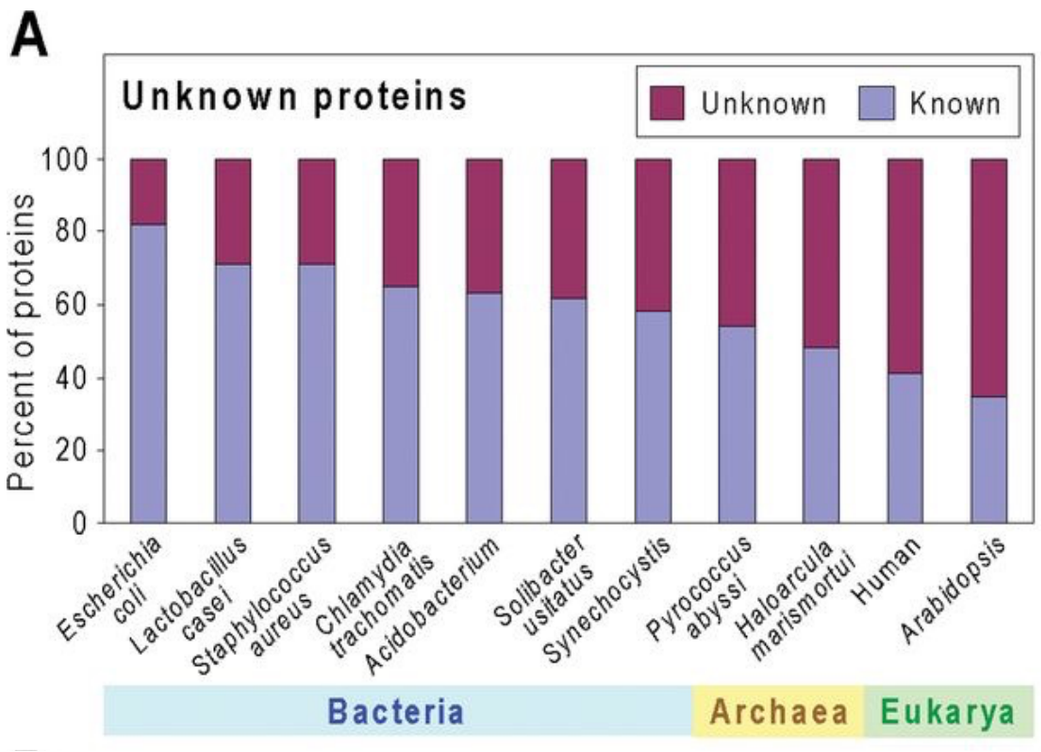
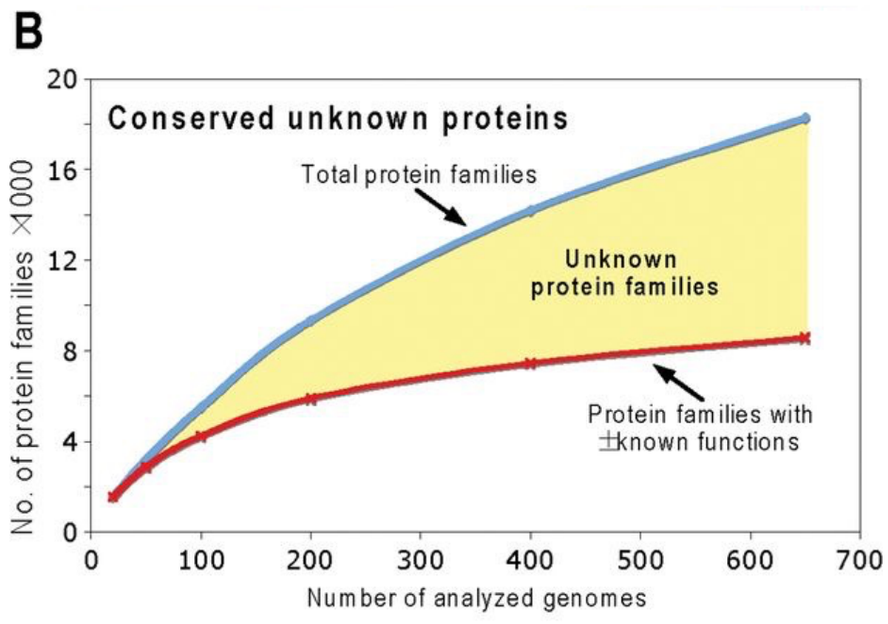
Análisis de expresión génica

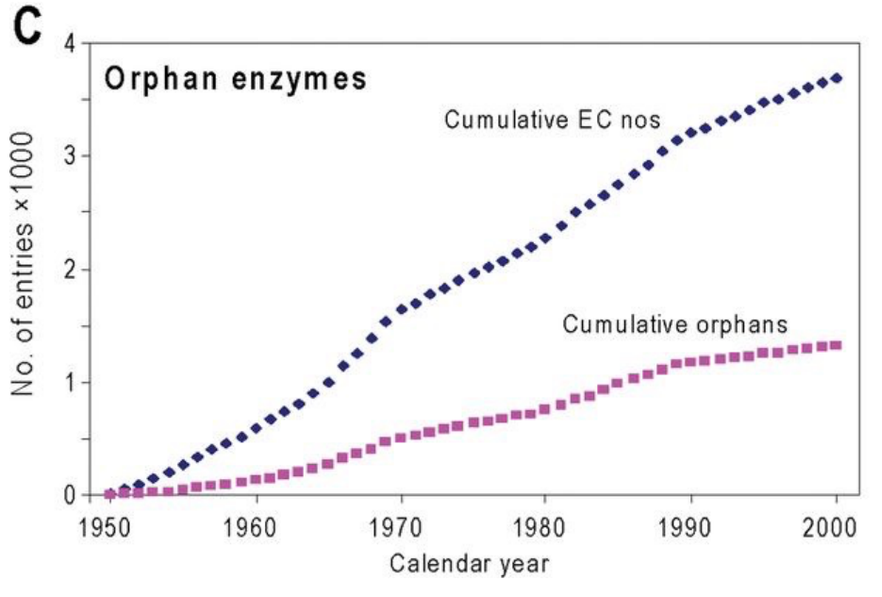
Bioinformática Genómica para Ingeniería en Bioinformática
19 de julio 2016
Eduardo Castro-Nallar, PhD
Center for Bioinformatics and Integrative Biology
www.cbib.cl
www.castrolab.org

La transcriptómica es el puente entre la variación genética y el mecanismo molecular

- Proteínas no caracterizadas (huérfanas; no se parecen a nada)

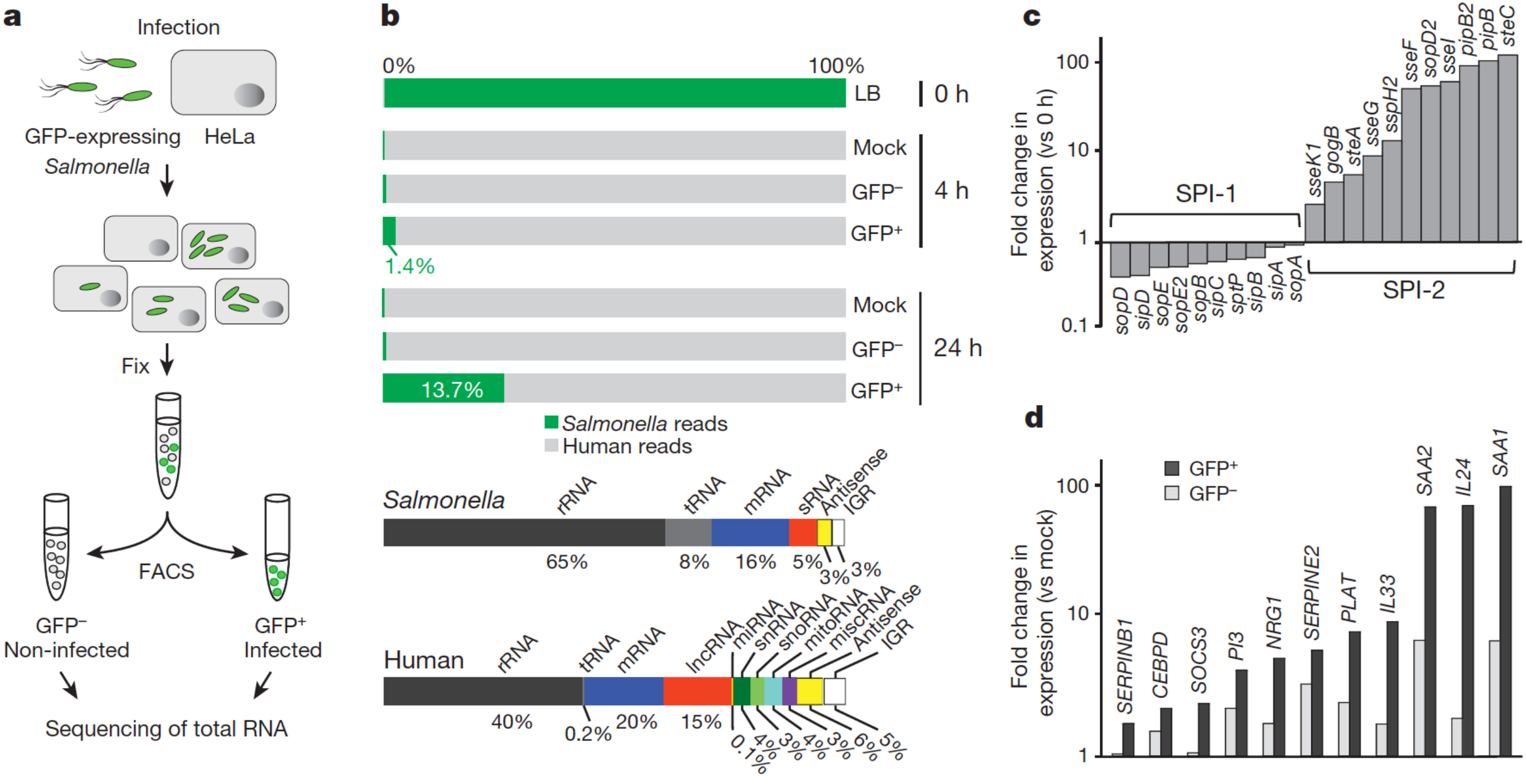






http://www.biochemj.org/content/425/1/1

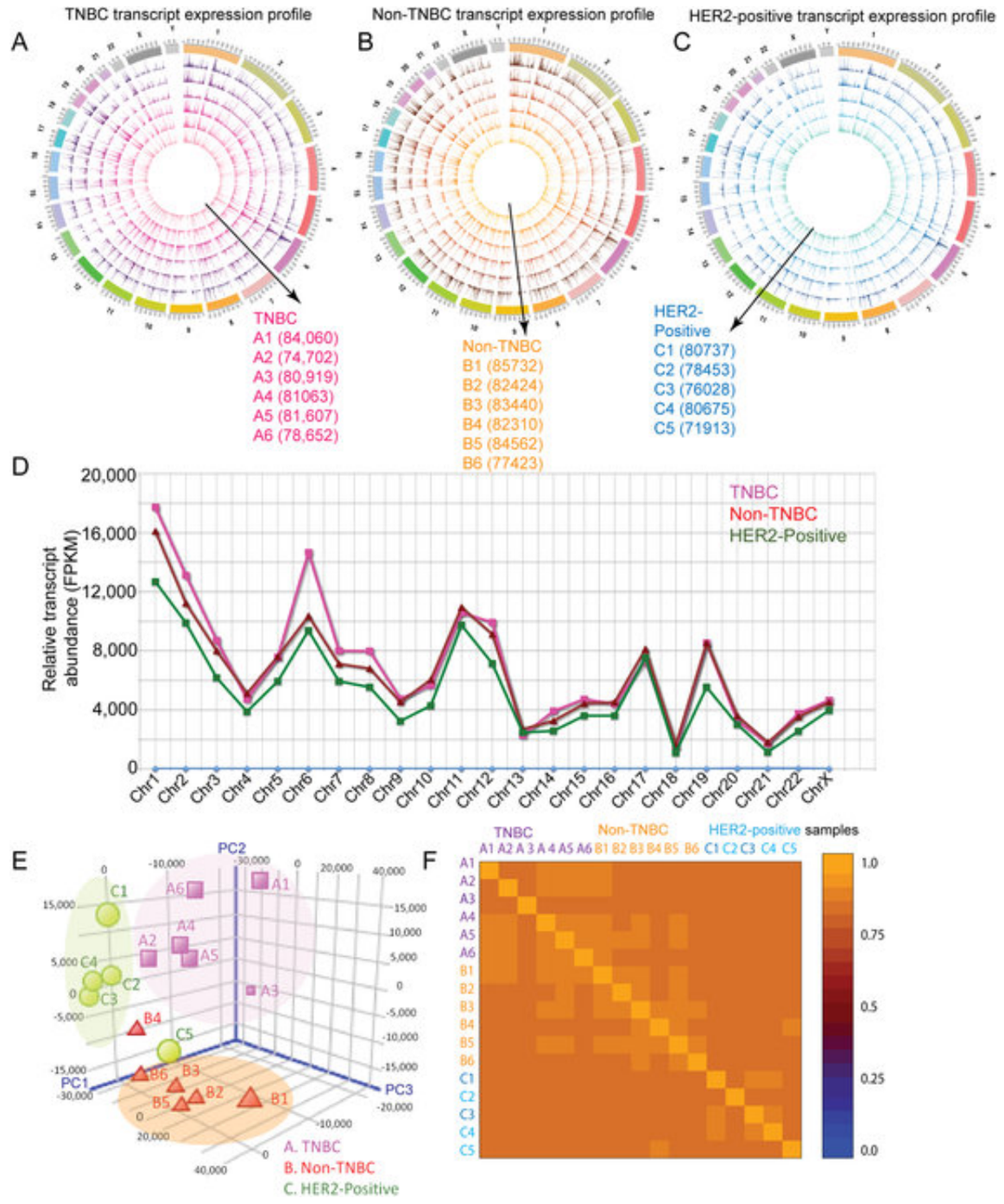
Transcriptómica en procariontes: Dual RNA-Seq



http://www.nature.com/nature/journal/v529/n7587/full/nature16547.html

Transcriptómica en cáncer

- Caracterización de tipos de cáncer
- ¿Son todos los cáncer iguales?
- En la figura, distintos tipos de cáncer de mamas



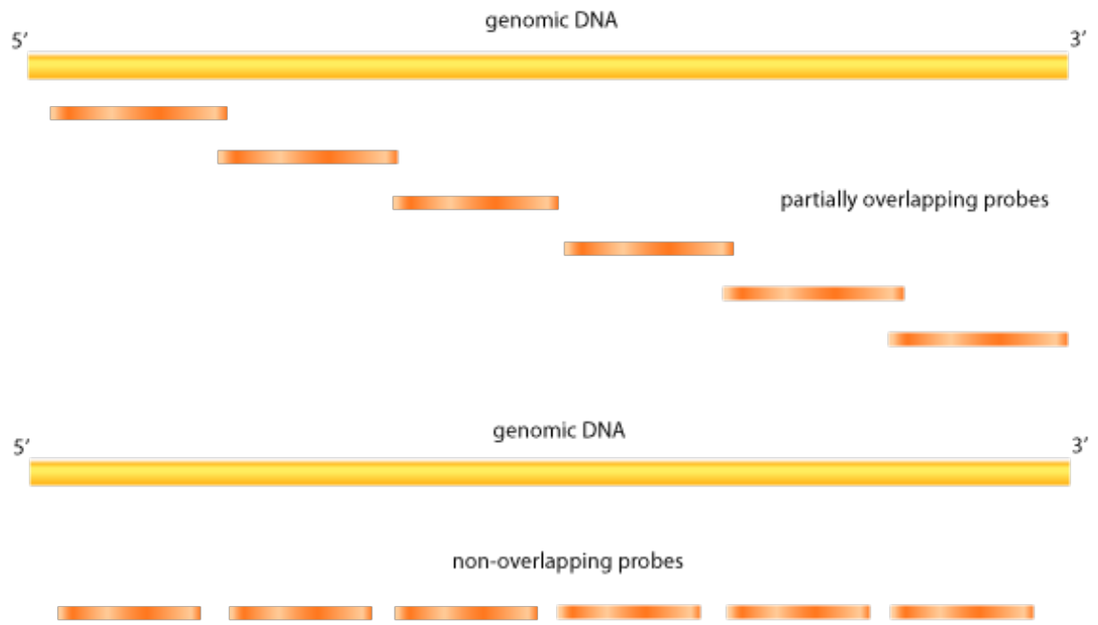
http://www.nature.com/articles/srep00264

Señal de fluorescencia como proxy de expresión

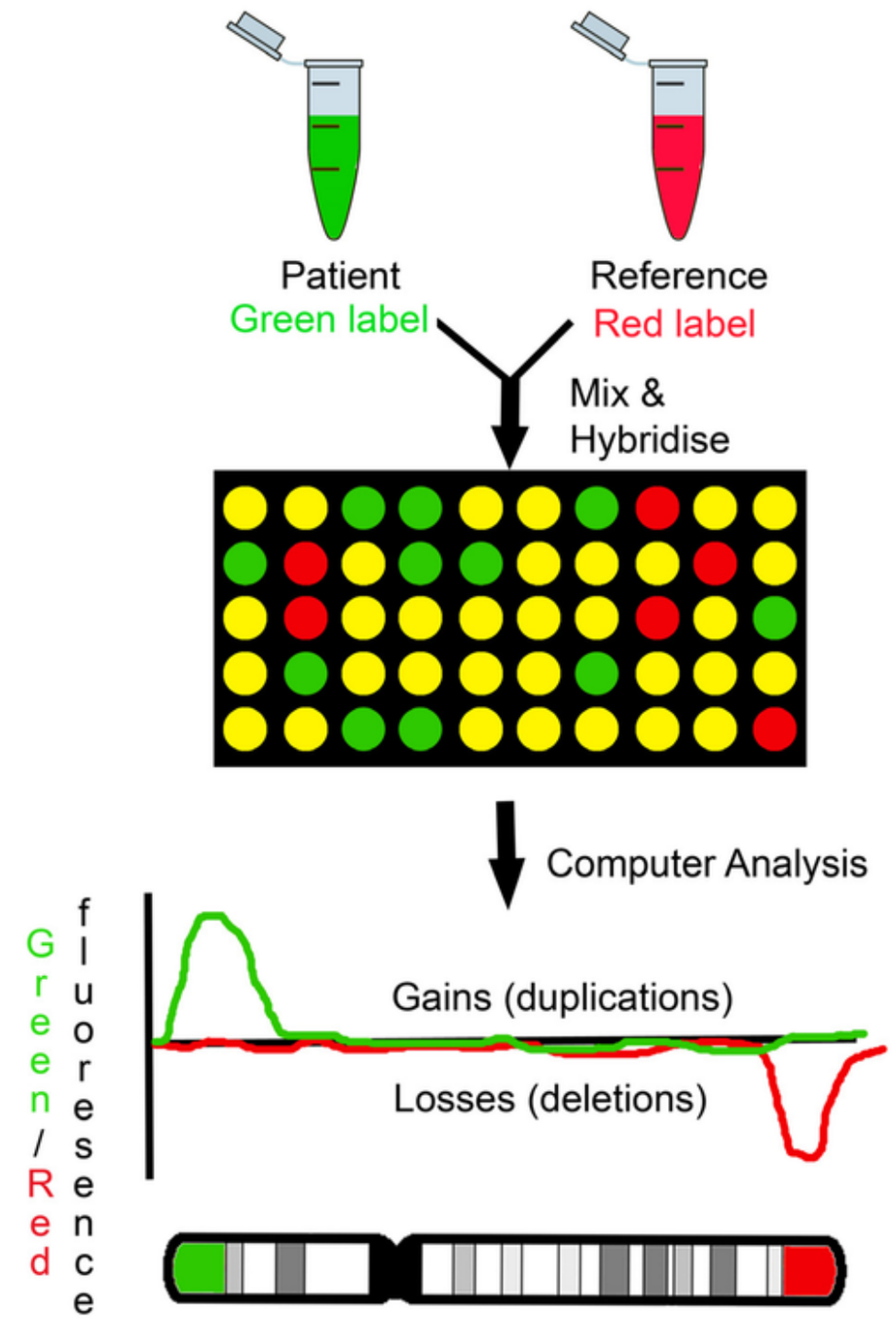
- Viejos tiempos, 2000s

- Inmovilizar o sintetizar oligos en una matriz inerte, vidrio u otro

- Tiling arrays para transcriptómica

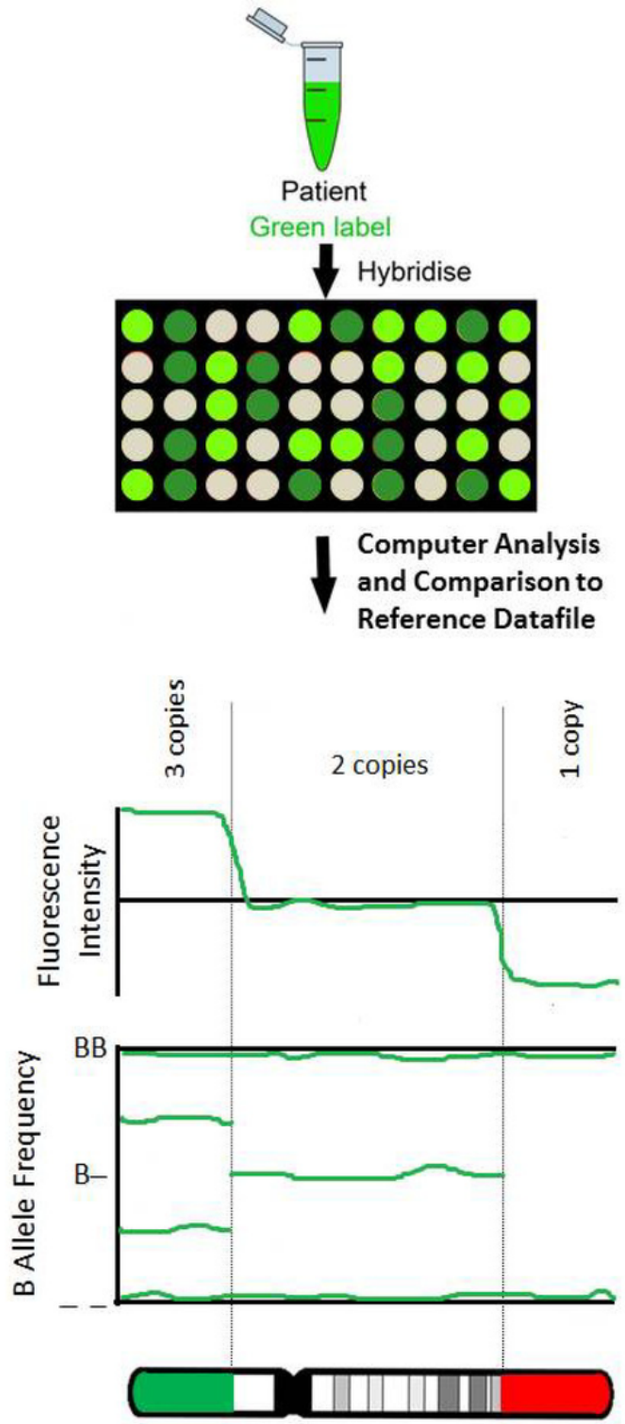


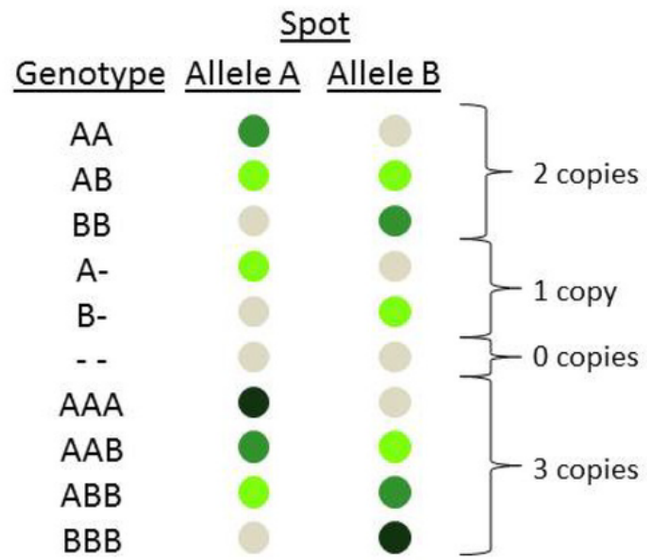
Microarreglos: Comparative Genomics Hybridization (CGH) Array



http://www.mdpi.com/2077-0383/3/2/663/htm

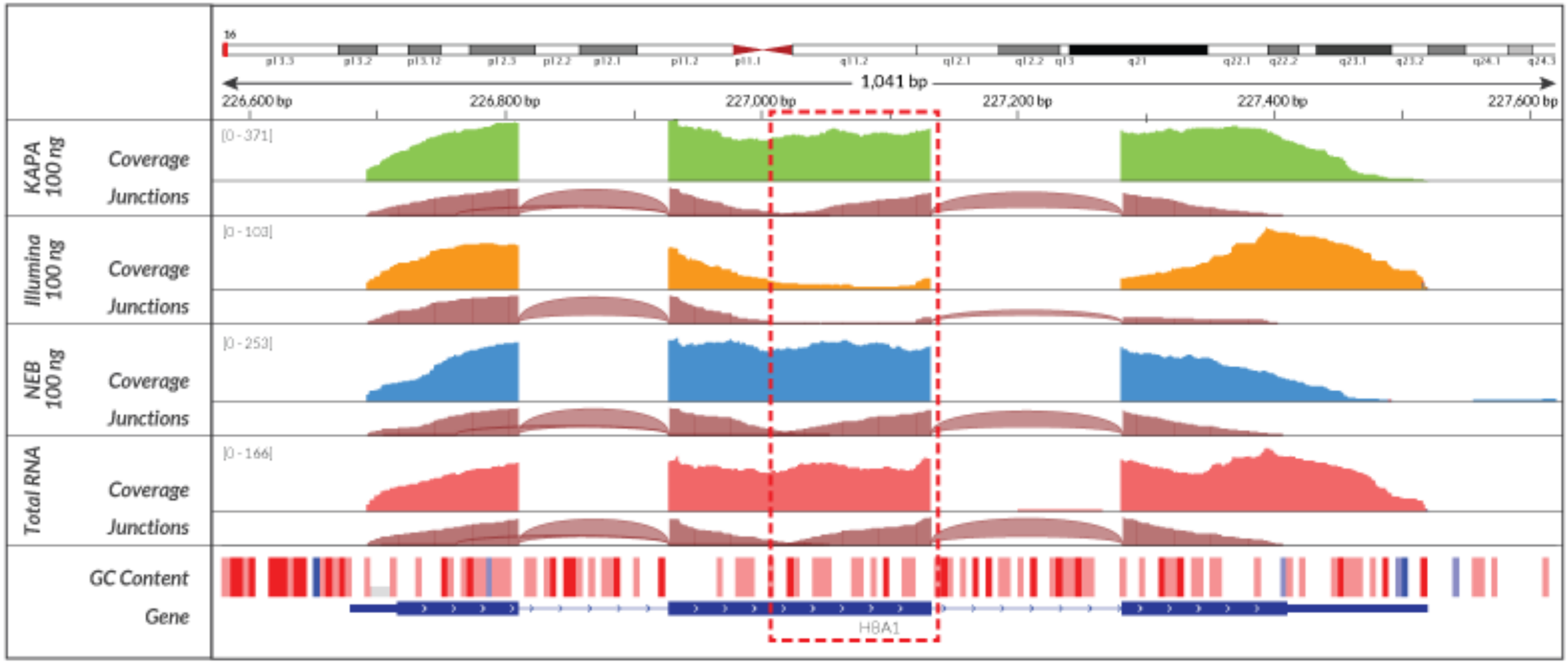
Microarreglos: SNP Array



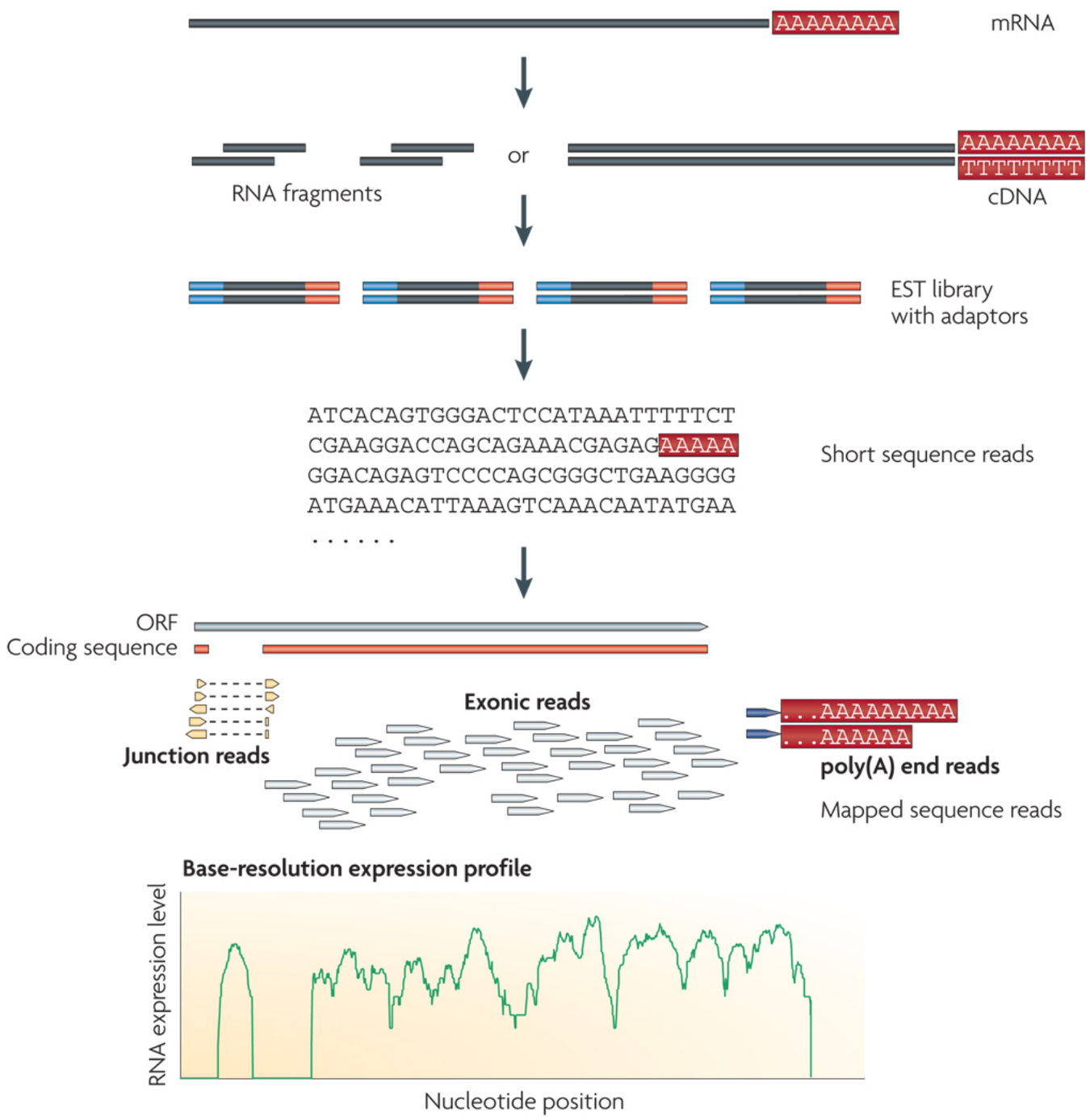


http://www.mdpi.com/2077-0383/3/2/663/htm

Abundancia de reads como proxy de expresión



Pasos clásicos en RNA-Seq



Algunas preguntas?

- ¿Cuál es el tipo de RNA más abundante en una célula?

- ¿Cómo se puede enriquecer una extracción de RNA para obtener mensajeros?

- ¿Podemos averiguar qué hebra está siendo transcrita?

Comparación

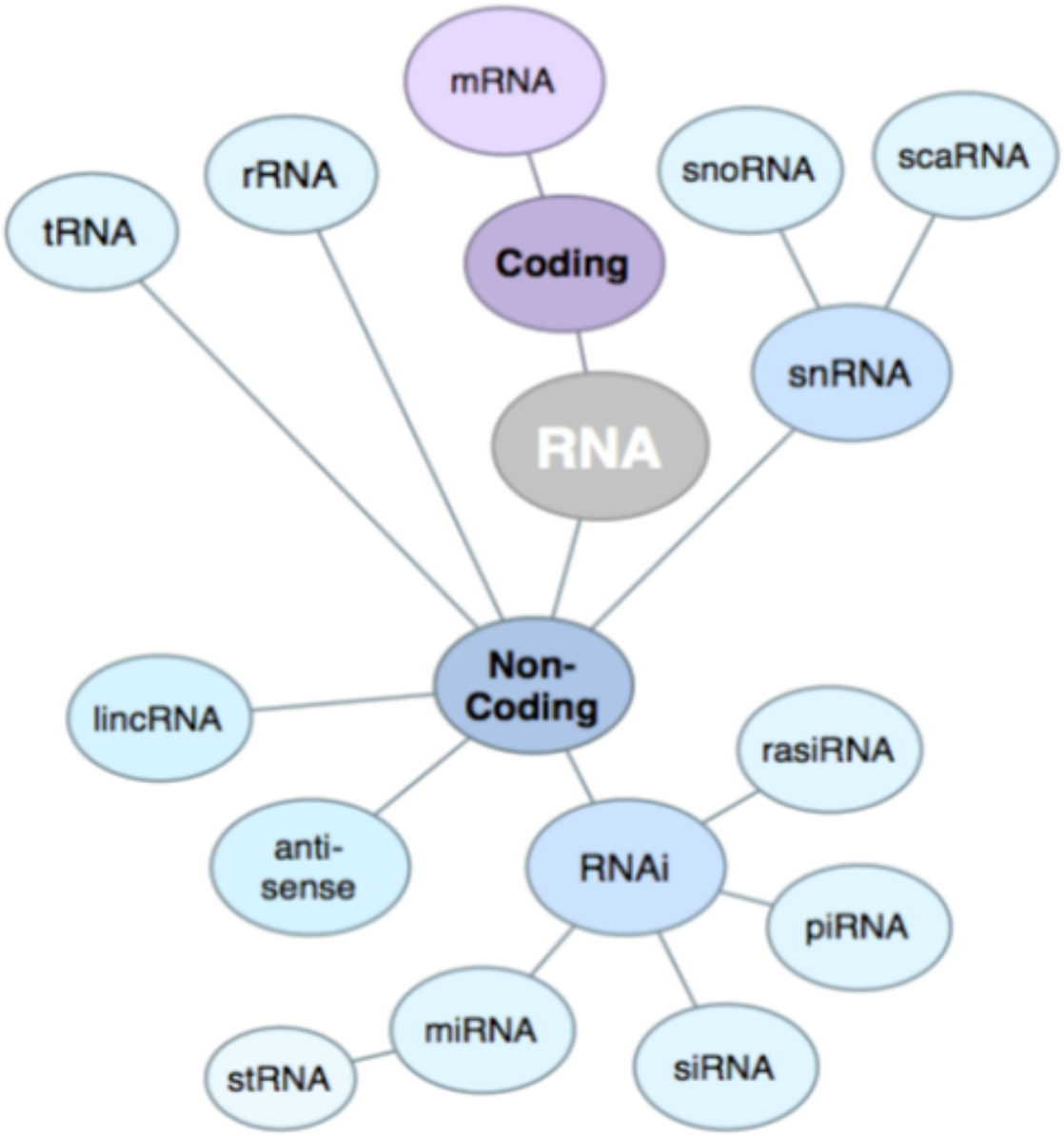
Technology	**Tiling microarray**	**cDNA or EST sequencing**	**RNA-seq**
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
cost for mapping transcriptomes of large genomes	High	High	Relatively low

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/

Aplicaciones prácticas en RNA-Seq

- Análisis diferencial de genes, identificar genes, exones, splicing alternativo, ncRNAs
- Redes de co-expresión
- Búsqueda de isoformas
- Reconstrucción de rutas metabólicas

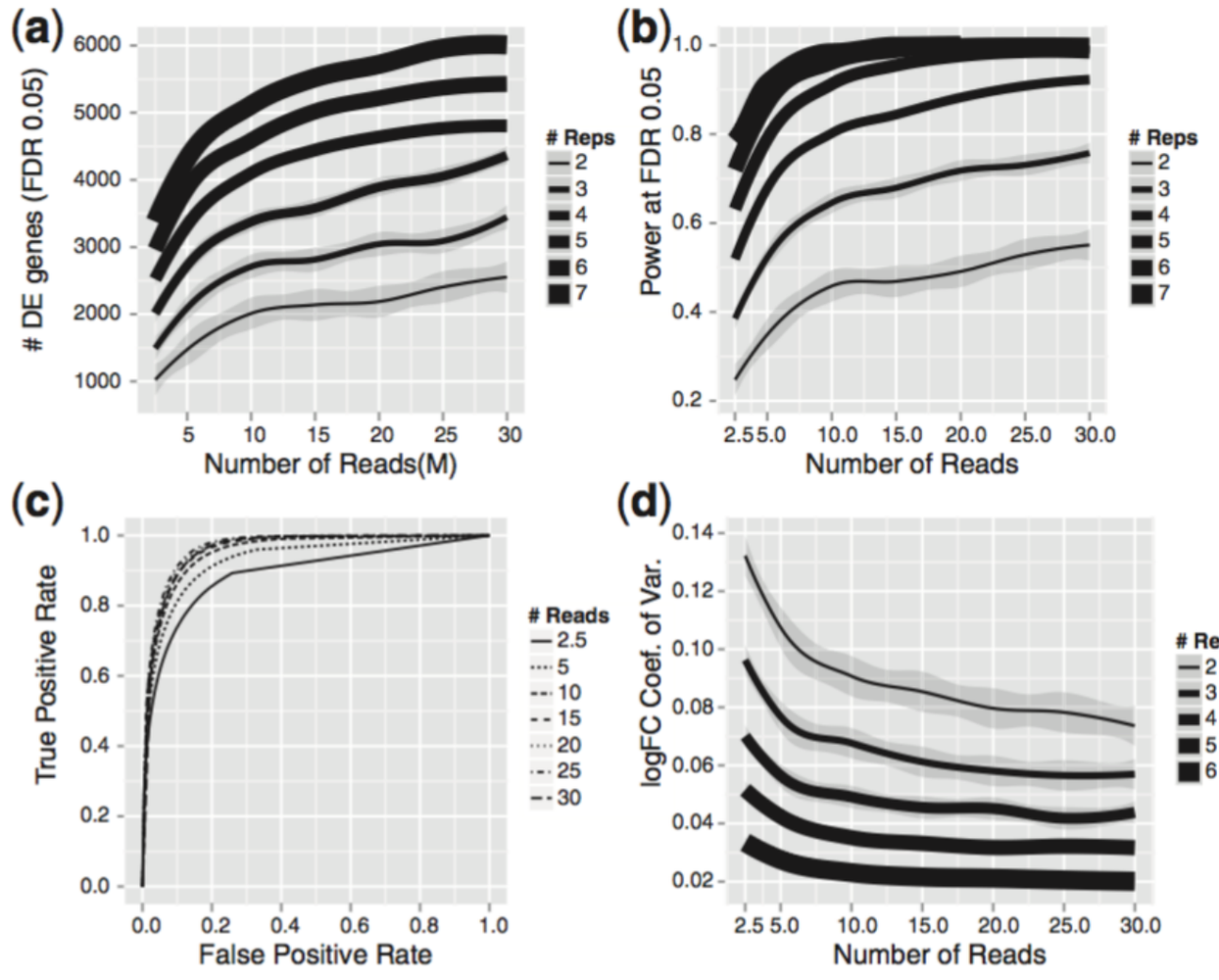
Aplicaciones prácticas en RNA-Seq



Análisis

Más secuencias o más muestras (réplicas)

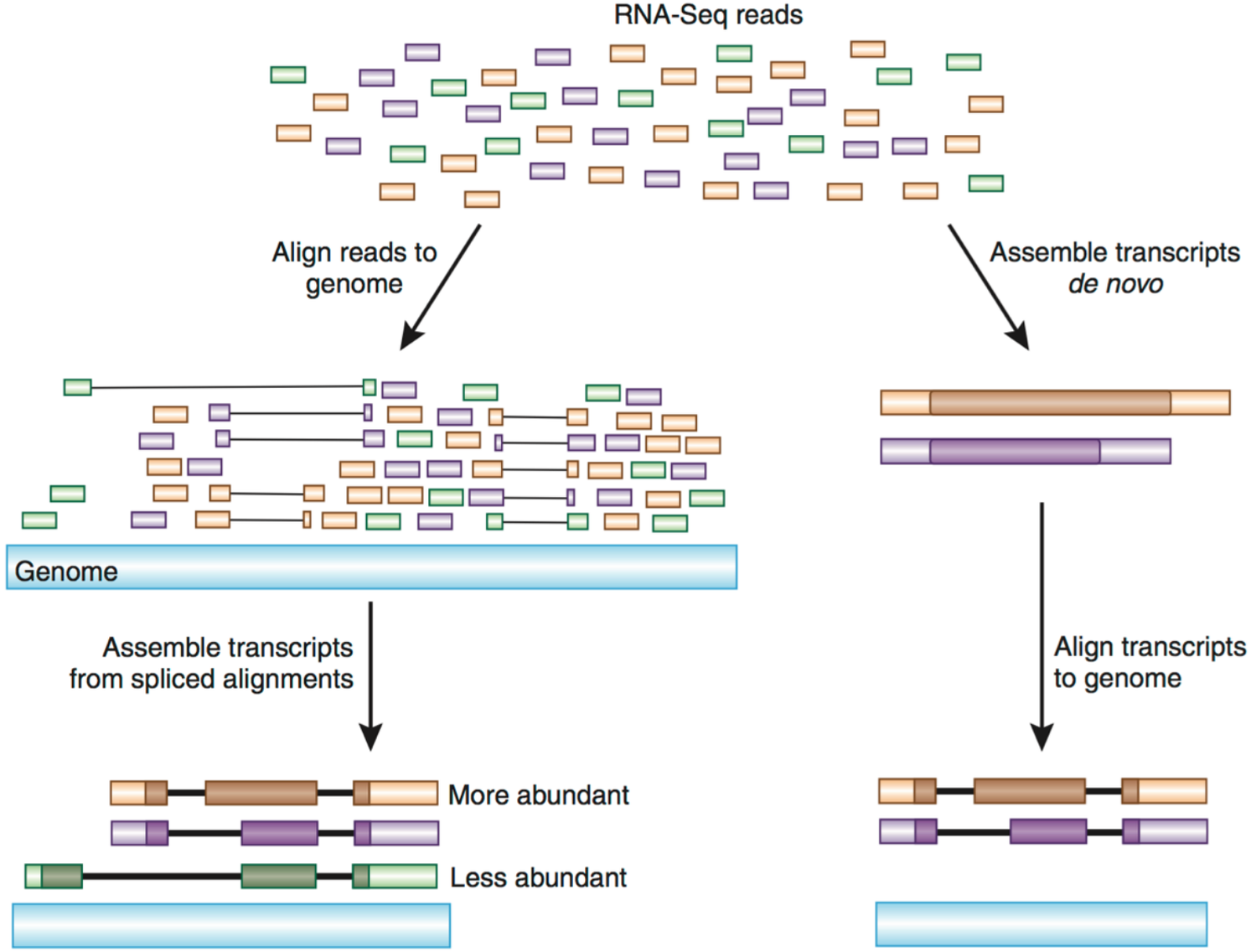
- Mientras más réplicas, mayor número de genes DE
- Mientras más réplicas, más poder estadístico
- Mayor profundidad no retorna mayor tasa de “verdaderos positivos”
- Mientras más réplicas menor es el coeficiente de variación



http://www.ncbi.nlm.nih.gov/pubmed/24319002

Estrategias de mapeo

- También se puede hacer *de novo*



http://guttmanlab.caltech.edu/pdf/HaasAndZody.pdf

¿Cómo cuantificar la abundancia?

- Reads / fragments per kilobase million (RPKM/ FPKM)
- Transcripts per million (TPM)
- Counts per million (CPM)
- Número de reads (counts) y posterior normalización

Problemas

- Genes de mayor longitud van a tener más reads
- Librerías con más reads van a sesgar la cuantificación
- Expresión génica en una condición no tiene la misma distribución ni varianza que en otras condiciones
- Normalizar por longitud de gen, por tamaño de librería

Reads per kilobase million

- Dividir el número total de reads por 1 millón (per million)
- Dividir las read counts de cada gen por el resultado de arriba (reads per million)
- Dividir el resultado por la longitud de los genes en kilobases (RPKM)
- FPKM es lo mismo pero para PE reads
- Estas medidas solo sirven para comparaciones INTRAmuestra

TPMs

- Contar las reads por cada gen y dividir por la longitud de los genes en kilobases (RPK)

- Sumar todos los valores de RPK en una muestra y dividir por 1 millón

- Dividir los valores de RPK por la constante obtenida en el punto anterior

- TPMs pueden ser usadas para comparaciones INTERmuestras porque la suma de todas las TPMs en una muestra es la misma que en otra muestra

- La excepción es cuando tienes pocos (decenas) de genes, lo cual no es realista

Ecuaciones

$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\widetilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\widetilde{l}_i N} \cdot 10^9$$

http://www.nature.com/nmeth/journal/v5/n7/abs/nmeth.1226.html

$$\mathrm{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}}\right) \cdot 10^6$$

http://bioinformatics.oxfordjournals.org/content/26/4/493.long

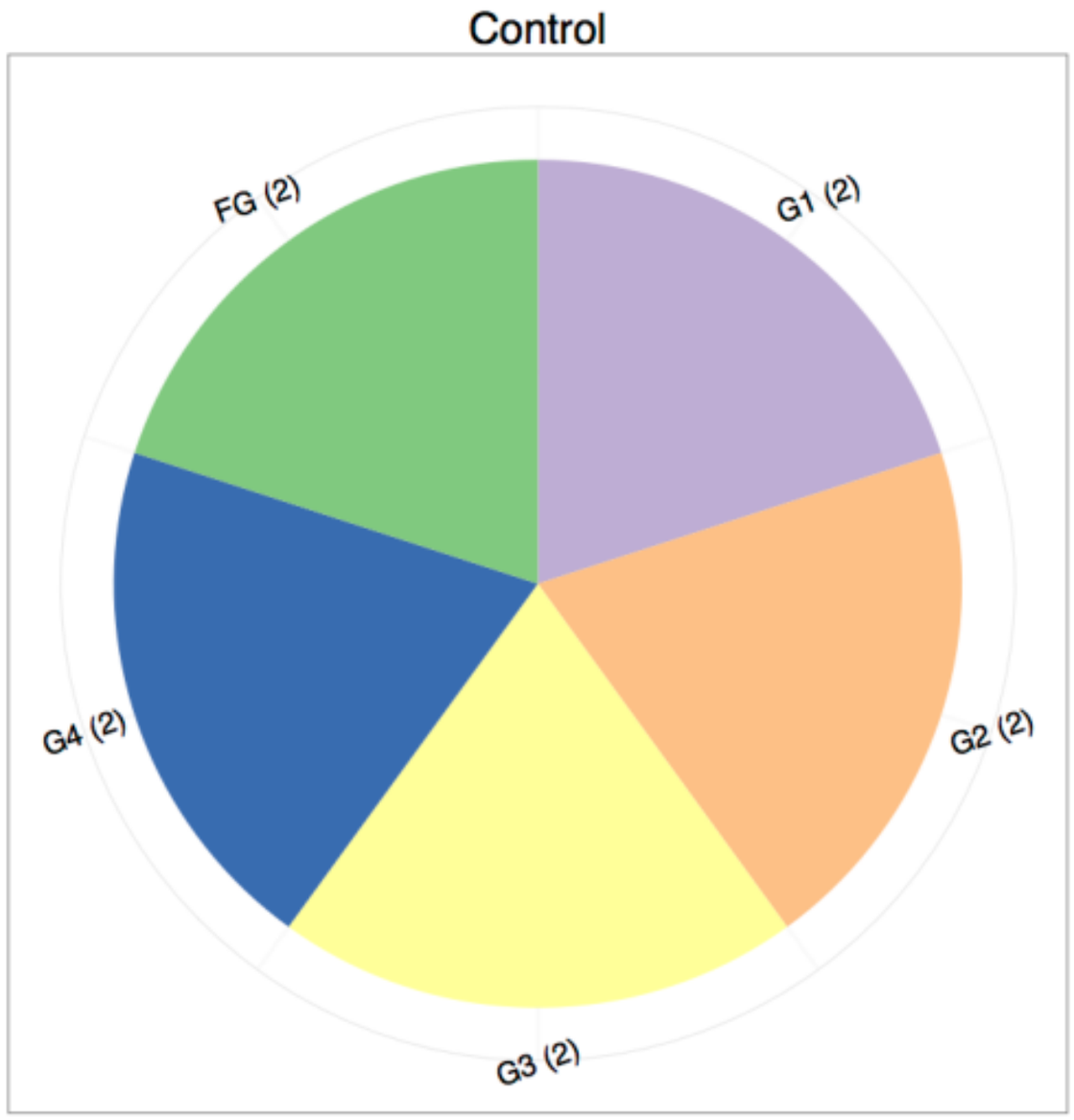
CPM

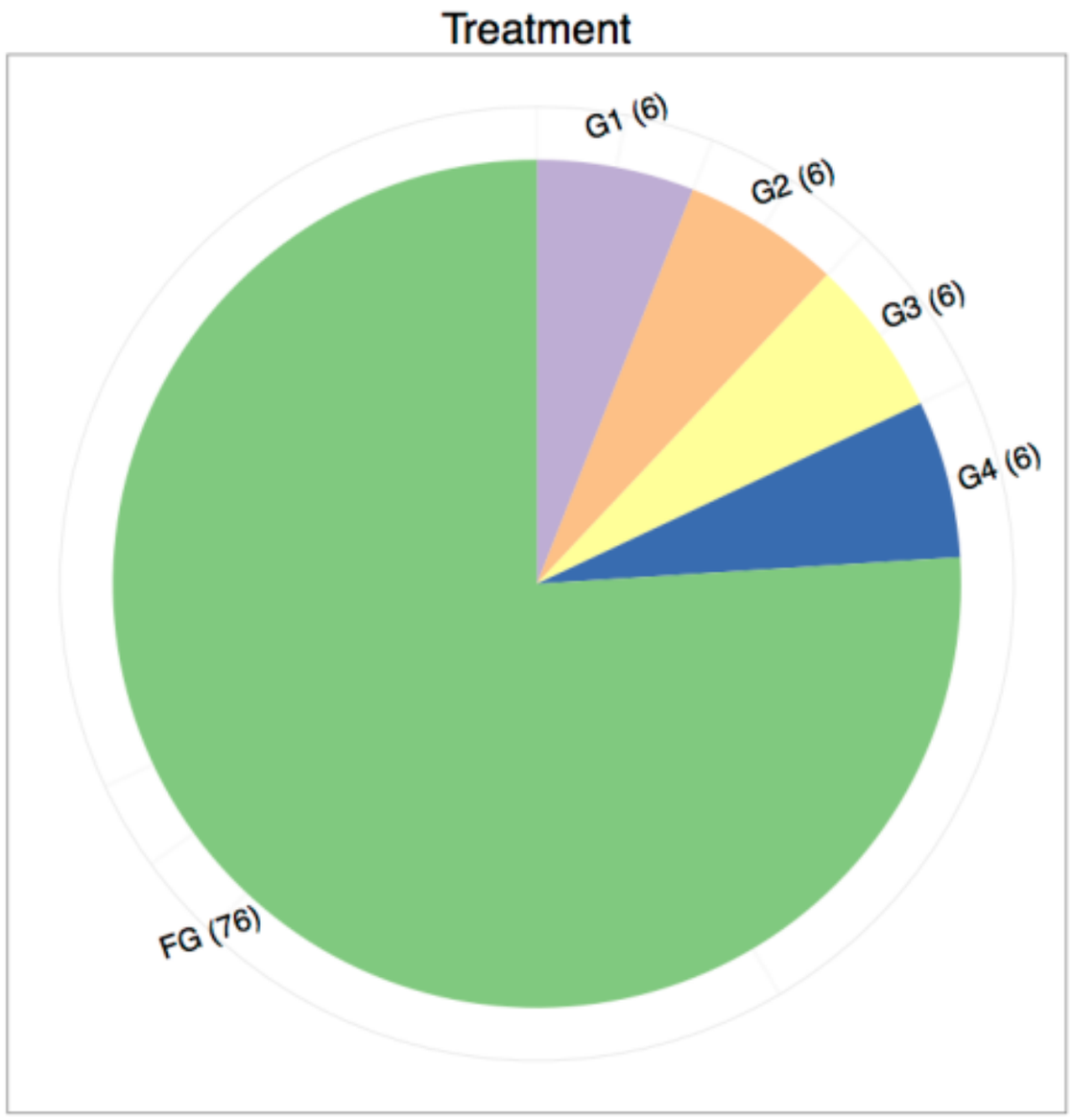
- Número de reads en un gen, dividido por el total de reads, multiplicado por 1 millón

- Solo para comparar INTRAmuestra

$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

Normalización entre muestras





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	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

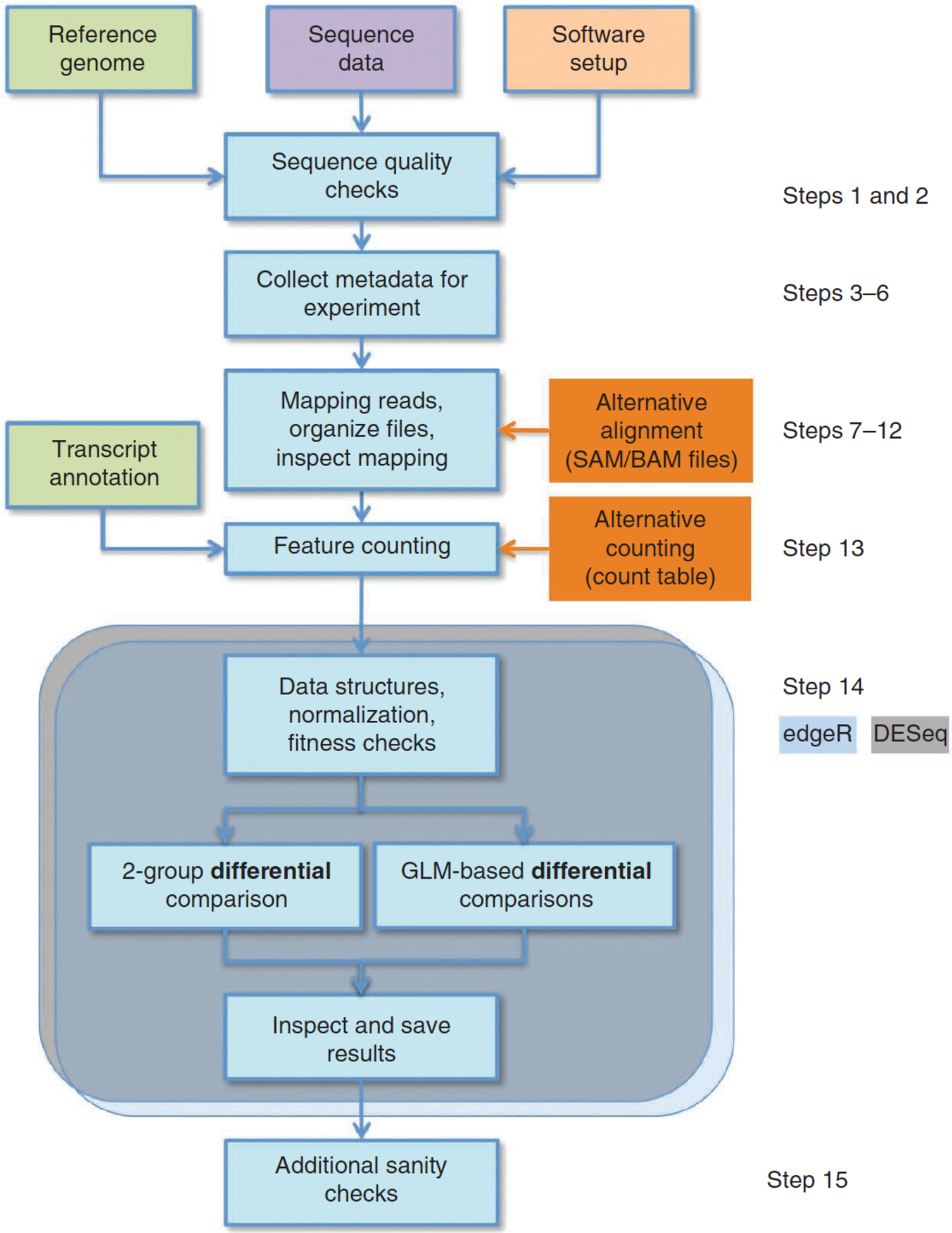
Gene	Control Counts	Treatment Counts	Control Normalized (-FG)	Treatment Normalized (-FG)
G1	2.00	6.00	0.25	0.25
G2	2.00	6.00	0.25	0.25
G3	2.00	6.00	0.25	0.25
G4	2.00	6.00	0.25	0.25
FG	2.00	76.00	0.25	3.17

- 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

Dos modelos más aceptados

- DESeq2 —> Negative Binomial normalization
- EdgeR —> Trimmed mean of M-values (TMM)
- Ambos métodos calculan “size factors” que se usan para hacer comparables distintas muestras


Reference genome
Sequence data
Software setup
Sequence quality checks
Steps 1 and 2
Collect metadata for experiment
Steps 3–6
Mapping reads, organize files, inspect mapping
Alternative alignment (SAM/BAM files)
Steps 7–12
Transcript annotation
Feature counting
Alternative counting (count table)
Step 13
Data structures, normalization, fitness checks
Step 14
edgeR
DESeq
2-group differential comparison
GLM-based differential comparisons
Inspect and save results
Additional sanity checks
Step 15

Revisar Nature Protocol

PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]

[1]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [2]Department of Statistics, University of Oxford, Oxford, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Bioinformatics Division, Walter and Eliza Hall Institute, Parkville, Victoria, Australia. [5]Department of Medical Biology, University of Melbourne, Melbourne, Victoria, Australia. [6]Functional Genomics Center UNI ETH, Zurich, Switzerland. [7]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, Australia. [8]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [9]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. Correspondence should be addressed to M.D.R. (mark.robinson@imls.uzh.ch) or W.H. (whuber@embl.de).

Published online 22 August 2013; doi:10.1038/nprot.2013.099

Otros papers relevantes

Models for transcript quantification from RNA-Seq https://arxiv.org/abs/1104.3889

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

http://bib.oxfordjournals.org/content/14/6/671.full

Tutoriales

http://www.gettinggeneticsdone.com/2015/12/tutorial-rna-seq-differential.html

http://www.bioconductor.org/help/workflows/rnaseqGene/

http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

https://github.com/crazyhottommy/RNA-seq-analysis

http://f1000research.com/articles/5-1408/v1