Biotecnología de Plantas
Posgrado

Cinvestav

Transcriptómica

Clase 6 - Análisis de expresión diferencial

Bionformática y Bioestadística 2023

Selene L. Fernández-Valverde
regRNAlab.github.io
@SelFdz

Objetivos de aprendizaje

En esta clase aprenderemos:

- A contar el número de alineamientos de lecturas
- Entender porque modelamos expresión usando una distribución binomial negativa
- A hacer análisis de expresión diferencial

¿Qué es un análisis de expresión diferencial?

- Es una prueba estadística que identifica genes que cambian entre dos o más condiciones con un cierto nivel de confianza

RNA-Seq mide abundancia **relativa**

- Sin datos adicionales solo provee información acerca de abundancias relativas

- Información adicional, por ejemplo, la adición de transcritos control o “spike-ins” a una muestra son necesarios para hacer mediciones absolutas entre genes que cambian entre dos o más condiciones con cierto nivel de confianza.

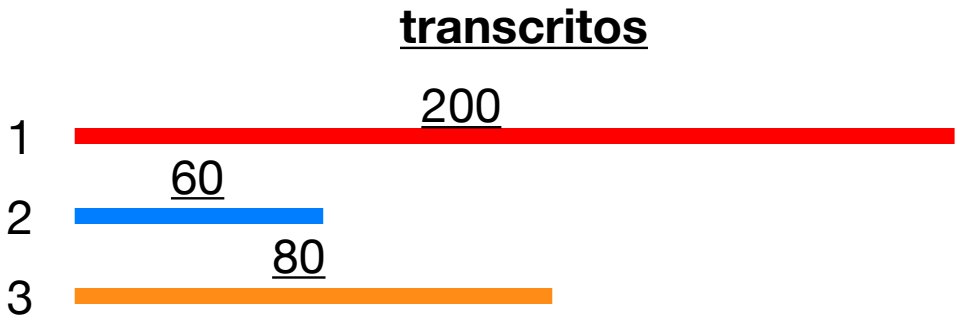
Problemas con la medidas de abundancia relativa

Gen	Muestra 1 abundancia absoluta	Muestra 1 abundancia relativa	Muestra 2 abundancia absoluta	Muestra 2 abundancia relativa
1	20	10%	20	5%
2	20	10%	20	5%
3	20	10%	20	5%
4	20	10%	20	5%
5	20	10%	20	5%
6	100	50%	300	75%

- Cambios en la expresión absoluta de genes altamente expresados causa problemas.
- Se requiere normalizar para realizar comparaciones válidas.

Principios básicos de cuantificación en RNA-Seq

- Para simplificar, asume que cada lectura tiene longitud 1.



copias	lecturas	depth	RPM
10	100	20	5
10	60	20	3
5	40	20	2

- ¿Qué abundancias relativas estimarías para estos genes?

¿Qué hace un software que detecta expresión diferencial?

- Dos tareas principales:

    1. Estimar la magnitud de expresión diferencial entre dos o más condiciones basados en el número de lecturas de muestras con réplicas.
    2. Estimar la significancia de esta diferencia y corregir por pruebas múltiples.

- Para determinar si las diferencias en número de lecturas entre dos condiciones es mayor a lo esperado por azar, las herramienta de ED deben asumir ciertas propiedades acerca de la distribución de las lecturas.

- La hipótesis nula - que la media de las lecturas de las muestra en condición A es igual a la media de lecturas en las muestra de la condición B - es probada para cada gen de manera individual.

RPKM - Reads per Kilobase per Million mapped reads

- Esta medida intenta normalizar usando la profundidad de la secuenciación y la longitud del gen.

- Para calcularla:
  1. Cuenta el número de lecturas en una muestra y divídelo por 1,000,000 - este será nuestro factor para escalar “por millón”.
  2. Divide el número de lecturas por este factor, normalizando así por profundidad de secuenciación y obteniendo lecturas por millón (Reads Per Million - RPM).
  3. Divide estos valores (RPM) por la longitud del gen en kilobases. Esto da RPKM.

- ¿Por qué querríamos no usar esta medida?

Es mejor no utilizar RPKMs

- Filtrar por RPKMs no es apropiado precisamente porque toma en cuenta el tamaño del gen.
- Considera un gen muy largo que se expresa a un nivel moderado. Dada su longitud, los RPKMs de este gen serán bajos y podría ser descartado por su baja expresión. Sin embargo, el número absoluto de cuentas para este gen (probablemente) sea alto. Esto significa que tendrá más que suficiente información para estimar cambios en su expresión (dispersión) y para análisis de expresión diferencial.

¿Podemos comparar las abundancias absolutas?

- Si, aunque, dado que las lecturas son absolutas no pueden distribuirse de manera normal (no puede haber -3 lecturas o 15.5 lecturas) se requieren distribuciones especiales: la **poisson** (que asume que la varianza y la media son iguales) y la **binomial negativa** (que no asume esto).

- Esto se vuelve un problema cuando tenemos pocas replicas biológica dado que es difícil estimar adecuadamente la varianza de datos absolutos si estamos cuantificando un solo gen y asumiendo que tenemos datos continuos y distribuidos normalmente. Un buen estimado de la varianza del gen es esencial para descartar cambios que están ocurriendo por azar.

¿Cómo lo solucionamos?

- La mayoría de las herramientas para RNA-Seq permiten que la varianza global (para un gen en general) pueda ser estimada con cierta abundancia de lecturas. Esto genera estimados más precisos de la varianza que observando cada gen en particular dado que puedes determinar algunos supuestos generales acerca de como varían los genes con expresión baja y alta.

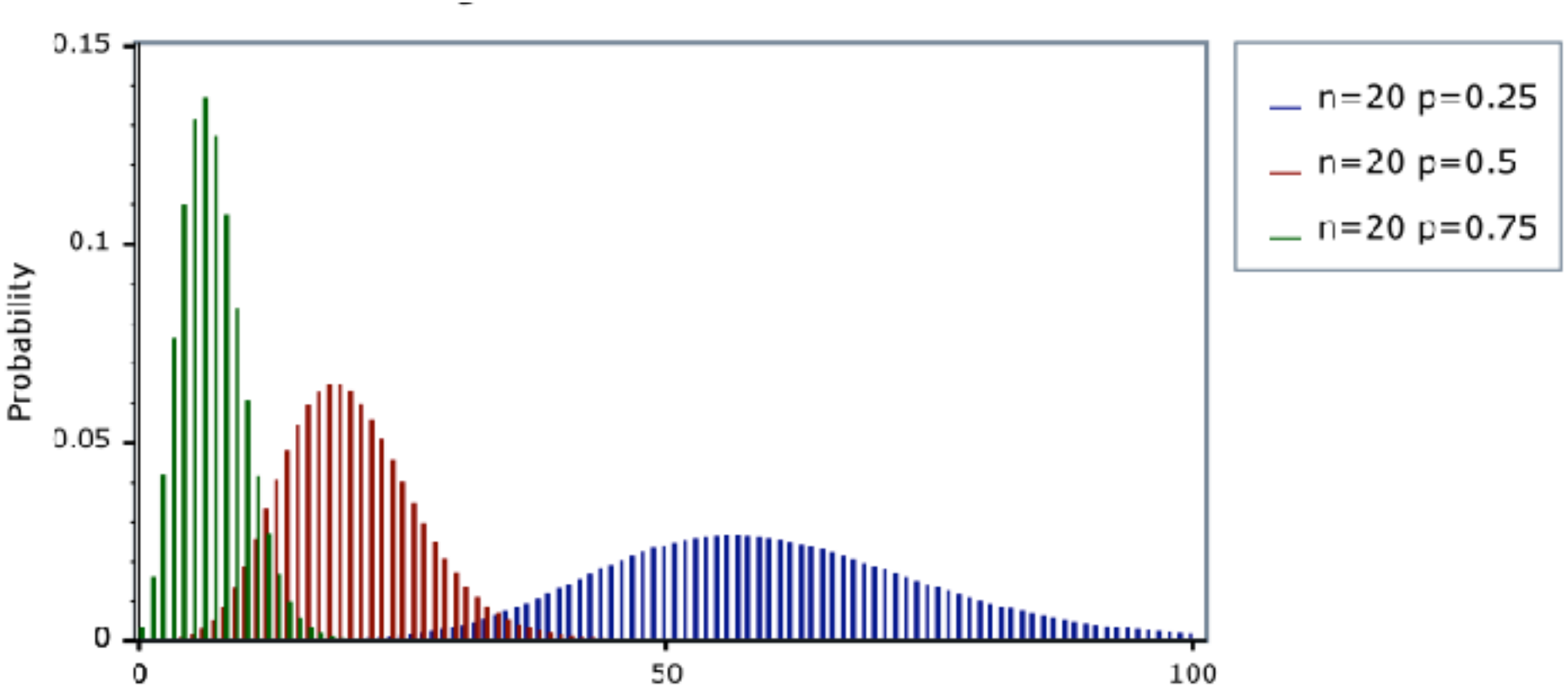
- Este parámetro se utiliza para modelar la distribución BN de cada gen y provee un mejor estimado de los errores. Conforme el número de muestras se incrementa la varianza local (específica a un gen) puede estimarse mejor (aun si la expresión presentará una distribución sesgada, estocástica y anormal).

La distribución binomial negativa

- Una distribución binomial negativa es una distribución de probabilidad discreta del número de pruebas exitosas en un número de pruebas independientes e idéntica tipo Bernoulli antes de que un número específico de fracaso (r) ocurra.

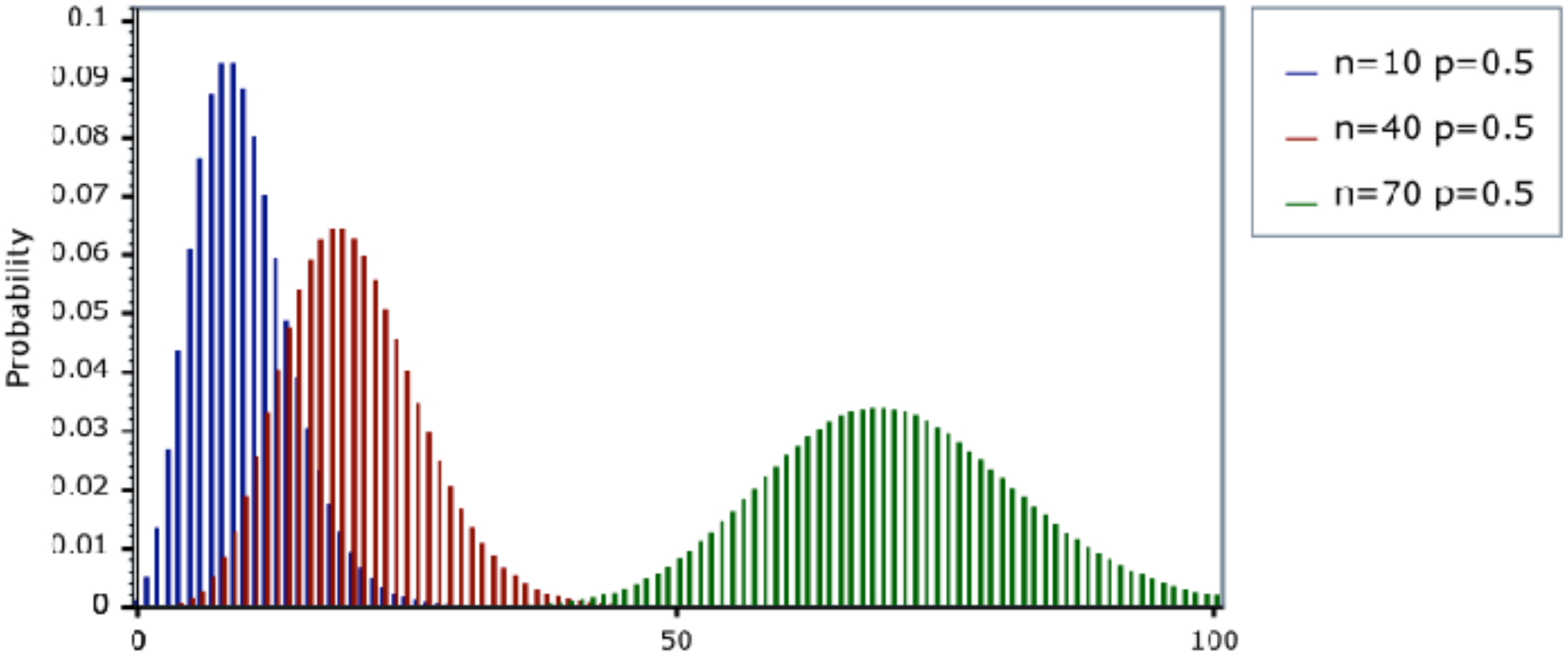
- Si definimos que si un dado cae en “1” es un fracaso y tiramos el dado hasta que el uno caiga tres veces (r= tres fallos), entonces la distribución de probabilidades de los números que no son “1” que obtuvimos tendrán una distribución binomial negativa.

Binomial negativa con cambio en fracción de éxitos



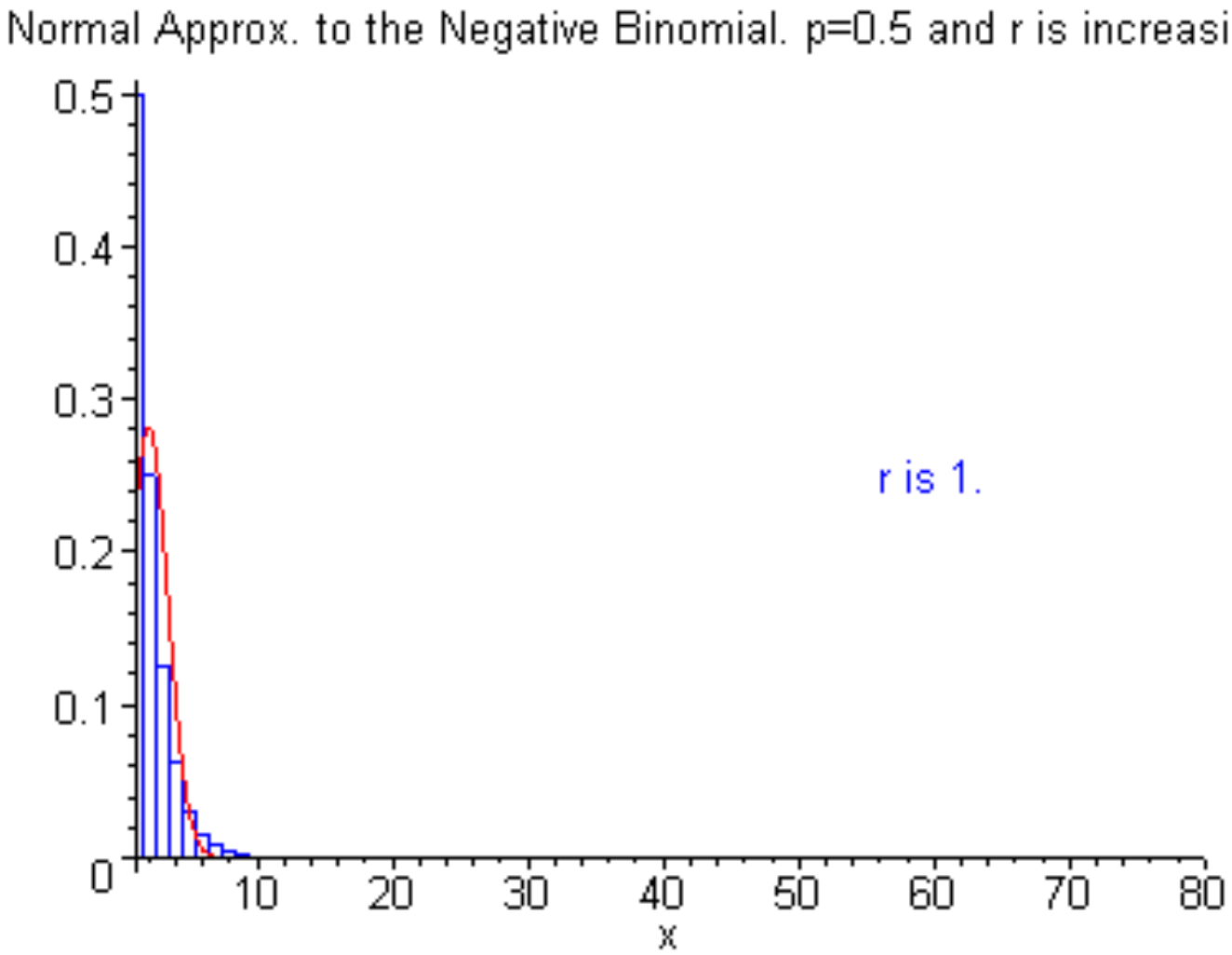
http://www.boost.org/doc/libs/1_36_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/negative_binomial_dist.html

Binomial negativa con cambio en número de éxitos

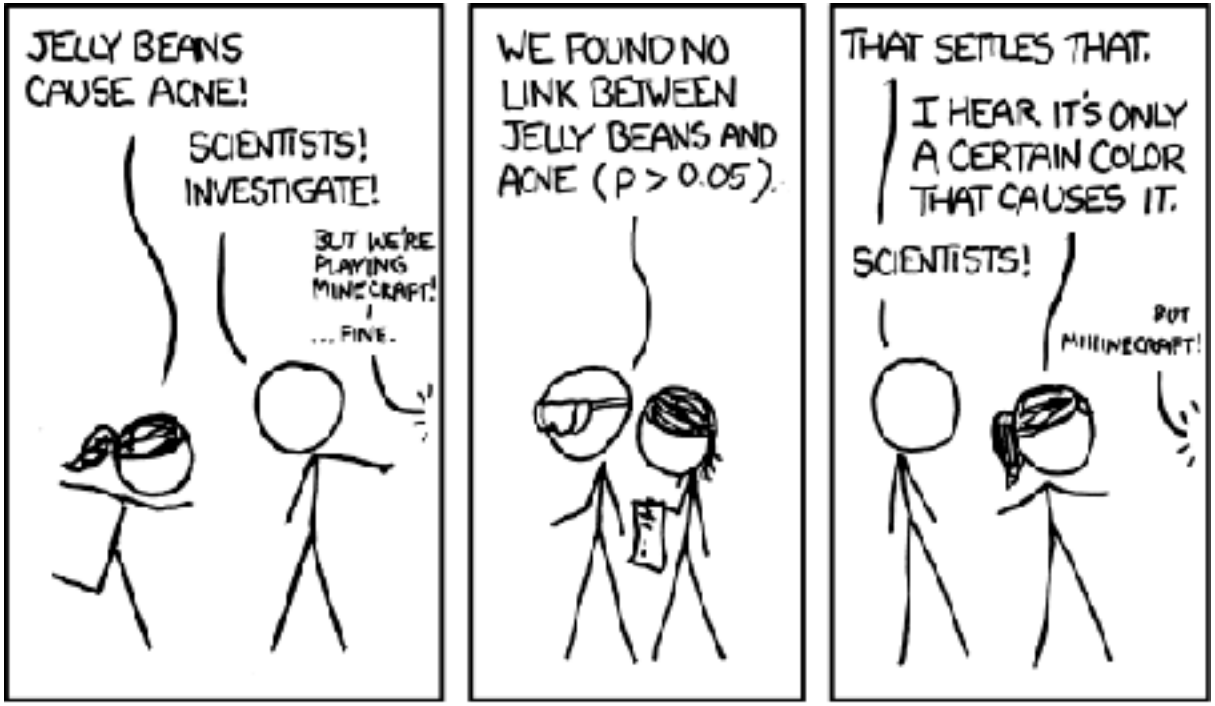


http://www.boost.org/doc/libs/1_36_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/negative_binomial_dist.html

Binomial negativa con cambio en número de fracasos

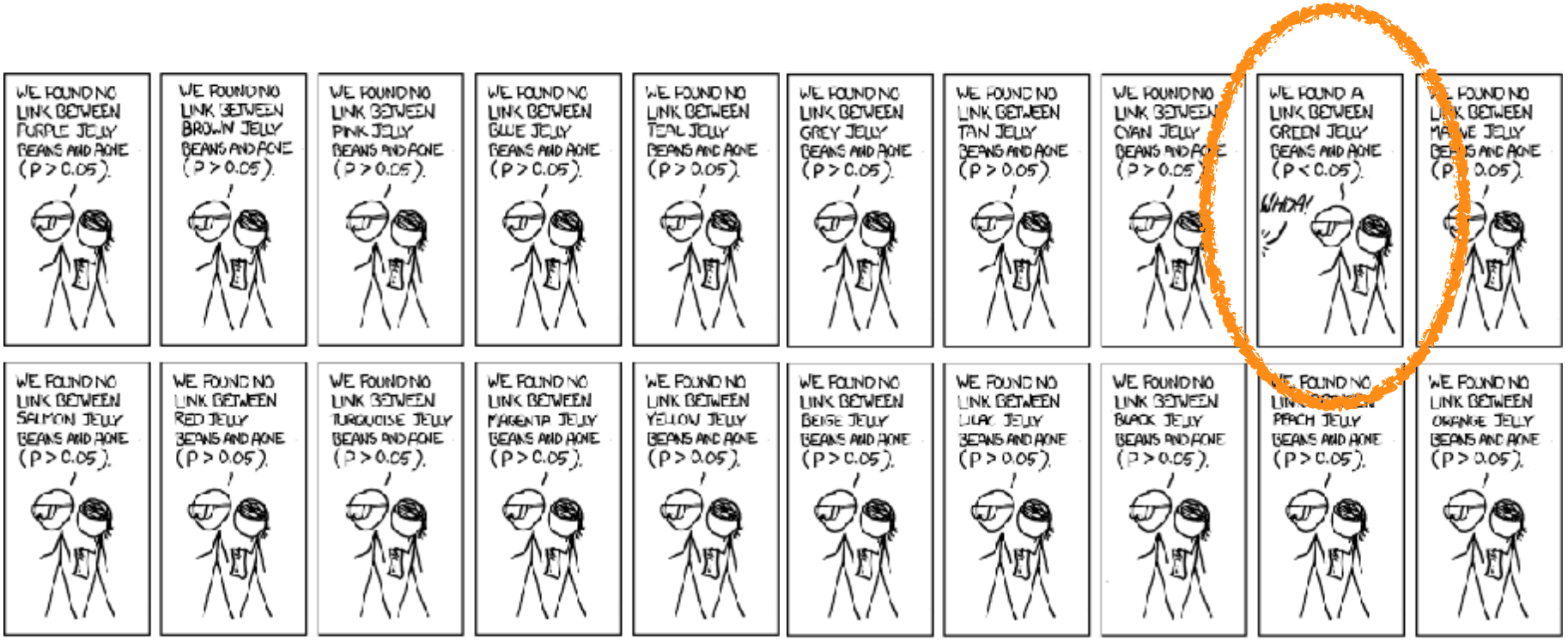


FDR - False discovery rate



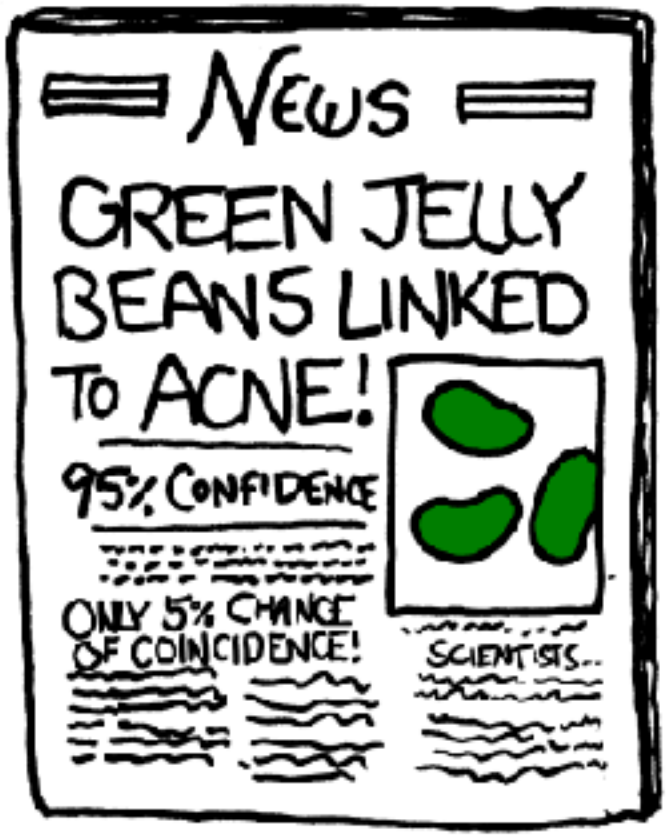
https://xkcd.com/882/

FDR - False discovery rate



https://xkcd.com/882/

FDR - False discovery rate



- La tasa de descubrimientos falsos (false discovery rate - FDR) es la proporción esperada de errores de **Tipo I**.
- Un error de tipo I es cuando uno rechaza la hipótesis nula de manera incorrecta, es decir, cuando se obtiene un **falso positivo**.
- Si se repite una prueba suficientes veces **siempre** habrá un número de falsos positivos.
- La FDR trata de estimar cuantos errores se deben al hecho de haber realizado pruebas múltiples.

https://xkcd.com/882/

Herramientas para análisis de expresión diferencial

Table 5: Comparison of programs for differential gene expression identification (Rapaport et al., 2013; Seyednasrollah et al., 2015; Schurch et al., 2015).

Feature	DESeq2	edgeR	limmaVoom	Cuffdiff
Seq. depth normalization	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
Assumed distribution	Neg. binomial	Neg. binomial	*log*-normal	Neg. binomial
Test for DE	Exact test (Wald)	Exact test for over-dispersed data	Generalized linear model	*t*-test
False positives	Low	Low	Low	High
Detection of differential isoforms	No	No	No	Yes
Support for multi-factored experiments	Yes	Yes	Yes	No
Runtime (3-5 replicates)	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours

Programas para realizar análisis de expresión diferencial

- **DESeq2** (https://bioconductor.org/packages/release/bioc/html/DESeq2.html)
- edgeR (https://bioconductor.org/packages/release/bioc/html/edgeR.html)
- cuffdiff + CummeRBund (http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/)
- sleuth (http://pachterlab.github.io/sleuth/)

Práctica - realizando un análisis de expresión diferencial

https://liz-fernandez.github.io/PBP_transcriptomics_2023/