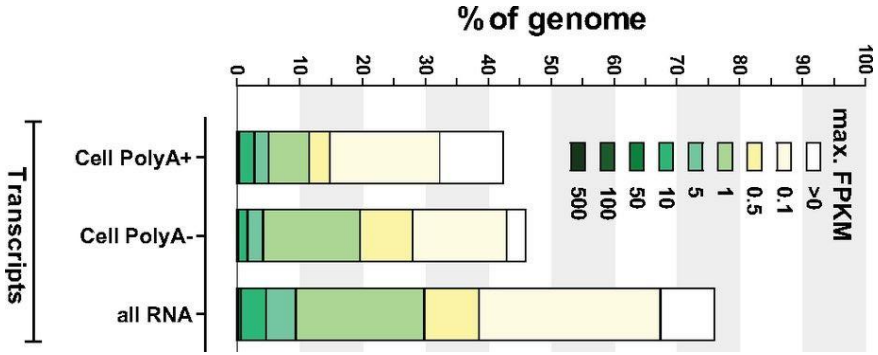
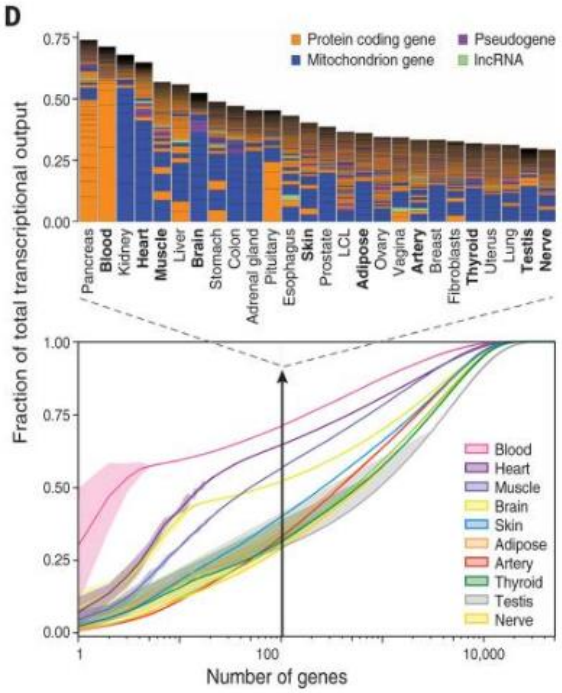
Análisis del transcriptoma

Guillermo Barturen Briñas
(gbarturen@ugr.es)

¿Qué es el transcriptoma?

- El transcriptoma de una célula es la colección de moléculas de ARN (transcritos) presentes en la célula. Actualmente y gracias a la evolución de las técnicas de secuenciación masiva se ha podido observar que más de un **75% del genoma puede transcribirse** (incluyendo elementos repetidos), aunque la gran mayoría presenta una transcripción residual.
- Del total de transcritos observados sólo **~1.5% codifica proteínas**, aunque eso no quiere decir que el resto no tengan función. Durante los últimos años se han identificado múltiples transcritos no-codificantes con función reguladora.





https://doi.org/10.1126%2Fscience.aaa0355

- A pesar de tener el mismo genoma, no todos los tejidos y/o células expresan los mismos genes. En promedio, **una célula puede expresar entre el 1-5%** del transcriptoma conocido dependiendo del estado de la misma y de su tipo celular.
- Por otro lado, una fracción pequeña de los genes transcritos en una célula (~1/100) monopolizan la gran mayoría de las moléculas presentes en la célula (~25%-70%). En la mayoría de los tejidos, estos genes pertenecen a la fracción mitocondrial. En el caso de la sangre, 3 genes de la hemoglobina contienen el 60% de los transcritos.

The GENCODE Project: Encyclopædia of genes and gene variants

https://www.gencodegenes.org/

General stats

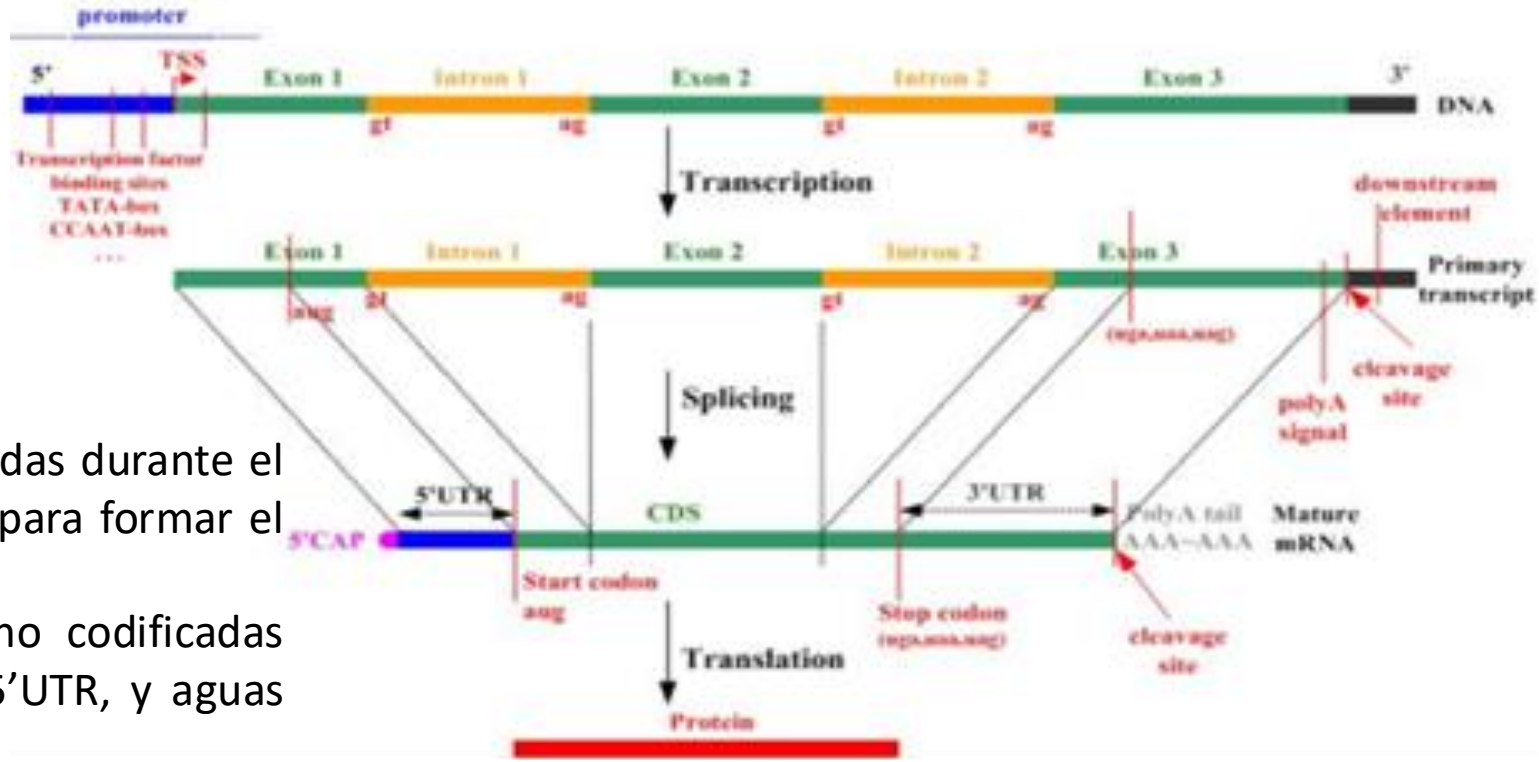
Total No of Genes	62700	Total No of Transcripts	252835
Protein-coding genes	19396	Protein-coding transcripts	89067
- readthrough genes (not included)	650	- full length protein-coding	63968
Long non-coding RNA genes	19922	- partial length protein-coding	25099
Small non-coding RNA genes	7566	Nonsense mediated decay transcripts	21384
Pseudogenes	14735	Long non-coding RNA loci transcripts	58246
- processed pseudogenes	10660		
- unprocessed pseudogenes	3570		
- unitary pseudogenes	254		
- pseudogenes	15	Total No of distinct translations	65342
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13594
- protein coding segments	411		
- pseudogenes	236		

- GENCODE, es un proyecto cuyo objetivo es mantener y mejorar la anotación de genes en el genoma humano y de ratón. Está anotación incluye curación manual, análisis computacionales y experimentos para validar ciertos transcriptos.

- Actualmente, GENCODE ha identificado ~62.000 genes de los cuáles ~19.000 tienen capacidad codificante. El barajamiento de exones (*splicing* alternativo) hace que el transcriptoma sea muy versátil presentando más de 250.000 isoformas a partir de esos genes.

Barajamiento de exones (*Splicing* alternativo)

- A diferencia de los genes procariotas, los eucariotas tenemos genes monocistrónicos (información para una sola cadena polipeptídica). Además, los genes eucariotas suelen contener secuencias que interrumpen su marco abierto de lectura (ORF), denominadas intrones. Estos intrones complican el proceso de traducción, debiendo ser eliminados durante el procesamiento del transcripto primario (alternative splicing) para dar lugar al transcrito maduro.

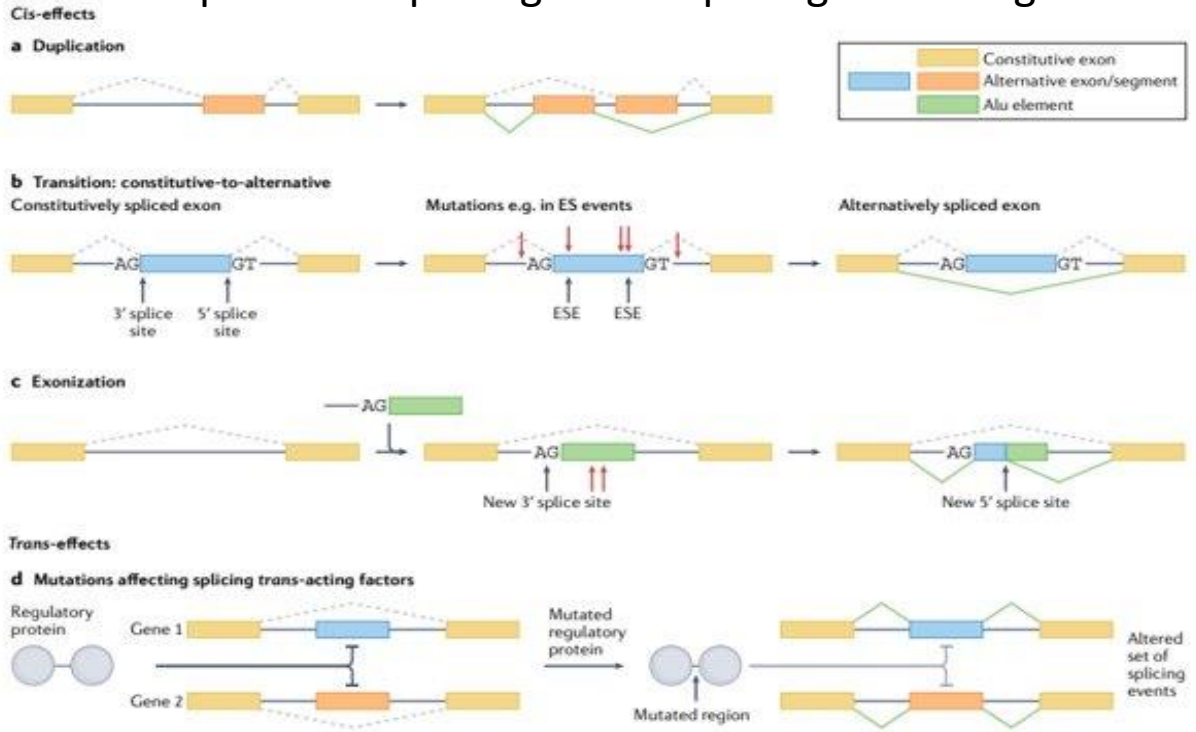


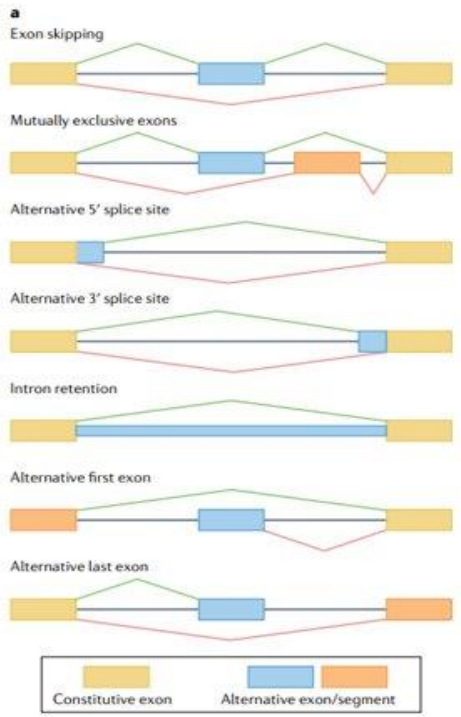
Gen eucariota

- Exones: transcrito maduro.
- Intrones: transcrito primario, eliminadas durante el proceso de barajamiento de exones para formar el transcrito maduro.
- Secuencias UTR, transcrito maduro, no codificadas (pueden encontrarse aguas arriba, 5'UTR, y aguas abajo del transcrito maduro, 3'UTR).
- CDS, regiones del gen que codifican a proteínas.

Evolución de la estructura génica

- El barajamiento de exones proporciona una inmensa plasticidad funcional y fenotípica al genoma, de esta manera un solo gen puede presentar múltiples isoformas con diferentes funcionalidades.
- Nuevas isoformas pueden aparecer mediante varios mecanismos:
  i. Duplicación de exones: dando lugar a exones mutuamente excluyentes.
  ii. Transición por mutaciones de secuencia de exones constitutivos a alternativos.
  iii. Exonización de elementos repetidos que tras mutación pueden generar nuevos sitios donadores o aceptores.
  iv. Mutaciones en proteínas que regulan el splicing de otros genes.





Análisis masivo de transcriptomas

- Originalmente los análisis masivos de transcriptoma se basaban principalmente en tecnología de *microarrays*. Sin embargo, la alta reproducibilidad de los resultados provenientes de secuenciación masiva y su caída de precio han llevado a que desaparezcan las plataformas de expresión basadas en *microarrays*.

- Los análisis de RNA-Seq son diversos e implican diferentes tecnologías y protocolos en función de lo que se quiera estudiar:
  i. Análisis de expresión
  ii. Análisis de isoformas y splicing alternativo
  iii. Análisis de variantes de secuencia
  iv. Análisis de modificaciones del ARN
  v. Ensamblado de *novo*
  vi. ARNs cortos

Análisis masivo de transcriptomas

- El estudio de ARNs largos puede abordarse con diferentes tecnologías con sus ventajas e inconvenientes.

- **Secuenciación de cDNA de lecturas cortas** ("Illumina")

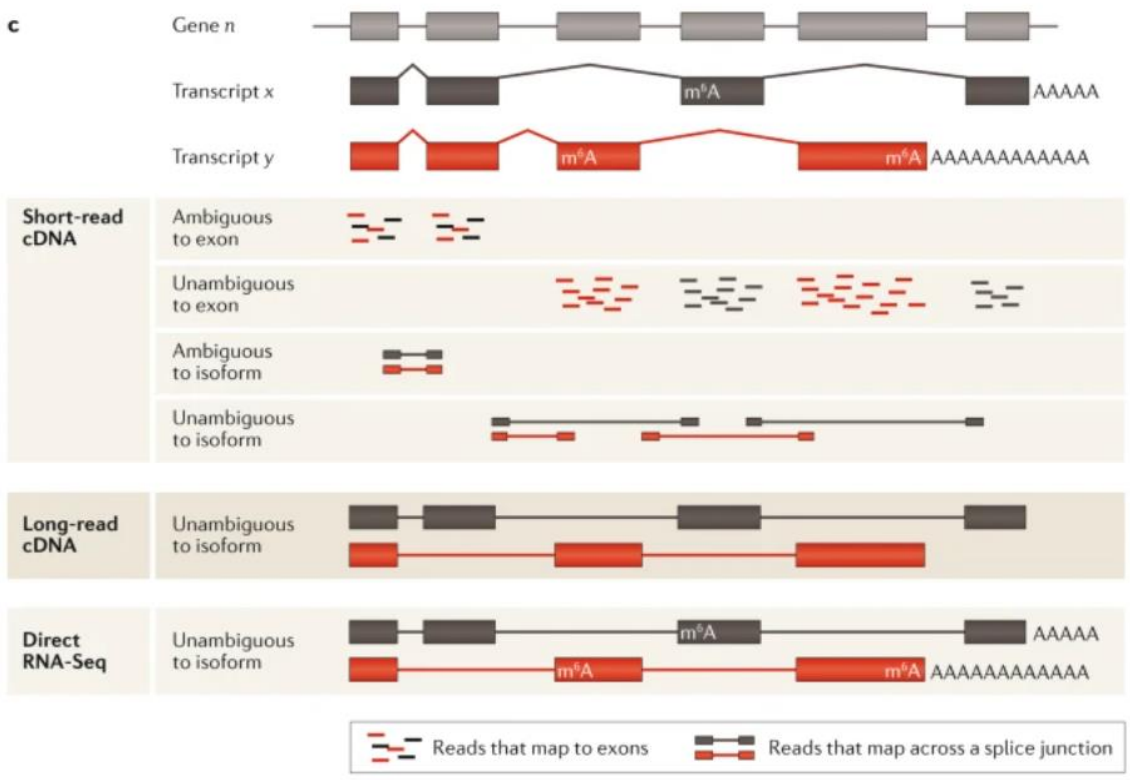
Es el tipo de secuenciación más utilizada y es la principal herramienta para realizar análisis de expresión diferenciales. Produce muchas más lecturas que la secuenciación de lecturas largas y permite trabajar con muestras degradadas. Elevada dificultad para identificar y cuantificar isoformas, muchos sesgos técnicos aunque conocidos.

- **Secuenciación cDNA de lecturas largas** ("PacBio")

Número de lecturas de secuenciación medio-bajo, lo que implica procesos más largos y costosos. Permite estudiar isoformas completas de manera no-ambigua, pero presenta sesgos durante el protocolo.

- **Secuenciación directa de lecturas largas** ("Nanopore")

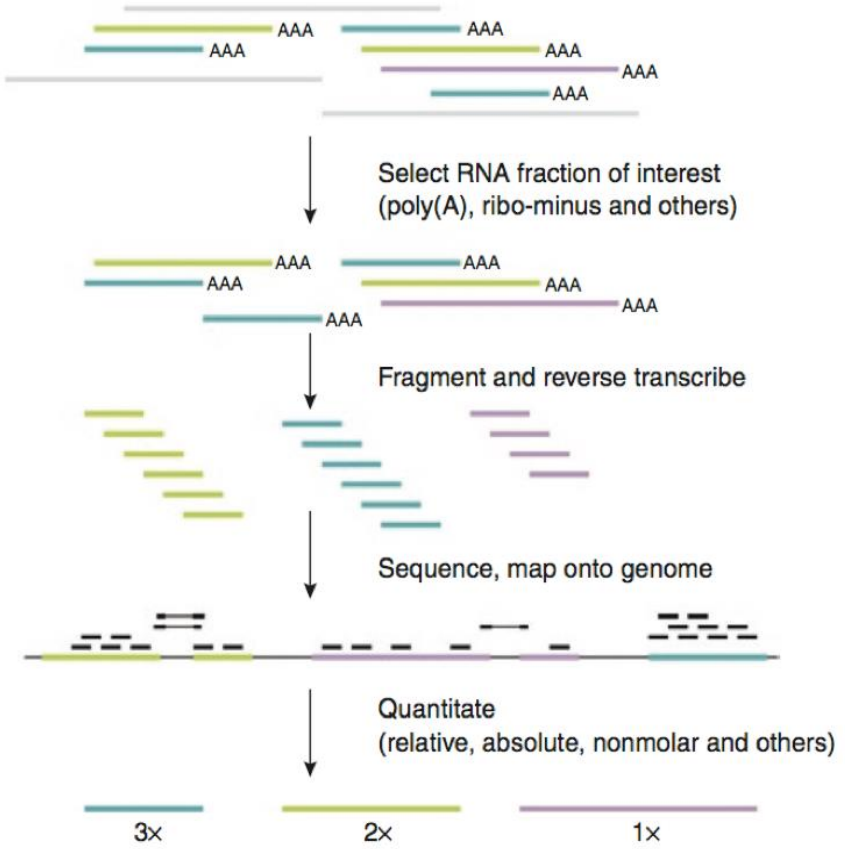
Número de lecturas de secuenciación bajo, no requiere de cDNA ni PCR lo que reduce los sesgos técnicos. Permite el estudio de modificaciones del ARN y estudio de colas de polyA.



https://doi.org/10.1038/s41576-019-0150-2

RNA-Seq de lecturas cortas de cDNA

- La secuenciación de transcriptomas basado en lecturas cortas de cDNA presenta dos procedimientos de selección de transcritos que implican diferentes sesgos técnicos que deben considerarse a la hora del análisis.



- El RNA-Seq de lecturas cortas se basa en la extracción de ARN de las células de interés, selección de los transcritos a secuenciar para evitar la secuenciación de transcritos mayoritarios sin interés (ribosómico, hemoglobina...), fragmentación aleatoria, cDNA, secuenciación y alineamiento con el genoma de referencia.

- Dentro del proceso de selección existen dos aproximaciones principales:

i. **Selección poly-A**: proporciona mayor cobertura en los exones y requiere de menor cobertura de secuenciación. Sin embargo, no detecta transcritos sin cola de poly-A, peores resultados en muestras degradadas y sesgo de cobertura hacia 3' de los transcritos.

ii. **Eliminación del ribosómico**: se necesita mayor cobertura de secuenciación, pero permite detectar cualquier ARN en las células y funciona bien en muestras degradas.

RNA-Seq de lecturas cortas de cDNA

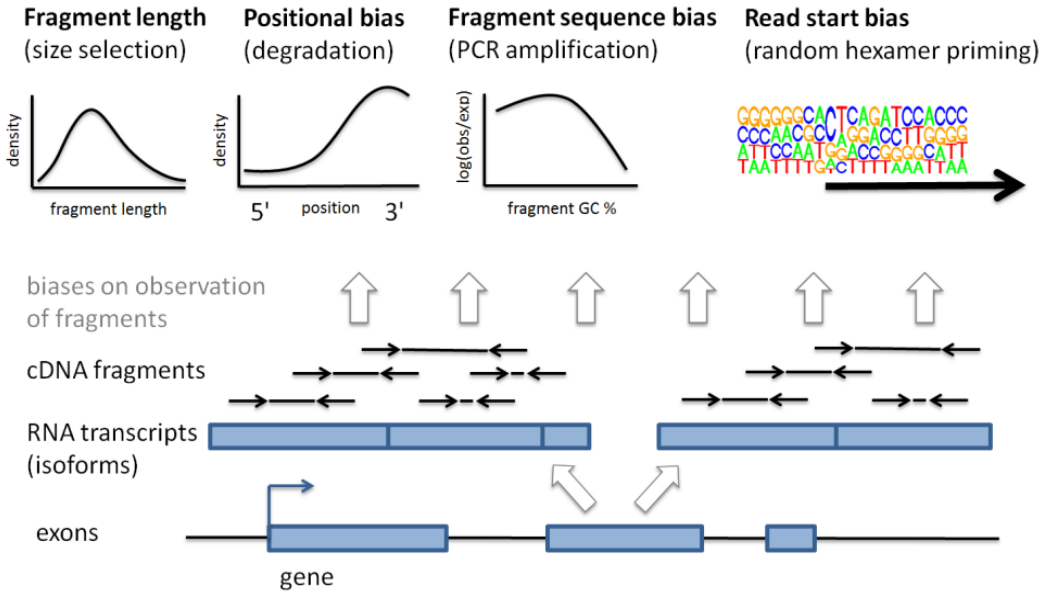
- Independientemente del método utilizado, este tipo de secuenciación presenta múltiples **sesgos técnicos** que deben ser tenidos en cuenta:

1. **Longitud de los fragmentos**: en estos protocolos de secuenciación se seleccionan los tamaños de los fragmentos de ARN a secuenciar. Esa selección no es perfecta por lo que fragmentos de diferentes tamaños son seleccionados y pueden generar sesgos en la cuantificación de la expresión.

2. **Sesgo posicional de los fragmentos**: la degradación del ARN y métodos de selección como el enriquecimiento en polyA generan preferencia por secuenciar fragmentos en el extremo 3' de los transcritos.

3. **Sesgo composicional de los fragmentos**: la amplificación por PCR tiende a amplificar más los fragmentos con contenido de GC promedio por encima de fragmentos con elevado o reducido contenido en GC.

4. **Sesgo en el inicio de la secuenciación**: a pesar de que los adaptadores deberían teóricamente unirse a fragmentos independientemente de su composición, existe cierta tendencia a unirse a fragmentos con una composición determinada.

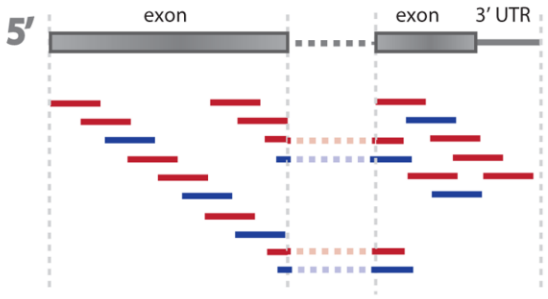


RNA-Seq de lecturas cortas de cDNA

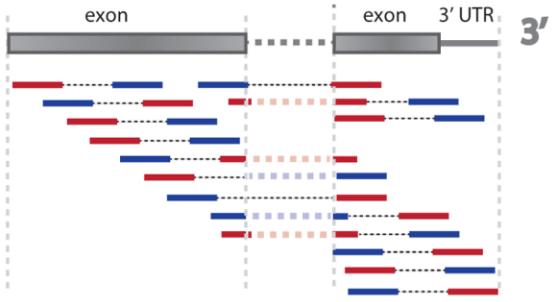
- Más allá de los sesgos técnicos del protocolo de secuenciación, existen otros sesgos que deben tenerse en cuenta para cuantificar la expresión de genes como la **longitud de los genes o la cobertura de secuenciación**.

- Existen tres **medidas para normalizar la expresión de los genes**:
  i. CPM (Count per Million): sólo normaliza por cobertura.
  ii. RPKM (Reads per Kilobase Million): normaliza por cobertura y longitud para single-end.
  iii. TPKM (Transcripts per Kilobase Million): normaliza por cobertura y longitud para paired-end.
  iv. TPM (Transcripts per Million): normaliza por longitud y cobertura. Una vez normalizado por longitud recalcula la cobertura, lo que implica que la cobertura total normalizada de todas las muestras es la misma.

Single-end sequencing



Paired-end sequencing



$$TPM = \frac{Ni/Li * 10^6}{sum(N1/L1 + N2/L2 + \ldots + Nn/Ln)}$$

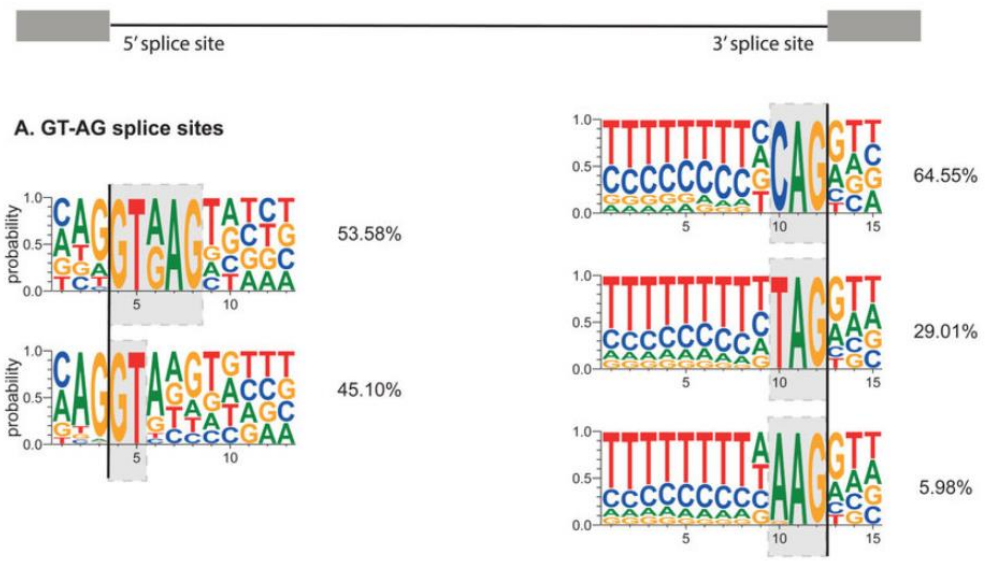
Ni is the number of reads compared to the i-th exon; Li is the length of the i-th exon; sum(N1/L1+N2/L2+.....+Nn/Ln) is the sum of the values of all (n) exons after normalization by length.

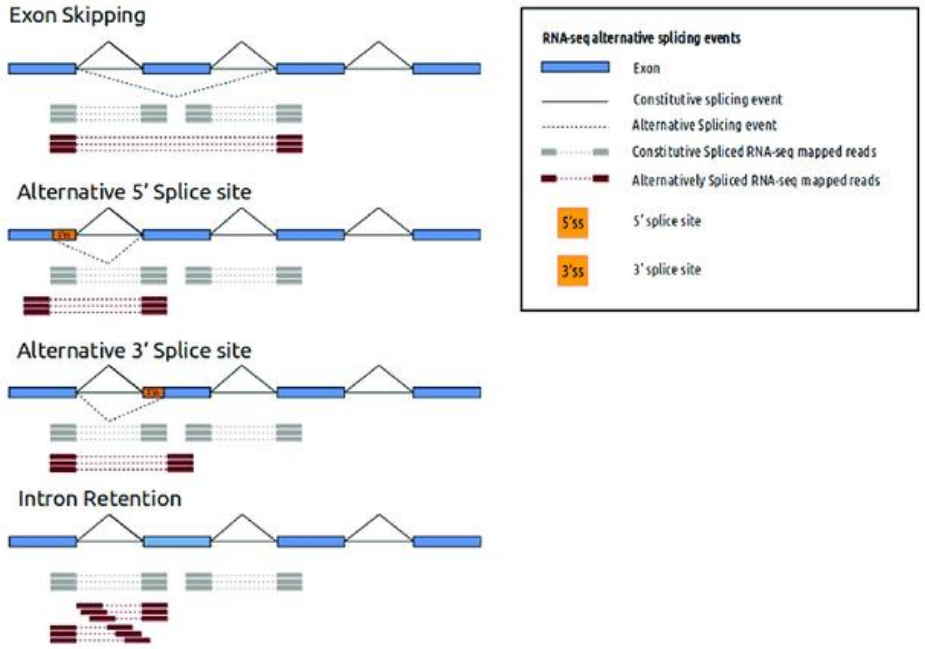
$$RPKM = \frac{ExonMappedReads * 10^9}{TotalMappedReads * ExonLength}$$

$$FPKM = \frac{ExonMappedFragments * 10^9}{TotalMappedFragments * ExonLength}$$

RNA-Seq de lecturas cortas de cDNA

- La cuantificación de isoformas basadas en lecturas cortas es compleja y suele basarse en el conocimiento previo de las isoformas. La identificación de nuevas isoformas se basa en la identificación de las secuencias donadoras (5' intron) y aceptoras (3' intron) del splicing alternativo. Esta identificación requiere de la existencia de múltiples lecturas que cubran la región donde se produce el splicing alternativo.

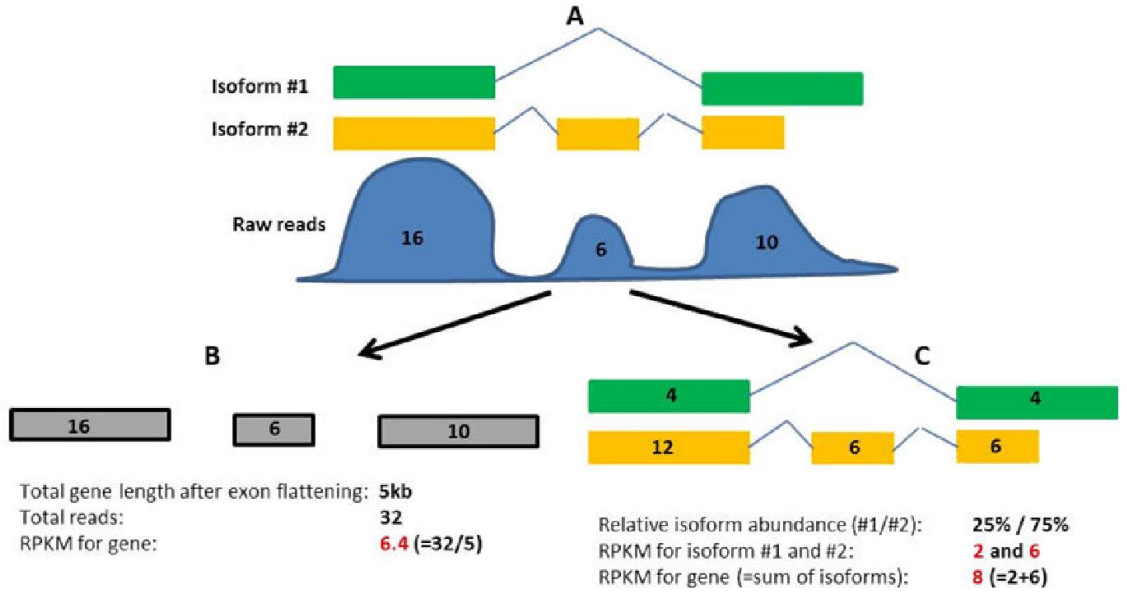




RNA-Seq de lecturas cortas de cDNA

- Actualmente, los métodos de estimación de la expresión de genes más utilizados se basan en la cuantificación de isoformas para estimar posteriormente la expresión del gen.

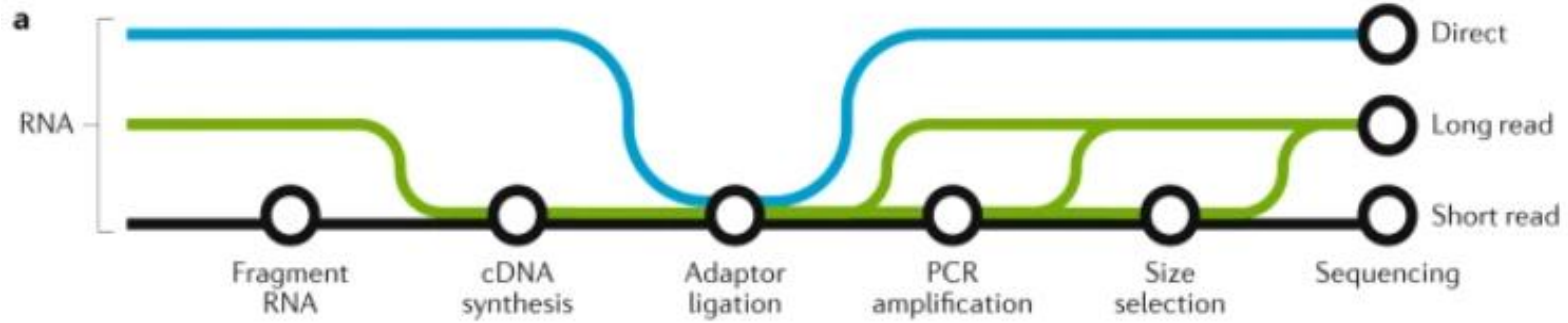
- En un gen con 3 exones de tamaños 2kb, 1kb y 2kb sin tener en cuenta las isoformas (A), si calculamos la expresión suponiendo que la cobertura total de secuenciación es de 1 millón de lecturas, obtendríamos un RPKM de 6.4 (B).



- Las isoformas se transcriben de manera independiente de la expresión total del gen, es decir que cada isoforma puede tener diferentes niveles de expresión. En este caso, observando el exón 2 de 1kb de la isoforma 2 con 6 lecturas y asumiendo que las lecturas se distribuyen de manera uniforme, podemos estimar que los exones 1 y 3 de dicha isoforma contribuyen con 12 y 6 lecturas respectivamente dado su tamaño. Por lo tanto, distribuyendo el resto de lecturas y calculando la expresión del gen en base a la suma de sus isoformas su expresión total sería de 8 RPKM (C).

RNA-Seq de lecturas largas

- La secuenciación de lecturas largas soluciona muchos de los problemas que encontramos en los experimentos de RNA-Seq con lecturas cortas. Sin embargo, son tecnologías relativamente recientes cuyos sesgos no se conocen tan bien como los de las lecturas cortas.
- Dada la excelente anotación del genoma humano, la identificación de **nuevas isoformas en muestras no cancerosas es relativamente rara**. Por lo que esta tecnología, principalmente se utiliza en muestras cancerosas o de especies no anotadas.
- La **tecnología de secuenciación cDNA** ("PacBio") sigue presentando muchos de los sesgos identificados en la secuenciación de lecturas cortas, principalmente el **sesgo composicional y la ligación de los adaptadores**. Además, la conversión a cDNA **limita la longitud** de los fragmentos a secuenciar e **impide identificar modificaciones del ARN**.
- La **fiabilidad de la secuenciación es relativamente baja** comparada con las secuencias cortas. Por lo que la identificación de variantes de secuencia es poco fiable, y complica el alineamiento con genomas de referencia.
- Además, es **difícil conseguir muestras de pacientes sin degradar**, aspecto fundamental para la secuenciación de lecturas largas.



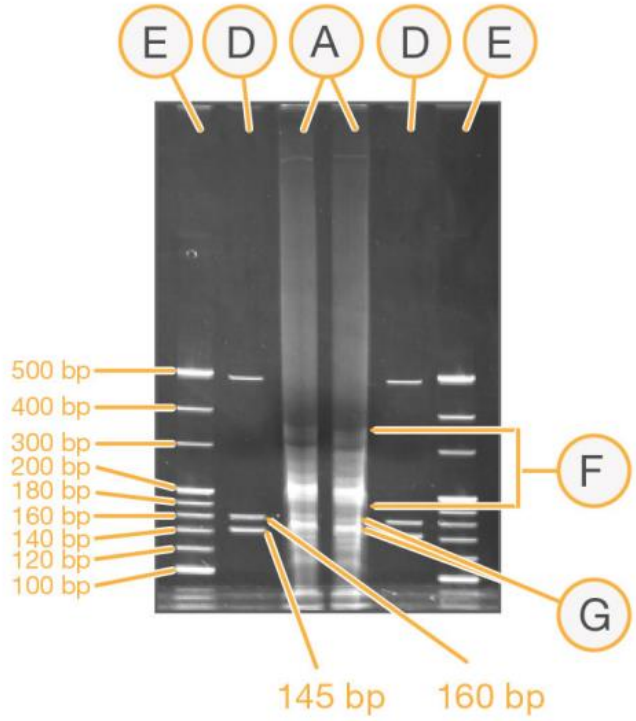
Small RNA-Seq

- La secuenciación de ARNs cortos no codificantes implica la **selección de fragmentos de ARN por debajo de los tamaños habituales de secuenciación**. El tamaño de estos fragmentos dificulta distinguir entre lecturas con ARN cortos y dímeros de adaptadores (fusión de adaptadores sin ARN).

- A pesar de que durante los últimos años se han desarrollado métodos para seleccionar estos fragmentos sin necesidad de geles, la migración en gel y el corte de la banda de interés sigue siendo parte de la mayoría de los protocolos.

- Algunos de los métodos propuestos incluyen la selección de tamaños mediante perlas magnéticas, pero la pequeña diferencia de tamaños entre los dímeros y las lecturas con ARN (~ 20pbs) hace que sean poco fiables.

- Por lo que la dependencia de geles y la intervención manual, convierte la secuenciación de **ARNs cortos en un proceso poco automatizable y que añade importantes sesgos en la selección de tamaños de las librerías**.



A: Cerebro humano
D y E: Ejemplo de tamalos (ladder)
F: Fracción de ARNs cortos no codificantes (tRNAs, rRNAs...)
G: Fracción de microARNs

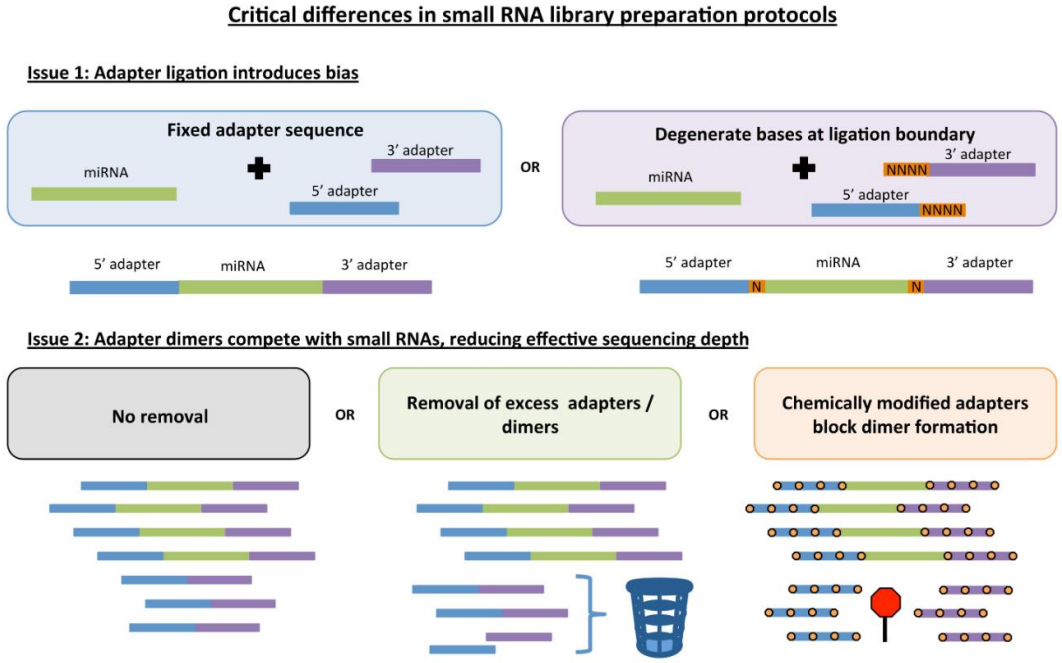
Small RNA-Seq

- Además de las dificultades a la hora de seleccionar los fragmentos a secuenciar, el tamaño de los ARNs de interés amplifica algunos de los sesgos observados en la secuenciación de RNA-Seq.

- La **presencia de dímeros en la librería** genera sesgos importantes en la profundidad de secuenciación. A misma cobertura, muestras con más dímeros tendrán menos lecturas efectivas. La presencia de dímeros a su vez se maximiza durante los ciclos de amplificación por PCR, lo que complica la secuenciación de ARNs cortos en muestras con poco material de partida.

- El **sesgo en la unión de adaptadores a los ARNs** es aún mayor que en los ARNs largos, ya que la unión diferencial de adaptadores en función de la composición del ARN corto afecta a la abundancia real del mismo.

- Para evitar estos sesgos se han desarrollado múltiples protocolos que incluyen la adición de secuencias degeneradas previa unión de los adaptadores y modificaciones químicas en los adaptadores que inhiben la formación de dímeros.

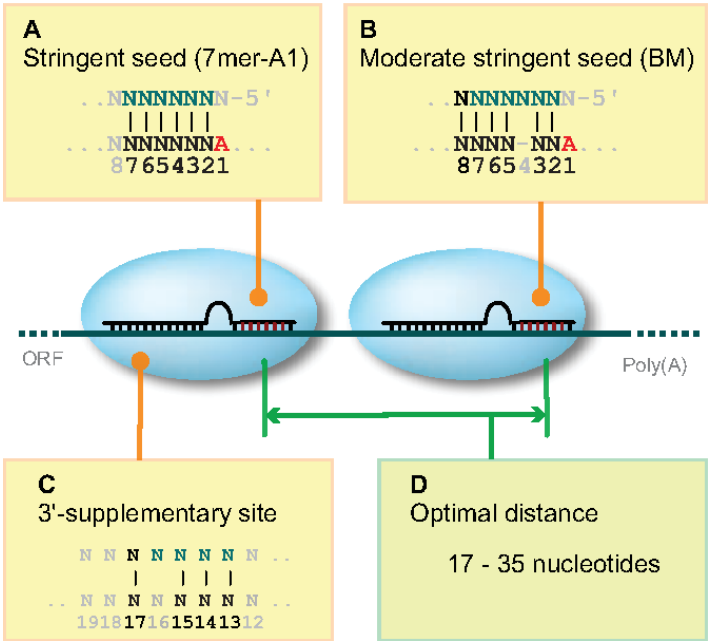


Small RNA-Seq

- Los resultados de un experimento de RNA-Seq pequeños requiere de análisis adicionales para interpretarlos. Todos suelen partir de análisis diferenciales, sin embargo, a diferencia del análisis de genes con una función más o menos conocida, los ARN pequeños no-codificantes requieren de interpretaciones biológicas no siempre conocidas, como piRNAs, snoRNAs, snRNAs y cuya validación experimental es compleja.

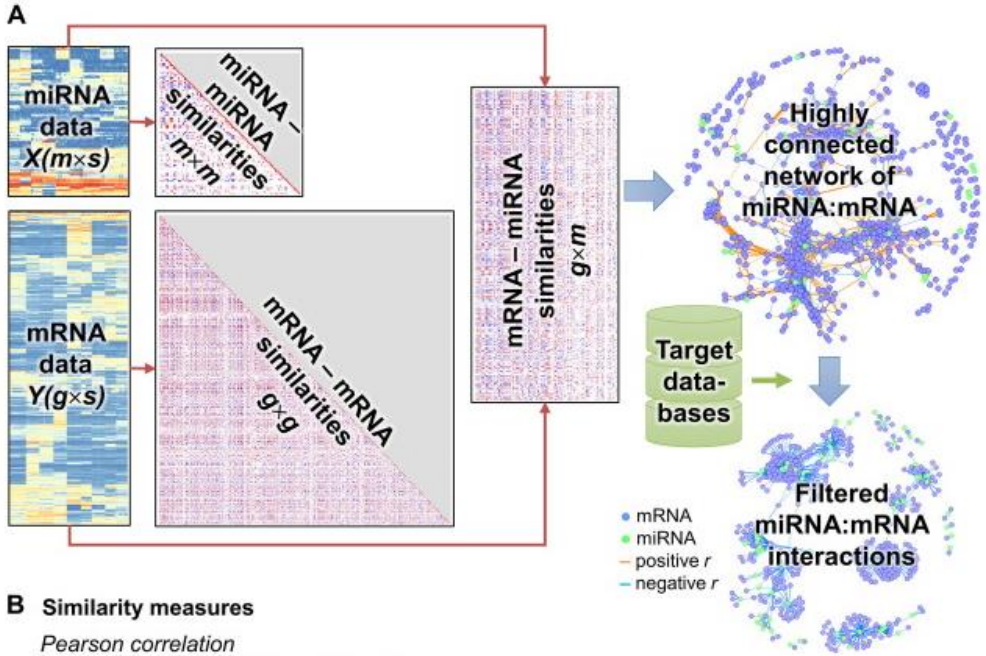
- En el caso de miRNAs, la interpretación suele realizarse mediante **predicciones de dianas** en la secuencia consenso de genes conocidos y/o su **integración con datos de mRNA** de las mismas muestras y el posterior estudio de redes de regulación.

- Algunas de las asunciones a la hora de identificar dianas de unión entre miRNAs y mRNAs, se basan en la complementariedad de las hebras, la distancia de las dianas al extremo 3' del mRNA, la estabilidad termodinámica del duplex miRNA-mRNA, la accesibilidad y/o la conservación del sitio de unión.



Small RNA-Seq

- La integración de experimentos RNA-Seq y RNA-Seq pequeños añaden fiabilidad a las predicciones de las dianas. De esta manera podemos **estudiar cambios coregulados entre mRNAs-miRNAs** y entre ellos que nos dan información adicional de los genes que pueden estar siendo regulados por los miRNAs de interés.
- Estos análisis se basan en la asunción de que generalmente un cambio de regulación biológicamente relevante no suele ocurrir por un solo miRNA, sino que diferentes miRNAs regulan el mismo o múltiples genes en una misma ruta funcional.
- Los análisis más sencillos de corregulación se basan en **medidas directas de similaridad** como el coeficiente de correlación de pearson o medidas no paramétricas como el índice de información mutua. Por supuesto, diferentes medidas pueden arrojar resultados diferentes lo que una vez más complica su interpretación.
- Este tipo de análisis también suelen abordarse utilizando **algoritmos de descomposición factorial**, que permiten agrupar mRNAs y miRNAs en bloques corregulados.



B Similarity measures

Pearson correlation

$$r(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

Spearman correlation

$$\rho(x,y) = r(\mathrm{rank}(x), \mathrm{rank}(y))$$

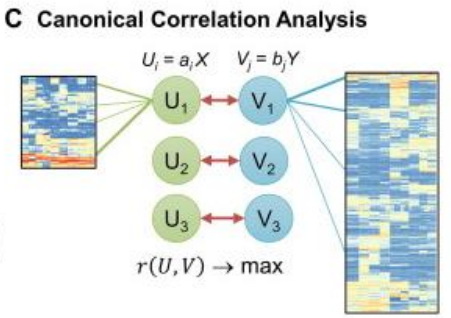
Cosine similarity

$$c(x,y) = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

Mutual information

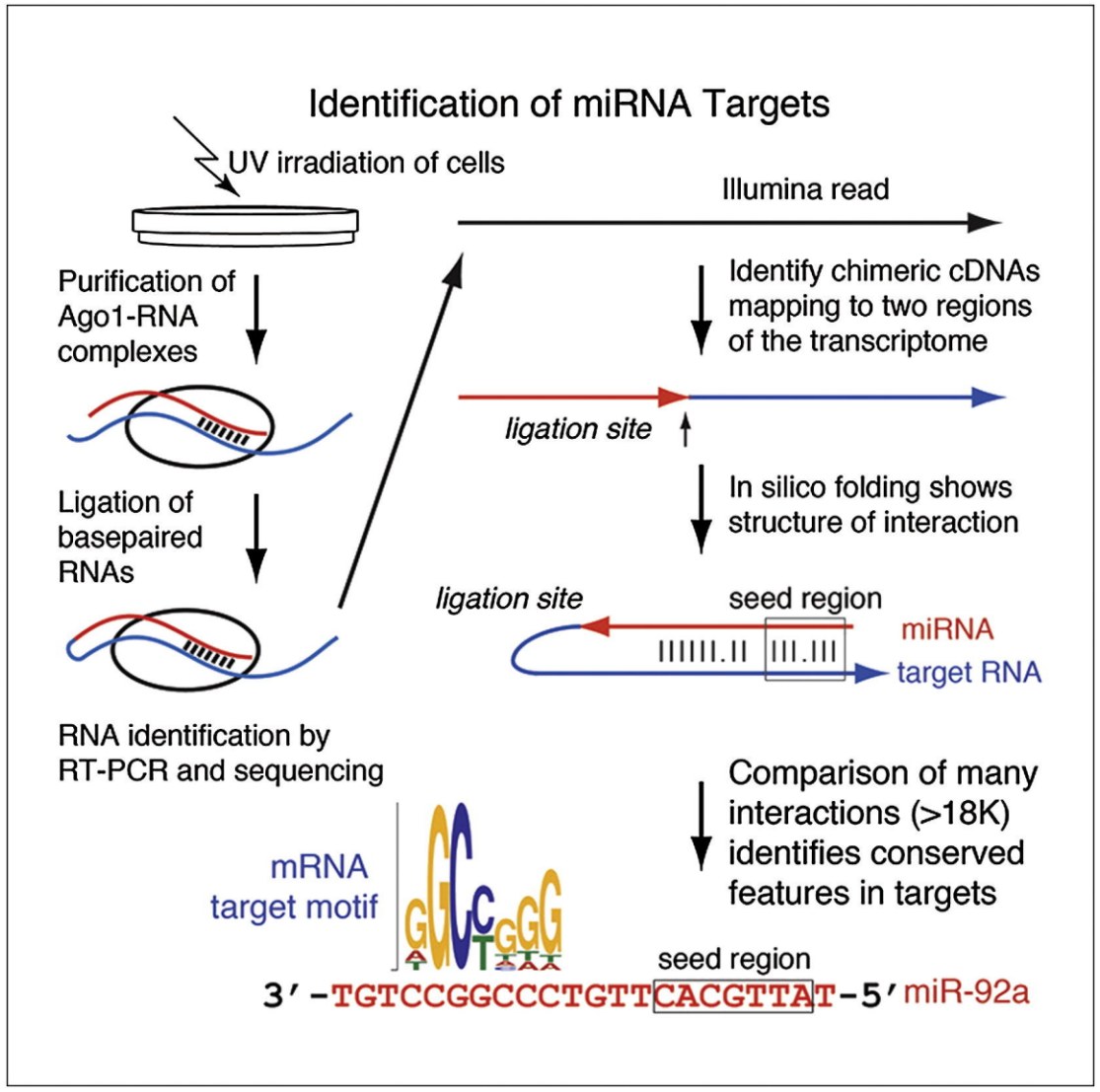
$$I(x,y) = \sum_{j\in y}\sum_{i\in x} p_{x,y}(i,j)\log\left(\frac{p_{x,y}(i,j)}{p_x(i)p_y(j)}\right)$$

$\bar{x}, \bar{y}$ – mean value of x and y
$\mathrm{rank}(..)$ – rank of the values
$p_{x,y}$ – joint probability distribution
p_x – marginal probability distribution



https://doi.org/10.1016/j.csbj.2021.01.029

CLASH-Seq (Validación dianas miRNAs)



- La funcionalidad de un miRNA en estudios de secuenciación masiva se basa en análisis indirectos o predicciones, lo que hace necesario validar experimentalmente la unión y el efecto de los miRNAs encontrados sobre los mRNAs dianas.

- La unión para miRNA y dianas individuales puede validarse mediante **ensayos de luciferasa**, pero no es un método óptimo para validaciones masivas. Mientras que la **secuenciación previo cross-linking** de la hebra doble de ARN y purificación de AGO1 (CLASH-Seq) requieren de una **gran cantidad de material de partida**, lo que los hace inviables en muestras in-vivo.

https://doi.org/10.1016/j.cell.2013.03.043

isomiRs

- La biogénesis de miRNAs puede dar lugar a isoformas de los mismos, dando lugar a subconjuntos de isomiRs de un mismo miRNA con dianas y por lo tanto funcionalidades diversas. Estas isoformas suelen encontrarse a **niveles de expresión inferiores a la secuencia canónica**.

- La escisión alternativa de drosha y/o dicer puede producir isomiRs con extremos 5' y/o 3' variables lo que puede afectar a lo localización de la semilla. Esto conlleva la adición o depleción de nucleótidos contenidos en el pri-miRNA.

- A su vez podemos encontrar isoformas con variantes de secuencias no contenidas en los pri-miRNAs, como pueden ser adiciones de nucleótidos (adenilación o uridilación) en el extremo 3' del miRNA. O modificaciones de adeninas a inosinas que pueden interferir en el procesado del pri-miRNA o directamente modificar la semilla, ya que la inosina es reconocida como guanina.