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.

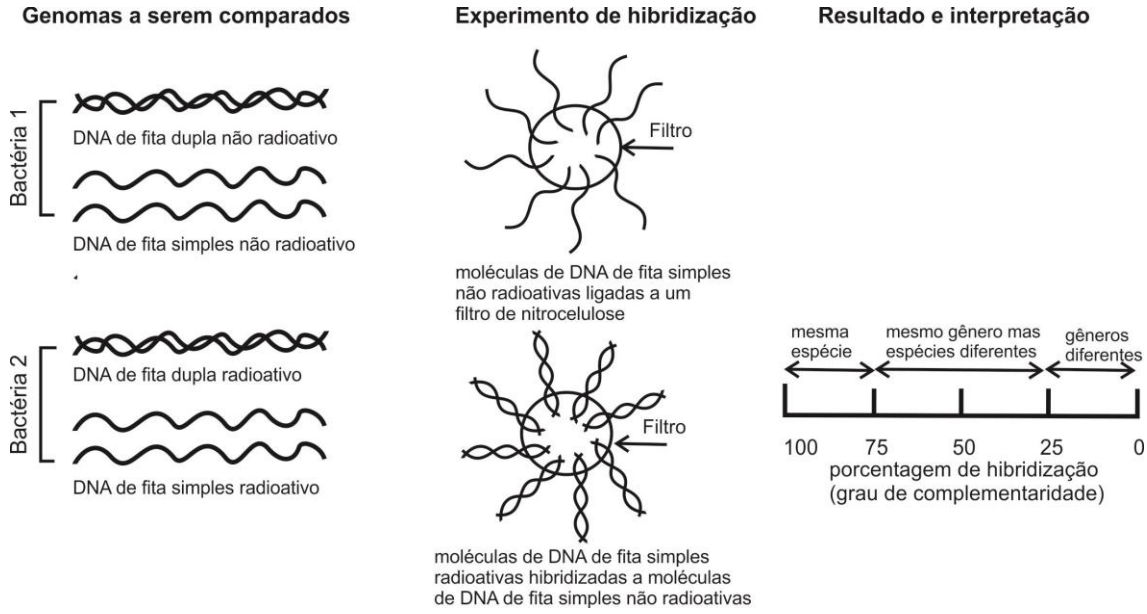


Figura 1: Exemplo de um experimento de hibridização entre DNAs de duas bactérias (adaptado de http://www.biologydiscussion.com/bacteria/bacterial-taxonomy/bacterial-taxonomy-meaning-importance-and-levels/54679, último acesso em: 10/02/2019)

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).

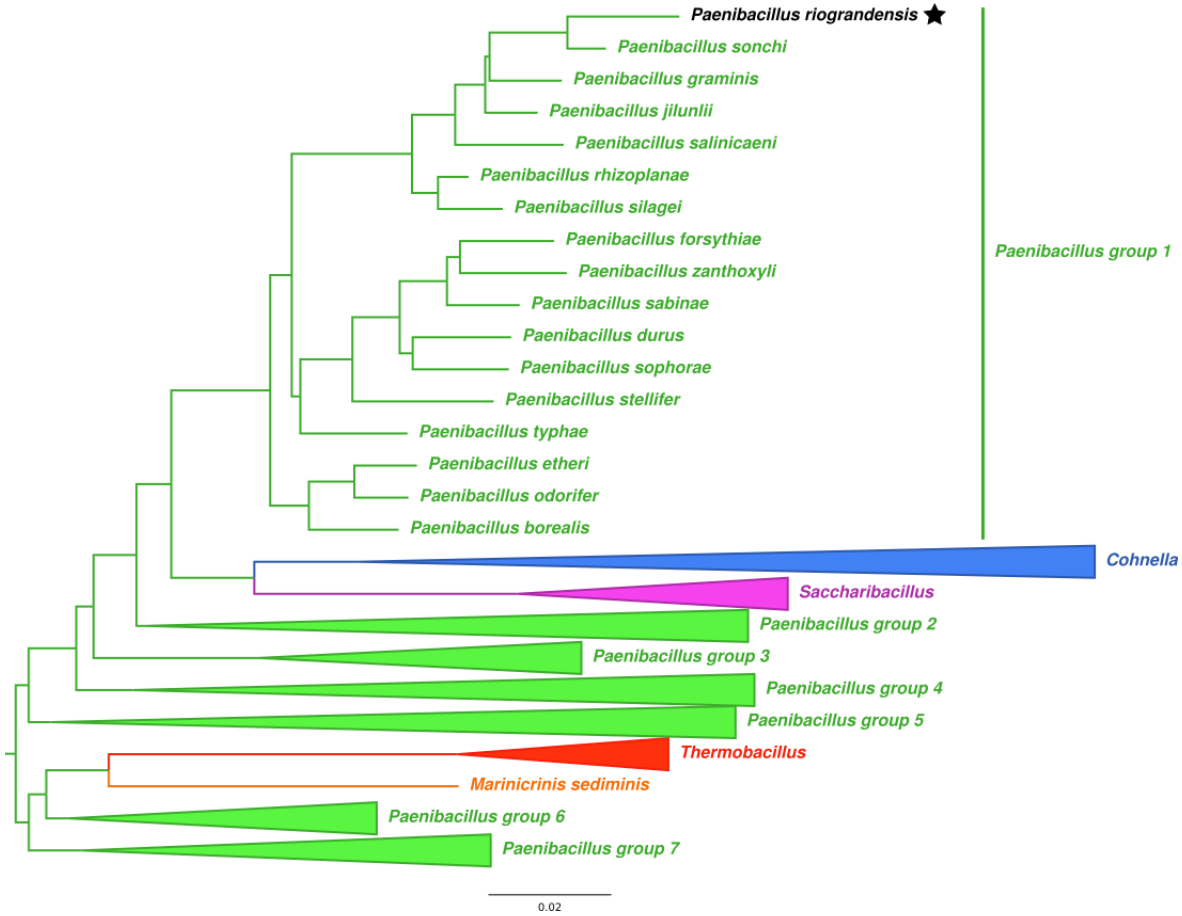


Figura 3: Reconstrução filogenética do gênero *Paenibacillus* e de gêneros próximos da mesma família.

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).

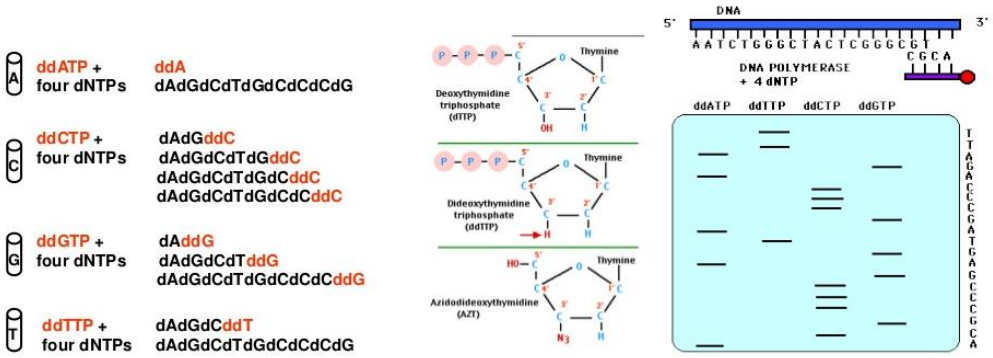


Figura 4: Exemplo do processo de sequenciamento pelo método de Sanger (recuperado de https://www.slideshare.net/SurenderRawat3/dna-sequencing-41318444/9 - último acesso 28/02/2019).

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*.

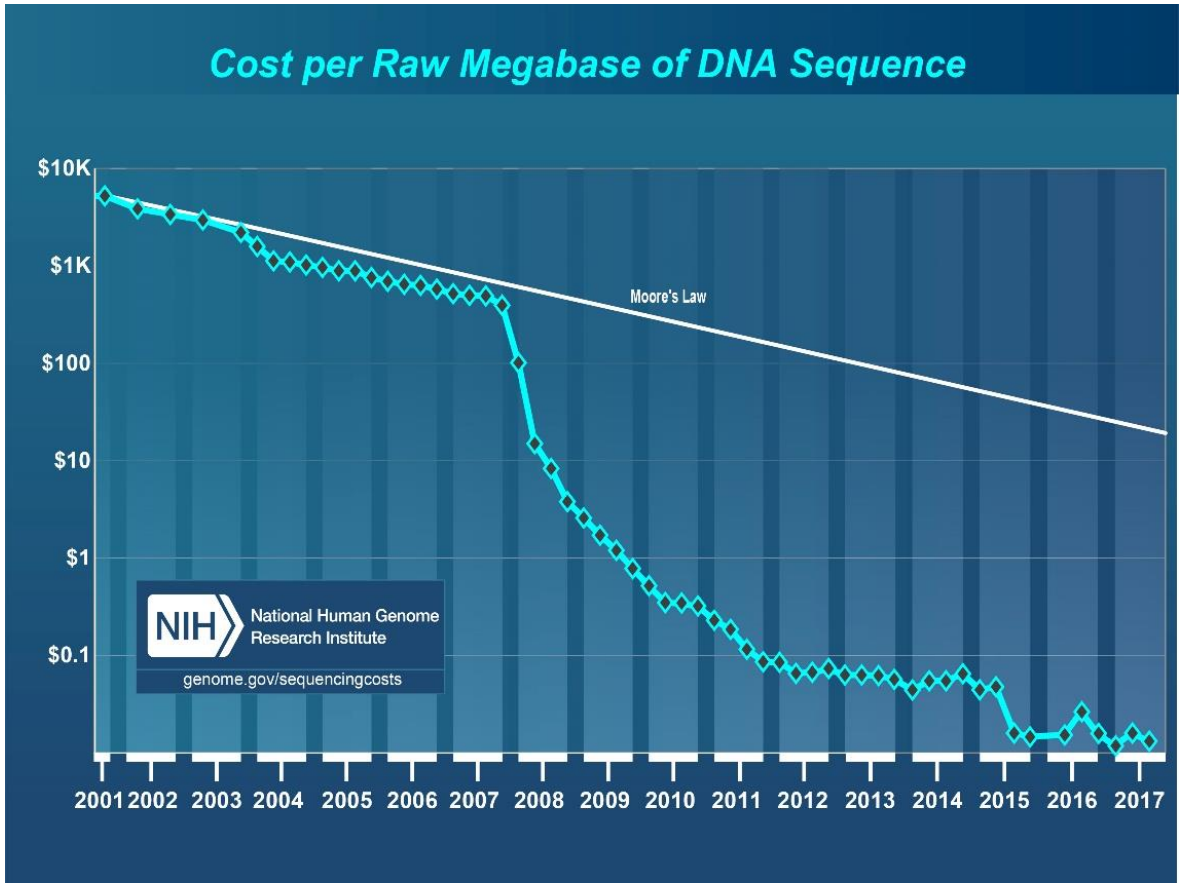


Figura 5: Variação do custo, em dólares, do sequenciamento de DNA por megabase entre os anos de 2001 e 2017. (retirado de https://www.genome.gov/27541954/dna-sequencing-costs-data/ — último acesso em 28/02/2019).

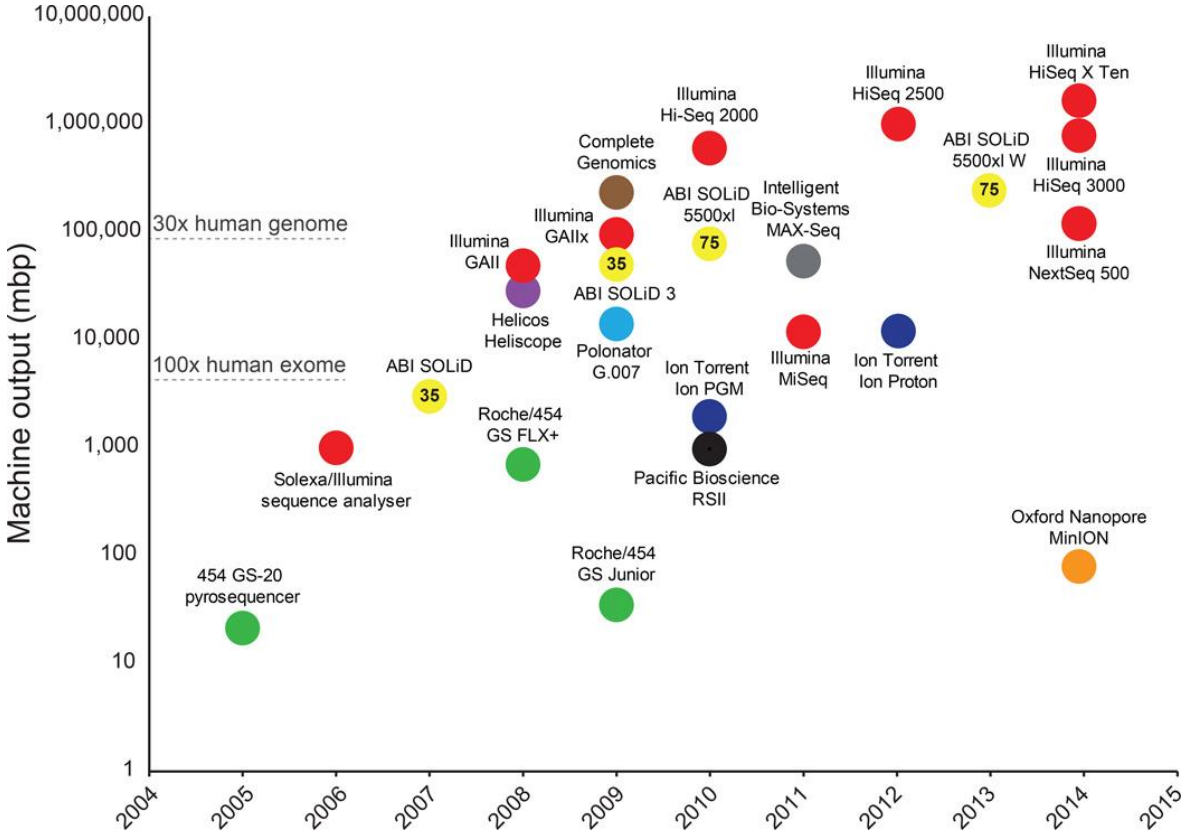


Figura 6: Gráfico de ano de lançamento comercial versus rendimento por execução de instrumentos de sequenciamento. [Retirado de (Reuter et al. 2015)].

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*.

3. **Capitulo 1**

Artigo 1

Título: Benchmark of algorithms for the computation of genome relatedness metrics

Em preparação

Revista para submissão: *Genome Biology*

Benchmark of algorithms for the computation of genome relatedness metrics

Felipe Guella[1], Fernando Hayashi Sant'Anna[1], Luciane Maria Pereira Passaglia[1#]

1 - Department of Genetics, Universidade Federal do Rio Grande do Sul, 9500, Bento Gonçalves Ave., Porto Alegre, 91501-970, Rio Grande do Sul, Brazil.

#Corresponding author

Email addresses:
FG - felipe.guella@gmail.com
FHS – fhsantanna@yahoo.com.br
LMPP - luciane.passaglia@ufrgs.br

Abstract

Background: Whole genome comparison (WGC) methodologies are becoming valuable tools for prokaryotic identification and classification. Although several metrics were created, as average nucleotide identity (ANI), maximum unique matches index (MUMi), tetranucleotide usage pattern (TETRA), and genome-to-genome distance (GGD), there is no present study comparing all these metrics. The study aims to benchmark all these metrics, comparing robustness, runtime, correlation, sensitivity and specificity, using ANI results as reference and *Paenibacillus* genomes as dataset.

Results: TETRA metric computation had the fastest runtime, followed by MUMi and ANIm computations. GGD had the longest processing time. ANIb, gANI and OrthoANI computations had similar runtime and they were only faster than GGD. All metrics presented correlation coefficients with ANI above 0.9, except TETRA (≈0.75). Specificity was high for all metrics (>0.9) and sensitivity was above 0.9 for ANIm, OrthoANI and TETRA, and between 0.7 and 0.8 for gANI, GGD and MUMi. Robustness tests had small variations of results on most alignment-based metrics, except for MUMi and TETRA.

Conclusions: Considering all parameters and conditions tested, ANIm was the best metric for genome similarity evaluation, since it was fast and had high sensitivity and sensibility. Other metrics based on ANI little differed among each other in terms of performance. Although MUMi and TETRA analyses are the fastest, they are not robust as ANI-based methods. On presenting the strengths and weaknesses of these genome metrics, this study might contribute for the standardization of using genome metrics for prokaryotic systematics.

Key words: ANI; dDDH; MUMi; bacterial classification; *Paenibacillus*; benchmark; tetranucleotide.

Background

In the late 1960s, the first genomic method for bacterial classification was developed, using DNA-DNA hybridization (DDH) (Brenner et al. 1969). Researchers found that different bacteria with phenotypical coherence tended to present DNA-DNA hybridization percentage higher than 70%. Therefore, this criterion was utilized for classifying bacterial species (Moore et al. 1987). Along with DDH, the phylogenetic reconstruction of the 16S rRNA gene — first used in 1980 to compare mycoplasma and other bacteria (Woese et al. 1980) — revolutionized bacterial systematics (Oren and Garrity 2014). In fact, these two methodologies became the pillars of bacterial classification (G. E. Murray et al. 1990; Oren and Garrity 2014). In spite of DDH being pivotal for the improvement of bacterial systematics, it has its limitations, having low reproducibility and being unable to provide cumulative data (Rosselló-Móra 2012), in addition to the necessity of depending on biological material (DNA) availability. Some researchers have been discussing that DDH should be replaced by more objective and reproducible methods (Stackebrandt et al. 2002; Oren and Garrity 2014).

In 1995, the *Haemophilus influenzae* was the first bacterium to have its genome sequenced (Fleischmann et al. 1995), and, since then, with the development of fast and affordable sequencing methods, more than 100.000 bacterial genomes were deposited in the GenBank database with over 15,000 described species (Parte 2018). Nevertheless, the conventional methods of classification, like DDH and 16S rRNA phylogeny, eventually are not enough to correctly determine species boundaries. The *Paenibacillus* genus, for example, despite being relatively new - it was created 25 years ago (Ash et al. 1993; 1994), has already presented some taxonomic issues: the Paenibacillaceae family phylogram has shown that *Paenibacillus* is paraphyletic with other genera from the family (Mayilraj and Stackebrandt 2013); furthermore, the 16S rRNA is also highly conserved among its species (Sant'Anna et al. 2017) and the genome size of its species range from 3 to 8.8 Mbp. Therefore, the use of 16S rRNA phylogeny and DDH results initially used for its species circumscription produced some misclassifications that needed to be corrected at a later date (Kim et al. 2011; Sant'Anna et al. 2017; Sant'Anna et al. 2018).

Moreover, the high availability of sequence data allied with increased computational power, allowed the designing of new genome wide methodologies for bacterial comparison (Richter and Rossello-Mora 2009; Oren and Garrity 2014). The first of those genomic

methodologies was the average nucleotide identity (ANI) computation (Konstantinidis and Tiedje 2005a). ANI uses the Basic Local Alignment Search Tool (BLAST) algorithm to emulate *in silico* DDH evaluation (Konstantinidis and Tiedje 2005a). Several derivations of the BLAST ANI (ANIb) were later developed: the MUMmer ANI (ANIm) (Richter and Rossello-Mora 2009), that uses the faster algorithm MUMmer (Kurtz et al. 2004), instead of BLAST; the orthologous ANI (OrthoANI), in which both bacterial genomes are fragmented and only the reciprocal best hits are used on the calculation (Lee et al. 2016); and the gene wide ANI (gANI), that uses only the coding region on its analysis (Varghese et al. 2015). Those metrics showed high correlation with DDH results, with identity scores of 95% or higher — 96.5% for gANI, whereas ANIb, ANIm and OrthoANI being equivalent to a DDH result of 70% or higher (Richter and Rossello-Mora 2009; Varghese et al. 2015; Lee et al. 2016).

Besides ANIb and its variations, other algorithms were designed based on whole genome comparison, such as maximum unique matches (MUMs) index (MUMi) (Deloger et al. 2009), Genome-to-Genome Distance Calculator (GGDC) (Meier-Kolthoff et al. 2013) and tetranucleotide usage pattern (TETRA) computations (Teeling et al. 2004a). MUMi calculates a distance index based on MUMs — varying between 0, to most closely related, and 1, to most distant related — resulted from the MUMmer algorithm — MUMs are identical and unique fragments of DNA sequences shared between two genomes (Kurtz et al. 2004). GGDC was developed as a direct substitute of DDH and is usually called digital DDH (dDDH) (Meier-Kolthoff et al. 2013). GGDC estimates a wet lab DDH result based on the Genome BLAST distance phylogeny (GBDP) method (Henz et al. 2005). The GGD final result is converted into a DDH estimation (Auch et al. 2006; Meier-Kolthoff et al. 2013). The GGD results can also be used as a metric for species circumscription (Auch et al. 2006). TETRA is different from the other genomic metrics, since it is not based on alignment (Teeling et al. 2004a). TETRA is a correlation index between the frequencies of all possible combinations of tetranucleotides of two genomes (Teeling et al. 2004b; Richter and Rossello-Mora 2009). The species threshold for MUMi, GGD and TETRA are ≤ 0.33, ≤ 0.045 and ≥ 0.989, respectively (Deloger et al. 2009; Richter and Rossello-Mora 2009; Meier-Kolthoff et al. 2013). With all these methodologies available, the use of whole genome comparison for bacterial classification has grown steadily in the last decade, especially ANIb. Genomic metrics are not only being used for novel taxa studies, but are

also being used to identify misclassifications (Kim et al. 2011; Sant'Anna et al. 2017; Sant'Anna et al. 2018).

Currently, the use of genomic metrics is being more and more accepted as a strong tool for prokaryote systematics by the scientific community (Konstantinidis and Tiedje 2005a; Richter and Rossello-Mora 2009; Mahato et al. 2017), while DDH is being considered obsolete (Stackebrandt et al. 2002; Richter and Rossello-Mora 2009). With the increased use and reduced price of whole genome sequencing methods, the International Journal of Systematic and Evolutionary Microbiology (IJSEM), the main journal of new prokaryotic species publication, started to demand sequenced genome data on new taxa publications from 2018 onwards (International Union of Microbiological Societies. et al. 2017) — which will further increase the use of genomic metrics in species circumscription.

Despite the large amount of available methodologies, the most utilized and recommended genomic metric for species delimitation so far is ANIb (Konstantinidis and Tiedje 2005a; Goris et al. 2007; Richter and Rossello-Mora 2009). Nevertheless, for a large dataset, the use of ANIb is unfeasible (Richter and Rossello-Mora 2009; Yoon et al. 2017) and alternatives are made necessary. Some studies already compared the performance of ANIb, ANIm and OrthoANI analyses (Richter and Rossello-Mora 2009; Yoon et al. 2017). However, other metrics such as MUMi, GGD and TETRA were not yet contrasted with ANI. Additionally, there is no study assessing the robustness of those metrics when the assumption of genome quality is violated, since contaminated sequence might be included on sequenced genomes eventually (Mukherjee et al. 2015).

With all that in consideration, this study aims to benchmark genome metrics, in order to evaluate robustness and runtime — given an increasing amount of data — and to verify the correlation between results for each metrics. For this *Paenibacillus* genomes available at the Refseq database (O'Leary et al. 2016) were used as dataset. *Paenibacillus* was specifically chosen in order to further use the results of this research to later identify possible misclassifications inside the genus.

Results

Runtime of genomic metrics analyses

Considering the progressive increase in the dataset, the time spent for calculating genomic metrics varied (Fig. 1). ANIb, OrthoANI and gANI had similar results, whereas

GGD had a significant increase in runtime from the dataset of 50 genomes onwards. MUMi and ANIm had similar runtimes, while TETRA was the fastest, as it seems to increase its runtime linearly (Fig. 1). The runtime difference between methods that use BLAST and methods that use MUMmer became larger as more genomes were added to the dataset (Fig. 1).

Statistical analyses

Absolute values for correlation results are shown on Fig. 2. Correlation on most methods in relation to ANIb were significantly high — with an exception of TETRA results that has shown a correlation value below 0.75. Specificity values were also high for all metrics (Fig. 3a), while sensitivity values (Fig. 3b) were below 0.8 for gANI, GGD and MUMi, and below 0.5 for TETRA, which means that gANI, GGD and MUMi have a slightly high amount of false negative results.

Robustness results on contaminated genomes

The overall variation results for each genomic metric (Fig. 4) shows that MUMi and TETRA were the only methods that presented considerable variation when comparing contaminated genomes with uncontaminated ones. However, MUMi values did not varied as those from TETRA analysis (Fig. 4), in which the values were distributed above and below the species boundary. Apart from MUMi, all alignment-based metrics had minimum to no variation on robustness tests.

Discussion

With the increase in computational power, the enhancement of genomic sequencing technologies and, consequently, the reduced cost and runtime of whole genome sequencing, the use of WGC for species circumscription is being frequently considered as the future of gold standard for prokaryotic systematics (Konstantinidis and Tiedje 2005b; Richter and Rossello-Mora 2009; Kim et al. 2014). Concomitantly, personal computers are increasingly more powerful and, for most cases, researchers do not need to rely on the use of super computers to run WGC comparisons. Even small budget labs could perform WGC analyses within an acceptable runtime. Methods faster than but with similar performance to ANIb computation might be of great assistance to those laboratories.

However, only few works carried out a comprehensive comparison of those genomic metrics, and, apart from the original works for each method (Deloger et al. 2009; Meier-Kolthoff et al. 2013; Lee et al. 2016), they were usually related to the ANI-based methods

(Yoon et al. 2017). With the advent of high-throughput sequencing, prokaryotic systematics is growing fast, therefore, uniformity and standardization on the methodology used is essential for a healthy development. In this sense, benchmark studies would be essential in order to point out the advantages and disadvantages of each WGC metric, an essential step for indicating the paths that the scientific area should take.

Most algorithms on this study uses BLAST search to run their calculations, but as it was shown here — and it was already discussed on other occasions (Kurtz et al. 2004; Edgar 2010) — this is not by far the fastest search method available. Our findings suggested that runtime of genomic metrics is related to the alignment algorithm used for its calculations. MUMi and ANIm were faster than ANIb, gANI, GGDC and OrthoANI due to the use of MUMmer instead of BLAST, and TETRA was the fastest because it does not use alignment tools. Although the gANI algorithm utilizes NSimScan, which is 10 to 100X faster than Blastn (Novichkov et al. 2016), its computation runtime was slower than those from ANIb and OrthoANI, due to its algorithm not allowing multi-threaded execution.

Sensitivity and specificity results shows that gANI, GGD and MUMi might not be reliable to determine if two genomes belong to the same species, when taking ANIb results as reference. Robustness tests findings suggest that for *de novo* sequenced genomes, with possible sample contaminations, any metrics, apart from TETRA, might be suitable for genomic comparison, while MUMi should be used with caution. Due to the way MUMi calculates distance — based on the average size of each genome —, when the size of one of the genomes is increased with the contaminant sequence, the average size of both genomes increases, which influences the final result. In spite of that, all alignment-based methods were less susceptible to variation when presented with contaminated data. TETRA, on the other hand, since it is based on genome composition and not in alignment, has shown extreme variations on its results following contamination, and almost all results were below the species threshold.

This study is the first one that attempts to do a thorough analysis on runtime, correlation, specificity, sensitivity and robustness of several WGC metrics. Although here we managed to give an overall insight on some characteristics — showing some pros and cons — of each methodology there were some limitations that should be taken in consideration. Firstly, the use of only a single bacterial genus to run all tests might not give a result that correctly represents the whole prokaryotic domain, especially on the

contamination analysis; however, this should not be an issue on the time consumption, considering that the *Paenibacillus* genomes have a wide range of sizes (Grady et al. 2016).

Conclusion

The use of WGC is becoming common practice on prokaryotic systematics analysis. In spite of that, using ANIb — the current recommended WGC method — to compare large clusters of data can be challenging even for the most powerful hardware (Yoon et al. 2017). In this study, nonetheless, we have shown that ANIm was the best alternative metric for genome similarity evaluation. Other metrics based on ANI, such as gANI and OrthoANI, had similar performance in relation to ANIb. Although MUMi and TETRA analyses were the fastest, they are not robust as ANI-based methods.

This work was an effort to evaluate current WGC methodologies in parameters which might help researchers into choosing the most suitable methodology for bacterial identification/classification.

Methods

All analyses were conducted on a computer with an Intel® Octa core i7-3770 CPU with 32gb of ram memory with Ubuntu 14.04 LTS Linux.

Genome selection

All 341 public available *Paenibacillus* genomes were downloaded from the NCBI Refseq database (O'Leary et al. 2016), in January 2018. A BLAST search of four housekeeping genes — *gyrB, recA, recN* and *rpoB* — was conducted in order to assess annotation completeness. From the 341 genomes selected, 22 had one or more housekeeping gene missing and were excluded from the dataset. All the remaining 319 genomes (Supplementary Table S1) were used for the analysis on this work, with some tests using only part of the dataset.

Time consumption analysis

From the 319 dataset, 100 genomic sequences (available in Supplementary Table S1) were randomly chosen to calculate progressive runtime of all metrics. Time consumption analysis of all metrics using a progressive amount of data in an all-against-all comparison was performed. Runtime tests started with two genomes on the dataset and, for each subsequent run, were increased by two, up until ten. From 10 genomes onwards, the number of genomes incremented by ten for each subsequent run, until reaching the 100 genomes. Runtime was measured using the "time" command from the Ubuntu terminal.

The list of the 100 genomes is available at Supplementary Table S1 and the genomes are shown in the order they were used.

Genetic Relatedness Measurement

All 319 *Paenibacillus* genomes had their genetic relatedness measured by ANIb, gANI, OrthoANI, ANIm, MUMi, GGDC and TETRA. ANIb, ANIm and TETRA were calculated using pyani (https://github.com/widdowquinn/pyani), according to Goris et al. (Goris et al. 2007), for ANIb, and according to Richter and Rosséllo-Móra (Richter and Rossello-Mora 2009), for ANIm and TETRA. Cutoff points for species delimitation were determined as ≥0.95 for ANIb and ANIm values and to ≥ 0.989 for TETRA values based on literature (Goris et al. 2007; Richter and Rossello-Mora 2009). For gANI calculation ANIcalculator was used (https://ani.jgi-psf.org/html/anicalculator.php), according to Varghese et al. (Varghese et al. 2015). Cutoff point for species delimitation was determined as ≥0.965 gANI values based on literature (Varghese et al. 2015). GGDC and OrthoANI were calculated using the Orthologous Average Nucleotide Identity Tool (OAT) (https://www.ezbiocloud.net/tools/orthoani), according Meier-Kolthoff et al. and Lee et al. (Meier-Kolthoff et al. 2013; Lee et al. 2016), respectively. Delimitation points for species delimitation were determined as ≥0.95 for OrthoANI values and as ≤0.044 for GGDC values based on literature (Auch et al. 2006; Meier-Kolthoff et al. 2013; Lee et al. 2016). MUMi was calculated using MUMi Python API (https://github.com/mb1511/MUMmer-MUMi), according to Deloger et al. (Deloger et al. 2009). Cutoff point for species delimitation was determined to ≤ 0.33 MUMi values based on literature (Deloger et al. 2009). All scripts were executed using standard parameters, as shown on Supplementary Table S2.

Statistical analyses

All statistical analysis were performed using R language (Ihaka and Gentleman 1996) in R Studio integrated development environment (IDE) (RStudio Team 2015), along with all resulting graphs and tables. The R libraries used in this study were ggpolt2 (Ginestet 2011), extrafonts (https://github.com/wch/extrafont) and APE (Paradis et al. 2004). In order to measure metrics sensitivity and specificity ANIb results were used as reference.

Correlation coefficient, specificity and sensitivity of all metrics were calculated using ANIb results as reference. Results considered as true positive, true negative, false positive

and false negative are explained on Fig. 5. Absolute values of correlation coefficients were considered in order to compare distance and similarity metrics.

All correlations were calculated using Spearman's method, since data results were not normally distributed (Supplementary Fig. 1).

Robustness Measurement

To assess how contamination might affect genomic metrics results, eight genomes of four *Paenibacillus* species were selected: *Paenibacillus sonchi* X19-5^T, *Paenibacillus riograndensis* SBR5^T — *P. riograndensis* is a genomovar of *P. sonchi* (Sant'Anna et al. 2017) —, *Paenibacillus darwinianus* BR^T, *Paenibacillus darwinianus* CE1, *Paenibacillus odorifer* DSM 15391^T, *Paenibacillus odorifer* FSL J3-0155, *Paenibacillus macquariensis* ATCC 23464^T and *Paenibacillus macquariensis* JCM 14954. All genomes were artificially contaminated with different proportions (5%, 10%, 15%, 20% and 25%) of *Escherichia coli* DSM 30083^T genome sequence. Sequence size of each percentage was measured on the original genome and then an equivalent size sequence from *E. coli* was inserted on it. The DNA fragment from *E. coli* was taken from random regions of its whole genome for each contamination process, making it so that no resulting contaminated genome had the same contaminant sequence. Five contamination iterations were executed for each percentage, adding up to 25 contaminated genomes for each strain. Later, genomic relatedness was calculated for the species strain pairs, comparing uncontaminated genome from one strain with the several contaminations of the other strain and vice-versa.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

FG received a scolarship from CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), and FHS. received a scholarship from Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil).

Authors' contributions

FG, FHS, and LMPP conceived and designed the analyses. FG developed the scripts, performed the analyses, and generated the artwork and tables. FHS participated in methodological decisions. FG wrote the manuscript. All authors discussed the data and reviewed the manuscript.

Acknowledgments

Not applicable.

References

Ash C, Priest FG and 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, Henz SR, Holland BR and Göker M (2006) Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. BMC Bioinformatics. doi: 10.1186/1471-2105-7-350

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

Deloger M, El Karoui M and 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

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics. doi: 10.1093/bioinformatics/btq461

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (80-). doi: 10.1126/science.7542800

G. E. Murray R, J. Brenner D, Colwell R, De Vos P, Goodfellow M, Grimont P, Pfennig N, Stackebrandt E and 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

Ginestet C (2011) ggplot2: Elegant Graphics for Data Analysis. J R Stat Soc Ser A (Statistics Soc. doi: 10.1111/j.1467-985x.2010.00676_9.x

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P and 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

Grady EN, MacDonald J, Liu L, Richman A and Yuan Z-C (2016) Current knowledge and perspectives of Paenibacillus: a review. Microb Cell Fact 15:203. doi: 10.1186/s12934-016-0603-7

Henz SR, Huson DH, Auch AF, Nieselt-Struwe K and Schuster SC (2005) Whole-genome prokaryotic phylogeny. Bioinformatics. doi: 10.1093/bioinformatics/bth324

Ihaka R and Gentleman R (1996) R: A Language for Data Analysis and Graphics. J Comput Graph Stat. doi: 10.1080/10618600.1996.10474713

International Union of Microbiological Societies., Society for General Microbiology., International Union of Microbiological Societies. Bacteria and Applied Microbiology Division., International Association of Microbiological Societies. International Committee on Bacteriological Nomenclature. Judicial Commission., International Association of Microbiological Societies. International Committee on Bacteriological Nomenclature., International Association of Microbiological Societies. International Committee on Nomenclature of Bacteria. Judicial Commission., International Association of Microbiological Societies. International Committee on Nomenclature of Bacteria., International Association of Microbiological Societies. International Committee on Systematic Bacteriology. and International Union of Microbiological Societies. International Committee on Systematic Bacteriology. (2017) News : Genome sequencing data required with Taxonomic Descriptions. https://ijs.microbiologyresearch.org/content/Genome_data_required_IJSEM. Accessed 20

Mar 2019

Kim KK, Lee KC and Lee JS (2011) Reclassification of Paenibacillus ginsengisoli as a later heterotypic synonym of Paenibacillus anaericanus. Int J Syst Evol Microbiol. doi: 10.1099/ijs.0.025650-0

Kim M, Oh HS, Park SC and 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 and Tiedje JM (2005a) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci 102:2567–2572. doi: 10.1073/pnas.0409727102

Konstantinidis KT and Tiedje JM (2005b) 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 and 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 and 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

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

Mayilraj S and Stackebrandt E (2013) The family paenibacillaceae. The Prokaryotes: Firmicutes and Tenericutes. doi: 10.1007/978-3-642-30120-9_354

Meier-Kolthoff JP, Auch AF, Klenk HP and 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

Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC and Pati A (2015) Large-scale

contamination of microbial isolate genomes by illumina Phix control. Stand Genomic Sci. doi: 10.1186/1944-3277-10-18

Novichkov V, Kaznadzey A, Alexandrova N and Kaznadzey D (2016) NSimScan: DNA comparison tool with increased speed, sensitivity and accuracy. Bioinformatics. doi: 10.1093/bioinformatics/btw126

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D et al. (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. doi: 10.1093/nar/gkv1189

Oren A and Garrity GM (2014) Then and now: A systematic review of the systematics of prokaryotes in the last 80 years. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol. doi: 10.1007/s10482-013-0084-1

Paradis E, Claude J and Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. doi: 10.1093/bioinformatics/btg412

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

Richter M and 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

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

RStudio Team (2015) RStudio: Integrated Development Environment for R.

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 and Passaglia LMP (2018) Genome-based reclassification of Paenibacillus dauci as a later heterotypic synonym of Paenibacillus shenyangensis. Int. J. Syst. Evol. Microbiol.

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

Teeling H, Meyerdierks A, Bauer M, Amann R and Glöckner FO (2004a) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol 6:938–947. doi: 10.1111/j.1462-2920.2004.00624.x

Teeling H, Waldmann J, Lombardot T, Bauer M and Glöckner FO (2004b) TETRA: A web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics. doi: 10.1186/1471-2105-5-163

Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC and Pati A (2015) Microbial species delineation using whole genome sequences. Nucleic Acids Res 43:6761–6771. doi: 10.1093/nar/gkv657

Woese CR, Manilofft J and Zablen LB (1980) Phylogenetic analysis of the mycoplasmas (evolution/clostridiumi/classification/oligonucleotide catalog/archaebacteria). Evolution (N. Y).

Yoon SH, Ha S min, Lim J, Kwon S and Chun J (2017) w. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol. doi: 10.1007/s10482-017-0844-4

Figures

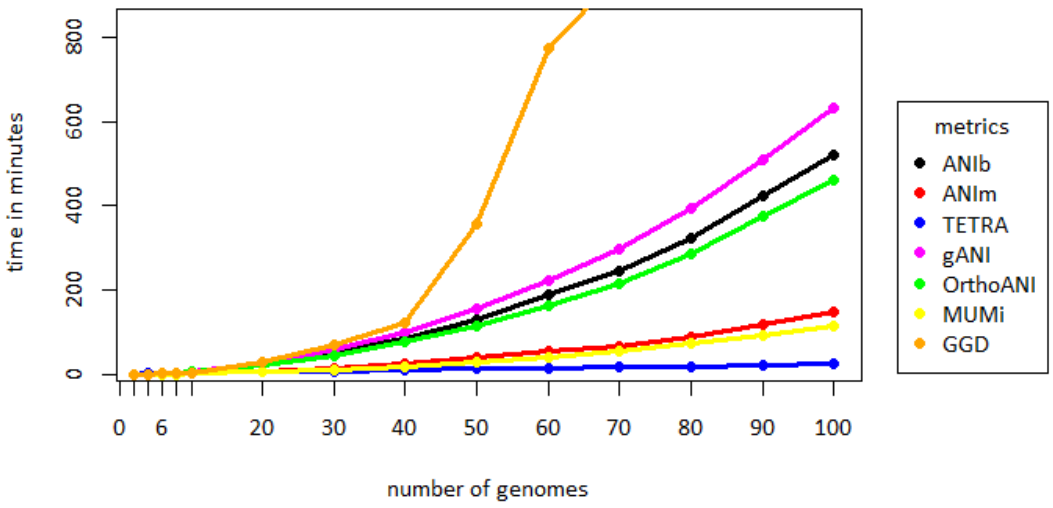


Fig 1: Runtimes of all metrics given a progressive increase in dataset. Each point on the graph represents the runtime for a given number of genomes in the dataset.

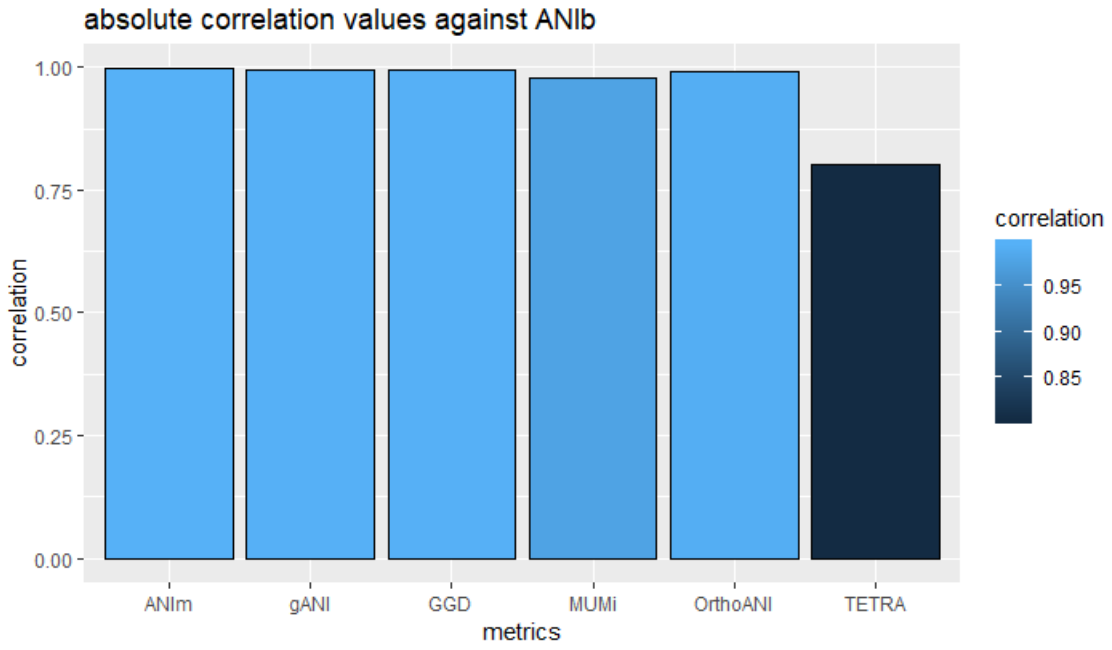


Fig 2: Correlation values of all metrics in relation to ANIb.

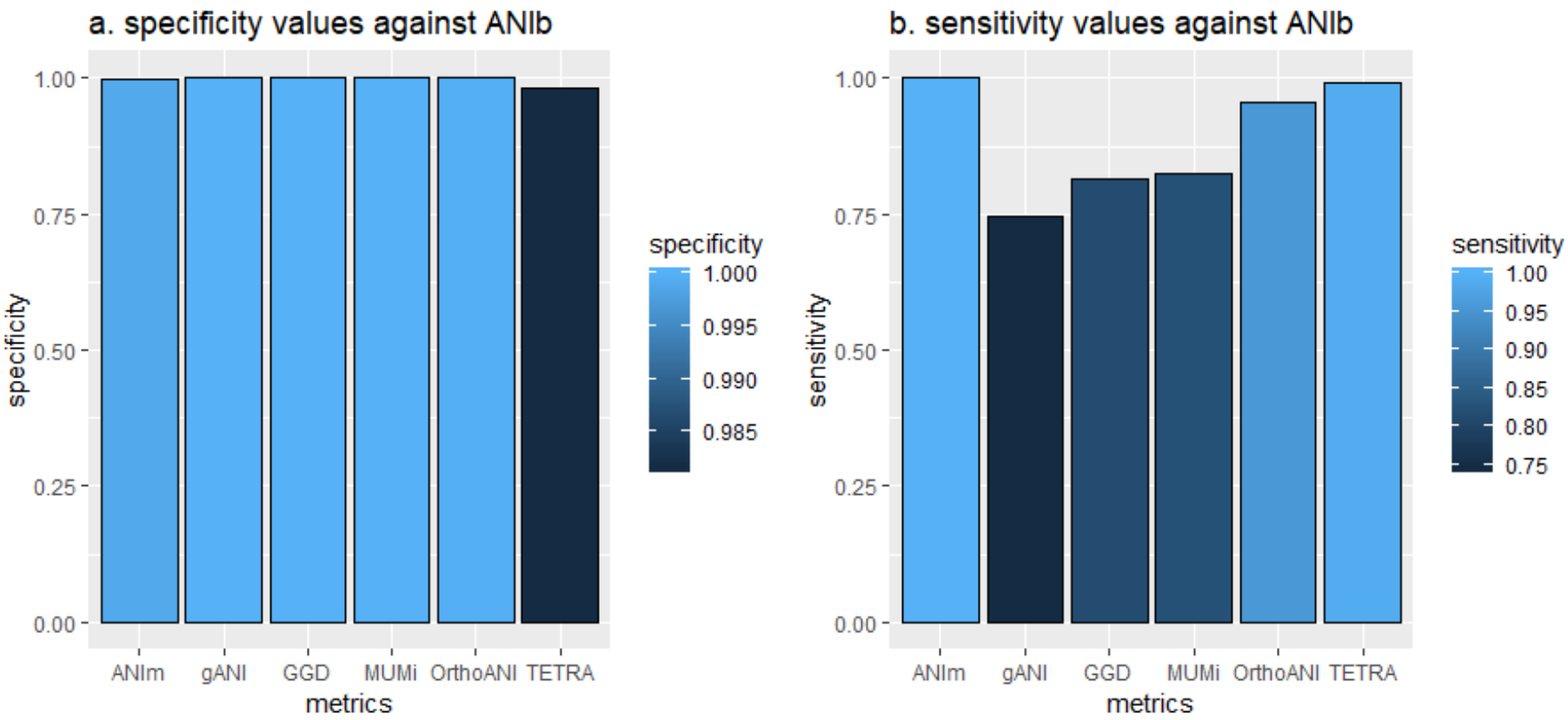


Fig 3: a. Specificity values for all metrics in relation to ANIb results. b. Sensitivity values for all metrics in relation to ANIb results.

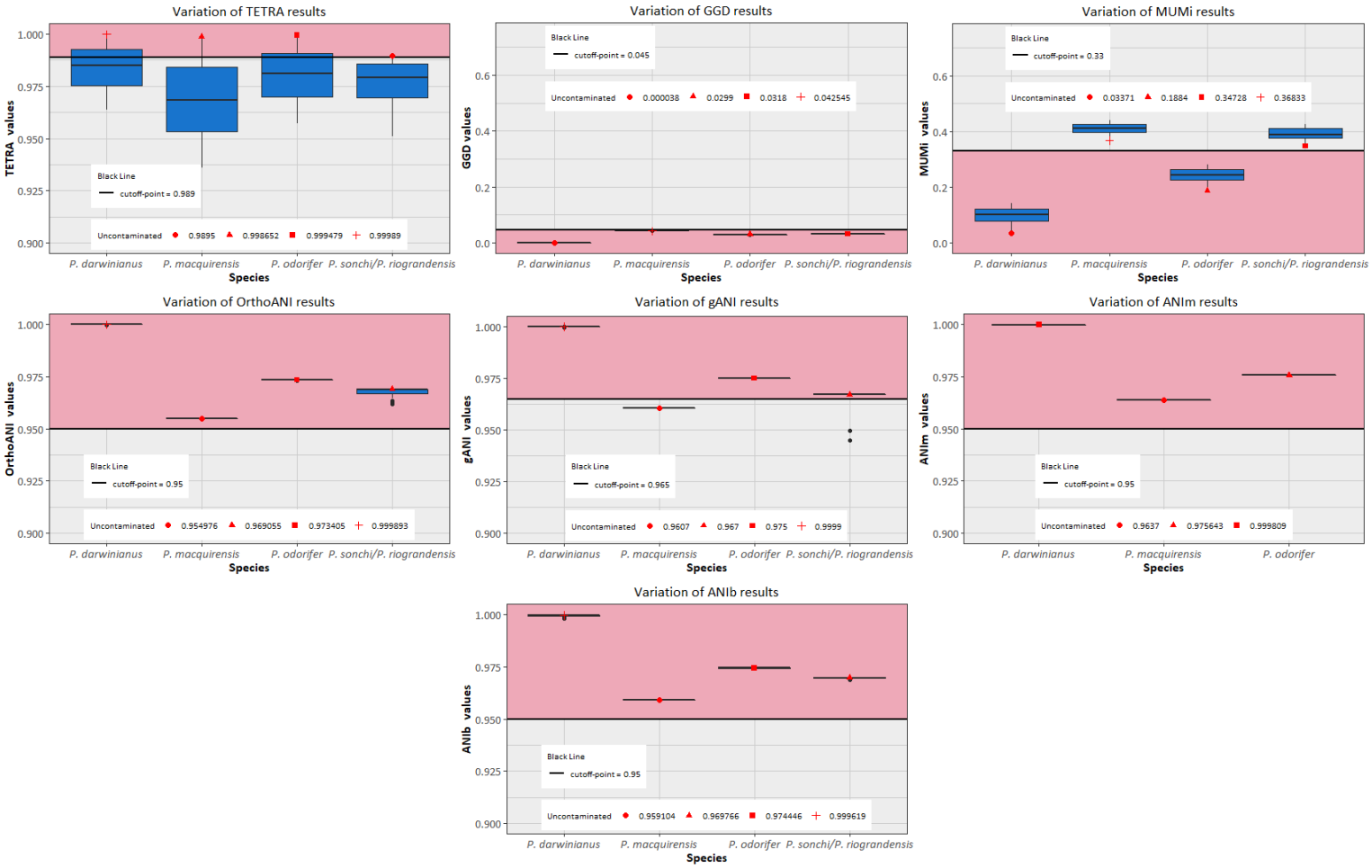
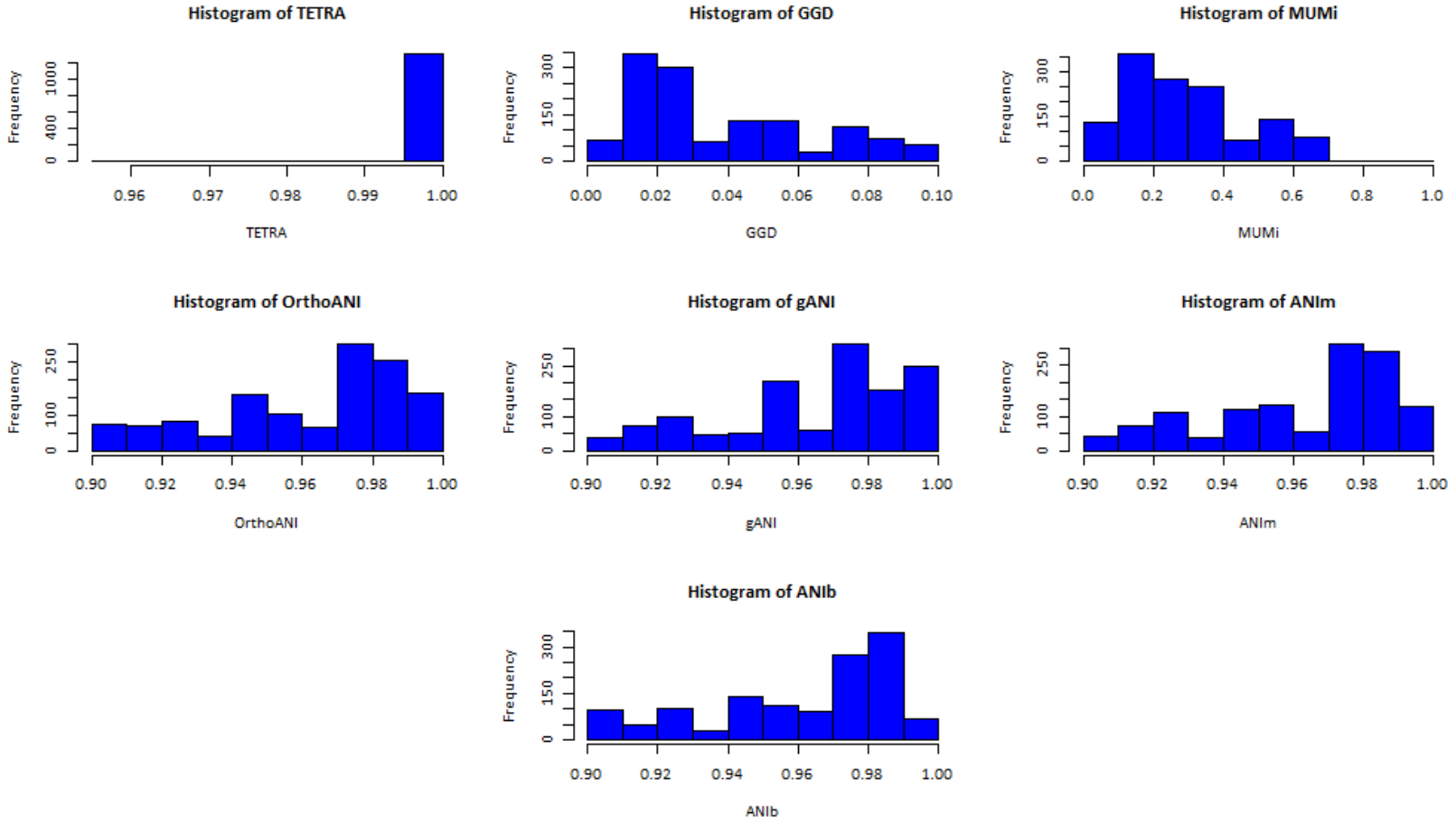


Fig 4: Boxplot of contamination results for each metric, the cross, triangle, square and circle represent the uncontaminated result for the *P. darwinianus*, *P. sonchi*/*P. riograndensis*, *P. odorifer* and *P. macquirensis* pairs, respectively. The black horizontal line in each plot represents the species circumscription threshold. The pink indicates the same species region.

	ANIb ≥ 0.95	ANIb < 0.95
≥ metric threshold*	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
< metric threshold*	FALSE NEGATIVE (FP)	TRUE NEGATIVE (TN)

Fig 5: Confusion matrix of the sensitivity and specificity tests. The threshold for the predicted values varies for each method: 0.95 for ANIm and OrthoANI, 0.965 for gANI, 0.989 for TETRA, 0.33 for MUMi and 0.044 for GGD. *for MUMi and GGD the inequality changes from ≥ and <, to ≤ and > respectively.

Additional files



Supplementary Fig 1: Distribution of all results of all metrics with a corresponding ANIb >0.90.

Supplementary Table S1: *Paenibacillus* genomes — and *E. coli* — used on this study with their respective assembly number.

Genome	Assembly
***Paenibacillus sp.* HGF5**	**GCF_000204455.1**
Paenibacillus elgii* B69	**GCF_000213315.1**
***Paenibacillus sp.* HGF7**	**GCF_000214295.1**
***Paenibacillus polymyxa* ATCC 842^{T}**	**GCF_000217775.1**
***Paenibacillus mucilaginosus* KNP414**	**GCF_000218915.1**
***Paenibacillus lactis* 154**	**GCF_000230915.1**
***Paenibacillus terrae* HPL-003**	**GCF_000235585.1**
***Paenibacillus peoriae* KCTC 3763**	**GCF_000236805.1**
***Paenibacillus polymyxa* M1**	**GCF_000237325.1**
***Paenibacillus dendritiformis* C454**	**GCF_000245555.1**
***Paenibacillus sp.* Aloe-11**	**GCF_000245715.1**
***Paenibacillus mucilaginosus* 3016**	**GCF_000250655.1**
***Paenibacillus mucilaginosus* K02**	**GCF_000258535.2**
***Paenibacillus polymyxa* OSY-DF**	**GCF_000265445.1**
***Paenibacillus sp.* OSY-SE OSY**	**GCF_000283315.1**
***Paenibacillus senegalensis* JC66^{T}**	**GCF_000285515.1**
***Paenibacillus alvei* DSM 29^{T}**	**GCF_000293805.1**
***Paenibacillus popilliae* ATCC 14706^{T}**	**GCF_000315235.1**
***Paenibacillus sp.* PAMC 26794**	**GCF_000316035.1**
***Paenibacillus sonchi* X19-5^{T}**	**GCF_000316285.1**
***Paenibacillus sp.* A9**	**GCF_000346635.1**
***Paenibacillus sp.* HW567**	**GCF_000374185.1**
***Paenibacillus sanguinis* DSM 16941**	**GCF_000374825.1**
***Paenibacillus terrígena* DSM 21567**	**GCF_000374845.1**
***Paenibacillus massiliensis* DSM 16942**	**GCF_000377505.1**
***Paenibacillus daejeonensis* DSM 15491^{T}**	**GCF_000378385.1**
***Paenibacillus ginsengihumi* DSM 21568**	**GCF_000380965.1**
***Paenibacillus fonticola* DSM 21315**	**GCF_000381905.1**
***Paenibacillus barengoltzii* G22**	**GCF_000403375.2**
***Paenibacillus sp.* HGH0039**	**GCF_000411255.1**
***Paenibacillus assamensis* DSM 18201**	**GCF_000422445.1**
***Paenibacillus harenae* DSM 16969^{T}**	**GCF_000422465.1**
***Paenibacillus pasadenensis* DSM 19293**	**GCF_000422485.1**
***Paenibacillus pinihumi* DSM 23905**	**GCF_000422505.1**
***Paenibacillus alginolyticus* DSM 5050^{T}**	**GCF_000422905.1**
***Paenibacillus taiwanensis* DSM 18679^{T}**	**GCF_000425125.1**
***Paenibacillus panacisoli* DSM 21345**	**GCF_000426545.1**
***Paenibacillus alvei* A6-6i-x**	**GCF_000442535.1**
***Paenibacillus alvei* TS-15**	**GCF_000442555.1**
***Paenibacillus antibioticophila* GD11^{T}**	**GCF_000455265.1**
***Paenibacillus polymyxa* WLY78**	**GCF_000463565.1**
***Paenibacillus sp.* P22**	**GCF_000469945.2**
***Paenibacillus sp.* MAEPY2**	**GCF_000499205.1**
***Paenibacillus sp.* MAEPY1**	**GCF_000499305.1**

***Paenibacillus polymyxa* CR1**	**GCF_000507205.3**
***Paenibacillus sp.* JCM 10914**	**GCF_000509425.1**
***Paenibacillus larvae* DSM 25719**	**GCF_000511115.1**
***Paenibacillus larvae* DSM 25430**	**GCF_000511405.1**
***Paenibacillus sp.* FSL R5-192**	**GCF_000517845.1**
***Paenibacillus sp.* FSL R7-269**	**GCF_000517865.1**
***Paenibacillus sp.* FSL R5-808**	**GCF_000517885.1**
***Paenibacillus sp.* FSL H7-689**	**GCF_000517905.1**
***Paenibacillus sp.* FSL H8-237**	**GCF_000517925.1**
***Paenibacillus sp.* FSL H8-457**	**GCF_000517945.1**
***Paenibacillus sp.* FSL R7-277**	**GCF_000517965.1**
***Paenibacillus sp.* J14**	**GCF_000518465.1**
***Paenibacillus durus* ATCC 35681**	**GCF_000520635.1**
Paenibacillus graminis* RSA19^{T}	**GCF_000520655.1**
***Paenibacillus massiliensis* T7**	**GCF_000520695.1**
***Paenibacillus forsythiae* T98^{T}**	**GCF_000520735.1**
***Paenibacillus sp.* 1-18**	**GCF_000520755.1**
***Paenibacillus polymyxa* 1-43**	**GCF_000520795.1**
***Paenibacillus sp.* 1-49**	**GCF_000520815.1**
***Paenibacillus polymyxa* SQR21**	**GCF_000597985.1**
***Paenibacillus darwinianus* Br^{T}**	**GCF_000598065.1**
***Paenibacillus darwinianus* CE1**	**GCF_000598085.1**
***Paenibacillus darwinianus* MB1**	**GCF_000598105.1**
***Paenibacillus ehimensis* A2**	**GCF_000612225.1**
***Paenibacillus sabinae* T27^{T}**	**GCF_000612505.1**
***Paenibacillus sp.* URHA0014**	**GCF_000620565.1**
***Paenibacillus sp.* UNCCL52**	**GCF_000686825.1**
***Paenibacillus sp.* UNC451MF**	**GCF_000686845.1**
***Paenibacillus sp.* UNC217MF**	**GCF_000686865.1**
***Paenibacillus polymyxa* DSM 365**	**GCF_000714835.1**
***Paenibacillus tyrfis* MSt1^{T}**	**GCF_000722545.1**
***Paenibacillus camerounensis* G4**	**GCF_000723885.1**
***Paenibacillus sp* TCA20**	**GCF_000732325.1**
***Paenibacillus polymyxa* CICC 10580**	**GCF_000735775.1**
***Paenibacillus chitinolyticus* NBRC 15660^{T}**	**GCF_000739915.1**
***Paenibacillus macerans* ATCC 8244^{T}**	**GCF_000746875.1**
***Paenibacillus durus* DSM 1735^{T}**	**GCF_000756615.1**
***Paenibacillus wynnii* DSM 18334**	**GCF_000757885.1**
***Paenibacillus sp.* FSL H7-0357**	**GCF_000758525.1**
***Paenibacillus sp.* FSL H7-0737**	**GCF_000758545.1**
***Paenibacillus sp.* FSL P4-0081**	**GCF_000758565.1**
***Paenibacillus sp.* FSL R5-0345**	**GCF_000758585.1**
***Paenibacillus sp.* FSL R5-0912**	**GCF_000758605.1**
***Paenibacillus sp.* FSL R7-0273**	**GCF_000758625.1**
***Paenibacillus sp.* FSL R7-0331**	**GCF_000758645.1**
***Paenibacillus borealis* DSM 13188^{T}**	**GCF_000758665.1**
***Paenibacillus stellifer* DSM 14472^{T}**	**GCF_000758685.1**
Paenibacillus graminis* DSM 15220^{T}	**GCF_000758705.1**

***Paenibacillus odorifer* DSM 15391^T**	**GCF_000758725.1**
***Paenibacillus polymyxa* CF05**	**GCF_000785455.1**
***Paenibacillus chondroitinus* OK414**	**GCF_000799595.1**
***Paenibacillus polymyxa* A18**	**GCF_000809185.2**
***Paenibacillus polymyxa* Sb3-1**	**GCF_000819665.1**
***Paenibacillus sp* VKM B-2647**	**GCF_000829455.1**
***Paenibacillus sp* E194**	**GCF_000935845.1**
***Paenibacillus polymyxa* NRRL B-30509**	**GCF_000943535.1**
Paenibacillus terrae NRRL B-30644	GCF_000943545.1
Paenibacillus sp. IHBB 10380	GCF_000949425.1
Paenibacillus polymyxa EBL06	GCF_000955925.1
Paenibacillus beijingensis DSM 24997	GCF_000961095.1
Paenibacillus wulumuqiensis Y24^T	GCF_000971965.1
Paenibacillus algorifonticola XJ259^T	GCF_000971975.1
Paenibacillus dauci H9^T	GCF_000971985.1
Paenibacillus riograndensis SBR5^T	GCF_000981585.1
Paenibacillus larvae MEX14	GCF_000988145.1
Paenibacillus durus ATCC 35681	GCF_000993825.1
Paenibacillus etheri SH7^T	GCF_001012825.1
Paenibacillus sp. VT-400	GCF_001029205.1
Paenibacillus sp. Mc5Re-14	GCF_001049835.1
Paenibacillus sp. D9	GCF_001188365.1
Paenibacillus peoriae HS311	GCF_001272655.2
Paenibacillus sp. FJAT-27812	GCF_001273905.1
Paenibacillus solani FJAT-22460^T	GCF_001277345.1
Paenibacillus sp. A59	GCF_001280595.1
Paenibacillus sp. JCM 10914 JCM 10914	GCF_001315105.1
Paenibacillus sp. FF9	GCF_001373415.1
Paenibacillus sp. A3	GCF_001399685.1
Paenibacillus ihumii AT5	GCF_001403875.1
Paenibacillus bovis BD3526^T	GCF_001421015.2
Paenibacillus sp. Leaf72	GCF_001422685.1
Paenibacillus sp. Root444D2	GCF_001426375.1
Paenibacillus sp. Root52	GCF_001426865.1
Paenibacillus sp. Soil724D2	GCF_001427935.1
Paenibacillus sp. Soil750	GCF_001428045.1
Paenibacillus sp. Soil766	GCF_001428105.1
Paenibacillus sp. Soil522	GCF_001428245.1
Paenibacillus sp. Soil787	GCF_001429545.1
Paenibacillus naphthalenovorans 32O-Y	GCF_001465255.1
Paenibacillus sp. 32O-W	GCF_001465275.1
Paenibacillus jamilae NS115	GCF_001477135.1
Paenibacillus polymyxa KF-1	GCF_001481575.1
Paenibacillus rubinfantis MT18	GCF_001486505.1
Paenibacillus senegalimassiliensis SIT18^T	GCF_001486585.1
Paenibacillus sp. FJAT-29882	GCF_001510645.1
Paenibacillus pabuli NBRC 13638	GCF_001514495.1
Paenibacillus sp. DMB5	GCF_001517085.1

Paenibacillus sp. FJAT-26967	GCF_001541095.1
Paenibacillus jilunlii DSM 23019^{T}*	GCF_001546055.1
Paenibacillus riograndensis CAS34	GCF_001546065.1
Paenibacillus amylolyticus Heshi-A3	GCF_001570725.1
Paenibacillus polymyxa CCI-25	GCF_001593085.1
Paenibacillus elgii M63	GCF_001619725.1
Paenibacillus jamilae CN9	GCF_001619755.1
Paenibacillus glucanolyticus 5162	GCF_001632305.1
Paenibacillus glucanolyticus SLM1	GCF_001633025.1
Paenibacillus sp. O199	GCF_001636635.1
Paenibacillus macquariensis DSM 2^{T}	GCF_001637165.1
Paenibacillus glacialis DSM 22343^{T}	GCF_001637205.1
Paenibacillus macquariensis subsp. defensor JCM 14954^{T}	GCF_001637215.1
Paenibacillus antarcticus CECT 5836^{T}	GCF_001637225.1
Paenibacillus crassostreae LPB0068^{T}*	GCF_001637335.1
Paenibacillus swuensis DY6^{T}	GCF_001644605.1
Paenibacillus sp. 1ZS3-15	GCF_001653565.1
Paenibacillus sp. AD87	GCF_001659845.1
Paenibacillus polymyxa ND25	GCF_001662815.1
Paenibacillus polymyxa ND24	GCF_001663585.1
Paenibacillus oryzae 1DrF-4^{T}	GCF_001675045.1
Paenibacillus sp. KS1	GCF_001680695.1
Paenibacillus yonginensis DCY84^{T}	GCF_001685395.1
Paenibacillus pectinilyticus KCTC13222^{T}	GCF_001700435.1
Paenibacillus kribbensis 6hRe76	GCF_001705305.1
Paenibacillus polymyxa CFSAN034343	GCF_001707685.1
Paenibacillus polymyxa CFSAN034341	GCF_001709075.1
Paenibacillus polymyxa CFSAN034342	GCF_001709135.1
Paenibacillus polymyxa J	GCF_001719045.1
Paenibacillus sp. TI45-13ar	GCF_001721045.1
Paenibacillus crassostreae LPB0068^{T}*	GCF_001857945.1
Paenibacillus sp. LC231	GCF_001860525.1
Paenibacillus sp. NAIST15-1	GCF_001894745.1
Paenibacillus xylanexedens PAMC 22703	GCF_001908275.1
Paenibacillus sp. P3E	GCF_001909045.1
Paenibacillus sp. P26E	GCF_001909055.1
Paenibacillus sp. P32E	GCF_001909085.1
Paenibacillus sp. P46E	GCF_001909095.1
Paenibacillus polymyxa ATCC 15970	GCF_001922145.1
Paenibacillus odorifer FSL R5-0883	GCF_001954285.1
Paenibacillus odorifer FSL J3-0159	GCF_001954295.1
Paenibacillus odorifer FSL H3-0280	GCF_001954335.1
Paenibacillus odorifer FSL F4-0152	GCF_001954345.1
Paenibacillus odorifer FSL H7-0604	GCF_001954365.1
Paenibacillus odorifer FSL R5-0636	GCF_001954375.1
Paenibacillus borealis FSL H7-0744	GCF_001954415.1
Paenibacillus odorifer FSL H7-0710	GCF_001954425.1
Paenibacillus odorifer FSL R5-0923	GCF_001954445.1

Paenibacillus odorifer FSL F4-0085	GCF_001954455.1
Paenibacillus odorifer FSL H8-0069	GCF_001954495.1
Paenibacillus odorifer FSL J3-0155	GCF_001954505.1
Paenibacillus odorifer FSL H8-0175	GCF_001954515.1
Paenibacillus odorifer FSL F4-0134	GCF_001954535.1
Paenibacillus odorifer FSL H8-0147	GCF_001954575.1
Paenibacillus odorifer FSL H7-0713	GCF_001954585.1
Paenibacillus odorifer FSL F4-0242	GCF_001954595.1
Paenibacillus odorifer FSL H7-0918	GCF_001954615.1
Paenibacillus odorifer FSL H7-0718	GCF_001954655.1
Paenibacillus odorifer FSL H3-0287	GCF_001954665.1
Paenibacillus odorifer FSL H7-0694	GCF_001954675.1
Paenibacillus odorifer FSL H3-0305	GCF_001955495.1
Paenibacillus sp. FSL H7-0326 FSL H7-0326	GCF_001955535.1
Paenibacillus odorifer FSL H7-0443	GCF_001955545.1
Paenibacillus odorifer FSL F4-0077	GCF_001955555.1
Paenibacillus odorifer FSL H7-0433	GCF_001955595.1
Paenibacillus odorifer FSL J3-0153	GCF_001955755.1
Paenibacillus odorifer FSL H3-0464	GCF_001955765.1
Paenibacillus odorifer FSL H3-0465	GCF_001955775.1
Paenibacillus odorifer FSL R5-0937	GCF_001955785.1
Paenibacillus odorifer FSL F4-0126	GCF_001955835.1
Paenibacillus odorifer FSL H8-0237	GCF_001955845.1
Paenibacillus sp. FSL R5-0765 FSL R5-0765	GCF_001955855.1
Paenibacillus glucanolyticus FSL R5-0817	GCF_001955865.1
Paenibacillus sp. FSL R7-0333 FSL R7-0333	GCF_001955915.1
Paenibacillus sp. FSL R7-0337 FSL R7-0337	GCF_001955925.1
Paenibacillus peoriae FSL A5-0030	GCF_001955935.1
Paenibacillus lautus FSL F4-0100	GCF_001955975.1
Paenibacillus pabuli FSL F4-0087	GCF_001955985.1
Paenibacillus amylolyticus FSL F4-0260	GCF_001955995.1
Paenibacillus amylolyticus FSL H7-0692	GCF_001956035.1
Paenibacillus sp. FSL H7-0331 FSL H7-0331	GCF_001956045.1
Paenibacillus amylolyticus FSL H8-0246	GCF_001956055.1
Paenibacillus sp. FSL H8-0548 FSL H8-0548	GCF_001956095.1
Paenibacillus sp. FSL H8-0259 FSL H8-0259	GCF_001956105.1
Paenibacillus peoriae FSL H8-0551	GCF_001956115.1
Paenibacillus peoriae FSL J3-0120	GCF_001956155.1
Paenibacillus amylolyticus FSL J3-0122	GCF_001956175.1
Paenibacillus rhizosphaerae FSL R5-0378	GCF_001956185.1
Paenibacillus sp. FSL R5-0490 FSL R5-0490	GCF_001956215.1
Paenibacillus peoriae FSL R7-0131	GCF_001956225.1
Paenibacillus peoriae FSL R7-0321	GCF_001956235.1
Paenibacillus sp. FSL A5-0031 FSL A5-0031	GCF_001956295.1
Paenibacillus sp. FSL R7-0273 FSL R7-0273	GCF_001957005.1
Paenibacillus ihbetae IHBB 9951	GCF_001996445.1
Paenibacillus larvae ATCC 9545^{T}	GCF_002003265.1
Paenibacillus selenitireducens ES3-24^{T}	GCF_002021565.1

Paenibacillus sp. VT-16-81	GCF_002027255.1
Paenibacillus ferrarius CY1^{T}	GCF_002027705.1
Paenibacillus sp. 32352 32352	GCF_002042965.1
Paenibacillus larvae CCM 38	GCF_002043025.1
Paenibacillus thiaminolyticus NRRL B-4156^{T}	GCF_002161855.1
Paenibacillus apiarius NRRL B-23460	GCF_002161865.1
Paenibacillus sp. MY03	GCF_002165585.1
Paenibacillus donghaensis KCTC 13049^{T}	GCF_002192415.1
Paenibacillus sp. SSG-1	GCF_002224835.1
Paenibacillus herberti R33^{T}	GCF_002233675.1
Paenibacillus rigui JCM 16352^{T}	GCF_002234615.1
Paenibacillus physcomitrellae XB^{T}	GCF_002240225.1
Paenibacillus kribbensis AM49^{T}	GCF_002240415.1
Paenibacillus sp. RUD330	GCF_002243345.1
Paenibacillus sp. XY044	GCF_002257645.1
Paenibacillus taichungensis VTT E-133285	GCF_002264305.1
Paenibacillus odorifer VTT E-133288	GCF_002264345.1
Paenibacillus sp. VTT E-133280	GCF_002264385.1
Paenibacillus sp. VTT E-133291	GCF_002264395.1
Paenibacillus campinasensis 7537-G1	GCF_002272015.1
Paenibacillus sp. 7884-2 7884-2	GCF_002272235.1
Paenibacillus sp. 7516 7516	GCF_002272715.1
Paenibacillus sp. 7541 7541	GCF_002276395.1
Paenibacillus sp. 7523-1 7523-1	GCF_002276415.1
Paenibacillus lautus BHU3	GCF_002407025.1
Paenibacillus sp. EZ-K15	GCF_002573715.1
Paenibacillus sp. BIHB4019	GCF_002741035.1
Paenibacillus ihbetae IHBB 9852^{T}	GCF_002741055.1
Paenibacillus sp. LK1	GCF_002750415.1
Paenibacillus amylolyticus GM1FR	GCF_002803315.1
Paenibacillus vortex GM2FR	GCF_002803325.1
Paenibacillus sp. BGI2013	GCF_002843485.1
Paenibacillus pasadenensis R16	GCF_002864175.1
Paenibacillus sp. lzh-N1	GCF_002872435.1
Paenibacillus castaneae DSM 19417^{T}	GCF_002884445.1
Paenibacillus polymyxa HY96-2	GCF_002893885.1
Paenibacillus sp. F4	GCF_002894905.1
Paenibacillus polymyxa KCCM 40454	GCF_002894925.1
Paenibacillus phocaensis mt24^{T}	GCF_900021165.1
Paenibacillus sp. GM2	GCF_900069005.1
Paenibacillus tuaregi Marseille-P2472^{T}	GCF_900086655.1
Paenibacillus amylolyticus Pamy_1	GCF_900095775.1
Paenibacillus typhae CGMCC 1.11012^{T}	GCF_900099765.1
Paenibacillus naphthalenovorans PR-N1^{T}	GCF_900099895.1
Paenibacillus tianmuensis CGMCC 1.8946^{T}	GCF_900100345.1
Paenibacillus sp. OK060	GCF_900101205.1
Paenibacillus sp. CF095	GCF_900101225.1
Paenibacillus sp. cl123	GCF_900101595.1

Paenibacillus polysaccharolyticus BL9^{T}	GCF_900102085.1
Paenibacillus sp. cl6col	GCF_900102125.1
Paenibacillus jilunlii CGMCC 1.10239^{T}*	GCF_900102965.1
Paenibacillus sp. yr247	GCF_900103825.1
Paenibacillus sp. GP183	GCF_900104695.1
Paenibacillus sp. PDC88	GCF_900106725.1
Paenibacillus sp. CF384	GCF_900106745.1
Paenibacillus sp. 276b	GCF_900107875.1
Paenibacillus sp. 181mfcol5.1	GCF_900109125.1
Paenibacillus sp. cl141a	GCF_900109305.1
Paenibacillus sp. OK003	GCF_900109515.1
Paenibacillus sp. OK076	GCF_900110055.1
Paenibacillus sp. OV219	GCF_900110075.1
Paenibacillus sophorae CGMCC 1.10238^{T}	GCF_900110315.1
Paenibacillus sp. NFR01	GCF_900111565.1
Paenibacillus catalpae CGMCC 1.10784^{T}	GCF_900112695.1
Paenibacillus algorifonticola CGMCC 1.10223	GCF_900112925.1
Paenibacillus sp. UNC496MF	GCF_900113915.1
Paenibacillus sp. 1 12	GCF_900114475.1
Paenibacillus sp. cl130	GCF_900116035.1
Paenibacillus sp. 453mf	GCF_900116105.1
Paenibacillus sp. BC26	GCF_900116125.1
Paenibacillus sp. UNCCL117	GCF_900119055.1
Paenibacillus sp. ov031	GCF_900143165.1
Paenibacillus sp. RU4X	GCF_900155905.1
Paenibacillus sp. RU4T RU4	GCF_900156355.1
Paenibacillus macquariensis subsp. macquariensis ATCC 23464^{T}	GCF_900156375.1
Paenibacillus sp. RU5A	GCF_900168305.1
Paenibacillus barengoltzii J12	GCF_900177265.1
Paenibacillus sp. J6	GCF_900177285.1
Paenibacillus aquistagni 11^{T}	GCF_900177815.1
Paenibacillus elgii B69*	GCF_900188505.1
Paenibacillus sp. St-s	GCF_900188525.1
Paenibacillus sp. RU26A	GCF_900221045.1
Paenibacillus sp. RU5M	GCF_900221055.1
Escherichia coli DSM 30083^{T}	GCF_000690815.1

*same organism, different assembly. Bold – genomes used on the runtime test.

Supplementary Table S2: Command line used for the execution of each metric.

Methods	**Parameters**
ANIb	averagenucleotideidentity.py -i PATH:to/Genomes -o PATH:to/Output -m ANIb -g –workers 8 –write_excel
ANIm	averagenucleotideidentity.py -i PATH:to/Genomes -o PATH:to/Output -m ANIm -g –workers 8 –write_excel
TETRA	averagenucleotideidentity.py -i PATH:to/Genomes -o PATH:to/Output -m TETRA -g –workers 8 –write_excel
gANI	./ANIcalculator -genome1fna PATH:to/genome1 -genome2fna PATH:to/genome2 -outfile PATH:to/outfile-outdir PATH:to/Outdir
OrthoANI	java -jar OATcmd.jar -fasta1 PATH:to/genome1 -fasta2 PATH:to/genome2 -method ani -numthreads 8
GGDC	java -jar OATcmd.jar -fasta1 PATH:to/genome1 -fasta2 PATH:to/genome2 -method ggdc -numthreads 8
MUMi	print mumi.batchrun(directory='PATH:to/Genomes', ext='fna', numthreads=8, k=19)

4. Capitulo 2

Artigo 2

Título: Genomic metrics analyses indicate that *Paenibacillus azotofixans* is not a later synonym of *Paenibacillus durus*

Submetido para: *International Journal of Systematic and Evolutionary Microbiology*

Genomic metrics analyses indicate that *Paenibacillus azotofixans* is not a later synonym of *Paenibacillus durus*

Felipe Guella[1], Renan Zanini Porto[1], Fernando Hayashi Sant'Anna[1], Adriana Ambrosini[1], Luciane Maria Pereira Passaglia[1*]

1 - Departamento de Genetica, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

*Corresponding author at Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul (UFRGS). Av. Bento Gonçalves, 9500, Caixa Postal 15.053, Prédio 43312, sala 207b, Porto Alegre, RS, CEP 91501-970, Brasil. Tel.: +55 51 3308 9813; Fax +55 51 3308 7311. E-mail address: luciane.passaglia@ufrgs.br (L. M. P. Passaglia)

Keywords: *Paenibacillus; Paenibacillus durus; Paenibacillus azotofixans*; ANI; *Paenibacillus zanthoxyli*; dDDH; genome metrics; taxonomy; reclassification
Wordcount: 1451 words
Subject category: Taxonomy note

ABSTRACT

Paenibacillus durus and *Paenibacillus azotofixans*, both Gram-stain-positive and endospore-forming bacilli have been considered to be a single species. However, a preliminary computation of their ANI values suggested that these species are not synonyms. Given this, the taxonomic attributions of these species were evaluated through different genomic and phylogenetic approaches. Although the identity of 16S rRNA gene sequences of *P. durus* DSM 1735 ^{T} and *P. azotofixans* ATCC 35681^{T} are above the circumscription species threshold, genomic metrics analyses indicate otherwise. ANI, gANI and OrthoANI values computed from their genome sequences were around 92%, below the species limits. dDDH and MUMi estimations also corroborated these observations. In fact, in all metrics, *Paenibacillus zanthoxyli* JH29^{T} seemed to be more similar to *P. azotofixans* ATCC 35681^{T} than *P. durus* DSM 1735 ^{T}. Phylogenetic analyses based on concatenated core-proteome and concatenated *gyrB*, *recA*, *recN*, and *rpoB* genes confirmed that *P. zanthoxyli* is the closest *Paenibacillus* species to *P. azotofixans.* A review of the phenotypic profiles from these three species revealed that their biochemical repertoires are very similar, although *P. azotofixans*

ATCC 35681^{T}can be differentiated from *P. durus* DSM1735T in 12 among more than 90 phenotypic traits. Considering phylogenetic and genomic analyses, *Paenibacillus azotofixans* should be considered as an independent species, and not as a later synonym of *Paenibacillus durus*.

Taxonomic note

The genus *Paenibacillus* is composed of Gram stain–positive endospore-forming and facultative anaerobic bacteria with heterogeneous characteristics. These bacteria inhabit a broad spectrum of environments (1), such as the Antarctic continent (2), volcanic soil (3), earthworm gut (4), and some species are known as honeybee pathogens (5). However, many of the species were found associated with plants (1), and some strains can act as plant growth promoters (6–8).

With the development of high-throughput sequencing tools, genome-based comparison of sequences is being utilized for describing new species and reclassifying wrongly identified organisms (9). In this sense, Richter et al. (9) suggested that wet lab methods might be replaced by average nucleotide identity (ANI) as gold standard for species classification.

Since the mid-90s, bacterial classification follows the polyphasic taxonomy approach, which integrates phylogenetic, genotypic and phenotypic data in order to define a species (10). Several species, however, were described before this classification method was implemented (11). While the *Paenibacillus* genus is relatively new, most of its species were relocated from other genera (12–14); therefore, some of its species were later reclassified (15–17).

Paenibacillus durus DSM 1735^{T} was initially classified as *Clostridium durum* (18), being relocated to the *Paenibacillus* genus (12, 13). *Paenibacillus azotofixans* ATCC 35681^{T}, on the other hand, was initially classified as *Bacillus azotofixans* (19), and later moved to the *Paenibacillus* genus (20). Rosado et al (21), using 16S rRNA gene similarity and DNA-DNA hybridization (DDH) analyses, concluded that both strains belonged to the same species, and reclassified them as belonging to the *Paenibacillus azotofixans* species. Finally, the Judicial Commission of the International Committee on Systematic of Prokaryotes (ICSP) determined that *P. durus* name had priority over *P. azotofixans* and therefore both strains should be named as *P. durus* (22), with *P.azotofixans* being considered as a later synonym of *P. durus*. A more thorough representation of both strains history can be seen at Figure 1.

In spite of that, a prior computation of Average Nucleotide Identity (ANI) values among all genome sequences of *Paenibacillus* strains deposited in Refseq indicated that *P. azotofixans* ATCC 35681^T is not a later synonym of *P. durus* DSM 1735^T (data not shown). Given this, the objective of this study is to clarify the taxonomic statuses of *P. azotofixans* ATCC 35681^T and *P. durus* DSM 1735^T using genomic and phylogenetic analyses.

In order to identify the closest *Paenibacillus* species to *P. azotofixans* and *P. durus*, phylogenetic reconstruction of 16S rRNA gene sequences from *Paenibacillus* type-strains was conducted. At first, all genes were downloaded from the GenBank database, using the List of Prokaryotic names with Standing in Nomenclature (LPSN) as reference (23). However, in a preliminary analysis, it was noted that the reference sequences of *P. azotofixans* ATCC 35681^T (X60608) and *P. durus* DSM 1735^T (X77846) significantly differed from the genome sequences available in Refseq, probably given to sequencing errors. Therefore, for the 16S rRNA gene phylogenetic reconstruction, considering the locus tag numbers, the first 16S rRNA copies from the genome sequences of *P. azotofixans* ATCC 35681^T and *P. durus* DSM 1735^T were utilized. Sequences were aligned using SINA aligner (24), and sequence gaps were excluded in Bioedit version 7.0.5 (25). The tree was built using Maximum Likelihood method with PhyML version 3.0, available at the Phylogeny.fr platform (26). The substitution model was Generalized Time Reversible (GTR), assuming an estimated proportion of invariant sites and four gamma-distributed rate categories for rate heterogeneity across sites. Gamma shape parameters were estimated from the data. The reliability of the clades was estimated using the approximate Likelihood Ratio Test (aLRT) (27).

The phylogenetic reconstruction of the 16S rRNA genes indicated that *P. azotofixans* ATCC 35681^T, *P. durus* DSM 1735^T, and *P. zanthoxyli* JH29^T form a monophyletic group (Figure 2). The 16S rRNA gene identity values between *P. azotofixans* ATCC 35681^T, *P. durus* DSM 1735^T and their closest species were computed in Bioedit using the SINA alignment without positions containing gaps. As demonstrated in Table 1, both *P. durus* DSM 1735^T and *P. zanthoxyli* JH29^T presented values that reached the species delimitation threshold (98.7%) in relation to *P. azotofixans* ATCC 35681^T (17). Intragenomic and intergenomic diversity between the copies of 16S rRNA genes of *P. azotofixans* ATCC 35681^T and *P. durus* DSM 1735^T was also investigated using the previous approach. *Paenibacillus durus* DSM 1735^T presented high sequence variability between its own copies, and some values were lower than the identity threshold for species circumscription (Figure 3). It is worth noting that depending on the pair of sequences compared, *P. azotofixans* ATCC

35681^T and *P. durus* DSM 1735^T could be considered to be of the same species. Nevertheless, for most of 16S rRNA intergenomic comparisons, identity values were below the species threshold. These observations illustrate how 16S rRNA identity analyses can be biased; therefore, they should be evaluated with caution.

For genome analyses methods, closely related genomes to those from both strains were selected and downloaded from RefSeq database (Table 2). The genome sequence of *P. polymyxa* ATCC 842^T, the type-species of the *Paenibacillus* genus, was utilized as outgroup. Core proteome analysis was conducted using GET_HOMOLOGUES version 03012019 (28), and the orthologous proteins were clustered using OrthoMCL version 1.4 (29) included in the GET_HOMOLGUES package. All of the 2,944 core proteins presented were aligned using MUSCLE version 3.8.31 (30). The resulting alignment was concatenated using MEGA-X version 10.0.5 software (31), resulting in 43,6947 amino acid positions. Positions containing gaps were then removed using Gblocks version 0.91b (32), resulting in 38,9145 amino acid positions, covering about 89% of the original alignment. The phylogeny of the core proteomics was reconstructed using Neighbor-joining method (33) with JTT substitution model and 1000 bootstrap replicates in MEGA-X version 10.0.5 software (31).

Multilocus sequence analysis was also conducted using four housekeeping genes: *gyrB*, *recA*, *recN* and *rpoB*. The genes were extracted from the genomes downloaded from the RefSeq database. Their Locus tags are available on Supplementary Table S1. Each sequence was aligned using MUSCLE software (30), then concatenated with MEGA-X version 10.0.5 (31). Columns containing gaps were removed using BioEdit version 7.0.5, and the maximum-likelihood (ML) estimation tree was reconstructed with MEGA-X version 10.0.5 (31), using the Tamura-Nei model and 1000 bootstrap replicates. The housekeeping genes identity matrix was conducted using BioEdit version 7.0.5 and identity values are available at Supplementary Table S2.

The core-proteome tree showed that *P. azotofixans* ATCC 35681^T and *P. zanthoxyli* JH29^T are the closest neighbors among closely related *Paenibacillus* species, both for core-proteome tree and housekeeping MLSA tree (Figures 4 and S1). These results contradict rRNA 16S phylogenetic reconstruction (Figure 2) and identity analysis (Table 1), and they suggest that the rRNA 16S genes could not represent properly the genomic diversity of these strains. On the other hand, these comparisons could be affected by distinct quality of sequences deposited in Genbank or by the high intragenomic variability of these genes (Figure 3).

In order to better solve this taxonomic problem, several genomic metrics were also used in order to validate phylogenetic reconstruction: ANIb (Average Nucleotide Identity that uses BLASTn+) (34), dDDH (35, 36), gANI (genome-wide ANI) (37), OrthoANI (36) and MUMi (maximal unique matches index) (38). For all genomic metrics, results corroborate the core proteome phylogeny and the MLSA, indicating that *P.azotofixans* ATCC 35681^T is more closely related to *P. zanthoxyli* JH29^T than to *P.durus* DSM 1735^T, although in none of these metrics the values surpassed the species threshold (Table 3).

After phylogenetic reconstructions and metrics results demonstrated strong relation between *P. zanthoxyli* JH29^T and *P. azotofixans* ATCC 35681^T, the phenotypic profiles of these type strains and *P. durus* DSM 1735^T were revised (see Supplementary Tables S3 and S4) (18, 19, 21, 39–46). Results show that, from more than dozen traits tested in all three organisms, *P. zanthoxyli* JH29^T and *P. azotofixans* ATCC 35681^T presented more similar profiles (Supplementary Table S3). *Paenibacillus durus* DSM 1735^T is positive for the dextrin production, oxidase activity and acid production from D-sorbitol, and negative for the growth in NaCl 3% (w/v), characteristics that differentiate it from *P. azotofixans* ATCC 35681^T and *P. zanthoxyli* JH29^T. Furthermore, from about 90 traits tested in *P. azotofixans* ATCC 35681^T and *P. durus* DSM 1735^T, 12 are differential (Supplementary Table S3). The major cellular fatty acids composition had no significant difference between the three strains, which are anteiso-branched $C_{15:0}$ and straight-chain $C_{16:0}$ (see Supplementary Table S4).

In conclusion, genomic and phylogenetic analyses demonstrate that *P. azotofixans* ATCC 35681^T and *P. durus DSM* 1735^T are members of different species. In fact, *P. zanthoxyli* is phylogenetically closer and genomically more similar to *P. azotofixans* than to *P. durus.* Therefore, we propose that *Paenibacillus azotofixans* should be considered an independent species, as previously described by Seldin et al. (1984) (19), and not as a later synonym of *Paenibacillus durus*.

AUTHOR STATEMENTS

Funding information

This work was funded by CNPq/INCT-FBN (Conselho Nacional de DesenvolvimentoCientífico e Tecnológico/Instituto Nacional de Ciência e Tecnologia da FixaçãoBiológica de Nitrogênio, Brazil). AA, FG, FHS, and RZP received scholarships from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil).

Conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ABBREVIATIONS

aLRT, approximate Likelihood-Ratio Test; ANI, Average Nucleotide Identity; ANIb, Average Nucleotide Identity based on Blast+; dDDH, digital DNA-DNA hybridization; gANI, genome-wide Average Nucleotide Identity; GTR, General Time-Reversible; LPSN, List of Prokaryotic names with Standing in Nomenclature; MUMi, Maximal Unique Matches index; OrthoANI, Orthologous Average Nucleotide Identity; TETRA, tetranucleotide frequency correlation coefficients.

REFERENCES

1. Grady EN, MacDonald J, Liu L, Richman A, Yuan Z-C. Current knowledge and perspectives of *Paenibacillus*: a review. Microb Cell Fact [Internet]. 2016;15(1):203. Available from: http://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-016-0603-7 DOI: 10.1186/s12934-016-0603-7
2. Montes MJ, Mercadé E, Bozal N, Guinea J. *Paenibacillus antarcticus* sp. nov., a novel psychrotolerant organism from the Antarctic environment. Int J Syst Evol Microbiol. 2004;54(5):1521–6. DOI: 10.1099/ijs.0.63078-0
3. Uetanabaro AP, Wahrenburg C, Hunger W, Pukall R, Spröer C, Stackebrandt E, et al. *Paenibacillus agarexedens* sp. nov., nom. rev., and *Paenibacillus agaridevorans* sp. nov. Int J Syst Evol Microbiol. 2003; DOI: 10.1099/ijs.0.02420-0
4. Validov S, Kamilova F, Qi S, Stephan D, Wang JJ, Makarova N, et al. Selection of bacteria able to control *Fusarium oxysporum* f. sp. radicis-lycopersici in stonewool substrate. J Appl Microbiol. 2007; DOI: 10.1111/j.1365-2672.2006.03083.x
5. Keller A, Brandel A, Becker MC, Balles R, Abdelmohsen UR, Ankenbrand MJ, Sickel1 W. Wild bees and their nests host *Paenibacillus* bacteria with functional potential of avail. Microbiome 2018; 6:229-38; DOI: 10.1186/s40168-018-0614-1
6. de Souza R, Meyer J, Schoenfeld R, da Costa PB, Passaglia LMP. Characterization of plant growth-promoting bacteria associated with rice cropped in iron-stressed soils. Ann Microbiol. 2015; DOI: 10.1007/s13213-014-0939-3
7. Fürnkranz M, Adam E, Müller H, Grube M, Huss H, Winkler J, et al. Promotion of

growth, health and stress tolerance of Styrian oil pumpkins by bacterial endophytes. Eur J Plant Pathol. 2012; DOI: 10.1007/s10658-012-0033-2

8. Ker K, Seguin P, Driscoll BT, Fyles JW, Smith DL. Switchgrass establishment and seeding year production can be improved by inoculation with rhizosphere endophytes. Biomass and Bioenergy. 2014; DOI: 10.1177/0885066616643529
9. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci [Internet]. 2009;106(45):19126–31. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0906412106 DOI: 10.1073/pnas.0906412106
10. Vandamme P, Pot B, Gillis M. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Mol Biol Rev. 1996; DOI: 10.1007/s12088-007-0022-x
11. Gürtler V, Mayall BC, Seviour R. Can whole genome analysis refine the taxonomy of the genus *Rhodococcus*? FEMS Microbiology Reviews. 2004. DOI: 10.1016/j.femsre.2004.01.001
12. Halim MA, Rahman AY, Sim K-S, Yam H-C, Rahim AA, Ghazali AHA, et al. Genome sequence of a gram-positive diazotroph, *Paenibacillus durus* type strain ATCC 35681. Genome Announc. 2016; DOI: 10.1128/genomea.00005-16
13. Collins MD, Lawson PA, Willems A, Cordoba JJ, Fernandez-Garayzabal J, Garcia P, et al. The phylogeny of the genus *Clostridium*: Proposal of five new genera and eleven new species combinations. Int J Syst Bacteriol. 1994; DOI: 10.1099/00207713-44-4-812
14. Ash C, Priest FG, Collins MD. 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. 1993;64(3–4):253–60. DOI: 10.1007/bf00873085
15. Sant'Anna FH, Ambrosini A, de Souza R, de Carvalho Fernandes G, Bach E, Balsanelli E, et al. Reclassification of *Paenibacillus riograndensis* as a genomovar of *Paenibacillus sonchi*: Genome-based metrics improve bacterial taxonomic classification. Front Microbiol. 2017; DOI: 10.3389/fmicb.2017.01849
16. Kim KK, Lee KC, Lee JS. Reclassification of *Paenibacillus ginsengisoli* as a later heterotypic synonym of *Paenibacillus anaericanus*. Int J Syst Evol Microbiol. 2011; DOI: 10.1099/ijs.0.025650-0
17. Sant'Anna FH, Ambrosini A, Guella FL, Porto RZ, Passaglia LMP. Genome-based

reclassification of *Paenibacillus dauci* as a later heterotypic synonym of *Paenibacillus shenyangensis*. Int J Syst Evol Microbiol [Internet]. 2018; Available from: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.003127.v1

18. Smith LD, Cato EP. *Clostridium durum*, sp. nov., the predominant organism in a sediment core from the Black Sea. Can J Microbiol. 1974; DOI: 10.1139/m74-214
19. Seldin L, Van Elsas JD, Penido EGC. *Bacillus azotofixans* sp. nov., a nitrogen-fixing species from brazilian soils and grass roots. Int J Syst Bacteriol. 1984; DOI: 10.1099/00207713-34-4-451
20. Heyndrickx M, Vandemeulebroecke K, Scheldeman P, Kersters P, De Vos P, Logan NA, et al. A polyphasic reassessment of the genus *Paenibacillus*, reclassification of *Bacillus lautus* (Nakamura 1984) as *Paenibacillus lautus* comb. nov. and of *Bacillus peoriae* (Montefusco et al. 1993) as *Paenibacillus peoriae* comb. nov., and emended descriptions of. Int J Syst Evol Microbiol [Internet]. 1996;46(4):988–1003. Available from: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-46-4-988
21. Rosado AS, Van Elsas JD, Seldin L. Reclassification of *Paenibacillus durum* (Formerly *Clostridium durum* Smith and Cato 1974) Collins et al. 1994 as a member of the species *P. azotofixans* (formerly *Bacillus azotofixans* Seldin et al. 1984) Ash et al. 1994. Int J Syst Evol Microbiol [Internet]. 1997;47(2):569–72. Available from: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-47-2-569
22. of the International Committee on Systematics of Prokaryotes JC. *Paenibacillus durus* (Collins et al. 1994, formerly *Clostridium durum* Smith and Cato 1974) has priority over *Paenibacillus azotofixans* (Seldin et al. 1984). Opinion 73. Int J Syst Evol Microbiol [Internet]. 2003;53(3):931. Available from: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.02496-0
23. Parte AC. LPSN - List of prokaryotic names with standing in nomenclature (Bacterio.net), 20 years on. Int J Syst Evol Microbiol. 2018. DOI: 10.1099/ijsem.0.002786
24. Pruesse E, Peplies J, Glöckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012; DOI: 10.1093/bioinformatics/bts252
25. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 1999; DOI: citeulike-article-id:691774

26. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008; DOI: 10.1093/nar/gkn180
27. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 2006; DOI: 10.1080/10635150600755453
28. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol [Internet]. 2013 Dec 15;79(24):7696 LP-7701. Available from: http://aem.asm.org/content/79/24/7696.abstract DOI: 10.1128/aem.02411-13
29. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; DOI: 10.1101/gr.1224503
30. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; DOI: 10.1093/nar/gkh340
31. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018; DOI: 10.1093/molbev/msy096
32. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; DOI: 10.1093/oxfordjournals.molbev.a026334
33. Nei M, Saitou N. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol [Internet]. 1987 Jul 1;4(4):406–25. Available from: https://dx.doi.org/10.1093/oxfordjournals.molbev.a040454 DOI: 10.1093/oxfordjournals.molbev.a040454
34. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci [Internet]. 2005;102(7):2567–72. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0409727102 DOI: 10.1073/pnas.0409727102
35. Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013; DOI: 10.1186/1471-2105-14-60
36. Lee I, Kim YO, Park SC, Chun J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. Int J Syst Evol Microbiol. 2016;66(2):1100–3. DOI: 10.1099/ijsem.0.000760

37. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, et al. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 2015;43(14):6761–71. DOI: 10.1093/nar/gkv657
38. Deloger M, El Karoui M, Petit MA. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J Bacteriol. 2009;91(1):91–9. DOI: 10.1128/jb.01202-08
39. Kong BH, Liu QF, Liu M, Liu Y, Liu L, Li CL, et al. *Paenibacillus typhae* sp. nov., isolated from roots of *Typha angustifolia* L. Int J Syst Evol Microbiol. 2013; DOI: 10.1099/ijs.0.042747-0
40. Yoon JH, Oh HM, Yoon BD, Kang KH, Park YH. *Paenibacillus kribbensis* sp. nov. and *Paenibacillus terrae* sp. nov., bioflocculants for efficient harvesting of algal cells. Int J Syst Evol Microbiol. 2003; DOI: 10.1099/ijs.0.02108-0
41. Seldin L, Penido EGC. Identification of *Bacillus azotofixans* using API tests. Antonie Van Leeuwenhoek. 1986; DOI: 10.1007/bf00393468
42. Elo S, Suominen I, Kämpfer P, Juhanoja J, Salkinoja-Salonen M, Haahtela K. Paenibacillus borealis sp. nov., a nitrogen-fixing species isolated from spruce forest humus in Finland. Int J Syst Evol Microbiol. 2001; DOI: 10.1099/00207713-51-2-535
43. Ma Y, Zhang J, Chen S. *Paenibacillus zanthoxyli* sp. nov., a novel nitrogen-fixing species isolated from the rhizosphere of *Zanthoxylum simulans*. Int J Syst Evol Microbiol. 2007; DOI: 10.1099/ijs.0.64652-0
44. Ma YC, Chen SF. *Paenibacillus forsythiae* sp. nov., a nitrogen-fixing species isolated from rhizosphere soil of Forsythia mira. Int J Syst Evol Microbiol. 2008; DOI: 10.1099/ijs.0.65238-0
45. Jin HJ, Lv J, Chen SF. *Paenibacillus sophorae* sp. nov., a nitrogen-fixing species isolated from the rhizosphere of *Sophora japonica*. Int J Syst Evol Microbiol. 2011; DOI: 10.1099/ijs.0.021709-0
46. Xie JB, Zhang LH, Zhou YG, Liu HC, Chen SF. *Paenibacillus taohuashanense* sp. nov., a nitrogen-fixing species isolated from rhizosphere soil of the root of *Caragana kansuensis* Pojark. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol. 2012; DOI: 10.1007/s10482-012-9773-4
47. Euzéby, J. P. Taxonomic note: necessary correction of specific and subspecific epithets according to Rules 12c and 13b of the International Code of Nomenclature of Bacteria (1990 Revision). Int J Syst Bacteriol 48, 1073–1075. 1998. DOI: 10.1099/00207713-48-3-1073

Figures and Tables

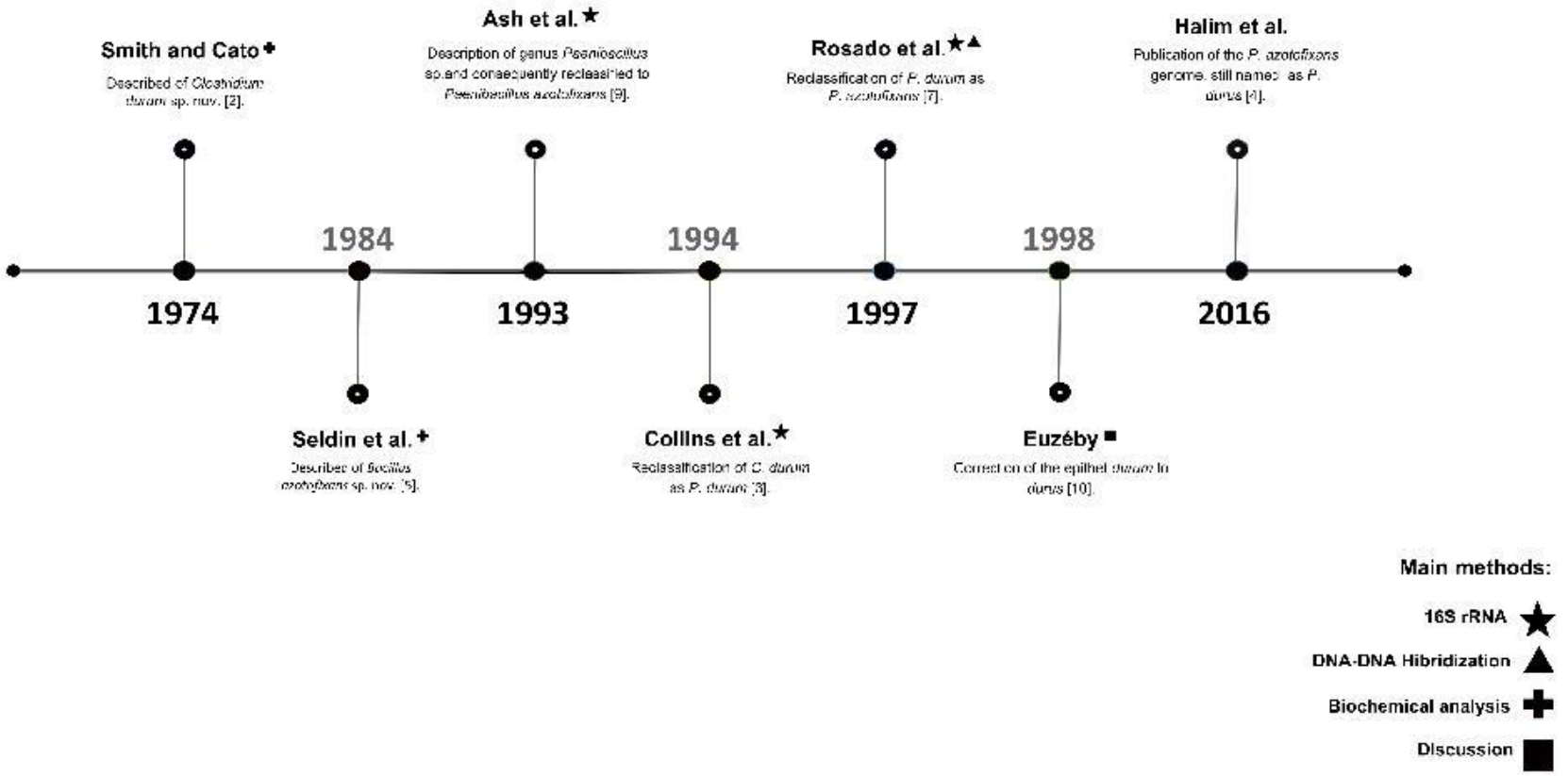


Figure 1: Classification and reclassification history of the current *Paenibacillus azotofixans* ATCC 35681^{T} and *Paenibacillus durus* DSM 1735^{T}. Main methods: 16S rRNA★, DNA-DNA Hibridization ▲, Biochemical analysis ✚, Discussion ■.

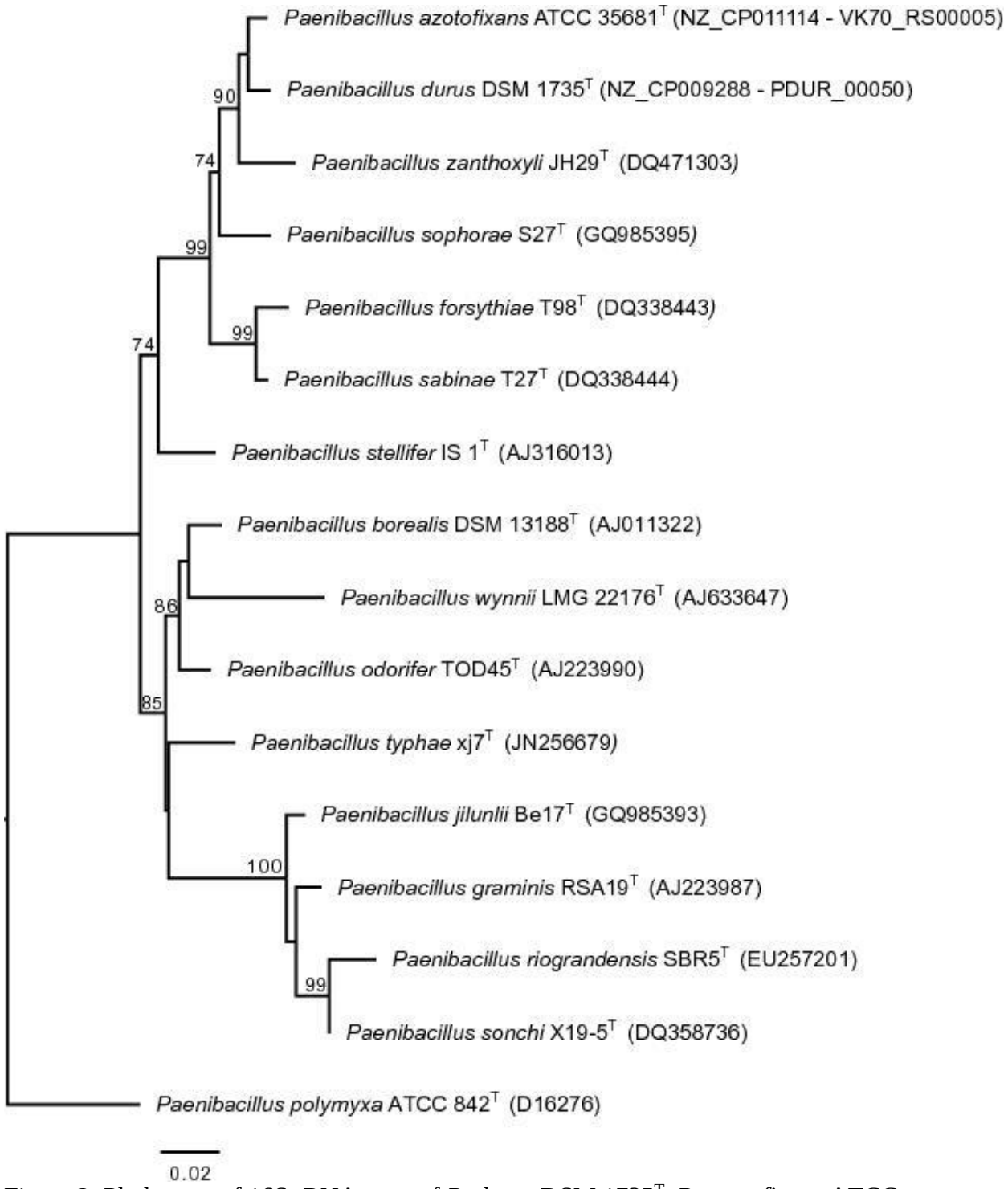


Figure 2: Phylogeny of 16S rRNA gene of *P. durus* DSM 1735^T, *P. azotofixans* ATCC 35681^T and their closest neighbors. The tree was built using the Neighbor-Joining method. Bootstrap values greater than 70 are shown next to the branches. The tree was drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. *Paenibacillus polymyxa* ATCC 842^T is the outgroup. Accession numbers of 16S rRNA gene sequences are between parentheses.

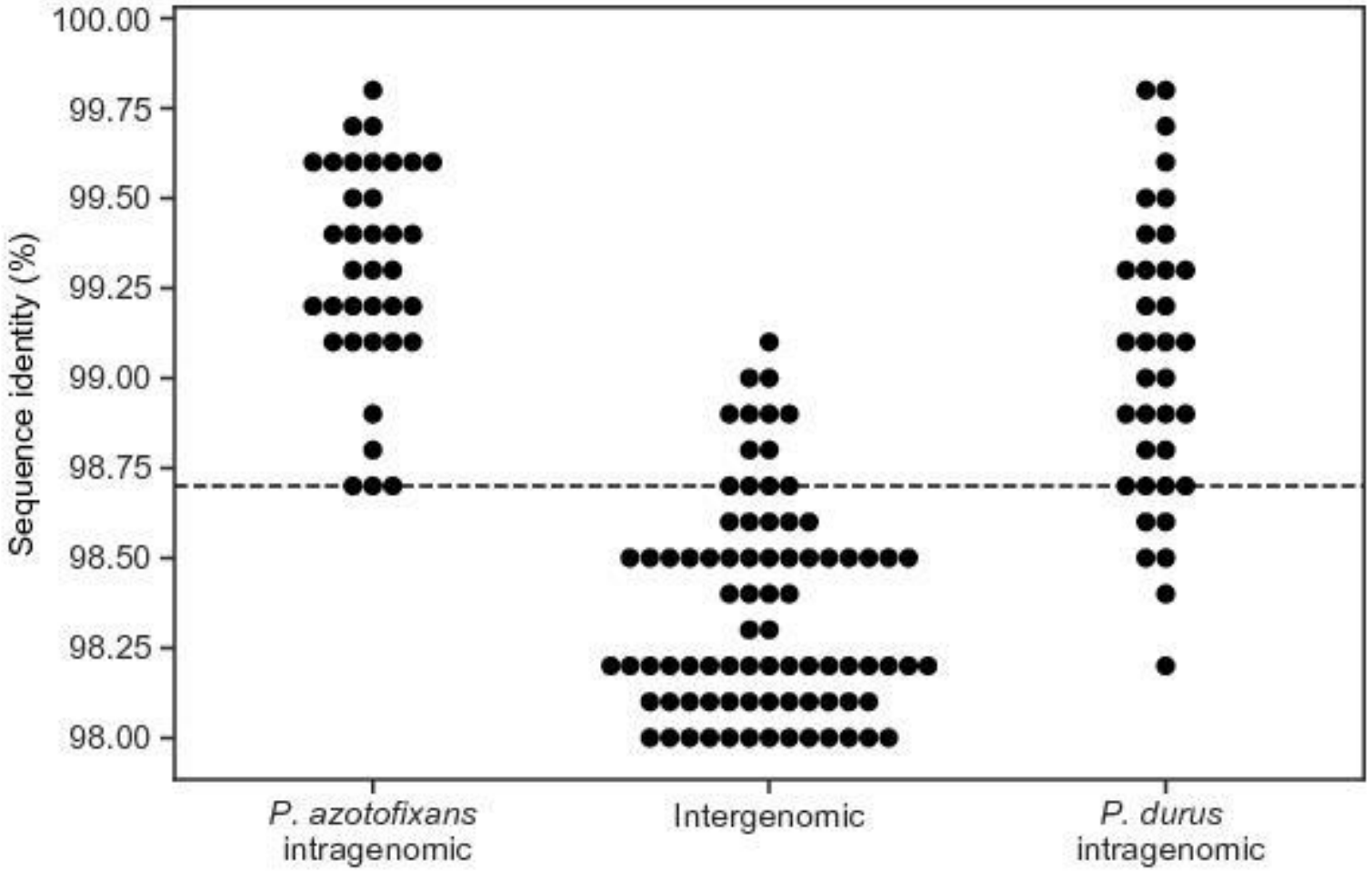


Figure 3: Identity scores of all 16S rRNA gene copies from *P. durus* DSM 1735^{T} and *P. azotofixans* ATCC 35681^{T}, both intra and intergenomic. Each point represents one identity score result.

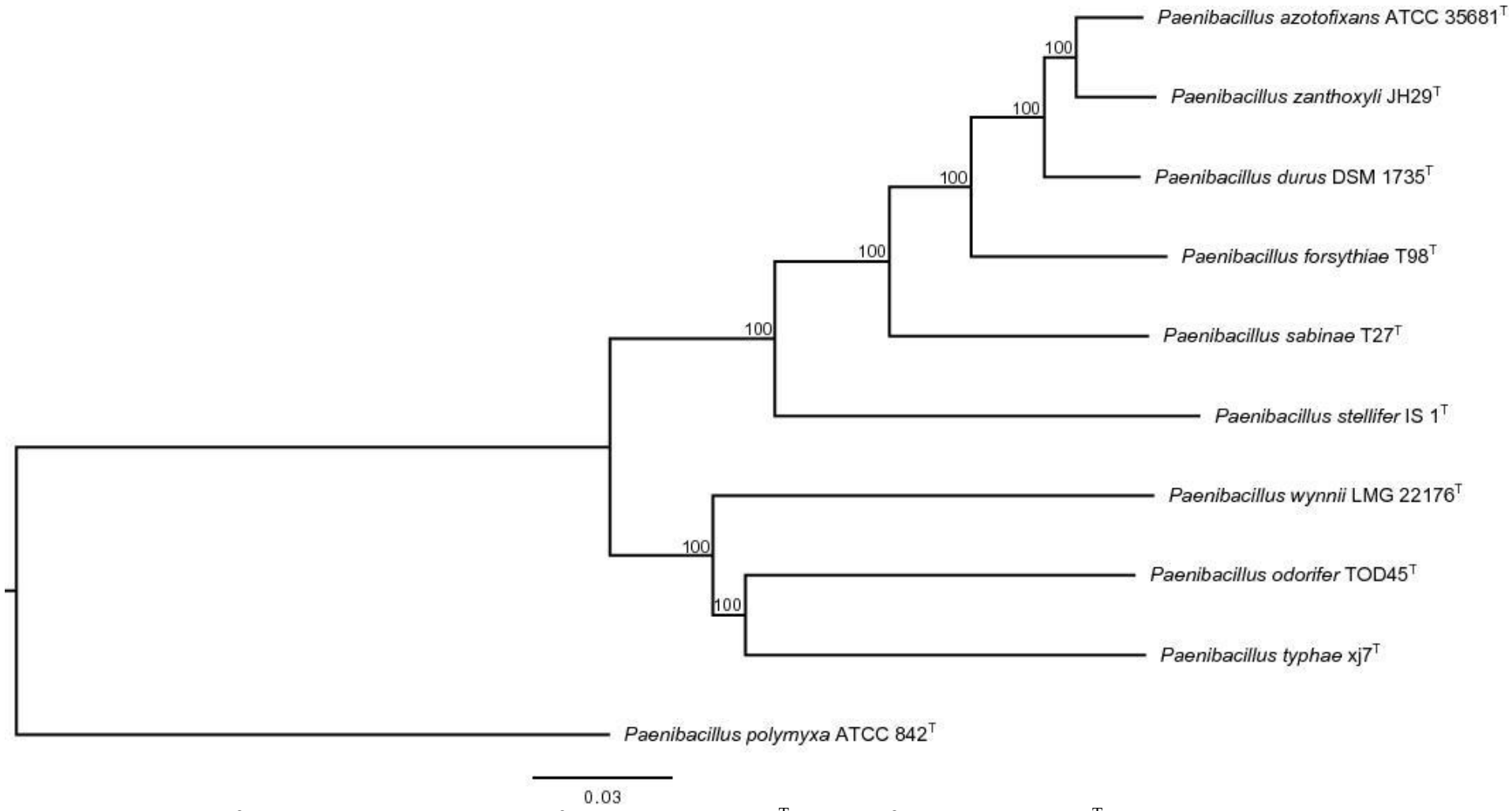
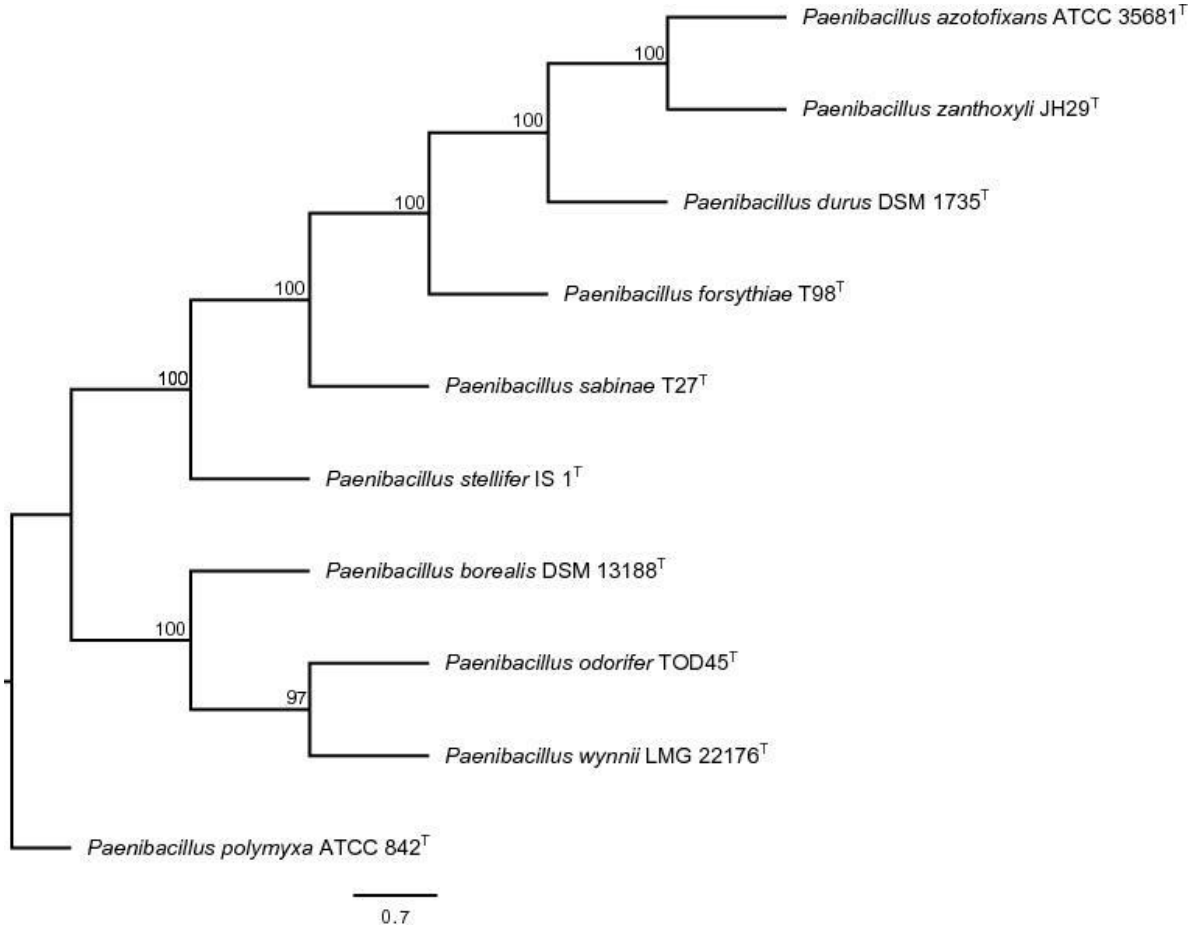


Figure 4: Phylogeny of concatenated core-proteome of *P. durus* DSM 1735^T, *P. azotofixans* ATCC 35681^{T}and their closest *Paenibacillus* type-strains. Details are as shown in Figure 2, unless specified otherwise.



Supplementary Figure S1: Phylogeny of housekeeping MLSA of *P. durus* DSM 1735^T, *P.azotofixans* ATCC 35681^{T}and their closest *Paenibacillus* type-strains. The tree was built using the Maximum Likelihood method. *Paenibacillus polymyxa* ATCC 842^T is the outgroup. Details are as shown in Figure 2, unless specified otherwise.

Table 1: Identity values of 16S rRNA genes of *Paenibacillus* type-strains in relation to their counterpart from *P. azotofixans* ATCC 35681^{T}.

Organism	Identity value (%)
Paenibacillus durus DSM 1735^{T}	**99.2**
Paenibacillus forsythia T98^{T}	96.9
Paenibacillus polymyxa ATCC 842^{T}	93.9
Paenibacillus sabinae T 27 ^{T}	97.0
Paenibacillus stellifer IS 1^{T}	97.4
Paenibacillus zanthoxyli JH29^{T}	98.3

Bold - Value above 98.7 of sequence identity. Accession numbers are provided in Figure 2.

Table 2: Genomes utilized in this study.

Organism	Accession number	Contig number	Size (Mbp)	Gene number	G+C content (%)
Paenibacillus azotofixans ATCC 35681^T	NZ_CP011114.1	1	5.57	5308	51.0
Paenibacillus durus DSM 1735^T	NZ_CP009288.1	1	6.03	5453	50.8
Paenibacillus forsythiae T98^T	NZ_ASSC00000000.1	896	5.08	5173	53.0
Paenibacillus odorifer TOD45^T	NZ_CP009428.1	1	6.81	5997	44.2
Paenibacillus polymyxa ATCC 842^T	NZ_AFOX00000000.1	13	5.90	5381	45.3
Paenibacillus riograndensis SBR5^T	NZ_LN831776.1	1	7.92	6793	51.0
Paenibacillus sabinae T27^T	NZ_CP004078.1	1	5.27	4904	52.7
Paenibacillus sophorae S27^T	NZ_FODH00000000.1	41	5.81	5610	51.2
Paenibacillus stellifer IS 1^T	NZ_CP009286.1	1	5.66	5232	53.5
Paenibacillus typhae xj7^T	NZ_FNDX00000000.1	92	6.74	5984	51.6
Paenibacillus wynnii LMG 22176^T	NZ_JQCR00000000.1	3	5.99	5575	44.9
Paenibacillus zanthoxyli JH29^T	NZ_ASSD00000000.1	714	5.04	5209	50.9

Table 3: Genomic metrics of *Paenibacillus* type strains in relation to *P. azotofixans* ATCC 35681^{T}.

Organism	ANIb (%)	gANI (%)	OrthoANI (%)	dDDH (%)	MUMi
Paenibacillus durus DSM 1735^{T}	92.19	92.99	92.60	47.90	0.542
Paenibacillus forsythia T98^{T}	87.22	88.24	87.67	33.50	0.755
Paenibacillus odorifer TOD45^{T}	74.15	74.21	71.83	20.00	0.982
Paenibacillus polymyxa ATCC 842^{T}	72.47	71.61	68.17	25.40	0.991
Paenibacillus riograndensis SBR5 ^{T}	75.93	75.69	73.55	20.80	0.970
Paenibacillus sabinae T27 ^{T}	84.32	85.22	84.72	28.70	0.814
Paenibacillus sophorae S27^{T}	91.15	91.97	91.32	44.00	0.599
Paenibacillus stellifer IS 1 ^{T}	77.96	78.27	76.72	22.70	0.946
Paenibacillus typhae xj7 ^{T}	75.29	75.31	73.40	19.80	0.972
Paenibacillus wynnii LMG 22176 ^{T}	74.11	74.17	71.93	19.10	0.981
Paenibacillus zanthoxyli JH29^{T}	94.30	94.77	94.44	58.60	0.434

Supplementary Table S1: Locus tag for each housekeeping gene from each genome utilized.

Organism	*gyrB*	*recA*	*recN*	*rpoB*
Paenibacillus azotofixans ATCC 35681^T	VK70_22230	VK70_11695	VK70_13840	VK70_26005
Paenibacillus borealis DSM 13188^T	PBOR_00035	PBOR_21090	PBOR_24505	PBOR_32900
Paenibacillus durus DSM 1735^T	PDUR_00030	PDUR_16170	PDUR_18350	PDUR_24495
Paenibacillus forsythiae T98^T	L692_RS0121380	L692_RS0124190	L692_RS0124870	L692_RS0104555
Paenibacillus odorifer TOD45^T	PODO_RS00030	PODO_RS17645	PODO_21510	PODO_RS27065
Paenibacillus polymyxa ATCC 842^T	PPT_RS0127605	PPT_RS0110780	PPT_RS0117095	PPT_RS0124415
Paenibacillus sabinae T27^T	PSAB_RS00030	PSAB_RS14450	PSAB_16580	PSAB_RS16600
Paenibacillus stellifer IS 1^T	PSTEL_00030	PSTEL_16060	PSTEL_18190	PSTEL_24175
Paenibacillus wynnii LMG 22176^T	PWYN_11990	PWYN_22895	PWYN_25770	PWYN_15180
Paenibacillus zanthoxyli JH29^T	L691_RS0116235	L691_RS0109760	L691_RS0102340	L691_RS0125345

Supplementary Table S2: Identity values of housekeeping genes of *Paenibacillus* type strains in relation to those of *P. azotofixans* ATCC 35681^T.

Strain	*gyrB* (%)	*recA* (%)	*recN* (%)	*rpoB* (%)
Paenibacillus durus DSM 1735^T	95.9	95.2	95.1	94.1
Paenibacillus forsythiae T98^T	91.5	91.2	89.3	93.4
Paenibacillus polymyxa ATCC 842^T	74.9	76.0	67.9	79.8
Paenibacillus sabinae T27^T	89.9	87.0	83.7	92.7
Paenibacillus stellifer DSM 14472^T	83.0	84.0	78.4	87.6
Paenibacillus zanthoxyli JH29^T	96.9	98.0	97.4	96.7

Supplementary Table S3: Phenotypic characteristics of the type strains DSM 1735^T, ATCC 35681^T, and JH29^T.

Strains: 1, *Paenibacillus durus* DSM 1735^T; 2, *Paenibacillus azotofixans* ATCC 35681^T; 3, *Paenibacillus zanthoxyli* JH29^T.

Characteristic	1	2	3
Isolation source	black sea sediment *	wheat roots (*Triticum aestivum* L.) †	prickly-ash shrub rhizosphere (*Zanthoxylum simulans*) ¶
General phenotypic traits:			
Cell size (diameter x length)	0.8 – 1.5 x 3.0 – 15 µm *	1.0 x 3.0 – 6.0 µm †	0.35 – 0.4 x 4.0 – 4.8 µm ¶
Colony shape	entire margin, granular, gray, translucent, glossy surface, 0.5 to 1.5 mm in diameter *	entire to undulate margins, butyrous texture, whitish, circular to slightly irregular, convex, mucoid, opaque, 1.0 to 2.0 mm in diameter †	entire margins, circular, convex, glossy ¶
Endospore shape	oval, subterminal to terminal *	oval to ellipsoidal and predominantly central to subterminal †	oval or ellipsoidal, subterminal or central, swollen sporangia ¶
Motility	+ *	w †	+ ¶
pH range	5.0–10.0 &	nd	4.2–10.0 ¶
Growth at 10°C	nd	− †	+ ¶
Growth at 45°C	− ‖	− †, ‖	− ¶
Optimum temperature (°C)	30 *, $, @, &	30 ¶	30 ¶, $, @
Temperature range (°C)	15 – 40 &	15 to 20 – 37 to 40 †	4 – 37 ¶
Biochemical characteristics:			
General biochemical traits:			
Catalase	+ §	+ †, §	+ ¶
Dextrin production	+ $, @	− †, ¶, #	− ¶, #, $, @

Growth in lysozyme 0.001%	nd	− †, ¶, #	− ¶, #
Growth in NaCl 3% (w/v)	− $, @	w †	+ ¶, $, @
Indole production	− *, §	− §	nd
Methyl-red reaction	− &, + @	nd	+ ¶
Nitrate reduction	− ‖, $, @, + &	− †, ‖, + ¶	+ ¶, $, @
Nitrogen fixation	+ §	+ †	+ ¶
Oxidase	+ &	− †	− ¶
Urease	− *, §	− §	nd
Voges-Proskauer test	+ ‖, $, @, &	+ †, ‖, ¶, #	+ ¶, #, $, @
Acid production from:			
β-Gentiobiose	− &	+ ‡	nd
D-Arabinose	− ‖, &	− †, ‡, ‖	nd
D-Arabitol	− ‖, + &	− ‡, ‖	nd
D-Fructose	− $, @	+ ‡	− $, @
D-Fucose	− ‖, + &	− ‡, ‖	nd
D-Lyxose	− ‖	− ‡, ‖	nd
D-Sorbitol	+ $, @	− ¶, #	− ¶, #, $, @
D-Tagatose	− ‖, &	− ‡, ‖	nd
D-Turanose	w &	+ ‡	nd
D-Xylose	− ‖, + $, @, &	− ‡, ‖, ¶	− ¶, $, @
Galactose	+ @, &	+ †, ‡	nd
Gluconate	+ ‖	+ ‖	nd
Glucose	+ $, @	+ †, ‡, ¶, #	− ¶, #, $, @
Glycerol	− ‖, + $, @	− †, ‡, ‖, ¶, #	− ¶, #, $, @
Glycogen	+ &	− ‡	nd
Inositol	− ‖, @	− ‡, ‖	nd
Inulin	+ ‖, &	+ †, ‖	nd
Lactose	− ‖, + $, @	− †, ‡, ‖, ¶, #	− ¶, #, @
L-Aspartate	nd	+ ¶, #	− ¶, #

L-Fructose	nd	− ¶	− ¶
L-Fucose	− ‖, &	− ‖	nd
Maltose	+ $, @	+ †, ‡, ¶, #	− ¶, #, $, @
Melezitose	+ ‖, &	+ ‡, ‖	nd
Methyl α-D-Glucoside	+ ‖	+ ‡, ‖	nd
Methyl α-D-Mannoside	− ‖	− ‖	nd
Methyl β-D-Xyloside	− ‖	− ‡, ‖	nd
N-Acetylglucosamine	− ‖	− ‖	nd
Raffinose	+ &	+ †, ‡	nd
Rhamnose	− ‖, &	− ‡, ‖	nd
Ribose	− ‖, + &	− †, ‡, ‖	nd
Sodium citrate	− $, @	− ¶	− ¶, $
Sodium succinate	− $, @	− ¶, #	− ¶, #, $, @
Sorbitol	+ ‖	+ ‡, ‖	nd
Sucrose	+ $, @, &	− ¶, #, + †, ‡	− ¶, #, $, @
Xylitol	− ‖, &	− ‡, ‖	nd
Assimilation of:			
3-Hydroxybenzoate	− ‖	− ‖	nd
4-Hydroxybenzoate	− ‖	+ ‖	nd
Acetate	+ ‖	− ‖	nd
Acetoin	+ §	+ §	nd
Adonitol	− ‖	− ‖	nd
α-D-Melibiose	+ ‖	+ ‖	nd
β-Arbutin	− ‖	− ‖	nd
cis-Aconitate	− ‖	− ‖	nd
Citrate	− ‖	− †, ‖	nd
D-Cellobiose	− ‖	+ ‖	nd
D-Fructose	+ *, ‖	+ ‖	nd
D-Galactose	+ ‖	+ ‖	nd

D-Glucose	+ *, ‖, &	+ ‖	nd
D-Maltose	+ *, ‖, &	+ ‖	nd
D-Mannitol	− ‖, + &	+ ‖	nd
D-Mannose	+ *, ‖, &	+ ‖	nd
D-Ribose	− *, ‖	− ‖	nd
D-Sorbitol	+ ‖	− ‖	nd
D-Sucrose	+ *, ‖	+ ‖	nd
D-Trehalose	+ *, ‖	+ ‖	nd
D-Xylose	− ‖	− ‖	nd
Fumarate	− ‖	− ‖	nd
Gluconate	+ ‖	+ ‖	nd
I-Inositol	− *, ‖	− ‖	nd
L-Alanine	− ‖	− ‖	nd
L-Arabinose	− ‖, + &	− ‖	nd
L-Aspartate	− ‖	− ‖	nd
L-Leucine	− ‖	− ‖	nd
L-Malate	− ‖	− ‖	nd
L-Rhamnose	− *, ‖	− ‖	nd
Maltitol	+ ‖	+ ‖	nd
N-Acetyl-D-glucosamine	− ‖, &	− ‖	nd
Oxoglutarate	− ‖	− ‖	nd
Pyruvate	+ ‖	+ ‖	nd
Salicin	+ ‖	+ ‖	nd
Hydrolysis of:			
2-Deoxythymidine-5'-pNP-Phosphate	− ‖	− ‖	nd
Aesculin	+ ‖	+ ‖	nd
Bis-pNP-phosphate	− ‖	+ ‖	nd
Casein	− *, §, ‖, $, @, + &	− †, §, ‖, ¶, #	− ¶, #, $, @
Gelatin	− *, §, ‖, @, &	− †, §, ‖, ¶	− ¶

pNP-α-D-Glucopyranoside	+ ‖	− ‖	nd
pNP-β-D-Galactopyranoside	− ‖	− ‖	nd
pNP-β-D-Glucopyranoside	+ ‖	+ ‖	nd
pNP-β-D-Glucuronide	− ‖	− ‖	nd
pNP-Phenyl-phosphonate	− ‖	− ‖	nd
pNP-Phosphoryl-choline	− ‖	− ‖	nd
Pectin	− ‖	− †, ‖	nd
Starch	− ‖, + $, @, &	− †, ‡, ‖, ¶	− ¶, $, @

+, positive. −, negative. w, weakly positive. nd, not determined. pNp: *para*-nitrophenyl. Bold, differential characteristics between *P. durus* DSM 1735T and *P. azotofixans* ATCC 35681T. Underlined, conflicting data. * Smith and Cato (1974)(18); † Seldin et al. (1984)(19); ‡ Seldin and Penido (1986)(41); § Rosado et al. (1997)(21); ‖ Elo et al. (2001)(42); ¶ Ma et al. (2007)(43); # Ma and Chen (2008)(44); $ Jin et al. (2011)(45); @ Xie et al. (2012)(46); & Kong et al. (2013)(39).

Supplementary Table S4: Chemotaxonomic profiles of the type strains DSM 1735^T, ATCC 35681^T, and JH29^T.

Strains: 1, *Paenibacillus durus* DSM 1735^T; 2, *Paenibacillus azotofixans* ATCC 35681^T; 3, *Paenibacillus zanthoxyli* JH29^T.

Fatty acid composition (%)	1	2	3
Saturated			
Straight-chain			
C14:0	4.6 *, 4.76 §, ‖, TR ¶	7.4 *, 5.0 †	4.87 ‡, 4.84 §
C15:0	7.7 *, 1.33 §	3.4 *, 2.2 †	2.07 ‡, 2.15 §
C16:0	17.4 *, 19.87 §, 31.59 ‖, 3.4 ¶	20.5 *, 15.5 †	12.83 ‡, 12.88 §
C16:0 *N* alcohol	nd	0.4 †	nd
C18:0	5.48 §, 3.08 ‖	0.3 †	6.74 ‡, 6.77 §
Iso-branched			
C13:0	< 0.2 *	< 0.2 *, 1.3 †	nd
C14:0	5.4 *, 5.31 §, 4.06 ‖, 4.7 ¶	6.9 *, 4.7 †	3.99 ‡, 3.79 §
C15:0	2.9 *, 7.98 §, 7.89 ‖, 16.6 ¶	6.4 *, 8.7 †	1.51 ‡, 1.63 §
C16:0	8.8 *, 10.58 §, ‖, 2.9 ¶	8.0 *, 5.3 †	14.85 ‡, 14.9 §
C17:0	0.9 *, 1.48 §, ‖, 1.9 ¶	0.8 *, 1.1 †	0.61 ‡, 0.59 §
C17:0 3OH	nd	2.9 †	nd
C18:0 H*	nd	0.8 †	nd
Anteiso-branched			
C13:0	0.8 *, TR ¶	1.0 *, 1.8 †	0.49 ‡
C15:0	47.6 *, 39.76 §, ‖, 50.9 ¶	42.4 *, 45.4 †	32.19 ‡, 32.25 §
C17:0	2.7 *, 1.94 §, ‖, 2.8 ¶	2.0 *, 2.1 †	3.89 ‡, 3.87 §
Unsaturated			
C16:1 ω7*c* alcohol	3.7 ¶	nd	nd
C16:1 ω11*c*	< 0.2 *, 4.5 ¶	< 0.2 *	nd
C18:1 ω5*c*	nd	1.8 †	2.02 ‡
C18:1 ω7*c*	nd	nd	1.78 ‡
C18:1 ω9*c*	< 0.2 *, 2.12 §, ‖	< 0.2 *	1.51 ‡, 1.54 §
iso-C15:1 ω9c	1.2 ¶	nd	nd

iso-C17:1 ω10c	2.2 ¶	nd	nd

nd, not determined; TR, Trace (< 0.5 %). * Elo et al. (2001)(42) ;† Yoon et al. (40)(2003); ‡ Ma et al. (2007)(43); § Jin et al. (2011) (45); || Xie et al. (2012)(46); ¶ Kong et al. (2013)(39).

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