



I-DELTA组会汇报

汇报人: 朱佳成

目录

01 结核病背景介绍和预期目标

02 数据集介绍

03 关键基因筛选的预处理

04 模型构建以及所遇到的问题

05 最终融合效果对比



PART ONE

结核病背景介绍和预期目标

重要事实

- 2023年总共有125万人死于结核病（其中包括16.1万名艾滋病毒感染者）。在世界范围内，继冠状病毒病（COVID-19）领先结核病三年后，现在结核病可能又重新成为全球由单一传染性病原体导致死亡的首要原因。结核病还是艾滋病毒感染者的头号杀手，以及抗微生物药物耐药性相关死亡的一个主要原因。
- 2023年，全球估计有1080万人罹患结核病，其中包括600万名男性、360万名女性和130万名儿童。所有国家和所有年龄组都有结核病感染。结核病是可以治愈和可以预防的。
- 耐多药结核病仍然是一项公共卫生危机和卫生安全威胁。2023年，只有约五分之二耐药结核病患者获得治疗。
- 自2000年以来，全球抗击结核病的努力估计已挽救了7900万人的生命。
- 要实现2023年联合国结核病问题高级别会议商定的2027全球目标，每年需要220亿美元用于结核病预防、诊断、治疗和护理。
- 到2030年终止结核病流行是联合国可持续发展目标的卫生相关具体目标之一。

概述

结核病是一种由细菌引起的传染病，大多影响肺部。结核病可在其患者咳嗽、打喷嚏或吐痰时通过空气传播。在结核病中，肺结核的发病率最高，约占结合病的80%,人们常以发病部位来称呼这种疾病，久而久之，肺结核就成为了结核病的一个常见别称。

结核病是可以预防和治愈的。

据估计，全球约四分之一的人口感染有结核杆菌。大约5%至10%的感染者最终会出现症状并发展为结核病。

那些受到感染但没有发病的人不具传染性。结核病通常使用抗生素进行治疗，如果不治疗，可能会致命。

在某些国家，为婴幼儿接种抗结核疫苗（卡介苗）以预防结核病。该疫苗可以防止结核病导致的死亡，并保护儿童免受严重结核病的侵害。

某些情况会增加一个人患结核病的风险：

- 糖尿病（高血糖）
- 免疫系统减弱（例如，艾滋病毒或艾滋病）
- 营养不良
- 烟草使用
- 有害使用酒精

预期目标

筛选尽量少的基因

只有用尽量少的基因来对结核病进行诊断这样才具有实际意义，用太多基因很难上临床

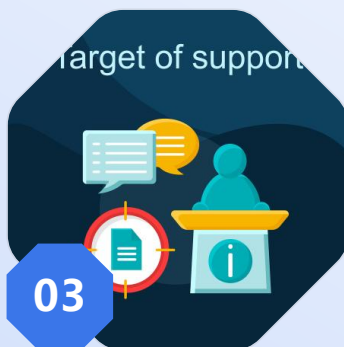


使用模型融合方法

以往都是直接构建一个TB VS HC二分类结核模型。这里要用融合模型构建两个或者多个分类器进行融合

和其它实验对比效果要好

只有效果比其它模型的效果好，这样方法才有意义



有理论支持

所选实验方法要有理论支撑















PART TWO

数据集介绍

7个研究, 12个数据集。435名TB,533名HC

Discovery Data (Microarray)								
Reference	GEO ID	TB	HC	Countries	Age range (yr)	HIV?	EndoA	EndoB
Anderson 2014	GSE39939	35	14	Kenya	< 15	Yes	23	12
Berry 2010	GSE19435	7	12	United Kingdom	21 – 51	No	6	1
	GSE19439	13	29	United Kingdom	19 – 72	No	9	4
	GSE19442**	20	31	South Africa	18 – 48	No	14	6
	GSE19444*	21	33	United Kingdom	18 – 78	No	11	10
Blankley 2016	GSE83456	45	61	United Kingdom	20 – 80	No	27	18
Bloom 2012	GSE40553	29	38	South Africa	> 17	No	22	7
Bloom 2013	GSE42825	8	23	United Kingdom	> 18	No	8	0
	GSE42826	11	52	United Kingdom, France	> 18	No	11	0
	GSE42830	16	38	United Kingdom	> 18	No	13	3
Kaforou 2014	GSE37250	195	167	South Africa, Malawi	19 – 53	Yes	108	87
Walter 2016	GSE73408	35	35	United States	20 – 86	No	17	18
Borstel Validation Data (Microarray)								
Heyckendorf 2021	GSE147689-91	121	14	Germany, Romania	18 - 85	No	64	57
RNA-Seq Validation Data								
Leong 2018	GSE101705	28	16	India	16 – 65	No	7	21
Singhania 2018	GSE107991*	21	33	United Kingdom	18 – 78	No	8	13
	GSE107992**	16	31	South Africa	18 – 48	No	10	6
	GSE107994	53	99	United Kingdom	16 – 84	No	40	13
*GSE107991 reanalyzed samples from GSE19444								
**GSE107992 reanalyzed samples from GSE19442								

	A	B	C	D	E	
1	SampleID	HIV. Status	Platform	Disease	Dataset	
2	GSM484368	Negative	GPL6947	HC	GSE19435	
3	GSM484369	Negative	GPL6947	HC	GSE19435	
4	GSM484370	Negative	GPL6947	HC	GSE19435	
5	GSM484371	Negative	GPL6947	HC	GSE19435	
6	GSM484372	Negative	GPL6947	HC	GSE19435	
7	GSM484373	Negative	GPL6947	HC	GSE19435	
8	GSM484374	Negative	GPL6947	HC	GSE19435	
9	GSM484375	Negative	GPL6947	HC	GSE19435	
10	GSM484376	Negative	GPL6947	HC	GSE19435	
11	GSM484377	Negative	GPL6947	HC	GSE19435	
12	GSM484378	Negative	GPL6947	HC	GSE19435	
13	GSM484379	Negative	GPL6947	HC	GSE19435	
14	GSM484380	Negative	GPL6947	TB	GSE19435	
15	GSM484383	Negative	GPL6947	TB	GSE19435	
16	GSM484386	Negative	GPL6947	TB	GSE19435	
17	GSM484387	Negative	GPL6947	TB	GSE19435	
18	GSM484388	Negative	GPL6947	TB	GSE19435	
19	GSM484395	Negative	GPL6947	TB	GSE19435	

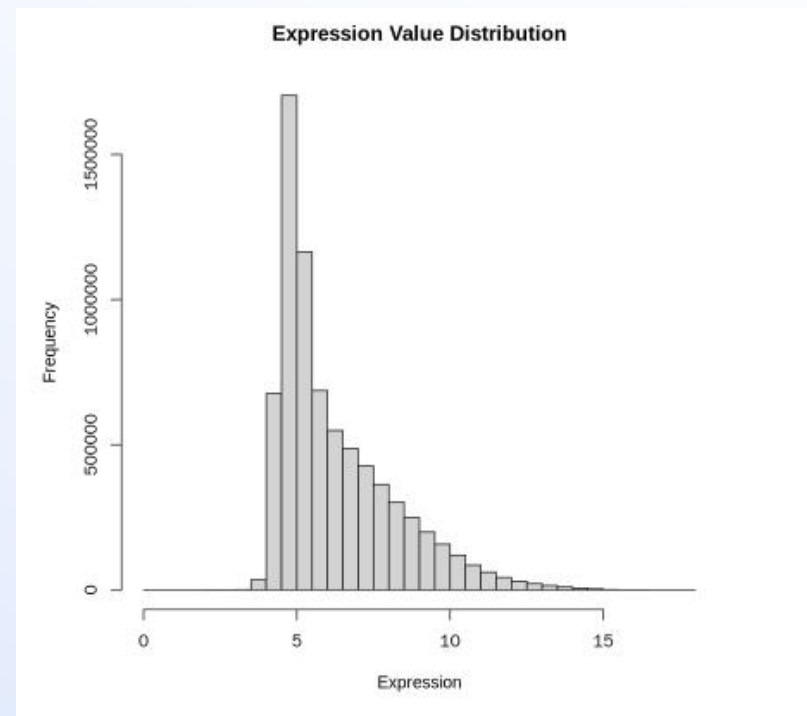
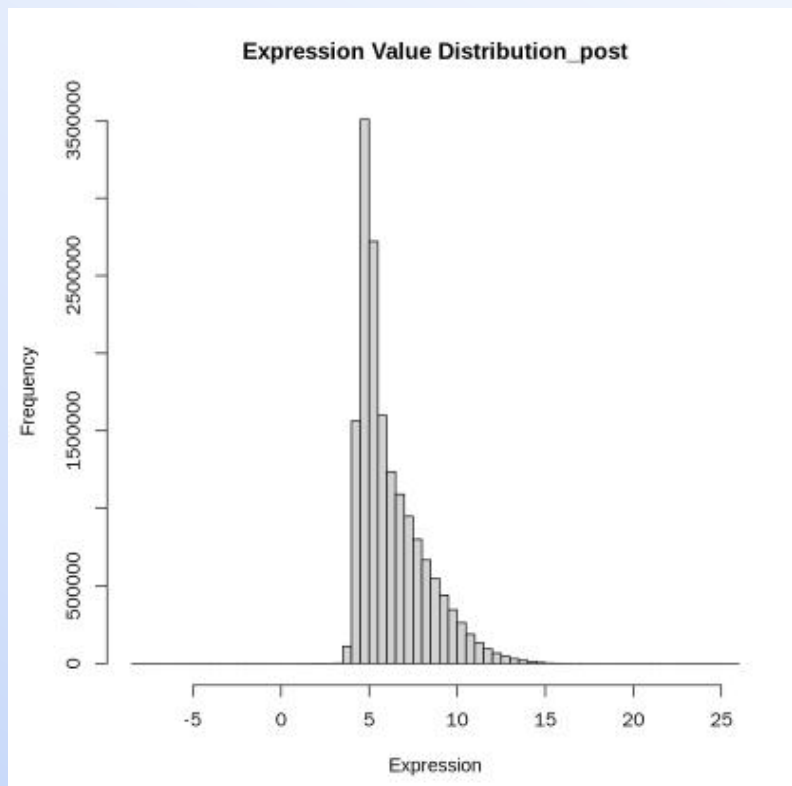
 endotype_phe.csv	XLS 工作表	2 KB	否	24 KB	93%
 endotypeA_limma_results.csv	XLS 工作表	944 KB	否	1,973 KB	53%
 endotypeA_vs_B_limma_results....	XLS 工作表	939 KB	否	1,970 KB	53%
 endotypeB_limma_results.csv	XLS 工作表	931 KB	否	1,966 KB	53%
 expr_combat.csv	XLS 工作表	124,229 KB	否	271,434 KB	55%
 fgsea_A_results.csv	XLS 工作表	8 KB	否	15 KB	50%
 fgsea_AB_results.csv	XLS 工作表	8 KB	否	15 KB	51%
 fgsea_B_results.csv	XLS 工作表	7 KB	否	13 KB	52%
 labels.csv	XLS 工作表	7 KB	否	64 KB	91%
 logistic_A_genes.csv	XLS 工作表	0 KB	否	0 KB	0%
 merged_expressiondata.csv	XLS 工作表	124,961 KB	否	270,947 KB	54%
 merged_phenotypedata.csv	XLS 工作表	4 KB	否	48 KB	94%

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
4. 89251966	GSM124917	GSM124917	GSM124917	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918	GSM124918
A1CF	5. 1599421	5. 1360120	5. 3984120	5. 3949558	5. 3268710	5. 1768131	5. 4495282	5. 6478724	5. 1376274	5. 2778742	5. 3133894	5. 4014854	5. 2683866	5. 3772249	5. 2017918	5. 3662614	5. 4517539	5. 3627051	5. 3159273	5. 3146
A2M	5. 2797692	4. 8504304	4. 9521822	4. 8384124	4. 9415731	4. 9049325	4. 9558575	4. 8094313	4. 9099042	4. 7835887	5. 0756932	4. 7519583	4. 7475468	5. 0336808	4. 9284595	4. 8635776	4. 9450261	4. 77	5. 2051	
A2ML1	4. 8314507	4. 8117835	4. 7613547	4. 7830728	4. 7645873	4. 8361120	4. 8732825	4. 8684295	4. 8735737	4. 8341274	4. 7795443	4. 8420506	4. 8070208	4. 7808976	4. 7956012	4. 8986500	4. 7607846	4. 8421632	4. 9406129	4. 8136
A3GALT2	5. 0163839	5. 1497354	4. 9422891	5. 0523205	5. 2730505	5. 2178572	5. 0334871	5. 2292797	4. 9935772	5. 1620622	5. 0210643	5. 2351364	5. 1319222	5. 0967429	5. 1956972	5. 1512059	5. 0803397	5. 0806133	5. 0710927	4. 9827
A4GALT	4. 9615745	5. 9239874	4. 8294528	4. 8527336	5. 0056341	5. 2444258	5. 7901186	5. 3161382	4. 8027996	16. 1534823	4. 9287853	5. 8259443	5. 6981728	4. 9317685	5. 2916104	5. 8229008	7. 5217262	5. 0415627	4. 8804237	4. 8339
A4GNT	5. 8826160	6. 1307460	6. 0705088	5. 9160807	6. 8581367	6. 8550738	5. 9687826	6. 3906293	6. 0263126	6. 4001411	6. 0688320	6. 0500668	5. 6740597	6. 0938705	5. 9085266	5. 7111448	6. 7785577	6. 6644878	5. 6437481	5. 9555
AAAS	6. 0734514	5. 4489628	6. 1222499	5. 9224011	5. 3954749	5. 0843918	5. 9505899	5. 0507698	6. 1306027	5. 2548423	6. 3969763	5. 2483511	5. 5645858	5. 3584328	5. 5676683	5. 3468635	5. 3286984	5. 4827663	5. 5215800	5. 4610
AACS	6. 4539097	5. 9102757	6. 0367000	6. 5446767	6. 2354817	5. 9347806	6. 4372661	5. 9221696	6. 5327110	5. 6462042	6. 9788257	6. 4070179	6. 3209544	6. 1232322	6. 7716678	5. 9128611	6. 3309888	6. 4922317	6. 3669917	5. 8223
AADAC	5. 3399274	5. 0522400	5. 1260408	5. 2069339	5. 5421235	5. 1126495	5. 0404694	5. 2225937	5. 4011692	5. 2170103	5. 2545703	5. 2684676	5. 0358026	5. 2005328	5. 2170103	5. 0364599	5. 0809556	4. 9558330	5. 1909918	5. 1526
AADACL2	5. 0876554	5. 3331056	5. 2263427	5. 1759100	5. 1658057	5. 1634494	5. 2638343	4. 7970629	5. 3418261	5. 1906879	4. 9001735	5. 1183350	5. 2072495	5. 2245954	5. 2497201	5. 1233072	5. 0658536	5. 0408385	4. 9993914	5. 1474
AADACL3	5. 3187705	5. 4583049	5. 7809586	5. 3708998	5. 4502366	5. 4296154	5. 6625989	5. 6362673	5. 2391310	5. 4710067	5. 3007795	5. 4575739	5. 4504809	5. 5902243	5. 3525988	5. 7536044	5. 3731390	5. 4459429	5. 5563360	5. 4075
AADACL4	5. 6955480	5. 3167258	5. 0616000	5. 5234675	5. 0273252	7. 1696339	5. 2587513	7. 0368907	5. 0935045	5. 5543583	5. 4373846	5. 5747795	5. 2500956	5. 3474230	5. 2319609	5. 7509136	5. 5209857	5. 5035624	5. 2602790	6. 3095
AADAT	5. 1283371	5. 0223150	4. 9982312	5. 0536442	5. 1896530	5. 0195891	5. 0211125	5. 0690515	5. 1695720	5. 0599158	5. 0107934	5. 1364451	5. 1921853	5. 1488354	5. 2467773	5. 0675573	5. 0858646	4. 9760738	5. 1902426	5. 1281
AAK1	6. 4939080	6. 6333664	6. 4665453	6. 5555168	5. 5284926	6. 7722230	6. 4671777	6. 7005940	6. 4228095	6. 3068160	6. 9965575	5. 5040325	6. 5965485	6. 6131568	6. 5259191	6. 7278067	6. 6089790	6. 7272359	6. 4852759	6. 6615
AAMDC	8. 1417395	8. 2659479	7. 9712104	8. 8292958	8. 4752235	8. 1447427	8. 1585666	8. 1033472	8. 3080633	8. 5930421	8. 3410675	8. 3309875	8. 3275063	8. 2100126	8. 4232689	8. 2211676	8. 0624347	8. 5782763	8. 1170801	7. 7312
AAMP	8. 1704786	7. 8144508	8. 1507738	7. 9108110	8. 5462911	7. 7045942	7. 8917079	7. 6942762	8. 3431014	8. 6085784	8. 5155425	8. 6740935	8. 5672526	7. 8150770	8. 2475418	8. 1524630	7. 9455036	8. 2034716	8. 6598937	7. 6955
AANAT	4. 7929585	4. 8874291	4. 7806606	4. 7578960	4. 7594133	4. 8438990	4. 9205722	4. 7953427	4. 9224836	4. 8233168	4. 8264607	4. 8830225	4. 7397235	4. 8037468	4. 7676663	4. 7193734	4. 7832154	4. 9024558	4. 7100922	4. 9476
AAR2	9. 838256	9. 5565295	9. 8347093	9. 3023581	9. 3833181	9. 3880552	9. 7545936	9. 4263385	10. 1921179	9. 4819283	10. 333894	9. 7158394	10. 084698	9. 0607043	9. 6427515	9. 5428972	9. 6009727	9. 6802542	9. 751805	8. 9536
AARD	5. 0190066	4. 9752917	4. 9456294	5. 1984805	5. 0564457	5. 0741103	4. 8142774	4. 9060383	5. 0524011	5. 0056134	4. 9339769	4. 9639040	4. 8390936	5. 0808592	4. 9301045	5. 2147879	4. 8984326	4. 9987242	5. 0318344	4. 9066
AARS	11. 115468	11. 358065	11. 330005	10. 718054	10. 306136	9. 6523782	11. 128591	10. 141234	10. 944349	10. 264913	11. 938809	10. 999204	11. 038585	10. 417240	10. 818326	10. 797144	10. 805552	10. 458181	11. 193961	10. 422
AARS2	7. 9091726	8. 3006950	7. 4885057	7. 7971178	7. 5517520	6. 3067512	7. 8053753	6. 6373338	8. 1933256	7. 0476738	8. 1975782	7. 6830497	7. 5893702	6. 2134939	7. 7787669	7. 7920331	7. 6324798	8. 1118522	7. 9815156	7. 8249
AASDH	8. 0145673	7. 3719735	7. 9607465	7. 9586027	7. 8929912	8. 1607984	7. 9648948	7. 3899472	7. 6989529	7. 7761749	6. 9982116	7. 5433183	7. 8441983	7. 4363144	7. 4820931	7. 8749230	7. 5899172	8. 0090131	7. 5357043	8. 9896
AASDPPT	8. 1620488	8. 4063887	8. 5620912	8. 5589110	8. 8224066	8. 3313329	8. 2805632	8. 644786	9. 1217376	8. 6470658	7. 7579127	7. 9663742	9. 0239414	8. 8024702	8. 8091694	8. 7749534	8. 1787817	8. 1427993	8. 6959113	9. 1183



数据集介绍

更具体的，数据集中有968个样本，经过Combat去批次，所有数据集**取交集**（GSE147689-91**除外**，每个样本有16992个基因。基因表达情况通过数据分析如下



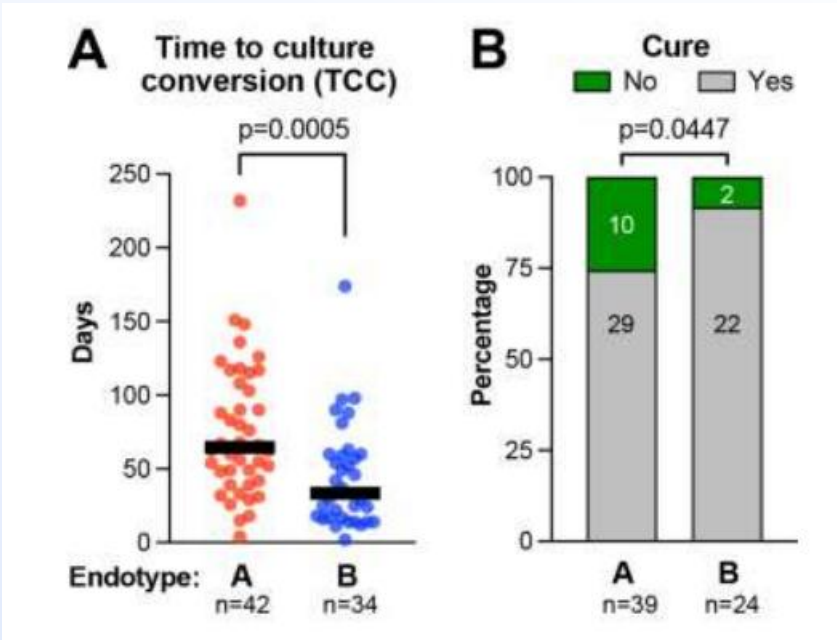
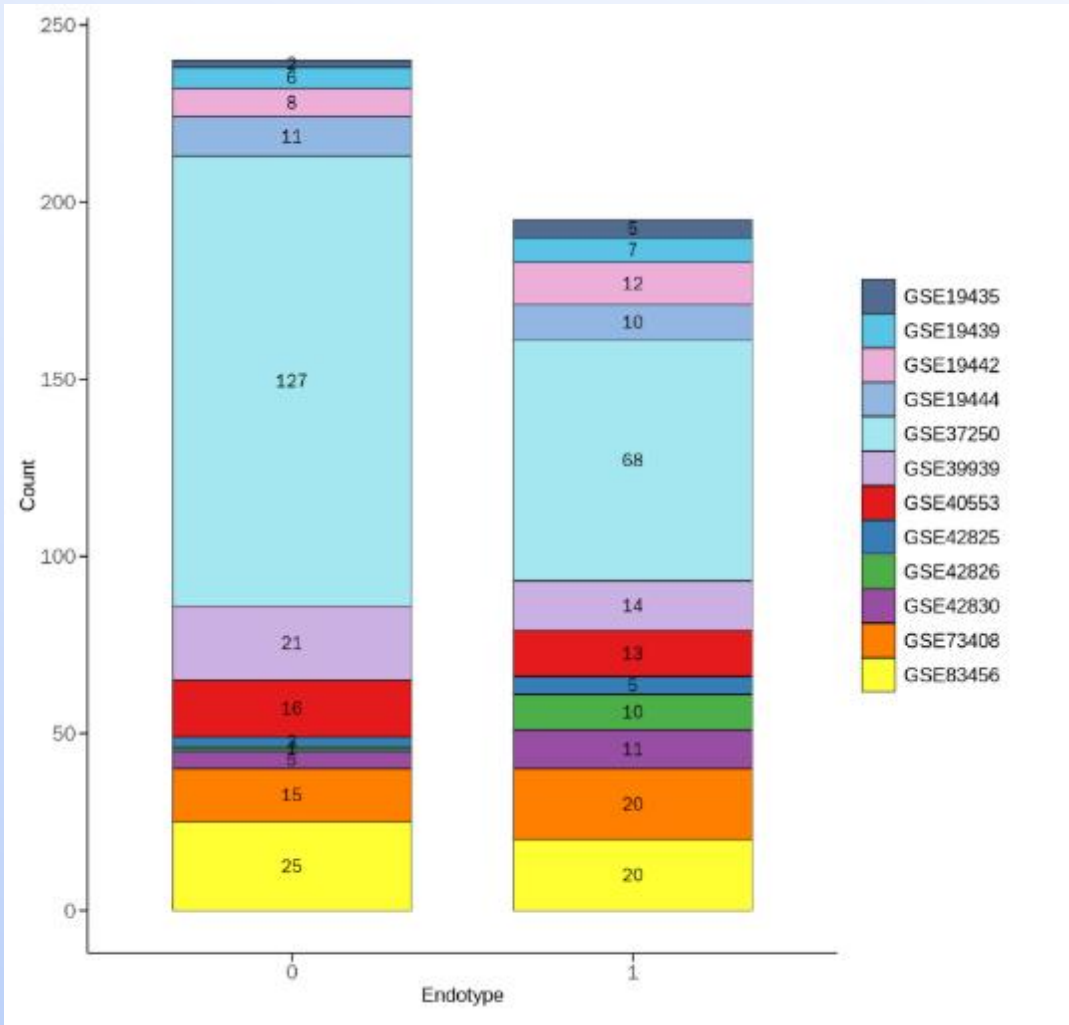
HC相对比较正常，主要集中在5-15之间，有一些异常值，可能是噪音引起的或者数据本身就有错误，当然也可能是高表达基因。

TB-endotype:仅保留TB的样本数据
435个



数据集介绍

更具体的，有一篇论文利用一系列机器学习方法对其进行了标签分配（论文：Gene expression signatures identify biologically and clinically distinct tuberculosis endotypes），我们的目标是利用建立两个对A和B进行分类的二分类模型，融合成一个TB VS HC，从而更好地进行识别结核病





PART THREE

关键基因筛选的预处理

基因组学分析的常用方法

所有真核物种的分析方法相同

1. 基因表达分析

- (1) 基因表达定量 (Count值和FPKM值)
- (2) 样本间皮尔逊相关性分析
- (3) 主成分分析 (PCA)
- (4) 基因表达分布 (箱型图)

2. 差异基因分析

- (1) Deseq2差异基因分析、统计、筛选
- (2) 差异基因火山图
- (3) 差异基因聚类热图

模式物种有现成的Orgdb注释包，其它非模式物种需要自行做GO和KEGG注释

3. 富集分析 (包含模式物种和非模式物种)

- (1) GO富集分析
- (2) KEGG富集分析

所有真核物种的分析方法相同

4. PPI蛋白互作网络分析

5. GSEA分析

差异分析:

基因名称		实验组样本			对照组样本			1.FC	2.P值	3.FDR		
#	A	B			C	D	E	F	G	H	I	J
	Genes	Treat1	Treat2	Treat3	Control1	Control2	Control3	FC	p	FDR		
1	Gene1	1.2768	1.3997	1.0724	0.8805	0.8286	0.6947	1.5596	0.0153	0.0406		
2	Gene2	0.0195	0.0271	0.0147	0.0442	0.0574	0.0594	0.3806	0.0051	0.0194		
3	Gene3	2.3951	3.0794	2.2560	2.4159	2.4866	2.4357	1.0535	0.6350	0.6981		
4	Gene4	0.0832	0.1432	0.1569	0.0833	0.1725	0.0685	1.1814	0.6460	0.7063		
5	Gene5	0.0274	0.0102	0.0206	0.0500	0.0732	0.0690	0.3027	0.0069	0.0237		
6	Gene6	0.0154	0.0145	0.0135	0.0089	0.0228	0.0121	0.9921	0.9800	0.9800		
7	Gene7	0.0272	0.0620	0.0839	0.7527	0.6673	0.6921	0.0819	0.0000	0.0008		
8	Gene8	0.5025	0.4246	0.4703	0.5337	0.3255	0.4817	1.0422	0.7910	0.8366		
9	Gene9	0.8548	1.0108	0.8316	1.2080	0.9226	1.0605	0.8452	0.1740	0.2546		

1.什么是FC?

基因表达量

翻译成中文是差异倍数(FoldChange), 也被称作Ratio。具体而言, FC是指基因在一组样品中的表达值的均值除以其在另一组样品中的表达值的均值。

2.什么是Pvalue?P值是统计检验中用于衡量是否存在统计差异显著的一个关键数值。它就像一个“信号灯”, 为我们指示数据差异的可信度。在科研领域, 约定成俗的标准是P-value<0.05为统计检验显著。

3.什么是FDR? FDR(falsediscoveryrate), 即校正后的P值, 中文一般译作错误发现率。在进行大规模的数据分析时, 我们往往会同时检验多个假设, 这就增加了误判的可能性, 也就是所谓的假阳性结果。

我们通常会综合考虑FC、P值或FDR这三个指标。一般来说, 一个基因要被判定为差异基因, 需要同时满足FC符合设定的阈值(如FC>2为上调或FC<1/2为下调1、P值小于0.05或FDR小于0.05。当然这些阈值要根据实际情况做相应的调整, 常见的FC会卡1.2,1.5.2等, p常见的有0.05.0.01等。注:P值和FDR卡一个就行了, 从严格意义上讲卡FDR, 会更好, 但是有时候卡FDR的话, 一个差异基因也筛选不出来, 要根据实际情况做相应的变动

我的处理

Input Files

The following files are used for limma analysis:

- limma_A_vs_NonA.csv
- limma_B_vs_NonB.csv
- limma_TB_vs_HC.csv

Key Steps

1. Extract significant differentially expressed genes based on the criteria:

$$|\log FC| > 1 \quad \text{and} \quad \text{adj.P.Val} < 0.05.$$

2. Calculate the importance score for each gene:

$$\text{Score}_{\text{limma}} = |\log FC| \times (-\log_{10}(\text{adj.P.Val})).$$

3. Normalize the scores to the range [0, 1] using Min-Max normalization:

$$\text{Normalized_Score}_{\text{limma}} = \frac{\text{Score}_{\text{limma}} - \min(\text{Score}_{\text{limma}})}{\max(\text{Score}_{\text{limma}}) - \min(\text{Score}_{\text{limma}})}.$$

4. Merge normalized scores from different comparisons into a single table:

Gene	Score_limma_A_vs_NonA	Score_limma_B_vs_NonB	Score_limma_TB_vs_HC
Gene1	0.8	0.6	0.9
Gene2	0.7	0.4	0.8

将A VS NonA,B vs NonB,TB VS HC的差异分析结果进行预处理：形成一个基因重要性矩阵，方便后续构建模型的时候筛选基因。

```
data_Pre_Processing_Code > limma_A_vs_NoA.csv > data
1  Gene,Score_limma_A_vs_NonA
2  ANKRD22,1.0
3  FCGR1A,0.9620650456163328
4  BATF2,0.8820884080732608
5  GBP6,0.8036131415092733
6  FCMR,0.7672955080633963
7  CACNA1E,0.671565224266308
8  LHFPL2,0.6689957951520057
9  KLHL3,0.6376860124376705
10 CD6,0.6356447306753965
11 TNFRSF25,0.6085304065521967
12 CD274,0.5947628801887305
13 GZMK,0.5863892567678611
14 PSTPIP2,0.5857586204312476
15 CD96,0.5855997035847926
16 CARD17,0.5798306394640453
17 ID3,0.5765697393342225
18 KLF12,0.5708487330440876
19 SKAP1,0.560660541333058
```



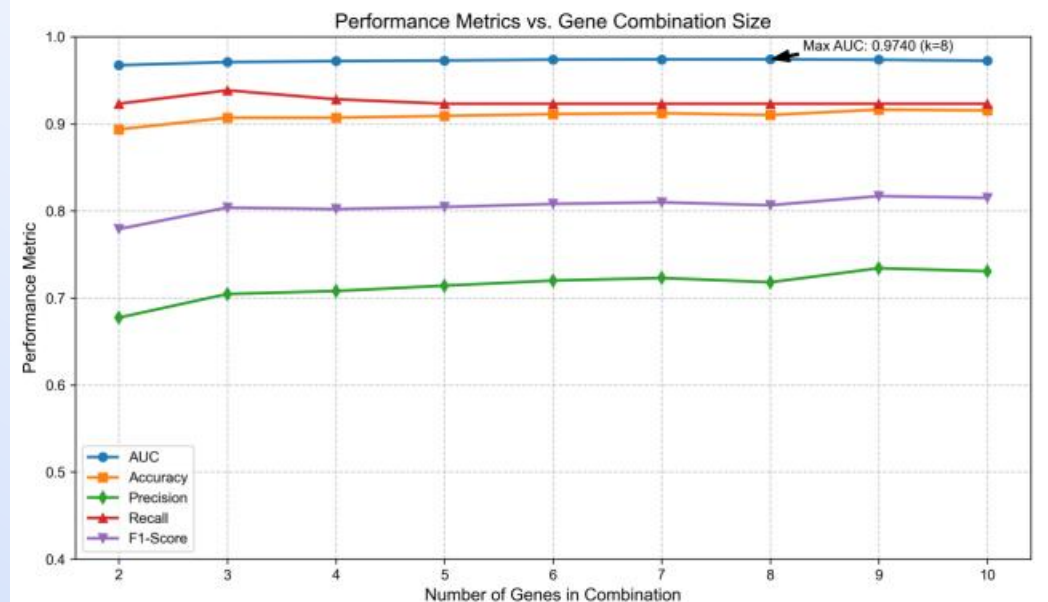
PART FOUR

模型构建

A模型构建 (A VS Non A 二分类模型)

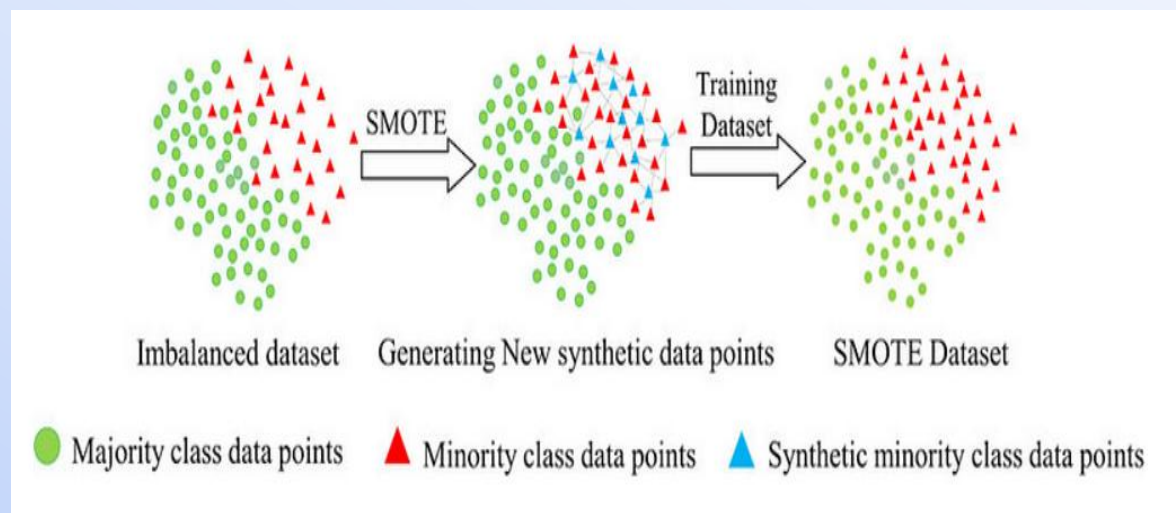
A和非A具有显著的差异，这里选用逻辑回归方法得到的效果比较好，主要的原因是A类别和（HC+B）差异比较明显,选取limma分析筛选的阈值等于0.5,0.6,0.7，甚至0.9都能得到比较好的效果，甚至只选取SCore_limma_A_vs_Non_A中的前两个基因都能得到比较好的效果。当然我们要的是使用尽可能少的基因去获取这个效果。因此这里我选取前40个基因，通过逻辑回归得到前10重要性基因，在对其进行组合C（10，K）（ $2 \leq K \leq 10$ ），从中挑选出最好的，最终发现仅用两个基因也可以达到比较好的结果。

一般情况下，逻辑回归会得到系数的值，
系数绝对值越大，
说明这个特征越重要,若系数为正，
则这个特征与目标值为1的概率正相关。
若系数为负,则说明这个特征与目标值为0的概率正相关。
最终挑选出来的基因是



B模型的构建

B基因在分类的时候尝试了逻辑回归，随机森林，支持向量机等方法，以及它们的集成方法（这种方法比较容易过拟合，在验证集效果非常好（可能分类完全正确那种），但在测试集中很差），最后选用了随机森林方法，随机森林在选用基因量还尝试了用组合方法拓展基因（比如选出的基因表达量两两之间相除作为一个新的特征作为输入，效果不是很好）。下面是最好的一次B的效果，AUC大概在0.87-0.94之间吗，但运用基因数量过多：



筛选到 28 个显著基因 (Score > 0.3)

类别分布：

0 730

1 242

应用SMOTE过采样...

过采样后类别分布：[730 730]

===== 模型评估 =====

准确率 (Accuracy): 0.8454

ROC AUC: 0.8923

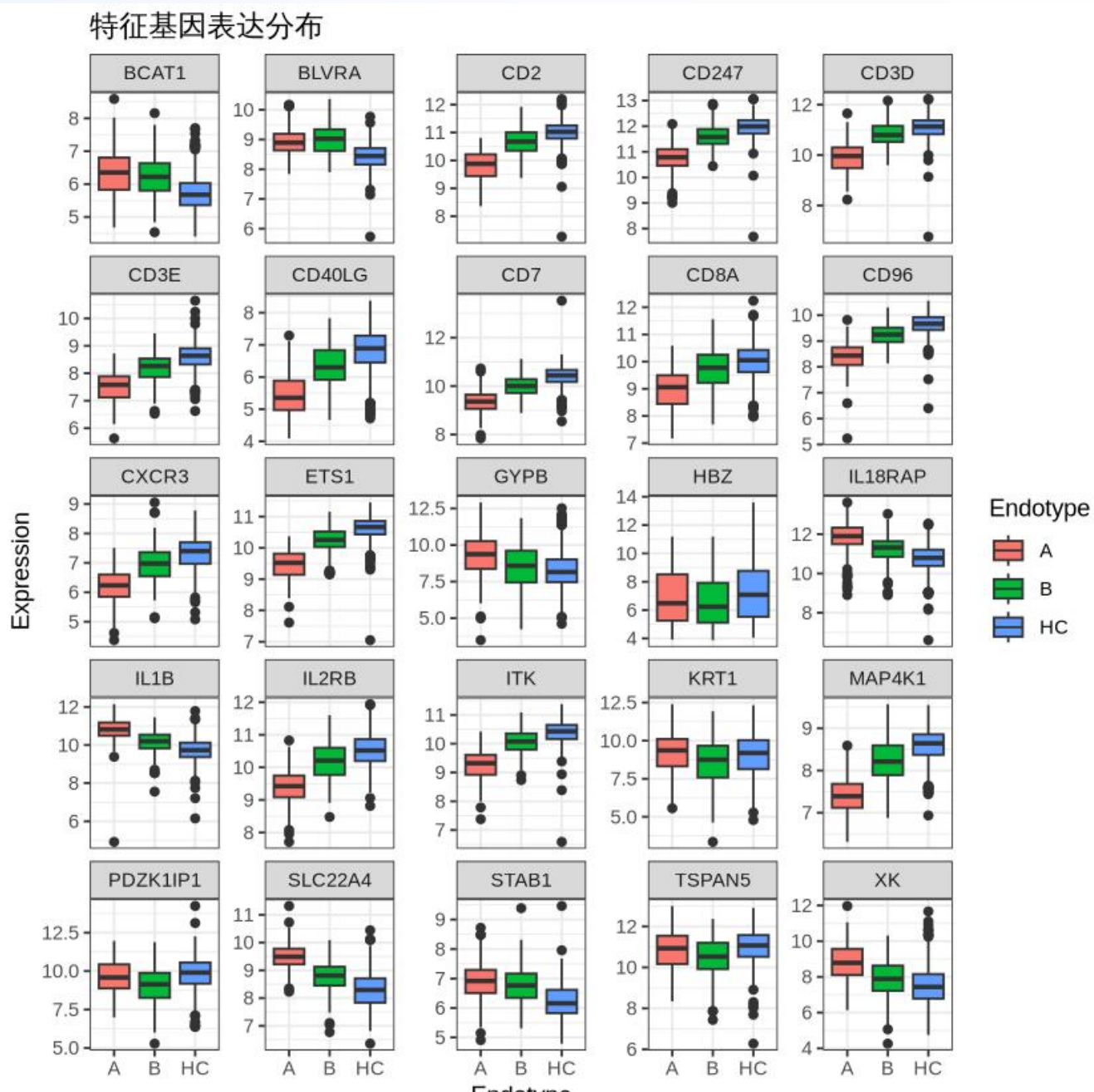
分类报告：

	precision	recall	f1-score	support
0	0.89	0.91	0.90	146
1	0.72	0.67	0.69	48

优化后分类报告 (阈值=0.3)：

	precision	recall	f1-score	support
0	0.91	0.85	0.88	146
1	0.65	0.77	0.70	48

B模型构建不好的原因分析



B模型构建不好的原因分析

误分类B 和 正确分类B 是两个不同的类别，它们分别代表了被错误地归类为B类和正确地归类为B类的样本。

从箱线图中分析出（这里没放，图放不下），误分类B的中位数大约在7.2左右，而正确分类B的中位数则略低，大约在7.0左右。

误分类B的四分位范围（IQR）较宽，表明其数据分布较为分散；而正确分类B的四分位范围相对较窄，说明其数据分布较为集中。

误分类B存在一个明显的异常值（标记为圆圈），数值接近9.5，这可能是一个极端值，对整体分布有一定的影响。

特征 GBP6（B类与HC类对比）

这里比较的是 误分类B 和 HC类 的特征GBP6分布。

误分类B的中位数大约在7.5左右，而HC类的中位数则显著较低，大约在5.8左右。

误分类B的数据分布相对HC类更为分散，四分位范围更宽，而HC类的数据分布较为集中。

HC类中也存在一些异常值，数值接近8.0，这些异常值可能会影响HC类的整体分布特征。

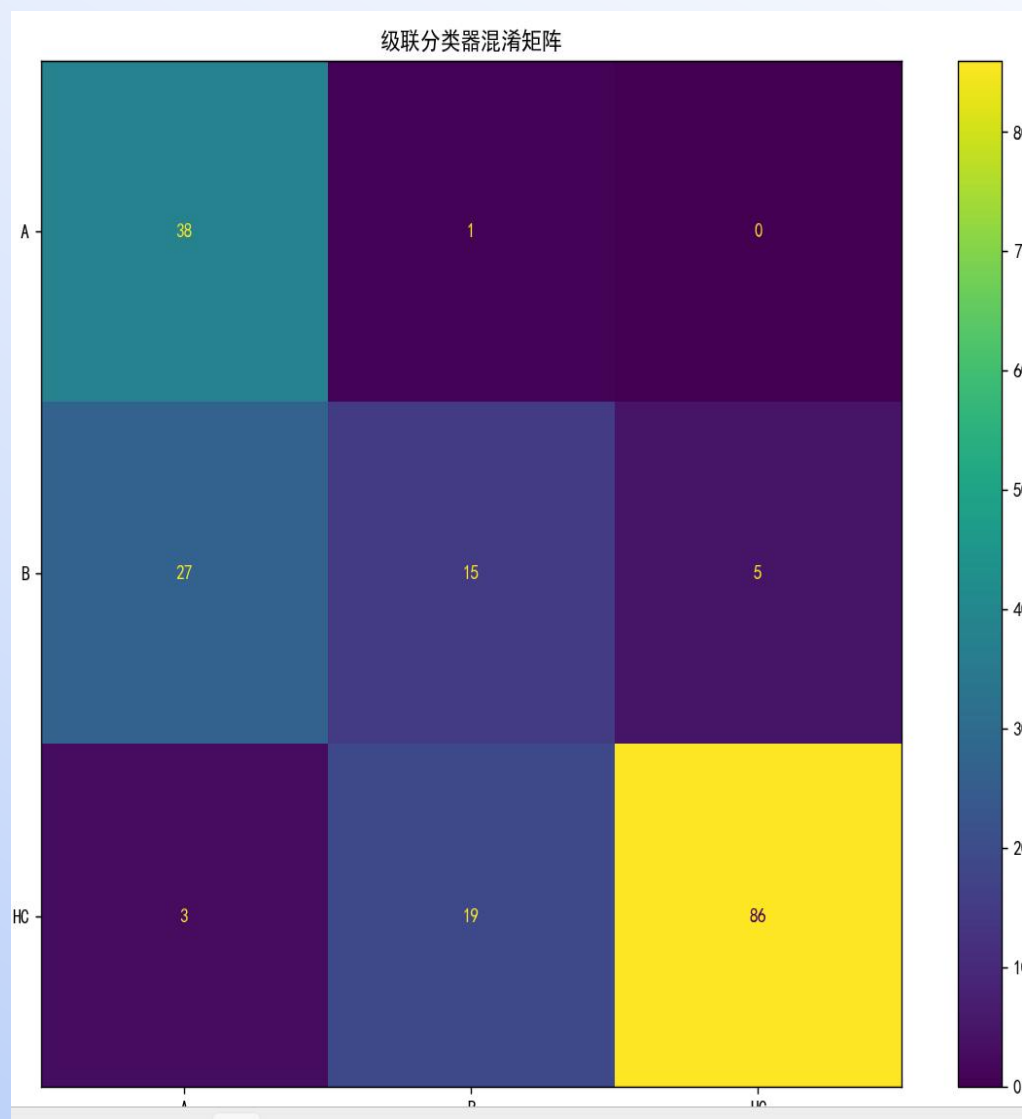
总结：主要是基因表达值和其它类别比较相似，不好区分，另外是异常值影响整体分布



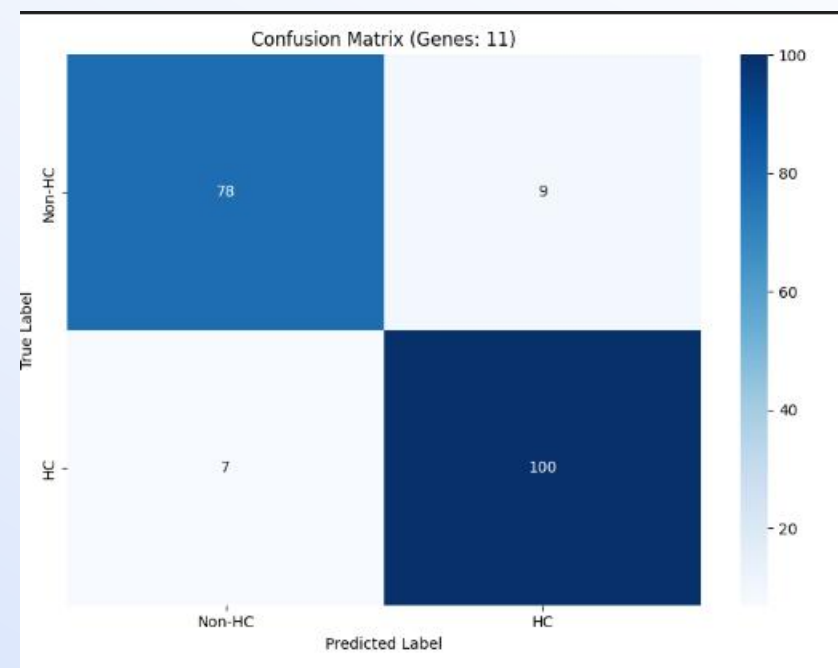
PART FIVE

最终融合效果对比

融合模型效果对比



融合之后的模型效果并不是很好，但是所选出来的11个基因直接去训练HC vs (A+B)具有相对较好的效果：

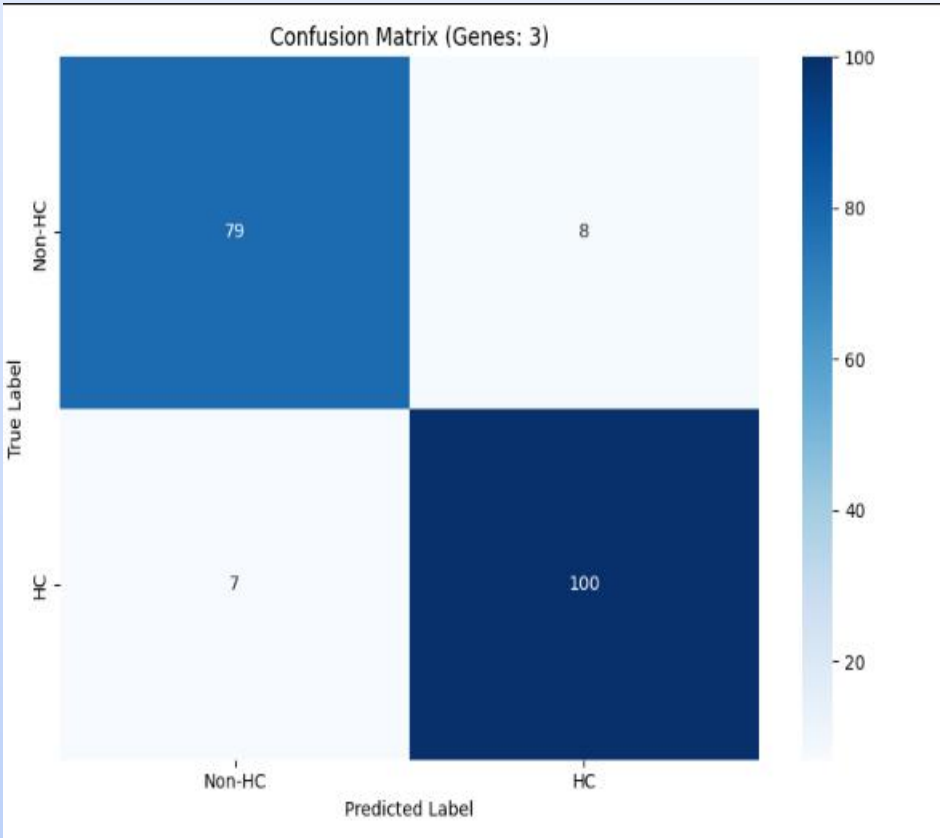


Accuracy,ROC_AUC

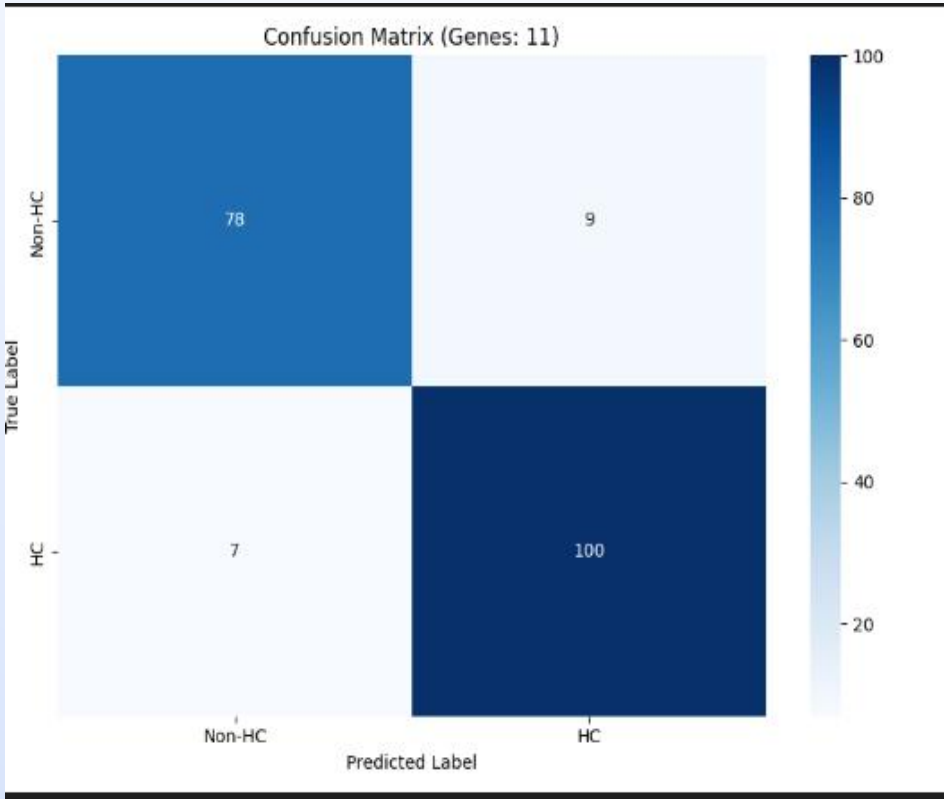
0.9175257731958762,0.970190138575572

已有论文筛选出来的基因效果验证

分别是3基因， 6基因（准确说是4基因， 这里缺失两个）， 12基因（其实是11基因， 因为A模型筛选出来的基因和B模型筛选出来的基因有一个重合）以及20基因的效果展示， 其中3基因的效果不论在哪种方法上都是最好的。 一共做4*4个16组实验， 其次是我们11基因的单模型， 但是我们融合模型效果比较差， 远远比不上三基因



三基因模型混淆矩阵



11基因模型混淆矩阵

谢谢

汇报人: 朱佳成