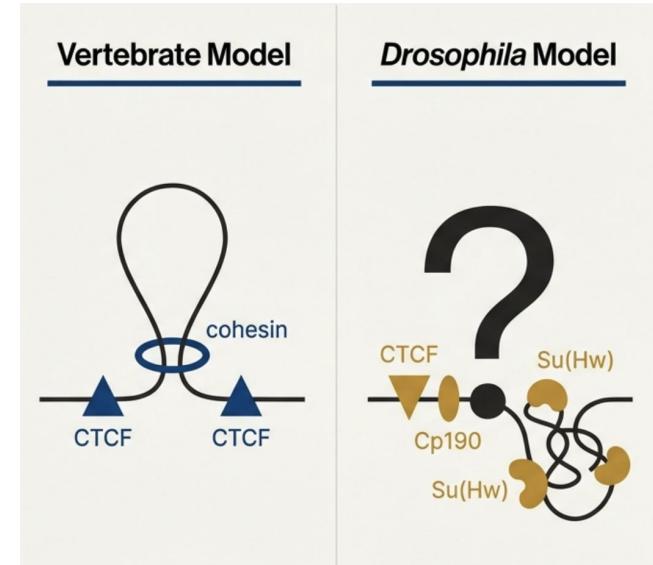


Topological screen identifies hundreds of Cp190- and CTCF-dependent Drosophila chromatin insulator elements

Group 1

Abstract

- *Drosophila* Insulator(染色質絶縁子) : first DNA elements found to regulate gene expression
 - Do not know how many?
 - How do they impact genome folding? Why is DESeq2 used?
- Experiment Design
 - Cultivating two specialized fly cell types using gene editing
 - Cp190-KO & CTCF-KO
 - Hi-C : Comparing the structure of control cells and cells lacking the protein



Abstract

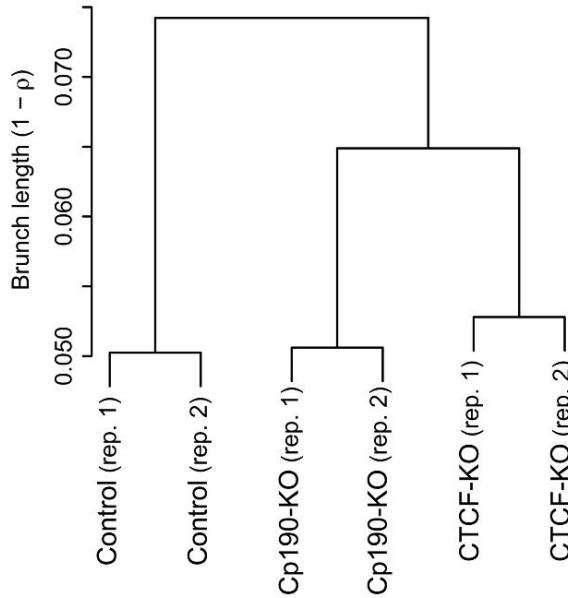
- Results
 - Insulators are responsible only for local structural regulation, with an influence spanning approximately 300 kb.
 - Cp190 promotes co-binding of other insulator proteins
 - *Drosophila* insulators block chromatin contacts by forming loops needs revision
 - Identify 745 newly-defined insulators (15% FDR)
 - Provides an important resource to study *Drosophila* genome folding

Fig 1C

Original

C

Hi-C experiment clustering



Reconstructed

Reproduction of Fig 1C (Normalized Log O/E)

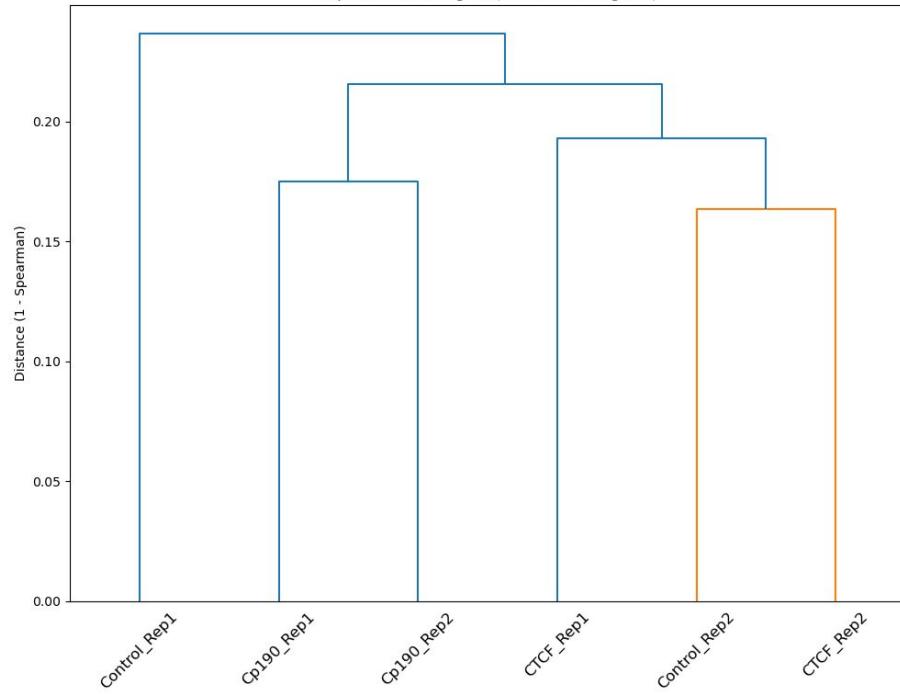


Fig 1C

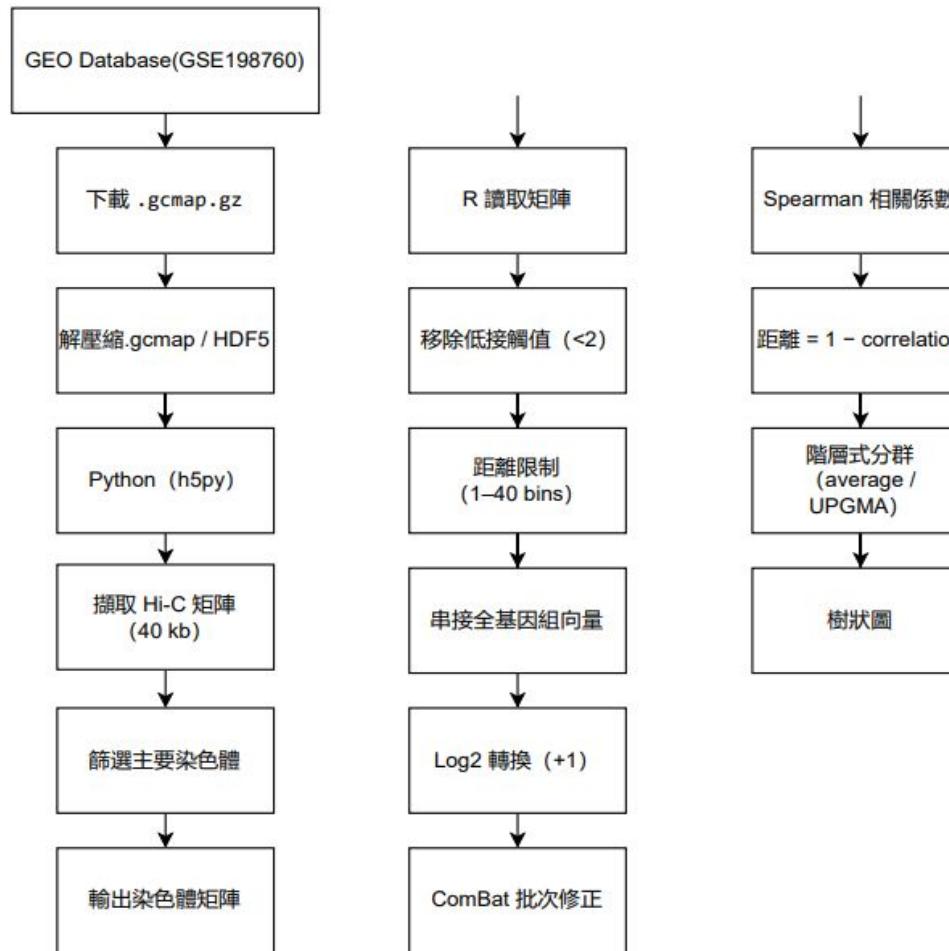


Fig 1C

程式語言與環境 (Programming Language & Environment)

- Python 3.13.7
- R 4.5.2

Python 函式庫 (Python Libraries)

- **h5py**: 用於讀取原始的 .gcmap 檔案(其底層為 HDF5 格式)。
- **numpy**: 用於高效的數值矩陣運算(如合併重複樣本時的矩陣相加)。
- **glob**: 用於檔案路徑比對, 自動搜尋資料夾 內所有的 .gcmap 檔案。
- **os**: 標準作業系統介面, 用於處理檔案系統路徑與名稱 。

Fig 1C

R 套件 (R Libraries)

- **sva**: 核心統計套件，執行 **ComBat 演算法** 以修正重複樣本間的批次效應 (Batch effects)。
- **limma**: 用於構建線性模型矩陣，定義樣本分組邏輯(如 Control vs. KO)以輔助 ComBat 運算。
- **dendextend**: 用於檔案路徑比對，自動搜尋資料夾 內所有的 .gcmap 檔案用於樹狀圖 (Dendrogram) 視覺化增強，實現節點旋轉與自定義標籤排列。
- **pheatmap**: 用於繪製交互式熱圖，視覺化呈現 Spearman 相關係數矩陣。
- **stats**: 執行基礎統計運算，包含 Spearman 相關性分析與階層式分群 (hclust)。

Fig 1C

Combat: 批次效應修正 (Batch Effect Correction), 還原生物學訊號

- 為什麼需要它？

- 批次效應 (Batch Effects): 因實驗時間、試劑批號或操作人員不同，產生的「非生物學」技術誤差。
- 雜訊干擾 : 在原始數據中，技術誤差往往大於生物差異 (Technical Variation > Biological Variation)。
- 後果 : 電腦會根據「哪一批做的」來分類，而不是根據「哪種細胞」來分類。

Fig 1C

Combat: 批次效應修正 (Batch Effect Correction), 還原生物學訊號

- Combat功用

- 核心算法 : 基於經驗貝葉斯 (Empirical Bayes) 框架的統計方法。
- 校正機制 :
 1. 中心化 (Center) : 校正不同批次間的平均值差異 (Location)。
 2. 縮放 (Scale) : 校正不同批次間的變異數差異 (Variance)。
- 目的 : 保留感興趣的生物變異 (Control vs. KO), 移除系統性的技術雜訊。

- 成效 (Outcome)

- 成功將不同批次的數據「對齊」到同一基準線上。
- 確保最終的聚類分析 (Clustering) 是真實反映細胞的基因組結構差異。

Fig 1C

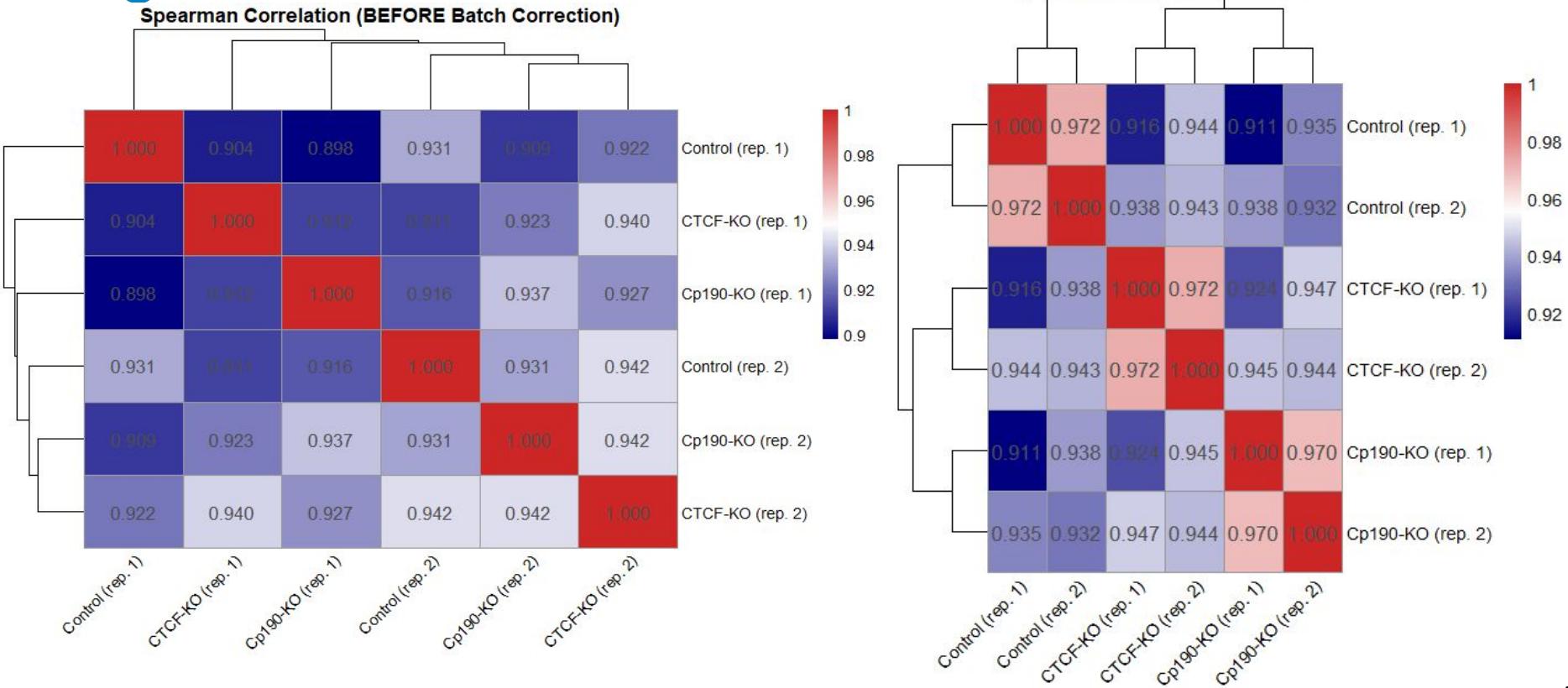
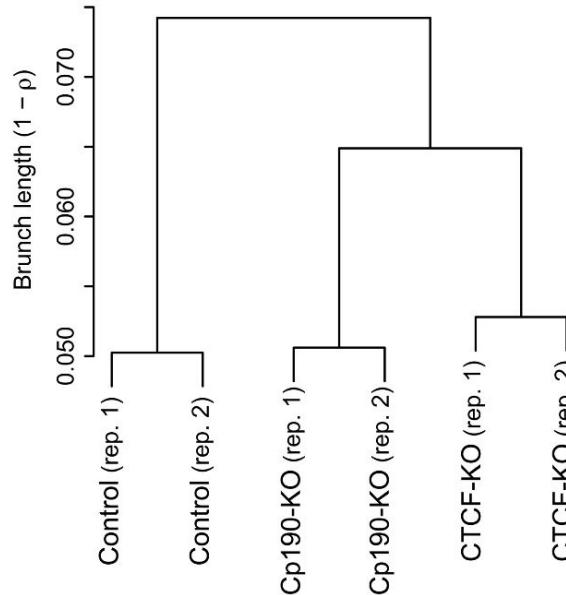


Fig 1C

Original

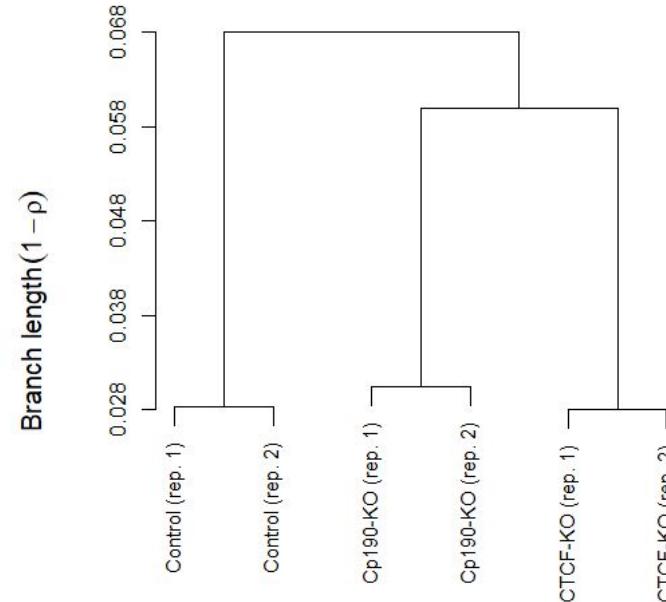
C

Hi-C experiment clustering



Original

Reconstructed
Reproduction of Hi-C experiment clustering



Reconstructed

Fig 1C

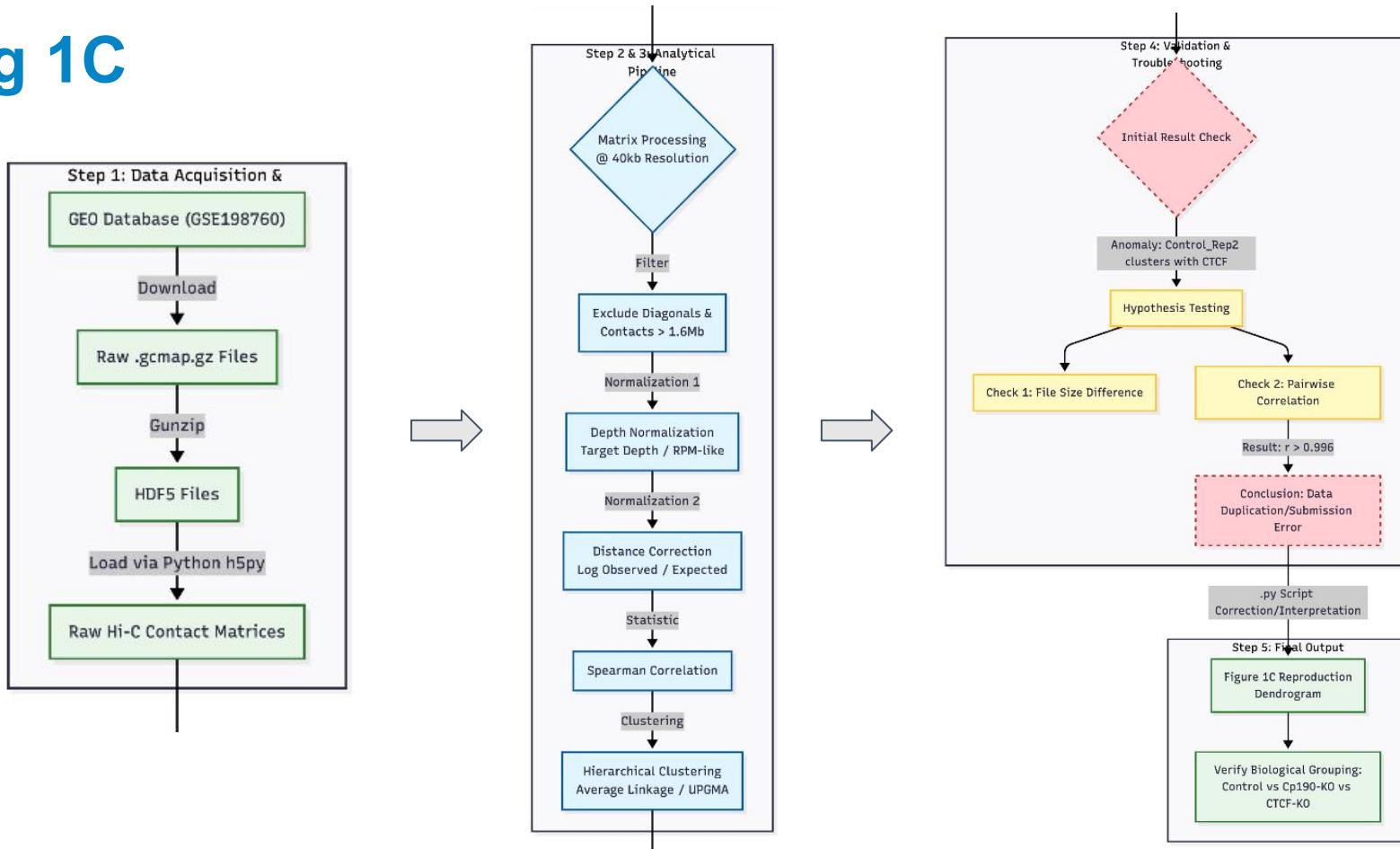
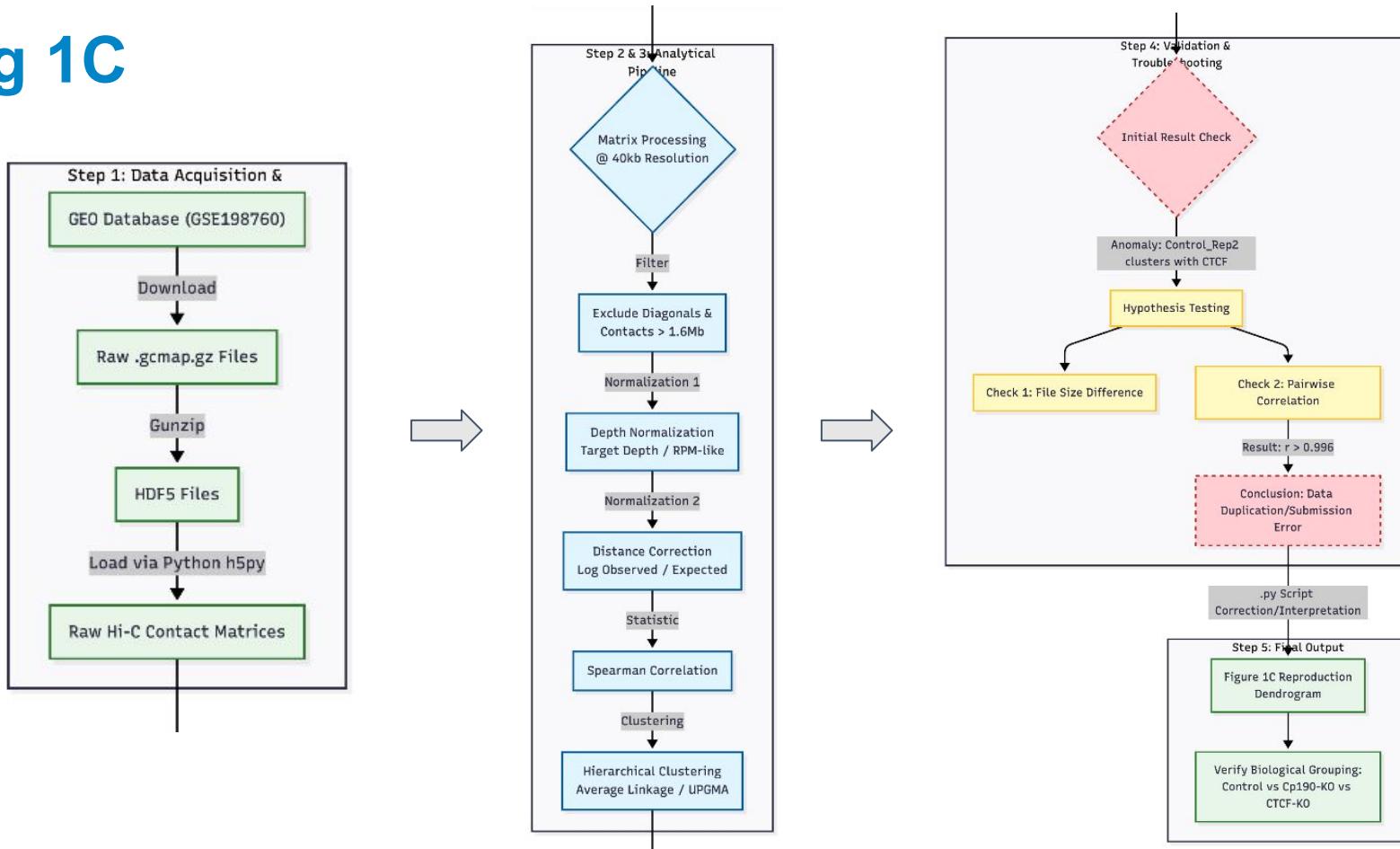


Fig 1C



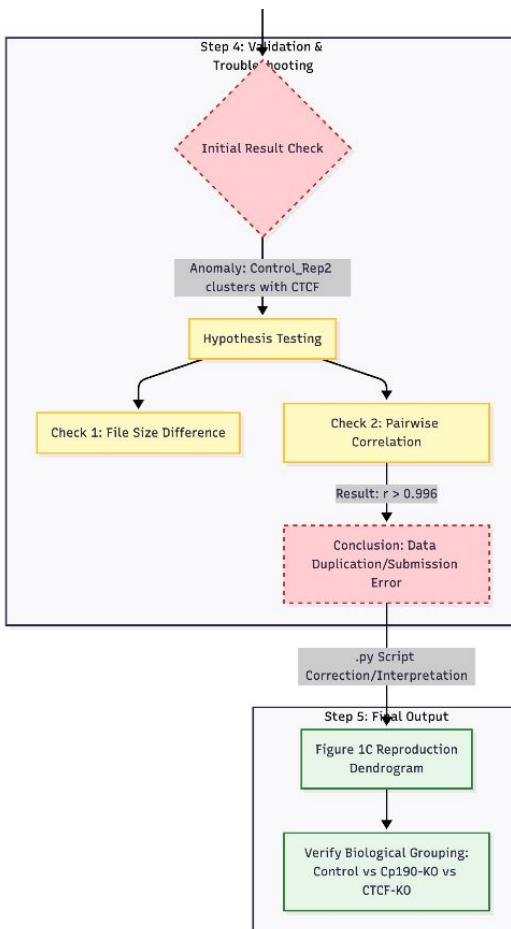
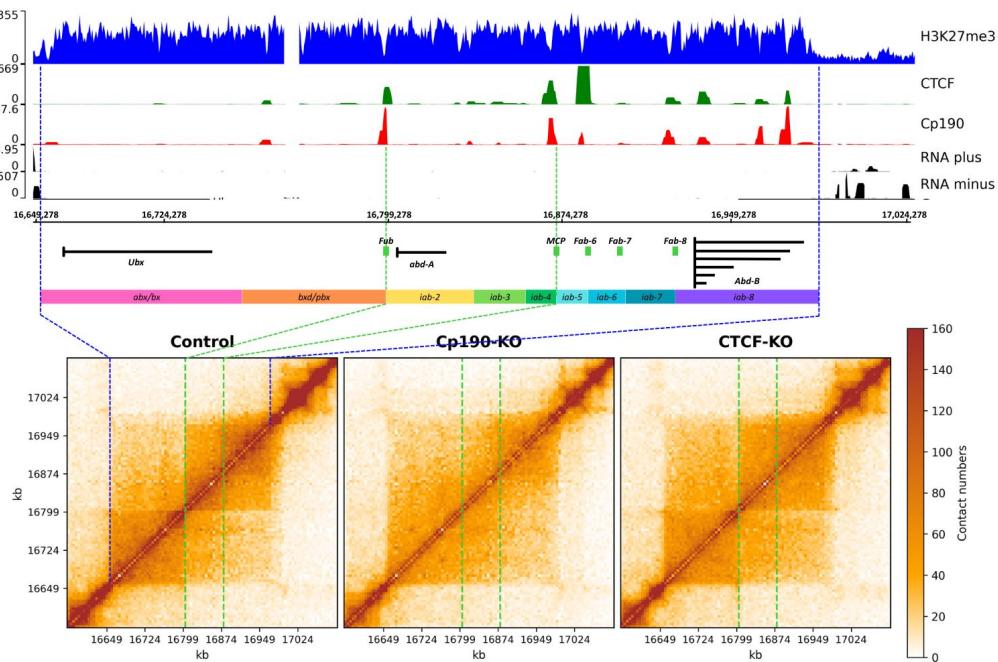
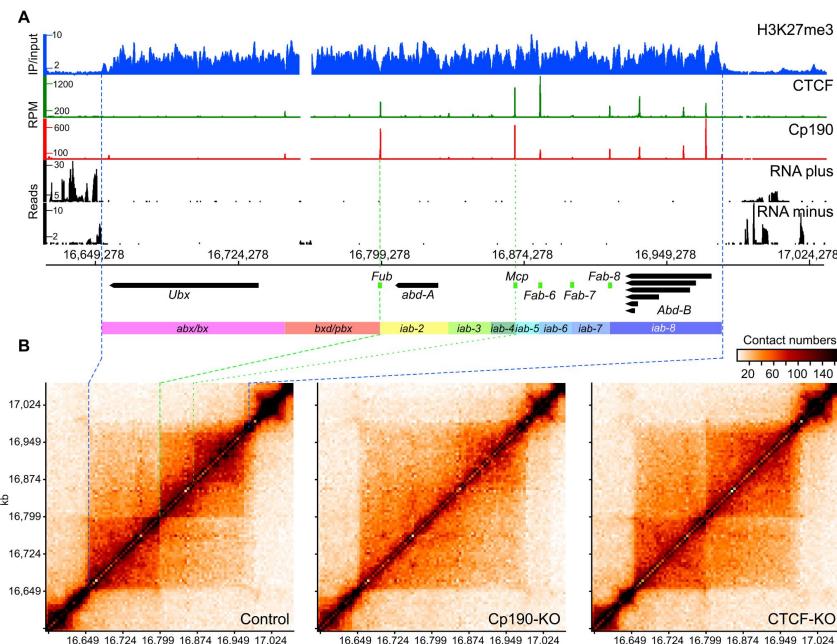


Fig 1C : Troubleshooting

```
(hic_analysis) brian@Brians-MacBook-Air-9 GSE198760_RAW % ls -lh *.gctmap
-rwx----- 1 brian staff 129M Mar 16 2022 GSM5956402_RAS3_rep1_05k_dm6.gctmap
-rwx----- 1 brian staff 156M Mar 17 2022 GSM5956403_RAS3_rep2_05k_dm6.gctmap
-rwx----- 1 brian staff 131M Mar 17 2022 GSM5956404_CPR6_rep1_05k_dm6.gctmap
-rwx----- 1 brian staff 139M Mar 17 2022 GSM5956405_CPR6_rep2_05k_dm6.gctmap
-rwx----- 1 brian staff 140M Mar 17 2022 GSM5956406_CTCF_rep1_05k_dm6.gctmap
-rwx----- 1 brian staff 143M Mar 17 2022 GSM5956407_CTCF_rep2_05k_dm6.gctmap
```



Fig 2A



Original

Reconstructed

Fig 2A

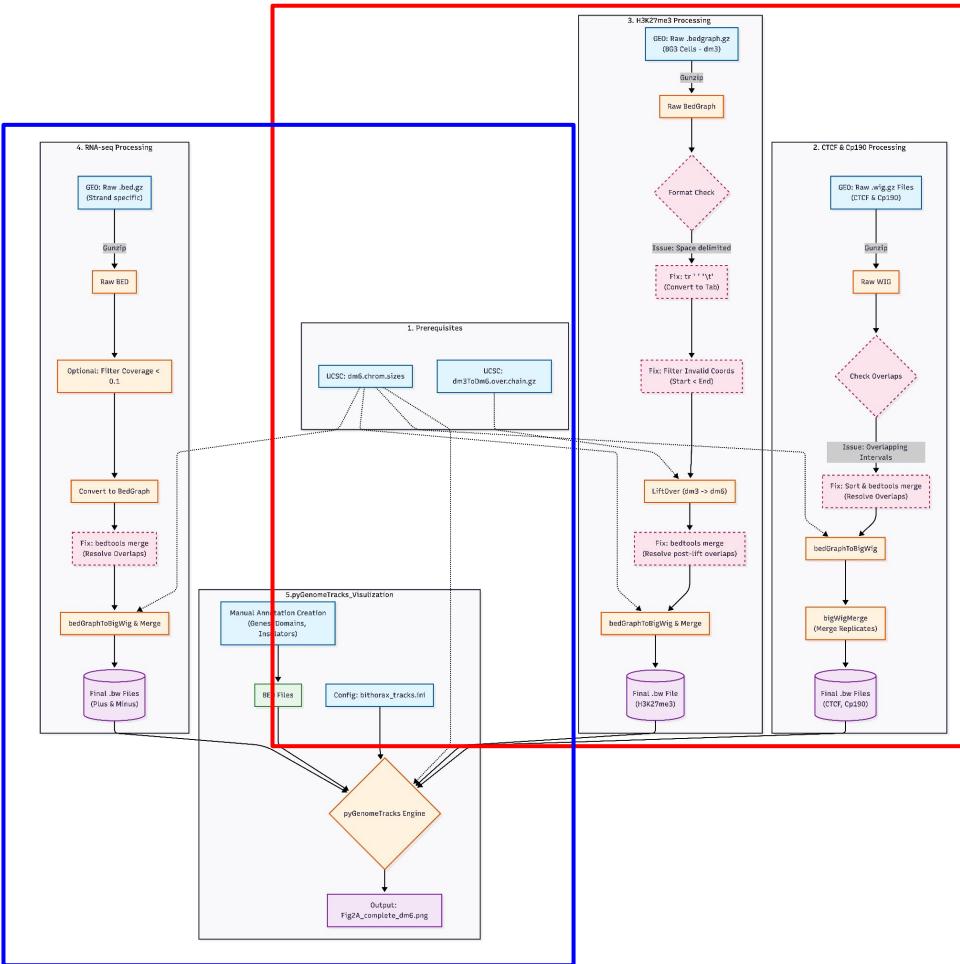


Fig 2A

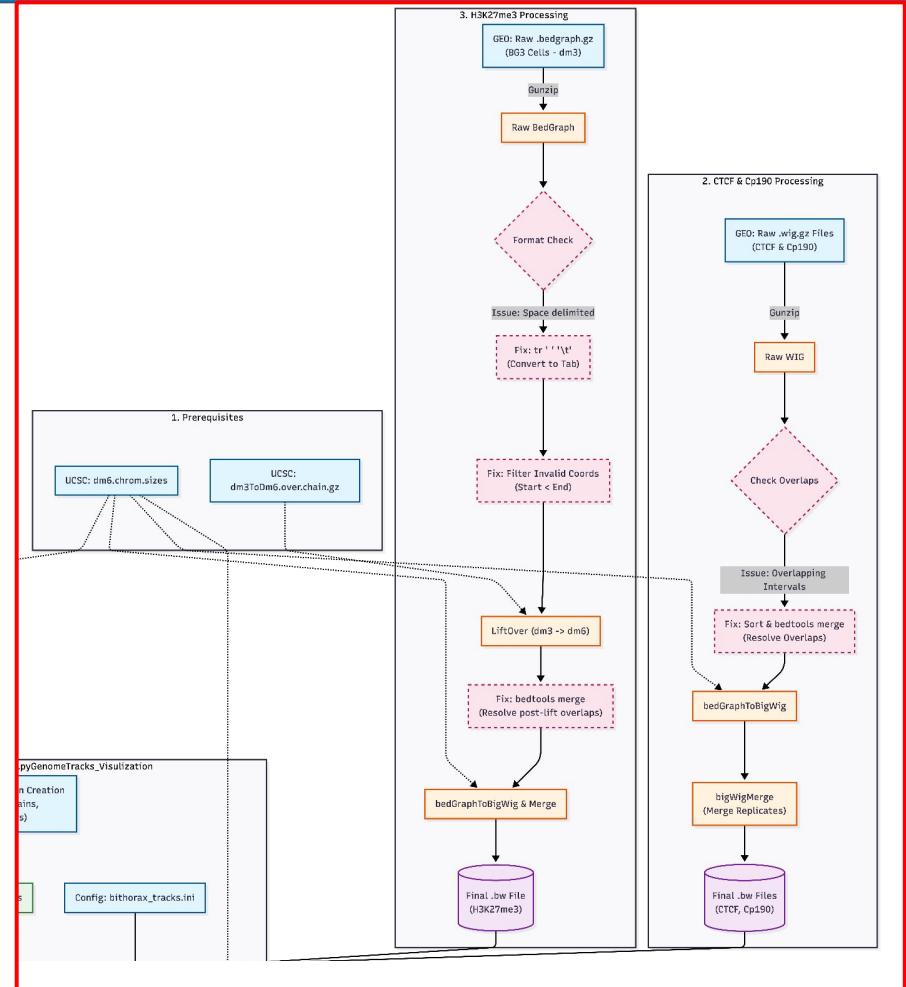
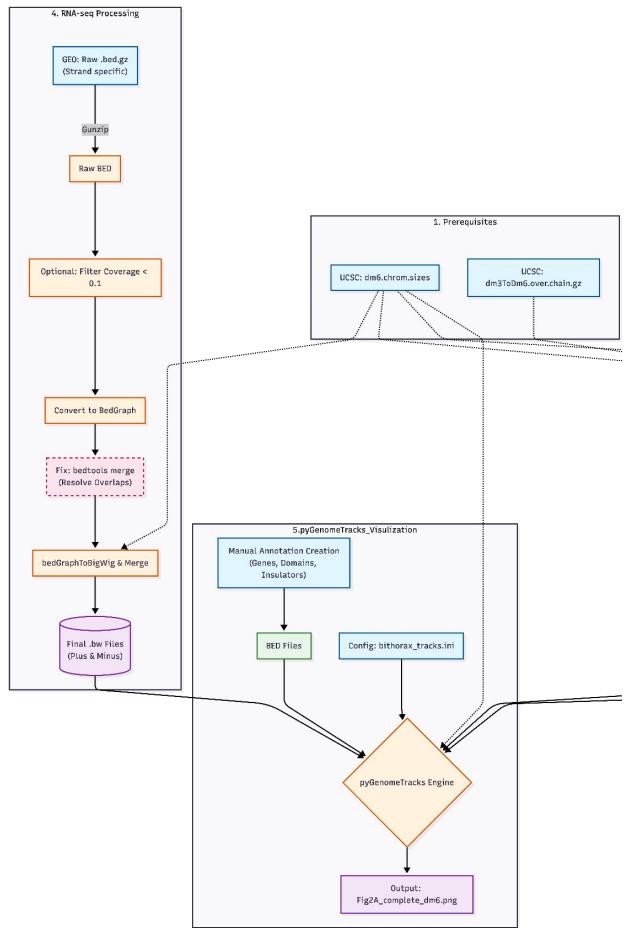


Fig 2A

環境與語言 (Environment & Language)

- Python 3.9

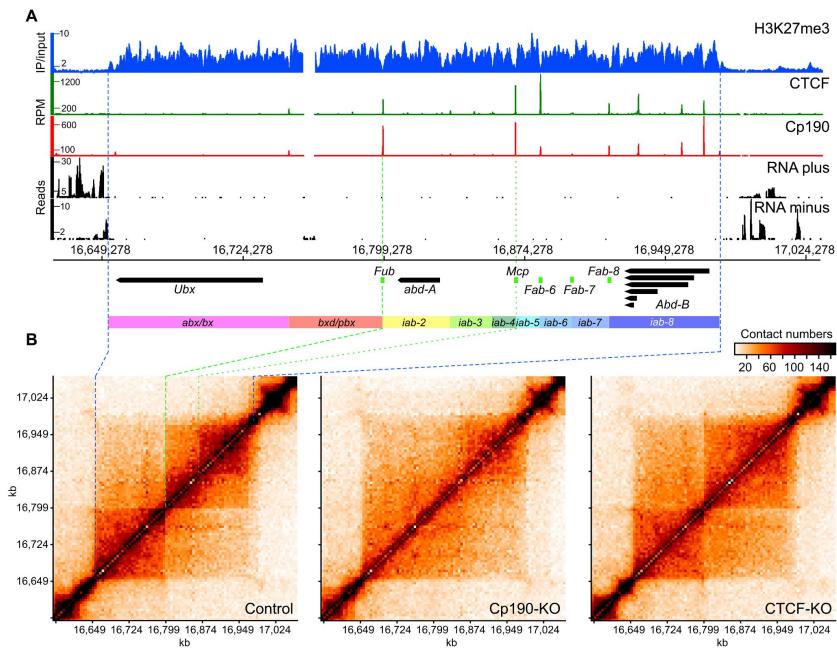
生物資訊專用工具 (Bioinformatics Tools)

- **pyGenomeTracks:**
 - 用途:核心繪圖工具。讀取設定檔(.ini) 與 bigWig/BED 檔案。
- **bedtools:**
 - 用途:處理基因體區間運算。
 - 具體指令:`bedtools merge` 用於合併重疊的訊號區間(處理wig 轉 bigWig 時的 overlap 問題)，並計算平均值。
- **UCSC Kent Utilities (UCSC 工具組):**
 - `bedGraphToBigWig`: 將文字格式的bedGraph 轉換為二進位索引格式bigWig (.bw), 供繪圖軟體快速讀取。
 - `bigWigMerge`: 將兩個重複樣本的bigWig 檔案合併為一個。
 - `liftOver`: 將 H3K27me3 數據從 `dm3` 轉換為 `dm6`。

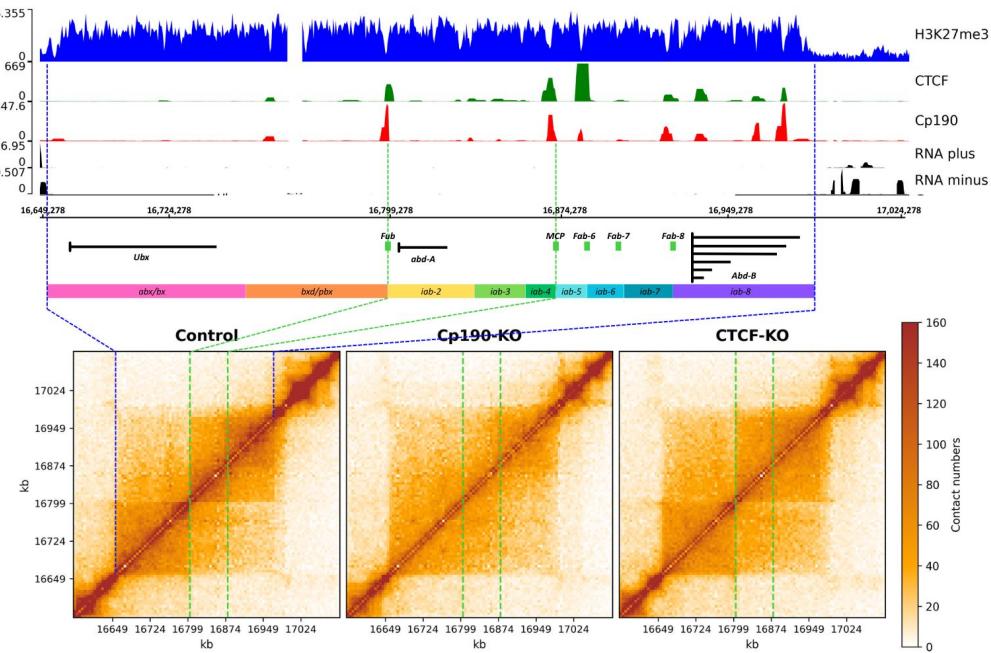
資料檔案格式 (Data Formats Processed)

- **WIG (.wig):** 原始下載數據。
- **bedGraph:** 中間處理格式，方便運算。
- **BigWig (.bw):** 最終繪圖格式。
- **BED (.bed):** 用於定義基因位置、絕緣子等註釋資訊。
- **Chain file (.chain):** 用於 LiftOver 座標轉換。

Fig 2B



Original



Reconstructed

Fig 2B - gcMapExplorer

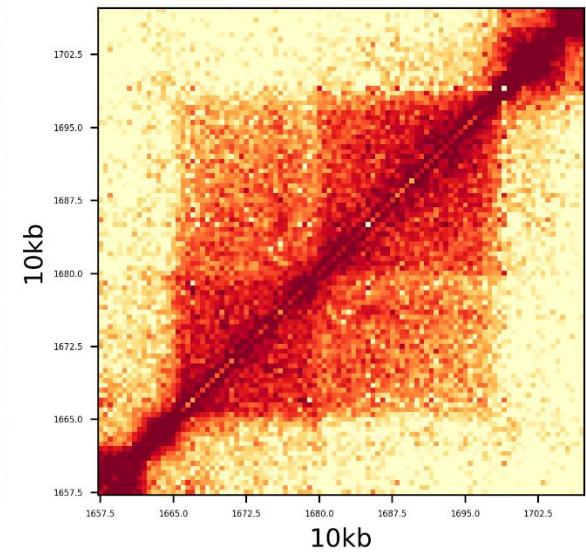
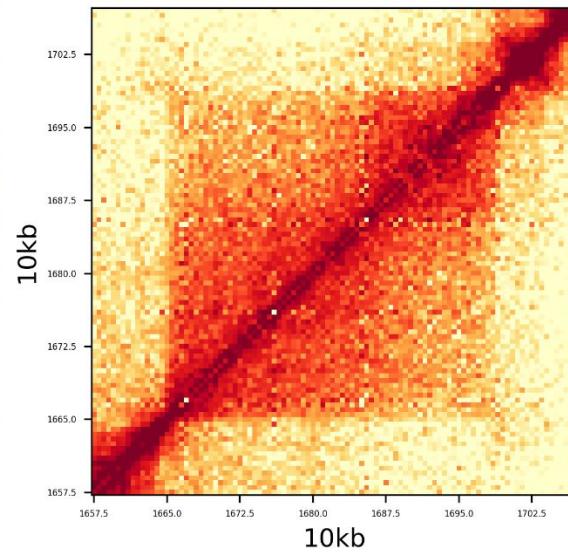
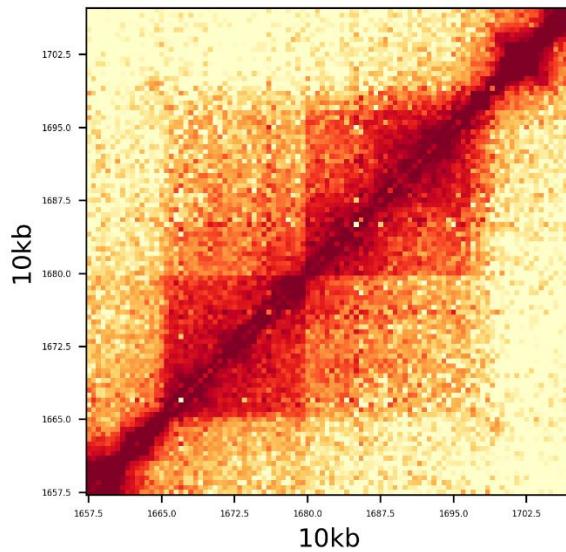


Fig 2B - gcMapExplorer

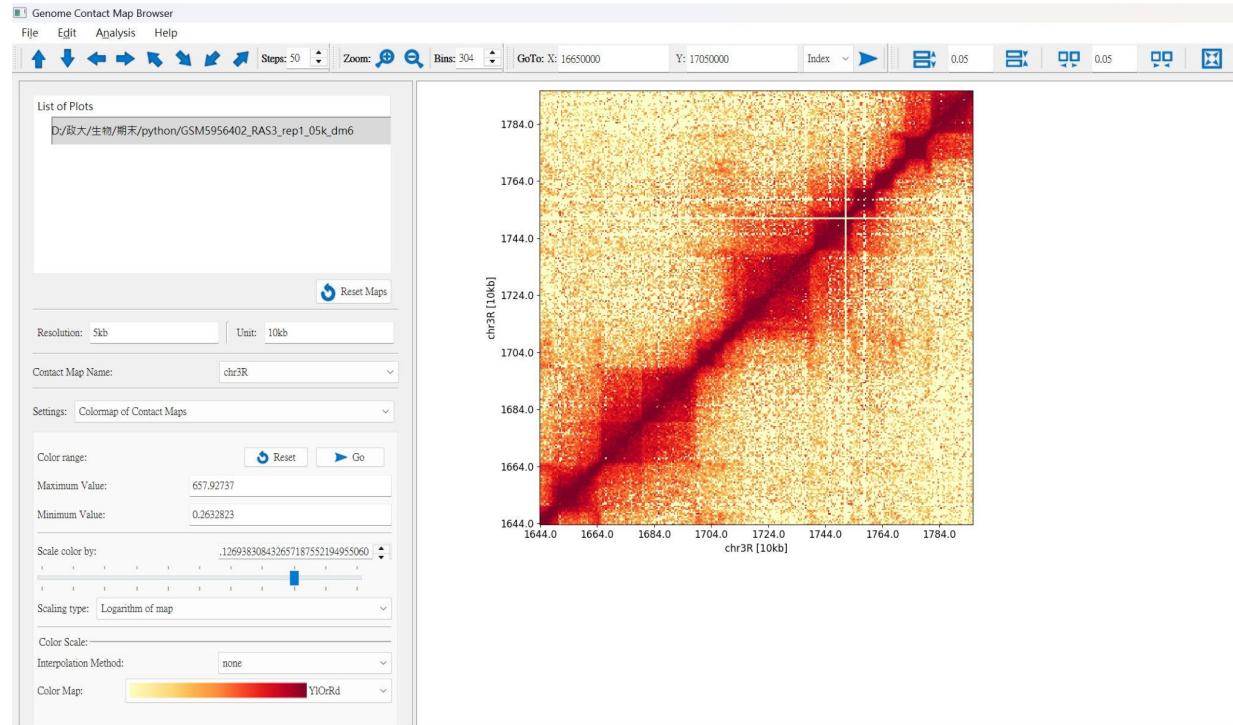


Fig 2B

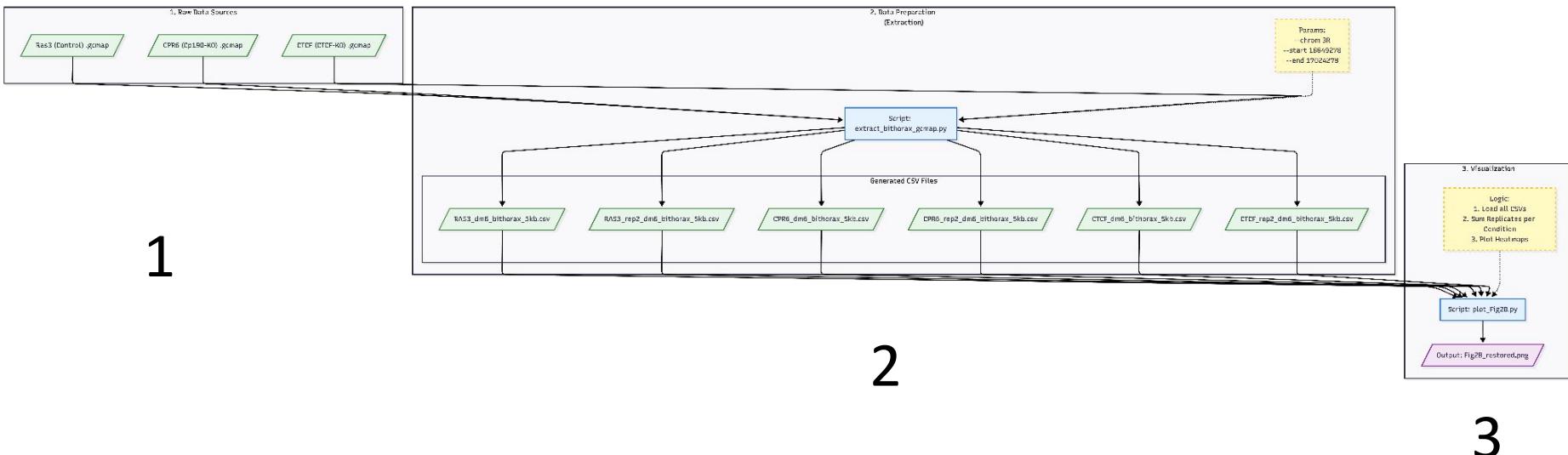


Fig 2B

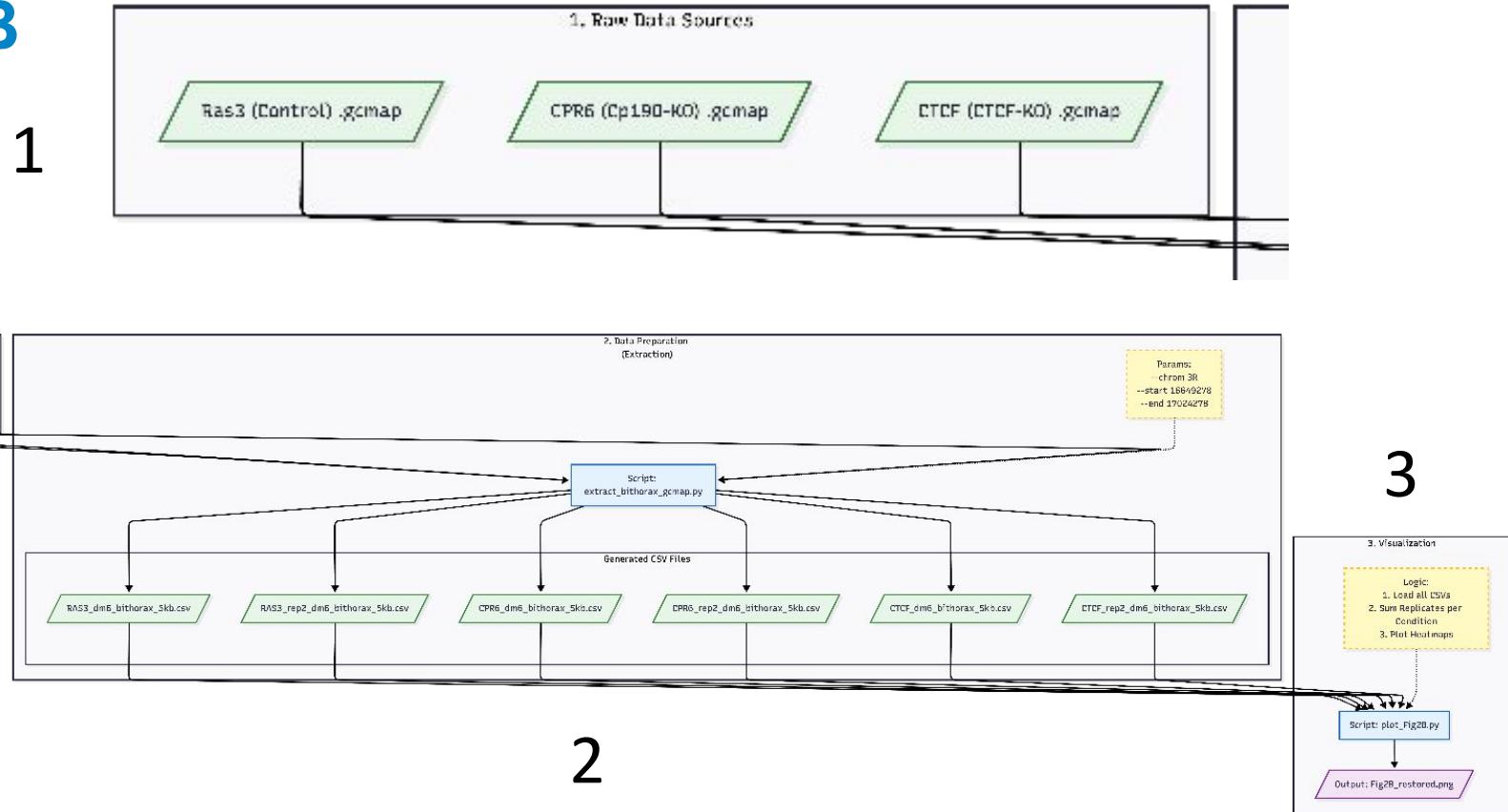


Fig 2B

程式語言與環境 (Programming Language & Environment)

- Python 3.10+

Python 函式庫 (Python Libraries)

- **h5py**: 用於讀取原始的 .gcmap 檔案(其底層為 HDF5 格式)。
- **pandas**: 用於處理結構化的矩陣數據，以及讀寫提取後的 .csv 檔案。
- **numpy**: 用於高效的數值矩陣運算(如合併重複樣本時的矩陣相加)。
- **matplotlib**: 視覺化工具，用於將接觸矩陣繪製成 Hi-C 熱圖 (Heatmaps) 並輸出最終圖片。

Fig 6B

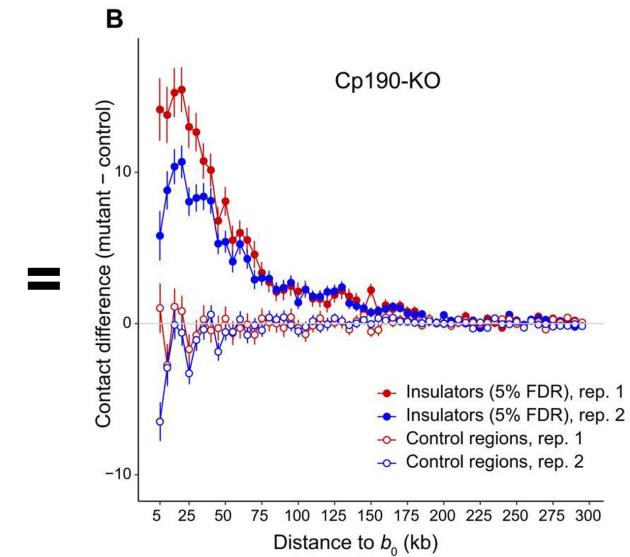
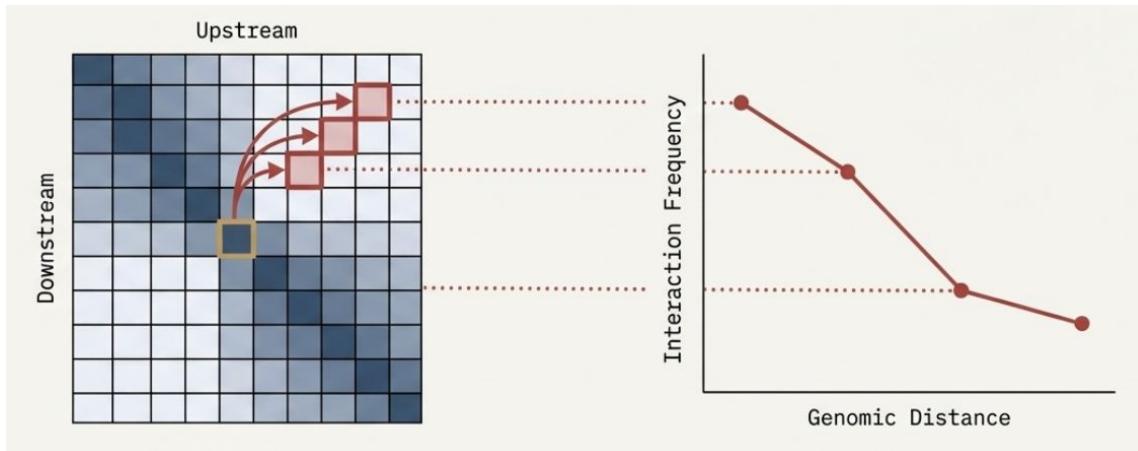
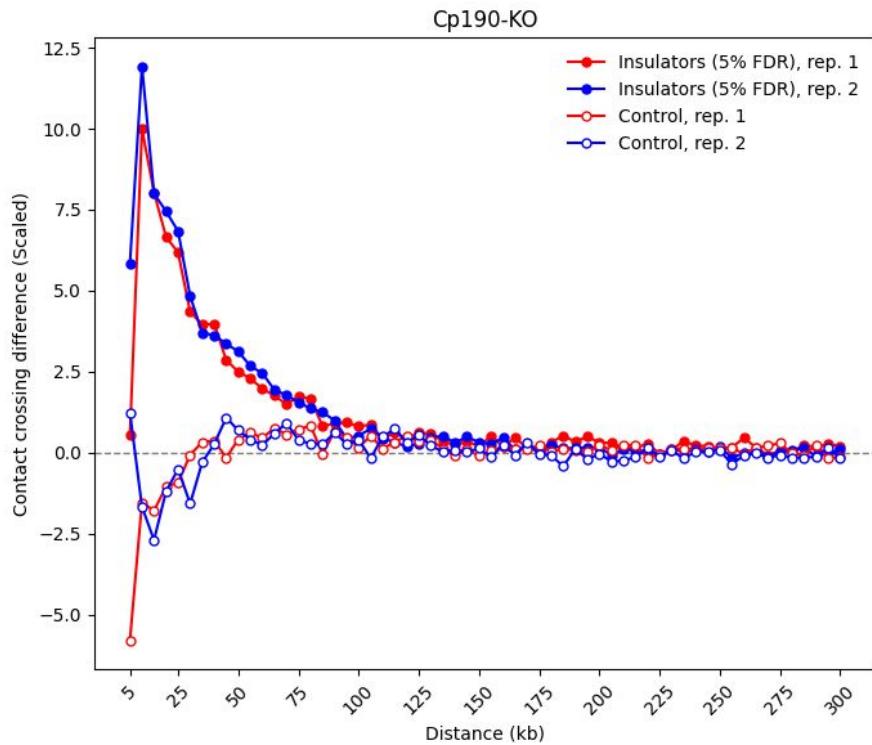


Fig 6B

Reconstructed



Original

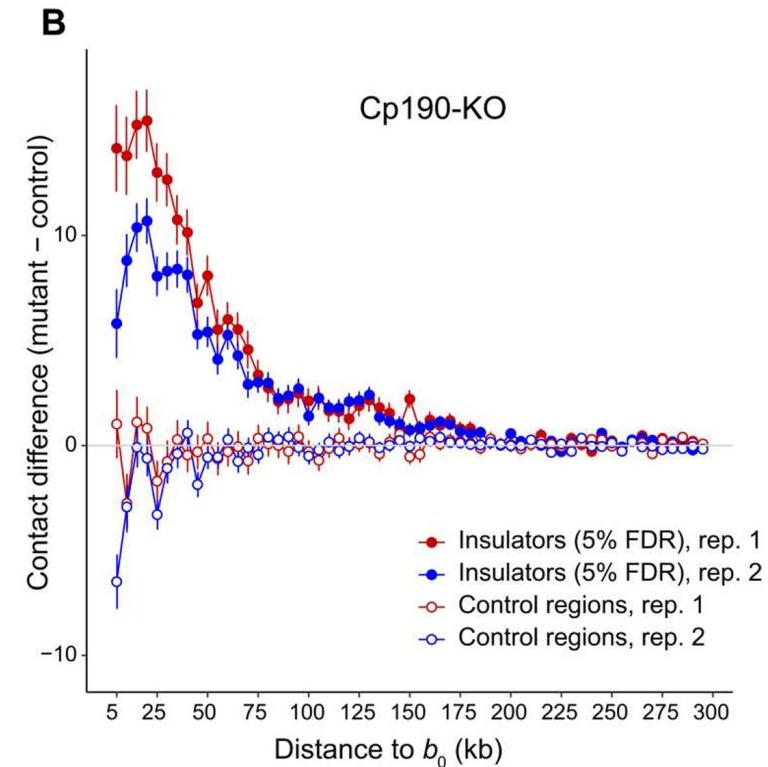
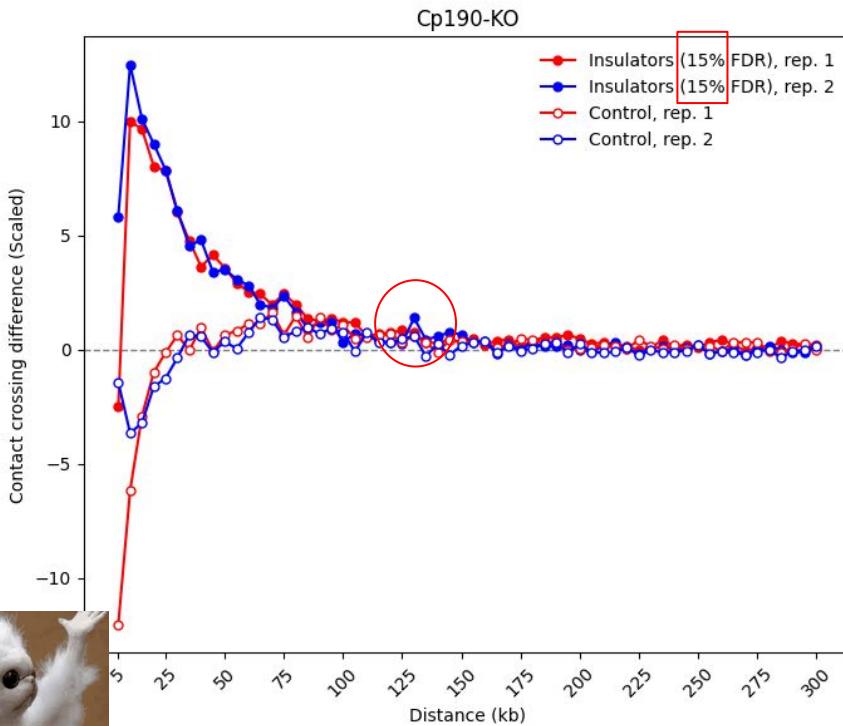
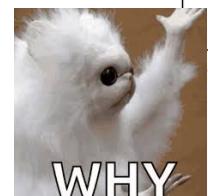
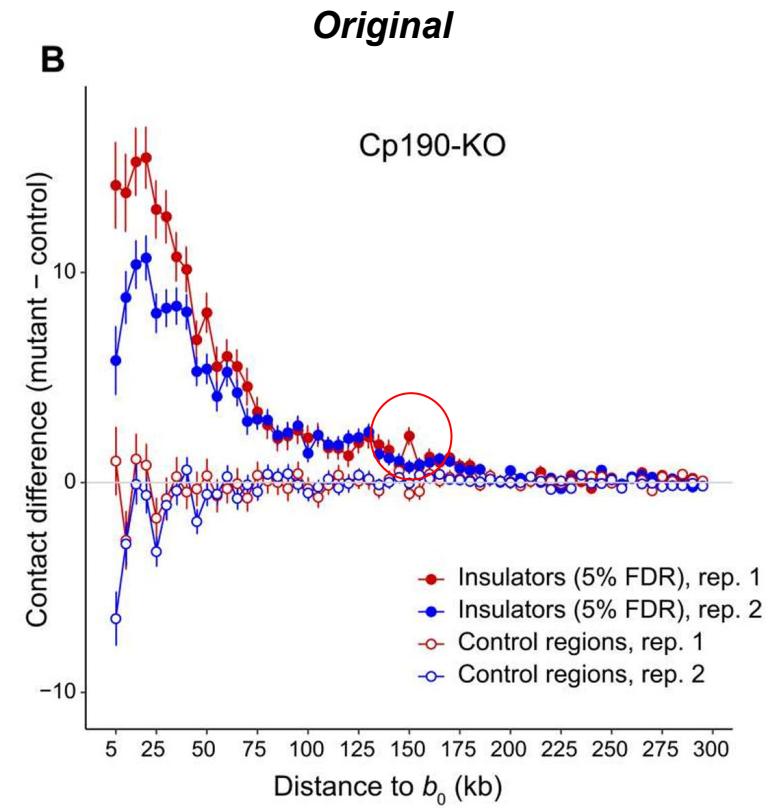


Fig 6B

Reconstructed



B



WHY

Figure S10

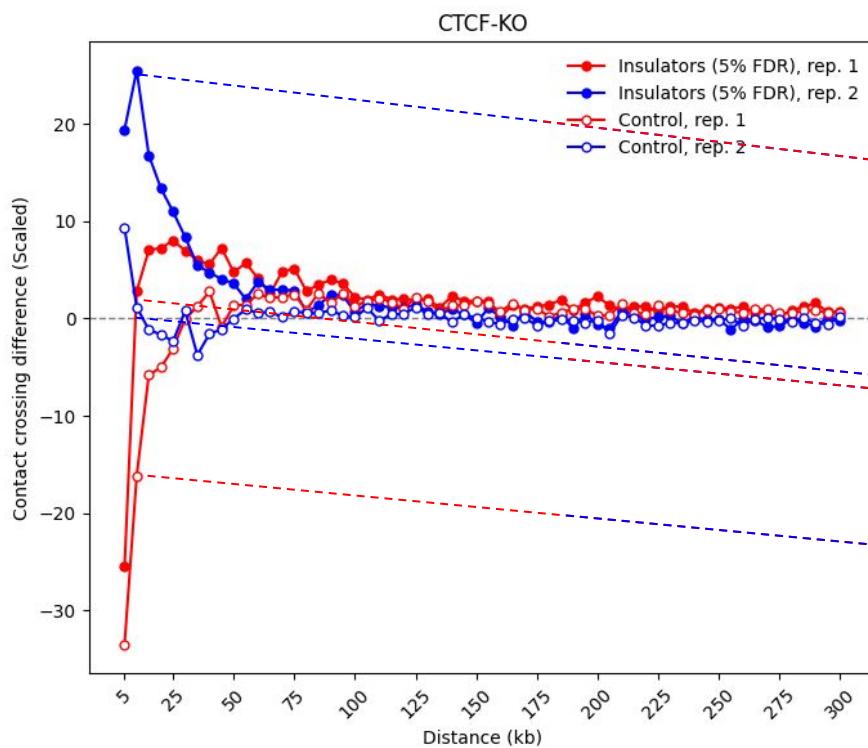
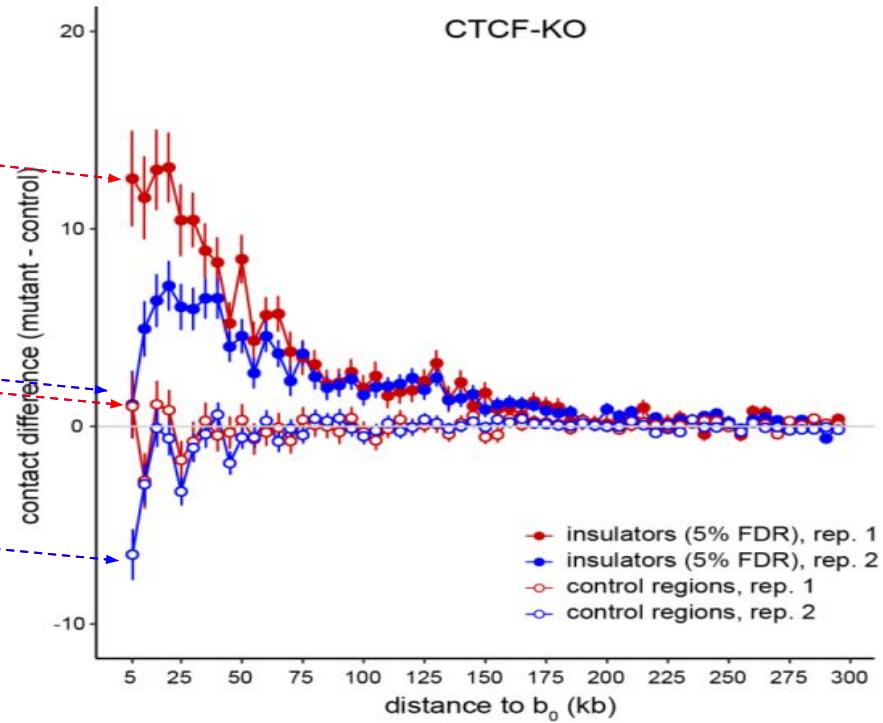
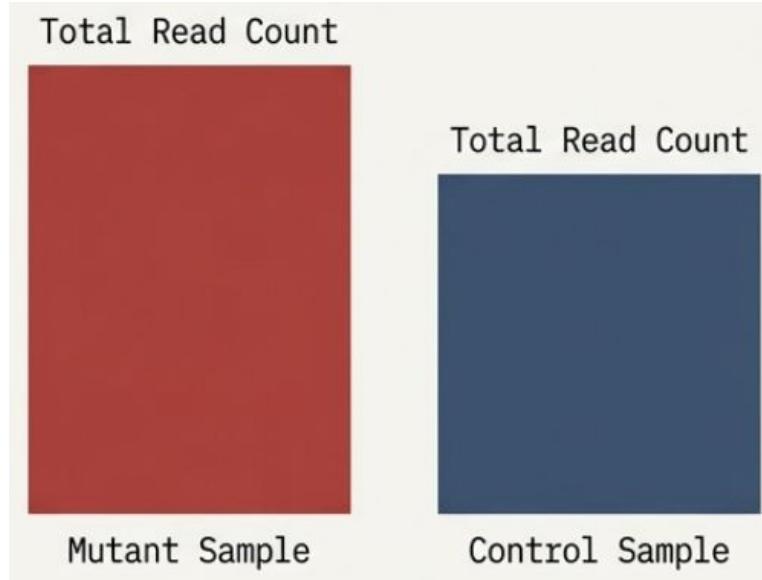
Fig 6B*Reconstructed**Original*

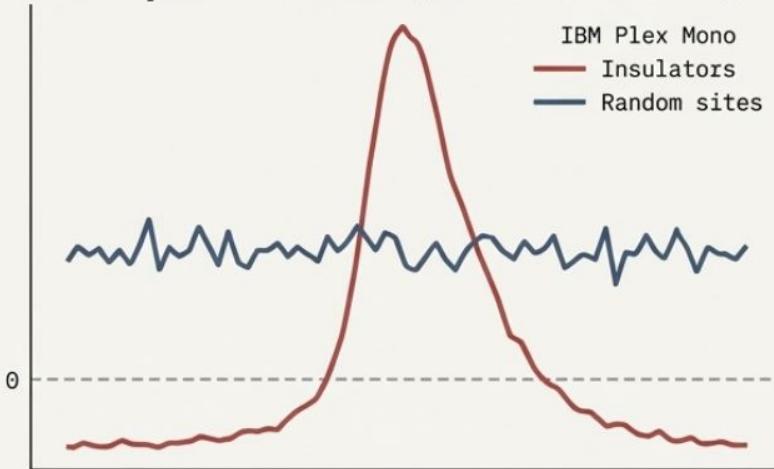
Fig 6B



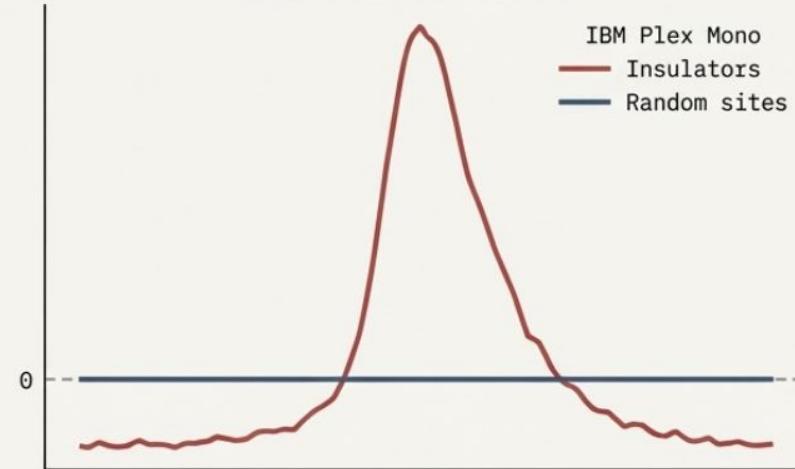
If the Control sample was sequenced less deeply, all its contact value will be systematically lower. We need to correct for this background noise.

Fig 6B

Simple Difference (Mutant - Control)



Normalized Difference



After normalization, the background signal at random sites is correctly centered at zero. The peak at insulator sites is now a high-confidence biological signal, not a technical artifact.

Fig 6B

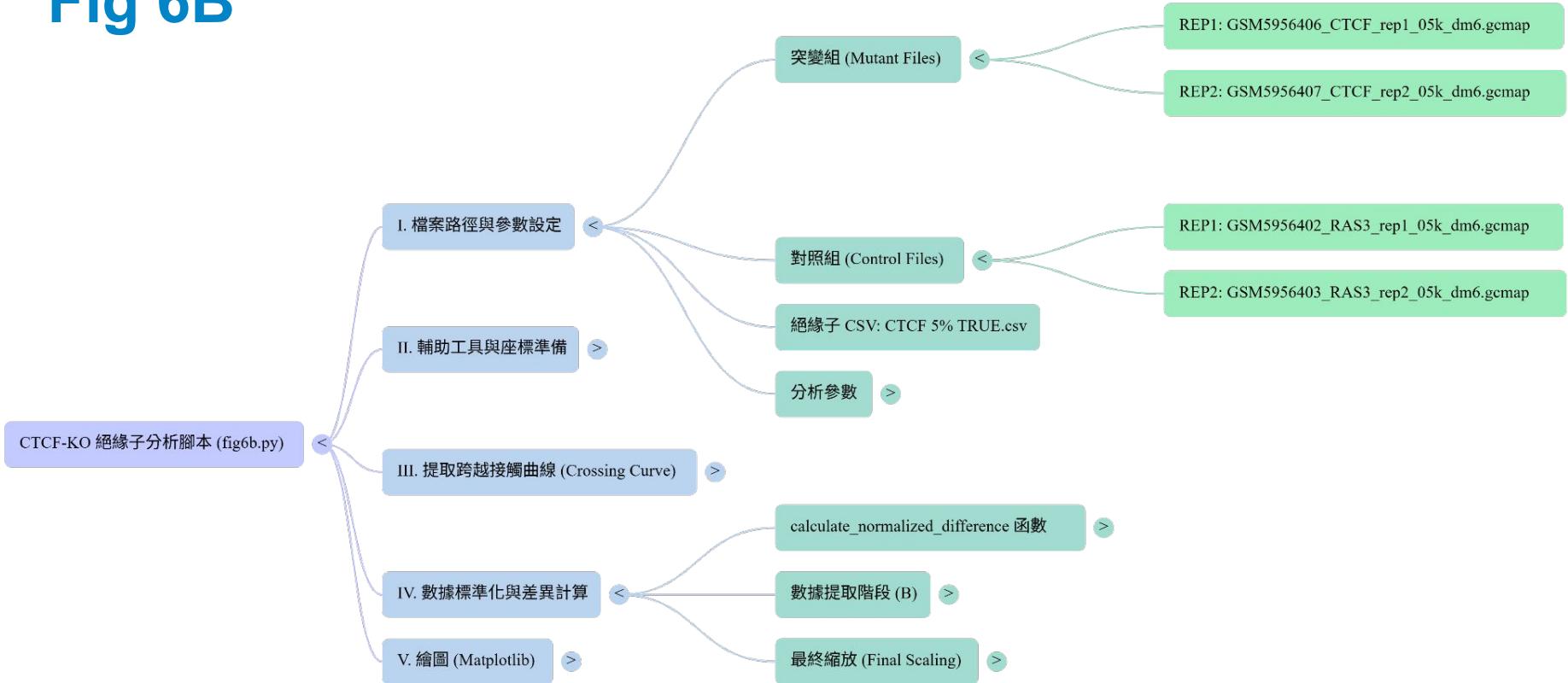


Fig 6B

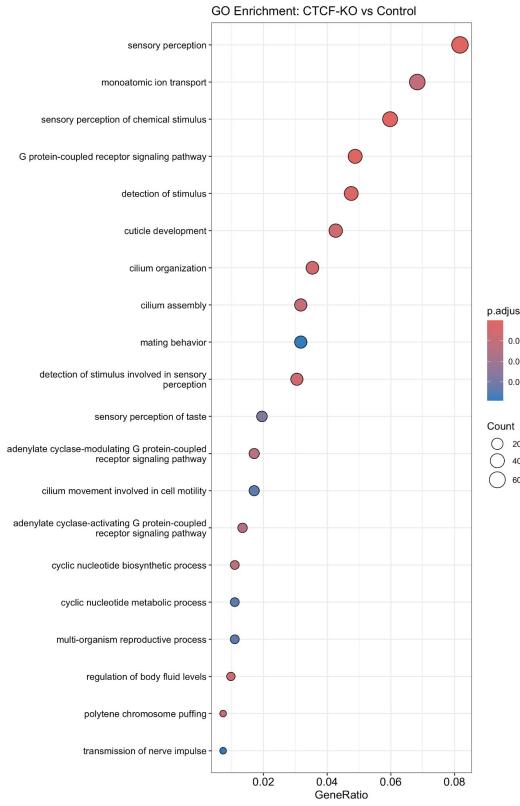
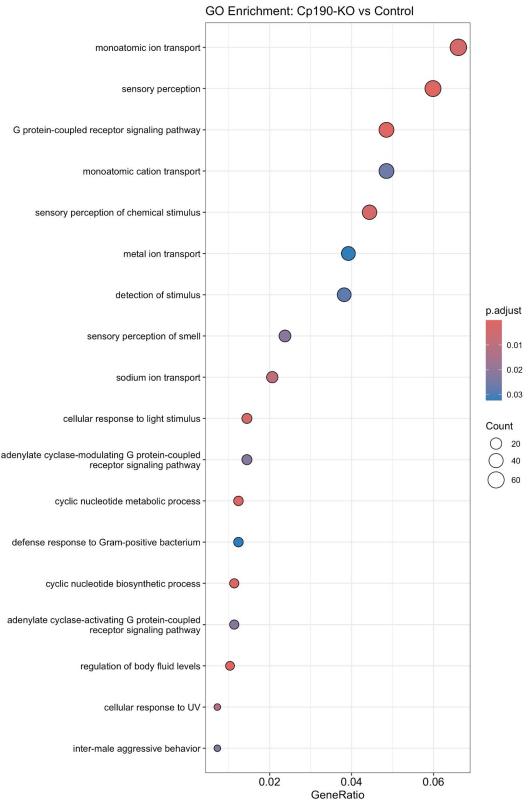
程式語言與環境 (Programming Language & Environment)

- Python 3.7.16

Python 函式庫 (Python Libraries)

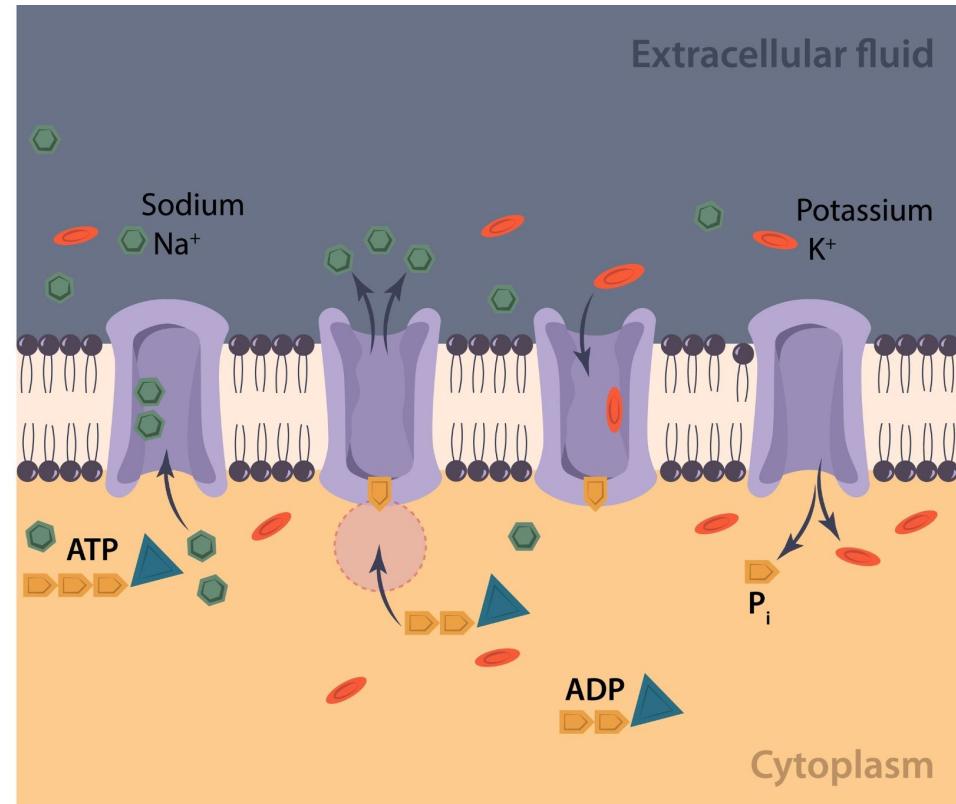
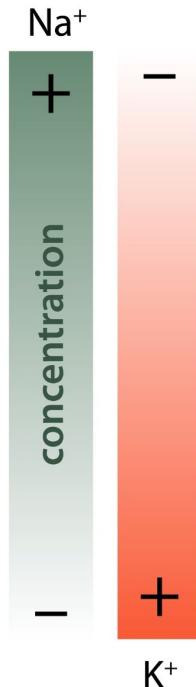
- **h5py**: 用於讀取原始的 `.gctmap` 檔案(其底層為 HDF5 格式), 用於提取染色體特定區域的矩陣數據。
- **pandas**: 用於處理結構化的矩陣數據, 以及讀寫提取後的 `.csv` 檔案。
- **numpy**: 用於高效的數值矩陣運算(如合併重複樣本時的矩陣相加)。
- **matplotlib**: 視覺化工具, 用於將接觸矩陣繪製成 折線圖設定 X/Y 軸標籤、圖例並輸出最終圖片。

Gene Ontology Analysis

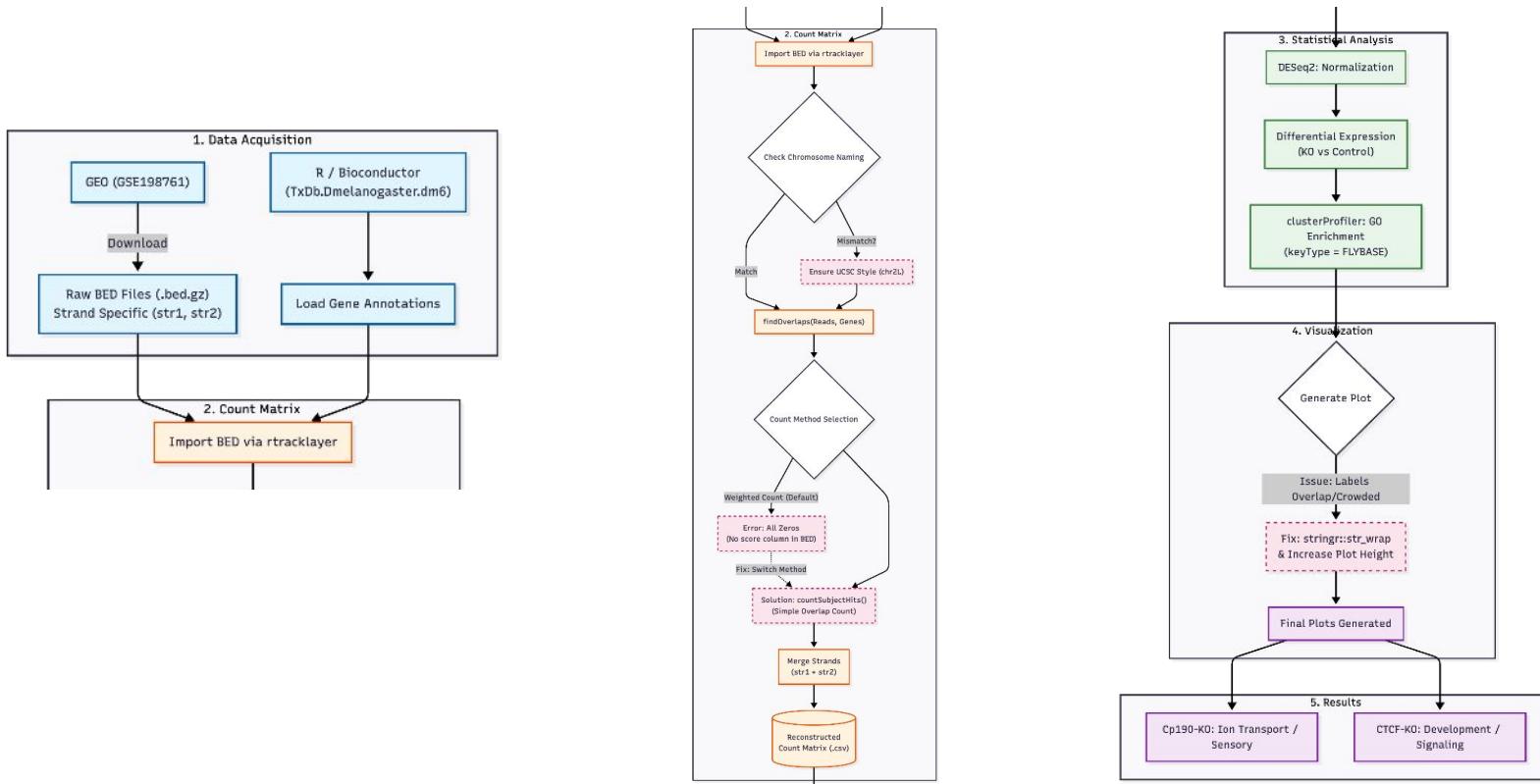


Gene Ontology Analysis

Monoatomic ion transport



Gene Ontology Analysis



Gene Ontology Analysis

1. 分析環境與語言 (Environment & Language)

- R

2. R 套件 - 數據處理與註釋 (Data Processing & Annotation Packages)

- **BiocManager**: 安裝與管理 Bioconductor 生物資訊相關套件。
- **rtracklayer**: 負責匯入/讀取原始的 BED 格式檔案 (`import` 指令), 這是將非標準格式數據載入 R 的第一步。
- **GenomicRanges & GenomicFeatures**: 處理基因體區間 (Genomic Intervals) 的核心套件, 支援尋找 Reads 與基因的重疊位置。
- **TxDb.Dmelanogaster.UCSC.dm6.ensGene**: 提供果蠅 dm6 版本的基因註釋資料庫。用來定義基因在染色體上的具體位置, 以便進行計數。
- **org.Dm.eg.db**: 果蠅的生物體層級註釋資料庫。用於將基因 ID (FlyBase ID) 轉換並對應到生物功能 (GO terms)。

Gene Ontology Analysis

3. R 套件 - 統計分析與富集 (Statistical Analysis & Enrichment)

- **DESeq2**: 進行差異表現分析 (Differential Expression Analysis)。負責對原始計數矩陣進行正規化 (Normalization)，並計算 Cp190-KO/CTCF-KO 與對照組之間的差異顯著性。
- **clusterProfiler**: 執行 GO 富集分析。計算差異基因在特定生物途徑 (Biological Process) 中是否顯著富集。

4. R 套件 - 視覺化與字串處理 (Visualization & Utilities)

- **ggplot2**: 繪製富集分析結果的氣泡圖。
- **stringr**: 字串處理工具。在此流程中特別用於 `str_wrap` 函式，將圖表中過長的 GO 途徑名稱自動換行，解決標籤遮擋座標軸的問題。

5. 關鍵函式與指令 (Key Functions & Commands)

- **findOverlaps**: 計算 BED 檔案中的 Reads 與基因註釋區域之間的重疊。
- **countSubjectHits**: 統計重疊次數。這是本流程中解決「沒有標準 Count Matrix」問題的關鍵步驟，將 BED 檔轉換為可分析的計數數據。

Thanks for Listening

功德圓滿

