UNIVERSIDADE DE SÃO PAULO

INTERUNIDADES EM BIOINFORMÁTICA

LUIZ THIBÉRIO LIRA DINIZ RANGEL

**O papel de Transferência Horizontal de Genes na
história evolutiva de duas classes de genes em
bactérias**

São Paulo

2017

LUIZ THIBÉRIO LIRA DINIZ RANGEL

**O papel de Transferência Horizontal de Genes na
história evolutiva de duas classes de genes em
bactérias**

Tese apresentada ao Programa de Pós-Graduação
Interunidades em Bioinformática da Universi-
dade de São Paulo para obtenção do título de
Doutor em Ciências.

Área de Concentração: Bioinformática

Orientador: Prof. João Carlos Setubal, PhD

Versão Final

São Paulo

2017

Agradecimentos

Impossível agradecer a todos que tiveram um impacto positivo durante este processo, mas não custa nada tentar! Mesmo aqueles que não estejam citados nesta página saibam que o mais importante é participar, seja ajudando ou atrapalhando!

À minha esposa, Lara, por todo suporte e carinho durante quase todo o processo, até mesmo quando não mereci.

Aos meus pais, por motivos óbvios... Deram-me tudo que alguém poderia precisar e muito mais.

À minha tia Flávia, minha família em SP!

Aos meus irmãos, que mesmo não sabendo têm grande importância em tudo que faço.

Agradecer aos meus amigos que estiveram presente em vários momentos, ajudando em discussões científicas ou sem o menor sentido. Amigos de João Pessoa (Gian, Yves, Olavo, Man, e Beretes). Amigos do 83b (Batata, 01, Cariri, e Binha). Várias pessoas que encontrei pela USP que tiveram grande valor não só durante o doutorado como desde o mestrado, e provavelmente por muito mais tempo (Caio Galado, Vargas, Incronha, Bisson, Pererê, etc).

Ao Professor João Carlos Setubal pela oportunidade e orientação, além de toda paciência (que deve ter sido bastante).

Todos os membros do Setulab, especialmente Rodrigo, Gian, George, e Dino. O convívio quase diário e o ambiente agradável teve fundamental importância no desenvolvimento deste trabalho.

A Ray Dixon e Patrícia dos Santos pela valiosa colaboração ao longo deste doutorado!

I am also very grateful to Peter and Maria Gogarten, who welcomed me in an awesome environment. Everyone I interacted with at UCONN/Storrs: Shannon, Fred, Celine, Ralph, Andrew, Jessica... it is a really long list!

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio financeiro através dos projetos 2012/17196-2 (regular) e 2014/23975-0 (BEPE).

“Study hard what interests you the most in the most undisciplined, irreverent and original manner possible.”

Richard Feynman

Resumo

A Transferência Horizontal de Genes (THG) é um dos principais mecanismos de evolução bacterianos, impactando a evolução de praticamente todas famílias gênicas. Neste trabalho identificamos e avaliamos padrões de possíveis transferências horizontais de genes pertencentes a duas classes funcionais de dois níveis taxonômicos distintos. Caracterizamos a ocorrência e evolução de 45 genes importantes para a fixação de N_2 em 479 genomas de Proteobacteria. Identificamos cinco potenciais aquisições de genes ligados a fixação de N_2 por linhagens de Proteobacteria, as quais foram identificadas consistentemente em 36 dos genes analisados. Realizamos predições de transferências horizontais dos 45 entre todos os 479 genomas de Proteobacteria e identificamos possíveis enriquecimentos de THG, provavelmente ligados à sinais filogenéticos e ecológicos. Desenvolvemos um *pipeline* para identificação semi-automática de efetores do Sistema Secretor do Tipo III em *Aeromonas*, o qual reportou 21 famílias de potenciais efetores presentes em 105 genomas. Entre os 21 efetores identificados 17 foram descritos pela 1^o vez em *Aeromonas*, corroborando a sensibilidade de nosso *pipeline*. Com o auxílio de nossos colaboradores foram realizados testes de citotoxicidade para efetores identificados *in silico*, e apenas quatro não inibiram o crescimento de *Saccharomyces cerevisiae*. Por fim, desenvolvemos um método para agrupamento de famílias gênicas com histórias evolutivas similares que não requer a reconstrução de árvores filogenéticas, aumentando a eficiência computacional. Aplicamos o método desenvolvido para reconstrução da filogenia de *Aeromonas*, o qual mostrou-se compatível com dados presentes na literatura.

Abstract

Horizontal Gene Transfer (HGT) is one of main mechanisms of bacterial evolution, affecting virtually all gene families. In this document we identified and assessed putative horizontal transfers of genes from two functional classes from two distinct taxonomic levels. We characterized the distribution and evolution of 45 genes important to N_2 fixation among 479 Proteobacteria genomes. We identified five potential distinct acquisitions of such genes by Proteobacteria lineages. The distinct origins are consistently identified in 36 out of the 45 assessed genes. We computed possible horizontal transfers of the 45 genes among the 479 Proteobacteria genomes, and we identified enrichments of HGT, likely related to phylogenetic and ecological signals. We developed a semi-automated pipeline to identify effectors of the Type III Secretion System within *Aeromonas*, which reported 21 putative effector families distributed among 105 genomes. Among the 21 likely effectors 17 have been described in *Aeromonas* for the first time, highlighting the sensibility of our pipeline. Our collaborators performed cytotoxicity tests for the 21 likely effector families identified by *in silico* analysis, and only four did not inhibited *Saccharomyces cerevisiae* growth. Lastly, we developed a method to cluster gene families according to shared evolutionary history, without the requirement of phylogenetic tree reconstruction, increasing computational efficiency. We applied this proposed method during *Aeromonas* phylogenetic reconstruction, and it showed up compatible with data available on the literature.

Lista de Figuras

1.1	Mecanismos de transferência horizontal de genes em procariotos	6
1.2	Análise composicional de <i>Magnetococcus marinus</i>	10
1.3	Exemplo de incongruências filogenéticas	11
1.4	Exemplo de grafo	12
3.1	MDS de distâncias entre árvores de housekeeping genes	16
3.2	Distribuição da frequência de genes adjacentes	21
3.3	Comparação de distâncias entre housekeeping genes e entre gene com evolução compartilhada	22
4.1	Filogenia de 87 Proteobacteria capazes de fixar N_2 e matriz de presença/ausência.	43
4.2	Rede de origem compartilhada entre genes ligados a fixação de N_2	44
4.3	Rede de transferência horizontal entre classes taxonômicas de Proteobacteria.	45
4.4	Rede de transferência horizontal de genes ligados a fixação de N_2	46
5.1	Pipeline para identificação de genes ligados ao SST3.	73
5.2	Filogenia de 105 <i>Aeromonas</i> spp. associada a matriz de presença/ausência de 21 efetores.	74
5.3	Teste de citotoxicidade dos 21 efetores de SST3 identificados.	75
5.4	Teste de citotoxicidade sob stress de 8 efetores de SST3 identificados.	76

Glossário

CDF Cumulative Density Function.

FDR *False Discovery Rate*.

GTA Gene Transfer Agents.

HGT Horizontal Gene Transfer.

HMM Hidden Markov Model.

ILS Incomplete Lineage Sorting.

MCL *Markov Clustering Algorithm*.

MNS Minimal NIF Set.

MSA Multiple Sequence Alignment.

PMF Probability Mass Function.

SGX Substituição por Gene Xenólogo.

SST3 Sistema Secretor do Tipo III.

THG Transferência Horizontal de Genes.

XGD Xenologous Gene Displacement.

Sumário

1	Introdução	5
1.1	Tendências e barreiras à transferência horizontal de genes	7
1.2	Métodos de identificação	9
1.2.1	Desvio de padrões composicionais	9
1.2.2	Incongruências filogenéticas	11
1.3	Redes filogenômicas	12
1.3.1	Porque representar a evolução procariótica utilizando redes	13
2	Objetivos	14
2.1	Objetivos gerais	14
3	Agrupamento de famílias gênicas por similaridade evolutiva	15
3.1	Motivação	15
3.2	Descrição do método desenvolvido	17
3.3	Implementação	19
3.4	Verificação	20
4	História evolutiva de genes de fixação de N_2 em Proteobacteria	23
4.1	Contextualização	23
4.2	Discussão	47
5	História evolutiva de efetores do Sistema Secretor do Tipo III em <i>Aeromonas</i> spp.	48
5.1	Contextualização	48

5.2	Discussão	77
6	Conclusões	78
6.1	Direções futuras de pesquisa	79
7	Outros trabalhos	81
7.1	Genoma de Candidatus <i>Liberibacter americanus</i>	81
7.2	Revisão sobre Ca. <i>Liberibacter</i> spp.	83

Capítulo 1

Introdução

A Transferência Horizontal de Genes (THG) é o principal mecanismo de evolução procariótico (Ochman *et al.*, 2000), possuindo um papel fundamental na evolução de praticamente todas as famílias gênicas existentes e apenas poucas famílias apresentam certa resistência (Baptiste *et al.*, 2009; Dagan and Martin, 2007; Gogarten *et al.*, 2002). A THG foi observada pela primeira vez em microrganismos em 1947 (Tatum and Lederberg, 1947), sendo apenas reconhecida como mecanismo adaptativo em 1971 (Went, 1971). Um gene transferido horizontalmente apresenta uma relação de *xenologia* com o gene a partir do qual foi transferido (Gray and Fitch, 1983; Patterson, 1988), sendo a *xenologia* um dos tipos de *homologia*, assim como a *ortologia* e a *paralogia* (Fitch, 1970, 2000). Com a atual compreensão da frequência com que eventos de transferência horizontal ocorrem sabemos que muitas das aparentes duplicações gênicas ocorreram através de THG, onde os aparentes parálogos possuem origens em diferentes linhagens (Treangen and Rocha, 2011; Swithers *et al.*, 2012).

Os três principais mecanismos de transferência horizontal são conjugação (Chen *et al.*, 2005) (Figura 1.1a), transformação (Chen and Dubnau, 2004) (Figura 1.1e), e transdução (Thomas and Nielsen, 2005) (Figura 1.1c). Mais recentemente foram descritos dois novos mecanismos de THG: Gene Transfer Agents (GTA) (Lang and Beatty, 2007; McDaniel *et al.*, 2010) (Figura 1.1d) e nanotubos (Dubey and Ben-Yehuda, 2011) (Figura 1.1b). Cada um destes mecanismos possui limitações e características próprias, e são capazes de transferências a partir de diferentes distâncias evolutivas e ecológicas (Popa and Dagan,

2011; Thomas and Nielsen, 2005). GTAs evoluíram a partir de profagos que perderam a capacidade de envelopar seu próprio DNA (Lang *et al.*, 2012) e são encontrados tanto em bactérias quanto arqueias. Apesar de sua alta taxa de ocorrência (McDaniel *et al.*, 2010), GTAs apresentam baixa eficiência já que envelopam fragmentos aleatórios de DNA, e muitas vezes nem mesmo seriam capazes de envelopar todo seu material genético (Lang *et al.*, 2012). O mais recente mecanismo de THG descrito ocorre através de nanotubos entre bactérias (Dubey and Ben-Yehuda, 2011), cujo comprimento chega a $1\mu\text{m}$ e sua espessura varia entre 30nm e 130nm. Estas estruturas possuem parede celular, membrana plasmática e citoplasma, sendo capazes de transferir diversos tipos de moléculas, grandes e pequenas, além de ácidos nucleicos (Dubey and Ben-Yehuda, 2011).

Figura 1.1: Esquema dos seis principais mecanismos de THG em procariotos: [a] conjugação; [b] nanotubos (Dubey and Ben-Yehuda, 2011); [c] transdução; [d] GTAs; [e] transformação. Adaptado de Soucy *et al.* (2015).

Apesar de sua reconhecida vasta importância na evolução de bactérias e arqueias, THG não é exclusiva de procariotos, tendo sido significativamente observada em eucari-

otos (Boto, 2014; Huang, 2013). Mesmo com relatos pontuais de THG em células germinativas e somáticas de eucariotos bem aceitos pela comunidade (Cornelis *et al.*, 2012; Riley *et al.*, 2013), sua importância e frequência com que THG participam a evolução de eucariotos multicelulares ainda não é bem compreendida (Boothby *et al.*, 2015; Crisp *et al.*, 2015; Salzberg, 2017; Koutsovoulos *et al.*, 2016). Ambos estudos que apresentaram uma quantidade significativa de THG em células germinativas de eucariotos multicelulares (e.g., humanos (Crisp *et al.*, 2015) e tardigrados(Boothby *et al.*, 2015)) foram veementemente desacreditados pela comunidade (Salzberg, 2017; Koutsovoulos *et al.*, 2016). Nestes casos específicos, as falsas identificações de genes recebidos horizontalmente em larga escala ocorreram devido contaminação (Boothby *et al.*, 2015) e/ou por falha nas inferência de homologia (Crisp *et al.*, 2015).

1.1 Tendências e barreiras à transferência horizontal de genes

As tendências e barreiras relacionadas a transferência lateral atuam em diversos níveis, desde a estabilidade do DNA no ambiente, à propensão de integrar-se ao genoma hospedeiro, até sua perpetuação na população.

A dissimilaridade entre o genoma hospedeiro e o doador constitui uma importante barreira para a ocorrência de THG, entretanto a intensidade desta barreira varia de acordo com o mecanismo de transferência em questão e as características do doador e receptor. Um estudo de THG utilizando redes filogenômicas (Popa *et al.*, 2011) apontou que 86% das transferências preditas entre bactérias ocorreram entre doadores e receptores com diferença de conteúdo GC menor que 5%. Este resultado pode ser melhor compreendido analisando o estudo de uma proteína similar à histona (H-NS), que suprime a transcrição de genes com baixo conteúdo GC (Navarre *et al.*, 2006). Este sistema, entretanto, pode ser inibido pela ação da proteína *sfh*, contida em um plasmídeo comum em diversas espécies de bactérias entéricas (Doyle *et al.*, 2007).

Transferências mediadas por transformação, conjugação, e nanotubos são afetadas

pela barreira de dissimilaridade genômica de uma maneira mais intensa que transferências mediadas por transdução ou GTAs. A integração do DNA adquirido por estes mecanismos geralmente ocorre por recombinação homóloga (Thomas and Nielsen, 2005), que requer uma grande similaridade entre as sequências recombinantes (Lovett *et al.*, 2002; Majewski and Cohan, 1999). Essa dependência de similaridade não ocorre com tamanha intensidade na recombinação não-homóloga, utilizada por integrases/transposases, já que sua função não requer similaridade entre longos fragmentos de DNA (Thomas and Nielsen, 2005). No caso da transdução, a integração do DNA viral não depende da sua similaridade com o genoma hospedeiro, e como barreiras a esse mecanismo podemos citar o próprio CRISPR (Marraffini and Sontheimer, 2008) e a especificidade de hospedeiros que o fago em questão é capaz de infectar (Popa and Dagan, 2011).

A principal barreira relacionada a transdução, e a GTAs, esta mais ligada a características da maquinaria de processamento de informação do que a similaridade de genomas em si, por utilizarem recombinação não-homóloga (Thomas and Nielsen, 2005; Popa *et al.*, 2017). Apesar da sobreposição de habitat ser um fator importante na transdução, esta barreira ecológica não é tão intensa quanto na conjugação e nanotubos, onde há a necessidade de contacto físico entre doador e receptor; e na transformação onde o DNA é sujeito a intempéries do meio (Popa and Dagan, 2011; Thomas and Nielsen, 2005; Popa *et al.*, 2017). Duas características da transdução sem correspondentes em outros mecanismos é a dependência da letalidade e especificidade do fago. A maioria dos fagos infecta apenas diferentes linhagens de uma mesma espécie, e portanto estão mais próximos de recombinação endógena que de THG (Popa *et al.*, 2017), apesar do viés de transferência entre organismos próximas ser comum a todos os mecanismos (Shapiro *et al.*, 2012; Andam *et al.*, 2010; Andam and Gogarten, 2011; Papke and Gogarten, 2012). Na rede de THG mediada por transdução estudada por Popa *et al.* (2017) apenas um evento de THG entre diferentes filos foi observado.

Para que uma ilha genômica seja perpetuada em um determinado organismo e, consequentemente, seja transferida para seus descendentes é necessário que ela forneça alguma vantagem adaptativa ao hospedeiro. Portanto, genes cujo produto não capacitem o hos-

pedreiro com certa vantagem, ou cuja expressão seja dificultada por algum motivo, tendem a ser perdidos ao longo da evolução do hospedeiro. Ou pelo menos este *era* o consenso. Sabe-se que vários dos genes transferidos horizontalmente entre organismos próximos possuem *quase* nenhum efeito em seu receptor, não conferindo nenhuma vantagem (Shapiro *et al.*, 2012; Gogarten and Townsend, 2005). Como exemplo disso Shapiro *et al.* (2012) avaliou 20 linhagens de *Vibrio cyclitrophicus*, divididos entre dois nichos, e concluiu que apenas 1% do “core genoma” desta espécie foi herdado verticalmente. Esta alta frequência de transferência de genes do “core genoma” exemplifica o baixo impacto que genes tão conservados possuem no *fitness* destes organismos, pode ser trocados quase que livremente (Shapiro *et al.*, 2012; Papke and Gogarten, 2012).

1.2 Métodos de identificação

Diversos métodos para identificação de genes transferidos horizontalmente foram desenvolvidos e vem estão sendo constantemente aperfeiçoados. Diferentes métodos possuem características distintas, sendo capazes de identificar genes transferidos através de diferentes perspectivas, as quais algumas vezes não seriam significantes a outros métodos (Lawrence and Ochman, 2002; Ragan, 2002; Ragan *et al.*, 2006; Ragan, 2001). Os dois principais tipos de métodos de identificação de genes transferidos horizontalmente baseiam-se em identificar regiões do genoma com diferentes padrões composicionais (Seção 1.2.1), ou observando incongruências filogenéticas entre a árvore de um determinado gene e a árvore das espécies onde ele se encontra (Seção 1.2.2).

1.2.1 Desvio de padrões composicionais

Genes transferidos horizontalmente tendem a apresentar padrões composicionais diferentes de seus genomas hospedeiros, e isto ocorre devido aos diferentes vieses mutacionais apresentados por diferentes organismos (Sharp and Li, 1987; Sharp *et al.*, 2010). A maioria dos métodos baseados em desvios composicionais procura por diferenças significativas na frequência de determinados *k-mers* em regiões do genoma em comparação com a distri-

buição esperada. As métricas mais comuns são em relação a frequência de G+C (Karlin *et al.*, 1998; Lawrence and Ochman, 1997), preferência no uso de codons (Sharp and Li, 1987; Karlin *et al.*, 1998; Lawrence, 1998), e dinucleotídeos (Karlin *et al.*, 1998), mas várias ferramentas usam *k-mers* de tamanho até 12 (Sandberg *et al.*, 2001; Tsirigos and Rigoutsos, 2005). As principais vantagens na identificação de THGs através de desvios composicionais é o baixo custo computacional em comparação aos métodos filogenéticos (Seção 1.2.2) e não precisar de vários genomas próximos para comparação. Esta última vantagem é particularmente interessante para estudos de organismos não-modelos, onde não há genomas de linhagens próximas para comparação. Os principais problemas dos métodos baseados em desvios composicionais é a alta taxa de erro (Koski *et al.*, 2001; Azad and Lawrence, 2005; Cortez *et al.*, 2005) e a limitação em identificar apenas genes recebidos recentemente. Esta limitação ocorre devido ao rápido processo de adaptação da sequência do gene adquirido ao genoma hospedeiro, processo chamado amelioração (Lawrence and Ochman, 1997). Exemplos das ferramentas mais comuns são: Dark Horse (Podell and Gaasterland, 2007), COLOMBO SIGI-HMM (Waack *et al.*, 2006), Island-Path (Hsiao *et al.*, 2003), e AlienHunter (Vernikos and Parkhill, 2006).

Figura 1.2: Três exemplos de análises de perfis composicionais do genoma de *Magnetococcus marinus*. Em azul a frequência de G+C, em verde as variações na preferência no uso de codons (como definida por Karlin *et al.* (1998)), e em preto uma outra métrica de preferência de codons (como definida por Sharp and Li (1987)). Destacados em vermelho encontram-se as posições de genes ligados a fixação de N_2 .

1.2.2 Incongruências filogenéticas

Este grupo compreende métodos que buscam identificar genes transferidos horizontalmente identificando incongruências entre a topologia observada na árvore filogenética de um gene com a topologia esperada pela árvore representante a dos organismos em questão. Como dito anteriormente, em geral as transferências horizontais são identificadas através de alterações significativas na topologia da árvore do gene, e esta abordagem é utilizada por diversas ferramentas (Bansal *et al.*, 2012; Zhaxybayeva *et al.*, 2006; Mao *et al.*, 2012). Métodos que utilizam esta abordagem tendem a necessitar de significativo tempo computacional, dado que a inferência filogenética é um fator crucial para a acurácia da análise. Para *burlar* tal barreira que impede análises acuradas com grande quantidade de taxa (*Quartet decomposition server* desenvolvido por Mao *et al.* (2012) recomenda o uso de 20 taxa, e no máximo 100) o método desenvolvido por Novichkov *et al.* (2004) compara matrizes de distância pareadas (Figura 1.3). Devido a métodos que utilizam esta abordagem possuírem maior acurácia, quando comparados a métodos baseados em desvios composicionais, decidimos utilizar a ferramenta de reconciliação filogenética desenvolvida por Bansal *et al.* (2012) no decorrer deste projeto. No capítulo 3 vamos mais afundo no método desenvolvido por Novichkov *et al.* (2004), e como o adaptamos para agrupar famílias gênicas com evolução compartilhada.

Figura 1.3: Exemplo da regressão esperada quando o gene em questão evolui exclusivamente através de herança vertical (A) e da incongruência entre distâncias pareadas entre *taxa* quando o gene em questão é transferido a partir de um ancestral de *X* para um ancestral de *BC* (B). Adaptado de Novichkov *et al.* (2004).

1.3 Redes filogenômicas

Uma rede, ou grafo, é um modelo matemático formado por um conjunto de vértices (nós) conectados por arestas (Fig. 1.4). As arestas de uma rede podem possuir pesos, representando a intensidade das conexões entre os vértices, ou direções. No caso de redes filogenômicas os vértices representam genomas, enquanto que as arestas o compartilhamento de genes entre genomas. O peso das arestas que ligam os genomas correspondem a quantos genes são compartilhados, e a direção das arestas pode representar a direção de possíveis eventos de THG envolvendo os genomas (a aresta é orientada do genoma doador para o genoma receptor). Neste modelo, o compartilhamento de genes entre genomas pode ocorrer por dois motivos: os genes foram herdados verticalmente de um ancestral comum, ou houve transferência lateral. Neste último caso não existe necessariamente a implicação de que a aresta representa transferência direta de um genoma para outro; na vasta maioria dos casos, está implícito que pode ter havido inúmeros passos intermediários que não estão representados no modelo. Acreditamos que pela sua versatilidade o modelo de redes é mais adequado para representar a evolução de procariotos do que o modelo dicotomial Feil *et al.* (2001); Gogarten *et al.* (2002); Baptiste *et al.* (2009); Martin (2011).

Figura 1.4: Exemplo de rede filogenômica composta pelos genomas hipotéticos: *a*, *b*, *c*, *d* e *e*. As arestas conectando os vértices representam o compartilhamento de genes entre os genomas, e a quantidade de genes compartilhados é expressa pelos números. Adaptado de Dagan Dagan (2011).

1.3.1 Porque representar a evolução procariótica utilizando redes

Como dito anteriormente, THGs possui grande importância na evolução de procariotos, sendo o principal mecanismo de inovação genômica, além de afetar virtualmente todas famílias gênicas. Dado isto, não é possível representar a evolução de genomas procarióticos utilizando modelos dicotomais (*i.e.*, árvores) já que os genomas são mosaicos cujos fragmentos foram obtidos de fontes distintas (Snel *et al.*, 2002).

O modelo de árvores dicotomais foi formulado levando em consideração apenas dois mecanismos de inovação: variações hereditárias e seleção natural, os quais modelam muito bem a evolução de eucariotos multicelulares. Entretanto, sabemos hoje a evolução não é *apenas* Darwinista (Koonin and Wolf, 2009) e que os mecanismos de variações naturais não são homogêneos a todas formas de vida, sendo particularmente distintos entre eucariotos e procariotos (Baptiste *et al.*, 2009). Diferentes dos mecanismos de evolução comuns entre eucariotos, eventos de THG podem introduzir genes, ou conjuntos de genes, que conferem funcionalidades complexas a organismos, capacitando-os a explorar novos nichos (Lawrence, 1998; Lawrence and Ochman, 1997; Hacker and Kaper, 2000).

Capítulo 2

Objetivos

2.1 Objetivos gerais

Estudar o impacto da Transferência Horizontal na evolução de genes ligados à fixação de N_2 em Proteobacteria e em efetores do Sistema Secretor do Tipo III em *Aeromonas* spp. utilizando análises computacionais.

Optamos por avaliar o impacto da THG envolvendo fixação de N_2 devido a experiência prévia do orientador (Setubal *et al.*, 2009; Dos Santos *et al.*, 2012; Setubal and Almeida, 2015). Além disto, o orientador é colaborador de dois pesquisadores de relevância na área Dr. Ray Dixon e Dr^a. Patrícia dos Santos, os quais colaboraram neste projeto.

Já os genes de efetores do Sistema Secretor do Tipo III (SST3) foram avaliados devido a nossa parceria com o Dr. Peter Gogarten e Dr. Joerg Graf. Entre Abril de 2015 e Março de 2016 o estudante obteve uma Bolsa Estágio de Pesquisa no Exterior (BEPE), concedida pela FAPESP, sob a supervisão do Dr. Gogarten. O Dr. Gogarten e seu colaborador, Dr. Graf, utilizam *Aeromonas* como modelo em estudos evolutivos (Silver *et al.*, 2011, 2007a,b; Colston *et al.*, 2014; Morandi *et al.*, 2005), e a evolução de efetores do SST3 mostrou-se bastante pertinente ao foco do nosso projeto.

Capítulo 3

Agrupamento de famílias gênicas por similaridade evolutiva

3.1 Motivação

Durante o estudo que levou a publicação do manuscrito “Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case” (Colston *et al.*, 2014), o grupo do Dr. Gogarten observou diferenças significativas entre as árvores filogenéticas de 16 genes housekeeping (*atpD*, *dnaJ*, *dnaK*, *dnaX*, *gltA*, *groL*, *gyrA*, *gyrB*, *metG*, *mdh*, *radA*, *recA*, *rpoC*, *rpoD*, *tsf*, e *zipA*) utilizados para reconstruir a filogenia de *Aeromonas* spp. utilizada no estudo em questão. Apesar de apresentarem incongruências, a árvore obtida a partir do concatenado dos 16 genes mostrou-se coerente com dados da literatura e bastante similar a árvore de core-genoma gerada no mesmo estudo (Colston *et al.*, 2014).

Fomos apresentados ao problema das incongruências entre árvores dos 16 genes housekeeping em *Aeromonas* spp. pelo Dr. Gogarten durante a execução do projeto apresentado no Capítulo 5. As diferenças entre as árvores dos genes housekeeping foram novamente identificadas entre homólogos dos 105 genomas de *Aeromonas*. A Figura 3.1 apresenta um **Multidimensional Scalling** das distâncias par-a-par entre árvores dos 16 genes housekeeping presentes nos 105 genomas apresentados no Capítulo 5, onde cada gene está

representado por uma cor. Cada ponto presente no gráfico apresentado na Figura 3.1 foi obtido através da distribuição *a posteriori* gerada pelo Mr. Bayes, 500 árvores foram amostradas para cada um dos genes housekeeping.

Figura 3.1: *Multidimensional scaling* das distâncias entre 16 genes housekeeping de 105 genomas de *Aeromonas* spp. descritos no Capítulo 5. Os diversos pontos representando topologias sub-ótimas foram amostradas da distribuição *a posteriori* gerada pelo Mr. Bayes.

Diversas abordagens foram propostas para agrupar genes que compartilham histórias evolutivas: ASTRAL (Mirarab *et al.*, 2014), CONCATERPILAR (Leigh *et al.*, 2008), MP-EST (Liu *et al.*, 2010), e treeCL (Gori *et al.*, 2016). Entretanto, várias destas abordagens modelam apenas certos tipos de processos evolutivos (*e.g.* ILS) ou são próprios para modelos eucarióticos (*i.e.* ASTRAL e MP-EST). Todas as ferramentas citadas necessitam que os genes que serão comparados possuam árvores filogenéticas determinadas, o que adiciona possíveis erros a abordagem devido a incerteza intrínseca a inferência filogenética, além do alto tempo computacional necessário para a determinação de uma árvore “confiável”.

Devido as limitações presentes nas ferramentas existentes desenvolvemos um método computacionalmente eficiente capaz de agrupar famílias gênicas com histórias evolutivas similares, o qual não modela processos evolutivos específicos. O desenvolvimento de uma ferramenta com estas características torna possível a reconstrução automática de árvores filogenômicas para grandes conjuntos de dados, e também para a identificação de grupos

de genes com comportamentos evolutivos similares, possivelmente constituindo *highways* de THG. Dado que genes com histórias evolutivas similares sejam agrupados com sucesso, esperamos que um dos grupos formados corresponda a genes cuja evolução se deu majoritariamente vertical, enquanto que os demais grupos de famílias gênicas representam variações da evolução vertical. A quantidade de famílias gênicas contidas no provável grupo representante da herança vertical pode sugerir o quão frequente é a THG entre os organismos estudados, enquanto que a diversidade de sub-grupos representando diferentes evoluções horizontais pode evidenciar o impacto da THG em certos *taxa*.

3.2 Descrição do método desenvolvido

Ao assumirmos o “relógio molecular relaxado” esperamos que genes homólogos acumulem substituições em frequências similares, as quais podem variar ao longo da evolução, mas de tal forma que estas mesmas variações também ocorram nos demais ramos da árvore (Ohta and Gillespie, 1996; Fay and Wu, 2001). Desta maneira, duas famílias gênicas evoluindo de maneira similar tendem a acumular substituições em taxas proporcionais, e portanto esperamos que as distâncias entre nós terminais de ambas famílias sejam altamente correlacionadas (Novichkov *et al.*, 2004). Variações na taxa de substituições em ramos específicos da árvore podem ocorrer, entretanto variações além do esperado são mais frequentes quando há neo-funcionalização (Bromham and Penny, 2003; Kimura, 1983).

Nosso método, até o momento, aplica-se apenas para comparar famílias gênicas sem duplicações. A comparação de histórias evolutivas utilizando matrizes de distância diminui o tempo computacional necessário, além de eliminar uma provável camada de erro devido a incerteza ligada a reconstrução filogenética. Dado que métricas para comparação de árvores filogenéticas que levam em consideração apenas comprimento de ramos tendem a ser tão precisos quanto métricas que avaliam topologia e comprimento de ramos (Gori *et al.*, 2016) não esperamos um impacto significativo na acurácia. A descrição dos passos utilizados encontra-se abaixo:

- (1) A partir da definição de grupos homólogos presentes em um determinado conjunto

de genomas, famílias gênicas com possíveis parálogos devem ser excluídos da cadeia de processamento

- (2) Construção de alinhamentos múltiplos para todos grupos homólogos;
- (3) Criação de matriz de distâncias de Máxima Verossimilhança entre todas sequências, uma matriz pra cada grupo homólogo;
 - A construção de árvores filogenéticas corresponde a maior parte do processamento computacional utilizado durante a comparação de histórias evolutivas. Como forma de aumentar a eficiência do método, e visto que não há perda em acurácia em não se utilizar topologias para comparações de árvores (Gori *et al.*, 2016), utilizamos apenas a matriz de distâncias de Máxima Verossimilhança como representação de cada grupo homólogo;
- (4) Calcular o coeficiente de correlação de Pearson para todas combinação de famílias gênicas;
 - a menor sobreposição necessária para o cálculo do coeficiente é de dois genomas, entretanto quanto maior o número de pontos utilizados maior a credibilidade da correlação obtida;
 - Um total de $\frac{n!}{2(n-2)!}$ testes de correlações são realizados caso todos os grupos sejam suficientemente sobrepostos, onde n é o número de grupos homólogos avaliados;
- (5) Calcular a significância dos coeficientes de correlação obtidos no passo anterior
- (6) Agrupar as famílias gênicas utilizando correlações significantes como métrica de similaridade entre histórias evolutivas

Para validar o métodos descrito acima, no futuro pretendemos utilizar dados simulados para conferir a acurácia do nosso método com outros presentes na literatura. A simulação de famílias gênicas sob diferentes níveis de recombinação, taxas de substituição, e retenção de caracteres ancestrais pode ser capaz de aferir a robustez do método, além de compara-lo com resultados obtidos por outros métodos.

3.3 Implementação

Afim de identificar similaridades evolutivas entre famílias gênicas presentes em 105 genomas de *Aeromonas* (Capítulo 5), implementamos o método descrito acima (Capítulo 3.2) com os seguintes parâmetros: do total de 25.518 grupos homólogos identificados entre os genomas de *Aeromonas*, 5.693 satisfizeram os requisitos necessários de [1] ausência de parálogos e [2] presente em ao menos dez genomas. Requisitamos no mínimo dez sequências em cada família gênica devido a tolerância entre o conjunto de genomas presente em cada família (Distância de Jaccard $\leq 0,3$, D_J). As matrizes de distâncias de Máxima Verossimilhança foram geradas utilizando a ferramenta **Tree-Puzzle** (Schmidt *et al.*, 2002), utilizando o modelo **GTR+GAMMA**. Calculamos os coeficientes de correlação de Pearson para todos os pares de grupos homólogos com sobreposição suficiente ($D_J < 0,3$), e suas respectivas significâncias utilizando o teste de Mantel (Mantel, 1967), desenvolvido para calcular correlações entre matrizes de distância.

A exigência de sobreposição mínima descartou 29,5% das comparações entre grupos homólogos. Coeficientes de correlação, e significância, foram calculados para 11.421.810 pares de grupos homólogos. Dada a grande quantidade de testes de significância realizados foi necessária corrigir os p-valores obtidos, esta correção foi realizada utilizando o *False Discovery Rate* (FDR) (Benjamini and Hochberg, 1995), e obtivemos 3.476.614 correlações significantes ($p \leq ,0.05$).

Utilizamos *Markov Clustering Algorithm* (MCL) (van Dongen, 2000; Enright *et al.*, 2002) para agrupar as famílias gênicas de acordo com as correlações entre suas respectivas matrizes de distância. Optamos por utilizar o MCL devido ao seu amplo uso agrupamento de sequências (Li *et al.*, 2003; Szklarczyk *et al.*, 2011; Brohee and Van Helden, 2006; Dunn *et al.*, 2008), apesar de não ser ideal para redes muito conectadas. Para diminuir o número de arestas presentes na rede utilizamos o limiar de correlação (0,7) e a transformação de correlações sugerida por van Dongen and Abreu-Goodger (2012), que reduziu o número de arestas para 2.219.472.

Esta implementação, e aplicação no conjunto de 105 genomas de *Aeromonas* (Capítulo 5) gerou 167 grupos de famílias gênicas com histórias evolutivas similares. O maior

cluster obtido abrange 1.678 famílias gênicas, o qual contém mais de duas vezes o número de famílias agrupadas no segundo maior *cluster* 738.

Detalhamos os passos desta implementação mais profundamente que no Capítulo 5 devido ao manuscrito científico que pretendemos desenvolver sobre esta metodologia.

3.4 Verificação

Além de comparar a similaridade entre as histórias evolutivas das famílias gênicas utilizando as correlações entre matrizes de distância de Máxima Verossimilhança, também utilizamos o critério de Fitch-Margoliash (Fitch *et al.*, 1967) para comparar as matrizes de distância. O critério de Fitch-Margoliash é bastante similar ao método de mínimos quadrados, e apresentou uma correlação de 0.89 com os valores obtidos através de nosso método.

Caso a rede de correlações de distâncias evolutivas corretamente reflita grupos de famílias gênicas com histórias evolutivas similares esperamos, entre outras coisas, que os genes de um mesmo *cluster* ocorram em regiões próximas uns dos outros. Por ocorrerem próximos, a probabilidade de eventos de reorganização genômica ocorrerem entre dois genes é menor, e portanto maior a probabilidade de ambos serem transferidos juntos, vertical ou horizontalmente. Verificamos esta hipótese calculando a proporção de pares de genes adjacentes de *A. veronii* HM21 presentes no maior *cluster*, total de 997. Comparamos este valor com a distribuição de genes adjacentes de *A. veronii* HM21 presentes em 1.000 permutações com reposição de 1.678 genes aleatórios entre os 5.693 utilizados para formar a rede, gerando uma distribuição nula da frequência de genes adjacentes (Figura 3.2). A frequência de genes adjacentes de *A. veronii* HM21 observados no maior *cluster*, em **vermelho**, é claramente maior que todos os valores da distribuição nula presente na Figura 3.2, em **azul**.

Figura 3.2: Distribuição da ocorrência de genes adjacentes observados em 1.000 amostras aleatórias de 1.678 genes de *A. veronii* HM21, em azul. O círculo em vermelho representa quantos pares de genes adjacentes de *A. veronii* HM21 foram observados no maior grupo obtido.

Como forma de mensurar a compatibilidade entre os genes do maior *cluster*, comparamos as distâncias entre as árvores dos 16 genes housekeeping (Figura 3.1) com distâncias entre árvores dos 16 genes mais conectados presentes no maior *cluster*. Para medir as distâncias entre membros dos diferentes conjuntos de genes reconstruímos suas árvores filogenéticas utilizando o RAxML, e calculamos as distâncias utilizando o método de Robinson-Foulds (Figura 3.3). Testamos a hipótese nula (H_0), de que as distâncias entre os 16 genes do maior *cluster* são maiores ou iguais as distâncias entre os 16 genes housekeeping, a qual foi rejeitada. Isto nos mostra que um conjunto de genes classicamente utilizados em estudos de MLSA não são tão compatíveis quanto previamente esperado, o que fica claro ao observarmos vários destes genes estão distribuídos em diversos *clusters*, portanto possuindo histórias evolutivas significantemente distintas.

Na árvore representativa dos genomas de *Aeromonas* presente na Figura 5.2 utilizamos os genes contidos no maior *cluster* obtido por este método. A árvore gerada mostrou-se altamente compatível com a árvore de core-genoma proposta por Colston *et al.* (2014).

O método apresentado mostrou-se eficiente em comparar histórias evolutivas, sendo

Figura 3.3: Distribuição das distâncias entre os 16 genes housekeeping, em roxo; e distâncias entre 16 genes selecionados do maior grupo obtido, em verde.

capaz de identificar as possíveis causas das incongruências identificadas entre os 16 genes housekeeping. O fato dos 16 genes housekeeping não estarem todos em um único grupo é provavelmente causado por diferentes eventos THG entre os mesmos.

Apesar de ainda necessitar validação mais incisiva o método proposto mostra-se é promissor, e pode ser aplicado em famílias gênicas pertencentes a diversos *taxa* por ser independente de características específicas das encontradas em *Aeromonas* spp.

Capítulo 4

História evolutiva de genes de fixação de N_2 em Proteobacteria

4.1 Contextualização

Neste capítulo, como resultado de nosso estudo sobre a história evolutiva da diazotrofia em Proteobacteria apresentamos um manuscrito científico em fase final de preparação. Foram amostrados 479 genomas completos de Proteobacteria, das quais 87 possuem os requisitos mínimos para fixar dinitrogênio (Dos Santos *et al.*, 2012). Comparamos as filogenias de homólogos de 45 genes ligados a diazotrofia em *Azotobacter vinelandii* DJ (Setubal *et al.*, 2009) com a filogenia esperada para os genomas do filo. Incongruências entre as filogenias foram identificadas utilizando o método de reconciliação filogenética, implementada sob o contexto de máxima parcimônia (Bansal *et al.*, 2012). Este estudo foi realizado com a participação de dois colaboradores: Dr^a. Patrícia dos Santos e Dr. Ray Dixon, cujas respectivas afiliações encontram-se no manuscrito a seguir.

Evolutionary history of nitrogen fixation genes within Proteobacteria

Luiz Thiberio Rangel^{1,2}, Patricia C. Dos Santos³, Ray Dixon⁴, João Carlos Setubal^{*,1,2}

¹Interunidades em Bioinformática, Universidade de São Paulo, São Paulo, Brazil

²Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil

³Department of Chemistry, Wake Forest University, Winston-Salem, NC, USA

⁴Department of Molecular Microbiology, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK

***Corresponding author:** E-mail: joao.c.setubal@gmail.com

Keywords: Horizontal Gene Transfer, nitrogen fixation, evolution, Proteobacteria.

Introduction

Diazotrophs are organisms capable of fixing nitrogen gas (N_2) into a biologically usable form, ammonia (NH_3). They are widely distributed across the bacterial and archaeal kingdoms and, up to date, no eukaryotic species have been found to sustain active nitrogen fixation functions. The predicted evolutionary origin of nitrogen fixation dates from early methanogens in its Mo-dependent form (Boyd, Anbar, et al. 2011; Boyd, Hamilton, et al. 2011). The distribution of diazotrophs is sparse among bacterial and archaeal lineages; for instance, it tends to not occur in all species within a given genus. An explanation for this distribution solely based on vertical inheritance and gene loss does not fit in a maximum parsimony framework. Therefore, horizontal gene transfer has likely played an important role in the evolution of diazotrophy-related genes (Hirsch et al. 1995; Henson et al. 2004; Gaby and Buckley 2014). While the capacity for fixing nitrogen is present in many species of bacteria and archaea, the Proteobacteria phylum represents the majority of completely sequenced genomes,

and constitutes roughly 53% of all diazotrophs available in IMG/DOE on January 2012 (Boyd et al. 2015). Thus, Proteobacteria is an attractive taxonomic group to investigate the evolutionary trajectory of genes involved in nitrogen fixation.

All N₂ fixing organisms identified to date produce molybdenum-containing nitrogenase. Some organisms also have additional paralogous enzymes that utilize other transition metals to complete their active site cofactor; these include the vanadium-containing nitrogenase and an iron-only nitrogenase (Eady 1996). Interestingly, alternative nitrogenases have only been found in species containing Mo-containing nitrogenase and are only expressed in environments where Mo is limiting. The existence of paralogous enzymes that utilize alternate metals has been attributed to gene duplication events in recent adaptive evolution in response to low molybdenum environments (Boyd, Hamilton, et al. 2011) and has not been explored in this study.

Mo-containing nitrogenases are composed of two oxygen sensitive metallo components designated the MoFe protein and the Fe protein, products of *nifDK* and *nifH* respectively. Buried within the MoFe protein, the N₂ binding and reduction site, the FeMo-cofactor, has a unique structure not identified in any other class of metalloprotein. The biosynthesis of FeMo-cofactor is critical for nitrogen fixation. That is, inactivation of FeMo-cofactor biosynthetic genes renders inactive nitrogenase. Likewise, species lacking the genes associated with the assembly of this cofactor are not able to sustain nitrogen fixation despite the presence of genes orthologous to nitrogenase components. To date, all nitrogen fixing species with sequenced genomes contain a minimum of six genes dedicated to nitrogen fixation, those encode both nitrogenase components (*nifH*, *nifD*, *nifK*) and two FeMo-cofactor biosynthetic enzymes (*nifB*, *nifE*, *nifN*) (Dos Santos et al. 2012). Thus the presence of this minimum gene set (defined in this study as MNS) has been used as an *in silico* predictor of nitrogen fixation. Important to note, however, that few exceptions of the MNS have been identified, but those were limited to a select group of hyperthermophilic diazotrophs (Dos Santos et al. 2012; Boyd et al. 2015) and are not found in any species of Proteobacteria known to fix nitrogen.

Biological nitrogen fixation in most species involves the recruitment of additional components other than the MNS. The number of genes associated with nitrogen fixation varies among different species and the involvement of accessory proteins depends on the metabolic and environmental

conditions encountered during growth. In some instances, these gene products also perform housekeeping functions not associated to nitrogen fixation. In other cases, this process involves the participation of paralogous genes dedicated to nitrogen fixation. The evolutionary origin of dedicated nitrogen fixation genes that function to complement the MNS is not fully understood.

Here we present a detailed analysis of the evolution of N_2 -fixation genes in Proteobacteria lineages. We examined 479 genomes, one from each proteobacterial species available on NCBI on GenBank on January 2013. The thorough representation of the diversity present in a single phylum allowed for the implementation of high resolution evolutionary methods, which would not be feasible among more diverse datasets (Hirsch et al. 1995; Henson et al. 2004; Gaby and Buckley 2014). We analyzed the evolution of 45 genes related to Mo-dependent N_2 -fixation, including the MNS. Here, we report five distinct origins of N_2 -fixation gene sets within Proteobacteria lineages, besides a diverse network-like process of horizontal gene exchange among lineages. The results provide new insights into the importance of HGT events in the evolution of N_2 -fixation.

Results and Discussion

Distribution of N_2 fixation-related genes within Proteobacteria

We started this investigation based on 83,803 groups of protein homologs from Proteobacteria obtained with orthoMCL and pre-compiled a list of 45 gene families important to diazotrophy (Supplementary Table S1). This gene list was based on previous genetic, transcriptomic, and biochemical studies using the γ -proteobacterium *Azotobacter vinelandii* (Setubal et al. 2009; Dos Santos et al. 2012), a model organism for the study of nitrogen fixation (Kennedy and Toukdarian 1987). In this study we have particularly targeted the evolutionary history of proteobacteria genes dedicated to nitrogen fixation, thus homologous genes performing housekeeping functions were not included in this analysis. We sought to establish the occurrence of gene products specialized in nitrogen fixation among proteobacterial lineages. By filtering out the homologous sequences performing additional functions other than nitrogen fixation, we could focus on the identification of evolutionary steps linked to nitrogen fixation.

While assignment of families corresponding to the six genes present in MNS was straight

forward, the identification of the remaining gene families was not. The 6 MNS genes are well-characterized and manual curation of the homolog groups showed to be able to identify and filter out homologs not directly related to nitrogen fixation. For such we manually assessed the completion of the MNS and the phylogenetic context of each homolog in its respective tree. The same approach was not applicable to homologs from the remaining 39 N_2 -fixation gene families given the occurrence of homologs not associated with nitrogen fixation. For example, genes located in the *rnf1* and *rnf2* operons are homologs (e.g., *rnfA1* and *rnfA2*, *rnfB1* and *rnfB2*, and so on) and therefore were clustered in the same homologous groups. However, while *rnf1* has a central role in diazotrophy (Schmehl et al. 1993), *rnf2* does not (Curatti et al. 2005). Therefore, in order to retain only genes dedicated to nitrogen fixation, we used a combination of the following criteria during the curation of non-MNS N_2 -fixation genes: genomic distance to MNS operons and phylogenetic context.

The genomic distance criterion was based on the observation that the MNS tends to occur in three gene regions: *nifHDK*, *nifEN*, and *nifB* (Fani et al. 2000; Raymond et al. 2004), and that N_2 -fixation genes also tend to occur in close genomic regions (Yan et al. 2008; Setubal et al. 2009). We used MNS well-defined regions as anchors to make comparisons with other genes. We observed that among N_2 -fixation homologous groups containing paralogs, only one copy in each diazotrophic genome is located close to MNS operons. In genomes with multiple N_2 -fixation candidate homologs, the genomic distance from each candidate to the MNS regions combined with their relative placement in the phylogenetic tree was shown to be a good indicator of a role in diazotrophy. We manually verified 121 proteins from *nifV*, *nifQ*, *nifS*, and *nifU* and only 3.3% of the verified proteins were more than 100 genes apart the MNS operons and its close homologs also do not occur the vicinity of these operons. The threshold of genomic distance between nitrogen fixing genes was first determined for well-known diazotrophs (e.g., *A. vinelandii*, *Azoarcus* sp., and *Rhodobacter capsulatus*). Based on these initial analyses, we established a distance threshold of 100 protein-coding genes, both downstream and upstream, on both strands as regions that could include genes required for nitrogen fixation.

The phylogenetic context criterion was then used to corroborate the genomic distance requirement. In cases where a candidate homolog was in the vicinity of an MNS region, but none of

its close homologs are located within the required range, we manually checked its relation to diazotrophy through the literature and gene annotation. N_2 -fixation candidate genes removed using this filtering process fell into one of two categories: [1] even though the candidate gene has diazotrophy-related homologs it does not possess a N_2 -fixation role in its respective organism (e.g., the *nifU* homolog Arnit_0876 in *Arcobacter nitrofigilis*); [2] the candidate gene has two copies, and only one is related to the N_2 fixation process, according to the distance criterion specified above. The list of N_2 -fixation genes obtained after this filtering process is presented in Supplementary Table S1.

A total of 479 Proteobacteria genomes, represented by α (29.5%), β (17%), γ (38.8%), δ (8.3%), ϵ (5.6%), and ζ -Proteobacteria (0.62%) were scanned for the presence MNS genes, used to infer diazotrophy. The distribution of N_2 -fixation genes was not uniformly distributed among distinct taxonomic classes of Proteobacteria. Within these groups, δ -Proteobacteria contained the highest prevalence of diazotrophs, with 40% of its members possessing MNS genes. While being the taxonomic class with the most sequenced species γ -Proteobacteria was the class with the least prevalence of diazotrophs with only 10.2% of its members possessing the complete MNS. The prevalence of diazotrophs was roughly similar among α , β , and ϵ -Proteobacteria (approximately 30%).

The occurrence of nitrogen fixation genes in 87 diazotrophic genomes is represented as a presence/absence matrix associated to the Proteobacteria phylogenetic tree in Figure 1. The larger concentration of genes related to N_2 -fixation at the bottom of the presence/absence matrix (Figure 1) is likely caused by the fact that all γ -Proteobacteria, and some β -Proteobacteria, have a more complex N_2 -fixation gene set than the other classes of Proteobacteria. The larger complexity of the N_2 -fixation orthologous genes in γ -Proteobacteria has also been discussed by Boyd *et al.* (2015) although not extended to β -Proteobacteria. It is also worth mentioning that the analysis performed in this study differs from the study of Boyd and collaborators because it surveyed a larger number of gene families (45) and included only genes dedicated to nitrogen fixation based on the criteria discussed above. This treatment excluded homologous genes that are unlikely to perform functions specifically related to nitrogen fixation. In this study, we defined as “complexity” the number of N_2 -fixation genes in an organism divided by the total number of surveyed gene families (45), which varied between 0.15 (

Desulfobacca acetoxidans and *Desulfobulbus propionicus*, both δ -Proteobacteria) to 1.0 (*A. vinelandii* and *Pseudomonas stutzeri*, both γ -Proteobacteria).

It has been suggested that the complexity of N_2 -fixation gene sets are associated with the occurrence of N_2 -fixation genes, likely due to stepwise improvements of the FeMo-co and MoFe clusters (Boyd et al. 2015). Such association is also present among N_2 -fixation genes present in Proteobacteria. When comparing the complexity of N_2 -fixation gene sets between sets of organisms with and without each of the non-MNS gene families only genomes possessing FesII did not show significantly larger N_2 -fixation gene set complexity than its counterpart ($p < 0.05$). Despite the positive association between the presence of N_2 -fixation genes and the gene set complexity, phylogenetic trees of MNS genes do not seem to explain the N_2 -fixation gene set complexity. Diazotrophic species within δ -Proteobacteria showed lower complexity of the N_2 -fixation gene set. Interestingly, nearly all representatives of this group display an anaerobic lifestyle.

Distinct origins of nitrogen fixation in Proteobacteria

The phylogenetic distribution of N_2 fixation gene sets allows the evaluation of HGT events during their evolution. First, we sought to determine whether homologs within each N_2 -fixation gene family have a single shared origin within the Proteobacteria. By recruiting homologs from non-Proteobacteria lineages and reconstructing the phylogenetic trees of each gene family, using both Proteobacteria and non-Proteobacteria sequences, we were able to assess how Proteobacteria homologs clustered with respect to each other. This analysis showed that 35 out of the 45 assessed N_2 -fixation gene families exhibit more than one Proteobacteria subtree. These results are in agreement with previous phylogenies of *nifHDK* suggesting that Proteobacteria homologs are paraphyletic, and are split in six monophyletic clades (Boyd et al. 2015). From all 45 N_2 -fixation gene families, only ten seem to have a single common origin within Proteobacteria shared by all their homologs (*nifT*, *clpX*, *rnfB1*, *rnfH*, *Avin_01360*, *Avin_01500*, *Avin_01540*, *Avin_01560*, *Avin_01570*, and *Avin_01720*). Amongst the remaining 35 gene families with multiple ancestral acquisitions, we predicted the following number of distinct acquisitions of MNS genes by Proteobacteria: five different origins within *nifHDK* (Supplementary Figure S1-S3), seven within *nifE* (Supplementary Figure S4), six

within *nifN* (Supplementary Figure S5), and four within *nifB* (Supplementary Figure S6).

We measured how consistent Proteobacteria genomes had their genes grouped in similar subtrees among distinct gene families using the Variation of Information (VI) metric (Meilă 2007). According to the normalized VI (VI_N), genes from similar sets of genomes were grouped in subtrees present in the majority of gene families. Among all 595 possible pairwise comparisons between the 35 gene families with multiple subtrees, 434 (75%) exhibited little difference in how genomes were grouped ($VI_N \leq 0.3$), while only 35 comparisons presented significant differences ($VI_N \geq 0.5$). When accounting for the subtrees only found in MNS gene families and considering no duplications among homologs, *nifH* had the most dissimilar genome clustering among its subtrees. The VI_N of *nifH* subtrees, when compared to *nifDKENB* subtrees, was 0.3, 0.4, 0.36, 0.4, and 0.35 respectively, which is a significantly large number than the average of 0.09 amongst *nifDKENB*. We attribute the differences amongst *nifH* subtrees to the grouping of Desulfuromonadales homologs with γ -Proteobacteria and Rhodocyclales, and the grouping of Burkholderiales homologs together with the main α -Proteobacteria subtree (Supplementary Figure S1). Overall, the consistency in which homologs from similar sets of genomes were independently grouped in the 35 assessed N_2 -fixation gene families provides additional evidence for multiple distinct origins of N_2 -fixation genes within Proteobacteria.

To identify sets of genomes sharing similar origins for their N_2 -fixation genes, we assembled a network based on shared ancestry of N_2 -fixation genes (Figure 2). Edges were placed between nodes (i.e., genomes) if homologs in both genomes are grouped in the same subtree, potentially sharing a common origin within Proteobacteria. The greater the edge weight between two genomes the more homologs in both genomes occur in the same subtree. We clustered the genomes in groups with similar origins for N_2 -fixation genes using the Louvain clustering algorithm (Blondel et al. 2008), which predicted five clusters of diazotrophic genomes. The assembly of genome clusters means that all members of the same cluster share the origin of a significant proportion of their N_2 -fixation gene set, suggesting five groups of Proteobacteria with distinct origins of gene sets. To assess the robustness of the clustering, we generated 100 random permutations of the edges, with repetitions, and compared the obtained clusters for each permutation. All 100 permutations generated the same

five clusters of genomes.

The five clusters of genomes sharing N_2 -fixation gene origins identified in this study are referred as clusters *A*, *B*, *C*, *D* and *E* (Figure 2). Interestingly, four of the five clusters correspond to the phylogenetic signal of their genomes. That is, clusters *A*, *C*, *E* and *D* are represented almost exclusively by δ , α , ϵ -Proteobacteria, and Burkholderiales genomes respectively. We hypothesize that the well-defined taxonomic delimitation of genome clusters with distinct origins of N_2 -fixation genes was caused by independent acquisitions of such genes by ancestors of each cluster. Alternatively, this distribution could be attributed to an independent acquisition by one member after speciation followed by horizontal radiation to other members of the same cluster. However, cluster *B* is represented by a large diversity of genomes from γ , α , ζ -Proteobacteria, and Rhodocyclales. Given the larger diversity of γ -Proteobacteria genomes, we propose N_2 -fixation genes were acquired by a γ -Proteobacteria ancestor and horizontally spread to other clades.

Influence of respiratory requirement on the evolution of nitrogen fixation genes

Nitrogenase metalloclusters are readily inactivated in the presence of O_2 (Gallon 1981); hence, organisms inhabiting oxic environments apply a distinct strategies to protect nitrogenases from O_2 -induced inactivation (Dixon et al. 1980; Stal and Krumbein 1985; Curatti et al. 2005). We sought to answer the following question: was the O_2 requirement a determining factor in the shared evolution of dedicated nitrogen fixing genes? To answer this question, we compared the respiratory requirement of the five clusters of Proteobacteria genomes, which had distinct origins for their N_2 -fixation gene sets (Figure 2). With the exception of genomes from cluster *B*, all groups exhibited a correlation between the shared origin of nitrogen fixation genes with the physiological lifestyle of the host organism. For example, all genomes represented by cluster *A* are anaerobic δ -Proteobacteria, with a N_2 -fixation gene set origin that has not been frequently shared with other classes of Proteobacteria. Similar scenarios of taxonomic and O_2 requirement homogeneity occur in two other genome clusters with shared origins of N_2 -fixation genes (*C* and *D*), both with minimal taxonomy discrepancies: Burkholderiales and *Dechlorosoma sulum* in cluster *D*, and α -Proteobacteria and *Acidithiobacillus* spp. in cluster *C*. In

both clusters, all members are either aerobes or facultatives, with the exception of *Rhodospirillum photometricum* in cluster *C*, which is able to grow only under very low oxygen tensions (Garrity 2006). Although genomes from cluster *B* exhibit significantly larger taxonomic diversity, represented mainly by γ -Proteobacteria and Rhodocyclales, only three of these organisms are anaerobes: *Halorhodospira halophila*, *Allochromatium vinosum*, and *Thiocystis violascens* (Figure 2). These three anaerobes from cluster *B* do not contain N_2 -fixation gene sets as complex as the aerobic/facultative members of cluster *B* (Figure 1), which highlights the influence of O_2 requirements on the complexity of the N_2 -fixation gene set. Collectively the taxonomic distribution along with physiological lifestyle (exploited in this study as O_2 requirement) observed in the five clusters provides support for the hypothesis of multiple origins of N_2 -fixation genes within Proteobacteria.

Horizontal Gene Transfers

The occurrence of HGT was investigated in the distinct clusters of N_2 -fixation genes within Proteobacteria. Each subtree was submitted to phylogenetic reconciliation with the Proteobacteria tree (Figure 1) using the Duplication-Transfer-Loss (DTL) approach to infer gene losses, duplication, and transfers among members of the subtree (Bansal et al. 2012). A total of 585 horizontal transfer events of genes related to N_2 fixation were predicted to have occurred within Proteobacteria. The majority of the predicted transfers were expected to have taken place between closely related donors and recipients. The sum of branch lengths corresponding to the longest predicted transfer corresponds to approximately 1.8 substitutions per site on the Proteobacteria tree (Figure 1). For example, this HGT corresponds to a possible transfer of *nifO* from an *Arcobacter* ancestor to *A. vinelandii*, which is equivalent to less than 30% of the branch length between the pair of most distant leaves on the Proteobacteria tree (approximately 6.05 substitutions per site). In fact, 80% of all predicted horizontal transfers are hypothesized to have occurred between donor and recipients separated by less than 20% the distance separating the most distant leaves (approximately 6.05 substitutions per site).

Predicted HGTs were inferred to have occurred with similar relative frequencies among α (8.7%), β (11.6%), and γ -Proteobacteria (13.1%) lineages. We defined such relative frequencies as the ratio of predicted HGTs in a given Proteobacteria class divided by the number of diazotrophic

genomes in the class. This ratio is not expected to depict the average number of predicted transfers in specific genomes, but roughly portrays how many transfers would be necessary for the current distribution of genes given the number of diazotrophs in each class. The relative frequency of predicted HGTs involving δ -Proteobacteria (3.2%) is significantly smaller than in β -proteobacteria, even though there are 16 diazotrophs in both classes. While 186 HGT events were predicted to involve β -Proteobacteria genomes, only 51 were predicted to involve δ -Proteobacteria genomes. We explain such disparity by hypothesizing that vertical inheritance is likely to have had a significantly greater importance in the radiation of diazotrophy among δ -Proteobacteria than among other taxonomic classes. We did not make assumptions regarding ε - and ζ -Proteobacteria due to the small number of diazotrophs in each class (five and one, respectively).

The majority of transfers between distinct classes of Proteobacteria were predicted between α , β , and γ -Proteobacteria (Figure 3). Among the 464 predicted gene transfers we identified an enrichment of transfers between β and γ -Proteobacteria ($p = 0.012$, Figure 3), characterizing a possible “highway” of N_2 -fixation gene transfers, similar to that previously reported in the literature for other bacterial genes (Beiko et al. 2005; Dagan 2011; Bansal et al. 2013). This possible highway of gene transfer involves 50 predicted transfers between β and γ -Proteobacteria internal and terminal nodes, which respectively have 19 and 16 diazotrophic genomes.

We limited our analysis to Proteobacteria due to the large proportion of diazotrophic organisms when compared to other phyla, accounting for the majority of diazotrophs with sequences genomes (Dos Santos et al. 2012; Boyd et al. 2015; Mukherjee et al. 2017). Recent reports regarding the evolution of nitrogen fixation provided valuable insight into the general evolutionary characteristics and its origins within the tree of life (Fani et al. 2000; Henson et al. 2004; Raymond et al. 2004; Boyd, Anbar, et al. 2011; Boyd, Hamilton, et al. 2011; Boyd et al. 2015). Here we focused strictly on the evolutionary processes driving the evolution of nitrogen fixation in Proteobacteria, allowing for more detailed analysis. Our restriction to a single phylum enabled the implementation of computationally intensive analysis, as the construction of homolog groups using orthoMCL and identification of HGTs using phylogenetic reconciliation, would not be feasible with a much larger dataset.

Differential transfers of *nifHDKENB*

The strict requirement of the MNS for the nitrogen fixation process (Dos Santos et al. 2012), prompted us to investigate if the evolutionary history of these six genes (*nifHDKENB*) is linked, especially for those gene pairs forming protein quaternary structures such as NifDK and NifEN. However, our analysis did not confirm this assumption. For example, only seven predicted transfers of both gene families were inferred between similar pairs of donors and recipients. In comparison, each individual *nifD* or *nifK* gene was predicted to have been transferred an additional 19 and 21 times, respectively. We considered predicted transfers of both genes together if both pairs of donors and recipients are in the same branch (e.g. donor of *geneA* is either immediate ancestor or immediate descendant of *geneB* donor). When predicted donors and recipients of both assessed genes are in the same branch, on average donors are separated by 0.08 bipartitions and recipients by 0.15 bipartitions. This implies evolutionary flexibility in the subunit composition of nitrogenase, which may allow enzyme fine tuning in response to different host requirements. A slightly different scenario was predicted regarding genes in the *nifEN* operon, whose products similarly form a E_2N_2 heterotetramer. Our phylogenetic reconciliation analysis predicted that both *nifEN* genes were horizontally transferred 15 times between similar pairs of donors and recipients. The predicted 15 shared transfers between *nifEN* account for roughly half of the predicted HGTs for each individual gene, which, given the close interactions between their products, is far less than anticipated.

From the 45 genomes predicted to have received at least one MNS gene, 64.4% are predicted as recipients of only three or fewer MNS genes, which can indicate a sequential refinement of the enzyme complexes, and consequently of the N_2 -fixation gene set. Protein products of *nifDK* form a dinitrogenase heterotetramer $\alpha_2\beta_2$. Despite displaying highly compatible phylogenetic trees and encoding subunits of the same protein complex, 14 genomes were predicted to have horizontally acquired either only *nifD* or only *nifK*. Differential transfers of *nifDK*, or *nifEN* lead to a protein complex whose subunits have diverging origins, an unexpected result given the close interactions required for the function of this heterotetramer.

On average, predicted horizontal transfers of the complete *nifHDKENB* gene set would have

occurred between similar pairs of donors and recipients less than 25% of the time each individual gene was predicted to have been transferred. This small percentage of shared predicted transfers among MNS genes highlights some level of interchangeability among MNS homologs. Given the highly conserved amino acid sequences among MNS homologs and the significant frequency in which they are horizontally transferred independently, one can infer that in the context of Proteobacteria the horizontal exchange of enzyme subunits and cofactor biosynthetic proteins does not cause a significant decrease in fitness. Besides the potential for exchange of MNS components, the small percentage of shared predicted transfers can be explained by the natural trend of genes being transferred in small sets, as are transfers of large chunks of genes rather rare (Kunin et al. 2005; Popa et al. 2011; Popa et al. 2017).

Recent transfers

Close to half of the predicted HGTs (44.5%) occurred between two terminal nodes of the Proteobacteria tree, which we treated as recent transfers given their predicted occurrence between two existing lineages. We plotted these predicted recent HGT events in the network presented in Figure 4. Similarly to the trend identified among the full set of horizontal transfers predicted by phylogenetic reconciliation analysis, few genomes were predicted to have participated in large numbers of transfers (thus constituting heavily connected nodes), either as donors or recipients. For example, about 20% of the nodes are connected to more than half the edges in Figure 4. A similar node connectivity distribution has been discussed for HGTs networks (Beiko et al. 2005; Kunin et al. 2005; Dagan 2011; Popa and Dagan 2011). The two most connected nodes are *Tolomonas auensis* (involved in 46 predicted HGTs) and *Pectobacterium atrosepticum* (involved in 55 predicted HGTs). Both are γ -Proteobacteria, whose connectivity is more than twice the predicted in the third most connected node (*Burkholderia vietnamiensis*, related to 19 predicted recent transfers). Although *T. auensis* and *P. atrosepticum* are predicted to have participated in similar numbers of transfers, they are likely to play distinct roles in the radiation of diazotrophy among Proteobacteria. While *P. atrosepticum* is heavily connected to only six other genomes (all γ -proteobacteria), *T. auensis* shares significantly less weighted adjacencies, with 12 nodes from α , γ , and δ -Proteobacteria. Among the 74 genomes related

to recent transfers 1five do not exhibit a MNS genes, thus theoretically not diazotrophs. Which despite not being capable of fixing nitrogen can still transfer/receive genes participating in additional pathways. *Thauera* sp. MZ1T is an example of a non-diazotroph that is predicted to have been the donor of *rnfACD1* to the diazotrophy *Halorhodospira halophila*.

A sizable amount of diazotrophs (23 genomes, 26.4% of diazotrophs present in our sample) were not involved in any of the recent transfers. This lack of predicted HGTs involving some genomes is likely due to two potential scenarios: [1] 100% vertical inheritance of the N_2 -fixation genes; or [2] possible horizontal acquisition from non-Proteobacterial lineages, which were not investigated in this study. For example, even though *Azonexus hydrophilus* was not predicted as recipient of any computed recent transfer, it is unlikely that members of its lineage are resistant to HGT given the number of predicted horizontal acquisitions by its ancestors (*nifDXMQ*, *FeSII*, *IscAnif*, *cysE1*, *rnfA1*, *Avin_01370*, *Avin_50900*). The closest diazotrophic genome to *A. hydrophilus* in the ribosomal tree (Figure 1), *Dechloromonas aromatica*, was predicted to have recently acquired eight genes through HGT. Among the genes predicted to have been recently acquired by *D. aromatica*, only the gene family called *orf* (homologs of *Avin_01430* in *A. vinelandii*) seems to be a completely new acquisition, since it was not identified in *A. hydrophilus* nor its horizontal acquisition was predicted by its ancestors. The remaining seven genes likely correspond to Xenologous Gene Displacement (XGD) of genes likely present *D. aromatica* ancestors (*nifHY* and *rnfABDGE1*). Beyond the 23 diazotrophs with no predicted involvement in recent HGTs, another 36 diazotrophs were implicated in five or fewer transfers, either as donors or recipients. Regarding genes from the MNS, only 45 out of the 87 diazotrophs in our sample are recipients of one or more genes. In support of our previous observation that the genes *nifHDKENB* in general are not transferred as a set, only three genomes were predicted to have recently acquired all six genes (*Cupriavidus taiwanensis*, *Pantoea* sp., and *Pectobacterium atrosepticum*).

Ecological impact on HGT of diazotrophy related genes

We identified a significant preference of gene exchange between genomes from diazotrophs that

exist in similar ecological environments, which we assessed through shared O_2 requirements (Figure 4). Given the 259 predicted recent transfers among Proteobacteria genomes, facultative anaerobes, aerobes, and anaerobic bacteria displayed significant preferences in exchanging N_2 -fixation genes with bacteria having similar lifestyles ($p < 0.05$). Along with the enrichment of predicted transfers between genomes with similar O_2 requirements, we identified an underrepresentation of HGT predictions between anaerobes and bacteria colonizing oxic environments, for both aerobes and facultative anaerobes ($p < 0.01$). These results show an overrepresentation of predicted HGTs within bacteria with similar O_2 requirements and underrepresentation of these events between aerobes and anaerobes. These observations suggest that distinct habits encountered by aerobes versus anaerobes limit the opportunities for gene exchange, whereas, overlapping niches provides more opportunities for HGT events to occur (Beiko et al. 2005; Popa et al. 2011).

Conclusions

In this study we have provided a detailed investigation into the evolutionary trajectory of genes dedicated to nitrogen fixation within species of Proteobacteria. Analysis of a large dataset within a single taxonomic phylum using genes whose function is proposed to be restricted to nitrogen fixation enabled identification of five independent lineages of N_2 -fixation gene sets (Figure 2). The paraphyly of Proteobacteria *nif* genes have been previously suggested in the literature with six Proteobacteria clades identified through analysis of the *nifHDK* genes (Figure 2 present in Boyd et al. 2015). Here we present strong evidence that Proteobacteria genomes were robustly partitioned into five distinct origins of N_2 -fixation genes according to shared ancestry and lifestyle. While we were not able to assess whether diazotrophy was independently acquired in each group or whether these origins have successively diverged by XGDs, both scenarios are supported by the HGT enrichment within taxonomic classes. HGT analyses indicated the occurrence of a highway of gene transfer relating γ -Proteobacteria genomes to β -Proteobacteria, which is supported by the similarity between their sets of N_2 -fixation genes.

The γ -Proteobacteria class displayed the most complex N_2 -fixation gene set (Figure 1, and Boyd 2015). This complexity also extends to a subset of β -Proteobacteria (*Sideroxydans lithotrophicus* +

Rhodocyclales, with the exception of *Dechlorosoma suillum*) as the number of N_2 -fixation genes in these genomes is significantly larger than all other Proteobacteria. The opposite side of the complexity distribution is represented by δ -Proteobacteria genomes, which have the smallest set of N_2 -fixation genes amongst Proteobacteria. All δ -Proteobacteria diazotrophs are anaerobes, and the simplicity of the N_2 -fixation gene set in organisms with anaerobic metabolism has been discussed in previous studies (Fani et al. 2000; Raymond et al. 2004; Boyd et al. 2015).

Another distinction of this study is our approach to HGT prediction, driven by shared origin inference prior to the robust phylogenetic reconciliation method. Besides yielding the multiple origins of N_2 -fixation gene sets among Proteobacteria, such analysis prevents an overestimation of horizontal transfers, since homologs independently acquired from non-Proteobacteria lineages were not reconciled in a single tree. Given the promising results presented in this study regarding distinct evolutionary paths of N_2 -fixation genes within Proteobacteria, we intend to extend the scope of our approach to include genomes from other phyla besides Proteobacteria in future analyses.

Materials and Methods

Dataset

Sequences from complete genomes of 468 Proteobacteria species available on GenBank on March 2013 were limited to one representative per species (Supplementary Table S2). Eleven additional Proteobacterial complete genomes were downloaded from IMG (Supplementary Table S3).

Protein coding sequences from the 468 genomes obtained from GenBank were clustered into 83,803 homolog groups using orthoMCL (Li et al. 2003; Fischer et al. 2011). Protein coding sequences from the 11 additional IMG genomes were mapped and included into the previously generated homolog groups.

Phylogenetic reconstruction

We used the concatenation of 35 ribosomal proteins (Supplementary Table S4) to re-construct the phylogeny of the sampled Proteobacteria genomes. Multiple Sequence Alignment (MSA) was

performed using MAFFT (Katoh et al. 2002) and automatically edited by GUIDANCE (Penn et al. 2010). The 7,063 concatenated sites were submitted to RAxML (Stamatakis 2006) for phylogenetic reconstruction using the PROTGAMEAAUTO substitution model and support assessed using aLRT SH-like method (Guindon et al. 2010).

Recruitment of non-Proteobacteria homologs

Amino acid sequences of all members of each homolog group were aligned with MAFFT and the resulting MSA submitted to HMMSCAN (Mistry et al. 2013) to generate HMM profiles. Each profile was searched against *nr* using HMMSEARCH (Mistry et al. 2013) and non- Proteobacteria hits above the default threshold were appended to the respective homolog group.

Variation of Information

We assessed the consistency of the possible multiple origins of *N*₂-fixation genes among Proteobacteria using the Variation in Information metric (Meilă 2007). The same metric was used to measure the consistency of the Louvain clustering used in the shared origins network (Figure 2).

HGT inferences

HGT events within Proteobacteria were inferred independently for each common origin subtree using the RangerDTL software (Bansal et al. 2012) with default penalties. We only accounted HGT events corroborated by 70% support on the gene tree.

HGT enrichment

We assessed the significance of HGT enrichment between and within groups of Proteobacteria using hypergeometric distribution. As such, we calculated the Cumulative Density Function (CDF) to test the probability of *k* transfers occurring between sets of genomes *i* and *j* using the following Probability Mass Function (PMF):

$$\frac{\binom{K_{ij}}{k_{ij}} \binom{N_{ij} - K_{ij}}{i * j - k_{ij}}}{\frac{N_{ij}}{i * j}}$$

where *N_{ij}* is how many pairwise combinations contain genomes from group *i* or *j*, *K_{ij}* is how many transfers involve members from *i* and/or *j*, and *k_{ij}* is the number of transfer connecting *i* and *j*.

Using a significance of 0.05 as cutoff on a two-tailed test, we assumed sub representation if $CDF \leq 0.025$ or super representation if $CDF \geq 0.975$.

Figure Legends

Figure 1: Phylogenetic tree of Proteobacteria obtained from concatenation of 35 ribosomal protein sequences. Nodes with aLRT support > 90% are highlighted in dark red. The tree was pruned to display only the 87 diazotrophic genomes from original sample. Leaves are colored based on taxonomic classes: blue for α , green for β , brown for γ , red for δ , orange for E^- , and purple for ζ -Proteobacteria. When present, colored circles preceding leaf names represent O_2 requirement: facultative anaerobes as green, aerobes as red, anaerobes as orange, and microaerophilic as blue. The binary heat map corresponds to the presence (black) or absence (white).

Figure 2: Visual representation of the network of shared origins of N_2 -fixation related genes. Nodes correspond to Proteobacterial genomes and edges connecting two nodes indicate they at least one of their genes have a common origin. All nodes on the network are diazotrophs, and their contour is colored based on clusters of genomes sharing the same gene origin. Shape, colors, and sizes of the nodes are representing O_2 requirement, taxonomic class, and degree centrality of each genome, respectively. **shape:** pentagon for facultative anaerobes, triangle-up for aerobes, triangle-down for anaerobes, squares for microaerophilics, and circle for unknown. **color:** blue for α , green for β , brown for γ , red for δ , orange for E^- , and purple for ζ -Proteobacteria.

Figure 3: Visual representation of the enrichment of predicted horizontal transfers between taxonomic classes of Proteobacteria. Nodes correspond to taxonomic classes of Proteobacteria and edges connecting two nodes indicate predicted transfers between genomes from both classes. The larger the width of the edges the more HGTs were predicted. **Color:** blue for α , green for β , brown for γ , red for δ , orange for E^- , and purple for ζ -Proteobacteria.

Figure 4: Visual representation of the network of horizontally transferred N_2 -fixation related genes. Nodes correspond to Proteobacterial genomes and edges connecting two nodes indicate at least

one HGT was predicted between them. Nodes corresponding to diazotrophic genomes are highlighted with black contours. Shape, colors, and sizes of the nodes are representing O₂ requirement, taxonomic class, and degree centrality of each genome, respectively. **shape:** pentagon for facultative anaerobes, triangle-up for aerobes, triangle-down for anaerobes, squares for microaerophilics, and circle for unknown. **color:** blue for α , green for β , brown for γ , red for δ , orange for E-, and purple for ζ -Proteobacteria.

Figures

Figure 1

Figure 2

Figure 3

Figure 4

4.2 Discussão

O agrupamento de genomas baseado na origem comum de genes ligados à fixação de N_2 gerou cinco grupos de genomas com origens distintas para suas maquinarias de fixação de N_2 (Figura 4.2). Cada um destes cinco grupos potencialmente representa aquisições independentes dos genes ligados a diazotrofia a partir de diferentes linhagens, com prováveis origens em outros filos. Genomas agrupados em quatro destes *clusters* mostraram-se consistentes com as classificações taxonômicas de Alpha, Delta, Epsilonproteobacteria, e Burkholderiales. Esta relação entre origem dos genes de fixação de N_2 e taxonomia sugere que estes genes foram obtidos por suas linhagens ancestrais. Apenas um *cluster* possui composição taxonômica diversificada, basicamente Gammaproteobacteria e Rhodocyclales. A maior diversidade de Gammaproteobacteria e a localização dos mesmos mais próximos das raízes de suas respectivas sub-árvores (*e.g.*, Figura Suplementar S1, capítulo 4) sugere a aquisição horizontal destes genes por uma linhagem de Gammaproteobacteria, posteriormente compartilhados com outros *taxa*.

Dadas as cinco origens distintas observadas, nossa predição de THG independente para cada sub-árvore observada produz um cenário potencialmente mais robusto, evitando falsas predições causadas por genes com origens distintas em uma mesma árvore. Neste cenário, inferimos 585 transferências horizontais entre genomas de Proteobacteria, não abordamos transferências envolvendo outros filos. As transferências diferenciadas entre genes codificando diferentes subunidades de um único complexo reflete a intercambialidade destas subunidades, provavelmente causada pelo auto grau de conservação dos genes.

Ao analisarmos as 260 transferências preditas entre nós terminais da árvore de Proteobacteria observamos que proximidade evolutiva está mais relacionada a frequência de THG que necessidades similares de O_2 , de acordo com o descrito na literatura (Popa *et al.*, 2017). Os enriquecimentos de THGs entre genomas de uma mesma classe taxonômica mostraram-se significantes, assim como a provável participação de genes ligados a diazotrofia na *highway* de transferência horizontal entre Beta e Gammaproteobacteria (Beiko *et al.*, 2005).

Capítulo 5

História evolutiva de efetores do Sistema Secretor do Tipo III em *Aeromonas* spp.

5.1 Contextualização

Neste capítulo, apresentamos um manuscrito também em fase final de preparação no qual detalhamos a identificação de 21 famílias de efetores do SST3 em 105 genomas de *Aeromonas*. Além da identificação *in silico* das famílias de prováveis efetores, junto com nossos colaboradores apresentamos validações da citotoxicidade observada em *saccharomyces cerevisiae*. Toda a validação experimental foi realizada pelo Dr. Joerg Graf e Dr. Jeremiah Marsden.

A evolução dos efetores do SST3 ainda não é profundamente compreendida. Genes que codificam efetores possuem características similares a genes ORFãos (Fischer and Eisenberg, 1999), e também são geralmente associados a fagos (Ehrbar and Hardt, 2005), ligando-os a THG. Ambas classes de genes possuem diversas complicações quando se estuda suas histórias evolutivas, tanto a evolução *de novo* em regiões não codificantes, quanto regiões que sofrem múltiplas THGs entre linhagens não relacionadas (Daubin and Ochman, 2004; Brown and Finlay, 2011).

Detection, Analysis and Experimental Assessment of Putative T3SS Effectors in 105 *Aeromonas* Strains

Authors: Luiz Thiberio Rangel^{1,2,3,\$}, Jeremiah N. Marden^{1,\$}, Sophie M. Colston¹, João Carlos Setubal^{2,3}, Joerg Graf^{1,4,*}, J. Peter Gogarten^{1,4,*}

Affiliations:

1 Department of Molecular and Cell Biology, University of Connecticut

2 Interunidades em Bioinformática, Universidade de São Paulo

3 Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo

4 Institute for Systems Genomics, University of Connecticut

\$ These authors contributed equally

* Corresponding authors

Address: 91 North Eagleville Rd, Storrs, CT 06269-3125, USA

Email addresses:

lthiberiol@gmail.com

jeremiah.marden@uconn.edu

sophie.colston.ctr@nrl.navy.mil

setubal@iq.usp.br

joerg.graf@uconn.edu

gogarten@uconn.edu and jpgogarten@gmail.com

Keywords: gene transfer, heterologous expression, cytotoxicity, type three secretion system,
T3SS effector

Abstract

This article describes a bioinformatics pipeline to identify Type III Secretion System (T3SS) effector proteins, applies this approach to 105 *Aeromonas* genomes, and evaluates the cytotoxicity of the identified effectors through heterologous expression in yeast. The genetic determinants of pathogenicity are highly variable between strains and species. A factor commonly associated with virulent bacteria, including many *Aeromonas* spp., is the presence of a T3SS, which injects effector proteins into target eukaryotic cells. Numerous types of effectors have been identified and characterized to date. This study describes a novel bioinformatics pipeline to examine the genomes of *Aeromonas* strains isolated from environmental, mutualistic, or pathogenic contexts to identify T3SS effectors. The developed pipeline uses a two-step approach: first candidate families are selected using HMM profiles with minimal similarity scores against the Virulence Factors DataBase (VFDB), followed by strict comparisons against positive and negative control datasets, greatly reducing the number of false positives. Our approach identified a total of 21 *Aeromonas* T3SS likely effector groups, of which 8 represented known or characterized effectors, while the remaining 13 were not previously described in *Aeromonas*. We experimentally validated our *in silico* findings by assessing the cytotoxicity of representative effectors in *Saccharomyces cerevisiae* BY4741. Out of 21 assayed proteins, 17 elicited a cytotoxic effect in yeast. This study demonstrates the utility of our approach, combining a novel *in silico* search method with *in vivo* experimental validation, and will be useful in future research aimed at identifying and authenticating bacterial effector proteins from other genera.

Introduction

Aeromonas spp. are Gram-negative γ -proteobacteria, many of which are of increasing clinical significance as emergent human pathogens (Hiransuthikul et al. 2005; Igbinosa et al. 2012). *Aeromonads* are present in diverse habitats and interact with a variety of organisms such as leeches, zebrafish, vultures, and humans, participating in both mutualistic and pathogenic symbiotic interactions with their host species (Hänninen et al. 1995; Janda and Abbott 1998; Graf 1999; Janda and Abbott 2010; Roeselers et al. 2011; Marden et al. 2016). *Aeromonas* species cause diseases in a variety of animals, and are particularly associated with fish diseases caused the psychrophilic species *A. salmonicida* as well as mesophilic ones such as *A. hydrophilia*, *A. sobria* and *A. veronii* (Wahli et al. 2005; Janda and Abbott 2010). In humans, *Aeromonas* strains are commonly identified as causative agents of traveler's diarrhea, septicemia, as well as other severe infections such as necrotizing fasciitis (Vila et al. 2003; Igbinosa et al. 2012).

Numerous important *Aeromonas* virulence factors have been identified, including aerolysin, exotoxins, and Type Three Secretion Systems (T3SS) (Chopra et al. 2009). The T3SS is one of the most important and best-studied bacterial virulence factors (Deane et al. 2010; Chatterjee et al. 2013; Notti and Stebbins 2016), and the diversity of niches inhabited by *Aeromonas* spp. may be due in part to many strains possessing one or more T3SSs (Braun et al. 2002; Burr et al. 2003; Burr et al. 2005; Preston 2007; Silver et al. 2007; Silver et al. 2011; Bomar et al. 2013). The T3SS is a molecular syringe that transfers effectors with a range of biochemical activities into target eukaryotic cells (Dean 2011). T3SSs are frequently associated with horizontal gene transfer (HGT) events and are commonly observed inside pathogenicity islands (Hacker et al. 1997; Hueck 1998; Ehrbar and Hardt 2005). The importance of T3SSs in bacterial-eukaryotic interactions and their frequent horizontal transfer, make them important players in niche adaptation processes (Dale and Moran 2006).

Next generation sequencing has driven remarkable advances in microbiology, both in fundamental research and clinical diagnostics (Lecuit and Eloit 2014; MacCannell 2016). A major objective in the analysis of whole genome sequence data is to better predict the pathogenic potential of bacterial strains

through the identification of virulence factor genetic determinants. The core components of the T3SS apparatus are homologous to the bacterial flagellum, and their sequences are sufficiently similar to cause cross identifications in homology searches. In contrast, the T3SS effectors are less conserved, often displaying uncharacterized domains, which also leads to difficult homology searches. Databases such as the Pathosystems Resource Integration Center (PATRIC) (Wattam et al. 2014), VICTORS (Xiang et al. 2007), EffectiveDB (Eichinger et al. 2016), and the Virulence Factor DataBase (VFDB) (Chen et al. 2016) contain T3SS effector sequences, with different levels of validation. One problem when attempting to identify T3SS components through homology searches is that many proteins and protein domains encoded in bacterial genomes are homologs of T3SS components, but not part of a T3SS system and encode other cellular components such as the bacterial flagellar components (Gophna et al. 2003), F-ATPase subunits (Poptsova and Gogarten 2007), and Type IV pili (Collins et al. 2004; Chami et al. 2005). In this study, we describe an *in silico* approach that solves this problem through the use of a negative control comprised of all protein sequences from *Vibrio fischeri* ES114 and *Escherichia coli* K12. Since these genomes are known to not encode a T3SS, their protein sequences can be used as non-T3SS references.

Here we discuss the distribution and cytotoxicity of 21 T3SS effector families within 105 *Aeromonas* genomes and describe the pipeline developed to identify them. The identified proteins were validated as potential T3SS effectors by expressing representative effectors from each family in the yeast *Saccharomyces cerevisiae* strain BY4741 and assessing cytotoxicity. Out of 21 assayed effectors 13 are newly described, and 17 exhibited cytotoxicity under the conditions assayed. Our findings extend the knowledge of the complement and distribution of T3SS effectors in *Aeromonas* strains, and our combination of a bioinformatics pipeline with verification through heterologous expression in yeast provides a template for studies in other bacterial genera.

Results and Discussion

Reference phylogeny

A phylogenomic study of 56 high quality genomes from *Aeromonas* strains was recently published together with an MLSA tree of 16 housekeeping genes from the same genomes (Colston et al. 2014). This study found incongruities between gene, MLSA, and core genome trees, which is expected since different sets of genes were used in each tree. Our current study expanded on this set with an additional 49 *Aeromonas* genomes. To arrive at a reference phylogeny that reflects the plurality of *Aeromonas* genes and minimizes the impact of HGT leading to conflicting evolutionary paths among the assessed genes, we constructed a phylogenomic tree using 1,678 genes with compatible evolutionary histories. Building phylogenies from concatenated genes is a widely used approach to reconstruct the representative history of a genome. However, the concatenation of genes with different histories can lead to unresolved trees and/or a phylogeny that represents neither the history of the organism, nor that of the individual genes used in the concatenation (Gogarten and Townsend 2005; Kubatko and Degnan 2007; Larget et al. 2010; Lewis et al. 2016). Distance metrics for grouping gene families that take branch lengths into consideration perform better than topology only metrics (Gori et al. 2016). Therefore, to assess the similarity between evolutionary histories of different genes, we calculated pairwise Pearson's correlation tests between all distance matrices of gene families present in at least 10 genomes. For comparison, we also measured distances between gene families using the Fitch-Margoliash criterion (Fitch and Margoliash 1967), i.e., in calculating the sum of squares of the distance differences between two matrices the square of each difference was multiplied by the inverse of the distance, thereby increasing emphasis on the difference in distance between more similar sequences. The two approaches to compare distance matrices (Pearson Correlation between distance matrices and Fitch-Margoliash weighted difference between the distance matrices) showed a high correlation ($r_s = -0.89$). From the 168 gene family clusters obtained through MCL clustering we considered the largest cluster, containing 1,678 gene families, as the main phylogenetic signal among the *Aeromonas* strains. This largest gene family cluster comprises more than twice the number of gene families in the second largest cluster, which contained 738

gene families. By using the largest gene cluster, which contains gene families with different functions and genomic locations but with compatible evolution, we ensure that our reference phylogeny represents the evolutionary history of the plurality of genes present in the genomes.

On average, each *Aeromonas* genome has homologs in 1,637 out of the 1,678 families used in the reference phylogeny ($std = 63.7$), representing 39% of coding sequences in the *Aeromonas* species assayed. The resulting phylogenomic tree (Figure 2) is highly supported, with all but three branches displaying an aLRT SH-like support (Guindon et al. 2010) greater than 0.90. Approximately 92.3% of all internal nodes have a aLRT SH-like support of at least 0.99. The concatenated gene tree presented in this study is largely consistent with the previously published *Aeromonas* spp. core-genome phylogeny (Colston *et al.* 2014), the only difference is the earliest branching of *A. salmonicida* strains in relation to the *A. hydrophyla* and *A. veronii* common ancestor, and *A. allosaccharophila* forming the the sister group to the *A. veronii* – *A. australiensis* clade.

***In silico* identification of putative effectors**

To identify putative T3SS effectors, all of the open reading frames were compared to each other to identify orthologous gene families. Out of 25,518 identified *Aeromonas* spp. gene families, 633 returned positive hits when compared to T3SS sequences from the VFDB. Some of these were clearly not T3SS related, but produced positive hits with e-values as low as 6×10^{-155} due to a common origin of other systems with T3SS components (e.g., flagellar components, F-ATPase subunits, and Type IV Secretion System components). Sequence comparisons of the 633 gene families against the VFDB sequences and against the negative control sequences identified 127 families as part of T3SSs. Through subsequent manual curation, we assessed the domains present in each putative family and their genomic locations (e.g., whether they are adjacent chaperones), and classified 21 gene families as encoding likely effectors (Table 1).

Eight of the 21 effector families had been previously identified in *Aeromonas* species, including: *aexT* (Braun et al. 2002; Burr et al. 2003; Vilches et al. 2008), *aexU* (Sha et al. 2007), *aopP* (Fehr 2006; Jones

et al. 2012), *aopH* and *aopO* (Dacanay 2006), *ati2* (Ebanks 2006), as well as *aopX* and *aopS*, which were initially observed in *A. salmonicida* as pseudogenes (Reith et al. 2008). We also identified *aopN*, but did not consider it further in this study as it was shown in *Bordetella bronchiseptica* to have a dual role in controlling the secretion of translocator proteins and suppressing host immunity but without being cytotoxic (Vanden Bergh et al. 2013).

The remaining 15 likely *Aeromonas* T3SS effector families had not been studied in any detail, and were designated as putative type III effectors: *pteA*, *pteB*, *pteC*, *pteD*, *pteD.1*, *pteE*, *pteF*, *pteG*, *pteH*, *pteI*, *pteJ*, *pteK*, *pteL*, *pteM*, and *pteM.1*. All sequences from homologs of these putative type III effectors and from the 8 previously characterized were submitted to InterProScan (Mulder and Apweiler 2007) for domain identification (Table 1). *pteD.1* and *pteM.1* were combined with the *pteD* and *pteM* families, respectively, because they showed significant sequence similarity in PRSS (Pearson 1996): using the PAM250 scoring matrix, the comparison between PteD.1 from *A. jandaei* Ho603 and PteD from *A. sp.* MDS8 resulted in a Z-score of 65.7, and the comparison between PteM.1 from *A. sobria* CECT4245T and PteM from *A. media* CECT4232T resulted in a Z-score of 1003. Besides the sequence similarities, homologs from both *pteD* and *pteD.1* are likely to related to the same chaperone family, whose members were correctly grouped within a single homolog protein cluster (data not shown). We assessed the cytotoxicity of the identified effectors by expressing representatives of each family in *S. cerevisiae* strain BY4741.

Screening of putative T3SS effector proteins in yeast

To assess the cytotoxicity of the putative T3SS effectors identified in *Aeromonas* spp. genomes we expressed representative proteins from each group (Table 1 and Table S1) in the yeast strain *S. cerevisiae* BY4741 and monitored for growth inhibition. Although eight of the identified effector families have previously been identified or studied, to our knowledge none have been assessed for causing cytotoxicity or growth inhibition in yeast. The expression of bacterial toxins in yeast is a common means of assessing the deleterious impact of such proteins in eukaryotic host cell processes (Curak et al. 2009; Salomon and Sessa 2010; Alam et al. 2011). We used pGREG533 to construct plasmids to express candidate effectors

as N-terminal hemagglutinin tagged (7-HA) proteins in a galactose inducible manner from the GAL1 promoter. Plasmids expressing representatives of each of the 21 established or putative effector protein families (Table S1) were separately introduced into yeast strain *S. cerevisiae* BY4741. Cells from overnight cultures grown under repressing conditions (2% glucose) were washed and 10x serial dilutions were spotted onto agar-solidified plates that repressed (SDM-His-Glu) or promoted (SDM-His-Gal) expression of the cloned protein. While all strains grew on media containing glucose (Figure 3A, B), those expressing *pteD*, *pteF*, and to a lesser extent *pteM*, exhibited growth inhibition under non-inducing conditions, generating no growth or colonies with reduced size relative to cells carrying the pGREG533 plasmid alone. Presumably, these effectors are so cytotoxic that even uninduced, basal protein expression is sufficient to inhibit growth.

Twelve of the 21 proteins assayed exhibited a dramatic growth phenotype when plated on SDM-His-Gal inducing media (Figure 3A, B). Strains expressing the proteins encoded by *aexT*, *aopX*, *ati2*, *aopS*, *pteA*, *pteB*, *pteC*, *pteD*, *pteF*, *pteG*, *pteI*, or *pteM* showed little or no growth on the lowest dilution tested (Figure 3A, B). The strain expressing *aexU* exhibited reduced growth to the 10^{-1} dilution, and in strains expressing *aopP*, *pteH*, *pteJ*, or *pteL* there were discernibly reduced colony sizes with either no or little growth on the 10^{-4} dilution. Strains expressing *aopO*, *aopH*, *pteE*, or *pteK* produced no growth inhibition phenotype in comparison to the control strain harboring pGREG533.

Previous studies have shown that some effectors may not elicit a growth defect phenotype in yeast if the targeted cellular process does not typically limit yeast growth (Slagowski et al. 2008), but that stressors (e.g., elevated salt concentrations or caffeine) can sensitize and to lead to observable growth inhibition phenotypes. The addition of NaCl to SDM-His-Gal media inhibited the growth of strains expressing *aopP* and *pteJ* on the 10^0 dilution, whereas *pteH* and *pteL* had small observable colonies on the 10^{-1} dilution compared to the observed growth on the 10^{-3} - 10^{-4} dilutions when plated on SDM-His-Gal media alone. A similar, but less dramatic pattern was seen in these strains when cells were plated on media containing 7 mM caffeine (Figure 4). One explanation for the observed lack of toxicity is that the expressed proteins were truncated or not expressed. For the constructs that did not produce a toxicity

phenotype, including *aopO*, *aopH*, *pteE*, or *pteK*, we confirmed that bacterial proteins of the expected fusion protein size were expressed in these strains by western blot (Figure S1).

Both AopH and PteK have similar domain structures (Table 1), and similarly did not produced a yeast phenotype. The AopH *Yersinia* homolog YopH has been previously assayed in yeast (Lesser 2001) and was not cytotoxic to cells. However, the AopO *Yersinia* homolog (YopO/YpkA) was shown to elicit a cytotoxic effect in yeast (Lesser 2001; Nejedlik et al. 2004), whereas the *Aeromonas* AopO homolog assayed in this study did not. Several scenarios could account for the lack of observed phenotypes for AopO, AopH, PteE, and PteK in these strains, including: a lack of activity due to the adjoined 7-HA tag; the lack of a target protein for the effector in the yeast strain assayed; the effector produced a non-cytotoxic effect but interacted with host cells in another manner; the protein was not expressed at a high enough level; or, the putative effector is simply was not toxic to cells.

These assays allowed for rapid evaluation of likely T3SS effectors identified in our bioinformatics analysis. These findings indicate that 17 out of the 21 tested proteins inhibited growth of *S. cerevisiae* BY4741, and that addition of NaCl, and to a lesser extent caffeine, to SDM-His-Gal plates sensitized yeast cells to 4 of these likely effectors. Further assessment of effectors that did not yield phenotypes should likely focus on generating C-terminal tagged proteins since the N-terminal tags generated in the current study could have caused misfolding, aberrant activity, or localization of proteins and may potentially mask a reduced growth or cytotoxicity phenotype. Additionally, alternate yeast strains could be tested.

Distribution and evolutionary history of putative effectors in *Aeromonas*

The scattered occurrence of the 21 likely effectors described in this study along the *Aeromonas* phylogenomic tree is evidence of the impact of HGT during their evolution. Only 15 out of the 21 likely effectors are present in at least four taxa, thus our HGT predictions using phylogenetic reconciliation are restricted to these 15 likely effector families. For effectors present in less than four genomes, we only considered their presence/absence in inferring gene transfers and loss events. The first occurrence of an effector inside the *Aeromonas* genus is most likely due to HGT from outside the Aeromonads. However,

in our reconstructions this first occurrence is not assigned to an HGT event. When no HGT was allowed during the reconciliation, a total of 1,177 gene loss and 105 duplication events are required to reconcile the 15 genes with four or more representatives with the *Aeromonas* phylogeny. When using the default penalties of RangerDTL2 (loss = 1, duplication = 2, and transfer = 3) a significantly more parsimonious scenario was obtained by reconciliation (Table 1). Given the uncertainty regarding the root placement in trees from the putative effectors, we ran RangerDTL2 multiple times for each root leading to the most parsimonious reconciliation. When accounting only for HGT events supported by at least 90% of the reconciliations for each effector a total of 38 HGT events are predicted to have occurred among *Aeromonads*. The sum of the mean number of transfer events predicted in the reconciliations for each effector family is 114. Horizontal exchange of the assessed effectors are more likely to occur between closely related *Aeromonads*. 75% of all predicted transfers, among all optimal resolutions, are expected to have occurred between nodes separated by less than 30% the total branch length separating *A. schubertii* CECT4240T and *A. sobria* 2014105092720, the two nodes most apart in the *Aeromonas* phylogeny (Figure 2). In contrast to the expectation that transfer rates are highest between closely related organisms (Andam and Gogarten 2011), we did not infer any within species transfers; however, the reason for this is that if a strong phylogenetic signal is absent, our approach to reconciliation assumes a phylogeny for the effectors that is most similar to the reference tree. Reconciliations also revealed a large variance of HGT events inferred among T3SS effector gene families, ranging from 0 to 15 (*aexU*) horizontal transfer events with at least 90% confidence within *Aeromonas* spp (Table 1).

The inclusion of homologs from other genera revealed that *Aeromonas* T3SS effectors were not frequently shared outside the genus. In all gene trees for the predicted T3SS effectors, the *Aeromonas* homologs grouped into clans, excluding homologs from other taxa. This suggests that the likely effectors each have a single origin in the *Aeromonas* strains, i.e., they entered the genus only once through a single transfer event. The disparity between intra-genus transfers and between genera is well described in the literature (Andam et al. 2010; Williams, Fournier, Lapierre, Swithers, Green, et al. 2011; Polz et al. 2013). The high number of inferred gene transfers within *Aeromonas* genomes is evidence of the

interchangeability of T3SS effectors among the genus, while the paucity of HGT events between genera may either reflect decreased transfer frequency with increasing evolutionary distance (Andam and Gogarten 2011; Williams, Fournier, Lapierre, Swithers, Grigoriev, et al. 2011; Polz et al. 2013) or the high degree of specificity between effectors and the secretion system apparatus. We recognize the possibility that some between genera transfers might have occurred with un-sampled lineages not represented in public databases, but if exchanges involving taxa outside the *Aeromonas* were common we would expect to observe at least some paraphyletic aeromonad clades.

Five likely effector families exhibited highly supported HGT events within *Aeromonas* spp. according to phylogenetic reconciliations. The most exchanged effector is *aexU*, with 15 inferred transfers within the genus. The reconciliation analysis inferred 14 HGT events for the evolution of *aopX* in *Aeromonas*. Three of the identified likely effectors, *pteA*, *pteB*, and *pteE* exhibited large differences between the number of highly supported transfers and the mean number of predicted transfer events (Table 1). Such differences were caused due to the small support of the reconstructed phylogenetic gene trees. The low support observed in these trees caused many possible tree topologies to lead to equally parsimonious reconciliations. Although the number of transfer predicted in each reconciliation is consistent, the same is not true for the mapping of the events themselves. Among the 200 reconciliations ran for the *pteB* tree, all results predicted 18 transfer events between *Aeromonas*, but there is no consistency regarding which nodes constitute the “correct” pairs of donors and recipients.

Aeromonas effectors tend to form clusters of co-occurring genes (Figure S2). In part this is due to shared ancestry (e.g., a clade of *A. salmonicida* strains share *aexT*, *ati2*, *aopP*, *aopH*, and *aopO*); however, in part the co-occurrence appears to be largely independent of phylogeny (e.g., *pteF*, *pteC*, *pteD*, *pteJ*, *pteL*, *pteI*) and may reflect structural or functional associations between the effectors. All putative effectors, with exception of *pteE* and *pteM*, significantly co-occur with other putative effectors in *Aeromonas* genomes ($p < 0.05$). One cluster of six putative effectors, *aexT*, *ati2*, *aopP*, *aopH*, and *aopO* occurs in the *A. salmonicida* branch (Figure 2). Interestingly, one of the effectors in this group, *aexT*, is also prevalent within the *A. veronii* strains. Another co-occurring cluster is comprised of 5 putative

effectors in 2 different branches of the phylogeny. The cluster is composed of *aexU*, *pteA*, *pteB*, *aopX*, and *pteK*, which are present in several *A. veronii*, *A. hydrophila* and *A. dhakensis* strains. While these five putative effectors are present in distinct *Aeromonas* species, they tend to be present in phylogenetically close strains (Fig 2). The phylogenetic signal observed in the effector's co-occurrence in part may reflect vertical inheritance within the group and in part be due the higher within-species HGT frequency; once the effector is acquired by a member of the species it is easily spread among closely related genomes (Andam and Gogarten 2011; Popa et al. 2017). The other two clusters shown in Figure S2 are composed of effectors that occur less frequently and thus we cannot reliably infer what led to their co-occurrence. *A. allosaccharophylla*, we infer the presence of three putative effectors (*aexU*, *pteK*, *pteE*); however, structural T3SS components were not detected in this genome (Colston 2016). A similar finding was previously discussed for *Pseudomonas aeruginosa* (Burstin et al. 2015). The presence in effectors in organisms that theoretically are not capable to secrete them via T3SSs could be due to such effectors possessing additional roles within distinct pathways, or due to a recent loss of the T3SS apparatus genes.

Based on our sampled genomes and their isolation data, we are unable to identify links between the putative effector occurrence and isolation sources. The lack of associations between effectors and habitats are not exclusive to aeromonads and have been previously discussed (Sadarangani et al. 2013). However, two of the co-occurring clusters of putative effectors reflect their genome's phylogeny. The great variation in effector occurrence among *Aeromonas* strains, from zero to nine effectors per genome, may reflect the diversity of lifestyles observed within the genus. *A. schubertii* possesses nine likely effectors in its genome, more than any other assayed strain, and is also the earlier branching genome. We are not able to determine if this large number of predicted T3SS effectors is related to the niche *A. schubertii* occupies (cultured from a clinical forehead abscess), since no other sampled strain was isolated from a similar source. At the other end of the spectrum, 19 strains have no predicted effector family. Among the likely effectors there are some with a wide distribution (*pteA* and *pteB*), and effectors with a very restricted distribution; *pteJ* only occurs in *A. schubertii*, and *pteI* and *pteL* are only found in one and

two additional strains, respectively. Strains lacking effectors are present in all branches of the *Aeromonas* tree, with the exception of *A. salmonicida* strains.

Conclusions

In this study we identified likely T3SS effectors present in the genomes of 105 *Aeromonas* strains and assessed their cytotoxicity in *S. cerevisiae*. *In silico* identification of effector sequences has been considered a complex task, since their short sequences and significant similarity with different cell systems constitute a barrier for accurate analysis (Panina et al. 2005; Arnold et al. 2009; Löwer and Schneider 2009). Our two-step comparison against positive and negative data sets greatly reduced the number of false positives and identified 13 new likely *Aeromonas* effector families and eight that were previously described. The expression of the likely T3SS effectors in *S. cerevisiae* provided strong evidence for the cytotoxicity of most of the identified effectors.

Based on both *in silico* and *in vitro* evidence addressed in this study we propose naming the 11 newly described effectors with validated cytotoxic activity with the prefix “*ate*”, for *Aeromonas* T3SS effectors. Consequently, the newly described *Aeromonas* effector families should be referred as *ateA*, *ateB*, *ateC*, *ateD*, *ateF*, *ateG*, *ateH*, *ateI*, *ateJ*, *ateL*, and *ateM*, while *pteE* and *pteK* should keep the putative designation since no yeast growth inhibition phenotype was detected.

The high frequency of horizontal transfer events of effectors within *Aeromonas* is reflected in their scattered distribution along the phylogenomic tree of the genus and reconciliations of each effector gene tree with the phylogenomic tree. Members of the *Aeromonas* genus are known for promiscuous gene and gene fragment exchange (Morandi et al. 2005; Silver et al. 2011), as are genes related to the T3SS (Naum et al. 2009; Brown and Finlay 2011). The comparison of the number of events necessary for gene tree reconciliations when not allowing HGT *versus* allowing HGT provided strong evidence for such high exchange rate. Most of the detected effector HGT among *Aeromonas* were inter-species transfers. This larger number of HGTs predicted between species than within species is likely due to the absence of a strong phylogenetic signal created by intra-species HGTs. In the context of our study we observed that a

shared isolation sources has a smaller impact on effector distribution compared to the phylogenetic signal. The lack of a strong isolation source signal could be due to *Aeromonas* strains interacting with a wide range of eukaryotic hosts, each requiring a different set of molecular tools.

Using a combination of bioinformatic and molecular approaches we were able to identify 11 new effectors that are toxic to yeast cells. Future studies focusing on the role of individual effectors in different animal models, e.g. fish, mice, wax worms, and leeches would further characterize the effectors and perhaps identify specific niches that some are associated with. In addition, the bioinformatic approach that we described can be used to identify effectors in other genera and also be applied to other gene families.

Materials and methods

Sequences

All protein coding genes sequences from 105 *Aeromonas* strains were sampled in this study, 40 were sequenced by our group and 65 obtained from public available databases (Supplementary Table 3 and 4). The genomes were sequenced, assembled and annotated as described in Colston et al 2014. Briefly, libraries were prepared using NexteraXT and sequenced on an Illumina MiSeq at the Microbial Analysis, Resources and Services Facility of the University of Connecticut. The reads were assembled using the CLC Genomics Workbench. Gene predictions and product annotations was performed using RAST (Overbeek et al. 2014).

Homology clustering

Protein sequences from the 105 *Aeromonas* genomes were clustered into homologous groups using the OrthoMCL algorithm (Li et al. 2003) as implemented in the Get_Homologues software package (Contreras-Moreira and Vinuesa 2013). A total of 25,518 homologous groups were assembled, 2,755 of which are present in at least 90% of surveyed genomes. The genes of this extended core comprise approximately 65% of the individual *Aeromonas* genome.

Identification of T3SS related proteins in *Aeromonas* spp. genomes

Reference amino acid sequences of T3SS apparatus and effector proteins were downloaded from the VFDB in April 2015. All *Aeromonas* spp. protein families were compared against the VFDB reference sequences using the HMMER suite (Mistry et al. 2013). HMMER profiles were generated for all identified homologous groups of *Aeromonas* proteins, whose alignments were performed using MAFFT (Katoh et al. 2002). Protein families without a positive VFDB hit (HMMER default parameters) were dismissed from further investigation. Proteomes of *Vibrio fischeri* ES114 (Ruby et al. 2005) and *Escherichia coli* K12 (Blattner et al. 1997; Riley et al. 2006) were used as negative controls as they are known to not possess T3SSs. We reasoned that by removing homologous groups matching equally well, or better, to the negative controls as to the VFDB sequences that inclusion of distant T3SS homologs, such as F-ATPase or flagellar components, would be avoided. All protein sequences from groups representing positive hits against the VFDB were used as queries in BLAST searches against the two different data sets, the VFDB reference sequences and the negative control. A t-test was performed on homologous groups that had at least three positive hits against both the VFDB and the negative control to evaluate, if the sequences were significantly more similar to those in the VFDB compared to the negative control. In cases where the homologous group had less than three hits against one of the datasets, we assessed if the lesser score against VFDB was at least 1.5 times the greater score against the negative control.

Gene families that were predicted to be related to the T3SS were manually curated, and were divided into separate clusters or merged into one as necessary. All putative effector families were manually identified from the set of 127 groups related to T3SS. The correlations between effector occurrences (Fig. S2) were assessed using Pearson's correlation tests.

Reference phylogeny

Multiple sequence alignments were generated for all 5,693 single copy gene families using MAFFT, and quality control and edits were automatically performed by GUIDANCE (Penn et al. 2010). Pairwise maximum likelihood distance matrices were calculated for gene families present in at least 10 genomes

using Tree-Puzzle (Schmidt et al. 2002). We applied Pearson correlation tests between all gene family combinations, and the significance of each correlation was assessed through Mantel tests and p-values corrected for multiple testing using Benjamini-Hochberg's False Discovery Rate (FDR) correction. The correlation-weighted network ($p \leq 0.05$ and $\rho \geq 0.7$) was submitted to Markov Clustering Algorithm (MCL) with 5.5 inflation parameter (Dongen 2000; Enright et al. 2002). The unpartitioned concatenation of the 1,678 gene families present in the largest cluster was submitted to RAxML (Stamatakis 2006) for phylogenetic reconstruction using the GTR+GAMMA substitution model, and support assessed with the aLRT SH-like method (Guindon et al. 2010).

Gene phylogenies and reconciliations

Nucleotide sequences from all putative effector gene-families were aligned using MAFFT, and phylogenetic trees were generated using RAxML (GTR+GAMMA+I model). The obtained trees were reconciled to the reference tree using the RangerDTL2 package (Bansal et al. 2012). All distinct rootings of the effector trees leading to the optimal reconciliation were accounted for using the OptRoot tool, and branches with bootstrap support less than 70% were collapsed previous to the reconciliation step using the OptResolution tool. The number of assessed optimal resolutions for a given rooting was limited to 100. RangerDTL2 was run 100 times for each reported resolution, and its output combined using the AggregateRanger tool. Supplementary Figure S3 gives cladograms for the 15 putative effector phylogenies with four or more taxa after applying the TreeFixDTL tool to obtain the most parsimonious root.

Addition of non-*Aeromonas* representatives

Amino acid sequences from putative T3SS effectors were used to query against GenBank to identify non-*Aeromonas* homologs. The top non-*Aeromonas* hit was downloaded if the aligned region spanned through at least 60% of the query sequence and had an identity of at least 30%. Once a non-*Aeromonas* hit fulfilling the requirements is identified, it was downloaded together with all other non-*Aeromonas* hits with bitscores of at least 85% of the top hit.

Strains and growth conditions

Bacterial and yeast strains, as well as plasmids generated in this study are listed in Table S1.

Aeromonas strains were grown on LB agar or in LB broth at 30°C (Sambrook et al. 2001). *E. coli* strain DH5α λ -*pir* was used for cloning of all plasmids, and selective growth of *E. coli* was done using LB broth and agar solidified plates containing 100 µg/ml ampicillin. Non-selective growth of the yeast strain *S. cerevisiae* BY4741 was using yeast extract-peptone-dextrose (YPD) medium (Sambrook et al. 2001). Selective growth of yeast was done using synthetic defined medium lacking histidine (SDM-His) and containing either 2% glucose (SDM-His-Glu), or 2% galactose and 1% raffinose (SDM-His-Gal). Some strains were additionally evaluated on SDM-His-Gal media containing a final concentration 7 mM caffeine or 0.5 M NaCl.

DNA isolation, genome sequencing, assembly and annotation.

DNA was isolated using MasterPure DNA purification kit (Epicenter, Madison, WI) as described by Colston et al (Colston et al. 2014). Libraries were prepared using the Illumina NexteraXT kit and sequenced on an Illumina MiSeq as described previously (Colston et al. 2014). The reads were quality trimmed and assembled using CLC Genomics Workbench (7.0.04) and the resulting sequences were annotated using RAST (Colston et al. 2014; Overbeek et al. 2014).

Strain and Plasmid Construction

All primers used in the present study are listed in Table S2. Putative T3SS effector genes were PCR amplified from *Aeromonas* genomic DNA using Phusion DNA polymerase (New England Biolabs; NEB) and appropriate primers. Primers were tailed to allow subsequent Gibson Assembly (NEB) cloning of amplicons at the *EcoRI* and *XhoI* sites of the shuttle vector pGREG533 (Euroscarf), a plasmid which allows for galactose inducible expression of cloned genes from the GAL1 promoter. Amplicons were purified using a commercial kit (Wizard SV Gel and PCR Clean up System, Promega) and *EcoRI/XhoI* digested pGREG533 was gel purified using a Qiaex II kit (Qiagen). The effector genes were cloned in-frame with the N-terminal 7x hemagglutinin tag (7-HA) of pGREG533. Constructs were tested for containing correct inserts by PCR analysis and DNA sequencing, and then were transformed into *S.*

cerevisiae strain BY4741. Yeast were transformed by mixing 0.1 µg of plasmid, 0.1 mg of salmon sperm carrier DNA (Invitrogen) and 0.1 ml of yeast cells resuspended in sterile 1x TE/LiAc buffer (0.1 M Tris-HCl, 10 mM EDTA, 0.1 M LiAc pH 7.5). After adding 0.6 ml of sterile 1x TE/LiAc buffer containing 40% polyethylene glycol 4000, cells were incubated at 30°C for 30 min with shaking (200 rpm). Cell suspensions were mixed with 70 µl dimethyl sulfoxide, heat shocked at 42°C for ≥ 15 min, then centrifuged and resuspended in sterile water. Cell dilutions were plated on SDM-His-Glu plates and incubated for two days at 30°C to obtain transformants.

Yeast growth inhibition assay

For each strain, 3 ml of SDM-His-Glu broth was inoculated with a single yeast colony and incubated at 30°C overnight with shaking. After centrifugation at 800 g for 5 min, the supernatant was decanted and the cell pellet resuspended in sterile water. This process was repeated to remove residual media, then cells were diluted to an OD₆₀₀ of 1.0 and four 10x serial dilutions were made. A 10 µl aliquot of each dilution was spotted onto SDM-His-Glu plates or SDM-His-Gal. Yeast strains presenting partial (*aopP*, *pteH*, *pteJ* & *pteL*) or no growth inhibition phenotypes (*aopO*, *aopH*, *pteE* & *pteK*) on SDM-His-Gal plates were further assessed on SDM-His-Gal plates containing 7 mM caffeine or 0.5 M NaCl. Plates were incubated at 30°C for two (SDM-His-Glu) or three days (SDM-His-Gal) and were photographed with a Nikon D7100 camera.

Western blot analysis

Expression of putative T3SS effectors in yeast strains presenting no growth inhibition phenotypes in any assayed conditions (*aopO*, *aopH*, *pteE* & *pteK*) was assessed by western blot. Overnight SDM-His-Glu cultures were washed and resuspended in SDM-His-Glu broth to an OD₆₀₀ of 0.1, then grown overnight at 30°C for with shaking (200 rpm). Cells were collected by centrifugation and washed prior to resuspension in 1x PBS. Cells were lysed in the presence of protease inhibitors (Halt, ThermoFisher) by beating 3x with 0.5 mm beads, with cooling on ice between lysis steps. Cell debris was pelleted by centrifugation and supernatants were mixed with an equal volume of 2x Laemmli sample buffer containing 2-Mercaptoethanol (BioRad) prior to boiling samples for 5 min. Protein extracts were separated by SDS-

PAGE and then transferred to a PVDF membrane. The pGREG533 based constructs produced N-terminal 7-HA epitope tagged proteins which were detected by using a rabbit polyclonal anti-HA antibody (Thermo Scientific) and a goat anti-rabbit IRDye 800CW 2° antibody (LiCOR). Imaging of blots was done with an Odyssey CLx Imaging system (LiCOR).

Acknowledgements

This work was supported by the USDA ARS agreement 58-1930-4-002, the NSF MCB program award 1616514, and by the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) award numbers 2012/17196-2 and 2014/23975-0. JCS is funded in part by a CNPq researcher fellowship. The genome sequencing was done at the Microbial Analysis, Research and Services facility of the University of Connecticut.

Table 1: Relevant characteristics of putative T3SS effector groups. Number of proteins gives the total number of protein encoding genes identified across all assayed genomes for a given group.

Effector	# of Proteins (# of T ^{\$})	Length in bp (max, median, min) ^{\$\$}	Pfam Domains ^{\$\$\$}	Domain Descriptions ^{\$\$\$}	GOterm IDs
AexT	20 (5-5)	1429, 1353, 1287	PF03496; PF03545	ADP-ribosyltransferase exoenzyme; Yersinia virulence determinant (YopE)	GO:0005576, GO:0009405
AexU	42 (15-15)	1545, 1537.5, 768	PF03545	Yersinia virulence determinant (YopE)	N/A
AopX	36 (14-14)	972, 951, 948	N/A	N/A	N/A
AopO	5 (0-1)	2188, 2187, 2187	PF09632; PF00069	Rac1-binding domain; Protein kinase domain	GO:0004672, GO:0005524, GO:0006468
AopP	6 (0-0)	897	PF03421	YopJ Serine/Threonine acetyltransferase	N/A
Ati2	6 (0-0)	1488, 1488, 1242	PF03372	Endonuclease/Exonuclease/ phosphatase family	N/A
AopH	9 (0-3)	1392, 1206, 1152	PF00102; PF09013	Protein-tyrosine phosphatase; YopH, N-terminal	GO:0004725, GO:0006470, GO:0004725, GO:0006470, GO:0009405
AopS	5 (0-1)	1155, 1155, 1149	PF02661	Fic/DOC family	N/A
PteA	27 (0-13.5)	2112, 2103, 1944	N/A	N/A	N/A
PteB	30 (4-18)	951, 903, 492	N/A	N/A	N/A
PteC	5 (0-2)	1044, 1044, 999	N/A	N/A	N/A

PteD	4 (0-2)	1089, 1083, 1083	N/A	N/A	N/A
PteD.1	1 (0 -0)	1032	N/A	N/A	N/A
PteE	51 (0-31.9)	1443, 1128 , 1083	PF13776; PF02661	Domain of unknown function (DUF4172); Fic/DOC family	N/A
PteF	3 (2-0)	1059, 969, 963	PF03543	Yersinia/Haemophilus virulence surface antigen	GO:0004197, GO:0009405
PteG	7 (0-2.79)	2010, 1872, 1833	PF01734	Patatin-like phospholipase	GO:0006629
PteH	1 (0-0)	1383	PF01734	Patatin-like phospholipase	GO:0006629
PteI	2 (1-1)	1134, 1132.5, 1131	PF03497	Anthrax toxin LF subunit	GO:0005576, GO:0008294, GO:0009405
PteJ	1 (0-0)	708	PF03536	Salmonella virulence-associated 28kDa protein	N/A
PteK	11 (0-4.86)	1422	PF09013	YopH, N-terminal	GO:0004725, GO:0006470, GO:0009405
PteL	3 (2-2)	1452, 1452, 1446	N/A	N/A	N/A
PteM	2 (1-1)	678, 615, 552	PF03496	ADP-ribosyltransferase exoenzyme	GO:0005576,G O:0009405
PteM.1	1 (0-0)	1098	PF03496	ADP-ribosyltransferase exoenzyme	GO:0005576,G O:0009405

§: Duplication, transfer, and loss events were estimated using Ranger-DTL (Bansal et al. 2012), see material and methods for details. No duplication was predicted within the putative effectors evolution. The number of transfers with 100% confidence and the mean number of transfer events is given in parentheses.

\$\$: In instances where the max, median and min nucleotide length of genes was identical a single value is given.

\$\$\$: Domains present in an effector group were identified using the Pfam database, and the PFAM associated Go term ID is given.

Figure Legends

Figure 1: Flowchart of the semi-automatic pipeline for T3SS effector identification.

Figure 2: Phylogeny of 105 *Aeromonas* spp. genomes. Tree branches with sh-like support of at least 90% are highlighted with red circles. The binary heat map represents the presence/absence of putative effectors in *Aeromonas* spp. isolates. The isolation source colors represent: green, veterinary; yellow, sick veterinary; blue, environment; orange, feces; red, blood; purple, wound; and pink, human.

Figure 3: Yeast growth inhibition assay. Strains carrying putative T3SS effectors cloned into pGREG533 were grown overnight in SDM-His-Glu, washed and serially diluted 10x. Aliquots from each dilution (10 μ l) were spotted onto SDM-His-Glu and SDM-His-Gal plates. The strain containing pGREG533 was used as a negative control and showed no growth inhibition. SDM-His-Glu plates were grown for two days at 30°C and SDM-His-Gal plates were grown for 3 days. T3SS effectors previously identified or biochemically characterized are presented in panel (A), whereas those identified in the present study are shown in panel (B).

Figure 4: Yeast growth inhibition assay under cellular stress. Strains carrying putative T3SS effectors cloned into pGREG533 were grown overnight in SDM-His-Glu, washed and serially diluted 10x.

Aliquots from each dilution (10 μ l) were spotted onto SDM-His-Gal plates containing either 0.5 M NaCl or 7 mM caffeine. The strain containing pGREG533 was used as a negative control and showed no growth inhibition. Plates were grown for were grown for 3 days at 30°C.

Figure 1:

The figure displays a phylogenetic tree on the left and a corresponding presence-absence matrix on the right. The tree shows relationships between 106 bacterial strains, with bootstrap values indicated at the nodes. The matrix columns represent 18 genes: *sdgA*, *sdgB*, *sdgC*, *sdgD*, *sdgE*, *sdgF*, *sdgG*, *sdgH*, *sdgI*, *sdgJ*, *sdgK*, *sdgL*, *sdgM*, *sdgN*, *sdgO*, *sdgP*, *sdgQ*, and *sdgR*. Each row corresponds to a strain, with colored dots indicating gene presence and black squares indicating absence.

Strains (from top to bottom):

- A. diversa* CECT4254T
- A. schuberti* CECT4240T
- A. simiae* CIP10798T
- A. bivalvium* CECT7113T
- A. rivuli* DSM2539T
- A. molluscorum* CIP108876T
- A. enchelela* CECT4342T
- A. eucrenophila* CECT4224T
- A. tecta* CECT7082T
- A. media* BAQ071013132
- A. hydrophila* BAQ071013115
- A. media* WS
- A. media* CECT4232T
- A. sanarelli* LMG24682T
- A. taiwanensis* LMG24683T
- A. caviae* CECT4221
- A. caviae* TCO22
- A. caviae* CECT838T
- A. caviae* Ae398
- A. popoffii* CIP105493T
- A. sp. nov.* AH4
- A. piscicola* LMG24783T
- A. bestiarium* CECT4227T
- A. salmonicida* 34mel
- A. salmonicida* pectinolytica CIP107036
- A. salmonicida* AS03
- A. salmonicida* masoucida CIP103210
- A. salmonicida* smithia CIP104757
- A. salmonicida* achromogenes CIP104001
- A. salmonicida* A449
- A. salmonicida* O18526
- A. salmonicida* CIP103209T
- A. dhakensis* 187
- A. dhakensis* 259
- A. dhakensis* CIP107500
- A. dhakensis* 173
- A. dhakensis* BVH65
- A. dhakensis* BVH69
- A. dhakensis* BVH68
- A. dhakensis* BVH70
- A. dhakensis* BVH43
- A. dhakensis* AAK1
- A. dhakensis* 277
- A. dhakensis* CECT7289T
- A. dhakensis* 116
- A. dhakensis* 014
- A. dhakensis* MDS8
- A. dhakensis* SSU
- A. dhakensis* 145
- A. hydrophila* SNUFPCA8
- A. hydrophila* ML09119
- A. hydrophila* CIP107985
- A. hydrophila* NF1
- A. hydrophila* PAQ09101412
- A. hydrophila* CECT839T
- A. hydrophila* BAQ071013136
- A. hydrophila* 226
- A. hydrophila* 2014105092827
- A. hydrophila* NF2
- A. hydrophila* PAQ09101421
- A. hydrophila* PAQ0910149
- A. hydrophila* ARS13114
- A. hydrophila* PAQ0910141
- A. enteropelogenes* CECT4487T
- A. trota* CECT4255T
- A. enteropelogenes* 1999lcr
- A. jandaei* CECT4228T
- A. jandaei* Hs603
- A. sobria* PAQ0910145
- A. sobria* 2014105092720
- A. sobria* ARS14514
- A. sobria* CECT4245T
- A. sobria* PAQ09101419
- A. fluvialis* LMG24681T
- A. allosaccharophila* BVH88
- A. allosaccharophila* ATCC35942
- A. allosaccharophila* CECT4199T
- A. australiensis* CECT8023T
- A. veronii* AMC34
- A. veronii* AK227
- A. veronii* BAQ116
- A. veronii* CIP107763
- A. veronii* CECT4486
- A. sp.* 159
- A. veronii* ADV102
- A. veronii* BVH47
- A. veronii* Hm22
- A. veronii* AMC25
- A. veronii* BVH46
- A. veronii* BVH37
- A. veronii* AK241
- A. veronii* AER39
- A. veronii* CECT7059
- A. veronii* BAQ135
- A. veronii* LMG13067
- A. veronii* G3C1
- A. veronii* Hm21
- A. veronii* CECT4902
- A. veronii* TCO21
- A. veronii* AMC35
- A. veronii* AER397
- A. veronii* B565
- A. veronii* CECT4257T
- A. veronii* AK247
- A. veronii* AK236

Scale bar: 0.14

Figure 3:

Figure 4:

5.2 Discussão

O pipeline que desenvolvemos para a identificação prováveis efetores do SST3 mostrou-se bastante robusto, dados os resultados dos testes de citotoxicidade realizados pelo Dr. Marsden. Dos 21 prováveis efetores identificados 13 não haviam sido previamente descritos em *Aeromonas*, e apenas dois destes novos efetores não inibiram o crescimento de *S. cerevisiae*.

Ao avaliarmos as árvores filogenéticas das 21 famílias de prováveis efetores observamos que em todas elas as sequências de *Aeromonas* compartilham um único ancestral comum dentro do gênero. As prováveis poucas transferências horizontais envolvendo linhagens de outros gêneros contrasta com a alta frequência de THG entre membros do gênero. Assim como em nosso estudo de fixação de N_2 em Proteobacteria, proximidade filogenética entre os genomas mostrou-se mais importante que o compartilhamento de características ambientais para as THGs preditas.

Capítulo 6

Conclusões

Neste trabalho apresentamos os dois principais projetos que realizamos no decorrer deste doutorado (Capítulos 4 e 5). Também descrevemos um novo método para agrupamento de famílias gênicas com histórias evolutivas simulares (Capítulo 3), que é útil para investigação de eventos de transferência horizontal de genes.

As contribuições alcançadas foram:

- um novo método para comparação de histórias evolutivas de genes bacterianos computacionalmente mais eficiente que métodos existentes na literatura
- novas hipóteses sobre a evolução de genes de fixação de N_2 em Proteobacteria, entre as quais estão cinco prováveis origens distintas dos genes de fixação de N_2 , presentes consistentemente na maioria dos genes avaliados
- identificação de potenciais efetores do SST3 em *Aeromonas*, entre os quais 13 prováveis efetores não foram previamente descritos no gênero e apenas dois destes não apresentaram citotoxicidade em *S. cerevisiae*

Há uma grande sobreposição entre as análises que realizados nos estudos apresentados nos Capítulos 4 e 5, apesar dos genes estudados apresentarem resultados significativamente diferentes. Ao avaliarmos a possível origem comum dos genes de fixação de N_2 em Proteobacteria identificamos cinco prováveis origens distintas, possivelmente ligadas a diferentes grupos taxonômicos, enquanto que uma análise similar referente a origem

comum de efetores do SST3 em *Aeromonas* indicou que cada um dos 21 efetores preditos possuem um único ancestral comum aos membros do gênero. Esta grande diferença no comportamento destes genes esta provavelmente relacionada a dois fatores: [1] variedade taxonômica dos genomas amostrados em cada estudo e [2] função exercida pelas classes de genes estudados. Em relação ao 1º, a maior variedade taxonômica presente em Proteobacteria (onde o gênero *Aeromonas* esta contido) esta automaticamente ligada a maior probabilidade de que diferentes linhagens obtenham genes de outros grupos taxonômicos.

Sobre o 2º ponto, o papel central que genes ligados a fixação de N_2 possuem no metabolismo procariótico faz com eles sejam pouco tolerantes a mutações, e conseqüentemente suas sequências são bastante conservadas. A grande similaridade entre homólogos dos genes de fixação de N_2 em uma grande diversidade de organismos (evidenciada pelas THGs diferenciais preditas entre *nifDK* apresentadas no Capítulo 4) faz com que transferências horizontais destes genes tenham um pequeno impacto no *fitness* do organismo receptor. Devido a este baixo impacto no *fitness*, genes altamente conservados tendem a ser mais “livremente” transferidos horizontalmente (Shapiro *et al.*, 2012; Papke and Gogarten, 2012), o que explica as possíveis cinco origens diferentes de genes de fixação de N_2 em Proteobacteria. Enquanto isso, genes que codificam efetores do SST3 possuem características inversas, sendo muito pouco conservados devido a constante adaptação ao ambiente colonizado pelo genoma hospedeiro. Por possuírem características específicas ao genoma hospedeiro e ao nicho que habitam transferências horizontais destes genes tendem a ter um maior impacto no *fitness* do recipiente (Shapiro *et al.*, 2012; Papke and Gogarten, 2012), o que ao mesmo tempo que aumenta o impacto evolutivo de sua transferência horizontal também diminui a frequência com que estes genes são fixados em diferentes populações.

6.1 Direções futuras de pesquisa

As seguintes possíveis direções de pesquisa são propostas com base no trabalho realizado:

- validação do método proposto para agrupamento de famílias gênicas utilizando sequências simuladas sob diferentes processos evolutivos
 - apesar de promissor, o método proposto necessita ser validado e comparado as diversas ferramentas presentes na literatura
- montagem de uma rede filogenômica de compartilhamento horizontal de todas famílias gênicas de Proteobacteria
 - o desenvolvimento dos passos necessários para a montagem da rede de compartilhamento horizontal apresentada no Capítulo 4 necessita pequenos ajustes referentes a escalabilidade dos dados utilizados

Capítulo 7

Outros trabalhos

7.1 Genoma de Candidatus *Liberibacter americanus*

Juntamente com o Prof. Setubal fui colaborador no artigo *The Complete Genome Sequence of ‘Candidatus Liberibacter americanus’, Associated with Citrus Huanglongbing* publicado na *Molecular Plant-Microbe Interactions Journal* (Wulff *et al.*, 2014).

Minhas contribuições foram análises de THGs no genoma de Candidatus *Liberibacter americanus* e em outros genomas do gênero. Além das análises de transferência horizontal sou responsável por parte das análises de genômica comparativa, inclusive a Figura 4, a qual representa o pan-genoma “aberto” dos quatro genomas disponíveis durante o estudo.

The Complete Genome Sequence of ‘*Candidatus Liberibacter americanus*’, Associated with Citrus Huanglongbing

Nelson A. Wulff,¹ Shujian Zhang,² João C. Setubal,³ Nalvo F. Almeida,⁴ Elaine C. Martins,¹ Ricardo Harakava,⁵ Dibyendu Kumar,⁶ Luiz Thiberio Rangel,³ Xavier Foissac,⁷ Joseph M. Bové,⁷ and Dean W. Gabriel²

¹Departamento Científico–Fundecitrus, Araraquara, SP, 14807-040, Brazil; ²Plant Pathology Department, University of Florida, Gainesville 32611, U.S.A.; ³Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, SP, 05508-000, Brazil; ⁴Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Campo Grande, MS, 79070-900, Brazil; ⁵Laboratório de Bioquímica Fitopatológica, Instituto Biológico, São Paulo, SP, 04014-002, Brazil; ⁶ICBR University of Florida, Gainesville 32610, U.S.A.; ⁷UMR-1332 Biologie du Fruit et Pathologie, INRA, Université Bordeaux Ségalen, Villenave d'Ornon, France

Submitted 30 September 2013. Accepted 25 October 2013.

Liberibacter spp. form a Rhizobiaceae clade of phloem-limited pathogens of limited host range. Two obligately parasitic species have been sequenced: ‘*Candidatus Liberibacter asiaticus*’, which causes citrus huanglongbing (HLB) worldwide, and ‘*Ca. L. solanacearum*’, which causes potato “zebra chip” disease. A third (proposed) species, *Liberibacter crescens*, was isolated from mountain papaya, grown in axenic culture, and sequenced. In an effort to identify common host determinants, the complete genomic DNA sequence of a second HLB species, ‘*Ca. L. americanus*’ strain ‘São Paulo’ was determined. The circular genome of 1,195,201 bp had an average 31.12% GC content and 983 predicted protein encoding genes, 800 (81.4%) of which had a predicted function. There were 658 genes common to all sequenced *Liberibacter* spp. and only 8 genes common to ‘*Ca. L. americanus*’ and ‘*Ca. L. asiaticus*’ but not found in ‘*Ca. L. solanacearum*’. Surprisingly, most of the lipopolysaccharide biosynthetic genes were missing from the ‘*Ca. L. americanus*’ genome, as well as *OmpA* and a key regulator of flagellin, all indicating a ‘*Ca. L. americanus*’ strategy of avoiding production of major pathogen-associated molecular patterns present in ‘*Ca. L. asiaticus*’ and ‘*Ca. L. solanacearum*’. As with ‘*Ca. L. asiaticus*’, one of two ‘*Ca. L. americanus*’ prophages replicated as an excision plasmid and carried potential lysogenic conversion genes that appeared fragmentary or degenerated in ‘*Ca. L. solanacearum*’.

Huanglongbing (HLB), commonly known as citrus greening disease, is one of the top three most damaging diseases of citrus in Africa, America, and Asia. HLB is naturally transmitted by

psyllids and experimentally by grafting or dodder (*Cuscuta* spp.). The disease was shown to be graft transmissible in 1956 (Lin 1956a and b) and, therefore, it was thought to be caused by a putative virus. However, in 1970, sieve-tube-restricted bacteria were discovered in affected trees. First thought to be mycoplasma-like (Laflèche and Bové 1970), they were soon recognized as walled bacteria (Bové and Saglio 1974; Saglio et al. 1971) of the gram-negative type (Garnier et al. 1984) and finally shown to be species of α -proteobacteria (Jagoueix et al. 1994). Two species were recognized: ‘*Candidatus Liberibacter asiaticus*’ for the disease in Asia and ‘*Ca. L. africanus*’ for the disease in Africa. In 2004, when HLB was seen for the first time in the Americas and, more precisely, in São Paulo State, Brazil, two *Liberibacter* spp. were identified: i) a new species, ‘*Ca. L. americanus*’, infecting most of the affected trees, and ii) the known Asian *Liberibacter*, ‘*Ca. L. asiaticus*’, present in a minority of trees (Teixeira et al. 2005). Over the last years, a shift in the prevalence of ‘*Ca. L. americanus*’ and ‘*Ca. L. asiaticus*’ has been observed, with ‘*Ca. L. asiaticus*’ being today far more prevalent than ‘*Ca. L. americanus*’ (Lopes et al. 2009a and b).

All three citrus *Liberibacter* spp. are uncultured and phloem-limited. ‘*Ca. L. asiaticus*’ is the most widely distributed by far. In the whole of Asia, from the Indian subcontinent to Papua-New Guinea, HLB is exclusively caused by ‘*Ca. L. asiaticus*’ and transmitted by the Asian citrus psyllid, *Diaphorina citri*. ‘*Ca. L. asiaticus*’ and *D. citri* are thought to be native to Asia (Nelson et al. 2013). Prior to 2004, ‘*Ca. L. asiaticus*’ was reported present only in Asia; it is now reported present in North, Central, and South America. In Africa and Madagascar, HLB is exclusively caused by ‘*Ca. L. africanus*’ and transmitted by the African citrus psyllid, *Trioza erythrae*. The “African” disease occurs in cool areas, often above 600 m in altitude, with temperatures below 30°C. Both ‘*Ca. L. africanus*’ and *T. erythrae* are native to Africa (Beattie et al. 2008; Bové 2013; Hollis 1984; Nelson et al. 2013) and both are heat sensitive (Bové et al. 1974; Catling 1969; Moran and Blowers 1967; Schwarz and Green 1972). In Brazil, both ‘*Ca. L. asiaticus*’ and ‘*Ca. L. americanus*’ are transmitted by *D. citri*, the Asian citrus psyllid. ‘*Ca. L. americanus*’ is significantly less heat tolerant than ‘*Ca. L. asiaticus*’ (Lopes et al. 2009b). Outside of Brazil, ‘*Ca. L. americanus*’ has been reported from a citrus

Nucleotide sequence data for the ‘*Ca. L. americanus*’ São Paulo genome has been deposited in GenBank as accession number CP006604.

Corresponding author: D. W. Gabriel; Telephone: +1.352.392.7329; Fax: +1.352.392.6532; E-mail: dgabr@ufl.edu

*The e-Xtra logo stands for “electronic extra” and indicates that six supplementary figures and one supplementary table are published online.

7.2 Revisão sobre Ca. *Liberibacter* spp.

Juntamente com o Prof. Setubal e o Dr. Joaquim Martins fui co-autor da revisão sobre *Liberibacter* spp., que será publicada na *Annual Review of Phytopathology*.

Fomos responsáveis pelas análises de genômica comparativa envolvendo 19 genomas de *Liberibacter* spp. disponíveis atualmente. Dentre as análises nós classificamos os genes preditos nos 19 genomas de acordo com suas atividades em vias do KEGG, identificamos diferentes enriquecimentos de vias em diferentes genomas, e a composição do pan e core-genoma.

The *Candidatus* Liberibacter–Host Interface: Insights into Pathogenesis Mechanisms and Disease Control

Annual Review of Phytopathology

Vol. 55: (Volume publication date September 2017)

DOI: 10.1146/annurev-phyto-080516-035513

Nian Wang,¹ Elizabeth A. Pierson,² João Carlos Setubal,³ Jin Xu,¹ Julien G. Levy,² Yunzeng Zhang,¹ Luiz Thiberio Rangel,³ and Joaquim Martins³

¹Department of Microbiology and Cell Science, University of Florida, Lake Alfred, FL 33850

²Department of Horticultural Sciences, Texas A&M University, College Station, TX 77803

³Department of Biochemistry, University of São Paulo, São Paulo, SP, 05508-000 Brazil

Article Metrics

Permissions | Reprints

Abstract

Candidatus Liberibacter species are associated with economically devastating diseases of citrus, potato, and many other crops. The importance of these diseases as well as the proliferation of new diseases on a wider host range is likely to increase as the insects vectoring the *Ca. Liberibacter* species expand their territories worldwide. Here, we review the progress on understanding pathogenesis mechanisms of *Ca. Liberibacter* species, and control approaches for diseases they cause. We discussed the virulence traits of Liberibacters including secretion systems and putative effectors, LPS, flagella, prophage, and salicylic acid hydroxylase. Two putative mechanisms for disease development are discussed: 1) Virulence factors are responsible for the disease symptom development by interfering with host targets in the phloem or in the companion cells. 2) Liberibacters cause disease symptoms by eliciting self-destructive plant responses via plant recognition of Liberibacter pathogen-associated molecular patterns (PAMPs, e.g., flagella, and LPS) or by altering phloem loading, unloading, and transport via their physical presence and metabolic activity. We also review disease management strategies including promising technologies. Currently, aggressive psyllid management enables potato production, although insecticide resistance is becoming an issue. Citrus production in the presence of Huanglongbing is also possible if the most promising management approaches are integrated. Meanwhile, new technologies such as CRISPR derived genome editing will provide an unprecedented opportunity to provide long-term solutions.

Expected final online publication date for the Annual Review of Phytopathology Volume 55 is August 4, 2017. Please see <http://www.annualreviews.org/catalog/pubdates.aspx> for revised estimates.

Referências Bibliográficas

- Alam, A., Miller, K. A., Chaand, M., Butler, J. S., and Dziejman, M. 2011. Identification of *Vibrio cholerae* type III secretion system effector proteins. *Infection and immunity*, 79(4): 1728–40.
- Andam, C. P. and Gogarten, J. P. 2011. Biased gene transfer and its implications for the concept of lineage. *Biology direct*, 6(1): 47.
- Andam, C. P., Williams, D., and Gogarten, J. P. 2010. Biased gene transfer mimics patterns created through shared ancestry. *Proceedings of the National Academy of Sciences of the United States of America*, 107(23): 10679–84.
- Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, H.-W., Horn, M., and Rattei, T. 2009. Sequence-based prediction of type III secreted proteins. *PLoS pathogens*, 5(4): e1000376.
- Azad, R. K. and Lawrence, J. G. 2005. Use of artificial genomes in assessing methods for atypical gene detection. *PLoS Comput Biol*, 1(6): e56.
- Bansal, M. S., Alm, E. J., and Kellis, M. 2012. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, 28(12): i283–i291.
- Bansal, M. S., Banay, G., Harlow, T. J., Gogarten, J. P., and Shamir, R. 2013. Systematic inference of highways of horizontal gene transfer in prokaryotes. *Bioinformatics (Oxford, England)*, 29(5): 571–9.
- Bansal, M. S., Wu, Y.-C., Alm, E. J., and Kellis, M. 2015. Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics (Oxford, England)*, 31(8): 1211–8.

- Baptiste, E., O'Malley, M. a., Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., Lapointe, F.-J., Dupré, J., Dagan, T., Boucher, Y., and Martin, W. 2009. Prokaryotic evolution and the tree of life are two different things. *Biology direct*, 4: 34.
- Batra, P., Mathur, P., and Misra, M. 2016. *Aeromonas* spp.: An emerging nosocomial pathogen. *Journal of Laboratory Physicians*, 8(1): 1.
- Beiko, R. G., Harlow, T. J., and Ragan, M. a. 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40): 14332–7.
- Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.
- Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science (New York, N.Y.)*, 277(5331): 1453–62.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.
- Boothby, T. C., Tenlen, J. R., Smith, F. W., Wang, J. R., Patanella, K. A., Osborne Nishimura, E., Tintori, S. C., Li, Q., Jones, C. D., Yandell, M., Messina, D. N., Glasscock, J., and Goldstein, B. 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1–6.
- Boto, L. 2014. Horizontal gene transfer in the acquisition of novel traits by metazoans. In *Proc. R. Soc. B*, volume 281, page 20132450. The Royal Society.
- Boyd, E. S., Anbar, a. D., Miller, S., Hamilton, T. L., Lavin, M., and Peters, J. W. 2011a. A late methanogen origin for molybdenum-dependent nitrogenase. *Geobiology*, 9(3): 221–32.

- Boyd, E. S., Hamilton, T. L., and Peters, J. W. 2011b. An alternative path for the evolution of biological nitrogen fixation. *Frontiers in microbiology*, 2(October): 205.
- Boyd, E. S., Garcia Costas, A. M., Hamilton, T. L., Mus, F., and Peters, J. W. 2015. Evolution of molybdenum nitrogenase during the transition from anaerobic to aerobic metabolism. *Journal of Bacteriology*, 197(9): 1690–1699.
- Braun, M., Stuber, K., Schlatter, Y., Wahli, T., Kuhnert, P., and Frey, J. 2002. Characterization of an adp-ribosyltransferase toxin (aext) from aeromonas salmonicida subsp. salmonicida. *Journal of bacteriology*, 184(7): 1851–1858.
- Brohee, S. and Van Helden, J. 2006. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, 7(1): 488.
- Bromham, L. and Penny, D. 2003. The modern molecular clock. *Nature Reviews Genetics*, 4(3): 216–224.
- Brown, N. F. and Finlay, B. B. 2011. Potential origins and horizontal transfer of type III secretion systems and effectors. *Mobile genetic elements*, 1(2): 118–121.
- Burr, S. E., Stuber, K., and Frey, J. 2003. The ADP-Ribosylating Toxin, AexT, from Aeromonas salmonicida subsp. salmonicida Is Translocated via a Type III Secretion Pathway. *Journal of Bacteriology*, 185(22): 6583–6591.
- Burr, S. E., Pugovkin, D., Wahli, T., Segner, H., and Frey, J. 2005. Attenuated virulence of an Aeromonas salmonicida subsp. salmonicida type III secretion mutant in a rainbow trout model. *Microbiology (Reading, England)*, 151(Pt 6): 2111–8.
- Chami, M., Guilvout, I., Gregorini, M., Remigy, H. W., Muller, S. A., Valerio, M., Engel, A., Pugsley, A. P., and Bayan, N. 2005. Structural Insights into the Secretin PulD and Its Trypsin-resistant Core. *Journal of Biological Chemistry*, 280(45): 37732–37741.
- Chatterjee, S., Chaudhury, S., McShan, A. C., Kaur, K., and De Guzman, R. N. 2013. Structure and Biophysics of Type III Secretion in Bacteria. *Biochemistry*, 52(15): 2508–2517.
- Cheesman, S. E., Neal, J. T., Mittge, E., Seredick, B. M., and Guillemin, K. 2011. Epithelial cell proliferation in the developing zebrafish intestine is regulated by the Wnt pathway and

- microbial signaling via Myd88. *Proceedings of the National Academy of Sciences*, 108(Supplement_1): 4570–4577.
- Chen, I. and Dubnau, D. 2004. Dna uptake during bacterial transformation. *Nature Reviews Microbiology*, 2(3): 241–249.
- Chen, I., Christie, P. J., and Dubnau, D. 2005. The ins and outs of DNA transfer in bacteria. *Science (New York, N.Y.)*, 310(5753): 1456–60.
- Chen, L., Zheng, D., Liu, B., Yang, J., and Jin, Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on. *Nucleic acids research*, 44(D1): D694–7.
- Collins, R. F., Frye, S. A., Kitmitto, A., Ford, R. C., Tonjum, T., and Derrick, J. P. 2004. Structure of the *Neisseria meningitidis* Outer Membrane PilQ Secretin Complex at 12 Å Resolution. *Journal of Biological Chemistry*, 279(38): 39750–39756.
- Colston, S. M., Fullmer, M. S., Beka, L., Lamy, B., Gogarten, J. P., and Graf, J. 2014. Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case. *mBio*, 5(6): e02136.
- Contreras-Moreira, B. and Vinuesa, P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and environmental microbiology*, 79(24): 7696–701.
- Cornelis, G., Heidmann, O., Bernard-Stoecklin, S., Reynaud, K., Véron, G., Mulot, B., Dupressoir, A., and Heidmann, T. 2012. Ancestral capture of syncytin-car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in carnivorans. *Proceedings of the National Academy of Sciences*, 109(7): E432–E441.
- Cortez, D. Q., Lazcano, A., and Becerra, A. 2005. Comparative analysis of methodologies for the detection of horizontally transferred genes: a reassessment of first-order markov models. *In silico biology*, 5(5, 6): 581–592.
- Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A., and Micklem, G. 2015. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome biology*, 16(1): 50.

- Curak, J., Rohde, J., and Stagljar, I. 2009. Yeast as a tool to study bacterial effectors. *Current opinion in microbiology*, 12(1): 18–23.
- Curatti, L., Brown, C. S., Ludden, P. W., and Rubio, L. M. 2005. Genes required for rapid expression of nitrogenase activity in *azotobacter vinelandii*. *Proceedings of the National Academy of Sciences of the United States of America*, 102(18): 6291–6296.
- Dacanay, A. 2006. Contribution of the type III secretion system (TTSS) to virulence of *Aeromonas salmonicida* subsp. *salmonicida*. *Microbiology*, 152(6): 1847–1856.
- Dagan, T. 2011. Phylogenomic networks. *Trends in Microbiology*, 19(10): 483–91.
- Dagan, T. and Martin, W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(3): 870–5.
- Dale, C. and Moran, N. A. 2006. Molecular Interactions between Bacterial Symbionts and Their Hosts. *Cell*, 126(3): 453–465.
- Daubin, V. and Ochman, H. 2004. Bacterial genomes as new gene homes: the genealogy of orphans in *e. coli*. *Genome Research*, 14(6): 1036–1042.
- Dean, P. 2011. Functional domains and motifs of bacterial type III effector proteins and their roles in infection. *FEMS Microbiology Reviews*, 35(6): 1100–1125.
- Deane, J. E., Abrusci, P., Johnson, S., and Lea, S. M. 2010. Timing is everything: the regulation of type III secretion. *Cellular and Molecular Life Sciences*, 67(7): 1065–1075.
- Dixon, R., Eady, R. R., Espin, G., Hill, S., Iaccarino, M., Kahn, D., and Merrick, M. 1980. Analysis of regulation of *Klebsiella pneumoniae* nitrogen fixation (*nif*) gene cluster with gene fusions. *Nature*, 286(5769): 128–32.
- Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C., and Dixon, R. 2012. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC genomics*, 13(1): 162.

- Doyle, M., Fookes, M., Ivens, A., Mangan, M. W., Wain, J., and Dorman, C. J. 2007. An H-NS-like stealth protein aids horizontal DNA transmission in bacteria. *Science (New York, N.Y.)*, 315(5809): 251–2.
- Dubey, G. P. and Ben-Yehuda, S. 2011. Intercellular nanotubes mediate bacterial communication. *Cell*, 144(4): 590–600.
- Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., *et al.* 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188): 745–749.
- Eady, R. R. 1996. Structure-Function Relationships of Alternative Nitrogenases. *Chemical Reviews*, 96(7): 3013–3030.
- Ebanks, R. O. 2006. Expression of and secretion through the *Aeromonas salmonicida* type III secretion system. *Microbiology*, 152(5): 1275–1286.
- Ehrbar, K. and Hardt, W. 2005. Bacteriophage-encoded type iii effectors in *Salmonella enterica* subspecies 1 serovar typhimurium. *Infection, Genetics and Evolution*, 5(1): 1–9.
- Enright, A. J., Van Dongen, S., and Ouzounis, C. A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7): 1575–84.
- Fani, R., Gallo, R., and Liò, P. 2000. Molecular evolution of nitrogen fixation: the evolutionary history of the *nifD*, *nifK*, *nifE*, and *nifN* genes. *Journal of molecular evolution*, pages 1–11.
- Fay, J. C. and Wu, C.-I. 2001. The neutral theory in the genomic era. *Current opinion in genetics & development*, 11(6): 642–646.
- Fehr, D. 2006. AopP, a type III effector protein of *Aeromonas salmonicida*, inhibits the NF- κ B signalling pathway. *Microbiology*, 152(9): 2809–2818.
- Feil, E. J., Holmes, E. C., Bessen, D. E., Chan, M. S., Day, N. P., Enright, M. C., Goldstein, R., Hood, D. W., Kalia, a., Moore, C. E., Zhou, J., and Spratt, B. G. 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 98(1): 182–7.

- Fischer, D. and Eisenberg, D. 1999. Finding families for genomic ORFans. *Bioinformatics*, 15(9): 759–762.
- Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., Shanmugam, D., Roos, D. S., and Stoeckert, C. J. 2011. Using orthomcl to assign proteins to orthomcl-db groups or to cluster proteomes into new ortholog groups. *Current Protocols in Bioinformatics*, pages 6–12.
- Fitch, W. 2000. Homology:: a personal view on some of the problems. *Trends in genetics*, 16(5).
- Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2): 99–113.
- Fitch, W. M., Margoliash, E., *et al.* 1967. Construction of phylogenetic trees. *Science*, 155(3760): 279–284.
- Gaby, J. C. and Buckley, D. H. 2014. A comprehensive aligned nifH gene database: a multipurpose tool for studies of nitrogen-fixing bacteria. *Database : the journal of biological databases and curation*, 2014: bau001.
- Gallon, J. 1981. The oxygen sensitivity of nitrogenase: a problem for biochemists and microorganisms. *Trends in Biochemical Sciences*, 6: 19–23.
- Garrity, G. 2006. *Bergey’s Manual® of Systematic Bacteriology: Volume Two: The Proteobacteria, Part A Introductory Essays*. Springer Science & Business Media.
- Gogarten, J. P. and Townsend, J. P. 2005. Horizontal gene transfer, genome innovation and evolution. *Nature reviews. Microbiology*, 3(9): 679–87.
- Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. 2002. Prokaryotic evolution in light of gene transfer. *Molecular biology and evolution*, 19(12): 2226–38.
- Gophna, U., Ron, E. Z., and Graur, D. 2003. Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene*, 312: 151–163.
- Gori, K., Suchan, T., Alvarez, N., Goldman, N., and Dessimoz, C. 2016. Clustering Genes of Common Evolutionary History. *Molecular biology and evolution*, 33(6): 1590–605.

- Gray, G. and Fitch, W. 1983. Evolution of antibiotic resistance genes: the dna sequence of a kanamycin resistance gene from staphylococcus aureus. *Molecular biology and evolution*, 1(1): 57–66.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3): 307–21.
- Hacker, J. and Kaper, J. B. 2000. Pathogenicity islands and the evolution of microbes. *Annual Reviews in Microbiology*, 54(1): 641–679.
- Hacker, J., Blum-Oehler, G., Mühldorfer, I., and Tschäpe, H. 1997. Pathogenicity islands of virulent bacteria: Structure, function and impact on microbial evolution. *Molecular Microbiology*, 23(6): 1089–1097.
- Hänninen, M., Salmi, S., Mattila, L., Taipainen, R., and Siitonen, A. 1995. Association of aeromonas spp. with travellers' diarrhoea in finland. *Journal of medical microbiology*, 42(1): 26–31.
- Henson, B. J., Watson, L. E., and Barnum, S. R. 2004. The evolutionary history of nitrogen fixation, as assessed by NifD. *Journal of molecular evolution*, 58(4): 390–9.
- Hirsch, A. M., McKhann, H. I., Reddy, A., Liao, J., Fang, Y., and Marshall, C. R. 1995. Assessing horizontal transfer of nifHDK genes in eubacteria: nucleotide sequence of nifK from Frankia strain HFPCcI3. *Molecular Biology and Evolution*, 12(1): 16–27.
- Hsiao, W., Wan, I., Jones, S. J., and Brinkman, F. S. L. 2003. IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, 19(3): 418–420.
- Huang, J. 2013. Horizontal gene transfer in eukaryotes: The weak-link model. *Bioessays*, 35(10): 868–875.
- Hueck, C. J. 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiology and molecular biology reviews : MMBR*, 62(2): 379–433.
- Igbinosa, I. H., Igumbor, E. U., Aghdasi, F., Tom, M., and Okoh, A. I. 2012. Emerging Aeromonas Species Infections and Their Significance in Public Health. *The Scientific World Journal*, 2012: 1–13.

- Janda, J. M. and Abbott, S. L. 1998. Evolving concepts regarding the genus aeromonas: an expanding panorama of species, disease presentations, and unanswered questions. *Clinical Infectious Diseases*, 27(2): 332–344.
- Janda, J. M. and Abbott, S. L. 2010. The genus aeromonas: taxonomy, pathogenicity, and infection. *Clinical microbiology reviews*, 23(1): 35–73.
- Jones, R. M., Luo, L., and Moberg, K. H. 2012. Aeromonas salmonicida-secreted protein AopP is a potent inducer of apoptosis in a mammalian and a Drosophila model. *Cellular Microbiology*, 14(2): 274–285.
- Karlin, S., Mrázek, J., and Campbell, A. M. 1998. Codon usages in different gene classes of the Escherichia coli genome. *Molecular microbiology*, 29(6): 1341–55.
- Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14): 3059–66.
- Kennedy, C. and Toukdarian, A. 1987. Genetics of azotobacters: applications to nitrogen fixation and related aspects of metabolism. *Annual review of microbiology*, 41: 227–58.
- Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press.
- Koonin, E. V. and Wolf, Y. I. 2009. Is evolution Darwinian or/and Lamarckian? *Biology direct*, 4: 42.
- Koski, L. B., Morton, R. A., and Golding, G. B. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. *Molecular biology and evolution*, 18(3): 404–412.
- Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A. A., and Blaxter, M. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade hypsibius dujardini. *Proceedings of the National Academy of Sciences*, 113(18): 5053–5058.
- Kubatko, L. S. and Degnan, J. H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic biology*, 56(1): 17–24.

- Kunin, V., Goldovsky, L., Darzentas, N., and Ouzounis, C. A. 2005. The net of life: reconstructing the microbial phylogenetic network. *Genome Research*, 15(7): 954–959.
- Lang, A. S. and Beatty, J. T. 2007. Importance of widespread gene transfer agent genes in α -proteobacteria. *Trends in microbiology*, 15(2): 54–62.
- Lang, A. S., Zhaxybayeva, O., and Beatty, J. T. 2012. Gene transfer agents: phage-like elements of genetic exchange. *Nature reviews. Microbiology*, 10(7): 472–82.
- Larget, B. R., Kotha, S. K., Dewey, C. N., and Ané, C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics (Oxford, England)*, 26(22): 2910–1.
- Lawrence, J. 1998. Molecular archaeology of the Escherichia coli genome. *Proceedings of the National*, 95(August): 9413–9417.
- Lawrence, J. G. and Ochman, H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *Journal of molecular evolution*, 44(4): 383–397.
- Lawrence, J. G. and Ochman, H. 2002. Reconciling the many faces of lateral gene transfer. *Trends in microbiology*, 10(1): 1–4.
- Lecuit, M. and Eloit, M. 2014. The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. *Frontiers in Cellular and Infection Microbiology*, 4.
- Leigh, J. W., Susko, E., Baumgartner, M., and Roger, A. J. 2008. Testing Congruence in Phylogenomic Analysis. *Systematic Biology*, 57(1): 104–115.
- Lesser, C. F. 2001. Expression of microbial virulence proteins in Saccharomyces cerevisiae models mammalian infection. *The EMBO Journal*, 20(8): 1840–1849.
- Li, L., Stoeckert, C. J., and Roos, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9): 2178–89.
- Liu, L., Yu, L., and Edwards, S. V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC evolutionary biology*, 10: 302.

- Lovett, S., Hurley, R., Sutura Jr, V., Aubuchon, R., and Lebedeva, M. 2002. Crossing over between regions of limited homology in *Escherichia coli*: RecA-dependent and RecA-independent pathways. *Genetics*, 160(3): 851–859.
- Löwer, M. and Schneider, G. 2009. Prediction of Type III Secretion Signals in Genomes of Gram-Negative Bacteria. *PLoS ONE*, 4(6): e5917.
- MacCannell, D. 2016. Next Generation Sequencing in Clinical and Public Health Microbiology. *Clinical Microbiology Newsletter*, 38(21): 169–176.
- Majewski, J. and Cohan, F. 1999. Dna sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics*, 153(4): 1525–1533.
- Mantel, N. 1967. The Detection of Disease Clustering and a Generalized Regression Approach. *CANCER RESEARCH*, 27(1): 209–220.
- Mao, F., Williams, D., Zhaxybayeva, O., Poptsova, M., Lapierre, P., Gogarten, J. P., and Xu, Y. 2012. Quartet decomposition server: a platform for analyzing phylogenetic trees. *BMC bioinformatics*, 13(1): 123.
- Marraffini, L. and Sontheimer, E. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *science*, 322(5909): 1843–1845.
- Martin, W. F. 2011. Early evolution without a tree of life. *Biology direct*, 6(1): 36.
- McDaniel, L. D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K. B., and Paul, J. H. 2010. High frequency of horizontal gene transfer in the oceans. *Science (New York, N.Y.)*, 330(6000): 50.
- Meilă, M. 2007. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5): 873–895.
- Mirarab, S., Bayzid, S. M., Boussau, B., and Warnow, T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215): 1250463.
- Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, 41(12): e121–e121.

- Morandi, A., Zhaxybayeva, O., Gogarten, J. P., and Graf, J. 2005. Evolutionary and diagnostic implications of intragenomic heterogeneity in the 16S rRNA gene in *Aeromonas* strains. *Journal of bacteriology*, 187(18): 6561–4.
- Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezemskaya, O., Isbandi, M., Thomas, A. D., Ali, R., Sharma, K., Kyrpides, N. C., and Reddy, T. B. K. 2017. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic acids research*, 45(D1): D446–D456.
- Mulder, N. and Apweiler, R. 2007. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in molecular biology (Clifton, N.J.)*, 396: 59–70.
- Naum, M., Brown, E. W., and Mason-Gamer, R. J. 2009. Phylogenetic evidence for extensive horizontal gene transfer of type III secretion system genes among enterobacterial plant pathogens. *Microbiology (Reading, England)*, 155(Pt 10): 3187–99.
- Navarre, W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S., and Fang, F. 2006. Selective silencing of foreign dna with low gc content by the h-ns protein in salmonella. *Science*, 313(5784): 236–238.
- Notti, R. Q. and Stebbins, C. E. 2016. *The Structure and Function of Type III Secretion Systems*. Ph.D. thesis.
- Novichkov, P. S., Omelchenko, M. V., Gelfand, M. S., Mironov, A. A., Wolf, Y. I., and Koonin, E. V. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *Journal of bacteriology*, 186(19): 6575–85.
- Ochman, H., Lawrence, J. G., and Groisman, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784): 299–304.
- Ohta, T. and Gillespie, J. H. 1996. Development of neutral and nearly neutral theories. *Theoretical population biology*, 49(2): 128–142.
- Panina, E. M., Mattoo, S., Griffith, N., Kozak, N. A., Yuk, M. H., and Miller, J. F. 2005. A genome-wide screen identifies a *Bordetella* type III secretion effector and candidate effectors in other species. *Molecular Microbiology*, 58(1): 267–279.

- Papke, R. T. and Gogarten, J. P. 2012. Ecology. How bacterial lineages emerge. *Science (New York, N.Y.)*, 336(6077): 45–6.
- Patterson, C. 1988. Homology in classical and molecular biology. *Molecular Biology and Evolution*, 5(6): 603–625.
- Penn, O., Privman, E., Landan, G., Graur, D., and Pupko, T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Molecular biology and evolution*, 27(8): 1759–1767.
- Podell, S. and Gaasterland, T. 2007. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome biology*, 8(2): R16.
- Polz, M. F., Alm, E. J., and Hanage, W. P. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in genetics : TIG*, 29(3): 170–5.
- Popa, O. and Dagan, T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Current opinion in microbiology*, 14(5): 615–23.
- Popa, O., Hazkani-Covo, E., Landan, G., Martin, W., and Dagan, T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome research*, 21(4): 599–609.
- Popa, O., Landan, G., and Dagan, T. 2017. Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. *The ISME journal*, 11(2): 543–554.
- Preston, G. M. 2007. Metropolitan microbes: type iii secretion in multihost symbionts. *Cell host & microbe*, 2(5): 291–294.
- Ragan, M. A. 2001. Detection of lateral gene transfer among microbial genomes. *Current opinion in genetics & development*, 11(6): 620–626.
- Ragan, M. A. 2002. Reconciling the many faces of lateral gene transfer. *Trends in Microbiology*, 10(1): 4.
- Ragan, M. A., Harlow, T. J., and Beiko, R. G. 2006. Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends in microbiology*, 14(1): 4–8.

- Raymond, J., Siefert, J. L., Staples, C. R., and Blankenship, R. E. 2004. The natural history of nitrogen fixation. *Molecular biology and evolution*, 21(3): 541–54.
- Reith, M. E., Singh, R. K., Curtis, B., Boyd, J. M., Bouevitch, A., Kimball, J., Munholland, J., Murphy, C., Sarty, D., Williams, J., Nash, J. H., Johnson, S. C., and Brown, L. L. 2008. The genome of *Aeromonas salmonicida* subsp. *salmonicida* A449: insights into the evolution of a fish pathogen. *BMC Genomics*, 9(1): 427.
- Riley, D. R., Sieber, K. B., Robinson, K. M., White, J. R., Ganesan, A., Nourbakhsh, S., and Dunning Hotopp, J. C. 2013. Bacteria-Human Somatic Cell Lateral Gene Transfer Is Enriched in Cancer Samples. *PLoS Computational Biology*, 9(6): e1003107.
- Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K. B., Blattner, F. R., Chaudhuri, R. R., Glasner, J. D., Horiuchi, T., Keseler, I. M., Kosuge, T., Mori, H., Perna, N. T., Plunkett, G., Rudd, K. E., Serres, M. H., Thomas, G. H., Thomson, N. R., Wishart, D., and Wanner, B. L. 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot–2005. *Nucleic acids research*, 34(1): 1–9.
- Ruby, E. G., Urbanowski, M., Campbell, J., Dunn, A., Faini, M., Gunsalus, R., Lostroh, P., Lupp, C., McCann, J., Millikan, D., Schaefer, A., Stabb, E., Stevens, A., Visick, K., Whistler, C., and Greenberg, E. P. 2005. Complete genome sequence of *Vibrio fischeri*: a symbiotic bacterium with pathogenic congeners. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8): 3004–9.
- Salomon, D. and Sessa, G. 2010. Identification of growth inhibition phenotypes induced by expression of bacterial type III effectors in yeast. *Journal of visualized experiments : JoVE*, (37).
- Salzberg, S. L. 2017. Horizontal gene transfer is not a hallmark of the human genome. *Genome Biology*, 18(1): 85.
- Sandberg, R., Winberg, G., Bränden, C. I., Kaske, A., Ernberg, I., and Cöster, J. 2001. Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome research*, 11(8): 1404–9.

- Schmehl, M., Jahn, A., Meyer zu Vilsendorf, A., Hennecke, S., Masepohl, B., Schuppler, M., Marxer, M., Oelze, J., and Klipp, W. 1993. Identification of a new class of nitrogen fixation genes in *rhodobacter capsalatus*: a putative membrane complex involved in electron transport to nitrogenase. *Molecular and General Genetics MGG*, 241(5): 602–615.
- Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics (Oxford, England)*, 18(3): 502–4.
- Setubal, J. C. and Almeida, N. F. 2015. The *Azotobacter vinelandii* Genome: An Update. In *Biological Nitrogen Fixation*, pages 225–234. John Wiley & Sons, Inc, Hoboken, NJ, USA.
- Setubal, J. C., dos Santos, P., Goldman, B. S., Ertesvåg, H., Espin, G., Rubio, L. M., Valla, S., Almeida, N. F., Balasubramanian, D., Cromes, L., Curatti, L., Du, Z., Godsy, E., Goodner, B., Hellner-Burris, K., Hernandez, J. a., Houmiel, K., Imperial, J., Kennedy, C., Larson, T. J., Latreille, P., Ligon, L. S., Lu, J., Maerk, M., Miller, N. M., Norton, S., O’Carroll, I. P., Paulsen, I., Raulfs, E. C., Roemer, R., Rosser, J., Segura, D., Slater, S., Stricklin, S. L., Studholme, D. J., Sun, J., Viana, C. J., Wallin, E., Wang, B., Wheeler, C., Zhu, H., Dean, D. R., Dixon, R., and Wood, D. 2009. Genome sequence of *Azotobacter vinelandii*, an obligate aerobe specialized to support diverse anaerobic metabolic processes. *Journal of bacteriology*, 191(14): 4534–45.
- Sha, J., Wang, S., Suarez, G., Sierra, J., Fadl, A., Erova, T., Foltz, S., Khajanchi, B., Silver, A., Graf, J., Schein, C., and Chopra, A. 2007. Further characterization of a type III secretion system (T3SS) and of a new effector protein from a clinical isolate of *Aeromonas hydrophila*—Part I. *Microbial Pathogenesis*, 43(4): 127–146.
- Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F., and Alm, E. J. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science (New York, N.Y.)*, 336(6077): 48–51.
- Sharp, P. and Li, W. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 15(3): 1281–95.
- Sharp, P. M., Emery, L. R., and Zeng, K. 2010. Forces that influence the evolution of codon

- bias. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1544): 1203–12.
- Silver, A. C., Rabinowitz, N. M., Küffer, S., and Graf, J. 2007a. Identification of aeromonas veronii genes required for colonization of the medicinal leech, hirudo verbana. *Journal of bacteriology*, 189(19): 6763–6772.
- Silver, A. C., Kikuchi, Y., Fadl, A. A., Sha, J., Chopra, A. K., and Graf, J. 2007b. Interaction between innate immune cells and a bacterial type iii secretion system in mutualistic and pathogenic associations. *Proceedings of the National Academy of Sciences*, 104(22): 9481–9486.
- Silver, A. C., Williams, D., Faucher, J., Horneman, A. J., Gogarten, J. P., and Graf, J. 2011. Complex evolutionary history of the Aeromonas veronii group revealed by host interaction and DNA sequence data. *PloS one*, 6(2): e16751.
- Slagowski, N. L., Kramer, R. W., Morrison, M. F., LaBaer, J., and Lesser, C. F. 2008. A functional genomic yeast screen to identify pathogenic bacterial proteins. *PLoS pathogens*, 4(1): e9.
- Snel, B., Bork, P., and Huynen, M. a. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome research*, 12(1): 17–25.
- Soucy, S. M., Huang, J., and Gogarten, J. P. 2015. Horizontal gene transfer: building the web of life. *Nature reviews. Genetics*, 16(8): 472–82.
- Stal, L. J. and Krumbein, W. E. 1985. Nitrogenase activity in the non-heterocystous cyanobacterium Oscillatoria sp. grown under alternating light-dark cycles. *Archives of Microbiology*, 143(1): 67–71.
- Stamatakis, A. 2006. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21): 2688–2690.
- Swithers, K. S., Soucy, S. M., and Gogarten, J. P. 2012. The role of reticulate evolution in creating innovation and complexity. *International journal of evolutionary biology*, 2012.

- Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguéz, P., Doerks, T., Stark, M., Müller, J., Bork, P., *et al.* 2011. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl_1): D561–D568.
- Tatum, E. and Lederberg, J. 1947. Gene recombination in the bacterium *Escherichia coli*. *Journal of bacteriology*, 53(6): 673.
- Thomas, C. M. and Nielsen, K. M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews. Microbiology*, 3(9): 711–21.
- Treangen, T. J. and Rocha, E. P. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet*, 7(1): e1001284.
- Tsirigos, A. and Rigoutsos, I. 2005. A new computational method for the detection of horizontal gene transfer events. *Nucleic acids research*, 33(3): 922–33.
- van Dongen, S. 2000. *Graph Clustering by Flow Simulation*. Ph.d. thesis, University of Utrecht.
- van Dongen, S. and Abreu-Goodger, C. 2012. Using MCL to extract clusters from networks. *Methods in molecular biology (Clifton, N.J.)*, 804: 281–95.
- Vanden Bergh, P., Heller, M., Braga-Lagache, S., and Frey, J. 2013. The *Aeromonas salmonicida* subsp. *salmonicida* exoproteome: global analysis, moonlighting proteins and putative antigens for vaccination against furunculosis. *Proteome Science*, 11(1): 44.
- Vernikos, G. S. and Parkhill, J. 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics (Oxford, England)*, 22(18): 2196–203.
- Vila, J., Ruiz, J., Gallardo, F., Vargas, M., Soler, L., Figueras, M. J., and Gascon, J. 2003. *Aeromonas* spp. and Traveler’s Diarrhea: Clinical Features and Antimicrobial Resistance. *Emerging Infectious Diseases*, 9(5): 552–555.
- Vilches, S., Wilhelms, M., Yu, H. B., Leung, K. Y., Tomás, J. M., and Merino, S. 2008. *Aeromonas hydrophila* AH-3 AexT is an ADP-ribosylating toxin secreted through the type III secretion system. *Microbial Pathogenesis*, 44(1): 1–12.

- Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., Surovcik, K., Meinicke, P., and Merkl, R. 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC bioinformatics*, 7: 142.
- Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J., Yoo, H. S., Zhang, C., Zhang, Y., and Sobral, B. W. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*, 42(D1): D581–D591.
- Went, F. 1971. Parallel evolution. *Taxon*, pages 197–226.
- Williams, D., Fournier, G. P., Lapierre, P., Swithers, K. S., Green, A. G., Andam, C. P., and Gogarten, J. P. 2011. A Rooted Net of Life. *Biology Direct*, 6(1): 45.
- Wulff, N. A., Zhang, S., Setubal, J. C., Almeida, N. F., Martins, E. C., Harakava, R., Kumar, D., Rangel, L. T., Foissac, X., Bové, J. M., *et al.* 2014. The complete genome sequence of ‘candidatus liberibacter americanus’, associated with citrus huanglongbing. *Molecular Plant-Microbe Interactions*, 27(2): 163–176.
- Xiang, Z., Tian, Y., and He, Y. 2007. PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biology*, 8(7): R150.
- Yan, Y., Yang, J., Dou, Y., Chen, M., Ping, S., Peng, J., Lu, W., Zhang, W., Yao, Z., Li, H., Liu, W., He, S., Geng, L., Zhang, X., Yang, F., Yu, H., Zhan, Y., Li, D., Lin, Z., Wang, Y., Elmerich, C., Lin, M., and Jin, Q. 2008. Nitrogen fixation island and rhizosphere competence traits in the genome of root-associated *Pseudomonas stutzeri* A1501. *Proceedings of the National Academy of Sciences of the United States of America*, 105(21): 7564–9.
- Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., and Papke, R. T. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome research*, 16(9): 1099–108.