

Universidade Federal de Minas Gerais

Instituto de Ciências Biológicas

Programa Interunidades de Pós-graduação em Bioinformática

Método filogenômico para inferência de árvore de espécie de organismos procariotos: uma nova abordagem filogenética para reconciliação de genes transferidos horizontalmente

Autor: Nilson Da Rocha Coimbra

Orientação: Dr. Aristóteles Góes Neto (UFMG)

Dra. Aïda Ouangraoua (Université de Sherbrooke – Canadá)

NILSON ANTONIO DA ROCHA COIMBRA

Método filogenômico para inferência de árvore de espécie de organismos procariotos: uma nova abordagem filogenética para reconciliação de genes transferidos horizontalmente

Tese de Doutorado apresentada ao Programa Interunidades de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais, para obtenção do título de Doutor em Bioinformática.

Orientação: Dr. Aristóteles Góes Neto
Dra. Aïda Ouangraoua

043

Coimbra, Nilson Antonio da Rocha.

Método filogenômico para inferência de árvore de espécie de organismos procariotos: uma nova abordagem filogenética para reconciliação de genes transferidos horizontalmente [manuscrito] / Nilson Antonio da Rocha Coimbra. - 2019.

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

Orientação: Dr. Aristóteles Góes Neto. Dra. Aïda Ouangraoua.

Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Filogenia. 2. Transferência Genética Horizontal. 3. Corynebacterium. 4. Actinobacteria. I. Góes Neto, Aristóteles. II. Ouangraoua, Aïda. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

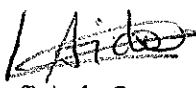


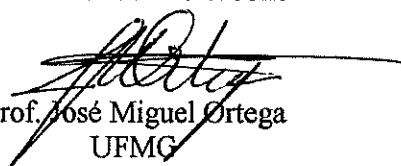
"Método filogenômico para inferência de árvore de espécie de organismos procariotos: uma nova abordagem filogenética para reconciliação de genes transferidos horizontalmente"

Nilson Antonio da Rocha Coimbra

Tese aprovada pela banca examinadora constituída pelos Professores:

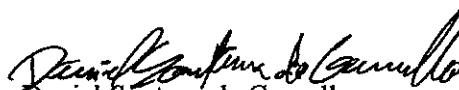

Prof. Aristóteles Góes Neto - Orientador
UFMG


Prof^a Aída Ouangraoua
Université de Sherbrooke


Prof. José Miguel Ortega
UFMG


Prof. Tetsu Sakamoto
UFRN


Prof^a Flavia Figueira Aburjaile
UFPE


Daniel Santana de Carvalho
UFMG


Prof^a Anne Cybelle Pinto
UFMG

Belo Horizonte, 10 de dezembro de 2019.

ABSTRACT

Classically, concatenated alignments of the small subunits of ribosomal RNA sequences (rDNA's) have been used to reconstruct bacteria phylogenies and identify new species. On the one hand and in the light of evolution, the genomic content of microorganisms is largely affected by Horizontal Gene Transfers - HGT, whereas, on the other hand, the reconstruction of microbial phylogenies using the classical sequence-based approach does not account for the presence of Horizontally transferred genes. In this work, we improve the methods of microbial phylogeny reconstruction, while accounting for the presence of HGT genes. We present a new gene tree-based method to correct putative transferred genes and applied this new method in the phylogenomic reconstruction of the Order *Corynebacteriales*, the largest clade in the Phylum *Actinobacteria*. We collected 360 genome records of *Corynebacteriales* from NCBI RefSeq Database, release 81 and estimate the phylogeny reconstruction using both concatenated-based (RaxML) and summary-based (ASTRAL and ASTRID) methods. Overall, all the reconstructed trees exhibited a highly resolved resolution and feasibility for the microbial classification, detecting speciation among the *Corynebacterium* genus.

KEYWORDS: phylogeny reconstruction; horizontal gene transfer; *Corynebacteriales*; *Actinobacteria*.

RESUMO

Classicamente, os alinhamentos concatenados das sequências da subunidade menor (16S) de RNA ribossômico (rDNAs) têm sido usados para reconstruir filogenias bacterianas e identificar novas espécies. Por um lado e à luz da evolução, o conteúdo genômico dos microrganismos é amplamente afetado pelas transferências gênicas horizontais, enquanto que, por outro lado, a reconstrução de filogenias microbianas usando a abordagem clássica baseada em sequências não leva em conta a presença de genes horizontalmente (ou lateralmente) transferidos. Neste trabalho, objetivamos melhorar os métodos de reconstrução de filogenia microbiana, considerando a presença de genes provavelmente transferidos lateralmente. Apresentamos um novo método baseado em árvore para corrigir genes putativos transferidos e aplicamos este novo método na reconstrução da Ordem *Corynebacteriales*, o maior clado do filo *Actinobacteria*. Foram coletados 360 genomas completos de espécies de *Corynebacteriales* do NCBI RefSeq Database, versão 81. A reconstrução filogenética foi realizada utilizando métodos baseados em árvores de genes (ASTRAL e ASTRID) e baseados em sequências (RaXML). No geral, todas as árvores reconstruídas exibiram alta resolução, e também, viabilidade para uso taxonômico, detectando, inclusive, o fenômeno de especiação dentro do gênero *Corynebacterium*.

Palavras-chave: reconstrução filogenômica; transferência horizontal de genes; *Corynebacteriales*; *Actinobacteria*.

LISTA DE FIGURAS

Figura 1: Representação de desenho de uma reconciliação12

Capítulo 1

FIG. 1. Overview of the method, which consists of 6 steps.....21

FIG. 2. Support values of internal branches of the ASTRID, ASTRAL, RAxML and Overall consensus trees26

FIG. 3. Illustration at the genus level of the overall consensus phylogeny reconstructed in our work.28

FIG. 4. Partition of *Corynebacterium* into two categories: non-pathogenic in yellow and pathogenic in orange.30

FIG. 5. Partition of *Corynebacterium pseudotuberculosis* into two biovars. Biovars *equi* and *ovis* are shown in green and blue, respectively.32

FIG. 6. Partition of *Mycobacterium* into two categories: slow growers in purple and fast-growers in pale pink.....33

Supplementary figures

FIG. S1. *Corynebacteriales* preliminary species tree estimated using the concatenation of multiple sequence alignments of 13 putative orthology groups and RAxML (Stamatakis, 2014).....54

FIG. S2. Illustration, at the genus level, of the 4 consensus trees estimated in our work.....55

FIG. S3. Classification of *Brevibacterium* inside *Corynebacterium glutamicum*.....56

FIG. S4. Estimated phylogeny for *Rhodococcus* genus, with 6 clusters.....56

FIG. S5.. Estimated phylogeny for *Nocardia* genus.56

FIG. S6. Estimated phylogeny for *Gordonia* genus.....56

FIG. S7. Illustration of the phylogenetic method used in Step 5 to reclassify some putative transferred genes as vertically inherited genes57

LISTA DE TABELAS

Capítulo 1

Table 1. Input dataset for Corynebacteriales phylogenetic tree estimation.....	22
Table 2. Square matrix of percentage of conserved clades between phylogenies estimated using RAxML (Step 4), and ASTRID and ASTRAL (Step 5).....	24
Table 3. Square matrix of the percentage of conserved clades between the consensus trees and the preliminary RAxML tree.....	27

Supplementary table

Table S1. Putative orthology groups used in Step 4 for computing the preliminary species tree	52
Table S2. Number of gene trees in the 5 collections of trees considered in Step 5.....	52
Table S3. Percentage of conserved clades between the consensus trees and the initial phylogenies reconstructed using ASTRID and ASTRAL.....	53
Supplementary File A1 – Number of genomes per species	58
Supplementary File A2 – Number of genomes per species	61
Supplementary File A3 – Number of Genomic Islands (GI) and Putative Transferred Genes (PTG) per genome	72
Supplementary File A4 – Statistics of homology groups	80

LISTA DE ABREVIACOES

HGT - *Horizontal gene transfer*

SUMÁRIO

ABSTRACT	I
RESUMO.....	II
LISTA DE FIGURAS.....	III
LISTA DE TABELAS	IV
LISTA DE ABREVIACOES.....	V
SUMRIO	1
APRESENTACO	2
ESTRUTURA DA TESE.....	3
INTRODUO	4
Mtodos para deteco de genes transferidos horizontalmente.....	7
Reconstruo filogentica baseada em genes horizontalmente transferidos.....	9
Reconciliao de rvores filogenticas.....	11
JUSTIFICATIVA.....	13
OBJETIVO GERAL.....	14
Objetivos especficos.....	15
Captulo 1 – <i>Reconstructing the phylogeny of Corynebacteriales while accounting for horizontal gene transfer</i>	16
DISCUSSO GERAL	81
CONCLUSO	83
PERSPECTIVAS.....	84
REFERNCIAS.....	85
APNDICES	91
APNDICE I – Comprovante do aceite para a publicao do artigo ao perdico <i>Genome Biology and Evolution</i>	93
APNDICE II – Cdigo Fonte: Easy Xenology Conciliation Tool.....	94
APNDICE III – Editorial: Second ISCB Latin American Student Council Symposium (LA-SCS) 2016	112
APNDICE VI – Editorial: Nurturing tomorrow’s leaders: The ISCB Student Council Symposia in 2018	120
APNDICE V – Editorial: Global network of computational biology communities: ISCB's Regional Student Groups breaking barriers.....	129

APRESENTAÇÃO

Este trabalho foi desenvolvido em colaboração com a *Université de Sherbrooke*, localizada em Sherbrooke, na província do Québec, no Canadá. A pesquisa apresentada neste documento foi desenvolvida sob a orientação da Dra. Aïda Ouangraoua (*Canada Research Chair* em Bioinformática Complexa e Biologia Computacional e professora adjunta da Universidade de Sherbrooke) e do Dr. Aristóteles Góes Neto (UFMG). Os cálculos computacionais realizados para a execução desta tese foram realizados em dois supercomputadores: *Mammoth* e *Graham*, ambos localizados na *Université de Sherbrooke* e *University of Waterloo*, respectivamente. O supercomputador *Mammoth* é gerenciado através das iniciativas *Calcul Québec* e *Compute Canada*. A operação do *Mammoth* é financiada pelo *Canada Foundation for Innovation* (CFI), pelo *Ministère de l'Économie, de la science et de l'innovation du Québec* (MESI) e também, através do *Fonds de Recherche du Québec – Nature et Technologies* (FRQ-NT). O desenvolvimento deste trabalho foi financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG BDSE-00401-16), pelo programa *Canada Research Chair in Computational and Biological Complexity* e também, pela *Université de Sherbrooke*.

ESTRUTURA DA TESE

Esta tese apresenta uma introdução (abordando os principais tópicos do tema proposto), a justificativa do trabalho, e os objetivos gerais e específicos. Em seguida, o artigo produzido será apresentado na forma de um capítulo único.

Capítulo 1: *Research Paper - "Reconstructing the phylogeny of Corynebacteriales while accounting for Horizontal Gene Transfer."*

Para finalizar, é apresentado uma breve discussão geral, a conclusão geral, e as perspectivas, seguidas das referências citadas no corpo do texto e os APÊNDICES, que acompanham esta tese.

INTRODUÇÃO

O século XX foi marcado por grandes revoluções tecnológicas, sendo a palavra “sequenciamento” de DNA inversamente proporcional ao seu entendimento. Ainda assim, foi devido ao alcance do marco do sequenciamento completo do genoma humano (porém não somente a ele), que houve o surgimento do termo ômicas, referindo a totalidade de informações extraídas a partir da extrapolação da informação presente no DNA nos organismos vivos (VENTER et al., 2001; DALY et al., 2001; DAOJING E BODOVITZ, 2010; BONNIE, PENG E SINGH, 2013).

Embora a velocidade e o volume de informações geradas pelas ômicas revolucionaram como interpretamos o código genético dos organismos, o tempo necessário para sua análise e compreensão é inversamente proporcional ao seu processamento (STÄHLER et al., 2006; TITMUS GURTOWSKI e SCHATZ, 2014). Dessa forma, novos campos de estudos surgiram com o estudo integrado de dados, combinando as informações geradas nas ômicas (i.e Biologia de sistemas, biologia sintética) (BENNER e SISMOUR, 2005). Assim, com a revolução das tecnologias de sequenciamento em larga escala, os dados gerados por elas, e as técnicas de processamento computacional para exploração dos dados; na década corrente, começamos a interpretar os alpes ômicos. Se o projeto de um genoma gera uma “montanha” de dados, o conjunto total de informações extraídas a partir do sequenciamento de todos os seres vivos, formam um complexo conglomerado de dados, atribuído neste documento, como os “alpes ômicos”. E, somente agora, no início do século XXI, que começaram a ser escalados.

Na ciência clássica, o estudo sistemático da evolução de espécies, no final dos anos 90, vislumbrou nos alpes ômicos, a quantidade de informações que poderiam ser utilizados/reutilizados para inferir a evolução das espécies de forma mais robusta, revisando os estudos já publicados e introduzindo novas metodologias e termos. Assim surgiu a Filogenômica, uma área do conhecimento que une filogenia/evolução com os dados de genômica.

Inicialmente, a Filogenômica é apresentada como um método para inferir a função de genes (EISEN, 1998). Entretanto, a área passou a ser revisitada com o aumento significativo de sequências de genomas completos depositado em banco de dados, e atualmente, o termo Filogenômica está associado à compreensão e à estimativa da história evolutiva de espécies utilizando a informação do conteúdo total de genes dentro de sequências de genomas completos (RABIEE, SAYYARI e MIRARAB, 2019).

Em Filogenômica, o ponto inicial observável se dá pela identificação de características compartilhadas entre as espécies (*i.e.*, genes ortólogos). Subsequentemente, se inferem distâncias evolutivas entre tais características, e, posteriormente, pelo agrupamento das menores distâncias, são formados os grupos-irmãos e então, feita a “reconstrução” da história evolutiva de espécies, genes e/ou funções biológicas (DESLUC, 2005).

Esse método é muito explorado para inferir o posicionamento taxonômico de novas espécies e explorar a diversidade de organismos em estudos de metagenômica. Entretanto, a diferença dos mecanismos de reprodução e aquisição de novos genes entre os procariotos e eucariotos, apresentam desafios e soluções diferentes na aplicação da

abordagem filogenômica para a inferência e a reconstrução evolutiva de espécies. Isso pois, a maioria dos métodos existentes disponíveis, principalmente para a inferência de árvore de espécies de procariotos, assume a evolução do sinal filogenético presente nos genes de forma vertical, com a informação genética (*gene flow*) passando da espécie ancestral, ignorando a quantidade de informação acumulada nas espécies modernas, adquiridas por vias de transferência horizontal (HACKER et al., 1997; DAVISON, 1999) .

Um dos grandes desafios presentes em estudos filogenômicos está na reconstrução evolutiva dos domínios *Bacteria* e *Archea* (BRINKMANN e PHILIPPE, 1999), pois, diferentemente dos organismos eucarióticos, nos quais a reprodução sexual é a principal forma de recombinação genética, nos organismos procarióticos, a principal forma de troca de material genético é mediada por mecanismos de transferência horizontal de genes (do inglês, *Horizontal Gene Transfer* - HGT) (GOGARTEN, PETER e TOWNSEND, 2005). Em procariotos, a aquisição de genes via HGT é mediada através de processos como: conjugação (transferência pelo contato direto entre organismos *i.e.* sistema de secreções tipo I, II, III e IV), transformação (aquisição de material genético do ambiente) e/ou transdução (infecção via bacteriófagos, que movem genes de uma célula para outra) (OCHMAN et al., 2000). Esses genes, uma vez transferidos, podem continuar a evoluir e, através de seleção positiva e recombinação homóloga, podem garantir alguma vantagem seletiva para a organismo receptor (GOGARTEN, 2005). Essa forma de adquirir novos genes, via transferência horizontal, amalgama o sinal filogenético durante a reconstrução evolutiva de procariotos, invalidando o uso de abordagens tradicionais como super-matrizes e super-alinhamentos (SOUCY *et al.*, 2015).

Há mais de 30 anos se utiliza o gene biomarcador que codifica a subunidade 16S rDNA para classificar e inferir a taxonomia de novos organismos (DAUBIN, 2002). Tal abordagem é proeminente na classificação taxonômica de espécies não relatadas, porém a baixa taxa de mutação dentro da sequência do 16S rDNA dificulta a identificação de espécies relativamente próximas (causadas por fenômenos de especiação), principalmente em organismos procariotos, sendo necessário, então, o uso de abordagens mais amplas, que visem à utilização de um conjunto maior de dados (genes) e de novos métodos para a determinação taxonômica de novas espécies (*Multilocus Sequence Analysis* - MLSA) (GLAESER and KÄMPFER, 2015).

Métodos para detecção de genes transferidos horizontalmente

Identificar e determinar genes transferidos horizontalmente é uma tarefa nada trivial (GALPERIN e KOONIN, 2000; SOREK et al., 2007). A diversidade genômica bacteriana é moldada através de processos balanceados entre a aquisição e perdas de genes, transferidos ou não, e, mediada por seleção positiva e/ou negativa. Bactérias detêm um aparato genômico relativamente pequeno, (quando comparadas com organismos eucariotos) e possuem uma maquinaria metabólica extremamente especializada (PARKINSON, 1993). Logo, a aquisição de novos genes, além de conferir determinada vantagem seletiva para que elas possam sobreviver no ambiente onde habitam, também está relacionada ao tempo de fixação desses genes adquiridos nos seus genomas (CHU, SPROUFFSKE E ANDREAS, 2018).

Como comentado anteriormente, genes recém-transferidos podem e são afetados pelos mesmos mecanismos que os transferiram, aumentando a complexidade de detecção do sinal filogenético e, conseqüentemente, da reconstrução filogenômica de bactérias (SOUCY *et al.*, 2015). Dado o fato que não há como realizar a confirmação *a posteriori* da transferência de genes ancestrais, a literatura descreve duas formas para identificar genes que provavelmente foram adquiridos via transferência horizontal: os métodos filogenéticos, que utilizam a discordância e conflitos entre árvores de genes e de espécies para inferir o evento de transferência, e, os métodos paramétricos, que utilizam assinaturas ao longo da sequência cromossômica para inferir regiões adquiridas via HGT (RAVENHALL *et al.*, 2015).

Os métodos paramétricos procuram por regiões ao longo do genoma (*i.e.*, ilhas genômicas) que divergem significativamente da média genômica, como por exemplo: as variações no conteúdo G+C, frequência da utilização e preferência por códons específicos, presença de genes relacionados a mobilidade, entre outros (LANGILLE *et al.*, 2010). Já os métodos filogenéticos usam a informação presente na árvore de espécie para detectar o sinal filogenético na árvore de genes, através de uma abordagem denominada de conciliação/reconciliação (DOYON *et al.*, 2011). Métodos filogenéticos são mais robustos que os métodos paramétricos, pois é possível identificar/confirmar pontos incongruentes presente nas árvores de genes, com base na árvore de espécies (PAGE e CHARLESTON, 1998; LUAY 2013) e, dessa forma, são exploradas relações evolutivas causadas por especiação (ortólogos), duplicação (parálogos) e transferências horizontais (xenólogos) (PHILIPPE e DOUADY, 2003; DALQUEN *et al.*, 2013).

Os métodos paramétricos, por se valerem apenas de características presentes na sequência do genoma, podem detectar um alto número de falsos-positivos (BECQ, CHURLAUD e DESCHAVANNE, 2010), e são homologos-dependentes, limitados pelo fato de necessitar da presença do número correto de ortólogos (o que não é o caso em genomas recentemente sequenciados). Para contornar esses vieses, podem ser utilizadas estratégias combinadas com métodos filogenéticos (principal objeto de estudo do presente trabalho). A combinação dessas duas famílias de métodos se mostra uma abordagem promissora para inferir árvores de referência de procariotos, com maior grau de resolução, quanto à disrupção do sinal filogenético, principalmente causado por genes transferidos horizontalmente.

Reconstrução filogenética baseada em genes horizontalmente transferidos

No século anterior com o vislumbre dos alpes ônicos, os estudos evolutivos se baseavam na identificação de características e particularidades compartilhadas entre os organismos através de estudos morfológicos, fisiológicos, ecológicos e comportamentais (DE QUEIROZ E DONOGHUE, 1988; LUDWIG E KLENK, 2005). Porém, durante a escalada dos alpes ônicos, foi possível detectar cada região genômica que codifica uma determinada característica fenotípica dos estudos supracitados e o sinal filogenético conservado presente na sequência do gene, e utilizá-las para inferir a história evolutiva com maior resolução. Logo, nos dias atuais, a inferência evolutiva esta diretamente relacionada à qualidade do sequenciamento dos genes marcadores e/ou do genoma que foi sequenciado (RANGEL E FURNIER, 2019).

A identificação da relação de homologia é um dos grandes gargalos em filogenômica, pois é diretamente dependente do número de genes presente nas espécies de estudo. Este fato também está relacionado à qualidade dos métodos para a inferência de homólogos diante da limitação de recursos computacionais disponíveis (GLOVER et al., 2019; SONNHAMMER et al., 2014). Por definição, homólogos são genes que descendem de um mesmo ancestral, e por divergência evolutiva, podem ser classificados em: ortólogos, parálogos e xenólogos (FITCH, 2000). Em termos gerais, ortólogos são genes que divergiram após um evento de especiação, enquanto que parálogos, são cópias de um gene que divergiu após um evento de especiação e genes xenólogos são genes que foram transferidos de uma espécie para outra, podendo ou não, ser produto de um evento de duplicação ou de especiação.

Em termos de filogenômica, o conjunto de dados de ortólogos é comumente utilizado para reconstruir a história evolutiva de determinado táxon. É comum também encontrar estudos concentrados em genômica comparativa que se referem ao conjunto de genes ortólogos do genoma central (*core genome*), e estimar a diversidade de organismos através da inferência filogenética pela concatenação a partir desse conjunto de genes. Em genômica comparativa, parálogos e xenólogos são classificados como genoma acessório (MEDINI et al., 2005; TETTELIN et al., 2008).

Reconciliação de árvores filogenéticas

As árvores filogenéticas ilustram a história evolutiva de genes e espécies (Figura 1). Entretanto, já é bem estabelecido que existem discordâncias entre as árvores de genes e as árvore de espécies. A informação discordante existente quando da sobreposição das topologias das árvores de genes, com a árvore de espécies, é o método mais robusto para se determinar eventos evolutivos como duplicação, perdas, ganhos e transferências. Essa sobreposição de topologias é conhecida como método de reconciliação de árvores filogenéticas, um amplo campo de estudo iniciado em meados dos anos 70 (GOODMAN et al., 1979), e também o principal método explorado neste trabalho para confirmar possíveis genes transferidos.

Métodos de reconciliação são bem utilizados, principalmente na identificação da relação de homologia entre genes e podem ser encontrados nas principais ferramentas de genômica comparativa como o Orthofinder (EMMS e KELLY, 2019; EMMS e KELLY, 2015). Identificar eventos de especiação e duplicação utilizando o método de reconciliação é bem definido pela comunidade de biologia computacional (EULENSTEIN, 1998; BININDA-EMONDS, 2004; BONIZZONI, DELLA VEDOVA e DONDI, 2005; ÅKERBORG et al., 2009; EULENSTEIN, HUZURBAZAR e LIBERLES, 2010). Entretanto, a aplicação do método de reconciliação, apesar de já ter sido descrita e matematicamente modelada em estudos anteriores (DARBY et al., 2017, ALTENHOFF, 2019), ainda sim, é pouco explorada diante da complexidade computacional demandada.

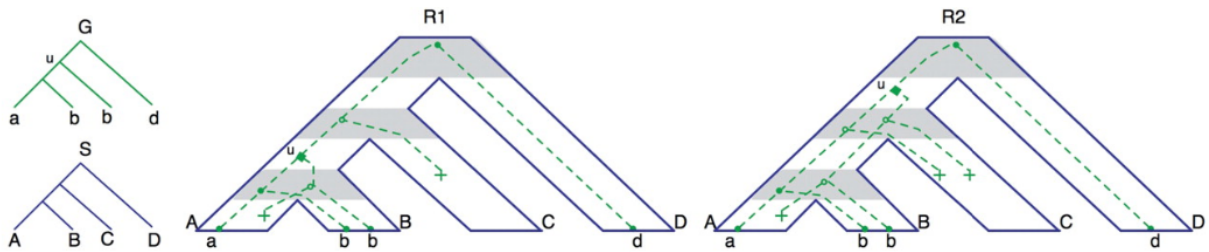


Figura 1: Representação de desenho de uma reconciliação. Uma árvore de espécies S e uma árvore de genes G, em que cada letra minúscula indica um gene observado de uma espécie existente (o gene 'a' pertence à espécie 'A', etc). R1 e R2 são duas reconciliações que incorporam G em S. Uma zona acinzentada (resp. Tubo) corresponde a um vértice (resp. Ramificação) de S e G é representada com linhas pontilhadas. Os nós da árvore incorporada representam: duplicação (losango), perda (+), especiação presente em G (círculo preenchido) ou não (círculo aberto). Observe que o nó u de G é uma duplicação para R1 e R2, embora localizado em diferentes ramos de S.

Fonte: Doyon *et. al.*, (2011).

Computacionalmente, os métodos de reconciliação são NP-difícil e demandam altos recursos de infraestrutura (memória e processamento) e suporte para se encontrar a solução ótima. Além disso, métodos de reconciliação dependem exclusivamente de uma árvore de espécies bem delimitada, tornando a sua usabilidade um grande desafio, diante da baixa coleção de árvores filogenéticas de referência, comparadas com o número de espécies já caracterizadas nos dias atuais.

Na mais recente literatura da área, Darby e colaboradores (DARBY, 2017) utilizaram a abordagem de reconciliação de árvores filogenéticas para a caracterização de divergência causadas por genes horizontalmente transferidos. Os autores também descrevem regras formais para cada associação de pares de genes, as classes de xenólogos, podendo ou não, estar associados a eventos de duplicação e de especiação.

JUSTIFICATIVA

A Filogenômica é um campo de estudo recente, tendo seu início no final dos anos 90 e ainda assim, pouco explorado, que utiliza a totalidade de informações produzidas nas ômicas para inferir a relação evolutiva entre espécies. Na evolução dos procariotos, a principal característica de herança genética adquirida é mediada através de processos de transferência horizontal, sendo um fato principal a ser considerado para a inferência da história evolutiva das espécies. Genes adquiridos por transferência horizontal, quando presentes, em conjuntos de prováveis ortólogos, causam ruído à informação filogenética enviesando a inferência da árvore de espécies e, conseqüentemente, o resultado final e possíveis análises futuras.

Diante do exposto, este trabalho visou desenvolver um *pipeline* computacional para reconstrução filogenômica de organismos procarióticos, identificando e tratando a informação de genes transferidos horizontalmente. Além disso, foi desenvolvido um novo método computacional, aplicando esse método de reconciliação entre árvores de genes e espécie para confirmar quais genes podem ter sido transferidos horizontalmente ao longo da história evolutiva do grupo de organismos estudado. Esse novo método foi aplicado na reconstrução evolutiva de *Corynebacteriales*, o maior clado em número de genomas completos sequenciados do filo *Actinobacteria*.

OBJETIVO GERAL

Desenvolvimento de um novo método computacional para reconciliação de genes transferidos horizontalmente e reconstrução de árvore de espécies de procariotos.

Objetivos específicos

- Identificar genes adquiridos via transferência horizontal, utilizando métodos paramétricos;
- Desenvolver um método filogenético para confirmar os genes transferidos horizontalmente identificados previamente por métodos paramétricos;
- Desenvolver um *pipeline* para reconstrução filogenômica de procariotos;
- Aplicar o método em um estudo de caso (Ordem *Corynebacteriales*) para reconstrução da árvore de espécies;
- Validar o método proposto com a literatura.

Capítulo 1 – *Reconstructing the phylogeny of Corynebacteriales while accounting for horizontal gene transfer*

Reconstructing the Phylogeny of *Corynebacteriales* While Accounting for Horizontal Gene Transfer

Nilson Da Rocha Coimbra,^{1,2} Aristoteles Goes-Neto,² Vasco Azevedo,² and Aïda Ouangraoua^{1*}

¹Department of Computer Science, University of Sherbrooke, 2500 Boul. de l'Université, Sherbrooke, Quebec, Canada, JK1 2R1

²Programa Interunidades de Pós-graduação em Bioinformática, Universidade Federal de Minas Gerais, 6627, Avenida Antônio Carlos, 31270-901, Belo Horizonte, Minas Gerais, Brazil

* Corresponding author: Aïda Ouangraoua, Department of Computer Science, University of Sherbrooke, aida.ouangraoua@usherbrooke.ca

Abstract

Horizontal gene transfer (HGT) is a common mechanism in Bacteria that has contributed to the genomic content of existing organisms. Traditional methods for estimating bacterial phylogeny, however, assume only vertical inheritance in the evolution of homologous genes, which may result in errors in the estimated phylogenies. We present a new method for estimating bacterial phylogeny that accounts for the presence of genes acquired by HGT between genomes. The method identifies and corrects putative transferred genes in gene families, before applying a gene tree-based summary method to estimate bacterial species trees. The method was applied to estimate the phylogeny of the order *Corynebacteriales*, which is the largest clade in the phylum *Actinobacteria*. We report a collection of 14 phylogenetic trees on 360 *Corynebacteriales* genomes. All estimated trees display each genus as a monophyletic clade. The trees also display several relationships proposed by past studies, as well as new relevant relationships between and within the main genera of *Corynebacteriales*: *Corynebacterium*, *Mycobacterium*, *Nocardia*, *Rhodococcus*, and *Gordonia*. An implementation of the method in Python is available on GitHub at <https://github.com/UdeS-CoBIUS/EXECT>.

Key words: Phylogeny estimation, Bacteria, *Corynebacterium*, *Mycobacterium*

INTRODUCTION

One of the major discoveries in the 20th century is the bacterial production of antibiotics, which are useful in treating bacterial infections (Bister et al., 2004; Fair and Tor, 2014). The ongoing evolution of bacteria, however, contributes to the appearance of new bacterial species, including new antibiotic-resistant pathogenic species (Fischbach and Walsh, 2009). In terms of genome structure, bacterial species differ from each other in the content and arrangement of genes in their genomes, which results from genome rearrangement, gene duplication, gene loss, and horizontal gene transfer (HGT) events (Gogarten et al., 2002). In particular, HGT has been shown to be a primary force underlying antibiotic resistance and virulent genes spreading in Bacteria (Ruiz et al., 2011; Zhi et al., 2017).

HGT is the transfer of genetic material through a process different from vertical inheritance (Soucy et al., 2015). The modules of genetic transfer are usually genes, but it was also shown that HGT can occur at the level of protein domains (Chan et al., 2009). The prevalence of HGT events in bacterial evolution limits the use of phylogenetic methods that assume only vertical inheritance evolutionary events. Traditionally, alignments of sequences of 16S rRNA genes have been used to estimate bacterial phylogenies and study bacterial diversity. This approach relies on the assumption that 16S rRNA genes constitute essential genes that are only vertically inherited. Several studies have, however, reported evidence for HGT of 16S rRNA genes (Kitahara and Miyazaki, 2013; Miyazaki et al., 2017; Schouls et al., 2003; Yap et al., 1999). Moreover, the identification and classification bacterial species based solely on 16S rRNA genes often lead to errors in phylogenetic estimations (Rajendhran and Gunasekaran, 2011). The reason could be the intragenomic heterogeneity in bacterial rRNA as well as the presence of mosaicism and multiple copies of 16S rRNA genes in genomes, which may result from HGT events (Klappenbach et al., 2000; Schouls et al., 2003). In this context, the main contribution of this work is a method for estimating bacterial phylogenies with sets of gene families but without assuming only vertical inheritance in the evolution of gene families.

Phylogenetic reconstruction usually relies on two steps: first, the identification of groups of orthologous sequences in genomes, and second, the construction of a tree explaining the evolution

within orthology groups by vertical inheritance (Felsenstein, 1985). Therefore, computing accurate orthology groups in the first step is a prerequisite for reconstructing accurate phylogenies in the second step. For the second step, phylogeny methods can be classified into three main approaches: alignment-based, gene order-based, and gene tree-based methods (Wolf et al., 2002).

Alignment-based methods infer the species tree based on a concatenation of multiple sequence alignments on the orthology groups. This approach has been widely used because it scales well to a large number of orthology groups and species (Ciccarelli et al., 2006; Rokas et al., 2003). Nonetheless, alignment-based methods do not allow for accounting for the impact of genome content and structure evolution in estimating species diversity (Saitou and Nei, 1987). Another class of alignment-based methods are whole genome SNP-based methods that start by removing signals from recombination and then build a species tree using whole genome alignments, SNPs and maximum likelihood approaches (Castillo-Ramírez et al., 2012 ; Comas et al., 2013).

Gene order-based methods infer a species tree based on the difference between genomes in terms of gene content and arrangement (Belda et al., 2005; Bourque et al., 2004; Sankoff and Blanchette, 1998). They allow for accounting for the evolution of gene content and arrangement as well as gene conservation or splitting. Gene order-based methods are suitable for reconstructing phylogenies of closely related species (Moret et al., 2013). Nevertheless, this approach is limited by the complexity in scaling up to large datasets and the lack of a well-defined model of gene order evolution (Moret et al., 2001).

Gene tree-based methods consist in using a set of gene trees—one for each orthology group—in order to estimate a species tree that could explain the evolution of gene families within the species tree (Suyama and Bork, 2001). Such methods are currently in limited use in estimating bacterial phylogeny because they are very sensitive to the presence of erroneous genes in orthology groups caused by HGT events, leading to a disruption in the phylogenetic signals (Ravenhall et al., 2015). Thus, the detection and discarding of transferred genes from orthology groups is a prerequisite for using gene tree-based methods for estimating bacterial phylogeny.

Computational methods for HGT detection can be classified into two main approaches: parametric methods and comparative methods. Parametric methods are intragenomic, and exploit sequence composition changes along a genome sequence to infer putative HGT regions. Comparative methods are intergenomic and include alignment-based and phylogeny-based methods. Alignment-based methods—such as MobilomeFINDER (Ou et al., 2007)—make use of the alignment between closely related genomes to infer HGT. Phylogeny-based methods—such as NOTUNG (Chen et al., 2000)—exploit the inconsistencies between gene trees and species trees to infer HGT (Lerat et al. 2005, Ravenhall et al., 2015 ; Jeong et al., 2019). On the one hand, parametric methods—such as IslandPath-DIMOB (Bertelli and Brinkman, 2018)—take advantage of relying solely on genome sequences by an intrinsic analysis. They achieve average recall rates with high precision rates. On the other hand, comparative methods are limited by their requirement of an accurate species tree, which, in turn, is challenging to build in the presence of HGT (Lasek-Nesselquist et al., 2012). Nevertheless, when a preliminary, partially resolved species tree is available, the congruence of a gene tree with this species tree can be used to correct the misclassification of some genes as transferred genes inferred by parametric methods.

This paper presents a gene tree-based method accounting for HGT events in estimating bacterial phylogenies (Figure 1). After collecting the input dataset, which consists of genome sequences with the locations of their CDS sequences representing genes (Step 1), the method starts by detecting putative transferred genes with a parametric method (Step 2). Putative transferred genes are identified using an intra-genomic genomic island detection method in order to avoid the circular argument of detecting HGT using a species tree, and then removing HGT from gene trees to compute a species tree. In parallel, genes are clustered into homology groups based on their CDS similarities (Step 3). Subsequently, putative orthology groups, containing a single gene per genome, are used to build a preliminary, partially resolved species tree with an alignment-based phylogenetic method (Step 4). The preliminary species tree is used to correct misclassified putative transferred genes in homology groups with a phylogenetic approach. The latter consists in comparing the gene tree of each homology group with the species tree in order to compare the phylogenetic position of putative transferred gene in the two trees. A putative transferred gene whose location induces no HGT in the reconciliation between the gene tree and the species tree is reclassified as a vertically inherited gene (Step 5). Lastly, the remaining transferred genes are

removed from homology groups. The latter are used to build gene trees and the final species trees using phylogenetic gene tree-based methods (Step 6).

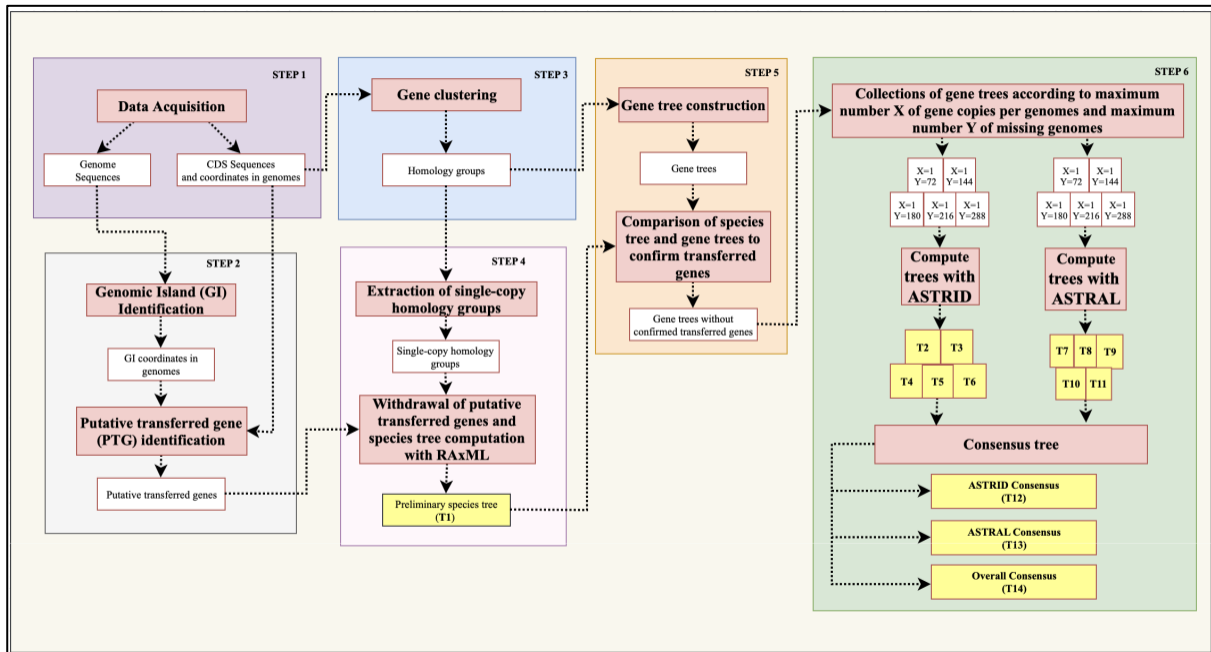


FIG. 1. Overview of the method, which consists of 6 steps.

The pipeline was applied to estimate the phylogeny of *Corynebacteriales*, the largest clade of the phylum *Actinobacteria* in terms of number of available genomes. The input dataset comprised 360 complete genome sequences obtained in Step 1, and the output consisted of 10 distinct phylogenetic trees on the 360 genomes that were estimated with two gene tree-based phylogenetic methods, ASTRID (Vachaspati and Warnow, 2015) and ASTRAL II (Mirarab and Warnow, 2015) in Step 6. The similarity of the estimated phylogenies was compared by computing the percentage of conserved clades between each pair of trees. The final phylogeny of *Corynebacteriales* was obtained by computing consensus trees using the majority rule consensus (Retief, 2000) in two phases. First, the collections of trees obtained with ASTRAL and ASTRID were reduced to two trees. Then, those two trees and the preliminary species tree from Step 4 were reduced to a single tree.

RESULTS

A new gene-tree based method applied to estimate Corynebacteriales phylogeny

We present a gene tree-based method that includes the detection and correction of putative horizontally transferred genes to estimate bacterial phylogenies using complete genome sequences (for an overview of the method, see Figure 1; for a detailed description of the six steps, see the methods section).

The phylogeny of *Corynebacteriales* was estimated using 360 records from NCBI Reference Sequence Database, release 81 (Step 1). The 360 genomes cover 101 species and 11 genera of *Corynebacteriales*, as presented in Table 01 and additional files A1 and A2.

Table 1. Input dataset for *Corynebacteriales* phylogenetic tree estimation

Genus	Number of Species	Number of Genomes
<i>Lawsonella</i>	1	2
<i>Hoyosella</i>	1	1
<i>Rhodococcus</i>	7	22
<i>Mycobacterium</i>	32	169
<i>Dietzia</i>	1	1
<i>Tsukamurella</i>	1	1
<i>Corynebacterium</i>	46	150
<i>Brevibacterium</i>	1	2
<i>Nocardia</i>	6	6
<i>Gordonia</i>	4	5
Total	101	360

Using parametric methods for HGT detection, 168 724 putative transferred genes (PTG) located into 2 874 genomic islands (GI) were detected (Step 2). Additional file A3 presents the number of GIs and PTGs detected per genome.

The gene clustering step resulted in the clustering of 1 356 782 genes (99.2% of genes) into 17 821 non-singleton homology groups (Step 3). Additional file A4 presents the details on the composition of the homology groups.

The homology groups containing exactly one gene from each of the 360 genomes were considered as putative orthology groups. After the PTGs were removed from these groups, they were used to build a preliminary species tree using the RAxML maximum likelihood phylogenetic method (Stamatakis, 2014) (Step 4). Table S1 (see the supplementary materials) presents the 13 putative orthology groups used in this step. Figure S1 shows the preliminary species tree.

The preliminary species tree was then used to check the PTGs in the homology groups using a phylogenetic approach that consists in comparing the gene trees of homology groups with the species tree. Using this approach, 13 966 PTGs (8.29% of PTGs) were reclassified as vertically inherited genes (Step 5).

The gene trees corresponding to homology groups with, at most, one gene per genome were clustered into 5 collections of trees according to the maximum proportion of genomes without any gene in the homology group: 20%, 40%, 50%, 60%, and 80%. For instance, gene trees in which 45% of genomes did not have a gene were included in the 50%, 60%, or 80% groups. Table S2 (see the supplementary materials) provides the number of trees in the 5 resulting collections.

Two gene tree-based phylogenetic methods—ASTRID (Vachaspati and Warnow, 2015) and ASTRAL II (Mirarab and Warnow, 2015)—were applied to the 5 collections to generate 10 phylogenies on the 360 input genomes. ASTRID and ASTRAL are methods motivated by, and statistically consistent with, the multispecies coalescent model such that there is free recombination between, but not within, loci. The use of ASTRID and ASTRAL is motivated by the presence of horizontally transferred genes in the data, not detected in Step 2. The trees obtained using ASTRID and ASTRAL were rooted with the outgroup method by including a homologous CDS from the species *Nostoc punctiforme*. In order to evaluate the similarity between the estimated phylogenies, the percentage of conserved clades between each pair of trees was computed (see Table 2). The average pairwise similarity between ASTRID trees is 75.89% with values ranging from 67.22% to 87.78%. ASTRAL trees display a higher average pairwise similarity of 80.14% with values ranging from 74.17% to 89.17%. The average pairwise similarity between ASTRID trees and ASTRAL trees is 67.89% with values ranging from 65.56% to 73.89%.

Considering the high similarity between the 5 phylogenies estimated using each of the two methods, the trees estimated with ASTRID, on one side, and with ASTRAL, on the other, were reduced to 2 consensus trees with the CONSENSE majority-rule consensus tool (Felsenstein, 1993). Table S3 (see the supplementary materials) presents the similarity between the 2 resulting consensus trees and the 10 initially estimated phylogenies. The 2 consensus trees for ASTRID and ASTRAL have a high percentage of conserved clades (78.27%), and 82.14% and 80.61%, respectively, of conserved clades with the preliminary species tree from Step 4 obtained with RAxML. Therefore, a final reduction of the 3 phylogenies—ASTRID consensus, ASTRAL consensus and RAxML—to a single consensus tree (referred to as overall consensus) was made. Figure S2 (see the supplementary materials) depicts the 4 trees viewed at the genus level. All 14 trees generated in this research are available on the iTOL webserver (Letunic and Bork, 2019), at https://itol.embl.de/shared/cobius_udes.

Table 2. Square matrix of percentage of conserved clades between phylogenies estimated using RAxML (Step 4), and ASTRID and ASTRAL (Step 5).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
RAxML (1)	100	82.65	82.14	83.16	82.65	82.14	81.12	81.12	80.10	81.12	81.63
ASTRID20 (2)		100	87.78	74.17	74.44	67.22	68.61	70.28	66.67	66.67	66.11
ASTRID40 (3)		–	100	76.11	75.56	68.61	67.5	69.17	67.22	67.78	66.67
ASTRID50 (4)		–	–	100	83.61	72.5	65.56	67.22	68.61	69.17	68.06
ASTRID60 (5)		–	–	–	100	78.89	66.11	67.22	68.06	70.56	70.28
ASTRID80 (6)		–	–	–	–	100	65.83	65.83	66.39	68.89	73.89
ASTRAL20 (7)		–	–	–	–	–	100	87.22	77.78	75.0	74.17
ASTRAL40 (8)		–	–	–	–	–	–	100	78.89	77.78	75.83
ASTRAL50 (9)		–	–	–	–	–	–	–	100	89.17	80.0
ASTRAL60 (10)		–	–	–	–	–	–	–	–	100	85.56
ASTRAL80 (11)		–	–	–	–	–	–	–	–	–	100

ASTRAL was used to compute the quartet supports of branches in the 4 trees. The quartet support of a branch is computed using the percentage of quartets in input gene trees that agree or disagree with the branch (Sayyari and Mirarab, 2016). For each of the ASTRID, ASTRAL and RAxML trees, the internal branches (non-trivial clades) were divided into 2 groups: those conserved in the overall consensus tree, and those that were not conserved. Figure 2 (Bottom-Left) presents the number of branches in each of the 6 groups, and Figure 2 (Top-Left) presents boxplots of the quartet supports of branches in the 6 groups. For RAxML, the boxplots of the bootstrap supports of branches are also depicted. We observe that the clades from ASTRID, ASTRAL and RAxML included in the overall consensus show high quartet support values, while the clades not included in the overall consensus show low quartet support values. The same observation holds for the RAxML bootstrap values. This means that the overall consensus tree is effective at retaining the clades of the 3 input trees which present the highest support values.

Table 3 presents the similarity measures between the 4 trees. The overall consensus has, respectively, 98.06%, 96.90% and 84.69% of conserved clades with the ASTRID, ASTRAL and RAxML trees. The internal branches of the overall consensus tree were divided into 4 groups depending on their presence in all 3 input trees (ASTRID, ASTRAL, RAxML) or in only 2 of the input trees. Figure 2 (Bottom-Right) presents the number of internal branches in each of the 4 groups and Figure 2 (Top-Right) presents boxplots of the quartet supports of branches in the 4 groups. The branches present in all 3 input trees (ASTRID-ASTRAL-RAxML) constitute the largest group with the highest support values, followed by the ASTRID-ASTRAL branches, a few ASTRID-RAxML branches, and finally a few ASTRAL-RAxML branches. This means that the largest contribution comes from the consensus between the 3 trees or between ASTRID and ASTRAL trees, and further branches are added thanks to the consensus with the RAxML tree.

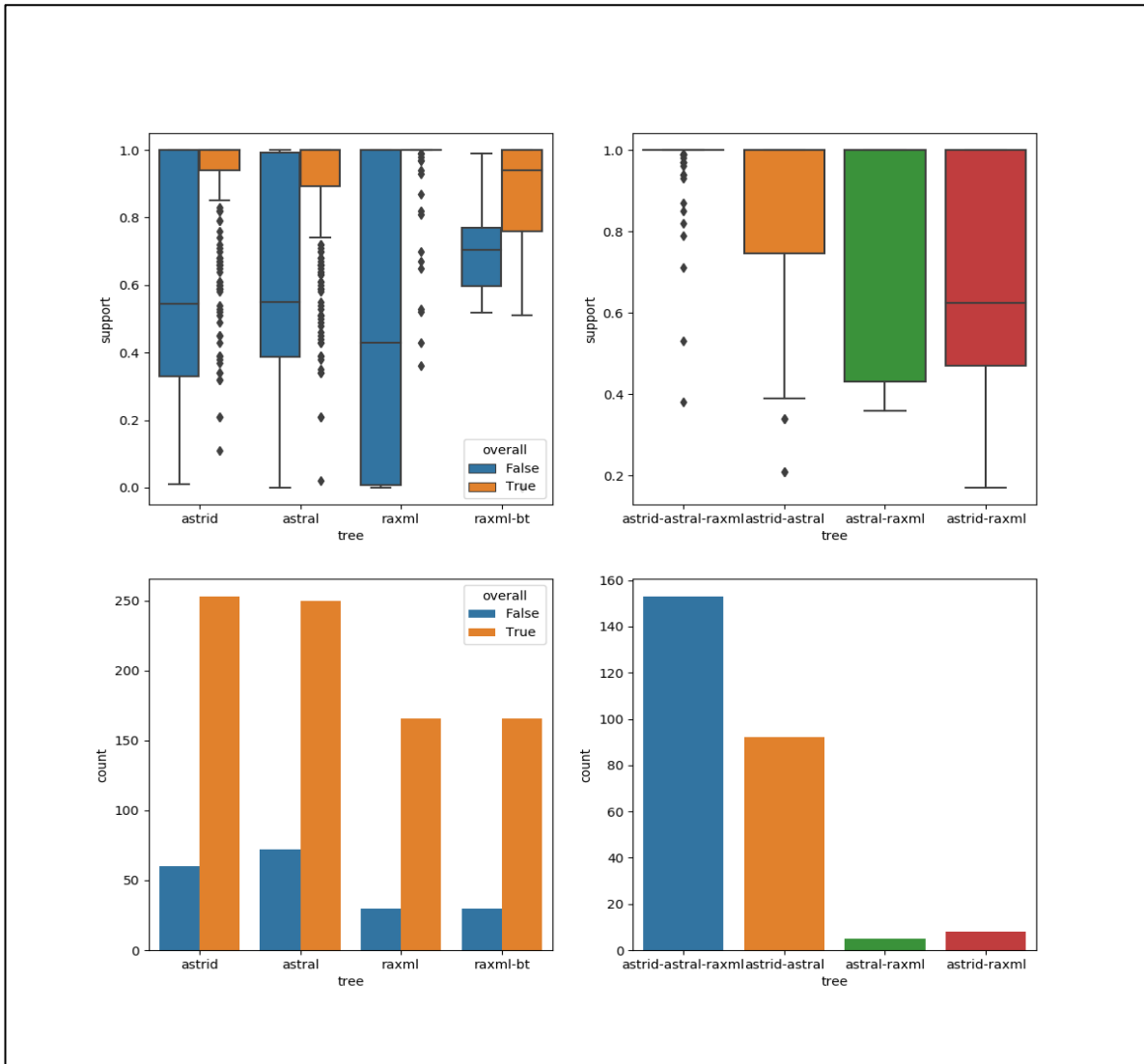


FIG. 2. Support values of internal branches of the ASTRID, ASTRAL, RAxML and Overall consensus trees. Bottom-Left: Numbers of branches in the ASTRID, ASTRAL, RAxML trees conserved in the overall consensus tree and numbers of branches not conserved. Top-Left: Boxplots of the quartet support values of branches in each group and the bootstrap support values for the RAxML tree. Bottom-Right: Numbers of branches in the overall consensus tree present in the ASTRID, ASTRAL and RAxML trees, or in only 2 of the 3 trees. Top-Right: Boxplots of the quartet support values of branches in each group.

Table 3. Square matrix of the percentage of conserved clades between the consensus trees and the preliminary RAxML tree.

	(1)	(2)	(3)	(4)
ASTRID consensus (1)	100	78.27	82.14	98.06
ASTRAL consensus (2)		100	80.61	96.90
RAxML (3)		–	100	84.69
Overall consensus (4)		–	–	100

Analysis at the genus level

All the trees estimated in our study place the genus *Brevibacterium* (BV) within *Corynebacterium* (CR), which supports reclassifying *Brevibacterium* as *Corynebacterium*, as proposed in a recent study (Yang and Yang, 2017). The *Corynebacteriales* phylogeny is still under debate. Past studies have reported various topologies for the phylogeny of this order. Gao and Gupta (Gao and Gupta, 2012) used the sequence alignments of 35 proteins with neighbor-joining methods to estimate the phylogeny of *Actinobacteria* that includes *Corynebacteriales*. Sen et al. (Sen et al., 2014) used 54 protein sequences aligned with RAxML to infer a phylogeny on 100 actinobacterial strains. They also reported a second phylogeny on the 100 actinobacterial strains obtained by applying RAxML to the alignment of 5 conserved genes identified with a multi-locus sequence analysis (MLSA). We compared the overall consensus tree obtained in our study with the *Corynebacteriales* phylogenies from (Gao and Gupta, 2012) and (Sen et al., 2014). Figure 3 depicts the compared phylogenies at the genus level. A strong consensus can be seen between the trees for the clade grouping *Nocardia* and *Rhodococcus* (6/6), and the clade regrouping *Gordonia* and *Tsukamurella* (5/6). We also observed a majority-rule consensus for a clade grouping *Nocardia*, *Rhodococcus*, and *Hoyosella* (3/5), and for placing *Corynebacterium* as the outgroup (3/6).

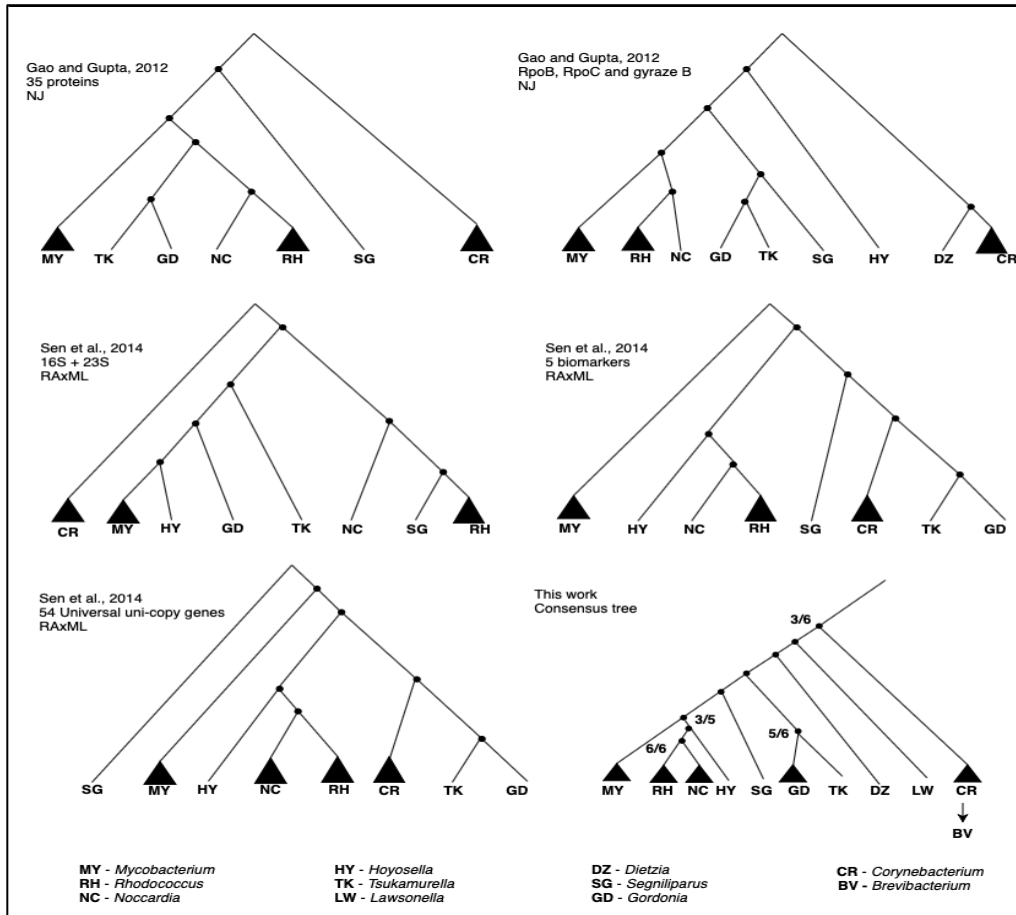


FIG. 3. Illustration at the genus level of the overall consensus phylogeny reconstructed in our work, 2 phylogenies reconstructed in (Gao and Gupta, 2012), and 3 phylogenies reconstructed in (Sen et al., 2014). For each tree, the genera for which the dataset contains more than one species are represented as triangles. Note that the sets of genera differ between trees. The ratio of trees displaying the clade is indicated for each conserved clade.

Systematic analysis of the phylogeny reported inside genera

We analyzed the phylogenies for each of the 5 genera with more than 1 species in the dataset: *Corynebacterium* (46 species), *Mycobacterium* (32 species), *Rhodococcus* (7 species), *Nocardia* (6 species), and *Gordonia* (4 species). In all the trees estimated in our study, the species of the same genus were grouped into monophyletic groups. We extracted the complete sub-tree corresponding to each genus in the overall consensus tree.

Corynebacterium. The genus *Corynebacterium* comprises a variety of bacterial species that includes potential pathogens for human and animals, as well as pathogens for normal microbiota (Von Graevenitz and Bernard, 2006). Most of the mechanisms underlying diseases caused by these species are still unclear; a few phylogenies of the genus have been reconstructed (Baek et al., 2018; Dangel et al., 2019; Pascual et al., 1995). In all the trees estimated in our study, we observed a division into two categories: non-pathogenic genomes and pathogenic genomes forming a monophyletic group (Figure 4). We noted a single exception: the classification of *C. jeikeium* among non-pathogenic genomes. *C. jeikeium* is a pathogen isolated from immunosuppressive patients highly exposed to antibiotic treatments (Tauch et al., 2005). The positioning of this pathogen species among non-pathogens is surprising and might be related to horizontal gene transfer for the acquisition of antibiotic-resistance genes.

In fact, as it was only isolated from immunosuppressive patients, this is a robust biological clue suggesting that this species does not act as a pathogen in healthy organisms (in this case, humans). Therefore, it is not a primary pathogen, which would corroborate our findings. More detailed studies are needed to refine this assumption. In the non-pathogenic group, *C. glutamicum* is the most thoroughly studied species due to its biotechnological applications in producing amino acids such as L-arginine, L-histidine, L-carnitine, L-lysine, and L-valine (Keilhauer et al., 1993). In all the trees estimated in our study, the genomes of *Brevibacterium flavum* strain ATCC 15168 (RefSeq. CP011309) and *Brevibacterium flavum* ZL 1 (RefSeq. CP004046) always appear in the same clade of the genomes of *C. glutamicum* (Figure S3 in the supplementary materials). This classification was recently proposed in the literature (Yang and Yang, 2017).

In the pathogenic group, we observe a clustering of *C. diphtheriae* and *C. ulcerans*. *C. diphtheriae* is the etiological agent of diphtheria in humans, an infectious disease caused by the exotoxin produced by this pathogen (Cerdeno-Tarraga et al., 2003). *C. ulcerans* is primarily reported for causing mastitis in cattle and humans due to the consumption of raw milk or unpasteurized dairy products in rural populations (Hommeiz et al., 1999). In the literature, *C. ulcerans* has been closely

related to *C. diphtheriae*, which produces a toxin causing symptoms similar to those caused by *C. ulcerans* (Riegel et al., 1995).

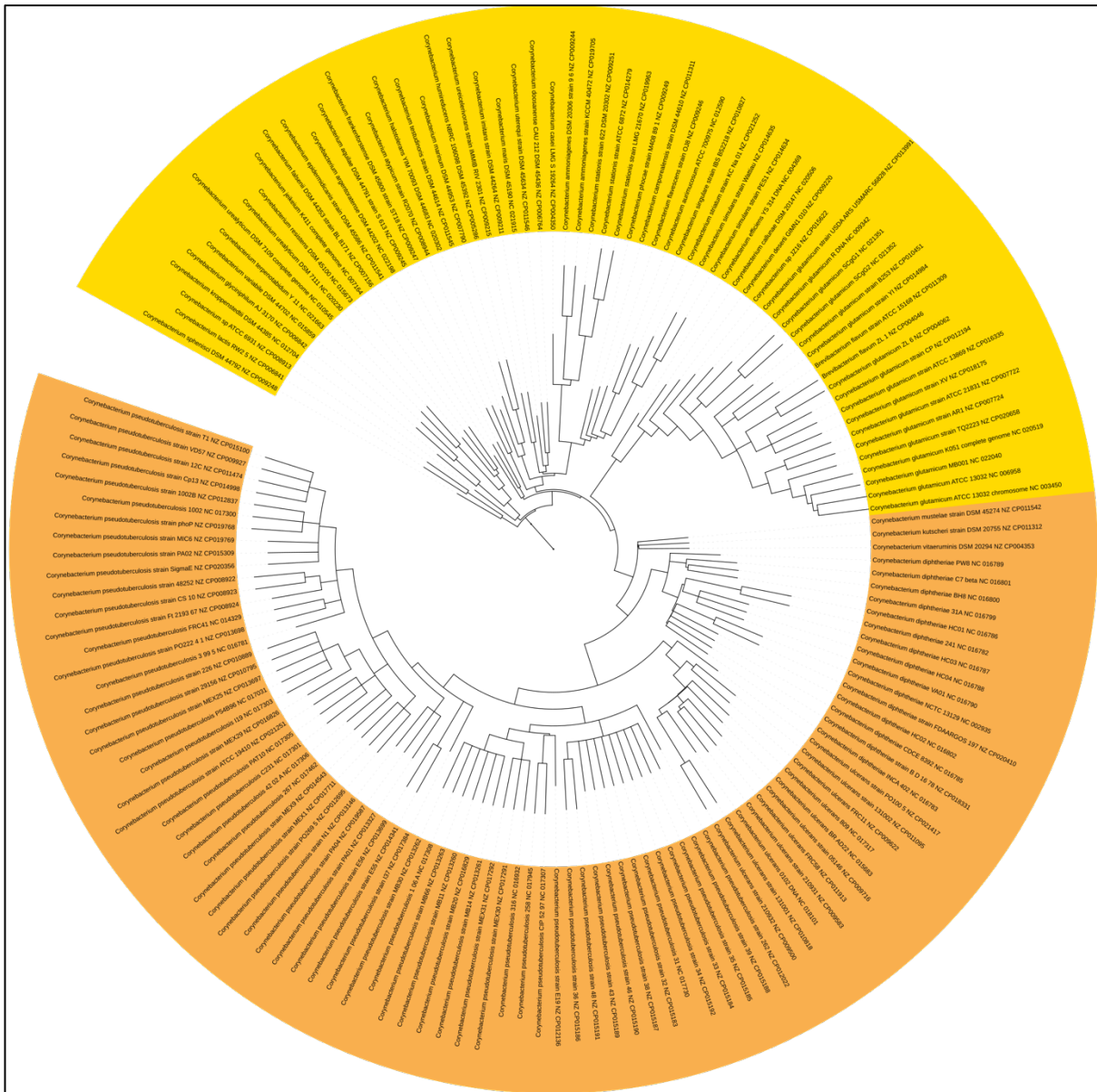


FIG. 4. Partition of *Corynebacterium* into two categories: non-pathogenic in yellow and pathogenic in orange. Detailed method used to estimate this phylogeny (overall consensus tree: RAxML + ASTRID consensus + ASTRAL consensus).

We also detected the recently proposed anagenesis of *C. pseudotuberculosis* (Oliveira et al., 2016). In this model, two biovars are described: equi and ovis. They mainly differ by the presence of the nitrate reductase enzyme present in biovar equi, which results in 1% of the nucleotide differences between biovars (Soares et al., 2013) (Figure 5). *C. pseudotuberculosis* is the etiological agent of caseous lymphadenitis (CLA), a highly prevalent chronic disease affecting sheep and goats. It is difficult to control and causes significant economic losses to farmers (Baird and Fontaine, 2007). Human infections caused by *C. pseudotuberculosis* are rare, but it has been reported as the agent of necrotizing lymphadenitis in human (Mills et al., 1997). Lastly, we observed that our phylogeny for *Corynebacterium* species is consistent with the phylogeny proposed by Gao and Gupta (Gao and Gupta, 2012).

Mycobacterium. The genus *Mycobacterium* comprises one the most dangerous human pathogens—*M. tuberculosis*—which causes tuberculosis (Gagneux, 2018; Koch and Mizrahi, 2018). This genus also comprises others important animal pathogens such as *M. leprae*, *M. bovis*, and *M. avium* (Frothingham and Wilson, 1993). The taxonomy of *Mycobacterium* solely relies in two categories: slow growers and fast growers. This poorly detailed taxonomy is due to the lack of descriptive features for taxonomic classification. A more detailed classification would help with global monitoring of disease outbreaks caused by species of this genus (ROGALL et al., 1990; Stahl and Urbance, 1990). All the estimated trees display a division into the 2 categories: 61 genomes of slow growers forming a monophyletic group, and 108 genomes of fast growers forming another monophyletic group (Figure 6).

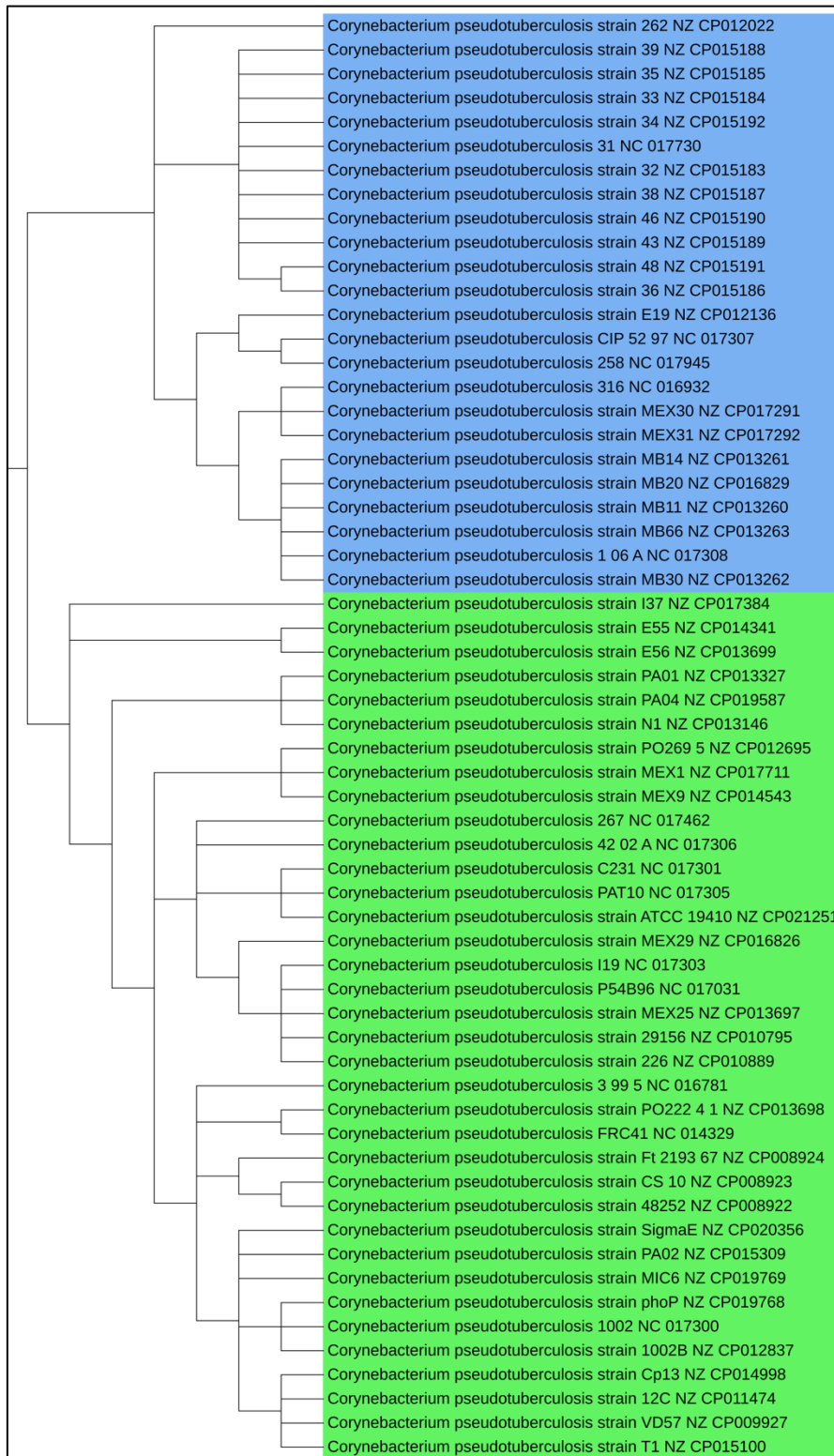


FIG. 5. Partition of *Corynebacterium pseudotuberculosis* into two biovars. Biovars *equi* and *ovis* are shown in green and blue, respectively. Detailed method used to estimate this phylogeny (overall consensus tree: RAxML + ASTRID consensus+ ASTRAL consensus)

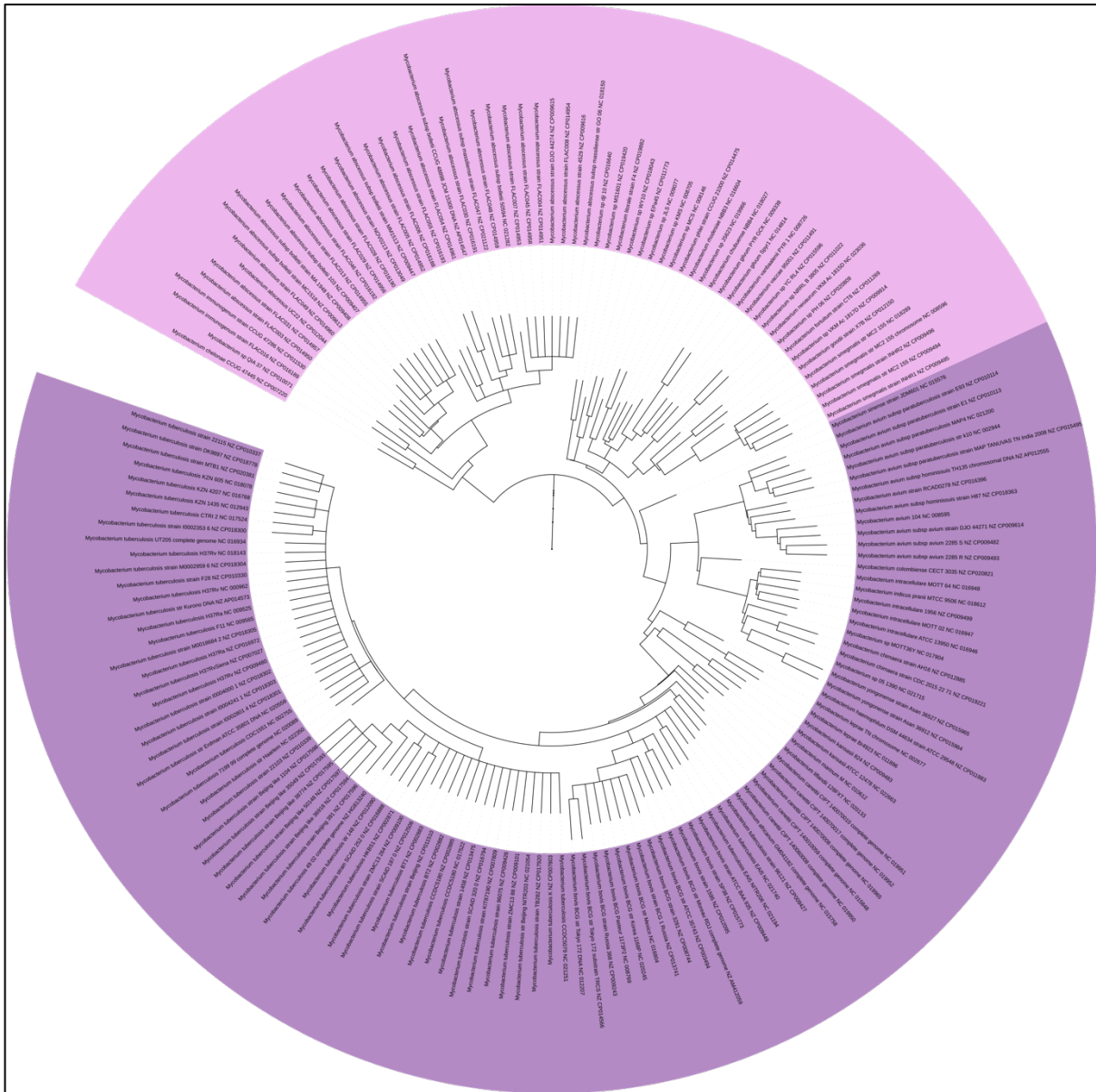


FIG. 6. Partition of *Mycobacterium* into two categories: slow growers in purple and fast-growers in pale pink. Detailed method used to estimate this phylogeny (overall consensus tree: RAxML + ASTRID consensus+ ASTRAL consensus).

Rhodococcus. The *Rhodococcus* species are used as versatile genetic tools in the biotechnological industry because of their capacity for remediation, biotransformation and biocatalysis, biodegradation of diverse metabolic compounds, adaptation and tolerance to solvents, and interactions with metals (Sangal et al., 2019). *Rhodococcus* species are distributed in soil, water, and marine sediments (Larkin et al., 2005). Some of them are also pathogens for humans, animals, and plants (Prescott, 1991). While new *Rhodococcus* genomes are still being sequenced because of their important biotechnological applications, the current phylogenies of *Rhodococcus* are only

estimated for closely related species using few biomarkers (Anastasi et al., 2016; Duquesne et al., 2017). We report a phylogeny of 22 *Rhodococcus* genomes corresponding to 6 species and 11 unclassified genomes, divided into 6 clusters (Figure S4 in the supplementary materials). The lack of estimated phylogenies for *Rhodococcus* at the species level in the literature makes it hard to conduct a proper comparison with past studies. We however observed partial agreement of the estimated phylogeny for the 6 *Rhodococcus* species with the phylogeny from (Anastasi et al., 2016): (*R. fascians*,(*R. pyridinovorans*, (*R. erythropolis*,(*R. opacus*, *R. jostii*))))).

Nocardia. *Nocardia* species are a complex group of organisms that cause serious human infections, especially in immunocompromised patients. Like *Rhodococcus* and the other genera of *Cornynebacteriales*, the taxonomy and phylogeny of *Nocardia* species are subject to open debate (Conville et al., 2018). The dataset contains 6 complete genomes of *Nocardia* corresponding to 6 species. All the estimated trees display the same phylogeny for the 6 species (Figure S5 in the supplementary materials). In an estimation of the phylogeny of *Nocardia* species, Conville et al. (Conville et al., 2018) described the complex history behind the taxonomy of the genus by reconstructing a phylogenetic tree using the 16S rRNA gene of 59 genomes of *Nocardia*. The intersection between their dataset and our dataset consists of only 4 species. The estimated phylogeny for these 4 species in (Conville et al., 2018) is ((*N. farcinica*, *N. brasilensis*), (*N. cyriacigeorgica*, *N. nova*)), which differs from the phylogeny estimated in this report.

Gordonia. *Gordonia* species have attracted interest from the biotechnological industry in recent years because of their ability to degrade environmental pollutants as well as natural polymers and compounds, making them potentially useful for environmental and industrial biotechnology (Arenskötter et al., 2004). Some species of *Gordonia* are reported to cause infections in humans (Ramanan et al., 2013; Sowani et al., 2017). Previous phylogenies of *Gordonia* were estimated with 16S rRNA genes, and the phylogeny is still under debate (Blaschke et al., 2007). Kang et al. (Kang et al., 2009) studied the phylogeny of 23 species using *gyrB*, *secA1*, and 16S rRNA genes. In our study, we collected 5 *Gordonia* genomes, corresponding to 3 species and 2 unclassified genomes. All the estimated trees display the same phylogeny for the 5 genomes (Figure S6 in the supplementary materials). The induced phylogeny for the 3 species included in our dataset—*G.*

polyisoprenivorans, *G. bronchialis*, and *G. terrae* — agrees with the induced phylogeny from (Kang et al., 2009).

DISCUSSION

As new bacterial genomes are still being sequenced, one of the major problems lies with identifying the main bacteria groups and recovering the phylogenetic relationships between these groups (Larson, 1998). Current modern molecular-biology techniques are still being redesigned to identify new species because the classical approaches based on sequence analysis are inefficient for discrimination (Glaeser and Kämpfer, 2015). Furthermore, characterizing the differences between closely related species remains challenging (Christensen and Olsen, 2018). Sequence-based phylogenies have been an active research field since the beginning of the 2000s. They grounded the current knowledge about the diversity of organisms on the Earth. Estimating bacterial phylogenies is not, however, a trivial problem. This is mainly because bacterial genomes are highly affected by the swapping of genetic material between genomes via HGT processes (Soucy et al., 2015). Through this mechanism, bacterial genomes acquire and spread genes that confer adaptive advantages, such as antibiotic-resistance genes leading to the rise of multidrug-resistant bacteria (van Duin and Paterson, 2016). Thus, accounting for horizontally transferred genes is necessary to accurately estimate bacterial phylogenies when using sequence-based phylogenetic methods. Nonetheless, current sequence-based phylogenetic methods do not include a step back to audit the datasets in order to identify and remove transferred genes. An alternative to identifying and removing transferred genes before estimating phylogenies is to infer ancestral recombination graphs (ARG) that record of all coalescence and recombination events in the evolution of a set of homologous sequences (D O'Fallon, 2013; Rasmussen et al., 2014). However, existing methods for ARG inference are computationally intensive and limited to small numbers of sequences. Herein, the phylogeny of *Corynebacteriales* was estimated while accounting for horizontal gene transfers (HGT), by detecting and removing a part of the HGT located in genomic islands (GI) using a parametric GI detection method, and by relying on phylogenetic reconstruction methods which are consistent the multispecies coalescent model with recombination within loci. The result is a species tree that displays all the genera as monophyletic clades. The estimated trees display several phylogenetic relationships proposed by previous studies: (i) the classification of

Brevibacterium flavum inside *Corynebacterium glutamicum* (Yang and Yang, 2017); (ii) the monophyletic group composed of pathogens *Corynebacterium ulcerans* and *Corynebacterium diphtheriae*; (iii) the biovar speciation inside *Corynebacterium pseudotuberculosis*; and (iv) the division between slow growers and fast growers in *Mycobacterium*. Finally, it is important to recall that the phylogenomics method devised in this paper presents the same limit as most phylogenetics and comparative genomics methods which reduce biological processes such as HGT to patterns, and thus investigate patterns (Nelson, 1970). One should always remember that phylogenetics methods are consistent under the hypothesis that there is a one-to-one correspondence between the target biological processes and the patterns investigated.

METHODS

Figure 1 depicts the entire method used to estimate the phylogeny of *Corynebacteriales*. The details and rationale underlying each step in the method are described below.

Step 1: Data acquisition

Step 1 consists in acquiring the *Corynebacteriales* genome and gene data. All complete *Corynebacteriales* genome sequences were retrieved from the REFSEQ NCBI database, release 81 (Maglott et al., 2005). The coding DNA sequences (CDSs) and gene coordinates were subsequently extracted using the genome annotations.

Step 2: Detecting genomic islands and identifying putative transferred genes

Identifying horizontally transferred genes in bacterial genomes is a prerequisite to computing a bacterial phylogenetic tree using gene tree-based phylogeny methods. Horizontally transferred genes can be located in genomic islands (GIs), which are large segments of DNA (10–200 kb) acquired by horizontal transfer (Langille et al., 2010). There are several approaches for detecting the GI regions in genomes. Some methods—such as MobilomeFINDER (Ou et al., 2007)—make use of a comparative genomics approach and identify GIs as deleted or inserted regions inferred by aligning closely related genomes. Other comparative genomics methods—such as NOTUNG (Chen et al., 2000)—identify HGT by detecting discordance between a gene tree and a species tree. Other methods, referred to as parametric, make use of a sequence composition-approach that defines GIs as regions with dinucleotide (G+C) bias or codon usage bias containing associated mobility genes. Compared to parametric methods, comparative approaches have the advantage of being able to detect old HGT events despite the process of sequence homogenization undergone by old GI regions. They however require the availability of closely related genomes or a reliable species tree for the input genomes. Since neither closely related genomes for all genomes of the *Corynebacteriales* dataset nor any reliable input species tree were available, GIs were detected with the parametric method IslandPath-DIMOB v1.0.0, which is currently the most accurate stand-alone method for GI prediction (Bertelli and Brinkman, 2018) (recall rate of 46.9% and high

precision rate of 87.4%). The default parameters of IslandPath-DIMOB were used. The genes contained in the detected GI regions were classified as putative transferred genes (PTGs). Note that, because of the recall rate of the method, there may be horizontally transferred genes located in GIs that were not detected by the method. They may also be transferred genes not detected by the method because they are not located in GIs. These undetected horizontally transferred genes are considered in Step 6 during the species trees construction.

Step 3: Clustering of genes into homology groups

The CDSs extracted in Step 1 were translated into protein sequences and clustered using Orthofinder1 (Emms and Kelly, 2015). This protein clustering tool was chosen because of its high accuracy compared to other currently available gene clustering methods (Emms and Kelly, 2015). Orthofinder solves gene length bias before constructing gene groups. An all-against-all BLASTP with a stringent cutoff e-value of 10^{-4} was applied between and within proteomes, and the result was used as input in Orthofinder to compute gene clusters. The resulting clusters of genes are called homology groups. The default parameters of Orthofinder were used. Note that a new version of Orthofinder, Orthofinder2 (Emms and Kelly, 2019) was released after the completion of the present study. The results presented in this study were obtained using Orthofinder1. However, the pipeline provided to reproduce the analysis on other datasets have been updated to include Orthofinder2.

Step 4: Preliminary species tree construction using single-copy homology groups

Single-copy homology groups were selected from the homology groups computed in Step 3. Single-copy homology groups are gene clusters containing exactly one gene from each genome. Such homology groups are considered as putative orthology groups that have evolved from a common ancestral gene without any gene duplication events. Thus, they can be used to infer a preliminary species tree using an alignment-based phylogeny estimation method. Due to HGT events, however, they may contain putative transferred genes that should be removed before using the groups for estimating the species tree. Still, putative transferred genes were removed from single-copy homology groups, and the remaining sequences in each group were aligned using the multiple sequence alignment software MAFFT (Katoh et al., 2002). The resulting alignments were

concatenated, and the concatenated multiple alignment was used as input to the phylogeny construction method RAxML (Stamatakis, 2014) to compute an initial phylogenetic tree with the set of *Corynebacteriales* genomes. The default parameters of MAFFT were used. RAxML was used with the following parameters: `raxml -s alignmentfile -p 123456 -m PROTGAMMAAUTO -b 123456 -N 100 -o Nostoc_punctiforme --asc-corr lewis`. The tree was rooted using the genome of *Nostoc punctiforme*—(Genbank ID: NC 010628), a symbiotic nitrogen-fixing cyanobacteria—as outgroup.

Step 5: Gene tree construction and discarding of confirmed transferred genes

Among the homology groups computed at Step 3, those that contained at most one gene per genome were extracted. The set of homology groups was restricted to this set because the gene tree-based methods for species tree estimation require that gene trees contain at most one gene per genome (Mirarab and Warnow, 2015; Vachaspati and Warnow, 2015). For each of the 9161 homology groups selected, a gene tree was built using the sequence alignment tool MAFFT (Katoh et al., 2002), and the phylogeny inference tool FastTree (Price et al., 2010). FastTree is a Maximum-likelihood (ML) method that only implements partially the ML approach. It was shown to be more accurate and faster than other ML approaches for applications on large datasets. FastTree was chosen for gene tree construction because of its effectiveness in computing trees on large datasets. The default parameters of FastTree were used. The gene trees were then rooted with homologous CDSs from *Nostoc punctiforme*, as in Step 4. A total of 631 homology groups without any homolog in *Nostoc punctiforme* were discarded, leaving 8 530 gene trees for the analysis. Each gene tree was compared to the preliminary species tree built in Step 4 in order to double-check the classification of putative transferred genes detected in Step 2, and correct false positives. The comparison method is as follows (see Figure S7 in the supplementary materials for an illustration). Given any maximum complete subtree T_1 of a gene tree G such that the leaves of T_1 were all putative transferred genes (PTGs), we considered T_2 , the sibling subtree of T_1 in G . The sets of species corresponding to the genes at leaves of T_1 and T_2 are denoted SA_1 and SA_2 , respectively. The putative transferred genes in T_1 were reclassified as vertically inherited genes if the lowest common ancestor (lca) node of SA_1 and the lca node of SA_2 in the species tree S were the same node or sibling nodes. The rationale is that if T_1 is the result of a HGT event from a donor branch

(a, b) to an acceptor branch (a', b') of the species tree such that b and b' are not sibling nodes, then the lca node of SA1 should be the node b' and the lca node of SA2 should be the node b. Thus, in the case where lca(SA1) and lca(SA2) are the same node or sibling nodes, the hypothesis that T1 is the result of a HGT event can be discarded. The putative transferred genes that were not reclassified were confirmed as transferred genes and removed from the homology groups and the corresponding gene trees.

Step 6: Species tree construction using gene tree-based methods

The gene trees obtained at the end of Step 5 were categorized into 5 collections of trees according to the maximum proportion of missing genomes in the gene trees: 20%, 40%, 50%, 60%, or 80% of missing genomes. Using each collection of trees, two species trees were constructed using the gene tree-based summary methods ASTRID (Vachaspati and Warnow, 2015) and ASTRAL (Mirarab and Warnow, 2015). ASTRID and ASTRAL were used in order to account for remaining horizontally transferred genes that were not detected in Step 2, either because they were missed by the GI detection method, or because they are not located in GIs. The default parameters of ASTRID and ASTRAL were used. Subsequently, each set of 5 trees estimated using the same gene tree-based method (ASTRID or ASTRAL) was reduced to a single consensus tree following the majority-rule consensus algorithm in CONSENSE (Felsenstein, 1993). Lastly, the ASTRID consensus tree, the ASTRAL consensus tree, and the RAxML preliminary species tree from Step 4 were reduced to single overall consensus tree.

Supplementary Materials

Supplementary files A1–A4, Tables S1–S3, and Figures S1–S7 are available *Genome Biology and Evolution* online.

All information to retrieve the data and the scripts used for the analysis are available on the CoBIUS lab GitHub (<https://github.com/UdeS-CoBIUS/EXECUT>).

Acknowledgments

This work was supported by the Brazilian funding agencies CNPq and FAPEMIG (24780/RTR/PRPG/ICB/DBG/BDS), and grants from the Canada Research Chair (CRC Tier 2 Grant 950-230577), and the Natural Sciences and Engineering Research Council of Canada (Discovery Grant RGPIN-2017-05552).

References

1. Anastasi, E., MacArthur, I., Scotti, M., Alvarez, S., Giguere, S., and Vazquez-Boland, J. A. 2016. Pangenome and phylogenomic analysis of the pathogenic actinobacterium *rhodococcus equi*. Genome biology and evolution, 8(10): 3140–3148.
2. Arenskötter, M., Bröker, D., and Steinbüchel, A. 2004. Biology of the metabolically diverse genus *gordonia*. Appl. Environ. Microbiol., 70(6): 3195–3204.
3. Baek, I., Kim, M., Lee, I., Na, S.-I., Goodfellow, M., and Chun, J. 2018. Phylogeny trumps chemotaxonomy: a case study involving *turicella otitidis*. Frontiers in Microbiology, 9: 834.
4. Baird, G. and Fontaine, M. 2007. *Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis. Journal of comparative pathology, 137(4): 179–210.
5. Belda, E., Moya, A., and Silva, F. J. 2005. Genome rearrangement distances and gene order phylogeny in γ -proteobacteria. Molecular Biology and Evolution, 22(6): 1456–1467.
6. Bister, B., Bischoff, D., Ströbele, M., Riedlinger, J., Reicke, A., Wolter, F., ... & Süssmuth, R. D. (2004). Abyssomicin C—A Polycyclic Antibiotic from a Marine *Verrucospora* Strain as an Inhibitor of the p-Aminobenzoic Acid/Tetrahydrofolate Biosynthesis Pathway. Angewandte Chemie International Edition, 43(19), 2574–2576.
7. Bertelli, C. and Brinkman, F. S. 2018. Improved genomic island predictions with *islandpath-dimob*. Bioinformatics, 34(13): 2161–2167.
8. Abyssomicin ca polycyclic antibiotic from a marine *verrucospora* strain as an inhibitor of the p-aminobenzoic acid/tetrahydrofolate biosynthesis pathway. Angewandte Chemie International Edition, 43(19): 2574–2576.
9. Blaschke, A. J., Bender, J., Byington, C. L., Korgenski, K., Daly, J., Petti, C. A., Pavia, A. T., and Ampofo, K. 2007. *Gordonia* species: emerging pathogens in pediatric patients that are identified by 16s ribosomal rna gene sequencing. Clinical infectious diseases, 45(4): 483–486.

10. Bourque, G., Pevzner, P. A., and Tesler, G. 2004. Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes. Genome Research, 14(4): 507–516.
11. Castillo-Ramírez, S., Corander, J., Marttinen, P., Aldeljawi, M., Hanage, W. P., Westh, H., ... & Holden, M. T. (2012). Phylogeographic variation in recombination rates within a global clone of methicillin-resistant *Staphylococcus aureus*. *Genome biology*, 13(12), R126.
12. Cerdeno-Tarraga, A., Efstratiou, A., Dover, L., Holden, M., Pallen, M., Bentley, S., Besra, G., Churcher, C., James, K., De Zoysa, A., et al. 2003. The complete genome sequence and analysis of *Corynebacterium diphtheriae* nctc13129. Nucleic acids research, 31(22): 6516–6523.
13. Chan, C. X., Darling, A. E., Beiko, R. G., & Ragan, M. A. (2009). Are protein domains modules of lateral genetic transfer?. *PloS one*, 4(2).
14. Chen, K., Durand, D., and Farach-Colton, M. 2000. Notung: a program for dating gene duplications and optimizing gene family trees. Journal of Computational Biology, 7(3-4): 429–447.
15. Christensen, H. and Olsen, J. E. 2018. Sequence-based classification and identification of prokaryotes. In Introduction to Bioinformatics in Microbiology, pages 121–134. Springer.
16. Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science, 311(5765): 1283–1287.
17. Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., ... & Yeboah-Manu, D. (2013). Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature genetics*, 45(10), 1176.
18. Conville, P. S., Brown-Elliott, B. A., Smith, T., and Zelazny, A. M. 2018. The complexities of nocardia taxonomy and identification. Journal of clinical microbiology, 56(1): e01419–17.
19. D O'Fallon, B. (2013). ACG: rapid inference of population history from recombining nucleotide sequences. *BMC bioinformatics*, 14(1), 40.

20. Dangel, A., Berger, A., Konrad, R., and Sing, A. 2019. Ngs-based phylogeny of diphtheria-related pathogenicity factors in different corynebacterium spp. implies species-specific virulence transmission. BMC microbiology, 19(1): 28.
21. Duquesne, F., Houssin, E., S'évin, C., Duytschaever, L., Tapprest, J., Fretin, D., H'ébert, L., Laugier, C., and Petry, S. 2017. Development of a multilocus sequence typing scheme for rhodococcus equi. Veterinary microbiology, 210: 64–70.
22. Emms, D. M. and Kelly, S. 2015. Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome biology, 16(1): 157.
23. Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, 20(1), 1-14.
24. Fair, R. J. and Tor, Y. 2014. Antibiotics and bacterial resistance in the 21st century. Perspectives in medicinal chemistry, 6: PMC–S14459.
25. Felsenstein, J. 1985. Phylogenies and the comparative method. The American Naturalist, 125(1): 1-15.
26. Felsenstein, J. 1993. PHYLIP (phylogeny inference package), version 3.5 c. Joseph Felsenstein.
27. Fischbach, M. A. and Walsh, C. T. 2009. Antibiotics for emerging pathogens. Science, 325(5944): 1089–1093.
28. Frothingham, R. and Wilson, K. H. 1993. Sequence-based differentiation of strains in the mycobacterium avium complex. Journal of Bacteriology, 175(10): 2818–2825.
29. Gagneux, S. 2018. Ecology and evolution of mycobacterium tuberculosis. Nature Reviews Microbiology, 16(4): 202.
30. Gao, B. and Gupta, R. S. 2012. Phylogenetic framework and molecular signatures for the main clades of the phylum actinobacteria. Microbiol. Mol. Biol. Rev., 76(1): 66–112.
31. Glaeser, S. P. and Kämpfer, P. 2015. Multilocus sequence analysis (mlsa) in prokaryotic taxonomy. Systematic and applied microbiology, 38(4): 237–245.
32. Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. 2002. Prokaryotic evolution in light of gene transfer. Molecular biology and evolution, 19(12): 2226–2238.

33. Hommeez, J., Devriese, L. A., Vaneechoutte, M., Riegel, P., Butaye, P., and Haesebrouck, F. 1999. Identification of nonlipophilic corynebacteria isolated from dairy cows with mastitis. Journal of clinical microbiology, 37(4): 954–957.
34. Jeong, H., Arif, B., Caetano-Anollés, G., Kim, K. M., & Nasir, A. (2019). Horizontal gene transfer in human-associated microorganisms inferred by phylogenetic reconstruction and reconciliation. *Scientific reports*, 9(1), 1-18.
35. Kang, Y., Takeda, K., Yazawa, K., and Mikami, Y. 2009. Phylogenetic studies of gordonia species based on gyrB and secA1 gene analyses. Mycopathologia, 167(2): 95– 105.
36. Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. 2002. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic acids research, 30(14): 3059–3066.
37. Keilhauer, C., Eggeling, L., and Sahm, H. 1993. Isoleucine synthesis in corynebacterium glutamicum: molecular analysis of the ilvB-ilvN-ilvC operon. Journal of bacteriology, 175(17): 5595–5603.
38. Kitahara, K. and Miyazaki, K. 2013. Revisiting bacterial phylogeny: natural and experimental evidence for horizontal gene transfer of 16s rRNA. Mobile genetic elements, 3(1): e24210.
39. Klappenbach, J. A., Dunbar, J. M., and Schmidt, T. M. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Applied and Environmental Microbiology , 66(4): 1328–1333.
40. Koch, A. and Mizrahi, V. 2018. Mycobacterium tuberculosis. Trends in microbiology, 26(6): 555–556.
41. Langille, M. G., Hsiao, W. W., and Brinkman, F. S. 2010. Detecting genomic islands using bioinformatics approaches. Nature Reviews Microbiology, 8(5): 373.
42. Larkin, M. J., Kulakov, L. A., and Allen, C. C. 2005. Biodegradation and rhodococcus—masters of catabolic versatility. Current opinion in Biotechnology, 16(3): 282–290.
43. Larson, A. 1998. The comparison of morphological and molecular data in phylogenetic systematics. In Molecular approaches to ecology and evolution, pages 275–296. Springer.

44. Lasek-Nesselquist, E., Gogarten, J. P., and Lapierre, P. 2012. The impact of HGT on phylogenomic reconstruction methods. Briefings in Bioinformatics, 15(1): 79–90.
45. Lerat, E., Daubin, V., Ochman, H., & Moran, N. A. (2005). Evolutionary origins of genomic repertoires in bacteria. *PLoS biology*, 3(5).
46. Letunic, I. and Bork, P. 2019. Interactive tree of life (itol) v4: recent updates and new developments. Nucleic acids research.
47. Maglott, D. R., Pruitt, K. D., and Tatusova, T. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 33(Suppl 1): D501–D504.
48. Mills, A. E., Mitchell, R. D., and Lim, E. K. 1997. *Corynebacterium pseudotuberculosis* is a cause of human necrotising granulomatous lymphadenitis. Pathology, 29(2): 231–233.
49. Mirarab, S. and Warnow, T. 2015. Astral-ii: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics, 31(12): i44–i52.
50. Miyazaki, K., Sato, M., and Tsukuda, M. 2017. Pcr primer design for 16s rrnas for experimental horizontal gene transfer test in escherichia coli. Frontiers in bioengineering and biotechnology, 5: 14.
51. Moret, B. M., Wang, L.-S., Warnow, T., and Wyman, S. K. 2001. New approaches for reconstructing phylogenies from gene order data. Bioinformatics, 17(suppl 1): S165–S173.
52. Moret, B. M., Lin, Y., and Tang, J. 2013. Rearrangements in phylogenetic inference: Compare, model, or encode? In Models and Algorithms for Genome Evolution, pages 147–171. Springer.
53. Nelson, G. J. (1970). Outline of a theory of comparative biology. *Systematic Zoology*, 19(4), 373-384.
54. Oliveira, A., Teixeira, P., Azevedo, M., Jamal, S. B., Tiwari, S., Almeida, S., Silva, A., Barh, D., Dorneles, E. M. S., Haas, D. J., et al. 2016. *Corynebacterium pseudotuberculosis* may be under anagenesis and biovar equi forms biovar ovis: a

- phylogenetic inference from sequence and structural analysis. BMC microbiology, 16(1): 100.
55. Ou, H.-Y., He, X., Harrison, E. M., Kulasekara, B. R., Thani, A. B., Kadioglu, A., Lory, S., Hinton, J. C., Barer, M. R., Deng, Z., et al. 2007. Mobilomefinder: web-based tools for in silico and experimental discovery of bacterial genomic islands. Nucleic acids research, 35(suppl 2): W97–W104.
56. Pascual, C., Lawson, P. A., Farrow, J. A., GIMENEZ, M. N., and Collins, M. D. 1995. Phylogenetic analysis of the genus corynebacterium based on 16s rna gene sequences. International journal of systematic and evolutionary microbiology, 45(4): 724–728.
57. Prescott, J. F. 1991. Rhodococcus equi: an animal and human pathogen. Clinical microbiology reviews, 4(1): 20–34.
58. Price, M. N., Dehal, P. S., and Arkin, A. P. 2010. Fasttree 2—approximately maximum-likelihood trees for large alignments. PloS one, 5(3): e9490.
59. Rajendhran, J. and Gunasekaran, P. 2011. Microbial phylogeny and diversity: small subunit ribosomal rna sequence analysis and beyond. Microbiological research, 166(2): 99–110.
60. Ramanan, P., Deziel, P. J., and Wengenack, N. L. 2013. Gordonia bacteremia. Journal of clinical microbiology, 51(10): 3443–3447.
61. Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. PLoS genetics, 10(5)
62. Ravenhall, M., kunca, N., Lassalle, F., and Dessimoz, C. 2015. Inferring horizontal gene transfer. PLOS Computational Biology, 11(5): 1–16.
63. Retief, J. D. 2000. Phylogenetic analysis using phylip. In Bioinformatics methods and protocols, pages 243–258. Springer.
64. Riegel, P., Ruimy, R., De Briel, D., Prevost, G., Jehl, F., Christen, R., and Monteil, H. 1995. Taxonomy of corynebacterium diphtheriae and related taxa, with recognition of corynebacterium ulcerans sp. nov. nom. rev. FEMS microbiology letters, 126(3): 271–276.
65. Rogall, T., Wolters, J., Flohr, T., and Bottger, E. C. 1990. Towards a phylogeny and definition of species at the molecular level within the genus mycobacterium.

- International Journal of Systematic and Evolutionary Microbiology, 40(4):323–330.
66. Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960), 798.
67. Ruiz, J. C., D'Afonseca, V., Silva, A., Ali, A., Pinto, A. C., Santos, A. R., Rocha, A. A., Lopes, D. O., Dorella, F. A., Pacheco, L. G., et al. 2011. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two corynebacterium pseudotuberculosis strains. PloS one, 6(4): e18551.
68. Sayyari, E., & Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. Molecular biology and evolution, 33(7), 1654-1668.
69. Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4): 406–425.
70. Sangal, V., Goodfellow, M., Jones, A. L., Seviour, R. J., and Sutcliffe, I. C. 2019. Refined systematics of the genus rhodococcus based on whole genome analyses. In Biology of Rhodococcus, pages 1–21. Springer.
71. Sankoff, D. and Blanchette, M. 1998. Multiple genome rearrangement and breakpoint phylogeny. Journal of Computational Biology, 5(3): 555–570.
72. Schouls, L. M., Schot, C. S., and Jacobs, J. A. 2003. Horizontal transfer of segments of the 16s rRNA genes between species of the streptococcus anginosus group. Journal of Bacteriology, 185(24): 7241–7246.
73. Sen, A., Daubin, V., Abrouk, D., Gifford, I., Berry, A. M., and Normand, P. 2014. Phylogeny of the class actinobacteria revisited in the light of complete genomes. the orders frankiales and micrococcales should be split into coherent entities: proposal of frankiales ord. nov., geodermatophilales ord. nov., acidothermales ord. nov. and nakamurellales ord. nov. International journal of systematic and evolutionary microbiology, 64(11): 3821–3832.
74. Soares, S. C., Silva, A., Trost, E., Blom, J., Ramos, R., Carneiro, A., Ali, A., Santos, A. R., Pinto, A. C., Diniz, C., et al. 2013. The pan-genome of the animal pathogen

- corynebacterium pseudotuberculosis reveals differences in genome plasticity between the biovar ovis and equi strains. PLoS One, 8(1): e53818.
75. Soucy, S. M., Huang, J., and Gogarten, J. P. 2015. Horizontal gene transfer: building the web of life. Nature Reviews Genetics, 16(8): 472.
 76. Sowani, H., Kulkarni, M., Zinjarde, S., and Javdekar, V. 2017. Gordonia and related genera as opportunistic human pathogens causing infections of skin, soft tissues, and bones. In The Microbiology of Skin, Soft Tissue, Bone and Joint Infections, pages 105–121. Elsevier.
 77. Stahl, D. A. and Urbance, J. 1990. The division between fast- and slow-growing species corresponds to natural relationships among the mycobacteria. Journal of bacteriology, 172(1): 116–124.
 78. Stamatakis, A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9): 1312–1313.
 79. Suyama, M. and Bork, P. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends in Genetics, 17(1): 10 – 13.
 80. Tauch, A., Kaiser, O., Hain, T., Goesmann, A., Weisshaar, B., Albersmeier, A., Bekel, T., Bischoff, N., Brune, I., Chakraborty, T., et al. 2005. Complete genome sequence and analysis of the multiresistant nosocomial pathogen corynebacterium jeikeium k411, a lipid-requiring bacterium of the human skin flora. Journal of bacteriology, 187(13): 4671–4682.
 81. Vachaspati, P. and Warnow, T. 2015. Astrid: Accurate species trees from internode distances. BMC Genomics, 16(10): S3.
 82. Van Duin, D. and Paterson, D. L. 2016. Multidrug-resistant bacteria in the community: trends and lessons learned. Infectious Disease Clinics, 30(2): 377–390.
 83. Von Graevenitz, A. and Bernard, K. 2006. The genus corynebacterium—medical. The Prokaryotes: Volume Archaea. Bacteria: Firmicutes, Actinomycetes, pages 819–842.
 84. Wolf, Y. I., Rogozin, I. B., Grishin, N. V., and Koonin, E. V. 2002. Genome trees and the tree of life. Trends in Genetics, 18(9): 472 – 479.

85. Yang, J. and Yang, S. 2017. Comparative analysis of *Corynebacterium glutamicum* genomes: a new perspective for the industrial production of amino acids. BMC genomics, 18(1): 940.
86. Yap, W. H., Zhang, Z., and Wang, Y. 1999. Distinct types of rna operons exist in the genome of the actinomycete *thermomonospora chromogena* and evidence for horizontal transfer of an entire rna operon. Journal of bacteriology, 181(17): 5201–5209.
87. Zhi, X.-Y., Jiang, Z., Yang, L.- L., and Huang, Y. 2017. The underlying mechanisms of genetic innovation and speciation in the family *corynebacteriaceae*: A phylogenomics approach. Molecular phylogenetics and evolution, 107: 246–255.

Supplementary Material

Table S1. Putative orthology groups used in Step 4 for computing the preliminary species tree

Group ID	Product
OG1103	50S Ribossomal protein L23
OG1104	50S Ribossomal protein L2
OG1105	30S Ribossomal protein S19
OG1106	50S Ribossomal protein L6
OG1107	50S Ribossomal protein L18
OG1108	30S Ribossomal protein S5
OG1113	ATP synthase subunit gamma
OG1116	phosphoglycerate kinase
OG1121	50S Ribossomal protein L35
OG1122	50S Ribossomal protein L20
OG1136	GTPase ObgE
OG1137	50S Ribosomal protein L27
OG1152	CarD TF regulator

Table S2. Number of gene trees in the 5 collections of trees considered in Step 5.

Max. number of missing genomes	Number of gene trees
72 (20%)	44
144 (40%)	46
180 (50%)	73
216 (60%)	93
288 (80%)	360

Table S3. Percentage of conserved clades between the consensus trees and the initial phylogenies reconstructed using ASTRID and ASTRAL.

	ASTRID 20	ASTRID 40	ASTRID 50	ASTRID 60	ASTRID 80
ASTRID Consensus	89.14	91.05	94.57	95.85	87.22

	ASTRAL 20	ASTRID 40	ASTRID 50	ASTRID 60	ASTRAL 80
ASTRAL Consensus	87.58	90.68	96.27	95.96	92.86

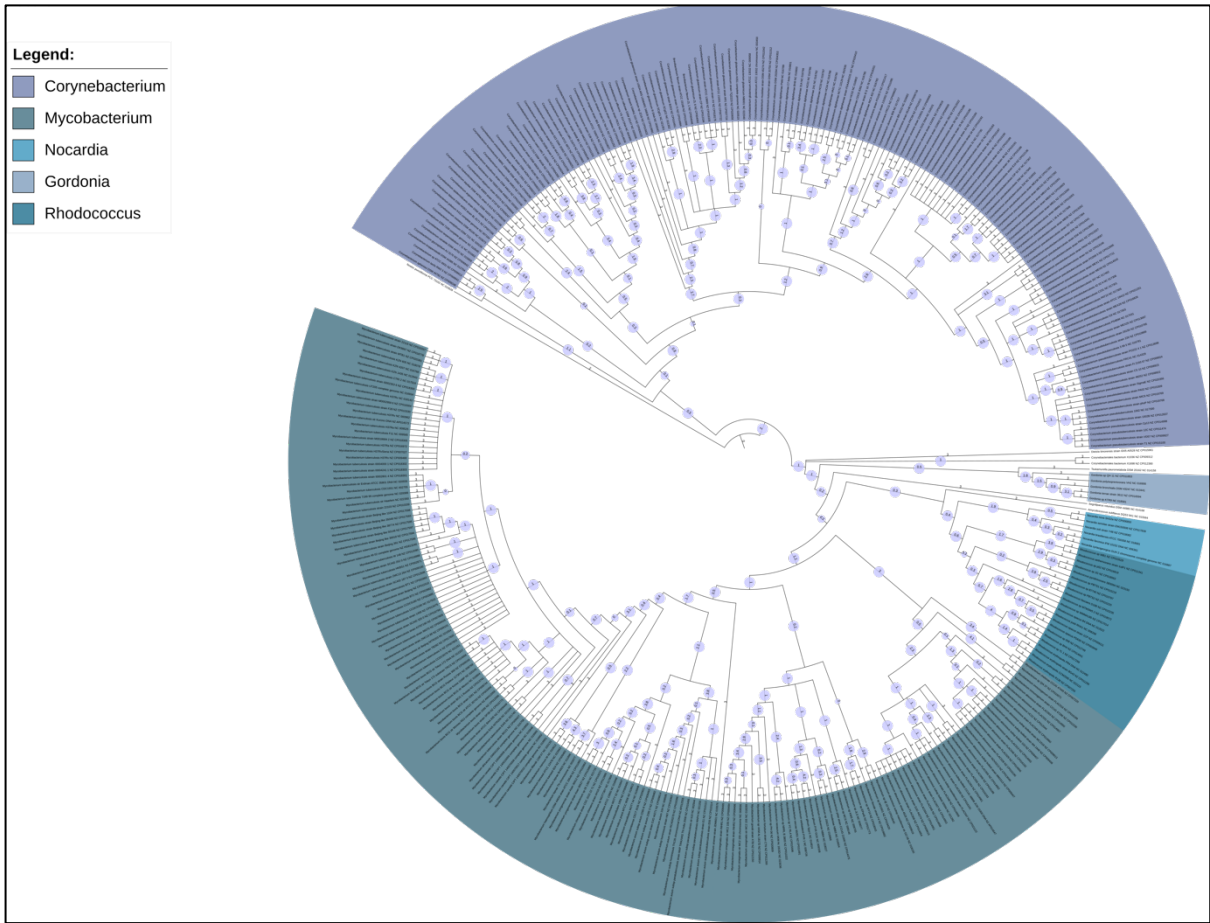


FIG. S1. Corynebacteriales preliminary species tree estimated using the concatenation of multiple sequence alignments of 13 putative orthology groups and RaXML (Stamatakis, 2014). The species belonging to the same genus are represented with the same color (Figure obtained using iTOL (Letunic and Bork, 2019)).

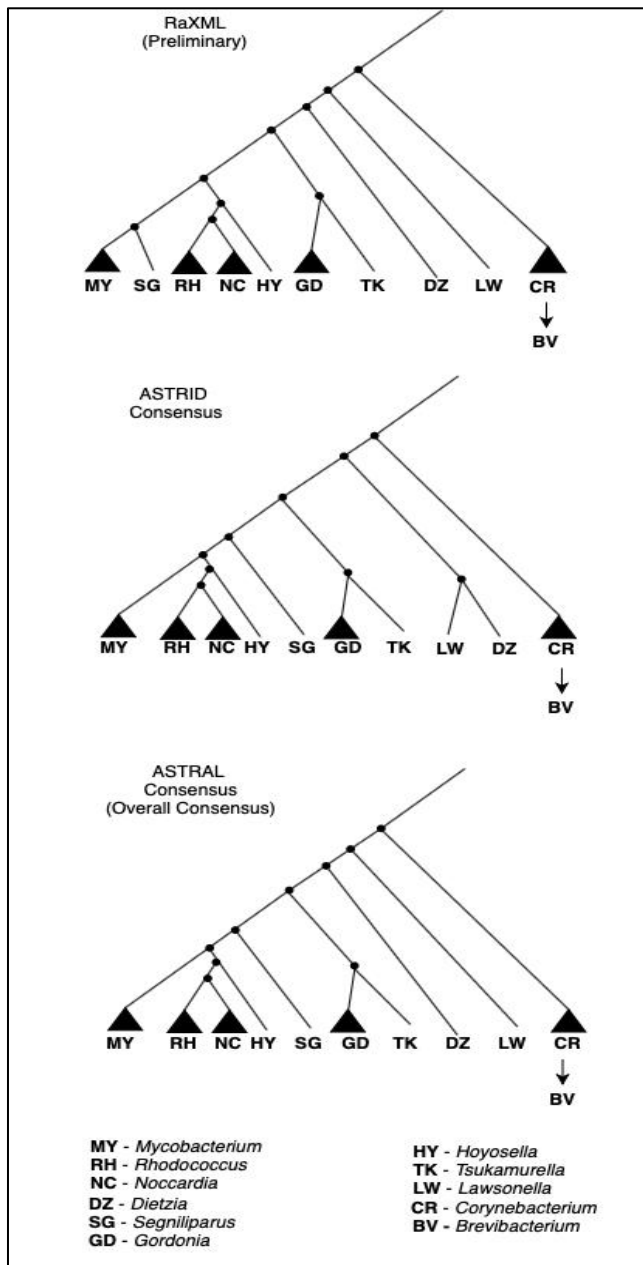


FIG. S2. Illustration, at the genus level, of the 4 consensus trees estimated in our work. A) phylogenomic tree reconstructed using the alignment of 13 single copy homology groups and RaXML; B) Consensus tree of ASTRID dataset; C) Consensus tree of ASTRAL dataset which is also the overall consensus between RaXML, ASTRID Consensus and ASTRAL Consensus. Genera for which the dataset contains more than one species are represented as triangles.

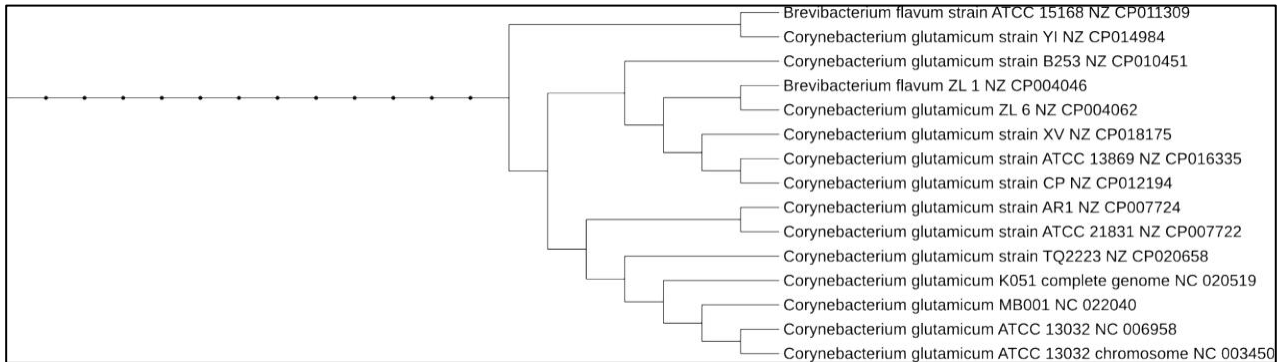


FIG. S3. Classification of *Brevibacterium* inside *Corynebacterium glutamicum*.

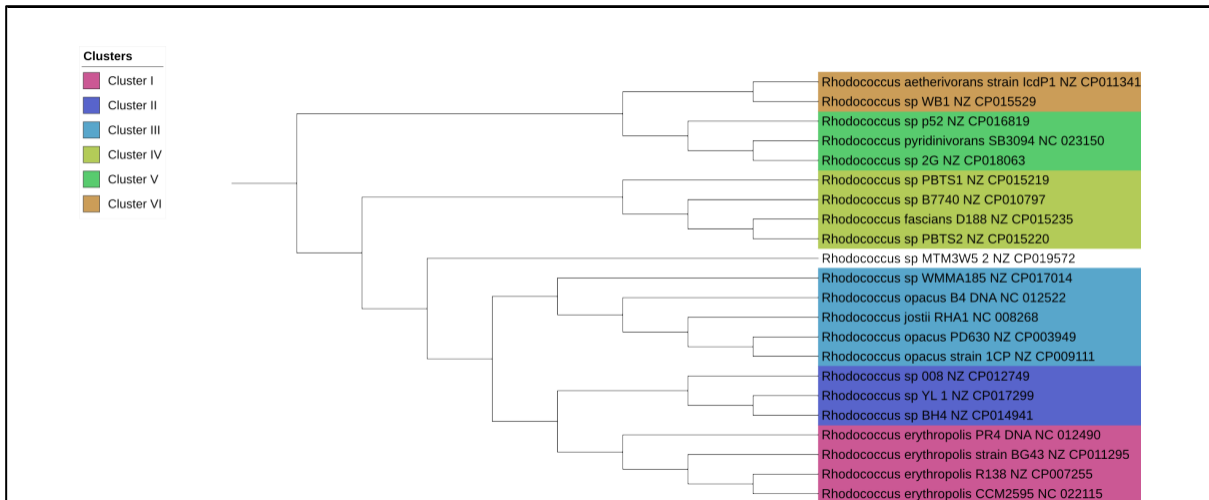


FIG. S4. Estimated phylogeny for *Rhodococcus* genus, with 6 clusters. Detailed method used to get this phylogeny (overall consensus tree: RaXML + ASTRID consensus+ ASTRAL consensus).

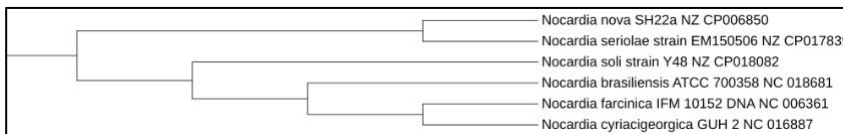


FIG. S5.. Estimated phylogeny for *Nocardia* genus.

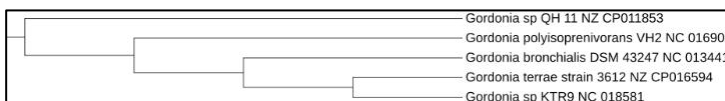


FIG. S6. Estimated phylogeny for *Gordonia* genus.

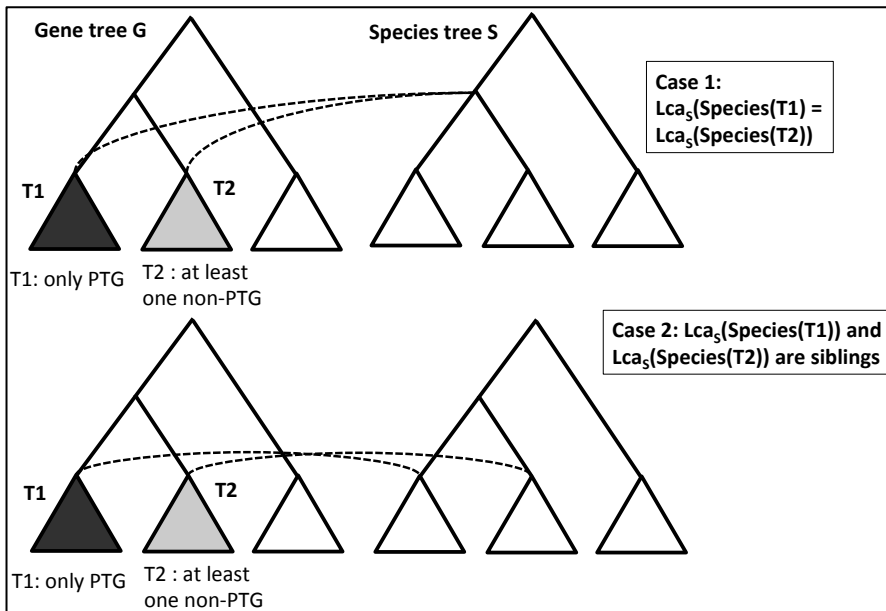


FIG. S7. Illustration of the phylogenetic method used in Step 5 to reclassify some putative transferred genes as vertically inherited genes. The two cases in which putative transferred genes are reclassified are depicted.

Supplementary File A1 – Number of genomes per species

Number of genomes per species		
Genus	Species	Number of genomes
<i>Lawsonella</i>	<i>clevelandensis</i>	2
<i>Hoyosella</i>	<i>subflava</i>	1
<i>Rhodococcus</i>	<i>erythropolis</i>	4
<i>Rhodococcus</i>	<i>fascians</i>	1
<i>Rhodococcus</i>	<i>jostii</i>	1
<i>Rhodococcus</i>	<i>pyridinivorans</i>	1
<i>Rhodococcus</i>	<i>aetherivorans</i>	1
<i>Rhodococcus</i>	<i>sp.</i>	11
<i>Rhodococcus</i>	<i>opacus</i>	3
<i>Mycobacterium</i>	<i>vaccae</i>	1
<i>Mycobacterium</i>	<i>tuberculosis</i>	57
<i>Mycobacterium</i>	<i>kansasii</i>	2
<i>Mycobacterium</i>	<i>avium</i>	12
<i>Mycobacterium</i>	<i>colombiense</i>	1
<i>Mycobacterium</i>	<i>intracellulare</i>	3
<i>Mycobacterium</i>	<i>goodii</i>	1
<i>Mycobacterium</i>	<i>sinense</i>	1
<i>Mycobacterium</i>	<i>leprae</i>	2
<i>Mycobacterium</i>	<i>abscessus</i>	29
<i>Mycobacterium</i>	<i>marinum</i>	2
<i>Mycobacterium</i>	<i>liflandii</i>	1
<i>Mycobacterium</i>	<i>chimaera</i>	2
<i>Mycobacterium</i>	<i>immunogenum</i>	2
<i>Mycobacterium</i>	<i>paraintracellulare</i>	1
<i>Mycobacterium</i>	<i>bovis</i>	13
<i>Mycobacterium</i>	<i>chubuense</i>	1
<i>Mycobacterium</i>	<i>africanum</i>	1
<i>Mycobacterium</i>	<i>yongonense</i>	3
<i>Mycobacterium</i>	<i>indicus</i>	1
<i>Mycobacterium</i>	<i>rhodesiae</i>	1
<i>Mycobacterium</i>	<i>smegmatis</i>	5
<i>Mycobacterium</i>	<i>haemophilum</i>	1
<i>Mycobacterium</i>	<i>sp.</i>	15
<i>Mycobacterium</i>	<i>phlei</i>	1
<i>Mycobacterium</i>	<i>vanbaalenii</i>	1
<i>Mycobacterium</i>	<i>canettii</i>	5
<i>Mycobacterium</i>	<i>gilvum</i>	2

<i>Mycobacterium</i>	<i>fortuitum</i>	1
<i>Mycobacterium</i>	<i>neoaurum</i>	1
<i>Mycobacterium</i>	<i>litorale</i>	1
<i>Mycobacterium</i>	<i>chelonae</i>	1
<i>Dietzia</i>	<i>timorensis</i>	1
<i>Tsukamurella</i>	<i>paurometabola</i>	1
<i>Corynebacterium</i>	<i>atypicum</i>	1
<i>Corynebacterium</i>	<i>doosanense</i>	1
<i>Corynebacterium</i>	<i>imitans</i>	1
<i>Corynebacterium</i>	<i>ulcerans</i>	11
<i>Corynebacterium</i>	<i>callunae</i>	1
<i>Corynebacterium</i>	<i>terpenotabidum</i>	1
<i>Corynebacterium</i>	<i>testudinoris</i>	1
<i>Corynebacterium</i>	<i>uterequi</i>	1
<i>Corynebacterium</i>	<i>epidermidicanis</i>	1
<i>Corynebacterium</i>	<i>ammoniagenes</i>	2
<i>Corynebacterium</i>	<i>kroppenstedtii</i>	1
<i>Corynebacterium</i>	<i>urealyticum</i>	2
<i>Corynebacterium</i>	<i>camporealensis</i>	1
<i>Corynebacterium</i>	<i>kutscheri</i>	1
<i>Corynebacterium</i>	<i>simulans</i>	2
<i>Corynebacterium</i>	<i>vitaeruminis</i>	1
<i>Corynebacterium</i>	<i>phocae</i>	1
<i>Corynebacterium</i>	<i>resistens</i>	1
<i>Corynebacterium</i>	<i>aquilae</i>	1
<i>Corynebacterium</i>	<i>efficiens</i>	1
<i>Corynebacterium</i>	<i>argenteratense</i>	1
<i>Corynebacterium</i>	<i>humireducens</i>	1
<i>Corynebacterium</i>	<i>glyciniphilum</i>	1
<i>Corynebacterium</i>	<i>crudilactis</i>	1
<i>Corynebacterium</i>	<i>diphtheriae</i>	15
<i>Corynebacterium</i>	<i>marinum</i>	2
<i>Corynebacterium</i>	<i>ureicelerivorans</i>	1
<i>Corynebacterium</i>	<i>lactis</i>	1
<i>Corynebacterium</i>	<i>deserti</i>	1
<i>Corynebacterium</i>	<i>casei</i>	1
<i>Corynebacterium</i>	<i>singulare</i>	1
<i>Corynebacterium</i>	<i>halotolerans</i>	1
<i>Corynebacterium</i>	<i>glutamicum</i>	17
<i>Corynebacterium</i>	<i>maris</i>	1
<i>Corynebacterium</i>	<i>frankenforstense</i>	1
<i>Corynebacterium</i>	<i>sphenisci</i>	1
<i>Corynebacterium</i>	<i>pseudotuberculosis</i>	60
<i>Corynebacterium</i>	<i>aurimucosum</i>	1
<i>Corynebacterium</i>	<i>striatum</i>	1
<i>Corynebacterium</i>	<i>flavescens</i>	1
<i>Corynebacterium</i>	<i>stationis</i>	3
<i>Corynebacterium</i>	<i>sp.</i>	2

<i>Corynebacterium</i>	<i>falsenii</i>	1
<i>Corynebacterium</i>	<i>variabile</i>	1
<i>Corynebacterium</i>	<i>jeikeium</i>	1
<i>Corynebacterium</i>	<i>mustelae</i>	1
[<i>Brevibacterium</i>]	<i>flavum</i>	2
<i>Nocardia</i>	<i>farcinica</i>	1
<i>Nocardia</i>	<i>seriolae</i>	1
<i>Nocardia</i>	<i>solii</i>	1
<i>Nocardia</i>	<i>nova</i>	1
<i>Nocardia</i>	<i>cyriacigeorgica</i>	1
<i>Nocardia</i>	<i>brasiliensis</i>	1
<i>Gordonia</i>	<i>polyisoprenivorans</i>	1
<i>Gordonia</i>	<i>bronchialis</i>	1
<i>Gordonia</i>	<i>terrae</i>	1
<i>Gordonia</i>	<i>sp.</i>	2
<i>Segniliparus</i>	<i>rotundus</i>	1

Supplementary File A2 – Number of genomes per species

Size and number of CDS per genome

Genome ID	Definition	Chromosome Size (bp)	Number of CDS
NZ_AM412059	Mycobacterium bovis BCG str. Moreau RDJ complete genome.	4340116	4158
NZ_CP010818	Corynebacterium ulcerans strain 131001	2483321	2196
NC_010545	Corynebacterium urealyticum DSM 7109 complete genome.	2369219	1996
NC_020302	Corynebacterium halotolerans YIM 70093 = DSM 44683	3135752	2794
NZ_CP016640	Mycobacterium sp. djl-10	6395946	6085
NC_016906	Gordonia polyisoprenivorans VH2	5669805	5014
NZ_CP015529	Rhodococcus sp. WB1	5924026	5301
NC_021054	Mycobacterium tuberculosis str. Beijing/NITR203	4411128	4244
NZ_CP015961	Dietzia timorensis strain ID05-A0528	3607892	3265
NZ_CP011883	Mycobacterium haemophilum DSM 44634 strain ATCC 29548	4235765	3976
NZ_CP017014	Rhodococcus sp. WMMA185	4444448	3949
NZ_CP015773	Mycobacterium bovis strain SP38	4347648	4165
NZ_CP018301	Mycobacterium tuberculosis strain I0002801-4	4376067	4203
NZ_CP009583	Corynebacterium ulcerans strain 210931	2503722	2238
NZ_CP016188	Mycobacterium abscessus strain FLAC006	4891993	4810
NC_009525	Mycobacterium tuberculosis H37Ra	4419977	4245
NZ_CP009480	Mycobacterium tuberculosis H37Rv	4396119	4232
NC_002935	Corynebacterium diphtheriae NCTC 13129	2488635	2310
NC_008769	Mycobacterium bovis BCG Pasteur 1173P2	4374522	4186
NC_021663	Corynebacterium terpenotabidum Y-11	2751233	2376
NZ_CP010113	Mycobacterium avium subsp. paratuberculosis strain E1	4781002	4504
NZ_CP020410	Corynebacterium diphtheriae strain FDAARGOS_197	2489000	2305
NC_021915	Corynebacterium maris DSM 45190	2787574	2536
NZ_CP016829	Corynebacterium pseudotuberculosis strain MB20	2370901	2128

NZ_CP009245	<i>Corynebacterium aquilae</i> DSM 44791 strain S-613	2926437	2397
NC_023150	<i>Rhodococcus pyridinivorans</i> SB3094	5227080	4748
NZ_CP013262	<i>Corynebacterium pseudotuberculosis</i> strain MB30	2364377	2124
NZ_CP013741	<i>Mycobacterium bovis</i> strain BCG-1 (Russia)	4370705	4185
NZ_CP020658	<i>Corynebacterium glutamicum</i> strain TQ2223	3306634	3039
NZ_CP005286	<i>Corynebacterium humireducens</i> NBRC 106098 = DSM 45392	2681312	2526
NZ_CP014279	<i>Corynebacterium stationis</i> strain ATCC 6872	2853666	2585
NC_016946	<i>Mycobacterium intracellulare</i> ATCC 13950	5402402	5065
NZ_CP014956	<i>Mycobacterium abscessus</i> strain FLAC029	5188507	5151
NZ_CP014635	<i>Corynebacterium simulans</i> strain Wattiau	2598702	2365
NZ_CP009244	<i>Corynebacterium ammoniagenes</i> DSM 20306 strain 9.6	2790185	2482
NZ_CP009615	<i>Mycobacterium abscessus</i> strain DJO-44274	4686331	4622
NZ_CP010889	<i>Corynebacterium pseudotuberculosis</i> strain 226	2337820	2079
NC_017303	<i>Corynebacterium pseudotuberculosis</i> I19	2337594	2079
NC_017462	<i>Corynebacterium pseudotuberculosis</i> 267	2337628	2077
NC_016604	<i>Mycobacterium rhodesiae</i> NBB3	6415739	6200
NZ_CP014951	<i>Mycobacterium abscessus</i> strain FLAC004	5242371	5281
NZ_CP014955	<i>Mycobacterium abscessus</i> strain FLAC013	5074222	5013
NC_019965	<i>Mycobacterium canettii</i> CIPT 140070008 complete genome.	4420197	4203
NC_017730	<i>Corynebacterium pseudotuberculosis</i> 31	2402956	2159
NC_020245	<i>Mycobacterium bovis</i> BCG str. Korea 1168P	4376711	4190
NZ_CP007255	<i>Rhodococcus erythropolis</i> R138	6236862	5688
NC_021352	<i>Corynebacterium glutamicum</i> SCgG2	3350619	3049
NC_018581	<i>Gordonia</i> sp. KTR9	5441391	4838
NZ_CP020809	<i>Mycobacterium</i> sp. PH-06	7595921	7225
NZ_CP004353	<i>Corynebacterium vitaeruminis</i> DSM 20294	2931780	2569
NZ_CP003494	<i>Mycobacterium bovis</i> BCG str. ATCC 35743	4334064	4175
NC_018681	<i>Nocardia brasiliensis</i> ATCC 700358	9436348	8434
NZ_CP018303	<i>Mycobacterium tuberculosis</i> strain I0004241-1	4386132	4207
NC_014168	<i>Segniliparus rotundus</i> DSM 44985	3157527	3033
NZ_CP007027	<i>Mycobacterium tuberculosis</i> H37RvSiena	4410911	4236
NC_000962	<i>Mycobacterium tuberculosis</i> H37Rv	4411532	3906

NZ_CP009312	Corynebacteriales bacterium X1036	1860551	1559
NZ_CP008744	Mycobacterium bovis BCG strain 3281	4410431	4217
NC_008726	Mycobacterium vanbaalenii PYR-1	6491865	6142
NZ_CP016192	Mycobacterium abscessus strain FLAC046	5214168	5180
NC_002944	Mycobacterium avium subsp. paratuberculosis str. k10	4829781	4510
NC_008146	Mycobacterium sp. MCS	5705448	5468
NZ_CP010827	Corynebacterium singulare strain IBS B52218	2830519	2556
NZ_CP004350	Corynebacterium casei LMG S-19264	3113488	2786
NZ_CP009211	Corynebacterium imitans strain DSM 44264	2565321	2336
NZ_CP015186	Corynebacterium pseudotuberculosis strain 36	2403412	2159
NZ_CP012695	Corynebacterium pseudotuberculosis strain PO269-5	2337124	2081
NZ_CP009449	Mycobacterium bovis strain ATCC BAA-935	4358088	4189
NC_007164	Corynebacterium jeikeium K411 complete genome.	2462499	2116
NZ_CP015495	Mycobacterium avium subsp. paratuberculosis strain MAP/TANUVAS/TN/India/2008	4829781	4501
NZ_CP009251	Corynebacterium stationis strain 622=DSM 20302	2808767	2520
NC_012207	Mycobacterium bovis BCG str. Tokyo 172 DNA	4371711	4183
NZ_CP019420	Mycobacterium sp. MS1601	6407860	6073
NZ_CP007722	Corynebacterium glutamicum strain ATCC 21831	3176076	2933
NC_016790	Corynebacterium diphtheriae VA01	2395441	2211
NZ_CP016335	Corynebacterium glutamicum strain ATCC 13869	3296500	3024
NC_002677	Mycobacterium leprae TN chromosome	3268203	1605
NZ_CP015187	Corynebacterium pseudotuberculosis strain 38	2403515	2161
NZ_CP014998	Corynebacterium pseudotuberculosis strain Cp13	2342237	2085
NZ_CP016888	Mycobacterium tuberculosis strain SCAID 252.0	4439387	4257
NZ_CP002882	Mycobacterium tuberculosis BT2	4401899	4228
NZ_CP009613	Mycobacterium abscessus subsp. bolletii strain MC1518	5049258	5003
NZ_CP013327	Corynebacterium pseudotuberculosis strain PA01	2337920	2077
NZ_CP012150	Mycobacterium goodii strain X7B	7105933	6662
NC_018101	Corynebacterium ulcerans 0102 DNA	2579188	2302
NZ_CP017291	Corynebacterium pseudotuberculosis strain MEX30	2368140	2122
NZ_CP010330	Mycobacterium tuberculosis strain F28	4421903	4253
NC_016787	Corynebacterium diphtheriae HC03	2478364	2284
NC_004369	Corynebacterium efficiens YS-314 DNA	3147090	2819
NC_020133	Mycobacterium liflandii 128FXT	6208955	5477
NC_013441	Gordonia bronchialis DSM 43247	5208602	4800
NC_008595	Mycobacterium avium 104	5475491	5147

NC_008268	<i>Rhodococcus jostii</i> RHA1	7804765	7197
NZ_CP014941	<i>Rhodococcus</i> sp. BH4	6314891	5787
NZ_CP009101	<i>Mycobacterium tuberculosis</i> strain ZMC13-88	4411515	4227
NZ_CP012837	<i>Corynebacterium pseudotuberculosis</i> strain 1002B	2335107	2078
NZ_CP018302	<i>Mycobacterium tuberculosis</i> strain I0004000-1	4365724	4201
NZ_CP007790	<i>Corynebacterium marinum</i> DSM 44953	2607268	2406
NZ_CP011541	<i>Corynebacterium epidermidicantis</i> strain DSM 45586	2692072	2417
NZ_CP009220	<i>Corynebacterium deserti</i> GIMN1.010	2972149	2682
NZ_CP015964	<i>Mycobacterium yongonense</i> strain Asan 36912	5445538	5074
NC_016768	<i>Mycobacterium tuberculosis</i> KZN 4207	4394985	4220
NZ_CP013698	<i>Corynebacterium pseudotuberculosis</i> strain PO222/4-1	2337508	2081
NZ_CP013697	<i>Corynebacterium pseudotuberculosis</i> strain MEX25	2337529	2079
NZ_CP019882	<i>Mycobacterium litorale</i> strain F4	6103712	5762
NC_009077	<i>Mycobacterium</i> sp. JLS	6048425	5794
NZ_CP017292	<i>Corynebacterium pseudotuberculosis</i> strain MEX31	2367880	2119
NC_008705	<i>Mycobacterium</i> sp. KMS	5737227	5512
NC_018612	<i>Mycobacterium indicus pranii</i> MTCC 9506	5589007	5214
NC_019950	<i>Mycobacterium canettii</i> CIPT 140060008 complete genome.	4432426	4219
NZ_CP013263	<i>Corynebacterium pseudotuberculosis</i> strain MB66	2372202	2127
NZ_CP018363	<i>Mycobacterium avium</i> subsp. <i>hominissuis</i> strain H87	5626623	5223
NZ_CP009246	<i>Corynebacterium flavescens</i> strain OJ8	2758653	2467
NC_015673	<i>Corynebacterium resistens</i> DSM 45100	2601311	2206
NZ_CP012390	<i>Corynebacteriales</i> bacterium X1698	1915154	1615
NZ_AP014547	<i>Mycobacterium abscessus</i> subsp. <i>bolletii</i> CCUG 48898 = JCM 15300 DNA	4978382	4944
NZ_CP017920	<i>Mycobacterium tuberculosis</i> strain TB282	4425860	4257
NC_016887	<i>Nocardia cyriacigeorgica</i> GUH-2 chromosome complete genome.	6194645	5544
NC_008596	<i>Mycobacterium smegmatis</i> str. MC2 155 chromosome	6988209	6717
NC_018143	<i>Mycobacterium tuberculosis</i> H37Rv	4411709	4235
NZ_CP016594	<i>Gordonia terrae</i> strain 3612	5701501	5035
NZ_CP011312	<i>Corynebacterium kutscheri</i> strain DSM 20755	2354065	2056
NZ_CP011295	<i>Rhodococcus erythropolis</i> strain BG43	6334075	5804
NZ_CP018305	<i>Mycobacterium tuberculosis</i> strain M0018684-2	4359825	4188
NC_019952	<i>Mycobacterium canettii</i> CIPT 140070017 complete genome.	4524466	4269
NC_009565	<i>Mycobacterium tuberculosis</i> F11	4424435	4237
NZ_CP014566	<i>Mycobacterium bovis</i> BCG str. Tokyo 172 substrain TRCS	4371707	4183
NZ_CP007156	<i>Corynebacterium falsenii</i> DSM 44353 strain BL 8171	2677607	2259

NZ_CP015622	Corynebacterium sp. JZ16	3047373	2733
NC_018150	Mycobacterium abscessus subsp. massiliense str. GO 06	4687873	4622
NC_017301	Corynebacterium pseudotuberculosis C231	2328208	2070
NZ_CP014961	Mycobacterium abscessus strain FLAC054	5330954	5355
NZ_CP011545	Corynebacterium testudinoris strain DSM 44614	2721226	2541
NZ_CP008922	Corynebacterium pseudotuberculosis strain 48252	2338139	2081
NZ_CP009716	Corynebacterium ulcerans strain 05146	2466435	2170
NC_017031	Corynebacterium pseudotuberculosis P54B96	2337657	2079
NZ_CP008924	Corynebacterium pseudotuberculosis strain Ft_2193/67	2338300	2081
NZ_CP008944	Corynebacterium atypicum strain R2070	2311380	2009
NC_015683	Corynebacterium ulcerans BR-AD22	2606374	2340
NZ_CP019221	Mycobacterium chimaera strain CDC 2015-22-71	6078402	5625
NZ_CP014341	Corynebacterium pseudotuberculosis strain E55	2335383	2076
NC_017945	Corynebacterium pseudotuberculosis 258	2369817	2126
NZ_CP009496	Mycobacterium smegmatis strain INHR2	6988302	6774
NC_017522	Mycobacterium tuberculosis CCDC5180	4405981	4233
NC_017904	Mycobacterium sp. MOTT36Y	5613626	5234
NC_012590	Corynebacterium aurimucosum ATCC 700975	2790189	2563
NC_016782	Corynebacterium diphtheriae 241	2426551	2257
NZ_CP015189	Corynebacterium pseudotuberculosis strain 43	2365075	2116
NC_018078	Mycobacterium tuberculosis KZN 605	4399120	4225
NZ_CP021252	Corynebacterium striatum strain KC-Na-01	2758551	2550
NZ_CP012194	Corynebacterium glutamicum strain CP	3342897	3072
NC_022115	Rhodococcus erythropolis CCM2595	6281198	5725
NZ_CP012044	Mycobacterium abscessus UC22	5257136	5226
NZ_CP010797	Rhodococcus sp. B7740	5341557	4956
NZ_CP019705	Corynebacterium ammoniagenes strain KCCM 40472	2808265	2498
NZ_CP015192	Corynebacterium pseudotuberculosis strain 34	2403454	2162
NZ_CP013475	Mycobacterium tuberculosis strain 1458	4402033	4226
NC_012943	Mycobacterium tuberculosis KZN 1435	4398250	4226
NC_018289	Mycobacterium smegmatis str. MC2 155	6988208	6787
NZ_CP016826	Corynebacterium pseudotuberculosis strain MEX29	2337866	2081
NC_017305	Corynebacterium pseudotuberculosis PAT10	2335323	2077
NZ_CP014634	Corynebacterium simulans strain PES1	2737971	2508
NZ_CP011510	Mycobacterium tuberculosis strain Beijing	4378588	4199
NZ_CP013049	Mycobacterium abscessus strain NOV0213	5173145	5156

NC_017524	Mycobacterium tuberculosis CTRI-2	4398525	4226
NZ_CP012022	Corynebacterium pseudotuberculosis strain 262	2361125	2099
NZ_CP015220	Rhodococcus sp. PBTS2	5179353	4784
NZ_CP014984	Corynebacterium glutamicum strain YI	3342103	3148
NZ_CP004062	Corynebacterium glutamicum ZL-6	3332458	3079
NZ_CP016972	Mycobacterium tuberculosis H37Ra	4426109	4248
NZ_CP018175	Corynebacterium glutamicum strain XV	3333639	3069
NZ_CP014952	Mycobacterium abscessus strain FLAC005	4869298	4772
NZ_CP017839	Nocardia seriolae strain EM150506	8304518	7637
NZ_CP013261	Corynebacterium pseudotuberculosis strain MB14	2370761	2127
NZ_CP016191	Mycobacterium abscessus strain FLAC030	4867257	4826
NC_011896	Mycobacterium leprae Br4923	3268071	2900
NZ_CP009483	Mycobacterium kansasii 824	6402301	5588
NZ_CP015965	Mycobacterium yongonense strain Asan 36527	5435152	5062
NC_021351	Corynebacterium glutamicum SCgG1	3350620	3050
NZ_CP010071	Mycobacterium sp. QIA-37	4855372	4673
NZ_CP007724	Corynebacterium glutamicum strain AR1	3145677	2912
NZ_CP011269	Mycobacterium fortuitum strain CT6	6254616	5950
NZ_CP020821	Mycobacterium colombiense CECT 3035	5581643	5236
NC_016804	Mycobacterium bovis BCG str. Mexico	4350386	4162
NZ_CP016396	Mycobacterium avium strain RCAD0278	4953610	4595
NZ_CP015191	Corynebacterium pseudotuberculosis strain 48	2403301	2160
NZ_CP009427	Mycobacterium tuberculosis strain 96121	4410945	4232
NZ_CP011022	Mycobacterium sp. NRRL B-3805	5421338	5049
NZ_CP012749	Rhodococcus sp. 008	6570200	6024
NC_009338	Mycobacterium gilvum PYR-GCK	5619607	5311
NZ_CP017299	Rhodococcus sp. YL-1	6367154	5845
NC_012704	Corynebacterium kroppenstedtii DSM 44385	2446804	2030
NZ_CP017384	Corynebacterium pseudotuberculosis strain I37	2370282	2108
NZ_HG813240	Mycobacterium tuberculosis 49-02 complete genome.	4412379	4234
NC_015564	Amycolicococcus subflavus DQS3-9A1	4738809	4371
NZ_CP006841	Corynebacterium lactis RW2-5	2769745	2405
NZ_CP011341	Rhodococcus aetherivorans strain IcdP1	5922748	5319
NZ_CP012136	Corynebacterium pseudotuberculosis strain E19	2367956	2120
NC_020089	Mycobacterium tuberculosis 7199-99 complete genome.	4421197	4237
NZ_CP009247	Corynebacterium frankenforstense DSM 45800 strain ST18	2604152	2193

NZ_CP014959	Mycobacterium abscessus strain FLAC048	4939234	4879
NZ_CP009215	Corynebacterium ureicelerivorans strain IMMIB RIV-2301	2279990	2174
NZ_CP018063	Rhodococcus sp. 2G	5231430	4866
NZ_CP009914	Mycobacterium sp. VKM Ac-1817D	6324222	6024
NC_002755	Mycobacterium tuberculosis CDC1551	4403837	4231
NC_022040	Corynebacterium glutamicum MB001	3079253	2825
NC_016785	Corynebacterium diphtheriae CDCE 8392	2433326	2270
NZ_CP019768	Corynebacterium pseudotuberculosis strain phoP	2339296	2088
NC_014814	Mycobacterium gilvum Spyr1	5547747	5235
NZ_CP009494	Mycobacterium smegmatis str. MC2 155	6988269	6774
NC_023036	Mycobacterium neoaurum VKM Ac-1815D	5421267	5048
NZ_CP009616	Mycobacterium abscessus strain 4529	4687494	4623
NZ_CP013699	Corynebacterium pseudotuberculosis strain E56	2335773	2077
NZ_CP015235	Rhodococcus fascians D188	5139988	4747
NZ_CP012506	Mycobacterium tuberculosis strain SCAID 187.0	4411829	4234
NZ_CP012885	Mycobacterium chimaera strain AH16	5852822	5410
NZ_CP010114	Mycobacterium avium subsp. paratuberculosis strain E93	4786065	4507
NZ_CP011546	Corynebacterium uterequi strain DSM 45634	2419437	2150
NZ_CP009407	Mycobacterium abscessus subsp. bolletii 103	5051394	4995
NZ_CP017594	Mycobacterium tuberculosis strain Beijing-like/36918	4441591	4260
NZ_CP016190	Mycobacterium abscessus strain FLAC028	5188101	5151
NC_022350	Mycobacterium tuberculosis str. Haarlem	4408224	4220
NZ_AP012555	Mycobacterium avium subsp. hominissuis TH135 chromosomal DNA	4951217	4581
NZ_CP013991	Corynebacterium glutamicum strain USDA-ARS-USMARC-56828	3245395	2976
NC_019951	Mycobacterium canettii CIPT 140070010 complete genome.	4525948	4270
NZ_CP007803	Mycobacterium tuberculosis K	4385518	4214
NZ_CP004046	[Brevibacterium] flavum ZL-1	3340941	3090
NC_017300	Corynebacterium pseudotuberculosis 1002	2335113	2078
NC_015758	Mycobacterium africanum GM041182 complete genome.	4389314	4192
NZ_CP016819	Rhodococcus sp. p52	4893347	4535
NC_017308	Corynebacterium pseudotuberculosis 1/06-A	2279118	2037
NC_016802	Corynebacterium diphtheriae HC02	2468612	2310
NZ_CP019587	Corynebacterium pseudotuberculosis strain PA04	2338093	2078
NZ_CP014950	Mycobacterium abscessus strain FLAC003	4826045	4714
NC_016783	Corynebacterium diphtheriae INCA 402	2449071	2279
NC_003450	Corynebacterium glutamicum ATCC 13032 chromosome	3309401	2959

NZ_CP014958	Mycobacterium abscessus strain FLAC045	5217908	5256
NZ_CP009495	Mycobacterium smegmatis strain INHR1	6988337	6774
NZ_CP021251	Corynebacterium pseudotuberculosis strain ATCC 19410	2337763	2084
NC_016947	Mycobacterium intracellulare MOTT-02	5409696	5059
NZ_CP010451	Corynebacterium glutamicum strain B253	3207539	2932
NZ_CP012095	Mycobacterium bovis strain 1595	4351712	4169
NZ_CP006842	Corynebacterium glyciniphilum AJ 3170	3509786	3201
NC_016934	Mycobacterium tuberculosis UT205 complete genome.	4418088	4195
NC_014329	Corynebacterium pseudotuberculosis FRC41	2337913	2080
NZ_CP009408	Mycobacterium abscessus subsp. bolletii strain MA 1948	5064190	5018
NZ_CP007220	Mycobacterium chelonae CCUG 47445	5029817	4867
NZ_CP009493	Mycobacterium avium subsp. avium 2285 (R)	5169415	4851
NZ_CP009482	Mycobacterium avium subsp. avium 2285 (S)	5197664	4856
NZ_CP017593	Mycobacterium tuberculosis strain Beijing-like/35049	4427062	4250
NZ_AP014573	Mycobacterium tuberculosis str. Kurono DNA	4415078	4240
NZ_CP006764	Corynebacterium doosanense CAU 212 = DSM 45436	2671798	2542
NC_006958	Corynebacterium glutamicum ATCC 13032	3282708	3017
NZ_CP018304	Mycobacterium tuberculosis strain M0002959-6	4386447	4208
NZ_CP011491	Mycobacterium vaccae 95051	6235754	5826
NC_021251	Mycobacterium tuberculosis CCDC5079	4414325	4234
NC_016948	Mycobacterium intracellulare MOTT-64	5501090	5158
NZ_CP009248	Corynebacterium sphenisci DSM 44792	2594799	2283
NZ_CP017595	Mycobacterium tuberculosis strain Beijing-like/38774	4431885	4259
NC_009342	Corynebacterium glutamicum R DNA	3314179	3046
NZ_CP018300	Mycobacterium tuberculosis strain I0002353-6	4385578	4208
NC_016786	Corynebacterium diphtheriae HC01	2427149	2257
NZ_CP009426	Mycobacterium tuberculosis strain 96075	4379376	4218
NC_016800	Corynebacterium diphtheriae BH8	2485519	2399
NC_021715	Mycobacterium sp. 05-1390	5521023	5142
NZ_CP009622	Corynebacterium ulcerans FRC11	2442826	2143
NZ_CP008923	Corynebacterium pseudotuberculosis strain CS_10	2338144	2082
NZ_CP020381	Mycobacterium tuberculosis strain MTB1	4433542	4255
NZ_CP010795	Corynebacterium pseudotuberculosis strain 29156	2338645	2082
NC_020559	Mycobacterium tuberculosis str. Erdman = ATCC 35801 DNA	4392353	4233
NZ_CP011311	Corynebacterium camporealensis strain DSM 44610	2451810	2223
NC_021194	Mycobacterium tuberculosis EAI5/NITR206	4390306	4221

NZ_CP015219	Rhodococcus sp. PBTS1	4251687	3861
NC_017306	Corynebacterium pseudotuberculosis 42/02-A	2337606	2079
NZ_CP014475	Mycobacterium phlei strain CCUG 21000	5349645	5090
NC_021282	Mycobacterium abscessus subsp. bolletii 50594	5000473	5011
NC_016781	Corynebacterium pseudotuberculosis 3/99-5	2337938	2080
NC_016801	Corynebacterium diphtheriae C7 (beta)	2499189	2372
NC_010612	Mycobacterium marinum M	6636827	5588
NZ_CP013260	Corynebacterium pseudotuberculosis strain MB11	2363423	2124
NZ_CP015188	Corynebacterium pseudotuberculosis strain 39	2403579	2159
NZ_CP015100	Corynebacterium pseudotuberculosis strain T1	2337201	2079
NZ_CP015596	Mycobacterium sp. YC-RL4	5801417	5519
NZ_CP009614	Mycobacterium avium subsp. avium strain DJO-44271	5011264	4683
NZ_CP019769	Corynebacterium pseudotuberculosis strain MIC6	2337147	2079
NZ_CP014954	Mycobacterium abscessus strain FLAC008	5166100	5177
NC_015848	Mycobacterium canettii CIPT 140010059 complete genome.	4482059	4242
NZ_CP011474	Corynebacterium pseudotuberculosis strain 12C	2337451	2080
NZ_CP018043	Mycobacterium sp. WY10	6041408	5846
NC_016932	Corynebacterium pseudotuberculosis 316	2310415	2057
NC_015859	Corynebacterium variabile DSM 44702	3433007	3087
NZ_CP006850	Nocardia nova SH22a	8348532	7504
NZ_CP017596	Mycobacterium tuberculosis strain Beijing/391	4406925	4246
NZ_CP015185	Corynebacterium pseudotuberculosis strain 35	2403502	2162
NZ_CP021122	Mycobacterium abscessus subsp. massiliense strain FLAC047	4936470	4851
NZ_CP014543	Corynebacterium pseudotuberculosis strain MEX9	2337578	2082
NZ_CP002871	Mycobacterium tuberculosis HKBS1	4407929	4235
NC_021740	Mycobacterium tuberculosis EAI5	4391174	4210
NZ_CP010339	Mycobacterium tuberculosis strain 22103	4399422	4222
NZ_CP015183	Corynebacterium pseudotuberculosis strain 32	2403533	2159
NC_017307	Corynebacterium pseudotuberculosis CIP 52.97	2369387	2124
NC_019966	Mycobacterium sp. JS623	6464916	6302
NZ_CP012090	Mycobacterium tuberculosis W-148	4418548	4240
NZ_CP007809	Mycobacterium tuberculosis strain KIT87190	4410788	4234
NC_020230	Corynebacterium urealyticum DSM 7111	2316065	1954
NZ_CP017711	Corynebacterium pseudotuberculosis strain MEX1	2337090	2081
NC_016789	Corynebacterium diphtheriae PW8	2530683	2383
NC_020506	Corynebacterium callunae DSM 20147	2839551	2558

NC_012490	Rhodococcus erythropolis PR4 DNA	6516310	6020
NZ_CP009499	Mycobacterium intracellulare 1956	5183048	4855
NZ_CP002885	Mycobacterium tuberculosis CCDC5180	4414346	4238
NZ_CP014960	Mycobacterium abscessus strain FLAC049	4799801	4704
NC_022663	Mycobacterium kansasii ATCC 12478	6432277	5579
NZ_CP018082	Nocardia soli strain Y48	7310115	6524
NZ_CP020356	Corynebacterium pseudotuberculosis strain SigmaE	2339255	2083
NC_016788	Corynebacterium diphtheriae HC04	2484332	2300
NZ_CP009500	Corynebacterium ulcerans strain 210932	2484335	2190
NZ_CP011913	Corynebacterium ulcerans FRC58	2542597	2267
NZ_CP018331	Corynebacterium diphtheriae strain B-D-16-78	2474151	2308
NC_020519	Corynebacterium glutamicum K051 complete genome	3309400	3051
NZ_CP011853	Gordonia sp. QH-11	4428727	4093
NZ_CP010337	Mycobacterium tuberculosis strain 22115	4401829	4243
NZ_CP011773	Mycobacterium sp. EPa45	6177406	5834
NZ_CP016189	Mycobacterium immunogenum strain FLAC016	5604845	5488
NZ_CP015184	Corynebacterium pseudotuberculosis strain 33	2403550	2160
NZ_CP011542	Corynebacterium mustelae strain DSM 45274	3391554	2982
NZ_CP009249	Corynebacterium phocae strain M408/89/1	2779609	2441
NZ_CP017597	Mycobacterium tuberculosis strain Beijing-like/50148	4444417	4266
NC_012522	Rhodococcus opacus B4 DNA	7913450	7222
NC_022198	Corynebacterium argentoratense DSM 44202	2031902	1842
NC_006361	Nocardia farcinica IFM 10152 DNA	6021225	5600
NZ_CP008913	Corynebacterium sp. ATCC 6931	2471920	2088
NZ_CP013146	Corynebacterium pseudotuberculosis strain N1	2337845	2078
NZ_CP011530	Mycobacterium immunogenum strain CCUG 47286	5573781	5466
NZ_CP009111	Rhodococcus opacus strain 1CP	7687653	7037
NZ_CP021417	Corynebacterium ulcerans strain PO100/5	2572413	2332
NZ_CP018778	Mycobacterium tuberculosis strain DK9897	4411511	4232
NZ_CP016193	Mycobacterium abscessus strain FLAC055	5331134	5356
NC_015576	Mycobacterium sinense strain JDM601	4643668	4358
NC_021200	Mycobacterium avium subsp. paratuberculosis MAP4	4829424	4500
NZ_CP011309	[Brevibacterium] flavum strain ATCC 15168	3338699	3141
NZ_CP011095	Corynebacterium ulcerans strain 131002	2434569	2140
NZ_CP017598	Mycobacterium tuberculosis strain Beijing-like/1104	4380156	4220
NZ_CP009927	Corynebacterium pseudotuberculosis strain VD57	2337177	2077

NZ_CP009447	Mycobacterium abscessus subsp. bolletii strain MM1513	4501725	4427
NZ_CP016794	Mycobacterium tuberculosis strain SCAID 320.0	4406628	4250
NC_018027	Mycobacterium chubuense NBB4	5583723	5226
NZ_CP015190	Corynebacterium pseudotuberculosis strain 46	2366565	2115
NZ_CP014953	Mycobacterium abscessus strain FLAC007	5064478	5053
NZ_CP003949	Rhodococcus opacus PD630	8376953	7603
NC_017317	Corynebacterium ulcerans 809	2502095	2208
NZ_CP019963	Corynebacterium stationis strain LMG 21670	2871159	2601
NZ_CP014957	Mycobacterium abscessus strain FLAC031	5146255	5061
NZ_CP009243	Mycobacterium bovis BCG strain Russia 368	4370138	4182
NC_016799	Corynebacterium diphtheriae 31A	2535346	2411
NZ_CP002883	Mycobacterium tuberculosis BT1	4399405	4227
NZ_CP019572	Rhodococcus sp. MTM3W5.2	5665081	5101
NC_014158	Tsukamurella paurometabola DSM 20162	4379918	4177
NZ_CP009100	Mycobacterium tuberculosis strain ZMC13-264	4411507	4225
NZ_CP015309	Corynebacterium pseudotuberculosis strain PA02	2328435	2070

Supplementary File A3 – Number of Genomic Islands (GI) and Putative Transferred Genes (PTG) per genome

Genome ID	Number of GI	Number of PTG
NC_002677	0	0
NC_014329	0	0
NC_016781	0	0
NC_016932	0	0
NC_017031	0	0
NC_017300	0	0
NC_017301	0	0
NC_017303	0	0
NC_017305	0	0
NC_017306	0	0
NC_017308	0	0
NC_017462	0	0
NZ_CP008922	0	0
NZ_CP008923	0	0
NZ_CP008924	0	0
NZ_CP009927	0	0
NZ_CP010795	0	0
NZ_CP010889	0	0
NZ_CP011474	0	0
NZ_CP012695	0	0
NZ_CP012837	0	0
NZ_CP013146	0	0
NZ_CP013327	0	0
NZ_CP013697	0	0
NZ_CP013698	0	0
NZ_CP013699	0	0
NZ_CP014341	0	0
NZ_CP014543	0	0
NZ_CP014998	0	0
NZ_CP015100	0	0
NZ_CP015309	0	0
NZ_CP016826	0	0
NZ_CP017711	0	0
NZ_CP019768	0	0

NZ_CP019769	0	0
NZ_CP020356	0	0
NZ_CP021251	0	0
NZ_CP011312	1	172
NZ_CP012022	1	942
NZ_CP012136	1	9
NZ_CP014960	1	23
NZ_CP017384	1	11
NZ_CP019587	1	13
NC_011896	2	38
NC_022115	2	37
NZ_CP009312	2	72
NZ_CP009447	2	35
NZ_CP009500	2	65
NZ_CP009622	2	40
NZ_CP009716	2	41
NZ_CP010071	2	2033
NZ_CP010818	2	65
NZ_CP011095	2	42
NZ_CP011913	2	107
NZ_CP012390	2	120
NZ_CP016829	2	18
NZ_CP017291	2	21
NC_006958	3	205
NC_012704	3	80
NC_015683	3	174
NC_017307	3	32
NC_017317	3	265
NC_017945	3	30
NC_018101	3	427
NC_020519	3	198
NC_022040	3	31
NC_022198	3	1039
NZ_CP008913	3	227
NZ_CP008944	3	143
NZ_CP009499	3	44
NZ_CP009583	3	247
NZ_CP011542	3	81
NZ_CP013260	3	31
NZ_CP013261	3	30
NZ_CP013262	3	107
NZ_CP013263	3	29
NZ_CP015183	3	348
NZ_CP015184	3	358
NZ_CP015186	3	355
NZ_CP015187	3	358
NZ_CP015188	3	354
NZ_CP015189	3	30

NZ_CP015190	3	29
NZ_CP015191	3	470
NZ_CP015192	3	355
NZ_CP015622	3	94
NZ_CP017292	3	33
NZ_CP019705	3	634
NC_009342	4	651
NC_015576	4	137
NC_016790	4	336
NC_017730	4	363
NZ_CP007255	4	155
NZ_CP009211	4	417
NZ_CP009244	4	69
NZ_CP009248	4	82
NZ_CP009251	4	1221
NZ_CP009407	4	580
NZ_CP009408	4	725
NZ_CP009613	4	680
NZ_CP011341	4	91
NZ_CP011545	4	122
NZ_CP014634	4	127
NZ_CP014635	4	61
NZ_CP014952	4	71
NZ_CP014955	4	616
NZ_CP015185	4	371
NZ_CP015235	4	46
NZ_CP016188	4	274
NZ_CP018331	4	127
NZ_CP021417	4	447
NC_012590	5	435
NC_016782	5	480
NC_016786	5	471
NC_016801	5	619
NC_018150	5	117
NC_018581	5	157
NC_021351	5	449
NC_021352	5	448
NC_021915	5	320
NZ_CP004350	5	654
NZ_CP009215	5	228
NZ_CP009220	5	123
NZ_CP009245	5	313
NZ_CP009247	5	63
NZ_CP009615	5	118
NZ_CP009616	5	117
NZ_CP010113	5	68
NZ_CP011311	5	257
NZ_CP013991	5	240
NZ_CP014941	5	3103
NZ_CP014950	5	171

NZ_CP014984	5	178
NZ_CP020410	5	267
NZ_CP020821	5	99
NC_002935	6	581
NC_014168	6	607
NC_016788	6	346
NC_016789	6	1190
NC_018612	6	263
NZ_CP006764	6	774
NZ_CP006842	6	88
NZ_CP007722	6	514
NZ_CP007724	6	489
NZ_CP009914	6	913
NZ_CP010451	6	161
NZ_CP011309	6	207
NZ_CP011541	6	359
NZ_CP011883	6	3115
NZ_CP012194	6	209
NZ_CP013049	6	809
NZ_CP014956	6	771
NZ_CP014959	6	302
NZ_CP015220	6	335
NZ_CP015529	6	94
NZ_CP015964	6	245
NZ_CP016190	6	771
NZ_CP016191	6	627
NZ_CP016335	6	219
NZ_CP020658	6	432
NC_003450	7	287
NC_016802	7	1030
NC_016948	7	301
NC_018027	7	205
NC_020230	7	111
NC_023036	7	1235
NZ_CP005286	7	171
NZ_CP006841	7	248
NZ_CP007156	7	179
NZ_CP007220	7	212
NZ_CP007790	7	158
NZ_CP009249	7	205
NZ_CP011295	7	144
NZ_CP011491	7	431
NZ_CP011546	7	104
NZ_CP011773	7	94
NZ_CP014279	7	850
NZ_CP014957	7	466
NZ_CP014961	7	1367
NZ_CP015961	7	928
NZ_CP015965	7	259
NZ_CP016193	7	1368

NZ_CP016396	7	105
NZ_CP017014	7	1671
NZ_CP018175	7	227
NZ_CP019963	7	593
NC_016604	8	298
NC_016783	8	380
NC_016785	8	704
NC_016787	8	628
NC_016799	8	909
NC_016946	8	154
NC_019966	8	268
NC_020302	8	678
NC_020506	8	337
NC_021663	8	396
NC_023150	8	379
NZ_AP014547	8	766
NZ_CP010797	8	101
NZ_CP011022	8	1244
NZ_CP012044	8	559
NZ_CP012885	8	353
NZ_CP016640	8	347
NZ_CP016819	8	356
NZ_CP017299	8	3111
NZ_CP019572	8	811
NZ_CP019882	8	266
NZ_CP021122	8	718
NC_000962	9	157
NC_006361	9	3807
NC_007164	9	401
NC_012490	9	3313
NC_012522	9	5191
NC_016800	9	850
NC_016947	9	116
NZ_AP012555	9	305
NZ_CP009482	9	544
NZ_CP009614	9	487
NZ_CP010114	9	203
NZ_CP014953	9	1121
NZ_CP015219	9	261
NZ_CP016594	9	736
NZ_CP018305	9	162
NC_008769	10	179
NC_009077	10	466
NC_014158	10	1178
NC_015564	10	473
NC_018681	10	436
NC_019965	10	210
NC_020245	10	181
NC_021282	10	1244
NZ_CP004046	10	282

NZ_CP004062	10	265
NZ_CP004353	10	671
NZ_CP009493	10	551
NZ_CP010827	10	596
NZ_CP012095	10	175
NZ_CP014475	10	397
NZ_CP014958	10	1298
NZ_CP016189	10	1593
NZ_CP018778	10	213
NZ_CP019221	10	648
NC_004369	11	738
NC_012207	11	216
NC_015673	11	1373
NC_015848	11	370
NC_016804	11	195
NC_016887	11	354
NZ_AM412059	11	214
NZ_CP003494	11	442
NZ_CP006850	11	251
NZ_CP009243	11	216
NZ_CP009449	11	212
NZ_CP011530	11	1755
NZ_CP014566	11	216
NZ_CP018300	11	219
NZ_CP021252	11	574
NC_002755	12	252
NC_009338	12	662
NC_015758	12	214
NC_021200	12	227
NC_021715	12	375
NC_021740	12	228
NC_022350	12	233
NZ_CP008744	12	232
NZ_CP009246	12	1027
NZ_CP009427	12	1019
NZ_CP009483	12	989
NZ_CP011269	12	362
NZ_CP011853	12	3630
NZ_CP012506	12	263
NZ_CP013741	12	231
NZ_CP014951	12	1372
NZ_CP014954	12	1199
NZ_CP015773	12	219
NZ_CP018302	12	248
NZ_CP018303	12	231
NZ_CP018304	12	232
NZ_CP020381	12	247
NC_002944	13	2262
NC_008595	13	793
NC_014814	13	1866

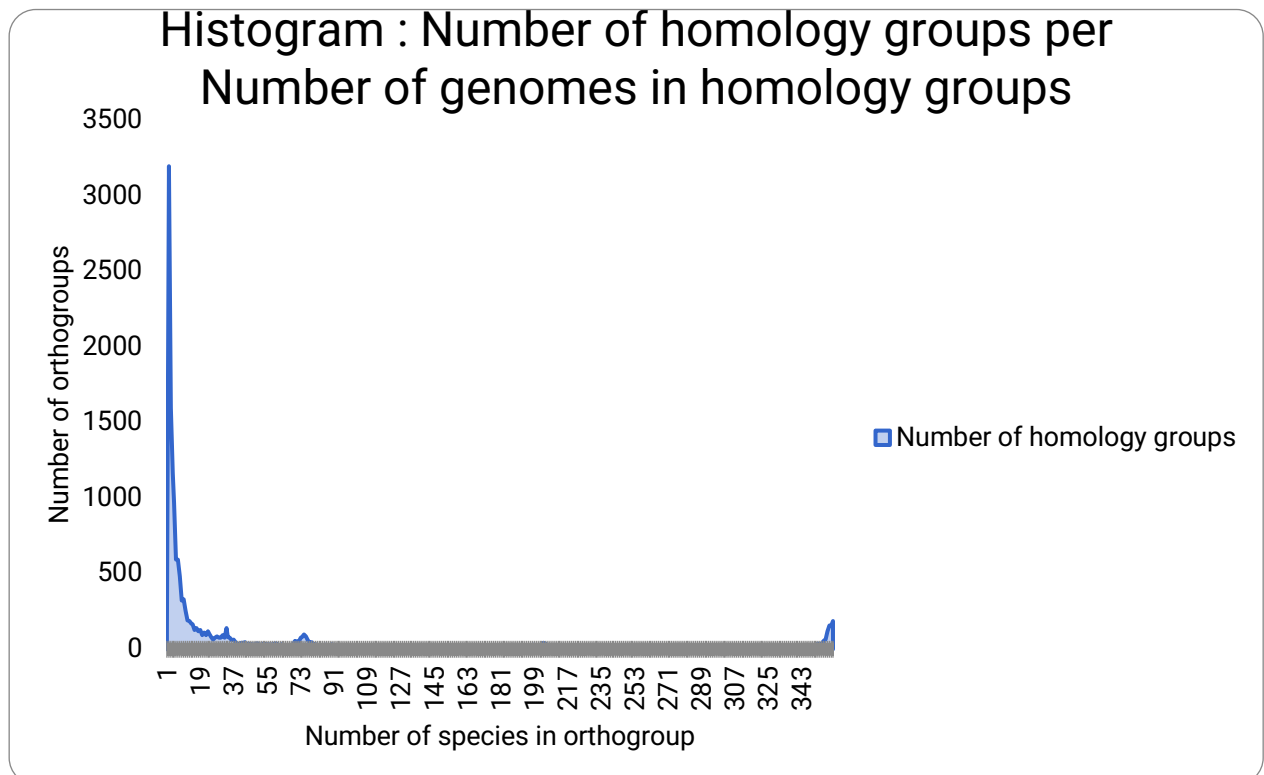
NC_015859	13	248
NC_016934	13	269
NC_017904	13	420
NC_018143	13	250
NC_019950	13	385
NC_020089	13	257
NC_021194	13	267
NC_022663	13	1001
NZ_AP014573	13	266
NZ_CP002882	13	285
NZ_CP003949	13	496
NZ_CP007027	13	249
NZ_CP007803	13	263
NZ_CP009100	13	228
NZ_CP009101	13	253
NZ_CP009426	13	276
NZ_CP010337	13	296
NZ_CP010339	13	267
NZ_CP012150	13	372
NZ_CP015495	13	2264
NZ_CP016888	13	267
NZ_CP016972	13	237
NZ_CP017597	13	292
NC_010612	14	272
NC_013441	14	2842
NC_021251	14	290
NZ_CP002871	14	296
NZ_CP002883	14	298
NZ_CP002885	14	293
NZ_CP009480	14	291
NZ_CP011510	14	599
NZ_CP012749	14	387
NZ_CP017594	14	289
NZ_CP017596	14	331
NZ_CP017598	14	292
NZ_CP017920	14	624
NZ_CP018301	14	283
NZ_HG813240	14	291
NC_009565	15	297
NC_010545	15	456
NC_016906	15	1176
NC_019951	15	212
NZ_CP007809	15	304
NZ_CP012090	15	313
NZ_CP015596	15	2072
NZ_CP017595	15	300
NZ_CP018363	15	409
NZ_CP020809	15	394
NC_008146	16	507
NC_012943	16	311

NC_018078	16	310
NC_020559	16	832
NC_021054	16	351
NZ_CP009111	16	553
NZ_CP010330	16	324
NZ_CP016192	16	914
NZ_CP016794	16	325
NZ_CP019420	16	647
NC_009525	17	287
NC_016768	17	330
NC_017522	17	362
NC_017524	17	310
NZ_CP013475	17	1364
NZ_CP017593	17	354
NC_008705	18	3046
NC_019952	18	311
NC_020133	19	682
NZ_CP018043	19	887
NC_008596	20	402
NC_008726	20	1277
NZ_CP018063	20	1017
NZ_CP009494	21	448
NZ_CP009495	21	448
NZ_CP009496	21	448
NC_018289	22	477
NZ_CP018082	22	846
NC_008268	26	2162
NZ_CP017839	46	1366
<hr/>		
Total	2874	168724
<hr/>		

Supplementary File A4 – Statistics of homology groups

Statistics on homology groups

Description	Numbers
Number of genes	1368128
Number of genes in homology groups	1356782
Number of unassigned genes	11346
Percentage of genes in homology groups	99,2
Percentage of unassigned genes	0,8
Number of homology groups	17821
Number of species-specific homology groups	142
Number of genes in species-specific homology groups	365
Percentage of genes in species-specific homology groups	0
Mean homology group size	76,1
Median homology group size	10
G50 (assigned genes)	352
G50 (all genes)	349
O50 (assigned genes)	1318
O50 (all genes)	1334
Number of homology groups with all species present	188
Number of single-copy homology groups	13



DISCUSSÃO GERAL

No capítulo 1, foi apresentado um *pipeline* para a reconstrução de filogenia de microrganismos que audita a presença de prováveis genes transferidos e remover esses genes, permitindo a reconstrução robusta da filogenômica de *Corynebacteriales*. A confirmação de prováveis genes transferidos ficou a cargo do desenvolvimento de uma ferramenta computacional para validar o conjunto de genes presentes em árvores de genes, anteriormente identificados como transferidos, e então, aplicar o método filogenético para inferir a árvore de espécies. A ferramenta foi desenvolvida nesta tese foi adicionada no passo 5 do *pipeline* de reconstrução filogenômica apresentado no capítulo 1. O *software* foi desenvolvido em linguagem Python, com as bibliotecas *ete3*, para manipulação de árvores filogenéticas e *Bio.SeqIO*, para o tratamento de sequências, além de os módulos: “*collections*”, para manipulação de estruturas de dados como listas, pilhas e dicionários e etc.; “*re*”, para criação de expressões regulares e o módulo “*argparse*”, para construir os parâmetros de entrada.

O EXECT - *Easy Xenology Conciliation Tool* (<https://github.com/UdeS-CoBIUS/EXECT>) é uma ferramenta filogenômica, de conciliação e corte de genes horizontalmente transferidos presentes em árvores de genes. O código do *software* é apresentado no Apêndice II. Para a execução do EXECT são necessários 4 parâmetros de entrada: 1 – Árvore de Gene; 2 – Árvore de espécie; 3 – Sequência de aminoácidos, em formato FASTA, correspondente a árvore de genes e 4 – Lista de genes identificados como transferidos. Como saída, a ferramenta apresenta quatro arquivos: a) um arquivo de índices estatísticos das operações realizadas; b) uma lista contendo os identificadores

de cada um dos genes removidos, c) uma lista contendo os identificadores de genes que foram reconciliados, e d) um arquivo de sequência sem os genes removidos.

Neste trabalho, foi apresentado um *pipeline* filogenômico para reconstrução de filogenias de procariotos a partir de levantamento de genes adquiridos por transferência horizontal. Genes transferidos horizontalmente contribuem para a diversidade e adaptação seletiva de micro-organismos, tendo alto impacto na plasticidade genômica de bactérias que, portanto, amalgamam o sinal filogenético do gene transferido com ao resto da sequência do genoma, através de processos como homogeneização e/ou recombinação homóloga. Ainda assim, sabendo que a transferência horizontal de genes pode resultar em um mosaico genômico composto por genes com diferentes origens, em casos de eventos ancestrais de HGT, a homogeneização resultante entre a sequência da espécie doadora e da sequência do genoma da espécie receptora pode não ser suficientemente diferente para ser detectar como sendo um gene horizontalmente transferido.

Sabendo-se das limitações metodológicas e computacionais, neste trabalho, os genes horizontalmente transferidos foram identificados primariamente em ilhas genômicas, cuja composição de nucleotídeos varia de acordo com a sequência do genoma (ex: conteúdo G+C) e, posteriormente, confirmados com o uso da ferramenta computacional desenvolvida, o EXECT.

Estudos comparativos, de qualquer natureza, em algum momento passam por análise de homologia, sendo essa uma fase crucial. Com o desenvolvimento do EXECT, percebeu-se a sua importância na curadoria de grupos de ortólogos, e, com o EXECT,

isso passa a ser integrado complementarmente aos resultados de ferramentas já disponíveis.

CONCLUSÃO

Neste trabalho de tese, apresentamos um *pipeline* filogenômico para reconstrução filogenômica de procariotos, confirmando e removendo a informação de genes horizontalmente transferidos pelo desenvolvimento de um novo método filogenômico, aplicado na reconstrução da filogenia de Ordem *Corynebacteriales*. A análise sistemática, além de validar a reclassificação taxonômica de espécies (*Brevibacterium* como *Corynebacterium*), também, permitiu a criação de uma árvore consenso que pode ser utilizada para estudar a aquisição/perda/troca de genes presentes entre as espécies dessa Ordem do Filo *Actinobacteria*.

PERSPECTIVAS

- Utilizar a metodologia desse trabalho na reconstrução filogenômica de outras ordens e de outros Filos de Bacteria e Archaea (i.e: proteobacteria)
- Averiguar o papel dos genes identificados como horizontalmente transferidos e a causa/efeito desses no estudo sobre a plasticidade de genomas (i.e: *Mycobacterium tuberculosis*, *Mycobacterium lepreae*, *Corynebacterium jeikeium*, *Corynebacterium pseudotuberculosis*, entre outros)

REFERÊNCIAS

1. Venter, J. Craig, et al. "The sequence of the human genome." *science* 291.5507 (2001): 1304-1351.
2. Daly, Mark J., et al. "High-resolution haplotype structure in the human genome." *Nature genetics* 29.2 (2001): 229.
3. Wang, Daojing, and Steven Bodovitz. "Single cell analysis: the new frontier in 'omics'." *Trends in biotechnology* 28.6 (2010): 281-290.
4. Berger, Bonnie, Jian Peng, and Mona Singh. "Computational solutions for omics data." *Nature reviews genetics* 14.5 (2013): 333-346.
5. Stähler, Peer, et al. "Another side of genomics: synthetic biology as a means for the exploitation of whole-genome sequence information." *Journal of biotechnology* 124.1 (2006): 206-212.
6. Titmus, Matthew A., James Gurtowski, and Michael C. Schatz. "Answering the demands of digital genomics." *Concurrency and Computation: Practice and Experience* 26.4 (2014): 917-928.
7. Benner, Steven A., and A. Michael Sismour. "Synthetic biology." *Nature Reviews Genetics* 6.7 (2005): 533.
8. Eisen, Jonathan A. "Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis." *Genome research* 8.3 (1998): 163-167.

9. Rabiee, Maryam, Erfan Sayyari, and Siavash Mirarab. "Multi-allele species reconstruction using ASTRAL." *Molecular phylogenetics and evolution* 130 (2019): 286-296.
10. Desluc, F., Brinkmann, H., Philippe, H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* 6: 361-375
11. Davison, John. "Genetic exchange between bacteria in the environment." *Plasmid* 42.2 (1999): 73-91.
12. Hacker, J., et al. "Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution." *Molecular microbiology* 23.6 (1997): 1089-1097.
13. Gogarten, J. Peter, and Jeffrey P. Townsend. "Horizontal gene transfer, genome innovation and evolution." *Nature Reviews Microbiology* 3.9 (2005): 679.
14. Ochman, Howard, Jeffrey G. Lawrence, and Eduardo A. Groisman. "Lateral gene transfer and the nature of bacterial innovation." *nature* 405.6784 (2000): 299.
15. Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten. "Horizontal gene transfer: building the web of life." *Nature Reviews Genetics* 16.8 (2015): 472.
16. Daubin, Vincent, Manolo Gouy, and Guy Perriere. "A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history." *Genome research* 12.7 (2002): 1080-1090.

17. Glaeser, Stefanie P., and Peter Kämpfer. "Multilocus sequence analysis (MLSA) in prokaryotic taxonomy." *Systematic and applied microbiology* 38.4 (2015): 237-245.
18. Sorek, Rotem, et al. "Genome-wide experimental determination of barriers to horizontal gene transfer." *Science* 318.5855 (2007): 1449-1452.
19. Galperin, Michael Y., and Eugene V. Koonin. "Who's your neighbor? New computational approaches for functional genomics." *Nature biotechnology* 18.6 (2000): 609.
20. Parkinson, John S. "Signal transduction schemes of bacteria." *Cell* 73.5 (1993): 857-871.
21. Chu, Hoi Yee, Kathleen Sprouffske, and Andreas Wagner. "Assessing the benefits of horizontal gene transfer by laboratory evolution and genome sequencing." *BMC evolutionary biology* 18.1 (2018): 54
22. Ravenhall, Matt, et al. "Inferring horizontal gene transfer." *PLoS computational biology* 11.5 (2015): e1004095.
23. Langille, Morgan GI, William WL Hsiao, and Fiona SL Brinkman. "Detecting genomic islands using bioinformatics approaches." *Nature Reviews Microbiology* 8.5 (2010): 373
24. Doyon, Jean-Philippe, et al. "Models, algorithms and programs for phylogeny reconciliation." *Briefings in bioinformatics* 12.5 (2011): 392-400.
25. Nakhleh, Luay. "Computational approaches to species phylogeny inference and gene tree reconciliation." *Trends in ecology & evolution* 28.12 (2013): 719-728.

26. Page, Roderic DM, and Michael A. Charleston. "Trees within trees: phylogeny and historical associations." *Trends in Ecology & Evolution* 13.9 (1998): 356-359.
27. Philippe, Hervé, and Christophe J. Douady. "Horizontal gene transfer and phylogenetics." *Current opinion in microbiology* 6.5 (2003): 498-505.
28. Dalquen, Daniel A., et al. "The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study." *PloS one* 8.2 (2013): e56925.
29. Becq, Jennifer, Cécile Churlaud, and Patrick Deschavanne. "A benchmark of parametric methods for horizontal transfers detection." *PLoS One* 5.4 (2010): e9989
30. De Queiroz, Kevin, and Michael J. Donoghue. "Phylogenetic systematics and the species problem." *Cladistics* 4.4 (1988): 317-338.
31. Ludwig, Wolfgang, and Hans-Peter Klenk. "Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics." *Bergey's manual® of systematic bacteriology*. Springer, Boston, MA, 2005. 49-66.
32. Glover, Natasha, et al. "Advances and Applications in the Quest for Orthologs." *Molecular biology and evolution* 36.10 (2019): 2157-2164.
33. Sonnhammer, Erik LL, et al. "Big data and other challenges in the quest for orthologs." (2014): 2993-2998.
34. Fitch, Walter M. "Homology: a personal view on some of the problems." *Trends in genetics* 16.5 (2000): 227-231.
35. Medini, Duccio, et al. "The microbial pan-genome." *Current opinion in genetics & development* 15.6 (2005): 589-594.

36. Tettelin, Hervé, et al. "Comparative genomics: the bacterial pan-genome." *Current opinion in microbiology* 11.5 (2008): 472-477
37. Hellmuth, Marc, et al. "Phylogenomics with paralogs." *Proceedings of the National Academy of Sciences* 112.7 (2015): 2058-2063.
38. Goodman, Morris, et al. "Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences." *Systematic Biology* 28.2 (1979): 132-163.
39. Emms, David M., and Steven Kelly. "OrthoFinder: phylogenetic orthology inference for comparative genomics." *BioRxiv* (2019): 466201.
40. Emms, David M., and Steven Kelly. "OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy." *Genome biology* 16.1 (2015): 157.
41. Eulenstein, Oliver. *Vorhersage von Genduplikationen und deren Entwicklung in der Evolution*. No. 20. GMD, 1998.
42. Bininda-Emonds, Olaf RP, ed. *Phylogenetic supertrees: combining information to reveal the tree of life*. Vol. 4. Springer Science & Business Media, 2004.
43. Bonizzoni, Paola, Gianluca Della Vedova, and Riccardo Dondi. "Reconciling a gene tree to a species tree under the duplication cost model." *Theoretical computer science* 347.1-2 (2005): 36-53.
44. Åkerborg, Örjan, et al. "Simultaneous Bayesian gene tree reconstruction and reconciliation analysis." *Proceedings of the National Academy of Sciences* 106.14 (2009): 5714-5719.

45. Eulenstein, Oliver, Snehalata Huzurbazar, and David A. Liberles. "Reconciling phylogenetic trees." *Evolution after gene duplication*(2010): 185-206.
46. Darby, Charlotte A., et al. "Xenolog classification." *Bioinformatics* 33.5 (2017): 640-649.
47. Altenhoff, Adrian M., et al. "OMA standalone: orthology inference among public and custom genomes and transcriptomes." *Genome Research* (2019).

APÊNDICES

Nesta sessão, o Apêndice I, dispõe o comprovante de aceite para a publicação do artigo apresentado no capítulo 1 ao periódico *Genome Biology and Evolution*. No Apêndice II, o código fonte do programa EXECT - *Easy Xenology Conciliation Tool*, apresentado na discussão geral. Os apêndices III, IV e V, é destinado às contribuições realizadas durante o período do doutorado do autor, entre os anos de 2015-2019.

Em 2014, o autor teve contato com a comunidade científica internacional através do congresso Latino-Americano da Sociedade Internacional de Biologia Computacional (do inglês, *International Society for Computational Biology – ISCB*). Este primeiro contato com a ISCB foi primordial, para o desenvolvimento de uma série de serviços estudantis prestados às comunidades nacional e internacionais de bioinformática e biologia computacional. A aproximação junto ao Conselho Estudantil (do inglês, *Student Council*), cuja principal missão é desenvolver a próxima geração de biólogos computacionais no mundo, motivou o autor e um pequeno grupo de entusiastas, a reestabeleceram a rede de alunos em biologia computacional no Brasil, a RSG-Brazil (do inglês, *Regional Student Group Brazil*). Com a reativação da RSG-Brazil, em 2016, ocorreu a primeira edição do Simpósio do Conselho Estudantil Brasileiro em Biologia Computacional (do inglês, *Brazilian Student Council Symposium – BR-SCS*), uma conferência estudantil, ao nível internacional, de modo a estabelecer uma rede de contatos entre a comunidade estudantil e facilitar a intercomunicação da comunidade estudantil brasileira em bioinformática e biologia computacional a apresentar seus resultados em conferências internacionais. Em 2019, o BR-SCS chegou a sua quarta edição, com 102 delegados, se

consolidando como a maior conferência estudantil em biologia computacional na América Latina e uma das maiores no mundo.

**APÊNDICE I – Comprovante do aceite para a publicação do artigo ao periódico
*Genome Biology and Evolution.***

From: **Genome Biology and Evolution** <onbehalf@manuscriptcentral.com>
Date: Mon, Mar 16, 2020, 7:52 AM
Subject: Genome Biology and Evolution - Decision on MS GBE-191144.R1 Accept
To: <aida.ouangraoua@gmail.com>
Cc: <golding@mcmaster.ca>, <a.c.eyre-walker@sussex.ac.uk>, <gbe.editorialoffice@oup.com>

16-Mar-2020

Dear Dr. Ouangraoua,

I am pleased to inform you that Dr. Brian Golding, the Associate Editor who handled your manuscript entitled "Reconstructing the phylogeny of Corynebacteriales while accounting for Horizontal Gene Transfer", has recommended that it be accepted for publication in *Genome Biology and Evolution*. I am following that recommendation and forwarding the file to the production office. You will receive further information from the Editorial Office and Production Department regarding the final version of your MS and author proofs shortly.

*** We are working to highlight some of our cutting-edge manuscripts by inviting authors to contribute images for the cover and for social media. To this end, we ask you to submit a photo/graphic/illustration, perhaps combining an organismal picture with some element of your data (e.g. a single panel from a figure). If you have or can generate a suitable colour figure, please email it to us at gbe.editorialoffice@oup.com and gbe.social.media@gmail.com using the subject line "GBE promotional image". The minimum size should be 1600 x 2100 pixels at 300 dpi, and it is essential that you have copyright permission. Please provide a brief legend and confirm that any third party who may claim rights to the image gives their permission for it to be used on social media. ***

Genome Biology and Evolution has an active social media presence, with Facebook, Twitter, and Instagram accounts. To promote your article, please follow us on social media, and like/retweet/share your article once it becomes available. Further, if you or your co-authors have a Twitter account, please send us your username to gbe.social.media@gmail.com so we can then tag you once we post the info on your article.

In order to publish your article, Oxford University Press requires that you complete a licence agreement online. A link to the online licensing system, and instructions on how to select and complete a licence, will be provided to you by the Production Editor at Oxford University Press in due course.

You will receive your official acceptance date from Oxford University Press once you have signed your licence to publish. This date should be used for deposit purposes.

On behalf of the entire Editorial Board of *Genome Biology and Evolution*, I wish to thank you for submitting this work to GBE. I look forward to receiving further contributions from you and your colleagues in the future.

Yours sincerely,

Adam Eyre-Walker
Editor in Chief, *Genome Biology and Evolution*
a.c.eyre-walker@sussex.ac.uk

APÊNDICE II – Código Fonte: Easy Xenology Conciliation Tool

```

1 from ete3 import Tree
2 from Bio import SeqIO
3 import argparse
4 import collections
5 import re
6
7 def build_arg_parser():
8
9     parser = argparse.ArgumentParser(description="EXECT - Easy
10 Xenology Conciliation Tool")
11     parser.add_argument('-f1', '--sourcetree', default = "/data/
12 darn2001/Projects/EXECT/data/data_test/OG0001103/
13 OG0001103_aling_tree_rooted_tree.nw")
14     parser.add_argument('-f2', '--targetTree', default = "/data/
15 darn2001/Projects/EXECT/data/tree_newick_rooted_tree.nw")
16     parser.add_argument('-f3', '--sequence', default = "/data/
17 darn2001/Projects/EXECT/data/data_test/OG0001103/OG0001103.fasta")
18     parser.add_argument('-f4', '--xenologlist', default = "/data/
19 darn2001/teste/list_of_proteins_in_island.txt")
20     return parser
21
22 def main():
23
24     parser = build_arg_parser()
25     arg = parser.parse_args()
26     filename = arg.sourcetree
27     filename2 = arg.targetTree
28     xenolog_file = arg.xenologlist
29     sequencefile = arg.sequence
30
31     xenolog_data = open(xenolog_file, "r")
32     xenologs = xenolog_data.readlines()
33
34     file = open(filename, "r")
35     data = file.readlines()[0]
36     source_tree = Tree(data)
37     print treeStatistics(source_tree, 0)
38
39     file2 = open(filename2, "r")
40     data2 = file2.readlines()[0]
41     target_tree = Tree(data2)
42
43     #count xenologs
44     xsource = countXenologs(xenologs, source_tree)
45     print xsource
46
47     ch = changeNames(source_tree.copy(),target_tree.copy())
48     ch_tree = ch[0]
49     dict = ch[1]
50
51     xtrees = conciledXenologs(xsource, ch_tree, target_tree.copy(), 0,
52 dict)
53     tree = xtrees[0]
54     xdict = xtrees[1]
55     stats = xtrees[2]
56     removed = xtrees[3]
57     conciled = xtrees[4]
58
59     writeStats(stats,removed,conciled,sequencefile)
60     fastadict = fastaWrite(sequencefile, xdict)

```



```

55
56 def returnNotMatches(list1, list2):
57     """ This Function compares two list and returns the elements not
    matched between them
58
59     Parameters
60     -----
61     list1: list of elements
62     list2: list of elements
63
64     Returns
65     -----
66
67     notmatches: list of non-matches between two list
68     """
69
70
71     notmatches = []
72     for item in list1:
73         if item not in list2:
74             notmatches.append(item)
75
76     return notmatches
77
78 def fastaWrite(sequence, dict):
79
80
81     """ This Function writes the fasta file without genes confirmed
    as tranfers
82
83     Parameters
84     -----
85     sequence: sequence file
86     dict: dictionary of confirmed transfered genes
87
88     Returns
89     -----
90
91     fastalist: list of fastaIDs
92     """
93
94     result = sequence.split('.')
95     result = result[0] + '_reconciled_xenologs.fasta'
96     newfasta = open(result, 'w')
97     fastalist = []
98     total = 0
99     file = open(sequence, 'rU')
100    #l = SeqIO.parse(file, 'fasta')
101    lines = file.readlines()
102
103    i = 0
104    while i < (len(lines)):
105        # fwrite.write('>' + lines[i])
106        # fwrite.write(lines[i+1])
107        tmp = lines[i]
108        tmp = tmp.strip('>')
109        if tmp.startswith('C'):
110            tmp = tmp
111        elif tmp.startswith('S'):
112            tmp = tmp
113        elif tmp.startswith('No'):

```

```

114         tmp = tmp
115     else:
116         tmp = tmp.split('_')
117         if len(tmp) == 4:
118             tmp = tmp[-2] + '_' + tmp[-1]
119         elif len(tmp) == 5:
120             tmp = tmp[-3] + '_' + tmp[-2] + '_' + tmp[-1]
121         elif len(tmp) == 2: #outgroup
122             tmp = tmp[0] + '_' + tmp[1].strip('\n') + '|' + tmp[0
] + '_' + tmp[1]
123         tmp = tmp.strip('\n')
124         print dict[tmp]
125         print i
126     tmp = tmp.strip('\n')
127     #print tmp
128     #print len(dict)
129
130     if tmp in dict:
131         #print tmp, 'alooo papai'
132
133         #print '>'+tmp
134         newfasta.write('>'+tmp+'\n')
135         #print lines[i+1].strip('\n')
136         newfasta.write(lines[i+1])
137         total = total + 1
138         fastalist.append(tmp)
139
140     i = i+2
141     #print total
142     #print result
143     return fastalist
144
145 def writeStats(list, removed, conciled, sequencefile):
146
147     """ This Function writes the statistics file, also, the list of
conciled and removed genes
148
149     Parameters
150     -----
151     list: list of id to be removed
152
153     removed: list of removed genes
154     conciled: list of conciled genes
155     sequencefile: fasta sequence as input
156
157     Returns
158     -----
159
160     """
161
162     num_lines = sum(1 for line in open(sequencefile))
163     numberOfCDS = ((num_lines-2)/2)
164     sequence = sequencefile
165     result = sequence.split('.')
166     generalOutput = result[0] + '_EXECT.stats'
167     removedOutput = result[0] + '_EXECT.removed'
168     conciledOutput = result[0] + '_EXECT.conciled'
169     orthogroupID = result[0].split('/')
170     orthogroupID = orthogroupID[-1]
171     exectStats = open(generalOutput, 'w')
172     exectStats.write(orthogroupID + '\t' + str(numberOfCDS) + '\t')

```

```

173     for element in list:
174         exectStats.write(str(element) + '\t')
175     exectStats.write('\n')
176
177     removedStats = open(removedOutput, 'w')
178     for geneRemoved in removed:
179         removedStats.write(str(geneRemoved) + '\n')
180     removedStats.close()
181
182     conciledStats = open(conciledOutput, 'w')
183     for geneConciled in conciled:
184         conciledStats.write(str(geneConciled) + '\n')
185     conciledStats.close()
186
187 def printDic(tree, dict):
188
189     #print len(dict)
190     list1 = []
191     list2 = []
192     i = 0
193     for leaf in tree:
194         i = i + 1
195         old = str(leaf)
196         old = old[3:]
197         #print old
198         for chave, valor in dict.iteritems():
199             if valor == old:
200                 if valor in list1:
201                     pass
202                 else:
203                     list1.append(old)
204                     list2.append(chave)
205     #print [item for item, count in collections.Counter(list1).items
206     () if count > 1]
207     return list2
208
209 def treeStatistics(tree, arg):
210     """ This Function returns the number of copies presented in a
211     phylogenetic tree.
212
213     Parameters
214     -----
215     tree: phylogenetic tree
216     arg: argument
217     1: print step by step
218     0: do not print
219
220     Returns
221     -----
222     1) repeat: numbers of copies elements in a phylonetic tree
223
224     """
225     arg = arg
226     leafs = 0
227     list = []
228     repeat2 = []
229     count = 0
230
231     """
231     for i in tree.iter_leaf_names():

```

```

232     #print i
233     s_id = i.split("_")
234     s_id = s_id[-2] + "_" + s_id[-1]
235     #print s_id
236     list.append(s_id)
237     my_repeat_dict = {i:list.count(i) for i in list}      # count the
number of repetitive IDs (paralogs)
238     for k, v in my_repeat_dict.iteritems():
239         if v > 1:
240             repeat.append(k)
241
242     """
243     repeat = countParalogs(tree, arg)
244     for i in tree.iter_leaf_names():
245         leafs = leafs + 1
246         list.append(i)
247         #print i
248         s_id = i.split("_")
249         s_id = s_id[-2] + "_" + s_id[-1]
250         if s_id in repeat:
251             repeat2.append(s_id)
252
253
254     print '\n\nNumber of Copies: ' + str(len(repeat2)) + '\n'
255     print 'Number of leafs: ' + str(len(list))
256
257     return repeat
258
259 def conciledParalogs(parap, sourceT, target_tree, arg, dict):
260
261     """ ATENTION: This is a non-recomended use function. unfinished.
262
263     This Function conciles the information of paralogs among the in
264     the species tree.
265
266     Parameters
267     -----
268     parap : list of paralogues
269     xsource: list of xenologes indentified among the source_tree
270     sourceT : Newick Tree gene tree containing xenologs
271     target_tree: Newick species tree
272     arg: argument
273         1: print step by step
274         0: do not print
275     dict: dictionary containing each instancie of the original IDs'
of the source_tree and t
276
277     Returns
278     -----
279     1) array:
280         paralogue: tree with conciled paralogues
281         dict: update dictionary
282
283     """
284
285     print "\n\nConcieling Paralogs"
286
287     paralogTotal = 0
288     removed = 0
289     mantained = 0

```

```

290     paralogtree = sourceT
291
292
293     for node in paralogtree.traverse("postorder"):
294         if node.is_leaf():
295             leaf_name = node.name
296             leaf_name = leaf_name.strip('\n')
297             species_leaf_name = leaf_name.split("_")
298             species_leaf_name = species_leaf_name[-2] + "_" +
species_leaf_name[-1]
299         if species_leaf_name in parap:                                     ## if is in
para-xenolog list
300             #print species_leaf_name
301             node_target = target_tree.search_nodes(name =
leaf_name)
302             #target_tree node
303             if len(node_target) != 1:
304                 print ("Node " + leaf_name + " is not found in
target tree")
305             pass
306         else:
307             paralogTotal = paralogTotal + 1
308
309             father_st = node.get_sisters()
310             father_st = father_st[0]
311             if not father_st:
312                 father_st = node.get_ancestors()
313                 father_st = father_st[0]
314             #print father_st
315
316             father_tt = node_target[0].get_children()
317             if not father_tt:
318                 father_tt = node_target[0].get_ancestors()
319                 father_tt = father_tt[0]
320             #print father_tt
321
322             compara = compareList(father_tt, father_st)
323             #print compara
324             if compara:
325                 if arg == 1:
326                     print 'Mantain: ' + leaf_name
327                     mantained = mantained + 1
328
329             else:
330                 node.delete()
331                 removed = removed + 1
332                 if arg == 1:
333                     print 'Removing: ' + leaf_name
334                     for k in dict.keys():
335                         if dict[k] == leaf_name:
336                             del dict[k]
337
338             """
339             #compareList sister groups
340             father_st = node.get_sisters()
341             father_tt = node_target[0].get_sisters()
342             #print 'fathers: '
343             #print father_st
344             #print father_tt
345             #print father_st.compareList(father_tt)
346             """

```

```

347
348         """
349         ### comparative between common ancestors
350         sourcenode = node.get_common_ancestor(leaf_name)
351         ancestor_st = sourcenode.get_leaf_names()
352         print ancestor_st
353
354         acestor_node_target = node_target[0].
355 get_common_ancestor(leaf_name)
356         ancestor_tt = acestor_node_target.get_leaf_names
357 ()
358         print ancestor_tt
359         """
360
361         """
362         ### comparative between common ancestors
363         for sourcenode in node.get_common_ancestor(
364 leaf_name):
365             ancestor_st = sourcenode.get_leaf_names()
366             print ancestor_st
367             acestor_node_target = node_target[0].
368 get_common_ancestor(leaf_name)
369             for ancestor_target in acestor_node_target:
370                 ancestor_tt = acestor_node_target.
371 get_leaf_names()
372                 print ancestor_tt
373                 """
374                 #compara = comparesister(ancestor_tt, ancestor_st
375 )
376                 """
377                 for sourcenode in node.get_sisters():
378                     sister_st = sourcenode.get_leaf_names()
379                     #print sister_st
380                     #print '#####'
381                     sisters_node_target = node_target[0].get_sisters
382 ()
383                 for sister in sisters_node_target:
384                     sister_tt = sister.get_leaf_names()
385                     #print sister_tt
386                     compara = comparesister(sister_tt,sister_st)
387
388                 print compara
389                 if compara:
390                     print 'Mantain: ' + leaf_name
391                     #mantained = mantained + 1
392                 else:
393                     node.delete()
394                     removed = removed + 1
395                     print 'Removing: ' + leaf_name
396                 """
397
398     print '\nParalogs: ' + str(paralogTotal)
399     print 'Removed paralogs: ' + str(removed)
400     print 'Conciled paralogs: ' + str(paralogTotal-removed)
401     return [paralogtree,dict]
402
403 def processParalogsList(list):
404     paralogslist = []
405     file = open(list, 'rU')
406     lines = file.readlines()
407     i = 1

```

```

401     while i < (len(lines)-1):
402         paralogid = lines[i].strip('\n')
403         paralogid = paralogid.split('\t')
404         paralogid = paralogid[0]
405         paraloglist.append(paralogid)
406         i = i+1
407     return paraloglist
408
409 def getFiles(path, list, suffix):
410
411     seqPathFile = []
412     for item in list:
413         seqPathFile.append(path + item + suffix)
414     return seqPathFile
415
416 def conciledXenologs(xsource, sourceT, target_tree, arg, dict):
417
418     """ this Function conciles the information of xenologs identified
419     by parametric approaches according their sister_group
420     in the species tree.
421
422     Parameters
423     -----
424     xsource: list of xenologs indentified among the source_tree
425     sourceT : Newick Tree gene tree containing xenologs
426     target_tree: Newick species tree
427     arg: argument
428         1: print step by step
429         0: do not print
430     dict: dictionary containing each instance of the original IDs'
431     of the source_tree and t
432
433     Returns
434     -----
435     1) statsList:
436         statsList[0]: conciled tree
437         statsList[1]: input dictionary
438         statsList[2]: statistics
439         statsList[3]: list of removed genes from source_tree
440         statsList[4]: list of conciled genes from source_tree
441
442     """
443     conciledList = []
444     removedList = []
445     new_dict = {}
446     orthodict = {}
447     xenolog = 0
448     removed = 0
449     mantained = 0
450     print "\n\nConcieling Xenologs"
451     tmp = []
452     statsList = []
453     for xs in xsource:
454         xx = xs.split('|')
455         velour = xx[0]
456         tmp.append(velour)
457         orthodict[xs] = velour
458     #print len(tmp) , 'vector of xenologs'
459     #print tmp

```

```

460     #print len(orthodict), 'length of diccionario'
461
462
463     conciledxenologtree = sourceT     # gene tree
464     targetT = target_tree           # species tree
465     leavesUnderNode = []
466     nodes = 0
467     for TreeNode in conciledxenologtree.traverse("postorder"):
468         if len(TreeNode) > 1:     # if node is a subtree
469             for leaf in TreeNode.iter_leaf_names():
470                 if leaf == '':
471                     pass
472                 else:
473                     node_target = target_tree.search_nodes(name =
leaf)     #T3node on the target leaf
474                     tmpleaf = leaf.split('_')
475                     newLeafName = tmpleaf[-2] + '_' + tmpleaf[-1]
476                     leavesUnderNode.append(newLeafName)
477                     #print leavesUnderNode
478                     cs = compareList(leavesUnderNode, tmp)
479                     if len(cs) == len(TreeNode):
480                         print cs, 'Putative transferred genes'
481                         print TreeNode, 'T1'
482                         sibilinglcaSourceTree = TreeNode.up     #lca source
tree before detach
483                         #print sibilinglcaSourceTree.up
484                         try:
485                             lcaSourceTree = sibilinglcaSourceTree.up.
get_closest_leaf()[0].up
486                             TreeNode.detach()
487                             print sibilinglcaSourceTree, 'T2'
488                             sibilinglcaSourceTree.detach()
489                             print lcaSourceTree, 'LCA T2 of T1'
490                         except:
491
492                             lcaSourceTree = []
493                             TreeNode.detach()
494                             print "Putative leaf at higher point"
495                         if node_target == []:
496                             pass
497                         else:
498                             print node_target[0].up, 'T3'
499                             sibilingNodeTarget = node_target[0].up,
500                             sibilingNodeTarget = sibilingNodeTarget[0].up
501                             node_target[0].up.detach()
502                             sibilingTreeLcaNodeTarget =
sibilingNodeTarget.get_closest_leaf()[0].up
503                             print sibilingTreeLcaNodeTarget, 'T4'
504                             lcaSibilingNodeTarget = sibilingNodeTarget.up
505                             node_target[0].up.detach()
506                             sibilingNodeTarget.detach()
507                             print lcaSibilingNodeTarget, 'LCA of T3 and
T4'
508
509                             try:
510
511                                 comparaLCA = (lcaSourceTree.
get_leaf_names(),lcaSibilingNodeTarget.get_leaf_names())
512                                 print comparaLCA, 'compara LCA'
513
514                             except:

```



```

515             comparaLCA = []
516
517             pass
518
519             checkSisterSibilingTrees = checkSister(
sibilinglcaSourceTree.get_leaf_names(),sibilingTreeLcaNodeTarget.
get_leaf_names())
520             print checkSisterSibilingTrees, 'check Sister
521
522             if comparaLCA or checkSisterSibilingTrees:#or
comparaSibilingTrees:
523                 print TreeNode
524                 print 'Tree node were not transferred
from the source tree ' + str(len(TreeNode))
525                 pass
526                 else:
527                     #print TreeNode
528                     #print leavesUnderNode
529                     for leaf_id in leavesUnderNode:
530                         for k, v in orthodict.iteritems():
531                             print k
532                             if leaf_id == v:
533                                 #print k, v, 'key on
orthodictionary'
534                                 if k in dict.keys():
535                                     #print dict[k], 'key on
dictionary'
536                                     removedList.append(k)
537                                     del dict[k]
538                                     removed = removed + 1
539
540                                 else:
541                                     pass
542
543                                     #removed = removed + len(leavesUnderNode)
544                                     maintained = maintained - len(TreeNode)
545                                     print 'Tree Node with ' + str(len(
leavesUnderNode)) + ' leafs were removed'
546                                     #for leaves in TreeNode.iter_leaf_names
():
547                                         # print leaves
548                                         print TreeNode
549                                         TreeNode.delete()
550                                     else: #len(cs) == len(TreeNode)
551                                         pass
552             else:
553                 if TreeNode.is_leaf():
554                     leaf_name = TreeNode.name
555                     if leaf_name == '':
556                         pass
557                     else:
558                         leaf_name = leaf_name.strip('\n')
559                         species_leaf_name = leaf_name.split("_")
560                         species_leaf_name = species_leaf_name[-2] + "_" +
species_leaf_name[-1]
561                         if species_leaf_name in tmp:
562                             # if the species name is in the xenolog list
xenolog = xenolog + 1
563                             node_target = target_tree.search_nodes(name =
leaf_name)
#search node in the target_tree node

```

```

564         if len(node_target) != 1:
565             print ("Node " + leaf_name + " is not
found in target tree")
566         pass
567     else:
568         for sisters in TreeNode.get_sisters():
569             sister_st = sisters.get_leaf_names()
570             sisters_node_target = node_target[0].
get_sisters()
571         for sister in sisters_node_target:
572             sister_tt = sister.get_leaf_names()
573             compara = compareList(sister_tt,
sister_st)
574         if compara:
575             mantained = mantained + 1
576             if arg == 1:
577                 print 'Mantain: ' + leaf_name
578             else:
579                 pass
580         else:
581             if arg == 1:
582                 print 'Removing: ' + leaf_name
583             TreeNode.delete()
584             for k, v in orthodict.iteritems():
585                 if species_leaf_name == v:
586                     if k in dict.keys():
587                         removedList.append(k)
588                         del dict[k]
589                         removed = removed + 1
590                     else:
591                         pass
592             else:
593                 pass
594         leavesUnderNode = []
595         conciled = returnNotMatches(xsource, removedList)
596         print '\n'
597         print 'Number of Xenologs: ' + str(len(tmp))
598         print 'Conciled Xenologs: ' + str(len(conciled))
599         print 'Removed Xenologs: ' + str(removed)
600         statsList.append(len(tmp))
601         statsList.append(str(len(conciled)))
602         statsList.append(removed)
603         #print conciledxenologtree
604
605
606     return [conciledxenologtree, dict, statsList, removedList, conciled]
607
608 def checkSister(list1, list2):
609     """
610     This Function compareList the contents of strings in two list
611
612     Parameters
613     -----
614     list1: list containing elements
615     list2: list containing elements
616
617     Returns
618     -----
619     1) list: list containing the elements of list1 into list2
620
621

```

```

622     """
623     result = []
624     for element in list2:
625         if element in list1:
626             result.append(element) # = [genome for genome in list1 if
        element in genome]
627     return result
628
629 def conciledParaXenologs(xenp, sourceT, target_tree):
630
631     """This Function conciles the sibling xenologs among the
        source_tree
632     To understand more detailed the relationship of sibling xenologs
        please check the following paper:
633
634     Darby, C. A., Stolzer, M., Ropp, P. J., Barker, D., & Durand, D
        . (2016).
635     Xenolog classification. Bioinformatics, 33(5), 640-649.
636
637     Parameters
638     -----
639     xenp: list containing sibling xenologs
640     sourceT: Newick Tree of Sibling Xenologs orthologous group
641     target_tree : Newick species tree
642
643
644     Returns
645     -----
646     1) tree: tree with sibling xenologs conciled
647
648     """
649
650     print "Conciling Xen-Paralogs"
651     sister_st = []
652     sister_tt = []
653     conciledxenologtree = sourceT
654     for node in conciledxenologtree.traverse("levelorder"):
655         if node.is_leaf():
656             leaf_name = node.name
657             leaf_name = leaf_name.strip('\n')
658             species_leaf_name = leaf_name.split("_")
659             species_leaf_name = species_leaf_name[-2] + "_" +
        species_leaf_name[-1]
660         if species_leaf_name in xenp:             ## if is in
        para-xenolog list
661             #print species_leaf_name
662             node_target = target_tree.search_nodes(name =
        leaf_name)
        #target_tree node
663             if len(node_target) != 1:
664                 exit ("Node " + leaf_name + " is not found in
        target tree")
665             else:
666                 for sourcenode in node.get_sisters():
667                     sister_st = sourcenode.get_leaf_names()
668                     #print sister_st
669                     #print '#####'
670                     sisters_node_target = node_target[0].get_sisters(
        )
671                 for sister in sisters_node_target:
672                     sister_tt = sister.get_leaf_names()
673                     #print sister_tt

```

```

674         compara = compareList(sister_tt, sister_st)
675         #print compara
676         if compara:
677             print 'Maintain: ' + leaf_name
678         else:
679             node.delete()
680             #print 'Removing: ' + leaf_name
681     return conciledxenologtree
682
683 def compareList(list1,list2):
684
685     """This function compareList two lists and return their matches
686
687     Parameters
688     -----
689     list1: list of contents
690     list2: list of contents
691
692
693     Returns
694     -----
695     1) list: List of matches in between list1 and list2
696
697     """
698     compare = set(list1) & set(list2)
699     return compare
700
701 def xenParalogs(xsource, source_tree):
702
703     """This Function counts the number of sibling xenologues (
704     paralogues as xenologs) among the source_tree
705
706     Parameters
707     -----
708     xsource: list of xenologes indentified among the source_tree
709     source_tree : Newick Tree of Multicopies orthologous group*
710
711     Returns
712     -----
713     1) list: List of sibling xenologs identified in the source_tree
714
715     """
716     print 'Xen-Paralogs analysis:'
717     tmp = []
718     r = []
719     polytomies = []
720     #print len(xsource)
721     for item in xsource:
722         tmps = str(item)
723         header = tmps.split('|')
724         tmp.append(header[0])
725
726     conciledxenologtree = source_tree
727     for node in conciledxenologtree.traverse("levelorder"):
728         if node.is_leaf():
729             leaf_name = node.name
730             leaf_name = leaf_name.strip('\n')
731             species_leaf_name = leaf_name.split("_")
732             species_leaf_name = species_leaf_name[-2] + "_" +
species_leaf_name[-1]

```

```

733         #print species_leaf_name
734         if species_leaf_name in tmp:           ## if is in
xenolog list
735             #xenolog = xenolog + 1
736             r.append(species_leaf_name)
737
738
739     my_xenologs_dict = {i:r.count(i) for i in r}
740     for k, v in my_xenologs_dict.iteritems():
741         if v > 1:
742             polytomies.append(k)
743     print str(len(polytomies)) + ' xen-paralogs were found among the
source tree'
744     return polytomies
745
746 def countParalogs(source_tree, arg):
747
748     """This Function counts the number of paralogues among the
source_tree
749
750     Parameters
751     -----
752     source_tree : Newick Tree of Multicopies orthologous group*
753     arg: flag
754         1=true
755         0=false
756
757     Returns
758     -----
759     1) list: List of Xenologs identified by parametric approaches
760
761     """
762
763     if arg == 1:
764         print 'Paralogs analysis:'
765     else:
766         pass
767
768     tmp = []
769     tmp2 = []
770     paralogs = []
771     source_leaves_tree_names = source_tree.get_leaf_names()
772     for leaf in source_leaves_tree_names:
773         #print leaf
774         s_id = leaf.split('_')
775         s = s_id[-2] + '_' + s_id[-1]
776         #s = s_id[-2].split('|') #reconstruct the sp NCBI_ID
777         #s = s_id[-3] + #species NCBI_id
778         #if len(s_id) > 4:           #reconstruct the gene ID
779         #     s_id = s_id[-3] + '_' + s_id[-2] + '_' + s_id[-1]
780         #else:
781         #     s_id = s_id[-2] + '_' + s_id[-1]
782         #print s_id[-3] + '_' + s[0], s_id[-3] + '_' + s_id[-2] + '_'
' + s_id[-1]
783         #s_id = s_id[-3] + '_' + s_id[-2] + '_' + s_id[-1]
784         #print s, s_id
785         #tmp.append([s, s_id])
786         tmp.append(s)
787         #for elemtns in tmp:           #parser NCBI_ID only
788         #     tmp2.append(elemtns[0])
789

```

```

790     my_paralogs_dict = {i:tmp.count(i) for i in tmp}           # count the
                        number of repetitive IDs (paralogs)
791     #print my_paralogs_dict
792     for k, v in my_paralogs_dict.iteritems():
793         if v > 1:
794             #print k,v
795             paralogs.append(k)
796     if arg == 1:
797         print str(len(paralogs)) + ' instances of paralogs were
found among the source tree'
798     else:
799         pass
800
801     return paralogs
802
803 def countXenologs(xenologs, source_tree):
804
805     """ Function count the number of xenologs among the source_tree
identified previously by parametric approaches
806
807     Parameters
808     -----
809     xenologs: List of Xenologs
810     source_tree : Newick Tree of homology group
811
812     Returns
813     -----
814     1) list: List of Xenologs among source_tree
815
816     """
817     print 'Xenologs counting analysis: \n'
818     tmp_list = []
819     for line in xenologs:
820         line = line.strip("\n")
821         tmp_list.append(line)
822
823     source_leaves_tree_names = source_tree.get_leaf_names()
824     s_id_array = []
825     xenologs_list = []
826     for leaf in source_leaves_tree_names:
827         s_id = leaf.split('_')
828         #print len(s_id)
829         if len(s_id) > 4:
830             s_id = s_id[-3] + '_' + s_id[-2] + '_' + s_id[-1]
831             s_id_array.append(s_id)
832         else:
833             s_id = s_id[-2] + '_' + s_id[-1]
834             s_id_array.append(s_id)
835         if s_id in tmp_list:
836             xenologs_list.append(s_id)
837     print str(len(xenologs_list)) + ' xenologs were found in the
source tree'
838
839     return xenologs_list #st #sta
840
841 def tmpTree(tree, arg):
842
843     """ Function change the leafs name among the source_tree
844
845     Parameters
846     -----

```

```

847     tree : source_tree
848     arg: flag
849         1=true
850         0=false
851
852     Returns
853     -----
854     1) tmptree: tree with new names
855     """
856
857     tmptree = tree
858     tmp_tree_names = tree.get_leaf_names()
859
860
861     for leaf in tmp_tree_names:
862         old = leaf
863         if arg == 1:
864             s_id = leaf.split('|')
865             s_id = s_id[0].split('_')
866             s_id = s_id[0] + '_' + s_id[1]
867         else:
868             s_id = leaf.strip('/n').split('_')
869             s_id = s_id[-2] + '_' + s_id[-1]
870             #print s_id_c
871         for node in tree:
872             if node.name == old:
873                 node.name = s_id
874
875     return tmptree
876
877 def changeNames(source_tree,target_tree): #stay
878
879     """ Function change names of source_tree based on the NCBI ID in
880     the end of target ID
881
882     Parameters
883     -----
884     source_tree : Newick Tree of Multicopies orthologous group*
885     target_tree: Newick tree of Multiple Sequence Alignment
886
887     *Default Orthofinder Output
888     Example:
889     >NCBI_ID_NCBI_ID_|PROTEIN_ID
890
891     >NZ_CP018778_NZ_CP018778|BTU11_RS15175
892
893     Returns
894     -----
895     1) source_tree: same source-tree with parsed NCBI ID
896
897     Example:
898     NZ_CP018778
899
900     2) dict: dictionary containing each instance of the original IDs
901     ' of the source_tree and t
902     the respectively leafname IDs'
903     {NZ_CP013741|BOVR_RS10860': '
904     Mycobacterium_bovis_strain_BCG_1_Russia__NZ_CP013741}
905
906     """

```

```

905     namesdict = {}
906     source_leaves_tree_names = source_tree.get_leaf_names()
907     target_leaves_tree_names = target_tree.get_leaf_names()
908
909     for leaf in source_leaves_tree_names:
910         old = leaf
911         s_id = leaf.strip('/n').split('|')
912         #print s_id
913         p_id = s_id[-1]
914         s_id = s_id[0].split('_')
915         s_id = s_id[0] + '_' + s_id[1]
916         if s_id.startswith('C'):
917             s_id_c = p_id
918         elif s_id.startswith('S'):
919             s_id_c = p_id
920         elif s_id.startswith('No'):
921             s_id_c = p_id
922         else:
923             s_id_c = s_id + '|' + p_id
924         s_id_c = s_id_c.strip("\n")
925         #print s_id_c
926         for s in target_leaves_tree_names:
927             if s_id in s:
928                 leaf = s
929         for node in source_tree:
930             if node.name == old:
931                 namesdict[s_id_c] = leaf
932                 node.name = leaf
933
934     return [source_tree,namesdict]
935
936 if __name__ == "__main__":
937     print 'EXECT - Easy Xenology Conciliation Tool - Version 1.0 \n' \
938           'Nilson Da Rocha Coimbra - CoBiUS Lab - Universite de
Sherbrooke \n' \
939           'Contact: darn2001@usherbrooke.ca' \
940           '(Da Rocha Coimbra and Ouangraoua, 2018, in prep)\n'
941     main()




```


**APÊNDICE III – Editorial: Second ISCB Latin American Student Council Symposium
(LA-SCS) 2016**



EDITORIAL

Second ISCB Latin American Student Council Symposium (LA-SCS) 2016 [version 1; peer review: not peer reviewed]

Alexander Miguel Monzon ¹, Marcia A. Hasenahuer¹, Estefanía Mancini²,
Nilson Coimbra ³, Fiorella Cravero⁴, Javier Cáceres-Molina⁵,
César A. Ramírez-Sarmiento⁶, Nicolas Palopoli^{1,7}, Pieter Meysman^{8,9},
R. Gonzalo Parra ¹⁰

¹Structural Bioinformatics Group, Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Buenos Aires, B1876BXD, Argentina

²Regulation of Alternative pre-mRNA Splicing during Cell Differentiation, Development and Disease, Centre for Genomic Regulation, Barcelona, Spain

³Department of Bioinformatics, Institute of Biological Sciences, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, 31270-90, Brazil

⁴Chemoinformatics Group, Process Engineering PLAPIQUI (UNS-CONICET), Bahía Blanca, 8000, Argentina

⁵Laboratorio de Fisiología y Genómica de Frutales, Centro de Biotecnología Vegetal (CBV), Universidad Andrés Bello, Santiago, Chile

⁶Institute for Biological and Medical Engineering, Schools of Engineering, Medicine and Biological Sciences, Pontificia Universidad Católica de Chile, Santiago, Chile

⁷Structural Bioinformatics Unit, Fundación Instituto Leloir, IIBBA-CONICET, Buenos Aires, C1405BWE, Argentina

⁸Advanced Database Research and Modelling (ADReM), Department of Mathematics and Computer Science, University of Antwerp, Antwerp, 2000, Belgium

⁹Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp, Antwerp, 2000, Belgium

¹⁰Quantitative and Computational Biology Group, Max Planck Institute for Biophysical Chemistry, Göttingen, 37077, Germany

v1 First published: 16 Aug 2017, 6(ISCB Comm J):1491 (<https://doi.org/10.12688/f1000research.12321.1>)

Latest published: 16 Aug 2017, 6(ISCB Comm J):1491 (<https://doi.org/10.12688/f1000research.12321.1>)

Not Peer Reviewed

This article is an Editorial and has not been subject to external peer review.

Any comments on the article can be found at the end of the article.

Abstract

This report summarizes the scientific content and activities of the second edition of the Latin American Symposium (LA-SCS), organized by the Student Council (SC) of the International Society for Computational Biology (ISCB), held in conjunction with the Fourth Latin American conference from the International Society for Computational Biology (ISCB-LA 2016) in Buenos Aires, Argentina, on November 19, 2016.

Keywords

bioinformatics, education, ISCB, Student Council, symposium



This article is included in the [International Society for Computational Biology Community Journal gateway](#).

Corresponding authors: Alexander Miguel Monzon (monzon.alexander@gmail.com), Nicolas Palopoli (nicopalo@gmail.com), Pieter Meysman (pieter.meysman@gmail.com), R. Gonzalo Parra (parra.gonzalo@gmail.com)

Author roles: **Monzon AM:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Hasenahuer MA:** Writing – Original Draft Preparation, Writing – Review & Editing; **Mancini E:** Writing – Original Draft Preparation; **Coimbra N:** Writing – Original Draft Preparation, Writing – Review & Editing; **Cravero F:** Writing – Original Draft Preparation, Writing – Review & Editing; **Cáceres-Molina J:** Writing – Original Draft Preparation, Writing – Review & Editing; **Ramírez-Sarmiento CA:** Writing – Original Draft Preparation, Writing – Review & Editing; **Palopoli N:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Meysman P:** Writing – Original Draft Preparation, Writing – Review & Editing; **Parra RG:** Conceptualization, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: The events mentioned in the article were partially supported by funds from uBiome, Universidad Nacional de San Martín and Asociación Argentina de Bioinformática y Biología Computacional (A2B2C). R.G.P. is a long-term EMBO postdoctoral fellow (ALTF 212-2016). A.M.M holds a PhD fellowship from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2017 Monzon AM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Monzon AM, Hasenahuer MA, Mancini E *et al.* **Second ISCB Latin American Student Council Symposium (LA-SCS) 2016 [version 1; peer review: not peer reviewed]** F1000Research 2017, 6(ISCB Comm J):1491 (<https://doi.org/10.12688/f1000research.12321.1>)

First published: 16 Aug 2017, 6(ISCB Comm J):1491 (<https://doi.org/10.12688/f1000research.12321.1>)

Introduction

The Student Council (SC), part of the International Society for Computational Biology (ISCB), is a global organization that aims to nurture and impulse the next generation of bioinformaticians and computational biologists. The SC is composed of young scientists at all levels, from undergraduate, masters and PhD students to postdocs; coming from all disciplines in the field. Among the different activities that are managed by the SC, the most important ones consist of a set of symposia that are organized as satellite events of the different ISCB conferences. For more than a decade, the Student Council Symposium (SCS) has been annually organized¹⁻⁹. More recently, as the organization became bigger and different Regional Student Groups (RSGs) were created in more and more countries, continental versions for the SCS started to be organized. Since 2010, the European Student Council Symposium (ESCS) is bi-annually organized as a satellite event of the European Conference of Computational Biology (ECCB)^{4,6,8}. In parallel, the Latin American SC community has steadily grown, with the creation of RSG Argentina in 2012. In 2014, the first edition for the Latin American Student Council Symposium (LA-SCS) was held in Belo Horizonte, Brasil¹⁰. Since then, a total of five RSGs have been created in the region: RSG Argentina, RSG Brasil, RSG Chile, RSG Colombia and RSG Mexico. Two years later, in 2016, a team headed by Alexander Monzon from RSG Argentina as chair and Javier Caceres from RSG Chile as vice chair, organized the second LA-SCS. This edition was held in Buenos Aires, Argentina on November 19, 2016 as a satellite event to the Fourth Latin American conference from the International Society for Computational Biology (ISCB-LA 2016)¹¹.

Second Latin America Student Council Symposium in Buenos Aires, Argentina

The format of the Latin American Student Council Symposium is a day-long meeting preceding the main ISCB-LA conference. The main goals of this meeting include creating opportunities for young researchers to meet peers from all over the world, promoting the exchange of ideas and providing networking opportunities. In total, there were 65 attendees from different countries in the region and other continents as well (Figure 1). Travel fellowships by uBiome allowed promising researchers to attend different events during the week.

The second LA-SCS had the pleasure to welcome two renowned scientists as keynote speakers: Dr. Seán O'Donoghue from the Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO) and Prof. Ruth Nussinov from the National Cancer Institute, Center for Cancer Research (CCR), United States of America and Tel Aviv University, Israel.

The symposium received 40 submissions from students that were peer-reviewed by 14 independent reviewers. 13 abstracts were selected for oral presentation and 27 additional abstracts were accepted for poster presentations.

Abstracts of poster presentations are available online in the symposium booklet (<http://lascs2016.iscbsc.org/lascs2016-booklet>).



Figure 1. Delegates who attended to the second LA-SCS 2016.

Keynote speakers

The first keynote speaker was Dr. Seán O'Donoghue who is a senior principal research scientist in Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO) as well as group leader at the Garvan Institute of Medical Research in Sydney, Australia. His talk was entitled "Bioinformatics: a happy hunting ground for data scientists" in which he elaborated upon interesting highlights and turning points in his career which brought him to become a bioinformatician. The aim of his talk was to give students a good piece of advice, specifically focused on how the decisions made during the career help to find one's own way in bioinformatics.

The second keynote presentation was delivered by Prof. Ruth Nussinov who is senior investigator in the Cancer and Inflammation Program, CCR, USA and a professor emeritus in the Department of Human Genetics, School of Medicine, Tel Aviv University, Tel Aviv, Israel. She also serves as Editor-in-Chief of the journal *PLoS Computational Biology*. She is a world reference in structural biology and bioinformatics and her talk was centered on modeling protein-protein interactions for peptide targeting. She went into detail about PRISM¹², an application that uses a novel prediction algorithm for protein-protein interactions.

Student and early-career researchers' presentations

The student presentations covered a wide range of topics in computational biology. The first student talk was presented by Yesid Cuesta, who shared an integrative method to unravel host-parasite interactomes based on an orthology approach, to understand parasite infection and local adaptation within the host. This could help to identify drug targets among genome sequence and provide a better understanding of parasite evolution behind infectious diseases.

Ariel Aptekmann reported a positive correlation between core promoter information content and optimal growth temperature in archaeal organisms, suggesting selective pressures towards binding sites with higher binding affinity to the proteins. Also,

Aptekmann suggested to extend the molecular information theory between the Rsequence and Rfrequency measures in order to take into account the effect of temperature.

Emilio Fenoy presented the NetPhosPan software, a phosphorylation site predictor. This tool is based on feed-forward and a long-short term memory neural network, which extracts information from both ligand and receptor sequences. It has a higher accuracy in small datasets and uncharacterized kinases, compared to other methods.

Osvaldo Burastero described how the autophosphorylation mechanism in histidine kinases could be studied using QM/MM hybrid technics, by the analysis of different quantum level approximations and evaluation of several possible reaction mechanisms to find a concerted one-step mechanism.

Finishing the morning session, Tadeo Enrique Saldaño explored the vibrations of Human transthyretin. He found that the thyroxine hormone generates a significant change in the vibration level of the tetramer that could be partly responsible for its stabilization.

After the lunch break Daniel Almonacid and Juan Pablo Cárdenas from uBiome, the second LA-SCS platinum sponsor, gave a tech talk sharing their experiences of working in a microbial genomics company and their research career leading up to it. UBiome is a pioneer company in the newborn era of the microbiome-based precision medicine. During the tech talk, delegates were invited to collect samples of their own saliva using provisional kits. Results of their microbiome analyses were made accessible through the QR code on the uBiome website.

After the tech talk, Maria Freiberger explored if enzyme's catalytic sites are enriched in energetically conflictive (frustrated) interactions. By applying the Frustratometer software on all Catalytic Site Atlas structures, she found that residues that surround the catalytic sites or are in close proximity to metal cofactor binding sites were enriched in highly frustrated interactions. This was shown in a well-controlled study of the beta-lactamase family, observing that residues at the catalytic site were systematically in an energetic conflict with their environment.

Diego Zea shared his study which showed how large the structural space, implicitly encoded in a multiple sequence alignment of a protein family is. This large structural space leaves evolutionary signals which can be misinterpreted when only one structure from the family is analysed. Using at least one structure from four different sequence clusters, at 62% identity, it is possible to get a better description of the structural space of the family that can help in the understanding of the evolutionary signals.

Franco Simonetti presented how protein families within super-families' can be clustered by coevolution residue networks. These findings provide a base to develop novel computational methods using these residues to better classify protein families with respect to their functionality.

Elin Teppa reported that conservation and coevolution at the protein-protein interface increase by the number of interacting partners, suggesting that constraints in a given position and changes in the sequence are directly related, providing novel information at the protein interface and protein-protein interaction.

Pieter Meysman presented a computational interaction model to study the affinity of the varicella-zoster virus (VZV) peptides under different HLA variants. His results strongly support the hypothesis that one of the possible underlying causes of the VZV disease severity and susceptibility is a suboptimal anti-VZV immune response due to weak HLA-binding peptide affinity.

Juan Pablo Bustamante showed the VarQ tool for structural analysis of protein variation. The software was built considering both available structural analysis tools and the Ruffus workflow system. The goal of VarQ is to help understand, analyze and discriminate the possible effect of annotated protein variations.

In the final oral presentation of the day, Soledad Ochoa reported a better cancer classification method based on mutational patterns of loci in different cancer types. This classification framework not only improves diagnosis but also has the potential for making treatment recommendations, based on similar cancer types.

Satellite workshop

Following the format of the first LA-SCS 2014, a satellite workshop was organised to accompany LA-SCS.

We consider these kind of activities an excellent opportunity to promote knowledge exchange between students coming from different areas, such as computer science and biological careers. Taking advantage of this heterogeneity, we give PhD. students, postdocs and researchers the chance to offer workshops arranged as basic one-day courses for students, before or after the symposium day.

This time, Dr. Seán O'Donoghue and Dr. Alan Bush (from Universidad Nacional de Buenos Aires) organized and presented an eight-hour workshop entitled: "DataViz Workshop: Data Visualisation Methods and Tools - A Practical Guide", which was held on November 18th. A total of 35 students attended the workshop. The course was structured in two blocks. First, Dr. Seán O'Donoghue offered a tour about the state-of-the-art methods and practices for turning data into insightful visualisations in order to tell compelling stories, using principles of human visual perception with modern methods and tools. He presented different tools for visualizing hierarchical, categorical, time-serial and multidimensional data, and strategies for tackling the problem of large and complex data. Then, Dr. Alan Bush presented an introduction to ggplot2, a powerful library programmed in the R programming language for generating graphics and data visualization. In a "hands-on" modality, the course went through the analysis of some examples, tutorials and practical exercises.

Award winners

Winners for the Best Poster/Presentation Awards sponsored by F1000Research were selected by non-compulsory plurality vote of the attendees who filled up a blank ballot with the names or

poster numbers of their candidates. The Outstanding Presentation Prizes were received by Emilio Fenoy (Universidad Nacional de San Martín - CONICET, Argentina), who presented “NetPhosPan: a pan-specific predictor for phosphorylation site predictions”; Lionel Uran Landaburu (Universidad Nacional de San Martín - CONICET, Argentina) for his work “Updates to the TDR Targets chemogenomics database”; and Elin Teppa (Fundación Instituto Leloir - CONICET, Argentina) who was selected for his oral talk on “Conservation and coevolution at the protein-protein interface increase with the number of interacting partners”. All winners received a certificate and discounts on article publications costs in the F1000Research open access platform.

Conclusions

In 2014, the LA-SCS was organized for the first time as an attempt to gather together the students, postdocs and young researchers in the field of Bioinformatics and Computational Biology from Latin America. Back then, organization was led by a few representatives from RSG Argentina, the only active one in the region, and supervised by external members from the SC executive team. Many doubts appeared along the process and several obstacles arose due to the lack of local collaborators in the organizing country and the surrounding ones. Thankfully, due to great effort and support from different newly discovered volunteers, everything went smoothly. Beyond the success for that event, the most important thing was to pave the road towards setting an international network of motivated students that aimed to increase the presence of the SC in the Latin American continent. Two years later, with five RSGs in Argentina, Brazil, Chile, Mexico and Colombia, and an extended network of collaborations through virtual channels, the horizons for this new edition were quite promising. As we envisioned in the highlights from the first LA-SCS¹⁰, many challenges appeared but thanks to the previous experience and the support of excellent regional collaborators, we were able to overcome most of them and minimize the impact of those that were out of our control. We have learned invaluable lessons from the successful decisions but maybe even more, from our mistakes¹³. The experience gained will help us to continue growing as a continental collective towards bigger projects in the near future.

The third LA-SCS will be organized in 2018 in Colombia, and although the bar is now set high, we are confident that a new team will take over and repeat the success from this edition. We encourage all students from the Latin American continent to get in touch with the ISCB Student Council representatives in their countries or close ones in order to participate in the third LA-SCS organization and become a part of the development of our regional Bioinformatics and Computational Biology community.

Competing interests

No competing interests were disclosed.

Grant information

The events mentioned in the article were partially supported by funds from uBiome, Universidad Nacional de San Martín and Asociación Argentina de Bioinformática y Biología Computacional (A2B2C). R.G.P. is a long-term EMBO postdoctoral fellow (ALTF 212-2016). A.M.M holds a PhD fellowship from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The LA-SCS Committee is greatly indebted to the ISCB-LA 2016 conference chair, Fernán Agüero, for giving us the opportunity of holding the symposium in the Universidad Nacional de San Martín and for his predisposition and assistance in organizing the LA-SCS 2016. We greatly appreciate the support of International Society for Computational Biology (ISCB) and Argentinian Association for Computational Biology and Bioinformatics (A2B2C) for giving us the opportunity to have the second LA-SCS 2016 in Buenos Aires. In addition, we would like to thank the ISCB President Prof. Alfonso Valencia, and the president of the A2B2C Dr. Gustavo Parisi for their unconditional support to the Student Council and the associated RSGs in charge of the organization.

The SC is grateful for the support and assistance of the ISCB Executive Director Diane Kovats and the ISCB Coordinator Belinda Hanson in the organization of the LA-SCS 2016.

The LA-SCS team would like to extend their gratitude toward the 2016 SC executive team members (Alexander Junge, Anupama Jigisha, Sayane Shome, Jakob Jespersen, Farzana Rahman) for their advice and assistance as and when required to make the symposium a great success.

The LA-SCS Committee would also like to thank our keynote speakers Prof. Ruth Nussinov and Dr. Seán O’Donoghue. Their participation contributed to the success of this Symposium and helped support the next generation of bioinformaticians.

We are extremely grateful for the financial support that we received from our sponsors uBiome and F1000 Research. Without their help it would have been impossible to cover organization costs, offer travel fellowships for students and grant presentations awards at the second LA-SCS 2016. Finally, we would like to thank all people who were involved in the organization and made the second LA-SCS possible.

References

1. Rahman F, Wilkins K, Jacobsen A, *et al.*: **Highlights from the tenth ISCB Student Council Symposium 2014.** *BMC Bioinformatics.* 2015; **16** Suppl 2: A1–10.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Di Domenico T, Prudence C, Vicedo E, *et al.*: **Highlights from the ISCB Student Council Symposium 2013.** *BMC Bioinformatics.* 2014; **15**(Suppl 3): A1.
[Publisher Full Text](#) | [Free Full Text](#)
3. Goncarencu A, Grynberg P, Botvinnik OB, *et al.*: **Highlights from the Eighth International Society for Computational Biology (ISCB) Student Council Symposium 2012.** *BMC Bioinformatics.* 2012; **13**(Suppl 18): A1.
[Free Full Text](#)
4. Grynberg P, Abeel T, Lopes P, *et al.*: **Highlights from the Student Council Symposium 2011 at the International Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology.** *BMC Bioinformatics.* 2011; **12**(Suppl 11): A1.
[Publisher Full Text](#) | [Free Full Text](#)
5. Klijn C, Michaut M, Abeel T: **Highlights from the 6th International Society for Computational Biology Student Council Symposium at the 18th Annual International Conference on Intelligent Systems for Molecular Biology.** *BMC Bioinformatics.* 2010; **11**(Suppl 10): 11.
[Publisher Full Text](#) | [Free Full Text](#)
6. Abeel T, de Ridder J, Peixoto L: **Highlights from the 5th International Society for Computational Biology Student Council Symposium at the 17th Annual International Conference on Intelligent Systems for Molecular Biology and the 8th European Conference on Computational Biology.** *BMC Bioinformatics.* BioMed Central; 2009; **10** Suppl 13: I1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Peixoto L, Gehlenborg N, Janga SC: **Highlights from the Fourth International Society for Computational Biology Student Council Symposium at the Sixteenth Annual International Conference on Intelligent Systems for Molecular Biology.** *BMC Bioinformatics.* BioMed Central; 2008; **9**(Suppl 10): I1.
[Publisher Full Text](#) | [Free Full Text](#)
8. Francescato M, Hermans SM, Babaei S, *et al.*: **Highlights from the Third International Society for Computational Biology (ISCB) European Student Council Symposium 2014.** *BMC Bioinformatics.* 2015; **16**(Suppl 3): A1–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Cuypers B, Jacobsen A, Siranosian B, *et al.*: **Highlights from the ISCB Student Council Symposia in 2016 [version 1; referees: not peer reviewed].** *F1000Res.* 2016; **5**: pii: ISCB Comm J-2852.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Parra RG, Simonetti FL, Hasenahuer MA, *et al.*: **Highlights from the 1st ISCB Latin American Student Council Symposium 2014.** *BMC Bioinformatics.* 2015; **16** Suppl 8: A1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Palopoli N, Monzon AM, Parisi G, *et al.*: **A report on the “International Society for Computational Biology - Latin America (ISCB-LA)” Bioinformatics Conference 2016.** *EMBnet.journal.* 2017; **23**: e883.
[Publisher Full Text](#)
12. Tuncbag N, Gursoy A, Nussinov R, *et al.*: **Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM.** *Nat Protoc.* 2011; **6**(9): 1341–1354.
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Mishra T, Parra RG, Abeel T: **The upside of failure: how regional student groups learn from their mistakes.** *PLoS Comput Biol.* 2014; **10**(8): e1003768.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com










F1000Research

**APÊNDICE VI – Editorial: Nurturing tomorrow’s leaders: The ISCB Student Council
Symposia in 2018**



EDITORIAL

Nurturing tomorrow's leaders: The ISCB Student Council Symposia in 2018 [version 1; peer review: not peer reviewed]

Daniele Parisi ^{1*}, Gabriel J. Olguín-Orellana^{2*}, Eli J. Draizen ^{3*},
Nilson Da Rocha Coimbra ⁴, Nikolaos Papadopoulos⁵, Susanne Kirchen⁶,
Yvonne Saara Gladbach ⁷⁻⁹, Numrah Fadra¹⁰, Nazeefa Fatima¹¹,
Aishwarya Alex Namasivayam⁶, Sayane Shome ¹², Dan DeBlasio ¹³,
Alexander M. Monzon ¹⁴, Farzana Rahman ¹⁵, R. Gonzalo Parra ¹⁶

¹ESAT-STADIUS, KU Leuven, Heverlee, Belgium

²Center for Bioinformatics and Molecular Simulations, Universidad de Talca, Talca, Chile

³Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, USA

⁴Bioinformatics Graduate Program, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

⁵Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

⁶Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg

⁷Institute for Biostatistics and Informatics in Medicine and Ageing Research, IBIMA - Rostock University Medical Center, Rostock, Germany

⁸Faculty of Biosciences, Heidelberg University, Heidelberg, Germany

⁹Division of Applied Bioinformatics, German Cancer Research Center and National Center for Tumor Diseases Heidelberg, Heidelberg, Germany

¹⁰Division of Biomedical Statistics and Informatics and Mayo Clinic and Department of Bioinformatics and Computational Biology, University of Minnesota, Minneapolis, MN, USA

¹¹Department of Biology, Lund University, Lund, Sweden

¹²Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA, USA

¹³Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA

¹⁴Department of Biomedical Sciences, University of Padova, Padova, Italy

¹⁵Genomics and Computational Biology Research Group, University of South Wales, Cardiff, UK

¹⁶Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

* Equal contributors

v1 First published: 09 Jan 2019, 8(ISCB Comm J):34 (<https://doi.org/10.12688/f1000research.17739.1>)

Latest published: 09 Jan 2019, 8(ISCB Comm J):34 (<https://doi.org/10.12688/f1000research.17739.1>)

Abstract

The Student Council of the International Society for Computational Biology (ISCB-SC) is a student-focused organization for researchers from all early career levels of training (undergraduates, masters, PhDs and postdocs) that organizes bioinformatics and computational biology activities across the globe. Among its activities, the ISCB-SC organizes several symposia in different continents, many times, with the help of the Regional Student Groups (RSGs) that are based on each region. In this editorial we highlight various key moments and learned lessons from the 14th Student Council Symposium (SCS, Chicago, USA), the 5th European Student Council Symposium (ESCS, Athens, Greece) and the 3rd Latin American Student Council Symposium (LA-SCS, Viña del Mar, Chile).

Not Peer Reviewed

This article is an Editorial and has not been subject to external peer review.

Any comments on the article can be found at the end of the article.

Keywords

symposia, ISCB, student council



This article is included in the [International Society for Computational Biology Community Journal gateway](#).



This article is included in the [KU Leuven collection](#).

This article is included in the [Iowa State University collection](#).

Corresponding authors: Farzana Rahman (frahmann@gmail.com), R. Gonzalo Parra (parra.gonzalo@gmail.com)

Author roles: **Parisi D:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Olguín-Orellana GJ:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Draizen EJ:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Da Rocha Coimbra N:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Papadopoulos N:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kirchen S:** Writing – Original Draft Preparation; **Gladbach YS:** Writing – Original Draft Preparation; **Fadra N:** Writing – Original Draft Preparation, Writing – Review & Editing; **Fatima N:** Writing – Original Draft Preparation, Writing – Review & Editing; **Alex Namasivayam A:** Writing – Original Draft Preparation, Writing – Review & Editing; **Shome S:** Writing – Original Draft Preparation, Writing – Review & Editing; **DeBlasio D:** Writing – Original Draft Preparation, Writing – Review & Editing; **Monzon AM:** Writing – Original Draft Preparation, Writing – Review & Editing; **Rahman F:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Parra RG:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: All authors from this article are members of the ISCB Student Council.

Grant information: This work and the events here described have been supported by funding provided by the International Society for Computational Biology (ISCB). The authors would also thank all our sponsors. SCS was sponsored by Elsevier, the Data Science Institute from University of Virginia, the Master of Biomedical Informatics programme from Harvard University, Oxford University Press and Paperpile. ESCS was sponsored by Paperpile, PLoS and the Jackson Laboratories. LA-SCS was sponsored by the IDPfun consortium.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Parisi D *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Parisi D, Olguín-Orellana GJ, Draizen EJ *et al.* **Nurturing tomorrow's leaders: The ISCB Student Council Symposia in 2018 [version 1; peer review: not peer reviewed]** F1000Research 2019, 8(ISCB Comm J):34 (<https://doi.org/10.12688/f1000research.17739.1>)

First published: 09 Jan 2019, 8(ISCB Comm J):34 (<https://doi.org/10.12688/f1000research.17739.1>)

Introduction

The International Society for Computational Biology Student Council (ISCB-SC) is composed of and led by student and post-doctoral researchers. Since its origin, the ISCB-SC has been instrumental in organizing symposia, workshops, and seminars each year to introduce the newest computational biologists to the field and vice versa, as well as to empower them to communicate and participate in the community. The cornerstone of these events are the four annual symposia that accompany the major ISCB conferences (ISMB, ECCB, ISMB-LA, ISCB Africa). The symposia are headlined by keynote lectures from renowned researchers, who, apart from discussing science, usually also provide valuable career advices for the young audience. Organizing such meetings requires collective work that spans several months, where the organizing committee needs to accomplish multiple tasks in a timely fashion in order to meet the quality standards set by previous organizers. Being part of such a team requires motivation, commitment, resilience and the development of different leadership and soft skills.

In this article we highlight three such symposia that happened in 2018: the 14th Student Council Symposium (SCS) in Chicago, USA; the 5th European Student Council Symposium (ESCS) in Athens, Greece and the 3rd Latin American Student Council Symposium (LA-SCS) in Viña del Mar, Chile. In doing so, we also summarize the main goals of the symposia, our general structure, and how each of the individual events is tailored for its specific purpose.

The networker scientist

The image of a scientist as a bespectacled genius with unruly hair, lab coat, clipboard in hand, isolated in a laboratory, surrounded by mysterious technology and fizzling test tubes seems ingrained in our culture¹. While this cliché may have held some truth in the past, a modern scientist needs much more than a sharp mind and good analytical skills in order to be successful in the current scientific system. Science, particularly at the interface of biological and computational disciplines, has become extremely complex and demands researchers from different fields to form interdisciplinary teams and to establish strategic and strong collaborations to reach sustainable, long-term success. Because of this, scientists need to develop a whole toolbox of both analytical and technical skills, but just as importantly, a set of soft skills that allow them to surf the complex waters of human interactions and leadership. Successful scientists are no longer fully enclosed in the laboratory performing experiments, but on the contrary spend a significant amount of their time networking, tweeting, blogging, writing divulgation books and so on. Keynote speakers at big conferences usually hold big personalities, they own the stage with their presence and captivate the audience with their communication skills. Regardless of how unreachable these scientific profiles might seem, all these people built their careers from scratch and had to work their way towards the position they occupy now. Additionally, in general, these researchers are hubs of large collaborative networks, and interacting with them benefits the student community by granting the opportunity to get in touch with closely and distantly related fields, expanding their minds to new career horizons^{2,3}.

Training a new generation of research leaders

Since 1997, the International Society for Computational Biology (ISCB) has organized meetings for gathering together the most world-renowned researchers in the fields of bioinformatics and computational biology. Early on, the ISCB directors realized that next to growing and strengthening the field, there lies a pressing need to nurture the next generation of computational biology researchers that would eventually take the lead from their hands. In 2005, the ISCB Student Council was created as a forum to bring together the next generation of computational biologists⁴. The ISCB-SC is composed of Regional Student Groups (RSGs) located all around the world, several committees to handle the functioning of the council as a whole, and an executive team (ET) coordinating and directing their efforts. Every year, the ISCB-SC organizes various networking and training activities, among which there is a series of symposia that constitute the flagships of the organization⁵. A chair and a co-chair lead each symposium's organization and are in charge of building a team that executes different tasks that are required to generate a successful event. Organization of an ISCB-SC Symposium typically takes more than half a year of preparation. During this time the team is responsible for a diverse set of activities including contacting potential keynotes, fundraising, outreach communication, abstracts collection and evaluation, and social event planning. Throughout this process the chair and co-chair are responsible for keeping a fluent communication, keeping motivation and the collaborative mood among the team members, and dealing with unforeseen/problematic situations. Additionally, the chair, co-chair, and other key team members are responsible to lead the event and putting themselves in front of an audience for an entire day. For more than a decade, the ISCB-SC Symposia have been a platform for the new generation of computational biologists and has served as a source for the emergence of new leaders in our community.

ISCB-SC Symposia in 2018

The ISCB-SC organizes four flagship events. The Student Council Symposium (SCS) is the only event held annually and has been organized since 2005⁶⁻⁸. The European Student Council Symposium (ESCS) is held biennially and started in 2010⁹. Every two years SCS is held in Europe jointly with the ESCS as a jointly event⁵. Since 2014 the Latin American Student Council Symposium (LA-SCS) has been biennial organized^{10,11} and since 2015 the African Student Council Symposium (SCS-Africa) has also taken place biennially^{12,13}. In 2018 the ISCB-SC organized the 14th SCS in Chicago, USA, the 5th ESCS in Athens, Greece and the 3rd LA-SCS in Viña del Mar, Chile.

Keynotes

Listening to a great keynote talk is always an inspiring experience. When attending a symposium, keynotes are one of the main attractions; it is a reason to pay the registration fee and spend an entire day in a given event. Summoning great keynote speakers is one of the most critical issues ISCB-SC symposia chairs have to face. Not only has the symposia series gathered top notch researchers in the field to deliver superb lectures, but it has

also offered a unique environment where students can deeply interact with them. In a normal conference, with hundreds of senior scientists as part of the audience, asking a question or addressing a keynote personally after the talk is an intimidating experience for most young people. At the student council symposia, keynotes not only are accessible to the students during their talks and after, but also they prepare lectures that touch on topics beyond their scientific work including anecdotes about their career path that led them to become successful researchers in their fields. The 2018 Symposia were no exception.

In Chicago, SCS 2018 featured a talk by Dr. Lucia Peixoto from Washington State University titled “Learning-dependent chromatin remodeling highlights non-coding regulatory regions linked to Autism”. In addition to an excellent presentation, one of the great contributions of Dr. Peixoto is that she served as a co-founder of our very own Student Council back in 2004; she and her colleagues from the first ISCB-SC executive team paved the way for developing this organization and hence having her as a keynote after all these years is a big inspiration for all of us. The second SCS keynote lecture was delivered by Dr. Philip Bourne and named “Eight (so far) things I wish I had thought 40 years ago.” Dr. Bourne has been a key figure in bioinformatics, from being a professor at the UC San Diego to becoming the director of the University of Virginia’s Data Science Institute. In the style of his “Ten Simple Rules” pieces, he shared his views on how to become a successful biological data scientist, which he learned over the years. He also emphasized that science is a team sport, and that collaboration, management, communication and administration are just as important as the science. Phil Bourne is a great example of ‘the networker scientist.’

In Athens, the ESCS 2018 began with a talk by Dr. Anna Zhukova (Institut Pasteur, Paris), who presented her work on “modelling the spread of HIV-1 resistance mutations” as well as her career path. The second keynote was delivered by Dr. Julio Saez-Rodriguez from the University of Heidelberg. His research combines mathematical models with biological data to explain disease mechanisms, such as logical models of signaling networks trained on data from mass spectrometry and antibody assays. Dr. Saez-Rodriguez also highlighted the importance of collaboration, both on a personal basis and on a community level: his group collaborates with many experimental groups and regularly participates in DREAM challenges, a community effort that promotes large-scale cooperation to answer specific biological questions. When asked for career advice, he commented that his career was not the result of meticulous planning, but rather the drive to research important questions, academic excellence, and actively looking for opportunities.

In Viña del Mar, for the 3rd LA-SCS, the oral session began with Dr. Wendy Gonzalez, head of the Center for Bioinformatics and Molecular Simulations at Universidad de Talca. In her talk “Molecules Son: the dance I am trying to learn,” she reflected on the history of her experience in studying ion channel associated diseases. The title refers to Son, the genre of music and dance of her home country, Cuba; e.g. the drugs “dance” as they move into channels associated with neoplasm and atrial

fibrillation. Later, Dr. Francisco Melo from the Faculty of Biological Sciences at Pontificia Universidad Católica de Chile, referred to his vision about past, present and future challenges in bioinformatics research from the perspective of being located in Latin America, through his talk titled: “A perspective about doing research in molecular Bioinformatics from Chile: from past to present (and future ...) challenges”. Finally, attendees at LA-SCS boarded a rocket to explore the Astrobiology field with Dr. David S. Holmes, founder of the Iberoamerican Society for Bioinformatics (SOI-BIO) and Head of the Center for Bioinformatics and Genome Biology at Fundación Ciencia y Vida. During his talk “Earth environments as analogs for extraterrestrial life”, he explored examples of sites that provide information on how physical and chemical conditions interact to form environments conducive to life and how metagenomics is being used to tease out essential genes and metabolisms needed to tailor-make microbes for applications in extra-terrestrial environments.

Student highlights

Student talks represent the core of the ISCB-SC symposia. Within this assigned time, young researchers in bioinformatics have the chance to practice their presentation skills and receive useful comments from other peers. The ISCB-SC symposia play an important role to prepare students for more prestigious, but also demanding, stages. This year the three ISCB-SC symposia contained extremely high caliber scientific talks from students coming from many different universities and institutes, thanks to the now well-known events that preceded these as well as the careful selection process.

SCS in Chicago featured 10 student talks and 34 poster presentations that were selected through a competitive review process. During the symposium, the presenters were judged by delegates using an anonymous voting scheme. Best presentation awards went to Carolin Loos (Helmholtz Zentrum München) and Ben Siranosian (Broad Institute); best poster awards went to Susanne Pieschner (Helmholtz Zentrum München), Michael Scherer (Max Planck Institute for Informatics), and Susanne Kirchen (University of Luxembourg). We consider useful to give our audience of students and researchers the opportunity to critically evaluate and take a main role in the review process.

For ESCS 2018, 11 projects were presented as full talks and six in a 5-minute flash talk format. The decision for awarding the best talk and the best poster was made by all 50 participants, again by anonymous vote. The best talk award was given to Melissa Adasme (BIOTEC TU Dresden), for her talk “From malaria to cancer, computational drug repositioning of amodiaquine using PLIP interaction patterns”. The two best poster prizes went to Neetika Nath (University Medicine Greifswald) and Dilip Ariyur Durai (Max Planck Institute for Informatics).

During LA-SCS, the Student Council awarded the three most outstanding presentations. The jury was confirmed by the chairs of the symposium, who selected the winners among the

9 student talks and 16 poster presentations. The best student talk award was conferred to Juanita Gil (Universidad de los Andes), with her talk titled “Accurate, efficient and user-friendly simulation and mutation calling for TILLING experiments”. The best poster award was received by Mauricio Bedoya (Universidad de Talca) for his work “Relevance of extracellular portals in the potassium K2P ion channel conduction mechanism”. An honor mention was given to David Medina (Universidad de Chile) for his talk “VHL-Hunter, a web service for classification of clinical relevance in single point mutations in Von Hippel-Lindau disease”.

A novelty in the student symposium structure: round table “Bioethics and bioinformatics”

For the first time in ISCB-SC symposia history, ESCS 2018 hosted a roundtable about bioethics and its importance in the field of bioinformatics. The roundtable aimed to increase students’ awareness about current popular topics such as data sharing, data protection and social hazards of technology. To start, Dr. Yves Moreau delivered a talk titled “‘Build it and they will come’, or how will we prevent abuses of clinical genomic databases?” followed by Dr. Mahsa Shabani who talked about “Ethical concerns associated with data sharing in biomedical sciences”. After, Dr. Moreau and Dr. Shabani were joined by the two ESCS keynotes, Dr. Saez Rodriguez and Dr. Zhukova, for an open discussion about bioethical issues in bioinformatics and computational biology in general. The students engaged in a stimulating and vivid discussion about the impact of bioinformatics in future society. We encourage future chairs and co-chairs to organize similar events that stimulate critical thinking about the next big trends in our field.

Challenges and lessons learned

Success is the result of planning, hard work, determination, foresight, and a little bit of luck. However, regardless of how much planning or how much experience the organizers or their advisors have, unforeseen situations will often arise and challenge the team’s resilience and versatility. Although these conflicting situations can be discouraging and time consuming they present valuable learning opportunities that serve to strengthen the operational skills of the members of the team. The ISCB-SC has previously discussed the important role of challenging situations in the road to success and this still holds true: “*difference between those who succeed and those who abandon their projects lies in their response to adversity*”¹⁴.

It is to be expected that problems will arise, possibly from the very beginning of the conference organization and, therefore, the team needs to come up with feasible and timely solutions. For many of the chairs and co-chairs, 2018 was their first time organizing a symposium. Here we share different stories about challenging situations and how we overcame them. We aim to encourage next chairs and co-chairs are prepared for when things do not go according to the initial plan. Many of these situations apply to all symposia while others are symposium specific.

Risk assessment and mitigation

The organization of a symposium is, on its whole, an exercise in risk assessment and mitigation. Things deviate from the

plan all the time; there are scheduling conflicts between team meetings and unexpected academic obligations, a team member cannot meet a deadline, delegates have forgotten their posters, a student speaker or keynote cannot attend due to health or visa issues, a keynote speaker goes missing last-minute and many other examples. The early setbacks teach you to take complications like those into account. By the end of the project you learned to have a contingency plan and you have extensively practiced on improvising for all those problems you could not foresee.

Importance of interpersonal cooperation

On the other hand, these setbacks foster team spirit and help you to improve as a team player. It is impossible for the symposium chairs to personally take care of everything - after all they are volunteers as well and have their own academic obligations. You learn to encourage and delegate responsibility. You cultivate flexibility by jumping in to cover for other team members when needed, and they do it in turn.

Team cooperation can only be achieved by mutual respect and focus on the common goal. Keeping this in mind will help you navigate differences in opinion and learn from them, effectively managing conflicts within the team and using them constructively instead of letting them derail the effort. This often requires the flexibility to accept solutions and initiatives that don’t agree with your own vision, as long as they benefit the team. It is important to establish objective criteria for success and make evaluations according to them, and not according to how you would have done things personally.

When coordinating a team with many members, time zones and obligations, fluent and organized communication is crucial. It is mandatory to have a centralized communication channel where messages can be easily delivered to all team members and information is quickly available to everyone at any time. This makes it easier to coordinate and keep track of progress, and also provides records of discussions and meetings, a valuable and easily accessible resource for future organizers.

Importance of promotion and outreach

One of the most crucial challenges is related to how to promote the event and reach a wide audience. At the end of the day, regardless of how good the keynotes, the venue and the organization are, much of the success is reflected on the amount of attendants that register and present their work at the event. The Student Council might not be on everyone’s radar, or people might not be aware of how the symposium is different from the main conference. While we mainly used social media (Twitter and Facebook), there is quite room for improvement. Most RSGs utilize social media and, therefore, could work in a more coordinated manner spreading the important information of the symposia. This requires an active role of the outreach committee, calling to the RSG community managers to share the announcements. Even though RSGs can cover large geographic regions, everyone needs to work together in the next years to find ways to promote participation of countries not present until now.

Finding the keynotes

This is perhaps the most important task the team needs to address. Over the years, we learned that if you aim to have the reputed scientists in the field as keynote speakers, you need to contact them as early as possible—it helps to network ahead of time so they know who you are before inviting. An early list of keynote speaker candidates is essential. Subsequently, just try to contact them one by one until the desired number of keynotes is confirmed. It is also important to have a backup plan in case things don't work out, such as week-of cancellations. While selecting keynotes, the ISCB-SC does its best to maintain gender balance and diversity, i.e. inviting women, people from different ethnicities, and members from other underrepresented communities. In the US, the field of bioinformatics is unfortunately dominated by white men. In order to change this, we believe that we need to highlight diversity and show newcomers into the field that everyone is accepted and welcome¹⁵.

Finding sponsors

Sometimes finding sponsors can be challenging. Some sponsors may not have heard of the Student Council and/or the parent organization ISCB. We suggest that the search for sponsors receives equal importance to keynote search and that the organizing committee also considers sponsors from outside the country where the symposium takes place. It is important to know that different companies of institutions have specific times in the year for when they need to decide on which sponsoring activities they will invest their money. Because of this an organized schedule for contacting sponsors as well as following up with them is needed. This year was a real challenge since we had to find sponsors for three different symposia. It is important to keep good relations with the sponsors and maintain collaborations with them after the event is over.

Unforeseen hurdles

Many unforeseen problems can and will occur during the symposium. Therefore, it is very important to be prepared: from lack of sponsors to student and speaker last minute changes/cancellations. Even satellite activities can give you a huge headache: e.g. when you realize you booked the social event for the wrong time or the catering forgets to bring the required food and drinks.

Problematic situations can arise from the most diverse internal or external sources. This year important economic processes happened in the Latin American continent, with many currencies being devalued in respect to the American dollar. This severely affected the organization and success of LA-SCS due to the registration cost. Since Chile is considered a High Income Country, based on World Bank ranking of economies¹⁶, the high symposium registration price was hard to pay for both national and international students from the region and had large consequences on the composition of attendants to the event. This forced the organizers to reduce the registration fees to a much lower level compared to standard costs on this type of symposium and to dedicate a substantial proportion of the budget to provide travel fellowships.

Conclusion

The importance of transferable skills such as teamwork and management, networking, leadership and effective communication, has been prioritized to achieve success both in the academic and scientific community¹⁷, with many researches showing a correlation between their networking and career outcomes¹⁸. In sight of this, it is crucial for young researchers to engage themselves in organizing and attending student symposia, where the development of transferable skills is proactively encouraged¹⁹, and the community increases its cohesiveness²⁰. These events represent important opportunities for students to present their own work to a community of peers, and gain skills that are indispensable on peer-review processes.

Co-organizing a symposium positively impacts on every member's networking opportunities. All team members are highly motivated, high achieving PhD students and postdocs who will become future colleagues and collaborators, future principal investigators or future board members. Symposia organizers get the chance to interact with high-profile scientists who are invited to give keynote talks. Cultivating relationships with sponsors increases your network and fosters collaboration at the same time that introduces the team members on how to deal with the fundamental task of fundraising for organizing scientific events. Certainly, the Student Council Symposia represent the major networking and education activities oriented to bioinformatics students around the world. Despite the geographical distance among the venues where the different events have been organized so far and the different cultures represented by the attendants, from its first edition they had managed its purpose: promote the develop good relationships between students, keynote speakers, sponsors, universities and scientific institutes.

Since the main ISCB conferences are large, it is important to help PhD students find peers early on so the main conference doesn't feel overwhelming. We strive to make it an accepting and welcoming community; ISCB-SC symposia are opportunities to practice giving talks in front of a large audience while fostering positive criticism and the establishment of new collaborations. This is further enhanced during the coffee breaks, the social event and a relaxed and stress-free atmosphere.

Time steadily advances and generational change is inexorable. In the next decades, new researchers will have to take over the command of our scientific community with the huge responsibility of keeping up to the level current board members have achieved. Conditions for international scientific development can become quite complex in a world where financial resources need to be carefully assigned, migration rules can change from one day to the other, gender policies are adapting to modern times where inclusion is a must and minorities need to be heard and respected. Moreover, individuals need to develop their skills in a much broader way to survive in a scientific system that becomes more competitive each day. Tomorrow's leaders need to become aware of all these topics and have to be able to correctly manage not only scientific related problems but also social related ones. These leading people will often need

to decide on subjects that are of political nature that need to be addressed with precision in order to defend the community's interests. So far, leaders cannot be collected from trees in a field or grown up in a petri dish, so we need to nurture them in the old-fashioned way by training, teaching and learning from each other. The ISCB-SC symposia have gathered the young generation of computational biology for more than a decade and have helped to produce researchers that are slowly getting into spotlight positions in the community. Many of tomorrow's leaders will necessarily emerge from the ISCB-SC and one of those can be you. Get involved and help us shape the future of this amazing community.

Future editions

The ISCB-SC Symposia series have become a successful platform to collaborate, train and nurture the future bioinformaticians who will pave the way to strengthen bioinformatics and computational biology community. Following the success of previous symposium, the future editions of symposia are planned for SCS2019 in Basel, Switzerland, SCS2020 in Toronto,

Canada, ESCS 2020 in Sitgens, Spain, plus editions of SCS-Africa in 2019 and LA-SCS in 2020.

Data availability

No data is associated with this article.

Grant information

This work and the events here described have been supported by funding provided by the International Society for Computational Biology (ISCB). The authors would also thank all our sponsors. SCS was sponsored by Elsevier, the Data Science Institute from University of Virginia, the Master of Biomedical Informatics programme from Harvard University, Oxford University Press and Paperpile. ESCS was sponsored by Paperpile, PLoS and the Jackson Laboratories. LA-SCS was sponsored by the IDPfun consortium.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Chambers DW: **Stereotypic images of the scientist: The draw-a-scientist test.** *Sci Educ.* 1983; **67**(2): 255–265.
[Publisher Full Text](#)
- Sciortino F: **Why organizing a scientific conference can produce huge benefits.** *Nature.* 2018; **559**(7714): 431.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mathé E, Busby B, Piontkivska H, *et al.*: **Matchmaking in Bioinformatics [version 1; referees: 2 approved].** *F1000Res.* 2018; **7**: pii: ISCB Comm J-171.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Corpas M: **Scientists & societies.** *Nature.* 2005; **436**(7054): 1204.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Hassan M, Namasivayam AA, DeBlasio D, *et al.*: **Reflections on a journey: a retrospective of the ISCB Student Council symposium series.** *BMC Bioinformatics.* 2018; **19**(Suppl 12): 347.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Grynberg P, Abeel T, Lopes P, *et al.*: **Highlights from the Student Council Symposium 2011 at the International Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology.** *BMC Bioinformatics.* 2011; **12**(Suppl 11): A1.
[Publisher Full Text](#) | [Free Full Text](#)
- Rahman F, Wilkins K, Jacobsen A, *et al.*: **Highlights from the tenth ISCB Student Council Symposium 2014.** *BMC Bioinformatics.* 2015; **16**(Suppl 2): A1–10.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wilkins K, Hassan M, Francescato M, *et al.*: **Highlights from the 11th ISCB Student Council Symposium 2015. Dublin, Ireland. 10 July 2015.** *BMC Bioinformatics.* 2016; **17**(Suppl 3): 95.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Francescato M, Hermans SM, Babaei S, *et al.*: **Highlights from the Third International Society for Computational Biology (ISCB) European Student Council Symposium 2014.** *BMC Bioinformatics.* 2015; **16**(Suppl 3): A1–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Parra RG, Simonetti FL, Hasenahuer MA, *et al.*: **Highlights from the 1st ISCB Latin American Student Council Symposium 2014. Introduction.** *BMC Bioinformatics.* 2015; **16**(Suppl 8): A1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Monzon AM, Hasenahuer MA, Mancini E, *et al.*: **Second ISCB Latin American Student Council Symposium (LA-SCS) 2016 [version 1; referees: not peer reviewed].** *F1000Res.* 2017; **6**: pii: ISCB Comm J-1491.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Souilmi Y, Allali I, Badad O, *et al.*: **Highlights of the first ISCB Student Council Symposium in Africa 2015 [version 1; referees: not peer reviewed].** *F1000Res.* 2015; **4**: pii: ISCB Comm J-569.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rafael CN, Ashano E, Moosa Y, *et al.*: **Highlights of the second ISCB Student Council Symposium in Africa, 2017 [version 1; referees: not peer reviewed].** *F1000Res.* 2017; **6**: pii: ISCB Comm J-2183.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mishra T, Parra RG, Abeel T: **The upside of failure: how regional student groups learn from their mistakes.** *PLoS Comput Biol.* 2014; **10**(8): e1003768.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Grogan KE: **How the entire scientific community can confront gender bias in the workplace.** *Nat Ecol Evol.* 2019; **3**(1): 3–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Fantom NJ, Serajuddin U: **The World Bank's classification of countries by income.** *The World Bank.* 2016; 1–52.
[Reference Source](#)
- Tomazou EM, Powell GT: **Look who's talking too: graduates developing skills through communication.** *Nat Rev Genet.* 2007; **8**(9): 724–726.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wolff HG, Moser K: **Do specific types of networking predict specific mobility outcomes? A two-year prospective study.** *J Vocat Behav.* 2010; **77**(2): 238–245.
[Publisher Full Text](#)
- de Ridder J, Meysman P, Oluwagbemi O, *et al.*: **Soft skills: an important asset acquired from organizing regional student group activities.** *PLoS Comput Biol.* 2014; **10**(7): e1003708.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ramdayal K, Stobbe MD, Mishra T, *et al.*: **Building the future of bioinformatics through student-facilitated conferencing.** *PLoS Comput Biol.* 2014; **10**(1): e1003458.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research

**APÊNDICE V – Editorial: Global network of computational biology communities:
ISCB's Regional Student Groups breaking barriers**



EDITORIAL

Global network of computational biology communities: ISCB's Regional Student Groups breaking barriers [version 1; peer review: not peer reviewed]

Sayane Shome ¹, R. Gonzalo Parra ², Nazeefa Fatima³,
Alexander Miguel Monzon ⁴, Bart Cuypers ^{5,6}, Yumna Moosa⁷,
Nilson Da Rocha Coimbra ⁸, Juliana Assis⁸, Carla Giner-Delgado⁹,
Handan Melike Dönertaş ¹⁰, Yesid Cuesta-Astroz ^{11,12}, Geetha Saarunya ¹³,
Imane Allali^{14,15}, Shruti Gupta ¹⁶, Ambuj Srivastava¹⁷, Manisha Kalsan¹⁶,
Catalina Valdivia¹⁸, Gabriel J. Olguin-Orellana¹⁹, Sofia Papadimitriou²⁰,
Daniele Parisi ²¹, Nikolaj Pagh Kristensen²², Leonor Rib²³, Marouen Ben Guebila²⁴,
Eugen Bauer²⁴, Gaia Zaffaroni ²⁴, Amel Bekkar ²⁵, Efejiro Ashano²⁶,
Lisanna Paladin⁴, Marco Necci⁴, Nicolás N. Moreyra ²⁷, Martin Rydén²⁸,
Jordan Villalobos-Solís²⁹, Nikolaos Papadopoulos³⁰, Candice Rafael ³¹,
Tülay Karakulak³², Yasin Kaya³³, Yvonne Gladbach ³⁴,
Sandeep Kumar Dhanda ³⁵, Nikolina Šoštarić ³⁶, Aishwarya Alex³⁷,
Dan DeBlasio³⁸, Farzana Rahman ^{39,40}

¹Bioinformatics and Computational Biology Program, Iowa State University, Iowa, USA

²Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

³Science for Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

⁴Department of Biomedical Sciences, University of Padova, Padova, Italy

⁵Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

⁶Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium

⁷KZN Research and Innovation Sequencing Platform, University of KwaZulu Natal, Durban, South Africa

⁸Graduate Program in Bioinformatics, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

⁹Institut de Biociències i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Spain

¹⁰European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK

¹¹School of Microbiology, Universidad de Antioquia, Medellín, Colombia

¹²Colombian Tropical Medicine Institute (ICMT), Universidad CES, Medellín, Colombia

¹³Department of Biological Sciences, University of South Carolina, South Carolina, USA

¹⁴Department of Biology, Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco

¹⁵Division of Computational Biology, Department of Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

¹⁶School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

¹⁷Department of Biotechnology, Indian Institute of Technology Madras, Chennai, India

¹⁸Ecosystem's Health Laboratory, Universidad Andres Bello, Santiago, Chile

¹⁹Center for Bioinformatics, Simulations and Modelling, Universidad de Talca, Talca, Chile

²⁰Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles-Vrije Universiteit Brussel, Brussels, Belgium

- ²¹ESAT-STADIUS KU Leuven, Heverlee, Heverlee, Belgium
- ²²DTU Health Technology, Technical University of Denmark, Lyngby, Denmark
- ²³The Bioinformatics Center, Biology and Biotech Research and Innovation Center, University of Copenhagen, Copenhagen, Denmark
- ²⁴Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg
- ²⁵Swiss Institute of Bioinformatics (SIB), University of Lausanne, Lausanne, Switzerland
- ²⁶Molecular Diagnostics, Laboratory Services, APIN Public Health Initiatives, Abuja, Nigeria
- ²⁷Genetics and Evolution of Buenos Aires (IEGEB), CONICET-UBA, Institute of Ecology, Buenos Aires, Argentina
- ²⁸Biomedical Centre, Faculty of Medicine, Lund University, Lund, Sweden
- ²⁹Laboratorio de Biotecnología de Plantas, Universidad Nacional de Costa Rica (UNA), Heredia, Costa Rica
- ³⁰Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany
- ³¹Research Unit for Bioinformatics, Rhodes University, Grahamstown, South Africa
- ³²Izmir Biomedicine and Genome Center, Dokuz Eylül University, Izmir, Turkey
- ³³Hacettepe University, Faculty of Science, Department of Biology, Ankara, Turkey
- ³⁴University Medical Center Rostock, University Heidelberg, Heidelberg, Germany
- ³⁵La Jolla Institute for Allergy and Immunology, La Jolla Institute for Immunology, California, USA
- ³⁶Centre of Microbial and Plant Genetics, KU Leuven, Leuven, Belgium
- ³⁷Roche Diagnostics Automation Solutions GmbH, Roche, Waiblingen, Germany
- ³⁸Computational Biology Department, Carnegie Mellon University, Pittsburgh, USA
- ³⁹Genomics and Computational Biology Research Group, University of South Wales, Pontypridd, UK
- ⁴⁰School of Human and Life Sciences, Canterbury Christ Church University, Kent, UK

v1 First published: 02 Sep 2019, 8(ISCB Comm J):1574 (<https://doi.org/10.12688/f1000research.20408.1>)

Latest published: 02 Sep 2019, 8(ISCB Comm J):1574 (<https://doi.org/10.12688/f1000research.20408.1>)

Abstract

Regional Student Groups (RSGs) of the International Society for Computational Biology Student Council (ISCB-SC) have been instrumental to connect computational biologists globally and to create more awareness about bioinformatics education. This article highlights the initiatives carried out by the RSGs both nationally and internationally to strengthen the present and future of the bioinformatics community. Moreover, we discuss the future directions the organization will take and the challenges to advance further in the ISCB-SC main mission: "Nurture the new generation of computational biologists".

Keywords

Student organizations, Symposia, Bioinformatics, Computational Biology, Workshops, Education, Virtual seminars, ISCB Student Council, Regional Student Groups, ISCB, early career bioinformaticians, collaboration, networking



This article is included in the [International Society for Computational Biology Community Journal gateway](#).

This article is included in the [Iowa State University collection](#).

Not Peer Reviewed

This article is an Editorial and has not been subject to external peer review.

Any comments on the article can be found at the end of the article.

Corresponding authors: Sayane Shome (sayaneshome.rsg@gmail.com), Farzana Rahman (frahmann@gmail.com)

Author roles: **Shome S:** Conceptualization, Data Curation, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Parra RG:** Conceptualization, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Fatima N:** Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Monzon AM:** Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Cuyppers B:** Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Moosa Y:** Writing – Original Draft Preparation, Writing – Review & Editing; **Coimbra NDR:** Writing – Original Draft Preparation, Writing – Review & Editing; **Assis J:** Writing – Original Draft Preparation, Writing – Review & Editing; **Giner-Delgado C:** Writing – Original Draft Preparation, Writing – Review & Editing; **Dönertaş HM:** Writing – Original Draft Preparation, Writing – Review & Editing; **Cuesta-Astroz Y:** Resources, Writing – Review & Editing; **Saarunya G:** Resources, Writing – Review & Editing; **Allali I:** Resources, Writing – Review & Editing; **Gupta S:** Resources, Writing – Review & Editing; **Srivastava A:** Resources, Writing – Review & Editing; **Kalsan M:** Resources, Writing – Review & Editing; **Valdivia C:** Resources, Writing – Review & Editing; **J. Olguin-Orellana G:** Resources, Writing – Review & Editing; **Papadimitriou S:** Resources, Writing – Review & Editing; **Parisi D:** Resources, Writing – Review & Editing; **Kristensen NP:** Resources, Writing – Review & Editing; **Rib L:** Resources, Writing – Review & Editing; **Guebila MB:** Resources, Writing – Review & Editing; **Bauer E:** Resources, Writing – Review & Editing; **Zaffaroni G:** Resources, Writing – Review & Editing; **Bekkar A:** Resources, Writing – Review & Editing; **Ashano E:** Resources, Writing – Review & Editing; **Paladin L:** Resources, Writing – Review & Editing; **Necci M:** Resources, Writing – Review & Editing; **Moreyra NN:** Resources, Writing – Review & Editing; **Rydén M:** Resources, Writing – Review & Editing; **Villalobos-Solís J:** Resources, Writing – Review & Editing; **Papadopoulos N:** Resources, Writing – Review & Editing; **Rafael C:** Resources, Writing – Review & Editing; **Karakulak T:** Resources, Writing – Review & Editing; **Kaya Y:** Resources, Writing – Review & Editing; **Gladbach Y:** Resources, Writing – Review & Editing; **Dhanda SK:** Resources, Writing – Review & Editing; **Šoštarić N:** Resources, Writing – Review & Editing; **Alex A:** Project Administration, Resources, Writing – Review & Editing; **DeBlasio D:** Conceptualization, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Rahman F:** Conceptualization, Project Administration, Resources, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: All authors are affiliated with the ISCB-SC Regional Student Group program.

Grant information: The events mentioned in the article were partially supported by funds from ISCB-Student Council, a subsidiary of the International Society for Computational Biology.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Shome S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Shome S, Parra RG, Fatima N *et al.* **Global network of computational biology communities: ISCB's Regional Student Groups breaking barriers [version 1; peer review: not peer reviewed]** F1000Research 2019, 8(ISCB Comm J):1574 (<https://doi.org/10.12688/f1000research.20408.1>)

First published: 02 Sep 2019, 8(ISCB Comm J):1574 (<https://doi.org/10.12688/f1000research.20408.1>)

Introduction

Regional Student Groups (RSGs) are student-oriented groups affiliated to the Student Council of the International Society of Computational Biology (ISCB-SC)¹. Aligned with the mission and objectives of the parent organization ISCB-SC, RSGs were formulated to promote networking amongst budding computational biologists in the local geographical regions. Since its formation in 2006 with four RSGs (Netherlands, India, Korea, and Singapore), the program has come a long way with 30 active RSGs operating around the world, together constituting a global network of over 2000 members. RSGs are completely autonomous and over the last decade, with the economic support of the ISCB-SC, have organized a large variety of activities according to their community needs.

Computational biology and bioinformatics are relatively new and multidisciplinary areas and hence undergraduate education on these topics is scarce. Instead, young researchers in these fields often come from other disciplines such as molecular biology, computer science or physics and need to complement their background education with knowledge from other disciplines. Additionally, the growing importance of computational biology in a wide range of biological fields is also motivating young researchers in pure experimental groups to get more expertise about this emerging field². For all these reasons, the RSGs are playing an instrumental role in promoting networking and knowledge transfer related to computational biology topics amongst student researchers. These are typically achieved by organizing offline and online networking and educational events such as workshops, symposiums, hackathons, online competitions, virtual seminars, and many others.

When the ISCB-SC was formulated in 2004, one of its main challenges was due to students from different geographical regions having different needs that could hardly be addressed by a single activity or event. As a consequence, the RSG-program was created in 2006 so people living in specific regions could articulate their own activities that will, in turn, enhance networking and the emergence of regional leaders that will later be potential successors to the ISCB-SC leadership.

Each RSG has a different organizational architecture depending on the local requirements, objectives and initial set up. For instance, some RSGs have been built from scratch, whereas some others have been created in collaboration with existing student organizations such as COMBINE (Australia) and SASBi (South African Student Bioinformatics Society). Irrespective of the organizational setup, there are a few requirements which have been kept mandatory for setting up an RSG at the region. The steering team requires a President and a Secretary who are primarily students, and a faculty advisor who is a member of the ISCB. Also, it is highly encouraged that this steering team has a representation from multiple universities/institutes to promote local collaborations in the field. Setting up an RSG involves taking into consideration operational and logistics aspects which can be a challenge at the start. However, this is a great learning experience for the involved young researchers as they can develop transferable soft skills such as conferences and symposia

organization, fundraising, conflict management, team building, project executions and many others. Some of the operational hurdles and related case-studies have been highlighted in our previous article³.

All 30 ISCB-SC RSGs have organized a diverse battery of events tailored to address specific needs and requirements from their target audiences and members. We will discuss some of the many successful ventures carried out by the different RSGs

Regional events are great networking platforms

Since 2005, the ISCB-SC organizes several events at distinct levels of national or international cooperation. One-day symposia, namely SCS^{4,5}, ESCS⁶, LA-SCS⁷ and SCS Africa⁸, are yearly organized as satellite events to the main ISCB official conferences: ISMB, ECCB, ISCB-LA and ISCB-Africa respectively. These symposia constitute the ISCB-SC flagship events and have itinerant locations for each edition.

Although these events constitute the perfect occasion where the general ISCB-SC leadership can get together and discuss in person the plans and balances for the current year, travel costs and accommodation make it difficult for all students to attend, venues are often located at capital cities and the official language is English which can intimidate and make it difficult for young students to participate.

To leverage these obstacles and inspired by the success of the aforementioned symposia, various RSGs have organized one-day symposia either under the aegis of a major conference or as a stand-alone event. In the next section, we revisit the highlights of some of the latest RSG organized events, such as symposia, workshops and social meetings, grouped by continent.

Latin America

Since 2016, RSG-Brazil has successfully organized three symposia editions in collaboration with the Brazilian Association of Bioinformaticians and Computational Biologists (AB3C), with an average of 70 delegates per year.

RSG-Colombia organized two regional meetings in Medellín and Bogotá in 2017. Their national meeting is held every two years during the biennial Colombian Congress in Bioinformatics (<http://ccbc.org/>) in collaboration with the Colombian Society of Bioinformatics and Computational Biology (<http://www.sc2b2.org/>). The first edition of this national meeting was held in Cali, which experienced an intense day of science and networking, with 12 student talks from 8 Colombian universities.

Similar to previous SAJIB versions, the 3rd Argentine Symposium of Young Bioinformatics Researchers (SAJIB) was held on July 28–29 2018 at Fundación Instituto Leloir (FIL), in Buenos Aires, Argentina. SAJIB⁹ was carried out for a period of two days; one day reserved for workshops and another day for the symposium. In this edition, they offered two courses: “Introduction to Python” and “Filtering, assembly, and assessment from NGS data.” Forty-five students and young researchers attended both. The second day included the symposium which gathered

34 people. Additionally, RSG Argentina has been working, joining efforts to expand the bioinformatics community over the country.

During 2017–2018, RSG-Colombia has been involved in the organization and teaching of different bioinformatics courses such as tutorial sessions on metagenomic data analysis at Universidad de Antioquia and Universidad del Valle. A similar approach has been employed by RSG-Chile for organizing tutorial workshops on topics of genomics, coarse-grained Molecular Dynamics, R programming etc. Recently initiated in 2017, RSG-Chile primarily has the presence in two campuses in Chile. They have been working on spreading it to other universities. Despite various logistic and technical hurdles, they have successfully managed two workshops in the past year and have received a good response from the audience. Technical-oriented seminars have also been organized by RSG-Brazil such as “Genomic data analysis using Python programming” and a hands-on course on bioinformatic analysis of data related to tropical diseases.

Europe

RSG-Spain has organized several Bioinformatics Student Symposium editions, either preceding the Spanish Bioinformatics Symposium or as standalone events, typically gathering over 50 attendees for selected scientific and/or career talks, talks, workshops and networking activities.

RSG-UK has organized several bioinformatics and life science student symposia since its inception along with several workshop and seminars gathering UK-wide community of bioinformatics students and scientists¹⁰.

In 2018, RSG-Turkey organized a one-day symposium as a satellite meeting to the most prominent international bioinformatics meeting in Turkey, 4th International Symposium on Health Informatics and Bioinformatics (HIBIT).

Collaborations between various RSGs have resulted in successful events such as the BeNeLux Bioinformatics Conference (which was organized jointly by RSG-Belgium, RSG-Netherlands, and RSG-Luxembourg).

Apart from formal sessions, informal setups also have proved to be good networking events well-received by youth computational biologists across Europe. In 2017, RSG-Luxembourg organized a series of science pub quiz events named ‘Sci-Pub,’ where scientists and citizens casually met to answer fun questions about science and win prizes. The invitations were extended to the broad public through Facebook adds as well as students from the University of Luxembourg from various backgrounds. After the end of each session, the assessment of knowledge was performed through written questionnaires. In total, for “Sci-Pub” events spanned the second semester of 2017 with an average attendance of 30 people per session.

Similarly, one of the recurring activities of RSG-Switzerland called “Bioinformatics in the Pub” has been very much appreciated

by the Swiss students. It is a monthly meeting that gathers bioinformaticians from different departments of the University of Lausanne (UNIL) and the polytechnic school EPFL in a friendly atmosphere. This event is planned to be extended in the future to Basel and Zurich. RSG-Switzerland is also closely related to the Swiss Institute of Bioinformatics (SIB). Encouraged by the SIB, RSG-Switzerland aims to enable students to discover other careers paths outside of academia. It is in this spirit; they were given the opportunity to organize a full session during the Basel Computational Biology conference BC2 named “Entrepreneurs’ stories: opportunities and challenges of starting your own company.”

RSG-Denmark focused on smaller events that facilitate networking between computational biologists around Copenhagen. They have organized workshop events and a regular bioinformatics coffee meetup. At CBio Coffee events students got a chance to ask questions on how to further their career for example “what elective courses should I focus on?”, “how can a foreign expat best further his/her career in Denmark?” or “How would it be to work in a specific company as a data scientist/bioinformatician?” On the other hand, their young and neighbouring community RSG-Sweden organized career events that have been much appreciated. Besides, RSG-Sweden community have organized journal several clubs and online networking hours.

Africa

RSG-South Africa organized a one-day student Symposium in association with the South African Bioinformatics (SASBi) and Genetics (SAGS) Societies¹¹ for two consecutive years.

Besides, to encourage student presenters to showcase their research work, RSGs also invited eminent speakers who are distinguished scientists in their area. RSG-Northern Africa organized a conference on “Personalized Medicine” and invited Prof. Peter Tonellato from Harvard Medical School as their keynote speaker in 2015.

RSG-South Africa organized a session of exhibitions and workshops at the National Science Festival, with prime focus at school-going scientists.

Irrespective of venue locations, events organized by RSGs have helped to facilitate interactions from students of other countries involved as well. For instance, most of the events hosted by RSG-Northern Africa 2013 generally have venues in different parts of Morocco. But, students of other neighbouring countries such as Tunisia and Mali could also participate in RSG-Northern Africa symposium as they were provided with travel fellowships to travel to the event venue in Nador, Morocco. In 2017, the 2nd RSG-Northern Africa Symposium was organized in December in Casablanca, Morocco. This event was co-founded by H3ABioNet that supported three travel fellowships for students coming from Tunisia and Mali. This was followed by a further meeting and student symposium alongside the International Society for Computational Biology (ISCB) and the African Society for Bioinformatics and Computational Biology’s (ASBCB) biennial conference in October of 2017 in Entebbe,

Uganda. Forty-four students from 9 different countries across the continent attended including participants from various levels of expertise. The symposium featured a keynote speaker, Dr Segun Fatumo, as well as student presentations and a poster session. These events are valuable for students in terms of experience as well as for networking within the continent not only the country¹².

USA

Some symposia have also extended to more than a one-day schedule and comprised various events. In December 2017, RSG-Southeastern USA organized their first research symposium for 2017–2018 in collaboration with University of South Carolina; University of South Florida, St. Petersburg; and University of Alabama, Birmingham. There were research talks from professors from across these universities, hands-on workshops on machine learning, designing pipelines for genetic analyses and three-dimensional modelling of biomolecules. Undergraduate and graduate students also had the opportunity to give talks and present posters at the symposium.

For students who search for initiatives where they can acquire hands-on experience with new concepts and techniques, stand-alone workshops typically appeal a lot. RSG-District of Columbia (USA) has been organizing summer workshops on “Bioinformatics, Genomics and Computational Biology” at the University of Maryland campus, where they have targeted to involve researchers from different programming backgrounds to benefit from the workshop (<https://iscb-dc-rsg.github.io/workshop2017/>).

Asia

For panel discussion at “Career Opportunities in Computational Biology and Bioinformatics” held during InCoB 2018 (International Conference of Bioinformatics), RSG-India invited various scientists and professionals who have considerable experience in their respective fields of academia and industry in India and abroad. The event held at Jawaharlal Nehru University, New Delhi focussed on the discussion about different career opportunities and skill sets required for a job profile in academia or industry.

Online platform for networking and knowledge transfer

The increasing number of social media platforms and usage of web resources have led to new communication channels for networking. The community of STEM researchers is expanding its presence on Twitter and other social media platforms to voice their opinions on crucial matters, share their research work, and promote science. Social media platforms allow people to get connected quickly and frequently irrespective of their locations. The majority of RSGs interact with their members via Twitter, curated mailing lists, online groups, and official Facebook pages.

Several RSGs are spreading their branches by launching online initiatives. The webinar project started a few years ago by the RSG-Turkey has reached the audience of more than 350 people (<https://www.bigmarker.com/communities/bioinfonet>) in over 30 countries and continues to grow. The RSG-Turkey team has

also initiated collaborative sessions with RSG-Colombia and RSG-Denmark. The primary goal of these webinars is to encourage researchers and mainly students to know more about computational biology in their countries as well as abroad. In collaboration with RSG-Turkey, the RSG-Colombia is inviting bioinformaticians and computational biologists working in Colombian universities as speakers. This is an essential step towards increasing the visibility of research work being carried out. Starting in 2019, RSG-Southeastern USA is also organizing online podcasts and talks to engage, foster and increase participation amongst the Bioinformatics student groups across Southeastern USA region.

Online platforms also benefit regions with limited access to resources, for instance in the case of RSG-Western Africa. H3A-BioNet, a Pan African Bioinformatics network comprising 32 Bioinformatics research groups distributed amongst 15 African countries with two partner institutions in the USA, is a major supporting network behind RSG-Western Africa. RSG Western Africa has primarily benefited from the H3ABioNet as they provide free content through webinars, funding tailored for students in resource-limited countries to attend career conferences and workshops and more recently, facilitating participation Bioinformatics development by inclusion in a variety of H3ABioNet projects.

In addition to a webinar series, online competitions have also organized programming challenges ‘CASPita’, by RSG-Italy, and ‘Research writing competition,’ by RSG-India. RSG-Italy organized a programming challenge inspired by the CASP (Critical Assessment of Structure Prediction)¹⁰ competition and hence named it ‘CASPita.’ The participants were challenged to write a parser for text output of BLAST. A few groups for all over Italy joined the competition and were evaluated by coding skill and biological accessibility. The winner was awarded a monetary prize funded by the ISCB.

In addition, a “Scientific Writing Competition” was organized by RSG-India at the end of 2018. The participants were asked to submit an essay entry in any of the three different topics provided. The topics were selected to encourage students to be creative and innovative while requiring a prerequisite knowledge of computational biology and bioinformatics and being up-to-date with the latest advancements and present status of the research area. The competition invited commendable participation from students from across the country and the best entries, graded by creativity, innovation, futuristic outlook and other aspects, were awarded.

Future directions and plans: Opportunities and obstacles

At present, the RSG program is extending further to new regions such as Lebanon, Czech Republic, Bangladesh, and other countries. Existing RSGs are attempting to expand further in their areas of operations; such as RSG-Spain, which has the intention to split into several local nodes to better cover the large region. Similar efforts have been carried out by RSG-Australia to have different divisions across the country. Many new RSGs started

in the past three years such as RSG-Costa Rica, Colombia, Chile, Greece, Bangladesh, Jordan, and Southeastern-USA.

The relatively new RSG-Sweden has also established itself as a bridging entity between the student community and industry in the field of computational biology in Sweden since their recent initiation in 2018. To further expand the community and make it inclusive, they have established the concept of branches starting with Lund and Stockholm and most recently Uppsala and Gothenburg. In the future, RSG-Sweden aims to collaborate with other RSGs across Europe to be able to share knowledge and ideas and strengthen the global community of the ISCB and its Student Council. Future plans include recruiting new members to the committee and branches and organizing seminars and hackathons.

Currently, RSG Germany is also in the process of re-establishing its connections to students by initiating monthly literature review events and in the planning process of organizing a student symposium in Heidelberg in 2019. Furthermore, the RSG plans to expand its network with German universities and arrange lunch meetings between interested participants along the lines of the Connect movement (<https://connected.mit.edu/about/connector/mit>).

RSG Chile has closely collaborated with biotechnologist leaders from all over Latin America who make up the Allbiotech Community. Allbiotech's purpose is to establish and promote a Latin American community ecosystem that includes all segments of the economy. This collaboration also resulted in the starting up of RSG-Costa Rica. Their focus is to unify the interested community in both disciplines through meetings, workshops and diffusion activities like the ones developed at the ISCB LA-SCS 2018 and Allbiotech 2018, and work together to be a part of the organization team for Allbiotech 2019 that will take place in Costa Rica. Although, new RSGs show zeal and enthusiasm to expand their ventures in the regions, they also face issues with the establishment or team transitions. For instance, to promote networking among the students and postdocs on the west coast, RSGs based in California and Nevada region of the United States was initiated in 2016. They had initial hurdles of expanding their membership, even though they tried by being one of the exhibitors at NCCB (North California Computational Biology)

symposium 2016. Later, RSG-California+Nevada was merged with the undergraduate student organization of the University of California, San Diego (UCSD) bioinformatics group for further development.

The way to success and growth when running an RSG is full of challenges. Irrespective of several operational hurdles and obstacles¹³, the RSGs have been putting efforts to expand the spirit of enthusiasm for computational biology research. In the future, the RSG program aims to expand to new regions, particularly in developing nations, promote collaborations between RSGs and also exploit virtual space via virtual seminar series programs.

Data availability

No data is associated with this article.

Grant information

The events mentioned in the article were partially supported by funds from ISCB-Student Council, a subsidiary of the International Society for Computational Biology.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to thank the ISCB and its Student Council for their continued support of the RSG program and the many volunteers all across the world for their contributions. In addition, the authors are very thankful to Dr Thomas Lengauer (President, ISCB), Dr Alfonso Valencia (Previous President, ISCB) and Diane Kovats (Executive Director, ISCB) for their continued support to the ISCB-Student Council initiatives including the RSG program. The authors also would like to thank the following content contributors: Ana Monzel, Apurva Badkas, Carlota Rubio-Perez, Dheeraj Reddy Bobbili, Federico Baldini, Işın Altınkaya, Juan Rodriguez-Rivas, Marta Coronado-Zamora, Miles McCabe, Neli Fonseca, Nikola DeLange, Qurratulain Khaleeq, Susanne Kirchen, Tariq Khaleeq, Tommaso Andreani, Anna Marcionetti, Leonardo de Oliveira Martins, Julia Kraemer, David Dylus, Sander Wuyts and Stijn Wittouck.

References

- Macintyre G, Michaut M, Abeel T: **The Regional Student Group Program of the ISCB Student Council: Stories from the Road.** *PLoS Comput Biol.* 2013; 9(9): e1003241.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yanai I, Chmielnicki E: **Computational biologists: moving to the driver's seat.** *Genome Biol.* 2017; 18: 223.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shome S, Meysman P, Parra RG, *et al.*: **ISCB-Student Council Narratives: Strategic development of the ISCB-Regional Student Groups in 2016 [version 1; peer review: not peer reviewed].** *F1000Res.* 2016; 5: pii: ISCB Comm J-2882.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hassan M, Namasivayam AA, DeBlasio D, *et al.*: **Reflections on a journey: a retrospective of the ISCB Student Council symposium series.** *BMC Bioinformatics.* 2018; 19(Suppl 12): 347.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Grynberg P, Abeel T, Lopes P, *et al.*: **Highlights from the Student Council Symposium 2011 at the International Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology.** *BMC Bioinformatics.* 2011; 12(Suppl 11): A1.
[Publisher Full Text](#)
- Francescato M, Hermans SM, Babaei S, *et al.*: **Highlights from the Third International Society for Computational Biology (ISCB) European Student**

- Council Symposium 2014.** *BMC Bioinformatics*. 2015; 16(Suppl 3): A1–A9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Parra RG, Simonetti FL, Hasenahuer MA, *et al.*: **Highlights from the 1st ISCB Latin American Student Council Symposium 2014. Introduction.** *BMC Bioinformatics*. 2015; 16(Suppl 8): A1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 8. Souilmi Y, Allali I, Badad O, *et al.*: **Highlights of the first ISCB Student Council Symposium in Africa 2015.** *F1000Res*. 2015; 4: pii: ISCB Comm J-569.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 9. Cravero F, Landaburu LU, Moreyra NN, *et al.*: **2nd Argentine Symposium of Young Bioinformatics Researchers (2SAJIB) organized by the ISCB-SC RSG-Argentina.** *PeerJ Preprints*. 2018; 6: e3504v2.
[Reference Source](#)
 10. White B, Fatima V, Fatima N, *et al.*: **Highlights of the 2nd Bioinformatics Student Symposium by ISCB RSG-UK [version 1; peer review: not peer reviewed].** *F1000Res*. 2016; 5(ISCB Comm J): 902.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 11. Rafael CN, Ambler J, Niehaus A, *et al.*: **Establishment of "The South African Bioinformatics Student Council" and Activity Highlights.** *EMBnet,journal*. 2018; 23: e903.
[Publisher Full Text](#)
 12. Rafael CN, Ashano E, Moosa Y, *et al.*: **Highlights of the second ISCB Student Council Symposium in Africa, 2017 [version 1; peer review: not peer reviewed].** *F1000Res*. 2017; 5: pii: ISCB Comm J-2183.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 13. Mishra T, Parra RG, Abeel T: **The Upside of Failure: How Regional Student Groups Learn from Their Mistakes.** *PLoS Comput Biol*. 2014; 10(8): e1003768.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research