.

MADRID, C.; NIETO, J. M.; JUÁREZ, A. Role of the Hha/YmoA family of proteins in the thermoregulation of the expression of virulence factors. **International journal of medical microbiology : IJMM**, v. 291, n. 6–7, p. 425–32, fev. 2002.

MAEDA, M.; SHIMADA, T.; ISHIHAMA, A. Strength and Regulation of Seven rRNA Promoters in Escherichia coli. **PLOS ONE**, v. 10, n. 12, p. e0144697, 30 dez. 2015.

MAGI, A. et al. Read count approach for DNA copy number variants detection. **Bioinformatics**, v. 28, n. 4, p. 470–478, 15 fev. 2012.

MAHILLON, J.; CHANDLER, M. Insertion sequences. **Microbiology and molecular biology reviews : MMBR**, v. 62, n. 3, p. 725–74, set. 1998.

MALOY, S. Bacterial Genetics. In: **Encyclopedia of Genetics**. [s.l.] Elsevier, 2001. p. 156–163.

MARINUS, M. G. DNA Mismatch Repair. **EcoSal Plus**, v. 5, n. 1, nov. 2012.

MARTINEZ, J. L.; BAQUERO, F. Mutation frequencies and antibiotic resistance. **Antimicrobial agents and chemotherapy**, v. 44, n. 7, p. 1771–7, 1 jul. 2000.

MARTINY, J. B. H. et al. Microbial biogeography: putting microorganisms on the map. **Nature Reviews Microbiology**, v. 4, n. 2, p. 102–112, 1 fev. 2006.

MAZEL, D. Integrons: agents of bacterial evolution. **Nature Reviews Microbiology**, v. 4, n. 8, p. 608–620, 1 ago. 2006.

MAZEL, D. et al. The Integron: Adaptation On Demand. **Microbiology Spectrum**, v. 3, n. 2, 1 abr. 2015.

MCDONNELL, G. Sterilization and Disinfection. **Encyclopedia of Microbiology**, p. 529–548, 2009.

MEADER, S. et al. Genome assembly quality: assessment and improvement using the neutral indel model. **Genome research**, v. 20, n. 5, p. 675–84, maio 2010.

MELL, J. C.; REDFIELD, R. J. Natural competence and the evolution of DNA uptake specificity.

Journal of bacteriology, v. 196, n. 8, p. 1471–83, 15 abr. 2014.

MIDDENDORF, B. et al. Instability of pathogenicity islands in uropathogenic Escherichia coli 536. **Journal of bacteriology**, v. 186, n. 10, p. 3086–96, maio 2004.

MIKI, T.; EASTON, A. M.; ROWND, R. H. Mapping of the resistance genes of the R plasmid NR1. **MGG Molecular & General Genetics**, v. 158, n. 3, p. 217–224, 1978.

MOON, S. et al. Predicting genes expressed via -1 and +1 frameshifts. **Nucleic acids research**, v. 32, n. 16, p. 4884–92, 2004.

MU, W. et al. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. **The Journal of Molecular Diagnostics**, v. 18, n. 6, p. 923–932, 1 nov. 2016.

MULVEY, M. A. Adhesion and entry of uropathogenic Escherichia coli. **Cellular Microbiology**, v. 4, n. 5, p. 257–271, 1 maio 2002.

MUNITA, J. M.; ARIAS, C. A. Mechanisms of Antibiotic Resistance. **Microbiology spectrum**, v. 4, n. 2, 2016.

MUSSER, J. M. et al. Genetic relationships of serologically nontypable and serotype b strains of Haemophilus influenzae. **Infection and immunity**, v. 52, n. 1, p. 183–91, abr. 1986.

NAKAYA, R.; NAKAMURA, A.; MURATA, Y. Resistance transfer agents in Shigella. **Biochemical and Biophysical Research Communications**, v. 3, n. 6, p. 654–659, 1 dez. 1960.

NASCIMENTO, A. M. A. **Clonagem de genes de resistência ao mercúrio em Escherichia coli**. [s.l.] Universidade Federal de Minas Gerais, 1987.

NASCIMENTO, A. M. A.; CHARTONE-SOUZA, E. Operon mer: Bacterial resistance to mercury and potential for bioremediation of contaminated environments. **Genetics and Molecular Research**, v. 2, n. 1, p. 92–101, 2003.

NCBI. **Prokaryotic and Eukaryotic Genomes Submission Guide**. Disponível em: <https://www.ncbi.nlm.nih.gov/genbank/genomesubmit/>. Acesso em: 28 fev. 2018.

NEEL, J. V. A revised estimate of the amount of genetic variation in human proteins: implications for the distribution of DNA polymorphisms. **American journal of human genetics**, v. 36, n. 5, p. 1135–48, set. 1984.

NEVO, E.; BEILES, A.; BEN-SHLOMO, R. The Evolutionary Significance of Genetic Diversity: Ecological, Demographic and Life History Correlates. In: [s.l.] Springer, Berlin, Heidelberg, 1984. p. 13–213.

NGUYEN, Y.; SPERANDIO, V. Enterohemorrhagic E. coli (EHEC) pathogenesis. **Frontiers in cellular and infection microbiology**, v. 2, p. 90, 2012.

NICHOLS, K. B. et al. Molecular Characterization of the Vacuolating Autotransporter Toxin in Uropathogenic Escherichia coli. **Journal of bacteriology**, v. 198, n. 10, p. 1487–98, 15 fev. 2016.

NIETO, J. . et al. Sequence, identification and effect on conjugation of the *rmoA* gene of plasmid R100-1. **FEMS Microbiology Letters**, v. 169, n. 1, p. 59–66, 1 dez. 1998.

NIKAIDO, H. Multidrug resistance in bacteria. **Annual review of biochemistry**, v. 78, p. 119–46, 2009.

NISHINO, K.; NIKAIDO, E.; YAMAGUCHI, A. Regulation of multidrug efflux systems involved in multidrug and metal resistance of Salmonella enterica serovar Typhimurium. **Journal of bacteriology**, v. 189, n. 24, p. 9066–75, 15 dez. 2007.

NIYOGI, S. K. Increasing antimicrobial resistance—an emerging problem in the treatment of shigellosis. **Clinical Microbiology and Infection**, v. 13, n. 12, p. 1141–1143, 1 dez. 2007.

NÚÑEZ, B.; DE LA CRUZ, F. Two atypical mobilization proteins are involved in plasmid CloDF13 relaxation. **Molecular Microbiology**, v. 39, n. 4, p. 1088–1099, 21 dez. 2001.

OCHOA, T. J.; CONTRERAS, C. A. Enteropathogenic escherichia coli infection in children. **Current opinion in infectious diseases**, v. 24, n. 5, p. 478–83, out. 2011.

OKEKE, I. N.; NATARO, J. P. Enteroaggregative Escherichia coli. **The Lancet Infectious Diseases**, v. 1, n. 5, p. 304–313, dez. 2001.

OLIVEIRA, L. DE C. **Análise do potencial probiótico de Lactococcus lactis subsp. lactis NCDO 2118 por meio de genômica comparativa**. [s.l.] Universidade Federal de Minas Gerias, 2014.

OLOOMI, M.; JAVADI, M.; BOUZARI, S. Presence of pathogenicity island related and plasmid encoded virulence genes in cytolethal distending toxin producing Escherichia coli isolates from diarrheal cases. **International journal of applied & basic medical research**, v. 5, n. 3, p. 181–6, 2015.

OOKA, T. et al. Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in Escherichia coli O157 genomes. **Genome research**, v. 19, n. 10, p. 1809–16, out. 2009.

OSHEROFF, W. P. et al. Base substitution specificity of DNA polymerase beta depends on interactions in the DNA minor groove. **The Journal of biological chemistry**, v. 274, n. 30, p. 20749–52, 23 jul. 1999.

OVERBEEK, R. et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). **Nucleic acids research**, v. 42, n. Database issue, p. D206-14, jan. 2014.

PARHAM, N. J. et al. Distribution of the Serine Protease Autotransporters of the Enterobacteriaceae among Extraintestinal Clinical Isolates of Escherichia coli. **Journal of Clinical Microbiology**, v. 43, n. 8, p. 4076–4082, 1 ago. 2005.

PARMAR, K. M. et al. Intriguing Interaction of Bacteriophage-Host Association: An Understanding in the Era of Omics. **Frontiers in microbiology**, v. 8, p. 559, 2017.

PASQUA, M. et al. The Intriguing Evolutionary Journey of EnteroinvasiveE. coli(EIEC) toward Pathogenicity. **Frontiers in microbiology**, v. 8, p. 2390, 2017.

PAULSEN, I. T. et al. The 3' conserved segment of integrons contains a gene associated with multidrug resistance to antiseptics and disinfectants. **Antimicrobial agents and chemotherapy**, v. 37, n. 4, p. 761–8, abr. 1993.

PAUWELS, K. et al. A strong carbon source effect is mediated by pUC plasmid sequences in a series of classical yeast vectors designed for promoter characterization. **Yeast**, v. 15, n. 12, p. 1269–1274, 15 set. 1999.

PECOTA, D. C. et al. Antimicrobial properties of the Escherichia coli R1 plasmid host killing peptide. **Journal of Biotechnology**, v. 100, n. 1, p. 1–12, 9 jan. 2003.

PEPE, C. M. et al. Regulation of the "tetCD" genes of transposon Tn10. **Journal of Molecular Biology**, v. 270, n. 1, p. 14–25, 4 jul. 1997.

PETERSON, J. W. **Bacterial Pathogenesis**. [s.l.] University of Texas Medical Branch at Galveston, 1996.

PHILLIPPY, A. M.; SCHATZ, M. C.; POP, M. Genome assembly forensics: finding the elusive mis-assembly. **Genome Biology**, v. 9, n. 3, p. R55, 2008.

PIRO, V. C. et al. FGAP: an automated gap closing tool. **BMC research notes**, v. 7, p. 371, 18 jun. 2014.

PRAY, L. A. Transposons: The jumping genes. **Nature Education**, v. 1, n. 1, p. 204, 2008.

PRENTICE, M. B. Bacterial comparative genomics. **Genome biology**, v. 5, n. 8, p. 338, 2004.

QADRI, F. et al. Enterotoxigenic Escherichia coli in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention. **Clinical microbiology reviews**, v. 18, n. 3, p. 465–83, jul. 2005.

RAJA, N. et al. Habituation to acid in Escherichia coli: conditions for habituation and its effects on plasmid transfer. **The Journal of applied bacteriology**, v. 70, n. 1, p. 59–65, jan. 1991.

RAMIREZ, M. S.; TOLMASKY, M. E. Aminoglycoside modifying enzymes. **Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy**, v. 13, n. 6, p. 151–71, dez. 2010.

RAMSAY, J. P.; FIRTH, N. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. **Current Opinion in Microbiology**, v. 38, p. 1–9, 1 ago. 2017.

RASTORGUEV, S. M.; ZAVILGELSKY, G. B.; TCHURIKOV, N. A. IncI1 plasmid R64 encodes the ArsR protein that alleviates type I restriction. **FEBS Letters**, v. 426, n. 1, p. 21–23, 10 abr. 1998.

RAYNER, E. et al. A panoply of errors: polymerase proofreading domain mutations in cancer. **Nature Reviews Cancer**, v. 16, n. 2, p. 71–81, 1 fev. 2016.

RAZ, Y.; TANNENBAUM, E. The influence of horizontal gene transfer on the mean fitness of unicellular populations in static environments. **Genetics**, v. 185, n. 1, p. 327–37, maio 2010.

REHM, H. L. Evolving health care through personal genomics. **Nature Reviews Genetics**, v. 18, n. 4, p. 259–267, abr. 2017.

REZNIKOFF, W. S. Transposon Tn *5*. **Annual Review of Genetics**, v. 42, n. 1, p. 269–286, 4 dez. 2008.

RICHARDSON, E. J.; WATSON, M. The automatic annotation of bacterial genomes. **Briefings in bioinformatics**, v. 14, n. 1, p. 1–12, jan. 2013.

ROBASKY, K.; LEWIS, N. E.; CHURCH, G. M. The role of replicates for error mitigation in next-generation sequencing. **Nature reviews. Genetics**, v. 15, n. 1, p. 56–62, 2014.

ROCHA, S. L. DA S. et al. Classification of Avian Pathogenic Escherichia coli (APEC) and Human Uropathogenic Escherichia coli (UPEC) in Phylogenetic Groups and Association with Pathogenicity In Vivo. **Acta Scientiae Veterinariae**, v. 45, n. 1, p. 8, 18 ago. 2017.

RODIONOV, O.; LOBOCKA, M.; YARMOLINSKY, M. Silencing of genes flanking the P1 plasmid centromere. **Science (New York, N.Y.)**, v. 283, n. 5401, p. 546–9, 22 jan. 1999.

ROSS, M. G. et al. Characterizing and measuring bias in sequence data. **Genome Biology**, v. 14, n. 5, p. R51, 29 maio 2013.

ROWND, R.; NAKAYA, R.; NAKAMURA, A. Molecular nature of the drug-resistance factors of the enterobacteriaceae. **Journal of Molecular Biology**, v. 17, n. 2, p. 376–393, 1 jun. 1966.

RUTHERFORD, K. et al. Artemis: sequence visualization and annotation. **Bioinformatics (Oxford, England)**, v. 16, n. 10, p. 944–5, out. 2000.

SAEDLER, H. et al. IS2, a genetic element for turn-off and turn-on of gene activity in E. coli. **MGG Molecular & General Genetics**, v. 132, n. 4, p. 265–289, 1974.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, v. 4, n. 4, p. 406–425, 1 jul. 1987.

SALZBERG, S. L.; YORKE, J. A. Beware of mis-assembled genomes. **Bioinformatics**, v. 21, n. 24, p. 4320–4321, 15 dez. 2005.

SAMBROOK, J. .; RUSSELL, D. . Plasmids and Their Usefulness In Molecular Cloning. In: **Molecular cloning: a laboratory manual**. 3. ed. NY: Cold Spring Harbor, 2001. p. 1.4-1.5.

SCHARBAAI-VÁZQUEZ, R.; CANDELAS, T.; TORRES-BAUZÁ, L. J. Mobilization of the gonococcal 5.2 kb β-lactamase plasmid pSJ5.2 into Escherichia coli by cointegration with several gram-conjugative plasmids. **Plasmid**, v. 57, n. 2, p. 156–164, 1 mar. 2007.

SCHMIDT, H.; HENSEL, M. Pathogenicity islands in bacterial pathogenesis. **Clinical microbiology reviews**, v. 17, n. 1, p. 14–56, jan. 2004.

SCHMOLL, T. et al. Analysis of genes coding for the sialic acid-binding adhesin and two other minor fimbrial subunits of the S-fimbrial adhesin determinant of Escherichia coli. **Molecular microbiology**, v. 3, n. 12, p. 1735–44, dez. 1989.

SCHNEEBERGER, K. et al. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 25, p. 10249–54, 21 jun. 2011.

SCHUBERT, M.; LINDGREEN, S.; ORLANDO, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. **BMC research notes**, v. 9, p. 88, 12 fev. 2016.

SEEMANN, T. **Genome Assembly Strategies**. Disponível em: <http://www.vicbioinformatics.com/documents/Genome Assembly Strategies - Torsten Seemann - IMB - 5 Jul 2010.pdf>.

SEIB, K. L.; ZHAO, X.; RAPPUOLI, R. Developing vaccines in the era of genomics: a decade of reverse vaccinology. **Clinical Microbiology and Infection**, v. 18, p. 109–116, out. 2012.

SELANDER, R. .; CAUGANT, D. .; WHITTAM, T. . Genetic Structure and Variation in Natural Populations of Escherichia coli. In: NEIDHARDT, F. . et al. (Eds.). . **Escherichia coli and Salmonella typhimurium: Celullar and Molecular Biology**. 2. ed. Washington: ASM Press, 1987. p. 1625–1648.

SHAW, W. . The Enzymatic Acetylation of Chloramphenicol by Extracts of R Factor-resistant Escherichiq coZi*. **THE JOURNAL OF BIOLOGICAL CHEMISTRY**, v. 242, n. 4, p. 687–693, 1967.

SHAW, W. V. et al. Primary structure of a chloramphenicol acetyltransferase specified by R plasmids. **Nature**, v. 282, n. 5741, p. 870–872, 20 dez. 1979.

SHELDON, I. M. et al. Specific strains of Escherichia coli are pathogenic for the endometrium of cattle and cause pelvic inflammatory disease in cattle and mice. **PloS one**, v. 5, n. 2, p. e9192, 12 fev. 2010.

SHPIGEL, N. Y.; ELAZAR, S.; ROSENSHINE, I. Mammary pathogenic Escherichia coli. **Current Opinion in Microbiology**, v. 11, n. 1, p. 60–65, fev. 2008.

SIGUIER, P. et al. ISfinder: the reference centre for bacterial insertion sequences. **Nucleic acids research**, v. 34, n. Database issue, p. D32-6, 1 jan. 2006.

SINGER, R. S. Urinary tract infections attributed to diverse ExPEC strains in food animals: evidence and data gaps. **Frontiers in microbiology**, v. 6, p. 28, 2015.

SKÖLD, O. Sulfonamide resistance: mechanisms and trends. **Drug Resistance Updates**, v. 3, n. 3, p. 155–160, 1 jun. 2000.

SKYBERG, J. A. et al. Acquisition of Avian Pathogenic Escherichia coli Plasmids by a Commensal E. coli Isolate Enhances Its Abilities To Kill Chicken Embryos, Grow in Human Urine, and Colonize the Murine Kidney. **Infection and Immunity**, v. 74, n. 11, p. 6287–6292, 1 nov. 2006.

SMILLIE, C. et al. Mobility of plasmids. **Microbiology and molecular biology reviews : MMBR**, v. 74, n. 3, p. 434–52, 1 set. 2010.

SMITH, D. K. et al. Escherichia coli has two homologous glutamate decarboxylase genes that map to distinct loci. **Journal of bacteriology**, v. 174, n. 18, p. 5820–6, set. 1992.

SMITH, M. A.; BIDOCHKA, M. J. Bacterial fitness and plasmid loss: the importance of culture conditions and plasmid size. **Canadian Journal of Microbiology**, v. 44, n. 4, p. 351–355, abr. 1998.

SOM, T.; TOMIZAWA, J. Regulatory regions of ColE1 that are involved in determination of plasmid copy number. **Proceedings of the National Academy of Sciences of the United States of America**, v. 80, n. 11, p. 3232–6, jun. 1983.

SORSA, L. J. et al. Characterization of an iroBCDEN gene cluster on a transmissible plasmid of uropathogenic Escherichia coli: evidence for horizontal transfer of a chromosomal virulence factor. **Infection and immunity**, v. 71, n. 6, p. 3285–93, jun. 2003.

STAWINSKI, P. et al. Basic Bioinformatic Analyses of NGS Data. In: **Clinical Applications for Next-Generation Sequencing**. [s.l.] Elsevier, 2016. p. 19–37.

STEINIGER-WHITE, M.; RAYMENT, I.; REZNIKOFF, W. S. Structure/function insights into Tn5 transposition. **Current Opinion in Structural Biology**, v. 14, n. 1, p. 50–57, 1 fev. 2004.

STONE, A. B. R factors: plasmids conferring resistance to antibacterial agents. **Science progress**, v. 62, n. 245, p. 89–101, 1975.

SUBASHCHANDRABOSE, S. et al. Genome-Wide Detection of Fitness Genes in Uropathogenic Escherichia coli during Systemic Infection. **PLoS Pathogens**, v. 9, n. 12, p. e1003788, 5 dez. 2013.

SUBASHCHANDRABOSE, S.; MOBLEY, H. L. T. Virulence and Fitness Determinants of Uropathogenic Escherichia coli. **Microbiology spectrum**, v. 3, n. 4, ago. 2015.

SUHL, J. A.; WARREN, S. T. Single-Nucleotide Mutations in *FMR1* Reveal Novel Functions and Regulatory Mechanisms of the Fragile X Syndrome Protein FMRP. **Journal of Experimental Neuroscience**, v. 9s2, n. Suppl 2, p. JEN.S25524, 8 jan. 2015.

SUZUKI, H. et al. Antibiotic resistance mutations induced in growing cells of Bacillus-related thermophiles. **The Journal of Antibiotics**, v. 71, n. 3, p. 382–389, 18 mar. 2018.

SWAMINATHAN, K.; VARALA, K.; HUDSON, M. E. Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. **BMC genomics**, v. 8, p. 132, 24 maio 2007.

TAMAZIAN, G. et al. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. **GigaScience**, v. 5, n. 1, p. 38, 22 ago. 2016.

TATUSOVA, T. et al. NCBI prokaryotic genome annotation pipeline. **Nucleic acids research**, v. 44, n. 14, p. 6614–24, 19 ago. 2016.

TENAILLON, O. et al. Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. **Research in microbiology**, v. 152, n. 1, p. 11–16, 2001.

TERLIZZI, M. E.; GRIBAUDO, G.; MAFFEI, M. E. UroPathogenicEscherichia coli(UPEC) Infections: Virulence Factors, Bladder Responses, Antibiotic, and Non-antibiotic Antimicrobial Strategies. **Frontiers in microbiology**, v. 8, p. 1566, 2017.

THISTED, T.; GERDES, K. Mechanism of post-segregational killing by the hok/sok system of plasmid R1: Sok antisense RNA regulates hok gene expression indirectly through the overlapping mok gene. **Journal of Molecular Biology**, v. 223, n. 1, p. 41–54, 5 jan. 1992.

TIWARI, S. et al. C. pseudotuberculosis Phop confers virulence and may be targeted by natural compounds. **Integr. Biol.**, v. 6, n. 11, p. 1088–1099, 12 set. 2014.

TODAR, K. **Corynebacterium diphtheriae and Diphtheria**. Disponível em: <http://textbookofbacteriology.net/diphtheria_3.html>. Acesso em: 9 fev. 2018.

TREANGEN, T. J.; SALZBERG, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. **Nature reviews. Genetics**, v. 13, n. 1, p. 36–46, 29 nov. 2011.

TRIVEDI, U. H. et al. Quality control of next-generation sequencing data without a reference. **Frontiers in genetics**, v. 5, p. 111, 2014.

TSUCHIMOTO, S.; OHTSUBO, H.; OHTSUBO, E. Two genes, pemK and pemI, responsible for stable maintenance of resistance plasmid R100. **Journal of bacteriology**, v. 170, n. 4, p. 1461–6, abr. 1988.

TUFARIELLO, J. et al. Specialized transduction: an efficient method for generating marked and unmarked targeted gene disruptions in Mycobacterium tuberculosis, M. bovis BCG and M. smegmatis. **Microbiology**, v. 148, n. 10, p. 3007–3017, 1 out. 2002.

UHLIN, B. E.; NORDSTRÖM, K. R plasmid gene dosage effects in Escherichia coli K-12: Copy mutants of the R plasmid R1drd-19. **Plasmid**, v. 1, n. 1, p. 1–7, 1 nov. 1977.

VAN BIESEN, T.; FROST, L. S. Differential levels of fertility inhibition among F-like plasmids are related to the cellular concentration of *finO* mRNA. **Molecular Microbiology**, v. 6, n. 6, p. 771–780, 1 mar. 1992.

VARSAKI, A. et al. Analysis of ColE1 MbeC unveils an extended ribbon-helix-helix family of nicking accessory proteins. **Journal of bacteriology**, v. 191, n. 5, p. 1446–55, 1 mar. 2009.

VEJBORG, R. M. et al. Comparative genomics of Escherichia coli strains causing urinary tract infections. **Applied and environmental microbiology**, v. 77, n. 10, p. 3268–78, 15 maio 2011.

VĚTROVSKÝ, T.; BALDRIAN, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. **PloS one**, v. 8, n. 2, p. e57923, 2013.

VIANA, M. V. C. et al. Comparative genomic analysis between Corynebacterium pseudotuberculosis strains isolated from buffalo. **PLOS ONE**, v. 12, n. 4, p. e0176347, 26 abr. 2017.

WACHINO, J. et al. Horizontal transfer of blaCMY-bearing plasmids among clinical Escherichia coli and Klebsiella pneumoniae isolates and emergence of cefepime-hydrolyzing CMY-19. **Antimicrobial agents and chemotherapy**, v. 50, n. 2, p. 534–41, fev. 2006.

WADDELL, T. E. et al. Generalized Transduction by Lytic Bacteriophages. In: **Methods in molecular biology (Clifton, N.J.)**. [s.l: s.n.]. v. 501p. 293–303.

WALDOR, M. K.; DAVIS, B. M. Mobile Genetic Elements and Bacterial Pathogenesis. In: **Mobile DNA II**. [s.l.] American Society of Microbiology, 2002. p. 1040–1059.

WANG, G. et al. Spontaneous mutations that confer antibiotic resistance in Helicobacter pylori. **Antimicrobial agents and chemotherapy**, v. 45, n. 3, p. 727–33, mar. 2001.

WATSON, R. J. et al. Characterization and nucleotide sequence of a colicin-release gene in the hie region of plasmid ColE3-CA38. **Gene**, v. 29, n. 1–2, p. 175–184, 1 jul. 1984.

WATTAM, A. R. et al. PATRIC, the bacterial bioinformatics database and analysis resource.

Nucleic acids research, v. 42, n. Database issue, p. D581-91, jan. 2014.

WELCH, R. A. Uropathogenic Escherichia coli-Associated Exotoxins. **Microbiology spectrum**, v. 4, n. 3, 2016.

WILANTHO, A. et al. Next generation sequencing (NGS) technologies and their applications in omics-research. **Thai J. Genet**, v. 5, n. 2, p. 104–129, 2012.

WILLETTS, N. S.; FINNEGAN, D. J. Characteristics of E. coli K12 strains carrying both an F prime and an R factor. **Genetical research**, v. 16, n. 1, p. 113–22, ago. 1970.

WOMBLE, D. D.; ROWND, R. H. Genetic and physical map of plasmid NR1: comparison with other IncFII antibiotic resistance plasmids. **Microbiological reviews**, v. 52, n. 4, p. 433–51, dez. 1988.

WU, R. et al. Expression of the repA1 gene of IncFII plasmid NR1 is translationally coupled to expression of an overlapping leader peptide. **Journal of bacteriology**, v. 174, n. 23, p. 7620–8, dez. 1992.

YAMADA, Y.; YAMADA, M.; NAKAZAWA, A. A ColE1-Encoded Gene Directs Entry Exclusion of the Plasmid. **JOURNAL OF BACTERIOLOGY**, v. 177, n. 21, p. 6064–6068, 1995.

YOSHIOKA, Y.; OHTSUBO, H.; OHTSUBO, E. Repressor gene finO in plasmids R100 and F: constitutive transfer of plasmid F is caused by insertion of IS3 into F finO. **Journal of bacteriology**, v. 169, n. 2, p. 619–23, fev. 1987.

ZANKARI, E. et al. Identification of acquired antimicrobial resistance genes. **The Journal of antimicrobial chemotherapy**, v. 67, n. 11, p. 2640–4, nov. 2012.

ZHANG, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. **Science (New York, N.Y.)**, v. 346, n. 6215, p. 1311–20, 12 dez. 2014.

ZHANG, J. et al. Interference of mRNA function by sequence-specific endoribonuclease PemK. **The Journal of biological chemistry**, v. 279, n. 20, p. 20678–84, 14 maio 2004.

ZHANG, W. et al. A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. **PLoS ONE**, v. 6, n. 3, p. e17915, 14 mar. 2011.

ZIMIN, A. V et al. Hybrid assembly of the large and highly repetitive genome ofAegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. **Genome**

research, v. 27, n. 5, p. 787–792, 27 jan. 2017.

ZURFLUH, K.; TASARA, T.; STEPHAN, R. Full-Genome Sequence of Escherichia coli K-15KW01, a Uropathogenic E. coli B2 Sequence Type 127 Isolate Harboring a Chromosomally Carried blaCTX-M-15 Gene. **Genome announcements**, v. 4, n. 5, 1 set. 2016.

ANEXOS

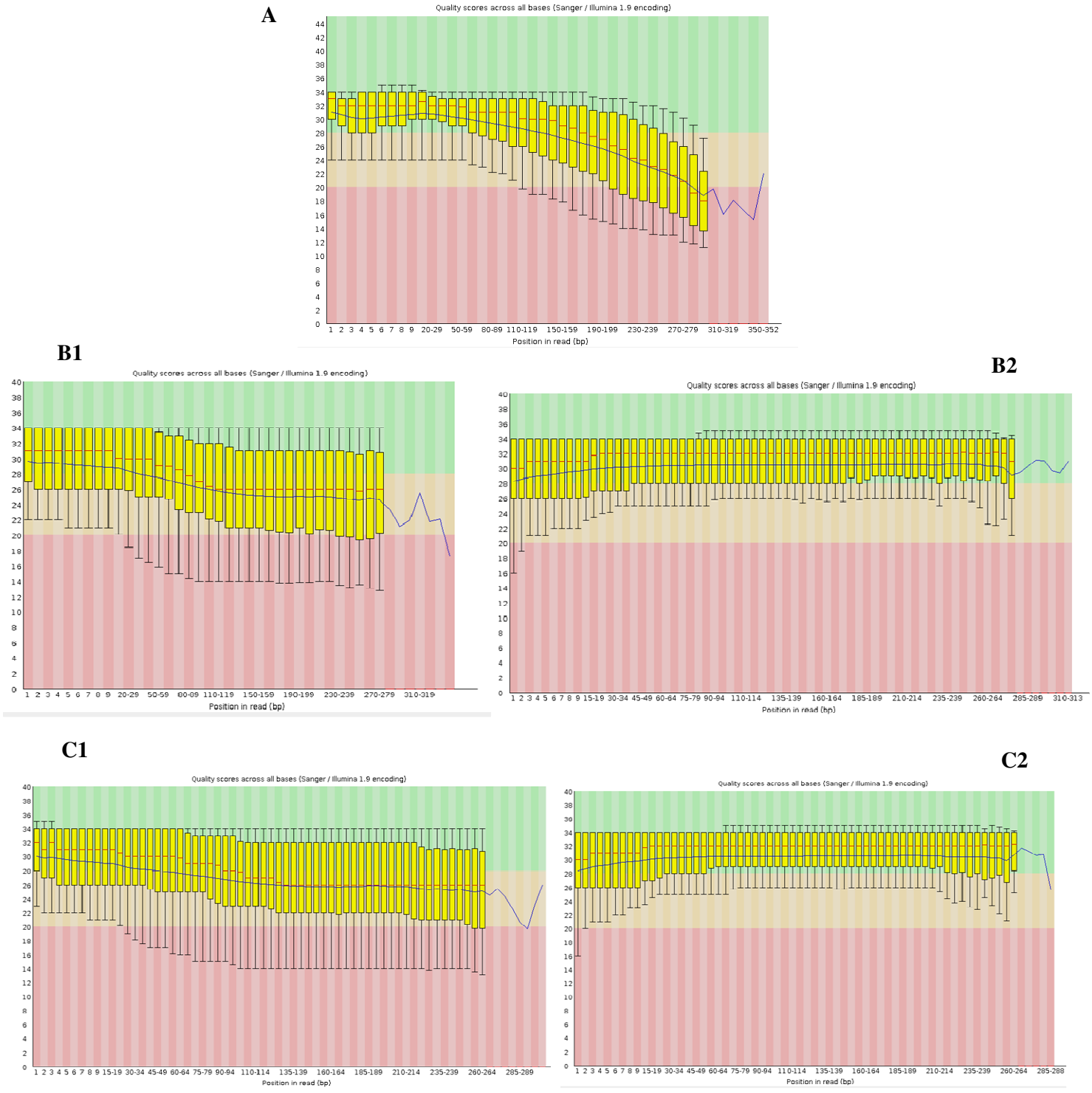


Figura S1 Histograma de qualidade das *reads* de sequenciamento da *E. coli* BH100 subtr. MG2014. A) 4.291.689 *reads fragment* com tamanhos variando entre 8 e 352 pb*;* B1) 1.830.545 *reads mate-pair* 3kb *forward* com tamanhos variando entre 1 e 345 pb*;* B2) 1.830.545 *reads mate-pair* 3kb *reverse* com tamanhos variando entre 1 a 313 pb; C1) 2.291.511 *reads mate-pair 6kb* forward com tamanhos variando entre 1 a 300 pb; C2) 2.291.511 *reads mate-pair* 6kb reverse com tamanhos variando entre 1 e 288 pb.

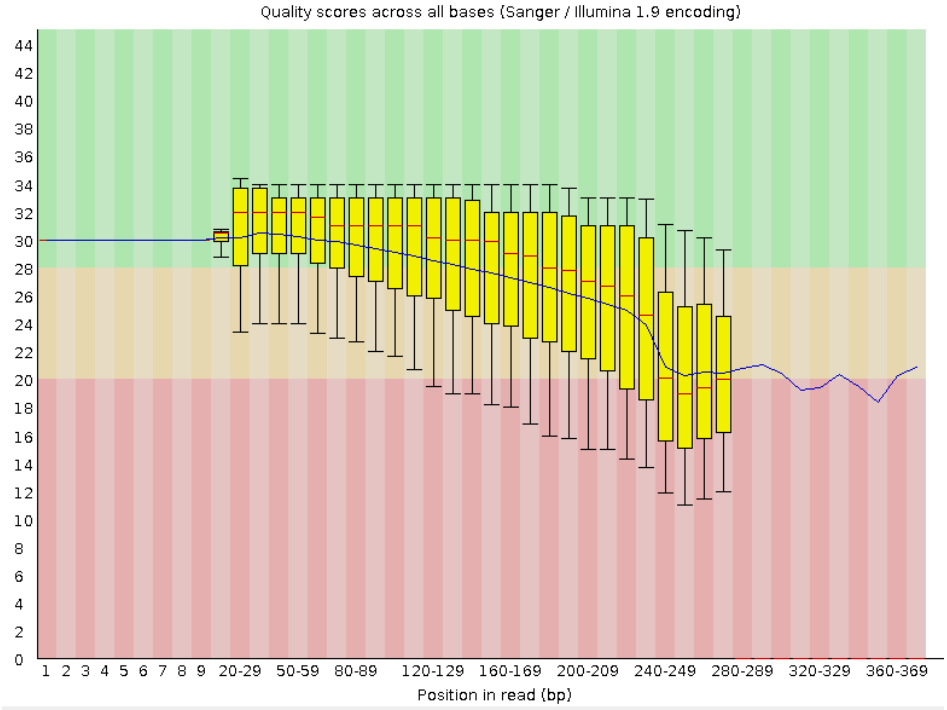


Figura S2 Histograma de qualidade de 2.509.671 *reads fragment* da *E. coli* BH100S subtr. MG2014, com tamanhos variando entre 25 e 373 pb.

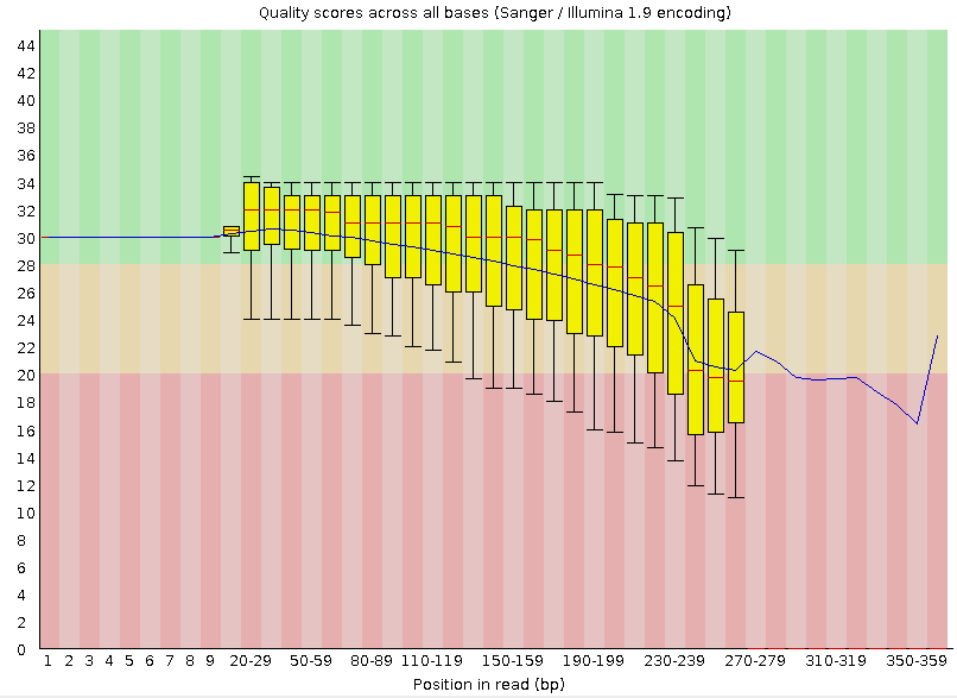


Figura S3 Histograma de qualidade de 2.247.497 *reads fragment* da *E. coli* BH100L subtr. MG2014, com tamanhos variando entre 25 e 368 pb.

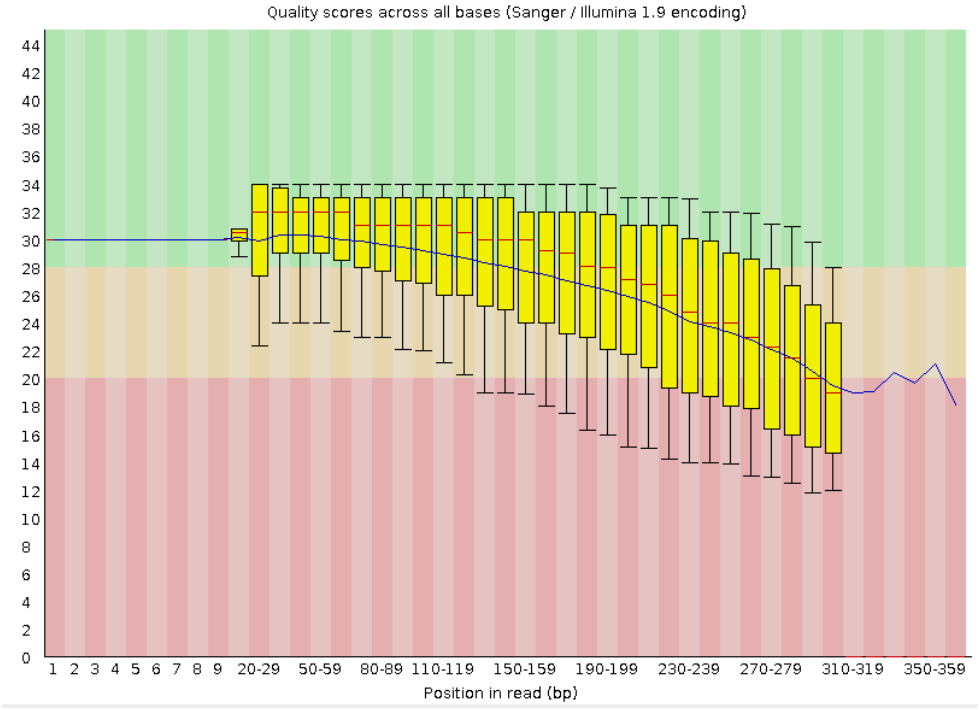


Figura S4 Histograma de qualidade de 2.253.026 *reads fragment* da *E. coli* BH100N subtr. MG2014, com tamanhos variando entre 25 e 367 pb.

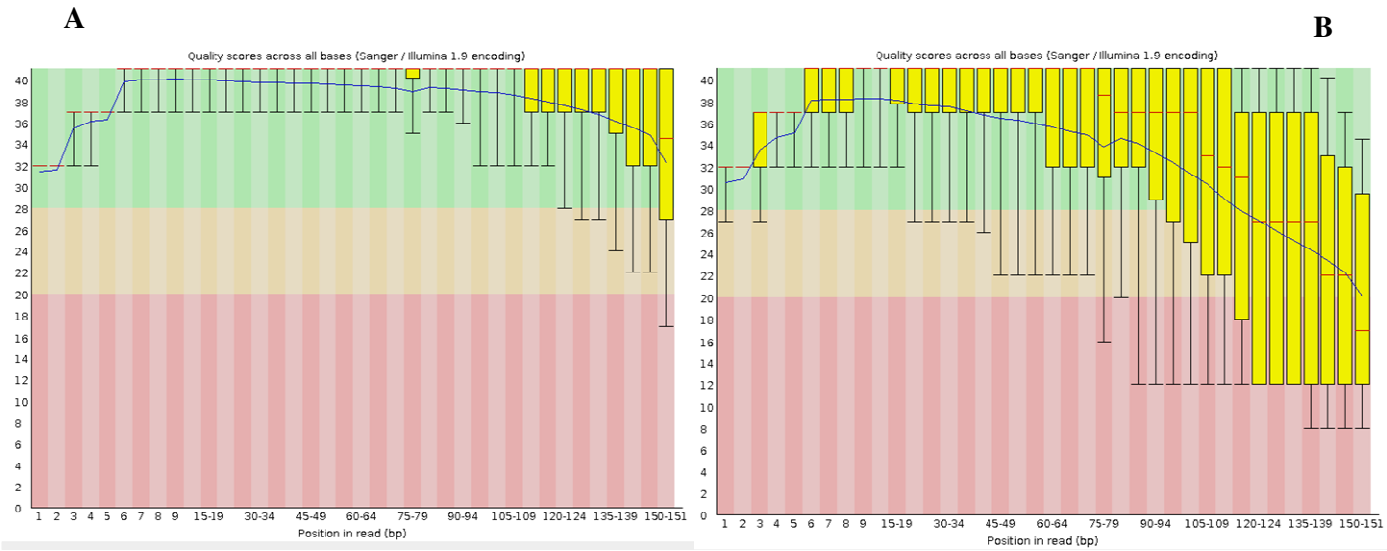


Figura S5 Histograma de qualidade de 8.194.154 *reads paired-end* (2x 151pb) da *E. coli* BH100 subtr. MG2017. A) *reads forward*; B) *reads reverse*.

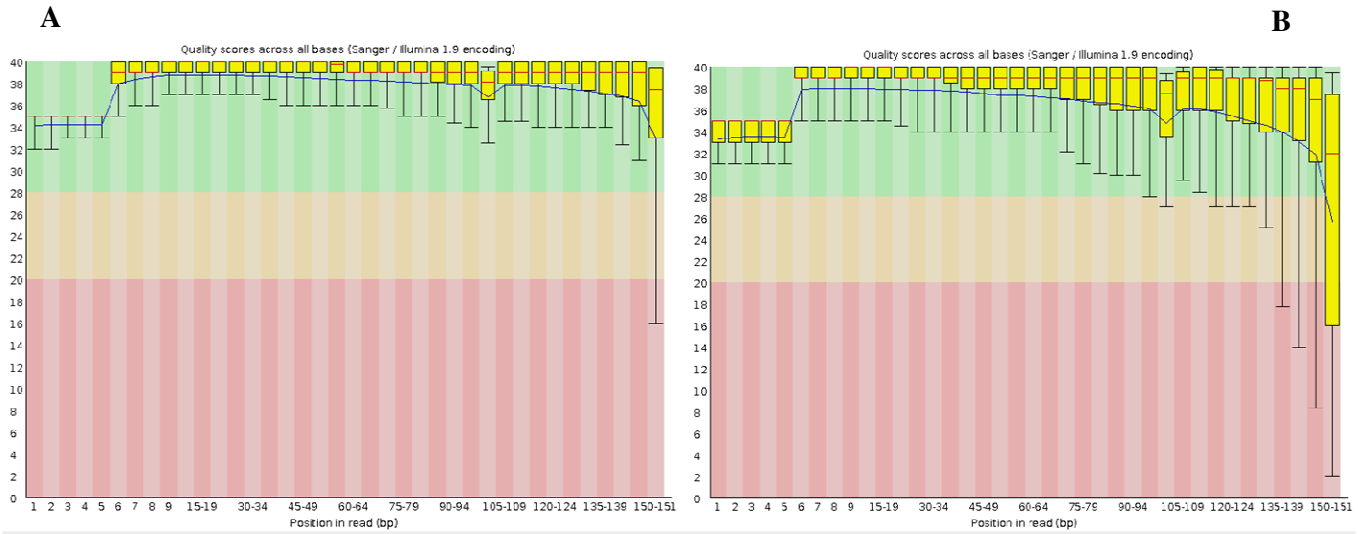


Figura S6 Histograma de qualidade de 6.275.281 *reads paired-end* (2x 151pb) da *E. coli* BH100N subtr. MG2017. A) *reads forward*; B) *reads reverse*.

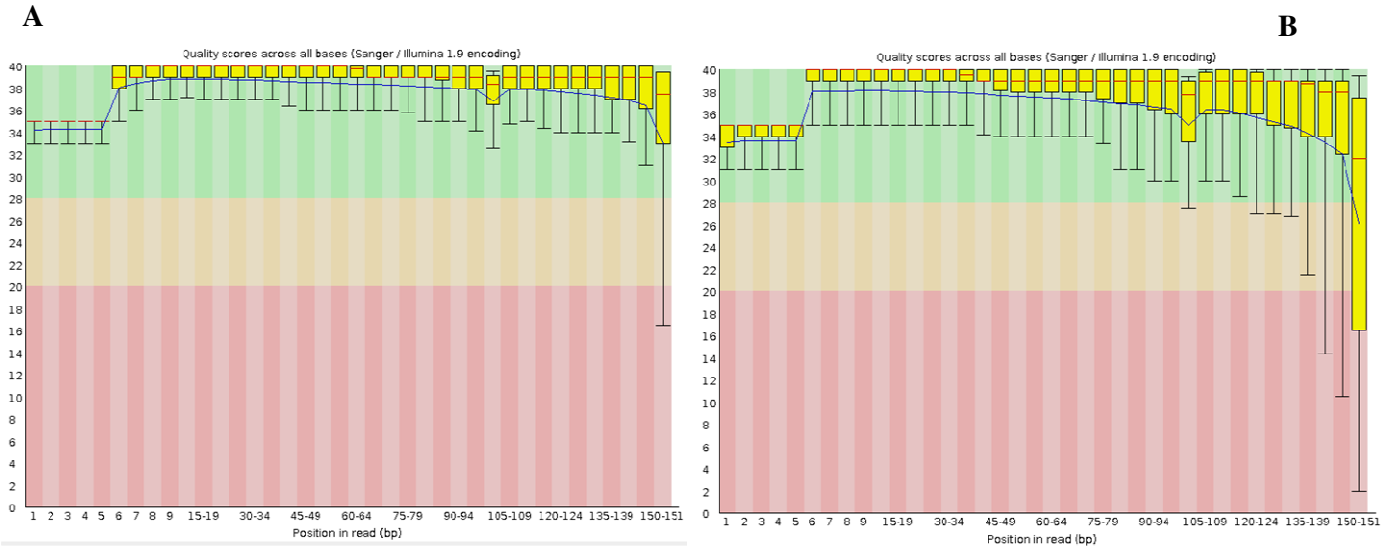


Figura S7 Histograma de qualidade de 6.685.569 *reads paired-end* (2x 151pb) da *E. coli* BH100L subtr. MG2017. A) *reads forward*; B) *reads reverse*.

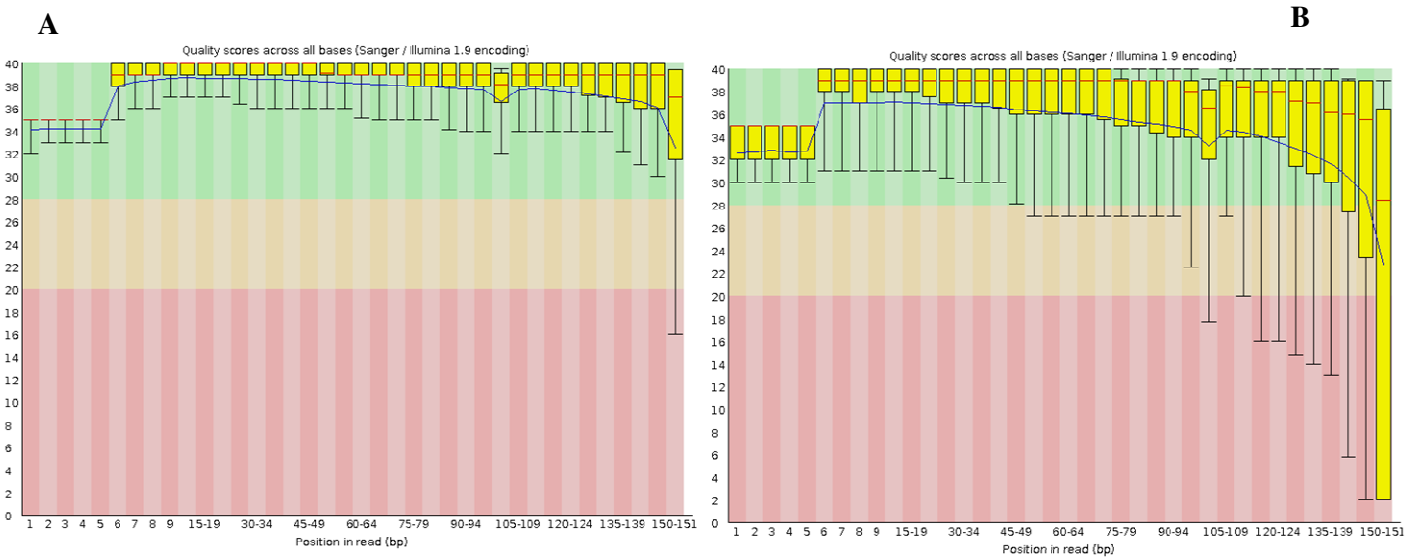


Figura S8 Histograma de qualidade de 7.156.086 *reads paired-end* (2x 151pb) da *E. coli* K12T subtr. MG2017. A) *reads forward*; B) *reads reverse*

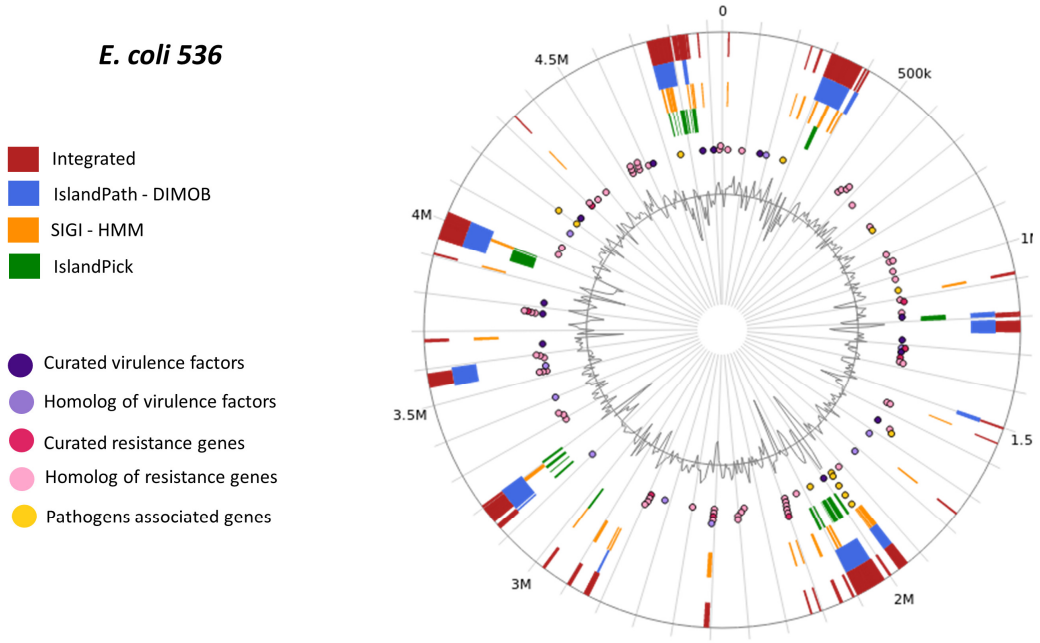


Figura S9 Ilhas genômicas identificadas no IslandViewer4 presentes na *E. coli* 536. Os quadriláteros indicam as diferentes estratégias de predição de ilhas. Os círculos representam genes com alguma associação com base de dados de fatores de virulência, resistência ou exclusivamente associados à patógenos.

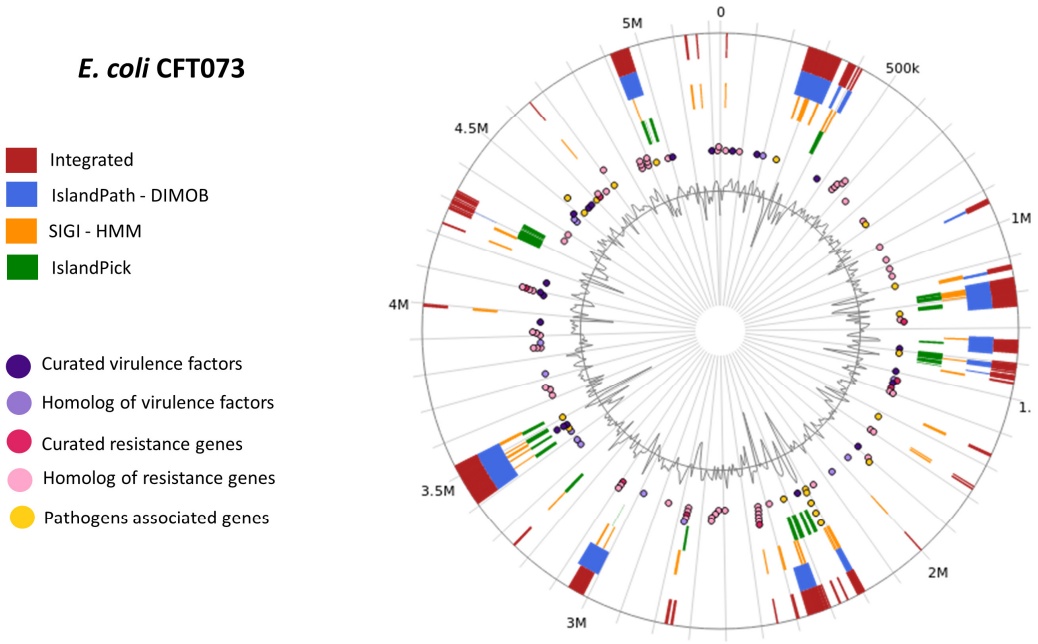


Figura S10 Ilhas genômicas identificadas no IslandViewer4 presentes na *E. coli* CFT073. Os quadriláteros indicam as diferentes estratégias de predição de ilhas. Os círculos representam genes com alguma associação com base de dados de fatores de virulência, resistência ou exclusivamente associados à patógenos.

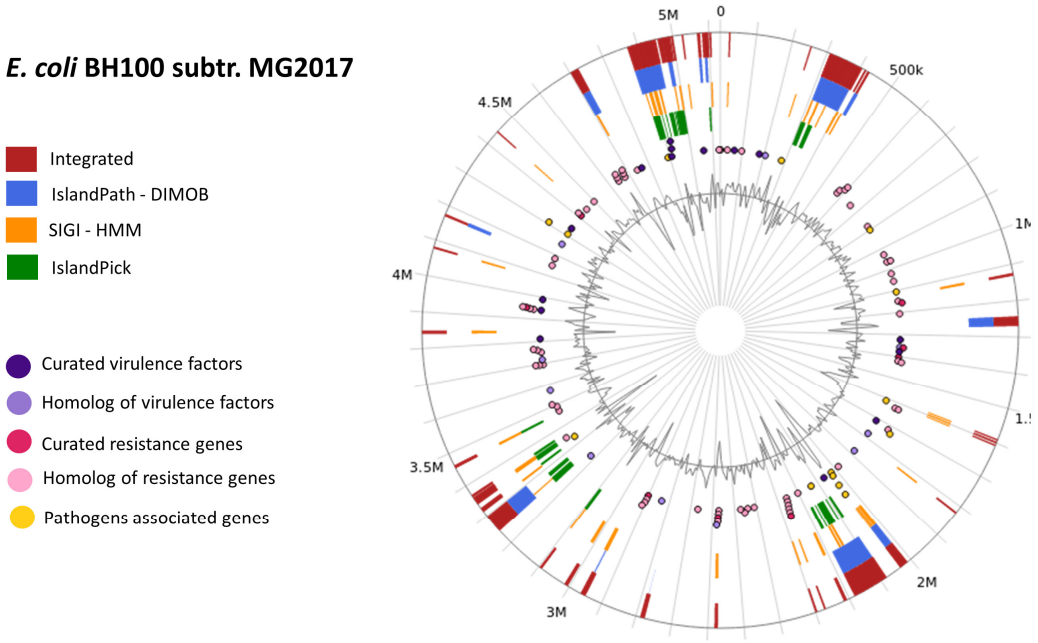


Figura S11 Ilhas genômicas identificadas no IslandViewer4 presentes na *E. coli* BH100 subtr. MG2017. Os quadriláteros indicam as diferentes estratégias de predição de ilhas. Os círculos representam genes com alguma associação com base de dados de fatores de virulência, resistência ou exclusivamente associados à patógenos.

APÊNDICES

APÊNDICE A: VARIABILIDADE FENOTÍPICA E GENOTÍPICA OBSERVADA E PREDITA NA *E. COLI* BH100

Desvios do esperado em experimentos de conjugação envolvendo os plasmídeos da *E. coli* BH100 levaram à hipótese da presença de marcadores de resistência em transposons. De fato, como aqui demonstrado, todos os marcadores de resistência, dos grupos pBH100 e pAp estão dispostas nesses elementos móveis. Desprezando-se possíveis duplicidades desses elementos num mesmo replicon, o marcador de resistência à estreptomicina e mecanismos de mobilização envolvendo ISs não vinculados aos respectivos transposons, pode-se estimar o número de combinações genéticas possíveis de cada marcador em cada replicon, como segue: 1) cada replicon (cromossomo, pBH100 ou pAp) pode abrigar um, dois, três, quatro ou cinco marcadores de resistência (mercúrio, ampicilina, canamicina, tetraciclina e cloranfenicol); 2) Há a possibilidade de cada um desses replicons não possuir marcador de resistência; 3) Assumindo os pressupostos 1 e 2 pode-se prever 326 combinações de marcadores de resistência para cada replicon; 4) Considerando que uma célula pode conter os três replicons (cromossomo, pBH100 e pAp), dois replicons (cromossomo e pBH100, ou cromossomo e pAp) ou apenas um replicon (cromossomo) pode-se prever o total de combinações para cada grupo; 5) em cada grupo deve-se subtrair uma combinação, referente à possível ausência de de marcadores de resistência em todos os replicons do grupo. Aplicando essa abordagem de análise combinatória (**Figura A1**) e considerando as premissas acima, pode-se estimar, ao todo, um total de 2.604 combinações genotípicas dos marcadores de resistência a antibióticos, salvo estreptomicina, ao longo de todos os possíveis variantes da *E. coli* BH100. Esse número de combinações são passiveis de gerar 325 possíveis combinações fenotípicas dos marcadores de resistência presentes na *E. coli* BH100.[35]

[35] Calculado de forma análoga às combinações genotípicas, desprezando-se a resistência à estreptomicina.

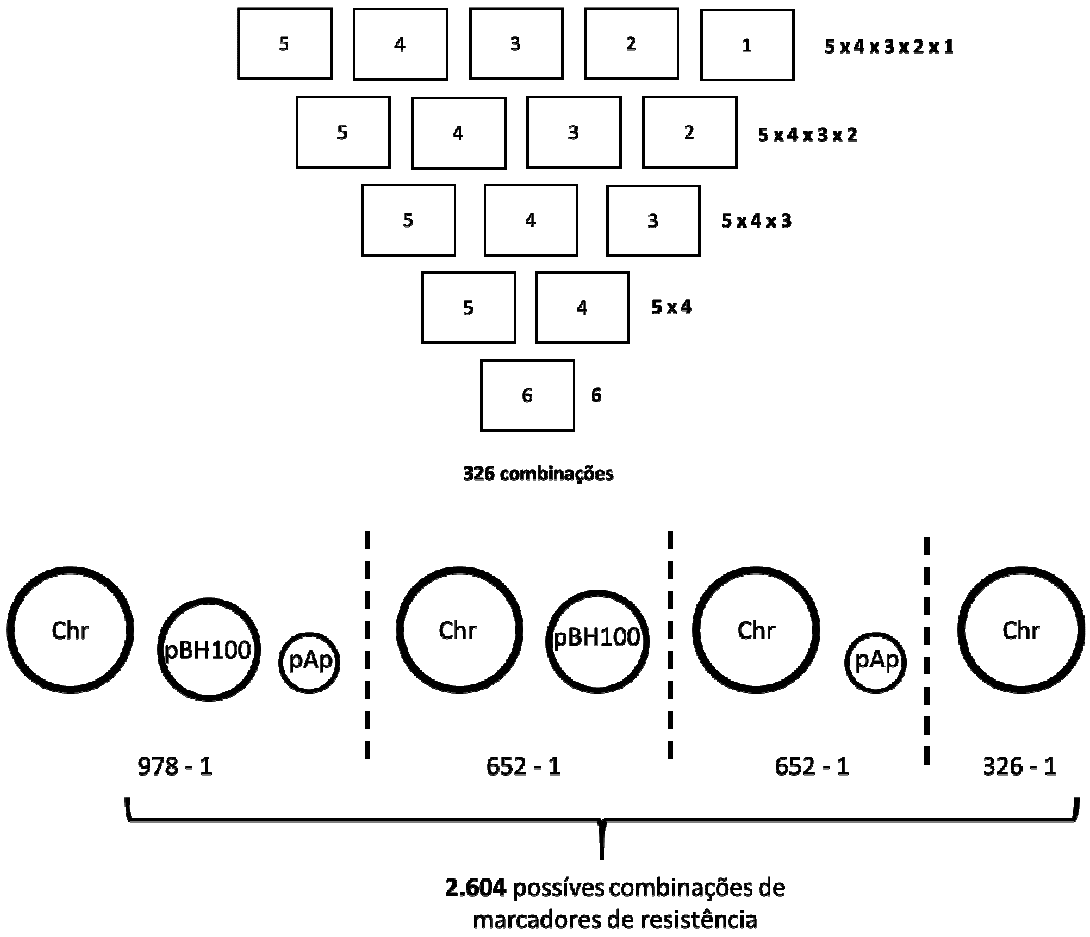


Figura A1 Representação esquemática do cálculo de análise combinatória para estimativa do número de combinações genéticas dos marcadores de resistência à antibióticos, salvo estreptomicina, considerando-os em cópia única por replicon.

De fato, observando para os desvios da incorporação dos marcadores de resistência em culturas de *E. coli* K-12, resistente à estreptomicina, transconjugantes, e obtidas de experimento de conjugação com a *E. coli* BH100, isoladas em meio contendo estreptomicina e tetraciclina (SmTc) e em estreptomicina e ampicilina (SmAp) (Tabelas A1 e A2) é possível averiguar grande variabilidade fenotípica. Tamanha variabilidade, justifica-se devido diferentes combinações de transposons ao longo dos diferentes replicons (cromossomo, plasmídeos pAp e plasmídeos pBH100). A presença desses elementos móveis, portanto, contabilizam para o surgimento de diversidade genotípica e fenotípica na *E. coli* BH100.

Tabela A1 Cresimento de colônias selecionadas em meio SmTc e crescidas em plascas contendo diferentes antimicrobianos.
(Replica 1)

Ant\Col.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
SmTc	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
SmCm	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+
SmKm	+	+	-	-	+	+	-	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	+
SmAp	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
SmHg	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+
EMB	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

(Replica 2)

Ant\Col.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
SmTc	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+
SmCm	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+
SmKm	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	-	-	-	+	+	+	+	-	+	+	+	+	+	+	-	+	-
SmAp	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	+	-	+	-
SmHg	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+
EMB	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Sm: estreptomicina; Tc: tetraciclina; Cm: cloranfenicol; Km: canamicina; Ap: ampicilina; Hg: mercúrio inorgânico; EMB: Meio eosina miosina; +: crescimento; -: não crecimento.

Tabela A2 Cresimento de colônias selecionadas em meio SmAp e crescidas em plascas contendo diferentes antimicrobianos.
(Replica 1)

Ant\Col.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
SmTc	-	-	-	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-
SmCm	-	-	-	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	+	-	-	-	+	-	-	-	-	-
SmKm	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SmAp	+	-	+	-	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+
SmHg	-	-	-	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	+	-	-	+	-	-	-	-	-	-
EMB	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	*	+	+	+	+	+	+

(Replica 2)

Ant\Col.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
SmTc	+	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	+	-	-	-	+	-	+	-	-	-	+	-	-
SmCm	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	-	-	-	-	+	-	-	-	+	-	-
SmKm	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SmAp	+	-	+	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SmHg	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	+	-	-	+	-	+	-	-	-	-	+	-	-
EMB	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Sm: estreptomicina; Tc: tetraciclina; Cm: cloranfenicol; Km: canamicina; Ap: ampicilina; Hg: mercúrio inorgânico; EMB: Meio eosina miosina; +: crescimento; -: não crecimento.

APÊNDICE B: LINHAGENS DE *E. COLI* EXTRAÍDAS DO NCBI E SEUS RESPECTIVOS PATÓTIPOS

Quadro S1 Linhagens obtidas no NCBI e utilizadas para fins comparativos neste trabalho.

RefSeq	Linhagem	Patótipo*
NC_007779.1	Escherichia coli str. K-12 sub-str. W3110	NONPATHOGENIC
NC_000913	Escherichia coli str. K-12 sub-str. MG1655	NONPATHOGENIC
NC_002695	Escherichia coli O157:H7 str. Sakai	EHEC
NC_004431	Escherichia coli CFT073	UPEC
NC_007946	Escherichia coli UTI89	UPEC
NC_008253	Escherichia coli 536	UPEC
NC_008563	Escherichia coli APEC O1	APEC
NC_009800	Escherichia coli HS	NONPATHOGENIC
NC_009801	Escherichia coli E24377A	ETEC
NC_010468	Escherichia coli ATCC 8739	NONPATHOGENIC
NC_010473	Escherichia coli str. K-12 sub-str. DH10B	NONPATHOGENIC
NC_010498	Escherichia coli SMS-3-5	ENVIROMENTAL
NC_011353	Escherichia coli O157:H7 str. EC4115	EHEC
NC_011415	Escherichia coli SE11	NONPATHOGENIC
NC_011601	Escherichia coli O127:H6 str. E2348/69	EHEC
NC_011741	Escherichia coli IAI1	NONPATHOGENIC
NC_011742	Escherichia coli S88	EXPEC
NC_011748	Escherichia coli 55989	EAEC
NC_011750	Escherichia coli IAI39	UPEC
NC_011993	Escherichia coli LF82	AIEC
NC_012759	Escherichia coli BW2952	NONPATHOGENIC
NC_012892	Escherichia coli BL21(DE3) (Prj:28965)	NONPATHOGENIC
NC_012947	Escherichia coli 'BL21-Gold(DE3)pLysS AG'	NONPATHOGENIC
NC_012967	Escherichia coli B str. REL606	APEC
NC_012971	Escherichia coli BL21(DE3)	NONPATHOGENIC
NC_013008	Escherichia coli O157:H7 str. TW14359	EHEC
NC_013353	Escherichia coli O103:H2 str. 12009	EHEC
NC_013361	Escherichia coli O26:H11 str. 11368	EHEC
NC_013364	Escherichia coli O111:H- str. 11128	EHEC
NC_013654	Escherichia coli SE15	

NC_013941	Escherichia coli O55:H7 str. CB9615	EHEC
NC_016902	Escherichia coli KO11FL	NONPATHOGENIC
NC_017625	Escherichia coli DH1	NONPATHOGENIC
NC_017626	Escherichia coli 042	EAEC
NC_017628	Escherichia coli IHE3034	EXPEC
NC_017631	Escherichia coli ABU 83972	UPEC
NC_017632	Escherichia coli UM146	AIEC
NC_017633	Escherichia coli ETEC H10407	ETEC
NC_017634	Escherichia coli O83:H1 str. NRG 857C	AIEC
NC_017635	Escherichia coli W	NONPATHOGENIC
NC_017638	Escherichia coli DH1	NONPATHOGENIC
NC_017641	Escherichia coli UMNK88	ETEC
NC_017646	Escherichia coli O7:K1 str. CE10	EXPEC
NC_017651	Escherichia coli str. 'clone D i2'	UPEC
NC_017652	Escherichia coli str. 'clone D i14'	UPEC
NC_017656	Escherichia coli O55:H7 str. RM12579	EPEC
NC_017660	Escherichia coli KO11FL	NONPATHOGENIC
NC_017663	Escherichia coli P12b	NONPATHOGENIC
NC_017664	Escherichia coli W	NONPATHOGENIC
NC_017906	Escherichia coli Xuzhou21	UPEC
NC_018650	Escherichia coli O104:H4 str. 2009EL-2050	EHEC
NC_018658	Escherichia coli O104:H4 str. 2011C-3493	EHEC
NC_018661	Escherichia coli O104:H4 str. 2009EL-2071	EHEC
NC_020163	Escherichia coli APEC O78	APEC
NC_020518	Escherichia coli str. K-12 sub-str. MDS42	NONPATHOGENIC
NC_022364	Escherichia coli LY180	NONPATHOGENIC
NC_022648	Escherichia coli JJ1886	EXPEC
NZ_CP005930	Escherichia coli APEC IMT5155	APEC
NZ_CP005998	Escherichia coli B7A	ETEC
NZ_CP006027	Escherichia coli O145:H28 str. RM13514	EHEC
NZ_CP006262	Escherichia coli O145:H28 str. RM13516	EHEC
NZ_CP006636	Escherichia coli PCN061	EXPEC
NZ_CP007133	Escherichia coli O145:H28 str. RM12761	EHEC

NZ_CP007136	Escherichia coli O145:H28 str. RM12581	EHEC
NZ_CP007149	Escherichia coli RS218	EXPEC
NZ_CP007265	Escherichia coli strain ST540	UPEC
NZ_CP007390	Escherichia coli strain ST540	ENVIROMENTAL
NZ_CP007391	Escherichia coli strain ST540	ENVIROMENTAL
NZ_CP007392	Escherichia coli strain ST2747	UNKNOWN
NZ_CP007393	Escherichia coli strain ST2747	ENVIROMENTAL
NZ_CP007394	Escherichia coli strain ST2747	ENVIROMENTAL
NZ_CP007442	Escherichia coli ACN001	APEC
NZ_CP007491	Escherichia coli strain ACN002	APEC
NZ_CP007592	Escherichia coli O157:H16 Santai	EHEC
NZ_CP007594	Escherichia coli strain SEC470	EHEC
NZ_CP007799	Escherichia coli Nissle 1917	NONPATHOGENIC
NZ_CP008697	Escherichia coli strain ST648	EXPEC
NZ_CP008801	Escherichia coli KLY	UNKNOWN
NZ_CP008805	Escherichia coli O157:H7 str. SS17	EHEC
NZ_CP008957	Escherichia coli O157:H7 str. EDL933	EHEC
NZ_CP009072	Escherichia coli ATCC 25922	UNKNOWN
NZ_CP009104	Escherichia coli RM9387	EHEC
NZ_CP009106	Escherichia coli 94-3024	EHEC
NZ_CP009166	Escherichia coli 1303	UNKNOWN
NZ_CP009273	Escherichia coli BW25113	NONPATHOGENIC
NZ_CP009644	Escherichia coli ER2796	NONPATHOGENIC
NZ_CP009685	Escherichia coli str. K-12 sub-str. MG1655	NONPATHOGENIC
NZ_CP009789	Escherichia coli K-12 ER3413	NONPATHOGENIC
NZ_CP009859	Escherichia coli ECONIH1	EXPEC
NZ_CP010116	Escherichia coli strain C1	UNKNOWN
NZ_CP010117	Escherichia coli strain C2	UNKNOWN
NZ_CP010119	Escherichia coli strain C3	UNKNOWN
NZ_CP010121	Escherichia coli strain C4	UNKNOWN
NZ_CP010122	Escherichia coli strain C5	UNKNOWN
NZ_CP010125	Escherichia coli strain C8	UNKNOWN
NZ_CP010129	Escherichia coli strain C9	UNKNOWN
NZ_CP010132	Escherichia coli strain C10	UNKNOWN
NZ_CP010133	Escherichia coli strain C11	UNKNOWN
NZ_CP010134	Escherichia coli strain D1	UNKNOWN
NZ_CP010137	Escherichia coli strain D2	UNKNOWN
NZ_CP010140	Escherichia coli strain D3	UNKNOWN
NZ_CP010143	Escherichia coli strain D4	UNKNOWN
NZ_CP010145	Escherichia coli strain D5	UNKNOWN
NZ_CP010148	Escherichia coli strain D6	UNKNOWN
NZ_CP010150	Escherichia coli strain D7	UNKNOWN

NZ_CP010151	Escherichia coli strain D8	UNKNOWN
NZ_CP010152	Escherichia coli strain D9	UNKNOWN
NZ_CP010157	Escherichia coli strain D10	UNKNOWN
NZ_CP010160	Escherichia coli strain H1	UNKNOWN
NZ_CP010163	Escherichia coli strain H2	UNKNOWN
NZ_CP010167	Escherichia coli strain H3	UNKNOWN
NZ_CP010169	Escherichia coli strain H5	UNKNOWN
NZ_CP010170	Escherichia coli strain H6	UNKNOWN
NZ_CP010171	Escherichia coli strain H7	UNKNOWN
NZ_CP010172	Escherichia coli strain H8	UNKNOWN
NZ_CP010176	Escherichia coli strain H10	UNKNOWN
NZ_CP010177	Escherichia coli strain H14	UNKNOWN
NZ_CP010178	Escherichia coli strain H15	UNKNOWN
NZ_CP010180	Escherichia coli strain M1	UNKNOWN
NZ_CP010183	Escherichia coli strain M3	UNKNOWN
NZ_CP010186	Escherichia coli strain M6	UNKNOWN
NZ_CP010191	Escherichia coli strain M8	UNKNOWN
NZ_CP010196	Escherichia coli strain M9	UNKNOWN
NZ_CP010200	Escherichia coli strain M10	UNKNOWN
NZ_CP010206	Escherichia coli strain M11	UNKNOWN
NZ_CP010213	Escherichia coli strain M15	UNKNOWN
NZ_CP010219	Escherichia coli strain M18	UNKNOWN
NZ_CP010221	Escherichia coli strain M19	UNKNOWN
NZ_CP010226	Escherichia coli strain S1	UNKNOWN
NZ_CP010228	Escherichia coli strain S3	UNKNOWN
NZ_CP010229	Escherichia coli strain S10	UNKNOWN
NZ_CP010230	Escherichia coli strain S21	UNKNOWN
NZ_CP010231	Escherichia coli strain S30	UNKNOWN
NZ_CP010235	Escherichia coli strain S40	UNKNOWN
NZ_CP010236	Escherichia coli strain S42	UNKNOWN
NZ_CP010237	Escherichia coli strain S43	UNKNOWN
NZ_CP010238	Escherichia coli strain S50	UNKNOWN
NZ_CP010240	Escherichia coli strain C7	UNKNOWN
NZ_CP010242	Escherichia coli strain S56	UNKNOWN
NZ_CP010304	Escherichia coli O157:H7 str. SS52	EHEC
NZ_CP010315	Escherichia coli 789	APEC
NZ_CP010344	Escherichia coli ECC-1470	UNKNOWN
NZ_CP010438	Escherichia coli K-12 strain ER3454	NONPATHOGENIC
NZ_CP010439	Escherichia coli K-12 strain ER3440	NONPATHOGENIC
NZ_CP010440	Escherichia coli K-12 strain ER3476	NONPATHOGENIC
NZ_CP010441	Escherichia coli K-12 strain ER3445	NONPATHOGENIC
NZ_CP010442	Escherichia coli K-12 strain ER3466	NONPATHOGENIC

NZ_CP010443	Escherichia coli K-12 strain ER3446	NONPATHOGENIC
NZ_CP010444	Escherichia coli K-12 strain ER3475	NONPATHOGENIC
NZ_CP010445	Escherichia coli K-12 strain ER3435	NONPATHOGENIC
NZ_CP010585	Escherichia coli C41(DE3)	NONPATHOGENIC
NZ_CP010816	Escherichia coli BL21 (TaKa-Ra)	NONPATHOGENIC
NZ_CP010876	Escherichia coli strain MNCRE44	UPEC
NZ_CP011018	Escherichia coli strain CI5	EPEC
NZ_CP011061	Escherichia coli str. Sanji	ENVIROMENTAL
NZ_CP011134	Escherichia coli VR50	NONPATHOGENIC
NZ_CP011320	Escherichia coli strain SQ37	NONPATHOGENIC
NZ_CP011321	Escherichia coli strain SQ88	NONPATHOGENIC
NZ_CP011324	Escherichia coli strain SQ2203	NONPATHOGENIC
NZ_CP011331	Escherichia coli O104:H4 str. C227-11	EAEC
NZ_CP011343	Escherichia coli K-12	NONPATHOGENIC
NZ_CP011416	Escherichia coli strain CFSAN029787	EIEC
NZ_CP011495	Escherichia coli strain NCM3722	NONPATHOGENIC
NZ_CP011938	Escherichia coli strain C43(DE3)	NONPATHOGENIC
NZ_CP012125	Escherichia coli strain DH1Ec095	UNKNOWN
NZ_CP012126	Escherichia coli strain DH1Ec104	UNKNOWN
NZ_CP012127	Escherichia coli strain DH1Ec169	UNKNOWN
NZ_CP012625	Escherichia coli strain SF-468	EXPEC
NZ_CP012631	Escherichia coli strain SF-173	EXPEC
NZ_CP012633	Escherichia coli strain SF-166	EXPEC
NZ_CP012635	Escherichia coli strain SF-088	EXPEC
NZ_CP012693	Escherichia coli strain FORC_028	EHEC
NZ_CP012802	Escherichia coli O157:H7 strain WS4202	EHEC
NZ_CP012868	Escherichia coli str. K-12 sub-str. MG1655	NONPATHOGENIC
NZ_CP012869	Escherichia coli strain K-12 substrain MG1655_TMP32XR1	NONPATHOGENIC
NZ_CP012870	Escherichia coli strain K-12 substrain MG1655_TMP32XR2	NONPATHOGENIC
NZ_CP013025	Escherichia coli strain 2009C-	STEC

	3133	
NZ_CP013029	Escherichia coli strain 2012C-4227	STEC
NZ_CP013031	Escherichia coli strain H1827/12	EXPEC
NZ_CP013112	Escherichia coli strain YD786	UPEC
NZ_CP013190	Escherichia coli strain FORC_031	ETEC
NZ_CP013253	Escherichia coli strain CQSW20	UNKNOWN
NZ_CP013483	Escherichia coli strain Y5	UPEC
NZ_CP013658	Escherichia coli strain uk_P46212	UPEC
NZ_CP013663	Escherichia coli strain GB089	UNKNOWN
NZ_CP013831	Escherichia coli strain CD306	EXPEC
NZ_CP013835	Escherichia coli strain JJ2434	EXPEC
NZ_CP014197	Escherichia coli strain MRE600	ENVIROMENTAL
NZ_CP014225	Escherichia coli str. K-12 sub-str. MG1655	NONPATHOGENIC
NZ_CP014269	Escherichia coli B strain C3029	NONPATHOGENIC
NZ_CP014270	Escherichia coli K-12 strain K-12 DHB4	NONPATHOGENIC
NZ_CP014272	Escherichia coli K-12 strain K-12 C3026	NONPATHOGENIC
NZ_CP014316	Escherichia coli JJ1887	UPEC
NZ_CP014348	Escherichia coli str. K-12 sub-str. MG1655	NONPATHOGENIC
NZ_CP014488	Escherichia coli strain G749	EXPEC
NZ_CP014492	Escherichia coli strain MVAST0167	EXPEC
NZ_CP014495	Escherichia coli strain SaT040	EXPEC
NZ_CP014497	Escherichia coli strain ZH193	EXPEC
NZ_CP014522	Escherichia coli strain ZH063	EXPEC
NZ_CP014583	Escherichia coli strain CFSAN004176	STEC
NZ_CP014667	Escherichia coli strain ECO-NIH2	UNKNOWN
NZ_CP014670	Escherichia coli strain CFSAN004177	STEC
NZ_CP015020	Escherichia coli strain 28RC1	STEC
NZ_CP015069	Escherichia coli strain Ecol_743	UNKNOWN
NZ_CP015076	Escherichia coli strain Ecol_448	UNKNOWN
NZ_CP015138	Escherichia coli strain Ecol_732	UNKNOWN
NZ_CP015159	Escherichia coli strain Eco889	UNKNOWN

NZ_CP015228	Escherichia coli strain 09-00049	UNKNOWN
NZ_CP015229	Escherichia coli strain 06-00048	UNKNOWN
NZ_CP015240	Escherichia coli strain 2011C-3911	STEC
NZ_CP015241	Escherichia coli strain 2013C-4465	STEC
NZ_CP015831	Escherichia coli O157 strain 644-PT8	STEC
NZ_CP015832	Escherichia coli O157 strain 180-PT54	STEC
NZ_CP015834	Escherichia coli strain MS6198	UPEC
NZ_CP015842	Escherichia coli O157:H7 strain FRIK2533	STEC
NZ_CP015912	Escherichia coli strain 210205630	UNKNOWN
NZ_CP015995	Escherichia coli strain S51	UNKNOWN
NZ_CP016007	Escherichia coli strain NGF1	NONPATHOGENIC
NZ_CP016018	Escherichia coli strain ER1821R	NONPATHOGENIC
NZ_CP016358	Escherichia coli strain K-15KW01	UPEC
NZ_CP016404	Escherichia coli strain 210221272	UNKNOWN
NZ_CP016497	Escherichia coli strain UPEC 26-1	UPEC
NZ_CP016546	Escherichia coli strain O177:H21	UNKNOWN
NZ_CP016625	Escherichia coli O157:H7 strain FRIK944	STEC
NZ_CP016628	Escherichia coli strain FORC_041	EHEC
NZ_CP017100	Escherichia coli strain K-12 NEB 5-alpha	NONPATHOGENIC
NZ_CP017249	Escherichia coli strain NADC 5570/86-24/6565	EHEC
NZ_CP017251	Escherichia coli strain NADC 5570/86-24/6564	EHEC
NZ_CP017434	Escherichia coli O157:H7 strain 1130	EHEC
NZ_CP017436	Escherichia coli O157:H7 strain 2149	EHEC
NZ_CP017438	Escherichia coli O157:H7 strain 2159	EHEC
NZ_CP017440	Escherichia coli O157:H7 strain 3384	EHEC
NZ_CP017442	Escherichia coli O157:H7 strain 4276	EHEC

NZ_CP017444	Escherichia coli O157:H7 strain 8368	EHEC
NZ_CP017446	Escherichia coli O157:H7 strain 9234	EHEC
NZ_CP017631	Escherichia coli strain SLK172	EPEC
NZ_CP017669	Escherichia coli strain PA20	STEC
NZ_CP017844	Escherichia coli strain FMU073332	ETEC
NZ_CP018103	Escherichia coli strain MRSN352231	UNKNOWN
NZ_CP018109	Escherichia coli strain MRSN346595	UNKNOWN
NZ_CP018115	Escherichia coli strain MRSN346638	UNKNOWN
NZ_CP018121	Escherichia coli strain MRSN346355	UNKNOWN
NZ_CP018206	Escherichia coli strain MRSN346647	UNKNOWN
NZ_CP018237	Escherichia coli strain 155	EHEC
NZ_CP018239	Escherichia coli strain 272	EHEC
NZ_CP018241	Escherichia coli strain 319	EHEC
NZ_CP018243	Escherichia coli strain 350	EHEC
NZ_CP018245	Escherichia coli strain 472	EHEC
NZ_CP018247	Escherichia coli strain 7784	EHEC
NZ_CP018250	Escherichia coli strain 10671	EHEC
NZ_CP018252	Escherichia coli strain 9000	EHEC
NZ_CP018801	Escherichia coli strain tolC-	UNKNOWN
NZ_CP018948	Escherichia coli strain Ecol_224	UNKNOWN
NZ_CP018953	Escherichia coli strain Ecol_276	UNKNOWN
NZ_CP018957	Escherichia coli strain Ecol_316	UNKNOWN
NZ_CP018962	Escherichia coli strain Ecol_422	UNKNOWN
NZ_CP018965	Escherichia coli strain Ecol_517	UNKNOWN
NZ_CP018970	Escherichia coli strain Ecol_542	UNKNOWN
NZ_CP018976	Escherichia coli strain Ecol_545	UNKNOWN
NZ_CP018979	Escherichia coli strain Ecol_656	UNKNOWN
NZ_CP018983	Escherichia coli strain Ecol_867	UNKNOWN
NZ_CP018991	Escherichia coli strain Ecol_AZ146	UNKNOWN
NZ_CP018995	Escherichia coli strain Ecol_AZ147	UNKNOWN

NZ_CP019000	Escherichia coli strain Ecol_AZ153	UNKNOWN
NZ_CP019005	Escherichia coli strain Ecol_AZ155	UNKNOWN
NZ_CP019008	Escherichia coli strain Ecol_AZ159	UNKNOWN
NZ_CP019012	Escherichia coli strain Ecol_AZ161	UNKNOWN
NZ_CP019015	Escherichia coli strain Ecol_AZ162	UNKNOWN
NZ_CP019020	Escherichia coli strain Ecol_244	UNKNOWN
NZ_CP019213	Escherichia coli strain WCHEC1613	UNKNOWN
NZ_CP019629	Escherichia coli strain MGY	NONPATHOGENIC
NZ_CP019777	Escherichia coli NU14 strain NU14	UPEC
NZ_CP019778	Escherichia coli strain NCTC86	NONPATHOGENIC
NZ_CP019903	Escherichia coli strain MDR_56	UNKNOWN
NZ_HF572917	Escherichia coli strain HU-SEC2011	EHEC
NZ_HG738867	Escherichia coli str. K-12 sub-str. MC4100	NONPATHOGENIC
NZ_LM993812	Escherichia coli strain K-12 substr. HMS174	NONPATHOGENIC
NZ_LM995446	Escherichia coli strain K-12 substr. RV308	NONPATHOGENIC
NZ_LN832404	Escherichia coli K-12	NONPATHOGENIC
NZ_LT601384	Escherichia coli strain NCTC86EC	NONPATHOGENIC
NZ_LT615377	Escherichia coli strain 106	UNKNOWN
NZ_LT615378	Escherichia coli strain 103	UNKNOWN
NZ_CP021202	Escherichia coli strain Z1002	UNKNOWN
NZ_CP021207	Escherichia coli strain strain Z247	UNKNOWN
NZ_CP011915	Escherichia coli strain PSUO2	APEC
NZ_CP012112	Escherichia coli strain PSUO78	APEC
NZ_CP020835	Escherichia coli strain CFSAN051542	UNKNOWN
NZ_CP023346	Escherichia coli strain ETEC-2265	ETEC
NZ_CP023349	Escherichia coli strain ETEC-2264	ETEC
NZ_CP023960	Escherichia coli strain FDAARGOS_448 strain Not applicable	UPEC

NZ_CP023899	Escherichia coli strain FDAARGOS_433 strain Not applicable	UNKNOWN
NZ_CP024228	Escherichia coli O25:NM strain 2014EL-1343-2	STEC
NZ_CP022689	Escherichia coli strain CDC#03-98	STEC
NZ_CP024260	Escherichia coli strain F5656C1	STEC
NZ_CP024273	Escherichia coli strain F9792	STEC
NZ_CP021175	Escherichia coli strain 5CRE51	UPEC
NZ_CP024289	Escherichia coli O178:H19 strain 2012C-4431	STEC
NZ_CP020092	Escherichia coli strain 13E0725	STEC
NZ_CP020106	Escherichia coli strain 13E0780	STEC
NZ_CP020107	Escherichia coli strain 13E0767	EPEC
NZ_CP014314	Escherichia coli strain JE-ONG-1266	UNKNOWN
NZ_CP015846	Escherichia coli O157:H7 strain FRIK2069	UNKNOWN
NZ_CP015843	Escherichia coli O157:H7 strain FRIK2455	UNKNOWN
NZ_CP022050	Escherichia coli O157 strain FDAARGOS_293	UNKNOWN
NZ_CP022164	Escherichia coli strain M160133	UPEC
NZ_CP024269	Escherichia coli strain F8111-1SC3	STEC
NZ_CP016828	Escherichia coli strain FORC_043	EHEC
NZ_CP023142	Escherichia coli strain CFSAN061770	ENVIROMENTAL
NZ_CP013662	Escherichia coli strain 08-00022	STEC
NZ_CP015074	Escherichia coli strain Ecol_745	UNKNOWN
NZ_CP016755	Escherichia coli strain FORC_044	UNKNOWN
NZ_CP021844	Escherichia coli strain EC1515	UPEC
NZ_CP021840	Escherichia coli strain EC974	UPEC
NZ_CP021179	Escherichia coli strain 81009	EXPEC
NZ_CP016182	Escherichia coli strain EC590	UPEC
NZ_CP023531	Escherichia coli strain FDAARGOS_401 strain Not applicable	UNKNOWN
NZ_CP022086	Escherichia coli O104:H4	EHEC

	strain FDAARGOS_348	
NZ_CP023535	Escherichia coli strain FDAARGOS_403 strain Not applicable	UNKNOWN
NZ_CP015244	Escherichia coli O91 str. RM7190	STEC
NZ_CP025048	Escherichia coli strain SC516	APEC
NZ_CP011342	Escherichia coli K-12 GM4792	NONPATHOGENIC
NZ_CP020543	Escherichia coli C strain C	NONPATHOGENIC
NZ_CP024997	Escherichia coli strain S17-13	UNKNOWN
NZ_CP014268	Escherichia coli B strain C2566	NONPATHOGENIC
NZ_CP011113	Escherichia coli strain RR1	NONPATHOGENIC
NZ_CP019944	Escherichia coli strain DSM 103246	UNKNOWN
NZ_CP020368	Escherichia coli strain BLR(DE3)	NONPATHOGENIC
NZ_CP018323	Escherichia coli strain KSC9	UNKNOWN
NZ_CP018840	Escherichia coli strain KSC64	UNKNOWN
NZ_CP019560	Escherichia coli strain KSC1031	UNKNOWN
NZ_CP019558	Escherichia coli strain KSC207	UNKNOWN
NZ_CP025036	Escherichia coli strain S17-20	UNKNOWN
NZ_CP022393	Escherichia coli strain E62	UNKNOWN
NZ_CP012379	Escherichia coli strain PAR	UNKNOWN
NZ_CP012380	Escherichia coli strain WAT	UNKNOWN
NZ_CP013952	Escherichia coli strain HST04	UNKNOWN
NZ_CP020933	Escherichia coli strain HB-Coli0	APEC
NZ_CP014641	Escherichia coli strain BLK9	NONPATHOGENIC
NZ_CP014642	Escherichia coli strain BLK16	NONPATHOGENIC
NZ_CP023815	Escherichia coli strain IMT16316	UNKNOWN
NZ_CP022154	Escherichia coli strain ABWA45	ENVIROMENTAL
NZ_CP023388	Escherichia coli strain 1105	UPEC
NZ_CP023353	Escherichia coli strain 746	UPEC
NZ_CP023386	Escherichia coli strain 1190	UPEC
NZ_CP023383	Escherichia coli strain 1223	UPEC
NZ_CP023377	Escherichia coli strain 127	UPEC
NZ_CP023371	Escherichia coli strain 1283	UPEC
NZ_CP023366	Escherichia coli strain 1428	UPEC
NZ_CP023364	Escherichia coli strain 144	UPEC
NZ_CP023359	Escherichia coli strain 1943	UPEC
NZ_CP023357	Escherichia coli strain 317	UPEC
NZ_CP021288	Escherichia coli strain PA45B	UPEC
NZ_CP024720	Escherichia coli strain NQ3	UNKNOWN
NZ_CP024717	Escherichia coli strain LS4	UNKNOWN

NZ_CP020058	Escherichia coli strain AR_0061	UNKNOWN
NZ_CP021879	Escherichia coli strain AR_0137	UNKNOWN
NZ_CP021722	Escherichia coli strain AR_0128	UNKNOWN
NZ_CP020048	Escherichia coli strain AR_0118	UNKNOWN
NZ_CP021935	Escherichia coli strain AR_0055	UNKNOWN
NZ_CP021535	Escherichia coli strain AR_0149	UNKNOWN
NZ_CP020055	Escherichia coli strain AR_0069	UNKNOWN
NZ_CP021736	Escherichia coli strain AR_0150	UNKNOWN
NZ_CP021691	Escherichia coli strain AR_0151	UNKNOWN
NZ_CP021532	Escherichia coli strain AR_0149	UNKNOWN
NZ_CP021683	Escherichia coli strain AR_0162	UNKNOWN
NZ_CP021689	Escherichia coli strain AR_0058	UNKNOWN
NZ_CP020116	Escherichia coli strain AR_0104	UNKNOWN
NZ_CP021732	Escherichia coli strain AR_0114	UNKNOWN
	Escherichia coli str. K-12 sub-str. MG1655	
NZ_AP017620	Escherichia coli strain MRY15-131	UNKNOWN
NZ_AP017610	Escherichia coli strain 20Ec-P-124	UNKNOWN
NZ_AP017617	Escherichia coli strain MRY15-117	UNKNOWN
NZ_CP021335	Escherichia coli strain 95JB1	EHEC
NZ_CP021339	Escherichia coli strain 95NR1	EHEC
NZ_CP021454	Escherichia coli strain H105	UNKNOWN
NZ_CP024223	Escherichia coli O169:H41 strain 2014EL-1345-2	ETEC
NZ_CP024263	Escherichia coli O169:H41 strain F6326-C1	ETEC
NZ_CP024232	Escherichia coli O6:H16 strain 2014EL-1346-6	ETEC
NZ_CP024275	Escherichia coli O6:H16 strain M9682-C1	ETEC
NZ_CP024240	Escherichia coli O114:H49 strain 90-9280	ETEC

NZ_CP024243	Escherichia coli O128:H27 strain 90-9281	ETEC
NZ_CP024239	Escherichia coli O15:H11 strain 90-9272	ETEC
NZ_CP024257	Escherichia coli O25:H16 strain F5505-C1	ETEC
NZ_CP024245	Escherichia coli O27:H7 strain B4103-1	ETEC
NZ_CP024256	Escherichia coli strain ATCC 43886	STEC
NZ_CP024278	Escherichia coli strain ATCC 43896	ETEC
NZ_CP006830	Escherichia coli APEC O18 strain APEC O18	APEC
NZ_CP006834	Escherichia coli APEC O2-211 strain APEC O2-211	APEC
NZ_CP007275	Escherichia coli strain NMEC O18	EXPEC
NZ_CP024090	Escherichia coli strain WG5	NONPATHOGENIC
NZ_LT838196	Escherichia coli strain WI1 isolate	UNKNOWN
NZ_LT906474	Escherichia coli strain NCTC122	UNKNOWN
NZ_CP006632	Escherichia coli PCN033	EXPEC
NZ_CP010371	Escherichia coli 6409	UNKNOWN
NZ_CP011124	Escherichia coli USML2	UNKNOWN
NZ_CP015023	Escherichia coli strain SRCC 1675	STEC
NZ_CP019029	Escherichia coli strain Ecol_881	UNKNOWN
NZ_CP022912	Escherichia coli 2011EL-1370-2	ETEC
NZ_CP023541	Escherichia coli O104:H21 str. CFSAN002236	UNKNOWN
NZ_HG941718	Escherichia coli O25b:H4-ST131 str. EC958	UPEC
NZ_CP017220	Escherichia coli strain FAM21845	UNKNOWN
NZ_CP024830	Escherichia coli strain CREC-532	UPEC
NZ_CP024851	Escherichia coli strain AR_0006	UNKNOWN
NZ_CP024859	Escherichia coli strain AR_0014	UNKNOWN
NZ_CP024815	Escherichia coli strain CREC-629	UPEC
NZ_CP024826	Escherichia coli strain CREC-544	UNKNOWN
NZ_CP024821	Escherichia coli strain CREC-	UNKNOWN

	591	
NZ_CP024862	Escherichia coli strain AR_0015	UNKNOWN
NZ_CP024855	Escherichia coli strain AR_0011	UNKNOWN
NZ_CP024889	Escherichia coli strain AR_0019	UNKNOWN
NZ_CP024886	Escherichia coli strain AR_0017	UNKNOWN
NZ_CP024801	Escherichia coli strain AMA1167	UNKNOWN
NC_003197	Salmonella enterica enterica	N/A
NC_011740	Escherichia fergusonii ATCC 35469	N/A

*Os patótipos foram extraídos do NCBI e/ou da literatura. Para certas linhagens derivadas o patótipo foi inferido a partir do patótipo da linhagem que as deram origem. N/A: Não aplicado.

APÊNDICE C: RELAÇÃO DE GENES NÃO PRESENTES EM CADA VARIANTE DA *E. COLI* BH100

Quadro S2 Relação de CDSs ausentes na *E. coli* BH100 subtr. MG2014

Region*	Uniprot Nomenclature	Uniprot Access	GSO terms	Category
A1	Histidine biosynthesis bi-functional protein HisIE	A0A0W3CTR7_ESCFE	-	BioProcess
B1	Uncharacterized protein	A0A029HS61_ECOLX	-	1
C1	Alcohol dehydrogenase in ethanolamine utilization	A0A0K5L8F1_ECOLX	general metabo-lism; nutrient use	4
C1	Ethanolamine utilization protein EutJ	A0A192ERG8_ECOLX	metabolosome formation	3
C1	Aldehyde dehydrogenase, ethanolamine utilization protein	T6H873_ECOLX	nutrient use	4
C1	Ethanolamine utilization protein eutN	L2VQ33_9ESCH	nutrient use	3
C1	BMC domain protein	D7ZF33_ECOLX	metabolosome formation	3
D1	Uncharacterized protein	A0A029IBS4_ECOLX	-	1
E1	Uncharacterized protein	W9A084_ECOLX	-	1
F1	Pectinesterase	A0A192EN52_ECOLX	-	BioProcess
F1	Kinase inhibitor	A0A192ELM1_ECOLX	general metabo-lism	4
F1	IS605 OrfB family trans-posase	A0A192ENI5_ECOLX	transposase	3
F1	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	A0A234IKF5_ECOLX	-	BioProcess
F1	Biotin synthase	A0A192ELS4_ECOLX	-	BioProcess
F1	8-amino-7-oxononanoate synthase	A0A029IA83_ECOLX	-	BioProcess

F1	Malonyl-[acyl-carrier protein] O-methyltransferase	A0A192ELM6_ECOLX	-	BioProcess
F1	ATP-dependent dethiobiotin synthetase BioD	A0A192ENJ0_ECOLX	-	BioProcess
F1	UvrABC system protein B	A0A029I972_ECOLX	-	BioProcess
F1	Putative gluconeogenesis factor	A0A192ELS9_ECOLX	-	BioProcess
F1	GTP 3',8-cyclase	A0A0E0V3H3_ECOLX	-	BioProcess
F1	Molybdenum cofactor biosynthesis protein B	A0A0E2L7Q7_ECOU3	-	BioProcess
F1	Cyclic pyranopterin monophosphate synthase	A0A0E2LM11_ECOU3	-	BioProcess
F1	B7MGP2 Molybdopterin synthase sulfur carrier subunit	B7MGP2_ECO45	-	BioProcess
F1	Molybdopterin synthase, large subunit	B7MGP3_ECO45	-	BioProcess
F1	Uncharacterized protein	B7MGP4_ECO45	transporter	2
F1	Inner membrane protein	T6HRD1_ECOLX	transport	4
F1	Uncharacterized protein	A0A0H2V5G9_ECOL6	antimicrobial peptide resistance	2
F1	T6IEW8 Cardiolipin synthase B	T6IEW8_ECOLX	-	BioProcess
F1	Endonuclease	S1PIP7_ECOLX	DNA modification	4
F1	Inner membrane protein ybhQ	I6BZF9_SHIFL	unknown	4
F1	I6BZG0 Transport permease protein	I6BZG0_SHIFL	-	BioProcess
F1	Inner membrane transport permease ybhS	A0A029I947_ECOLX	-	BioProcess
F1	ABC transporter ATP-	T6IEX5_ECOLX	transport	4

	binding protein			
F1	UPF0194 membrane protein YbhG	A0A0W3CXN9_ESCFE	transport	4
F1	Putative DNA-binding transcriptional regulator	B7MGQ4_ECO45	-	BioProcess
F1	S0ZRX0 ATP-dependent RNA helicase RhlE	S0ZRX0_ECOLX	-	BioProcess
F1	Swarming motility protein YbiA	A0A243V5X9_ECOLX	intercellular communication	3
F1	ATP-dependent DNA helicase	B7MGQ7_ECO45	-	BioProcess
F1	YbiB	A0A192CKD4_ECOLX	nucleotide metabolism	4
F1	Dehydrogenase	A0A192ELU8_ECOLX	general metabolism	4
F1	Uncharacterized protein	V0VBC8_ECOLX	-	1
F1	Uncharacterized protein	X0W3S4_9ZZZZ	-	1
F1	Uncharacterized protein	A0A234IK59_ECOLX	unknown	2
F1	PKHD-type hydroxylase YbiX	A0A234I5J9_ECOLX	antimicrobial peptide resistance	4

*Regiões idênticas indicam que os genes ausentes nessa linhagem se encontram próximos entre si em outras.

Quadro S3 Relação de CDSs ausentes na *E. coli* BH100 subtr. MG2017

Region*	Uniprot Nomenclature	Uniprot Access	G.S.O terms	Category
A2	Uncharacterized protein	A0A234FE73_ECOLX	-	1
A2	Uncharacterized protein	A0A234HQJ4_ECOLX	-	1
A2	Evolved beta-galactosidase subunit beta	A0A0E2L2I5_ECOU3	intercellular communication	3
A2	Putative oxidoreductase	B6I450_ECOSE	transport	4
A2	Alpha-glucosidase	A0A192EUD2_ECOLX	-	BioProcess

*Regiões idênticas indicam que os genes ausentes nessa linhagem se encontram próximos entre si em outras.

Quadro S4 Relação de CDSs ausentes na *E. coli* BH100L subtr. MG2017

Region*	Uniprot Nomenclature	Uniprot Access	GSO terms	Category
A3	Inner membrane protein YcfZ	A0A192EP21	unknown	4
A3	Inner membrane protein ymfA	A0A029I5G0	unknown	4
A3	Putrescine-binding periplasmic protein	B7MJB0	-	BioProcess
A3	Spermidine/putrescine transport system permease potC	S0YQF4	-	BioProcess
A3	Spermidine/putrescine ABC transporter membrane protein	A0A0E0Y0M9	-	BioProcess
A3	Spermidine/putrescine import ATP-binding protein PotA	H4J718	transport; gene expression	4
A3	Uncharacterized protein	I2RNV4		1
A3	Peptidase T	A0A1Z3V3B5		BioProcess
A3	Cupin family protein	A0A0E1SVM4	gene expression	3
A3	Sensor protein phoQ	T6HJ99	signal transduction	4
A3	DNA-binding transcriptional regulator PhoP	A0A168HXG8	-	BioProcess
A3	Integrase	A0A192EMN8	-	BioProcess
A3	Uncharacterized protein	D6J991	-	BioProcess
B3	Iron transporter, periplasmic-binding protein SitA	S1PQ23	-	BioProcess
B3	Uncharacterized protein	A0A177NVK5	-	1
C3	Uncharacterized protein	A0A0E0U817	-	BioProcess
D3	Protein rof	A0A029J3E1	gene expression	3
D3	UPF0253 protein YaeP	A0A0A0F5B8	unknown	3
D3	Uncharacterized protein	A0A029ID31	-	1
D3	Peptidyl-tRNA hydrolase ArfB	A0A192EKP3	gene expression	4
D3	Lipoprotein NlpE	A0A0E2LA46	intercellular communication	3
E3	Uncharacterized protein	B7MF56	-	1
E3	Uncharacterized protein	U9XGY1	unkown	2

E3	Dihydropyrimidine dehydrogenase	A0A0K6AE33	nucleotide metabolism	4
E3	Dihydropyrimidine dehydrogenase subunit B	A0A0W3CTP2	nucleotide metabolism	4
F3	Class II glutamine amidotransferase	V0USP9	-	BioProcess
F3	ErfK/YbiS/YcfS/YnhG superfamily protein	F4UYH3	cell wall formation	4
F3	Uncharacterized protein	B7MC78	-	BioProcess
F3	Putative exported hydrolase	B7MC77	general metabolism	4
F3	Uncharacterized protein	A0A029IBS4	-	1
F3	Uncharacterized protein	F5NZ10	unkown	2
F3	Carbamoyltransferase HypF	A0A192EU22	-	BioProcess
F3	Formate dehydrogenase-H, [4Fe-4S] ferredoxin subunit	B7MKI2	electron transfer	3
F3	Uncharacterized protein	V0VLQ1	-	1
F3	AscBF operon repressor	A0A1X3K504	-	BioProcess
F3	PTS cellobiose transporter subunit IIBC	A0A192ESB8	-	BioProcess
G3	Mannitol-1-phosphate 5-dehydrogenase	A0A1X3J6L9	general metabolism	3
H3	Phosphonate transport ATP-binding protein PhnK	L3PSW3	nutrient use; transport	4
H3	Alpha-D-ribose 1-methylphosphonate 5-phosphate C-P lyase	L3PTU0	-	BioProcess
H3	Carbon-phosphorus lyase complex subunit	A0A0E0VDK0	-	BioProcess
H3	Phosphonate C-P lyase system protein PhnH	V8KNQ8	-	BioProcess
H3	Alpha-D-ribose 1-methylphosphonate 5-triphosphate synthase subunit PhnG	A0A192EXE8	-	BioProcess

H3	Phosphonate metabolism transcriptional regulator PhnF	A0A029IHW0	-	BioProcess
H3	Phosphonate ABC transporter, permease PhnE	A0A234EK20	nutrient use; transport	3
H3	Phosphonate ABC transporter substrate-binding protein	A0A192EXY0	-	BioProcess
H3	Phosphonates import ATP-binding protein PhnC	S1EHA1	nutrient use	4
H3	Uncharacterized protein	D6J583	nutrient use	2
H3	Putative alkylphosphonate utilization operon protein PhnA	E1J3C6	nutrient use	3
H3	Clamp-binding protein CrfC	A0A192EXF3	DNA segregation	4
H3	Uncharacterized protein	A0A192EWK0	-	1
H3	Proline/glycine betaine transporter	A0A170YLW8	transport	4
H3	PmrR protein	A0A1E2VGX6	signal transduction	4
H3	Sensor protein BasS	S0YNT3	signal transduction	4
H3	DNA-binding response regulator BasR	D3QKB9	-	BioProcess
H3	Hydrolase	A0A192EXF7	-	BioProcess
H3	Arginine/agmatine antiporter	S1E544	transport	4
H3	Transcriptional regulator	V8KPJ5	-	BioProcess
H3	Biodegradative arginine decarboxylase (ADC)	A0A1X3KME9	-	BioProcess
H3	Melibiose operon regulatory protein	V0VGR2	-	BioProcess
H3	Alpha-galactosidase	A0A029HC41	-	BioProcess
H3	Inner membrane protein YjdF	A0A192EXZ6	unknown	4
H3	Fumarate hydratase class I	A0A0W3D0X3	-	BioProcess
H3	Anaerobic C4-dicarboxylate transporter	A0A0E2KY63	transport	4

H3	Transcriptional regulatory protein	H4USH5	-	BioProcess
H3	Sensory box histidine kinase	H4J0Q0	-	BioProcess
H3	Uncharacterized protein	S0UY16	-	1
H3	Uncharacterized protein	W1YF16	general metabo-lism	2
H3	Uncharacterized protein	B7MK01	unkown	2
H3	Uncharacterized protein	E9XRQ7	toxin-antitoxin system	2
H3	Peptide permease	A0A0W3D0Y6	-	BioProcess
H3	Lysine decarboxylase, inducible	T6GRJ8	-	BioProcess
H3	Transporter, basic amino acid/polyamine antiporter (APA) family	E1JCI2	transport	4
H3	DNA-binding transcriptional activa-tor	B7MK06	-	BioProcess
H3	Uncharacterized protein	A0A125WV58	-	1
H3	Prophage CP4-57 regulatory family protein	A0A070UKH5	phage-related	3
H3	Uncharacterized protein	A0A125WVC0	-	1
H3	Uncharacterized protein	A0A1Q6AF93	-	1
H3	Uncharacterized protein	A0A125WV32	-	1
H3	Uncharacterized protein	A0A125WV30	-	1
H3	Uncharacterized protein	A0A070UKH1	-	1
H3	Uncharacterized protein	A0A155P6M7	unkown	2
H3	Uncharacterized protein	A0A0K3PB65	unkown	2
H3	Uncharacterized protein	A0A125WVB7	-	1
H3	Uncharacterized protein	A0A070UX86	-	1
H3	Uncharacterized protein	L2UZY3	-	1
H3	Uncharacterized protein	A0A205K7V9	-	1
H3	tRNA_anti-like family protein	A0A070ULG1	unknown	3
H3	DNA-directed RNA polymerase subunit beta	A0A209N5N7	gene expression	4

H3	Uncharacterized protein	W8UI25	-	1
H3	Phage protein	W8VFT0	phage-related	3
H3	Uncharacterized protein	A0A070UHF3	-	1
H3	Uncharacterized protein	W8UFM6	-	1
H3	MobA/MobL family protein	A0A125WUR6	DNA segregation	3
H3	Uncharacterized protein	A0A070UKG1	-	1
H3	Conjugal transfer TraD family protein	A0A125WVD0	-	BioProcess
H3	Uncharacterized protein	C8TCG7	-	1
H3	Modification methylase Eco57IB	A0A070UHE9	DNA methylation	4
H3	Uncharacterized protein	A0A286NU72	-	1
H3	Uncharacterized protein	C8TCH1	-	1
H3	Integrase	A0A250I3M5	phage-related	3
H3	HTH-type transcriptional regulator yjdC	I6B6R7	gene expression	4
H3	Thiol:disulfide interchange protein DsbD	V0VK79	-	BioProcess
H3	Divalent-cation tolerance protein CutA	A0A0W3D0T7	-	BioProcess
H3	Anaerobic C4-dicarboxylate transporter	A0A0E2KWX3	transport	4
H3	Aspartate ammonia-lyase	A0A0W3D0N6		BioProcess
H3	Uncharacterized protein	I6B6R2	phage-related	2
H3	Transporter	A0A0W3D0R6	transport	4
H3	Sugar ABC transporter permease	A0A0W3D0H2	transport	4
H3	Inner membrane ABC transporter permease protein YjfF	A0A1X3M2C5	transport	4
H3	Fructose-1,6-bisphosphatase class 1	A0A0E1T1M8	-	BioProcess
H3	UDP-N-acetylmuramate--L-alanyl-gamma-D-glutamyl-meso-2,6-	S1CFK4	-	BioProcess

	diaminoheptandioate ligase			
H3	UPF0307 protein YjgA	T5Z3L8	ribosome-related	3
H3	Peptidase PmbA	A0A192EXT2	Toxin-antitoxin system	4
H3	Soluble cytochrome b562	W1WVL1	electron transfer	4
H3	Anaerobic ribonucleoside-triphosphate reductase-activating protein	A0A029HC14	anaerobic metabolism	4
H3	Anaerobic ribonucleoside-triphosphate reductase	T6H2Q2	-	BioProcess
H3	Glucohydrolase	A0A192EXT7	-	BioProcess
H3	PTS system trehalose-specific EIIBC component	A0A234HS07	-	BioProcess
H3	HTH-type transcriptional regulator TreR	A0A0G9FR67	-	BioProcess
H3	Magnesium transporter	B7MSX7	transport	4
I3	Uncharacterized protein	V0YXL0	-	1
I3	DNA-binding transcriptional regulator of mntH	B7MGS5	-	BioProcess
I3	Anion transporter	A0A192ELR8	-	BioProcess
I3	Uncharacterized protein	B7MGS7	cell wall formation	2
I3	Uncharacterized protein	A0A0E0TVC3	transport	2
I3	Sugar phosphatase SupH	H4J5D7	sugar metabolism	4
I3	Formate acetyltransferase	A0A192EN99	-	BioProcess
I3	Putative AdoMet-dependent glycyl radical activating enzyme	B7MQS7	anaerobic metabolism	4
I3	Molybdopterin-synthase adenylyltransferase	A0A192EMF2	-	BioProcess
I3	Molybdopterin biosynthesis protein MoeA	A0A192ELX2	-	BioProcess

I3	Isoaspartyl peptidase	A0A192ENA4	aminoacid metabolism	4
I3	Glutathione ABC transporter ATP-binding protein	A0A192ELS8	-	BioProcess
I3	Glutathione ABC transporter substrate-binding protein	A0A192ENP0	-	BioProcess
I3	Glutathione ABC transporter, permease protein GsiC	A0A1X3LVJ1	-	BioProcess
I3	Glutathione ABC transporter permease	A0A0W3CXK9	-	BioProcess
I3	C-di-GMP phosphodiesterase	A0A192ENA9	signal transduction	4
I3	Lipoprotein YliF	A0A0W3CXZ6	signal transduction	4
I3	Ribosomal protein S12 methylthiotransferase RimO	A0A0W3CXS1	-	BioProcess
I3	Putative receptor	I6BZL1	intercellular communication	3
I3	Aldose dehydrogenase	A0A0W3CXK1	general metabolism	4
I3	Transferase	K4VZ93	general metabolism	4
I3	D-alanyl-D-alanine carboxypeptidase dacC	L4VX71	cell wall formation	4
I3	DeoR-like helix-turn-helix domain protein	E3XVB1	-	BioProcess
I3	PAP2 family protein	I2UKE3	general metabolism	4
I3	Multidrug transporter MdfA	S1PIV2	-	BioProcess
I3	Uncharacterized protein	E9XNB4	-	1
I3	Sugar-phosphatase	A0A0W3CXJ4	sugar metabolism	4
I3	Putative transporter	B7MGV2	-	BioProcess

I3	DEOR-type transcriptional regulator	A0A192ENQ6	-	BioProcess
I3	Putative transport protein YbjL	V0UMA4	-	BioProcess
I3	Putative membrane protein	H4KE96	unknown	4
I3	Glutaredoxin, GrxA family	D7XET1	-	BioProcess
I3	Uncharacterized protein	B7MHF4	unkown	2
I3	Nitroreductase A, NADPH-dependent, FMN-dependent	B7MHF5	nutrient use	4
I3	Ribosomal protein S6--L-glutamate ligase	L3QCJ1	-	BioProcess
I3	Uncharacterized protein	I6BZN4	signal transduction	2
I3	Putrescine-binding periplasmic protein	V0UQE2	-	BioProcess
I3	Spermidine/putrescine import ATP-binding protein PotA	T6IAH7	transport; gene expression	4
I3	Putrescine transport system permease PotH	A0A192ENR5	-	BioProcess
I3	Putrescine transport system permease potI	T6HWX5	-	BioProcess
I3	Inner membrane protein ybjO	A0A0E2L853	unknown	4
I3	23S rRNA (uracil(747)-C(5))-methyltransferase RlmC	A0A0E2L7Y4	ribosome-related	4
I3	Arginine ABC transporter, periplasmic arginine-binding protein ArtJ	A0A1X3LVA5	transport	4

*Regiões idênticas indicam que os genes ausentes nessa linhagem se encontram próximos entre si em outras.

Quadro S5 Relação de CDSs ausentes na *E. coli* BH100N subtr. MG2017

Region*	Uniprot Nomenclature	Uniprot Access	GSO terms	Category
A4	WSSV410	Q8QTC0_WSSV	-	BioProcess
B4	Non-ribosomal peptide synthetase	A0A0W3CTY0_ESCFE	Non-ribosomal	4

			synthesis	
C4	Chain length determination protein	A0A0W3D1F8_ESCFE	-	BioProcess
C4	6-phosphogluconate dehydrogenase, decarboxylating	A0A0E0Y0N6_ECO1C	-	BioProcess
C4	Mannose-1-phosphate guanyltransferase	A0A0W3CU10_ESCFE	-	BioProcess
C4	Glycosyl transferase family 1	A0A0W3CTM6_ESCFE	general metabolism	4
C4	UDP-glucose 4-epimerase	A0A0W3CTM2_ESCFE	-	BioProcess
C4	Glycosyl transferase family 1	A0A0W3CTN4_ESCFE	general metabolism	4
C4	Glycosyl transferase family 1	A0A0W3CTM9_ESCFE	general metabolism	4
C4	Glycosyl transferase family A	A0A0W3CTN1_ESCFE	general metabolism	4
C4	Polymerase	A0A0W3CTN7_ESCFE	cell wall formation	4
C4	Polysaccharide biosynthesis protein	A0A0W3CTM8_ESCFE	cell wall formation	4
C4	Regulatory protein GalF	I6BKJ9_SHIFL	-	BioProcess
C4	Colanic acid biosynthesis protein WcaM	S0Z9S4_ECOLX	colanic acid biosynthesis	3
C4	Colanic acid biosynthesis glycosyltransferase WcaL	A0A0W3CTM3_ESCFE	colanic acid biosynthesis	4
C4	Colanic acid biosynthesis pyruvyl transferase WcaK	A0A0W3CTX6_ESCFE	colanic acid biosynthesis	4
D4	Putative yqhC protein	A0A029HJ29_ECOLX	gene expression	3
E4	S1QAJ3 Carbamate kinase	S1QAJ3_ECOLX	-	BioProcess
E4	Deaminase	A0A192EL64_ECOLX	general metabolism	4

E4	Putative chlorophyll synthesis pathway protein BchC	V0V241_ECOLX	signal transduction	4
E4	LysE family translocator	A0A0K3CT78_ECOLX	-	BioProcess
E4	Propionate catabolism operon regulatory protein	A0A192EMH1_ECOLX	-	BioProcess
E4	2-methylisocitrate lyase	A0A192CIZ7_ECOLX	-	BioProcess
E4	D7ZAZ0 Citrate synthase	D7ZAZ0_ECOLX	-	BioProcess
E4	2-methylcitrate dehydratase	A0A192EM51_ECOLX	-	BioProcess
E4	Propionate--CoA ligase	A0A1E3N5S1_9ENTR	-	BioProcess
E4	Cytosine permease	D7ZAY6_ECOLX	transport	4
E4	Cytosine deaminase	T6IDS8_ECOLX	nucleotide metabolism	4
F4	Phosphonate C-P lyase system protein PhnK	D7ZE23_ECOLX	transport; nutrient use	4
F4	L3PTU0 Alpha-D-ribose 1-methylphosphonate 5-phosphate C-P lyase	L3PTU0_ECOLX	-	BioProcess
F4	Protein phnI	T6LKF1_ECOLX	-	BioProcess
F4	Phosphonate C-P lyase system protein PhnH	V0V7Y2_ECOLX	-	BioProcess
F4	Alpha-D-ribose 1-methylphosphonate 5-triphosphate synthase subunit PhnG	A0A192EXE8_ECOLX	-	BioProcess
F4	Putative DNA-binding transcriptional regulator of phosphonate uptake and biodegradation	B7MRA8_ECO81	-	BioProcess
F4	Phosphonate ABC transporter, permease PhnE	A0A234EK20_ECOLX	transporter; nutrient use	3
F4	Phosphonate ABC transporter substrate-binding protein	A0A192EXY0_ECOLX	-	BioProcess

F4	I2RRS1 Phosphonates import ATP-binding protein PhnC	I2RRS1_ECOLX	transport; nutrient use	4
F4	Protein PhnB	D7ZE32_ECOLX	general metabolism; nutrient use	3
G4	Uncharacterized protein	Q1R2N4_ECOUT	-	BioProcess
G4	Na+/H+ antiporter NhaC	A0A192EYM2_ECOLX	transport	4
G4	Na+/H+ antiporter	A0A192EY57_ECOLX	transport	4
H4	Uncharacterized protein	V0VBC8_ECOLX	-	1

*Regiões idênticas indicam que os genes ausentes nessa linhagem se encontram próximos entre si em outras.

APÊNDICE D: *CORE* DE PSEUDOGENES

Quadro S6 Relação de pseudogenes comuns aos cromossomos da *E. coli* BH100 subtr. MG2014, *E. coli* BH100 subtr. MG2017, *E. coli* BH100L subtr. MG2017 e *E. coli* BH100N subtr. MG2017.

Produto	**Causa**
ferric-rhodotorulic acid/ferric-coprogen receptor FhuE	frameshifted
hypothetical protein	frameshifted
Putative transport protein	frameshifted
fructuronate reductase	frameshifted
Hydrolase, inner membrane	frameshifted
acyl CoA:acetate/3-ketoacid CoA transferase	internal stop codon
membrane protein	internal stop codon
Endo-1,4-beta-xylanase A precursor	frameshifted
LysR family transcriptional regulator	frameshifted
enterotoxin	frameshifted
hypothetical protein	frameshifted
methyl-accepting chemotaxis protein	frameshifted
glutathione-regulated potassium-efflux system protein KefB	internal stop codon
multifunctional 2',3'-cyclic-nucleotide 2'-phosphodiesterase/5'-nucleotidase/3'-nucleotidase	internal stop codon
autotransporter outer membrane beta-barrel domain-containing protein	frameshifted
PTS N-acetylgalactosamine transporter subunit IIC	internal stop codon
glutathione S-transferase	frameshifted
Dehydrogenase	frameshifted
formate dehydrogenase-N subunit alpha	internal stop codon
type II toxin-antitoxin system HipA family toxin	frameshifted
Inner membrane protein YcfT	internal stop codon
serine/threonine protein phosphatase	frameshifted
cell division protein	internal stop codon
hypothetical protein	frameshifted

APÊNDICE E: *AVERAGE NUCLEOTIDE IDENTITY* (ANI)

Tabela S7 Similaridade entre cromossomos de algumas linhagens de *E. coli* e enterobactérias próximas, a partir da análise de ANI.

	B14	**B17**	**BL17**	**BN17**	**536**	**CFT036**	**K12**	**042**	**Ferg**	**Sal**
B14		99,96	99,97	99,97	99,78	98,99	96,99	96,90	89,82	80,64
B17			99,96	99,97	99,75	98,96	96,93	96,85	89,77	80,60
BL17				99,98	99,78	98,94	96,91	96,93	89,99	80,65
BN17					99,76	98,97	96,98	97,00	89,76	80,68

B14: *E. coli* BH100 subtr. MG2014; B17: *E. coli* BH100 subtr. MG2017; BL17: *E. coli* BH100L subtr. MG2017; BN17: *E. coli* BH100N subtr. MG2017; 536: *E. coli 536; CFT036: E. coli* CFT073; 042: *E. coli* 042; Ferg: *E. fergusonii; Sal: S.* enterica

APÊNDICE F: GENES ASSOCIADOS A ILHAS DE PATOGENICIDADE[36]

Quadro S8 CDSs presentes em diversas ilhas identificadas pelo IslandPath-DIMOB e compiladas no IslandViewer4 da *E. coli* 536. Potenciais paralogos foram identificados via tBLAStn. utilizando cobertura de 70%, identidade acima de 40% e *e-value* inferior a 10^{-15}.

Quadro S9 CDSs presentes em diversas ilhas identificadas pelo IslandPath-DIMOB e compiladas no IslandViewer4 da *E. coli* CFT073. Potenciais paralogos foram identificados via tBLAStn. utilizando cobertura de 70%, identidade acima de 40% e *e-value* inferior a 10^{-15}.

Quadro S10 CDSs presentes em diversas ilhas identificadas pelo IslandPath-DIMOB e compiladas no IslandViewer4 da *E. coli* BH100 subtr. MG2017. Potenciais paralogos foram identificados via tBLAStn. utilizando cobertura de 70%, identidade acima de 40% e *e-value* inferior a 10^{-15}.

[36] Material em mídia digital.

APENDICE G: GBKFINISHER – A TOOL FOR GBK FILES REFINEMENT

After genome assembly and annotation, usually a lot of effort is taken in order to check for potential pseudogenes as well as in the refinement of qualifiers set to submission. Those processes are usually manually performed in Artemis and are time consuming. Different filters are set in Artemis, which help in the identification of potential issues, as the lacking of stop codons in CDSs, incorrect start codons, duplication and overlap of features, etc. For potential pseudogenes, however, a manual inspection must be perfomed. The user must check for frameshifts corrections by observing for possible indels, the presence of premature stop codons or even gene fragments. The user, then frequently uses BLASTp as a guide for possible correction of those different issues. Based on the need for a more automatic process we introduce GBKFinisher. This tool was developed as an effort to help users to save time and more accurately fix annotation issues. GBKFinisher is based on three major packages: GBKChecker, GBKParser and GBKSolver. GBKChecker attempts to give the basic stats, like qualifiers count and nucleotide composition, as well as a diagnosis of possible issues, like the number of possible frameshifts, fragments and splitted ORFs that could be safely joined, without a stop codon in between. GBKParser was created to rewrite the gbk file, as desired by users. GBKSolver attempts to automatically solve for detected issues by GBKChecker. GBKFinisher allows a very convenient way for users to filter out unwanted qualifiers, as well as to annotate locus_tag differently for CDSs, tRNAs, and rRNAs. It is also possible for users to manually correct potential issues detected in GBKChecker. A log file with all possible issues encountered by GBKChecker is maintained and can be accessed for detailed troubleshooting. Likewise, a gbk file is created with a color scheme that assists users in the curation process. The primary output of GBKFinisher is a user defined Genbank file, that can ultimately be used for guiding the curation process or even for sequin and GenBank submission

APÊNDICE H: SCRIPTS IN-HOUSE[37]

BPGPAfiles2tab.pl

intersect2faa.pl

IStable2gff.pl

multigbk2gff.pl

patricmultifasta_splitter.pl

phastertable2gff.pl

tBLASTnParser.pl

[37] Material em mídia digital