Bioinformática y análisis de datos ómicos

Tema 9: Identificación de alteraciones transcriptómicas

Ignacio Varela Egocheaga
Departamento de Biología Molecular

Tema 9: Identificación de alteraciones transcriptómicas

9.1 Identificación de cambios en los patrones de expresión

9.2 Estudios de enriquecimiento de rutas moleculares

9.3 Clasificación y agrupamiento de muestras

9.4 Generación de modelos de clasificación automática

9.5 Identificación de nuevos transcritos

9.6 Estudios de RNAs pequeños

Tema 9: Identificación de alteraciones transcriptómicas

9.1 Identificación de cambios en los patrones de expresión

9.2 Estudios de enriquecimiento de rutas moleculares

9.3 Clasificación y agrupamiento de muestras

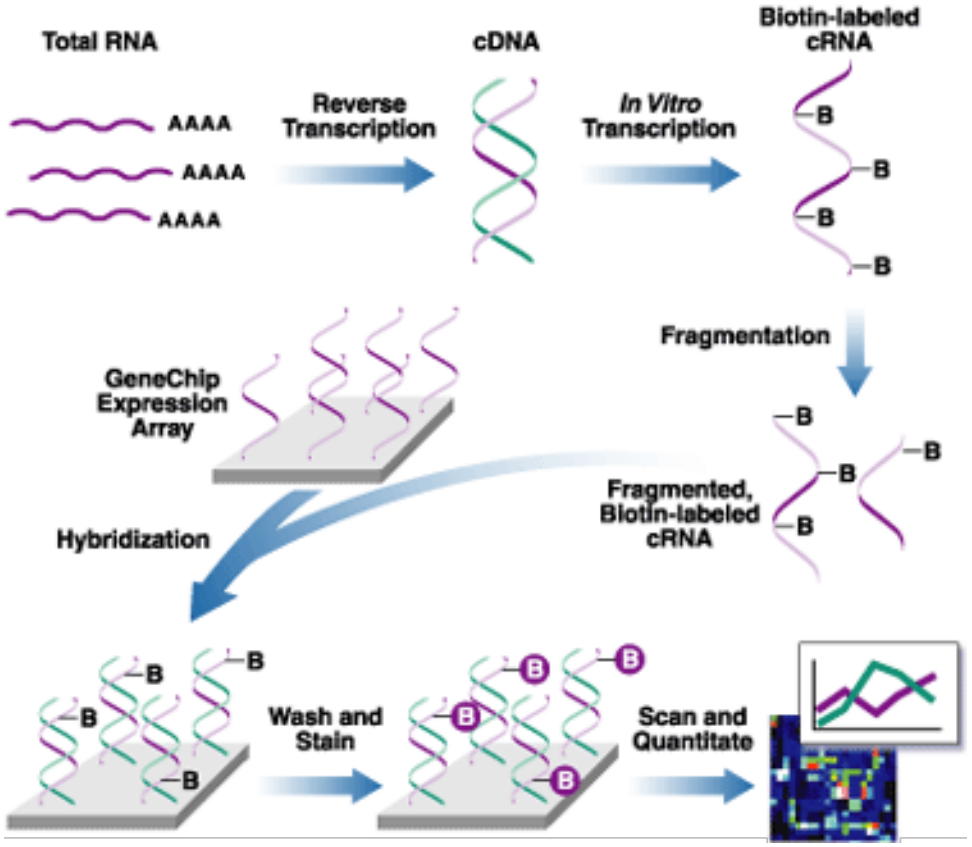
9.4 Generación de modelos de clasificación automática

9.5 Identificación de nuevos transcritos

9.6 Estudios de RNAs pequeños

Cuantificación del cDNA

Chip de expresión génica



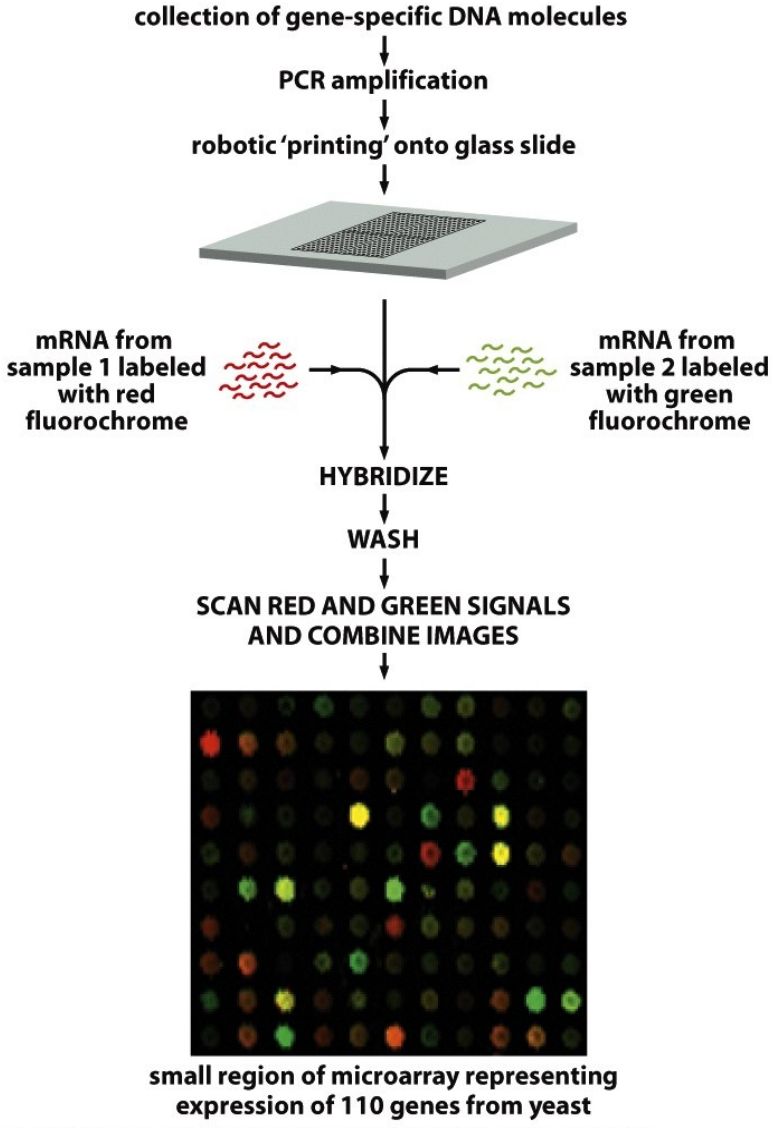
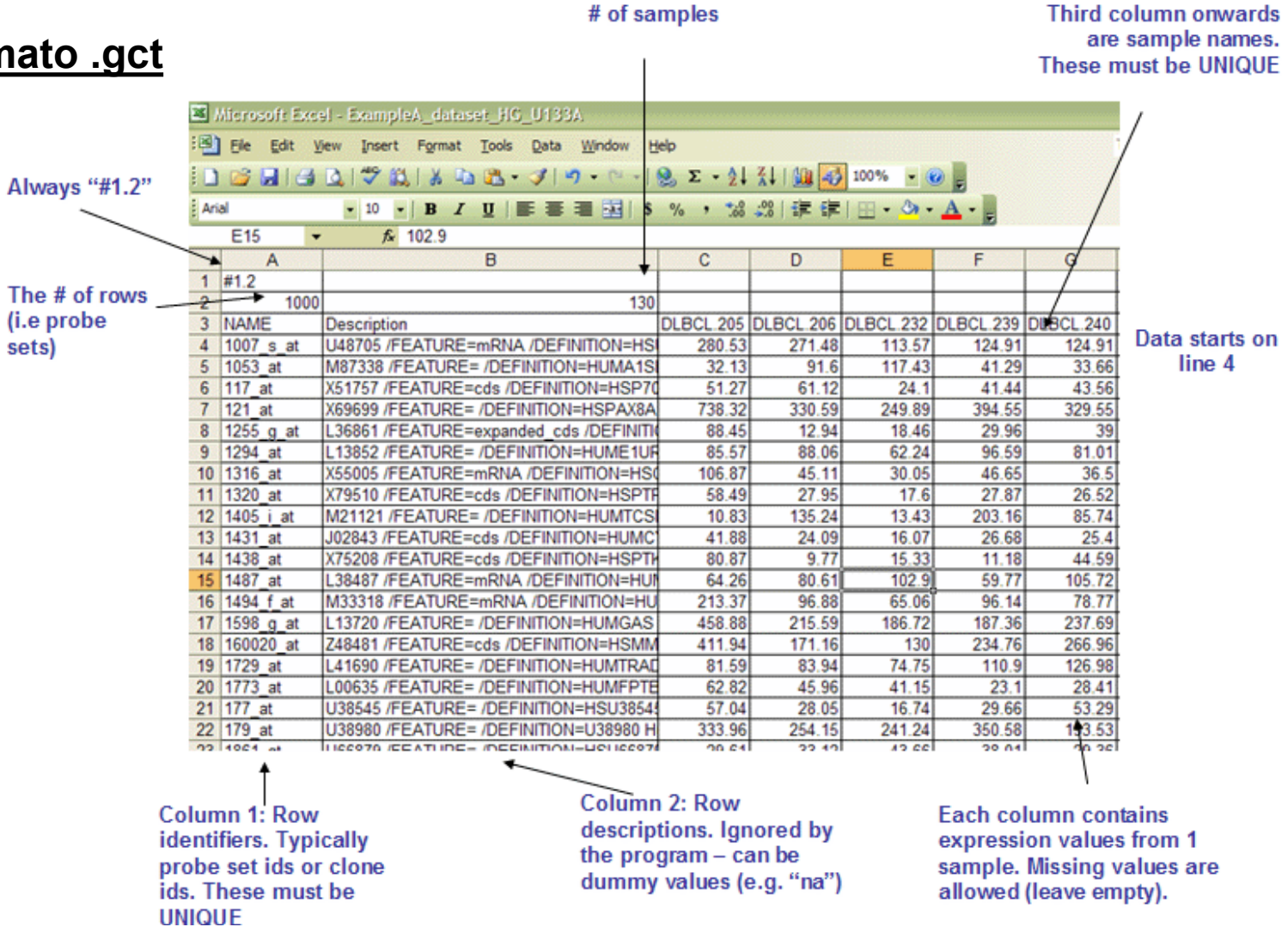


Figure 8-73 Molecular Biology of the Cell 5/e (© Garland Science 2008)

http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat545/Notes/microarray.html

Formato de resultados del chip

Formato .gct



	A	B	C	D	E	F	G
1	#1.2						
2	1000	130					
3	NAME	Description	DLBCL 205	DLBCL 206	DLBCL 232	DLBCL 239	DLBCL 240
4	1007_s_at	U48705 /FEATURE=mRNA /DEFINITION=HS	280.53	271.48	113.57	124.91	124.91
5	1053_at	M87338 /FEATURE= /DEFINITION=HUMA1S	32.13	91.6	117.43	41.29	33.66
6	117_at	X51757 /FEATURE=cds /DEFINITION=HSP7	51.27	61.12	24.1	41.44	43.56
7	121_at	X69699 /FEATURE= /DEFINITION=HSPAX8A	738.32	330.59	249.89	394.55	329.55
8	1255_g_at	L36861 /FEATURE=expanded_cds /DEFINIT	88.45	12.94	18.46	29.96	39
9	1294_at	L13852 /FEATURE= /DEFINITION=HUME1UR	85.57	88.06	62.24	96.59	81.01
10	1316_at	X55005 /FEATURE=mRNA /DEFINITION=HS	106.87	45.11	30.05	46.65	36.5
11	1320_at	X79510 /FEATURE=cds /DEFINITION=HSPTF	58.49	27.95	17.6	27.87	26.52
12	1405_i_at	M21121 /FEATURE= /DEFINITION=HUMTCS	10.83	135.24	13.43	203.16	85.74
13	1431_at	J02843 /FEATURE=cds /DEFINITION=HUMC	41.88	24.09	16.07	26.68	25.4
14	1438_at	X75208 /FEATURE=cds /DEFINITION=HSPTH	80.87	9.77	15.33	11.18	44.59
15	1487_at	L38487 /FEATURE=mRNA /DEFINITION=HU	64.26	80.61	102.9	59.77	105.72
16	1494_f_at	M33318 /FEATURE=mRNA /DEFINITION=HU	213.37	96.88	65.06	96.14	78.77
17	1598_g_at	L13720 /FEATURE= /DEFINITION=HUMGAS	458.88	215.59	186.72	187.36	237.69
18	160020_at	Z48481 /FEATURE=cds /DEFINITION=HSMM	411.94	171.16	130	234.76	266.96
19	1729_at	L41690 /FEATURE= /DEFINITION=HUMTRAD	81.59	83.94	74.75	110.9	126.98
20	1773_at	L00635 /FEATURE= /DEFINITION=HUMFPTE	62.82	45.96	41.15	23.1	28.41
21	177_at	U38545 /FEATURE= /DEFINITION=HSU3854	57.04	28.05	16.74	29.66	53.29
22	179_at	U38980 /FEATURE= /DEFINITION=U38980 H	333.96	254.15	241.24	350.58	193.53

https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats

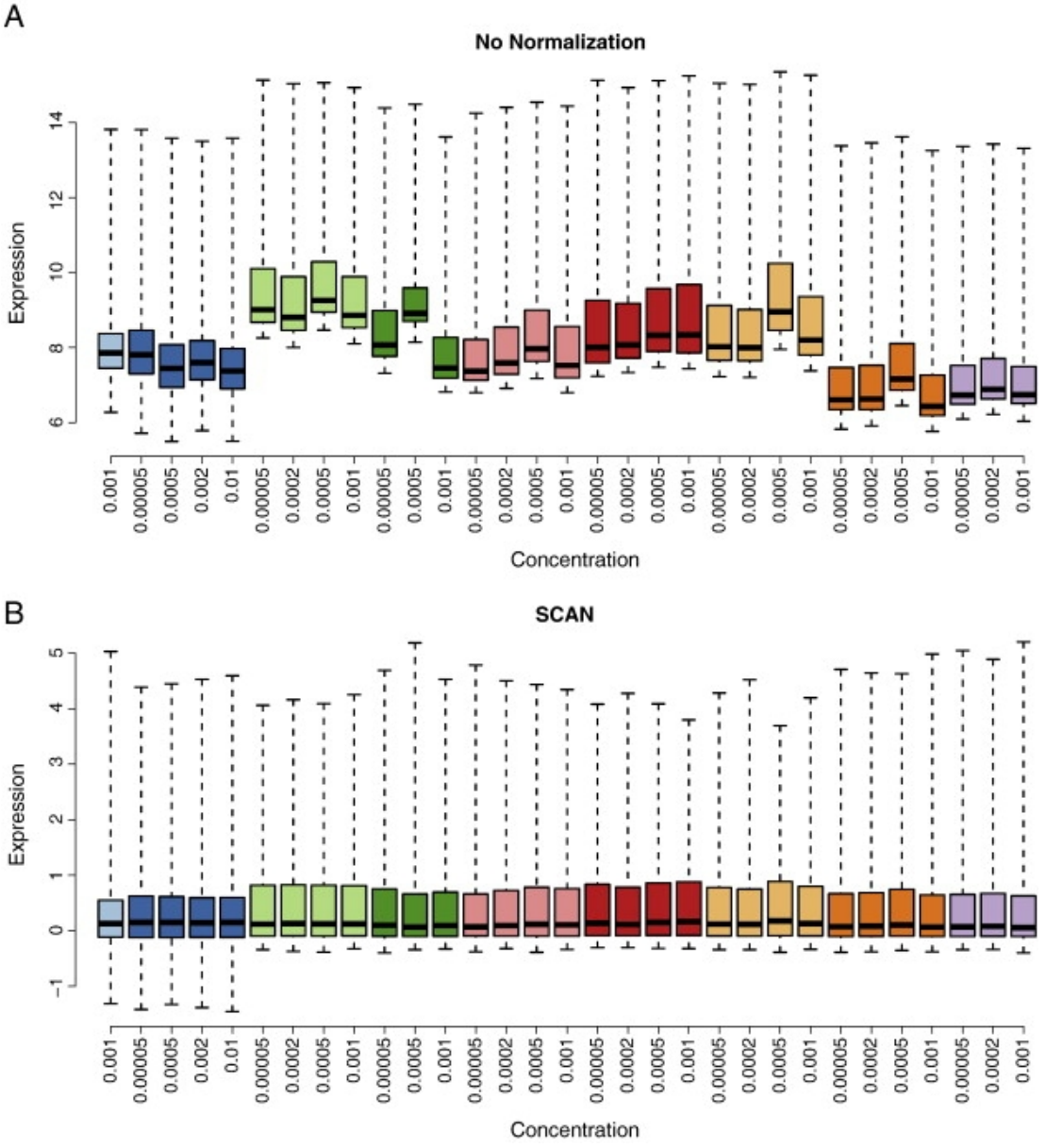
Existen diversos formatos de expresión. En uno de los más utilizados, de la compañía affimetrix, la información de la expresión (.cel) y de las sondas del chip (.chp) vienen en archivos distintos.

Normalización de los datos

1.- Diferencias en la cantidad total de RNA utilizado en cada chip.

2.- Diferencias en las reacciones de hibridación o generación del color.

3.- Diferencias en los parámetros de detección



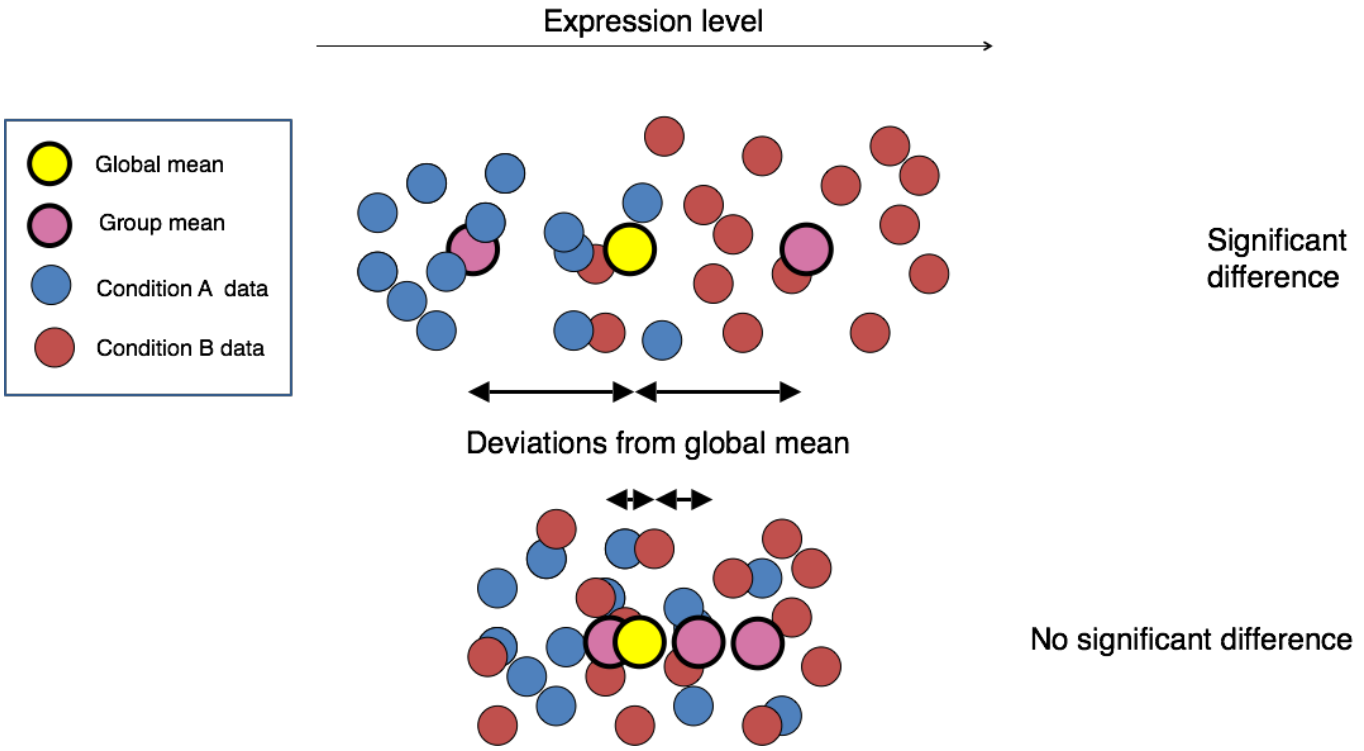
Existen diversas estrategias de normalización, basadas en la idea de que la mayoría de los RNAs muestran una expresión similar entre las distintas muestras.

Explicación de la normalización por cuartiles:

https://www.youtube.com/watch?v=ecjN6Xpv6SE

Picolo SR et al. Genomics 100: 337-344

Identificación de diferencias de expresión



https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html

Corrections for multiple testing

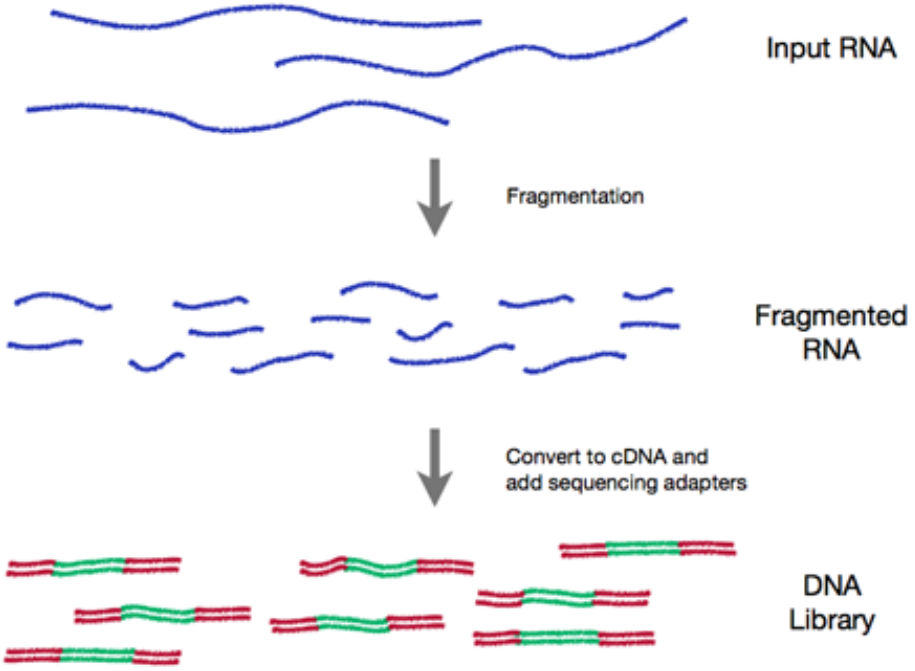
Bonferroni correction:

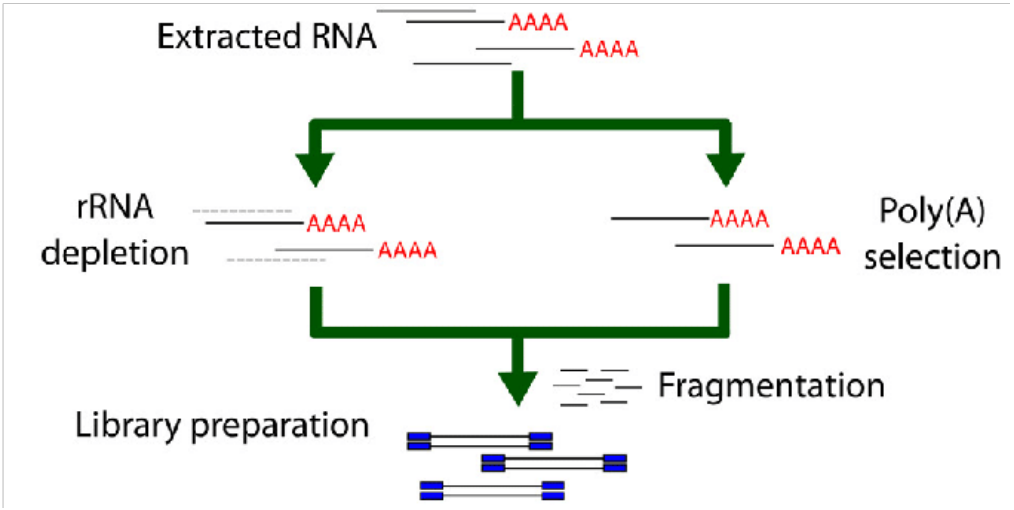
- Reject H_0 when: $m \cdot p\text{-value} \leq \alpha$
  where m is the total number of hypothesis tests performed
- Bonferroni correction implies $\mathrm{FWER} \leq \alpha$

Taken from MITx 6.419x Data Analysis: Statistical Modeling and Computation in Applications

https://www.youtube.com/watch?v=HLzS5wPqWR0

RNA-seq library preparation

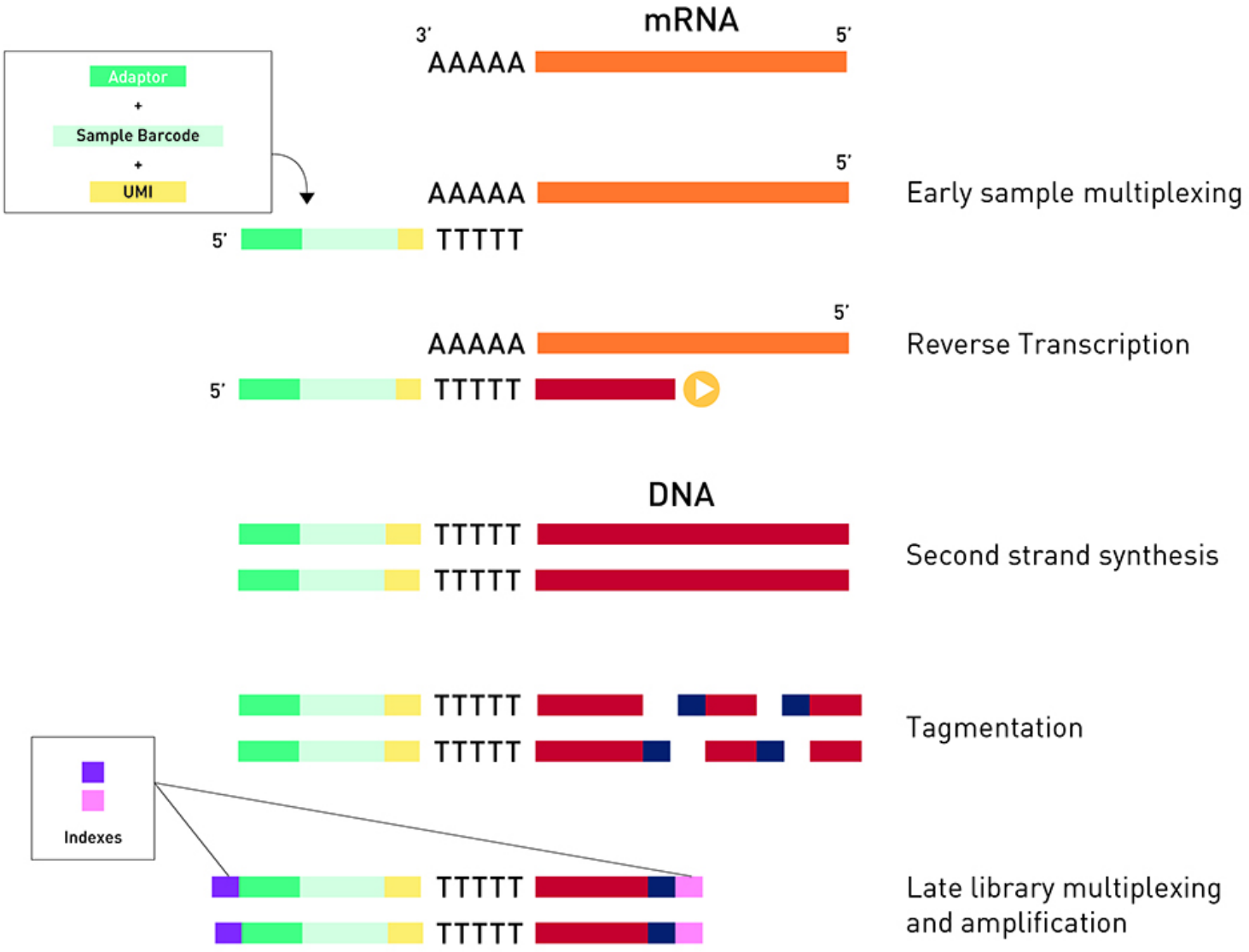




Namjoshi SV et al. Front Mol Neur (2017)

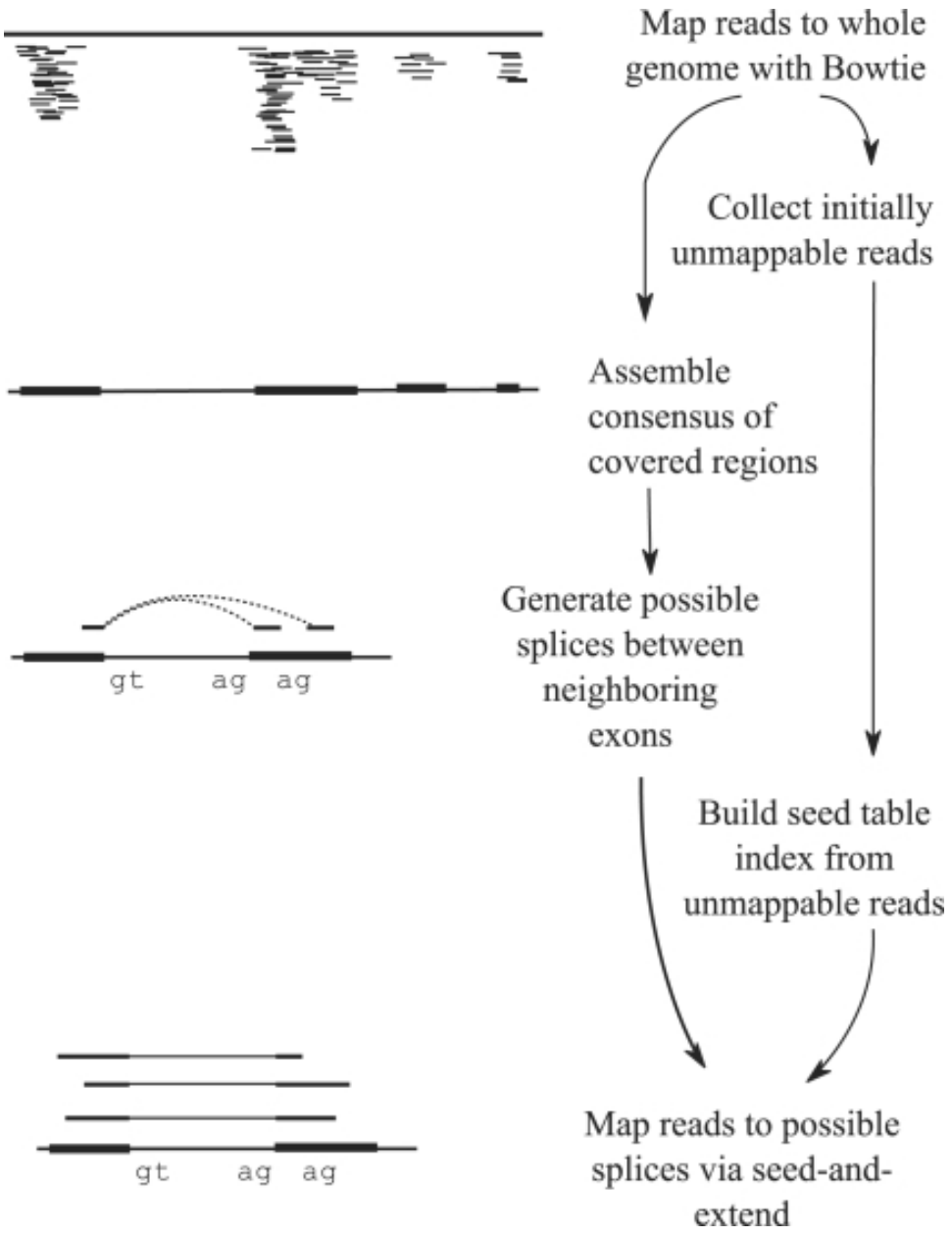
https://rnaseq.uoregon.edu/

3'-Seq



https://www.diagenode.com/en/p/high-throughput-3mrna-seq-services

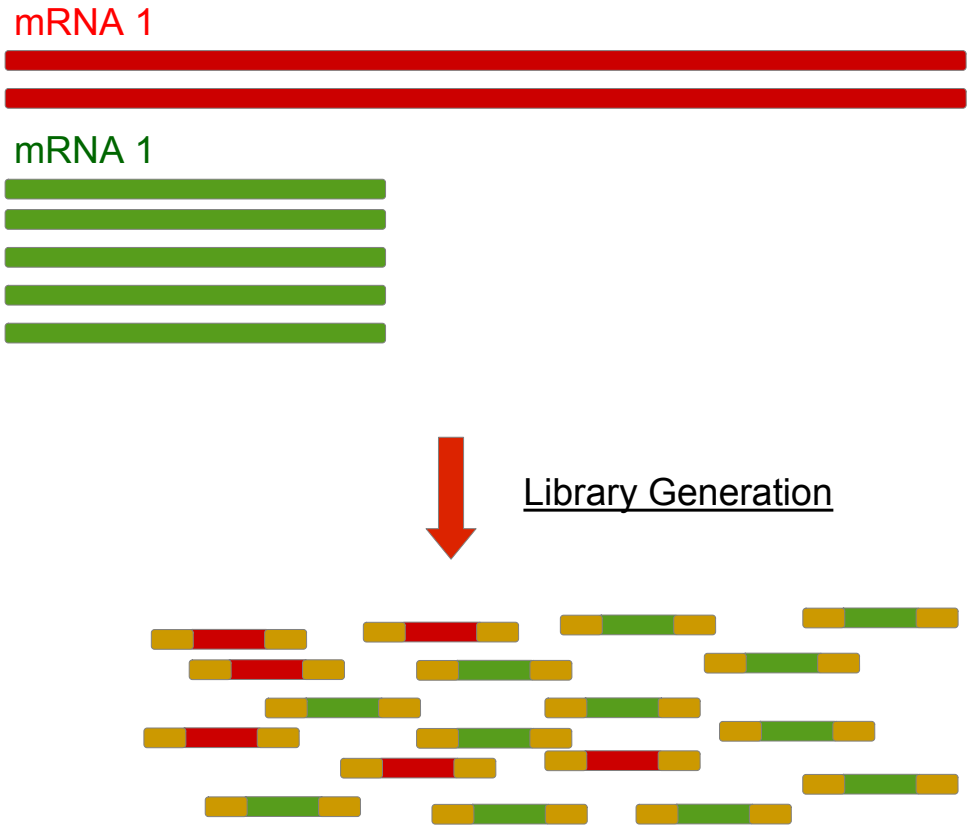
Alineamiento de secuencias de mRNA

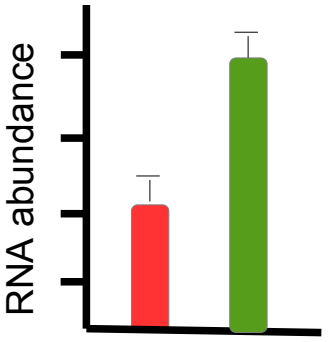


El alineamiento de las secuencias de RNA-seq requiere de alineadores especiales como ***Tophat*** o ***Hisat***.

Trapnell C. et al. Bioinformatics (2009) 25:1105-1111

Cuantificación de la abundancia de mRNAs

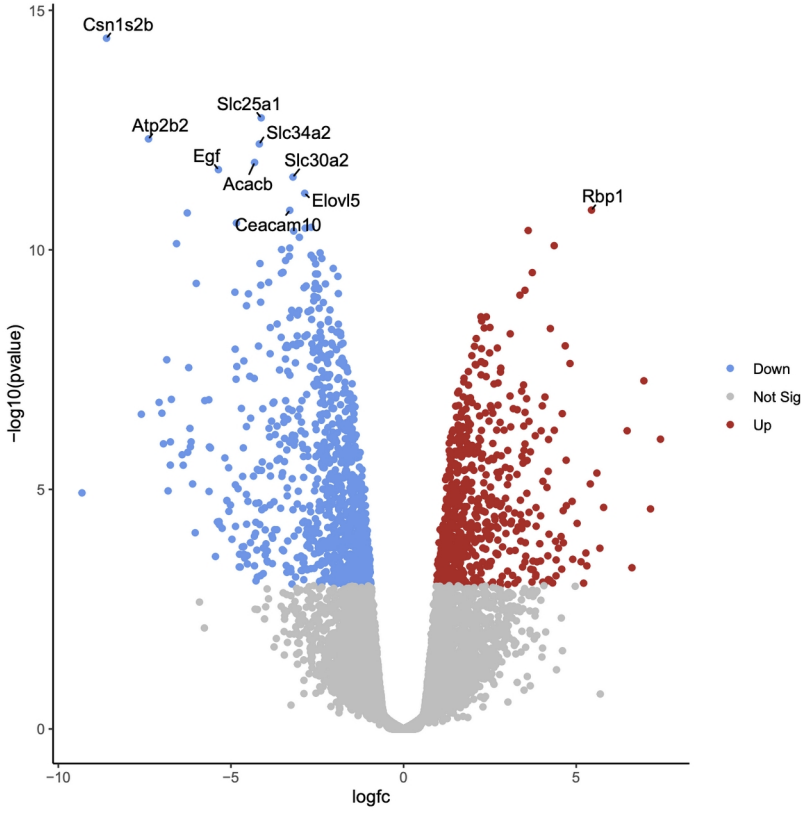




Herramientas como ***HTSeq*** permiten calcular el número de lecturas/secuencias correspondientes a cada mRNA.

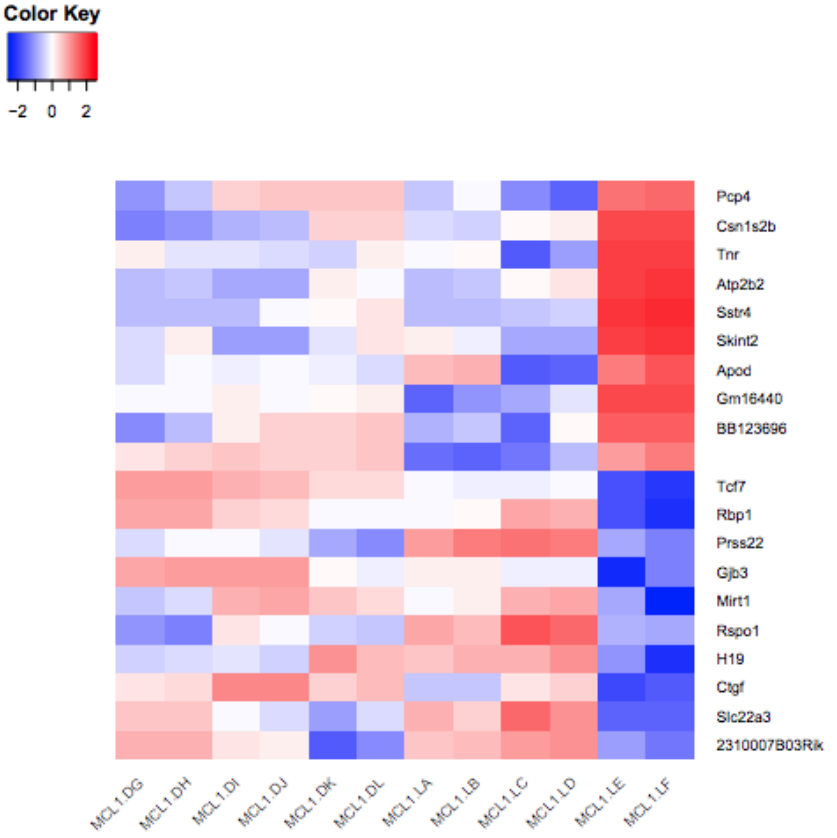
Representaciones habituales en estudios transcripcionales

Volcano plot



Representa la magnitud del cambio en el eje X (*fold-change*) frente a la significancia estadística (*pvalue*)

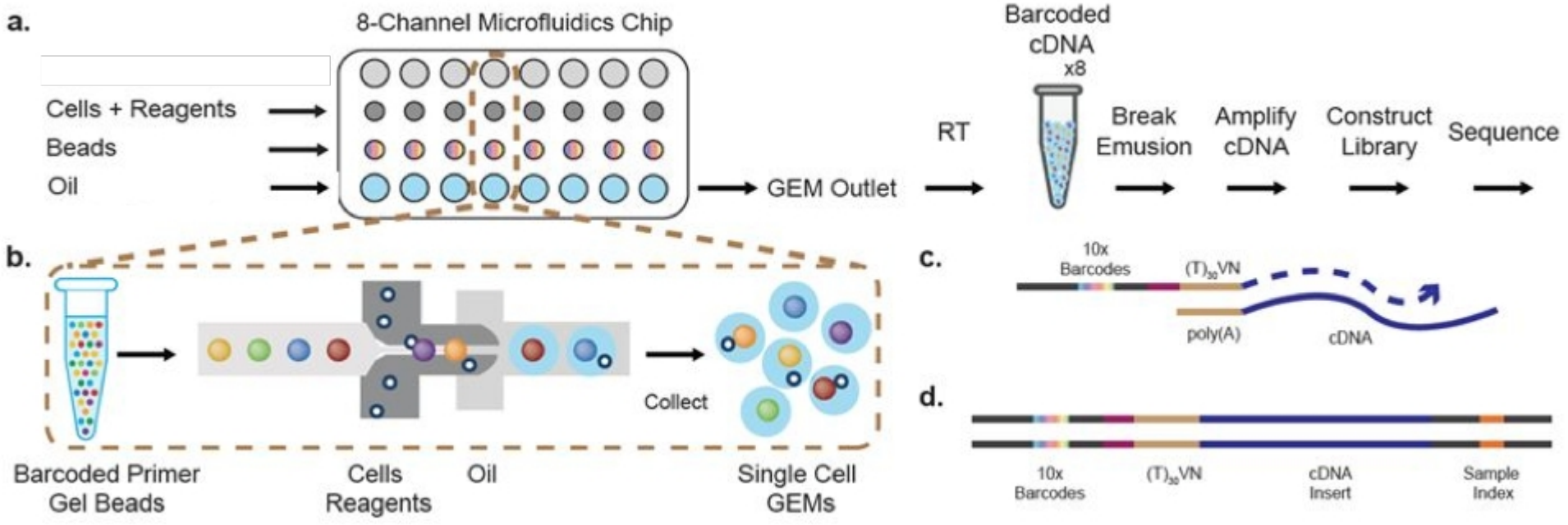
Heatmap



Representa con un mapa de colores la magnitud del cambio de un gen en una muestra relativizado a la media de expresión de ese gen en todas las muestras. Las filas no son comparables entre si.

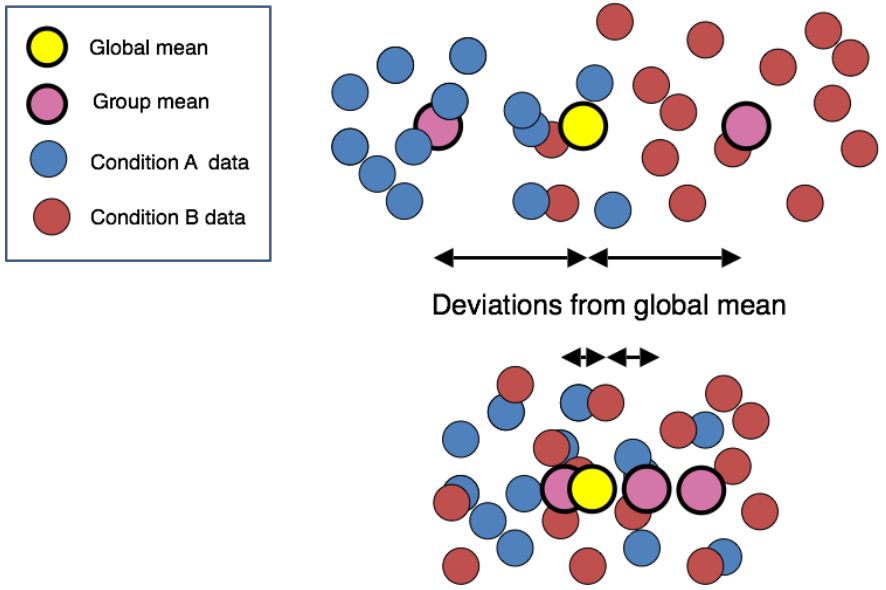
Secuenciación de célula única

En las técnicas de secuenciación de célula única todos los cDNAs de una misma célula se marcan con un código de barras único que permite luego analizar el transcriptoma individualizado de cada célula



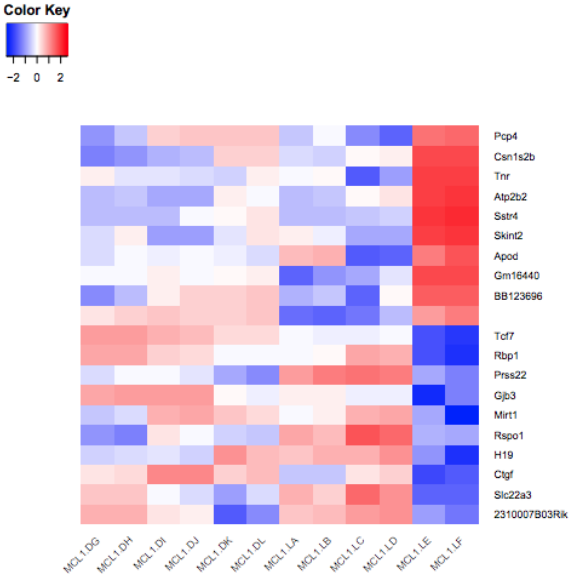
https://medicine.uiowa.edu/humangenetics/genomics-division/genome-sequencing/single-cell-expression-analysis-scrna-seq

Identificación de diferencias de expresión



Significant difference

No significant difference



https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-viz-with-heatmap2/tutorial.html

https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html

Librerías en R como ***limma*** (chips), ***DESeq*** (RNA-seq) o ***Seurat*** (scRNA-seq) permiten realizar los análisis de expresión diferencial en datos transcriptómicos.

Tema 9: Identificación de alteraciones transcriptómicas

9.1 Identificación de cambios en los patrones de expresión

9.2 Estudios de enriquecimiento de rutas moleculares

9.3 Clasificación y agrupamiento de muestras

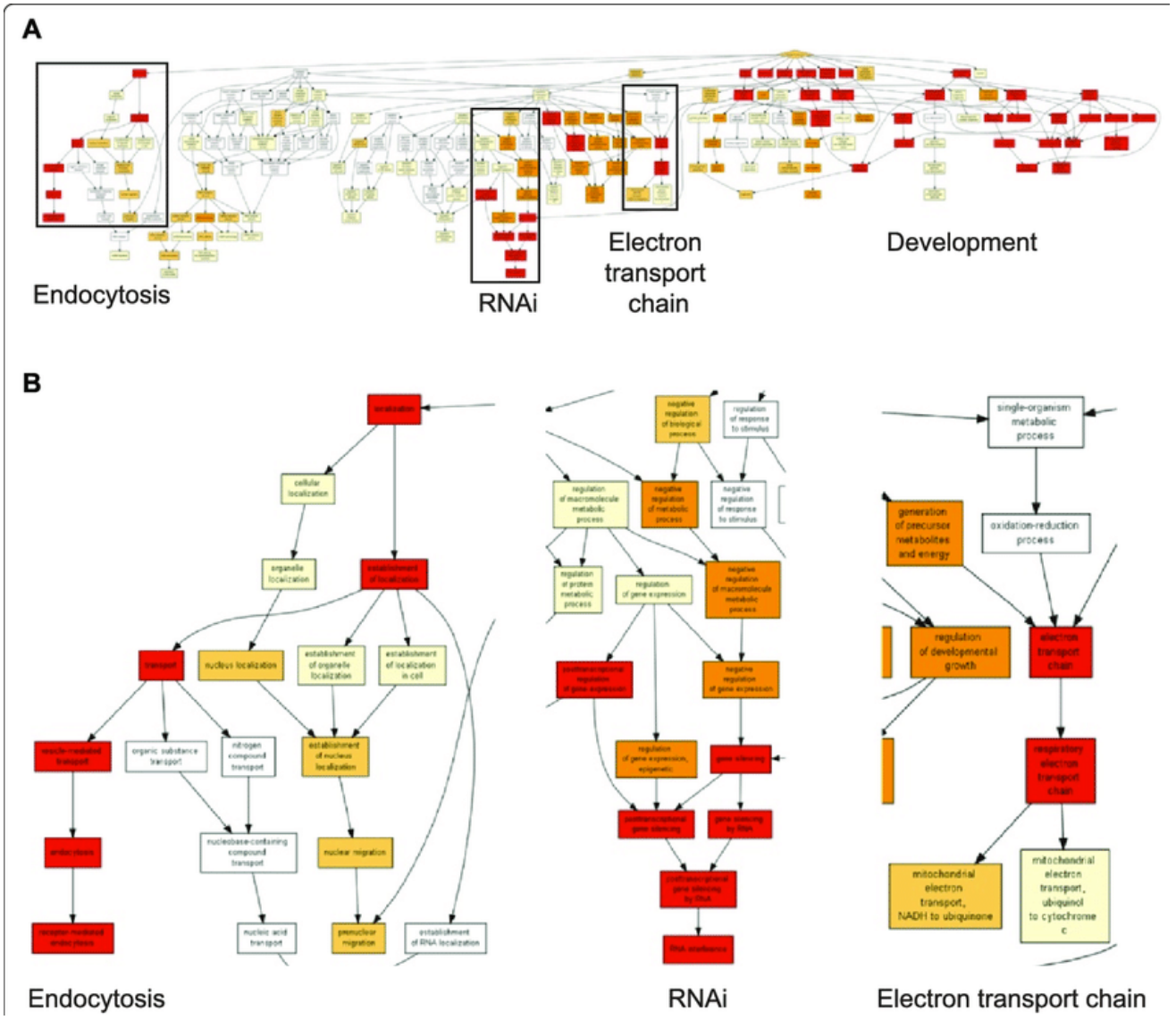
9.4 Generación de modelos de clasificación automática

9.5 Identificación de nuevos transcritos

9.6 Estudios de RNAs pequeños

Rutas moleculares

Gene Ontologies (GO)



Biological Process

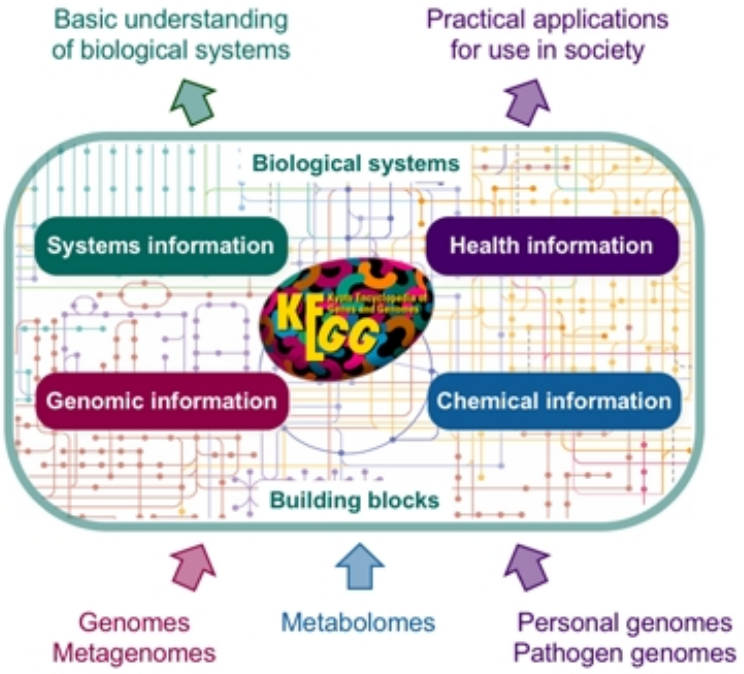
Molecular Function

Cellular Component

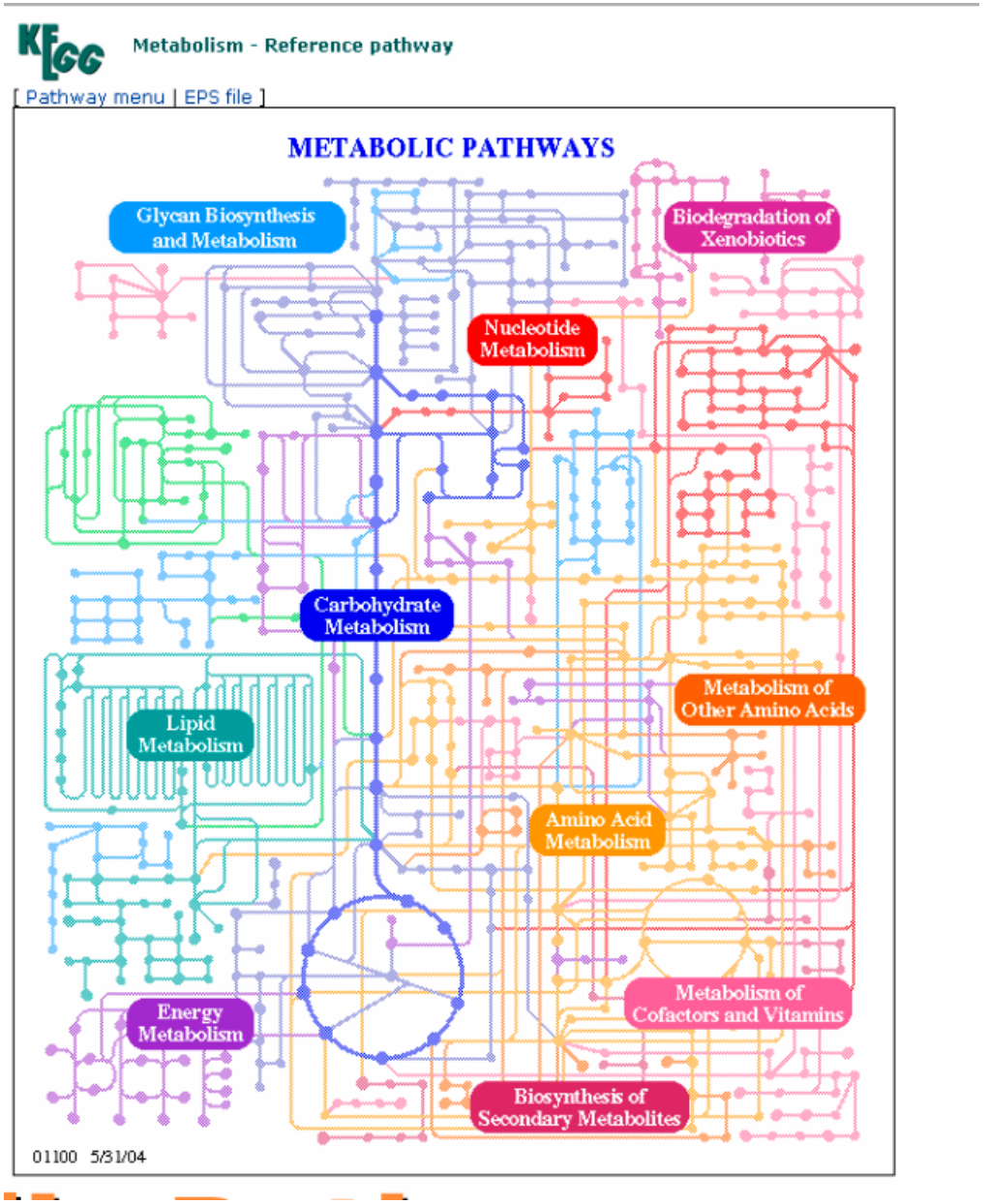
Zugasti O. et al. BMC Biology (2016) 14:35

Rutas moleculares clásicas

Kegg



https://www.genome.jp/kegg/pathway.html



Enriquecimiento de genes en rutas en listados de genes

GORILLA

Gene Ontology enRIchment anaLysis and visuaLizAtion tool

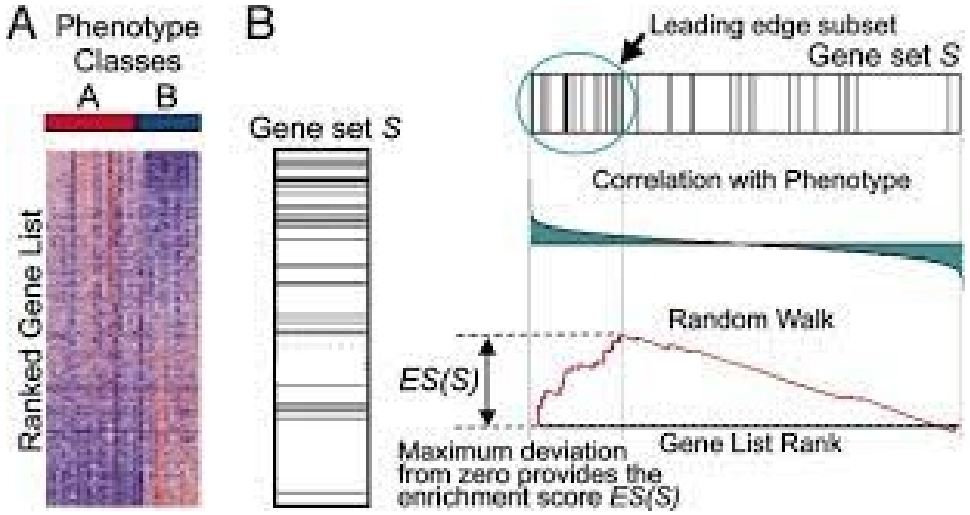
http://cbl-gorilla.cs.technion.ac.il/

AmiGO 2

http://amigo.geneontology.org/amigo

Diversas herramientas online como ***Gorilla*** o ***Amigo*** permiten identificar un enriquecimiento de genes pertenecientes a una determinada ruta u ontología (GO) de lo que cabría esperar por azar dentro de una lista de genes que muestran expresión diferencial.

Sesgo en la colocación de genes de rutas concretas en listas ordenadas de expresión



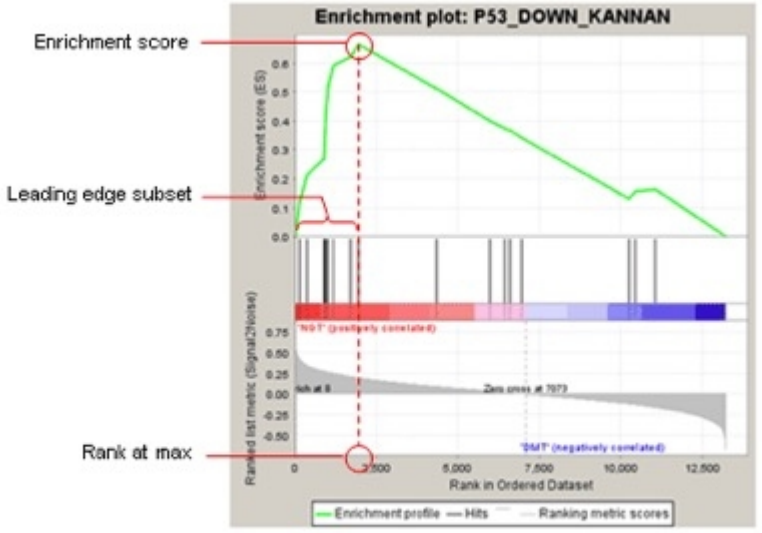


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Herramientas como ***GSEA*** permiten detectar desviaciones de la colocación aleatoria de los genes de una ruta dentro de un listado ordenado de genes de acuerdo a su expresión. Esto permite una mayor sensibilidad a la hora de detectar alteraciones en rutas u ontologías concretas.

Tema 9: Identificación de alteraciones transcriptómicas

9.1 Identificación de cambios en los patrones de expresión

9.2 Estudios de enriquecimiento de rutas moleculares

9.3 Clasificación y agrupamiento de muestras

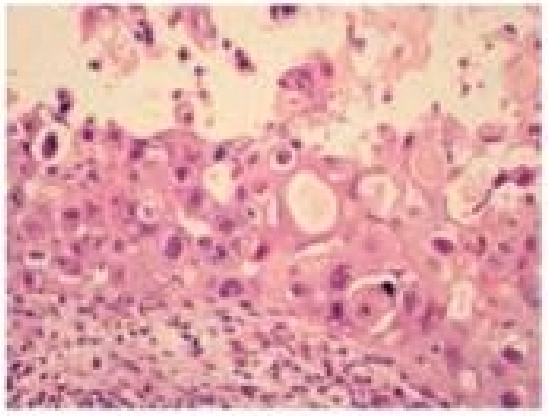
9.4 Generación de modelos de clasificación automática

9.5 Identificación de nuevos transcritos

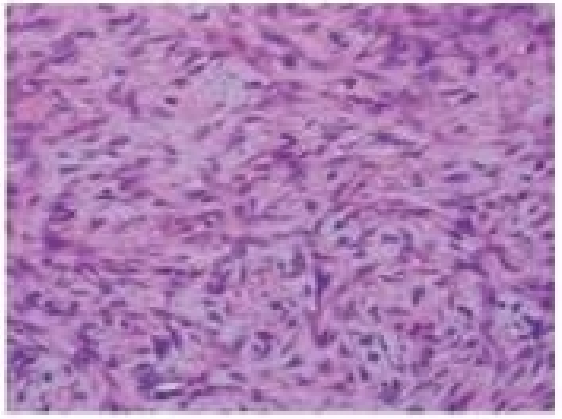
9.6 Estudios de RNAs pequeños

Variabilidad entre pacientes

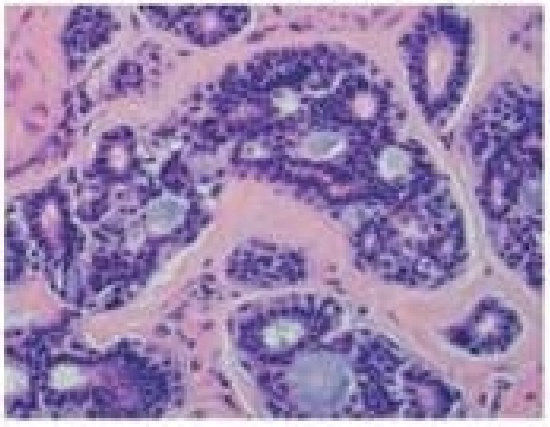
Metaplastic carcinoma with squamous differentiation



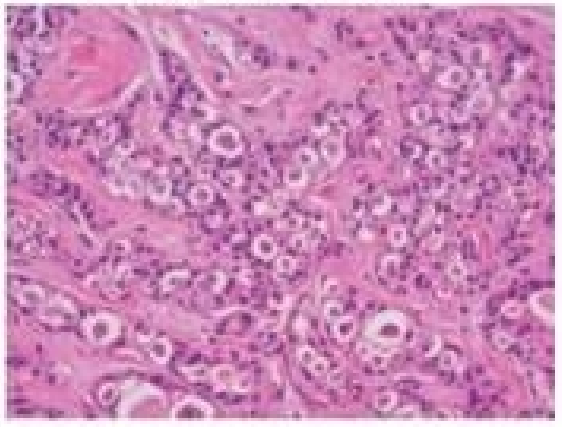
Spindle-cell metaplastic carcinoma



Adenoid cystic carcinoma

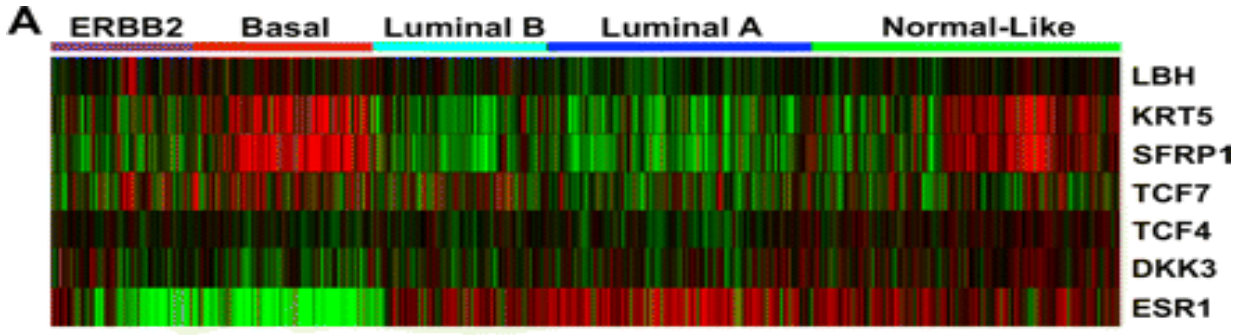


Secretory carcinoma



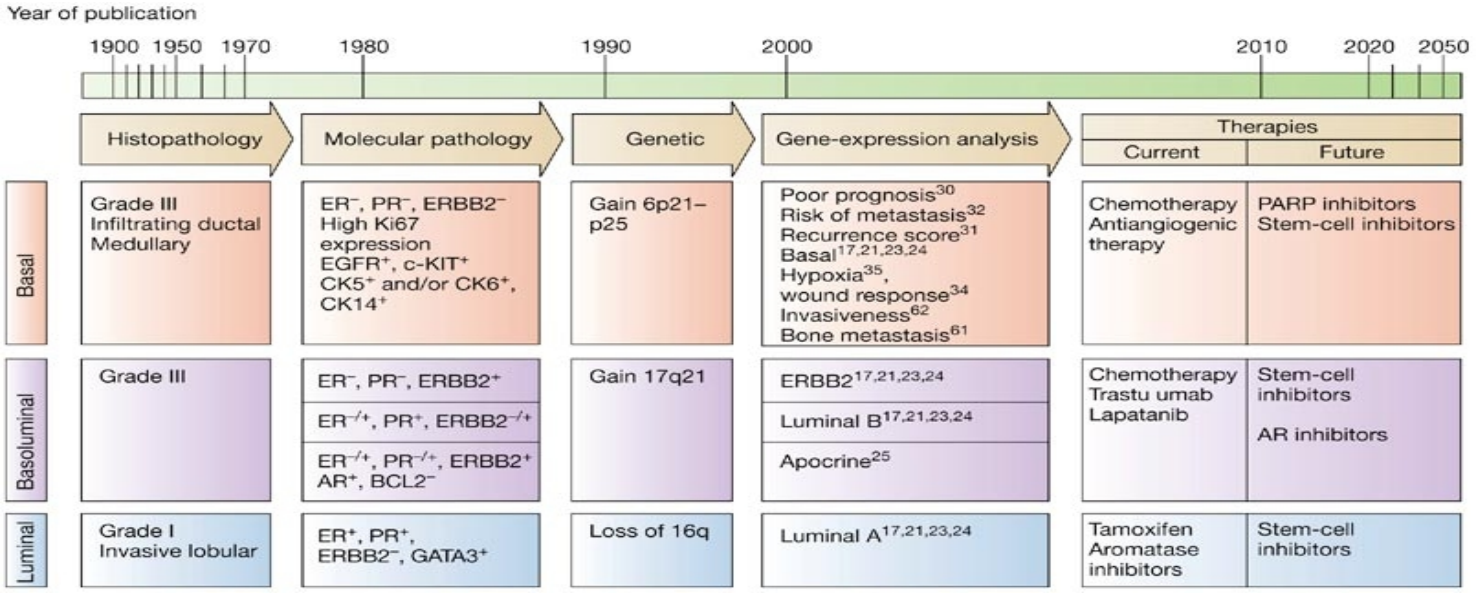
Nat Rev Clin Oncol. 2016 Nov;13(11):674-690

Utilidad de la clasificación molecular en tumores



B

Tumor type	Total number of cases	Number with high-level LBH expression	Proportion of subtype with high LBH expression
Normal-like	195	47	24%
Luminal A	286	46	16%
Luminal B	109	25	23%
ERBB2	154	42	27%
Basal	112	50	45%

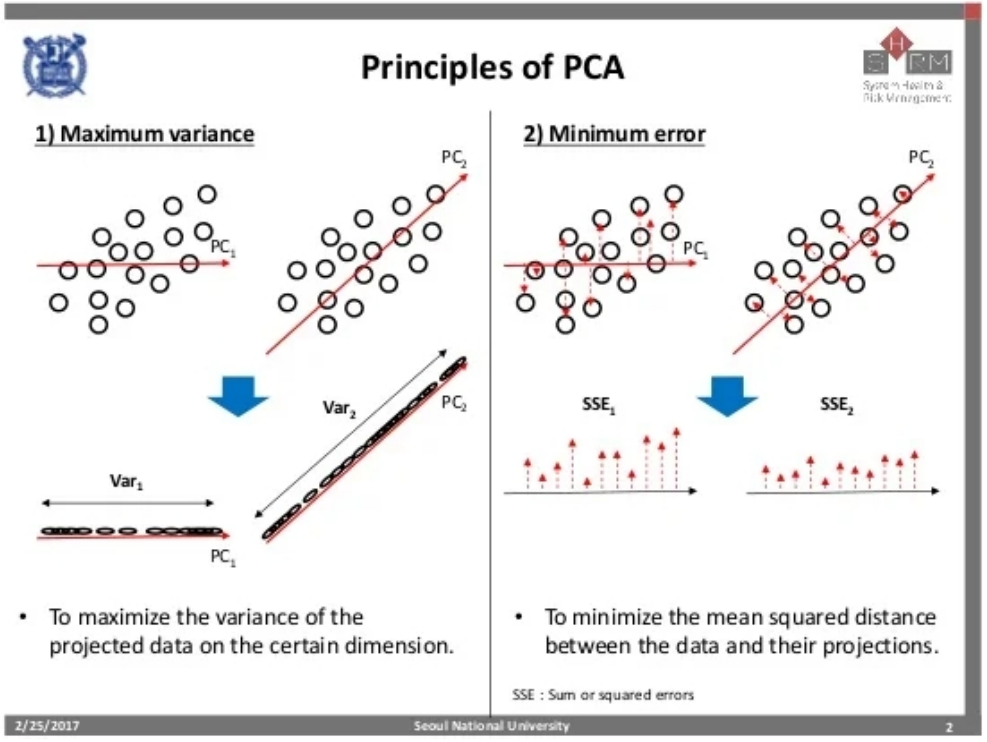


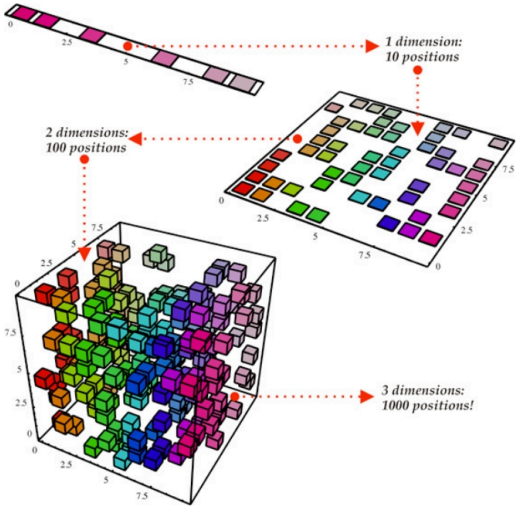
http://www.nature.com/nrclinonc/journal/v4/n9/fig_tab/ncponc0908_F1.html

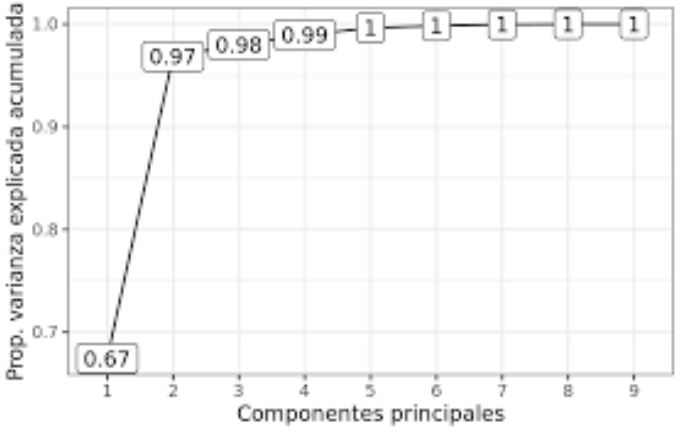
Reducción de la dimensionalidad

El objetivo de la reducción de la dimensionalidad es generar menos dimensiones combinando las dimensiones originales pero intentando mantener al máximo la varianza entre las muestras minimizando el error producido en las proyecciones.

Análisis de componentes principales



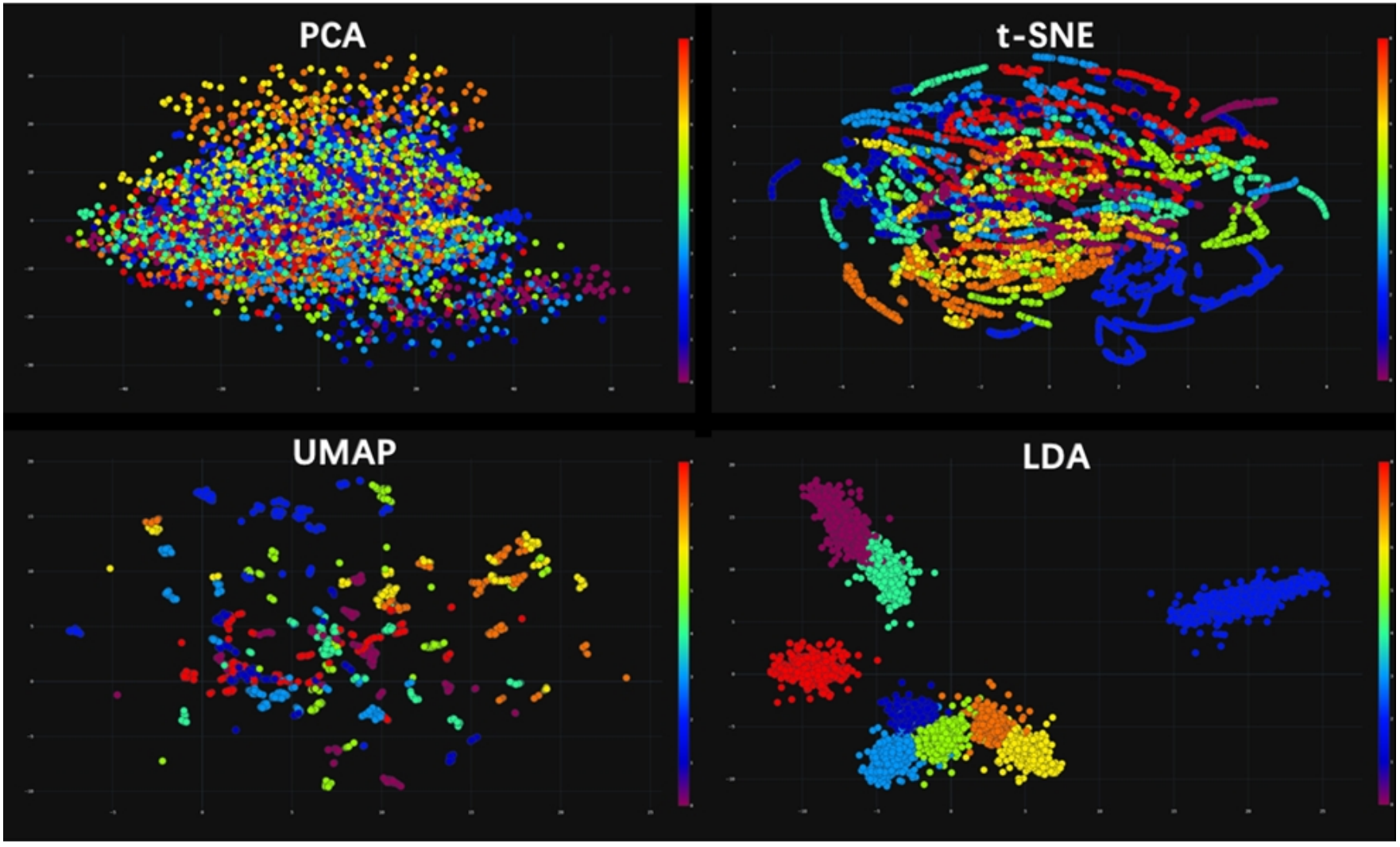




https://www.freecodecamp.org/news/an-overview-of-principal-component-analysis-6340e3bc4073/
https://pt.slideshare.net/hihijungho/pca-principal-component-analysis
https://www.cienciadedatos.net/documentos/35_principal_component_analysis

Reducción de la dimensionalidad

Distintos métodos de reducción de la dimensionalidad



https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29

Clusterizado jerarquizado

El clusterizado jerarquizado es quizás la estrategia más utilizada para agrupar las muestras de acuerdo a la expresión de sus genes. Se puede hacer en ambas direcciones.

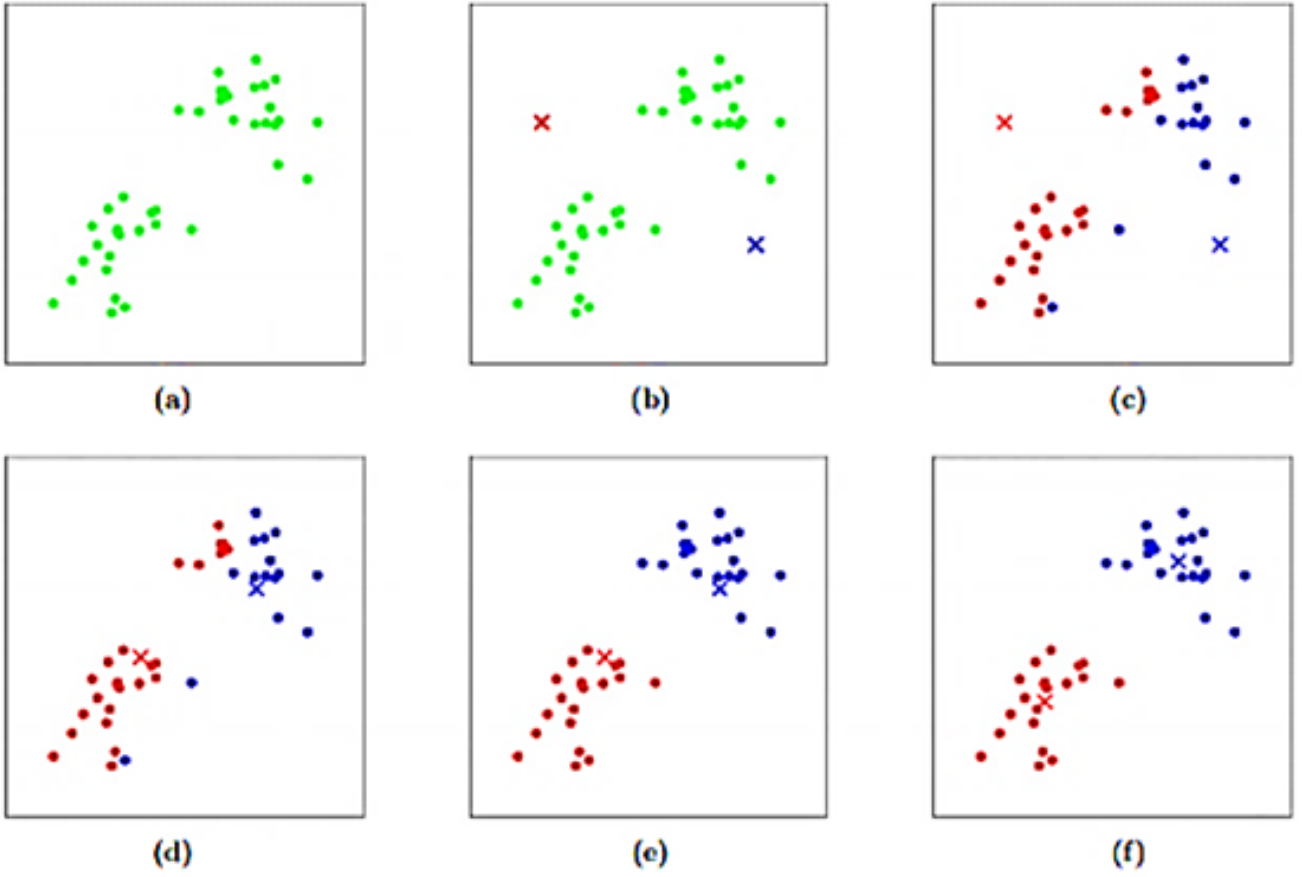
Se inicia con un único clúster que se va dividiendo a medida que se hacen las comparativas

Se inicia con un clúster por muestra que se van agrupando a medida que se realizan las comparativas

La medida clave para decidir las muestras que deben ir en el mismo clúster se denomina **distancia** y en muchos casos corresponde a la diferencia de expresión entre las variables normalizadas, bien la expresión directa de los genes o bien las nuevas dimensiones generadas en la reducción de la dimensionalidad.

Otras estrategias de clusterizado: k-means

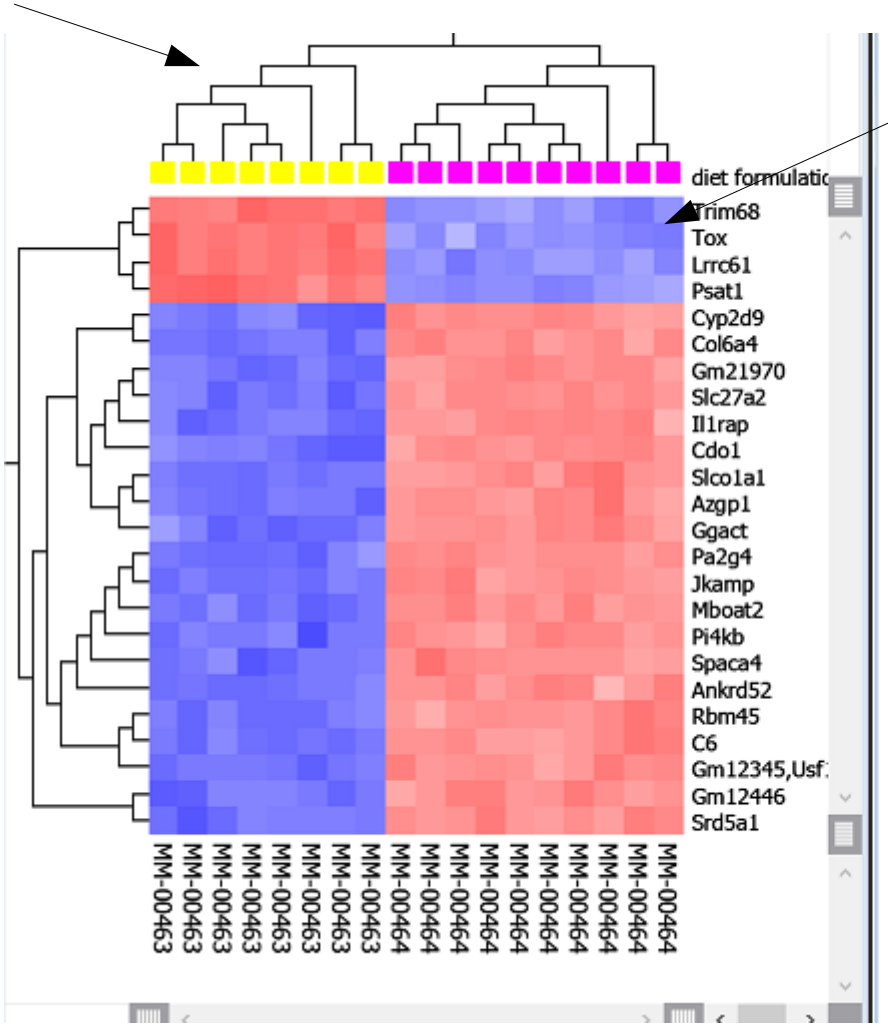
La técnica de medias k o k-means utiliza un método interactivo de agrupación de las muestras y cálculo de centros de los grupos con reasignación posterior de acuerdo a proximidad. El algoritmo repite la operación hasta que “converge” y no se producen más cambios de asignación



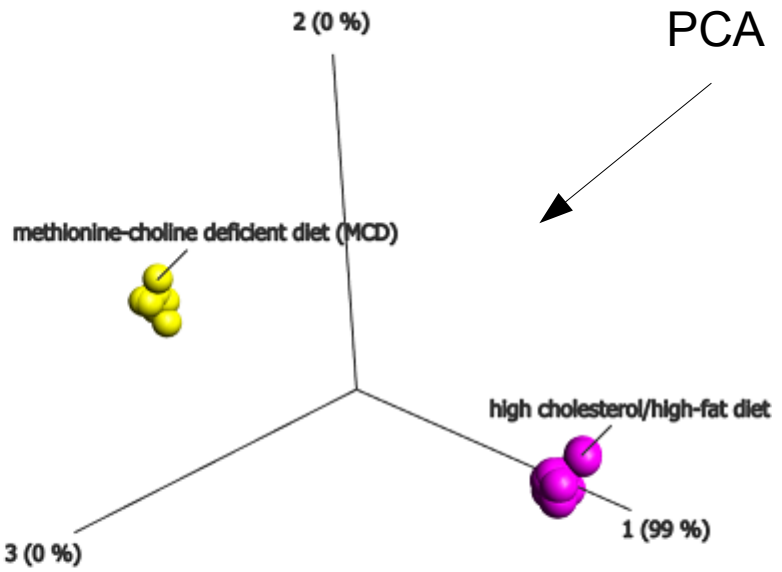
https://www.jparzival.com/blog/como-funciona-k-means/

Representación de datos transcriptómicos

Clusterizado



Heatmap

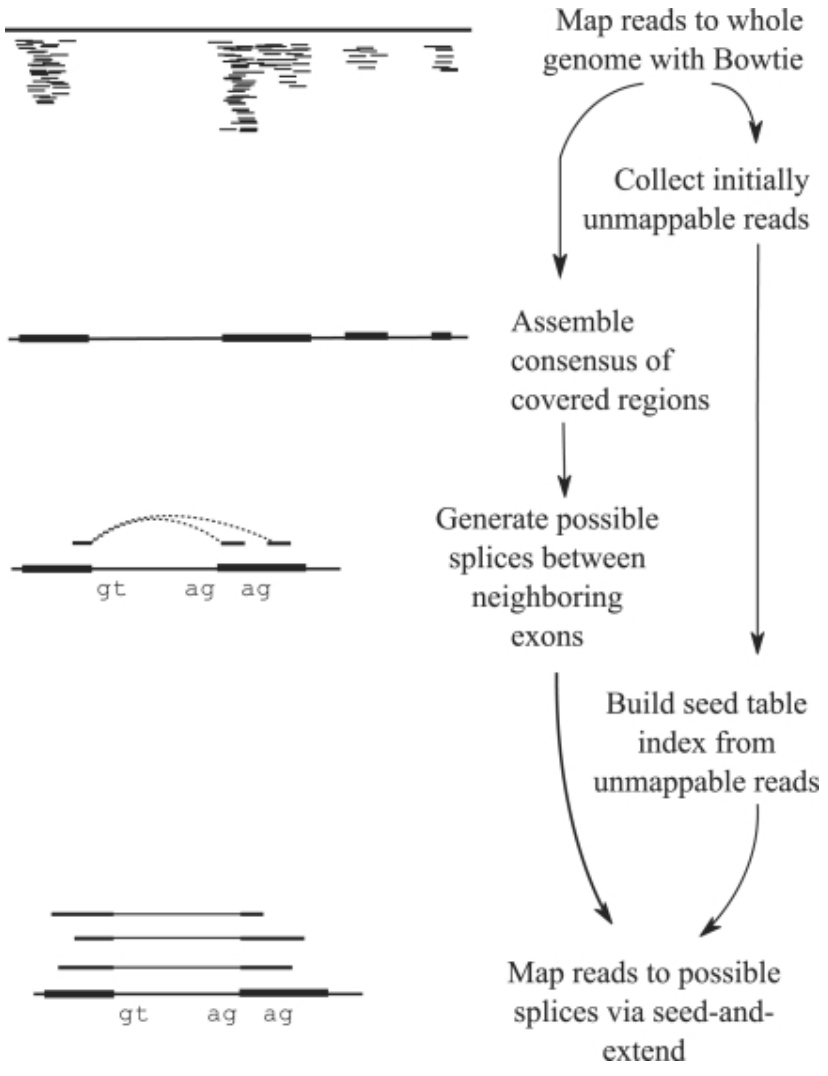


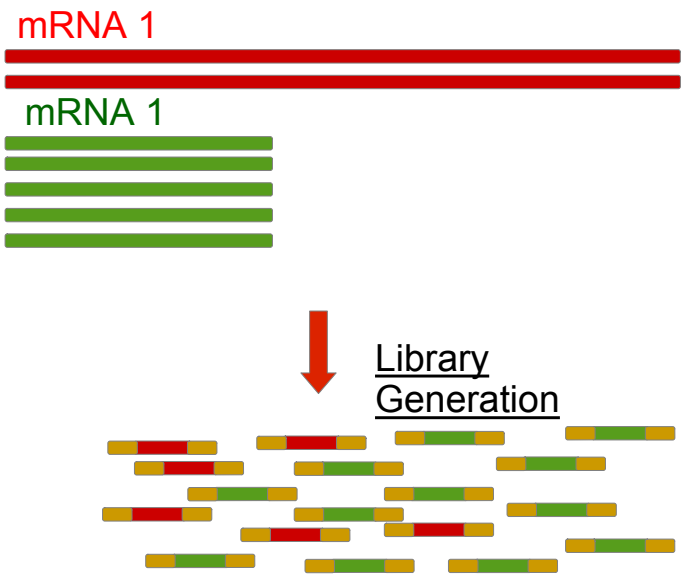
PCA

https://qlucore.com/news/qlucore-newsletter-analyze-your-rna-seq-data-in-the-light-of-public-data

Análisis de datos de célula única

El alineamiento y el contaje de transcritos, teniendo en cuenta los códigos de barras de cada célula individual es muy similar al análisis de datos de RNA-seq normal



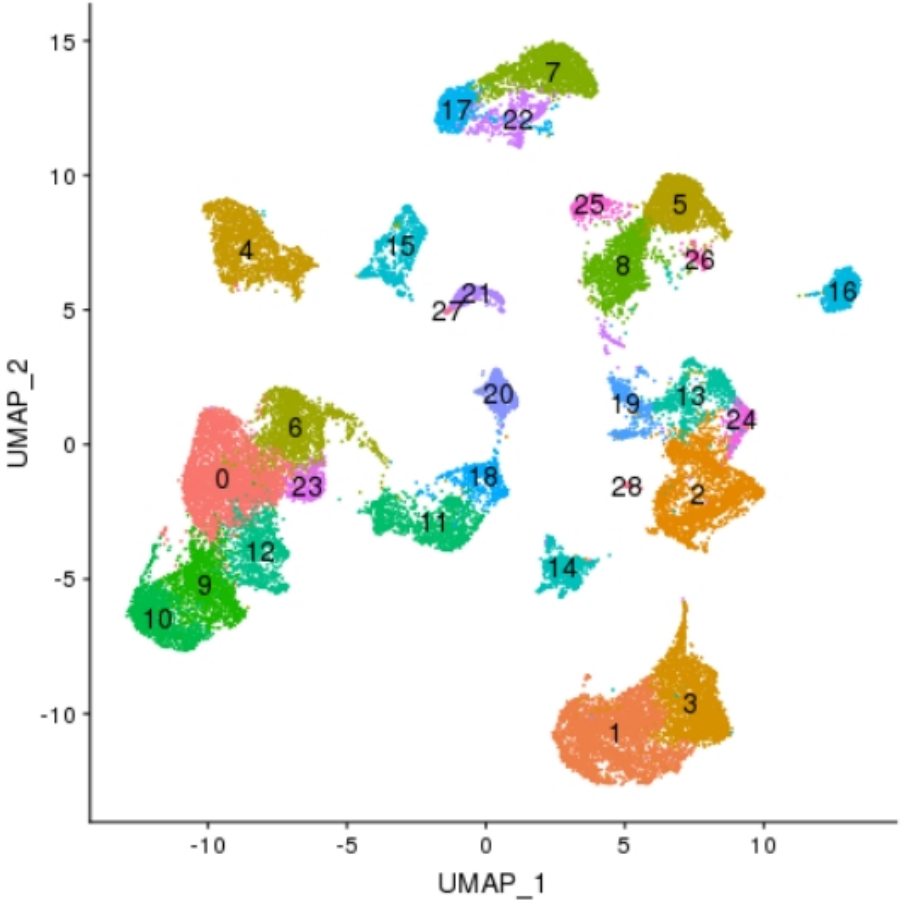


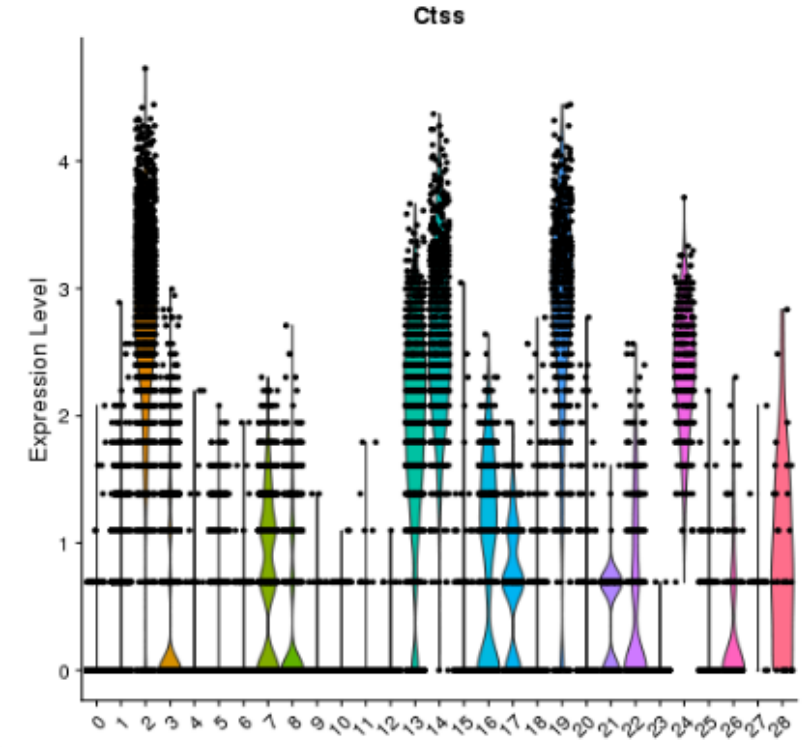
	Cell 1	Cell 2	Cell 3	...
Gene 1	100	50	0	
Gene 2	300	25	30	
Gene 3	0	65	80	
...				

Trapnell C. et al. Bioinformatics (2009) 25:1105-1111

Análisis de datos de célula única

Las estrategias de reducción de la dimensionalidad y clusterizado son especialmente importante en el análisis de datos de secuenciación de célula única donde tenemos frecuentemente información de miles de muestras (células) con miles de variables (genes)





Tema 9: Identificación de alteraciones transcriptómicas

9.1 Identificación de cambios en los patrones de expresión

9.2 Estudios de enriquecimiento de rutas moleculares

9.3 Clasificación y agrupamiento de muestras

9.4 Generación de modelos de clasificación automática

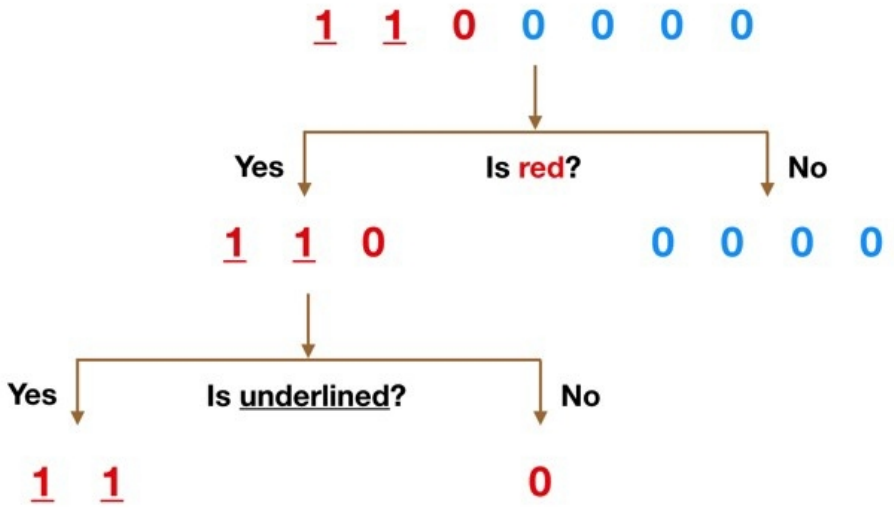
9.5 Identificación de nuevos transcritos

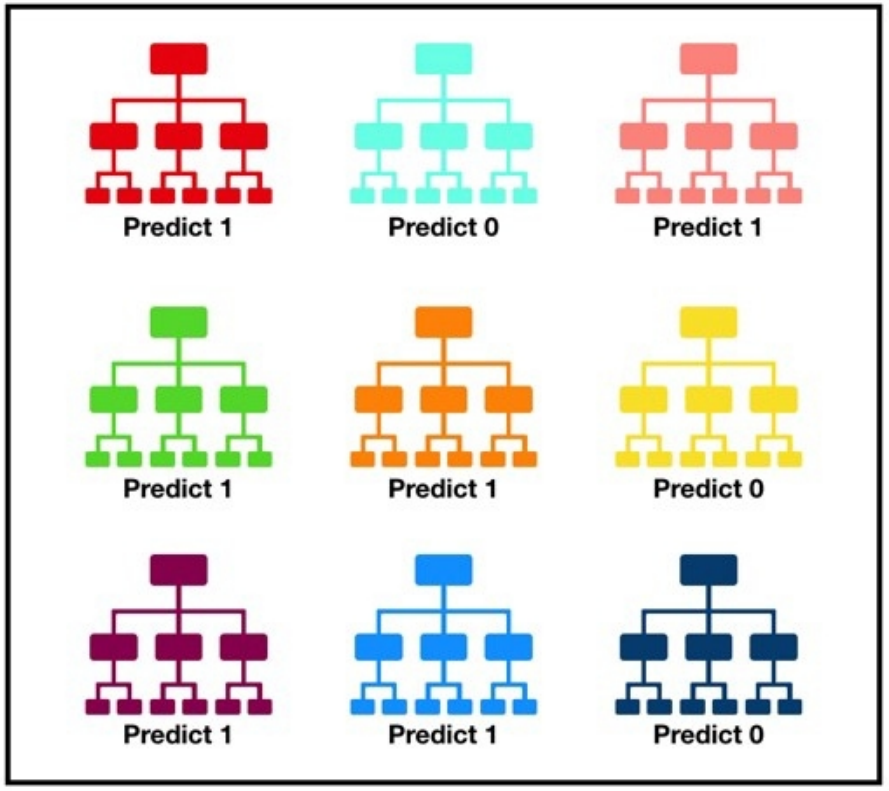
9.6 Estudios de RNAs pequeños

Modelos automáticos de clasificación

Árboles aleatorios o *Random Forests*

En un grupo de datos de entrenamiento en donde se conoce la clasificación, el ordenador construye árboles de decisión para separar las muestras de acuerdo a un grupo aleatorio de variables (expresión de genes).





https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree.

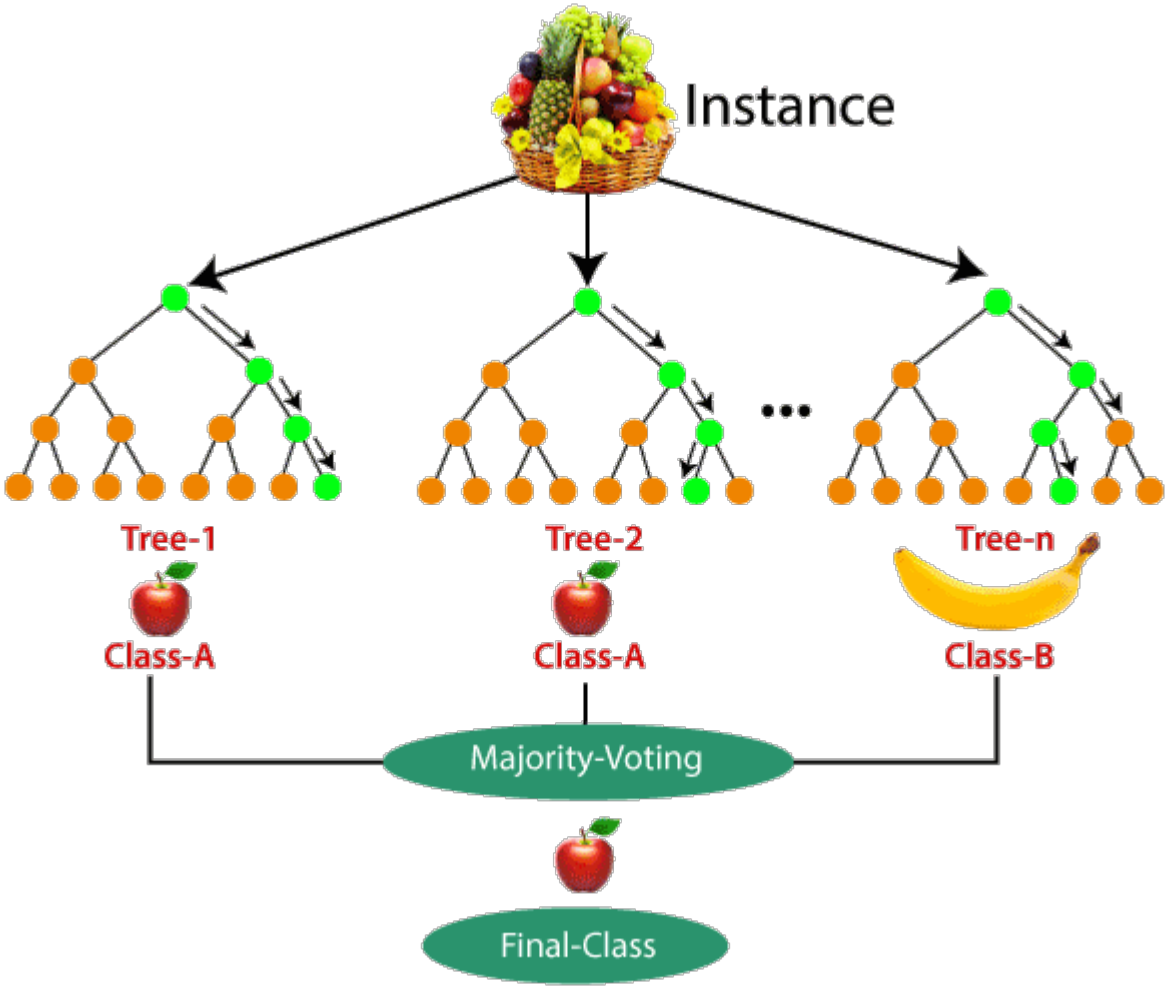
Tally: Six 1s and Three 0s

Prediction: 1

Aunque algunos de los árboles de decisión clasifiquen incorrectamente, si este error no es sistemático o relacionado entre distintos árboles de decisión, la suma de la actuación conjunta de los árboles tiene una mayor probabilidad de acertar en la clasificación.

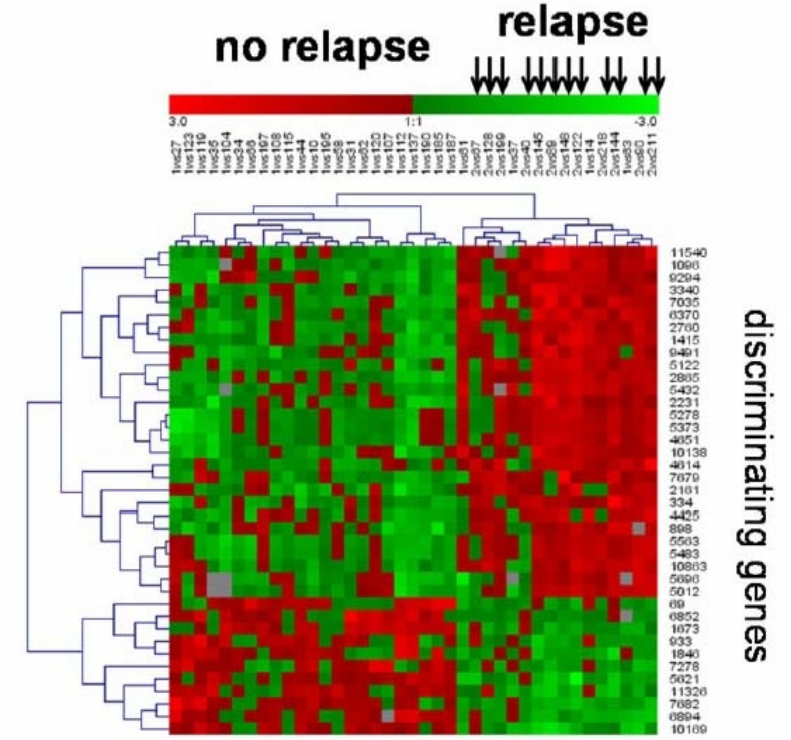
Modelos automáticos de clasificación

Árboles aleatorios o *Random Forests*

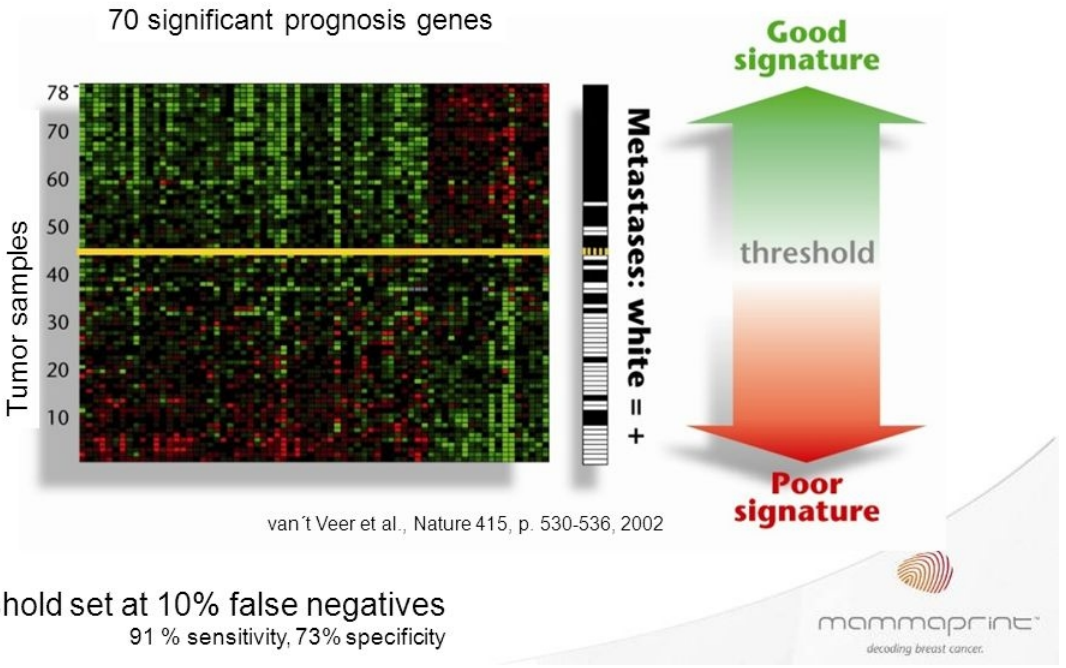


No hay control sobre las variables que se están usando para clasificar los eventos. Por lo tanto los modelos dependen mucho de los datos de entrenamiento. Cuanto más diversos (ausencia de sesgos) y más parecidos a los casos problema, más acertado será el modelo predictivo.

Modelos de predicción de riesgo



https://www.radioncologa.com/2013/02/cancer-de-mama-mammaprint-y-oncotype/



https://slideplayer.com/slide/5863447/

Otros métodos de *machine learning*

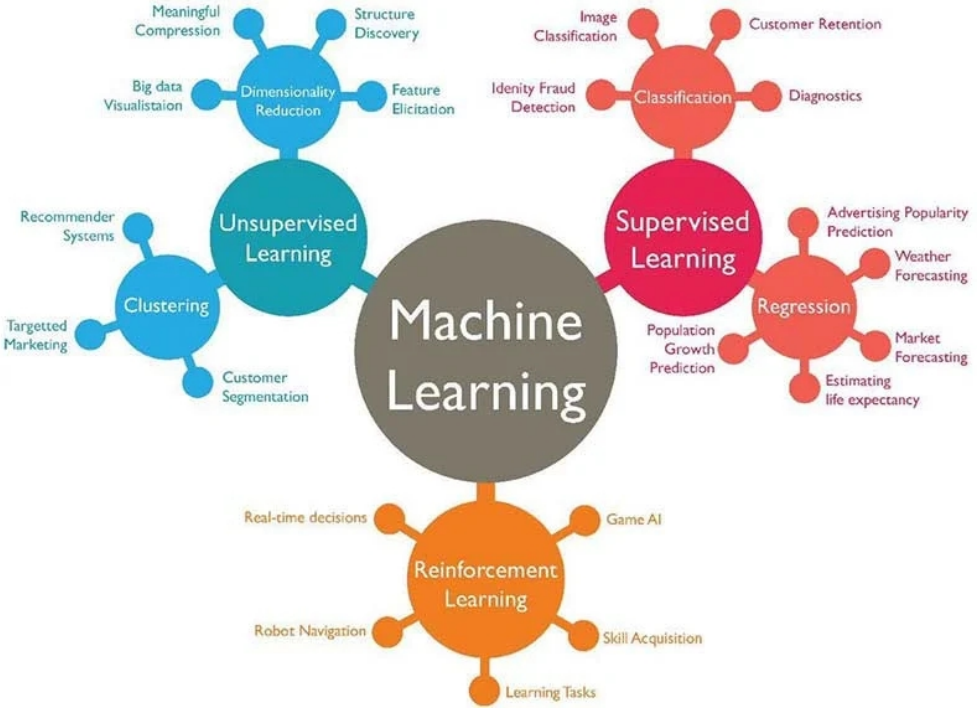
Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised

Continuous	• Clustering & Dimensionality Reduction
	  ○ SVD
	  ○ PCA
	  ○ K-means
		• Regression
		  ○ Linear
		  ○ Polynomial
		• Decision Trees
		• Random Forests

Categorical	• Association Analysis
	  ○ Apriori
	  ○ FP-Growth
	• Hidden Markov Model
		• Classification
		  ○ KNN
		  ○ Trees
		  ○ Logistic Regression
		  ○ Naive-Bayes
		  ○ SVM

https://www.researchgate.net/figure/Machine-learning-problems-classification-and-useful-algorithms_fig3_316991321



https://www.fiverr.com/rkhalid95/solve-machine-learning-problems

Tema 9: Identificación de alteraciones transcriptómicas

9.1 Identificación de cambios en los patrones de expresión

9.2 Estudios de enriquecimiento de rutas moleculares

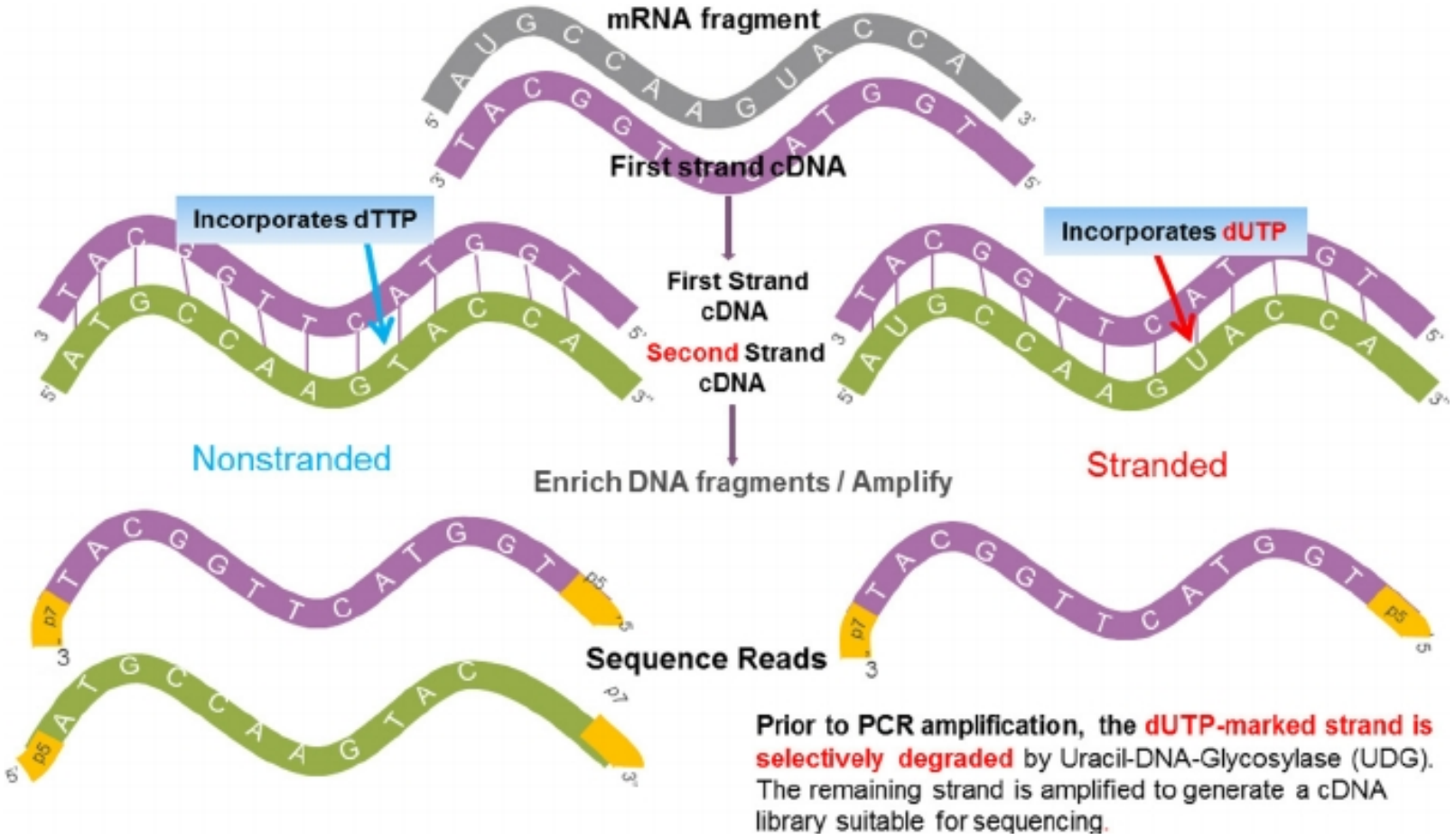
9.3 Clasificación y agrupamiento de muestras

9.4 Generación de modelos de clasificación automática

9.5 Identificación de nuevos transcritos

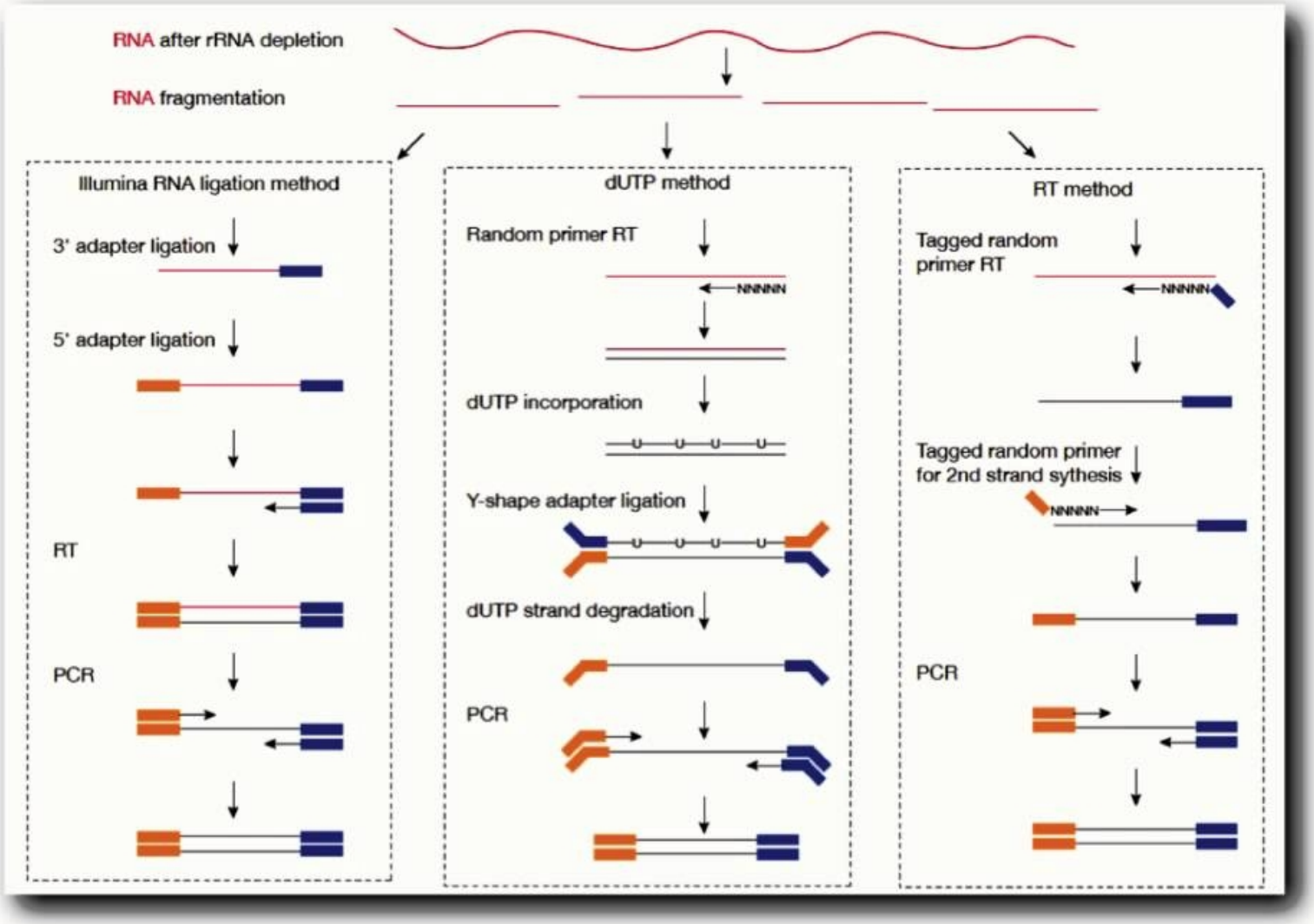
9.6 Estudios de RNAs pequeños

Información de hebra en las librerías de RNA-seq



Zhao S et al. BMC Genomics 16:675

Generación de librerías con información de hebra

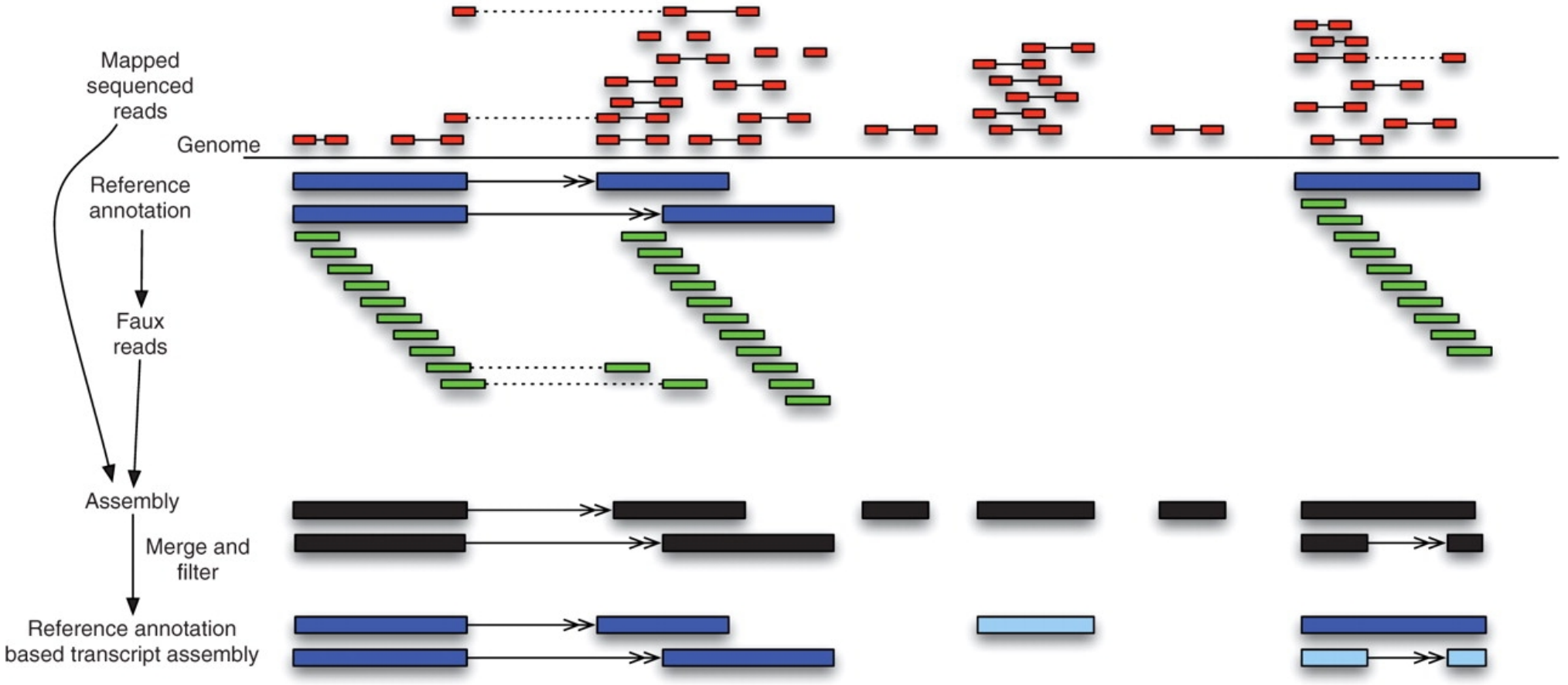


Strand-specific library protocols (Credit: Zhao Zhang)

Generación de librerías con información de hebra

https://www.ecseq.com/support/ngs/how-do-strand-specific-sequencing-protocols-work

Descubrimiento de nuevos mRNAs o isoformas



Herramientas como ***Cufflinks*** combinan alineamiento frente a un genoma de referencia y ensamblaje de novo para identificar nuevas isoformas o transcritos.

Tema 9: Identificación de alteraciones transcriptómicas

9.1 Identificación de cambios en los patrones de expresión

9.2 Estudios de enriquecimiento de rutas moleculares

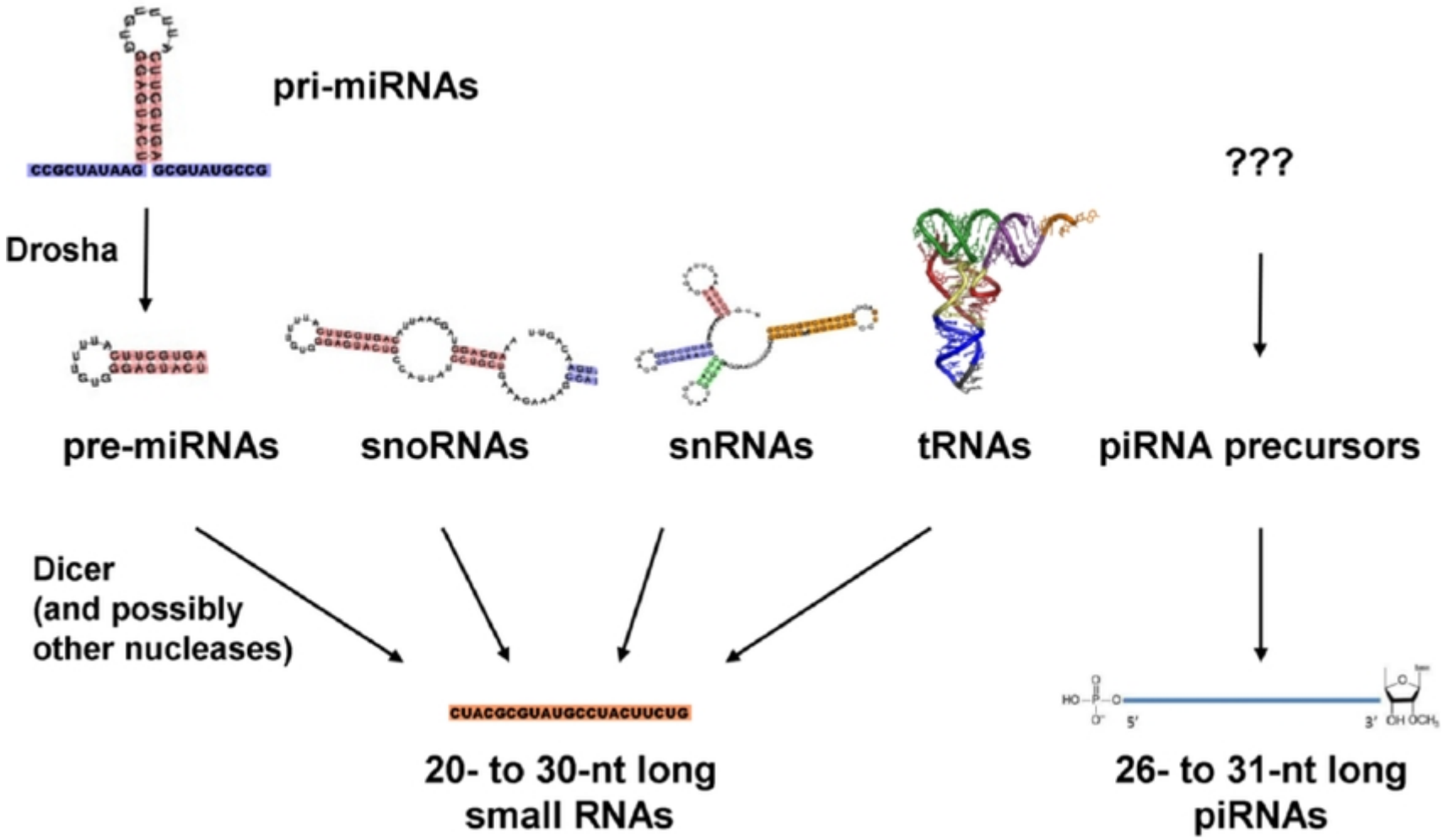
9.3 Clasificación y agrupamiento de muestras

9.4 Generación de modelos de clasificación automática

9.5 Identificación de nuevos transcritos

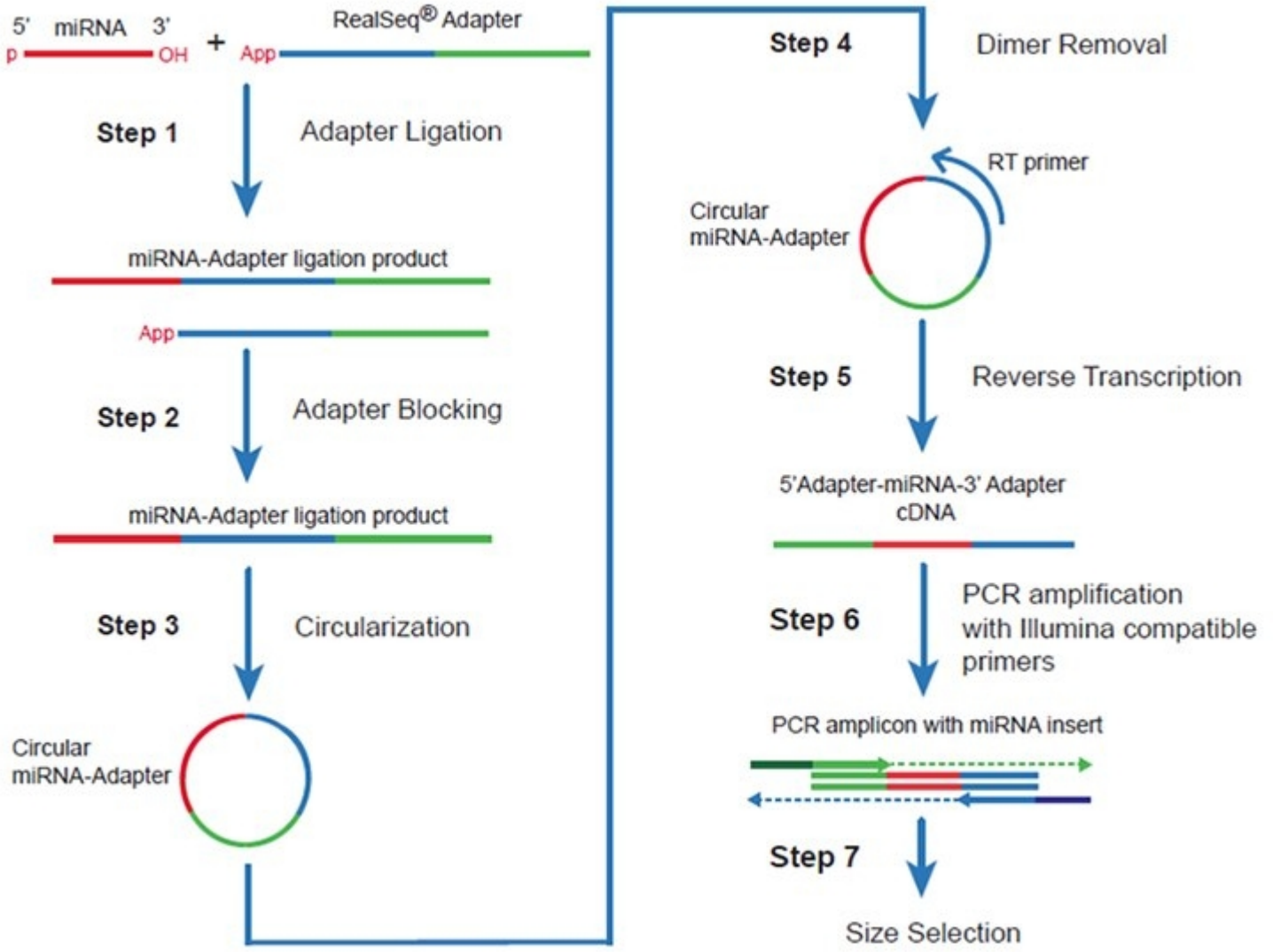
9.6 Estudios de RNAs pequeños

RNAs pequeños



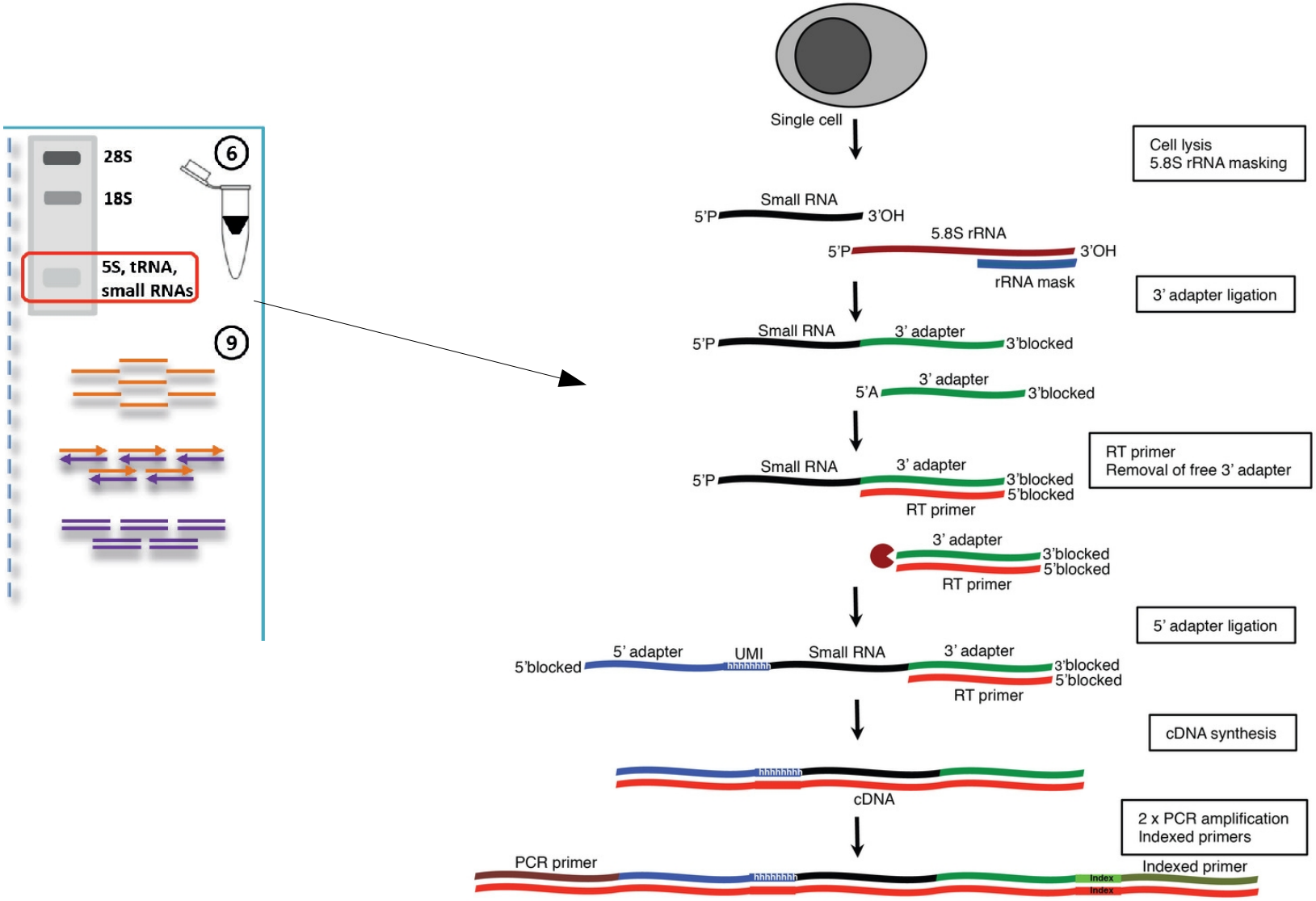
https://www.researchgate.net/figure/A-variety-of-RNA-classes-is-processed-into-20-to-30-nt-long-small-RNAs-For-details-see_fig1_45093158

Generación de librerías de RNAs pequeños



https://www.sigmaaldrich.com/ES/es/technical-documents/technical-article/genomics/gene-expression-and-silencing/small-rna-sequencing

Generación de librerías de RNAs pequeños



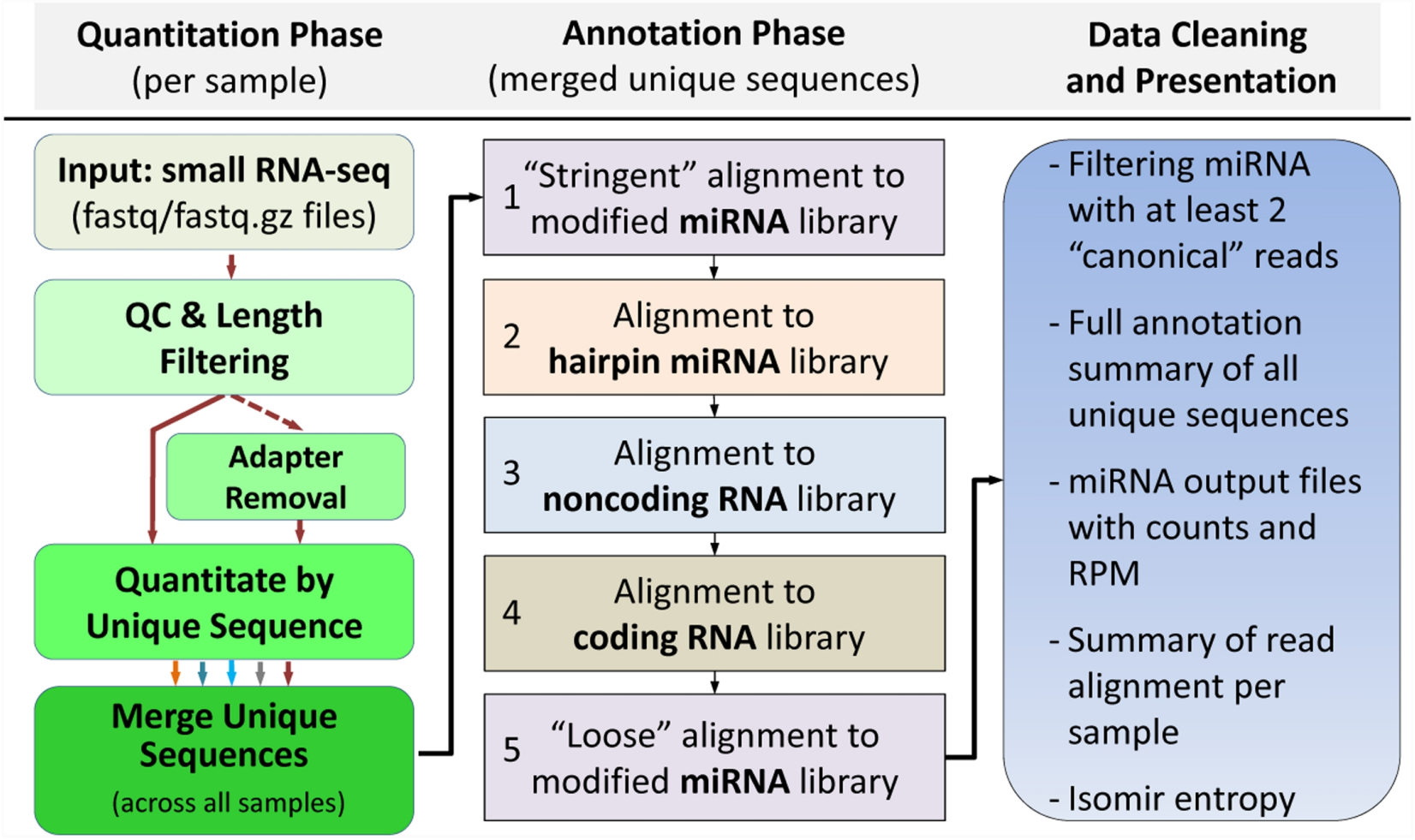
https://www.rna-seqblog.com/small-rna-sequencing-of-single-cells/

Problemas en el alineamiento de lecturas de RNAs pequeños

Solapamiento con mRNAs

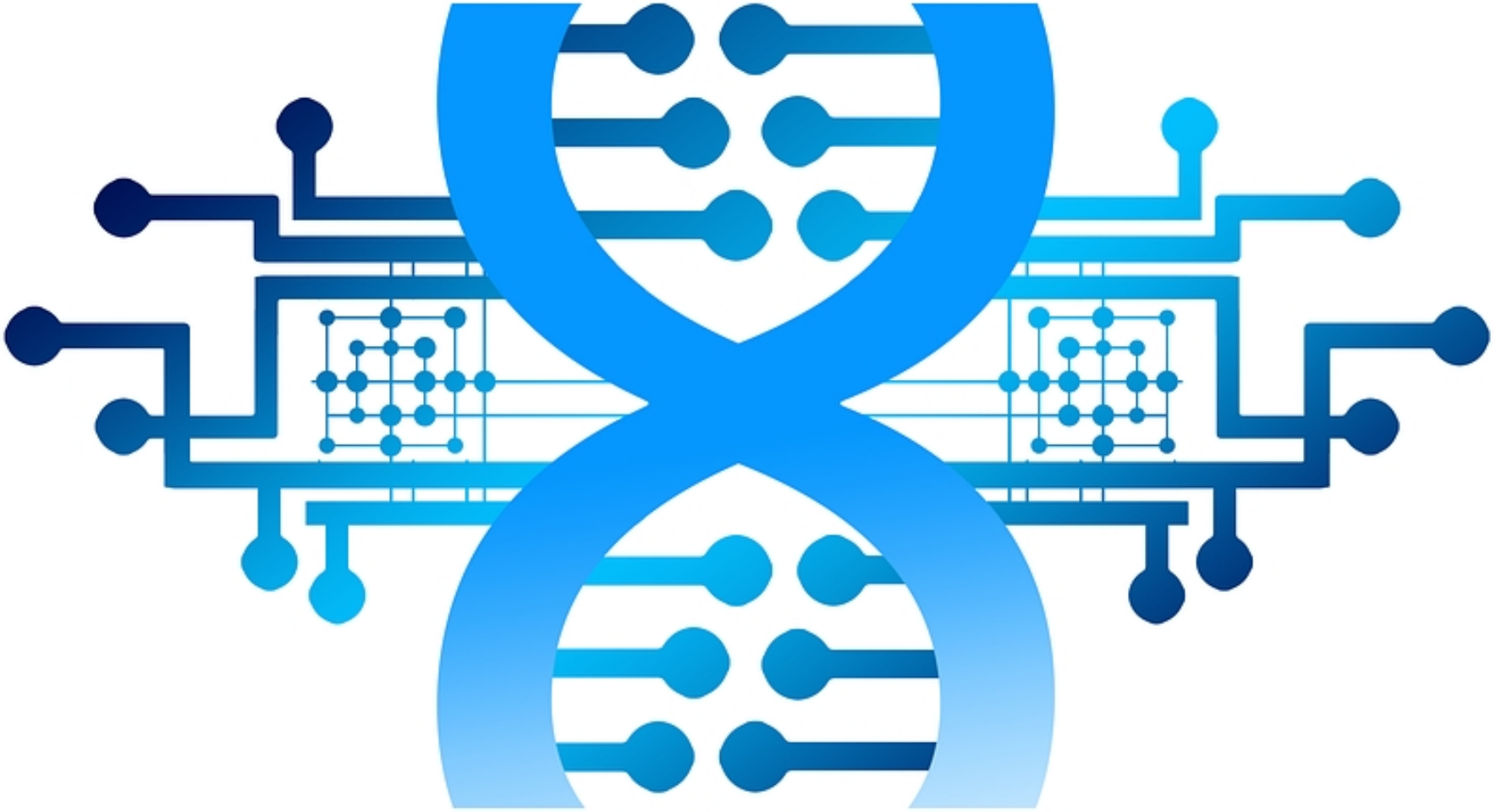
Procesamiento postranscripcional

Gran homología de secuencias



Baras AS et al. Plos One :10.1371/journal.pone.0143066

Bioinformática y análisis de datos ómicos



Tema 9: Identificación de alteraciones transcriptómicas

ignacio.varela@unican.es