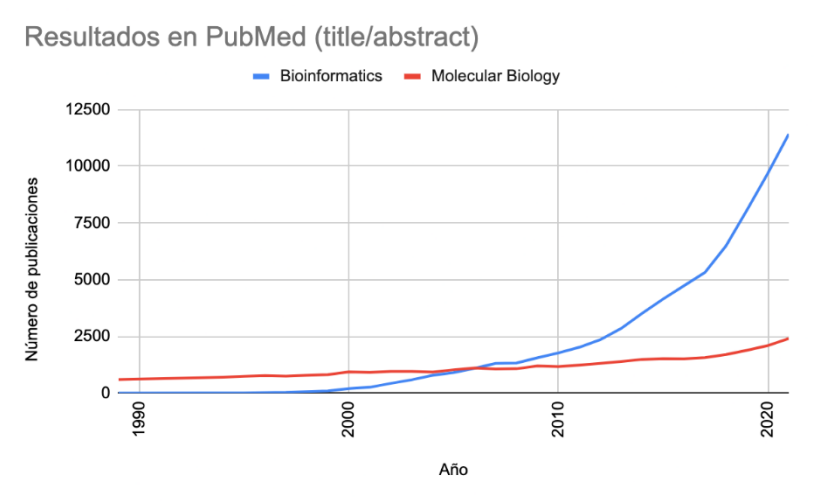
Introducción

Bioinformática
Máster en Biotecnología

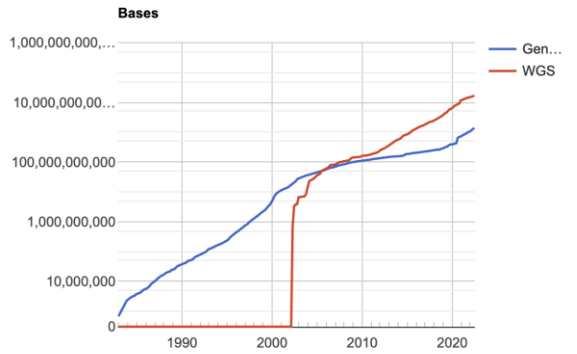
La bioinformática en números

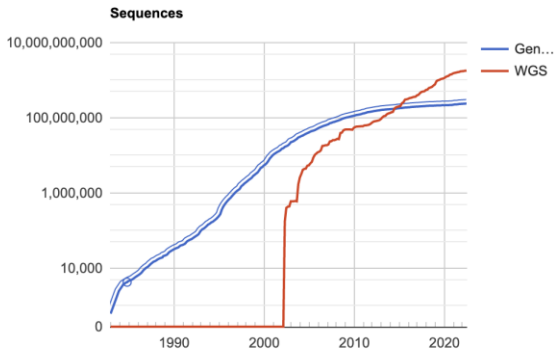
Número de publicaciones



- La disponibilidad de datos moleculares y publicaciones 'bioinformáticos' crecen exponencialmente
- El contenido de GeneBank se duplica cada 18 meses

Tamaño bases de datos





Notes on GenBank statistics

The following table lists the number of bases and the number of sequence records in each release of GenBank, beginning with Release 3 in 1982. CON-division records are not represented in these statistics: because they are constructed from the non-CON records in the database, their inclusion here would be a form of double-counting. Also note that this table is limited to 'traditional', non-set-based (WGS/TSA/TLS) GenBank records. From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months.

Una biología cada vez más computacional

¿Por que tanta computación?

"[B]iology adapted itself to the computer, not the computer to biology," (Hallam Stevens)

Ordenadores y métodos computacionales permiten responder a preguntas nuevas

Métodos computacionales permiten obtener una versión global del problema

Convertir ideas en hipótesis testeables


PLOS BIOLOGY

OPEN ACCESS

RESEARCH MATTERS

All biology is computational biology

Florian Markowetz

Published: March 9, 2017 • https://doi.org/10.1371/journal.pbio.2002050

'Ideas equivocadas' y recelos

La bioinformática y la biología computacional prestán meramente servicios 'técnicos' pero no hay visión cientifica
-- Biología usando metodos computacionales

Resuso de datos publicos es 'parasitismo cientifico' (Data Sharing. 2016 Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D. NEJM)

Definición: Bioinformática, Biología computaciónal

Técnicas computacionales para resolver problemas biológicos/biomédicos

- Secuencias de genes y proteínas
- Secuencias de genomas completos
- Datos de expresión génica
- Estructura de proteínas y RNA no-codificante
- Interacción entre proteínas y rutas metabólicas
- Información epi-genética
- Regulación de la expresión génica (TFBS, islas CpG)

Biología molecular
Genética
Genómica
Biomedicina
…

Datos experimentales
Bases de Datos

Computación
Algorítmica

Conocimiento biológico

- Salud
- Biotecnología
- Medio ambiente
- Evolución

El ejemplo 'clásico': BLAST

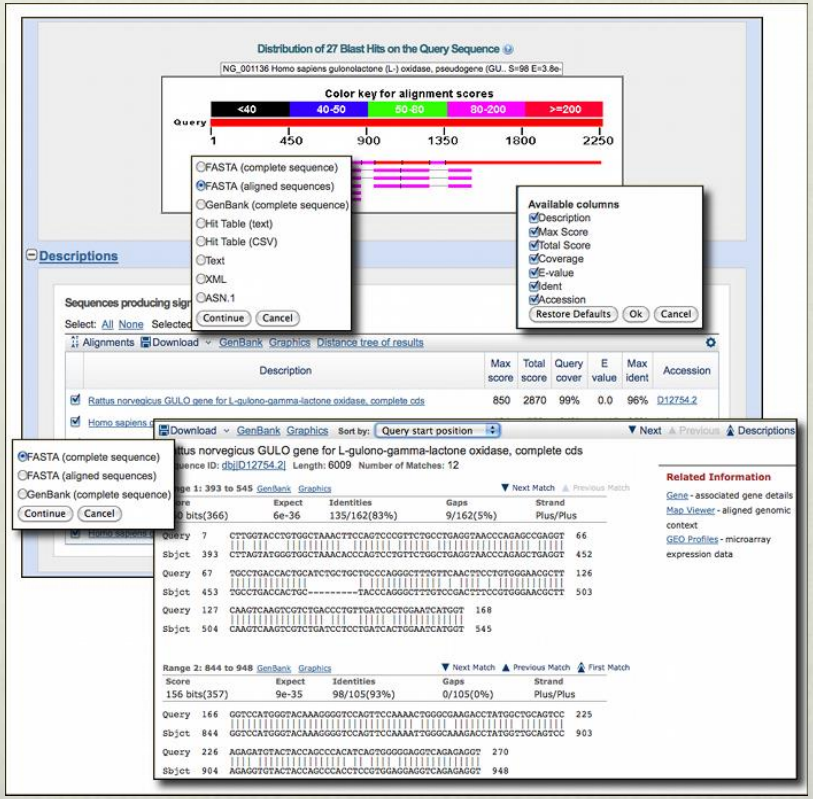
Tenemos una secuencia anónima de DNA o de proteína

¿A que gene/proteína corresponde la secuencia anónima?
→*probablemente hay un gen bien caracterizado en la base de datos con una secuencia "similar" lo que permitiría obtener conocimiento acerca del gen estudiado*

¿En que otras especies existe mi gen o la proteína?
→*responder a preguntas evolutivas*

¿De que especie es?
→ *en experimentos de metagenómica o interacciones entre especies – parasito/huésped*

- Comparar la secuencia de entrada con **todas** las secuencias de una base de datos
- Encontrar las entradas en la base de datos con las que comparte un **alto grado de similitud**



Bioinformática vs. Laboratorio húmedo – nivel usuario

Bioinformática

¿Que función podría tener una secuencia anónima dada?

Laboratorio húmedo

¿El paciente está infectado por el SARS-CoV-2?

BLAST

Buscar (mediante alineamiento) si existe una secuencia similar con función conocida en una base de datos

PCR

Intentar amplificar una secuencia del parásito para detectar su presencia

Bioinformática vs. Laboratorio húmedo: Misma finalidad (responder a preguntas biológicas) pero diferentes métodos

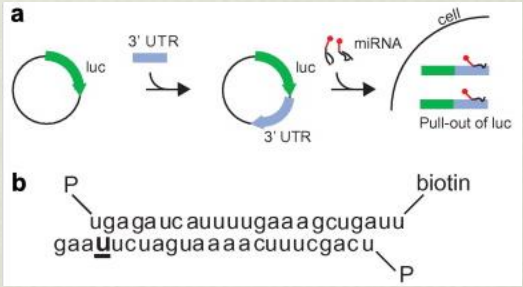
Bioinformática vs. Laboratorio húmedo: desarrollo de nuevos métodos

¿A que genes regula un microRNA dado?

- **Predicción de dianas a partir de las secuencias basándose en modelos termodinámicos**
- **Implementación del modelo en un algoritmo computacional**

```
HTT 3' UTR     5' ...GAUUGCUUUUGUUUU--CCUGCUGG...
                          |||     |||||||
miR-214        3'       UGACGGACAGACACGGACGACA
```

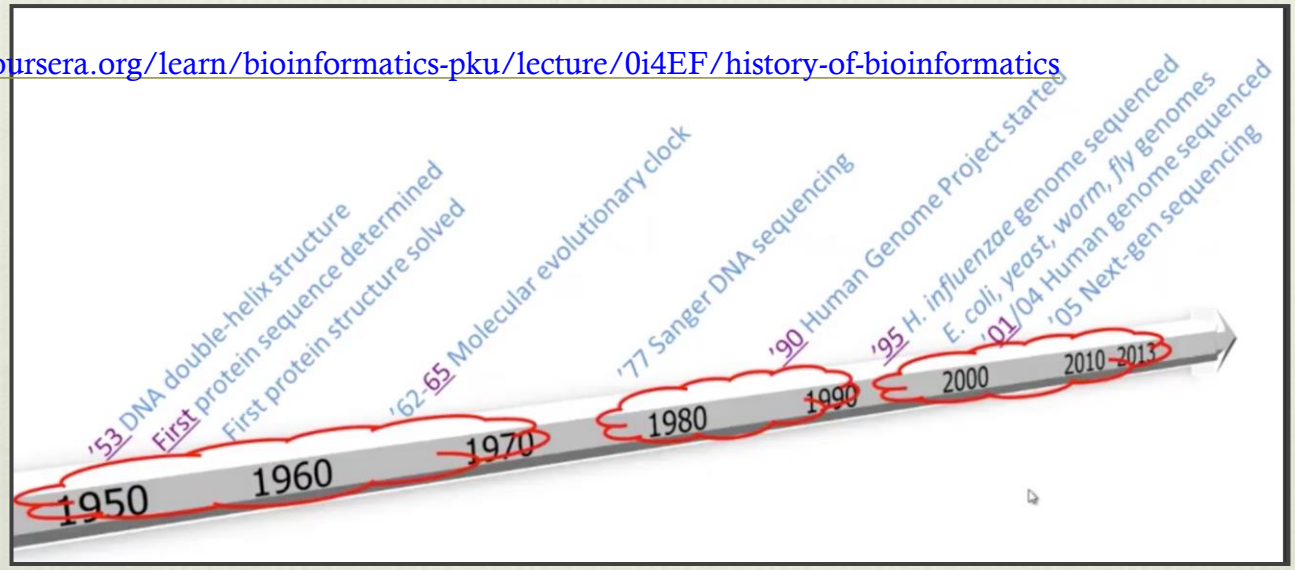
- **Biotinilar los microRNAs y capturar las uniones entre microRNAs biotinilados y los mRNAs**
- **Desarrollar nuevos métodos bioquímicos**



Bioinformática vs. Laboratorio húmedo: Misma finalidad (responder a preguntas biológicas) pero diferentes métodos

Orígenes y evolución del la bioinformática

https://www.coursera.org/learn/bioinformatics-pku/lecture/0i4EF/history-of-bioinformatics



La bioinformática tiene sus orígenes en la biología molecular de los años 60 del siglo pasado (aunque en aquel entonces nadie hablaba de bioinformática)

Dos eventos tienen especial relevancia para la evolución la bioinformática:

- Proyecto Genoma Humano
- La aparición de métodos de secuenciación masiva

1. Hogeweg P, Hesper B. Interactive instruction on population interactions. *Comput Biol Med.* 1978;8:319–327. [PubMed] [Google Scholar]

2. Hogeweg P. Simulating the growth of cellular forms. *Simulation.* 1978;31:90–96. [Google Scholar]

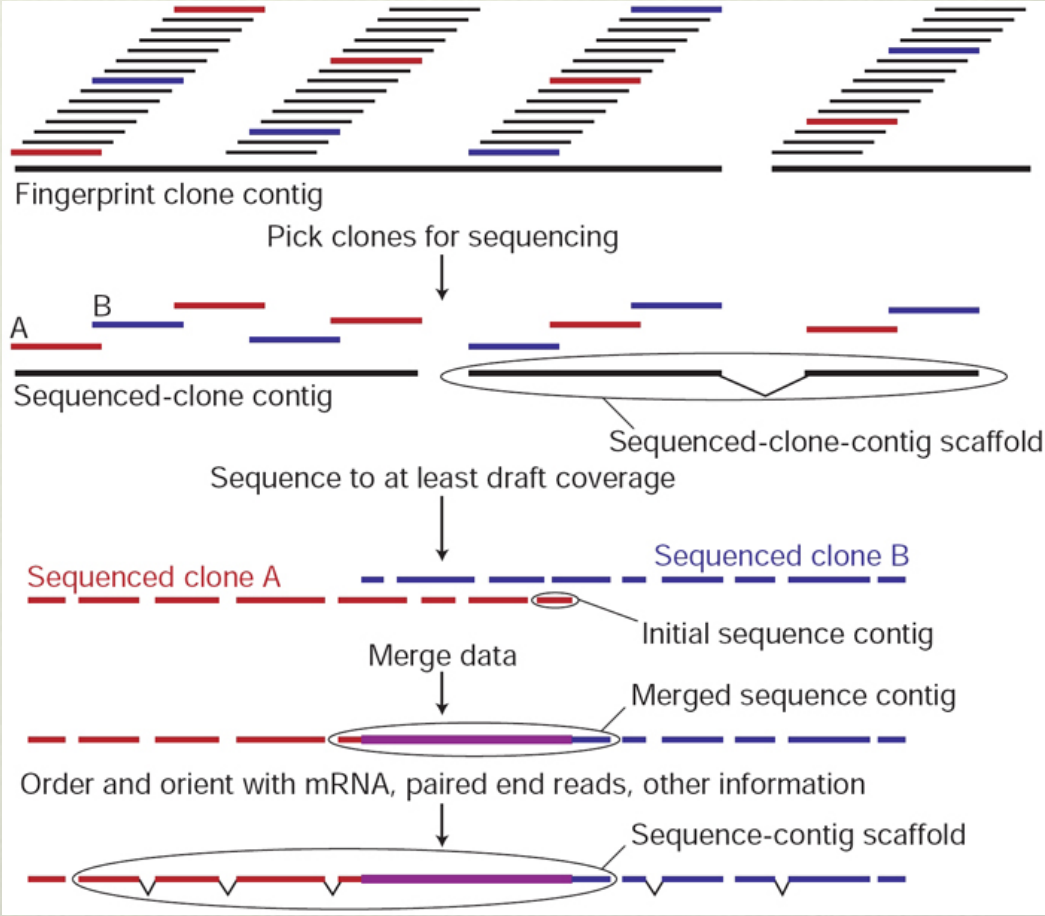
Proyecto genoma humano

El proyecto empezó en 1984 y terminó en 2003 con la publicación de la secuencia de la parte de eucromatina del genoma humano

- Determinación de la secuencia de bases nitrogenadas que forman el ADN humano
- Identificación de los genes en el genoma humano
- Hacer los resultados accesibles libremente

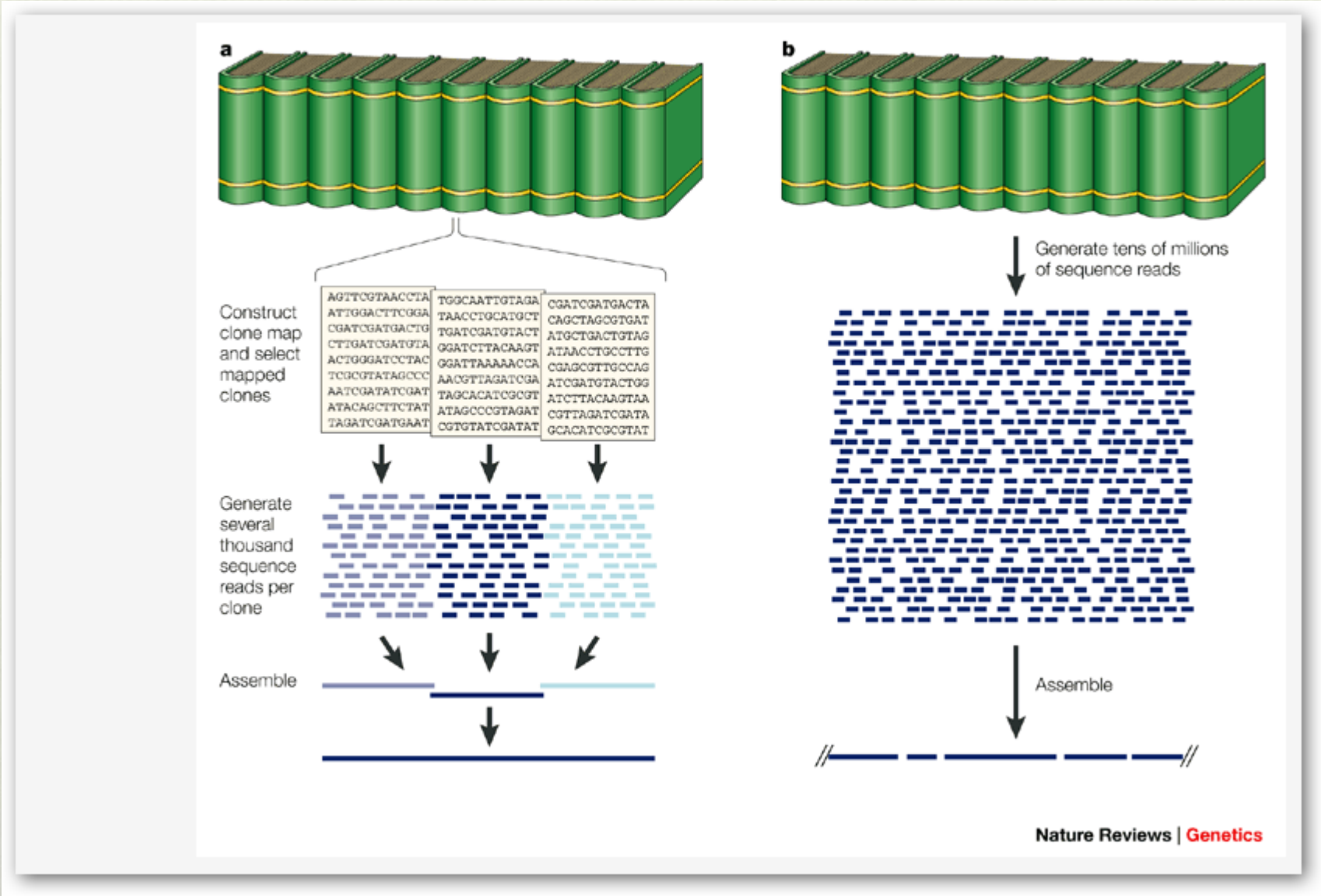
Importancia de la bioinformática

- Selección de los clones
- Ensamblar los contigs



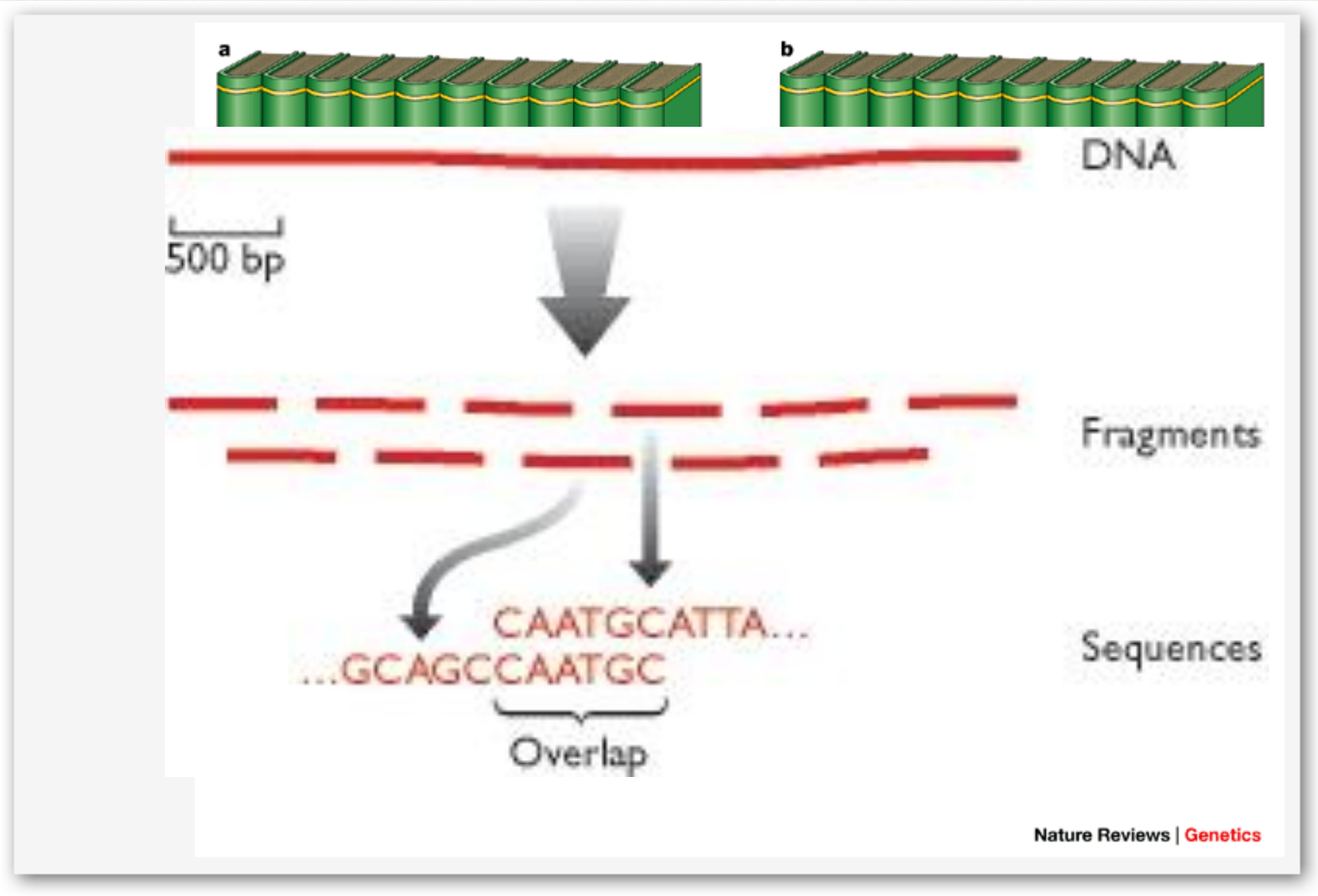
https://www.nature.com/scitable/content/Initial-sequencing-and-analysis-of-the-human-16729

Proyecto genoma humano



http://www.nature.com/nrg/journal/v2/n8/full/nrg0801_573a.html

Proyecto genoma humano



Secuenciación masiva

- Lecturas cortas (< 500 pb)
- Alto número de lecturas

Lecturas mas largas (10kb-25kb)
'Bajo' número de lecturas

illumina®

https://www.illumina.com/

Oxford NANOPORE Technologies

https://nanoporetech.com/products

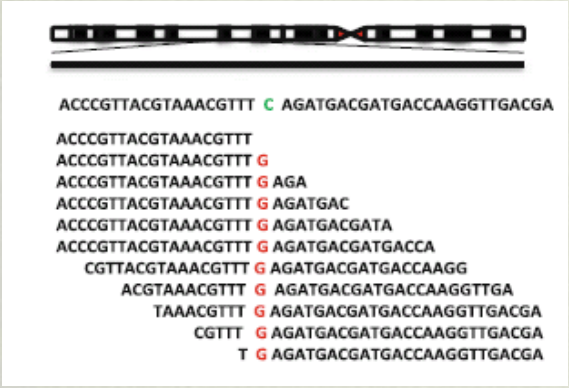
PacBio https://www.pacb.com/

Secuenciación masiva: la revolución molecular

	SANGER	SECUENCIACIÓN MASIVA		
	Di-deoxy terminator	**Roche 454 GS FLX (PS)**	**Illumina HiSeq 2000 (RT)**	**SOLID V4 (SBL)**
Salida por proceso	1.6 Mb	600 Mb	200 GB	100 GB
Tiempo/Proceso	1h	10 h	9 d	11 d
Longitud media "reads"	800 pbs	400 pb	100 pb	75 pb
Salida por día	38.4 Mb	1.44 GB	22.2 GB	9 GB
Usos frecuentes	-	Secuenciación de novo Captura de exones	Resecuenciación Captura de exones Metagenómica	Resecuenciación Captura de exones Metagenómica

Secuenciación masiva: diferentes aplicaciones

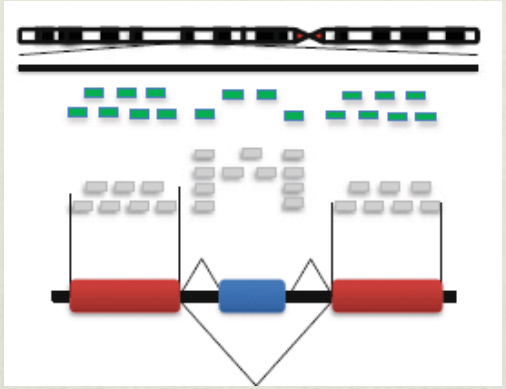
DNA sequencing



Detect sequence variants (SNV/SNPs, indels)

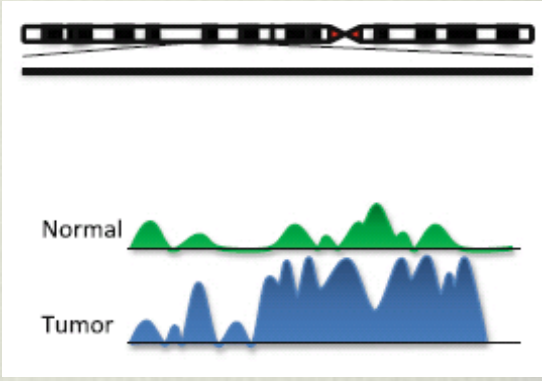
- Individual genomes
- Cancer genomes

Gene Expression



- Transcription
- small RNAs
- TF binding to DNA

Epigenomics

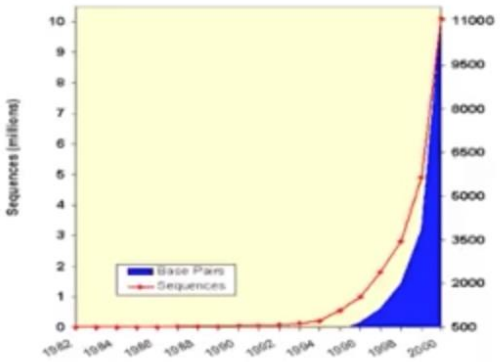


- DNA methylation
- Histone modifications

Retos técnicos

Opportunities and challenges hand-in-hand: the driving forces of bioinformatics

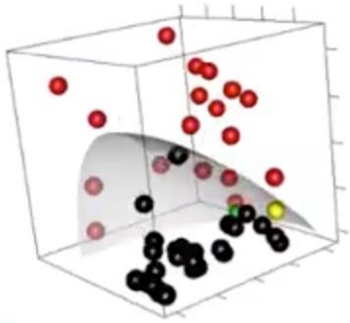
- High-throughput data
  - Huge amount
  - Explosive growth
  - Low signal-to-noise ratio
  - Multiple types



ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

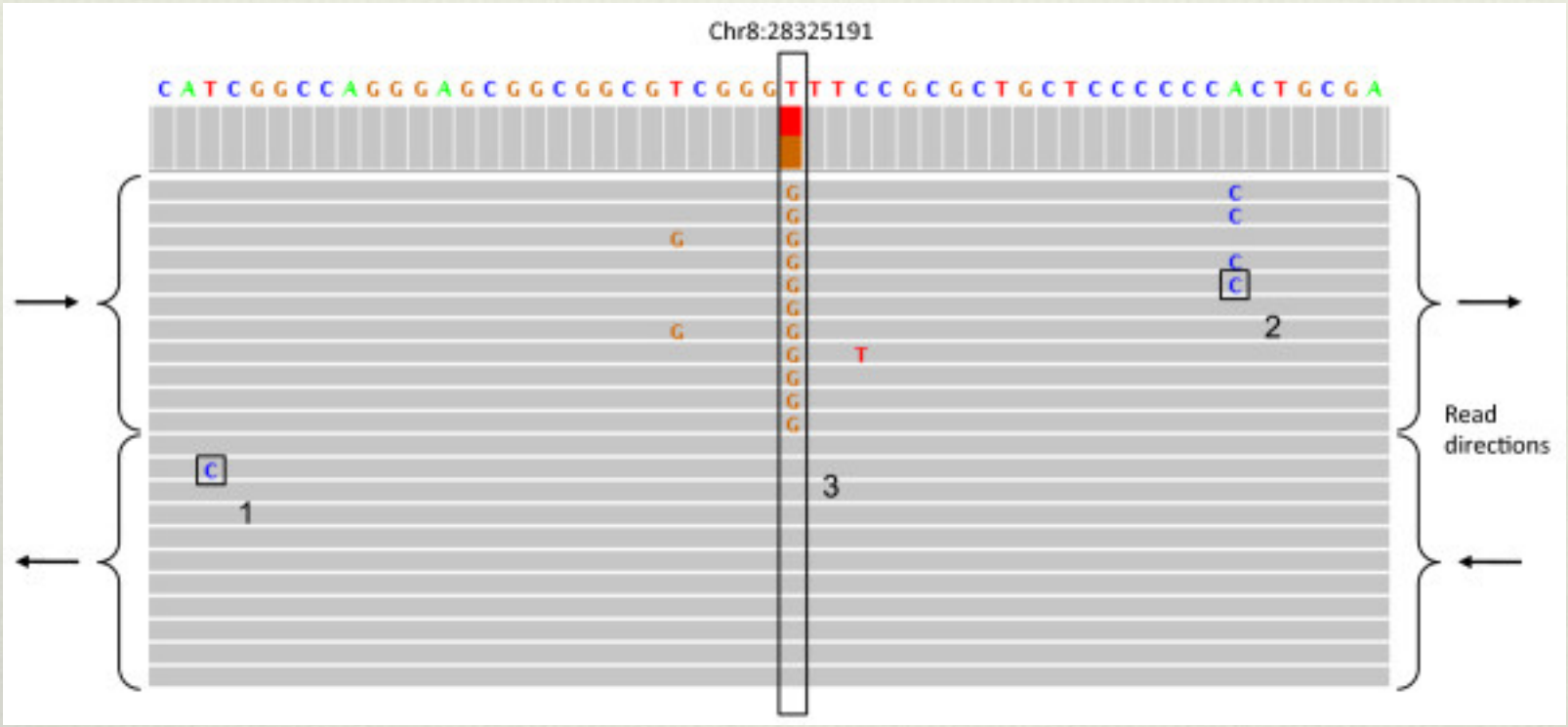
- Requirements for the methods
  - Data needs to be stored in efficient ontology-based database systems
  - The huge amount of data requires efficient algorithms
  - Exponential growth requires scalable methods
  - The low signal-to-noise ratio requires accurate methods
  - Multiple types of data require data integrative methods

12:08



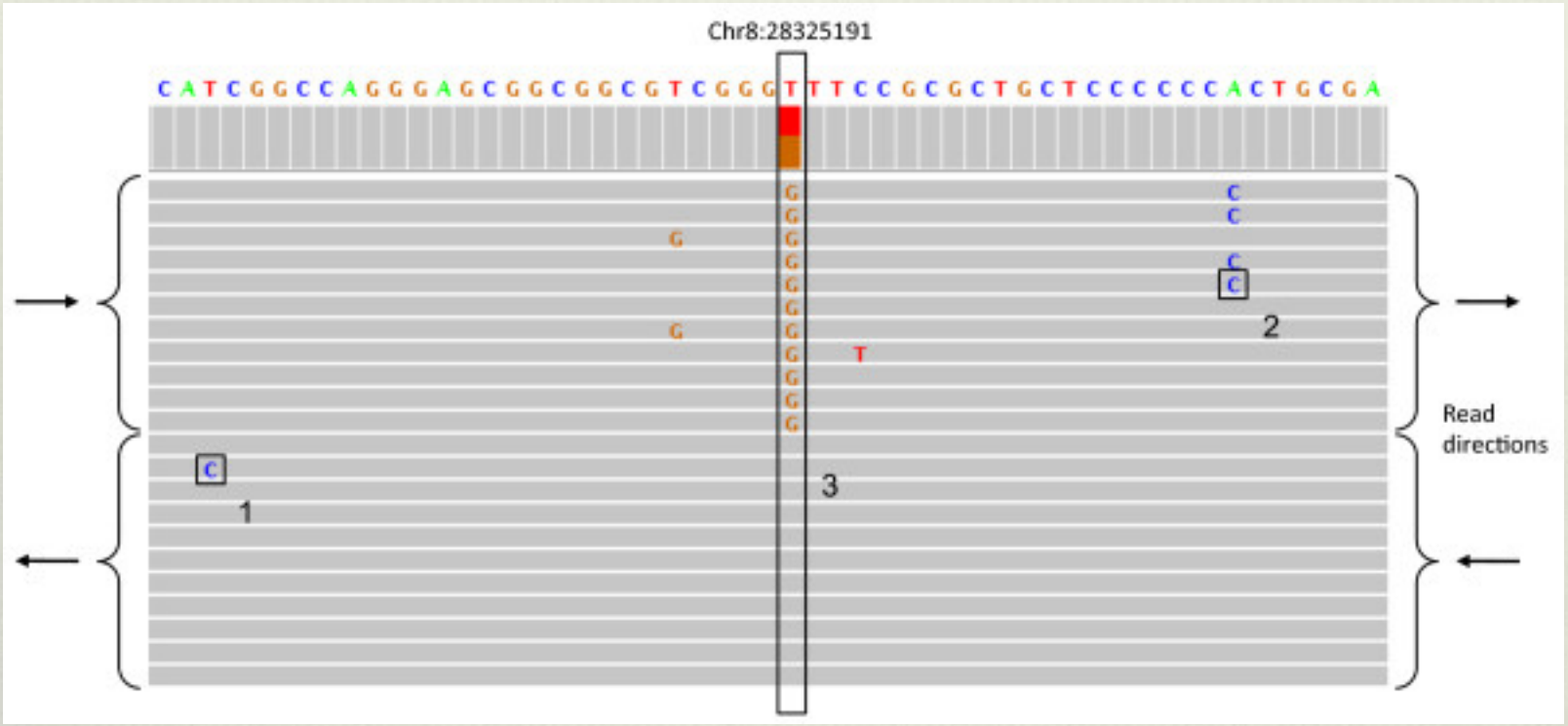
https://www.coursera.org/learn/bioinformatics-pku/lecture/vQo1i/what-is-bioinformatics

Retos técnicos



https://www.coursera.org/learn/bioinformatics-pku/lecture/vQo1i/what-is-bioinformatics

Retos técnicos



Heterozygote G/T

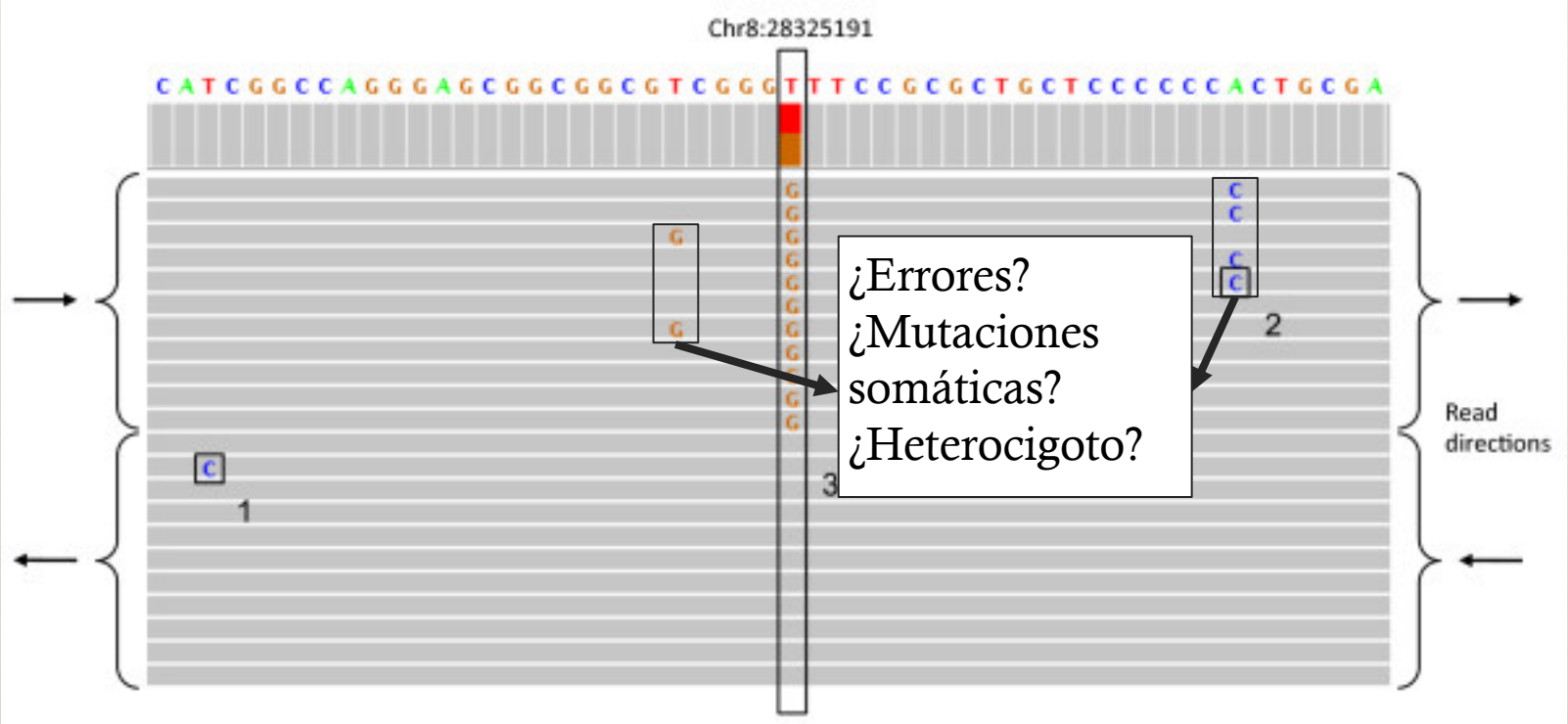
Detect germline variants

Inferential statistics

- Bayesian
- Frequentist

https://www.coursera.org/learn/bioinformatics-pku/lecture/vQo1i/what-is-bioinformatics

Retos técnicos



Se requiere métodos altamente específicos para no producir un elevado número de falsos positivos

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3295828/

Nuevos retos – nuevas habilidades, destrezas y conocimientos

Conocimientos básicos: Genética, Bioquímica, Biología molecular
Matemática, Estatística

<u>El entorno Linux</u>

- La resolución de muchos problemas requieren servidores de calculo intensivo
- Existen programas que no están disponibles para Windows

<u>Programación</u>

Emplear programas o código (R, Python) para analizar los datos: visualización, exploración (PCA, agrupamiento jerarquico, etc.)

Implementar nuevos métodos: alinear millones de lecturas frente al genome, expresión diferencial, visualización, encontrar picos en el genoma (ChIP-Seq), etc. etc. etc.

Análisis de datos: hacer mucho con poco código

Matrix de expresión: Niveles de expresión génica para todos los genes y muestras

gene	SG_unfed_R1	SG_unfed_R2	SG_unfed_R3	SG_unfed_R4	SG_unfed_R5	SG_first_12h_R1	SG_first_12h_R2
MG_TRINITY_DN0_c0_g2_i2	0.0	9.16	3.39	2.89	2.15	11.49	0.0
MG_TRINITY_DN0_c0_g2_i3	0.54	0.0	0.0	4.97	10.54	23.4	0.0
MG_TRINITY_DN0_c0_g2_i5	0.5	0.0	0.0	2.07	2.87	0.0	0.0
MG_TRINITY_DN0_c0_g2_i6	0.0	0.0	0.03	0.04	0.0	0.0	0.0
MG_TRINITY_DN0_c0_g2_i7	0.0	0.0	0.0	0.89	1.82	0.0	0.0
MG_TRINITY_DN0_c0_g2_i8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MG_TRINITY_DN0_c0_g3_i1	0.45	0.0	0.0	0.0			
MG_TRINITY_DN0_c0_g3_i2	1.48	14.62	7.05	4.7			
MG_TRINITY_DN10001_c0_g1_i3	7.74	15.88	6.09	5.89			

Rows 25011

Columns 89

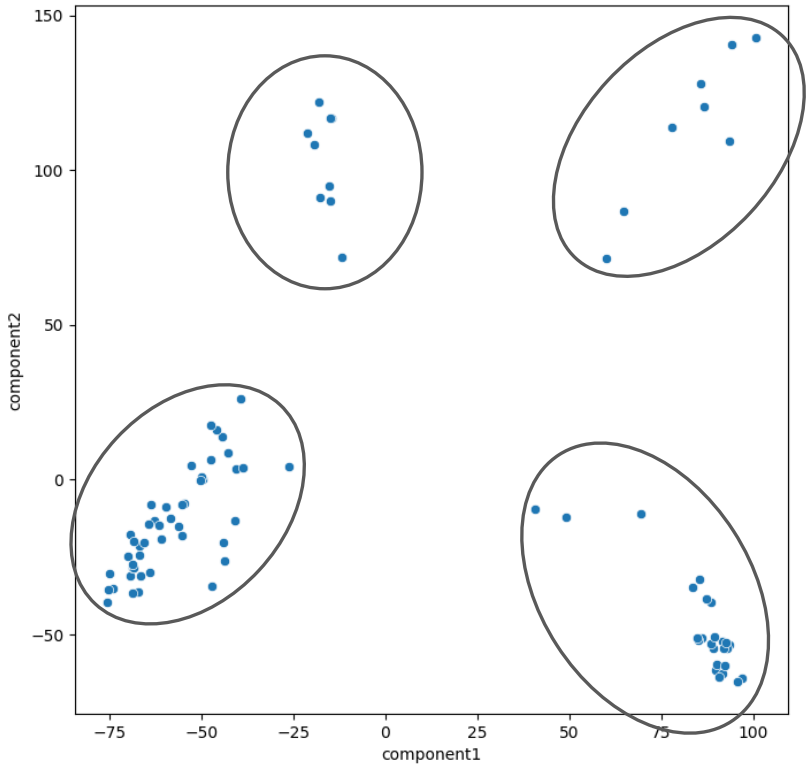
- ¿Que relación hay entre las muestras? – ¿que condiciones se parecen más?
- ¿Existen muestras ‘problematicas’?

Análisis de datos: hacer mucho con poco código

Código en python

```python
import argparse
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

parser = argparse.ArgumentParser()
parser.add_argument("-f", "--file", help="file")
parser.add_argument("-o", "--output", help="output filename")
args = parser.parse_args()

# Reading the matrix
df = pd.read_csv(args.file, sep="\t", index_col=0)
df_transposed = df.transpose()

#PCA Analysis
df_standarized = StandardScaler().fit_transform(df_transposed)
PCA_data = PCA(n_components=2).fit_transform(df_standarized)
df_PCA = pd.DataFrame(data=PCA_data, index=df_transposed.index, c

#Plot
plt.figure(figsize=(8, 8))
ax = sns.scatterplot(data=df_PCA, x="component1", y="component2")
fig = ax.get_figure()
fig.savefig(args.output)
plt.close()
```

Resultado: PCA



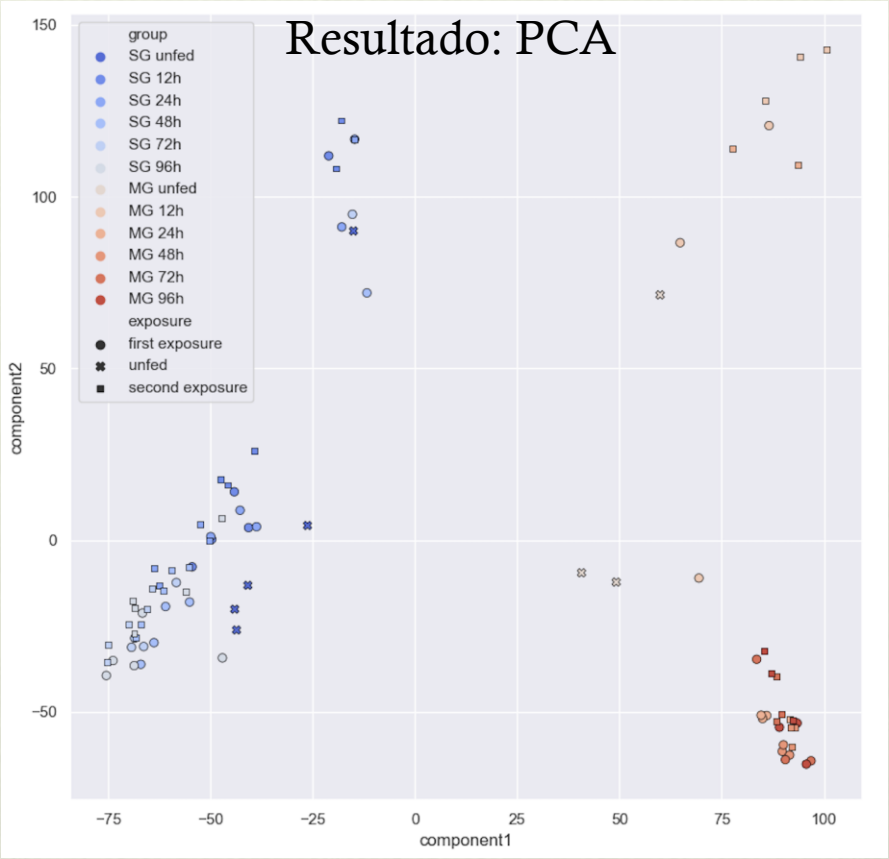
Se puede distinguir 4 grupos pero no sabemos que muestras son

Análisis de datos: hacer mucho con poco código

Código en python

```
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

def get_vectors(df_transposed):
    color_vector = []
    form_vector = []
    for index, row in df_transposed.iterrows():
        if "first" in index:
            form_vector.append("first exposure")
            color_vector.append(" ".join([index.split("_")[0], index.sp
        elif "second" in index:
            form_vector.append("second exposure")
            color_vector.append(" ".join([index.split("_")[0], index.sp
        elif "unfed" in index:
            form_vector.append("unfed")
            color_vector.append(" ".join(index.split("_")[0:2]))
    return color_vector, form_vector


parser = argparse.ArgumentParser()
parser.add_argument("-f", "--file", help="file")
parser.add_argument("-o", "--output", help="output filename")
args = parser.parse_args()

# Reading the matrix
df = pd.read_csv(args.file, sep="\t", index_col=0)
df_transposed = df.transpose()

#PCA Analysis
df_standarized = StandardScaler().fit_transform(df_transposed)
PCA_data = PCA(n_components=2).fit_transform(df_standarized)
df_PCA = pd.DataFrame(data=PCA_data, index=df_transposed.index, columns=

#Adding information to plot
color_vector, form_vector = get_vectors(df_transposed)
df_PCA["group"] = color_vector
df_PCA["exposure"] = form_vector

#Plotting
sns.set_theme()
plt.figure(figsize=(10, 10))
ax = sns.scatterplot(data=df_PCA, x="component1", y="component2", hue="
                     style_order=["first exposure", "unfed", "second exp
ax.legend(loc=2)
fig = ax.get_figure()
fig.savefig(args.output)
plt.close()
```

Resultado: PCA



- Diferencias entre tubo digestivo (MG) y glandulas salivales (SG)
- Diferencias en función de la duración de alimentación

Desarrollo de algoritmos

Resultado de secuenciación: fichero en formato fastq

Convertir en valores de expresión

↓

Expresión diferencial

↓

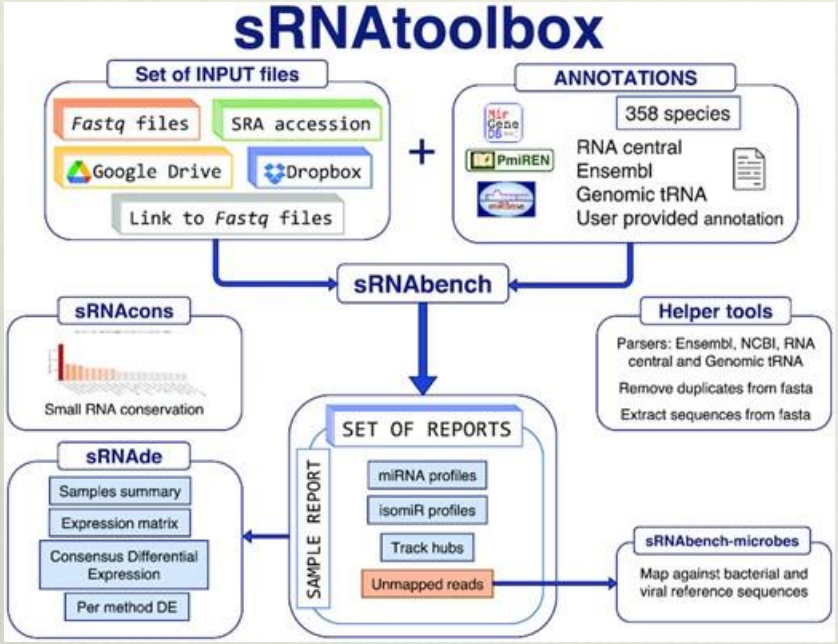
Visualizar los resultados

- Preprocesar los datos
- Alinear frente a una referencia
- Tomar en cuenta sesgos tecnicos
- Alineamientos ambiguos

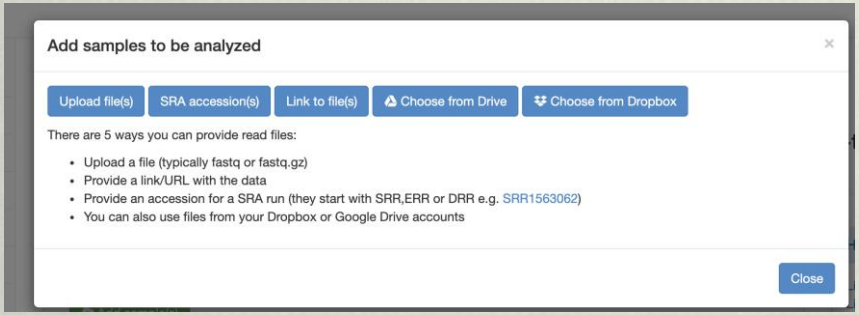
- Emplear un test estatistico teniendo en cuanta el diseño del experimento

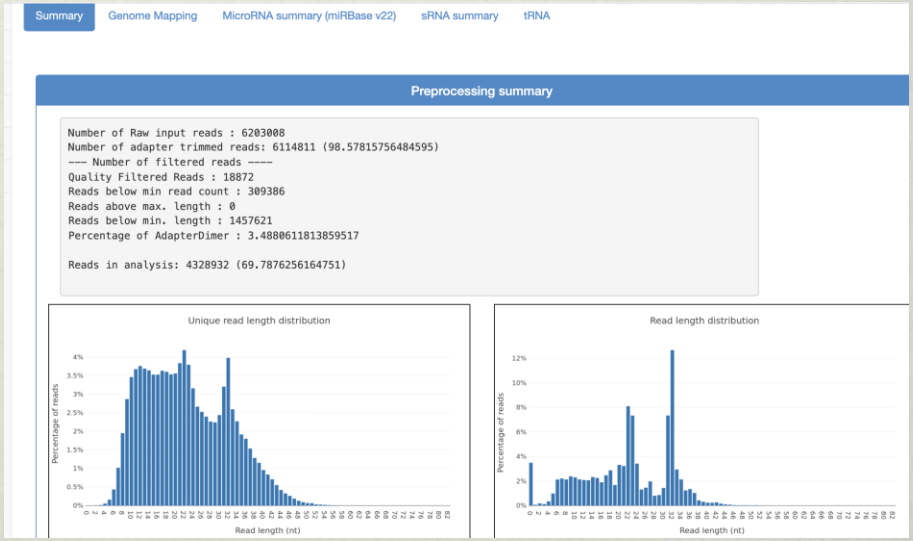
- PCA, agrupamiento jerárquicom heatmaps
- Vulcano plots, RLE
- Reproducibilidad, expresión consenso
- …

Desarrollo de algoritmos



- Java, Python, R
- Django, Ajax, javascript
- Bases de datos (mySQL)
- Json
- Docker, Apache
- Gestor de colas

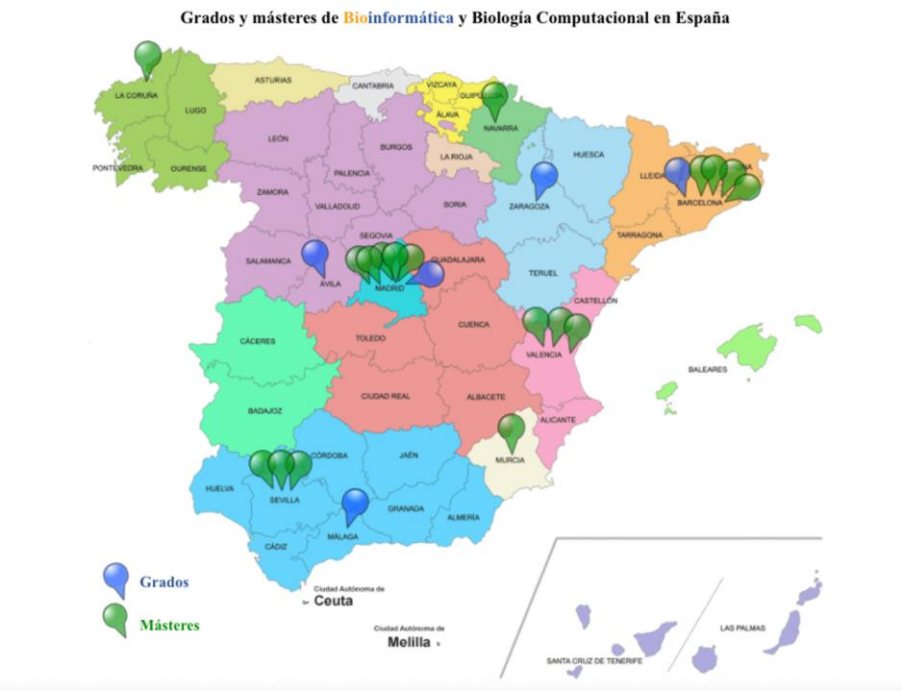




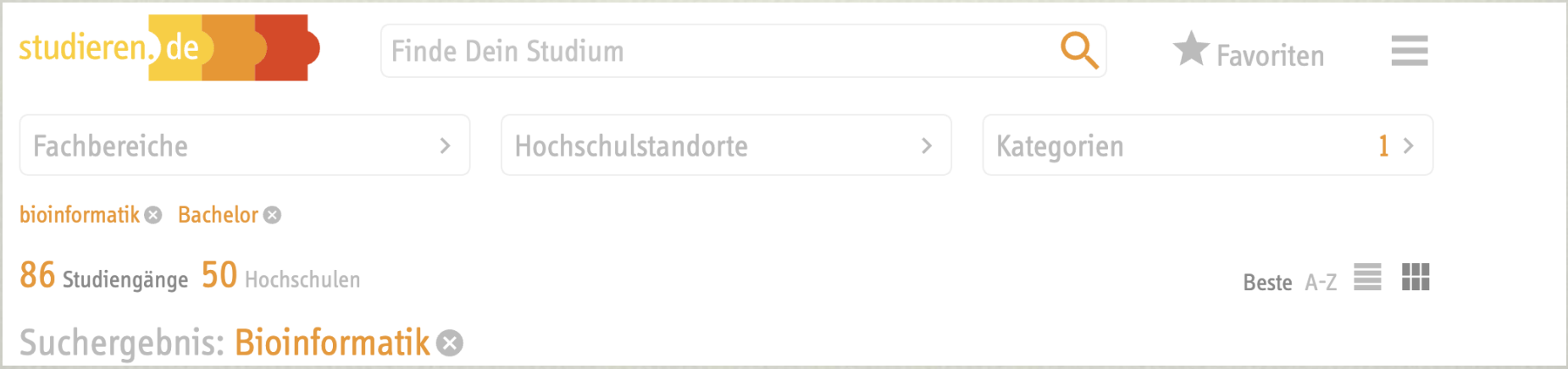
¿Como puedo iniciarme?

BIOINFO GRX

https://bioinformaticsgrx.es/



Pocas opciones de formación reglada en las Universidades españolas



¿Como puedo iniciarme?

BIOINFO GRX

Asociación de jóvenes bioinformáticos con base en Granada.

Nodo de ISCB RSG-Spain.

Inicio Agenda Aprende BioInfo Bolsas de trabajo Actividades Sobre nosotros Redes Sociales EventBrite Contacto

- Cursos (online)
- Talleres
- Autoaprendizaje