UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

Benchmark de algoritmos para a computação de métricas de similaridade genômica

FELIPE LHYWINSKH GUELLA

Orientadora: Prof[a]. Dra. Luciane Maria Pereira Passaglia
Co-orientador: Dr. Fernando Hayashi Sant'Anna

Porto Alegre, abril de 2019.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

Benchmark de algoritmos para a computação de métricas de similaridade genômica

FELIPE LHYWINSKH GUELLA

Dissertação submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para a obtenção do grau de Mestre em Genética e Biologia Molecular.

Orientadora: Profª. Dra. Luciane Maria Pereira Passaglia
Co-orientador: Dr. Fernando Hayashi Sant'Anna

Porto Alegre, abril de 2019.

“Look again at that dot. That's here. That's home. That's us. On it everyone you love, everyone you know, everyone you ever heard of, every human being who ever was, lived out their lives.”

Carl Sagan, Pale Blue Dot, 1994

AGRADECIMENTOS

São muitas as pessoas às quais devo agradecer por ser possível completar esta dissertação. A jornada foi longa e sei que não poderia ter feito nada disso sozinho.

Primeiramente agradeço à minha companheira, Mariana Teles, por todo apoio e por aguentar minhas crises de ansiedade e variações constantes de humor. Agradeço aos meus irmãos por me fornecerem suporte e apoio nos piores e melhores momentos da minha vida. Agradeço à minha mãe, Luiza, que, apesar de todos os seus defeitos, sempre teve minha felicidade como seu objetivo.

Agradeço, em especial, ao meu falecido pai, por ter acreditado em mim mesmo quando eu fraquejei, por ter me dado o suporte financeiro para poder chegar onde cheguei, mesmo que tenha demorado bastante. Agradeço por ter me tratado sempre com respeito e igualdade, por ter me ensinado a ser honesto e resiliente em frente às adversidades. Tu foste e sempre serás meu maior exemplo de ser humano.

Agradeço à minha orientadora, Luciane, por ter sido extremamente paciente com minhas falhas e minhas constantes faltas e remarcações de reuniões, fruto das minhas crises constantes de ansiedade.

Agradeço também ao meu co-orientador, Fernando, que me criticou quando necessário e me deu os frequentes empurrões e discursos motivacionais para que eu desse o melhor de mim — talvez eu não tenha conseguido chegar nos 100%, juro eu tentei. Sua disposição para me explicar tanto de uma área que eu sabia tão pouco, e, quando necessário, para me apontar as referências para que eu encontrasse as repostas sozinho me tornou um pesquisador melhor do que eu imaginei que pudesse ser, ainda que ainda estou longe de me sentir totalmente seguro e independente.

Agradeço ao Renan por me ajudar tanto na esfera pessoal quanto profissional, espero conseguirmos realizar metade do que discutimos construir.

Aos demais membros do laboratório, agradeço por serem tão queridos e amáveis comigo. Tive um acolhimento que não é fácil de se encontrar em qualquer lugar, sinto orgulho de fazer parte deste grupo excepcional e espero fazer jus aos grandes cientistas que se encontram neste laboratório.

Agradeço aos meus amigos, Zachow, Bombardelli, Fetter, Maurício, Aline e Caroline pelas noites de diversão, jogos e boemia. Sem vocês certamente teria perdido a sanidade.

Agradeço ao PPGBM e ao CNPq pelo suporte financeiro e a oportunidade a mim concedida para realizar este trabalho.

Agradeço, por fim, aos membros da minha banca, que disponibilizaram um pouco de seu tempo para a leitura e avaliação desta dissertação.

SUMÁRIO

Resumo

A hibridização de DNA-DNA (DDH) é ainda considerada a principal técnica para classificação procariótica, apesar de ter certas limitações e já ser considerada obsoleta por muitos pesquisadores. A redução significativa nos custos de sequenciamento genômico, por sua vez, está abrindo espaço para novas métricas baseadas na comparação *in silico* de sequências genômicas. A análise computacional de genomas não apresenta as mesmas limitações da DDH e, desde o desenvolvimento de métricas genômicas, houve um aumento paulatino e constante em seu uso na descrição e reclassificação de espécimes bacterianos. O paradigma da sistemática procariótica está mudando e a tendência é que métricas genômicas se tornem o padrão ouro da área. A métrica mais utilizada é o ANI (*average nucleotide identity*), mas, além dela, surgiram outras métricas que convergem para o mesmo objetivo de comparar genomas bacterianos para delimitação de espécie. Não obstante, poucos estudos de fato compararam essas métricas entre si em termos de performance e intercambialidade. É necessário, portanto, uma análise abrangente que possibilite uma padronização de diversas métricas, com o objetivo de se desenvolver um esquema de classificação e identificação padrão baseado nas métricas genômicas mais eficientes na discriminação de espécies bacterianas. A primeira parte da dissertação envolveu a avaliação de métricas genômicas em relação a diversos parâmetros, utilizando os resultados de ANIb como referência e genomas de *Paenibacillus* como conjunto de dados. Os resultados de tempo de execução indicam que o TETRA é a métrica mais rápida, seguido do MUMi e do ANIm, enquanto o GGD, ANIb, gANI e OrthoANI exigiram maior tempo de computação. Todas as métricas tiveram valores de coeficiente de correlação elevado ($\geq$ 0.9), com exceção do TETRA ($\approx$0.75). A especificidade, em relação aos resultados do ANIb, foi elevada para todas as métricas ($\geq$0.9), enquanto a sensibilidade foi elevada para todas ($\geq$0.9), exceto para o gANI, GGD e MUMi (entre 0.7 e 0.8). Em relação a testes de robustez, utilizando genomas artificialmente contaminados, houve uma variação mínima entre as métricas que utilizam cálculos baseados em alinhamento, exceto com o MUMi, que apresentou variação significativa nos resultados. O TETRA, em contrapartida, teve a maior variação das métricas testadas, resultados que poderiam comprometer a definição de espécie. Considerando todos os parâmetros e condições testadas, o ANIm foi uma das melhores métricas testadas, devido a sua robustez, seu tempo de execução e sua

elevada similaridade de resultados com o ANIb. As outras métricas que derivaram do ANIb — OrthoANI e gANI — tiveram pouca variação em termos de performance. Apesar da grande velocidade das análises do MUMi e do TETRA, eles não apresentam a mesma robustez que as outras métricas. A segunda parte da dissertação foi um estudo derivado dos dados gerados na primeira parte e envolveu a reclassificação das espécies bacterianas *Paenibacillus durus* e *Paenibacillus azotofixans*. Os resultados das métricas, aliados às análises filogenéticas — como MLSA e reconstrução do proteoma *core* — e características morfofisiológicas e quimiotáxicas, possibilitaram a reclassificação dessas espécies. Excetuando o resultado da análise de identidade do gene do rRNA 16S — que definia ambos como da mesma espécie —, todos resultados indicaram a separação desses dois micro-organismos em duas espécies independentes. A dissertação apresentou as qualidades e limitações de diversas métricas disponíveis atualmente e um exemplo prático de como esses dados quantitativos podem ser úteis na área de sistemática procariótica.

Abstract

DNA-DNA hybridization (DDH) is still considered the main method for genomic prokaryotic classification, despite having certain limitations and already being considered as an obsolete approach by several researchers. The significant reduction in genomic sequencing costs, on the other hand, allowed that several metrics based on comparative genomics were more utilized in prokaryotic taxonomy. The most utilized metric is ANI (average nucleotide identity), but besides it, many other genomic metrics were developed. Nevertheless, few studies compared these metrics among each other with respect to performance and interchangeability. Therefore, it is necessary a broad analysis that allows the standardization of these metrics, aiming the development of a classification and identification scheme based on efficient genomic metrics for the discrimination of prokaryotic species. The first part of our study is related to the evaluation of several parameters of genomic metrics, using ANIb results as reference and *Paenibacillus* genomes as dataset. Runtime results shows that TETRA is the fastest metric, followed by MUMi and ANIm, while GGD, ANIb, gANI and OrthoANI were significantly slower. All metrics had high correlation coefficients (≥0.9), except for TETRA (≈0.75). Specificity values, when comparing to ANIb results, were high for all metrics (≥0.9), while sensitivity

values were high for almost all metrics (≥0.9), apart from gANI, GGD and MUMi — that were between 0.7 and 0.8. When comparing artificially contaminated genomes for robustness evaluation, the variation on alignment-based had minimum variation between results, with the exception of. TETRA, on the other hand, had the highest variation of results on all tested metrics. Considering all parameters and tested conditions, ANIm was one of the most reliable and efficient metrics tested, due to its robustness, runtime and similarity to ANIb results. All other metrics derived from ANIb — OrthoANI and gANI — had little difference on performance compared to ANIb. Despite their fast runtime analysis, MUMi and TETRA do not have the same robustness as the other metrics. The second part of the study utilized the data derived from the first one, and it was the reclassification of the bacterial species *Paenibacillus durus* and *Paenibacillus azotofixans*. All metrics results, combined with phylogenetic analysis — like MLSA and core proteome reconstruction — and morphophysiological and chemiotaxis results, allowed the reclassification of *P. durus* and *P. azotofixans*. Excluding 16S rRNA gene phylogeny — that defined both bacteria as the same species —, all results indicate that both microorganisms belong to two independent species. Our study presented qualities and limitations of several metrics currently available, and a practical example of how these metrics can be useful in the prokaryotic systematic field.

Lista de abreviaturas

AAI — *Average Aminoacid Identity*

ANI — *Average Nucleotide Identity*

ANIb — BLAST ANI

ANIm — MUMmer ANI

BLAST — *Basic Alignment Search Tool*

BMSB — *Bergey's Manual of Systematic Bacteriology*

dDDH — digital DDH

DDH — Hibridização DNA-DNA

gANI — *Genome-wide ANI*

GBDP — *Genome BLAST Distance Phylogeny*

GGDC — *Genome-to-genome distance calculator*

HGT — *Horizontal Gene Transfer*

HTS — *High-throughput sequencing*

ICSP — *International Committee on Systematics of Prokaryotes*

IJSEM — *International Journal of Systematics and Evolutionary Microbiology*

MUMi — *Maximum unique matches index*

NCBI — *National Center for Biotechnology Information*

NGS — *Next Generation Sequence*

OrthoANI — *Orthologous* ANI

PCR — Reação de cadeia de polimerase

TETRA — *Tetranucleotide usage pattern*

MLSA — *Multi-locus Sequence analysis*

1. INTRODUÇÃO

1.1. Delineamento de Espécie em Bactérias

Diferentemente de organismos sexuados, onde o delineamento de espécie se dá especialmente por fatores reprodutivos, bactérias e outros seres vivos assexuados apresentam certos desafios na demarcação do que se caracteriza uma espécie e quais fatores devem qualificar se dois indivíduos próximos pertencem a duas espécies diferentes ou à mesma espécie. Ao longo dos anos, o modelo de classificação de bactérias foi se atualizando e se especificando, à medida que novas tecnologias permitiam uma observação mais complexa desses organismos.

A primeira bactéria foi observada e descrita em 1676 por Antonie van Leeuwenhoek e nomeada, na época, de "animacules" (Parker 1965). Como o termo protista só foi cunhado em 1866 (Haeckel 1866) por Ernst Haeckel, as bactérias foram previamente classificadas como plantas. Após diversos sistemas de classificação e subdivisão dos reinos e domínios dos seres vivos (Chatton 1925; Copeland 1938; Whittaker 1969), foi apenas em 1990 que Carl Woese e colaboradores incluíram as bactérias em um domínio próprio na taxonomia (Woese et al. 1990). Em seu artigo, o autor discute que diferentemente da classificação dos organismos eucariotos — onde seu grupo filético é definido por características específicas e complexas — os procariotos foram todos unidos pela falta da presença das características que definem os eucariotos. Dessa forma, o sistema de classificação dos procariotos era todo baseado em fatores negativos, tornando-se um modelo extremamente superficial, para não se dizer completamente errado. O autor conclui dividindo o domínio protista em dois: o domínio Bacteria e o domínio Archea (Woese et al. 1990). Esse histórico de divisões e rearranjos na taxonomia, não obstante, só foi possível pela evolução dos estudos moleculares e genéticos desses micro-organismos. Apesar de todos os avanços na delimitação de espécie em bactérias, contudo, até o presente momento, não há um modelo oficial de classificação bacteriana (Euzéby 1997), apenas regras fixadas de nomenclatura (Parte 2018). Os primeiros campos de pesquisa microbiológica, além disso, não tinham interesse na classificação desses organismos, eles eram puramente focados na relação do organismo com certos campos de interesse, como medicina, engenharia sanitária, medicina veterinária e agricultura (Winslow 1914). Dessa forma, não havia um sistema recomendado de classificação bacteriana até o início do século 20 (Winslow 1914), pois

cada área de estudo classificava de acordo com as características de interesse para seu campo de trabalho.

De acordo com o "Bergey's Manual of Systematic Bacteriology" (BMSB) (Brenner et al. 2005), a taxonomia bacteriana se divide em três áreas fortemente correlacionadas: classificação, nomenclatura e identificação. O passo inicial, antes de nomear e classificar um espécime bacteriano, é o de identificação, pois "nenhum organismo pode ser classificado antes que suas características morfológicas, culturais, fisiológicas e patogênicas tenham sido determinadas através de estudos detalhados" ("Bergey's Manual of Determinative Bacteriology") (Perkins 2008). Para a classificação, é necessário ter-se conhecimento do que se define como uma espécie bacteriana, que é estabelecida como "uma ou mais estirpes distintas que apresentam elevada similaridade em suas características de organização essenciais" (Brenner et al. 2005). Estas estirpes, por sua vez, apresentam uma estirpe-tipo — a representante do grupo — que, normalmente, é primeira bactéria isolada da espécie e não necessariamente representa a estirpe mais comum do grupo (Brenner et al. 2005). Dessa forma, tanto a definição de espécie bacteriana é subjetiva à interpretação do que seriam as características essenciais, quanto a estirpe-tipo não representa, obrigatoriamente, a mediana da espécie.

Os primeiros esforços para delimitação de espécie baseados em características fenotípicas dos isolados surgiram de um trabalho pioneiro de 1962, que tinha como objetivo classificar espécies de enterobactérias em ambientes clínicos (Edwards e Ewing 1962). Os organismos foram separados quanto ao tipo de carboidrato consumido, ao consumo ou não de ureia, à motilidade, dentre outras características (Ewing et al. 1969), que possibilitaram, para o objetivo dos autores, uma separação satisfatória. Quase todas as análises bacterianas, até meados da década de 70, seguiram modelos semelhantes de classificação, se baseando puramente em análises sorológicas. Outro modelo que teve um papel essencial para a taxonomia bacteriana — e atualmente é um dos pontos centrais dela — foi a filogenia, que, inicialmente, se baseou em características fenotípicas para suas reconstruções; entretanto, devido ao fato de procariotos não possuírem fosseis para uma análise de seus ancestrais, como no caso dos eucariotos, muitos biólogos inicialmente abordaram essa metodologia com certa desconfiança (Stanier e Van Niel 1941).

No final da década de 60 o primeiro modelo genômico para delimitação de espécies bacterianas foi idealizado: o grau de hibridização DNA-DNA (DDH) (Brenner et al. 1969).

Tornando-se, também, a primeira "característica" com uma métrica fixa: quando o percentual de hibridização dos genomas de dois organismos ultrapassar o valor de 70%, ambos são considerados membros da mesma espécie (Moore et al. 1987) (Figura 1). A partir do desenvolvimento desta técnica, a similaridade entre dois organismos ao nível de espécie pode ser objetivamente mensurada, não dependendo mais de abordagens subjetivas. O impacto do DDH se comprova com o fato de que, em 1987, aproximadamente 60% dos artigos de descrição de novas espécies incluíram esta técnica para sedimentarem seus achados, chegando à 75% em 1993 (Stackebrandt e Goebel 1994). Para classificações de níveis hierárquicos superiores ao gênero, entretanto, permaneceram as metodologias convencionais.

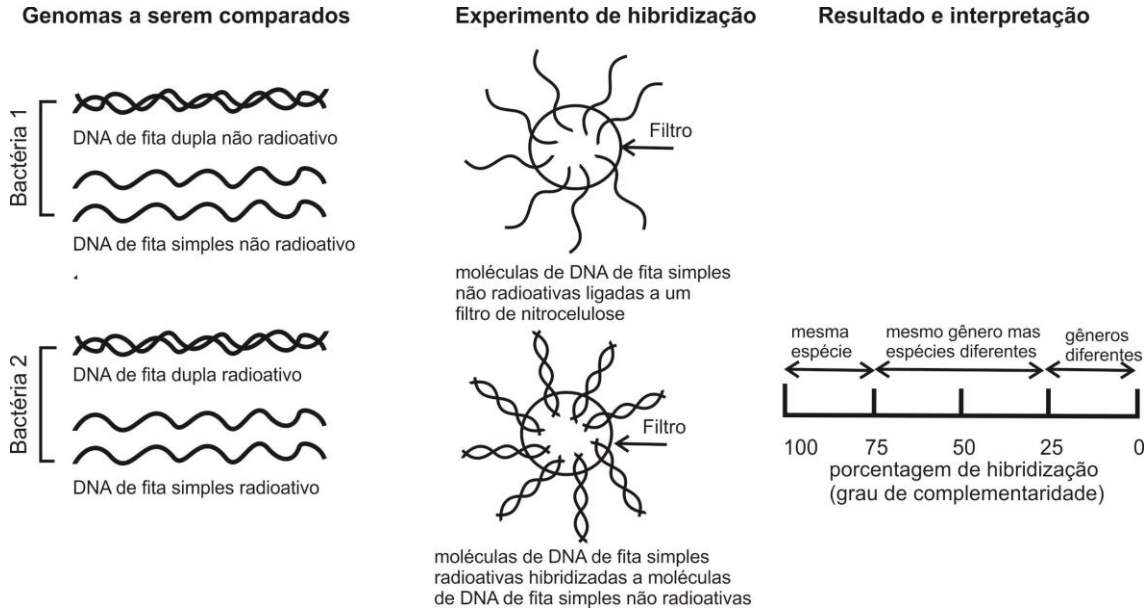

Figura 1: Exemplo de um experimento de hibridização entre DNAs de duas bactérias (adaptado de http://www.biologydiscussion.com/bacteria/bacterial-taxonomy/bacterial-taxonomy-meaning-importance-and-levels/54679, último acesso em: 10/02/2019)

Em 1990, o "Ad Hoc Committee on Approaches to Taxonomy within the Proteobacteria" já sugeria a análise da similaridade gênica e filogenia do rRNA 16S para a delimitação de espécie bacterianas (Murray et al. 1990). Em 1994, Stackebrandt e Goebel (Stackebrandt e Goebel 1994) definiram que um valor inferior a 97,5% de identidade dos genes rRNA 16S entre dois organismos indicaria que eles não pertenceriam à mesma espécie, pois com essa identidade seria muito improvável que eles atingissem um valor de DDH de 70% ou superior. Em 2014, este valor foi reajustado para 98,65% (Kim et al.

2014). Como dito anteriormente, em 75% das novas espécies descritas em 1993 se utilizou o DDH para sua descrição, mas, muitos desses estudos utilizaram concomitantemente a análise filogenética do gene do rRNA 16S para corroborar seus achados, e 14% dos 25% restantes utilizaram rRNA16S exclusivamente como método genômico de classificação, totalizando 89% das novas espécies descritas que utilizaram um ou mais métodos genômicos de classificação. Com a sedimentação do DDH e da análise do rRNA 16S — atualmente considerados, em conjunto, os testes de referência para classificação genômica de bactérias — na demarcação de espécie bacteriana, surgiu o consenso de que a sistemática bacteriana deveria seguir um modelo polifásico, utilizando conjuntamente métodos genômicos e fenotípicos (Vandamme et al. 1996).

1.2. Desafios da classificação bacteriana

Os desafios para classificação de espécies de seres vivos são muitos, abrangendo tanto discussões filosóficas, quanto biológicas. No caso dos procariotos, a dificuldade é dobrada, pois além de não se reproduzirem de forma sexuada, muitos de seus representantes são de difícil isolamento ou ainda não foram isolados com sucesso. Um recente estudo demonstrou que ao sequenciar todos os genomas procarióticos provenientes de diversificadas amostras ambientais, grande parte das linhagens sequenciadas não continham representantes previamente isolados e descritos (Hug et al. 2016). Isso sugere que a diversidade dos procariotos em geral é gravemente subestimada. Estima-se que existam, aproximadamente, 4-6 X 10^{30} células procarióticas no planeta (Whitman et al. 1998) e, entre essas células, em torno de 10^5 a 10^7 espécies bacterianas (Finlay et al. 1997), sendo que apenas uma pequena fração já foi isolada e classificada — em torno de 1,5 X 10^4 (Parte 2018). Até 2004, apenas 4500 espécies haviam sido descritas (Garrity et al. 2004), e, por isso, o potencial biológico das bactérias foi frequentemente subestimado e subvalorizado, dificultando ainda mais a classificação de novas espécies. O aumento do interesse econômico no uso de bactérias e seus derivados para indústria e agricultura, entretanto, levou ao aumento também dos recursos destinados à sua pesquisa.

Outro fator que ocorre com freqüência nos procariotos, que se apresenta como um dos maiores obstáculos para delimitação de espécie bacteriana, é a transferência horizontal de genes (“horizontal gene transfer” — HGT) (Syvanen 1994). Esse fenômeno só foi efetivamente confirmado após o desenvolvimento de técnicas de sequenciamento genômico e, por isso, foi desconsiderado por um longo período na sistemática bacteriana,

tendo um profundo impacto na organização filética dos procariotos (Pennisi 1998). Ao se comparar a reconstrução filogenética do gene do rRNA 16S com as reconstruções de outros genes, por exemplo, diversas discrepâncias são encontradas entre as árvores (Christensen et al. 2004; Case et al. 2007) (Figura 2). Isso se deve ao fato de que, mesmo sendo um gene essencial para a manutenção da vida da célula, há estudos comprovando que, em alguns casos, pode ocorrer transferência horizontal de genes de rRNA, inclusive do 16S (Schouls et al. 2003; Tian et al. 2015; Sato e Miyazaki 2017).

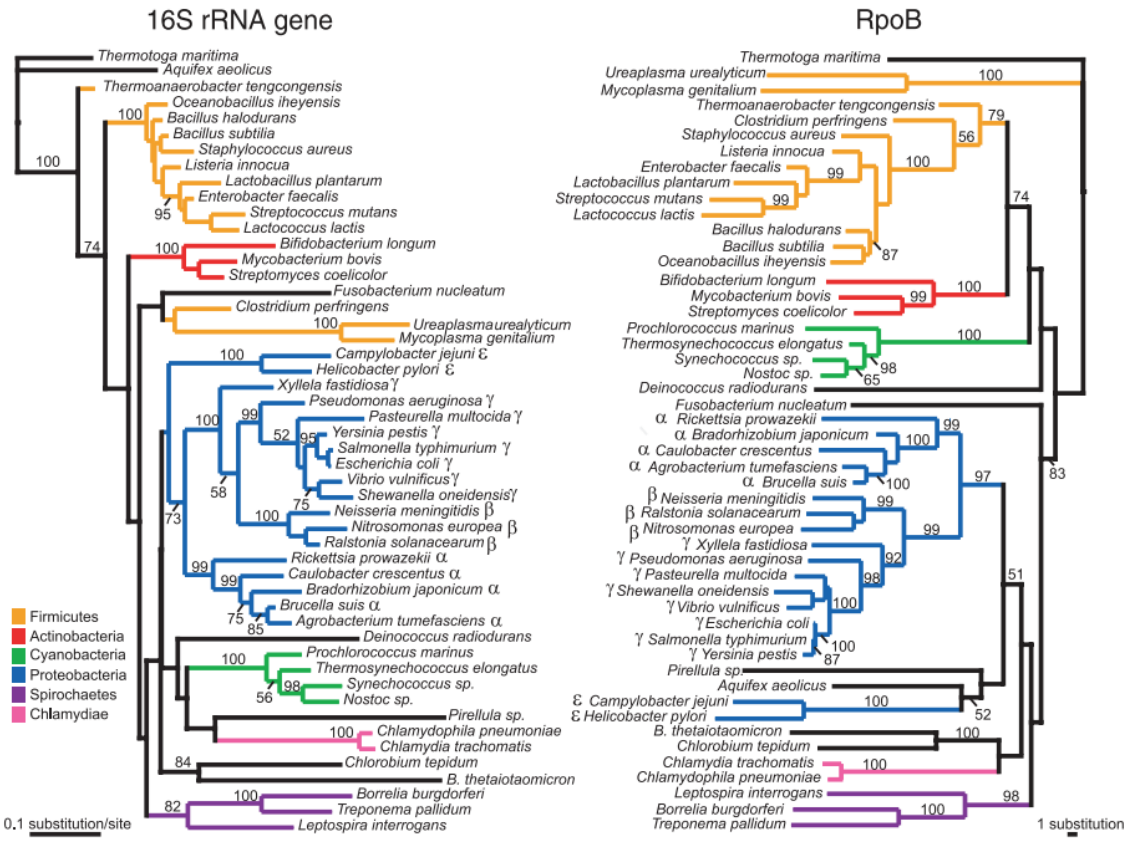

Figura 2: Diferença entre a reconstrução filogenética do gene rRNA 16S e do gene *rpoB* de diversos filos bacterianos [Figura retirada de (Case et al. 2007)].

1.3. Inconsistências nos testes de referência

As dificuldades na classificação de espécies bacterianas podem ser vistas ao analisar, por exemplo, a taxonomia do gênero *Paenibacillus*. Separado do gênero *Bacillus* em 1994 (Ash et al. 1993; 1994) e em 2009 — em conjunto com mais 7 outros gêneros — tornou-se o gênero tipo de sua própria família: Paenibacillaceae (De Vos et al. 2009). Uma análise da reconstrução filogenética usando o gene do rRNA 16S da família, inicialmente, já indicou problemas de taxonomia dentro do gênero, devido à sua organização parafilética (Zeigler 2016) (Figura 3). O uso da metodologia de referência da análise de identidade do rRNA 16S não apresenta resolução suficiente para delimitação de suas espécies (Sant'Anna et al. 2017; Ambrosini et al. 2018; Sant'Anna et al. 2018), assim como em outros grupos bacterianos (Fox et al. 1992). Isso se deve à alta taxa de conservação dos genes de rRNA, comparado à diversidade gênica do resto do genoma dessas espécies.

As metodologias genômicas de referência para delimitação de espécie bacteriana atuais apresentam certas limitações e inconsistências que podem prejudicar a correta classificação de espécies bacterianas. O DDH é um procedimento de difícil execução, com mais de um tipo de metodologia disponível e, portanto, de baixa reprodutibilidade entre laboratórios, e, devido a isso, não gera dados cumulativos (Rosselló-Móra 2012). Dessa forma, apesar de ter auxiliado nos esforços iniciais para a taxonomia bacteriana, a necessidade de substituição para esta metodologia já é um consenso na comunidade científica (Stackebrandt et al. 2002).

A análise da identidade do gene do rRNA 16S, por sua vez, apresenta outras dificuldades — além das já citadas anteriormente — como a presença de mais de uma cópia do gene dentro do genoma, que eventualmente pode apresentar valores de identidade intragenômicos abaixo do limiar de espécie de 98,7% (Stackebrandt 2011; Guella et al. 2019), o que poderia dificultar uma comparação adequada entre dois genomas. Diversas vezes o ponto de corte para delimitação de espécie foi alterado (Stackebrandt e Goebel 1994; Stackebrandt 2011; Kim et al. 2014), e a única garantia atualmente desta metodologia é que para valores abaixo de 98,7% de identidade há divergência de espécie (Stackebrandt 2011). Valores de identidade acima deste limiar, necessitam de outras técnicas para confirmar a delimitação de espécie, devido à sua baixa resolução a partir deste ponto (Rosselló-Móra 2012; Sant'Anna et al. 2017; Ambrosini et al. 2018; Sant'Anna

et al. 2018); É proposto, atualmente, o uso de genes estruturais ("housekeeping genes") como forma de contornar esta falha metodológica (Stackebrandt et al. 2002).

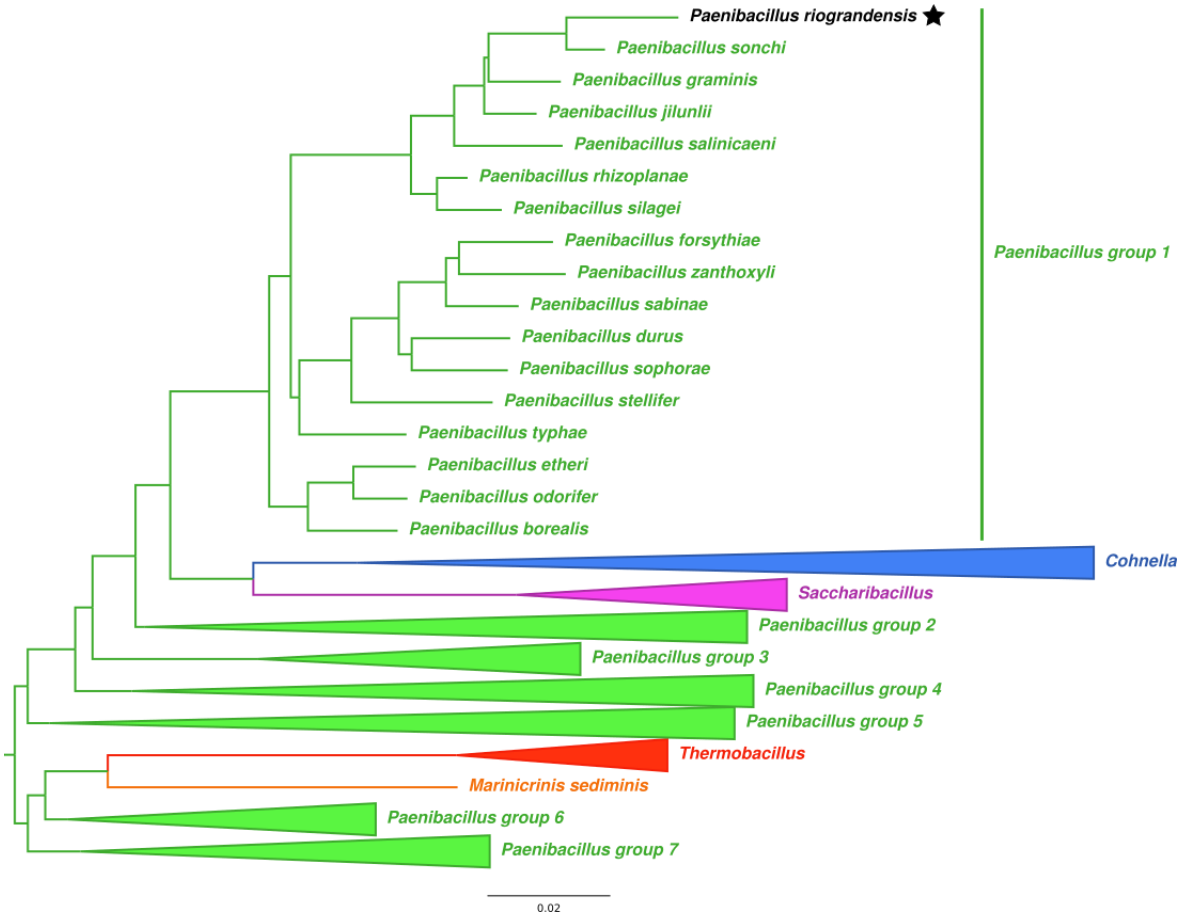

Figura 3: Reconstrução filogenética do gênero *Paenibacillus* e de gêneros próximos da mesma família.

1.4. Classificação genômica empregando a bioinformática

Apesar da análise do escore de identidade genômica do gene do rRNA 16S utilizar ferramentas computacionais para classificação genômica, esta técnica pouco aproveita a capacidade de processamento e o potencial desta ferramenta que está se tornando cada vez mais essencial para o estudo da ciência em diversas áreas de conhecimento (Lunteren 2016). Na área de genômica bacteriana o processo de inclusão da bioinformática iniciou-se com a implementação e aprimoramento das técnicas de sequenciamento genômico. Um exemplo do impacto do sequenciamento genômico no estudo dos procariotos se apresenta pelo aumento vertiginoso de espécies descritas entre 1980 e 2017, pois, apesar de as bactérias terem sido descobertas há mais de três séculos, de acordo com o "International Committee on Systematics of Prokaryotes" (ICSP), em 1980, havia oficialmente pouco mais de 1800 espécies bacterianas corretamente nomeadas (1980), e, em 2017, esse valor já ultrapassava 15000 espécies bacterianas (Parte 2018) (Tabela 1).

Tabela 1: Inclusão de novas espécies ou níveis taxonômicos entre 1980 e início de 2017 (modificado de LPSN[1] – último acesso 28/02/2019).

Ano	Classe	Subclasse	Ordem	Subordem	Família	Gênero	Espécie	Subesp.	Total
Ant	7	1	21	3	66	290	1792	131	2311
1980	-	-	-	-	2	10	49	1	62
1981	-	-	3	-	5	22	103	7	140
1982	-	-	1	-	3	16	102	13	135
1983	-	-	-	-	2	27	166	17	212
1984	-	-	1	-	4	31	161	23	220
1985	-	-	-	-	-	29	125	14	168
1986	-	-	1	-	3	27	176	16	223
1987	-	-	2	-	2	19	100	11	134
1988	5	-	1	-	1	30	144	8	189
1989	-	-	2	-	5	23	167	19	216
1990	-	-	-	-	3	21	148	30	202
1991	-	-	-	-	5	21	145	12	183
1992	-	-	-	-	1	12	122	16	151
1993	-	-	1	-	2	36	178	6	223
1994	-	-	-	-	-	42	161	6	209
1995	-	-	1	-	2	37	217	11	268
1996	-	-	1	-	3	46	232	20	302
1997	1	5	6	10	19	42	223	4	310
1998	1	-	-	-	1	55	256	6	319
1999	-	-	-	-	4	79	273	14	370
2000	-	-	-	-	8	76	275	11	370
2001	-	-	-	-	1	68	356	8	433
2002	42	-	25	-	14	72	350	9	512
2003	1	-	1	-	3	75	372	20	472
2004	-	-	3	-	13	80	435	13	544
2005	1	-	12	-	20	105	528	6	672
2006	6	-	19	1	38	118	593	4	779
2007	3	-	5	3	8	135	631	19	804
2008	1	-	5	-	11	116	597	8	738
2009	2	-	7	2	12	112	663	15	813
2010	7	1	9	-	25	105	611	13	771
2011	2	-	3	1	11	105	619	13	754
2012	9	-	8	-	26	100	655	11	809
2013	16	-	9	4	16	152	666	12	875
2014	4	-	14	-	22	135	811	10	996
2015	1	-	22	-	21	237	1009	8	1298
2016	3	-	8	-	26	186	1056	13	1292
2017*	4	-	5	-	7	38	181	3	238
Total	116	7	196	24	415	2930	15448	581	19717

1.5. Do sequenciamento de Sanger aos “high-throughput sequencing” e suas consequências

O sequenciamento genômico iniciou-se de maneira simples: o método de Sanger utilizava nucleotídeos modificados, que causavam a interrupção do alongamento da cadeia de DNA, juntamente com um iniciador de DNA, uma DNA polimerase, nucleotídeos normais e DNA de fita simples como molde. O processo de alongamento era executado em quatro tubos de ensaio — cada um com um tipo de nucleotídeo interruptor — e, depois, as amostras eram aplicadas em um gel de eletroforese e a posição dos nucleotídeos era descoberta pela banda de cada fragmento de DNA interrompido pela adição do nucleotídeo terminador correspondente (Figura 4) (Sanger e Coulson 1975). Essa metodologia foi aprimorada com a aplicação de corantes diferentes para cada nucleotídeo, permitindo a execução da reação em um único tubo de ensaio, com o resultado da eletroforese sendo lido e interpretado através de cromatografia (Prober et al. 1987). Em 1981, o primeiro modelo automatizado de sequenciamento, utilizando capilares, foi introduzido (Jorgenson e Lukacs 1981).

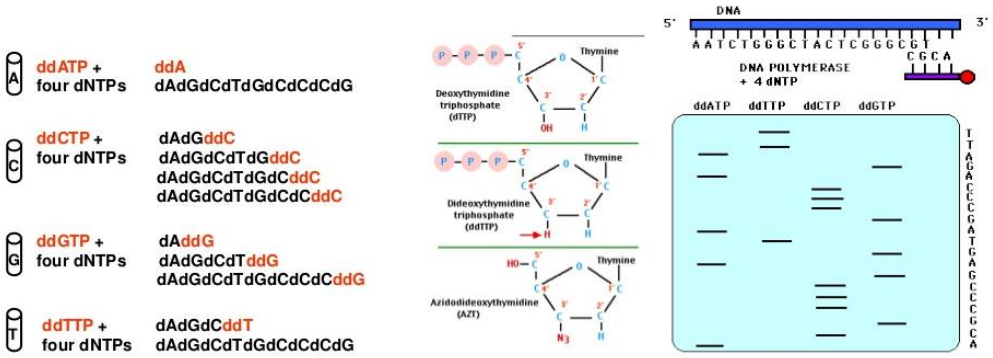

Figura 4: Exemplo do processo de sequenciamento pelo método de Sanger (recuperado de https://www.slideshare.net/SurenderRawat3/dna-sequencing-41318444/9 - último acesso 28/02/2019).

Em 1987 a empresa “Applied Biosciences” comercializou o primeiro sequenciador automatizado, baseado no método de Sanger com a utilização dos capilares. Entretanto, apesar de ter se tornado um marco no sequenciamento genômico, seu custo e tempo de execução ainda tornavam seu uso inviável para a maioria dos laboratórios de pesquisa. A descoberta da reação de cadeira de polimerase (PCR) (Mullis et al. 1986) e a consequente automatização da amplificação de fragmentos de DNA ajudou a mudar esse quadro, pois levou ao desenvolvimento de diversas técnicas que viriam a ser chamadas de “Next

Generation Sequencing" (NGS). A "Illumina dye sequencing", uma dessas técnicas e a mais utilizada atualmente, foi desenvolvida em 2006 (Bentley et al. 2008) e — em conjunto com o pirosequenciamento (Margulies et al. 2005) — abriu espaço para o sequenciamento em largar escala a preços acessíveis, como evidenciado nas Figuras 5 e 6. A metodologia Illumina se divide, simplificadamente, em três passos: amplificação, sequenciamento e análise. Na amplificação, o genoma de interesse é fragmentado em pequenas sequencias de DNA e é amplificado em uma placa, gerando diversas cópias. O sequenciamento se dá após a lavagem da placa, onde um tipo de nucleotídeo com fluorescência é adicionado de cada vez em diversos ciclos. Dessa forma, a ordem dos nucleotídeos de cada fragmento é mapeada. No processo de análise, os pontos de sobreposição entre os fragmentos resultantes (ou "contigs") são alinhados e o genoma é montado como um quebra-cabeça. O genoma montado é comparado, posteriormente, com um genoma de referência, caso houver algum disponível (Morozova e Marra 2008) — caso não haja genoma de referência é utilizada a estratégia de montagem *de novo*.

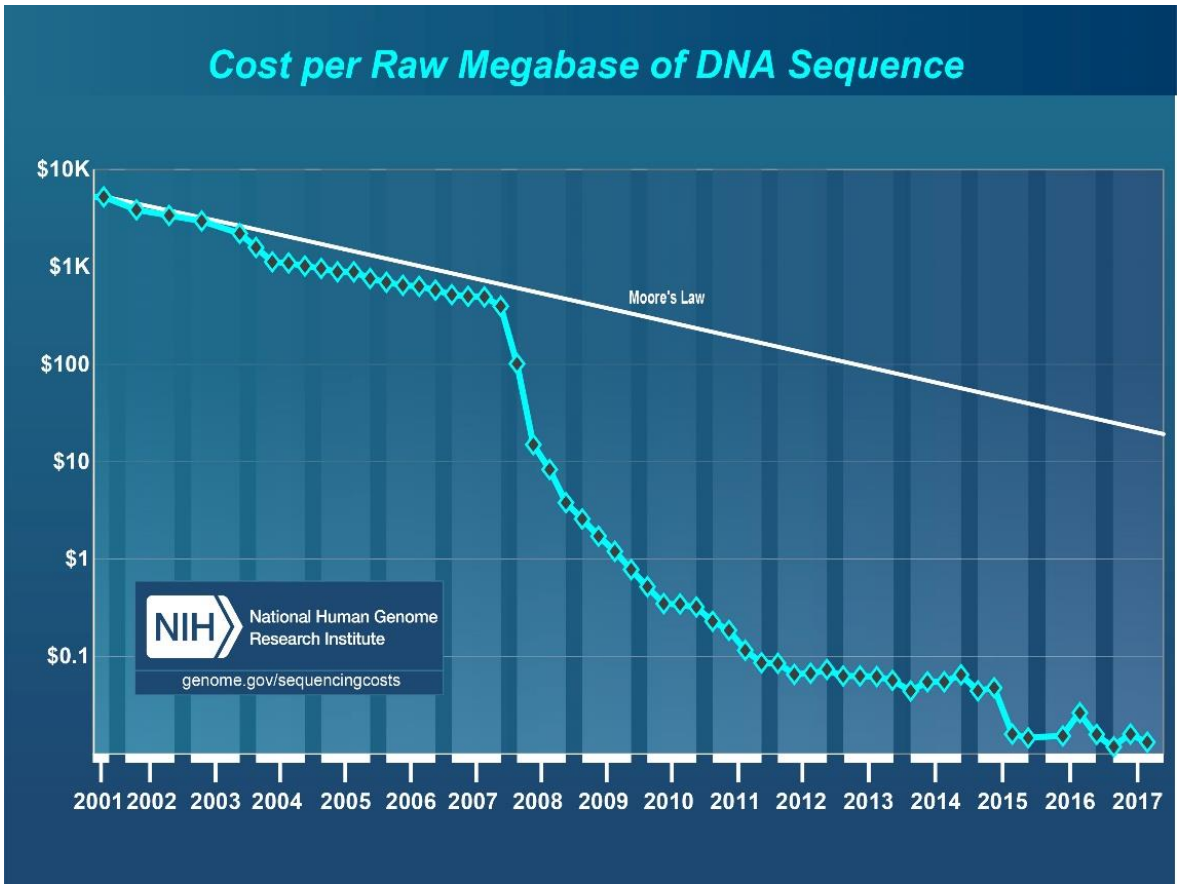

Figura 5: Variação do custo, em dólares, do sequenciamento de DNA por megabase entre os anos de 2001 e 2017. (retirado de https://www.genome.gov/27541954/dna-sequencing-costs-data/ — último acesso em 28/02/2019).

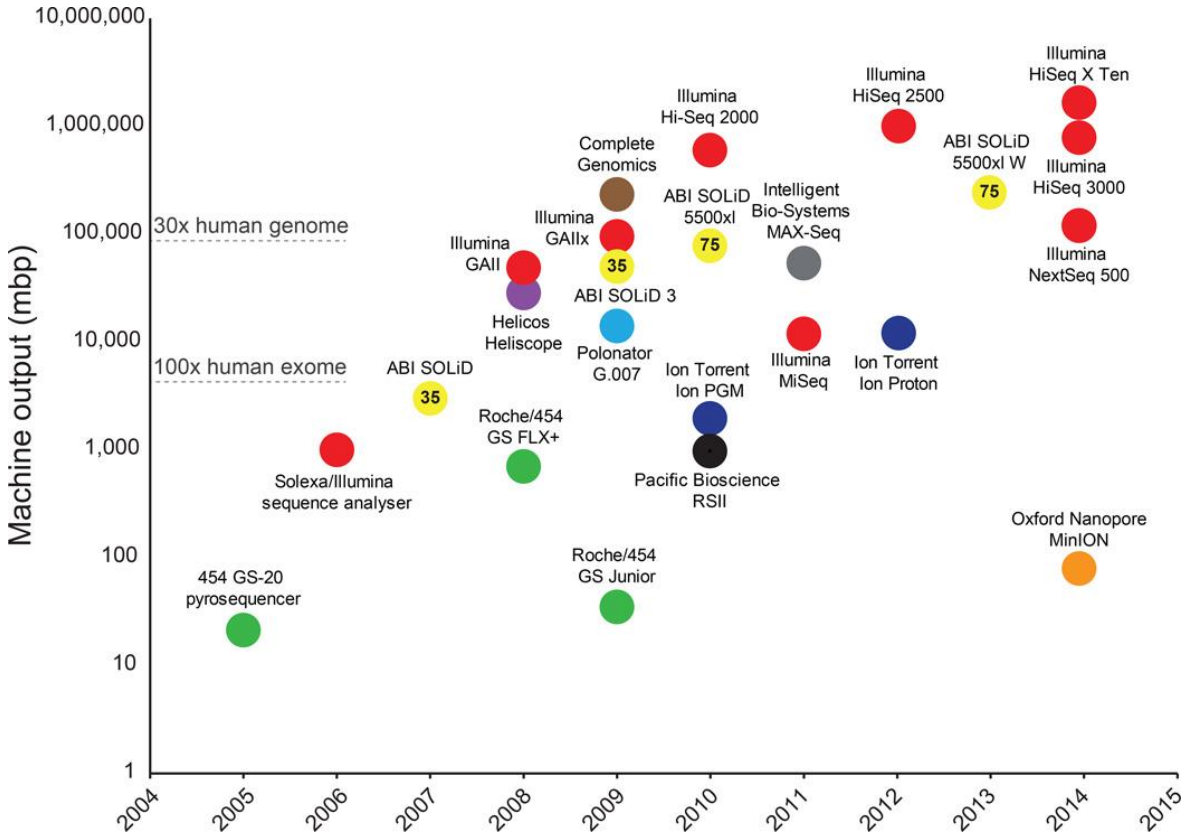

Figura 6: Gráfico de ano de lançamento comercial versus rendimento por execução de instrumentos de sequenciamento. [Retirado de (Reuter et al. 2015)].

O aprimoramento da metodologia Illumina e o advento de novas técnicas avançadas de sequenciamento levaram à mudança de nomenclatura para sequenciamento de alto desempenho ("High-throughput sequencing" — HTS), devido à diferença de rendimento entre a segunda geração (NGS) e a geração atual. Como visto nas Figuras 5 e 6, a redução do custo de sequenciamento, aliado ao aumento de desempenho dos equipamentos comerciais, possibilitaram, especialmente no estudo de procariotos, um uso mais abrangente do sequenciamento do genoma completo de espécimes bacterianas e sua deposição em bancos de dados de amplo acesso, como o GenBank[2]. De acordo com o banco de dados RefSeq do "National Center for Biotechnology Information" (NCBI) — um banco de dados apenas com genomas curados — em novembro de 2013 havia 31.646 organismos diferentes depositados[3]. Em janeiro de 2019 este número já havia mais do que duplicado — 86,867 organismos diferentes depositados[4]. Devido a isso, a "International Journal of Systematic and Evolutionary Microbiology" (IJSEM), a mais importante revista científica de sistemática microbiana, passou a recomendar fortemente, desde 2018, o deposito do genoma completo dos organismos descritos em artigos de novas espécies para assegurar sua publicação[5].

Apesar da recomendação da IJSEM vir somente a partir de 2018, o sequenciamento e depósito de genomas já têm sido uma prática relativamente comum para outros estudos, além da sistemática bacteriana: como, por exemplo, no estudo de novos alvos de antibióticos no estudo de patógenos (Donkor 2013), na análise de genes de interesse e seu ambiente metabólico (Fernandes et al. 2014), no estudo da formação de sabor em alimentos (McAuliffe et al. 2018) e no estudo de microbiota humana (Qin et al. 2010). O depósito constante de sequencias em bancos de dados gera uma grande quantidade de informações — os chamados "Big Data" — que podem ser exploradas pela sistemática bacteriana para aprimorar o modelo de classificação bacteriana, usando grandes partes ou a totalidade do genoma dos micro-organismos para comparar sua similaridade, em oposição ao uso de apenas o gene do rRNA 16S. Esse, como discutido anteriormente, por mais conservado que seja, pode apresentar variações intraespecíficas e, em alguns casos, transferência horizontal de organismos distantes evolutivamente.

1.6. Transição do uso do DDH para o uso "overall genome relatedness indices" (índices de relação genômica geral — OGRIs)

Com o objetivo de entrar na nova era do "Big Data" e desfrutar da grande quantidade de genomas disponíveis publicamente, foi desenvolvida a "average nucleotide identity" (ANI) (Konstantinidis e Tiedje 2005), que calcula a identidade média entre dois genomas bacterianos — metodologia que atualmente está sendo consideradaa sucessora do DDH como padrão ouro de classificação de espécies bacterianas (Goris et al. 2007; Richter e Rossello-Mora 2009; Mahato et al. 2017). Para comparar dois organismos, essa metodologia divide o genoma da primeira bactéria em diversos fragmentos de 1020 nucleotídeos e é realizada uma busca com cada fragmento ("query"), utilizando o algoritmo "Basic Local Alignment Search Tool" (BLAST), no genoma da segunda bactéria ("reference"). É feito uma média de todos os "hits" que apresentarem um valor de identidade de sequência acima de 30% e região alinhável de pelo menos 70% com o genoma "reference". O processo é repetido com a inversão dos genomas das bactérias analisadas — o genoma da primeira bactéria é usado como "reference" enquanto que o genoma da segunda bactéria é fragmentado e usado como "query". Faz-se, então, uma média dos dois resultados, configurando o valor final de ANI (Goris et al. 2007). Valores de ANI acima de 90% demonstraram boa correlação com resultados de DDH (Goris et al. 2007), sendo que o ponto de corte inicial de valor de ANI determinado para delimitação de espécie, que corresponderia a um valor de DDH de 70% ou superior, foi de 94% ou superior entre os dois genomas envolvidos (Richter e Rossello-Mora 2009). Uma metodologia semelhante foi desenvolvida para comparar a identidade média de aminoácidos, o "average aminoacid identity" (AAI) (Konstantinidis e Tiedje 2005).

Apesar de ANI ter sido a primeira metodologia robusta de comparação genômica para delimitação de espécie, e a mais popularmente usada, outras métricas foram desenvolvidas através dos anos. Algumas destas são diretamente derivadas de ANI, como ANIm (Richter e Rossello-Mora 2009), que utiliza o algoritmo MUMmer (Kurtz et al. 2004) como substituto do BLAST; OrthoANI (Lee et al. 2016), que, diferente de seu sucessor, fragmenta ambos genomas, ao invés de usar um "query" e um "reference", e considera apenas os fragmentos que apresentam "best hit" recíproco para calcular a identidade média dos genomas, utilizando tanto o algoritmo BLAST como o algoritmo USEARCH (Edgar 2010); e o genome-wide ANI (gANI) (Varghese et al. 2015), que

utiliza apenas as regiões codificantes dos genomas em sua análise. Como forma de especificar que a metodologia ANI original utiliza o algoritmo BLAST em sua análise, esta é comumente intitulada ANIb (BLAST ANI). Os pontos de corte atuais de delimitação de espécie para ANIb, ANIm e OrthoANI ficam no limiar entre 95% e 96% (Richter e Rossello-Mora 2009; Kim et al. 2014; Lee et al. 2016), enquanto que para gANI o limiar mínimo é de 96,5% (Varghese et al. 2015), visto que essa metodologia utiliza apenas regiões codificantes, que são mais conservadas.

Outras métricas foram desenvolvidas sem ter relação direta com ANIb, como "Genome-to-Genome Distance Calculator" (GGDC) (Meier-Kolthoff et al. 2013), "MUM index" (MUMi) (Deloger et al. 2009) e o "Tetra-nucleotide signature correlation index" (TETRA) (Teeling et al. 2004). O GGDC — substituto digital direto do DDH— calcula a distância intergenômica entre dois organismos para determinar a relação entre as espécies utilizando a metodologia baseada no programa "Genome Blast Distance Phylogeny" (GBDP) (Henz et al. 2005). Com o resultado se calcula um escore que é convertido no valor esperado, caso fosse realizado em bancada o DDH entre as duas espécies. Este valor final seria o DDH digital (dDDH) entre ambos os organismos, com o ponto de corte para espécie igual ao DDH, de 70% (Meier-Kolthoff et al. 2013). O escore calculado pelo GGDC determina a distância entre os dois genomas, que varia entre 0, para mais similar, e 1, para mais distante, (Meier-Kolthoff et al. 2013). Esse escore também pode ser usado para delimitação de espécie e apresenta um ponto de corte para valores iguais ou inferiores a 0,044 ou 0,045 (Auch et al. 2010).

O MUMi é um índice que calcula a distância entre dois genomas baseada nas "maximal unique exact matches" (correspondências exatas únicas máximas — MUM), resultado do algoritmo MUMmer (Kurtz et al. 2004). Os MUMs são fragmentos de sequência idênticos e únicos entre dois genomas (Kurtz et al. 2004). Para calcular o MUMi, todos os fragmentos que apresentarem um tamanho mínimo de nucleotídeos — que pode ser escolhido empiricamente, mas apresentou um resultado ideal com tamanho mínimo de 19 nucleotídeos (Deloger et al. 2009) — tem seu comprimento somado e posteriormente dividido pela média de comprimento dos dois genomas que estão sendo comparados (Deloger et al. 2009). O valor resultante é subtraído de 1, gerando o resultado final do MUMi, como apresentado na fórmula: $\mathrm{MUMi} = 1 - \frac{\mathrm{Cmum}}{\mathrm{Cmed}}$ (Deloger et al. 2009). O

resultado pode variar de 0, para mais similar, a 1, para mais distante, sendo 0,33 o ponto de corte para delimitação de espécie (Deloger et al. 2009).

O TETRA é um modelo de delimitação de espécie que, diferente das outras métricas, não é baseado em alinhamento de sequencias (Teeling et al. 2004). Para o cálculo do TETRA a sequência de DNA dos genomas bacterianos é estendida pelo seu complemento reverso — com o objetivo de compensar padrões discrepantes entre a fita "Watson" e a fita "Crick" — e as frequências das 256 possíveis combinações de tetranucleotídeos são calculadas (Teeling et al. 2004). Posteriormente, é feito um gráfico de dispersão com as frequências de tetranucleotídeos de ambos os genomas e um coeficiente de correlação de Pearson é calculado a partir do gráfico (Teeling et al. 2004). O valor da correlação determina se ambos os genomas podem pertencer à mesma espécie, onde o limite para delimitação de espécie varia entre $r \geq 0,989$ e $r \geq 0,999$ (Teeling et al. 2004) — o valor menor indica forte possibilidade de ambos organismos pertencerem a mesma espécie, enquanto o valor maior é confirmatório (Richter et al. 2015).

Apesar do uso do ANIb já estar sendo sedimentando na literatura como o melhor substituto do DDH para delimitação de espécie, para análises de grande escala, com centenas ou milhares de genomas para serem comparados entre si, essa métrica se torna computacionalmente inviável (Yoon et al. 2017). As demais métricas apresentadas acima são boas candidatas para servir como alternativa para delimitação de espécie bacteriana quando o conjunto de dados é grande demais para viabilizar o uso do ANIb, pois apresentam boa correlação tanto com os resultados de DDH quanto com os resultados de ANIb (Teeling et al. 2004; Deloger et al. 2009; Richter e Rossello-Mora 2009; Meier-Kolthoff et al. 2013; Lee et al. 2016). Portanto, uma avaliação comparativa ("benchmarking") dessas métricas é recomendada para determinar qual a melhor alternativa, baseado no tamanho do conjunto de dados e qualidade dos resultados.

2. OBJETIVOS

2.1. Objetivo Principal

Comparar diferentes métricas genômicas de delimitação de espécie bacteriana, com o propósito de elencar a metodologia mais recomendada baseado em sua especificidade, sensibilidade, robustez e poder computacional disponível.

2.2. Objetivos Específicos

- Comparar todas as métricas utilizando os genomas de *Paenibacillus* presentes no banco refseq disponível no sítio do NCBI;
- Aplicar os resultados das métricas genômicas na resolução de problemas de classificação de espécies dentro do gênero *Paenibacillus*.

5. DISCUSSÃO

Ao contrário dos eucariotos, a delimitação de espécie em procariotos não é regida por um conceito teórico, tende a ser mais pragmática, embora mais arbitrária e antropocêntrica (Gevers et al. 2005). Ainda que esteja estabelecido que um valor de DDH acima de 70% seja o principal critério para determinar que duas bactérias pertençam a mesma espécie (Moore et al. 1987), ainda não há uma definição oficial do que uma espécie bacteriana representa, apenas regras de nomenclatura (Parte 2018). Considerando que o DNA contém a informação evolutiva primordial, a introdução de métricas de comparação genômica poderá ser um passo importante para o estabelecimento de uma regra geral para definição de espécie bacteriana (Henz et al. 2005; Richter e Rossello-Mora 2009; Meier-Kolthoff et al. 2013; Kim et al. 2014; Varghese et al. 2015). Apesar disso, embora cada vez mais surjam novas métricas, não há um consenso na comunidade taxonômica a respeito sobre quais devem ser adotadas além do ANIb em esquemas de classificação. De fato, somente muito recentemente essas métricas estão sendo utilizadas como critérios taxonômicos. Por isso, os esforços para se comparar as diversas métricas que foram desenvolvidas desde 2005 (Konstantinidis e Tiedje 2005) ainda são praticamente nulos. Algumas métricas, ainda, foram testadas apenas em seus artigos de origem, sem estudos subsequentes validando seus resultados. Para piorar, muitas ferramentas possuem uma documentação muito limitada, o que dificulta seu uso por não-especialistas. Portanto, são necessários estudos que objetivem padronizar essas métricas, visando identificar quais são as mais apropriadas para cada tipo de dado e para cada condição de análise. Isso reduziria a produção de resultados redundantes ou o uso de métricas não compatíveis com determinados estudos.

Métricas dependentes do BLAST, como ANIb, OrthoANI e GGD, são muito mais lentas do que as que utilizam MUMmmer, como ANIm e MUMi, e do que as que verificam composição de sequência, como a análise de tetranucleotídeos (Richter e Rossello-Mora 2009; Yoon et al. 2017). Entretanto, a velocidade de processamento tem um custo, visto que os métodos que não se baseiam em BLAST tendem a ser menos precisos.

A reclassificação do *Paenibacillus durus* e do *Paenibacillus azotofixans* foi um trabalho paralelo derivado dos resultados gerados a partir da avaliação das métricas genômicas e ilustra a praticidade do uso desses dados quantitativos na resolução de problemas de classificação bacteriana que perduram há anos. De forma geral, os resultados

são congruentes, mas ao mesmo tempo mostram a falta de padronização das regras na avaliação de métricas genômicas na classificação bacteriana. Por isso, os resultados encontrados no estudo de *benchmarking* permitem auxiliar na escolha da metodologia mais adequada e mais eficiente, baseando-se nos tipos de dados disponíveis e na disponibilidade de poder computacional do pesquisador.

No presente estudo foi demonstrado que existe muita redundância nos resultados entre as diferentes métricas genômicas, embora o tempo de computação varie. Além disso, uma aplicação prática dessas métricas foi demonstrada na reclassificação das espécies *P. durus* e *P. azotofixans*.

É uma questão de tempo para que métricas genômicas sejam compulsórias, embora sua utilização seja ainda incipiente por parte da comunidade científica. O presente estudo gerou resultados que poderão ser futuramente utilizados na determinação de um esquema de classificação padronizado e eficiente na taxonomia procariótica.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Ambrosini A, Sant'Anna FH, Heinzmann J, de Carvalho Fernandes G, Bach E e Passaglia LMP (2018) Paenibacillus helianthi sp. nov., a nitrogen fixing species isolated from the rhizosphere of Helianthus annuus L. Antonie Van Leeuwenhoek 111:2463–2471. doi: 10.1007/s10482-018-1135-4

Ash C, Priest FG e Collins MD (1993) Molecular identification of rRNA group 3 bacilli (Ash, Farrow, Wallbanks and Collins) using a PCR probe test - Proposal for the creation of a new genus Paenibacillus. Antonie Van Leeuwenhoek 64:253–260. doi: 10.1007/BF00873085

Auch AF, von Jan M, Klenk HP e Göker M (2010) Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci. doi: 10.4056/sigs.531120

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature. doi: 10.1038/nature07517

Brenner DJ, Fanning GR, Rake A V. e Johnson KE (1969) Batch procedure for thermal elution of DNA from hydroxyapatite. Anal Biochem. doi: 10.1016/0003-2697(69)90199-7

Brenner DJ, Staley JT e Krieg NR (2005) Classification of Procaryotic Organisms and the Concept of Bacterial Speciation. Bergey's Manual® Syst Bacteriol. doi: 10.1007/0-387-28021-9_4

Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF e Kjelleberg S (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol. doi: 10.1128/AEM.01177-06

Chatton E (1925) Pansporella perplexa: amœbien à spores protégées parasite des daphnies : réflexions sur la biologie et la phylogénie des protozoaires. Masson et cie

Christensen H, Kuhnert P, Olsen JE e Bisgaard M (2004) Comparative phylogenies of the housekeeping genes atpD, infB and rpoB and the 16S rRNA gene within the Pasteurellaceae. Int J Syst Evol Microbiol. doi: 10.1099/ijs.0.03018-0

Copeland HF (1938) The Kingdoms of Organisms. Q Rev Biol 13:383–420. doi: 10.1086/394568

De Vos P, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA, Karl-Heinz S e Whitman WB (2009) Bergey's Manual of Systematic Bacteriology - Vol 3: The

Firmicutes. Springer-Verlag New York Inc. doi: 10.1007/978-0-387-68489-5
Deloger M, El Karoui M e Petit MA (2009) A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol 91:91–99. doi: 10.1128/JB.01202-08
Donkor ES (2013) Sequencing of bacterial genomes: Principles and insights into pathogenesis and development of antibiotics. Genes (Basel). doi: 10.3390/genes4040556
Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics. doi: 10.1093/bioinformatics/btq461
EDWARDS PR e EWING WH (1962) Identification of Enterobacteriaceae. Minneapolis 15: Burgess Publishing Co., 426, South Street, Minn., U.S.A.
Euzéby JP (1997) List of Bacterial Names with Standing in Nomenclature. In: Int. Syst. Bacteriol.
Ewing WH, Wilfert JN, Kunz LJ, Dumoff M e Isenberg HD (1969) Roundtable: How Far to Go with Enterobacteriaceae? J Infect Dis 119:197–213.
Fernandes G de C, Trarbach LJ, De Campos SB, Beneduzi A e Passaglia LMP (2014) Alternative nitrogenase and pseudogenes: Unique features of the Paenibacillus riograndensis nitrogen fixation system. Res Microbiol. doi: 10.1016/j.resmic.2014.06.002
Finlay BJ, Maberly SC e Cooper JI (1997) Microbial Diversity and Ecosystem Function. Oikos. doi: 10.2307/3546587
Fox GE, Wisotzkey JD e Jurtshuk P (1992) How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. Int J Syst Bacteriol. doi: 10.1099/00207713-42-1-166
G. E. Murray R, J. Brenner D, Colwell R, De Vos P, Goodfellow M, Grimont P, Pfennig N, Stackebrandt E e A. Zavarzin G (1990) Report of the Ad Hoc Committee on Approaches to Taxonomy within the Proteobacteria. Int J Syst Bacteriol. doi: 10.1099/00207713-40-2-213
Garrity GM, Bell JA e Lilburn TG (2004) TAXONOMIC OUTLINE OF THE PROKARYOTES BERGEY'S MANUAL ® OF SYSTEMATIC BACTERIOLOGY. Bergey's Man Trust. doi: 10.1007/bergeysoutline200405
Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL et al. (2005) Re-evaluating prokaryotic species. Nat Rev Microbiol. doi: 10.1038/nrmicro1236

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P e Tiedje JM (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91. doi: 10.1099/ijs.0.64483-0
Haeckel Ernst (1866) Generelle morphologie der organismen. Allgemeine grundzüge der organischen formen-wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte descendenztheorie,. Berlin,G. Reimer,
Henz SR, Huson DH, Auch AF, Nieselt-Struwe K e Schuster SC (2005) Whole-genome prokaryotic phylogeny. Bioinformatics. doi: 10.1093/bioinformatics/bth324
Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K et al. (2016) A new view of the tree of life. Nat Microbiol. doi: 10.1038/nmicrobiol.2016.48
Jorgenson JW e Lukacs KD (1981) Free-zone electrophoresis in glass capillaries. Clin. Chem.
Kim M, Oh HS, Park SC e Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol. doi: 10.1099/ijs.0.059774-0
Konstantinidis KT e Tiedje JM (2005) Towards a genome-based taxonomy for prokaryotes. J Bacteriol 187:6258–6264. doi: 10.1128/JB.187.18.6258-6264.2005
Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C e Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol. doi: 10.1186/gb-2004-5-2-r12
Lee I, Kim YO, Park SC e Chun J (2016) OrthoANI: An improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol 66:1100–1103. doi: 10.1099/ijsem.0.000760
Lunteren F van (2016) Clocks to Computers: A Machine-Based “Big Picture” of the History of Modern Science. Isis. doi: 10.1086/689764
Mahato NK, Gupta V, Singh P, Kumari R, Verma H, Tripathi C, Rani P, Sharma A, Singhvi N, Sood U et al. (2017) Microbial taxonomy in the era of OMICS: application of DNA sequences, computational tools and techniques. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol. doi: 10.1007/s10482-017-0928-1
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. (2005) Genome sequencing in microfabricated

high-density picolitre reactors. Nature. doi: 10.1038/nature03959
McAuliffe O, Kilcawley K e Stefanovic E (2018) Symposium review: Genomic investigations of flavor formation by dairy microbiota. J Dairy Sci. doi: 10.3168/jds.2018-15385
Meier-Kolthoff JP, Auch AF, Klenk HP e Göker M (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. doi: 10.1186/1471-2105-14-60
Moore WEC, Stackebrandt E, Kandler O, Colwell RR, Krichevsky MI, Truper HG, Murray RGE, Wayne LG, Grimont PAD, Brenner DJ et al. (1987) Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. Int J Syst Evol Microbiol. doi: 10.1099/00207713-37-4-463
Morozova O e Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics. doi: 10.1016/j.ygeno.2008.07.001
Mullis K, Faloona F, Scharf S, Saiki R, Horn G e Erlich H (1986) Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. Cold Spring Harb Symp Quant Biol. doi: 10.1101/SQB.1986.051.01.032
Parker V (1965) Antony van Leeuwenhoek. Bull Med Libr Assoc. doi: 10.1213/XAA.0000000000000421
Parte AC (2018) LPSN - List of prokaryotic names with standing in nomenclature (Bacterio.net), 20 years on. Int J Syst Evol Microbiol. doi: 10.1099/ijsem.0.002786
Pennisi E (1998) Genome data shake tree of life. Science (80-). doi: 10.1126/science.280.5364.672
Perkins RG (2008) Bergey's Manual of Determinative Bacteriology. Am J Public Heal Nations Heal. doi: 10.2105/ajph.20.5.565-a
Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA e Baumeister K (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science (80-). doi: 10.1126/science.2443975
Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature. doi: 10.1038/nature08821
Reuter JA, Spacek D V. e Snyder MP (2015) High-Throughput Sequencing Technologies. Mol Cell. doi: 10.1016/j.molcel.2015.05.004

Richter M e Rossello-Mora R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci 106:19126–19131. doi: 10.1073/pnas.0906412106
Richter M, Rosselló-Móra R, Oliver Glöckner F e Peplies J (2015) JSpeciesWS: A web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics. doi: 10.1093/bioinformatics/btv681
Rosselló-Móra R (2012) Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environ Microbiol. doi: 10.1111/j.1462-2920.2011.02599.x
Sanger F e Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. doi: 10.1016/0022-2836(75)90213-2
Sant'Anna FH, Ambrosini A, de Souza R, de Carvalho Fernandes G, Bach E, Balsanelli E, Baura V, Brito LF, Wendisch VF, de Oliveira Pedrosa F et al. (2017) Reclassification of Paenibacillus riograndensis as a genomovar of Paenibacillus sonchi: Genome-based metrics improve bacterial taxonomic classification. Front Microbiol. doi: 10.3389/fmicb.2017.01849
Sant'Anna FH, Ambrosini A, Guella FL, Porto RZ e Passaglia LMP (2018) Genome-based reclassification of Paenibacillus dauci as a later heterotypic synonym of Paenibacillus shenyangensis. Int. J. Syst. Evol. Microbiol.
Sato M e Miyazaki K (2017) Phylogenetic network analysis revealed the occurrence of horizontal gene transfer of 16S rRNA in the genus Enterobacter. Front Microbiol. doi: 10.3389/fmicb.2017.02225
Schouls LM, Schot CS e Jacobs JA (2003) Horizontal Transfer of Segments of the 16S rRNA Genes between Species of the Streptococcus anginosus Group. J Bacteriol. doi: 10.1128/JB.185.24.7241-7246.2003
Stackebrandt E (2011) Taxonomic parameters revisited: tarnished gold standards. Microbiol Today. doi: 10.1007/978-3-642-30782-9_3
Stackebrandt E, Frederiksen W, Garrity GM, Grimont PADD, Kämpfer P, Maiden MCJJ, Nesme X, Rosselló-Mora R, Swings J, Trüper HG et al. (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol 52:1043–1047. doi: 10.1099/ijs.0.02360-0
STACKEBRANDT E e GOEBEL BM (1994) Taxonomic Note: A Place for DNA-DNA

Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int J Syst Evol Microbiol. doi: 10.1016/S0140-6736(01)43317-4
Stanier RY e Van Niel CB (1941) The Main Outlines of Bacterial Classification. J. Bacteriol.
Syvanen M (1994) Horizontal Gene Transfer: Evidence and Possible Consequences. Annu Rev Genet. doi: 10.1146/annurev.ge.28.120194.001321
Teeling H, Meyerdierks A, Bauer M, Amann R e Glöckner FO (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol 6:938–947. doi: 10.1111/j.1462-2920.2004.00624.x
Tian RM, Cai L, Zhang WP, Cao HL e Qian PY (2015) Rare events of intragenusand intraspecies horizontal transfer of the 16S rRNA gene. Genome Biol Evol. doi: 10.1093/gbe/evv143
Vandamme P, Pot B e Gillis M (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Mol Biol Rev. doi: 10.1007/s12088-007-0022-x
Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC e Pati A (2015) Microbial species delineation using whole genome sequences. Nucleic Acids Res 43:6761–6771. doi: 10.1093/nar/gkv657
Whitman WB, Coleman DC e Wiebe WJ (1998) Prokaryotes: The unseen majority. Proc Natl Acad Sci. doi: 10.1073/pnas.95.12.6578
Whittaker RH (1969) New concepts of kingdoms of organisms. Science (80-). doi: 10.1126/science.163.3863.150
WINSLOW C-EA (1914) THE CHARACTERIZATION AND CLASSIFICATION OF BACTERIAL TYPES. Science (80-) 39:77–91. doi: 10.1126/science.39.994.77
Woese CR, Kandler O e Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci 87:4576–4579. doi: 10.1073/pnas.87.12.4576
Yoon SH, Ha S min, Lim J, Kwon S e Chun J (2017) w. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol. doi: 10.1007/s10482-017-0844-4
Zeigler D (2016) The Family Paenibacillaceae. doi: 10.13140/RG.2.1.1949.5289
(1994) Validation of the Publication of New Names and New Combinations Previously Effectively Published Outside the IJSB: List No. 51†. Int J Syst Evol Microbiol 44:852.
(1980) Approved lists of bacterial names. Med J Aust. doi: 10.1099/00207713-30-1-225