Guião Laboratório

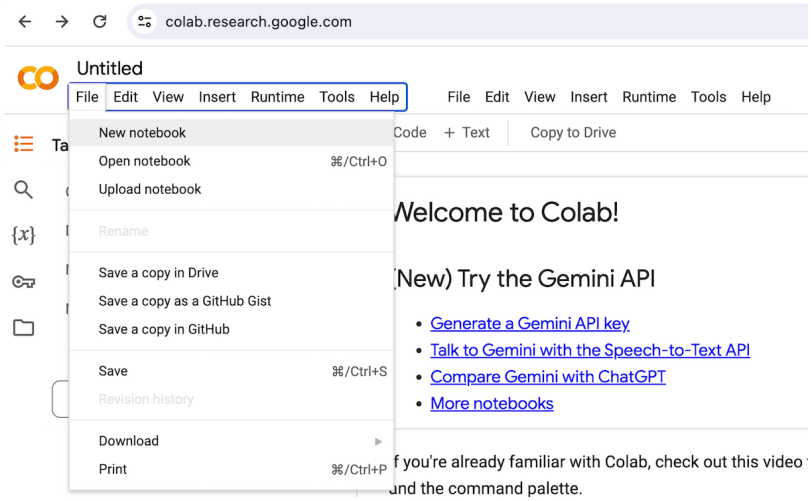
Parte 1 - Google Colab

- O Google Colab é uma ferramenta para quem deseja utilizar a linguagem de programação Python em um ambiente colaborativo e baseado em nuvem.
- O Colab permite armazenar seus notebooks e arquivos de dados no Google Drive
- O Colab já vem pré-configurado com várias bibliotecas populares do ecossistema Python, como NumPy, Pandas, Matplotlib e TensorFlow.

Passos a realizar:

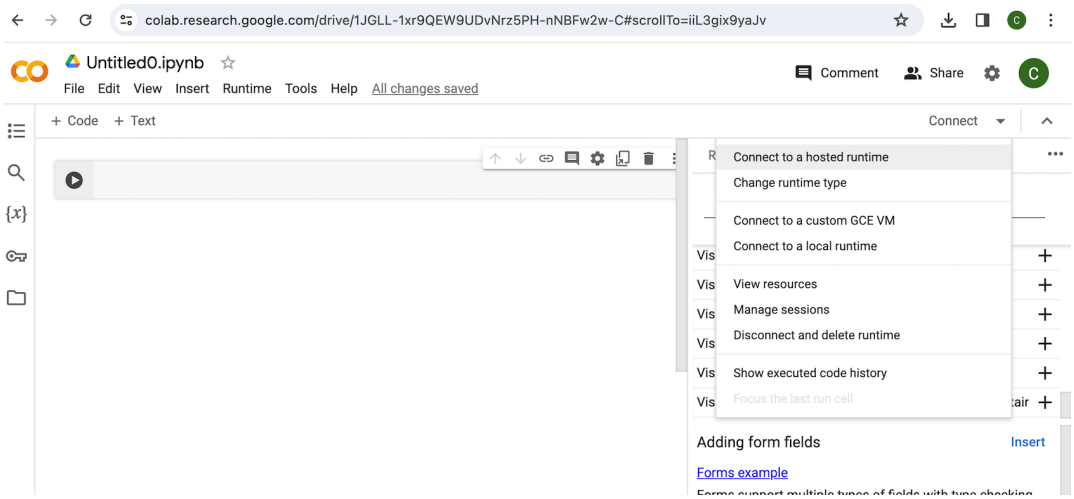
1) Registro no Google Colab: para começar, é preciso ter uma conta no Google. Se já tiver uma conta, basta aceder ao site do Google Colab e fazer login.

2) Criar um novo notebook: Após fazer login, será direcionado para a página inicial do Google Colab. Clique em "Novo notebook" para criar um novo arquivo de código.

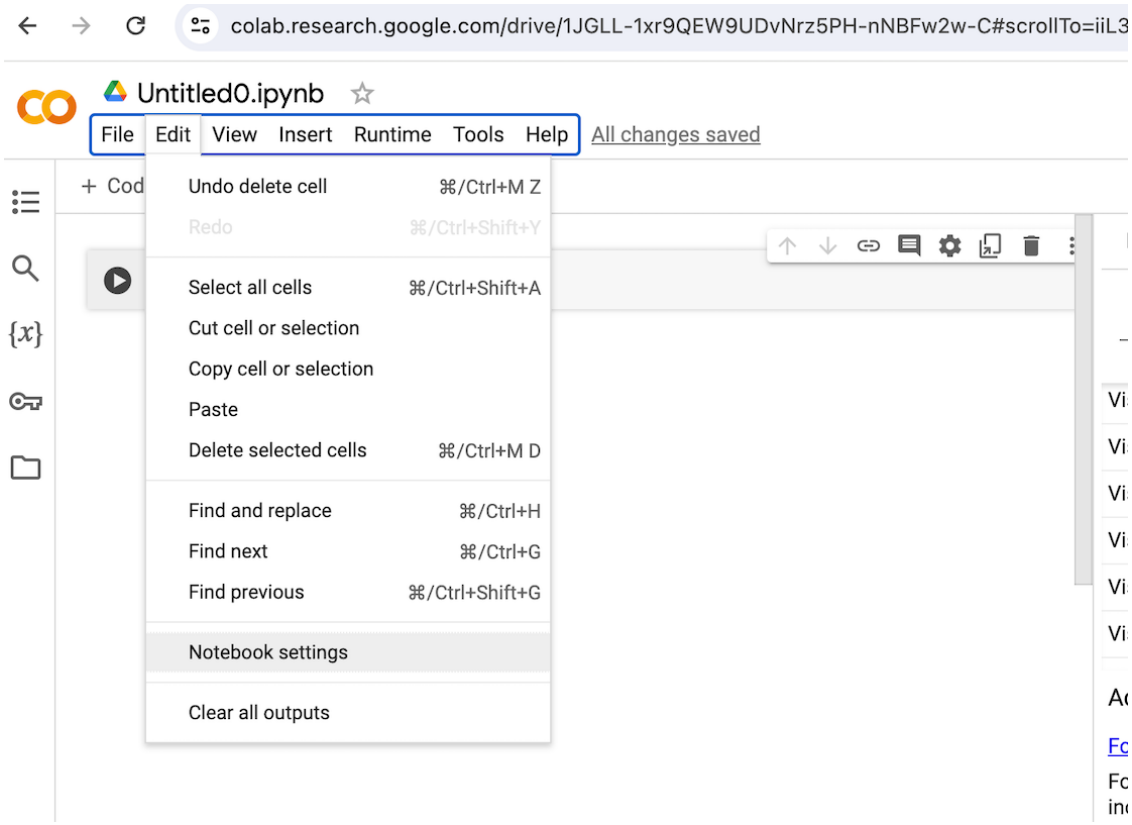


3)Entender a interface do Colab: A interface do Colab é dividida em células, onde você pode escrever e executar seu código. Existem dois tipos principais de células: células de código e células de texto. Pode-se alternar entre esses tipos usando o menu suspenso na parte superior.

4)Conectar a um "hosted runtime"



4. Selecionar o acelerador de hardware.



5) Colocar na google drive os dados “TESTX_H7YRLADXX_S1_L001_R1_001.fastq”
EX: My Drive/Colab Notebooks/Data_Lab1/

6) Realizar mount da Google Drive

```
from google.colab import drive
drive.mount('/content/drive')
```

ou então deixar os ficheiros localmente e carregar a parte de uma diretoria local

```
from google.colab import files
pe1 = files.upload()
pe1_filename, pe1_data = next(iter(pe1.items()))
with open(pe1_filename, 'wb') as f:
    f.write(pe1_data)
```

7) Selecionar a no menu esquerdo a opção “directory” para aceder aos conteúdos do google drive

8) Por exemplo, para listar os ficheiros

```
# After executing the cell above, Drive
# files will be present in "/content/drive/MyDrive/Colab
Notebooks/Data_Lab1".
!ls "/content/drive/MyDrive/Data_Lab1"
```

Parte 2- Ferramenta SPAdes

9) Obter a ferramenta SPAdes

```
#Fix: Use precompiled SPAdes that works with Colab
!wget https://github.com/steventango/colab-
spades/releases/download/v3.15.5/SPAdes-3.15.5-Colab.tar.gz
!tar -xzf SPAdes-3.15.5-Colab.tar.gz
```

10) Executar a ferramenta

```
import subprocess
process = subprocess.run(
    f'python ./bin/spades.py -s "/content/drive/MyDrive/Colab
Notebooks/Data_Lab1/SRR396636.sra_1.fastq" -o
"/content/drive/MyDrive/Colab Notebooks/Data_Lab1/" ',
    capture_output=True,
    text=True,
    shell=True
)

print(process.stdout)
print(process.stderr)
```

11) Durante a execução, analise o ficheiro spaces.log. Quais os parâmetros que foram definidos e o que significa?

12) Qual é o tamanho das reads deste exemplo?

13) O que é o *Read error correction*?

14) Durante a execução, o que é o processo *K-mer Splitting*?

15)Analise/realize o download do ficheiro contigs.fasta

16) E os ficheiros de extensão .gfa? Tenha em conta o exemplo de formato presente neste link. https://github.com/GFA-spec/GFA-spec/blob/master/GFA1.md
EXEMPLO:

```
H       VN:Z:1.0
S       11      ACCTT
S       12      TCAAGG
S       13      CTTGATT
L       11      +       12      -       4M
L       12      -       13      +       5M
L       11      +       13      +       3M
P       14      11+,12-,13+     4M,5M
```

The resulting path is:

```
11 ACCTT
12  CCTTGA
13   CTTGATT
14 ACCTTGATT
```

Parte 3 – ferramenta flye

17)Instalar o gestor de pacotes Conda

```
!pip install -q condacolab
import condacolab
condacolab.install()
```

18) Verificar os canais que são estão disponíveis.

```
!conda config --show channels
```

19) Se não tiveres os seguintes 3 , adicionar

```
!conda config --add channels defaults
!conda config --add channels bioconda
!conda config --add channels conda-forge
!conda config --set channel_priority strict
```

20) instalar a ferramenta flye como package

```
!conda install flye
```

21) Obter dados de https://github.com/fenderglass/Flye/blob/flye/docs/USAGE.md#examples, como por exemplo:

```
!wget https://zenodo.org/record/1172816/files/E.coli_PacBio_40x.fasta
```

22) Executar o comando

```
!flye --pacbio-raw "E.coli_PacBio_40x.fasta" -o "/content/Res/"
```

23) Qual a diferença entre os formatos FASTA e FASTAQ?

24) Analise flye.log

25) Analise o ficheiro assembly.fasta e assembly.info. Quantos contigs estão presentes?

Parte 4- Implementações no contexto do genome assembly

26) Implemente um programa que tenha como input uma coleção de k-mers e como output o **De BruijnGraph. Exemplos:**

a) Input:
GAGG CAGG GGGG GGGA CAGG AGGG GGAG

Output:
GAG: AGG
CAG: AGG AGG
GGG: GGG GGA
AGG: GGG

GGA: GAG

b)
Input:
GCAAG CAGCT TGACG Output:
GCAA: CAAG
CAGC: AGCT
TGAC: GACG

27) Implemente o algoritmo DFS sobre o grafo gerado na alínea anterior

28) Implemente o algoritmo Hierholzer para procurar um caminho de Euler