

人生好難

有夠常態的 Normal distribution (Gaussian distribution)

常態分布 (高斯分布)

- 將一連續變項之觀察值發生機率以圖呈現其分布情形
- 中線會是平均值
- 範圍：負無窮大 ~ 無窮大

References

homepage.ntu.edu.tw/~clhsieh/biostatistic/4/4-1.htm

統計學中的 P-value 你確定你會？

p值其實就是在分佈常態分佈下 $\geq t$ 值的機率密度值(probability density value , $P(x \geq t)$)

T-value

越接近 0 代表差異越沒有差異，相反的，越遠離差異就越大

$$t = \frac{(X_1 - X_2)}{\sqrt{\frac{(S_1)^2}{n_1} + \frac{(S_2)^2}{n_2}}}$$

(平均數1 - 平均數2) / 兩組資料的標準差

How to calculate P-value

從 T-value 積分到無窮大之面積，就為 P-value

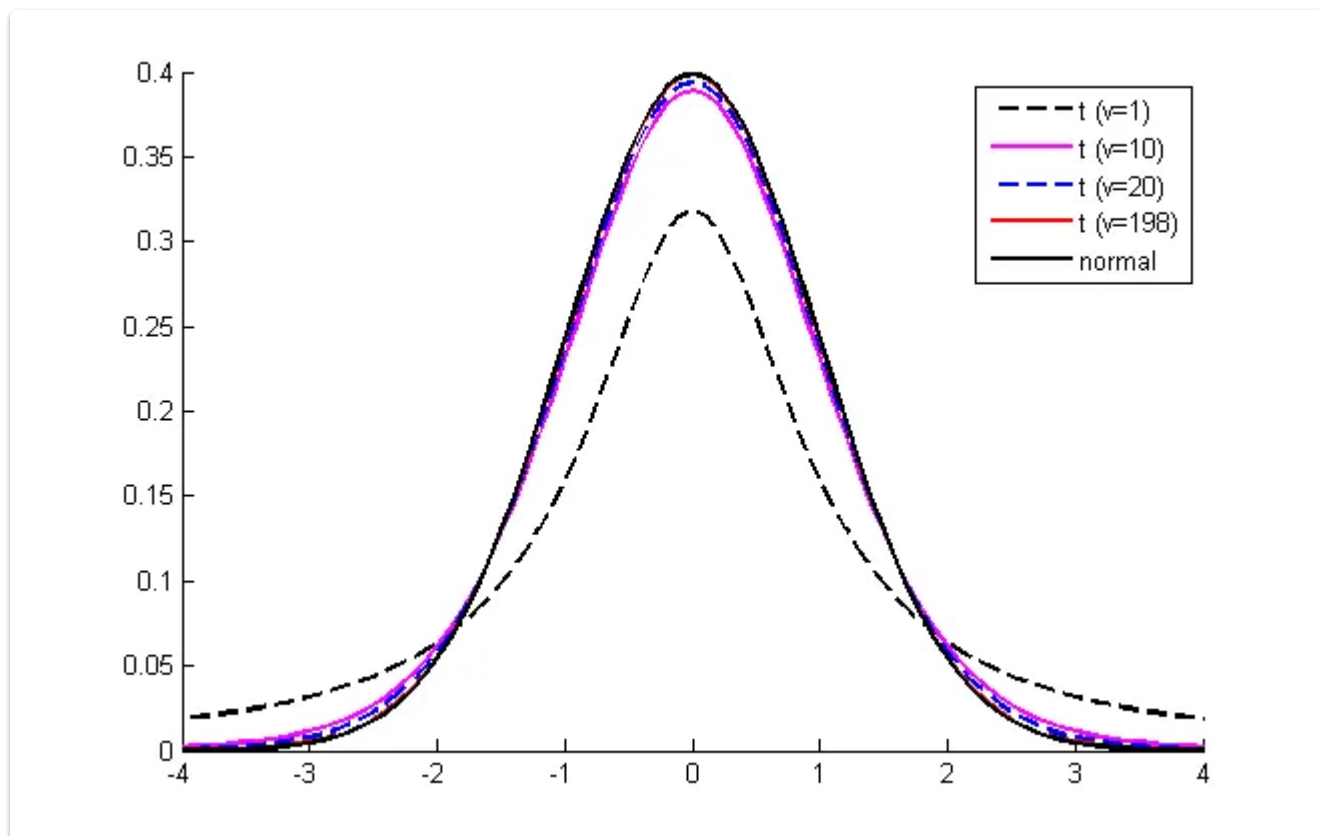
有兩邊無窮大，所以正常來說兩邊會加起來，也可以面積 * 2

T分布 (Student's t-distribution)

抽樣下的常態分佈

自由度越高，會越接近該樣本的常態分佈

$v \rightarrow$ 自由度 ($N1 + N2 - 2$)



Null hypothesis

- 又稱**虛無假說**，通常被標寫為 H_0
- 假設兩筆資料集相似、差不多

P-value < 0.05

意義就在於要當下 T-value 到無窮大的面積大於 0.05 的自由度
而就是說我們**接受虛無假說**，又可以說兩筆資料是有差異的

↔

意義就在於要當下 T-value 到無窮大的面積小於 0.05 的自由度
而就是說我們**拒絕虛無假說**，有可以說兩筆資料差異不大

References

[統計學:大家都喜歡問的系列-p值是什麼. 常常看到有人說p-value... | by Tommy Huang | Medium](#)

你真的懂甲基化嗎？

甲基化

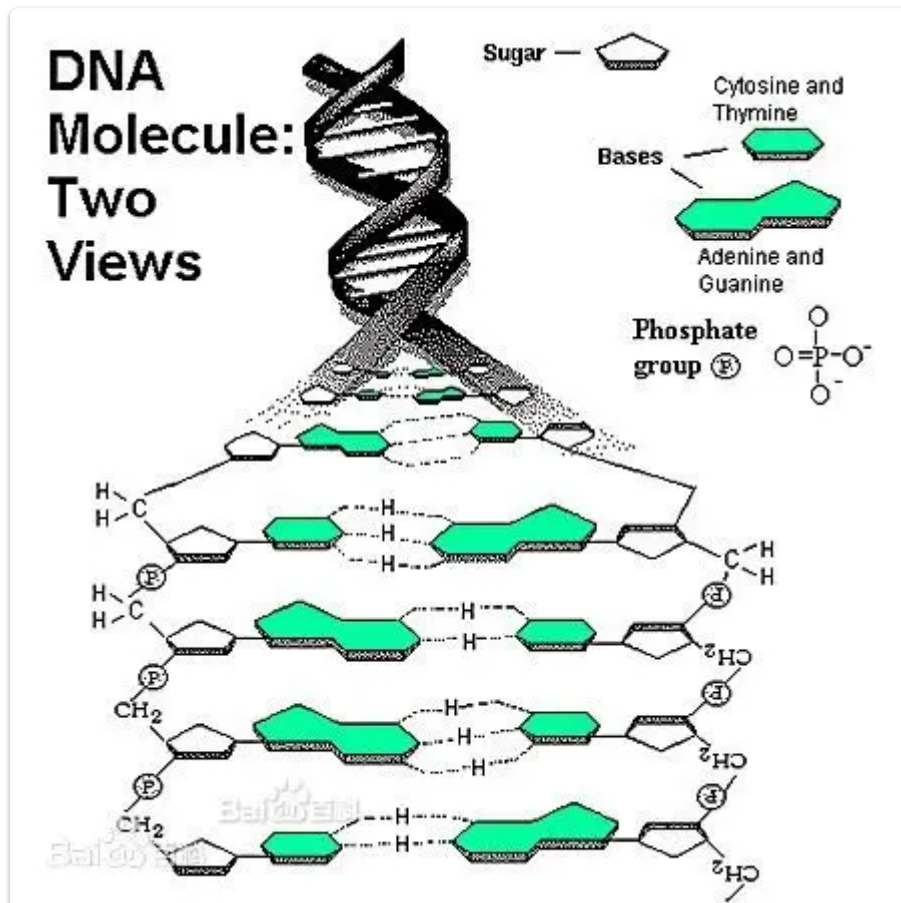
簡單說：DNA 分子上多了一個甲基

而這個甲基只會結合在一個特定位置 → 胞嘧啶 (Cytosine)

如果要仔細一點.....，你可能要自己查，網路很多資料

CpG

胞嘧啶 (C , Cytosine)、磷酸 (p , phosphoric acid)、鳥嘌呤 (G , Guanine)
其實就是配對的核苷酸



CpG位點 (CpG site)：就是每一個這樣的核苷酸

CpG島 (CpG island)：包含有很多個CpG位點的DNA片段。

CpG

人類 CpG 特徵：

- 有 25000 個 CpG 島
- 50% 在啟動子區域，25% 在基因內，剩下....我也不知道
- 人體內約 60~70% 基因的啟動子中都具有 CpG 島

啟動子 (Promoter)

啟動子未必在起始密碼子的上游
所以有些甲基化位點會在基因內部

甲基化如何影響基因表達

1. 基因啟動子區域存在 CpG 島
2. CpG 島上的 CpG 位點通常會有甲基
→
3. CpG 島被甲基化時會抑制基因的表達
4. CpG 島上有很多 CpG 位點，CpG 島上的多少 CpG 位點被甲基化，才算這個 CpG 島被甲基化

轉錄相當於表達？！

References

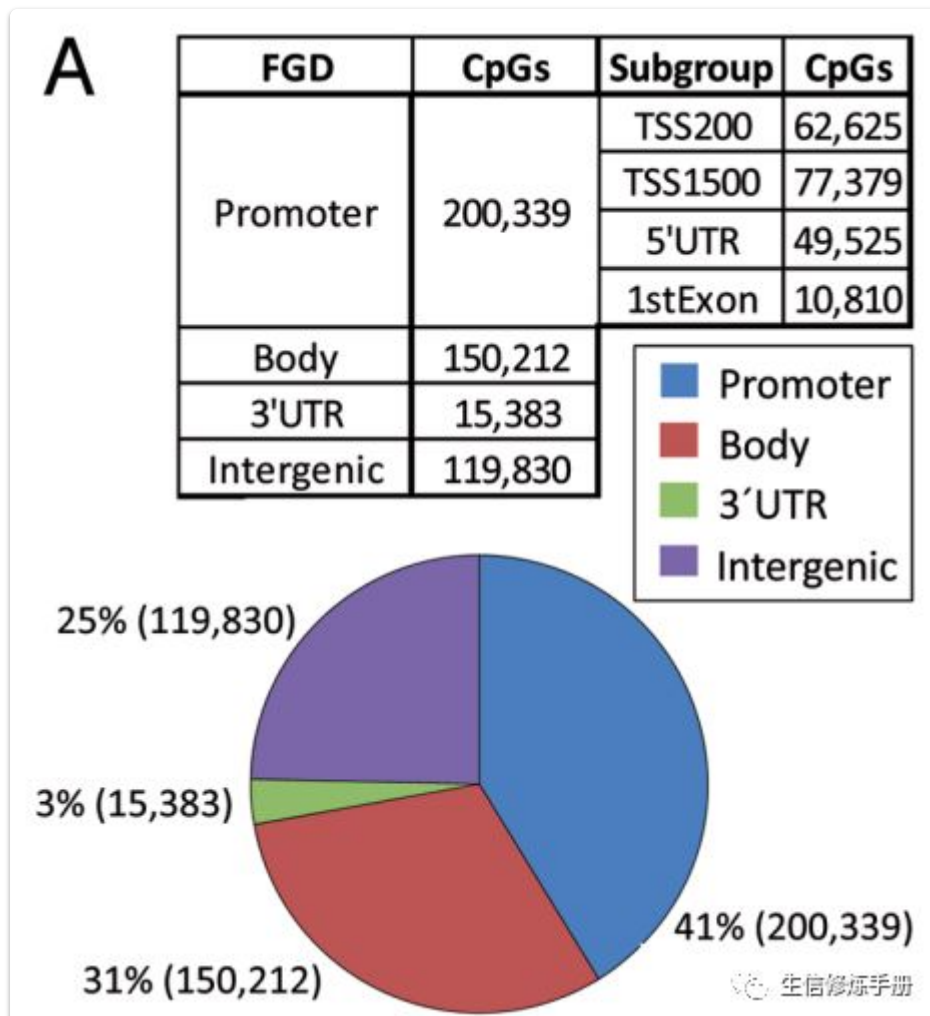
[人的甲基化、基因结构、启动子的一些概念 - 简书 \(jianshu.com\)](http://jianshu.com).

[甲基化芯片学习记录 \(qq.com\)](http://qq.com). 看看這個....我是真的看不下去 太詳細了

功能註釋中的字你真的懂嗎？

基因上的分布

基因組可以劃分為下列 4 個區域

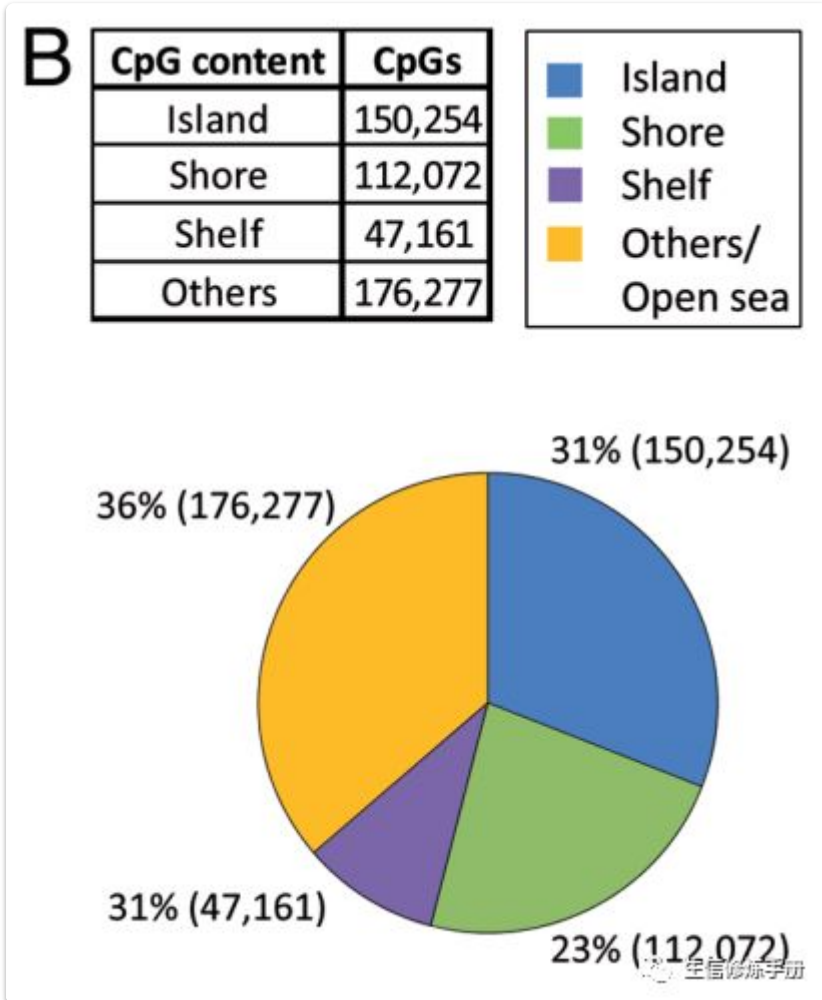


- Promoter 啟動子
- Intergenic 基因區間

在各種 CpG 區域的分布

把 CpG Island 當作圓心，往外的距離擴張

CpG Island → 2kb → CpG Shore → 2kb → CpG Shelf → 2kb → Open sea



讓我們聽聽 ChatGPT 怎麼說

甲基化晶片註釋中的 CpG island、CpG shores、CpG shelf 和 open sea 是描述 DNA 甲基化區域的四個術語。

1. CpG 島 (CpG islands) 指的是一段 GC 含量豐富、長度大於 200bp 的 DNA 區域，其中 CpG 二核苷酸密度高於整個基因組的平均水平。CpG 島通常位於基因的啟動子區域，對於基因表達和調控具有重要作用。
2. CpG shores 指的是靠近 CpG 島但不包括 CpG 島的區域，通常距離 CpG 島不超過 2kb。這些區域的 GC 含量較低，CpG 二核苷酸密度也較低。CpG shores 對基因表達的調控具有微小的影響。
3. CpG shelf 指的是位於 CpG 島和 CpG shores 之間的 DNA 區域，通常距離 CpG 島不超過 2kb，但 CpG 二核苷酸密度較低。這些區域與 CpG 島和 CpG

shores 一樣也可能對基因表達具有一定的影響。

4. Open sea 指的是與 CpG 島、CpG shores 和 CpG shelf 距離較遠的 DNA 區域，其中 CpG 二核苷酸密度很低，不容易被甲基化。這些區域相對於 CpG 島、CpG shores 和 CpG shelf 來說基因表達的影響更小。

註釋這些甲基化區域可以幫助我們了解基因表達的調控機制以及某些疾病的發病機制。

References

[甲基化芯片注释中的CpG shores, open sea 是什么 - 简书 \(jianshu.com\)](#).
[有問題就問他 \(ChatGPT \)](#).

甲基化類型

類型可以分為兩種

維持甲基化 (maintenance DNA methylation)、重新甲基化 (de novo methylation)

- 維持甲基化
- 重新甲基化

維持甲基化

- 在甲基轉移酶的作用下，DNA 的半保留複製過程中，會在子鏈的相應位置進行甲基化修飾過程
- 細胞在DNA複製過程中，將原有的甲基化狀態精確地複製到新合成的DNA分子中。這是由一些甲基轉移酶（如 DNMT1）和甲基化調節蛋白（如 UHRF1）協同作用完成的。
- 對於保持細胞的表觀基因組穩定性和遺傳穩定性至關重要。

重新甲基化

- 在甲基轉移酶的作用下，原本沒有甲基化的 DNA 雙鏈上，進行甲基化過程，之後由維持甲基化酶維持穩定的 DNA 甲基化狀態
- 在某些情況下，DNA 的甲基化狀態可能會受到外部環境、生活方式或疾病等因素的影響而被重新調整，這可能會對基因表達和細胞功能產生影響。
- DNA上甲基化狀態的調整，可能是從未被甲基化的區域重新甲基化，也可能是已經被甲基化的區域被去甲基化或調整甲基化水平。這些調整通常是由DNA甲基轉移酶、DNA脫甲基酶或其他調節因子所引起的。

References

[表观遗传学: DNA甲基化 - 简书 \(jianshu.com\)](http://jianshu.com)

好想懂甲基化芯片的分析流程

芯片

1. 主要分為 450k 和 850k
2. 都是採用兩種探針 Infinium I 和 Infinium II
 - |：甲基化 M 和非甲基化 U，會有綠紅螢光探測訊號
 - ||：兩者會在一起

β value

β -value = 甲基化珠粒類型的強度值 / (甲基化珠粒類型的強度值 + 未甲基化珠粒類型的強度值 + 100)

M 對應到甲基化強度，U 對應到未甲基化強度

$$\beta = \text{Max} (M, 0) / (\text{Max} (M, 0) + \text{Max} (U, 0) + 100)$$

使用 P-value 對數據進行質量過濾， $P\text{-value} > 0.01 * \beta\text{-value}$ 被認為低於最小強度，閥值顯示為 "NA"

如果想知道 β value 的細節，可以讀下面這篇論文

[通過微陣列分析定量甲基化水準的 \$\beta\$ 值和M值方法的比較 - PMC \(nih.gov\)](http://pmc.nih.gov)

描述了 M-value 和 β -value 的權衡

一般來說，具體的 β value 的意義是：

- 任何 ≥ 0.6 的 β value 都被認為完全甲基化
- $0.6 > \beta \text{ value} > 0.2$ 都被認為部分甲基化
- 任何 ≤ 0.2 的 β value 都被認為完全未甲基化

差異甲基化區域鑑定及可視化

甲基化信號值矩陣有 3 個層次的差異分析：DMP、DMR、DMB

- DMP 代表找出 Differential Methylation Probe (差異化 CpG 位點)
- DMR 代表找出 Differential Methylation Region (差異化 CpG 區域)
- DMB 代表找出 Differential Methylation Block (更大範圍的差異化區域)

DMP

就是一個個的差異甲基化 CpG 位點

DMR

科學家認為，DMR 這樣連續差異的片段，對於基因的影響會更加明顯，可以使打擊精度更為精確，可使結論的數據更少，以便挑選

DMB

有人認為 DMR 不夠顯著，DNA 上的甲基化出現變化，可能是綿延幾千點的，所以有了 DMB

References

[甲基化芯片的一般分析流程 - 知乎 \(zhihu.com\).](#)

Data References

GSE62336 [GEO Accession viewer \(nih.gov\).](#)