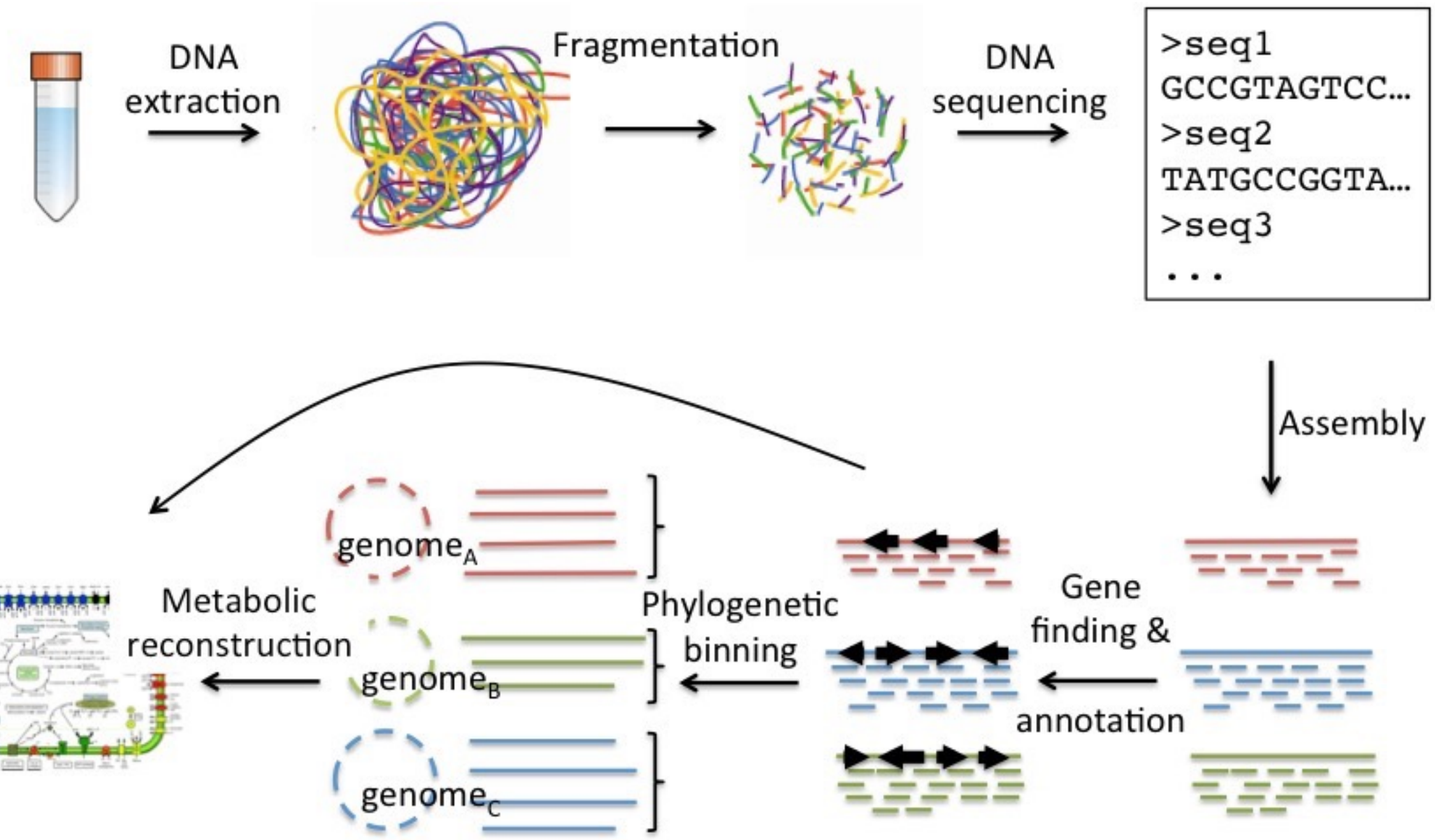
Metagenómica parte 2

Bioinformática Genómica para Ingeniería en Bioinformática
9 de agosto de 2016
Eduardo Castro-Nallar, PhD
Center for Bioinformatics and Integrative Biology
www.cbib.cl
www.castrolab.org

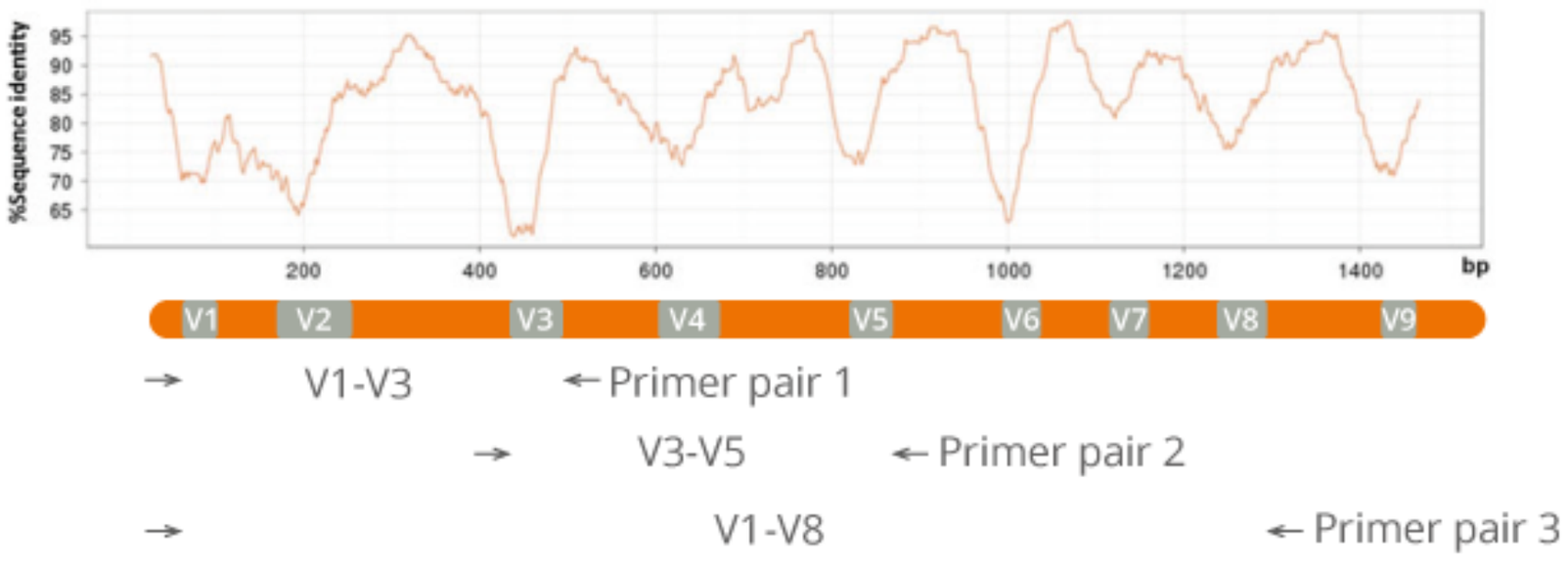
Shotgun Sequencing

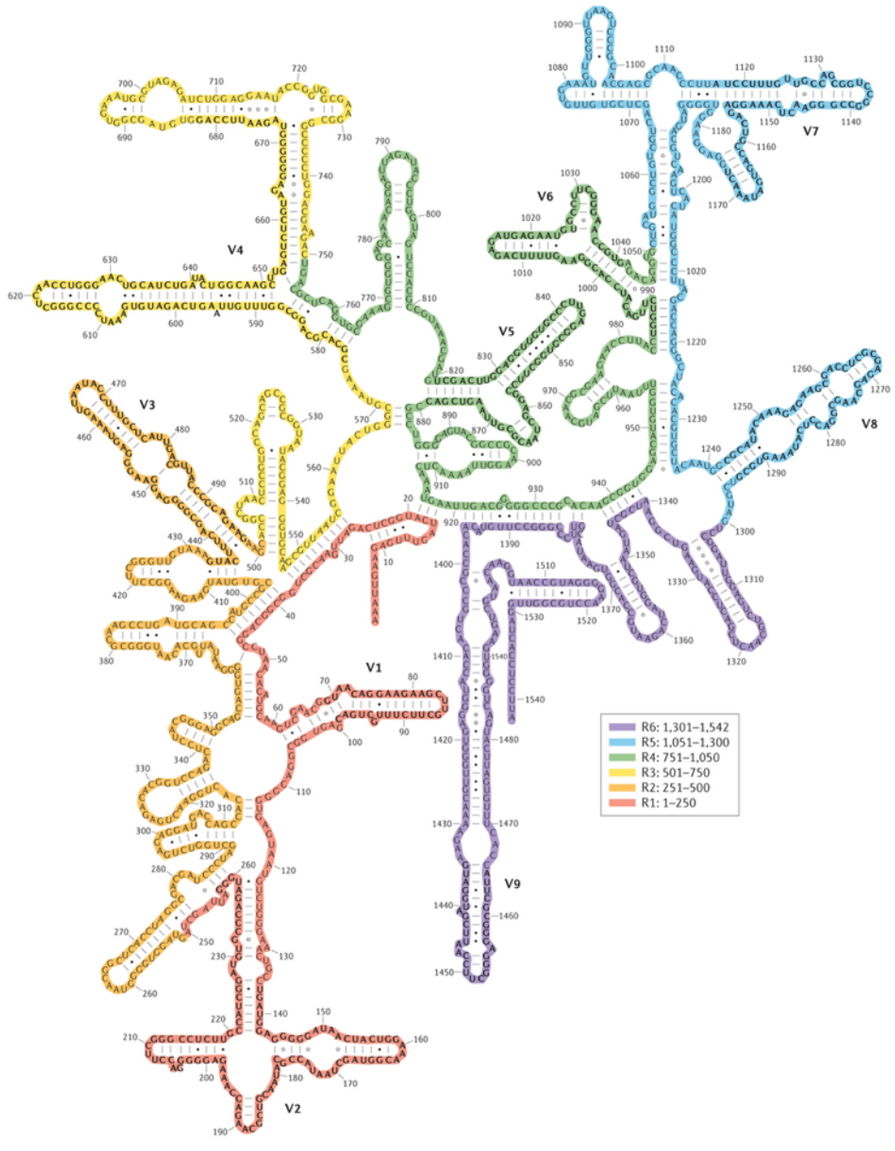


http://envgen.github.io/metagenomics.html

Metataxonómica

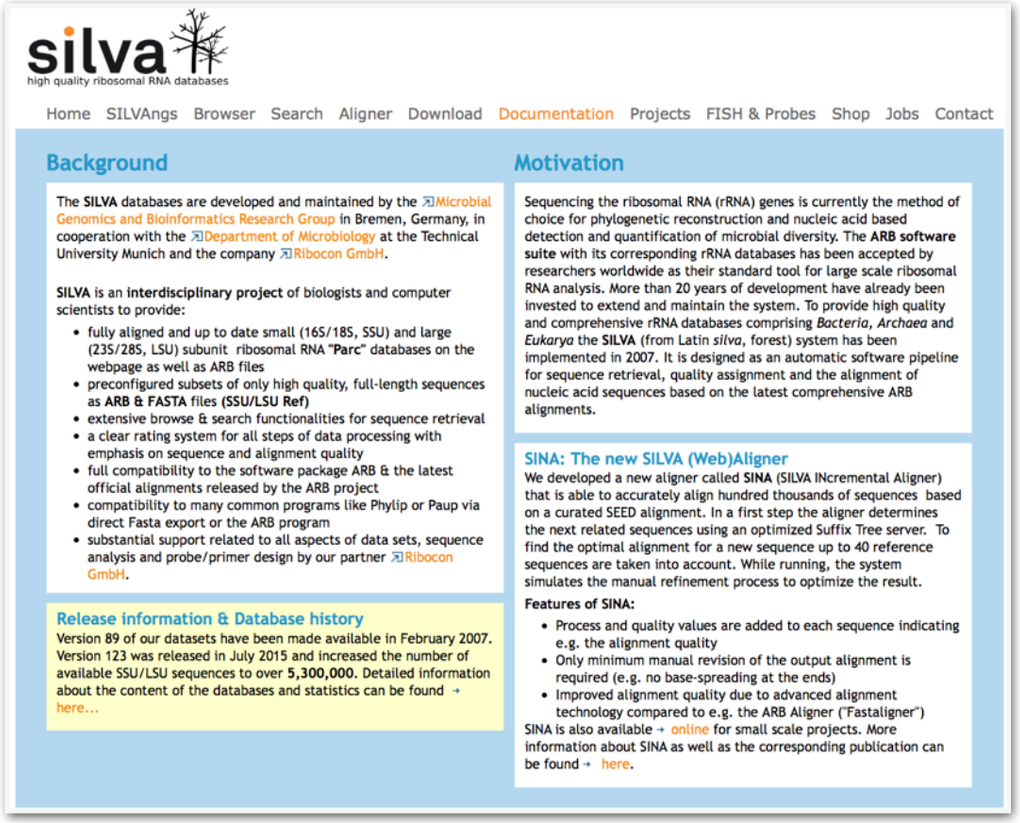
Regiones variables y constantes componen el gen 16S rRNA

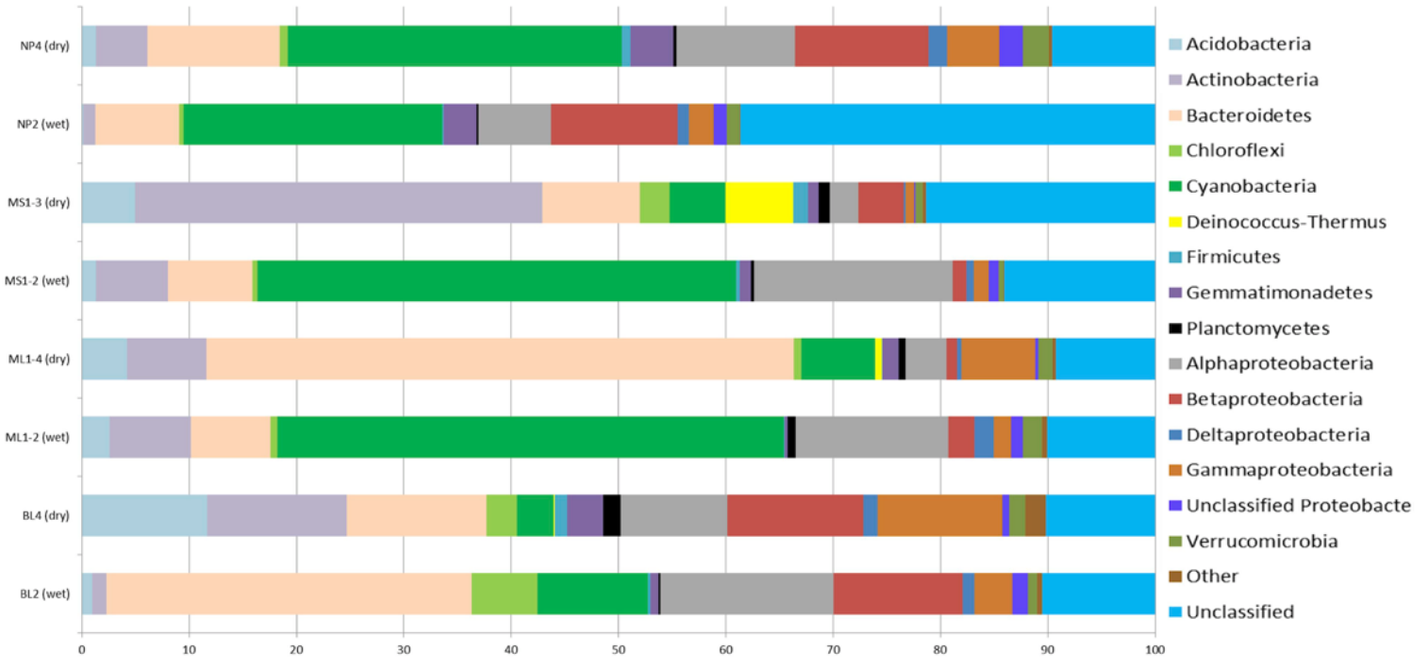




Nature Reviews | Microbiology

Metataxonómica





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

Estructura de metagenomas

Ensamblaje *de novo*

anotación taxonómica y funcional x alineamiento a BD

genomas \ muestras	muestras
genomas	reads

→ Análisis estadísticos

Estrategias analíticas para datos de metagenomas

Estructura de metagenomas

Ensamblaje *de novo*

anotación taxonómica y funcional x alineamiento a BD

datos sin ensamblar

	muestras
genomas	reads

Análisis estadísticos

Estrategias analíticas para datos de metagenomas

Estructura de metagenomas

Ensamblaje *de novo*

anotación taxonómica y funcional x alineamiento a BD

datos sin ensamblar

	muestras
genomas	reads

Muestras humanas

Análisis estadísticos

Estrategias analíticas para datos de metagenomas

Estructura de metagenomas

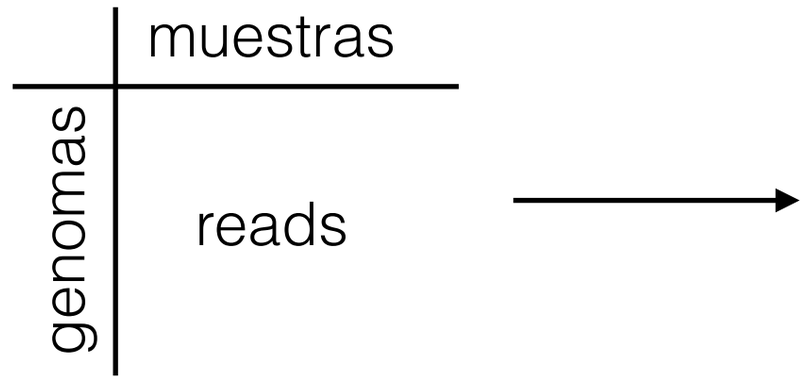
Ensamblaje *de novo*

datos sin ensamblar

anotación taxonómica y funcional x alineamiento a BD

Muestras ambientales

Muestras humanas



Análisis estadísticos

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?

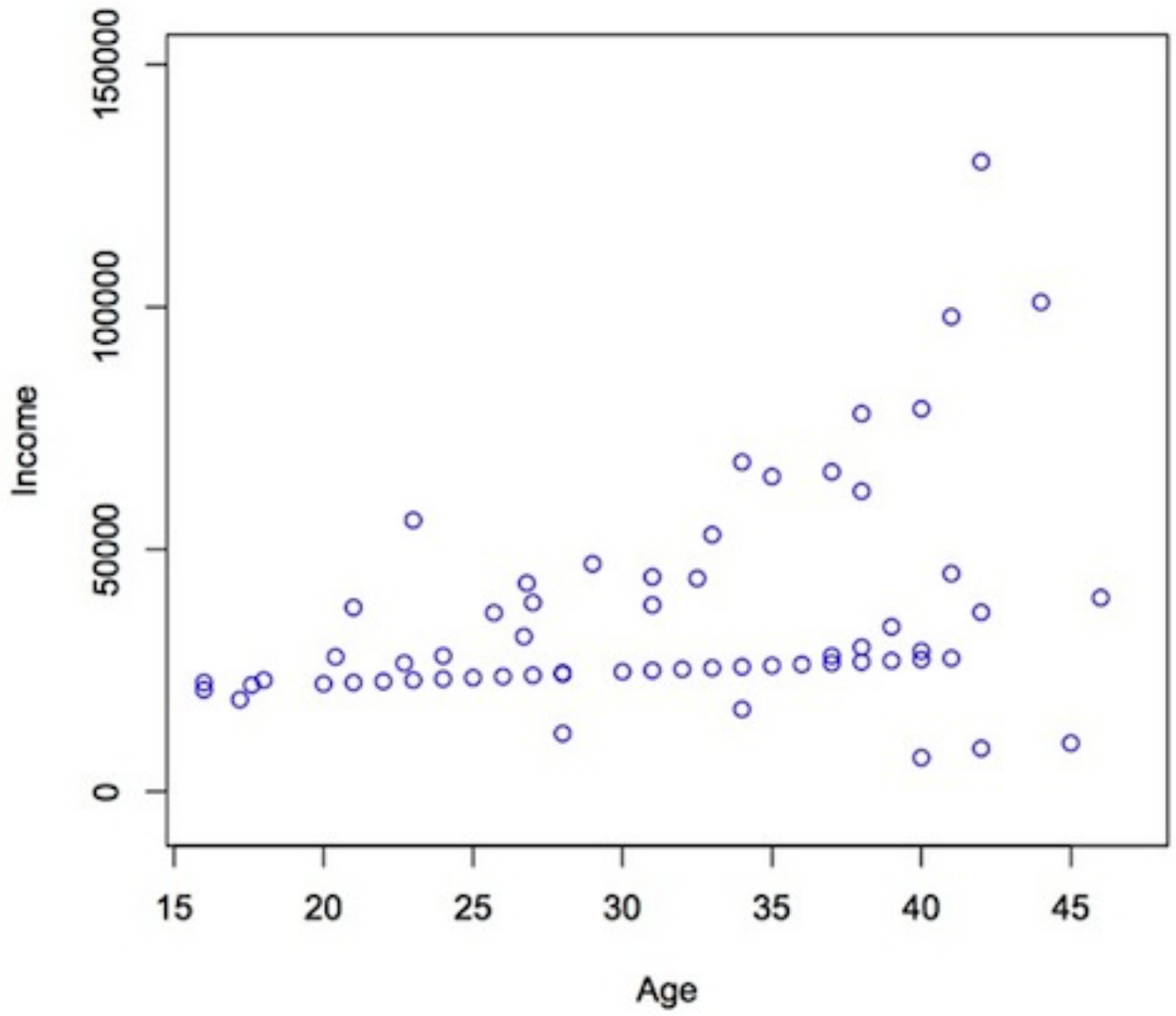
- Genoma bacteriano, 5 Mb; genoma viral 10 Kb
- Si tenemos un genoma bacteriano y un genoma viral, ¿cuántos fragmentos de 400 bp van a generar cada genoma?

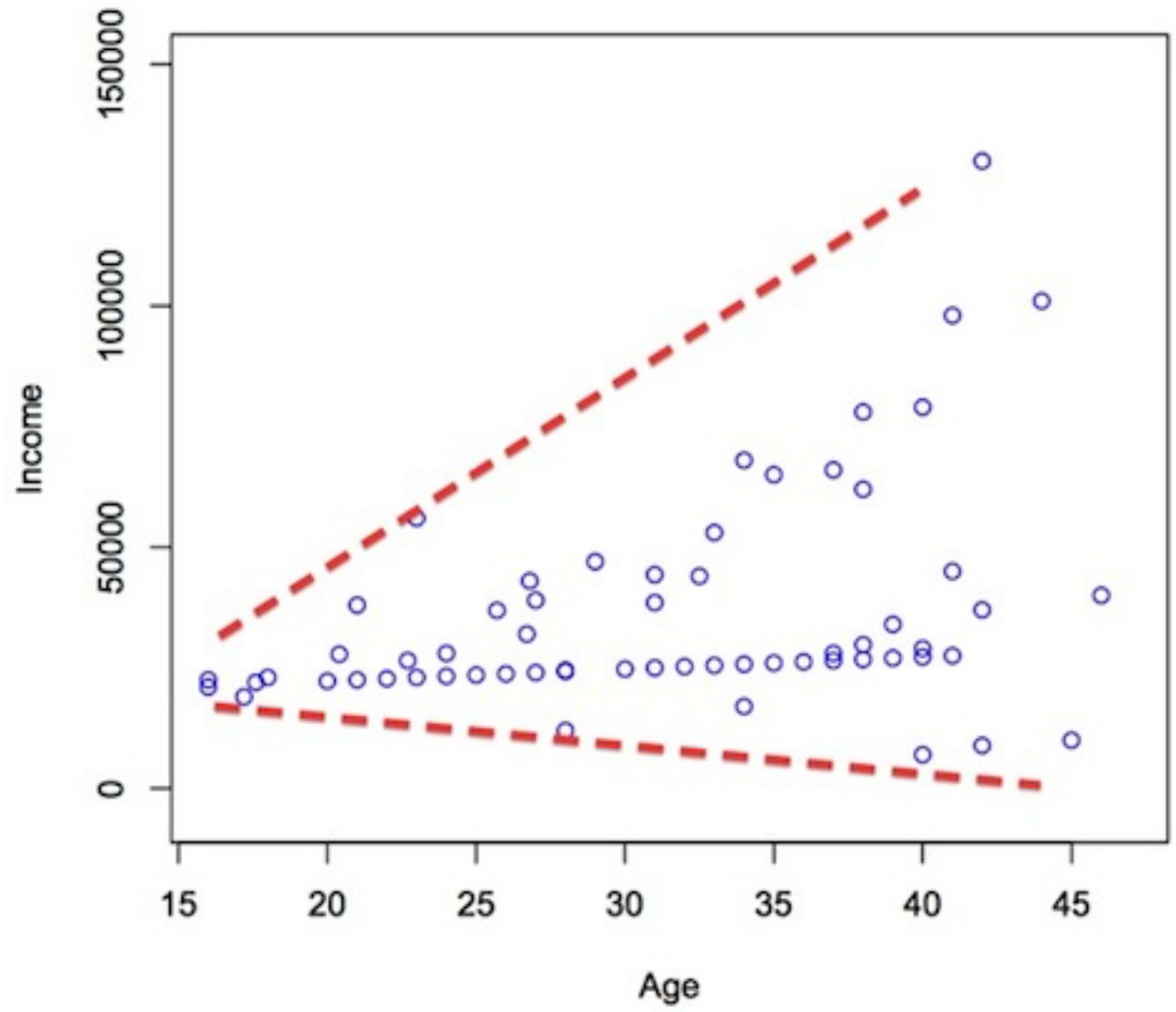
Tarea

- Heteroscedasticity
- Rarefaction
- Negative Binomial distribution
- Mixture Model

Heteroscedasticity

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





http://www.statsmakemecry.com/smmctheblog/confusing-stats-terms-explained-heteroscedasticity-heteroske.html

¿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

https://haroldpimentel.wordpress.com/2014/12/08/in-rna-seq-2-2-between-sample-normalization/#footnotes

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

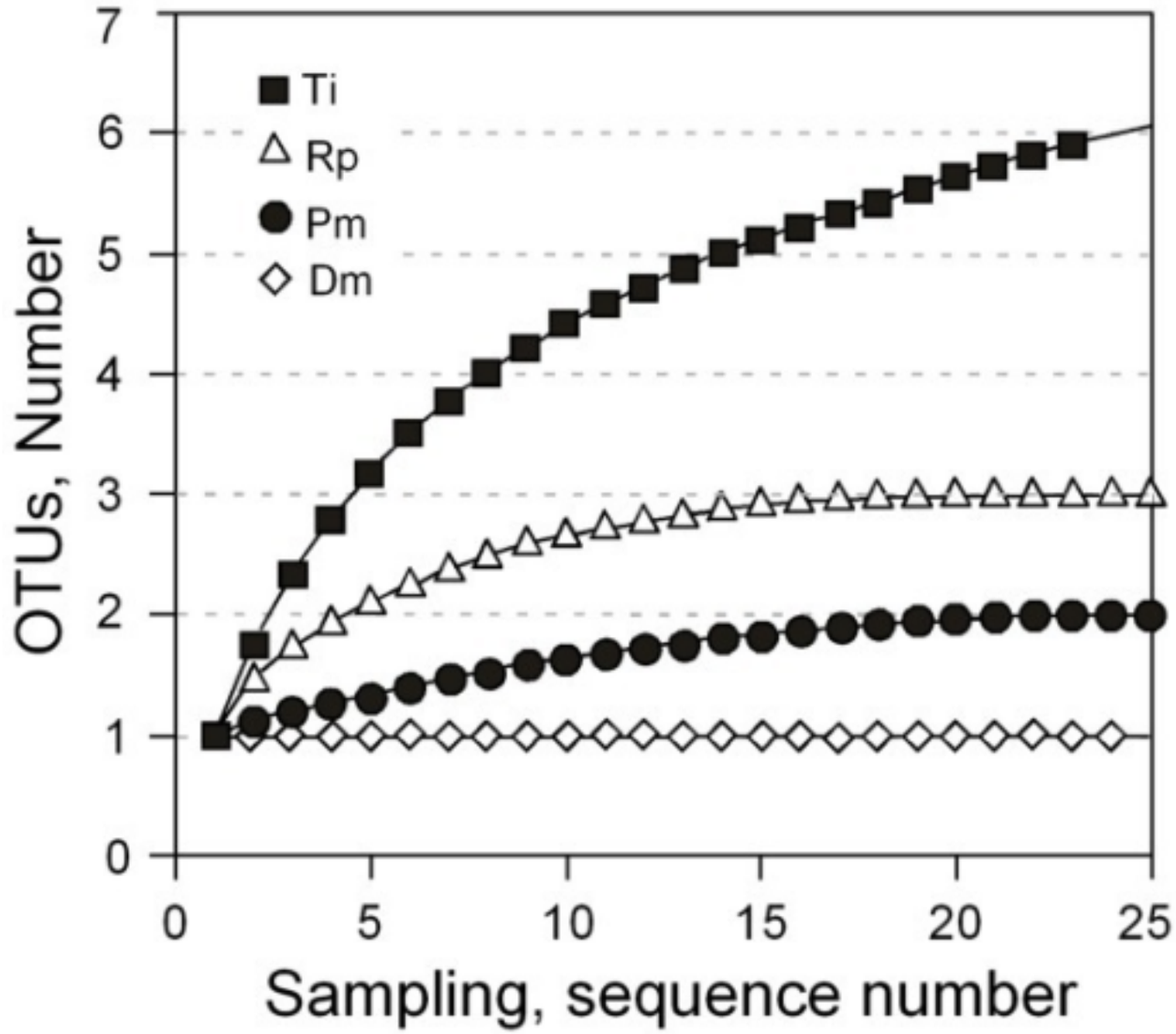
- La clave está en encontrar un conjunto de genes entre las muestras que sirvan para normalizar todas las muestras

https://haroldpimentel.wordpress.com/2014/12/08/in-rna-seq-2-2-between-sample-normalization/#footnotes

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

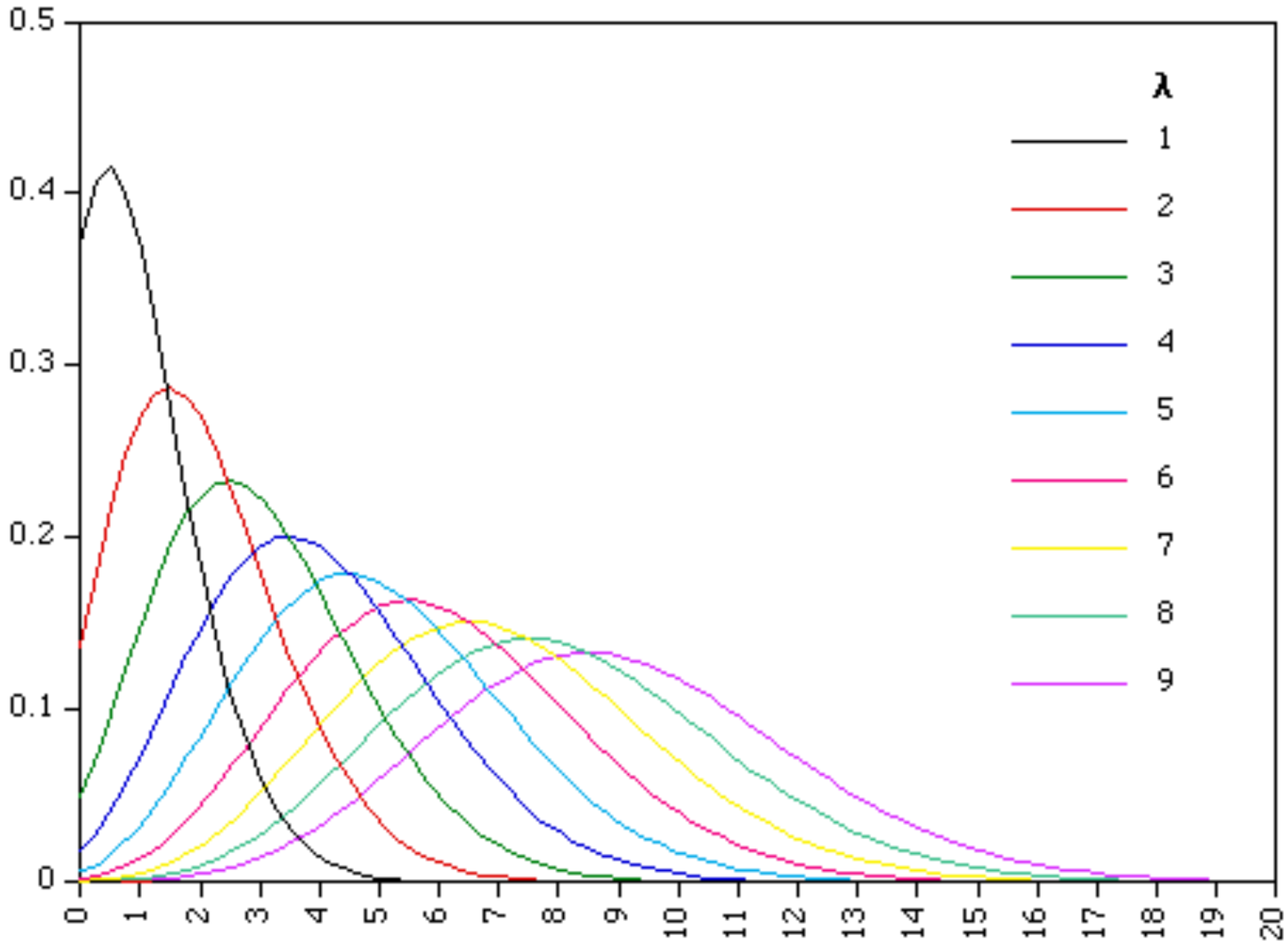
Negative Binomial

- 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:

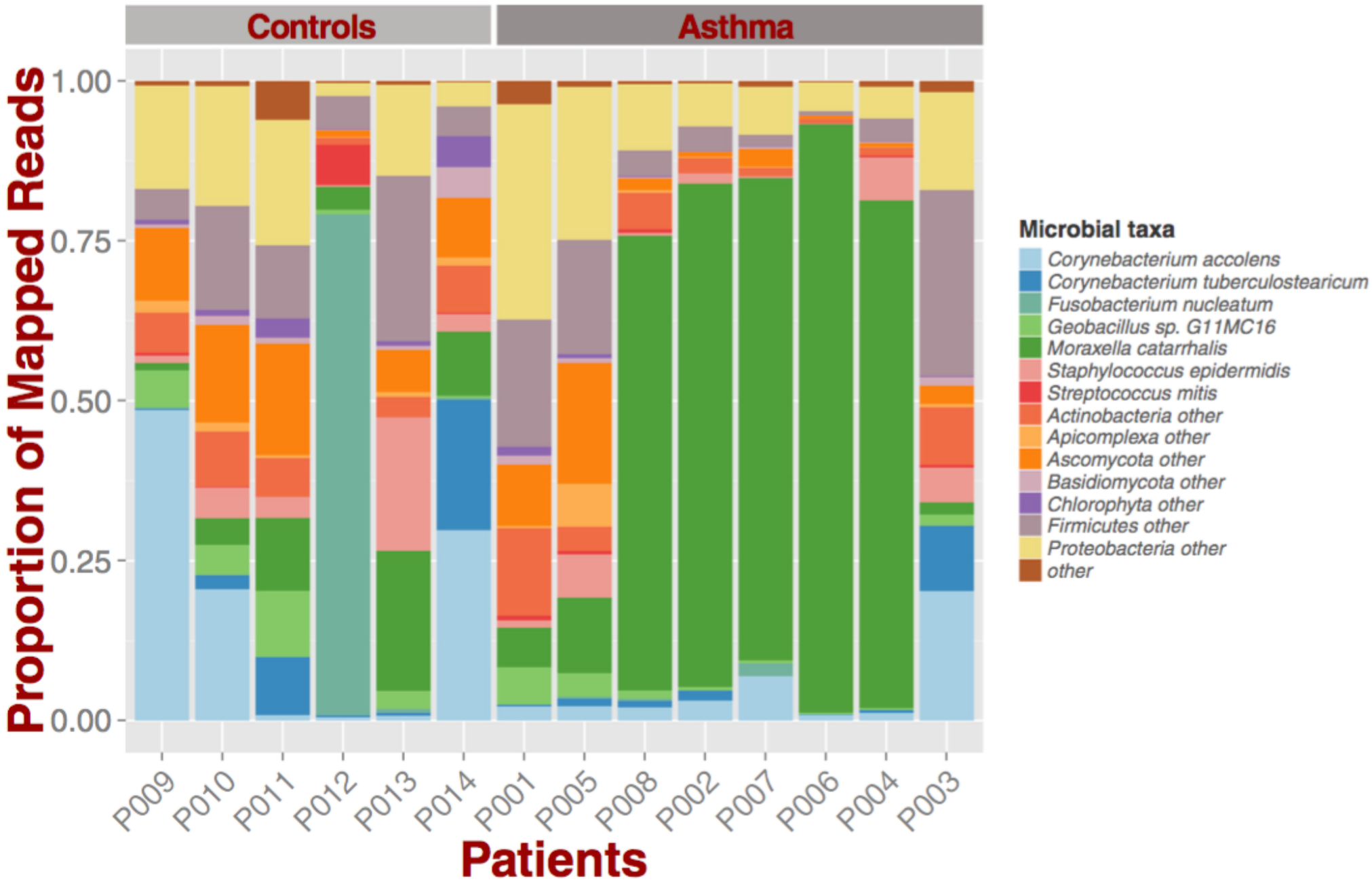
$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$
$$\mu_{ij} = s_j q_{ij}$$
$$\log_2(q_{ij}) = x_{j\cdot}\beta_i$$

where counts K_{ij} for gene i, sample j are modeled using a negative binomial distribution with fitted mean μ_{ij} and a gene-specific dispersion parameter α_i. The fitted mean is composed of a sample-specific size factor s_j[5] and a parameter q_{ij} proportional to the expected true concentration of fragments for sample j. The coefficients β_i give the log2 fold changes for gene i for each column of the model matrix X.

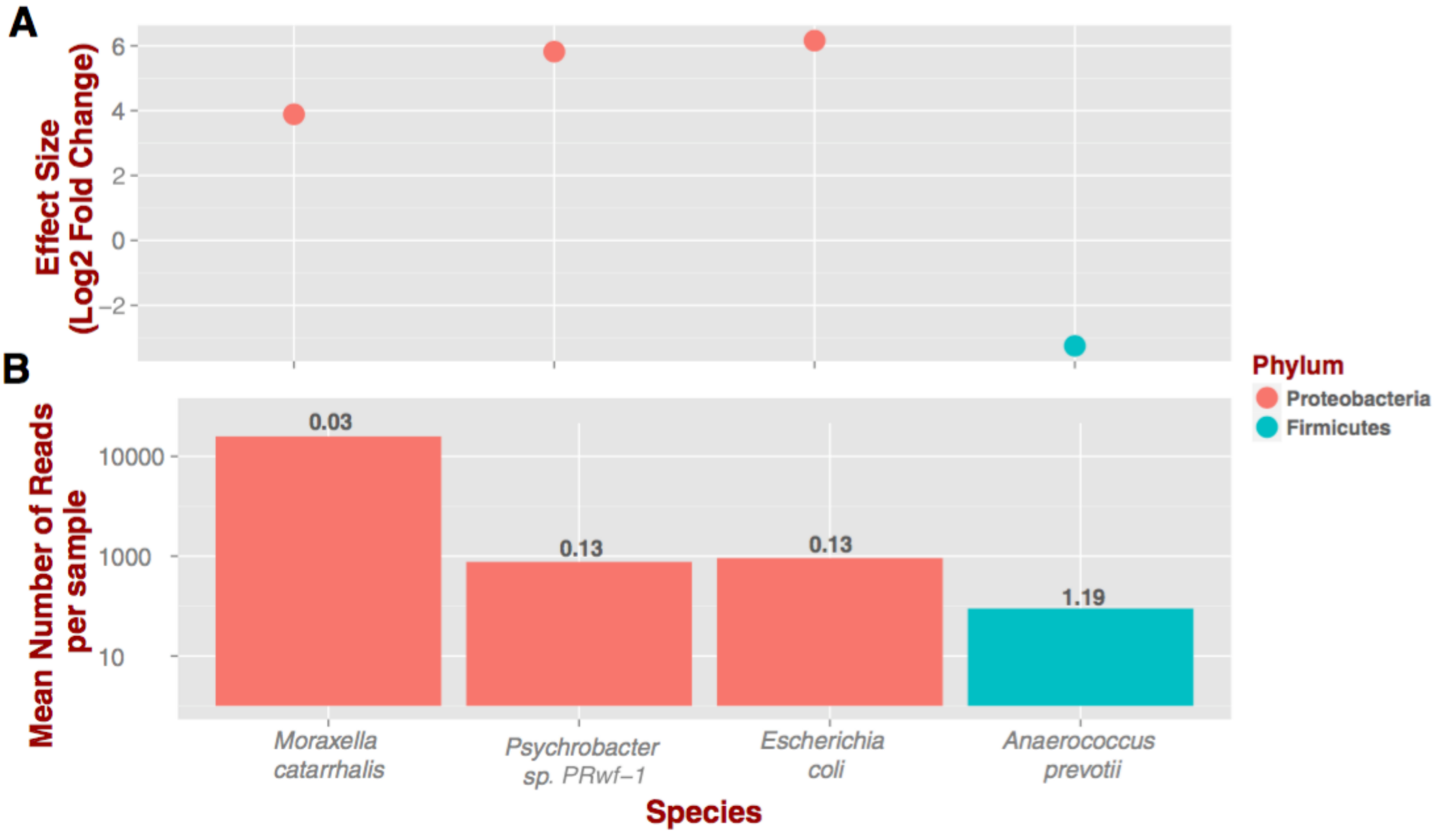
The dispersion parameter α_i defines the relationship between the variance of the observed count and its mean value. In other words, how far do we expected the observed count will be from the mean value, which depends both on the size factor s_j and the covariate-dependent part q_{ij} as defined above.

https://www.bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

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