

I TALLER DE BIOINFORMÁTICA: DISEÑO

Y ANÁLISIS DE PRIMERS E INFERENCIA FILOGENÉTICA MOLECULAR

Diciembre 16 - 19 de 2019. Chiclayo, Perú





















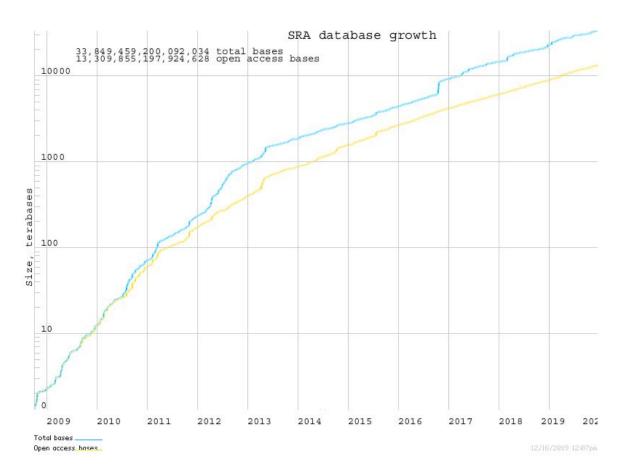
# ¿ Qué es la Bioinformática?

Fernanda Luz P Costa

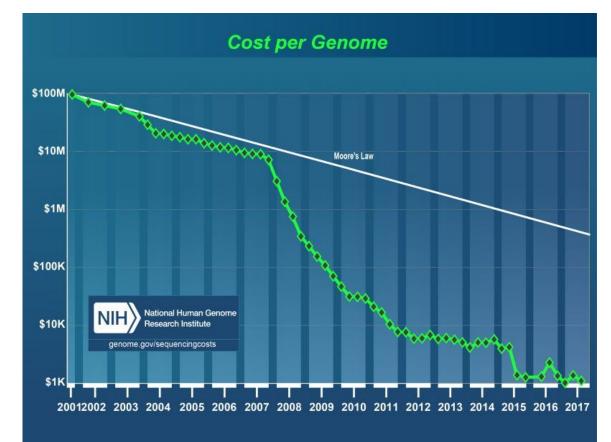




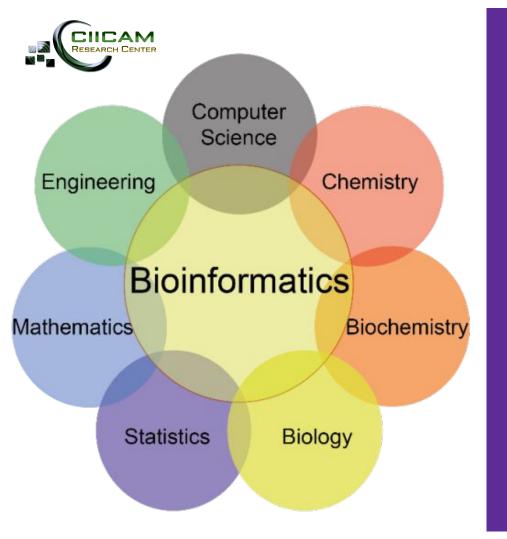
"Un libro escaneado normal de 200 a 300 páginas ocuparía 1 megabyte. En nuestro laboratorio tenemos un espacio de almacenamiento de 100 terabytes, serían 100 millones de libros." João Setubal



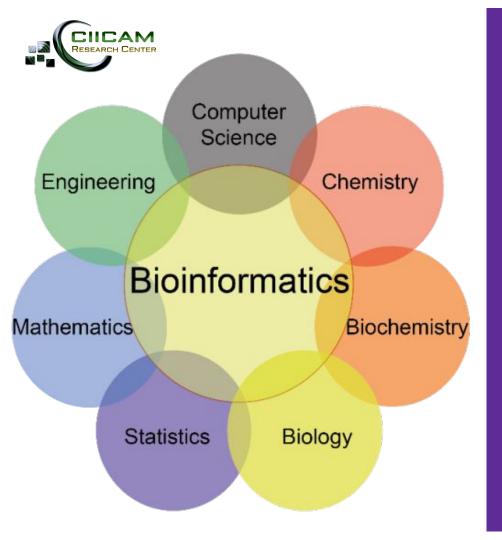




A principios de la década de 2000, "para secuenciar el genoma de una bacteria, el costo era de \$ 1 por base, es decir, de \$ 5 millones a \$ 5 millones de bases. Hoy en día, esta misma secuencia cuesta 100 dólares para ser secuenciada", ejemplifica.



La bioinformática es conceptualizar la biología en términos de macromoléculas (en el sentido de la química física) y luego aplica técnicas de "informática" (derivadas de disciplinas como las matemáticas aplicadas, la informática y la estadística) para comprender y organizar la información asociada con estas moléculas, a gran escala



"Es un matrimonio feliz entre las áreas." João Setubal



## Tres partes de la bioinformática

Creación de bases de datos

La organización, el almacenamiento y la gestión de los conjuntos de datos biológicos

2. Desarrollo de algoritmos y estadísticas

Desarrollo de herramientas y recursos para determinar la relación entre los miembros de grandes conjuntos de datos

3. Análisis de datos e interpretación

El uso apropiado de las partes 1 y 2 para analizar los datos e interpretar los resultados de una manera biológicamente significativa



# Tipos de datos

Estructuras macromoleculares, secuencias del genoma y los resultados de experimentos de genómica funcional (por ejemplo, datos de expresión). La información adicional incluye el texto de artículos científicos y "datos de relación" de rutas metabólicas, árboles de taxonomía y redes de interacción proteína-proteína.

Luscombe NM. et al., 2001



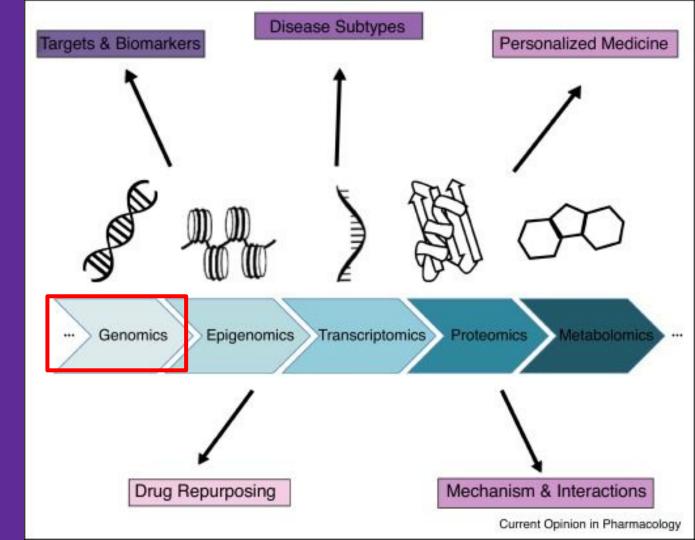
Biologia Molecular

+

Big Data

=

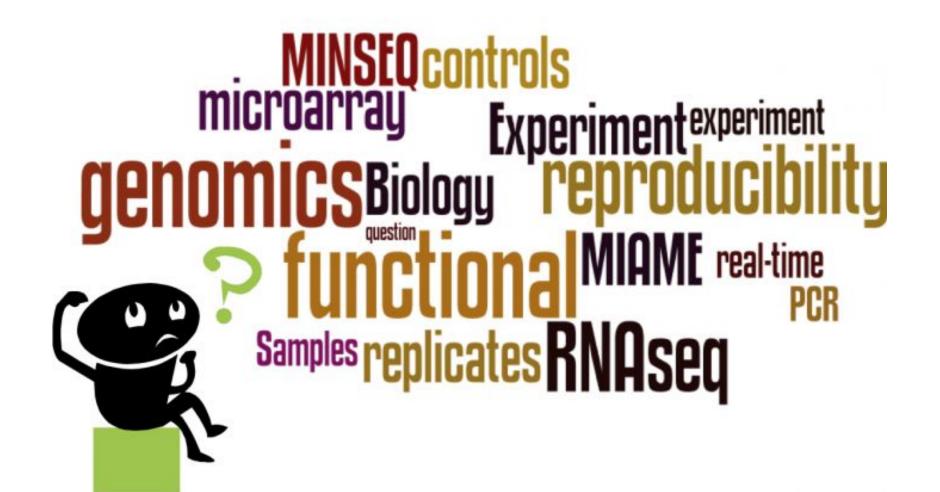
**OMICS** 



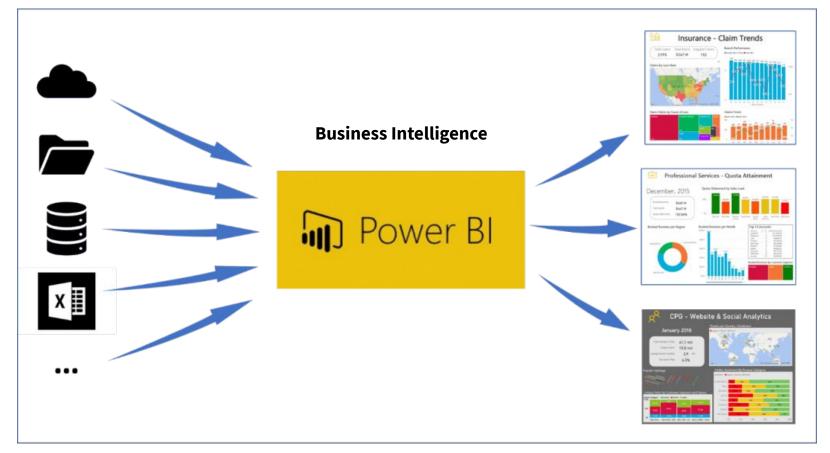


# El proceso de Análisis de Datos

Fernanda Luz P Costa

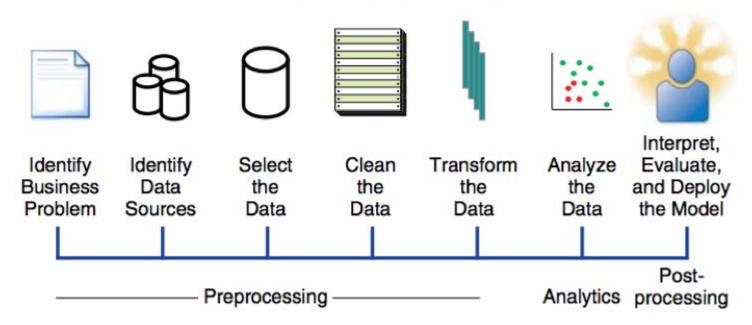








#### Overview of the Analytics Process Model

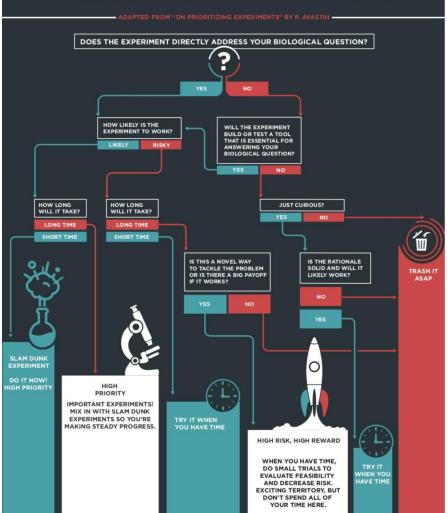


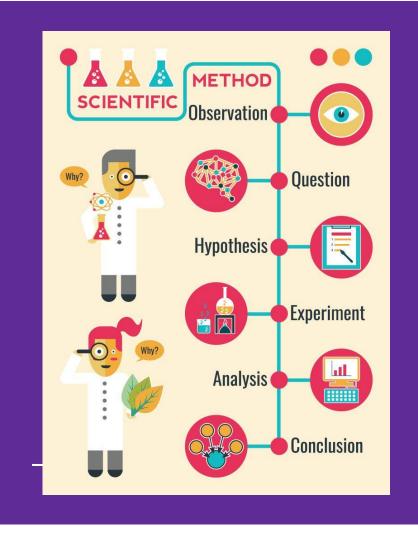




¿Cómo aplicar los métodos de inteligencia empresarial a la bioinformática?

#### WHAT EXPERIMENT TO DO NEXT







# Conceptos básicos de secuenciación masiva

Fernanda Luz P Costa







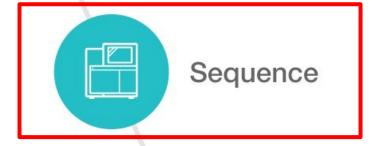
Sequence



Analyze





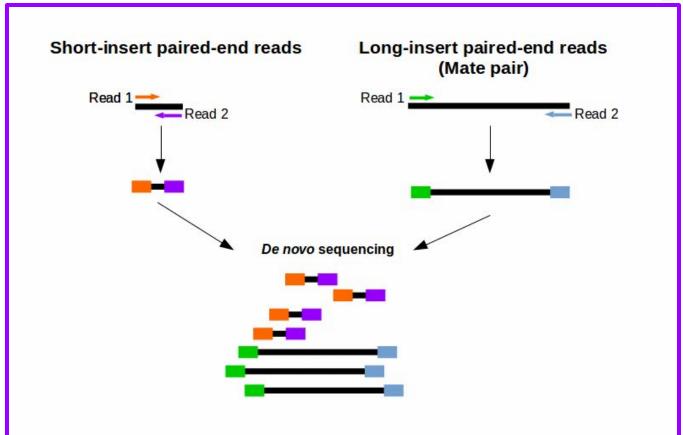




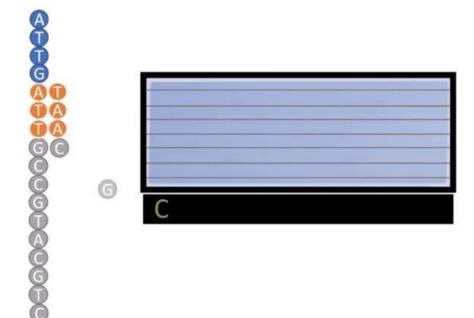


#### Para recordar ...





# 4 Principales pasos Secuenciación



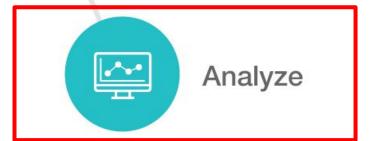






Sequence







```
CBBFFFFFHHHHHGIJIFHIJJJJIIJJJJJJJJJJJJJGGEIIIJJJJJJJJJJ
@SRR2145247.124 SBSUSER:379:C4C9GACXX:6:1101:17421:2199 length=101
GAGAGAGTTAGATTGTATGTTCGTGGAACAGTCCTCGGCTACAAGAGGTC
+SRR2145247.124 SBSUSER:379:C4C9GACXX:6:1101:17421:2199 length=101
CCCFFFFDHHGHHJIHIJJJJJJJHIJJJJJJJJJJJJJGIJIHIJJJJJFH
@SRR2145247.160 SBSUSER:379:C4C9GACXX:6:1101:1590:2381 length=101
CTTGCCTTGACAATTGGCAAGCTATCAGAAAGCTTCATTCGACAAAGCAT
+SRR2145247.160 SBSUSER:379:C4C9GACXX:6:1101:1590:2381 length=101
@BCFFEDDHHGHGIHJJJIIIIGJGIIIJJJJGIGIIJIIJJJJJJJIJIIGG
@SRR2145247.169 SBSUSER:379:C4C9GACXX:6:1101:2075:2333 length=101
AACAATCTGACCAATTCAATGAAGGCAATTGGTGTTGACATATTGACTGG
+SRR2145247.169 SBSUSER:379:C4C9GACXX:6:1101:2075:2333 length=101
CCCFFFFFGHHGHIGIJJIEHGIIJIJIIII@:CG@GGIIJGGGEHGEHB
CCGAACCAAGGAAACCCTTCAGAACTTGTGCTTGGAGTTAGCAACCATGG
+SRR2145247.261 SBSUSER:379:C4C9GACXX:6:1101:7948:2392 length=101
@SRR2145247.268 SBSUSER:379:C4C9GACXX:6:1101:8452:2352 length=101
AACGTGGGCTTGTGACCCGTGTGGGGACAATAAGTTGGTGGAACACTGGC
+SRR2145247.268 SBSUSER:379:C4C9GACXX:6:1101:8452:2352 length=101
@SRR2145247.377 SBSUSER:379:C4C9GACXX:6:1101:16126:2356 length=101
CTTGAAACCGCCATGCGTCCTACTCAGATTAGGACTCATGCTATAAGGAA
+SRR2145247.377 SBSUSER:379:C4C9GACXX:6:1101:16126:2356 length=101
CCCFFFFFHHDHGGIJJCGHIJJIGHH@FIGGGIIJIIIJIJGIEFGBFH
@SRR2145247.410 SBSUSER:379:C4C9GACXX:6:1101:17766:2265 length=101
CCTTGACATTCCAGCTACTAAATCACATTCTACTTTCAAAGTTGTGAAAC
+SRR2145247.410 SBSUSER:379:C4C9GACXX:6:1101:17766:2265 length=101
@SRR2145247.418 SBSUSER:379:C4C9GACXX:6:1101:18324:2283 length=101
GCTTACTGTACAGAGGAGAGCGACTGAGAATAGCGCTGTAAGAAAGGTTA
+SRR2145247.418 SBSUSER:379:C4C9GACXX:6:1101:18324:2283 length=101
```

+SRR2145247.86 SBSUSER:379:C4C9GACXX:6:1101:13109:2200 length=101

GTTTTATCCTCCGAAAGTCTCAACCTTTTTCTTATCCTCAACAAGGAGAA

#### Output de NGS

**Archivo FastQ** 

??

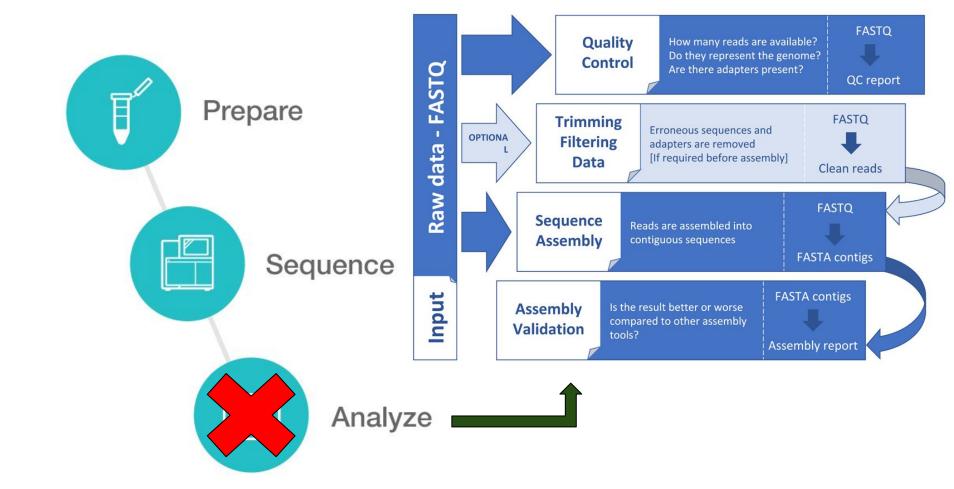


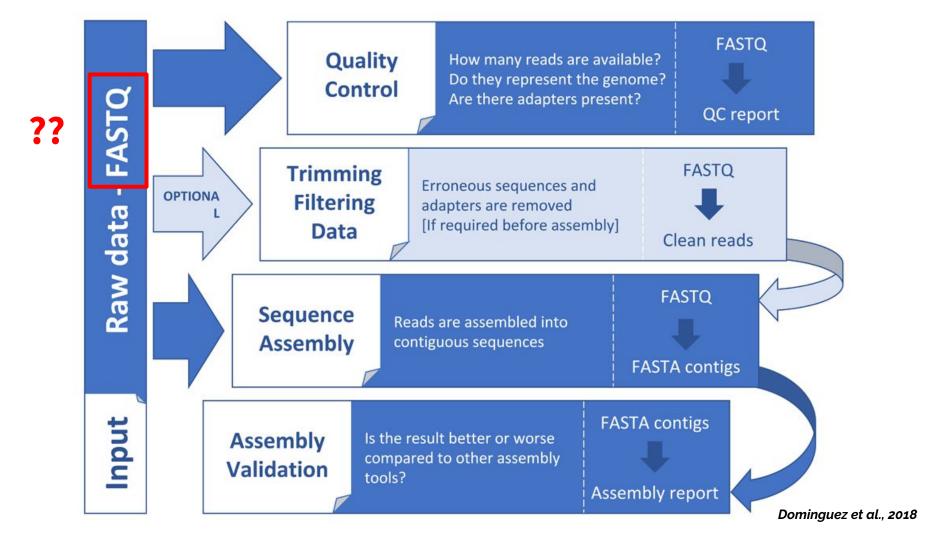


# Ensamblaje de genomas

Fernanda Luz P Costa



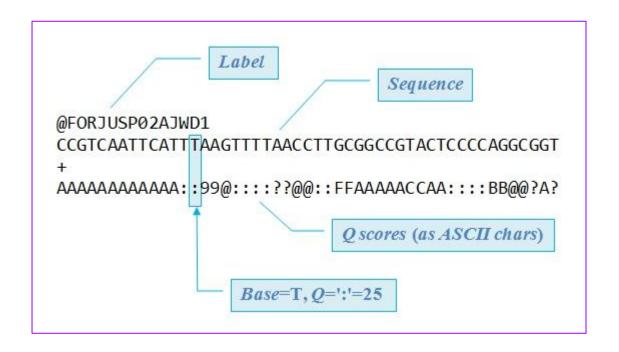






#### **Archivo FastQ**

Secuencia y calidad por base





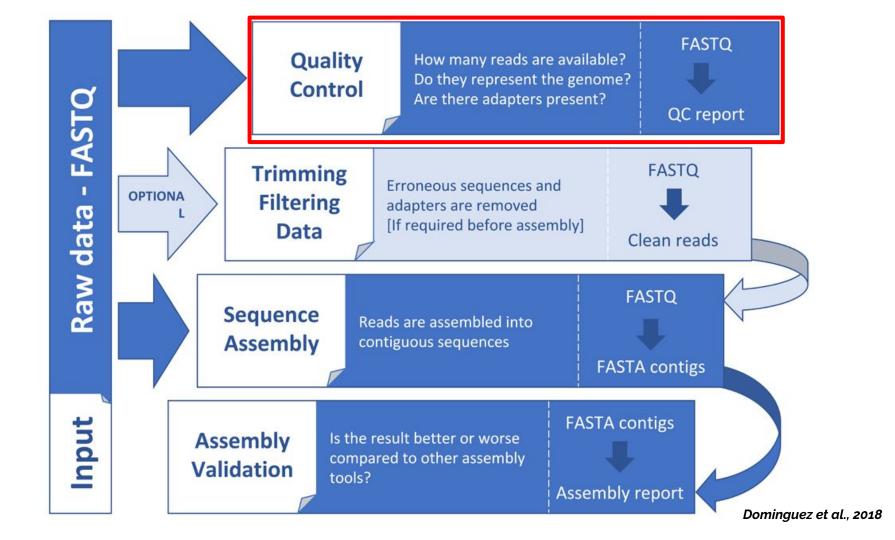
#### **Phred Score**

Calidad de las bases de los fragmentos (reads)

A quality value Q is an integer representation of the probability p that the corresponding base call is incorrect.

$$Q = -10 \log_{10} P$$
  $\longrightarrow$   $P = 10^{\frac{-Q}{10}}$ 

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

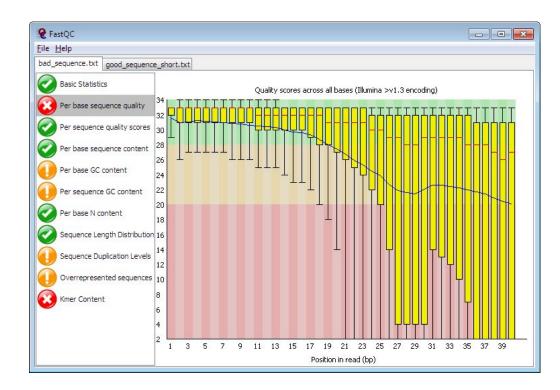


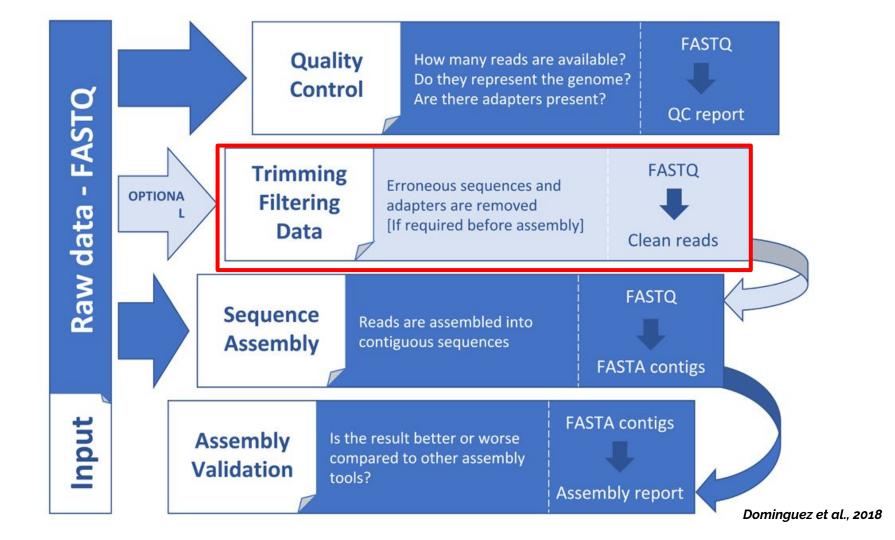


## **Actividad 2 - FastQC**

#### Instalar FastQC

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

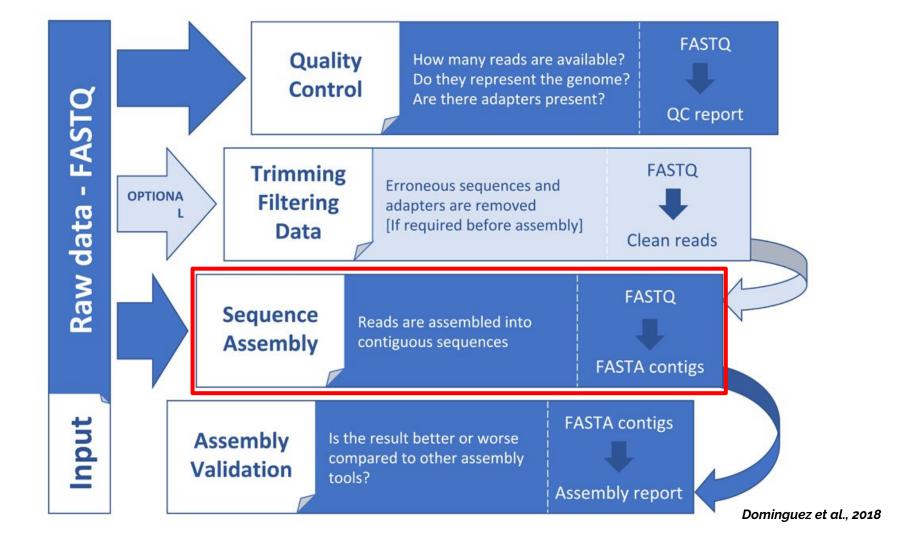


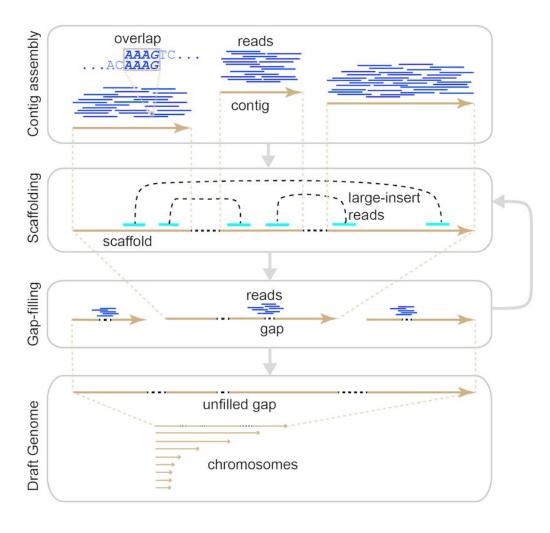


## **Trimmomatic**

A flexible read trimming tool for Illumina NGS data

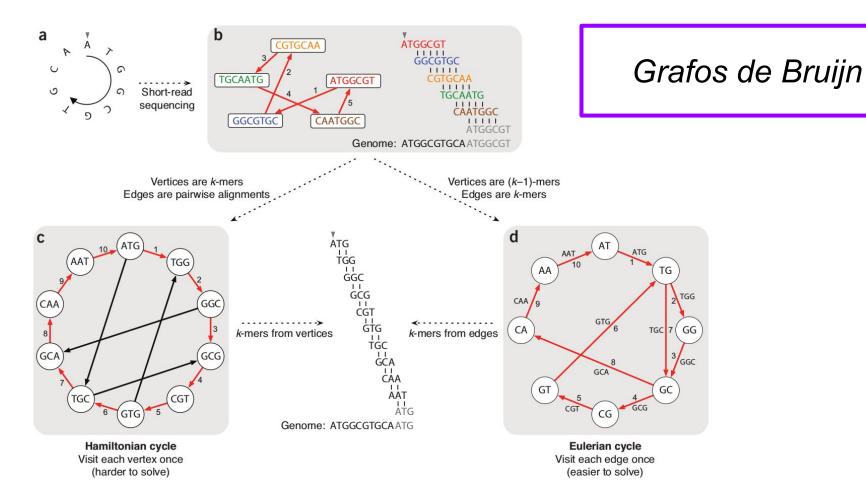
http://www.usadellab.org/cms/?page=trimmomatic

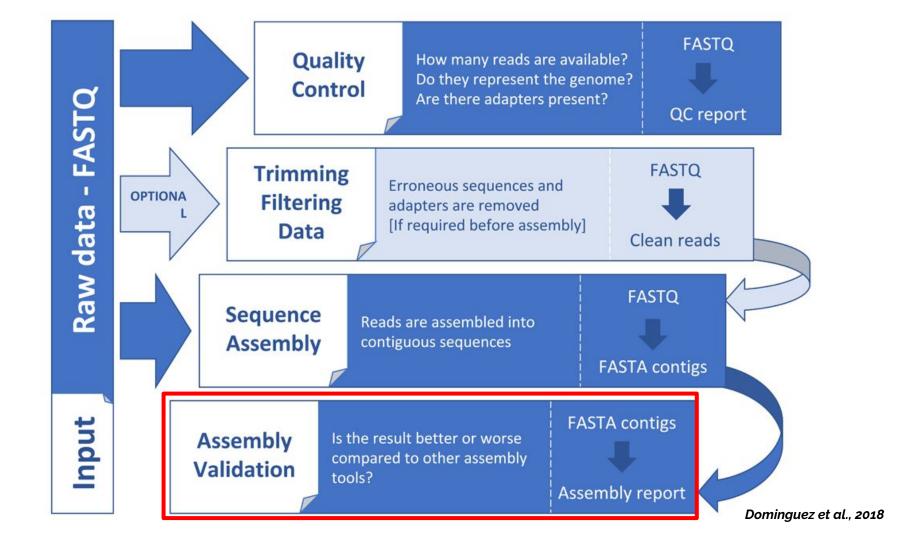




## De Novo Ensamblaje

¿Cuáles son las ventajas y limitaciones de esta técnica?





## **QUAST**

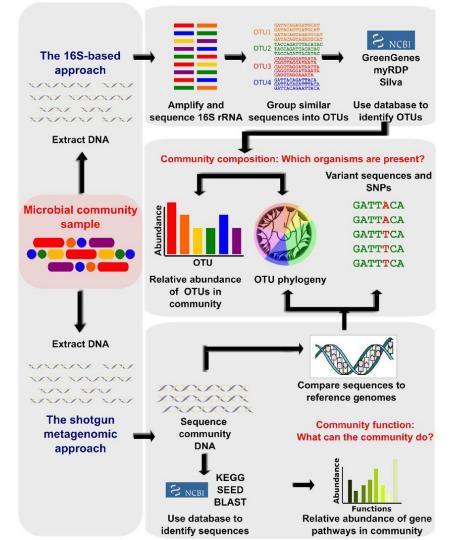
Quality Assessment Tool for Genome Assemblies

http://bioinf.spbau.ru/quast

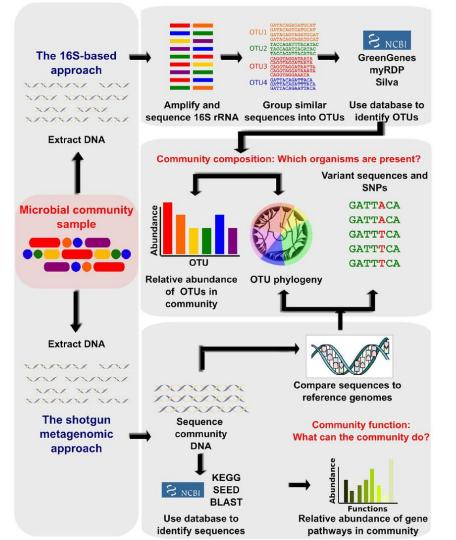


# Introducción al análisis de Metagenómica

Fernanda Luz P Costa



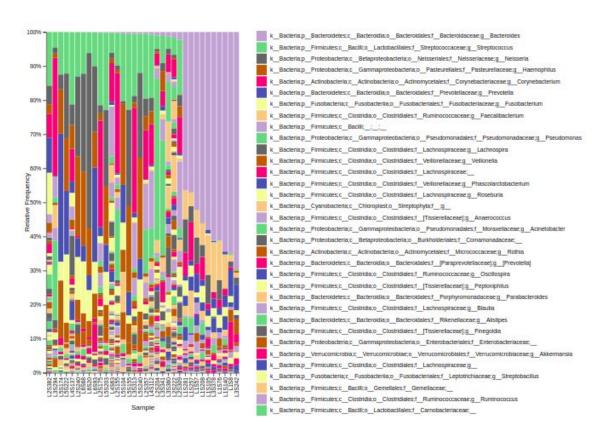
La metagenómica es una herramienta molecular utilizada para analizar el ADN adquirido de muestras ambientales, con el fin de estudiar la comunidad de microorganismos presentes, sin la necesidad de obtener cultivos puros.



- diversidad de la vida;
- nuevas moléculas de utilidad biotecnológica como biomarcadores, nuevos antibióticos, nuevos medicamentos naturales;
- comportamiento microbiótico y su huésped o entorno;
- identificación de nuevas especies.









#### Fernanda Luz P Costa

fernandalpcosta@gmail.com consultoriabiocom@gmail.com





https://www.consultoriabiocom.com/

### Referencias Bibliograficas

- Dominguez Del Angel V, Hjerde E, Sterck L et al. Ten steps to get started in Genome
  Assembly and Annotation [version 1; peer review: 2 approved]. F1000Research 2018,
  7(ELIXIR):148 (https://doi.org/10.12688/f1000research.13598.1)
- Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med. 2001;40(4):346–358.
- Kedaigle A, Fraenkel E. Turning omics data into therapeutic insights. Current Opinion in Pharmacology. 2018; 4:95-101.
- Sohn, jang-il & Nam, Jin-Wu. (2016). The present and future of de novo whole-genome assembly. Briefings in Bioinformatics. 19. bbw096. 10.1093/bib/bbw096.
- Morgan, Xochitl & Huttenhower, Curtis. (2012). Chapter 12: Human Microbiome Analysis.
   PLoS computational biology. 8. e1002808. 10.1371/journal.pcbi.1002808.