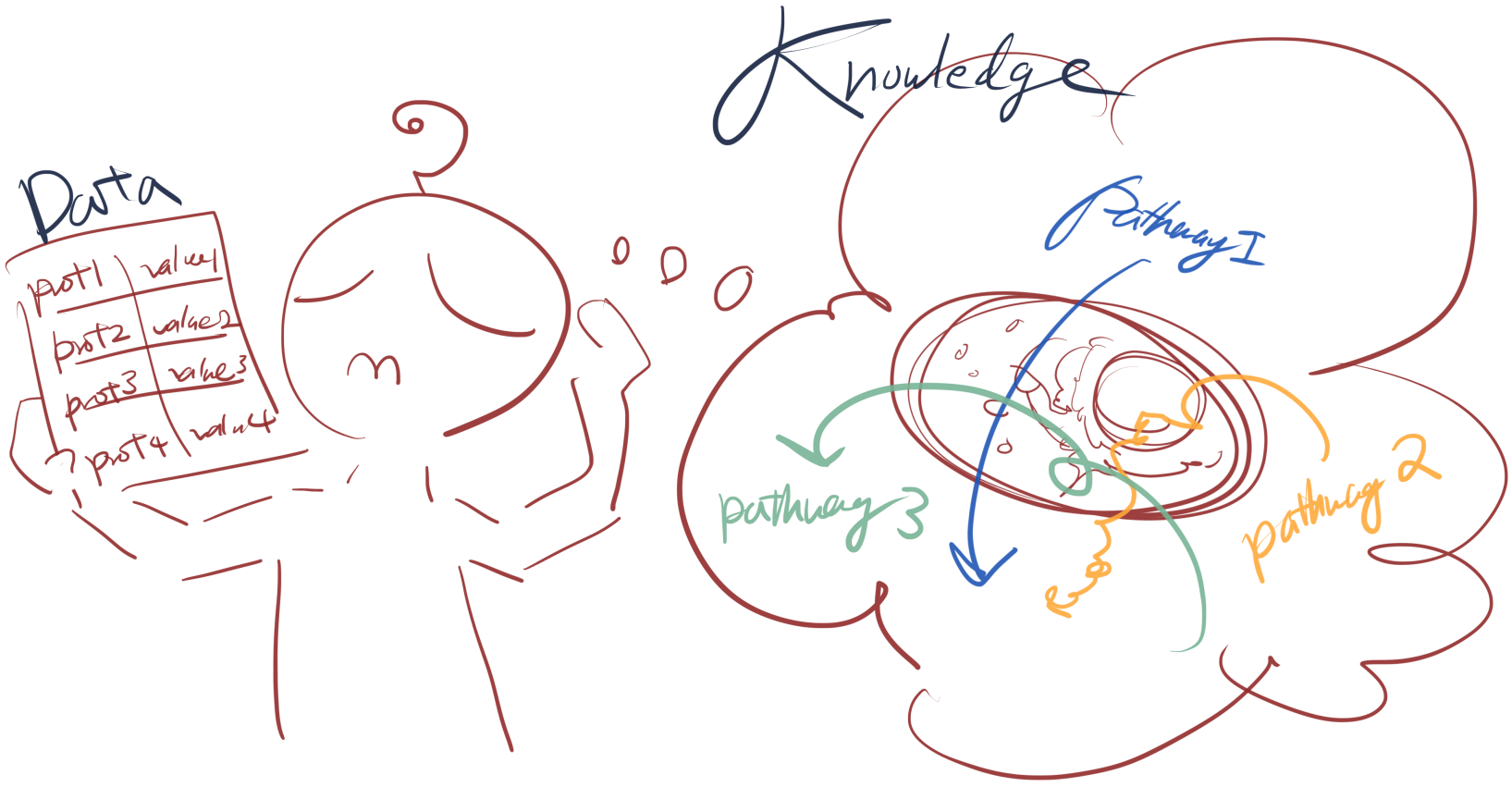
嘴砲 (x) 資料詮釋 (o) 的藝術系列，之一
Annotation and Enrichment Analysis

June Lai

Main Theme — guessing mechanism from data



« Background Check!!! 身家調査 »

Outline

First hour:

- Definition: Set-based analysis (non-graphical)
- Underlying statistical hypotheses: Competitive vs. Self-contained
- Common statistical tests: Hypergeometric / T^2 / GSEA (KS)

  Recall p-values and multiple testing problem: FDR

Second hour:

- Common databases: GO / KEGG / MSigDB / STRING / Reactome
- Requirements:

  Dataset Gene set (with/without values)

  Knowledge Database (hierarchical or not) & 「脳補之力」
- Demonstration: STRING / STAGEs / WebGestalt
- Afterword: The **art** of fine-tuning

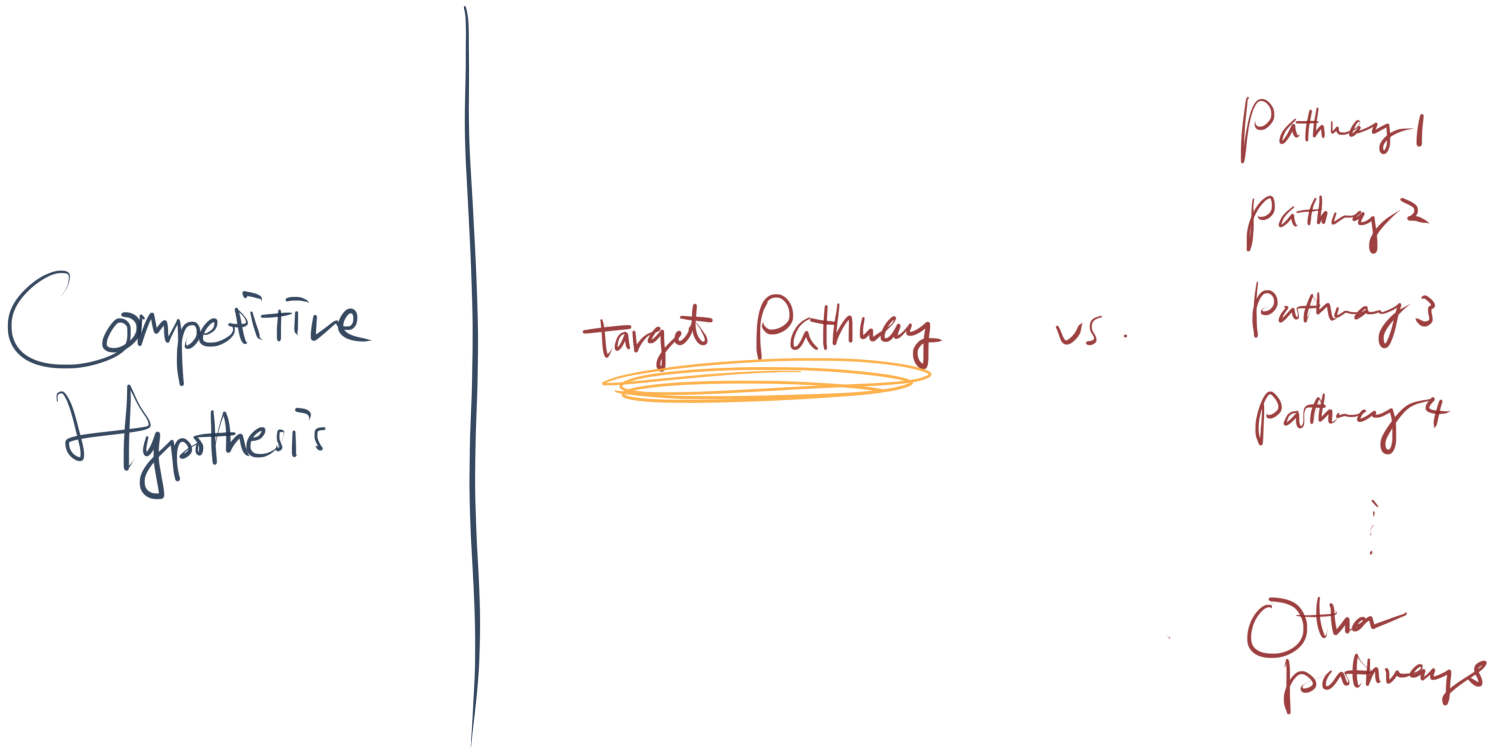
---

reference → https://www.nature.com/articles/s41596-018-0103-9

Competitive vs. Self-contained

H_0：大家一樣強

H_A：我比你們強

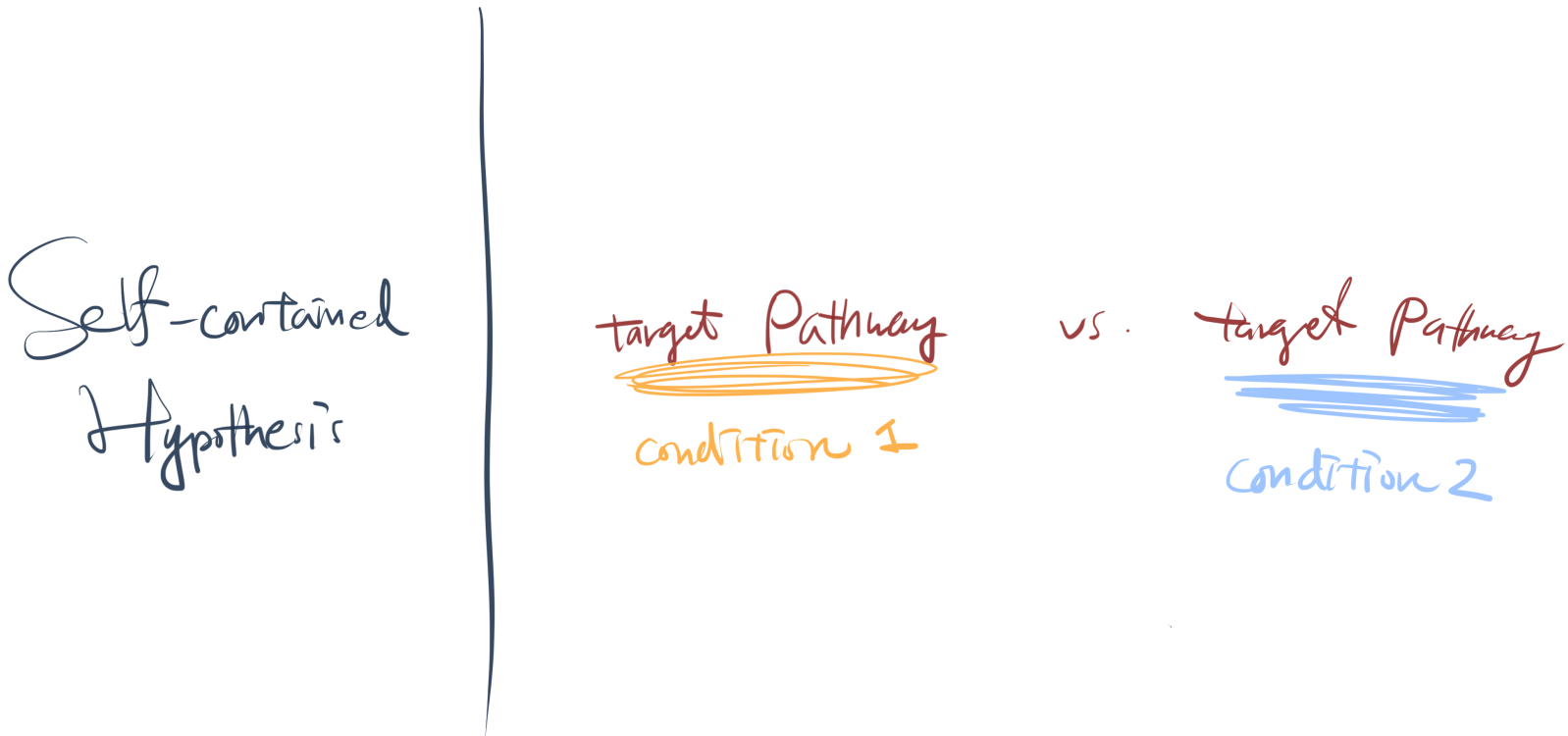


p-value: 在大家應該一樣強的前提下，我長得這副模樣的機率有多少

Competitive vs. Self-contained

H_0：我在狀態一與狀態二是一樣的

H_A：我在狀態一與狀態二是不同的



p-value: 在狀態一與狀態二對我來說應該是一樣的前提下，
我長得這副模樣的機率有多少

Fisher's exact test (gene set)(competitive)

$$p = \frac{\binom{10+20}{10}\binom{300+3560}{300}}{\binom{310+3580}{310}}$$

$$H_0 : \pi_{\text{target pathway}} = \pi_{\text{other pathways}}$$

number of proteins	target pathway	other pathways	total (data)
proteins in the data	10	300	310
absent proteins	20	3560	3580
total (pathway)	30	3860	3890

Hotelling's T^2 test (gene set with values)(self-contained)

$$T^2 = \underbrace{\begin{bmatrix} 2.5 & 3 & -0.5 \end{bmatrix}}_{\text{normalized expression}} \begin{bmatrix} \quad \text{Covariance} \quad \end{bmatrix} \begin{bmatrix} 2.5 \\ 3 \\ -0.5 \end{bmatrix}$$

$$H_0 : \boldsymbol{\mu} = \mathbf{0}$$

---

reference $\rightarrow$
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005601

GSEA (gene set with values)(self-contained/competitive)

- Step 1: Calculation of an Enrichment Score (ES).
  1. Rank order the N genes in D to form $L = \{g_1, \ldots, g_N\}$ according to the correlation, $r(g_j) = r_j$, of their expression profiles with phenotype.
  2. Evaluate the fraction of genes in S weighted by their correlation and the fraction of genes not in S present up to a given position i in L.

$$P_{\text{hit}}(S, i) = \sum_{g_j \in S, j \leq i} \frac{|r_j|^p}{N_R}, \text{ where } N_R = \sum_{g_j \in S} |r_j|^p; \quad P_{\text{miss}}(S, i) = \sum_{g_j \notin S, j \leq i} \frac{1}{N - N_H}$$

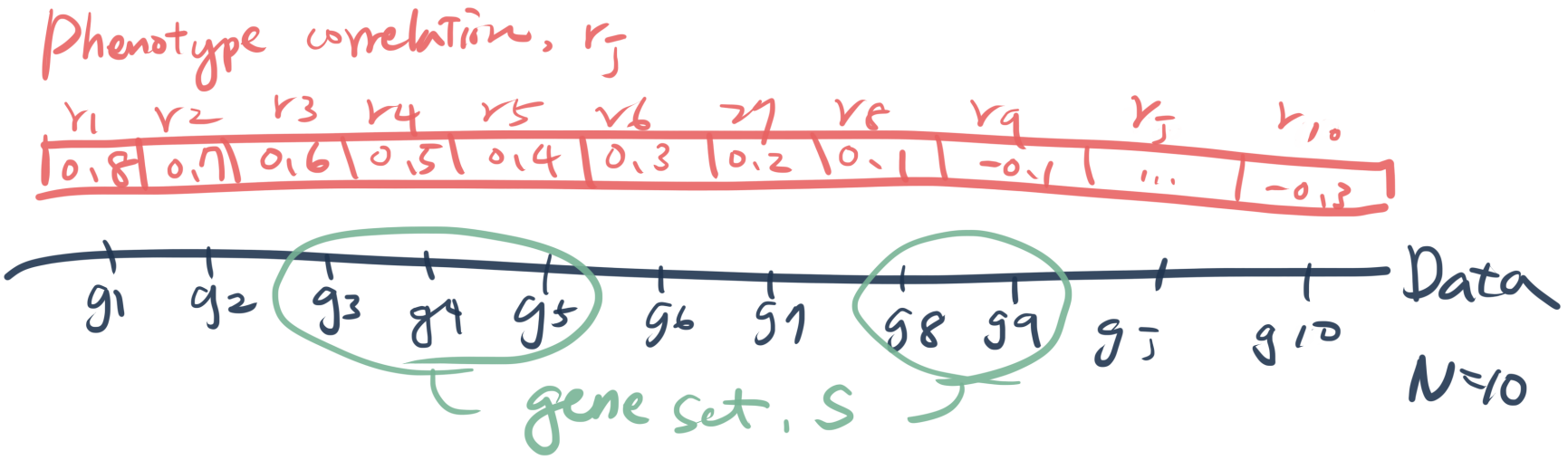
$p = 0$ standard Kolmogorov–Smirnov statistic

$p = 1$ weighted Kolmogorov–Smirnov-like statistic

---

reference → https://www.pnas.org/doi/10.1073/pnas.0506580102

GSEA (gene set with values)(self-contained/competitive)



$N_H = |S| = 5$ $N_R = \Sigma|r_j| = 1.7$

$ES(S)$

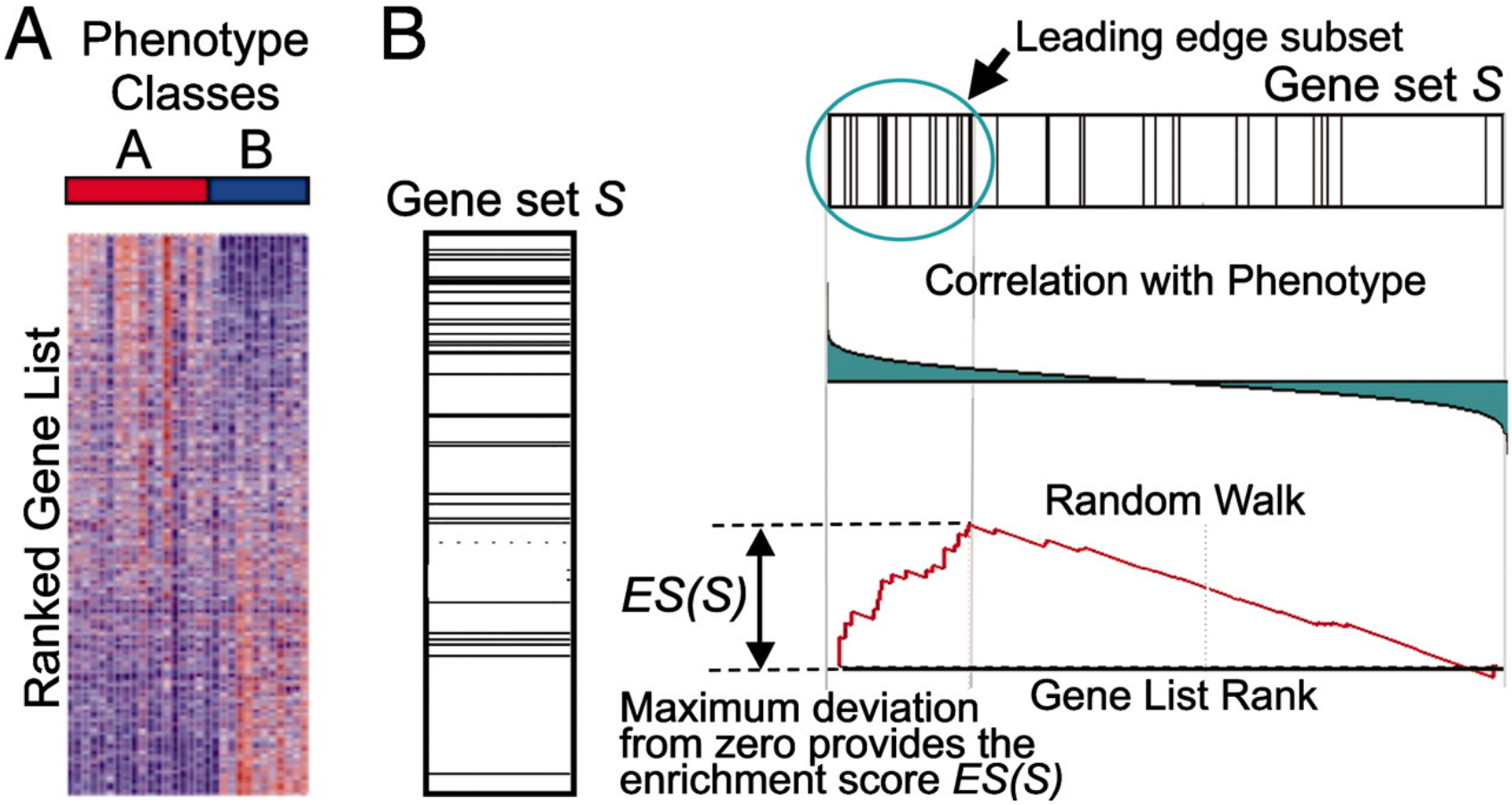
① $i=1$, miss, $P_{miss} = \frac{1}{10-5} = \frac{1}{5}$, $P_{sum} = -0.2$ $= \max\{P_{sum}\}$

② $i=2$, miss, $P_{miss} = \frac{1}{5} + \frac{1}{5} = \frac{2}{5}$, $P_{sum} = -0.4$

③ $i=3$, hit, $P_{hit} = 0.6/1.7$, $P_{sum} = 0.35 - 0.4 = -0.05$

★④ $i=4$, hit, $P_{hit} = (0.6+0.5)/1.7$, $P_{sum} = 0.64 - 0.4 = 0.24$

GSEA (gene set with values)(self-contained/competitive)



GSEA (gene set with values)(self-contained/competitive)

- Step 2: Estimation of Significance Level of ES (self-contained).
  1. Randomly assign the original *phenotype* labels to samples, reorder genes, and re-compute $ES(S)$.
  2. Repeat step 1 for 1,000 permutations, and create a histogram of the corresponding enrichment scores ES_{NULL}.
  3. Estimate nominal P value for S from ES_{NULL} by using the positive or negative portion of the distribution corresponding to the sign of the observed $ES(S)$.
- 沒寫在 Step 2 裡的隱藏資訊：(gene_set permutation, competitive)

4. How many samples do I need for GSEA?

This depends on your specific problem and data characteristics; however, as a general recommendation, if there are fewer than 7 samples per phenotype, GSEA should be run with gene_set rather than phenotype permutation. 3 samples per phenotype are the minimum for the GSEA default signal2noise, and the tTest, ranking metrics.

If you have technical replicates, you generally want to remove them by averaging or some other data reduction technique. For example, assume you have five tumor samples and five control samples each run three times (three replicate columns) for a total of 30 data columns. You would average the three replicate columns for each sample and create a dataset containing 10 data columns (five tumor and five control).

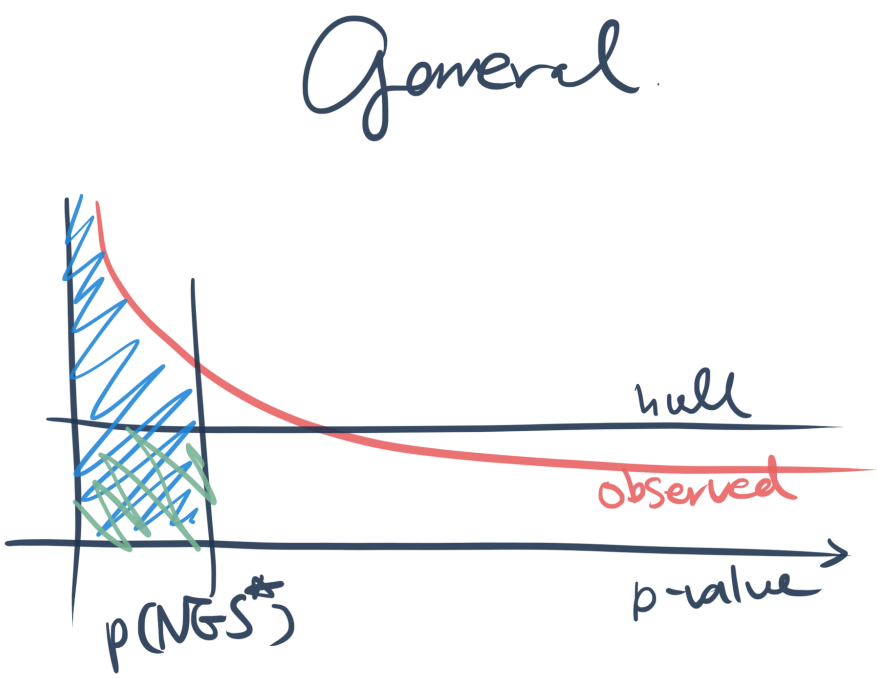
---

reference → https://docs.gsea-msigdb.org/#GSEA/GSEA_FAQ/

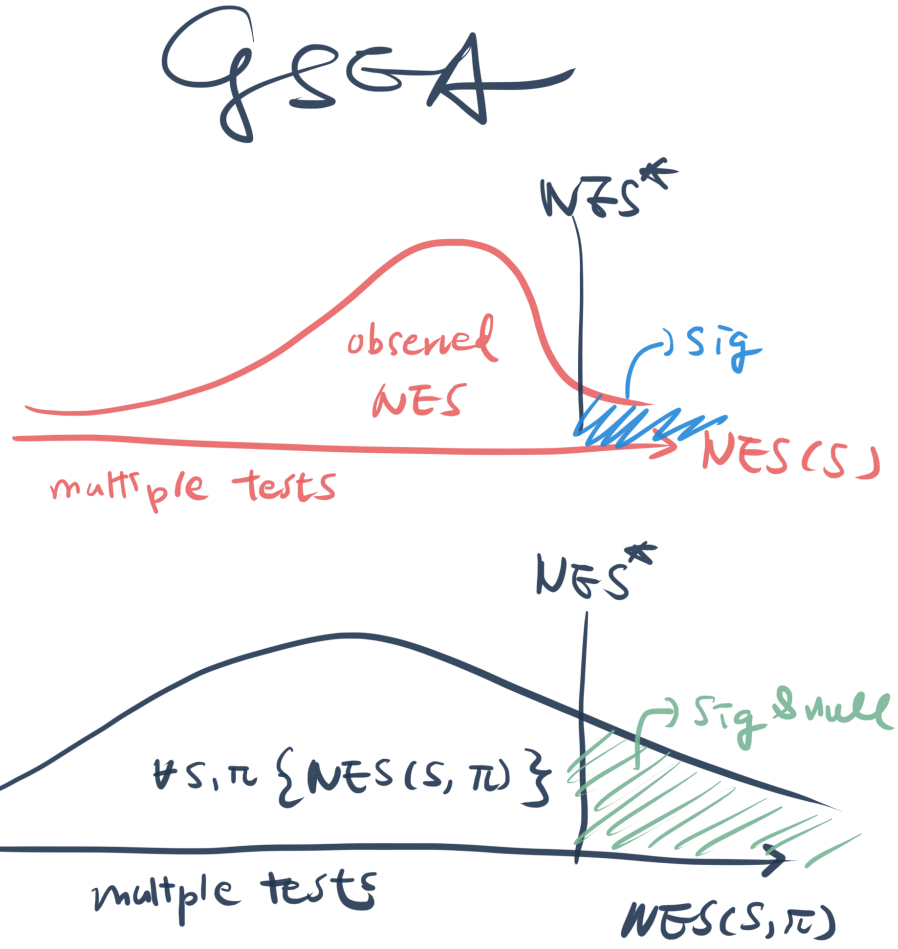
GSEA (gene set with values)(self-contained/competitive)

- Step 3: Adjustment for Multiple Hypothesis Testing.
  1. Determine $ES(S)$ for each gene set in the database.
  2. For each S and 1000 fixed permutations π of the phenotype labels, reorder the genes in L and determine $ES(S, \pi)$.
  3. Adjust for variation in gene set size. Normalize the $ES(S, \pi)$ and the observed $ES(S)$, separately rescaling the positive and negative scores by dividing by the mean of the $ES(S, \pi)$ to yield the normalized scores $NES(S, \pi)$ and $NES(S)$.
  4. Compute false discovery rate (FDR). Create a histogram of all $NES(S, \pi)$ over all S and π. Use this null distribution to compute an FDR q value, for a given $NES(S) = NES^* \geq 0$. The FDR is the ratio of the percentage of all (S, π) with $NES(S, \pi) \geq 0$, whose $NES(S, \pi) \geq NES^*$, divided by the percentage of observed S with $NES(S) \geq 0$, whose $NES(S) \geq NES^*$, and similarly if $NES(S) = NES^* \leq 0$.

GSEA (gene set with values)(self-contained/competitive)



$$FDR = \frac{\#\,sig\ \&\ null}{\#\,significance}$$



Multiple testing adjustment

p-value 從虛無假說 (self-contained or competitive) 生成這樣的 pathway 的機率有多高 (single test)

$\Rightarrow$ 所以只要多試幾次，遲早能抽到顯著的結果 www

FDR 這個看似顯著的 pathway 其實「無關緊要，只是賽到」的機率有多高 (multiple test)，所以 FDR 越高則越可能是假貨

$\Rightarrow$ 做 multiple comparison 是一種「思維」，試的次數越多則越為必要

\(0w0)/ Intermission 中場休息 \(0w0)/

Outline

First hour:

- Definition: Set-based analysis (non-graphical)
- Underlying statistical hypotheses: Competitive vs. Self-contained
- Common statistical tests: Hypergeometric / T^2 / GSEA (KS)

  Recall p-values and multiple testing problem: FDR

Second hour:

- Common databases: GO / KEGG / MSigDB / STRING / Reactome
- Requirements:

  Dataset Gene set (with/without values)

  Knowledge Database (hierarchical or not) & 「脳補之力」
- Demonstration: STRING / STAGEs / WebGestalt
- Afterword: The **art** of fine-tuning

---

reference → https://www.nature.com/articles/s41596-018-0103-9

STRING v12 Demonstration

Usage scenario:

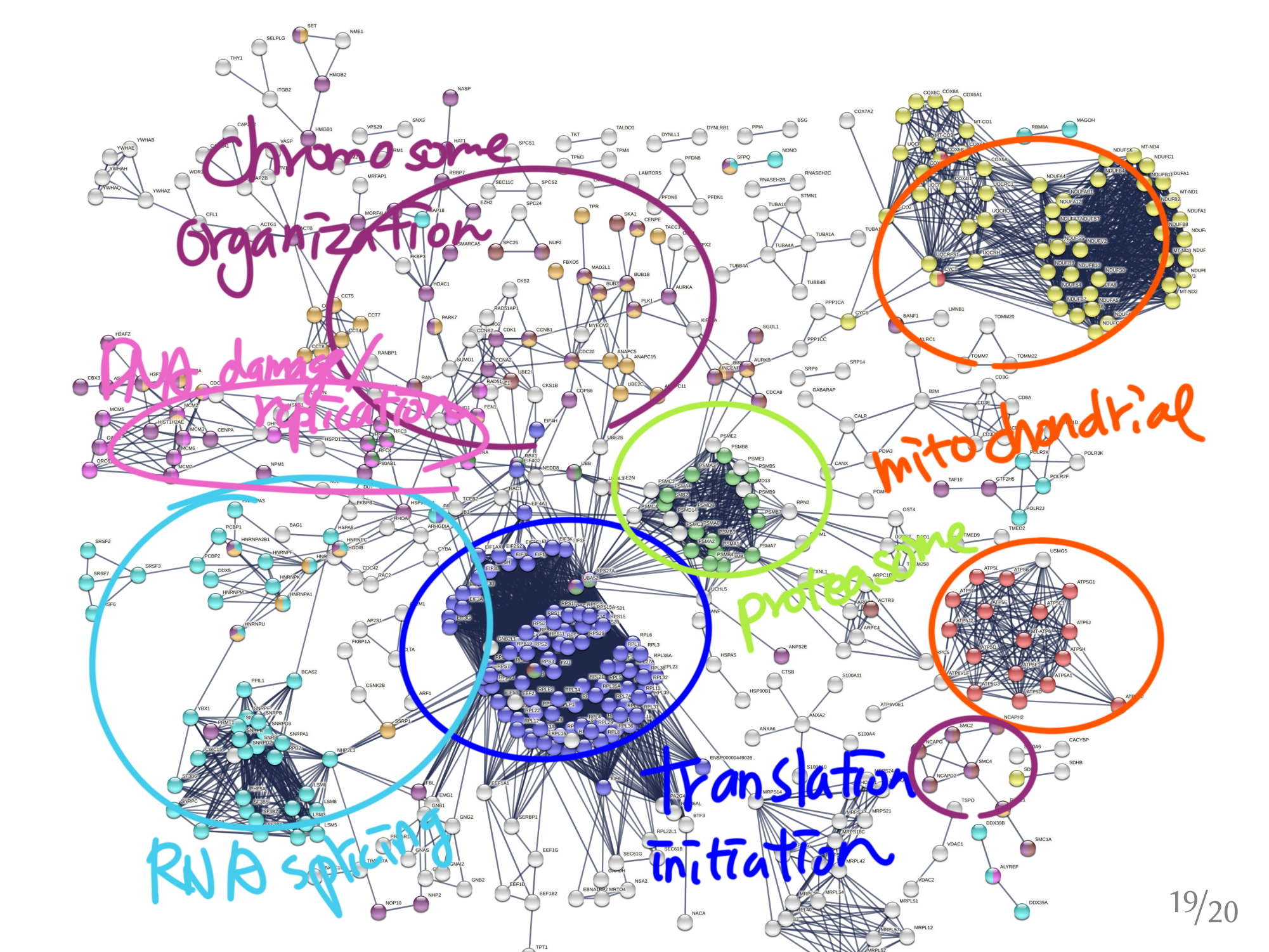
- Single protein
- **Multiple proteins**
- Multiple proteins with value
- Search known pathways

腦補之力：

- Hub proteins / Bottleneck
- **Pathway highlighting**
  點點看，用最少的顏色把圖面上最大坨的 clusters 都個別解釋好
  https://version-11-5.string-db.org/cgi/network?networkId=bQ32wU7U2Dbk
- Clustering

---

reference →
https://string-db.org/cgi/about?footer_active_subpage=references


Chromosome organization
DNA damage/ replication
Mitochondrial
Protesome
Translation initiation
RNA splicing

STAGEs (2023) & WebGestalt (2024)

STAGEs Usage scenario:

- Gene set with values (time-series)
- Web-based pipelines

APP https://kuanrongchan-stages-stages-vpgh46.streamlitapp.com/

REF https://www.nature.com/articles/s41598-023-34163-2

WebGestalt Usage scenario:

- Gene set / Gene set with values (GSEA)
- Web-based pipelines

APP https://www.webgestalt.org/

REF https://academic.oup.com/nar/article/52/W1/W415/7684598