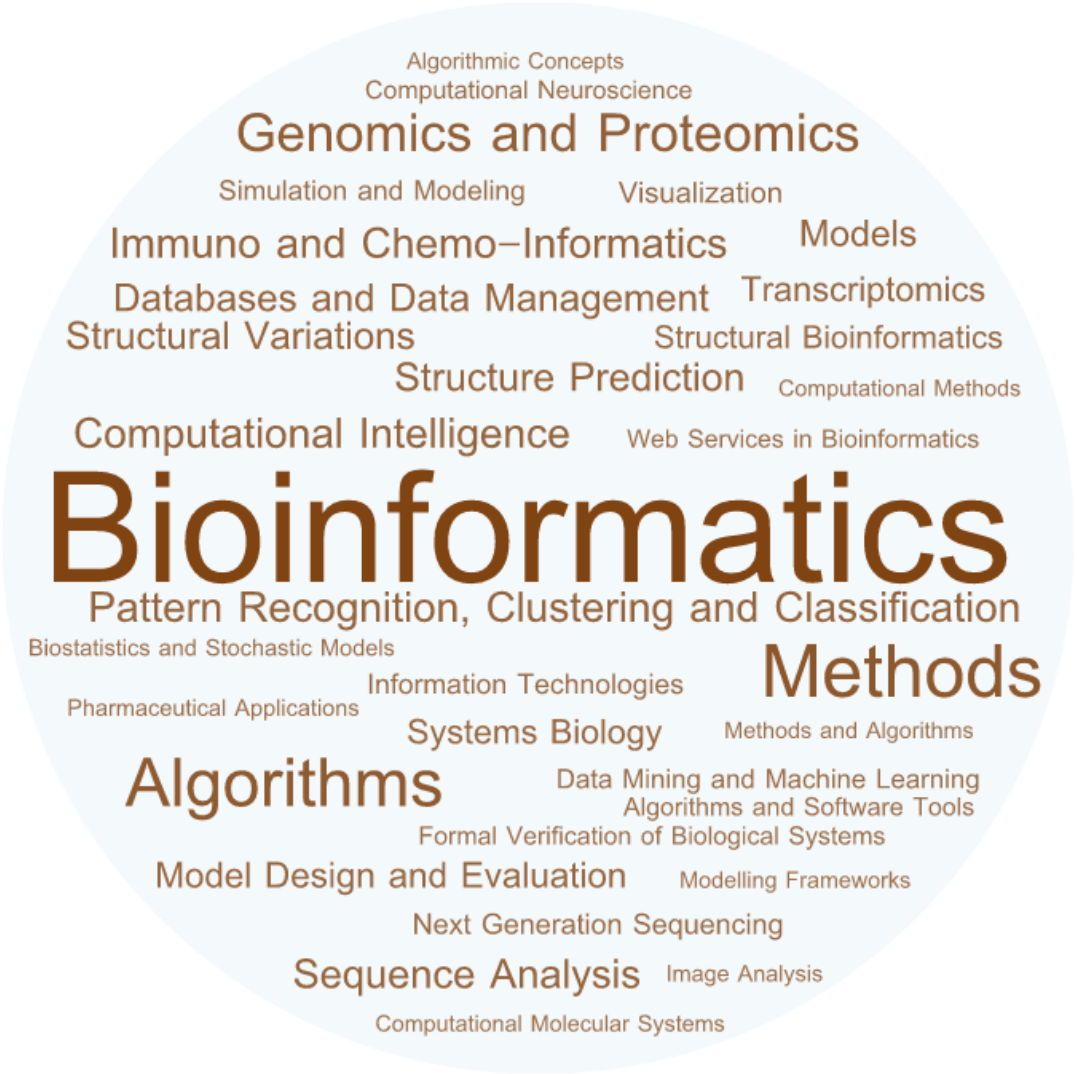
Ciências Ômicas em Doenças Infecciosas (ICB5747)

Apostila Transcriptômica



2022

1 Introdução a transcriptômica

Com o recente avanço das técnicas de sequenciamento de genomas e com o desenvolvimento do Sequenciamento de Nova Geração (NGS), muitas possibilidades e caminhos se abriram, tendo início a era genômica em larga escala. Várias plataformas de NGS estão disponíveis e o RNA-seq (do inglês RNA *sequencing*) é um exemplo de técnica que aplica essas plataformas para análise de transcriptoma. Transcriptoma refere-se ao conjunto total de transcritos (incluindo mRNAs e ncRNAs como miRNAs) de uma célula e suas quantidades em um determinado instante, representando assim uma "foto" do conjunto de transcritos do momento estudado. Esse tipo de análise permite identificar e caracterizar os transcritos quanto a níveis de expressão, inferir modificações pós-transcricionais, como padrão diferente de *splicing* e encontrar mutações e polimorfismos de nucleotídeo único (SNPs). Todos os eventos biológicos que ocorrem dentro de uma célula são governados, principalmente, por mudanças na expressão de genes importantes, e essa habilidade de ativar e reprimir a expressão dos genes é um fator importante para regular todas as funções e atividades biológicas. Tendo em mãos o perfil de expressão de uma célula e os softwares necessários, é possível realizar a quantificação das modificações dos níveis do transcrito em diferentes ambientes ou situações em que se encontra (análises de expressão diferencial), como, por exemplo, encontrar transcritos diferencialmente expressos em célula de tecido atingido por tumor em relação ao mesmo tecido saudável. Portanto, análises de transcriptoma tornaram-se uma ferramenta importantíssima na identificação de genes ou grupos de genes que desempenham papel-chave no desenvolvimento de doenças e de marcadores de diagnóstico e prognóstico.

1.1 O que é o sequencianto de RNA?

O sequenciamento de RNA não é uma coisa nova. Na verdade, antes mesmo de existirem técnicas para análise de sequenciamento de DNA, como os métodos descritos por Sanger e Maxam & Gilbert (ambos de 1977), sequências de RNAs não-codificantes (como tRNAs) já haviam sido publicadas, sendo obtidas por degradação química ou enzimática.

Entre os anos 80 e começo dos anos 2000, diferentes abordagem foram desenvolvidas para a análise do transcriptoma (conjunto de transcritos de um organismo em uma condição) através do sequenciamento de Sanger, como EST (*expressed sequence tags*) e SAGE (*serial analysis of gene expression*). Esses métodos, no entanto, eram limitados pelo baixo desempenho (*throughput*) das plataformas de Sanger, que, por "rodada", permitiam no

máximo que 96 amostras fossem sequenciadas por vez. Por conta disso, neste período os microarranjos (*microarray*) se tornaram uma alternativa útil para o estudo em larga escala da expressão gênica, mas estes apresentavam uma outra limitação: apenas regiões conhecidas poderiam ser analisadas.

Com o advento do sequenciamento de nova geração (NGS), o *throughput* gerado por cada rodada de sequenciamento passou de algumas centenas de kilobases (Kbs) para vários gigabases (Gbs). Apesar de ter sido inicialmente direcionado para o estudo de sequências genômicas, o NGS foi rapidamente adotado também por muitas outras ciências "ômicas", dentre elas a transcriptômica. Para fins de diferenciação, adotou-se o termo RNA-seq para se referir à análise do transcriptoma através destas novas tecnologias, apesar dos primeiros trabalhos produzidos com esta abordagem ainda terem usado o nome EST.

A análise do transcriptoma com RNA-seq pode ser realizada com ou sem uma sequência genômica de referência. No primeiro caso, chamamos esta análise de *reference-guided*, e esta abordagem geralmente é empregada quando estamos analisando genomas microbianos ou eucarióticos "modelo" (por exemplo, *Homo sapiens*, *Mus musculus*). Já a análise sem referência (*de novo*) é geralmente realizada para organismos não-modelo.

No caso da análise *reference-guided*, os resultados do sequenciamento, as leituras (*reads*) são alinhadas contra o genoma e a contagem de leituras mapeadas em cada gene é usada para pedir a sua expressão. Já no caso *de novo*, é necessário fazer primeiro uma montagem dos transcritos, para que depois seja possível realizar a sua quantificação.

Como a expressão gênica varia de acordo com as condições na qual o organismo se encontra, é possível utilizar a análise de RNA-seq para mensurar o efeito de diferentes tratamentos na expressão de diferentes genes simultaneamente. Este tipo de análise, denominada expressão gênica diferencial, permite entender como o perfil de expressão de um determinado organismo é alterado ao ser submetido a uma determinada condição. Entretanto, diferente de métodos clássicos de biologia molecular, como os baseados em PCR, no caso do RNA-seq não precisamos escolher um gene normalizador.

Após identificarmos genes com expressão aumentada (*up-regulated*) ou diminuída (*down-regulated*), podemos mapear processos biológicos que estão sendo modulados em resposta às mudanças na qual o organismo se encontra. Este tipo de abordagem é extremamente útil para se entender diferentes processos biológicos, como a dinâmica de expressão gênica em células neoplásicas submetidas a certos medicamentos, ou em plantas submetidas a uma determinada condição de estresse, por exemplo.

1.2 NCBI e GEO

O *National Center for Biotechnology Information* (NCBI) possui diferentes bancos de dados integrados, entre eles, o *BioProjects*, que constitue um meta recurso para dados biológicos depositados em repositórios de arquivos mantidos por membros da *International Nucleotide SequenceDatabase Consortium* (INSDC), que inclui o *DNA DataBank* do Japão (DDBJ), o *European Nucleotide Archive* (ENA) do *European Molecular Biology Laboratory* (EMBL) e o *GenBank* do NCBI.

O NCBI também possui um banco de dados de bioinformática que fornece um repositório público para dados de sequenciamento de DNA e RNA, como um armazenamento para dados brutos de sequenciamento gerados por tecnologias de próxima geração, incluindo Ilumina, IonTorrent, Complete Genomics, entre outras plataformas de sequenciamento. Atualmente, o NCBI armazena mais de 3 milhões de estudos, cobrindo mais de 8,6 milhões de amostras, incluindo 2.597.223 experimentos de acesso público e outros 674.522 estudos controlados. O *Gene Expression Omnibus* (GEO) é um repositório de dados genômicos funcional, internacional e público que arquiva e distribui gratuitamente *microarrays*, sequenciamentos de próxima geração e outras formas de dados genômicos funcionais de alto desempenho, oferecendo suporte a envios de dados em conformidade com o padrão “informações mínimas sobre um experimento de *microarray*” (do inglês *Minimum Information About a Microarray Experiment - MIAME*).

Dados baseados em matriz e sequência também são aceitos, além de serem fornecidas ferramentas para ajudar os usuários a consultar, analisar e fazer o download de experimentos e perfis de expressão gênica para fins como curadoria. Hoje, o repositório conta com mais de 5 milhões de amostras, oriundas de diversos organismos, das quais 4.510.937 são públicas, dentro de uma variedade de tecnologias distribuídas entre variados estudos, tais como perfil de expressão gênica, perfil de metilação por *array*, entre outros.

2 Introdução ao R

O R é um software livre(gratuito) e colaborativo para computação estatística e construção de gráficos. Além disso, o R também é uma linguagem de programação e por isso está em constante atualização, gerado por sua comunidade ativa ao redor do mundo. Linguagem de Programação colaborativa é quando suas atualizações são feitas pelos próprios usuários. No R, essas atualizações são denominadas “pacotes”.

2.1 Como instalar o R:

O R pode ser baixado diretamente de seu site http://www.r-project.org e está disponível para as plataformas Linux, Windows e MacOS. Para instalar , siga as instruções a seguir, considerando a sua plataforma:

-- Windows:

1. Acesse o site do R.
2. Clique em **CRAN**.
3. Escolha o *mirror* de sua preferência.
4. Clique na opção **Download R for Windows**
5. Clique na opção **base** ou **install R for the first time**.
6. Escolha a opção **Download R-x.x.x for Windows** (versão atual).
7. Execute o instalador.

-- macOS:

1. Acesse o site do R.
2. Clique em **CRAN**.
3. Escolha o *mirror* de sua preferência.
4. Clique na opção **Download R for macOS.**
5. Escolha a versão mais atual adequada ao seu sistema.
6. Salve o arquivo .pkg, abra e siga as instruções de instalação.

-- Linux:

1. Acesse o site do R.
2. Clique em **CRAN**.
3. Escolha o *mirror* de sua preferência.
4. Clique na opção **Download R for Linux.**

5. Clique na distro adequada ao seu sistema.
6. Siga as instruções de instalação.

2.2 RStudio

O RStudio é um ambiente de desenvolvimento integrado (IDE) para o R e está disponível em duas edições: RStudio Desktop e RStudio Server. Utilizaremos a versão gratuita para Desktop. O RStudio diferente do R, possui um apelo visual muito maior e busca melhorar a experiência do usuário com o ambiente R. Para entender melhor compare as duas imagens a seguir:

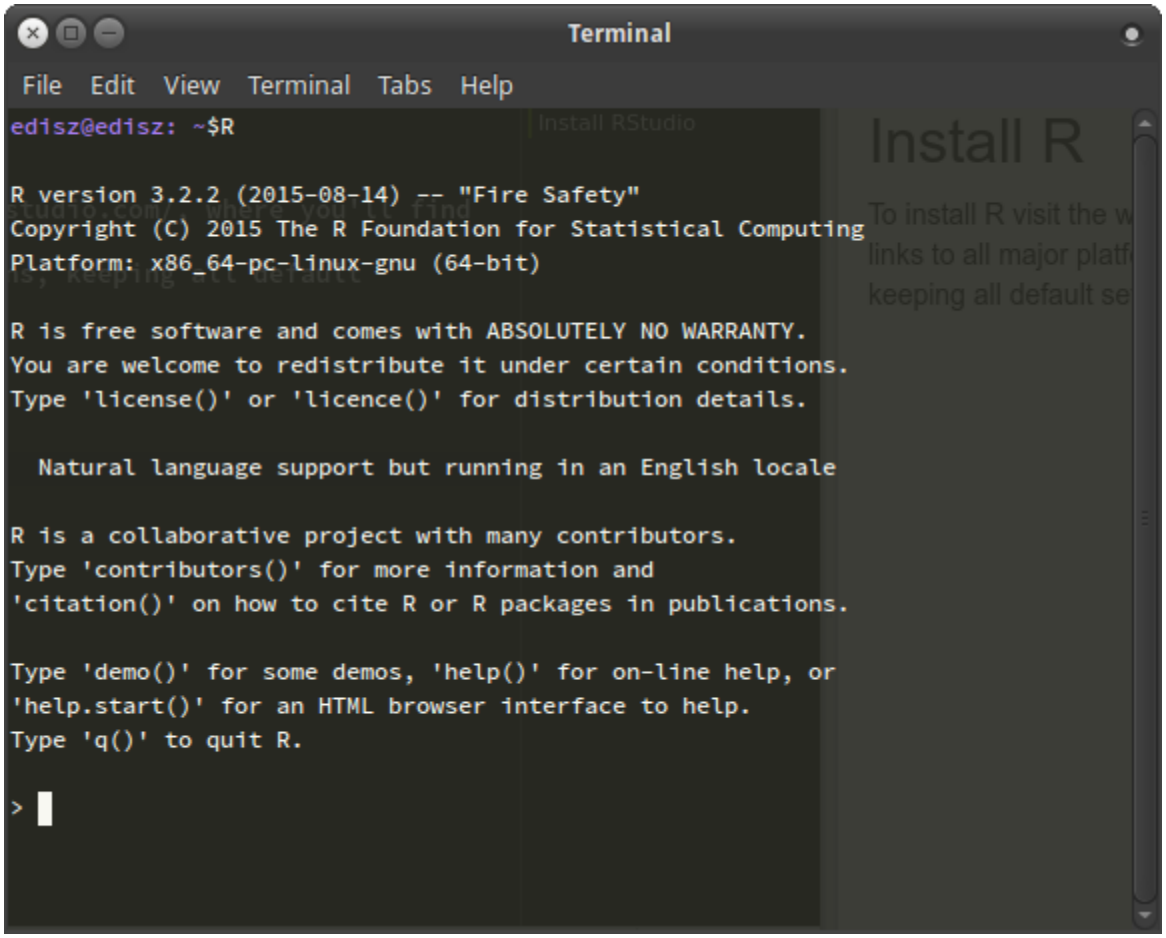


Figura 1 - R na linha de comando em um terminal Linux

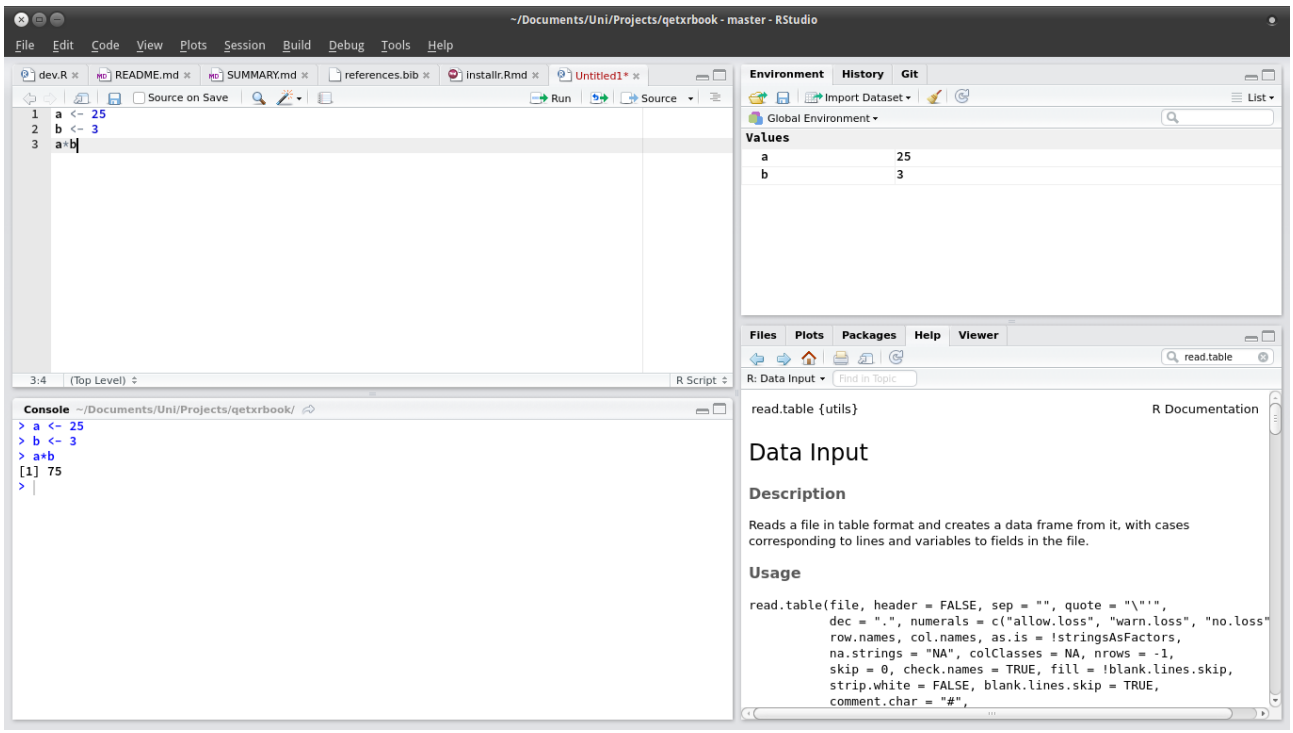


Figura 2 - Interface do RStudio

Note que o RStudio é muito mais organizado e com muito mais funcionalidades, tornando a experiência do usuário muito mais rica e menos cansativa. A seguir algumas funcionalidades que o RStudio traz ao R:

- Autocompletar de funções.
- Sugestão de funções ou objetos ao escrever palavras similares nas linhas de código.
- Abas específicas para determinada ação, tal como gráficos, bases de dados etc.
- Histórico de códigos utilizados.
- Executor de linhas de código sequenciais.
- Identificação de erros antes de executá-los.

2.3 Como instalar o RStudio

1. Acesse a página https://www.rstudio.com/products/rstudio/download/#download dentro do site do RStudio.
2. Procure a seção **All Installers** e escolhe a opção de arquivo adequada ao seu sistema operacional.
3. Faça o download e execute o arquivo de instalação.

2.4 RStudio Cloud

Caso não queira passar por todos esses processos de instalação, você tem a opção de utilizar o RStudio Cloud, que é uma versão "*cloud-based*" do RStudio diretamente no seu navegador.

1. Acesse o site https://rstudio.cloud.
2. Clique em **Get started for free**, caso ainda não tenha uma conta ou faça seu login.
-- Caso não tenha uma conta:
3. Clique na aba **Cloud Free** e em seguida, clique no botão **Sign Up.**
4. Agora é só criar sua conta e fazer seu login.

Note que a opção free do RStudio Cloud tem algumas limitações como 1GB de RAM e 1 CPU por projeto, dependendo da análise, isso pode não ser suficiente para a execução.

2.5 Instalação de Pacotes

Ao instalar o R, apenas as configurações mínimas para sua operação básica são instaladas (pacotes que acompanham a instalação "base"). Para o passo-a-passo que seguiremos em nossa aula, será necessária a instalação de alguns pacotes adicionais.

Para instalar pacotes, usamos o repositório CRAN (https://cloud.r-project.org) ou Bioconductor (http://www.bioconductor.org), que possui pacotes voltados para a Bioinformática.

A instalação de pacotes é feita diretamente através do RStudio. Pacotes contidos no repositório CRAN são instalados utilizando o comando a seguir:

```
install.packages("nome_do_pacote")
```

Baseado neste comando, instale os pacotes:

- webshot
- dplyr
- tidyverse
- ggplot2

Para instalar pacotes a partir do Bioconductor, utilize o comando abaixo:

1. Apenas da primeira vez para verificar se o pacote “BiocManager” está instalado e caso não esteja, instalá-lo:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
```

2. Para instalar os pacotes personalizados:

```
BiocManager::install("nome_do_pacote")
```

Baseado nestes comandos, instale os pacotes:

- DESeq2
- biomaRt

Recomendados fortemente que você faça as instalações antes do início da aula prática (13/07), pois alguns pacotes demoram para ser instalados. Caso tenham algum problema com a instalação de algum pacote, anote o nome do pacote e nos comunique ao início da aula prática para resolvermos seu problema.

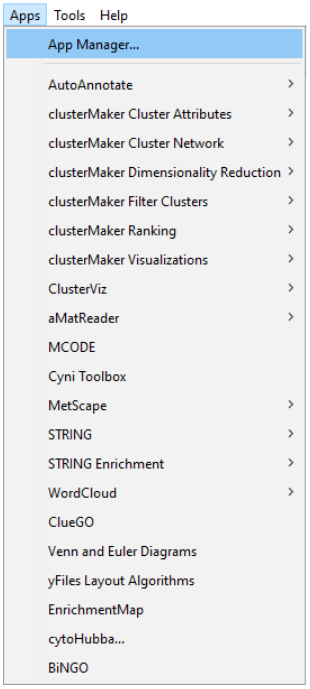
3 Cytoscape

O Cytoscape é um programa para manipulação, visualização e análise topológica de redes. Vamos utilizá-lo na aula do dia 12/07 e recomendamos que a instalação, apesar de simples, seja feita previamente.

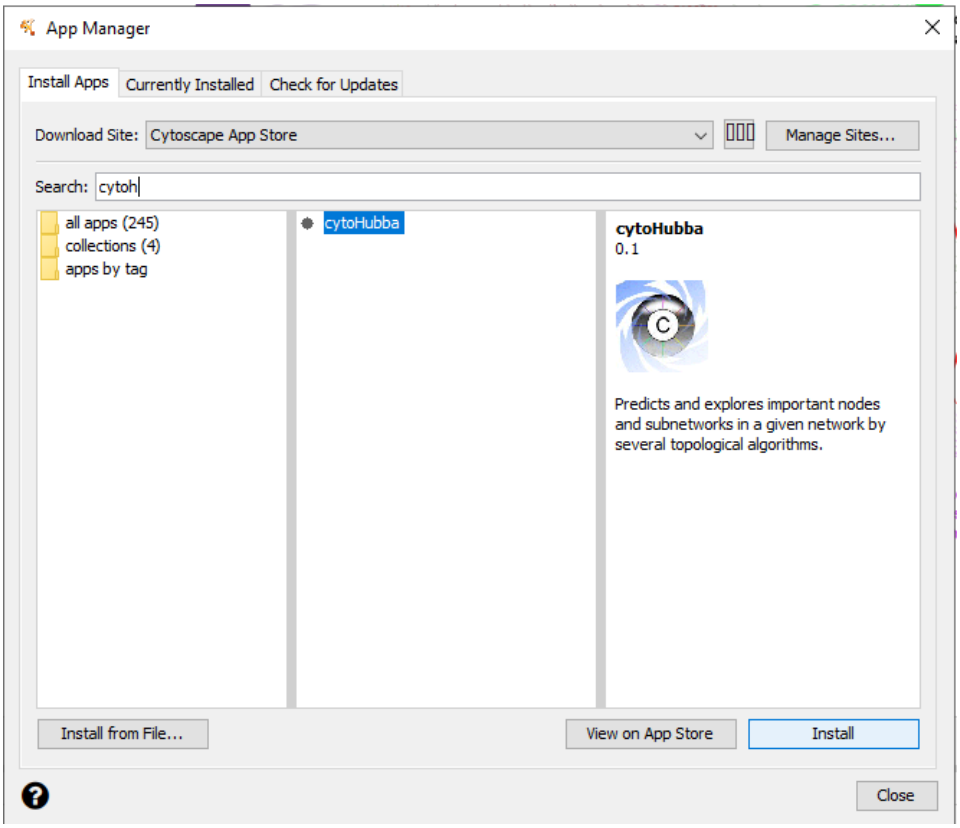
3.1 Instalação Cytoscape

1. Acessar o site do Cytoscape (https://cytoscape.org) e clicar no botão **Download x.x.x** (versão atual).
2. Fazer o download do arquivo adequado para o seu sistema operacional e executar o instalador.

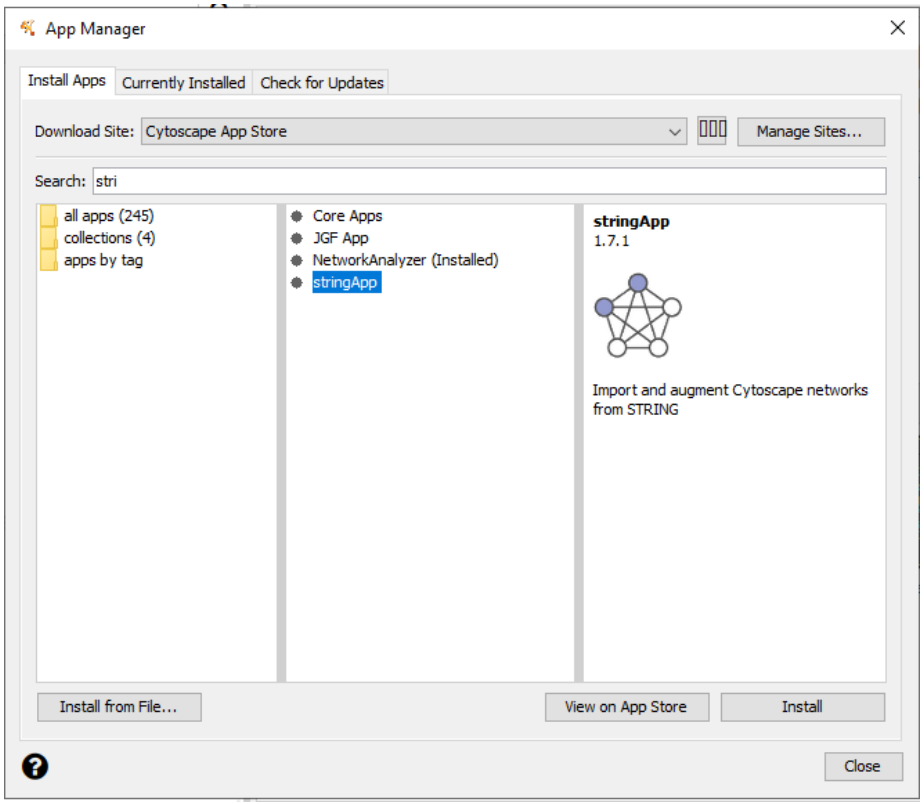
2. Uma vez instalado o Cytoscape, abrir e selecionar o menu **Apps/ App Manager**.



3. Na aba **Install Apps**, digitar no campo **Search** o nome **cytoHubba,** selecionar e clicar em instalar.



4. Repetir o passo 3 procurando e instalando **stringApp.**



5. Fazer o download do arquivo: https://github.com/csbl-inovausp/RNAseq_DESeq2_R_tutorial/raw/main/data/GSE131282_Frontal_cortex_Lesion_control.xlsx.