



# Metagenómica parte 2

Bioinformática Genómica para Ingeniería en Bioinformática

9 de agosto de 2016

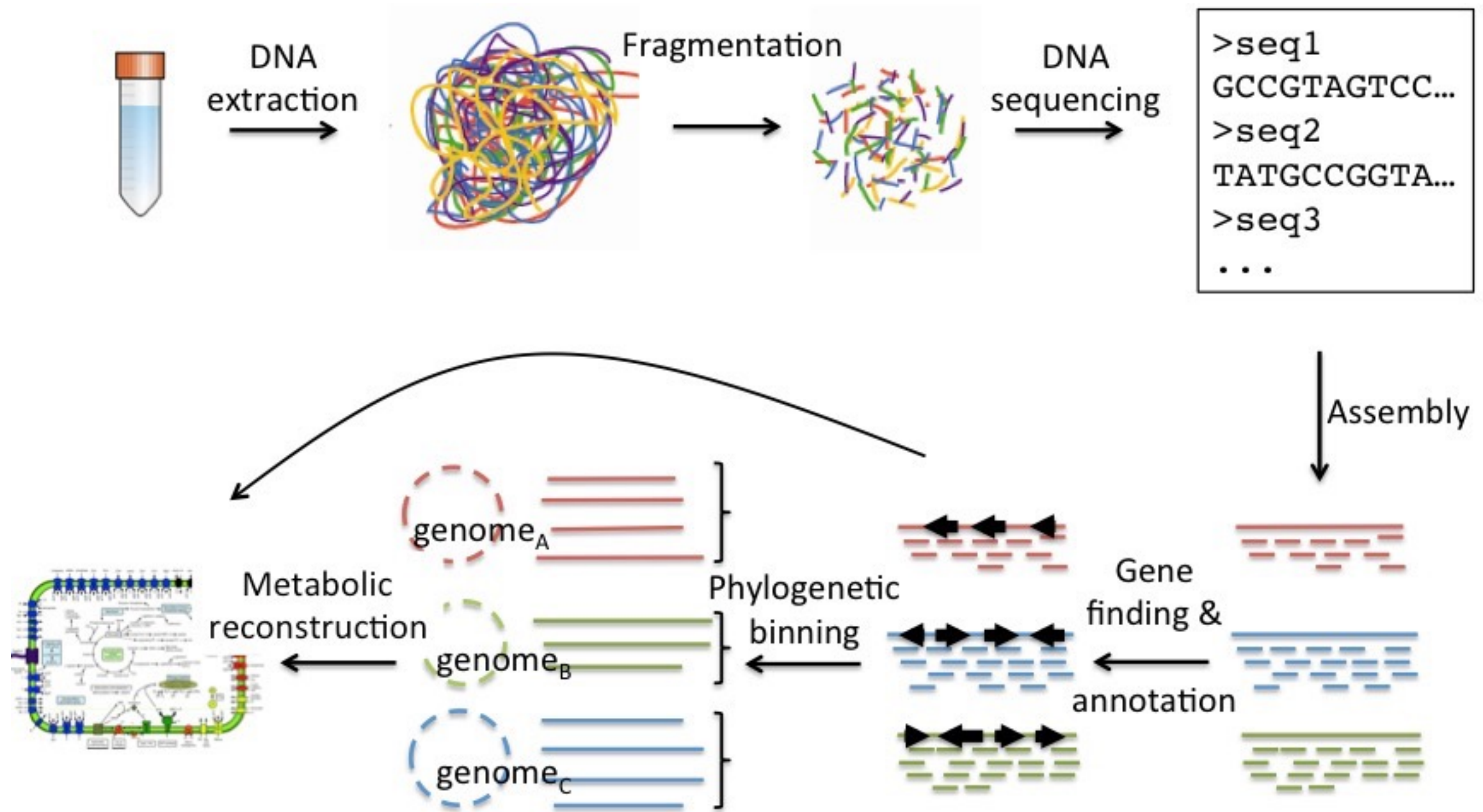
Eduardo Castro-Nallar, PhD

Center for Bioinformatics and Integrative Biology

[www.cbib.cl](http://www.cbib.cl)

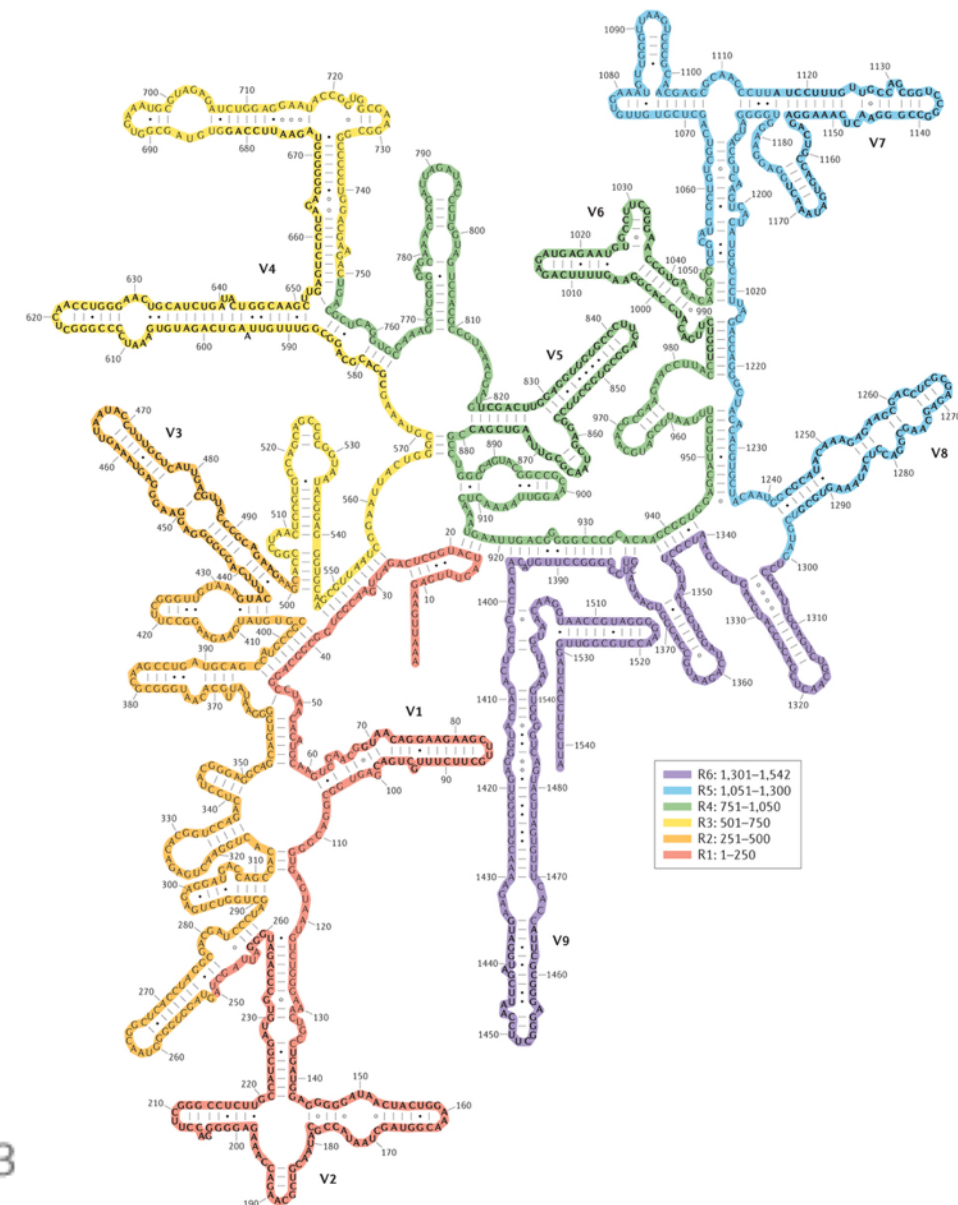
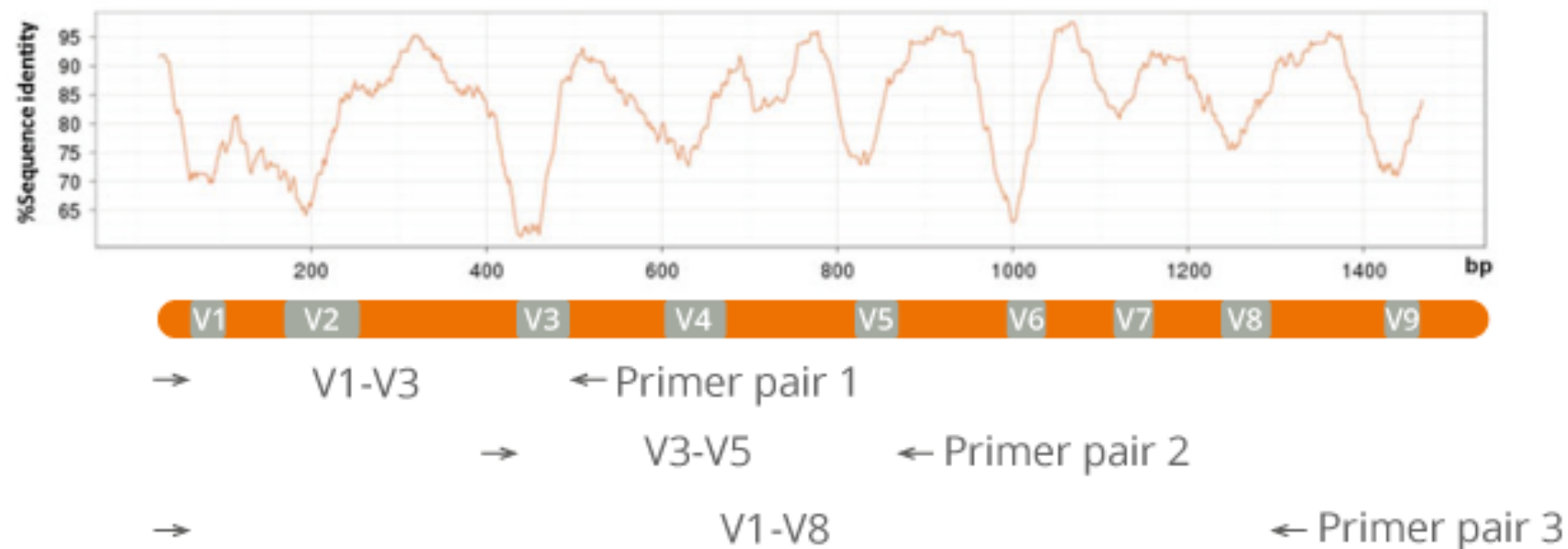
[www.castrolab.org](http://www.castrolab.org)

# Shotgun Sequencing



# Metataxonómica

Regiones variables y constantes componen el gen 16S rRNA



# Metataxonómica



HomeSILVAngsBrowserSearchAlignerDownloadDocumentationProjectsFISH & ProbesShopJobsContact

### Background

The **SILVA** databases are developed and maintained by the [Microbial Genomics and Bioinformatics Research Group](#) in Bremen, Germany, in cooperation with the [Department of Microbiology](#) at the Technical University Munich and the company [Ribocon GmbH](#).

**SILVA** is an interdisciplinary project of biologists and computer scientists to provide:

- fully aligned and up to date small (16S/18S, SSU) and large (23S/28S, LSU) subunit ribosomal RNA "Parc" databases on the webpage as well as ARB files
- preconfigured subsets of only high quality, full-length sequences as **ARB & FASTA** files (SSU/LSU Ref)
- extensive browse & search functionalities for sequence retrieval
- a clear rating system for all steps of data processing with emphasis on sequence and alignment quality
- full compatibility to the software package ARB & the latest official alignments released by the ARB project
- compatibility to many common programs like Phylip or Paup via direct Fasta export or the ARB program
- substantial support related to all aspects of data sets, sequence analysis and probe/primer design by our partner [Ribocon GmbH](#).

#### Release information & Database history

Version 89 of our datasets have been made available in February 2007. Version 123 was released in July 2015 and increased the number of available SSU/LSU sequences to over **5,300,000**. Detailed information about the content of the databases and statistics can be found [here...](#)

### Motivation

Sequencing the ribosomal RNA (rRNA) genes is currently the method of choice for phylogenetic reconstruction and nucleic acid based detection and quantification of microbial diversity. The **ARB** software suite with its corresponding rRNA databases has been accepted by researchers worldwide as their standard tool for large scale ribosomal RNA analysis. More than 20 years of development have already been invested to extend and maintain the system. To provide high quality and comprehensive rRNA databases comprising *Bacteria*, *Archaea* and *Eukarya* the **SILVA** (from Latin *silva*, forest) system has been implemented in 2007. It is designed as an automatic software pipeline for sequence retrieval, quality assignment and the alignment of nucleic acid sequences based on the latest comprehensive ARB alignments.

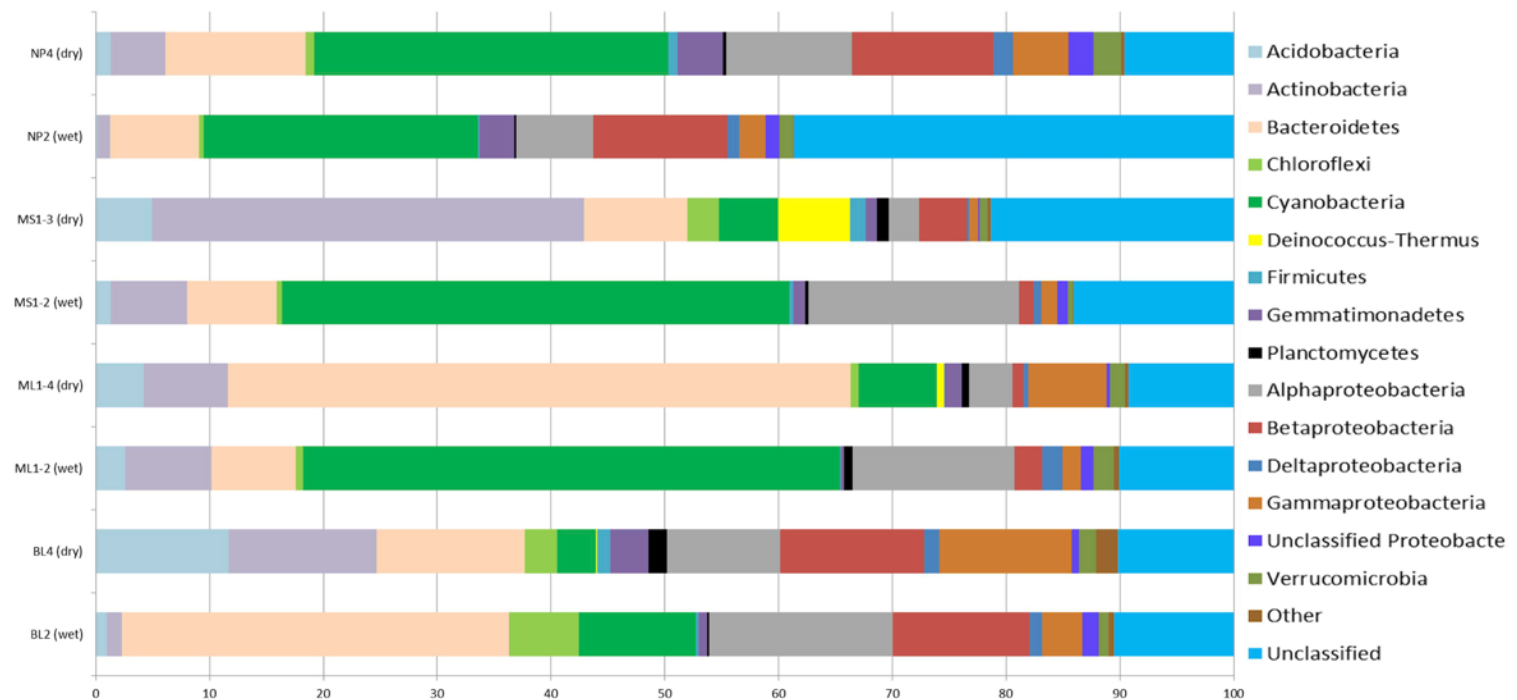
#### SINA: The new SILVA (Web)Aligner

We developed a new aligner called **SINA** (SILVA INcremental Aligner) that is able to accurately align hundred thousands of sequences based on a curated SEED alignment. In a first step the aligner determines the next related sequences using an optimized Suffix Tree server. To find the optimal alignment for a new sequence up to 40 reference sequences are taken into account. While running, the system simulates the manual refinement process to optimize the result.

**Features of SINA:**

- Process and quality values are added to each sequence indicating e.g. the alignment quality
- Only minimum manual revision of the output alignment is required (e.g. no base-spreading at the ends)
- Improved alignment quality due to advanced alignment technology compared to e.g. the ARB Aligner ("Fastaligner")

SINA is also available [online](#) for small scale projects. More information about SINA as well as the corresponding publication can be found [here](#).



Bases de datos —> Perfil taxonómico

# Dos estrategias de análisis

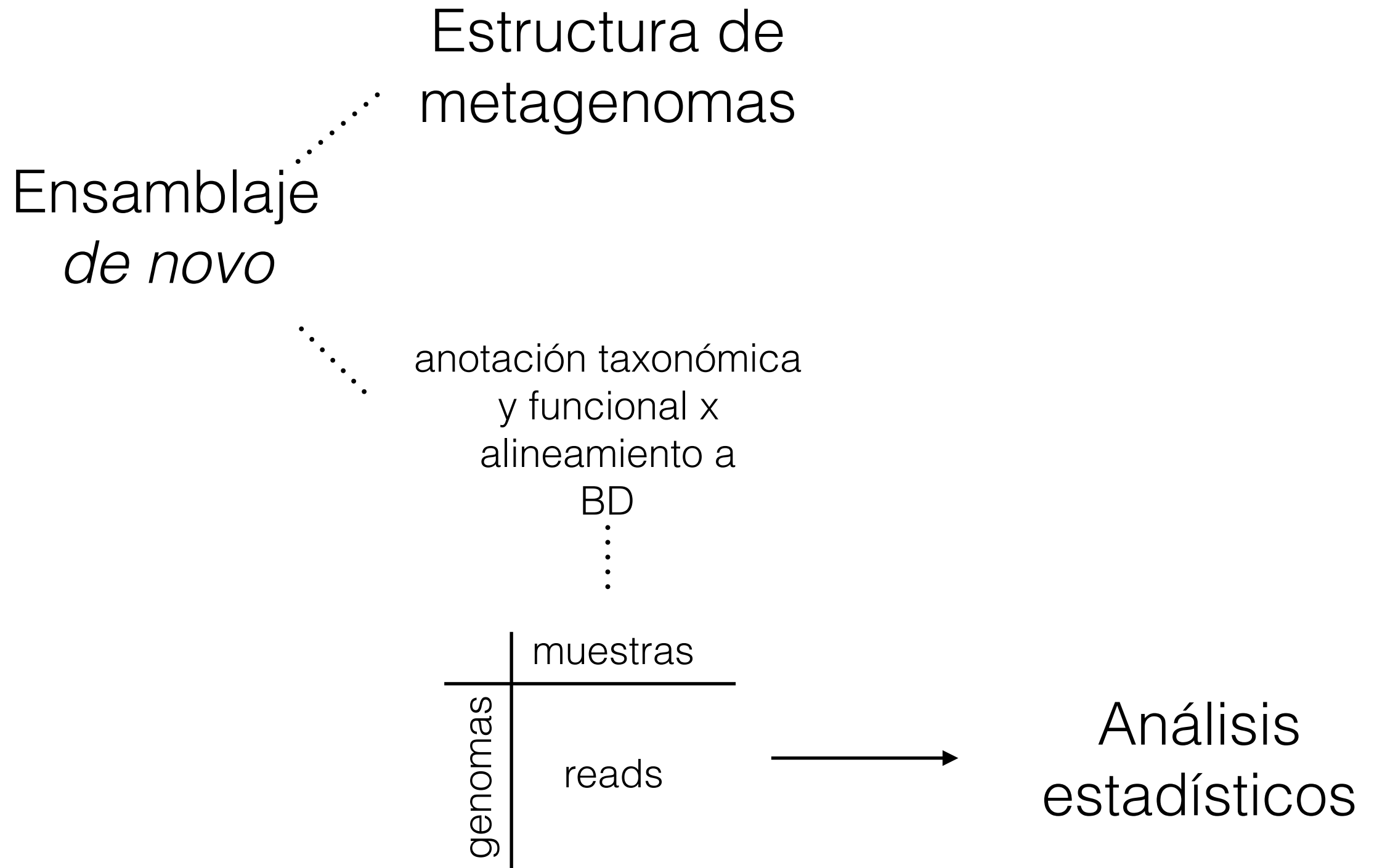
- Basado en ensamblaje *de novo*
- Basado en mapeo en contra de referencias

# Estrategias analíticas para datos de metagenomas

# Estrategias analíticas para datos de metagenomas

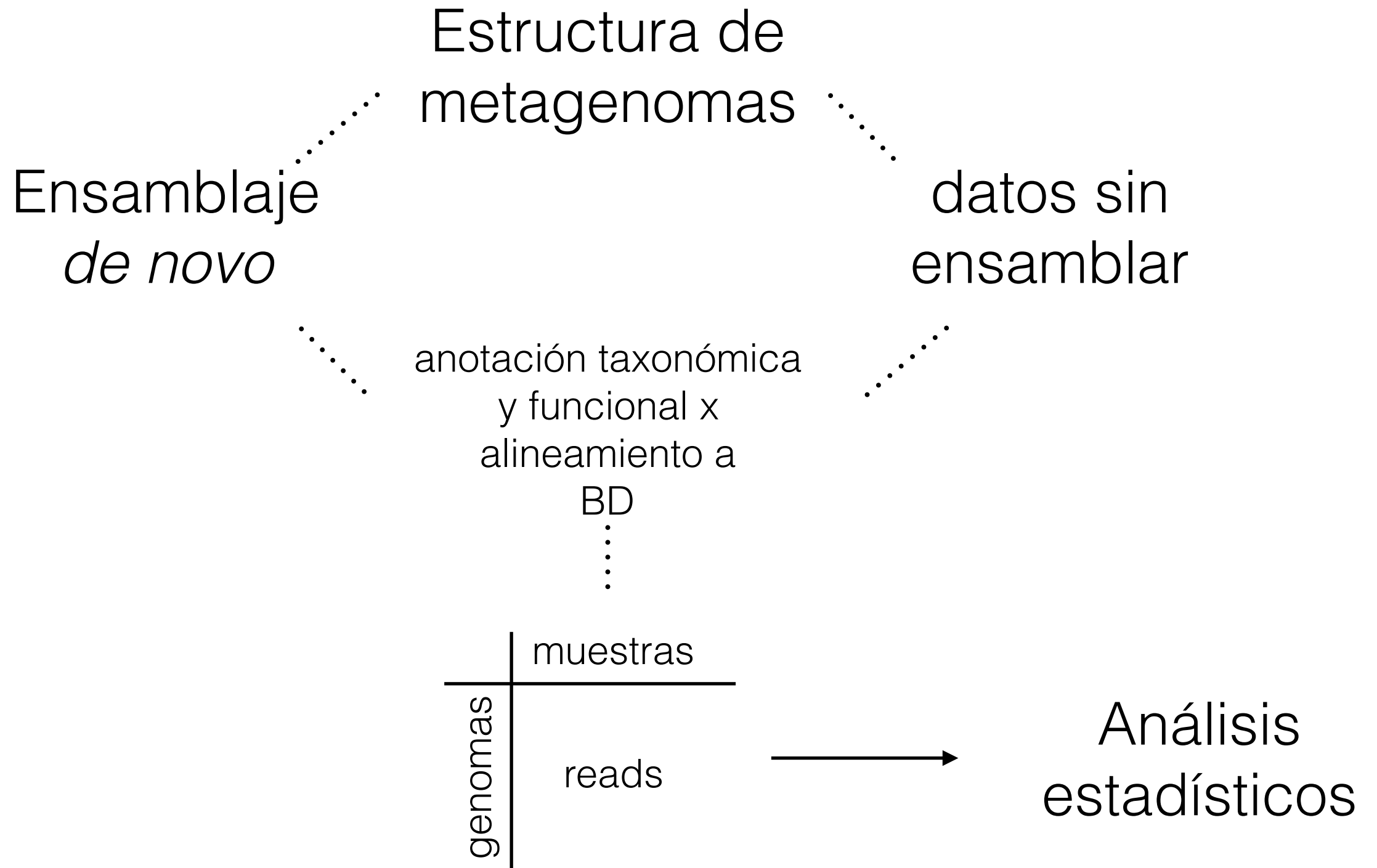
Estructura de  
metagenomas

# Estrategias analíticas para datos de metagenomas

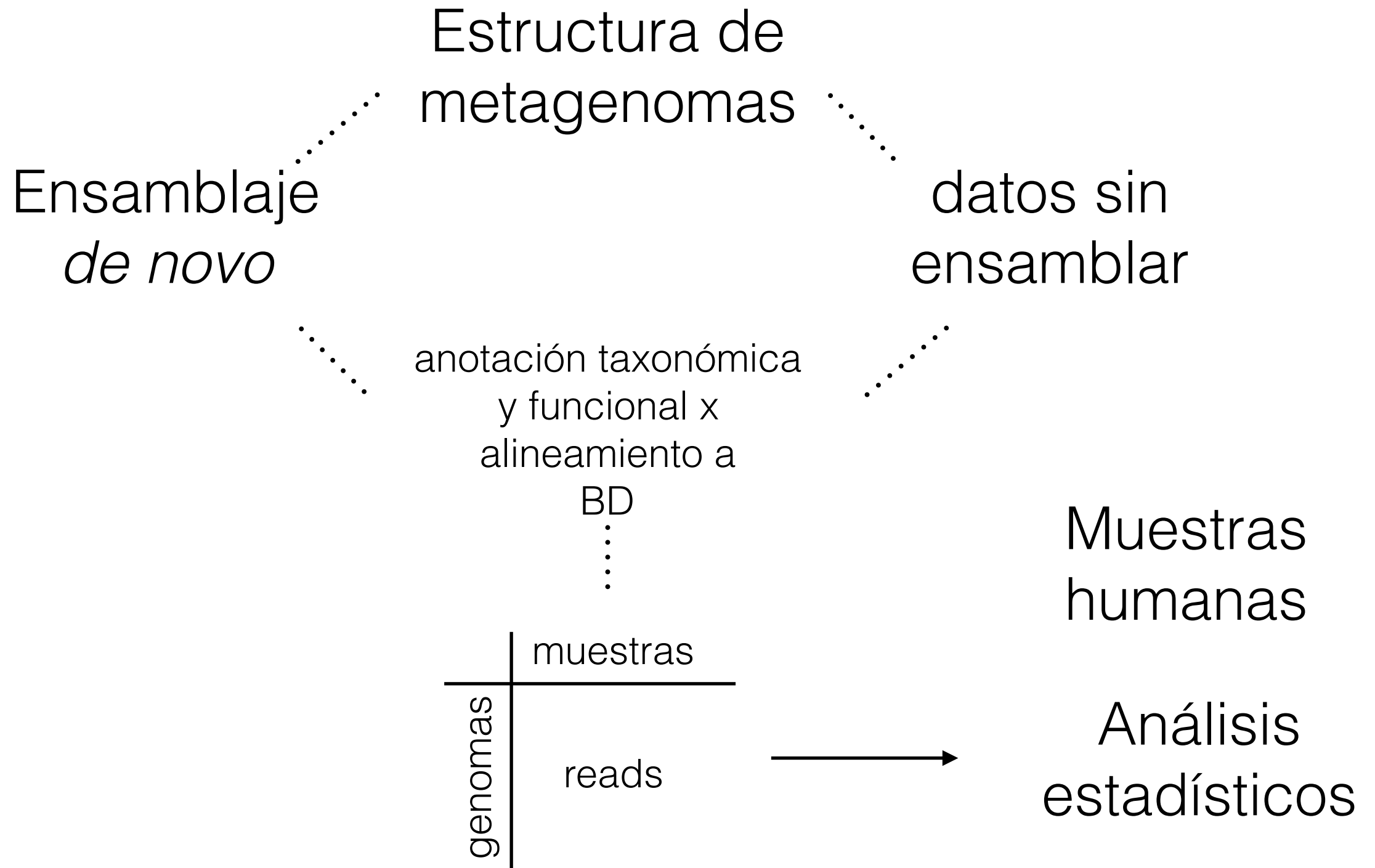




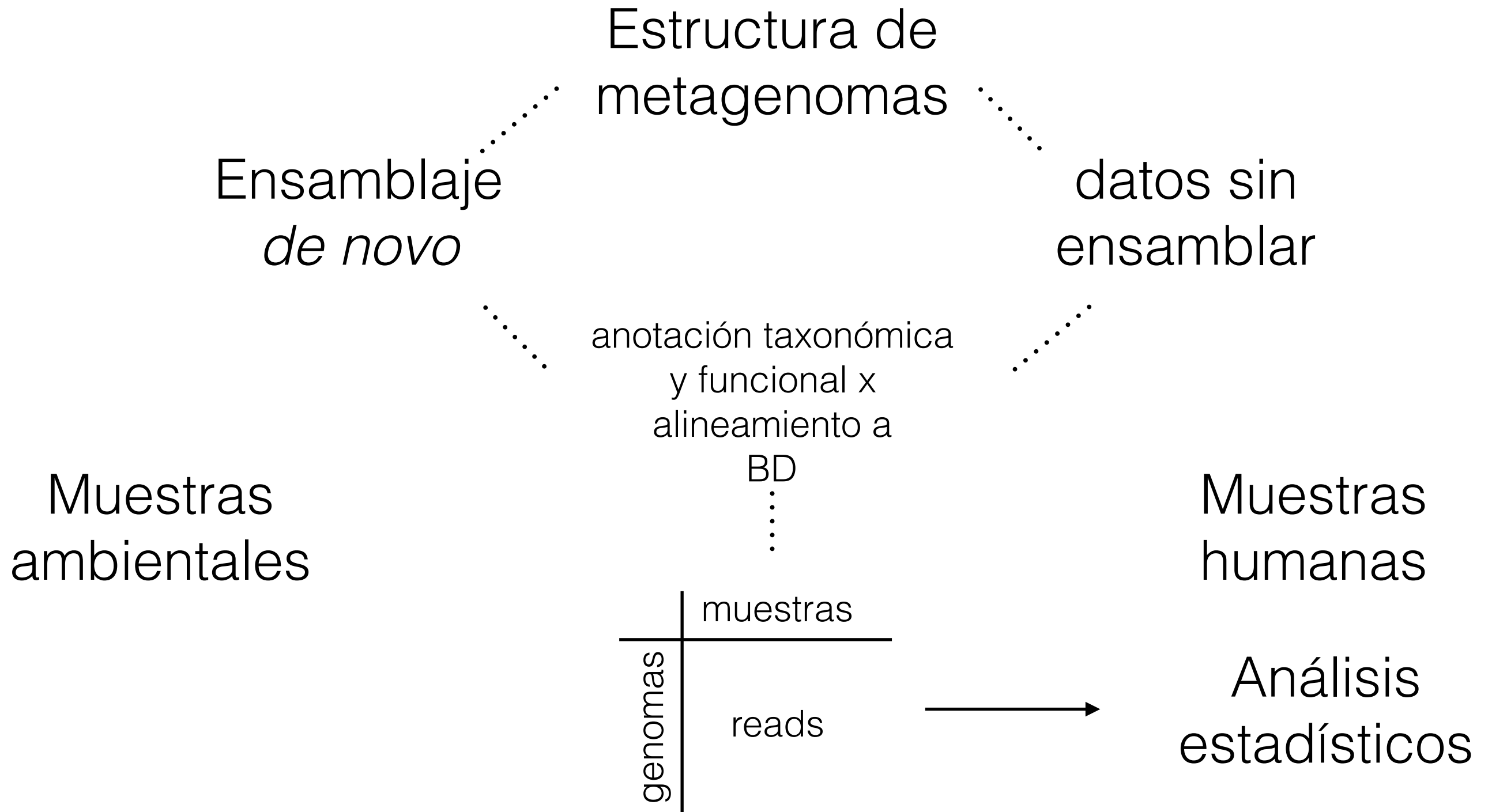
# Estrategias analíticas para datos de metagenomas



# Estrategias analíticas para datos de metagenomas



# Estrategias analíticas para datos de metagenomas



# Metagenómica

- Secuenciar todo el DNA (cromosomal, plasmidial, etc.)
- Generar un perfil de miembros del metagenóma
- Qué hay y en qué proporción

# Metatranscriptómica

- Secuenciar todo el RNA
- Generar un perfil de expresión de genes en la comunidad microbiana
- Qué genes se expresan y en qué medida

# ¿Podemos secuenciar virus desde metagenomas?

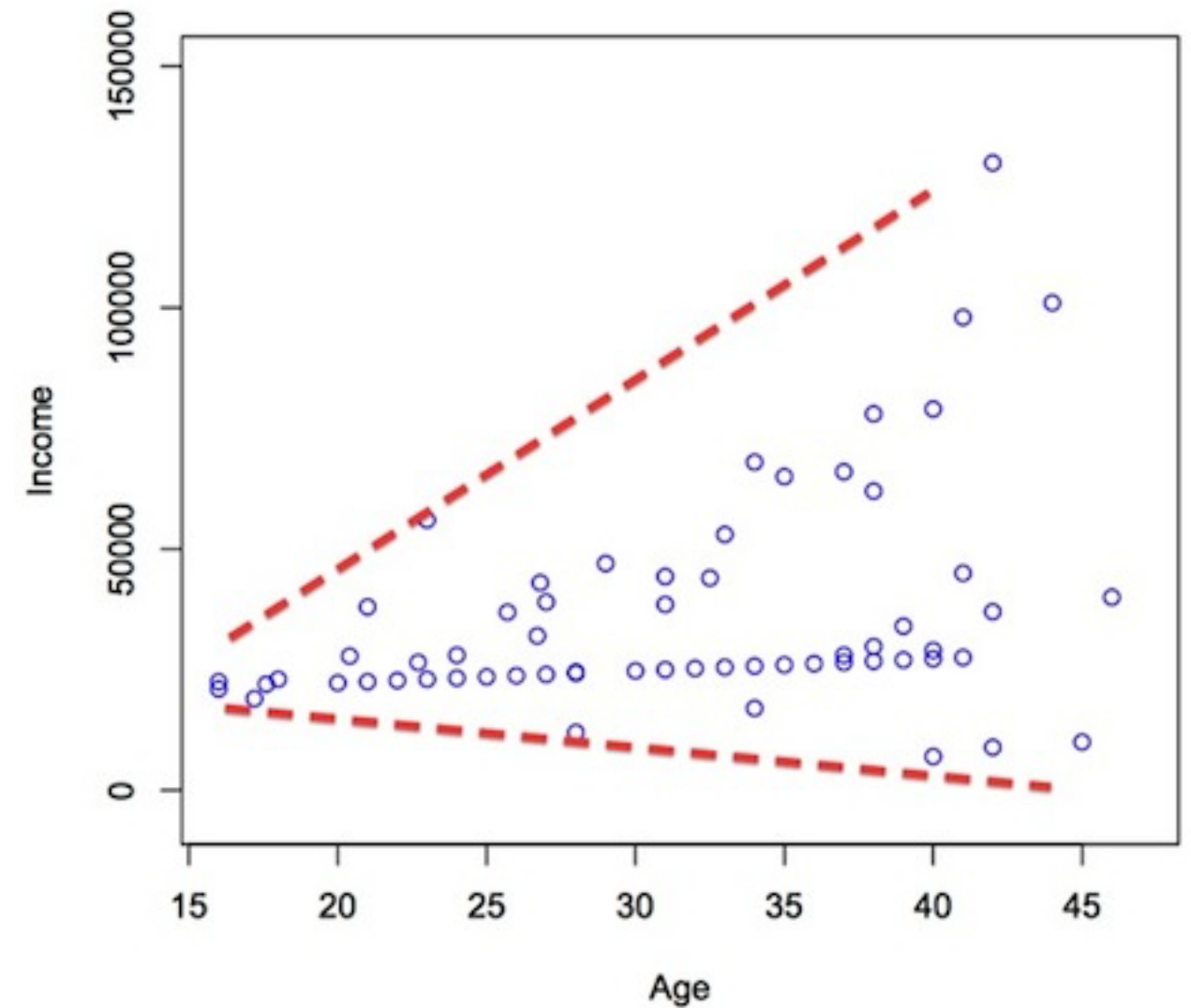
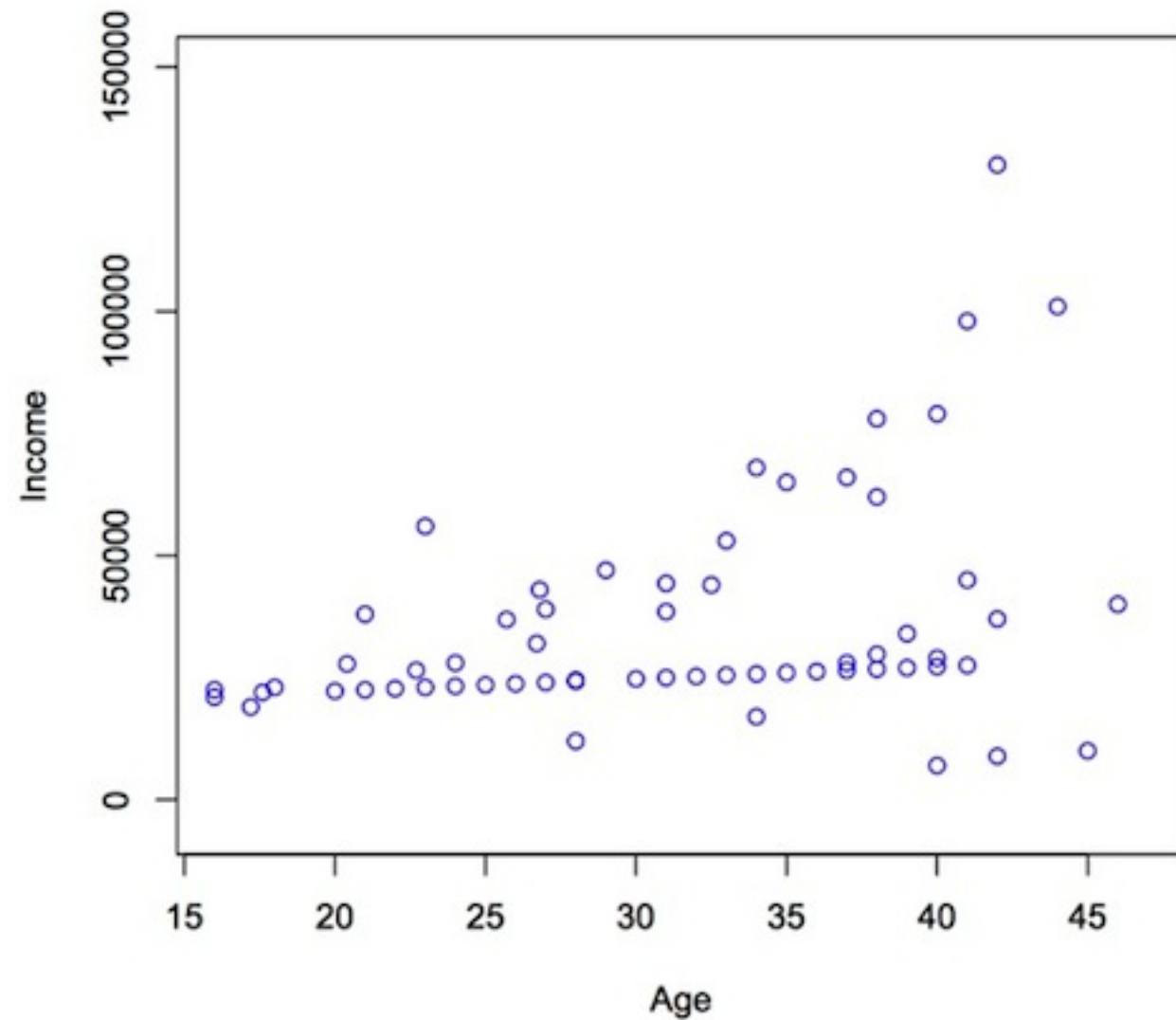
- Genoma bacteriano, 5 Mb; genoma viral 10 Kb
- Si tenemos un genoma bacteriano y un genoma viral, ¿cuántos fragmentos de 400 bp van a generar cada genoma?

# Tarea

- Heteroscedasticity
- Rarefaction
- Negative Binomial distribution
- Mixture Model

# Heteroscedasticity

La variabilidad de la variable dependiente se amplía o contrae en función de la variable independiente





# ¿Por qué no calculamos proporciones simples?

- Genomas que son muy abundantes pueden distorsionar la proporción total de reads

# Normalización entre muestras

Gene	Control Counts	Treatment Counts	Control Normalized	Treatment Normalized	
G1		2.00	6.00	0.20	0.06
G2		2.00	6.00	0.20	0.06
G3		2.00	6.00	0.20	0.06
G4		2.00	6.00	0.20	0.06
FG		2.00	76.00	0.20	0.76

# Normalización entre muestras

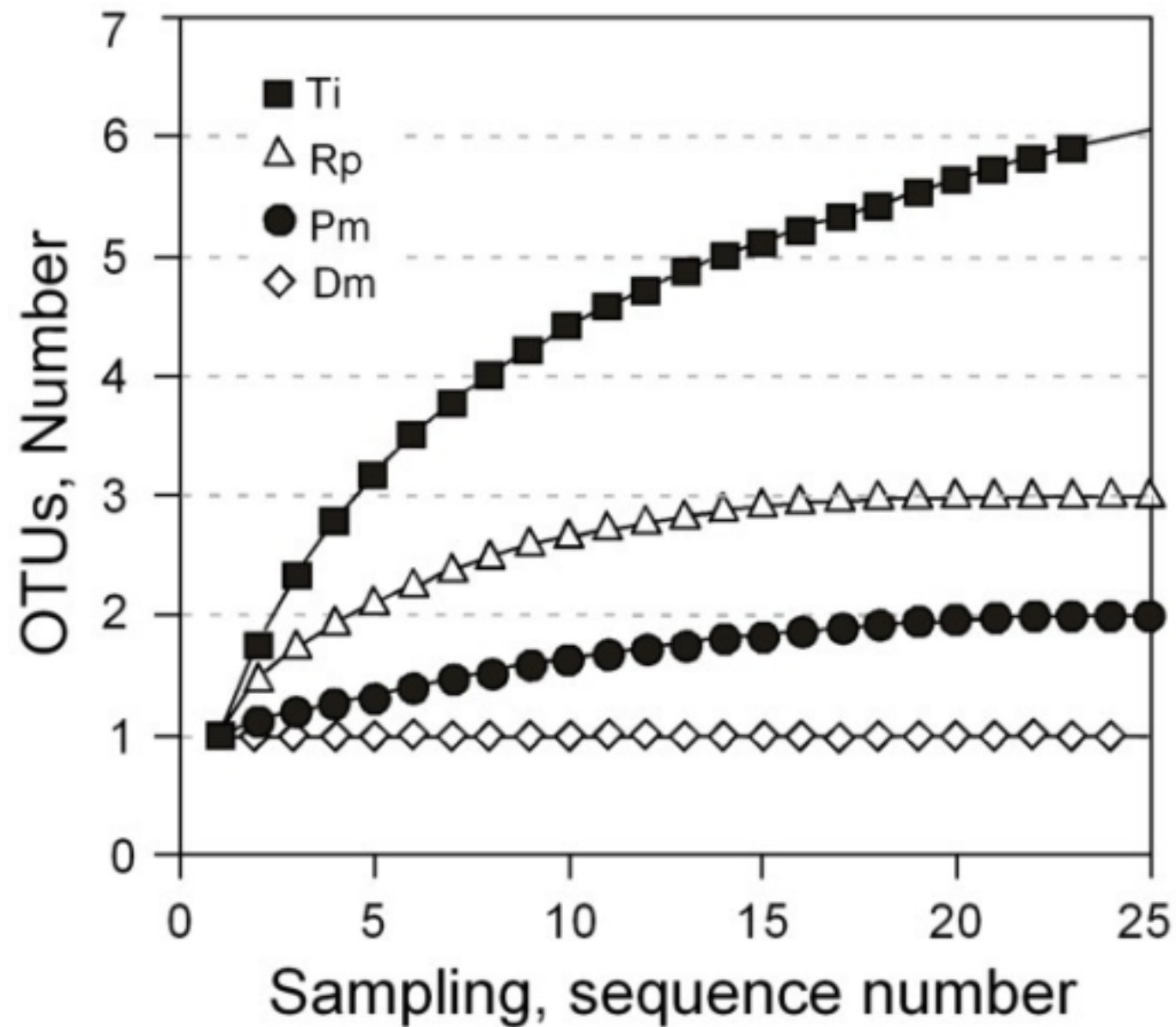
Gene	Control Counts	Treatment Counts	Control Normalized (-FG)	Treatment Normalized (-FG)
G1	2.00	6.00	0.25	0.25
G2	2.00	6.00	0.25	0.25
G3	2.00	6.00	0.25	0.25
G4	2.00	6.00	0.25	0.25
FG	2.00	76.00	0.25	3.17

- La clave está en encontrar un conjunto de genes entre las muestras que sirvan para normalizar todas las muestras

# Entonces ¿cómo normalizamos?

- Tomamos una referencia virtual
- Calculamos la media de las cuentas a través de todas las muestras
- Normalizamos cada muestra con respecto a un factor (scale factor, size factor)

# Rarefaction (disminuir en densidad)



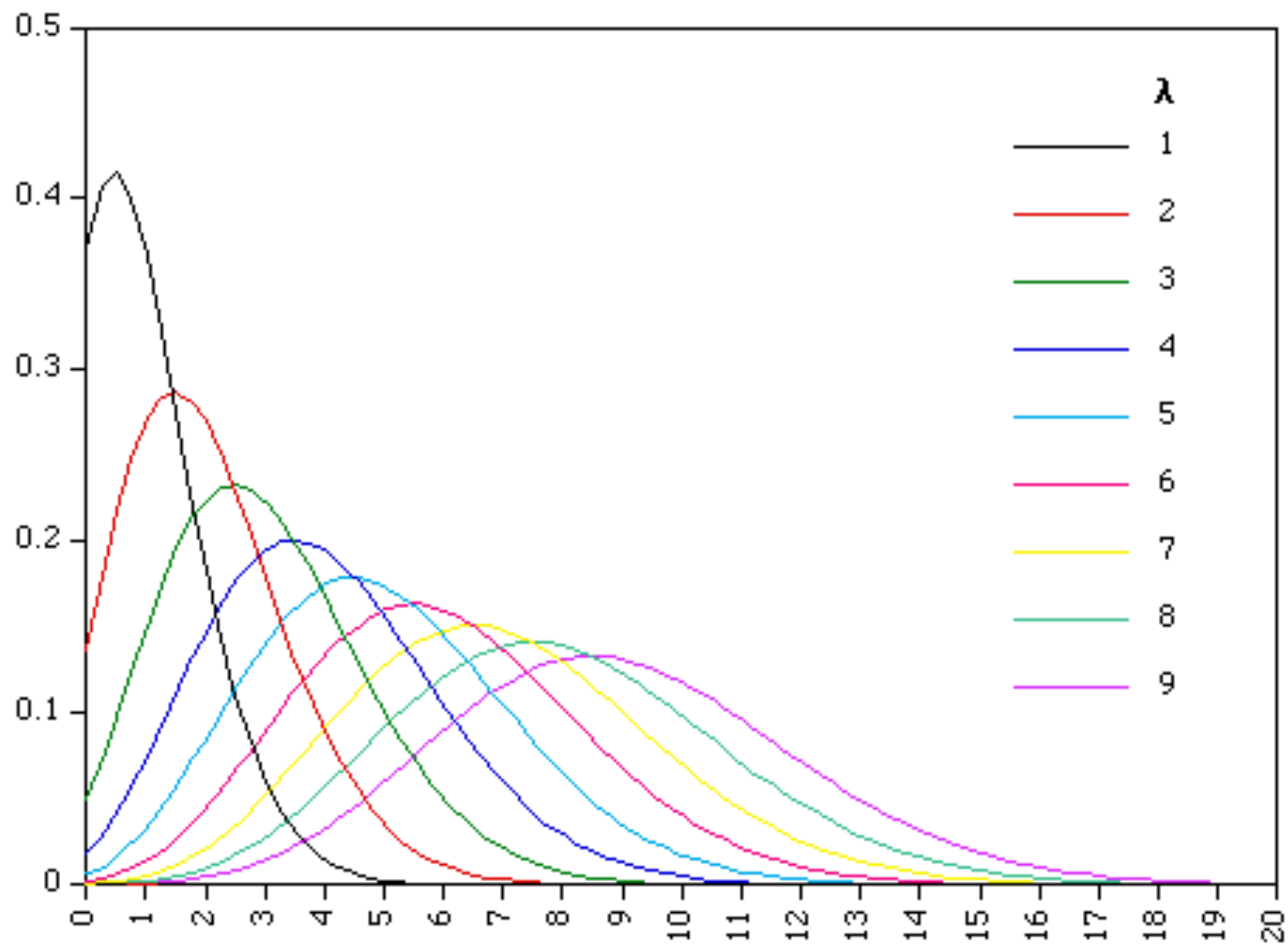
# Negative Binomial

- Es una generalización de la distribución de Poisson pero con dos parámetros
- Asumimos que los datos de RNASeq o de Metagenómica siguen este tipo de distribución
- Con esto podemos hacer una prueba de hipótesis para determinar qué genes (RNASeq) o genomas (metagenómica) están diferencialmente expresados o abundantes

# Negative Binomial

- Una distribución de Poisson modela la tasa de que algo ocurra
- Mientras la tasa de que algo ocurra es alta, la probabilidad será también alta
- Tiene un solo valor, la media

# Poisson distribution





# Mixture Model

- Modelo probabilístico usado para representar sub poblaciones dentro de una población
- Read counts de genes son como subpoblaciones
- Lo mismo para metagenomas, read counts de genomas son como subpoblaciones

# Abundancia diferencial de OTUs/genomas

- Análogo a Análisis Diferencial de Genes en RNASeq
- Se utiliza el modelo de DESeq2
- Se utiliza una ecuación lineal o de la recta, generalized linear model

# Abundancia diferencial de OTUs/genomas

## The DESeq2 model

The *DESeq2* model and all the steps taken in the software are described in detail in our publication [1], and we include the formula and descriptions in this section as well. The differential expression analysis in *DESeq2* uses a generalized linear model of the form:

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$$

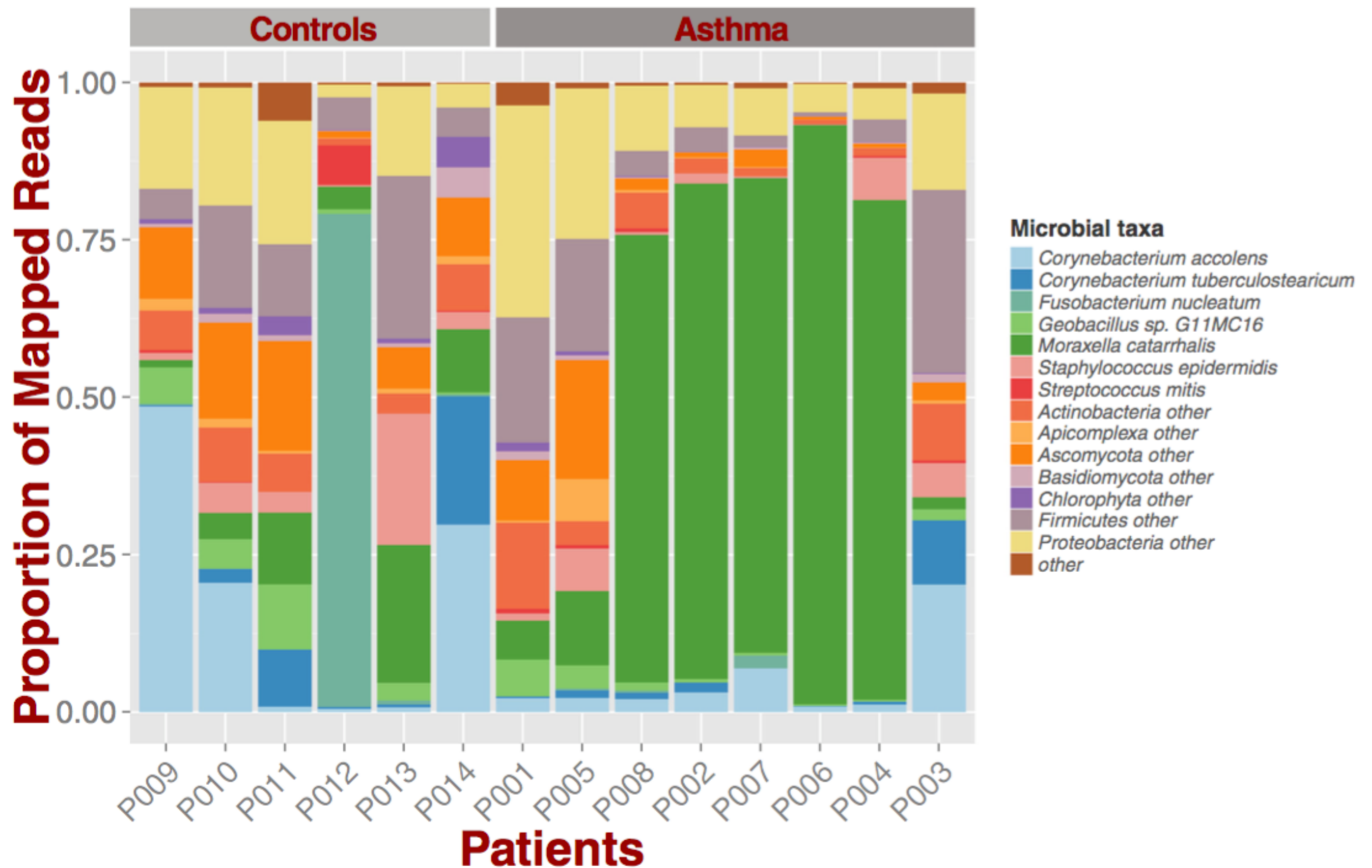
$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = x_j \beta_i$$

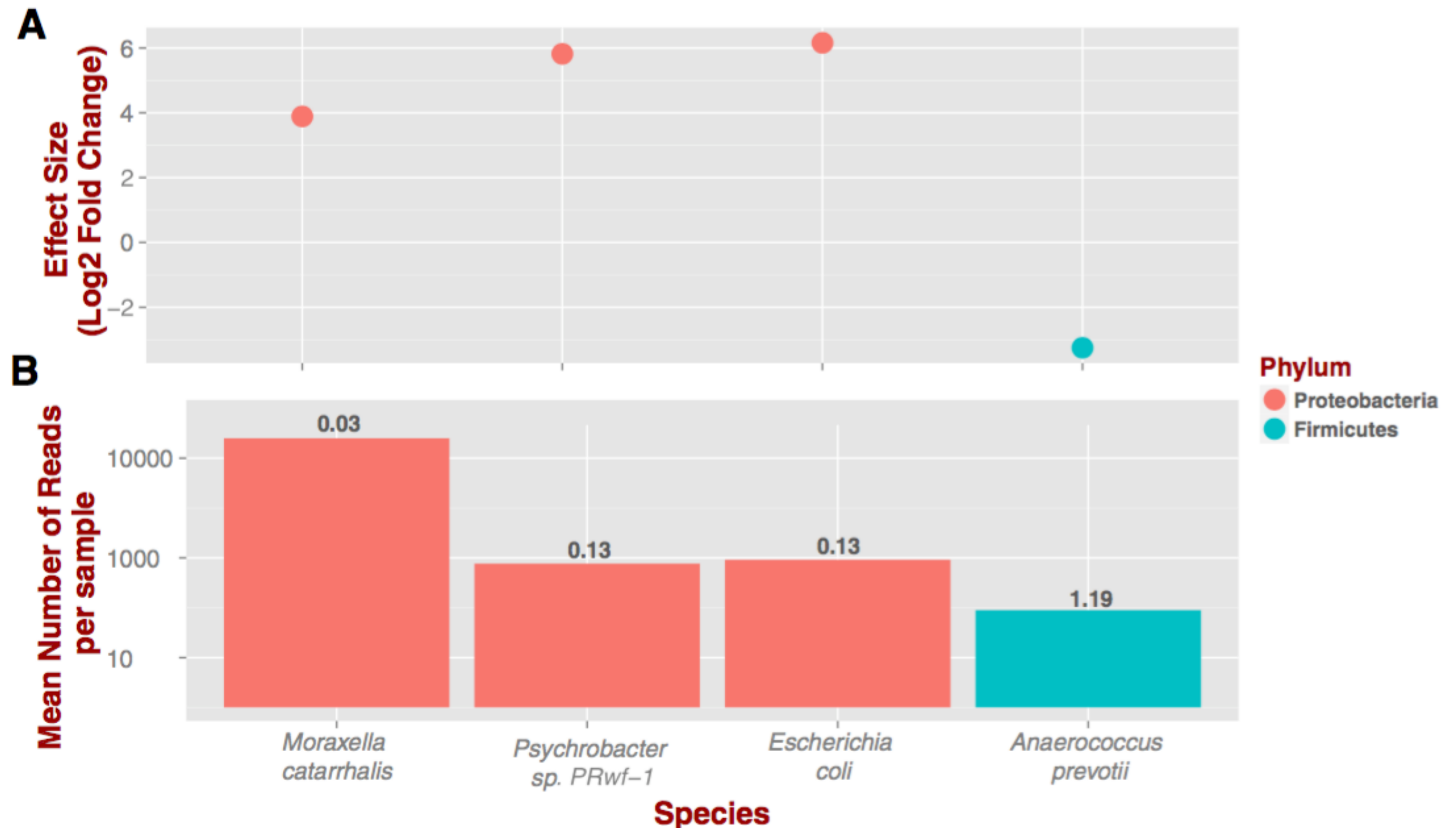
where counts  $K_{ij}$  for gene  $i$ , sample  $j$  are modeled using a negative binomial distribution with fitted mean  $\mu_{ij}$  and a gene-specific dispersion parameter  $\alpha_i$ . The fitted mean is composed of a sample-specific size factor  $s_j$ <sup>5</sup> and a parameter  $q_{ij}$  proportional to the expected true concentration of fragments for sample  $j$ . The coefficients  $\beta_i$  give the log2 fold changes for gene  $i$  for each column of the model matrix  $X$ .

The dispersion parameter  $\alpha_i$  defines the relationship between the variance of the observed count and its mean value. In other words, how far do we expected the observed count will be from the mean value, which depends both on the size factor  $s_j$  and the covariate-dependent part  $q_{ij}$  as defined above.

# Abundancia diferencial de OTUs/genomas



# Abundancia diferencial de OTUs/genomas



# Laboratorio de Genética de poblaciones

- <http://membres-timc.imag.fr/Olivier.Francois/tutoRstructure.pdf>
- La tarea es reproducir los análisis que se muestran en el tutorial, explicando en español cada paso y porqué
- Al igual que los otros labs, esto es parte del libro que van a producir