Your Title

Group 1

尹子維、李承祐、黃亮臻、張立勳

today

Table of contents

1 Introduction

本次報告使用的資料來自於 DepMap Portal，其屬於 DepMap Consortium (DMC)，此機構致力於加速癌症精準醫療的發展，建構了具系統性的資料集，並提供多種工具可進一步分析與視覺化。

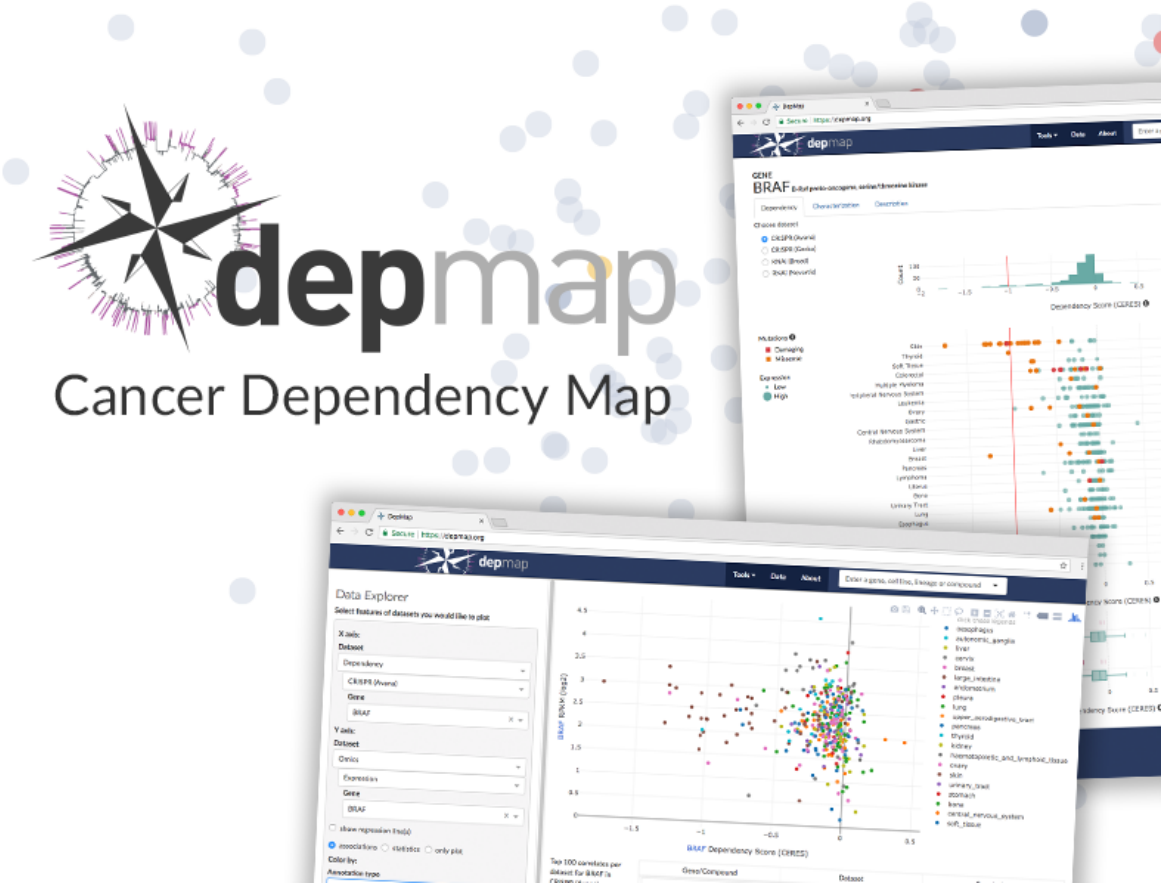


Figure 1: DepMap Portal

本次報告使用的資料可分為基因表達量、代謝體濃度、癌細胞轉移方向，分別有 1406、264、264 筆 (cell lines) 資料。資料是基於以下實驗流程取得:

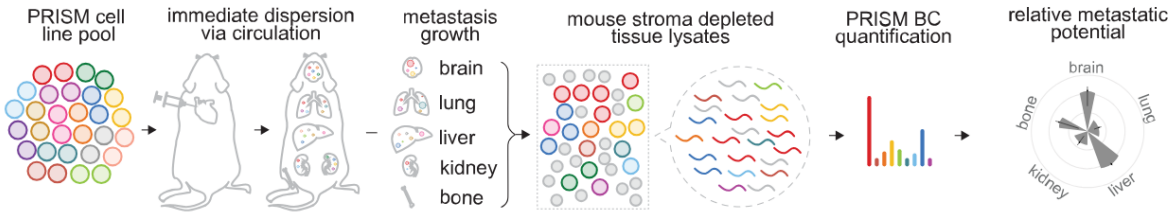


Figure 2: Metastasis Map of Human Cancer Cell Lines

在基因表達量此資料中，包含 1406 筆 cell lines、19221 個基因表達量 (經正規化後的 RNA-seq 資料)，資料整理如下:

	A	B	C	D	E	F	G	H
1		TSPAN6 (7105)	TNMD (64102)	DPM1 (8813)	SCYL3 (57147)	C1orf112 (55732)	FGR (2268)	CFH (3075)
2	ACH-001113	4.331991778	0	7.364397337	2.792855352	4.470536865	0.02856915	1.22650853
3	ACH-001289	4.566815154	0.584962501	7.106536769	2.543495883	3.504620392	0	0.18903382
4	ACH-001339	3.150559677	0	7.379031732	2.333423734	4.227278994	0.05658353	1.31034012
5	ACH-001538	5.085339669	0	7.154109243	2.545968369	3.084064265	0	5.86814348
6	ACH-000242	6.729144865	0	6.537606691	2.456806149	3.867896464	0.79908731	7.20838069
7	ACH-000708	4.272023189	0.189033824	7.022922589	2.555816155	3.841973119	0	0.0976108
8	ACH-000327	3.337711092	0	5.927185358	1.944858446	2.678071905	0.01435529	3.08915913
9	ACH-000233	0.056583528	0	6.093602429	3.970853654	3.731183242	0.02856915	6.09296851
10	ACH-000461	4.016139703	0	6.533874777	2.22650853	3.021479727	0.02856915	0.08406426
11	ACH-000705	4.411426246	0	6.412442825	2.364572432	4.275007047	0.04264434	0.20163386
12	ACH-001794	3.871843649	0.056583528	6.76394266	1.937344392	3.152183419	0.62293035	6.94462422
13	ACH-002023	5.264911693	0	6.756889855	2.283921772	3.813524689	0	4.0591822
14	ACH-000528	4.512226887	0	7.099821179	2.843983844	4.672425342	0.01435529	0.81557543
15	ACH-001655	3.592158002	0	6.747118816	0.925999419	1.839959587	0.02856915	0.05658353
16	ACH-000167	0.042644337	0	6.710117632	2.353323291	3.98550043	5.85972113	0.27500705
17	ACH-000792	3.279471296	0	6.390254956	1.748461233	3.435628594	0.08406426	0.422233
18	ACH-001098	5.909293086	0	6.783980414	3.292781749	3.087462841	0	4.89820835
19	ACH-000570	5.264911693	0	6.830483516	2.835924074	4.177917792	0.17632277	5.84046323
20	ACH-000351	3.972692654	0	8.161333468	2.350497247	4.097610797	1.60880924	0.3448285
21	ACH-000769	4.713145902	0	6.193771743	2.440952198	4.176322773	0	0.11103131

Figure 3: Overview of Genes Dataset

	Cell Lines			Genes		
	Complete	Contains 0	Total	Complete	Contains 0	Total
Number	0	1406	1406	7959	11262	19221

```
Welcome to package ztable ver 0.2.3
```

Table 1. Overview of Genes Dataset

Genes

Cell Lines

Number

0

1406

1406

7959

11262

19221

在代謝體濃度此資料中，包含 265 筆 cell lines、225 個代謝體濃度 (經 log 轉換後)，其中有 41 筆重複的 cell lines，資料整理如下:

	A	B	C	D	E	F	G	H
1	ID	X2.aminoadipate	X3.phosphoglycerate	alpha.glycerophosphate	X4.pyridoxate	aconitate	adenine	adipate
2	ACH-000007	5.6375969	5.6066161	5.4012427	5.7402012	5.926973	5.339019	6.099346
3	ACH-000013	5.6068532	5.7304864	5.7031022	6.0965971	5.579302	5.753198	5.843439
4	ACH-000019	5.8008271	5.8371813	5.8600887	6.1057	5.800354	5.753776	5.783954
5	ACH-000028	5.7034675	6.7284286	6.2191608	5.5067244	6.33621	5.448056	5.638264
6	ACH-000048	6.1231825	6.1441561	5.6746983	5.5176405	5.845675	5.802	6.052461
7	ACH-000091	5.7862107	5.8886067	6.382852	5.9960417	5.901736	5.873012	5.838974
8	ACH-000097	5.666087	6.2128096	6.2231168	6.1166083	6.021599	6.018505	5.875382
9	ACH-000117	5.8631621	5.9425659	6.0331119	6.3172582	5.862481	5.951045	5.96952
10	ACH-000123	5.8827883	5.6281984	6.1188585	5.4592921	5.552678	5.701121	5.938682
11	ACH-000132	5.9590368	5.624752	5.6923298	6.2411233	5.939079	5.216474	6.112304
12	ACH-000147	5.5957775	5.9552412	7.0280981	5.4274208	5.993546	5.751236	5.73649
13	ACH-000192	5.3647557	6.074358	6.0786267	5.4877845	6.179539	6.058146	5.626932
14	ACH-000212	5.5956755	5.8619824	5.8839763	5.4988265	5.962996	5.88797	5.79393
15	ACH-000223	6.2001239	5.8289575	4.4268051	7.2222478	5.129517	5.742008	6.094408
16	ACH-000237	6.3257534	6.0040484	6.5698974	6.1583628	5.896452	5.455284	5.714727
17	ACH-000252	5.7088673	5.7843453	5.4976081	5.4595956	6.090508	5.861861	6.150766
18	ACH-000255	6.217646	6.006074	6.2388867	5.915646	5.925503	6.603285	5.909973
19	ACH-000276	5.7345491	5.7412427	5.7269337	5.5580139	5.837319	6.174542	5.922278
20	ACH-000277	5.3957414	6.074301	6.1999301	6.3559646	5.953233	6.242203	5.683543
21	ACH-000278	5.7149653	5.0300495	6.1042009	6.2278942	5.501309	6.071313	5.917487

Figure 4: Overview of Metabolites Dataset

	Cell Lines			Metabolites		
	Complete	Contains 0	Total	Complete	Contains 0	Total
Number	224	0	224	225	0	225

Table 2. Overview of Metabolites Dataset

Metabolites

Cell Lines

Number

224

0

224

225

0

225

在癌細胞轉移方向此資料中，包含 265 筆 cell lines、5 種器官 (癌細胞是否轉移) 以及癌細胞所屬主要癌症，其中有 41 筆重複的 cell lines，且皆無遺失值。

	A	B	C	D	E	F	G
1	bone	brain	ID	kidney	liver	lung	primary_disease
2	NonMetastasis	NonMetastasis	ACH-000007	NonMetastasis	NonMetastasis	NonMetastasis	Colon/Colorectal Cancer
3	NonMetastasis	NonMetastasis	ACH-000013	NonMetastasis	NonMetastasis	NonMetastasis	Ovarian Cancer
4	NonMetastasis	NonMetastasis	ACH-000019	NonMetastasis	NonMetastasis	Metastasis	Breast Cancer
5	NonMetastasis	NonMetastasis	ACH-000028	NonMetastasis	Metastasis	Metastasis	Breast Cancer
6	NonMetastasis	NonMetastasis	ACH-000048	NonMetastasis	NonMetastasis	NonMetastasis	Ovarian Cancer
7	NonMetastasis	NonMetastasis	ACH-000091	NonMetastasis	NonMetastasis	NonMetastasis	Ovarian Cancer
8	NonMetastasis	NonMetastasis	ACH-000097	Metastasis	NonMetastasis	Metastasis	Breast Cancer
9	NonMetastasis	NonMetastasis	ACH-000117	NonMetastasis	NonMetastasis	NonMetastasis	Breast Cancer
10	NonMetastasis	NonMetastasis	ACH-000123	NonMetastasis	NonMetastasis	NonMetastasis	Ovarian Cancer
11	NonMetastasis	NonMetastasis	ACH-000132	NonMetastasis	NonMetastasis	NonMetastasis	Ovarian Cancer
12	NonMetastasis	Metastasis	ACH-000147	Metastasis	Metastasis	Metastasis	Breast Cancer
13	NonMetastasis	NonMetastasis	ACH-000192	NonMetastasis	NonMetastasis	NonMetastasis	Endometrial/Uterine Cancer
14	NonMetastasis	Metastasis	ACH-000212	NonMetastasis	NonMetastasis	NonMetastasis	Breast Cancer
15	NonMetastasis	NonMetastasis	ACH-000223	NonMetastasis	NonMetastasis	NonMetastasis	Breast Cancer
16	NonMetastasis	NonMetastasis	ACH-000237	Metastasis	Metastasis	NonMetastasis	Ovarian Cancer
17	NonMetastasis	NonMetastasis	ACH-000252	NonMetastasis	NonMetastasis	NonMetastasis	Colon/Colorectal Cancer
18	Metastasis	NonMetastasis	ACH-000255	NonMetastasis	Metastasis	Metastasis	Gastric Cancer
19	NonMetastasis	NonMetastasis	ACH-000276	NonMetastasis	NonMetastasis	NonMetastasis	Breast Cancer
20	NonMetastasis	NonMetastasis	ACH-000277	NonMetastasis	NonMetastasis	NonMetastasis	Breast Cancer
21	NonMetastasis	NonMetastasis	ACH-000278	Metastasis	NonMetastasis	NonMetastasis	Ovarian Cancer

Figure 5: Overview of Metastasis Dataset

以 DepMap ID 將三種資料中共有的 cell lines 進行整合，整合的資料共 224 筆 (cell lines)，包含 19221 個基因 (表達量)、225 個代謝體 (濃度)、5 種器官 (癌細胞是否轉移) 以及癌細胞所屬主要癌症症，且皆無遺失值。

2 Quality Control

在基因表達量此資料中，移除 1526 個低表達量的基因 (> 0.5 的個數至少有 2 個)，接著移除 13268 個有極值表現量的基因 (> 3 倍標準差)，移除後的資料包含 224 筆 cell lines、4427 個基因表達量。

	A	B	C	D	E	F	G	H
1		TSPAN6 (7105)	TNMD (64102)	DPM1 (8813)	SCYL3 (57147)	C1orf112 (55732)	FGR (2268)	CFH (3075)
2	ACH-001113	4.331991778	0	7.364397337	2.792855352	4.470536865	0.02856915	1.22650853
3	ACH-001289	4.566815154	0.584962501	7.106536769	2.543495883	3.504620392	0	0.18903382
4	ACH-001339	3.150559677	0	7.379031732	2.333423734	4.227278994	0.05658353	1.31034012
5	ACH-001538	5.085339669	0	7.154109243	2.545968369	3.084064265	0	5.86814348
6	ACH-000242	6.729144865	0	6.537606691	2.456806149	3.867896464	0.79908731	7.20838069
7	ACH-000708	4.272023189	0.189033824	7.022922589	2.555816155	3.841973119	0	0.0976108
8	ACH-000327	3.337711092	0	5.927185358	1.944858446	2.678071905	0.01435529	3.08915913
9	ACH-000233	0.056583528	0	6.093602429	3.970853654	3.731183242	0.02856915	6.09296851
10	ACH-000461	4.016139703	0	6.533874777	2.22650853	3.021479727	0.02856915	0.08406426
11	ACH-000705	4.411426246	0	6.412442825	2.364572432	4.275007047	0.04264434	0.20163386
12	ACH-001794	3.871843649	0.056583528	6.76394266	1.937344392	3.152183419	0.62293035	6.94462422
13	ACH-002023	5.264911693	0	6.756889855	2.283921772	3.813524689	0	4.0591822
14	ACH-000528	4.512226887	0	7.099821179	2.843983844	4.672425342	0.01435529	0.81557543
15	ACH-001655	3.592158002	0	6.747118816	0.925999419	1.839959587	0.02856915	0.05658353
16	ACH-000167	0.042644337	0	6.710117632	2.353323291	3.98550043	5.85972113	0.27500705
17	ACH-000792	3.279471296	0	6.390254956	1.748461233	3.435628594	0.08406426	0.422233
18	ACH-001098	5.909293086	0	6.783980414	3.292781749	3.087462841	0	4.89820835
19	ACH-000570	5.264911693	0	6.830483516	2.835924074	4.177917792	0.17632277	5.84046323
20	ACH-000351	3.972692654	0	8.161333468	2.350497247	4.097610797	1.60880924	0.3448285
21	ACH-000769	4.713145902	0	6.193771743	2.440952198	4.176322773	0	0.11103131

Figure 6: Demonstration of Low Expression Gene

3 Summary Statistics & Data Visualization

3.1 Plots of Statistics

3.1.1 Bar Chart of Metastasis on 5 Organs

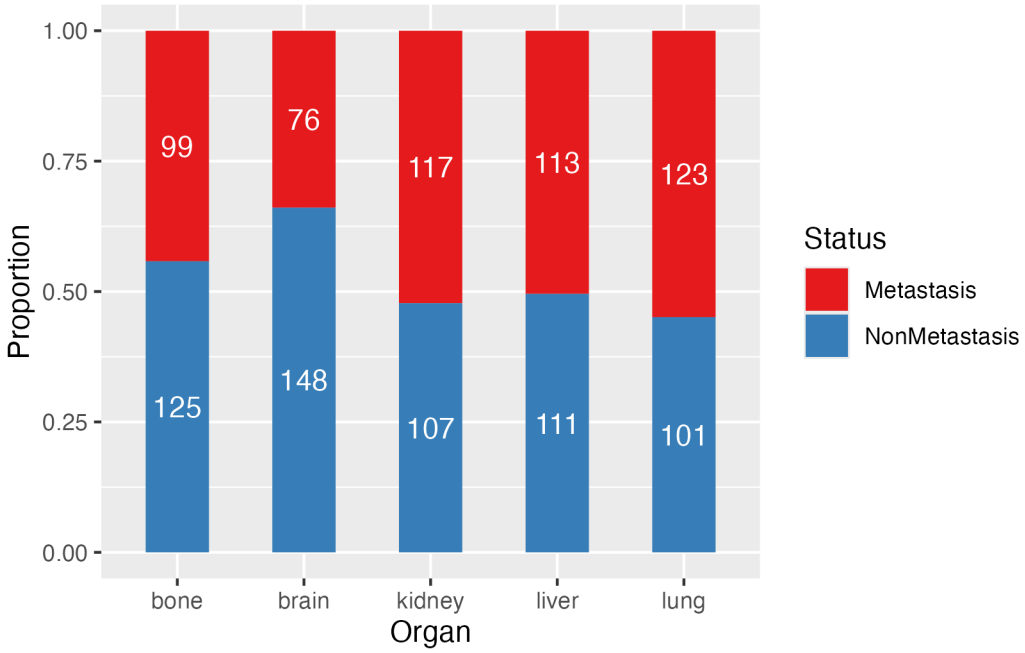


Figure 7: Stacked Bar Chart of Metastasis on 5 Organs

3.1.2 Bar Chart of Statistics about Gene Expression

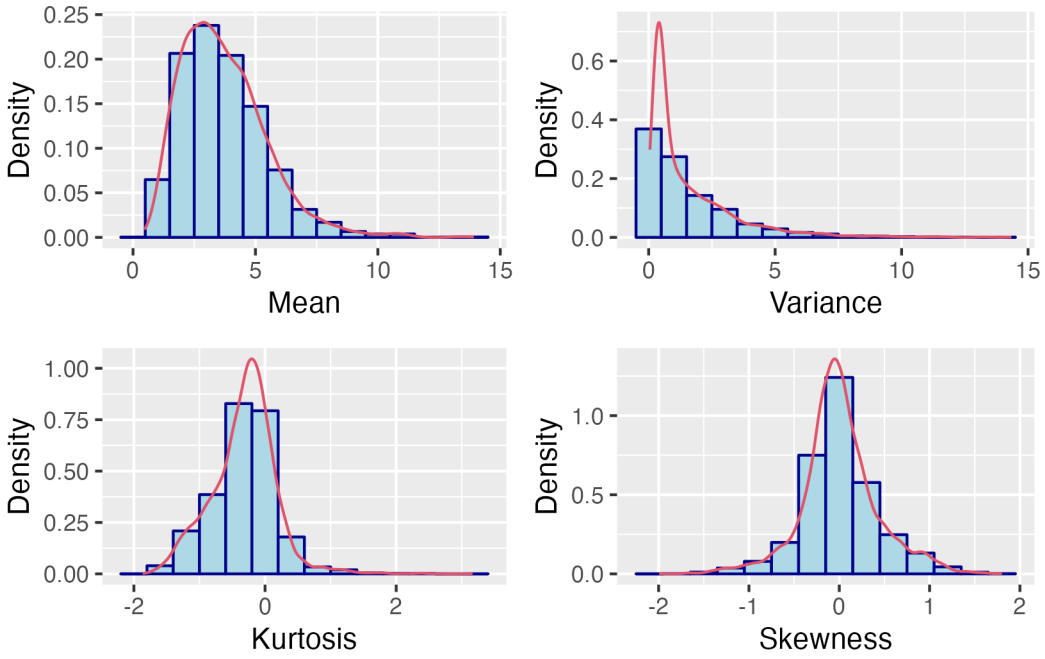


Figure 8: Bar Chart of Statistics about Gene Expression

3.1.3 Bar Chart of Statistics about Metabolites

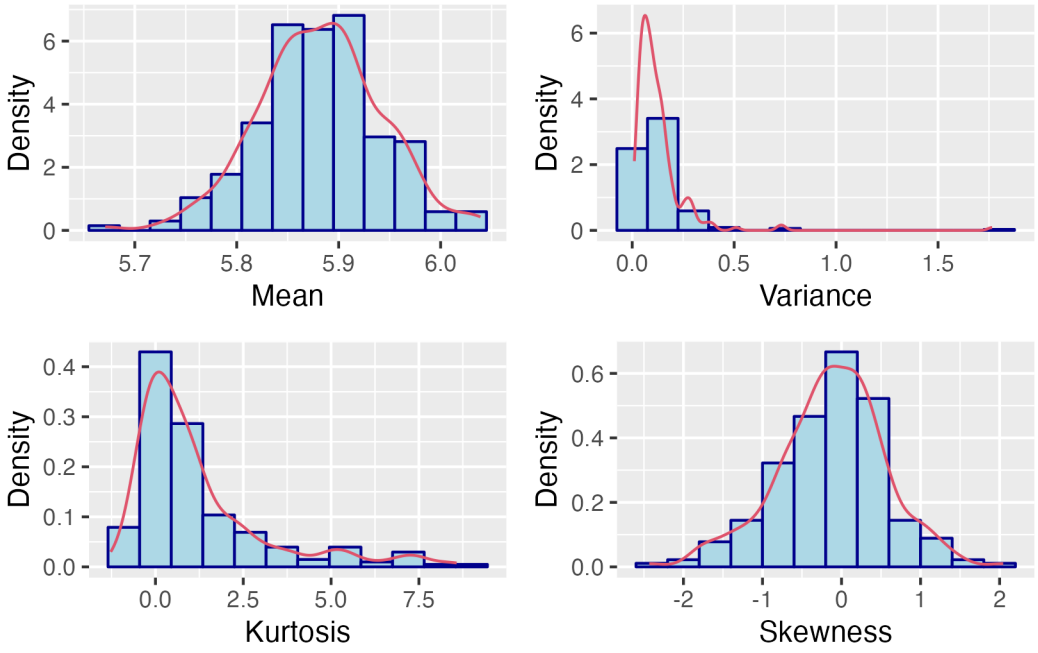
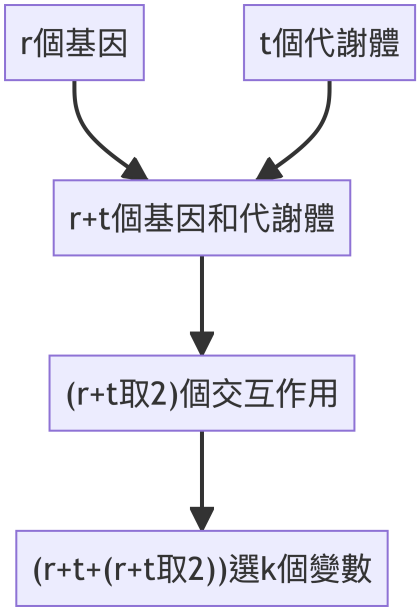


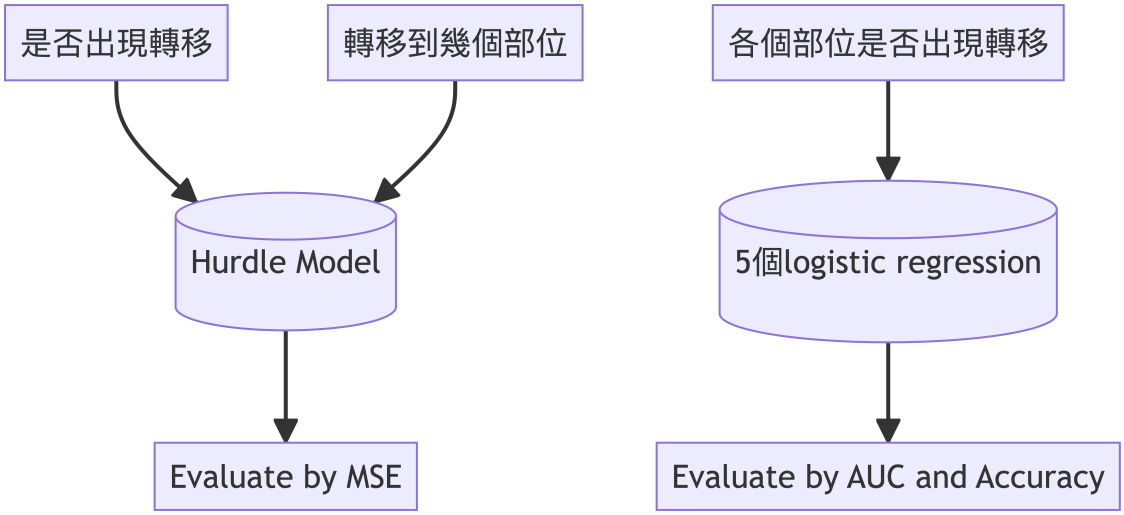
Figure 9: Bar Chart of Statistics about Metabolites

4 分析流程

1. 篩選變數



2. 建立模型



1. 關於篩選變數：

   在這份資料中，由於 p>n 的問題，篩選變數是一個需要我們仔細思考的問題，另外，我們也從生物資訊相關文獻裡面知道了，除了個別基因和代謝體，基因和基因之間的交互作用，代謝體和代謝體之間的交互作用，基因和代謝體之間的交互作用都會影響反應變數，因此，如何有效的將上述三種交互作用放入模型中是我們考量的方向。當然，我們可以將上述的三種交互作用一起放入模型中做篩選，但我們的

變數本就很大了，再放進去顯然是一種不效率的做法。為了解決這個問題，我們發想了以下流程：

1. 先從所有基因中篩選出 r 個最重要的基因
2. 再從所有代謝體中篩選出 t 個最重要代謝體
3. 接下來，我們只用 (r+t) 個基因和代謝體來看交互作用，共有 $\binom{r+t}{2}$ 個交互作用。
4. 最後，從 (r+t+$\binom{r+t}{2}$) 個變數中篩選出最重要的 k 個變數

2. 建模

由這分資料的 response，我們可以定義出下列三種反應變數：

1. 反應變數為一個細胞是否轉移
2. 反應變數為一個細胞轉移到幾個部位
3. 反應變數為 5 個部位分別是否有轉移

由於我們主要關心是第三種反應變數，前兩者我們會使用 Hurdle model 來配適模型，可以同時看出一些前兩者反應變數的資訊，這邊會使用。第三種反應變數，我們會使用 logistic regression 來配適模型，並且透過切割 train test 和 AUC 來驗證模型是否有效。

5 機器學習

5.0.1 目標

預測五個變數 (Bone, Brain, Kidney, Liver, Lung) 是否有轉移。

5.0.2 方法

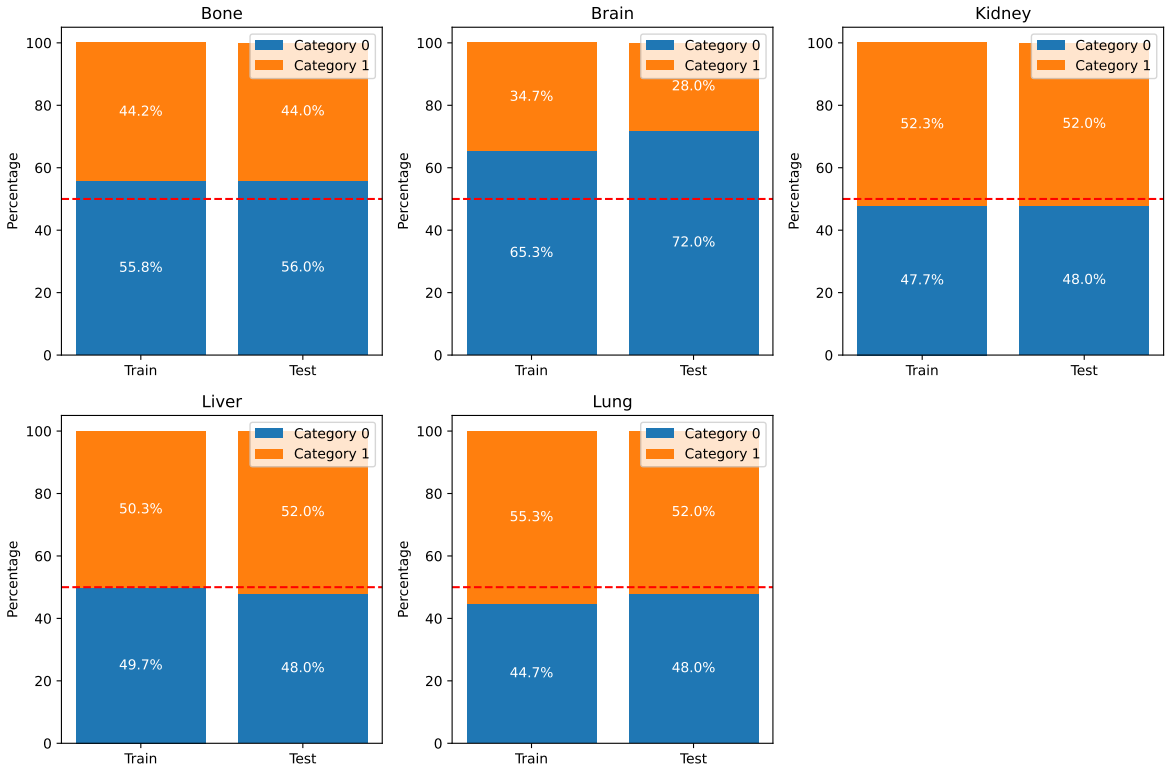
- 深度學習: Multi-head Neural Network
- 機器學習: Decision Tree, Random Forest, LightGBM, CatBoost

5.0.3 預處理

劃分資料集

將資料按照 9:1 切分訓練集與測試集。針對五個目標變數的轉移狀態，此資料共有 28 種可能的狀態組合。採用分層切分的方式，確保各個組合在訓練集與測試集的比例盡量保持一致。
::: {.cell}

:::



由上圖發現，`Brain` 存在些許資料不平衡問題 (1.89 倍)，其餘變數則不是很明顯，後續分析將嘗試針對 `Brain` 變數做上採樣 (SMOTE)。

5.0.4 Multi-head Neural Network

目標

希望做到一個模型同時預測多個目標變數。

模型設計

共享層

包含一個全連接層，配合 ReLU 激活函數，其主要作用是提取輸入資料中的通用特徵。這層的輸出維度設定為 128 維。

五個獨立的輸出層

在共享層之後，架構分出五個獨立的輸出層，每個輸出層都由一個全連接層構成。每個輸出層都專門負責預測一個特定的目標變數。這種設計允許模型對每個目標進行專門的學習和預測，同時基於共享層的特徵，加強模型對各目標之間可能存在的隱含關聯的理解。

損失函數

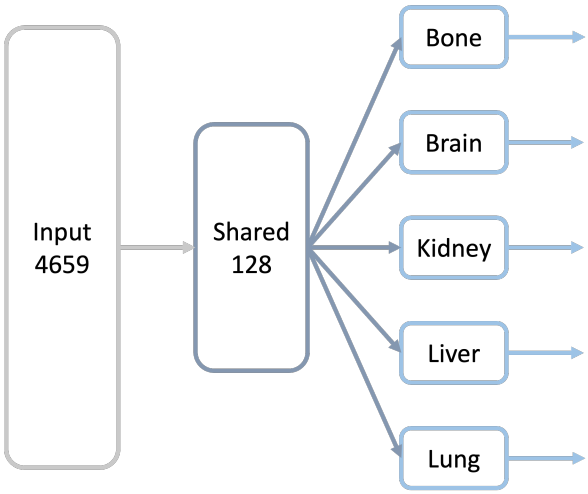
採用二元交叉熵（Binary Cross-Entropy）作為損失函數，對每個獨立輸出層的預測結果計算損失。由於這是一個多目標的預測任務，模型會計算所有輸出層的損失總和，以此來進行梯度下降並更新網路的權重。

優化器

Adam

學習率

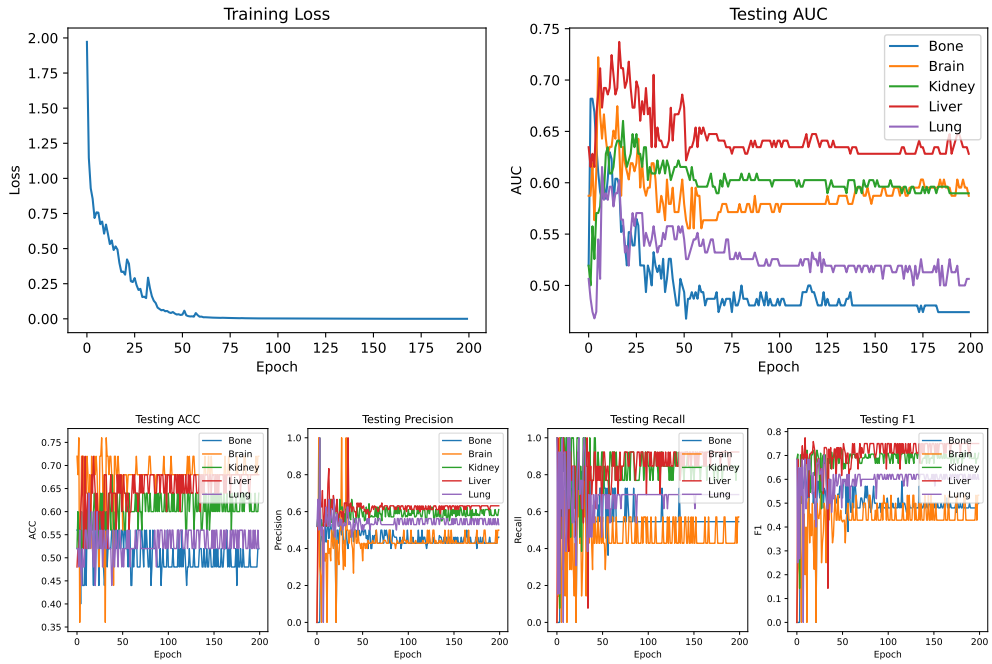
0.0005



共有 $(4659+1)\times 128+(128+1)\times 5=597125$ 個參數待估計。

模型表現

```
<torch._C.Generator object at 0x352d80650>
```



深度學習模型對隨機初始值特別敏感，尤其當樣本數較少時，這一點更為明顯。在此次實驗的五個目標變數的預測中，AUC 值的波動範圍從 0.4 到 0.7 不等。但若是初始值設定不佳，模型有時會在初期傾向將所有樣本預測為全正或全負，導致預測結果極不穩定。因此，在樣本量有限的情況下，依賴深度學習可能不是理想的選擇。

5.0.5 Machine Learning Method

目標

為目標變數: Bone, Brain, Kidney, Liver, Lung 分別建立五個模型，並觀察對這五個目標最有影響的變數。

方法

由於希望考慮交互作用項，以及後續方便解釋特徵重要性，這裡皆使用 Tree-based 模型。

- Decision Tree

- Random Forest
- LightGBM
- CatBoost
  ::: {.cell}

:::

Decision Tree

	Accuracy	Precision	Recall	F1	AUC
Bone	0.48	0.3750	0.2727	0.3158	0.4578
Brain	0.52	0.2222	0.2857	0.2500	0.4484
Kidney	0.32	0.3750	0.4615	0.4138	0.3141
Liver	0.60	0.6154	0.6154	0.6154	0.5994
Lung	0.52	0.5385	0.5385	0.5385	0.5192

Random Forest

	Accuracy	Precision	Recall	F1	AUC
Bone	0.64	0.6250	0.4545	0.5263	0.6883
Brain	0.68	0.3333	0.1429	0.2000	0.4206
Kidney	0.64	0.6111	0.8462	0.7097	0.7115
Liver	0.52	0.5333	0.6154	0.5714	0.6538
Lung	0.40	0.4444	0.6154	0.5161	0.3622

LightGBM

	Accuracy	Precision	Recall	F1	AUC
Bone	0.56	0.5000	0.6364	0.5600	0.5455
Brain	0.48	0.1250	0.1429	0.1333	0.4206
Kidney	0.52	0.5238	0.8462	0.6471	0.5577
Liver	0.60	0.6154	0.6154	0.6154	0.5833
Lung	0.48	0.5000	0.7692	0.6061	0.5577

CatBoost

	Accuracy	Precision	Recall	F1	AUC
Bone	0.56	0.5000	0.5455	0.5217	0.6299
Brain	0.64	0.3333	0.2857	0.3077	0.4921
Kidney	0.48	0.5000	0.8462	0.6286	0.5897
Liver	0.52	0.5333	0.6154	0.5714	0.5897
Lung	0.48	0.5000	0.7692	0.6061	0.5641

SMOTE

僅針對 Brain 變數做上採樣，將少類的樣本補到多數類的 80%。::: {.cell} ::: {.cell-output .cell-output-stdout}

```
Brain
```

:::

```
原資料: 轉移 130, 沒轉移 69
```

```
SMOTE後: 轉移 130, 沒轉移 104
```

Model	Accuracy	Precision	Recall	F1	AUC
DecisionTree	0.48	0.3750	0.2727	0.3158	0.4578
RandomForest	0.72	0.8333	0.4545	0.5882	0.7695
LightGBM	0.60	0.5714	0.3636	0.4444	0.6558
CatBoost	0.68	0.7143	0.4545	0.5556	0.6234

:::

將做過 SMOTE 上採樣的資料進行訓練，所有模型的表現皆提高了。除了 DecisionTree 只有微幅上升以外，其餘模型 RandomForest, LightGBM, CatBoost 皆提高了 0.2 以上。

各個目標變數對應 **AUC** 最高之模型

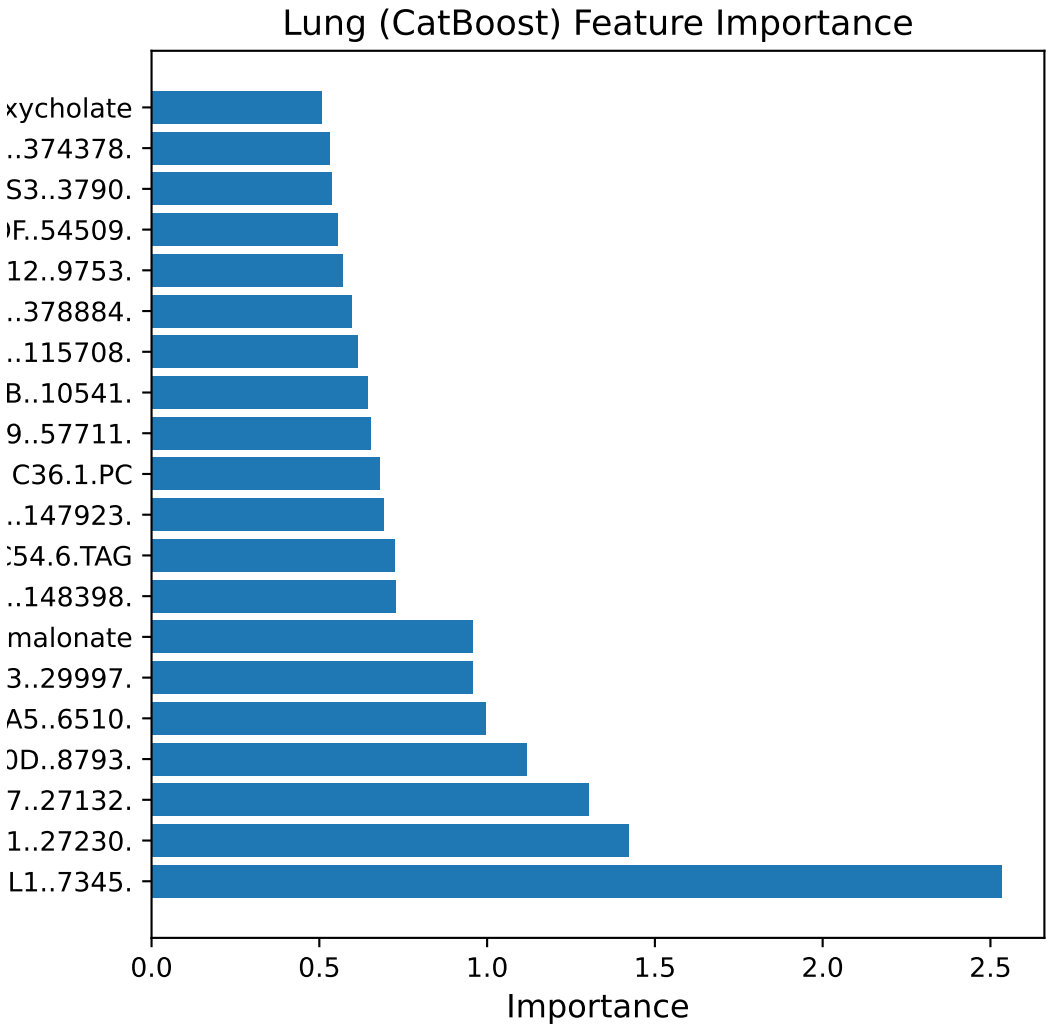
	Bone	Brain	...	Liver	Lung
Accuracy	0.64	0.72	...	0.52	0.48
Precision	0.625	0.8333	...	0.5333	0.5
Recall	0.4545	0.4545	...	0.6154	0.7692
F1	0.5263	0.5882	...	0.5714	0.6061
AUC	0.6883	0.7695	...	0.6538	0.5641
Source	RandomForest	RandomForest_Smote	...	RandomForest	CatBoost

[6 rows x 5 columns]

以下為針對不同目標變數預測轉移的最佳模型選擇：

- 對於是否轉移到「骨頭」，隨機森林模型表現最佳，AUC 可達 66.56%。
- 對於是否轉移到「大腦」，結合 SMOTE 與隨機森林模型能夠達到最佳效果，AUC 可達 76.95%。
- 對於是否轉移到「腎臟」，隨機森林模型表現最佳，AUC 可達 71.15%。
- 對於是否轉移到「肝臟」，「隨機森林模型表現最佳，AUC 可達 65.38%。
- 對於是否轉移到「肺臟」，CatBoost 模型表現最佳，AUC 為 56.41%。

5.0.6 對是否轉移影響最大的前 20 個變數



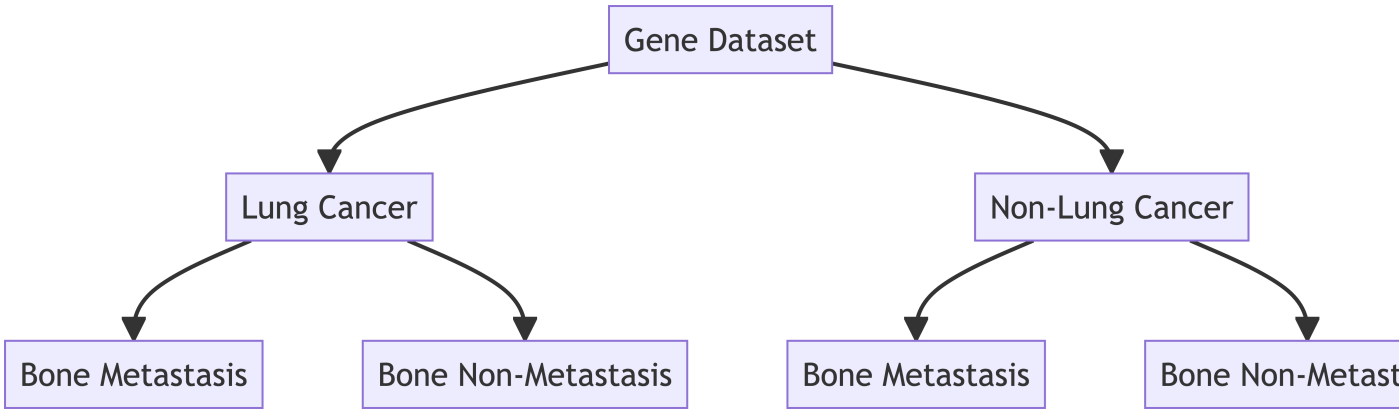
- 預測是否轉移至「骨頭」，重要的代謝體為: C34.4.PC
- 預測是否轉移至「大腦」，重要的代謝體為: C54.6.TAG, C58.8.TAG, asparagine, C56.6.TAG, C54.4.TAG
- 預測是否轉移至「腎臟」，重要的代謝體為: homocysteine, cytidine
- 預測是否轉移至「肝臟」，重要的代謝體為: GABA, putrescine, cytidine
- 預測是否轉移至「肺臟」，重要的代謝體為: taurodeoxycholate.taurochenodeoxycholate, C36.1.PC, C54.6.TAG, succinate.methylmalonate

6 篩選變數

6.1 Differential Analysis

以 Primary Disease 的 Lung Cancer 作為切割點，將資料區分成 Lung Cancer 與 Non-Lung Cancer，再分別對各器官癌細胞是否轉移配適羅吉斯迴歸模型，並比對基因表現量和代謝體濃度在各器官癌細胞是否轉移上是否有顯著不同，以 bone 為例，$T = log(\frac{Mean(Metastasis|Lung,Gene1)}{Mean(NonMetastasis|Lung,Gene1)})$。以 t 檢定統計量進而計算各基因和各代謝體的 $p - value$，再輔以 fold change，最終挑選出 50 個基因與代謝體作為後續交互作用項產生所使用的變數，並保留前 20 個顯著的基因與代謝體在最終挑選的變數。由此 50 個基因與代謝體所產生的交互作用項變數透過前述流程可得出約 30 個顯著的交互作用項變數，再加上先前所保留 20 個基因與代謝體變數，最終可得出約 50 個變數供後續模型使用。

若針對預測至少 1 種器官發生癌細胞轉移，則將 5 種器官癌細胞是否轉移整合成至少 1 種器官癌細胞是否轉移，並依上述流程可得出約 50 個變數供後續模型使用。



6.1.1 Volcano Plot (Bone|Gene)

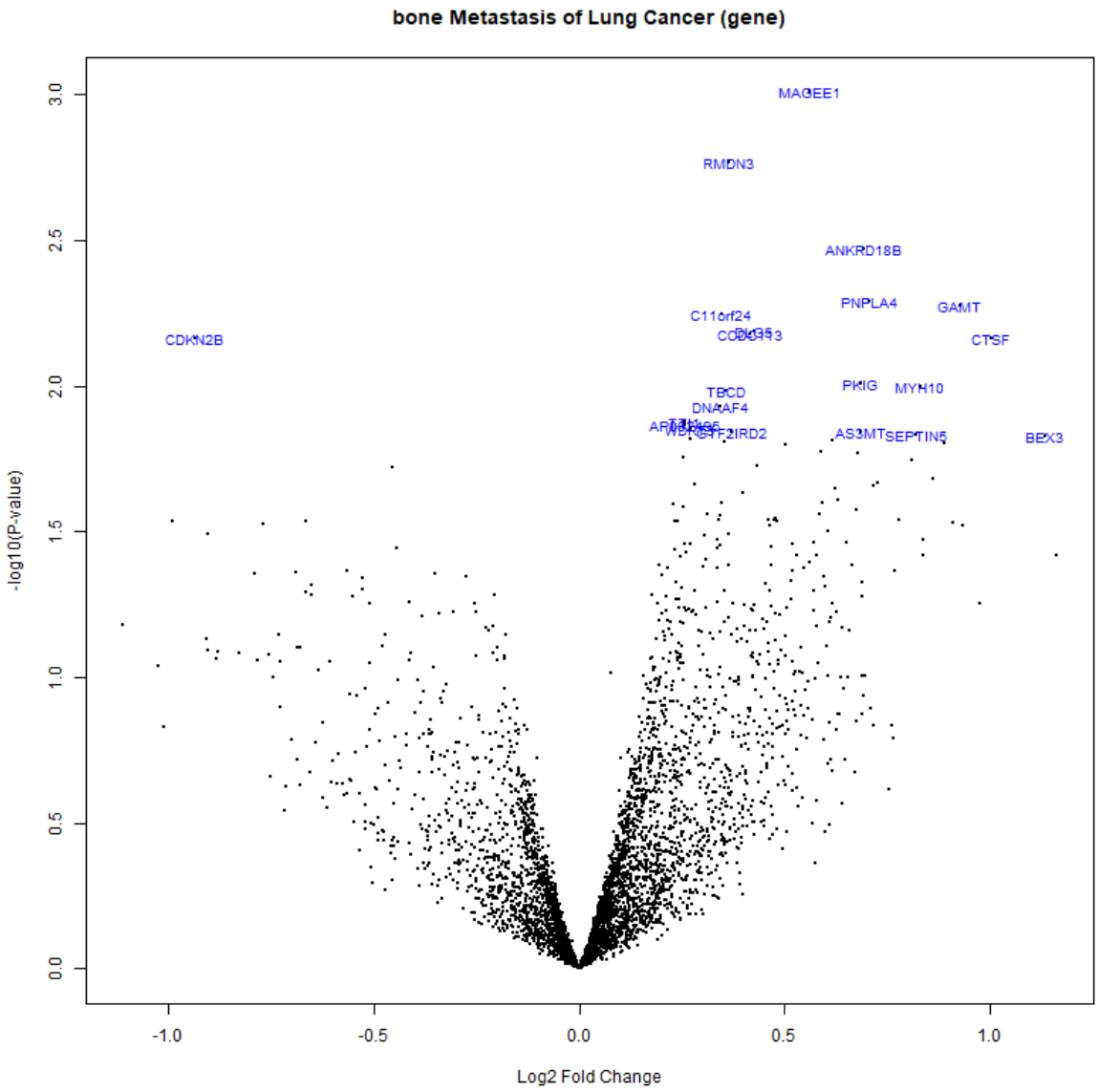


Figure 10: Volcano Plot of Bone Metastasis on Gene Expression under Lung Cancer

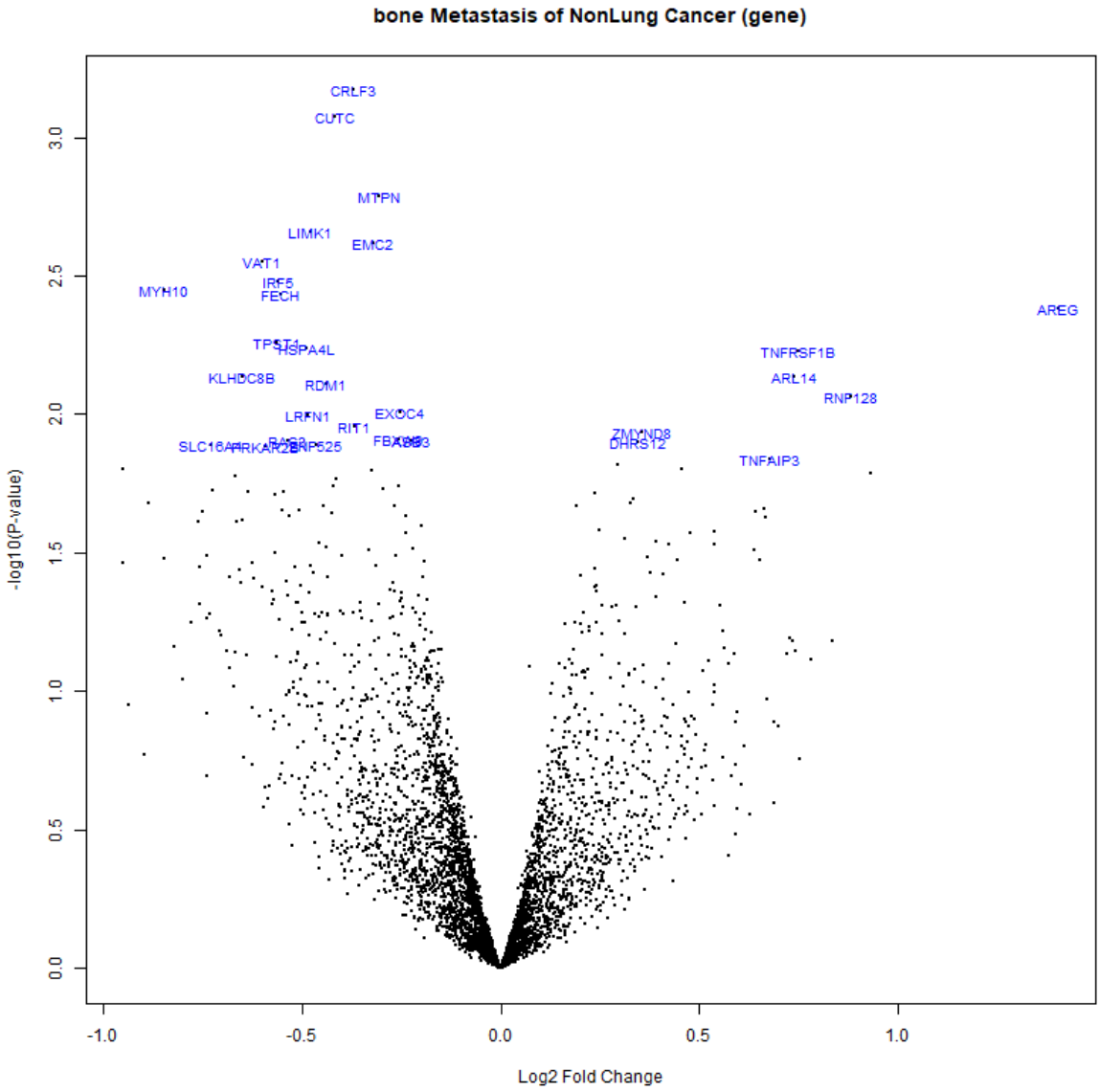


Figure 11: Volcano Plot of Bone Metastasis on Gene Expression under NonLung Cancer

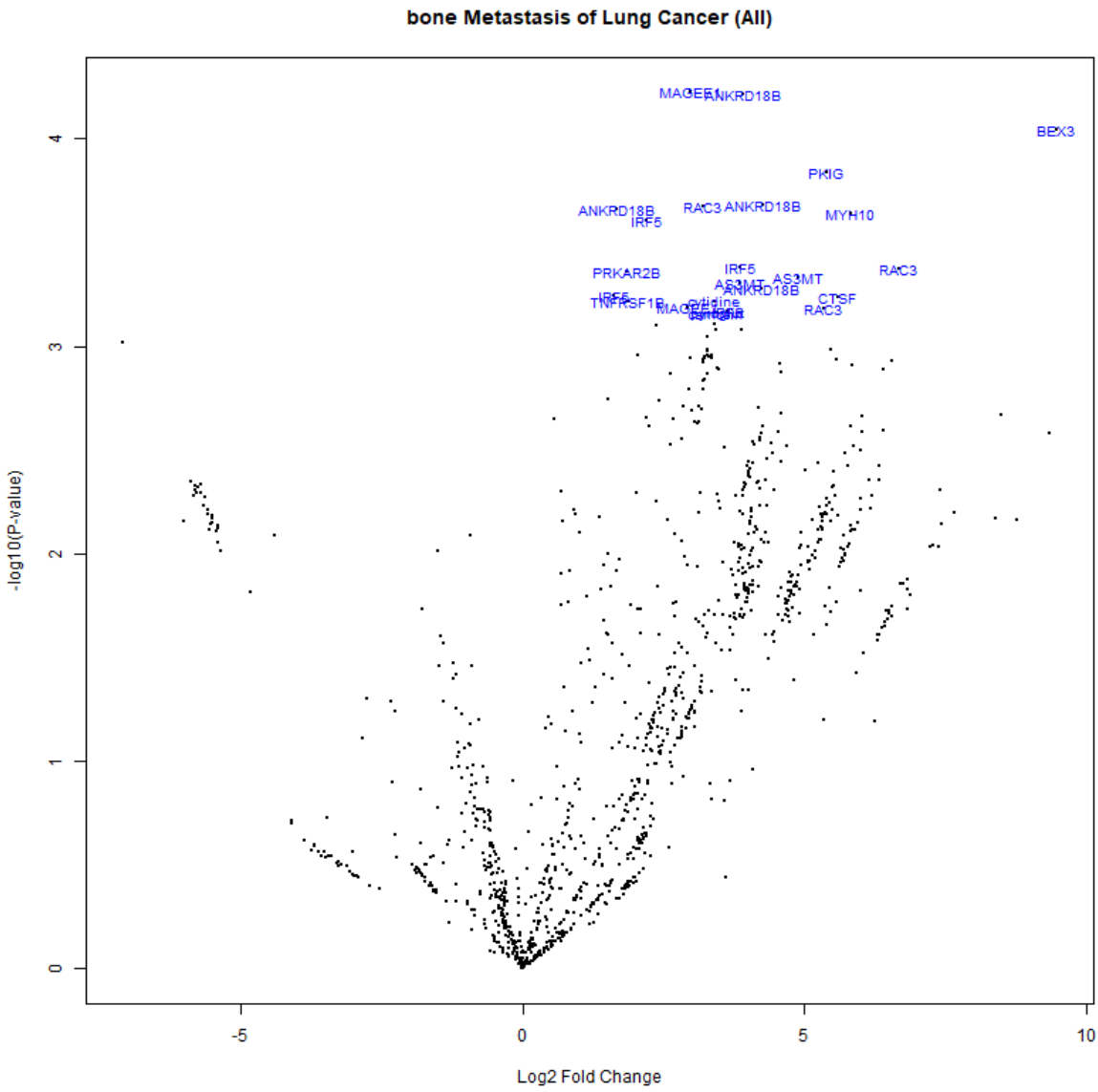


Figure 12: Volcano Plot of Bone Metastasis on Interaction Effect under Lung Cancer

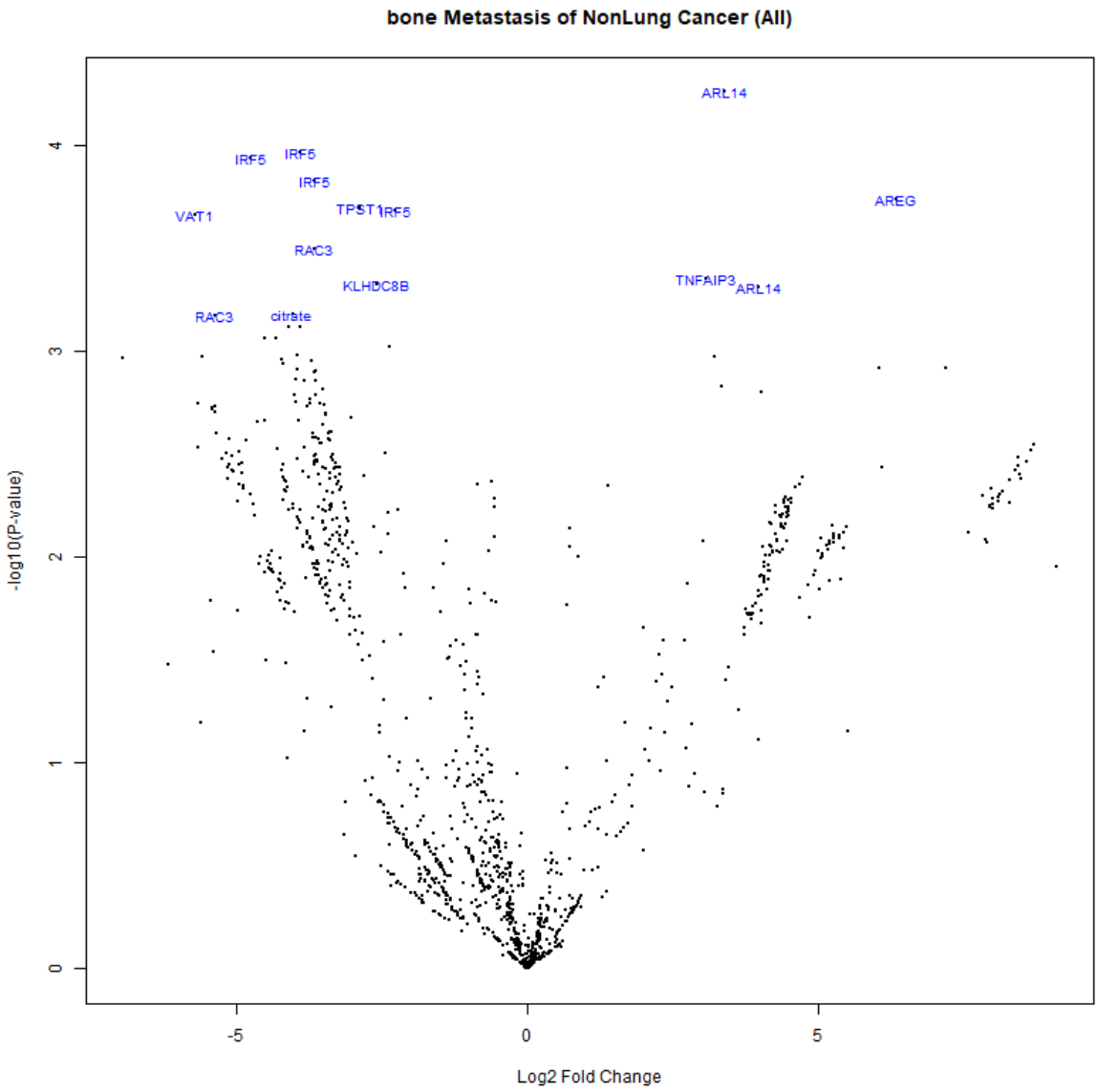


Figure 13: Volcano Plot of Bone Metastasis on Interaction Effect under NonLung Cancer

Table 3. Criteria of Differenial Analysis on Bone without Interaction Effect

bone

Lung Cancer

Non-Lung Cancer

Total

Genes

Metabolites

Genes

Metabolites

Criteria

0.015

0.500

0.025

0.100

0.015

0.500

0.025

0.100

48.000

Table 4. Criteria of Differenial Analysis on Bone with Interaction Effect

bone

Lung Cancer

Non-Lung Cancer

Total

Criteria

0.0007

3.0000

0.0007

3.0000

47.0000

最終挑選出的變數整理如下:

variables_bone	variables_brain	variables_kidney	variables_liver	variables_lung	variables_binary
MAGEE1..57692.	ZSL2..64843.	SMAP1..60682.	PNMA8A..55228.	BBS5..129880.	AMIGO1..57463.
ANKRD18B..4417.	IZUMO1..22873.	NLRP1..22861.	H2BC5..3017.	DBNDD2..55861.	NECAB3..63941.
PNPLA4..8228.	FMNL1..752.	CYBA..1535.	H2BC4..8347.	CTSF..8722.	SMAP1..60682.
GAMT..2593.	RAB38..23682.	BEX3..27018.	H2AC6..8334.	DOCK4..9732.	SEPTIN5..5413.
CTSF..8722.	SIMC1..375484.	LRP3..4037.	CCDC106..29903.	AUTS2..26053.	MAPK8IP1..9479.
VAT1..10493.	AREG..374.	ERMP1..79956.	GATA2..2624.	ARL4D..379.	ADCY9..115.
IRF5..3663.	ZNF607..84775.	SERPINB8..5271.	SERPINB8..5271.	SEMA3B..7869.	CBX2..84733.
MYH10..4628.	STEAP1..26872.	SLC29A3..55315.	TRAV1..7409.	ERMP1..79956.	SIK1B..102724428.
FECH..2235.	KAZN..23254.	AREG..374.	CPT1B..1375.	SPACA6..147650.	GAS6..2621.
AREG..374.	MID1..4281.	CXCL8..3576.	TRAF5..7188.	KCNMA1..3778.	homocysteine
acetylcarnitine	C46.0.TAG	C40.6.PC	thymine	trimethylaminecytoxide	N.oxide
aconitate	C40.6.PC	homocysteine	C18.0.CE	C40.6.PC	C32.2.PC
isocitrate	C58.7.TAG	C18.2.LPC	arachidonyl_carnitine	homocysteine	C32.0.PC
C18.2.LPC	butyrobetaine	NADP	butyrobetaine	C46.0.TAG	guanosine
C18.0.CE	X3.phosphoglycerate.	C34.4.PC	X2.aminoadipate	uracil	C36.1.PC
C36.1.DAG	C52.5.TAG	uridine	C36.1.DAG	phosphocreatine	phosphocreatine
cytidine	C56.6.TAG	thiamine	cytidine	thiamine	C18.0.LPE
pyroglutamic.acid	C58.8.TAG	C18.0.LPE	C18.0.LPE	C36.1.PC	C36.1.DAG
inositol	C54.4.TAG	cytidine	C48.3.TAG	niacinamide	SMAP1..60682.
					PHYH..5264.

variables_bone	variables_brain	variables_kidney	variables_liver	variables_lung	variables_binary
C58.8.TAG	DZIP1..22873. _ RAB38..23682.	CYBA..1535. _ DEF6..50619.	H2BC5..3017. _ PLEK2..26499.	lysine	MAPK8IP1..9479. _ SMAP1..60682.
MYH10..4628. _ PN-PLA4..8228.	ISL2..64843. _ SPARC..6678.	SMAP1..60682. _ GPC1..2817.	H2BC5..3017. _ OAS3..4940.	LRP3..4037. _ GPC1..2817.	NECAB3..63941. _ SMAP1..60682.
BEX3..27018. _ PN-PLA4..8228.	ISL2..64843. _ RAB38..23682.	CTNNAL1..8727. _ CYBA..1535.	H2BC5..3017. _ EFNB2..1948.	GSTM3..2947. _ IL1R1..3554.	GSN..2934. _ PHYH..5264.
PKIG..11142. _ PN-PLA4..8228.	ISL2..64843. _ DZIP1..22873.	MT2A..4502. _ CYBA..1535.	PTGER4..5734. _ H2BC5..3017.	GSTM3..2947. _ LRP3..4037.	GSN..2934. _ SMAP1..60682.
AS3MT..57412. _ PN-PLA4..8228.	TLR3..7098. _ DZIP1..22873.	LRP3..4037. _ GPC1..2817.	EPS8L2..64787. _ H2BC5..3017.	GSN..2934. _ IL1R1..3554.	PKIG..11142. _ NECAB3..63941.
AS3MT..57412. _ CTSF..8722.	TLR3..7098. _ GJA1..2697.	DIPK1B..138311. _ GPC1..2817.	H2AC6..8334. _ OAS3..4940.	AUTS2..26053. _ GPC1..2817.	PKIG..11142. _ GSN..2934.
ANKRD18B..44459. _ PN-PLA4..8228.	SLC1..375484. _ RAB38..23682.	BEX3..27018. _ ATP9A..10079.	H2AC6..8334. _ PTGER4..5734.	AUTS2..26053. _ GSTM3..2947.	AMIGO1..57463. _ GAMT..2593.
ANKRD18B..44459. _ GAMT..2593.	GOLGA8A..23015. _ RAB38..23682.	BEX3..27018. _ GPC1..2817.	H2AC6..8334. _ EPS8L2..64787.	AUTS2..26053. _ GSN..2934.	AMIGO1..57463. _ GSN..2934.
IRF5..3663. _ GAMT..2593.	GOLGA8A..23015. _ ISL2..64843.	BEX3..27018. _ SMAP1..60682.	H2BC4..8347. _ OAS3..4940.	BBS5..129880. _ GSTM3..2947.	B4GALNT4..338707. _ AMIGO1..57463.
RAC3..5881. _ MYH10..4628.	MID1..4281. _ GCH1..2643.	BEX3..27018. _ FAXC..84553.	H2BC4..8347. _ H2BC5..3017.	CTSF..8722. _ GPC1..2817.	SEPTIN5..5413. _ SMAP1..60682.
RAC3..5881. _ MAGEE1..57692.	COL6A1..1291. _ TFPI..7035.	SMTN..6525. _ CYBA..1535.	H2BC4..8347. _ PTGER4..5734.	CTSF..8722. _ PCSK1N..27344.	FITM2..128486. _ GAMT..2593.
phosphocreatine _ MAGEE1..57692.	STEAP1..26872. _ COL6A1..1291.	RASA3..22821. _ CYBA..1535.	H2BC4..8347. _ EPS8L2..64787.	CTSF..8722. _ LRP3..4037.	FITM2..128486. _ GSN..2934.
acetylcarnitine _ MAGEE1..57692.	WNK2..65268. _ MID1..4281.	RASA3..22821. _ EFNB2..1948.	GPR39..2863. _ H2BC5..3017.	CTSF..8722. _ GSN..2934.	FITM2..128486. _ SEPTIN5..5413.

variables_bone	variables_brain	variables_kidney	variables_liver	variables_lung	variables_binary
C18.0.CE _ MAGEE1..57692.	CXXC5..51523. _ LRFN1..57622.	EVL..51466. _ GPC1..2817.	GPR39..2863. _ H2BC4..8347.	CTSF..8722. _ AUTS2..26053.	CBX2..84733. _ SMAP1..60682.
beta.alanine _ MAGEE1..57692.	KAZN..23254. _ MID1..4281.	EVL..51466. _ FAXC..84553.	CPT1B..1375. _ H2BC5..3017.	EVL..51466. _ GPC1..2817.	CBX2..84733. _ PKIG..11142.
VAT1..10493. _ FECH..2235.	ZNF525..170958. _ WNK2..65268.	SRC..6714. _ ATP9A..10079.	CPT1B..1375. _ H2BC4..8347.	SHISA4..149345. _ AUTS2..26053.	CBX2..84733. _ SHISA4..149345.
AREG..374. _ TN-FRSF1B..7133.	ZNF525..170958. _ KAZN..23254.	SRC..6714. _ GPC1..2817.	thymine _ H2BC5..3017.	PCDHGC3..5098. _ GPC1..2817.	homocysteine _ PHYH..5264.
TNFAIP3..7128. _ TN-FRSF1B..7133.	C48.3.TAG _ C58.7.TAG	SRC..6714. _ MAP2..4133.	thymine _ H2AC6..8334.	PCDHGC3..5098. _ GSN..2934.	homocysteine _ SMAP1..60682.
IRF5..3663. _ MYH10..4628.	C52.5.TAG _ C58.7.TAG	SRC..6714. _ SMAP1..60682.	cytidine _ H2BC5..3017.	PCDHGC3..5098. _ AUTS2..26053.	homocysteine _ NECAB3..63941.
IRF5..3663. _ FECH..2235.	C52.4.TAG _ C58.7.TAG	SRC..6714. _ LRP3..4037.	cytidine _ H2BC4..8347.	PCDHGC3..5098. _ CTSF..8722.	homocysteine _ AMIGO1..57463.
IRF5..3663. _ VAT1..10493.	C56.6.TAG _ C48.3.TAG	SRC..6714. _ DIPK1B..138311.	taurodeoxycholate.taurodeoxycholate _ H2BC5..3017.	LTBP3..4054. _ AUTS2..26053.	acetylcholine _ NECAB3..63941.
RAC3..5881. _ IRF5..3663.	C58.8.TAG _ C48.3.TAG	SRC..6714. _ BEX3..27018.	C58.7.TAG _ H2BC5..3017.	SEMA3B..7869. _ EVL..51466.	acetylcholine _ AMIGO1..57463.
ARL14..80117. _ TN-FRSF1B..7133.	C58.8.TAG _ C52.5.TAG	SRC..6714. _ EVL..51466.	VAV1..7409. _ BEX3..27018.	SMARCD3..6604. _ SEMA3B..7869.	C32.2.PC _ AMIGO1..57463.
ARL14..80117. _ TN-FAIP3..7128.	C58.8.TAG _ C52.4.TAG	homocysteine _ BEX3..27018.	CPT1B..1375. _ SERPINB8..5271.	DENND3..22898. _ SEMA3B..7869.	C32.0.PC _ AMIGO1..57463.
cytidine _ FECH..2235.	NA	SERPINB8..5271. _ AREG..374.	NA	NCOA7..135112. _ SEMA3B..7869.	C34.4.PC _ AMIGO1..57463.
cytidine _ VAT1..10493.	NA	cytidine _ ERMP1..79956.	NA	LTBP3..4054. _ ERMP1..79956.	C18.0.LPE _ AMIGO1..57463.

variables_bone	variables_brain	variables_kidney	variables_liver	variables_lung	variables_binary
pyroglutamic.acid _ VAT1..10493.	NA	uridine _ ERMP1..79956.	NA	VASN..114990. _ ERMP1..79956.	C24.1.SM _ AMIGO1..57463.
C58.8.TAG _ VAT1..10493.	NA	C18.0.LPE _ SERPINB8..5271.	NA	ARL4D..379. _ SEMA3B..7869.	C24.0.SM _ AMIGO1..57463.
NA	NA	C46.2.TAG _ ERMP1..79956.	NA	ARL4D..379. _ ERMP1..79956.	C36.1.DAG _ AMIGO1..57463.
NA	NA	C48.3.TAG _ ERMP1..79956.	NA	ARL4D..379. _ NCOA7..135112.	SIK1B..102724428. _ ADCY9..115.
NA	NA	C50.3.TAG _ ERMP1..79956.	NA	ARL4D..379. _ LTBP3..4054.	SIK1B..102724428. _ CBX2..84733.
NA	NA	NA	NA	cytidine _ ARL4D..379.	cytidine _ ADCY9..115.
NA	NA	NA	NA	lactose _ ARL4D..379.	NA

6.2 Lasso

由於變數過多所以我們對基因資料和代謝體資料做降維，而我們第一個嘗試的降維方法為 PCA，我們分別對兩筆資料做 PCA，但因為兩筆資料的樣本數都遠少於變數個數，前幾個主成分也不能解釋大部分原始資料的變異，從特徵值來看也沒有一個有大於 1。所以我們認為做 PCA 的效益不大，後續再使用別種方法進行降維。

這些是由基因和代謝體各用 lasso 挑出 25 個變數，再由 50 個變數加上所有二階交互作用中用 lasso 挑出其中 50 個變數。其中 bone/brain/kidney/liver/lung 是分別對五種疾病各建一個羅吉斯迴歸模型。binary 是只要其中一種疾病有轉移，就 coding 成 1 的羅吉斯迴歸模型。而 count 為轉移到幾個地方的 Poisson 迴歸模型。可以注意到的是不管是哪個模型，我們用 lasso 挑出來的變數都是二階交互作用項，且每個模型中都有基因對基因和基因對代謝體還有代謝體對代謝體的交互作用。

7 Statistical Model

7.1 Hurdle Model(Lasso)

$$P(Y_i = y_i) \begin{cases} p_i & y_i = 0 \\ (1-p_i)\frac{p(y_i;\mu_i)}{1-p(y_i=0;\mu_i)} & y_i > 0 \end{cases}$$

$$log(\mu_i) = x_i^T\alpha \Rightarrow \mu_i = e^{x_i^T\alpha}$$

$$logit(p_i) = z_i^T\beta \Rightarrow p_i = \frac{e^{z_i^T\beta}}{1+e^{z_i^T\beta}}$$

- 假設 0 來自一個系統性地來源
- 非零的觀測值來自不同的分配

7.1.1 統計報表

7.1.2 分析結果

7.2 Logistic(Lasso)

7.2.1 統計報表

以下報表整理了五個器官針對各個變數的係數，() 內的數字是 standard error，* 號代表此變數 p-value<0.05，若某變數針對某部位是空格，代表在篩選變數階段，此變數就已經被篩選掉了。

我們可以從表格中看到，在篩選變數階段，五個部位篩選出來的變數就有明顯的不同，最多只有兩兩重複的變數，在進一步看顯著性，可以看出，沒有任何一個變數對於五個部位的轉移是有重複的，因此可以看出五個部位的轉移與否並沒有共用的變因。

7.2.2 AUC

在 liver 和 lung 的 auc 上，都為 1，因為運算問題，在使用 glm 模型時，程式會回傳"waring:fitted probabilities numerically 0 or 1 occurred"(Bobbitt 2024)，這個問題需要我們重新挑選變數才能解決，但是在這邊，我們進一步的切割 train-test 來看看模型是否有效。

7.2.3 Accuracy

7.3 Hurdle(p-value)

7.3.1 統計報表

7.3.2 分析結果

7.4 Logistic(p-value)

7.4.1 統計報表

7.4.2 分析結果

8 Reference

A metastasis map of human cancer cell lines ("Depmap" 2024)

RNA-seq analysis in R (Phipson 2024)

Bobbitt, Zach. 2024. "How to Handle: Glm.fit: Fitted Probabilities Numerically 0 or 1 Occurred." 2024. https://www.statology.org/glm-fit-fitted-probabilities-numerically-0-or-1-occurred/.

"Depmap." 2024. 2024. https://depmap.org/metmap/.

Phipson, Belinda. 2024. "RNA-Seq Analysis in r." 2024. https://combine-australia.github.io/RNAseq-R/06-rnaseq-day1.html#Quality_control.