Classificação de Sequências e Análise de Diversidade em Metagenômica

Bruno Malveira Peixoto
Zanoni Dias e Guilherme Pimentel Telles

Instituto de Computação, UNICAMP

19 de Abril de 2013

Agenda

O Que é Metagenômica

Genômica: Estudo do material genético dos organismos.

O Que é Metagenômica

Genômica: Estudo do material genético dos organismos.

- A genômica microbiana tradicional depende do isolamento e da cultura de um organismo em laboratório para então identificar o seu material genético.

O Que é Metagenômica

Genômica: Estudo do material genético dos organismos.

- A genômica microbiana tradicional depende do isolamento e da cultura de um organismo em laboratório para então identificar o seu material genético.
- No entanto, nem sempre é possível cultivar um organismo em laboratório. Espécies microbianas interagem entre si e com o ambiente, muitas vezes criando uma relação de dependência que é fatal quando desfeita pelo isolamento.

O Que é Metagenômica

Metagenômica: Estudo do material genético obtido diretamente de amostras ambientais.

O Que é Metagenômica

Metagenômica: Estudo do material genético obtido diretamente de amostras ambientais.

- A metagenômica aplica técnicas modernas da genômica para estudar comunidades microbianas diretamente de seu ambiente natural, sem a necessidade do isolamento e cultura de espécies individuais.

O Que é Metagenômica

Metagenômica: Estudo do material genético obtido diretamente de amostras ambientais.

- A metagenômica aplica técnicas modernas da genômica para estudar comunidades microbianas diretamente de seu ambiente natural, sem a necessidade do isolamento e cultura de espécies individuais.
- Estudos metagenômicos revelam que grande parte da biodiversidade microbiana não foi estudada por métodos baseados em cultura.

Motivação

- A metagenômica é uma área nova de estudo que vem crescendo bastante na comunidade científica.

Motivação

- A metagenômica é uma área nova de estudo que vem crescendo bastante na comunidade científica.
- Permite a descoberta de novos genes e espécies.

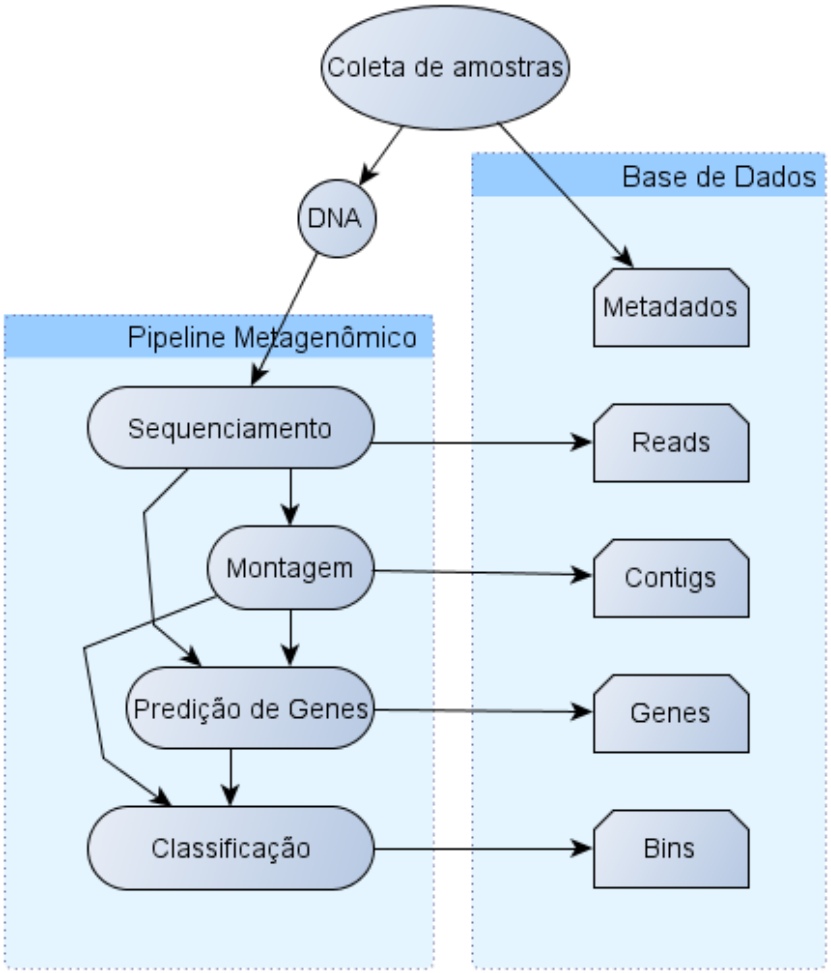
Motivação

- A metagenômica é uma área nova de estudo que vem crescendo bastante na comunidade científica.
- Permite a descoberta de novos genes e espécies.
- Compreensão das interações entre organismos e ambiente.

Motivação

- A metagenômica é uma área nova de estudo que vem crescendo bastante na comunidade científica.
- Permite a descoberta de novos genes e espécies.
- Compreensão das interações entre organismos e ambiente.
- Análise da diversidade de amostras de compostagem do Zoológico de São Paulo.

Fluxo de Trabalho de um Projeto Metagenômico



Classificação

- Classificação é a associação de sequências ou genes à sua espécie de origem.

Classificação

- Classificação é a associação de sequências ou genes à sua espécie de origem.
- Altamente desejado para uma boa interpretação do ambiente, capaz de fornecer estatísticas de composição da comunidade.

Classificação

- Classificação é a associação de sequências ou genes à sua espécie de origem.
- Altamente desejado para uma boa interpretação do ambiente, capaz de fornecer estatísticas de composição da comunidade.
- Duas estratégias mais conhecidas:
  - Classificação baseada em similaridade.
  - Classificação baseada em composição.

Classificação baseada em similaridade

- Encontrar similaridades com sequências de referência, geralmente fazendo buscas com o BLAST, e os resultados podem ser usadas para construir uma árvore.

Classificação baseada em similaridade

- Encontrar similaridades com sequências de referência, geralmente fazendo buscas com o BLAST, e os resultados podem ser usadas para construir uma árvore.
- Útil quando a maioria das sequências da amostra possuem similaridades significantes com espécies conhecidas.

Classificação baseada em similaridade

- Encontrar similaridades com sequências de referência, geralmente fazendo buscas com o BLAST, e os resultados podem ser usadas para construir uma árvore.
- Útil quando a maioria das sequências da amostra possuem similaridades significantes com espécies conhecidas.
- Sequências que não tem similaridades significantes com alguma outra de referência, são adicionados em seus próprios nós isolados na árvore.

Classificação baseada em composição

- Análises estatísticas das sequências. Buscam-se similaridades entre as próprias sequências observadas, agrupando as mais próximas entre si.

Classificação baseada em composição

- Análises estatísticas das sequências. Buscam-se similaridades entre as próprias sequências observadas, agrupando as mais próximas entre si.
- Pode-se usar métodos de aprendizado de máquina para identificar características relevantes que distinguem uma população em particular das outras.

Classificação baseada em composição

- Análises estatísticas das sequências. Buscam-se similaridades entre as próprias sequências observadas, agrupando as mais próximas entre si.
- Pode-se usar métodos de aprendizado de máquina para identificar características relevantes que distinguem uma população em particular das outras.
- Processos celulares como uso de códons e mecanismos de consertos de DNA produzem assinaturas de composição de sequências que são distintas em diferentes genomas.

Análise de Dados

- Descrição da composição da comunidade, identificando as espécies mais abundantes.

Análise de Dados

- Descrição da composição da comunidade, identificando as espécies mais abundantes.
- Atribuição de algumas atividades metabólicas para membros individuais do ecossistema.

Análise de Dados

- Descrição da composição da comunidade, identificando as espécies mais abundantes.
- Atribuição de algumas atividades metabólicas para membros individuais do ecossistema.
- Comparações de amostras diferentes do mesmo ambiente ou similar, podem revelar a influência de fatores ambientais particulares em comunidades microbianas.

Análise de Dados

- Descrição da composição da comunidade, identificando as espécies mais abundantes.
- Atribuição de algumas atividades metabólicas para membros individuais do ecossistema.
- Comparações de amostras diferentes do mesmo ambiente ou similar, podem revelar a influência de fatores ambientais particulares em comunidades microbianas.
- A comparação de metagenomas de diferentes habitats permite a descoberta de tendências gerais que ligam as características dos ambientes com as propriedades das comunidades.

Rarefação

- Com o conhecimento das espécies contidas na amostra, curvas de rarefação são feitas para estimar quão bem amostrado foi o ambiente durante o processo de amostragem.

Rarefação

- Com o conhecimento das espécies contidas na amostra, curvas de rarefação são feitas para estimar quão bem amostrado foi o ambiente durante o processo de amostragem.
- Curvas de rarefação são gráficos do número de espécies observadas em função do número de amostras coletadas. No caso, o número de amostras pode ser substituído pelo número de sequências consideradas.

Rarefação

- Com o conhecimento das espécies contidas na amostra, curvas de rarefação são feitas para estimar quão bem amostrado foi o ambiente durante o processo de amostragem.
- Curvas de rarefação são gráficos do número de espécies observadas em função do número de amostras coletadas. No caso, o número de amostras pode ser substituído pelo número de sequências consideradas.
- É natural que no início haja um crescimento rápido dessa curva, pois ao escolher uma sequência de DNA aleatória, a probabilidade dela pertencer a uma espécie diferente é alta.

Rarefação

- Com o conhecimento das espécies contidas na amostra, curvas de rarefação são feitas para estimar quão bem amostrado foi o ambiente durante o processo de amostragem.
- Curvas de rarefação são gráficos do número de espécies observadas em função do número de amostras coletadas. No caso, o número de amostras pode ser substituído pelo número de sequências consideradas.
- É natural que no início haja um crescimento rápido dessa curva, pois ao escolher uma sequência de DNA aleatória, a probabilidade dela pertencer a uma espécie diferente é alta.
- Quando o ambiente é exaustivamente amostrado, no entanto, novas espécies tem menor probabilidade de aparecer e a curva se aproxima de um platô.

Curvas de rarefação

- Verde: Maior parte das espécies foram amostradas.
- Azul: Ambiente não foi exaustivamente amostrado.
- Vermelho: Ambiente muito rico em espécies, necessita de mais amostras.

Estimadores de Diversidade

- Curvas de rarefação descrevem a riqueza das amostras coletadas.

Estimadores de Diversidade

- Curvas de rarefação descrevem a riqueza das amostras coletadas.
- Estimadores de diversidade estimam a riqueza total de uma comunidade a partir de uma amostra, ou uma coleção delas.

Estimadores de Diversidade

- Curvas de rarefação descrevem a riqueza das amostras coletadas.
- Estimadores de diversidade estimam a riqueza total de uma comunidade a partir de uma amostra, ou uma coleção delas.
- Duas abordagens para o cálculo dessa estimativa.
  - **Estimadores Paramétricos**: Modelam os dados observados em distribuições teóricas para então fazer as estimativas baseadas na modelagem.
  - **Estimadores Não-Paramétricos**: Consideram a proporção de espécies que foram observadas apenas uma vez relativa às espécies que foram observadas repetidas vezes.

Dados do Zoológico

- Os dados analisados foram amostrados de uma unidade de compostagem do Zoológico de São Paulo.

Dados do Zoológico

- Os dados analisados foram amostrados de uma unidade de compostagem do Zoológico de São Paulo.
- Compostagem é um processo que transforma restos de matéria orgânica em fertilizante para ser usado no próprio Zoológico.

Dados do Zoológico

- Os dados analisados foram amostrados de uma unidade de compostagem do Zoológico de São Paulo.
- Compostagem é um processo que transforma restos de matéria orgânica em fertilizante para ser usado no próprio Zoológico.
- As sequências de DNA foram obtidas a partir do método de pirosequenciamento 454.

Dados do Zoológico

- Os dados analisados foram amostrados de uma unidade de compostagem do Zoológico de São Paulo.
- Compostagem é um processo que transforma restos de matéria orgânica em fertilizante para ser usado no próprio Zoológico.
- As sequências de DNA foram obtidas a partir do método de pirosequenciamento 454.
- Dois conjuntos independentes de amostras foram analisados:
  - Zoo Compost 1 (ZC1) - 836 Mbp.
  - Zoo Compost 2 (ZC2) - 842 Mbp.

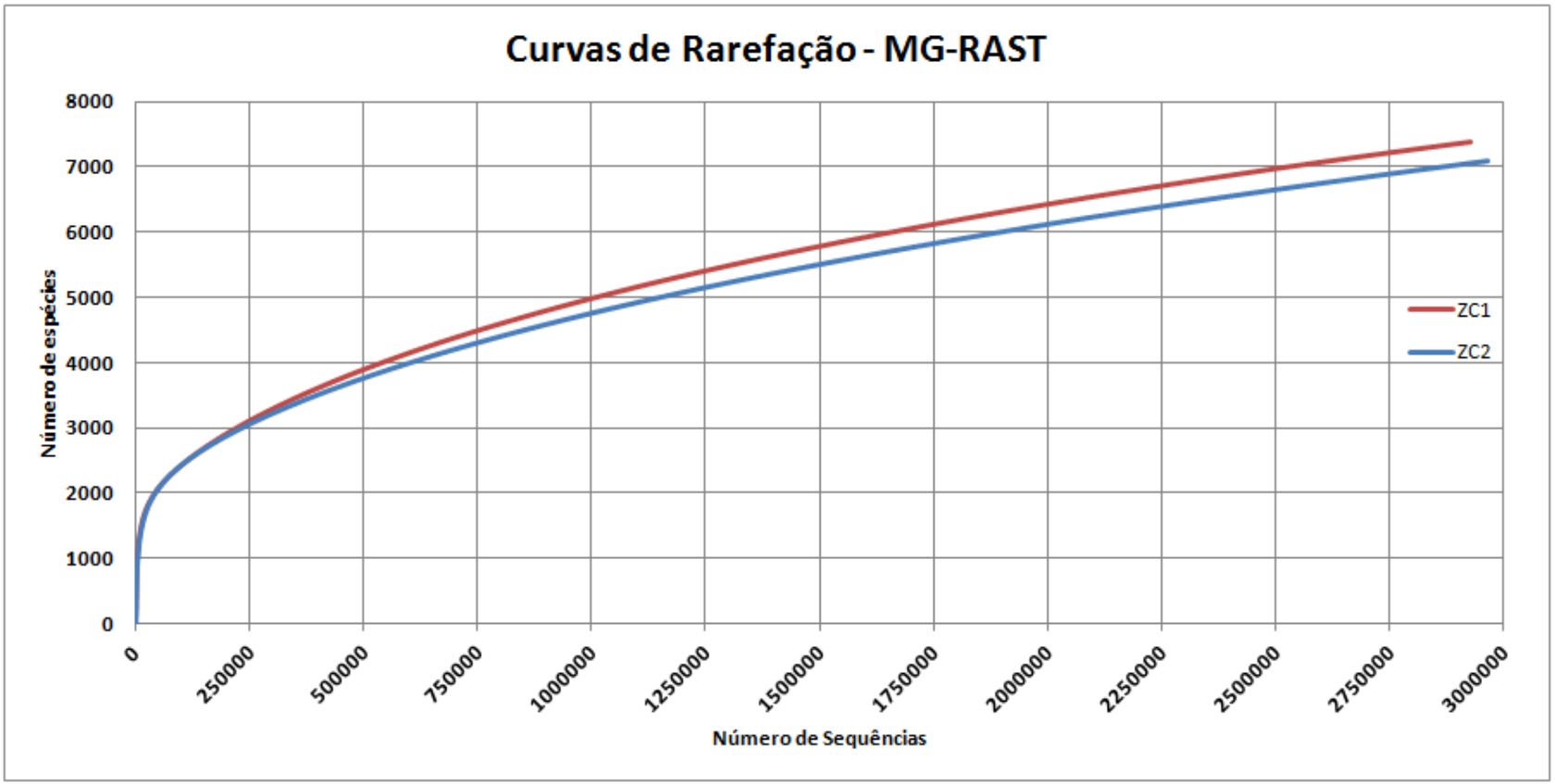
MG-RAST

- Os dados do Zoológico foram submetidos ao MG-RAST para uma análise inicial.

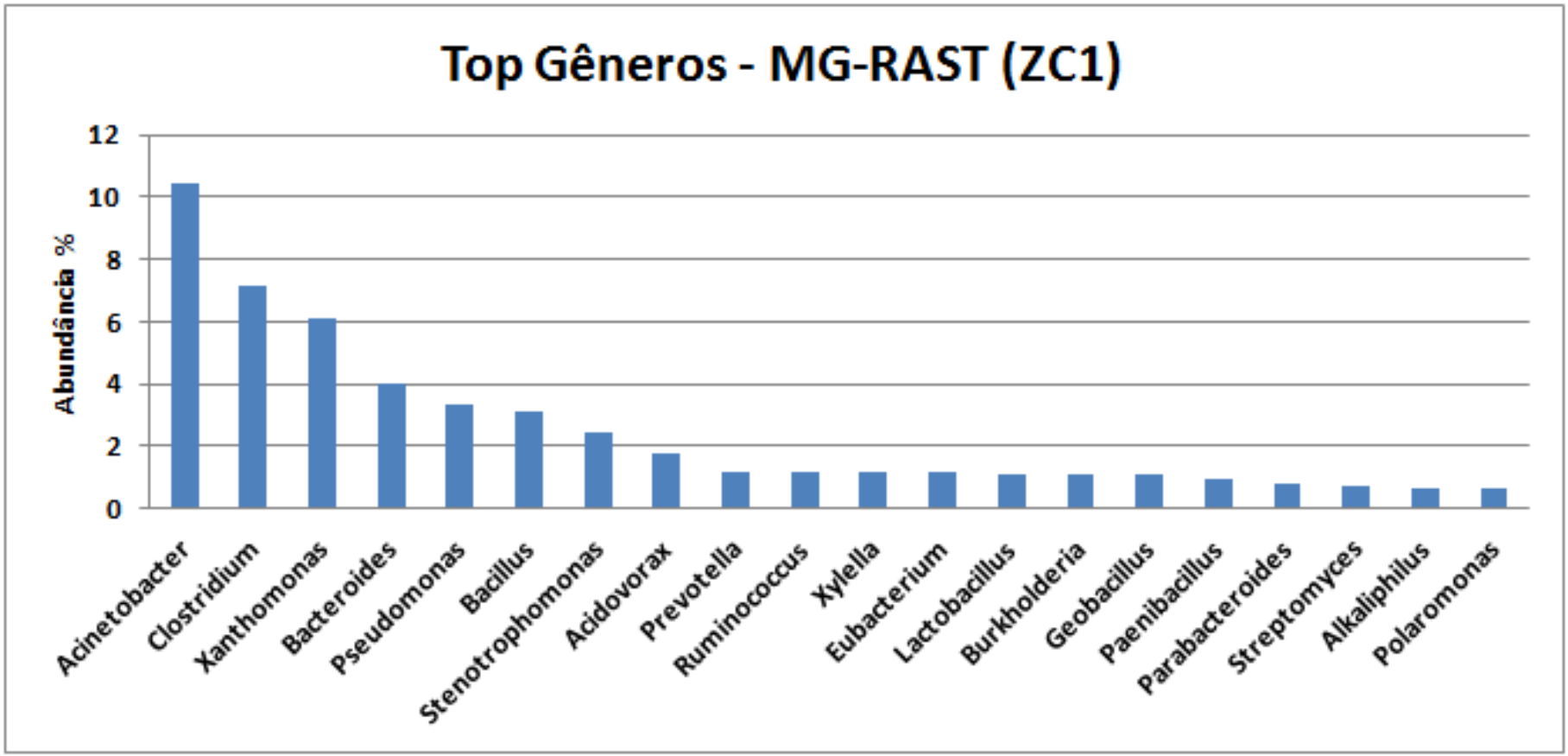
MG-RAST

- Os dados do Zoológico foram submetidos ao MG-RAST para uma análise inicial.
- O servidor metagenômico MG-RAST é um sistema desenvolvido para processar automaticamente as sequências de metagenomas.

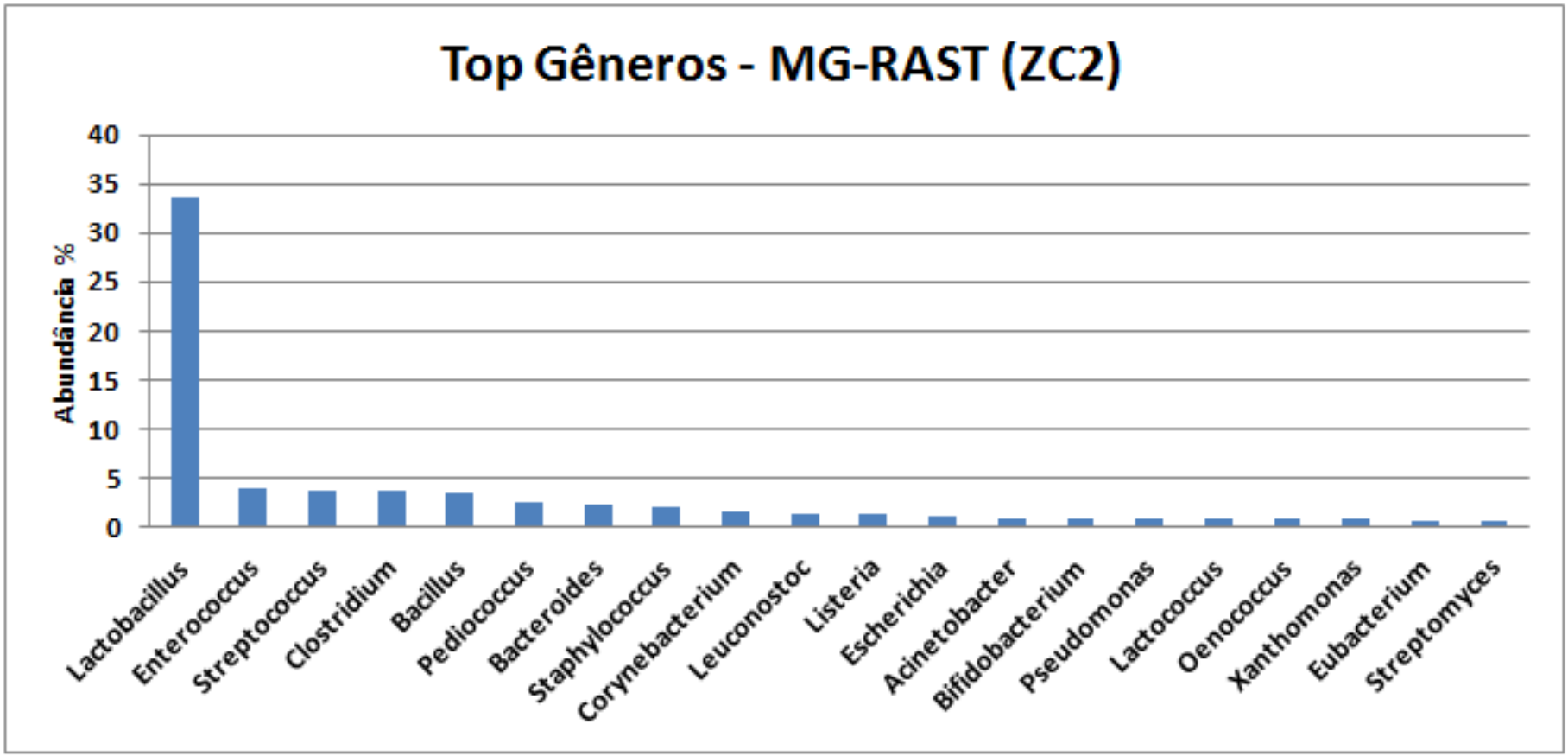
MG-RAST - Rarefação



MG-RAST - ZC1


Top Gêneros - MG-RAST (ZC1)
Abundância %
12
10
8
6
4
2
0
Acinetobacter
Clostridium
Xanthomonas
Bacteroides
Pseudomonas
Bacillus
Stenotrophomonas
Acidovorax
Prevotella
Ruminococcus
Xylella
Eubacterium
Lactobacillus
Burkholderia
Geobacillus
Paenibacillus
Parabacteroides
Streptomyces
Alkaliphilus
Polaromonas

MG-RAST - ZC2


Top Gêneros - MG-RAST (ZC2)
Abundância %
40
35
30
25
20
15
10
5
0
Lactobacillus
Enterococcus
Streptococcus
Clostridium
Bacillus
Pediococcus
Bacteroides
Staphylococcus
Corynebacterium
Leuconostoc
Listeria
Escherichia
Acinetobacter
Bifidobacterium
Pseudomonas
Lactococcus
Oenococcus
Xanthomonas
Eubacterium
Streptomyces

Naive Bayes Classifier (NBC)

- Um programa de classificação baseado no método de classificação bayesiano ingênuo.

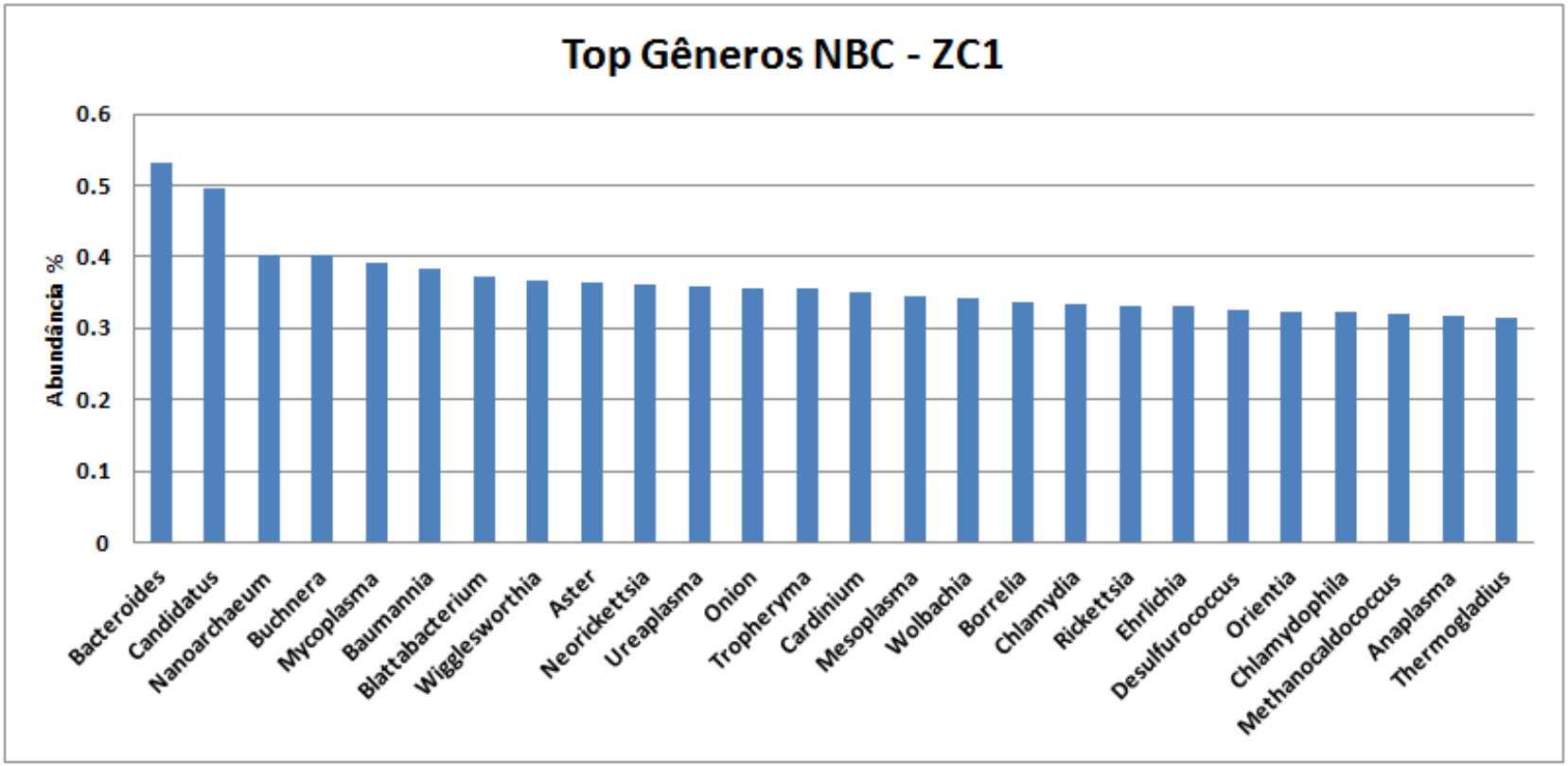
Naive Bayes Classifier (NBC)

- Um programa de classificação baseado no método de classificação bayesiano ingênuo.
- Um banco de dados de treinamento de 2262 genomas completos foi fornecido, e características sobre esses genomas foram aprendidos pelo programa.

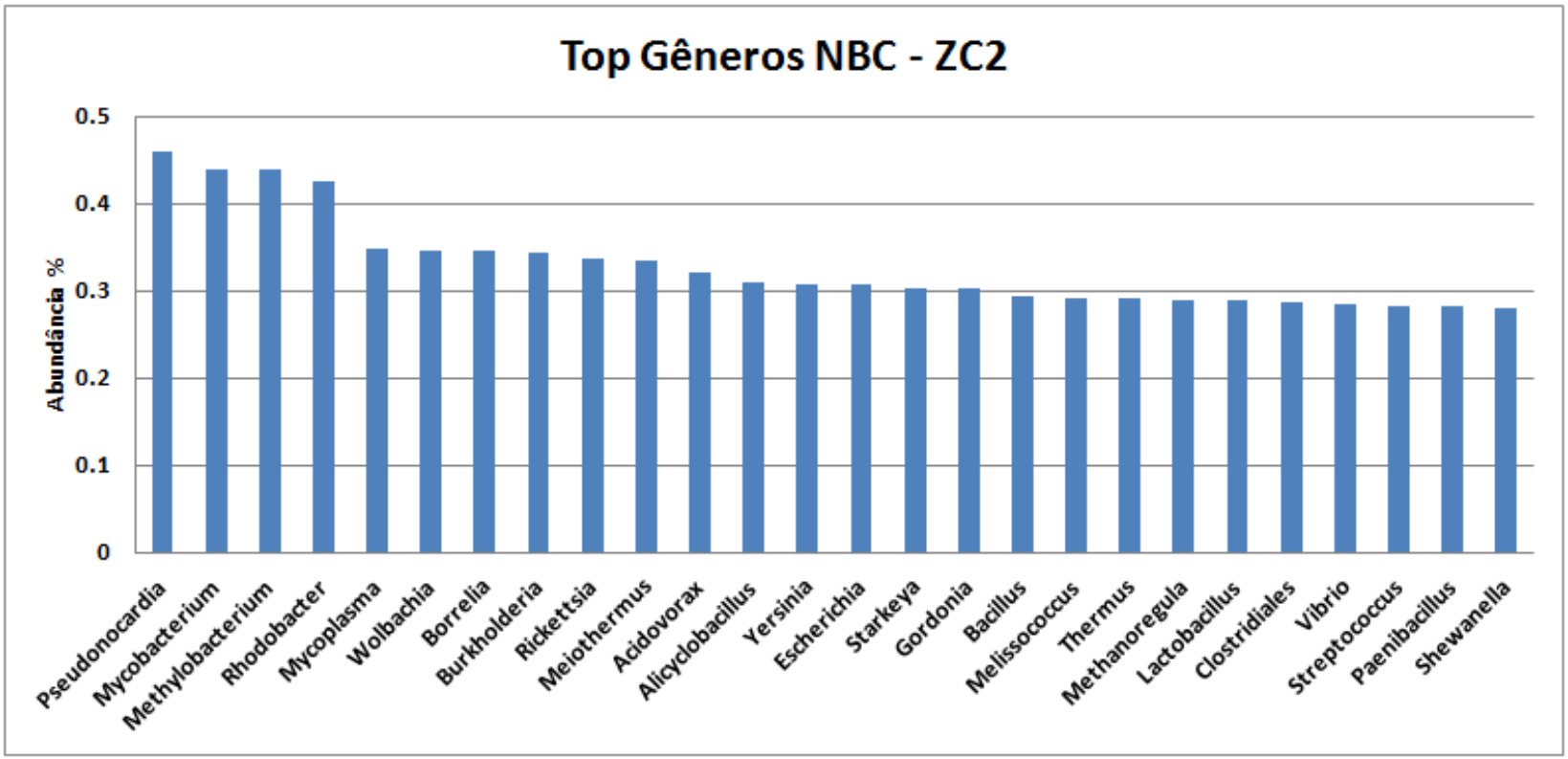
Naive Bayes Classifier (NBC)

- Um programa de classificação baseado no método de classificação bayesiano ingênuo.
- Um banco de dados de treinamento de 2262 genomas completos foi fornecido, e características sobre esses genomas foram aprendidos pelo programa.
- As sequências do Zoológico foram então submetidas ao classificador. Essas sequências não foram comparadas diretamente com as dos genomas fornecidos no treinamento, mas sim com as características extraídas deles.

Naive Bayes Classifier (NBC) - ZC1



Naive Bayes Classifier (NBC) - ZC2



Ribossomal Database Project (RDP)

- Banco de Dados de RNA ribossomal, usado para treinar o Infernal.

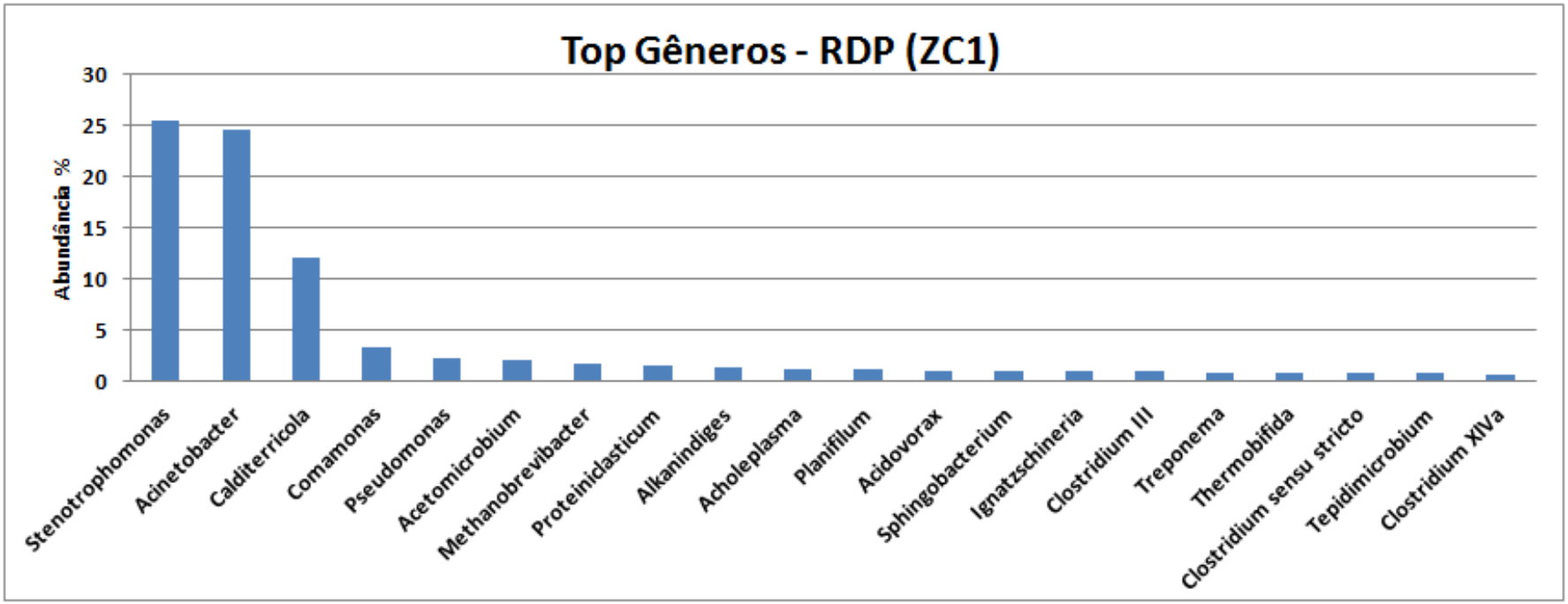
Ribossomal Database Project (RDP)

- Banco de Dados de RNA ribossomal, usado para treinar o Infernal.
- Infernal monta modelos estatísticos para representar consensos de estruturas secundárias de RNA e utiliza esses modelos para buscar RNA similares em outros bancos de dados.

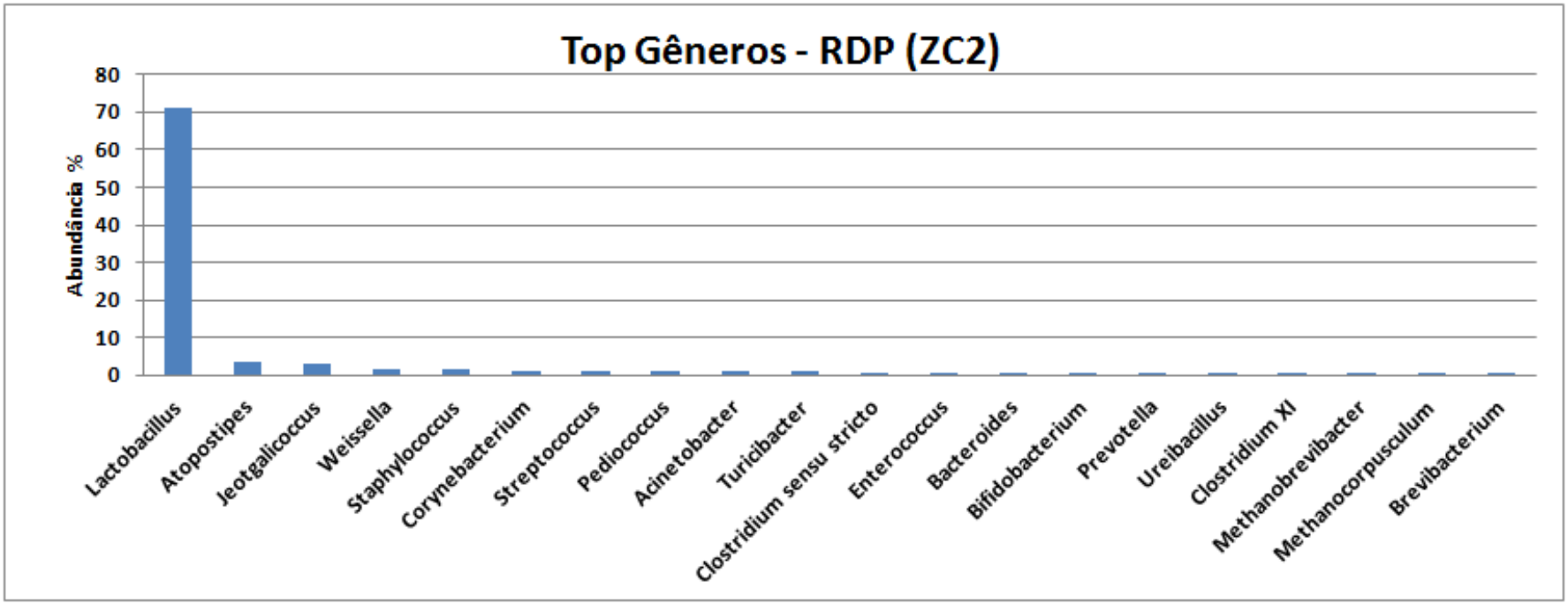
Ribossomal Database Project (RDP)

- Banco de Dados de RNA ribossomal, usado para treinar o Infernal.
- Infernal monta modelos estatísticos para representar consensos de estruturas secundárias de RNA e utiliza esses modelos para buscar RNA similares em outros bancos de dados.
- Sequências do Zoológico foram comparadas com o Infernal treinado com o banco de dados do RDP.

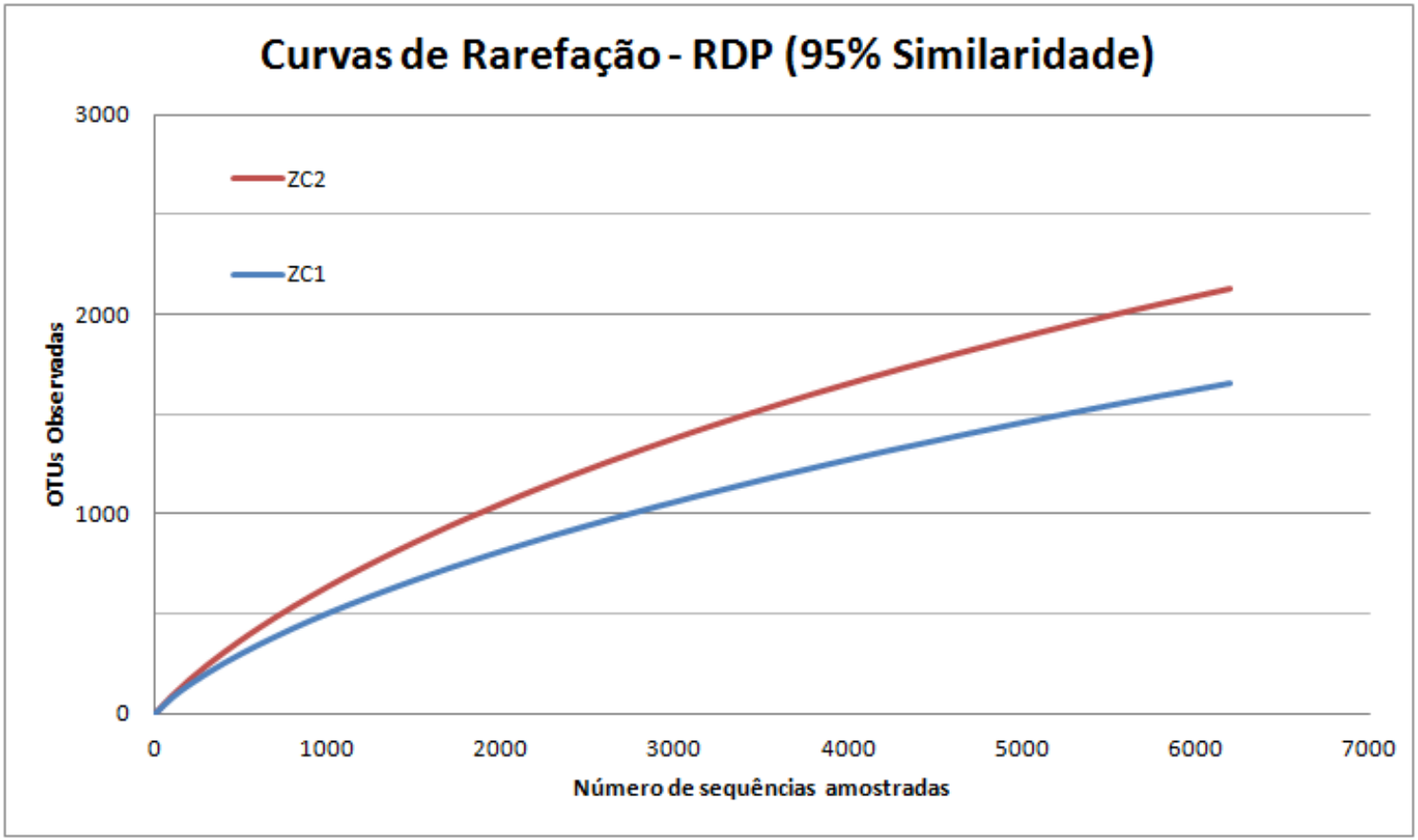
Ribossomal Database Project (RDP) - ZC1



Ribossomal Database Project (RDP) - ZC2



Ribossomal Database Project (RDP) - Rarefação



Swiss-Prot

- Banco de dados de proteínas, mais geral que o RDP.

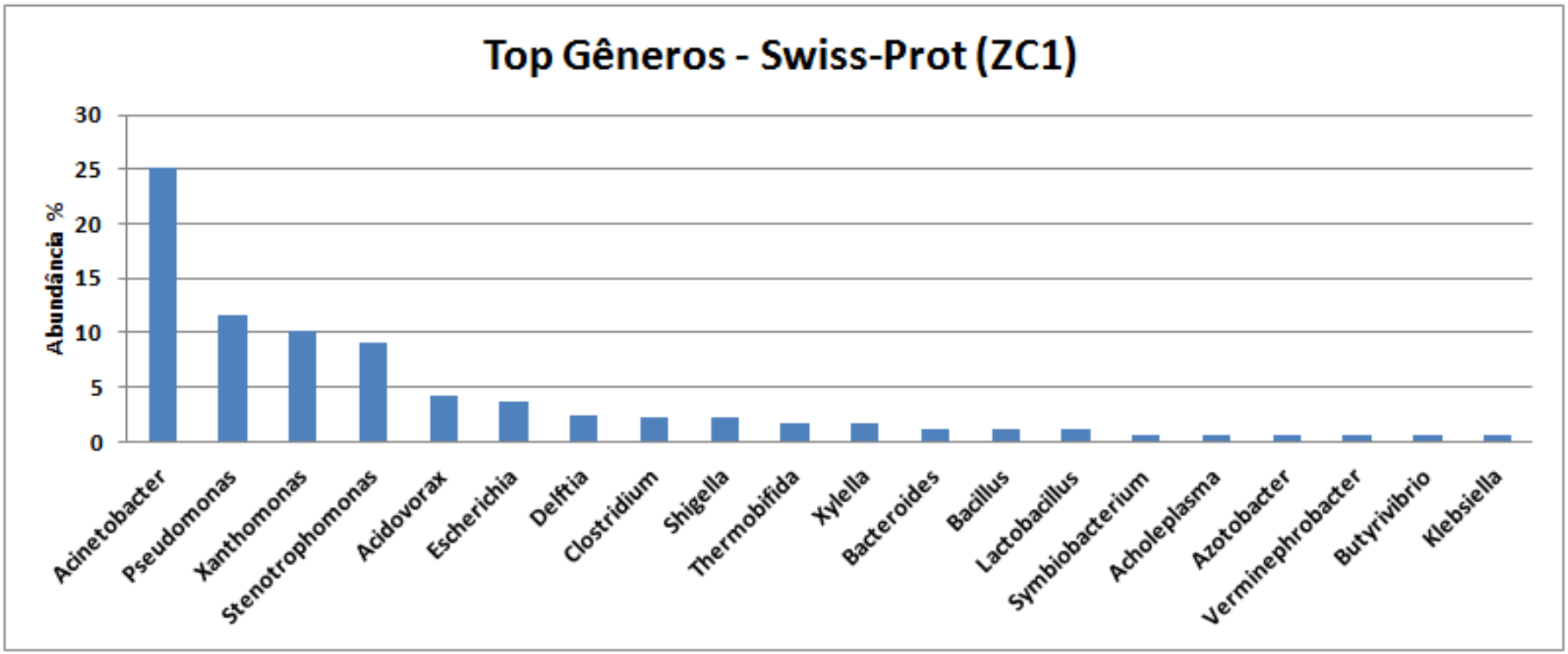
Swiss-Prot

- Banco de dados de proteínas, mais geral que o RDP.
- Busca feita através do BLAST, na versão blastx.

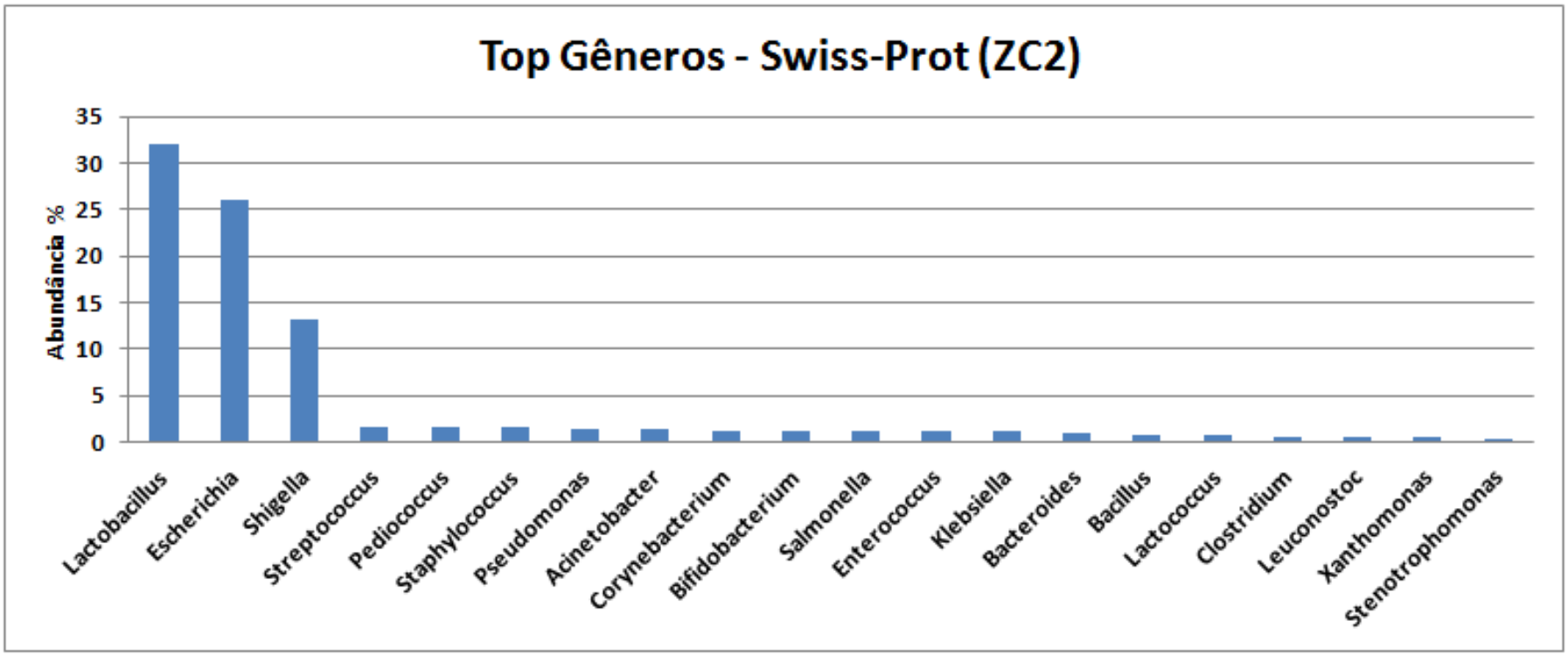
Swiss-Prot

- Banco de dados de proteínas, mais geral que o RDP.
- Busca feita através do BLAST, na versão blastx.
- Consideramos o primeiro hit do BLAST com um alinhamento de pelo menos 40 bases e 95% de identidade para determinar o gênero da sequência buscada.

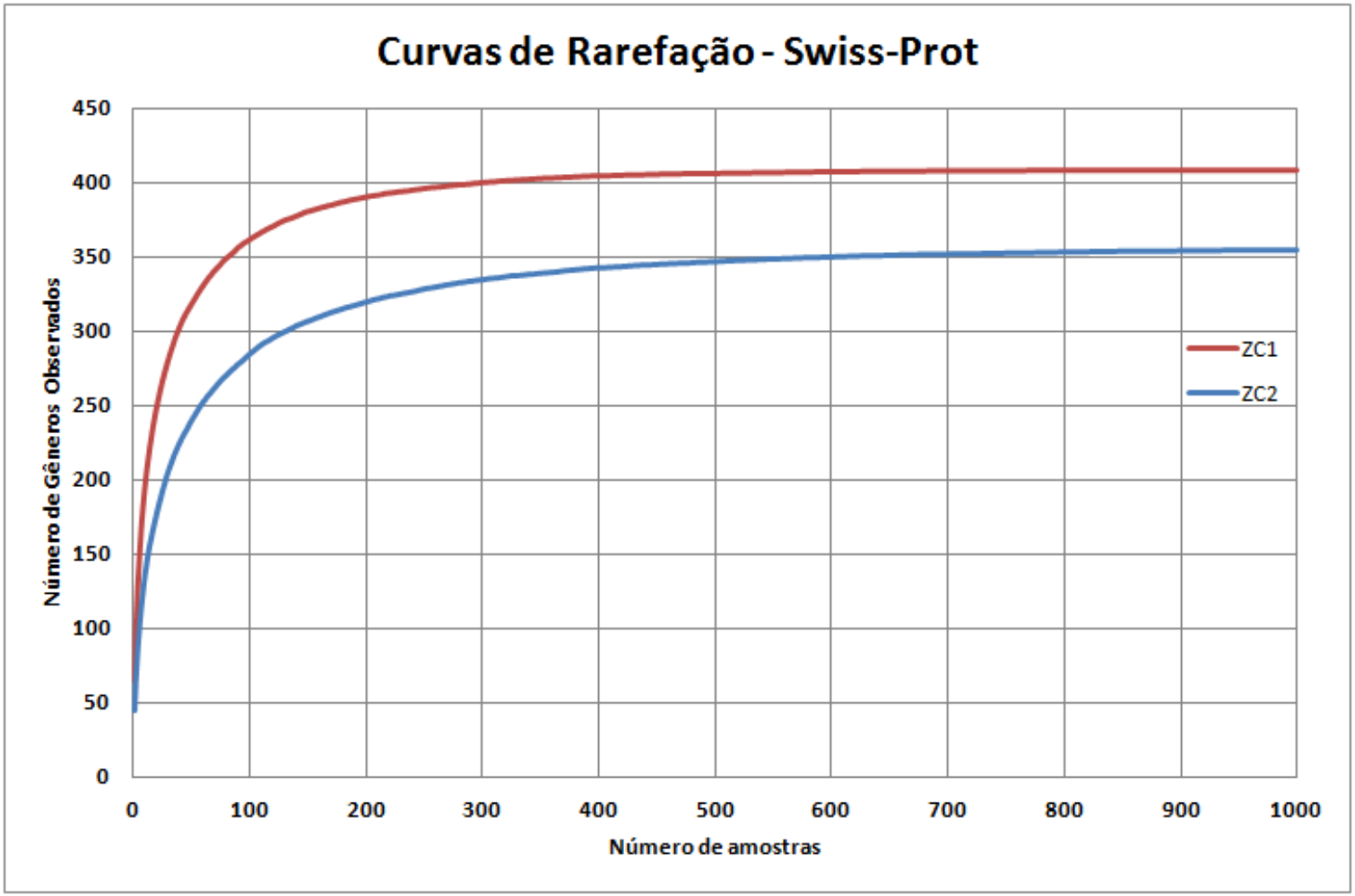
Swiss-Prot - ZC1



Swiss-Prot - ZC2



Swiss-Prot - Rarefação



NCBI-NR

- O banco de dados não-redundante de proteínas do NCBI também foi utilizado, para uma análise mais abrangente.

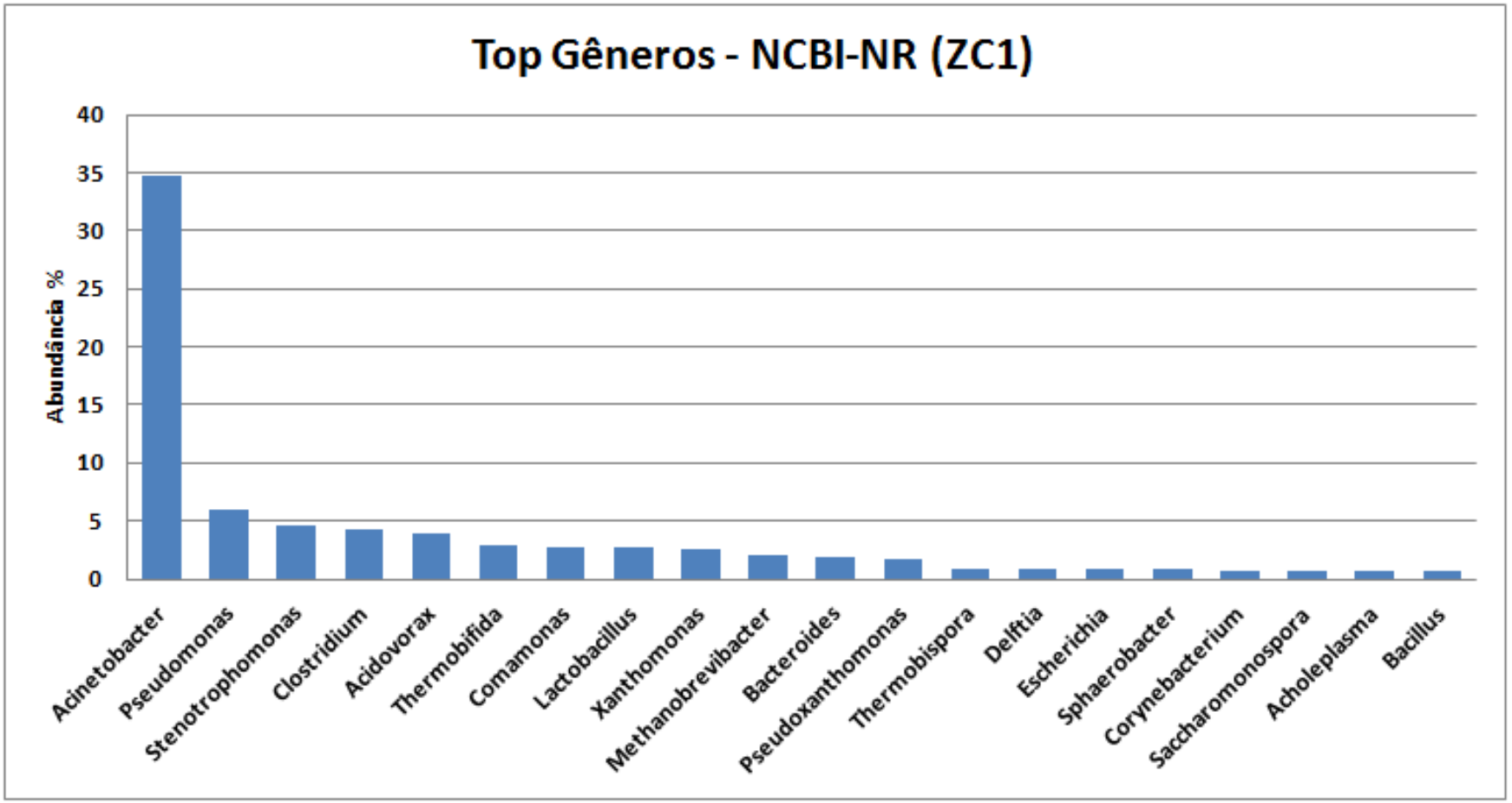
NCBI-NR

- O banco de dados não-redundante de proteínas do NCBI também foi utilizado, para uma análise mais abrangente.
- Também foram considerados os primeiros hits do BLAST, com um alinhamento de pelo menos 40 bases e 95% de identidade para determinar o gênero da sequência buscada.

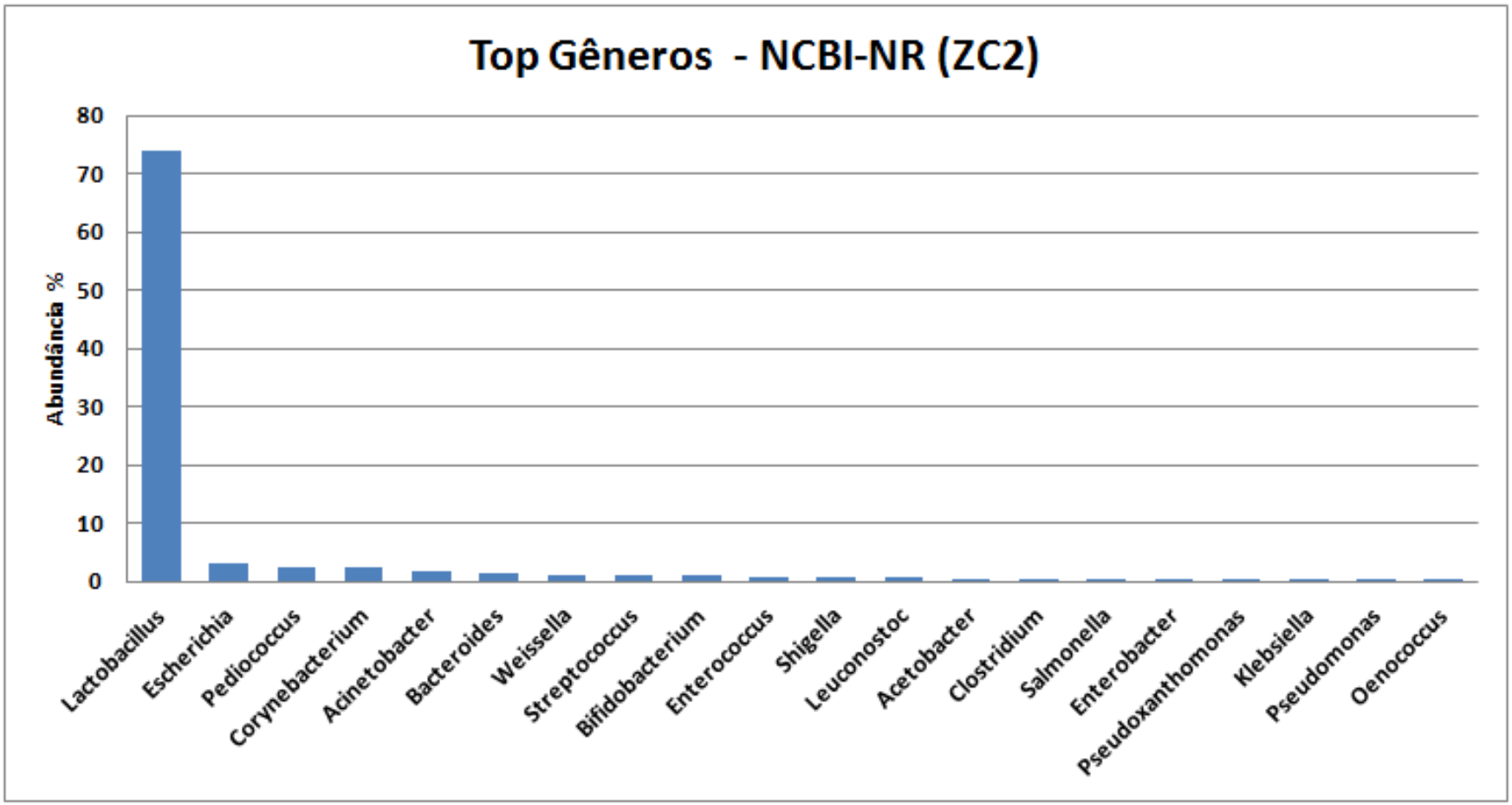
NCBI-NR

- O banco de dados não-redundante de proteínas do NCBI também foi utilizado, para uma análise mais abrangente.
- Também foram considerados os primeiros hits do BLAST, com um alinhamento de pelo menos 40 bases e 95% de identidade para determinar o gênero da sequência buscada.
- Devido ao tamanho do banco, apenas 10% das sequências do zoológico foram processadas. Foram 40 dias para esse processamento, com quatro máquinas Intel Core 2 Duo, 2.13GHz rodando em paralelo.

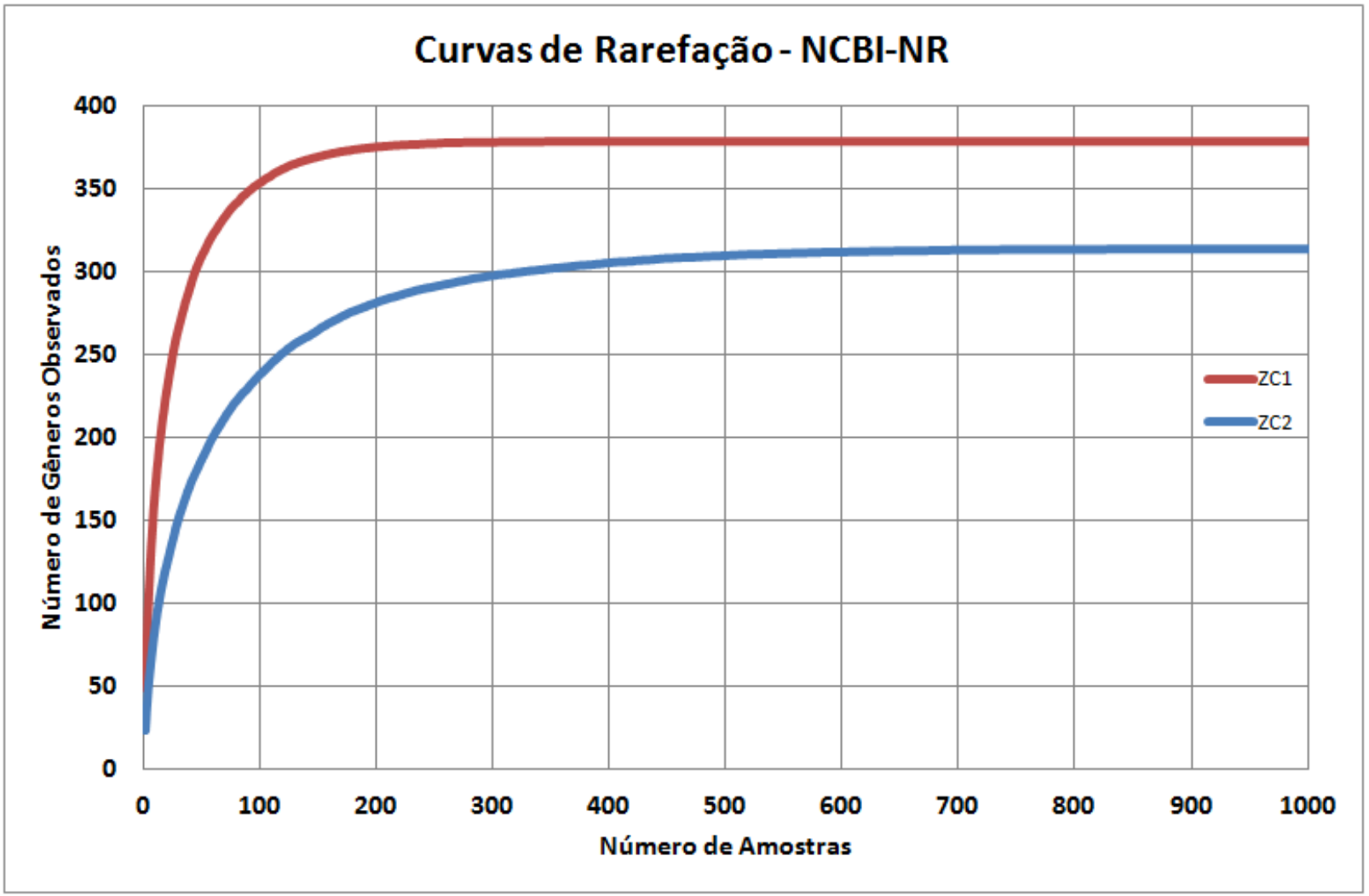
NCBI-NR - ZC1


Top Gêneros - NCBI-NR (ZC1)
Abundância %
40
35
30
25
20
15
10
5
0
Acinetobacter
Pseudomonas
Stenotrophomonas
Clostridium
Acidovorax
Thermobifida
Comamonas
Lactobacillus
Xanthomonas
Methanobrevibacter
Bacteroides
Pseudoxanthomonas
Thermobispora
Delftia
Escherichia
Sphaerobacter
Corynebacterium
Saccharomonospora
Acholeplasma
Bacillus

NCBI-NR - ZC2


Top Gêneros - NCBI-NR (ZC2)
Abundância %
80
70
60
50
40
30
20
10
0
Lactobacillus
Escherichia
Pediococcus
Corynebacterium
Acinetobacter
Bacteroides
Weissella
Streptococcus
Bifidobacterium
Enterococcus
Shigella
Leuconostoc
Acetobacter
Clostridium
Salmonella
Enterobacter
Pseudoxanthomonas
Klebsiella
Pseudomonas
Oenococcus

NCBI-NR - Rarefação



Valores dos estimadores de diversidade

- Comparando os estimadores de cada método para os resultados com 95% de similaridade

	Bancos de Dados								
	RDP			Swiss-Prot			NCBI-NR		
Amostra	OTUs	ACE	Chao1	OTUs	ACE	Chao1	OTUs	ACE	Chao1
ZC1	2129	3956	3606	409	475	477	379	508	511
ZC2	1856	5264	3351	357	473	478	314	452	457

Simulador MetaSim

- Quão confiável são os resultados com esses diversos bancos de dados?

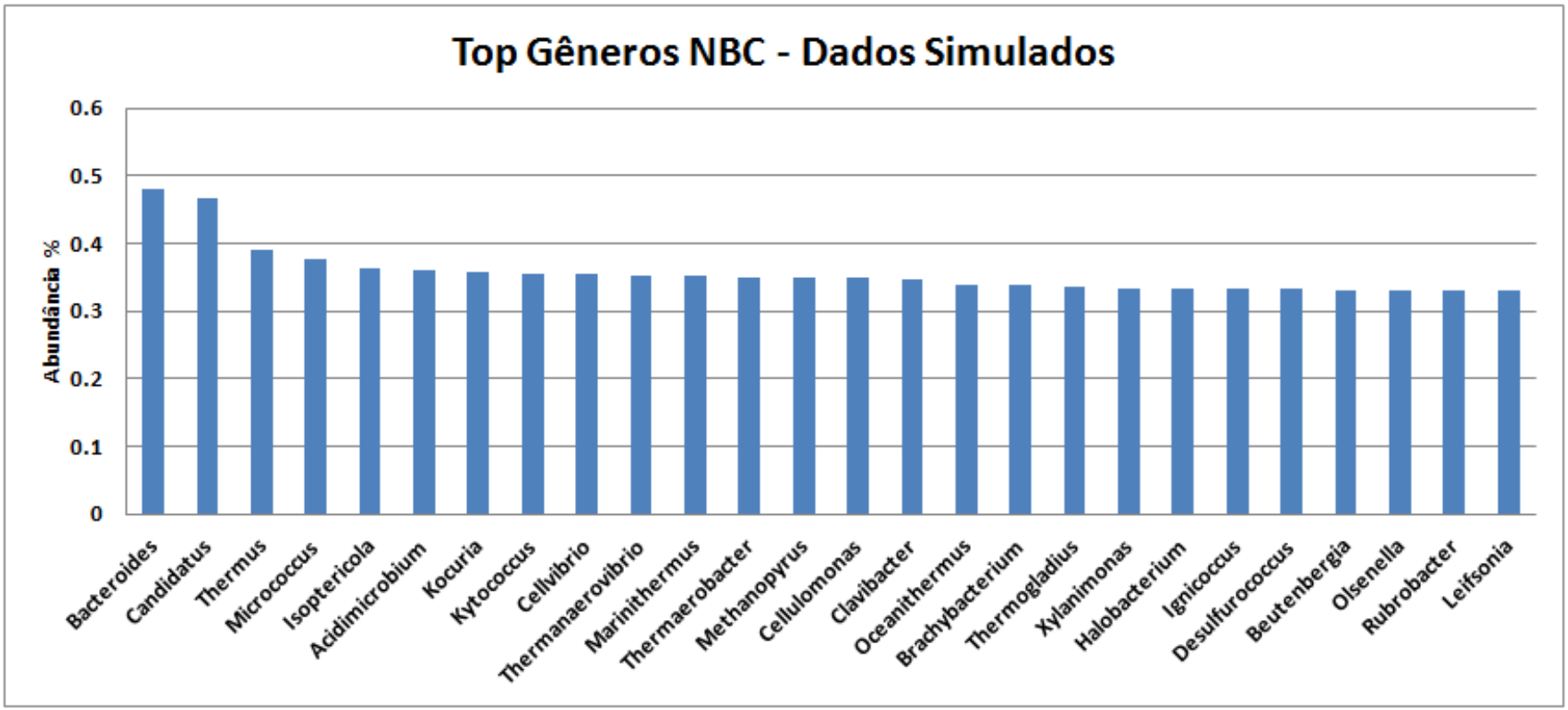
Simulador MetaSim

- Quão confiável são os resultados com esses diversos bancos de dados?
- Simulamos um metagenoma baseado no trabalho de Brulc et. al. sobre rúmen de vaca. Utilizamos os dados desse projeto disponível no MG-RAST.

Simulador MetaSim

- Quão confiável são os resultados com esses diversos bancos de dados?
- Simulamos um metagenoma baseado no trabalho de Brulc et. al. sobre rúmen de vaca. Utilizamos os dados desse projeto disponível no MG-RAST.
- Selecionamos os 50 gêneros mais abundantes e geramos um conjunto de 100 mil sequências com uma representatividade de cada gênero proporcional à abundância identificada pelo MG-RAST.

Simulação NBC



Simulação NBC

- Distribuição muito parecida com a dos dados do Zoológico.

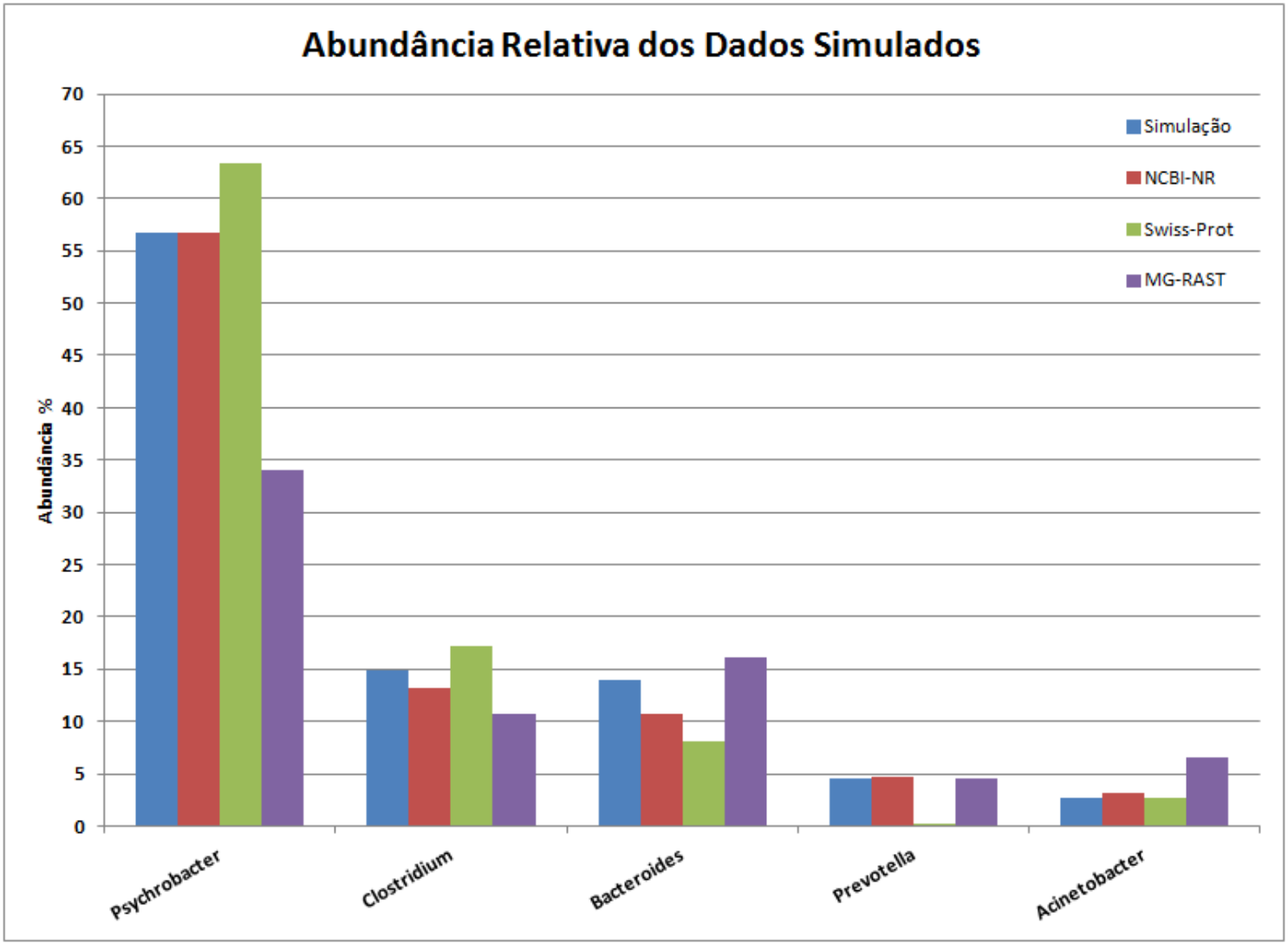
Simulação NBC

- Distribuição muito parecida com a dos dados do Zoológico.
- Gêneros identificados bem diferentes dos simulados.

Simulação NBC

- Distribuição muito parecida com a dos dados do Zoológico.
- Gêneros identificados bem diferentes dos simulados.
- O NBC não forneceu uma análise confiável para os nossos dados.

Simulação


Abundância Relativa dos Dados Simulados
Abundância %
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Simulação
NCBI-NR
Swiss-Prot
MG-RAST
Psychrobacter
Clostridium
Bacteroides
Prevotella
Acinetobacter

Simulação

- Porcentagens de abundância identificadas pelo banco NCBI-NR foram muito próximas das reais. Os 13 gêneros mais abundantes na simulação são exatamentes os 13 mais abundantes identificados pelo NCBI-NR. Sendo que os 8 primeiros também estão na mesma ordem de abundância.

Simulação

- Porcentagens de abundância identificadas pelo banco NCBI-NR foram muito próximas das reais. Os 13 gêneros mais abundantes na simulação são exatamentes os 13 mais abundantes identificados pelo NCBI-NR. Sendo que os 8 primeiros também estão na mesma ordem de abundância.
- O MG-RAST apresentou uma distribuição distante da real, além de ter identificado um número maior de sequências do que realmente existia em determinados gêneros, apesar de ser o único banco que identificou todos os gêneros utilizados na simulação.

Simulação

- Porcentagens de abundância identificadas pelo banco NCBI-NR foram muito próximas das reais. Os 13 gêneros mais abundantes na simulação são exatamentes os 13 mais abundantes identificados pelo NCBI-NR. Sendo que os 8 primeiros também estão na mesma ordem de abundância.
- O MG-RAST apresentou uma distribuição distante da real, além de ter identificado um número maior de sequências do que realmente existia em determinados gêneros, apesar de ser o único banco que identificou todos os gêneros utilizados na simulação.
- O NCBI-NR identificou 96% dos gêneros simulados e o Swiss-Prot 62%.

Comparação dos Dados Simulados

Reads Simulados		Abundância NCBI-NR		Abundância Swiss-Prot		Abundância MG-RAST	
Gênero	(%)	Gênero	(%)	Gênero	(%)	Gênero	(%)
Psychrobacter	56.738	Psychrobacter	56.753	Psychrobacter	63.379	Psychrobacter	34.033
Clostridium	14.860	Clostridium	13.242	Clostridium	17.166	Bacteroides	16.130
Bacteroides	13.943	Bacteroides	10.670	Bacteroides	8.112	Clostridium	10.734
Prevotella	4.530	Prevotella	4.659	Acinetobacter	2.726	Acinetobacter	6.601
Acinetobacter	2.686	Acinetobacter	3.134	Bacillus	1.885	Prevotella	4.603
Ruminococcus	2.152	Ruminococcus	2.017	Pseudomonas	1.380	Enhydrobacter	1.995
Eubacterium	1.259	Eubacterium	1.317	Eubacterium	0.505	Ruminococcus	1.891
Butyrivibrio	0.947	Butyrivibrio	0.922	Escherichia	0.370	Moraxella	1.825
Bacillus	0.688	Pseudomonas	0.499	Parabacteroides	0.370	Pseudomonas	1.442
Pseudomonas	0.367	Bacillus	0.388	Vibrio	0.303	Bacillus	1.139
Parabacteroides	0.232	Moraxella	0.263	Shigella	0.303	Eubacterium	0.863
Moraxella	0.192	Streptococcus	0.215	Streptococcus	0.269	Shewanella	0.581
Streptococcus	0.145	Parabacteroides	0.146	Shewanella	0.236	Parabacteroides	0.540

Comparação dos Dados Simulados

Reads Simulados	NCBI-NR	Swiss-Prot	MG-RAST
Psychrobacter	Psychrobacter	Psychrobacter	Psychrobacter
Clostridium	Clostridium	Clostridium	Bacteroides
Bacteroides	Bacteroides	Bacteroides	Clostridium
Prevotella	Prevotella	Acinetobacter	Acinetobacter
Acinetobacter	Acinetobacter	Bacillus	Prevotella
Ruminococcus	Ruminococcus	Pseudomonas	Enhydrobacter
Eubacterium	Eubacterium	Eubacterium	Ruminococcus
Butyrivibrio	Butyrivibrio	Escherichia	Moraxella
Bacillus	Pseudomonas	Parabacteroides	Pseudomonas
Pseudomonas	Bacillus	Vibrio	Bacillus
Parabacteroides	Moraxella	Shigella	Eubacterium
Moraxella	Streptococcus	Streptococcus	Shewanella
Streptococcus	Parabacteroides	Shewanella	Parabacteroides

Comparação dos Dados Simulados

Reads Simulados	NCBI-NR	Swiss-Prot	MG-RAST
Psychrobacter	Psychrobacter	Psychrobacter	Psychrobacter
Clostridium	Clostridium	Clostridium	Bacteroides
Bacteroides	Bacteroides	Bacteroides	Clostridium
Prevotella	Prevotella	Acinetobacter	Acinetobacter
Acinetobacter	Acinetobacter	Bacillus	Prevotella
Ruminococcus	Ruminococcus	Pseudomonas	Enhydrobacter
Eubacterium	Eubacterium	Eubacterium	Ruminococcus
Butyrivibrio	Butyrivibrio	Escherichia	Moraxella
Bacillus	Pseudomonas	Parabacteroides	Pseudomonas
Pseudomonas	Bacillus	Vibrio	Bacillus
Parabacteroides	Moraxella	Shigella	Eubacterium
Moraxella	Streptococcus	Streptococcus	Shewanella
Streptococcus	Parabacteroides	Shewanella	Parabacteroides

Comparação dos Dados Simulados

Reads Simulados	NCBI-NR	Swiss-Prot	MG-RAST
Psychrobacter	Psychrobacter	Psychrobacter	Psychrobacter
Clostridium	Clostridium	Clostridium	Bacteroides
Bacteroides	Bacteroides	Bacteroides	Clostridium
Prevotella	Prevotella	Acinetobacter	Acinetobacter
Acinetobacter	Acinetobacter	Bacillus	Prevotella
Ruminococcus	Ruminococcus	Pseudomonas	Enhydrobacter
Eubacterium	Eubacterium	Eubacterium	Ruminococcus
Butyrivibrio	Butyrivibrio	Escherichia	Moraxella
Bacillus	Pseudomonas	Parabacteroides	Pseudomonas
Pseudomonas	Bacillus	Vibrio	Bacillus
Parabacteroides	Moraxella	Shigella	Eubacterium
Moraxella	Streptococcus	Streptococcus	Shewanella
Streptococcus	Parabacteroides	Shewanella	Parabacteroides

Comparação dos Dados Simulados

Reads Simulados	NCBI-NR	Swiss-Prot	MG-RAST
Psychrobacter	Psychrobacter	Psychrobacter	Psychrobacter
Clostridium	Clostridium	Clostridium	Bacteroides
Bacteroides	Bacteroides	Bacteroides	Clostridium
Prevotella	Prevotella	Acinetobacter	Acinetobacter
Acinetobacter	Acinetobacter	Bacillus	Prevotella
Ruminococcus	Ruminococcus	Pseudomonas	Enhydrobacter
Eubacterium	Eubacterium	Eubacterium	Ruminococcus
Butyrivibrio	Butyrivibrio	Escherichia	Moraxella
Bacillus	Pseudomonas	Parabacteroides	Pseudomonas
Pseudomonas	Bacillus	Vibrio	Bacillus
Parabacteroides	Moraxella	Shigella	Eubacterium
Moraxella	Streptococcus	Streptococcus	Shewanella
Streptococcus	Parabacteroides	Shewanella	Parabacteroides

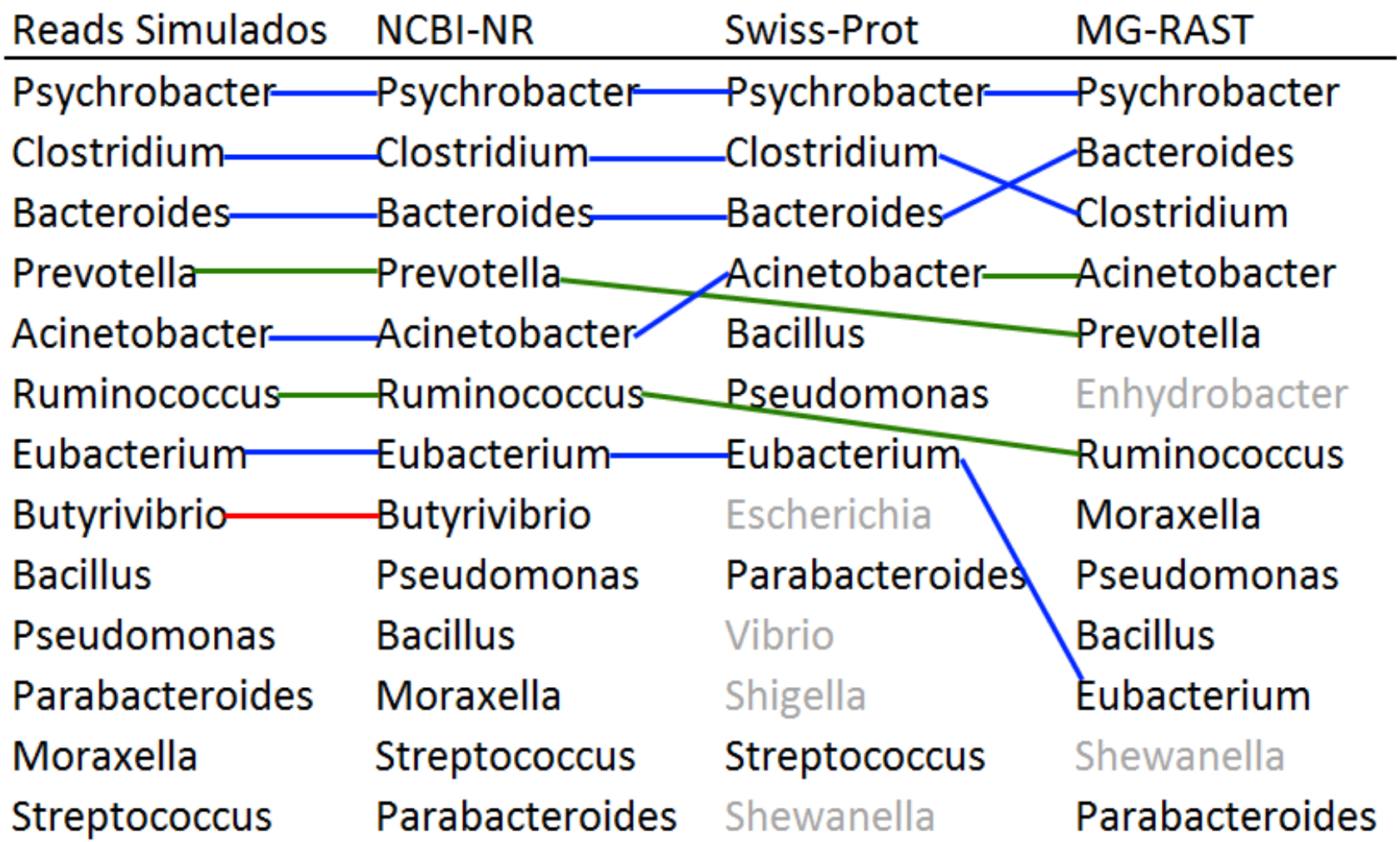
Comparação dos Dados Simulados

Reads Simulados	NCBI-NR	Swiss-Prot	MG-RAST
Psychrobacter	Psychrobacter	Psychrobacter	Psychrobacter
Clostridium	Clostridium	Clostridium	Bacteroides
Bacteroides	Bacteroides	Bacteroides	Clostridium
Prevotella	Prevotella	Acinetobacter	Acinetobacter
Acinetobacter	Acinetobacter	Bacillus	Prevotella
Ruminococcus	Ruminococcus	Pseudomonas	Enhydrobacter
Eubacterium	Eubacterium	Eubacterium	Ruminococcus
Butyrivibrio	Butyrivibrio	Escherichia	Moraxella
Bacillus	Pseudomonas	Parabacteroides	Pseudomonas
Pseudomonas	Bacillus	Vibrio	Bacillus
Parabacteroides	Moraxella	Shigella	Eubacterium
Moraxella	Streptococcus	Streptococcus	Shewanella
Streptococcus	Parabacteroides	Shewanella	Parabacteroides

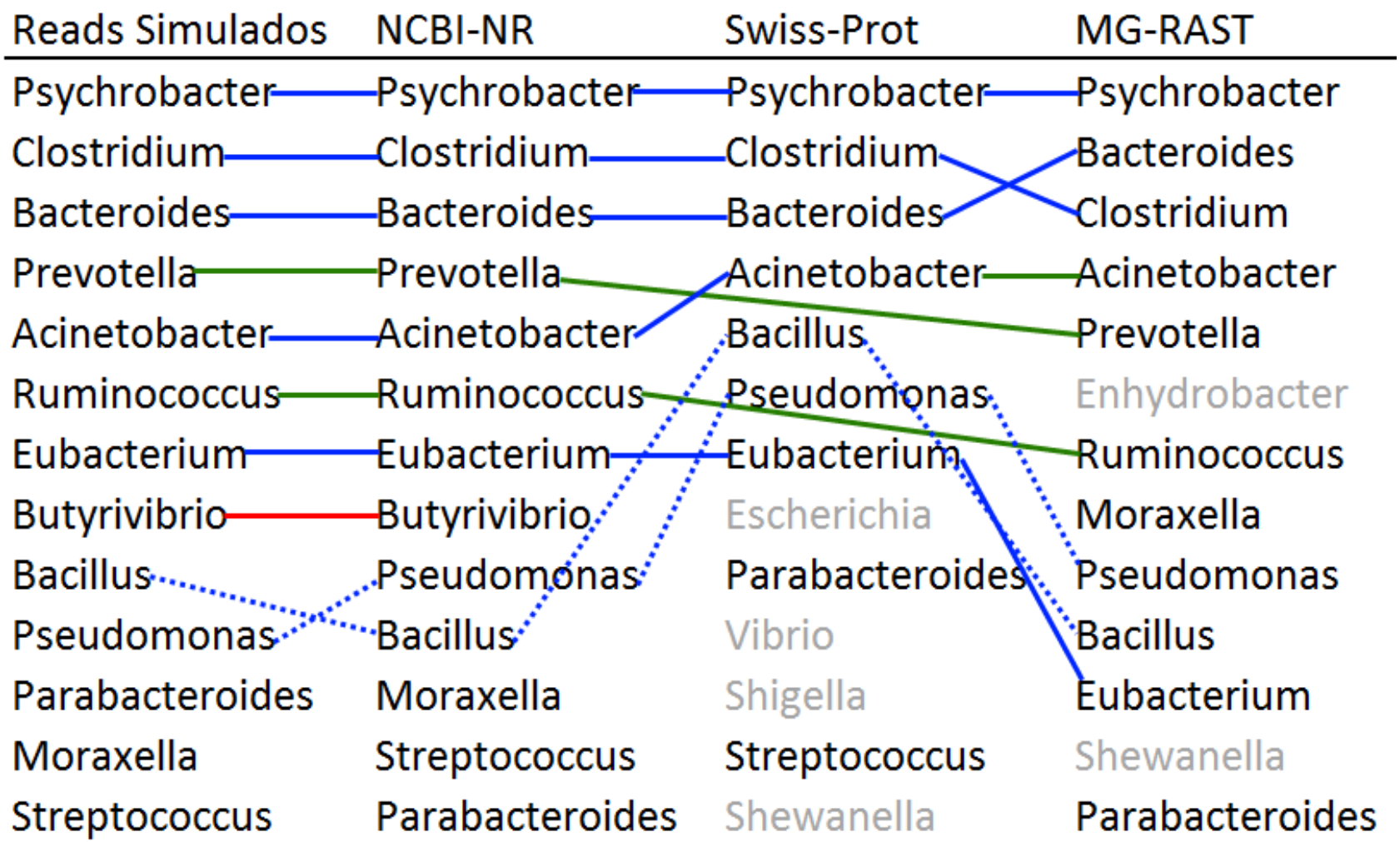
Comparação dos Dados Simulados

Reads Simulados	NCBI-NR	Swiss-Prot	MG-RAST
Psychrobacter	Psychrobacter	Psychrobacter	Psychrobacter
Clostridium	Clostridium	Clostridium	Bacteroides
Bacteroides	Bacteroides	Bacteroides	Clostridium
Prevotella	Prevotella	Acinetobacter	Acinetobacter
Acinetobacter	Acinetobacter	Bacillus	Prevotella
Ruminococcus	Ruminococcus	Pseudomonas	Enhydrobacter
Eubacterium	Eubacterium	Eubacterium	Ruminococcus
Butyrivibrio	Butyrivibrio	Escherichia	Moraxella
Bacillus	Pseudomonas	Parabacteroides	Pseudomonas
Pseudomonas	Bacillus	Vibrio	Bacillus
Parabacteroides	Moraxella	Shigella	Eubacterium
Moraxella	Streptococcus	Streptococcus	Shewanella
Streptococcus	Parabacteroides	Shewanella	Parabacteroides

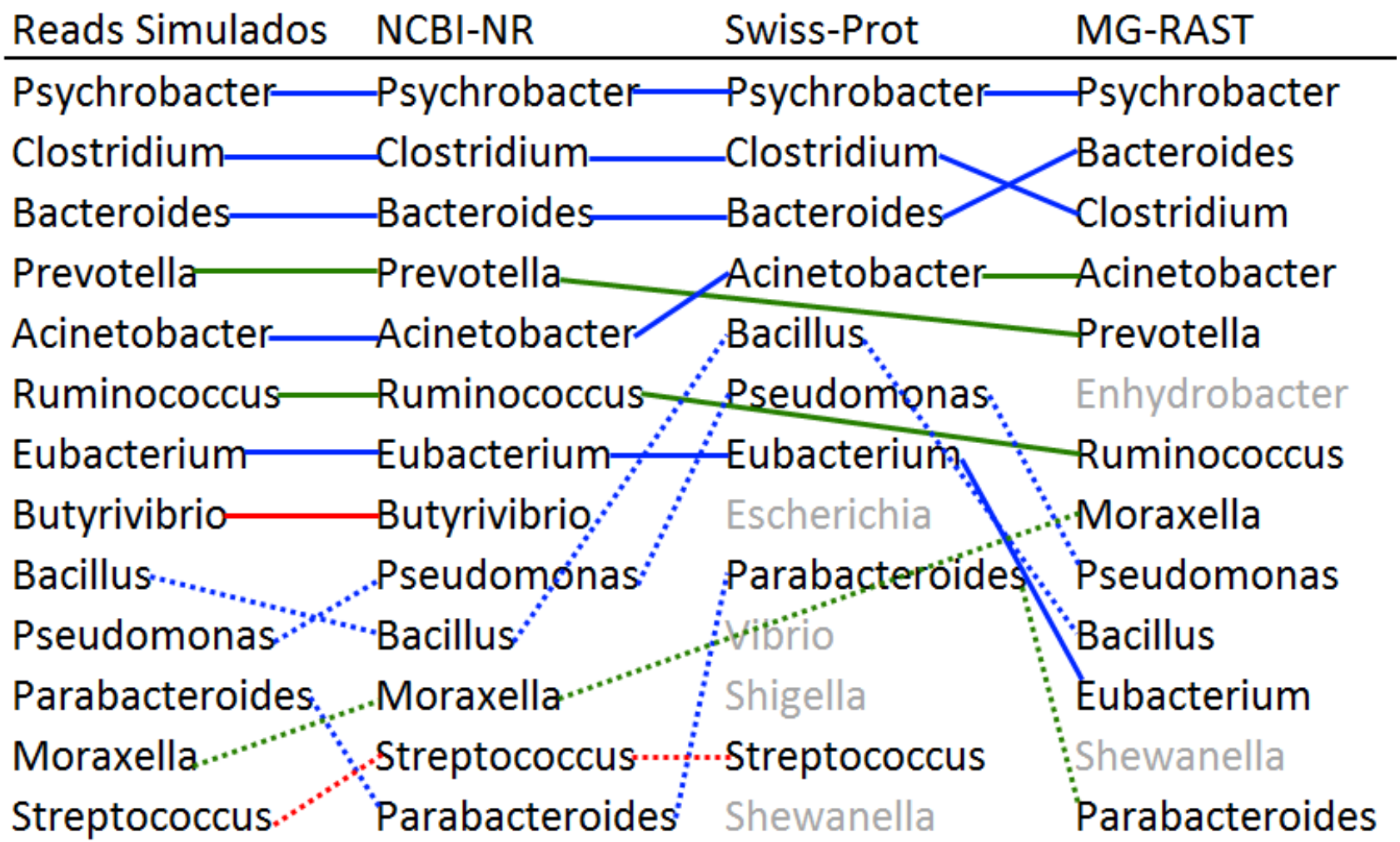
Comparação dos Dados Simulados



Comparação dos Dados Simulados



Comparação dos Dados Simulados



Comparação dos Dados do Zoológico

- Apesar do MG-RAST ter sido o único que identificou todos os gêneros existentes na simulação, o NCBI-NR apresentou uma maior semelhança na distribuição das abundâncias.

Comparação dos Dados do Zoológico

- Apesar do MG-RAST ter sido o único que identificou todos os gêneros existentes na simulação, o NCBI-NR apresentou uma maior semelhança na distribuição das abundâncias.
- Diante disso, podemos voltar aos dados do Zoológico e comparar os gêneros mais abundantes identificados por cada método, tomando por base o resultado do NCBI-NR.

Comparação dos Dados do Zoológico - ZC1

NCBI-NR	MGRAST	Swiss-Prot	RDP
Acinetobacter	Acinetobacter	Acinetobacter	Stenotrophomonas
Pseudomonas	Clostridium	Pseudomonas	Acinetobacter
Stenotrophomonas	Xanthomonas	Xanthomonas	Calditerricola
Clostridium	Bacteroides	Stenotrophomonas	Comamonas
Acidovorax	Pseudomonas	Acidovorax	Pseudomonas
Thermobifida	Bacillus	Escherichia	Acetomicrobium
Comamonas	Stenotrophomonas	Delftia	Methanobrevibacter
Lactobacillus	Acidovorax	Clostridium	Proteiniclasticum
Xanthomonas	Prevotella	Shigella	Alkanindiges
Methanobrevibacter	Ruminococcus	Thermobifida	Acholeplasma
Bacteroides	Xylella	Xylella	Planifilum
Pseudoxanthomonas	Eubacterium	Bacteroides	Acidovorax
Thermobispora	Lactobacillus	Bacillus	Sphingobacterium

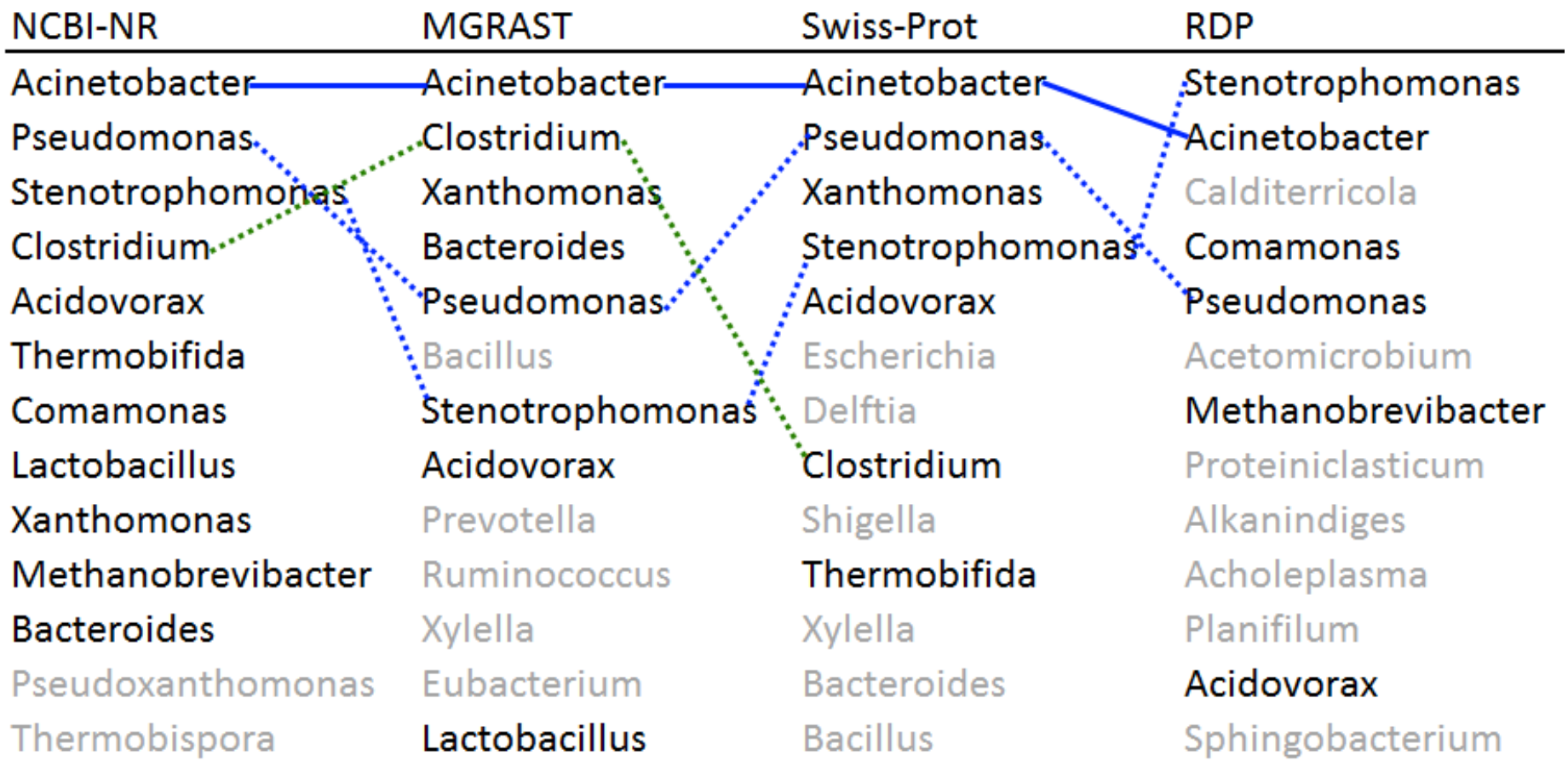
Comparação dos Dados do Zoológico - ZC1

NCBI-NR	MGRAST	Swiss-Prot	RDP
Acinetobacter	Acinetobacter	Acinetobacter	Stenotrophomonas
Pseudomonas	Clostridium	Pseudomonas	Acinetobacter
Stenotrophomonas	Xanthomonas	Xanthomonas	Calditerricola
Clostridium	Bacteroides	Stenotrophomonas	Comamonas
Acidovorax	Pseudomonas	Acidovorax	Pseudomonas
Thermobifida	Bacillus	Escherichia	Acetomicrobium
Comamonas	Stenotrophomonas	Delftia	Methanobrevibacter
Lactobacillus	Acidovorax	Clostridium	Proteiniclasticum
Xanthomonas	Prevotella	Shigella	Alkanindiges
Methanobrevibacter	Ruminococcus	Thermobifida	Acholeplasma
Bacteroides	Xylella	Xylella	Planifilum
Pseudoxanthomonas	Eubacterium	Bacteroides	Acidovorax
Thermobispora	Lactobacillus	Bacillus	Sphingobacterium

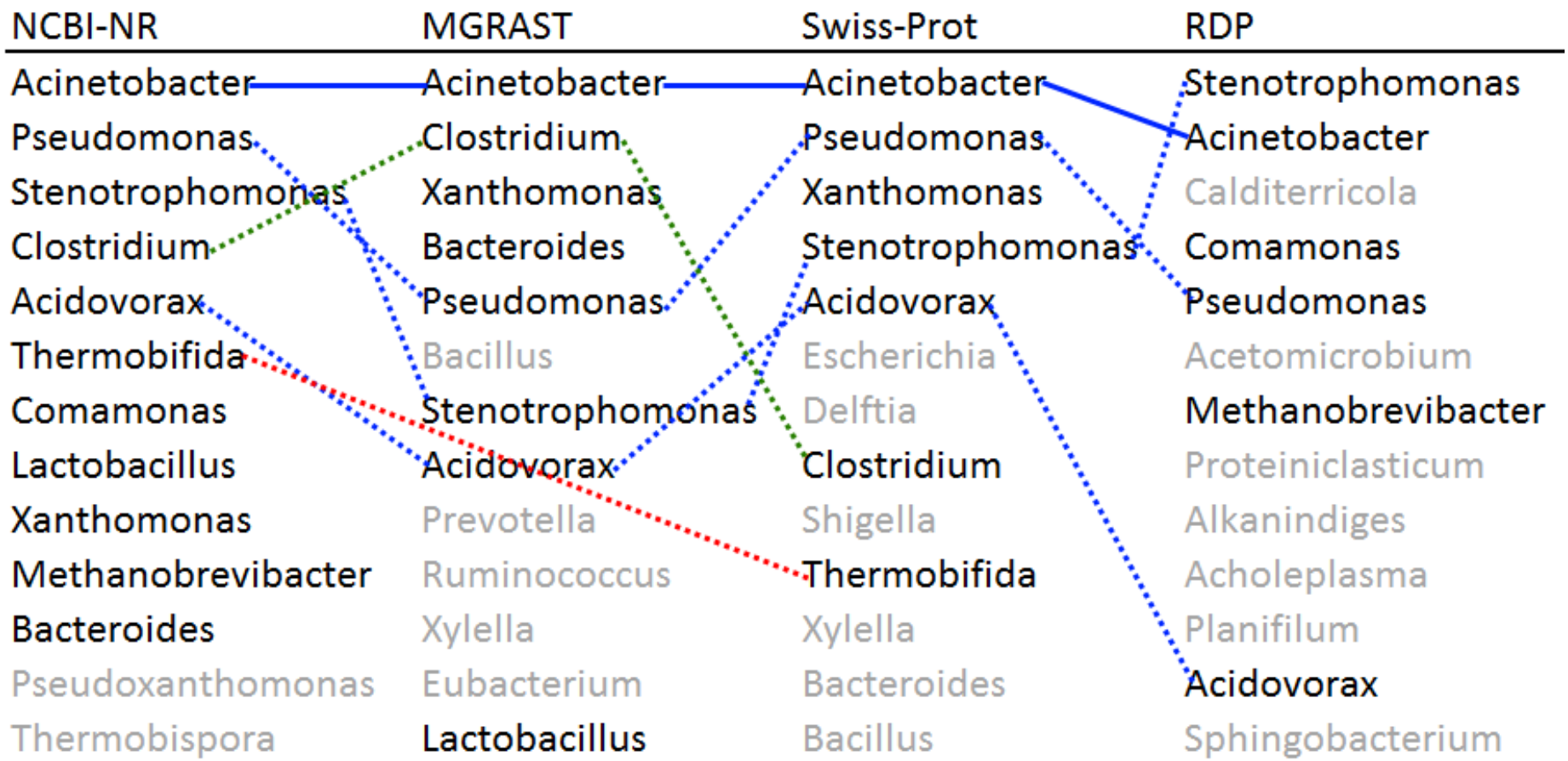
Comparação dos Dados do Zoológico - ZC1

NCBI-NR	MGRAST	Swiss-Prot	RDP
Acinetobacter	Acinetobacter	Acinetobacter	Stenotrophomonas
Pseudomonas	Clostridium	Pseudomonas	Acinetobacter
Stenotrophomonas	Xanthomonas	Xanthomonas	Calditerricola
Clostridium	Bacteroides	Stenotrophomonas	Comamonas
Acidovorax	Pseudomonas	Acidovorax	Pseudomonas
Thermobifida	Bacillus	Escherichia	Acetomicrobium
Comamonas	Stenotrophomonas	Delftia	Methanobrevibacter
Lactobacillus	Acidovorax	Clostridium	Proteiniclasticum
Xanthomonas	Prevotella	Shigella	Alkanindiges
Methanobrevibacter	Ruminococcus	Thermobifida	Acholeplasma
Bacteroides	Xylella	Xylella	Planifilum
Pseudoxanthomonas	Eubacterium	Bacteroides	Acidovorax
Thermobispora	Lactobacillus	Bacillus	Sphingobacterium

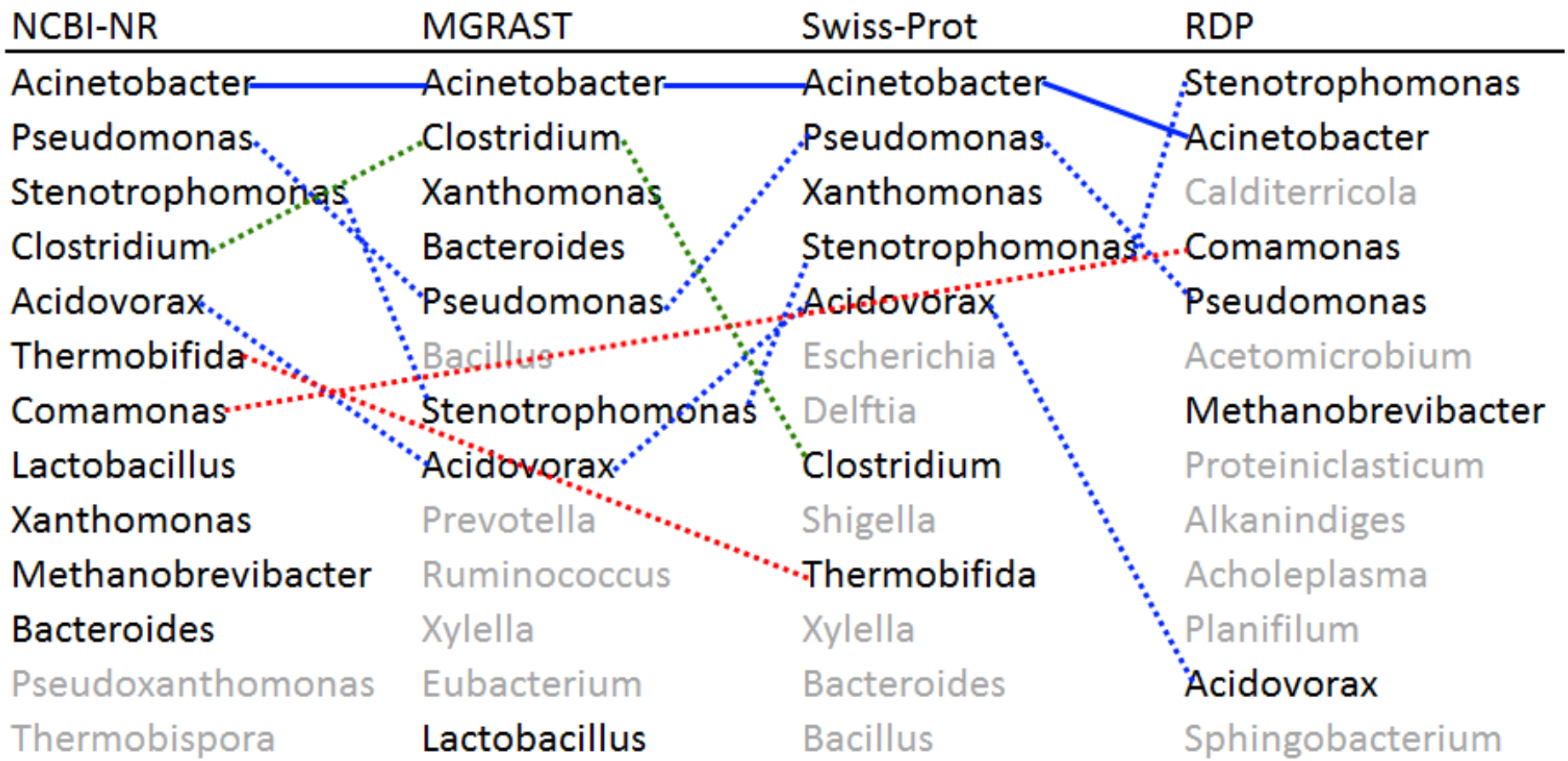
Comparação dos Dados do Zoológico - ZC1



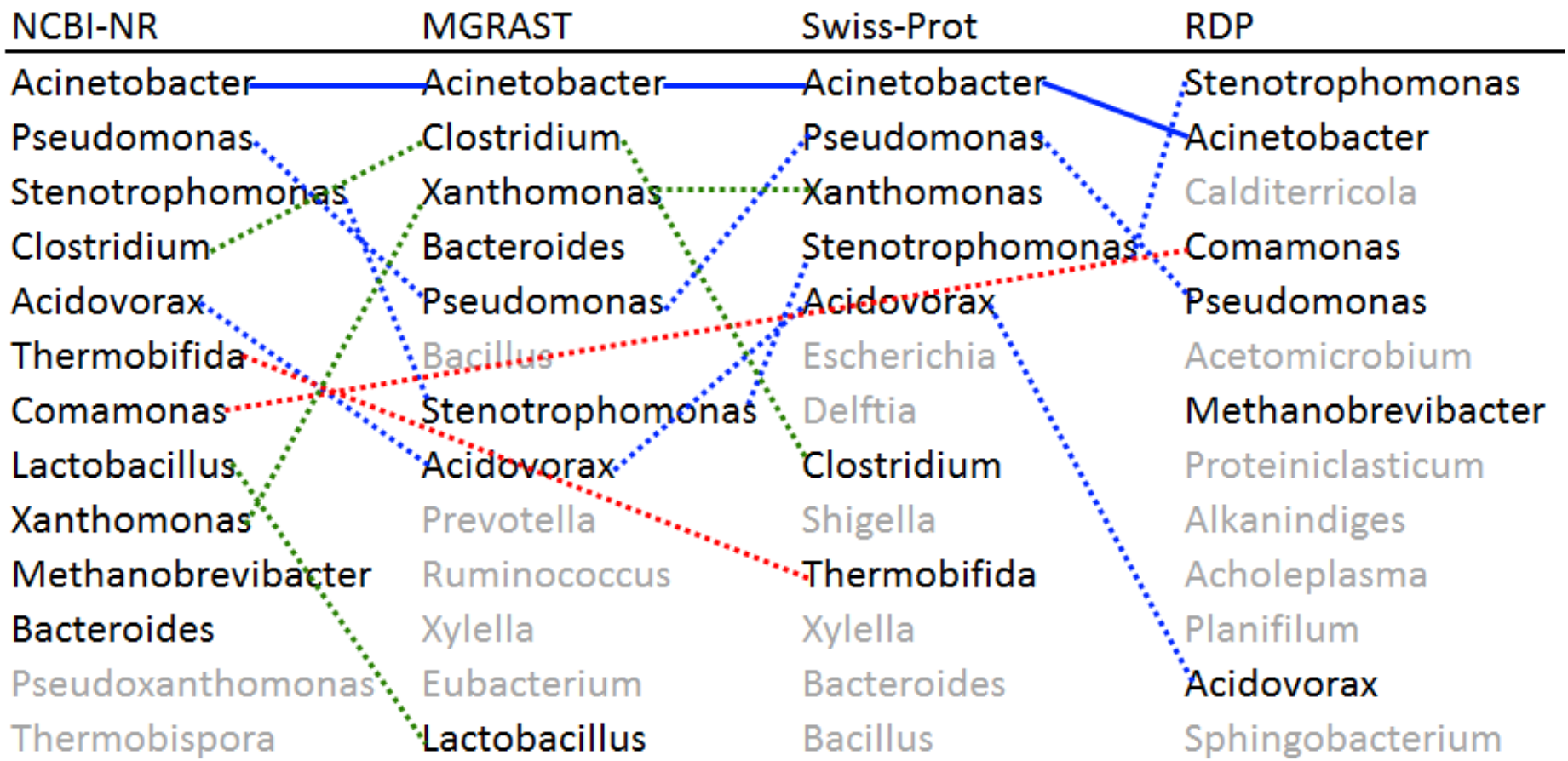
Comparação dos Dados do Zoológico - ZC1



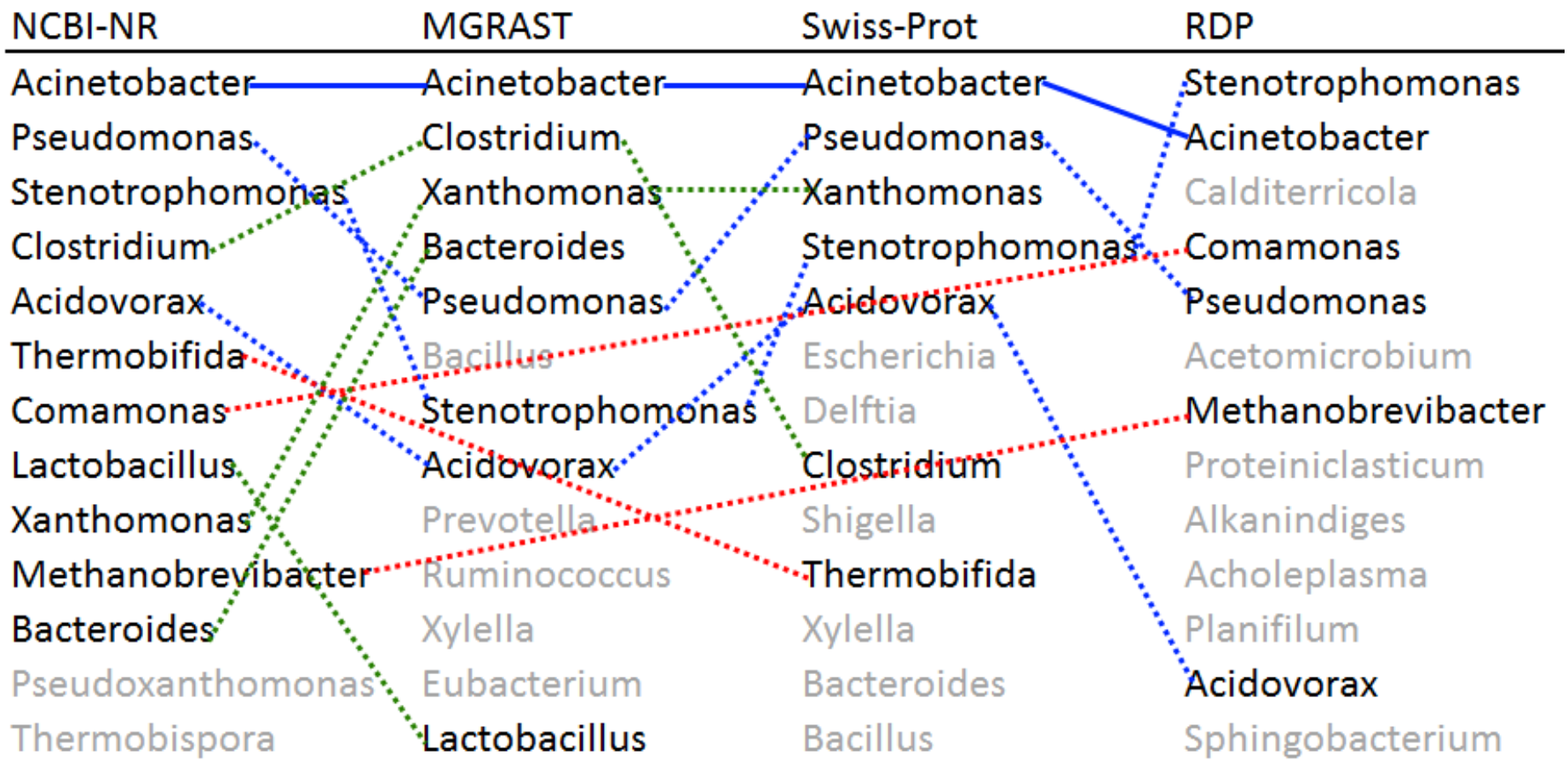
Comparação dos Dados do Zoológico - ZC1



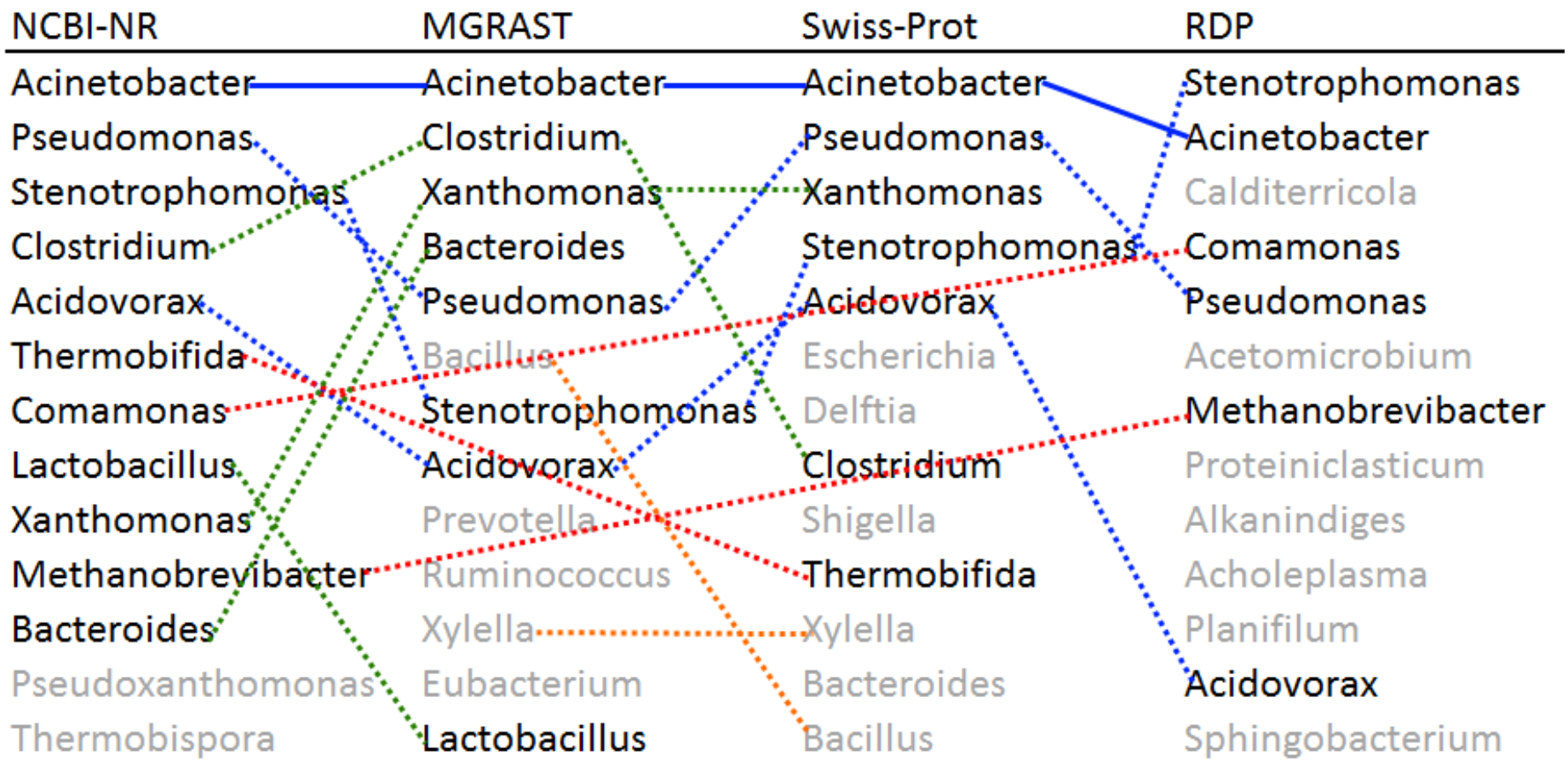
Comparação dos Dados do Zoológico - ZC1



Comparação dos Dados do Zoológico - ZC1



Comparação dos Dados do Zoológico - ZC1



Comparação dos Dados do Zoológico - ZC2

NCBI-NR	MG-RAST	Swiss-Prot	RDP
Lactobacillus	Lactobacillus	Lactobacillus	Lactobacillus
Escherichia	Enterococcus	Escherichia	Atopostipes
Pediococcus	Streptococcus	Shigella	Jeotgalicoccus
Corynebacterium	Clostridium	Streptococcus	Weissella
Acinetobacter	Bacillus	Pediococcus	Staphylococcus
Bacteroides	Pediococcus	Staphylococcus	Corynebacterium
Weissella	Bacteroides	Pseudomonas	Streptococcus
Streptococcus	Staphylococcus	Acinetobacter	Pediococcus
Bifidobacterium	Corynebacterium	Corynebacterium	Acinetobacter
Enterococcus	Leuconostoc	Bifidobacterium	Turicibacter
Shigella	Listeria	Salmonella	Clostridium
Leuconostoc	Escherichia	Enterococcus	Enterococcus
Acetobacter	Acinetobacter	Klebsiella	Bacteroides
Clostridium	Bifidobacterium	Bacteroides	Bifidobacterium

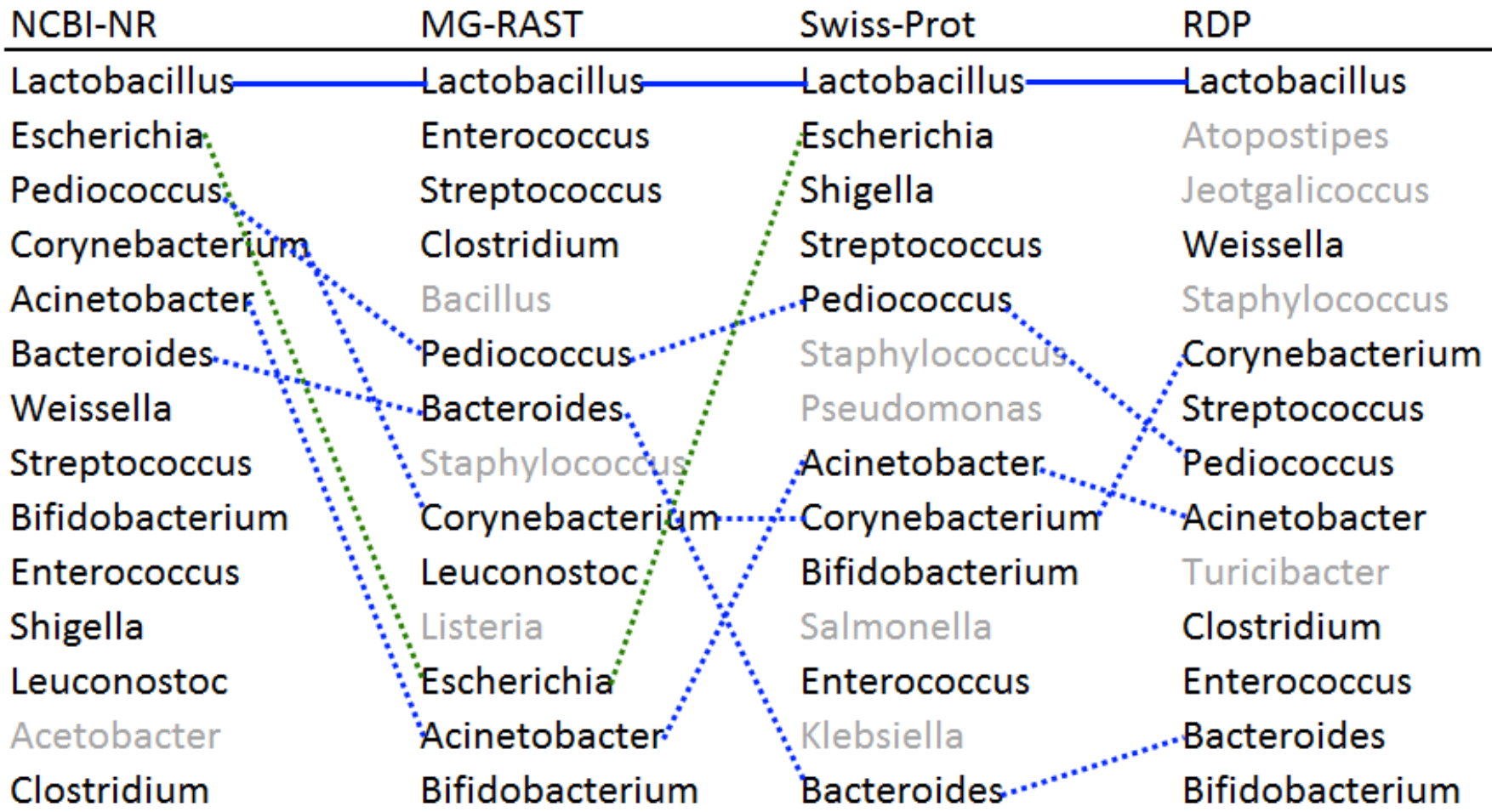
Comparação dos Dados do Zoológico - ZC2

NCBI-NR	MG-RAST	Swiss-Prot	RDP
Lactobacillus	Lactobacillus	Lactobacillus	Lactobacillus
Escherichia	Enterococcus	Escherichia	Atopostipes
Pediococcus	Streptococcus	Shigella	Jeotgalicoccus
Corynebacterium	Clostridium	Streptococcus	Weissella
Acinetobacter	Bacillus	Pediococcus	Staphylococcus
Bacteroides	Pediococcus	Staphylococcus	Corynebacterium
Weissella	Bacteroides	Pseudomonas	Streptococcus
Streptococcus	Staphylococcus	Acinetobacter	Pediococcus
Bifidobacterium	Corynebacterium	Corynebacterium	Acinetobacter
Enterococcus	Leuconostoc	Bifidobacterium	Turicibacter
Shigella	Listeria	Salmonella	Clostridium
Leuconostoc	Escherichia	Enterococcus	Enterococcus
Acetobacter	Acinetobacter	Klebsiella	Bacteroides
Clostridium	Bifidobacterium	Bacteroides	Bifidobacterium

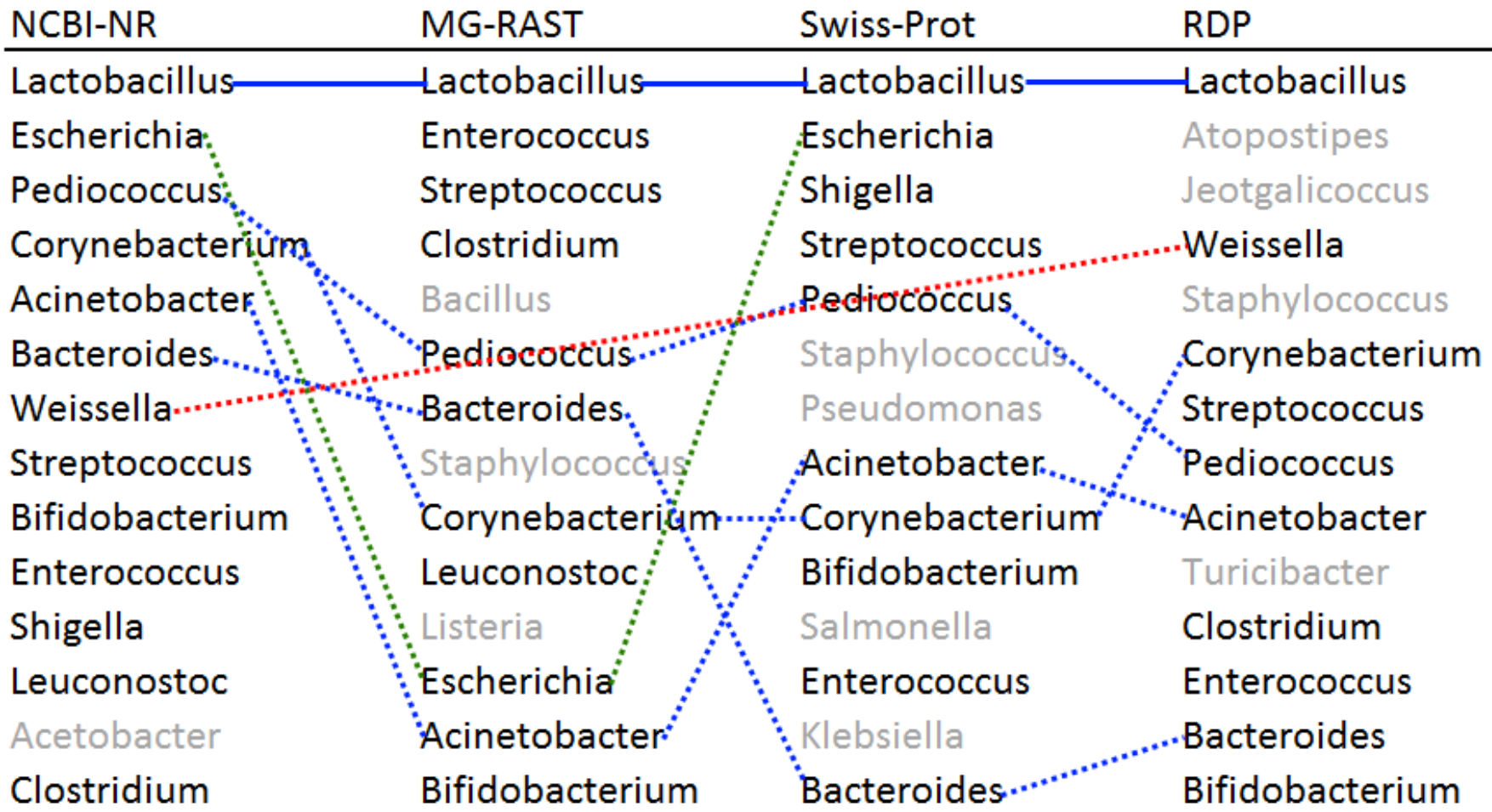
Comparação dos Dados do Zoológico - ZC2

NCBI-NR	MG-RAST	Swiss-Prot	RDP
Lactobacillus	Lactobacillus	Lactobacillus	Lactobacillus
Escherichia	Enterococcus	Escherichia	Atopostipes
Pediococcus	Streptococcus	Shigella	Jeotgalicoccus
Corynebacterium	Clostridium	Streptococcus	Weissella
Acinetobacter	Bacillus	Pediococcus	Staphylococcus
Bacteroides	Pediococcus	Staphylococcus	Corynebacterium
Weissella	Bacteroides	Pseudomonas	Streptococcus
Streptococcus	Staphylococcus	Acinetobacter	Pediococcus
Bifidobacterium	Corynebacterium	Corynebacterium	Acinetobacter
Enterococcus	Leuconostoc	Bifidobacterium	Turicibacter
Shigella	Listeria	Salmonella	Clostridium
Leuconostoc	Escherichia	Enterococcus	Enterococcus
Acetobacter	Acinetobacter	Klebsiella	Bacteroides
Clostridium	Bifidobacterium	Bacteroides	Bifidobacterium

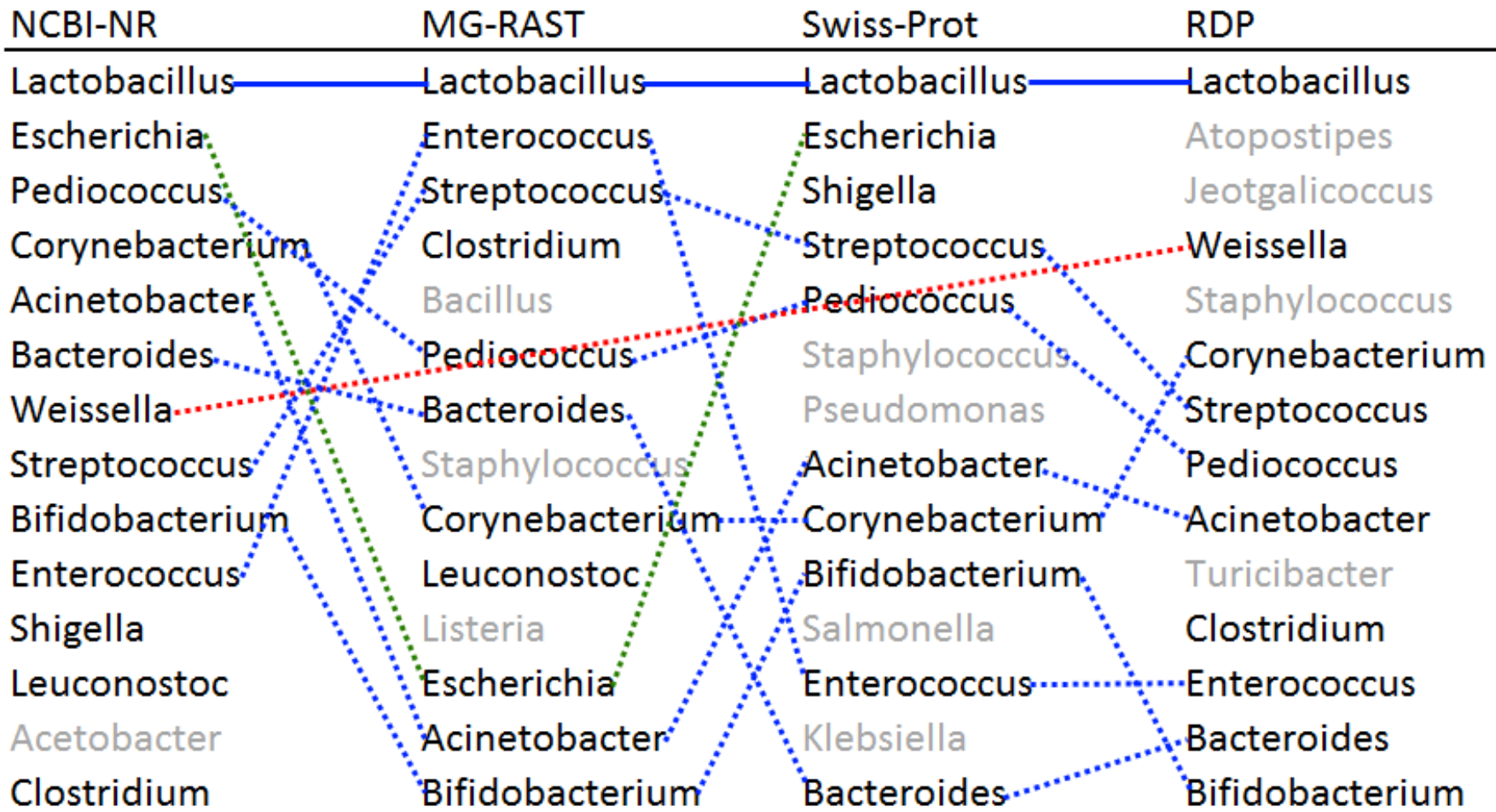
Comparação dos Dados do Zoológico - ZC2



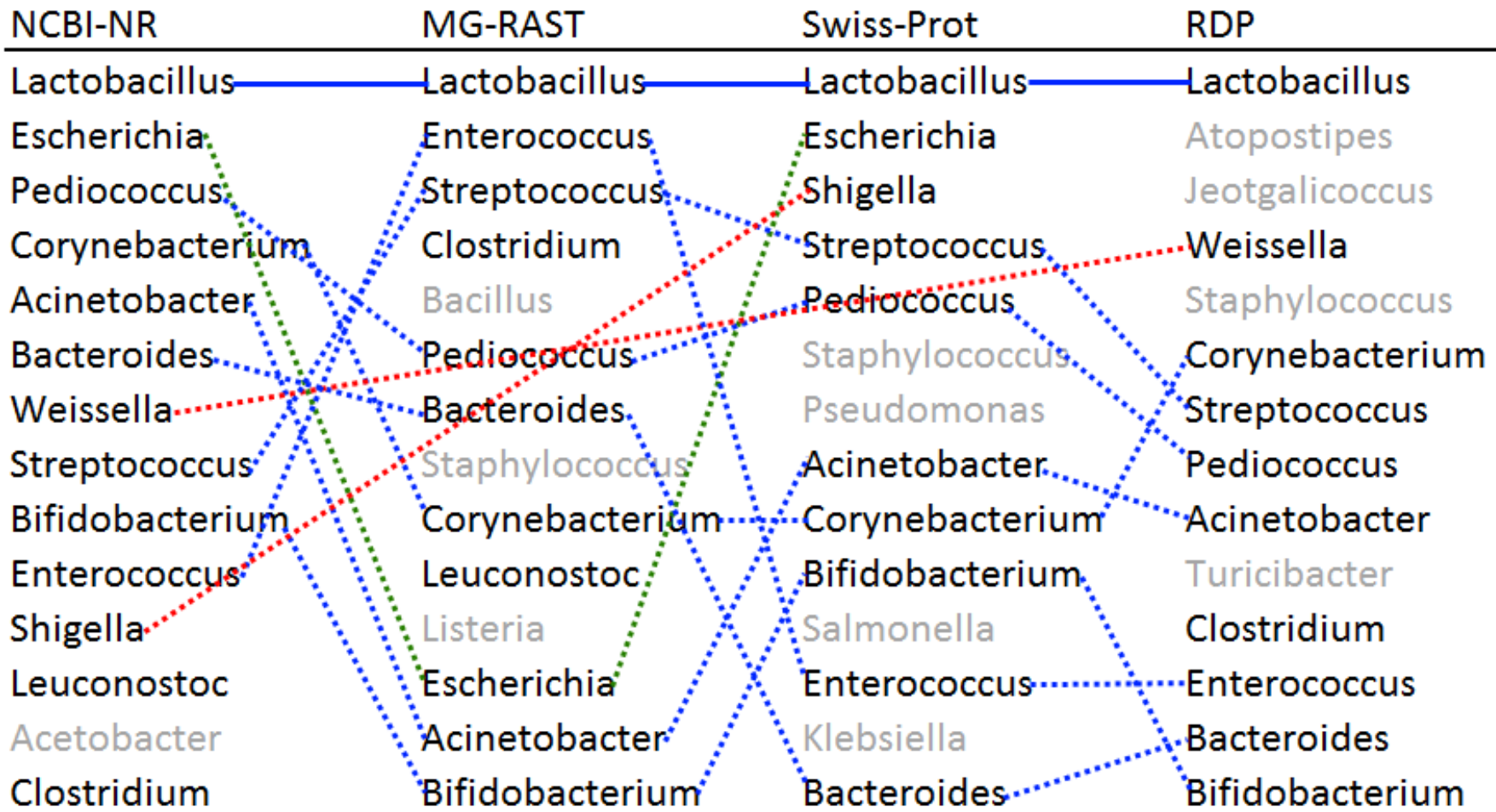
Comparação dos Dados do Zoológico - ZC2



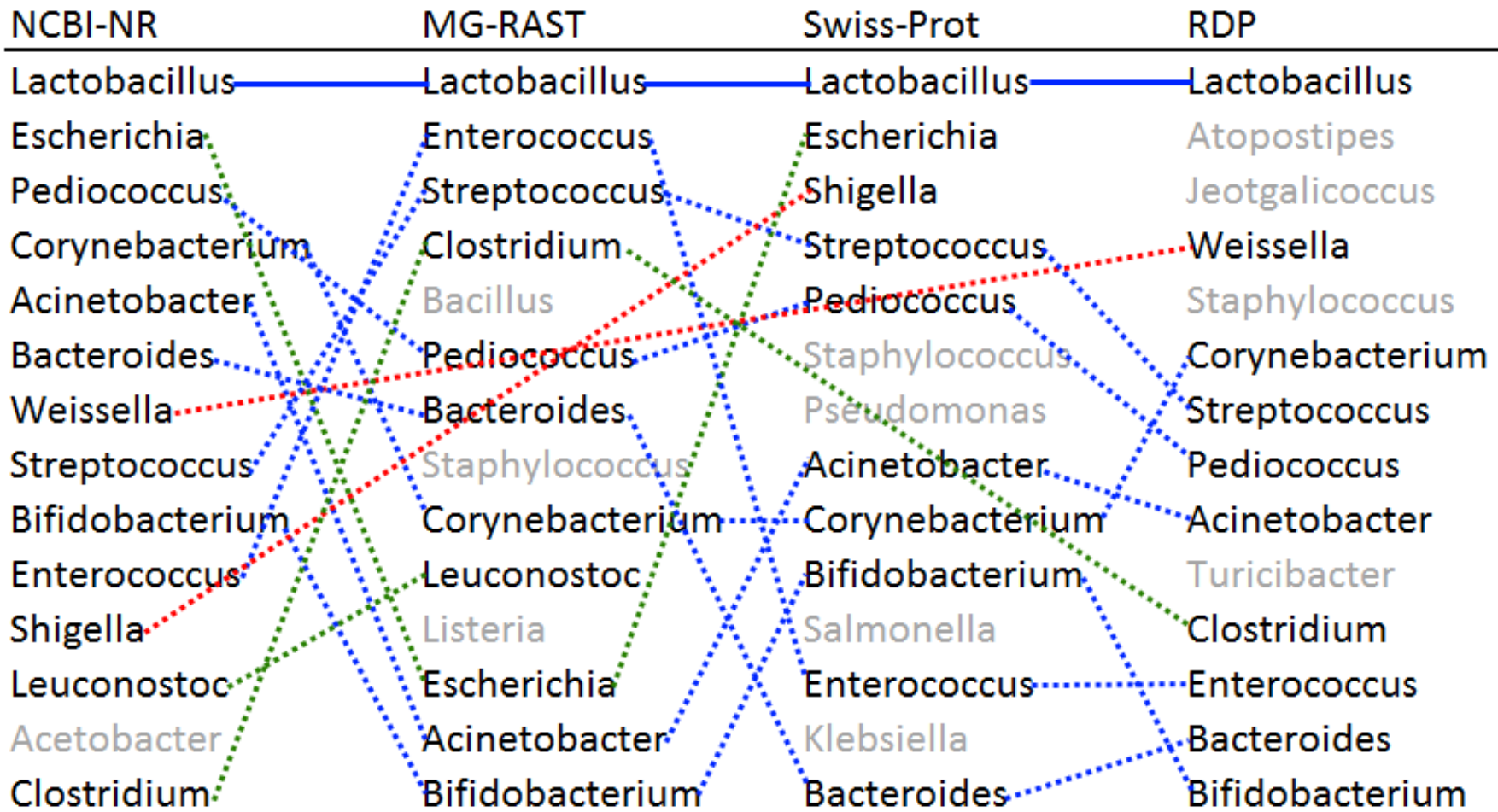
Comparação dos Dados do Zoológico - ZC2



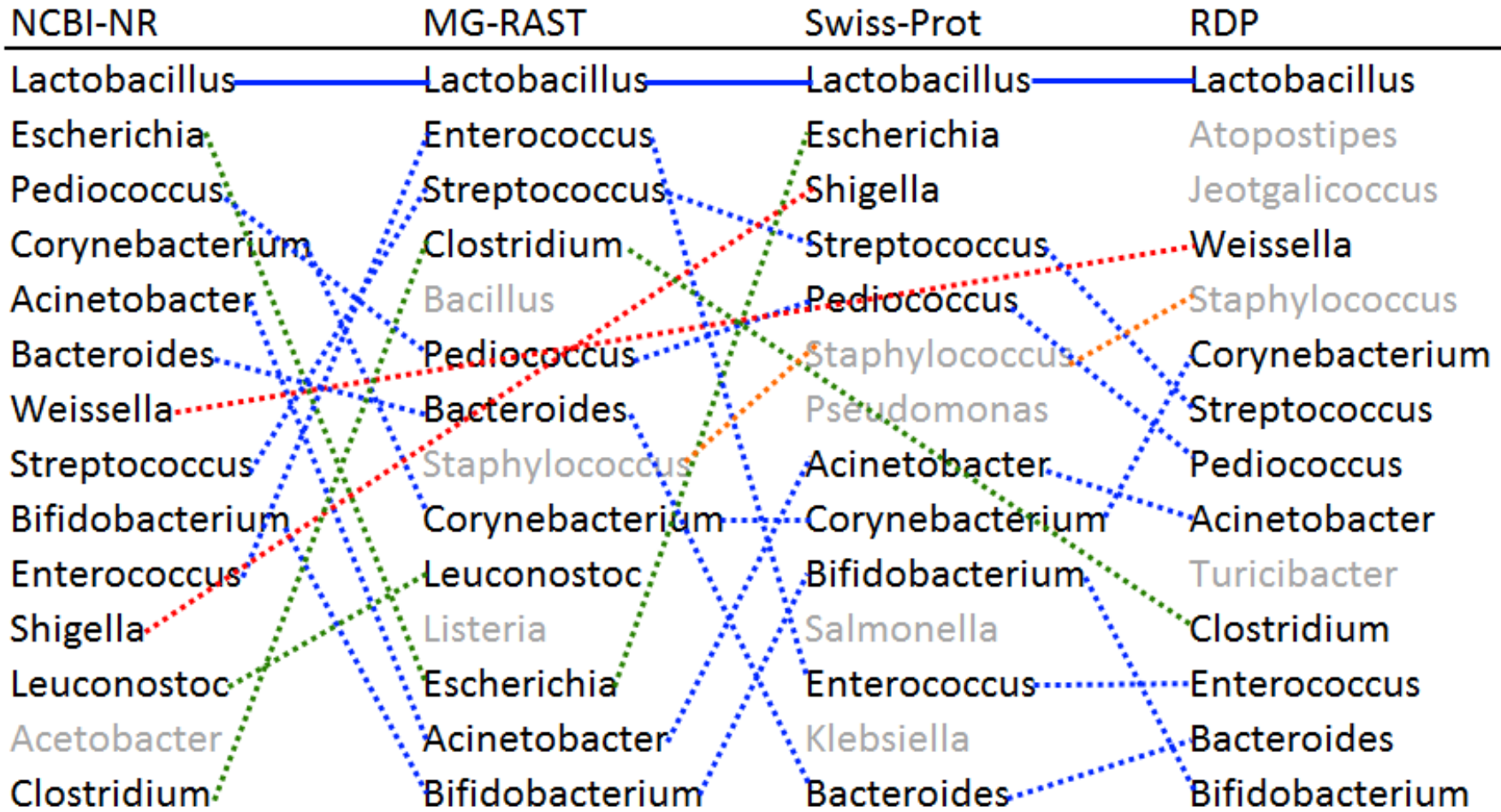
Comparação dos Dados do Zoológico - ZC2



Comparação dos Dados do Zoológico - ZC2



Comparação dos Dados do Zoológico - ZC2



Conclusões

- Com os dados simulados, o NCBI-NR e o Swiss-Prot apresentaram grandes concordâncias com o resultado esperado, sugerindo que as análises com esses dois bancos são confiáveis.

Conclusões

- Com os dados simulados, o NCBI-NR e o Swiss-Prot apresentaram grandes concordâncias com o resultado esperado, sugerindo que as análises com esses dois bancos são confiáveis.
- Os resultados do RDP para os dados do Zoológico concordam, principalmente na amostra ZC2, com os resultados do NCBI-NR, que se mostrou o melhor indicador da composição da população.

Conclusões

- Com os dados simulados, o NCBI-NR e o Swiss-Prot apresentaram grandes concordâncias com o resultado esperado, sugerindo que as análises com esses dois bancos são confiáveis.
- Os resultados do RDP para os dados do Zoológico concordam, principalmente na amostra ZC2, com os resultados do NCBI-NR, que se mostrou o melhor indicador da composição da população.
- Apesar do MG-RAST também se mostrar confiável através da simulação, ele identificou uma quantidade significativa de gêneros que não existiam no conjunto original. Esse ruído pode atrapalhar uma análise de diversidade aprofundada.

Trabalhos Futuros

- Uma comparação mais completa das sequências do Zoológico com o banco NCBI-NR pode revelar resultados mais precisos.

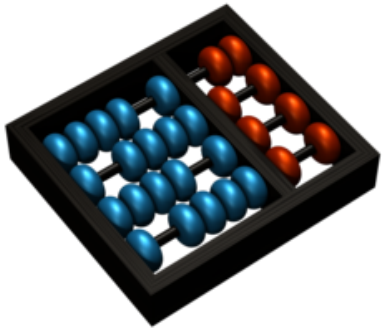
Trabalhos Futuros

- Uma comparação mais completa das sequências do Zoológico com o banco NCBI-NR pode revelar resultados mais precisos.
- O estudo da diversidade de uma população não é apenas uma análise das espécies que a compõe. Uma análise mais detalhada dos genes encontrados nesse metagenoma pode mostrar o funcionamento da população e sua interação com o ambiente.

Trabalhos Futuros

- Uma comparação mais completa das sequências do Zoológico com o banco NCBI-NR pode revelar resultados mais precisos.
- O estudo da diversidade de uma população não é apenas uma análise das espécies que a compõe. Uma análise mais detalhada dos genes encontrados nesse metagenoma pode mostrar o funcionamento da população e sua interação com o ambiente.
- As amostras utilizadas ao longo desse trabalho não possuem relações diretas entre si. Uma análise de amostras coletadas ao longo de um mesmo processo de compostagem pode fornecer mais detalhes sobre a dinâmica microbiológica desse processo.

Agradecimentos



UNICAMP

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico