問都給你問

1. 除了complete linkage，你是否考慮使用其他分群方法？若有，可以談談你為什麼最終選擇了complete linkage 作為我們最終的分群方法，以及為什麼最終選擇這個方法而不是其他方法。

- 在我們的研究初期，我們使用了single linkage階層式分群法，而後也嘗試了K-means作為分群方式。然而，最後我們選擇了complete linkage作為最終的分群方法，主要原因在於 complete linkage 能提供更關聯性的結果，這有助於更全面地理解疾病間的相互關係。

2. 為什麼你最後選擇使用complete linkage 這種分群方法？有特定的理由或前人研究支持這個決定嗎？

- 相對於single linkage，complete linkage和average linkage方法更能考慮各疾病之間的相似性，有助於更深入地瞭解疾病之間的關聯性。我們進行了complete linkage 和average linkage的比較，發現complete linkage 所分類出的結果有助於呈現疾病之間的相關性。

3. 為什麼你採用分群方式來分類疾病，而不是其他方法，比如監督學習分類模型？這個方法在這個研究中的優勢是什麼？

- 我們選擇使用分群方式來分類疾病，主要是為了探索不同疾病之間的共病關聯性。我們將不同疾病的發病情況視為特徵，通過計算它們之間的相似度，以得知它們之間的相關性。利用由相關性轉換的距離矩陣進行階層式分群，有助於將相似的疾病分為同一組，從而提供對不同疾病之間相互關係的更深入理解。

4. 你選擇分成五個群組，有特定的理由嗎？是否有考慮過其他不同數量的群組，並為什麼最終選擇五個？

- 分成五個群組的主要原因是基於我們的初始分群結果，當使用single linkage分群方法時，我們最初分成七個群組。因此，在改用目前的方式時，我們也從七個群組的基礎上開始，然後根據實際情況進行了調整，最終分為五個群組，這在結果呈現和人數分布方面最為合適。如果分為六或七個群組，那麼結果將是以分為五個群組時的結果為基礎，進一步將人數稀少的群組再細分，這樣不會為我們的分析提供實質幫助。

5. 你提到有使用Kmeans，為什麼最終不採用這個方法？有特別的原因或優勢，讓你選擇complete linkage 分群？

- 我們確實嘗試了K-means作為分群方法，但最終未選擇它的原因在於 K-means 在初始化時需要指定分群數量，而階層式分群方法不需要這樣的指定，能夠根據結果自動調整分群數量。我們的目標是為各群找到具有相關主題或高相似性的疾病，所以最終選擇了階層式分群。

6. 差異甲基化分析與內部控制基因位點中需要的低差異，是否矛盾，兩者的差異是否一樣
    - 不一樣
    - 差異甲基化分析的主要目的是比較不同條件或組別間特定基因位點的甲基化差異。**通過統計分析，我們可以過濾出在統計上有顯著差異的基因位點**。
    - 而內部控制基因位點的特點是，無論在**何種時間、空間或實驗方法下，它都應該展示出高度的一致性和穩定性**。這意味著這些位點的甲基化狀態**在各種條件下都是相對固定的**，因此它們可以作為一個基准或參考。
7. 過濾後的DMPs已有顯著的差異性，為什麼還適合最為候選內部控制基因位點的參考
    - 首先，我們遵循了**標準的差異甲基化分析通用流程**。這意味著我們的分析方法是根據**已經公認且被廣泛應用的標準方法**

進行。這不僅確保了我們的結果的可靠性，而且意味著**我們的結果可以與其他使用相同方法的研究進行比較**。

- 我們在**多次重覆的實驗中**觀察到這些DMPs都展示了**高度的穩定性**。這種一致性為我們提供了信心，認為這些DMPs可以作為DNA甲基化實驗的內部控制。畢竟，**內部控制的主要特性就是它們應該在不同的實驗或條件下保持相對的穩定性**。
- 更重要的是，我們觀察到這些DMPs在**每次實驗中的表達量都是相對穩定的**。這確保了這些DMPs的甲基化水平不會因為**實驗中的微小變化而受到太大的影響**，**從而提供了一個穩定的基線**，用於比較其他可能的變異性或差異性。
- 結合這三點，我們認為都符合理想內部控制所需要的特性

8. 為什麼用Accuracy作為評判標準
    - 在資料集當中，我們所挑選的疾病發生率幾乎是常態分佈，這意味著**罹患**疾病和**未罹患**疾病的分布相對均勻。在這樣的情境下，Accuracy是一個合理且有代表性的指標，因為模型的隨機猜測不太可能獲得很高的Accuracy。
    - **如果他們問起為什麼沒有使用其他指標作為評判標準 (ex. Precision, Recall, F1 Score)**
        - 雖然Precision、Recall 和F1 Score都是非常有價值的指標，尤其當考慮模型在特定情境下的表現或針對特定類型的錯誤時，但在我們的情境中，我們主要關心的是利用挑選的內部控制基因特徵，整體上對於疾病的預測能力如何。我們主要關注模型能夠正確預測疾病的整體機率，而不是深入到特定的預測錯誤。因此，Accuracy成為了一個直觀且合適的指標，能夠最佳地反映我們研究的核心目標。
        - 我們希望結果報告能夠簡單、直觀地傳達給大眾，**而Accuracy是一個容易理解的指標**。它直接告訴我們模型

正確預測的百分比，對於非專業背景的聽眾來說，這是一個很直觀的描述。

- 我們在此研究中主要使用Accuracy作為評估指標的根本原因在於研究的核心目的。**雖然Precision, Recall, 和 F1 Score都是非常有價值的指標**，尤其當**考慮到模型在特定情境下的表現或是針對特定類型的錯誤**時。但在我們的情境中，我們特別關心的是利用挑選的內部控制基因特徵，整體上對於疾病的預測能力如何。換句話說，我們**主要關注的是模型能夠正確預測疾病的整體機率，而不是深入到特定的預測錯誤**。因此，Accuracy成為了一個直觀且合適的指標，**能夠最佳地反映我們研究的核心目標。**

9. 為什麼你選擇使用SVM (支持向量機) 而不使用Random Forest來建立疾病預測模型？這個選擇有特定的原因或優勢嗎？你可以解釋一下你的決策背後的考慮。
    - 我們選擇使用SVM的原因是，SVM具有最大邊界的特性，特別適合於小型數據集。隨機森林則通常依賴於更大的樣本量來確保多個樹的多樣性。尤其在特徵較少的情況下，SVM通過調整正則化參數，可以更容易地避免過擬合。
    - 隨機森林雖然通常具有很好的泛化性能，但當特徵數量非常少時，它可能難以建立多棵多樣性的樹。這可能使得森林中的樹在少量的特徵上過擬合。而 SVM，通過調整正則化參數，可以更容易地避免過擬合。
    - **解釋性**: 此外，SVM提供了良好的解釋性，能夠直觀地顯示每個特徵的重要性，這在我們的研究中非常有價值。雖然隨機森林也提供特徵重要性的度量，但在特徵較少的情況下，SVM的權重通常更容易理解。
    - 總之，儘管隨機森林是一個非常強大的算法，但在特徵數量較少的情況下，SVM 可能會提供更為簡單且高效的解決方

案。但是，最好的方法仍然是在具體的數據集上試驗不同的模型，並根據實際性能來做出決策。