Trabalho I

IMD0601 - Bioestatística - Instituto Metrópole Digital - UFRN

Instruções:

- Este trabalho consiste de uma série de exercícios que avaliará o conhecimento do aluno sobre análise de manipulação dos dados e de estatística descritiva em ambiente R;
- Realize todos os procedimentos em ambiente R;
- Apresente os comandos em R utilizados para responder o exercício;
- Se o exercício pedir um gráfico para responder a pergunta, utilize a biblioteca ggplot2 do R para tal.
- Este trabalho deve ser entregue até o dia 10/04/2019 às 23:59

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. (1 ponto) 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. (1 ponto) Quais são as instituições que mais depositaram dados genômicos bacterianos neste banco de dados? Faça um gráfico para auxiliar na resposta deste exercício.

3. (1 ponto) Os dados sobre o tamanho do genoma segue uma distribuição normal? Faça testes e gráficos que justifiquem a sua resposta.

4. (1 ponto) Como é o perfil taxonômico em nível de gênero dos genomas listados nesta tabela (em outras palavras, qual a frequência de cada classe na coluna Genus)? Existe um viés de sequenciamento para algum gênero? Faça gráficos que justifiquem sua resposta.

5. (2 pontos) Calcule a média do tamanho do genoma de cada gênero presente na tabela e refaça os testes e gráficos para verificar a normalidade dos dados. Os dados desta vez seguem uma distribuição normal? Se não, qual procedimento você adotaria para que seus dados aproximem mais de uma distribuição normal?

6. (1 ponto) Digamos que você tenha interesse em estudar organismos que estejam listados na tabela e que tenham menores proporções de proteínas com função predita. Cite 5 organismos de diferentes gêneros que seriam candidatos para seu estudo.

7. (3 pontos) Como foi o uso dos sequenciadores utilizados para o sequenciamento dos organismos listados na tabela ao longo do tempo? Faça um gráfico onde é possível verificar o uso das diferentes tecnologias de sequenciamento ao longo do tempo. Nesta análise, considere apenas a marca do sequenciador (por exemplo, para “Illumina GAIIx”, considere apenas “Illumina”). Se uma mesma amostra foi sequenciada por dois ou mais modelos de sequenciadores da mesma marca (por exemplo, “Illumina HiSeq 2000” e “Illumina HiSeq 2500”), considere apenas uma contagem para a marca repetida. E se uma mesma amostra foi sequenciada por dois ou mais modelos de sequenciadores de marcas diferentes, considere uma contagem para cada uma das marcas.