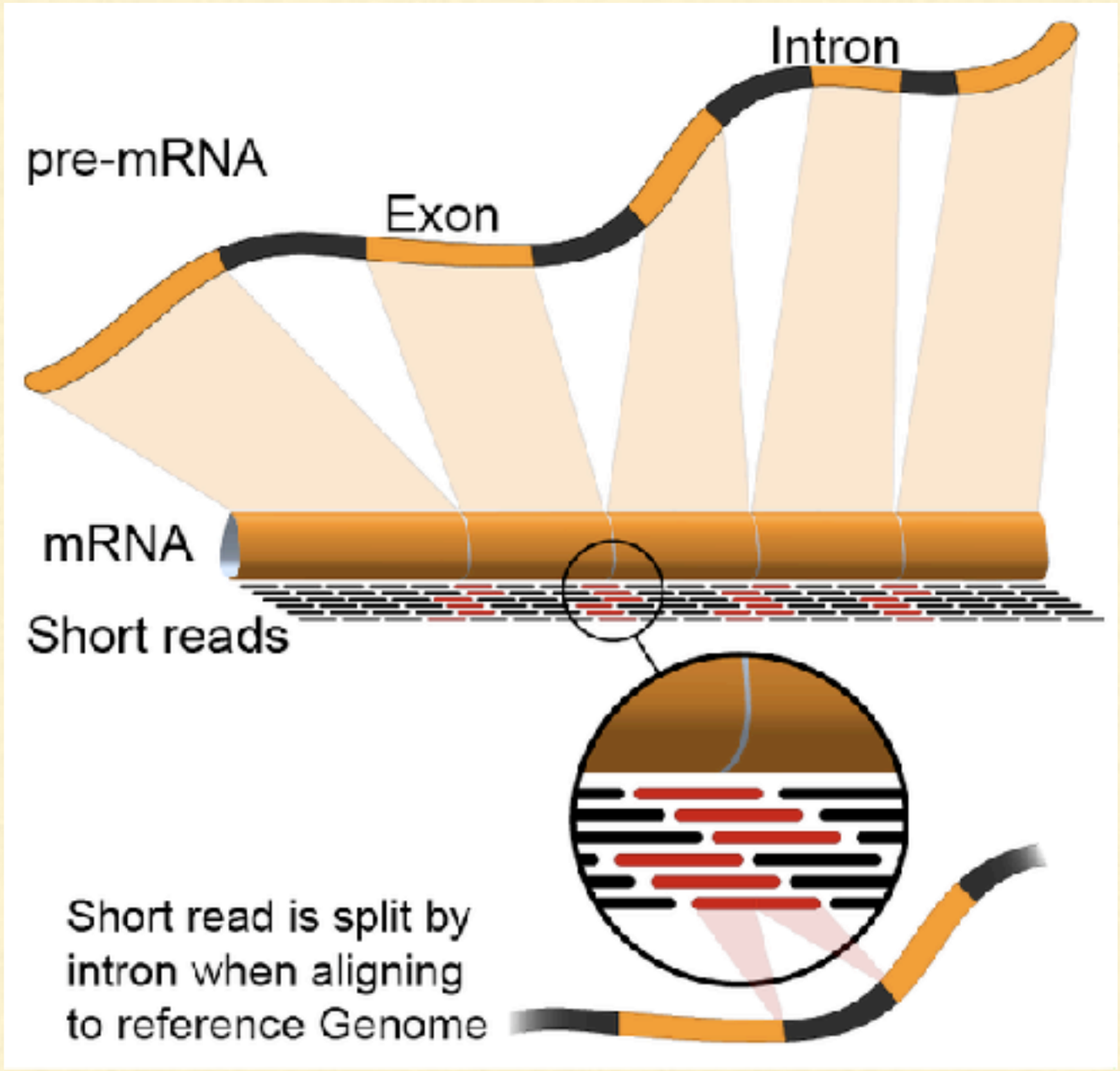
RNA-SEQ

Guillermo G Torres

Pasto 2017

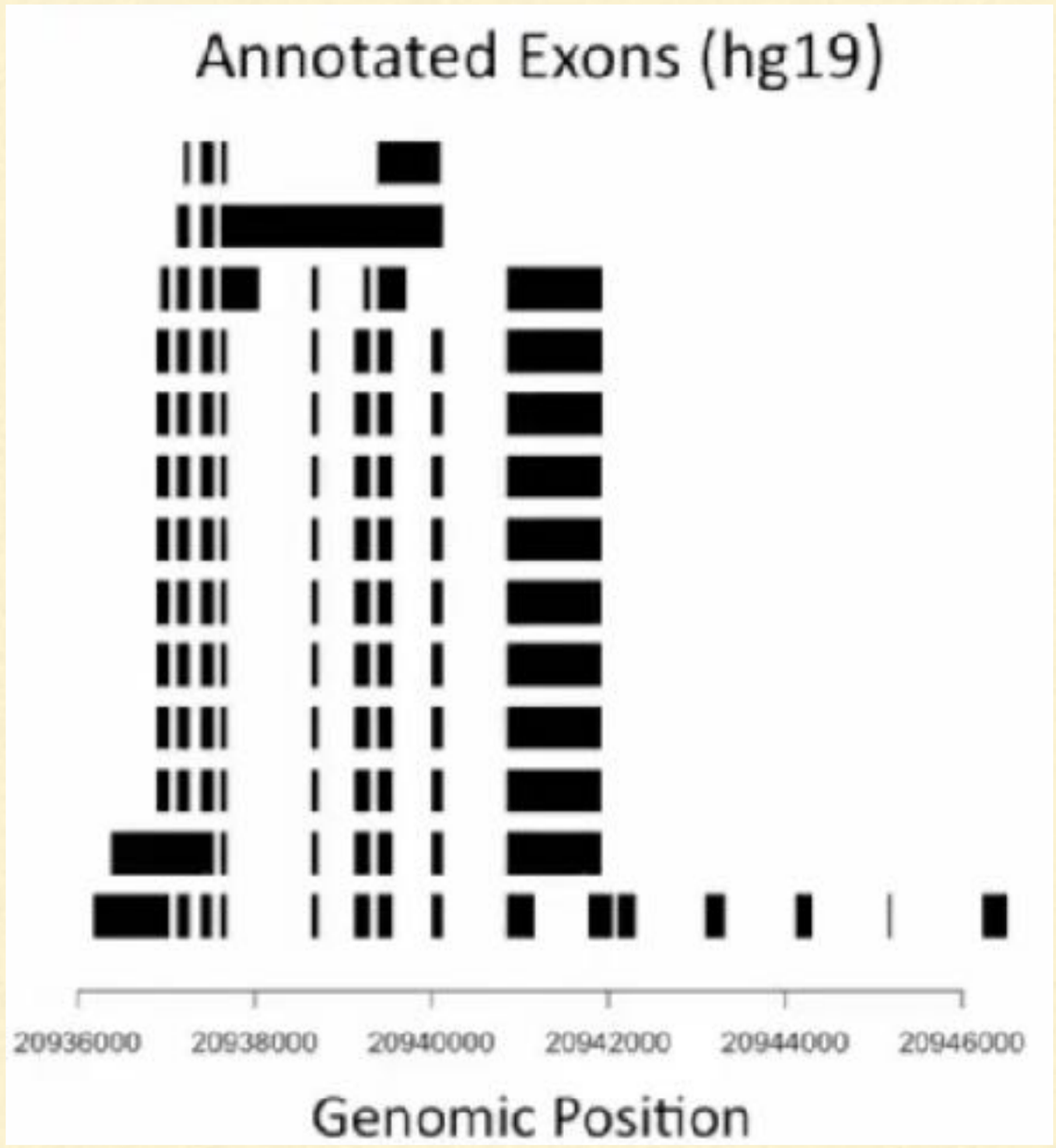
CUAL ES LA IDEA

COMPLICACIONES



No hay reads que mapeen intrones

COMPLICACIONES



Los genes tienen un comienzo y final pero existen diferentes versiones: **variantes de splacing**

FLUJO DE TRABAJO

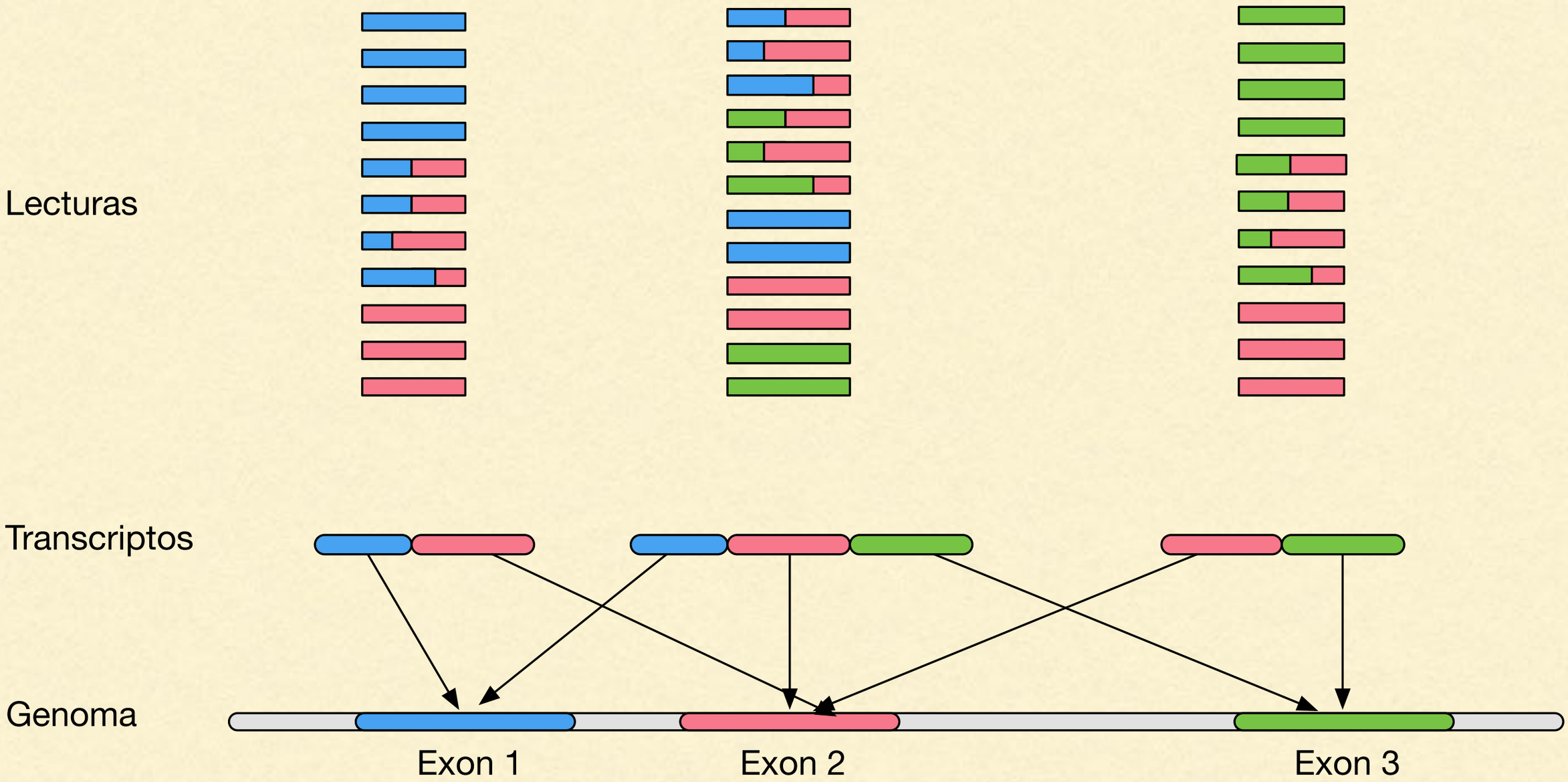
1. Ensamblaje de los transcriptos (Trinity, Soap, Cufflinks, Scripture)

    A. De novo

    B. Referencia o mapeo

2. Conteo de genes o exones (edgeR, DESeq, DEXSeq)

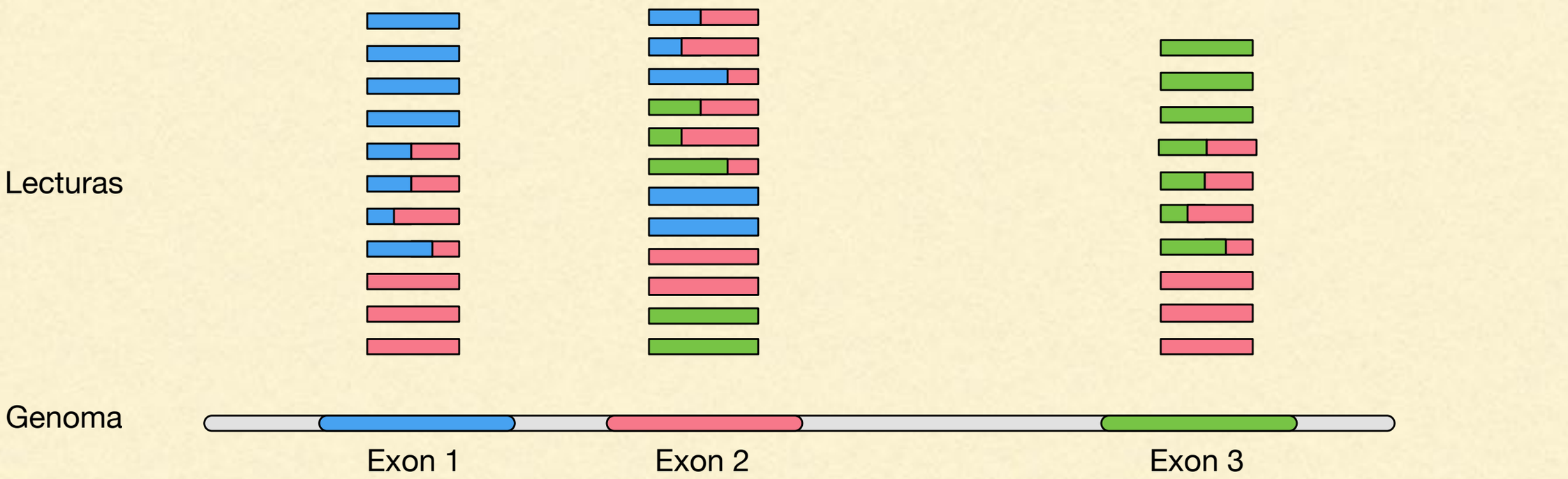
GENERACIÓN DE DATOS



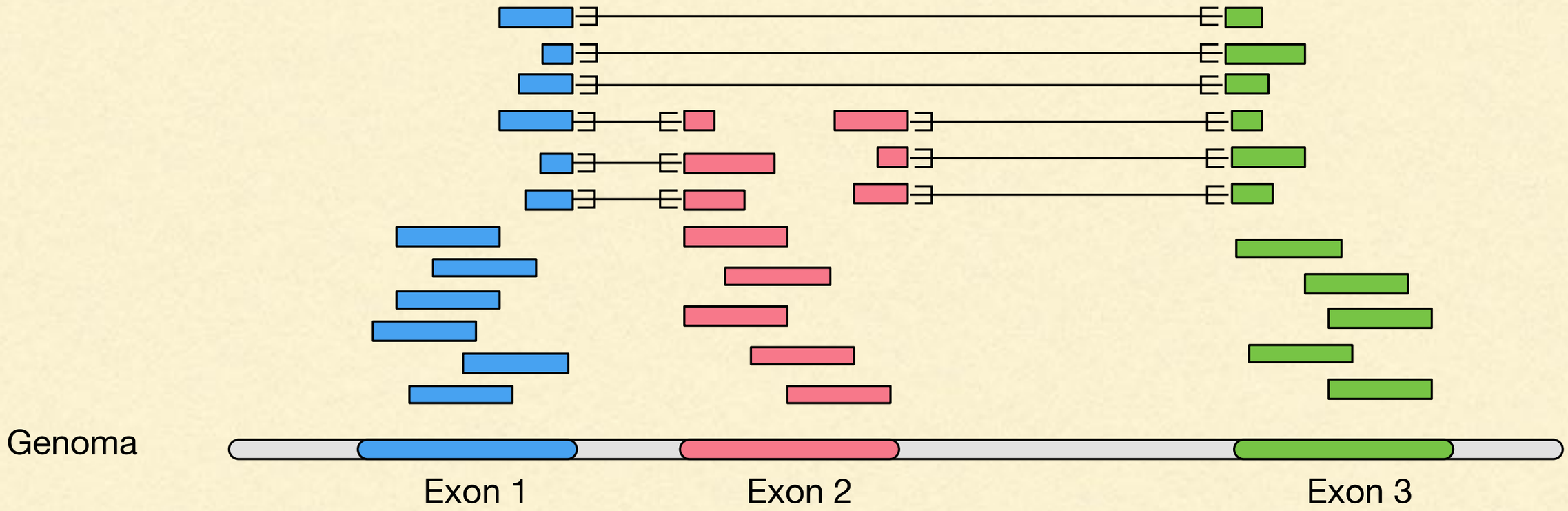
RETOS

Con esta información: lecturas e información genómica (no siempre)

- Reconstruir los transcriptos
  - Con la ayuda del genoma (mapeo o ensamblaje por referencia)
  - Sin la ayuda de un genoma (de novo)
- Cuantificar cuan abundantes o presentes estan estos transcriptos.



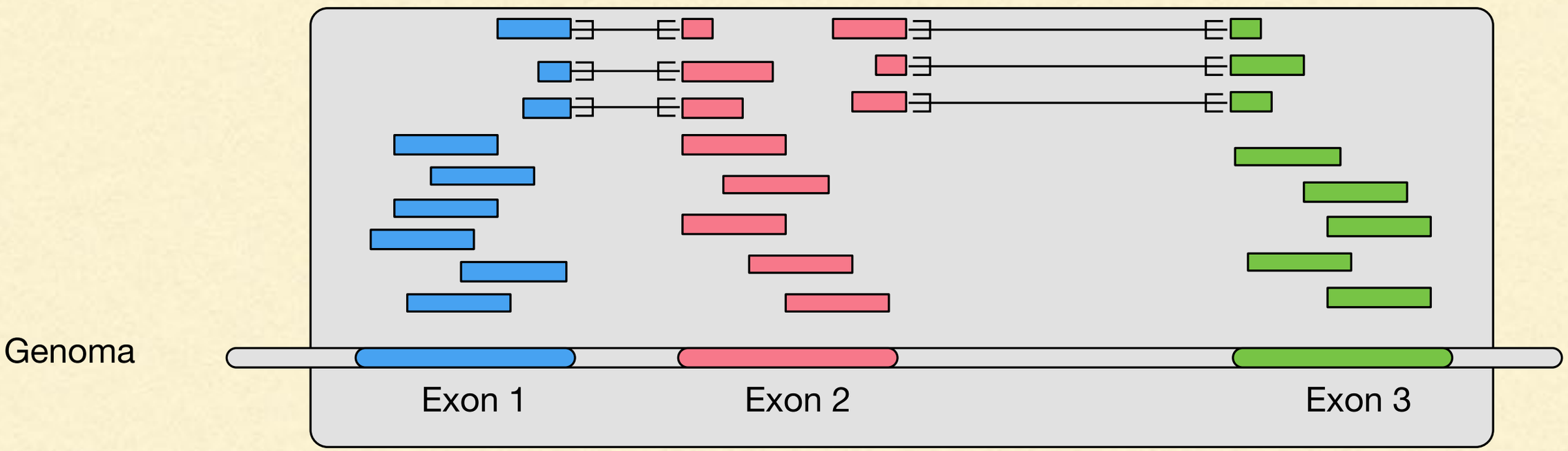
MAPEO - PRIMER ALINEAMIENTO



- TopHat, GSNAP, MApSplice, STAR, PALMapper y otros
  - Para comparación de estos métodos ver: Steijger et al. 2013 Assessment of transcript reconstruction methods for RNA-seq, Nature methods

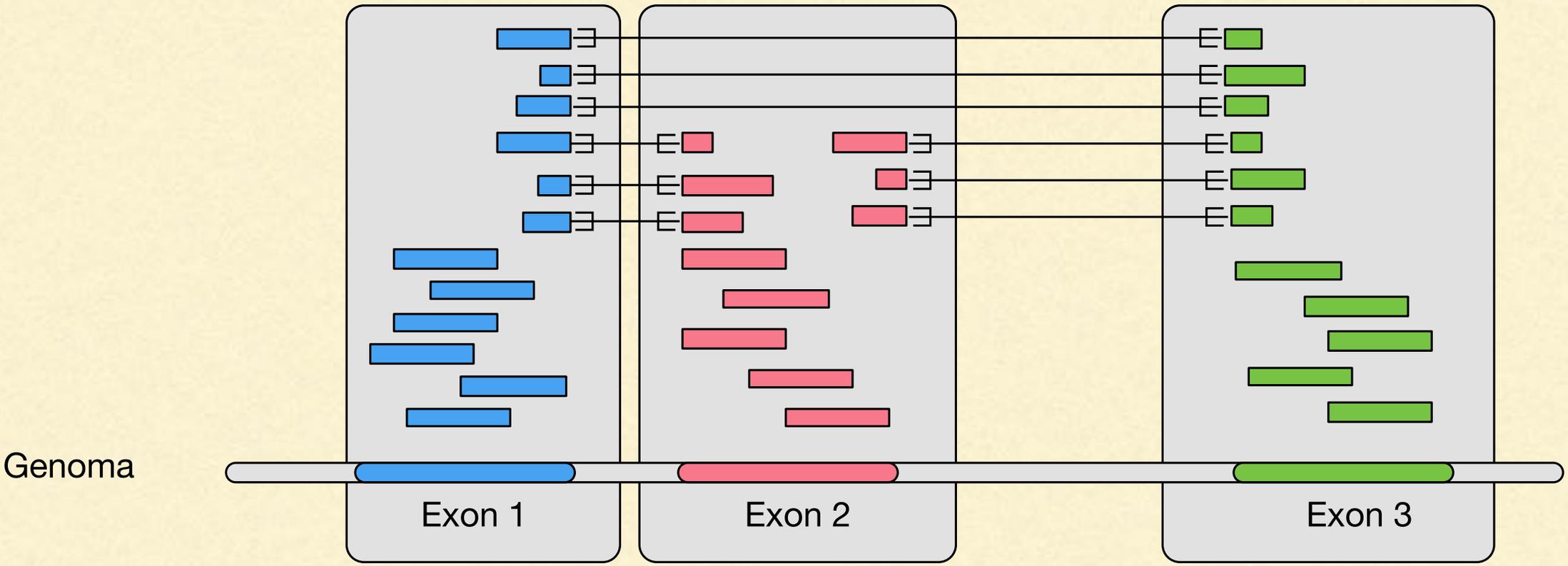
TABLAS DE CONTEOS

- Usando los alineamientos para contar genes. Union de todos los expones



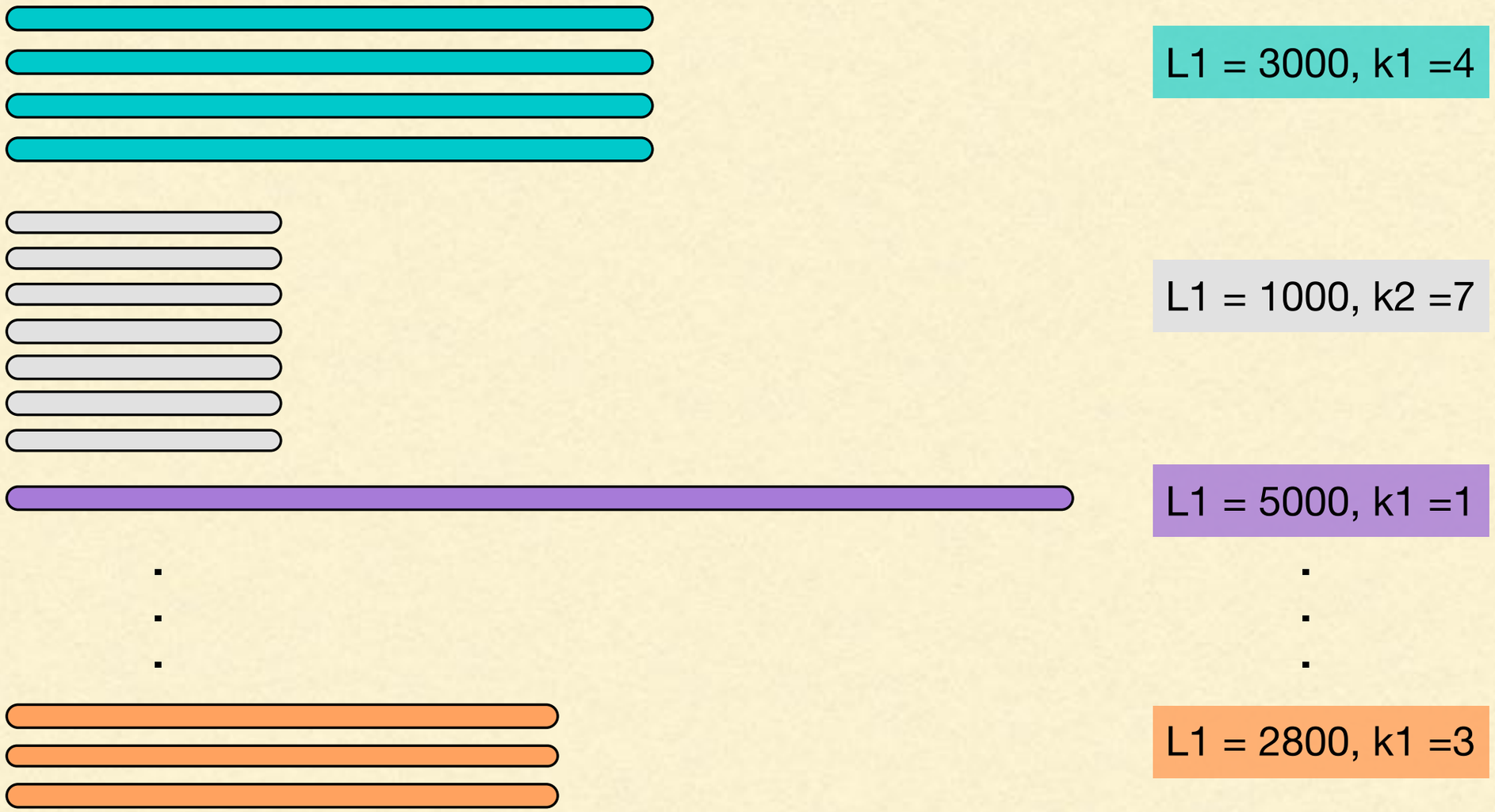
TABLAS DE CONTEOS

- Usando los alineamientos o para contar exones.



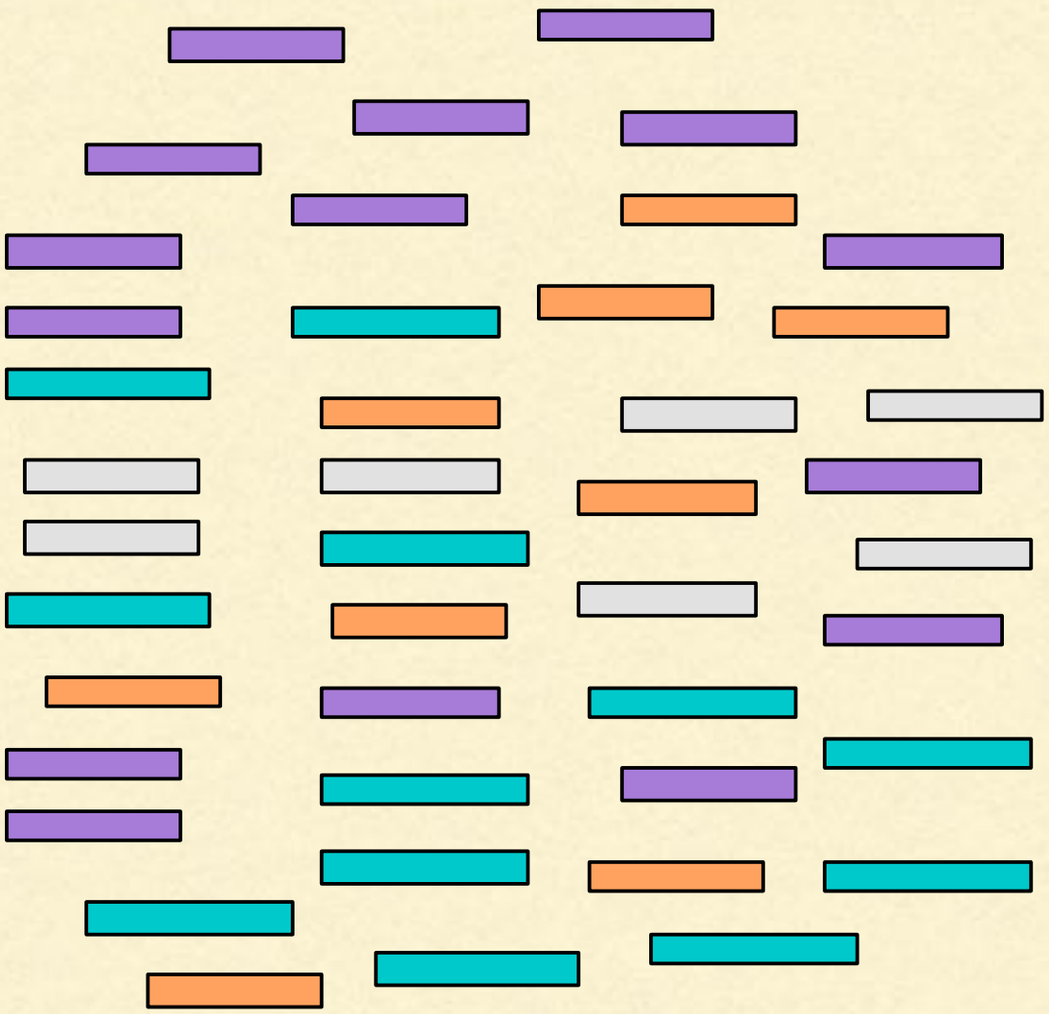
MODELO DE CUANTIFICACIÓN

- Todos los transcriptos en la muestra



MODELO DE CUANTIFICACIÓN

- Todos los transcriptos en la muestra - en términos de las lecturas



L1 = 3000, k1 =4

L1 = 1000, k2 =7

L1 = 5000, k1 =1

L1 = 2800, k1 =3

MODELO DE CUANTIFICACIÓN

- Pf = proporción de la cantidad total de secuencias. (profundidad del transcripto)

Transcript expression
(relative to total expression)

Total amount of sequence for transcript f

$$p_f = \theta_f l_f, \ \theta_f = \frac{k_f}{\sum_{i=1}^{M} k_i l_i}$$

Total amount of sequence

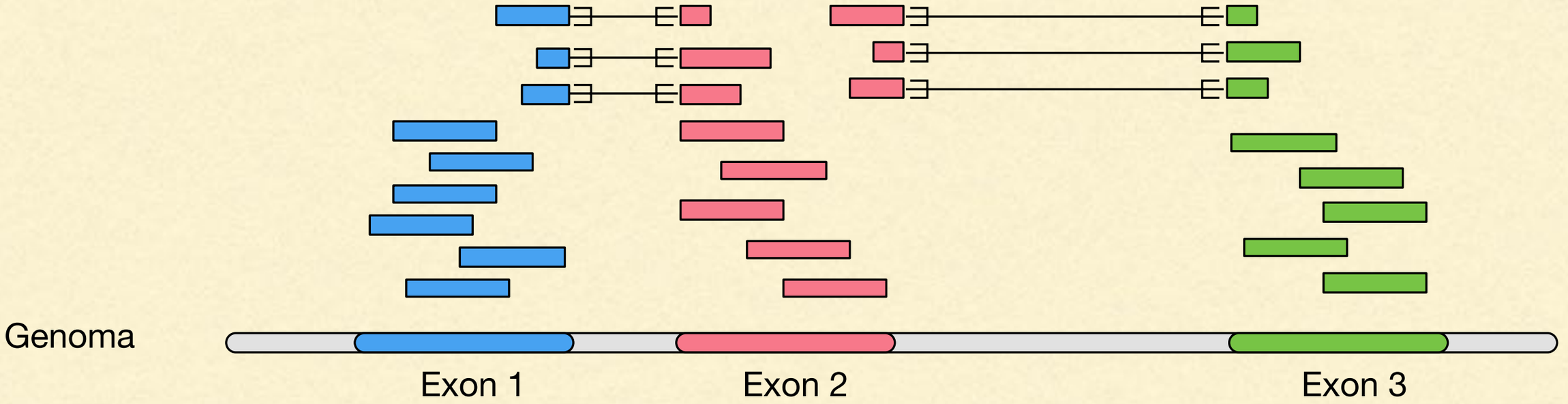
MODELO DE CUANTIFICACIÓN - DISTRIBUCIÓN

- Y_f = Numero de lecturas que provienen de la expresión del transcripto

$$Y_f \sim \text{Poisson}(\theta_f l_f)$$

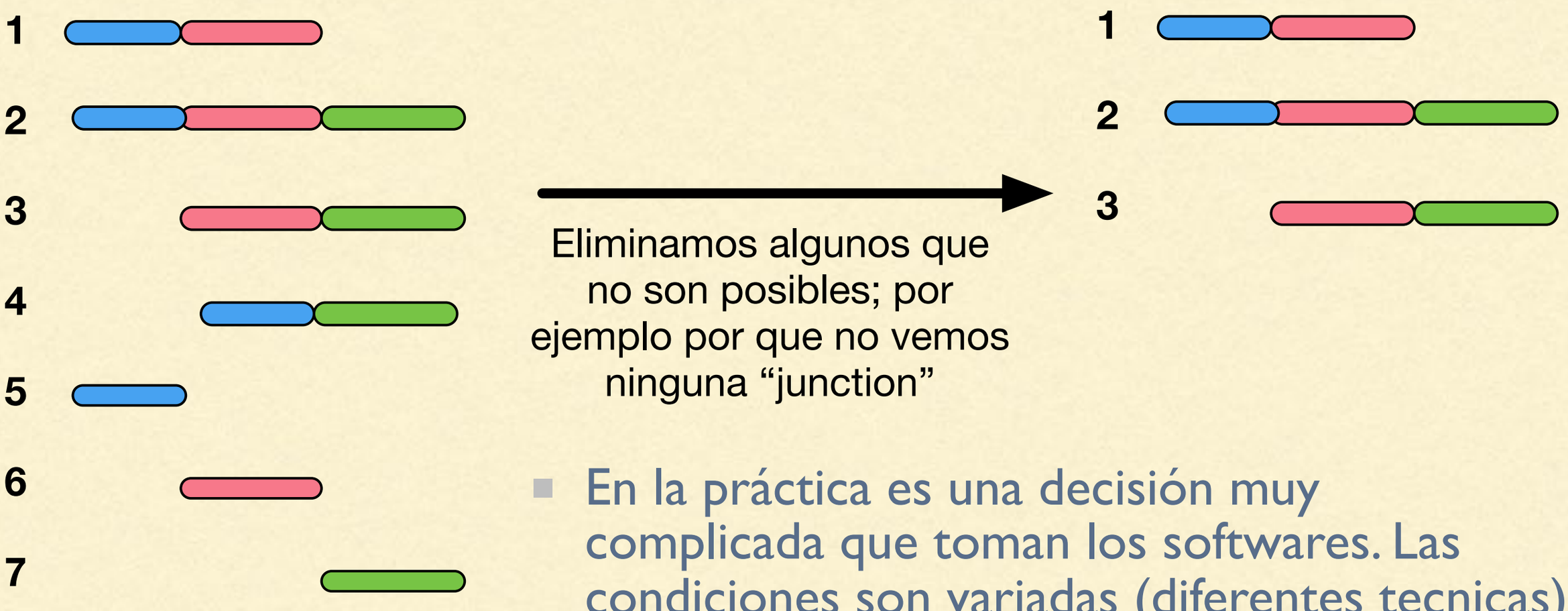
- El número observado de lecturas Y para el transcripto f es la suma, sobre todas las lecturas, de as salidas binarias (1 si la lectura proviene del transcripto f , 0 si no).
- Puesto que N (número de lecturas totales) es largo y *p*f (proporción del transcripto f) es pequeño, el modelo de Poisson se ajusta muy bien. (Marioni et.al 2008, Genome Research; Jiang and Wong 2009, Bioinformatics)

CUANTIFICANDO LOS TRANSCRIPTOS



CUANTIFICANDO LOS TRANSCRIPTOS

- Todos los posibles transcriptos:



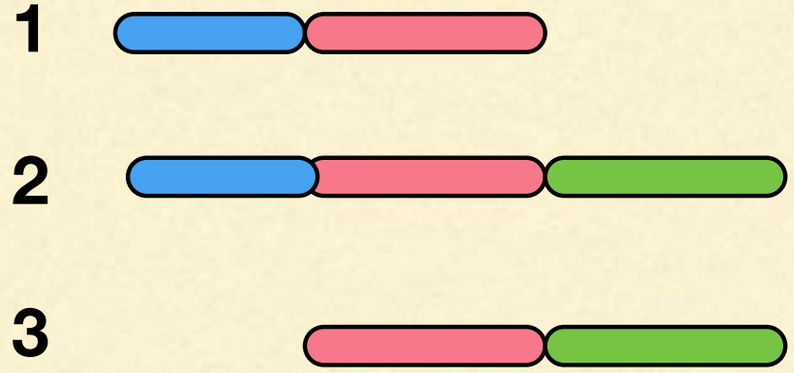
- En la práctica es una decisión muy complicada que toman los softwares. Las condiciones son variadas (diferentes tecnicas)

CUANTIFICANDO LOS TRANSCRIPTOS

- Modelo estadístico (Modelos lineales). θ_i = expresión del exón

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_{1,2} \\ Y_{2,3} \end{pmatrix} = \begin{pmatrix} \text{Exon 1 count} \\ \text{Exon 2 count} \\ \text{Exon 3 count} \\ \text{Junction 1,2 count} \\ \text{Junction 2,3 count} \end{pmatrix}$$

$$Y_1 \sim \text{Poisson}(wl_1\theta_1 + wl_1\theta_2)$$

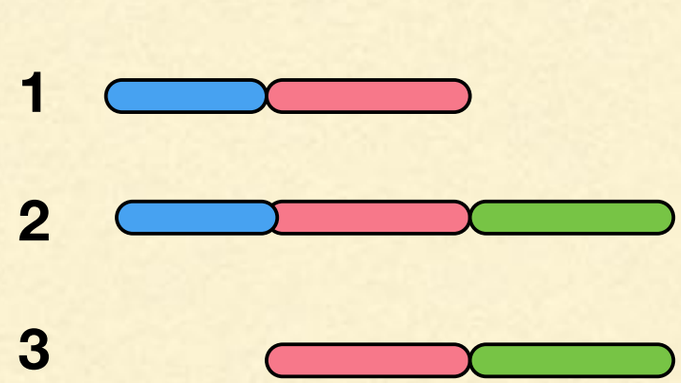


Jiang and Wong 2009, Bioinformatics

CUANTIFICANDO LOS TRANSCRIPTOS

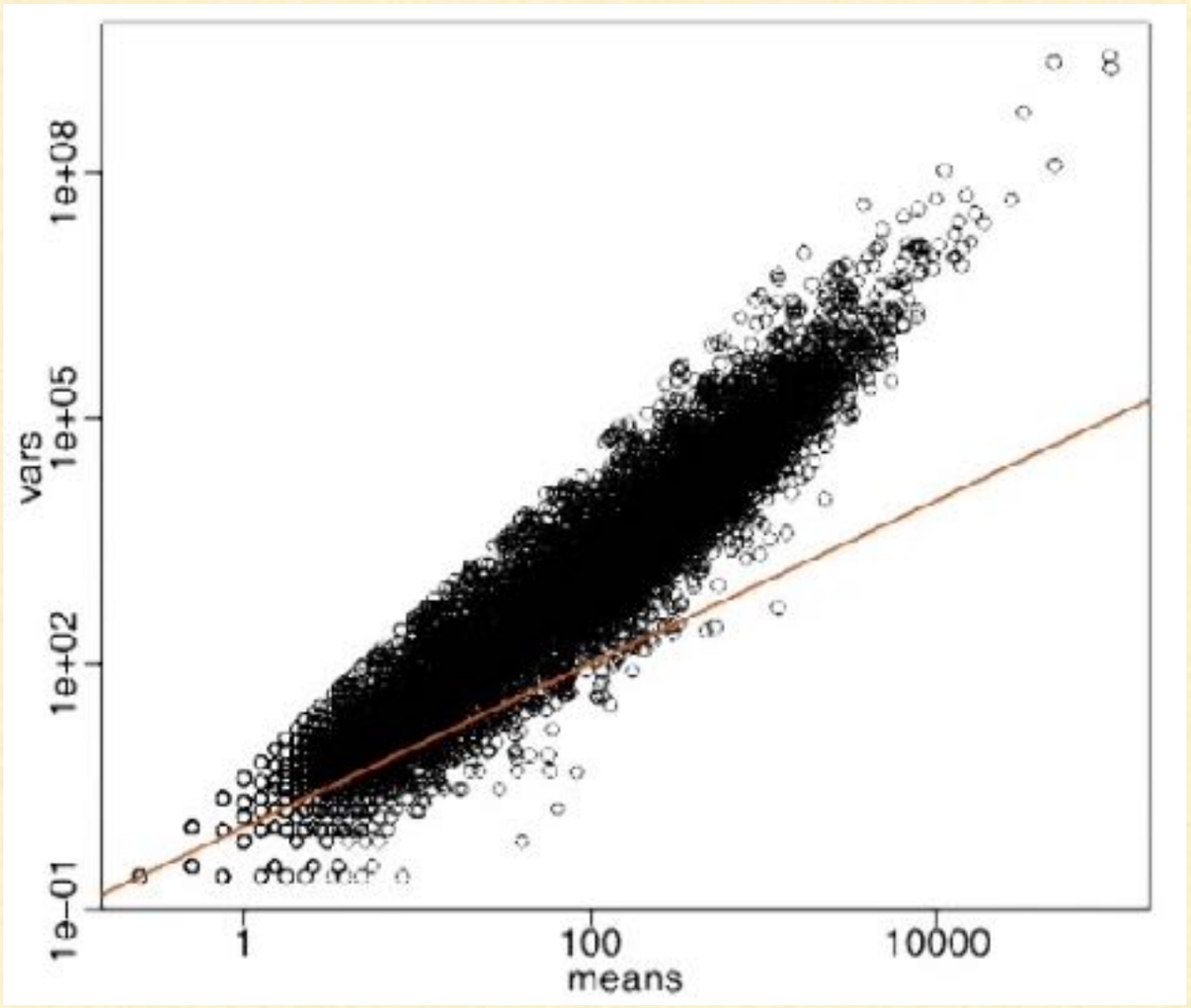
- Y's son Poisson independientes y la cuantificación de los transcriptos la hacemos calculando Maximum Likelihood Estimation (MLE) de los θ.

$$\mathrm{E}\begin{pmatrix} Y_1/l_1 \\ Y_2/l_2 \\ Y_3/l_3 \\ Y_{1,2}/l_{1,2} \\ Y_{2,3}/l_{2,3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}$$



Jiang and Wong 2009, Bioinformatics

BINOMIAL NEGATIVA



- Los datos reales no se ajustan bien a una Pisson
- Uso de Binomial negativa - Es una Poisson la cual acepta una variabilidad mayor a la estimada
- La mayor variabilidad de los datos es debida en gran medida a la variabilidad biológica