Workshop UAustral: Práctico 1

# Ensamble de genomas, anotación y búsqueda de homólogos

Dr. Jonathan Maldonado

Researcherid J-6429-2012

PSBL
PLANT SYSTEMS BIOLOGY LAB

iBio
Millennium Institute for Integrative Biology

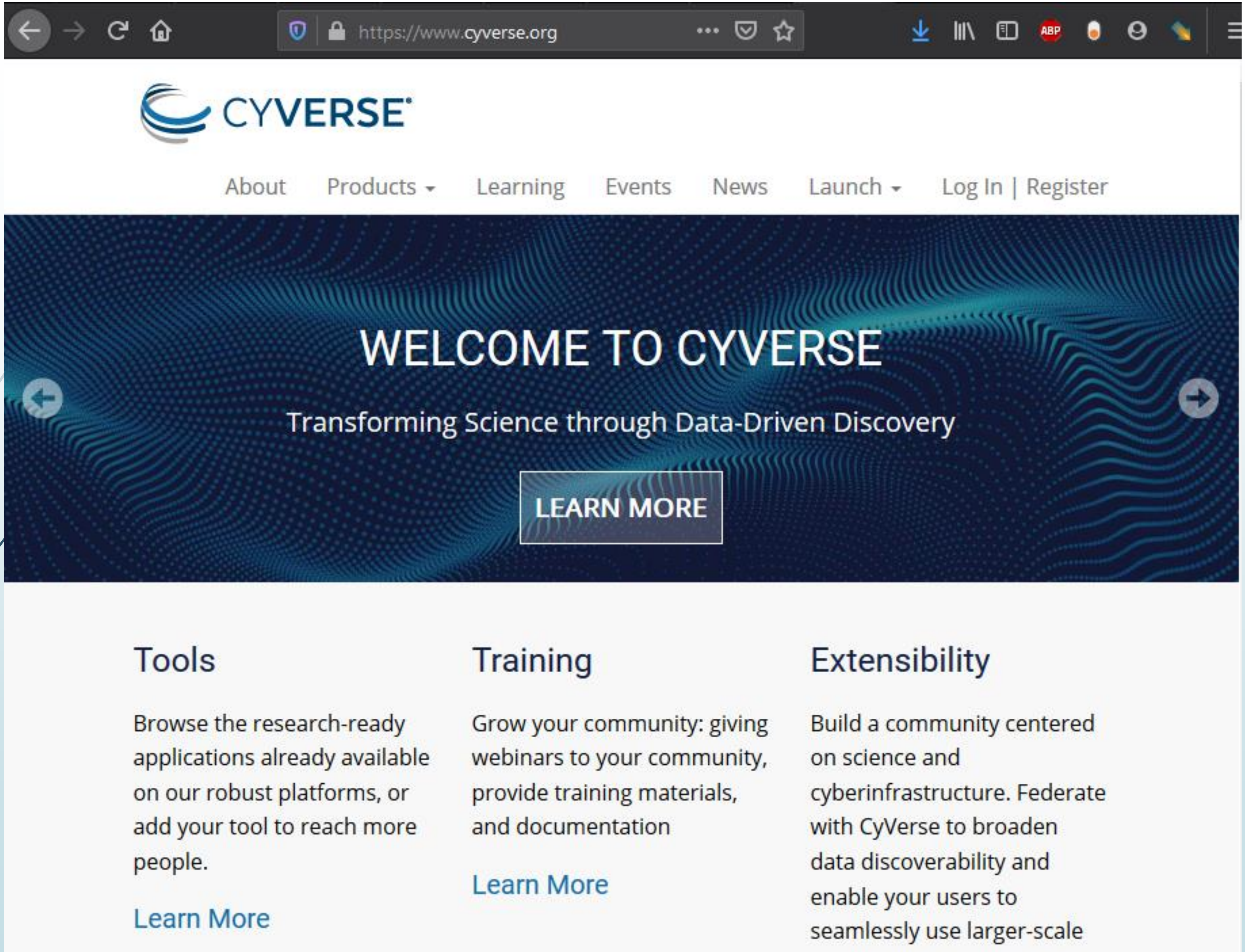
PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

LIBERTAS CAPITUR
UNIVERSIDAD AUSTRAL DE CHILE

jomaldon@gmail.com

jomaldon

@JM_bioinfo



# WELCOME TO CYVERSE

Transforming Science through Data-Driven Discovery

## Tools

Browse the research-ready applications already available on our robust platforms, or add your tool to reach more people.

Learn More

## Training

Grow your community: giving webinars to your community, provide training materials, and documentation

Learn More

## Extensibility

Build a community centered on science and cyberinfrastructure. Federate with CyVerse to broaden data discoverability and enable your users to seamlessly use larger-scale

# History and Context

***iPlant 2008***
*Empowering a*
*New Plant Biology*

***iPlant 2013***
Cyberinfrastructure
for Life Science

***CyVerse 2016***
Transforming Science
Through Data-Driven
Discovery

THE UNIVERSITY OF ARIZONA

TACC
TEXAS ADVANCED COMPUTING CENTER

UNCW

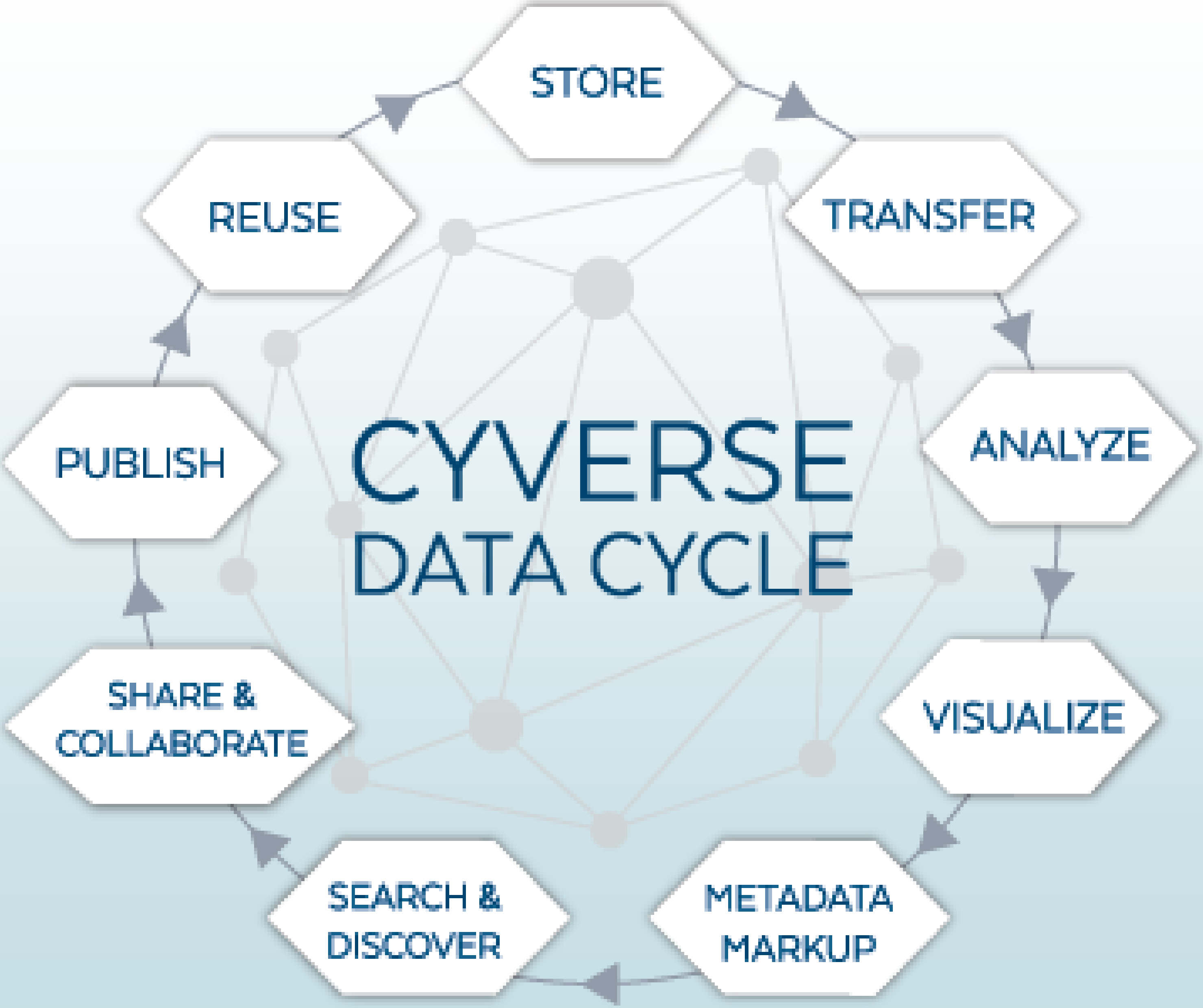
CSH Cold Spring Harbor Laboratory

NSF

**Plant Science Cyberinfrastructure Collaborative**
A "new type of organization" that is "community-driven" uniting "biologists, computer and information scientists and experts from other disciplines working in an integrated team" to provide "computational and cyberinfrastructure capabilities and expertise that are capable of handling large and heterogeneous plant biology data sets"

~ $100m direct NSF investment over 10 years

Currently working to sustain its successes beyond 2018



CYVERSE DATA CYCLE

STORE

TRANSFER

ANALYZE

VISUALIZE

METADATA MARKUP

SEARCH & DISCOVER

SHARE & COLLABORATE

PUBLISH

REUSE

# Tipos de archivos que usaremos

- FASTA
- FASTQ

# Archivos FASTA

```
cat SRR1811524_1M.fasta | head -n 12
```

```
>SRR1811524.2 HWI-ST615:645:H7P8JADXX:1:1101:1414:2351/1
CTTAGCTTGTCGAAACAACATCCAAGAGTTCATGCTTAACCAACTCGGAAT
>SRR1811524.3 HWI-ST615:645:H7P8JADXX:1:1101:1416:2377/1
CCGACTTGTTCTGCGTATCTAGACTAACTCTGGTTTCCCTCTTAGACAAAG
>SRR1811524.4 HWI-ST615:645:H7P8JADXX:1:1101:1479:2437/1
CAAACACTTTCCAGGATACCACTCAATCTCCGTCCCGAGGGCCTTCTCAAG
```

```
cat SRR1811524_1M.qual | head -n 12
```

```
>SRR1811524.2 HWI-ST615:645:H7P8JADXX:1:1101:1414:2351/1
5 10 15 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40 40 404 04 04 040 40 40 40 40 40 40 30 20 10 5
>SRR1811524.3 HWI-ST615:645:H7P8JADXX:1:1101:1416:2377/1
5 10 15 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40 40 404 04 04 040 40 40 40 40 40 40 30 20 10 5
>SRR1811524.4 HWI-ST615:645:H7P8JADXX:1:1101:1479:2437/1
5 10 15 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40 40 404 04 04 040 40 40 40 40 40 40 30 20 10 5
```

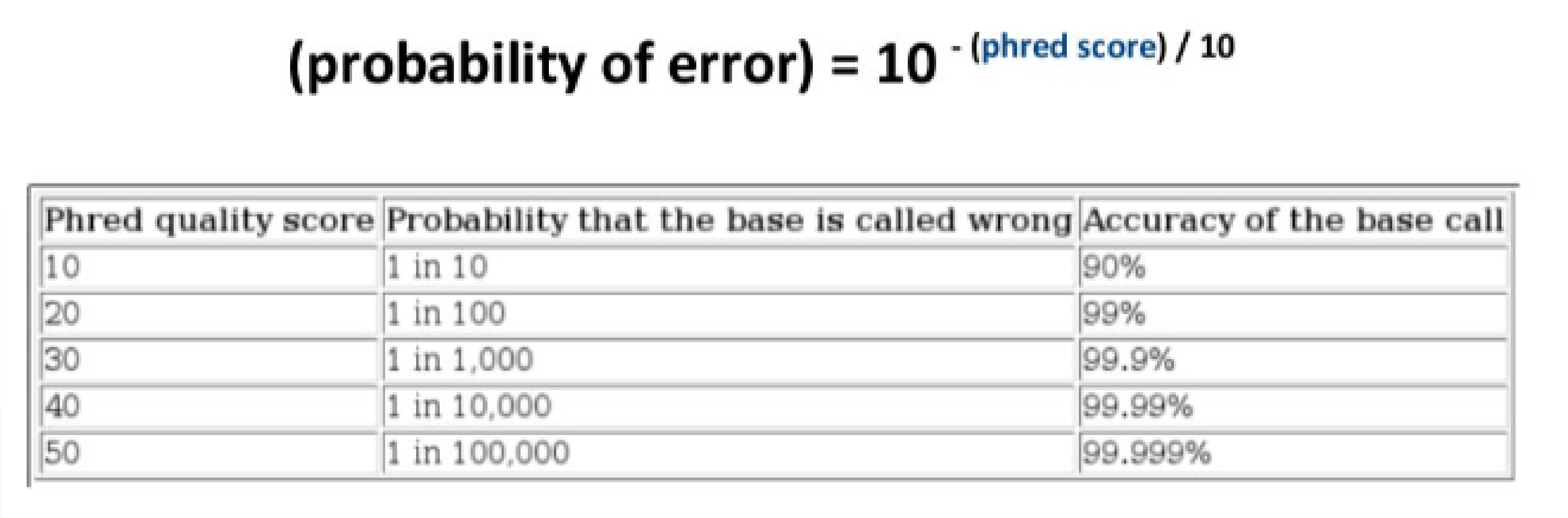
# Archivos FASTA

```
cat SRR1811524_1M.fasta | head -n 12
```

```
                  >SRR1811524.2 HWI-ST615:645:H7P8JADXX:1:1101:1414:2351/1
                  CTTAGCTTGTCGAAACAACATCCAAGAGTTCATGCTTAACCAACTCGGAAT
identificador →   >SRR1811524.3 HWI-ST615:645:H7P8JADXX:1:1101:1416:2377/1
    secuencia →   CCGACTTGTTCTGCGTATCTAGACTAACTCTGGTTTCCCTCTTAGACAAAG
                  >SRR1811524.4 HWI-ST615:645:H7P8JADXX:1:1101:1479:2437/1
                  CAAACACTTTCCAGGATACCACTCAATCTCCGTCCCGAGGGCCTTCTCAAG
```

```
cat SRR1811524_1M.qual | head -n 12
```

```
                  >SRR1811524.2 HWI-ST615:645:H7P8JADXX:1:1101:1414:2351/1
                  5 10 15 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40 40 404 04 04 040 40 40 40 40 40 40 30 20 10 5
identificador →   >SRR1811524.3 HWI-ST615:645:H7P8JADXX:1:1101:1416:2377/1
      calidad →   5 10 15 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40 40 404 04 04 040 40 40 40 40 40 40 30 20 10 5
                  >SRR1811524.4 HWI-ST615:645:H7P8JADXX:1:1101:1479:2437/1
                  5 10 15 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40 40 404 04 04 040 40 40 40 40 40 40 30 20 10 5
```



$$(\text{probability of error}) = 10^{-(\text{phred score})/10}$$

| Phred quality score | Probability that the base is called wrong | Accuracy of the base call |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |

```
cat SRR1811524_1M.qual | head -n 12
```

```
                >SRR1811524.2 HWI-ST615:645:H7P8JADXX:1:1101:1414:2351/1
                5 10 15 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40 40 404 04 04 040 40 40 40 40 40 40 30 20 10 5
identificador → >SRR1811524.3 HWI-ST615:645:H7P8JADXX:1:1101:1416:2377/1
      calidad → 5 10 15 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40 40 404 04 04 040 40 40 40 40 40 40 30 20 10 5
                >SRR1811524.4 HWI-ST615:645:H7P8JADXX:1:1101:1479:2437/1
                5 10 15 20 30 40 40 40 40 40 40 40 40 40 40 40 40 40 40 404 04 04 040 40 40 40 40 40 40 30 20 10 5
```

**Codificación**

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
0 1 2 3 ..                                                                                  93
```

# Archivos FASTQ

```
zcat SRR1811524_1M.fastq.gz | head -n 12
```

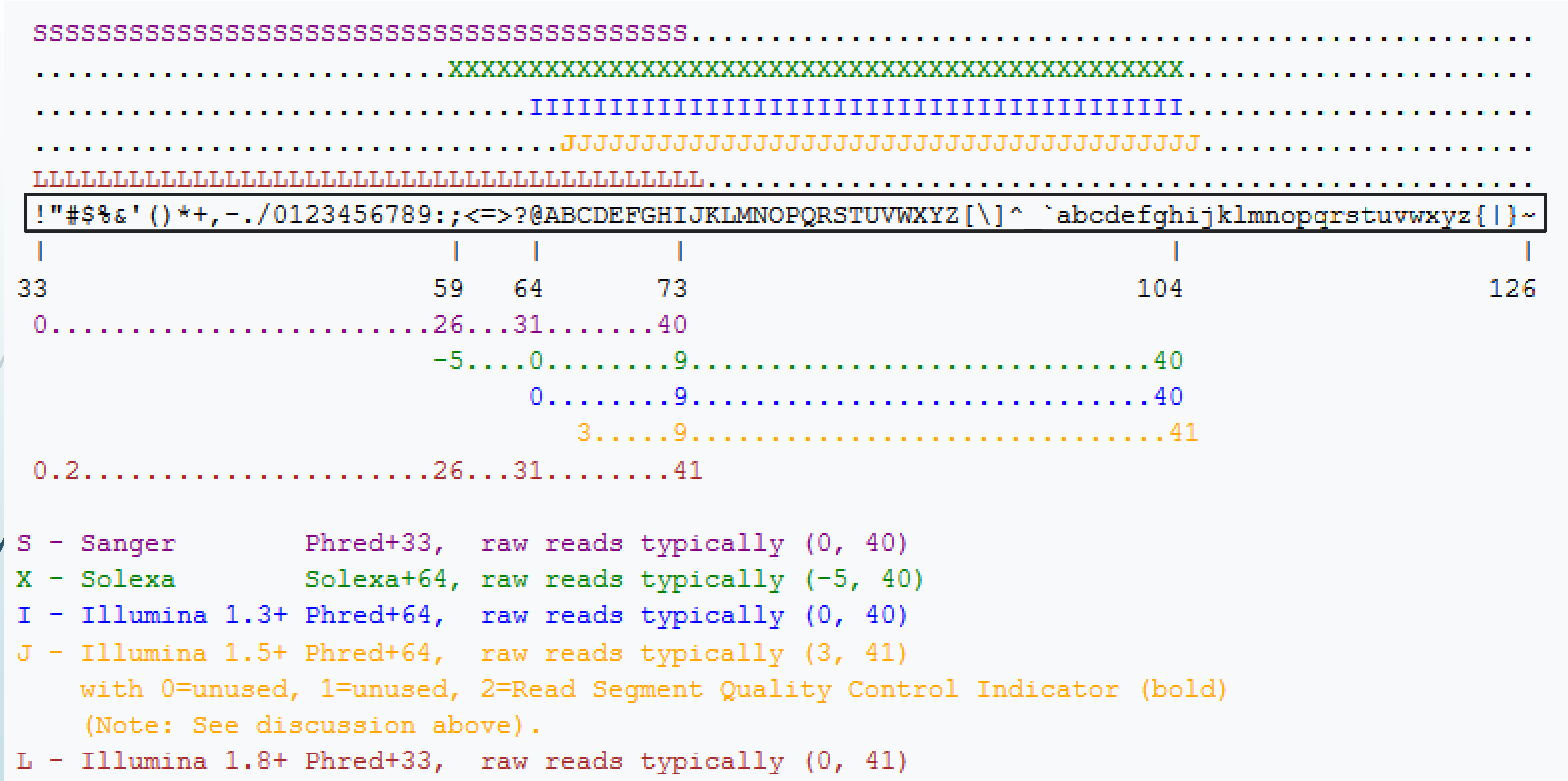
```
@SRR1811524.1 HWI-ST615:645:H7P8JADXX:1:1101:1015:2373/1
NAACACTCAACTTGAACGCAATGAACCAGGCACCAAACCCATGGCACGTGT
+
#4BDFFFFHHHHHIJJJJJJJJJJJJJIJJJJJIJJHIJJIJJIIGIJJAH
@SRR1811524.2 HWI-ST615:645:H7P8JADXX:1:1101:1414:2351/1
CTTAGCTTGTCGAAACAACATCCAAGAGTTCATGCTTAACCAACTCGGAAT
+
CCCFFFFFHFHHAHHIIIIEHIGIIHIIIIGEEIIIIEGGIGHIEFHGIIG
@SRR1811524.3 HWI-ST615:645:H7P8JADXX:1:1101:1416:2377/1
CCGACTTGTTCTGCGTATCTAGACTAACTCTGGTTTCCCTCTTAGACAAAG
+
@??DD?DDHFDF?GEEHIIJJEHEHIIAFHFHCGGGCBF@HGI@?FBDGED
...
```

# Archivos FASTQ

```
zcat SRR1811524_1M.fastq.gz | head -n 12
```

```
                  @SRR1811524.1 HWI-ST615:645:H7P8JADXX:1:1101:1015:2373/1
                  NAACACTCAACTTGAACGCAATGAACCAGGCACCAAACCCATGGCACGTGT
                  +
                  #4BDFFFFHHHHHIJJJJJJJJJJJJJIJJJJJIJJHIJJIJJIIGIJJAH
identificador →   @SRR1811524.2 HWI-ST615:645:H7P8JADXX:1:1101:1414:2351/1
    secuencia →   CTTAGCTTGTCGAAACAACATCCAAGAGTTCATGCTTAACCAACTCGGAAT
    separador →   +
      calidad →   CCCFFFFFHFHHAHHIIIIEHIGIIHIIIIGEEIIIIEGGIGHIEFHGIIG
                  @SRR1811524.3 HWI-ST615:645:H7P8JADXX:1:1101:1416:2377/1
                  CCGACTTGTTCTGCGTATCTAGACTAACTCTGGTTTCCCTCTTAGACAAAG
                  +
                  @??DD?DDHFDF?GEEHIIJJEHEHIIAFHFHCGGGCBF@HGI@?FBDGED
                  ...
```

# Codificación de calidad



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                 0........9.............................40
                                    3.....9..............................41
  0.2......................26...31........41

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

https://en.wikipedia.org/wiki/FASTQ_format

*E. coli*
https://www.biostars.org/p/160377/
https://www.ebi.ac.uk/ena/data/view/ERX008638
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR022/ERR022075/

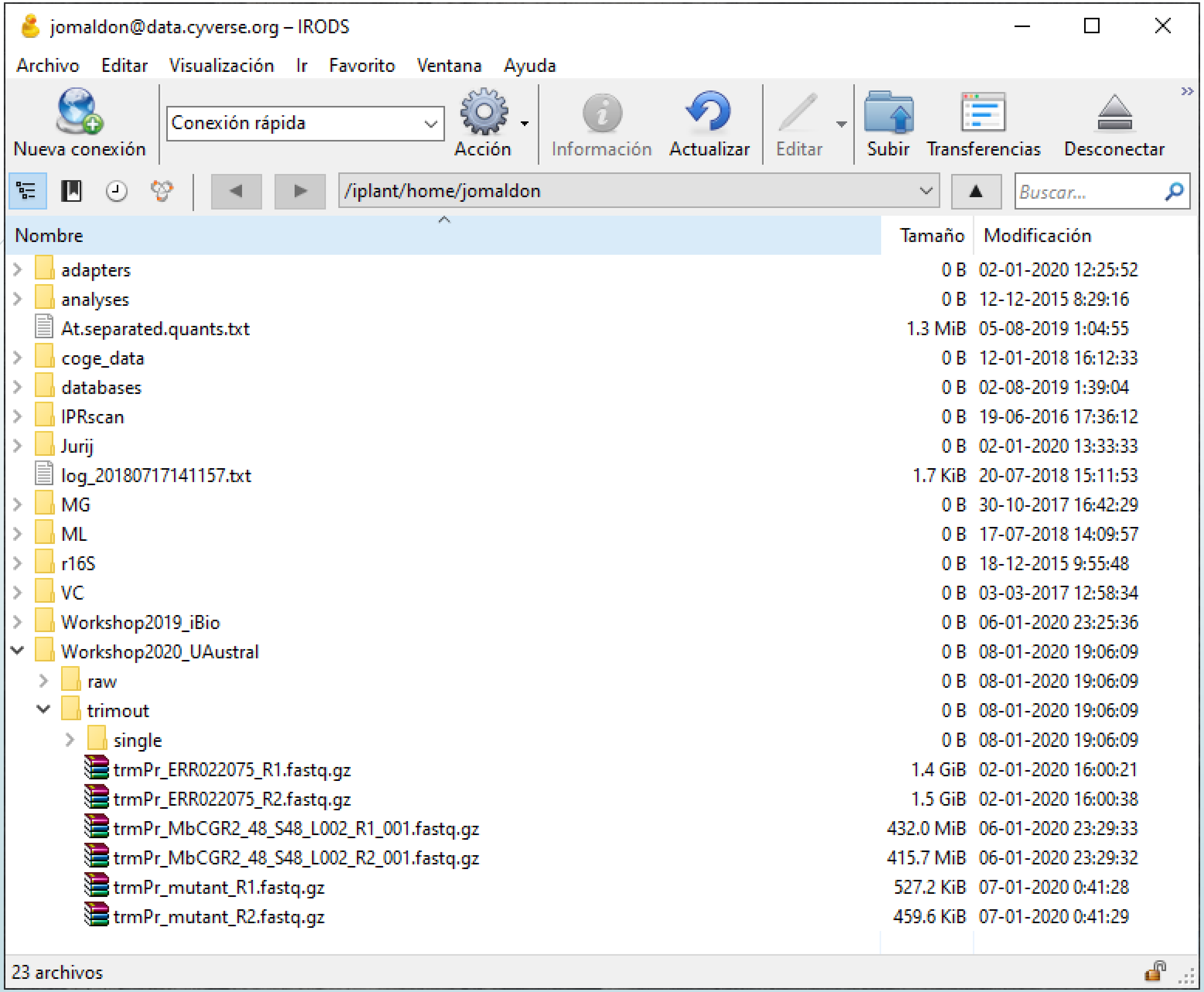
Microbacterium sp. CGR2
https://www.ncbi.nlm.nih.gov/genome/14190?genome_assembly_id=410147

ABRPI-Training
https://sepsis-omics.github.io/tutorials/modules/data-dna/

Cyberduck is a libre server and cloud storage browser for Mac and Windows with support for **FTP**, **SFTP**, **WebDAV**, **Amazon S3**, **OpenStack Swift**, **Backblaze B2**, **Microsoft Azure & OneDrive**, **Google Drive** and **Dropbox**.

https://wiki.cyverse.org/wiki/display/DS/Using+Cyberduck+for+Uploading+and+Downloading+to+the+Data+Store

https://cyberduck.io/download/



/iplant/home/jomaldon

| Nombre | Tamaño | Modificación |
|---|---|---|
| adapters | 0 B | 02-01-2020 12:25:52 |
| analyses | 0 B | 12-12-2015 8:29:16 |
| At.separated.quants.txt | 1.3 MiB | 05-08-2019 1:04:55 |
| coge_data | 0 B | 12-01-2018 16:12:33 |
| databases | 0 B | 02-08-2019 1:39:04 |
| IPRscan | 0 B | 19-06-2016 17:36:12 |
| Jurij | 0 B | 02-01-2020 13:33:33 |
| log_20180717141157.txt | 1.7 KiB | 20-07-2018 15:11:53 |
| MG | 0 B | 30-10-2017 16:42:29 |
| ML | 0 B | 17-07-2018 14:09:57 |
| r16S | 0 B | 18-12-2015 9:55:48 |
| VC | 0 B | 03-03-2017 12:58:34 |
| Workshop2019_iBio | 0 B | 06-01-2020 23:25:36 |
| Workshop2020_UAustral | 0 B | 08-01-2020 19:06:09 |
| raw | 0 B | 08-01-2020 19:06:09 |
| trimout | 0 B | 08-01-2020 19:06:09 |
| single | 0 B | 08-01-2020 19:06:09 |
| trmPr_ERR022075_R1.fastq.gz | 1.4 GiB | 02-01-2020 16:00:21 |
| trmPr_ERR022075_R2.fastq.gz | 1.5 GiB | 02-01-2020 16:00:38 |
| trmPr_MbCGR2_48_S48_L002_R1_001.fastq.gz | 432.0 MiB | 06-01-2020 23:29:33 |
| trmPr_MbCGR2_48_S48_L002_R2_001.fastq.gz | 415.7 MiB | 06-01-2020 23:29:32 |
| trmPr_mutant_R1.fastq.gz | 527.2 KiB | 07-01-2020 0:41:28 |
| trmPr_mutant_R2.fastq.gz | 459.6 KiB | 07-01-2020 0:41:29 |

**Los datos**

El set de reads que usaremos son del genoma de una bacteria imaginaria de *Staphylococcus aureus* secuenciados en un instrumento Illumina. Corresponden a un subconjunto del genoma verdadero.

•Los archivos que ensamblaremos son mutant_R1.fastq y mutant_R2.fastq.

•(No es necesario un genoma de referencia).

•Son paired-end.

•Cada read tiene 150 pares de bases de largo.

•El número de bases secuenciadas equivale a un 19x de cobertura de la cepa original lo cual es muy bajo pero útil para el ejemplo.

# FIN