Trabalho I

Bioestatística - IMD0601 - DEB1010 - UFRN

professores: Tetsu Sakamoto e Beatriz Stransky

Instruções

- Este trabalho consiste de uma série de exercícios que avaliará o conhecimento sobre manipulação e análise exploratória dos dados e estatística descritiva em ambiente R.
- Realize todos os procedimentos em ambiente R.
- Submeta um script em R com os comandos utilizados para responder cada um dos exercícios.
- Se o exercício pedir um gráfico, utilize a biblioteca ggplot2.
- O trabalho é **individual** e deve ser submetido até o dia **16/07/2021**, via SIGAA.
- Valor total do trabalho: 10 pontos (questões de 2 a 6, 2 pontos cada).

Sobre o conjunto de dados

O conjunto de dados que será utilizado para realizar este trabalho são dados genômicos de bactérias que possui o status de completo (Sequencing status: Finished) que estão disponíveis no site do IMG/G (https://img.jgi.doe.gov/cgi-bin/m/main.cgi). Células vazias são dados faltantes. A tabela consiste nas seguintes colunas:

nome da coluna	**descrição**
taxon_oid	identificador único da amostra sequenciada
Domain	Domínio taxonômico
Sequencing Status	Status do sequenciamento
Study Name	Nome do estudo relacionado a amostra
Genome Name / Sample Name	Nome da amostra sequenciada
Sequencing Center	Local onde a amostra foi sequenciada
IMG Genome ID	Identificador único da amostra próprio do IMG

Phylum	Filo taxonômico da amostra
Class	Classe taxonômica da amostra
Order	Ordem taxonômica da amostra
Family	Família taxonômica da amostra
Genus	Gênero taxonômico da amostra
Species	Espécie taxonômica da amostra
Assembly Method	Método utilizado para montagem do genoma
Release Date	Data de disponibilização do genoma
Biotic Relationships	Relação biótica do organismo
Cell Shape	Forma da célula do organismo
Energy Source	Fonte de energia utilizada pelo organismo
Oxygen Requirement	Tipo de respiração utilizada pelo organismo
Sequencing Method	Método de sequenciamento utilizado
Sporulation	Se o organismo esporula ou não
Genome Size * assembled	Tamanho do genoma em pb
Gene Count * assembled	Contagem de genes
CRISPR Count * assembled	Contagem de CRISPR
GC Count * assembled	Contagem de GC no genoma
CDS Count * assembled	Contagem de CDS
RNA Count * assembled	Contagem de RNA
16S rRNA Count * assembled	Contagem de 16S rRNA
23S rRNA Count * assembled	Contagem de 23S rRNA
Pseudo Genes Count	Contagem de Pseudogenes
Unchar Count	Contagem de genes não caracterizados
w/ Func Pred Count * assembled	Contagem de genes com uma função predita
w/o function prediction * assembled	Contagem de genes sem uma função predita
Paralogs Count	Contagem de parálogos

Exercícios

1. Carregue o arquivo *data.tab* inteiro no ambiente R. (Parece um processo simples, mas você pode precisar consultar o manual da função que carrega arquivos no R e o conteúdo de *data.tab* para conseguir realizar esta tarefa)

2. Qual a classe taxonômica de bactérias (coluna Class) possui maior número de espécies distintas?

3. Crie uma tabela que indique a média, o valor máximo e o valor mínimo, da contagem de CDS encontrados em cada família bacteriana. Mantenha as colunas Phylum e Class e ordene a tabela primeiro por família, depois por classe e finalmente por filo.

4. Faça um gráfico de barra empilhado que indique a proporção das diferentes formas de bactérias quanto ao requerimento do oxigênio (coluna Oxygen Requirement) entre os diferentes filos bacterianos (coluna Phylum).

5. Como o uso das tecnologias de sequenciamento evoluiu ao longo dos anos? Organize os dados e mostre em um gráfico a frequência de utilização de cada tecnologia, a cada ano. OBS: Considere apenas a marca do sequenciador (Illumina) e não o modelo (por exemplo, “Illumina HiSeq 2000” e “Illumina GAIIx”).

6. Analise a correlação entre as variáveis e mostre, graficamente e utilizando o coeficiente de correlação adequado, as 2 correlações mais fortes e as 2 mais fracas em relação ao tamanho do genoma (coluna Genome Size * assembled).