Analysis

Huang LiChuang of Wie-Biotech

Contents

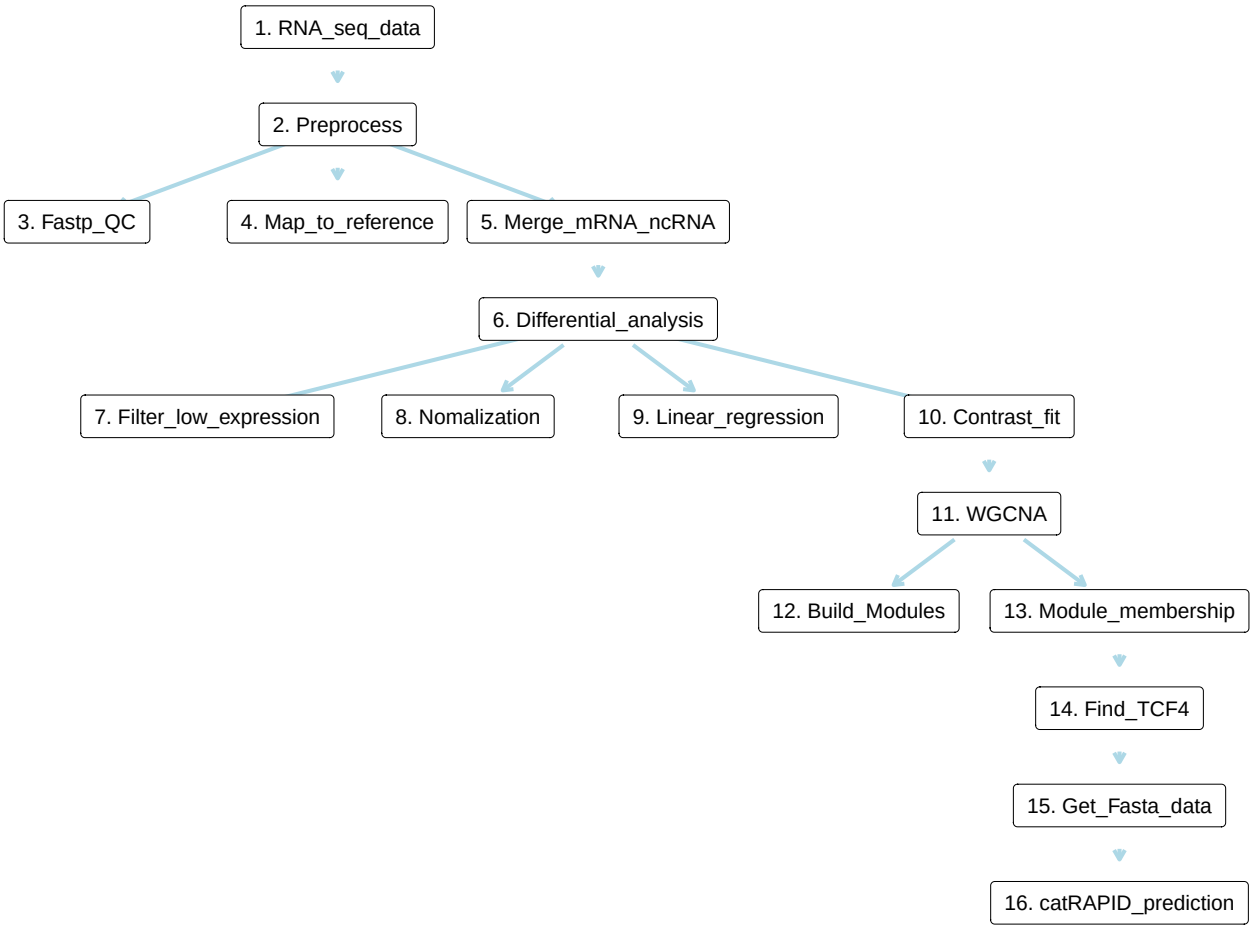
List of Figures

List of Tables

1 摘要

根据客户需求和提供的数据，筛出（癥痕增生）能够与 TCF-AS1 结合又能与 TCF4 结合的 RNA 结合蛋白。结果请参考 5

2 研究设计流程图



3 材料和方法

1. 获取参考注释基因。
2. 初步处理客户提供的数据（fastp 质控、kallisto[1] 对比到参考基因座等）。
3. 使用 limma 差异分析[2]。
4. 使用 WGCNA 方法[3]，从差异基因中筛选与 TCF4-AS1 lncRNA 和 TCF4mRNA 具有共表达关系的基因。
5. 视情况选择合适的预测工具[4–7]，预测蛋白和 RNA 的结合程度，并可视化为图表。

4 分析结果

4.1 下载参考基因组注释文件

下载 cDNA 和 ncRNA 参考基因注释。https://ftp.ensembl.org/pub/release-110/fasta/homo_sapiens/

4.2 fastq 预处理

4.2.1 数据质控

使用 fastp 去接头和去低质量的碱基

此为 fastp 处理时生成的报告文件。'Reports fastq files processed with fastp' 数据已全部提供。

(对应文件为 `./fastp_report`)

> 注：文件夹./fastp_report 共包含 6 个文件。
>
> 1. CT1-CT1_combined_R.html
> 2. CT2-CT2_combined_R.html
> 3. CT3-CT3_combined_R.html
> 4. CUR1_R.html
> 5. CUR2_R.html
> 6. ...

4.3 使用 kallisto 比对 fastq 到参考基因座

kallisto 提供了快速且准确的 fastq 比对到参考基因座的方法[1] (http://pachterlab.github.io/kallisto/manual.html)。

4.3.1 鉴定 mRNA

使用 kallisto 将 fastq 与 hg38 的 cDNA 数据比对。

主要为子目录下的 abundance.tsv 文件。'Refer to mRNA' 数据已全部提供。

(对应文件为 `./quant_hg38_mrna`)

> 注：文件夹./quant_hg38_mrna 共包含 6 个文件。
>
> 1. CT1-CT1_combined_R
> 2. CT2-CT2_combined_R
> 3. CT3-CT3_combined_R
> 4. CUR1_R
> 5. CUR2_R
> 6. ...

4.3.2 鉴定 ncRNA

使用 kallisto 将 fastq 与 hg38 的 ncRNA 数据比对。

主要为子目录下的 abundance.tsv 文件。'Refer to ncRNA' 数据已全部提供。

(对应文件为 `./quant_hg38_ncrna`)

注：文件夹./quant_hg38_ncrna 共包含 6 个文件。

1. CT1-CT1_combined_R
2. CT2-CT2_combined_R
3. CT3-CT3_combined_R
4. CUR1_R
5. CUR2_R
6. ...

4.4 差异分析

4.4.1 读取并合并不同样本 RNA 定量数据

Table 1为表格 merged mrna 概览。

(对应文件为 `Figure+Table/merged-mrna.csv`)

注：表格共有 207249 行 7 列，以下预览的表格可能省略部分数据；表格含有 207249 个唯一 'target_id'。

Table 1: Merged mrna

targe...	CT1-C...	CT2-C...	CT3-C...	CUR1_R	CUR2_R	CUR3_R
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
...	...	...	...	...	...	...

Table 2为表格 merged ncrna 概览。

(对应文件为 `Figure+Table/merged-ncrna.csv`)

注：表格共有 68492 行 7 列，以下预览的表格可能省略部分数据；表格含有 68492 个唯一 'target_id'。

Table 2: Merged ncrna

targe…	CT1-C…	CT2-C…	CT3-C…	CUR1_R	CUR2_R	CUR3_R
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	2496.34	3798.65	11014.8	2845.62	3811.47	10121.3
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0.166667	0	0.333333	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0.5	0	0	0	0
ENST0…	3	7	5	4	6	9
ENST0…	0	0	0	0	0	0
…	…	…	…	…	…	…

4.4.2 合并 mRNA 和 ncRNA 数据

在这里，将 mRNA 数据和 ncRNA 数据按照列（样品）合并。

Table 3为表格 merged data of mRNA and ncRNA 概览。

(对应文件为 `Figure+Table/merged-data-of-mRNA-and-ncRNA.csv`)

注：表格共有 275741 行 7 列，以下预览的表格可能省略部分数据；表格含有 275741 个唯一 'target_id'。

Table 3: Merged data of mRNA and ncRNA

targe…	CT1-C…	CT2-C…	CT3-C…	CUR1_R	CUR2_R	CUR3_R
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0
ENST0…	0	0	0	0	0	0

targe...	CT1-C...	CT2-C...	CT3-C...	CUR1_R	CUR2_R	CUR3_R
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
ENST0...	0	0	0	0	0	0
...	...	...	...	...	...	...

4.4.3 使用 `biomaRt` 获取基因注释

使用 R 包 `biomaRt` 获取 mRNA 和 ncRNA 的注释。

Table 4为表格 annotation mRNA 概览。

(对应文件为 `Figure+Table/annotation-mRNA.tsv`)

> 注：表格共有 275741 行 8 列，以下预览的表格可能省略部分数据；表格含有 275741 个唯一 'ensembl_transcript_id'。

Table 4: Annotation mRNA

ensem......1	ensem......2	entre...	hgnc_...	chrom...	start...	end_p...	descr...
ENST0...	ENSG0...	4535	MT-ND1	MT	3307	4262	mitoc...
ENST0...	ENSG0...	4536	MT-ND2	MT	4470	5511	mitoc...
ENST0...	ENSG0...	4512	MT-CO1	MT	5904	7445	mitoc...
ENST0...	ENSG0...	4513	MT-CO2	MT	7586	8269	mitoc...
ENST0...	ENSG0...	4509	MT-ATP8	MT	8366	8572	mitoc...
ENST0...	ENSG0...	4508	MT-ATP6	MT	8527	9207	mitoc...
ENST0...	ENSG0...	4514	MT-CO3	MT	9207	9990	mitoc...
ENST0...	ENSG0...	4537	MT-ND3	MT	10059	10404	mitoc...
ENST0...	ENSG0...	4539	MT-ND4L	MT	10470	10766	mitoc...
ENST0...	ENSG0...	4538	MT-ND4	MT	10760	12137	mitoc...
ENST0...	ENSG0...	4540	MT-ND5	MT	12337	14148	mitoc...
ENST0...	ENSG0...	4541	MT-ND6	MT	14149	14673	mitoc...
ENST0...	ENSG0...	4519	MT-CYB	MT	14747	15887	mitoc...
ENST0...	ENSG0...	10272...		KI270...	4612	29626	
ENST0...	ENSG0...	10272...		KI270...	4612	29626	

ensem……1	ensem……2	entre…	hgnc_…	chrom…	start…	end_p…	descr…
…	…	…	…	…	…	…	…

4.4.4 使用 `limma` 差异分析

Figure 1为图 filter low expression genes 概览。

(对应文件为 `Figure+Table/filter-low-expression-genes.pdf`)

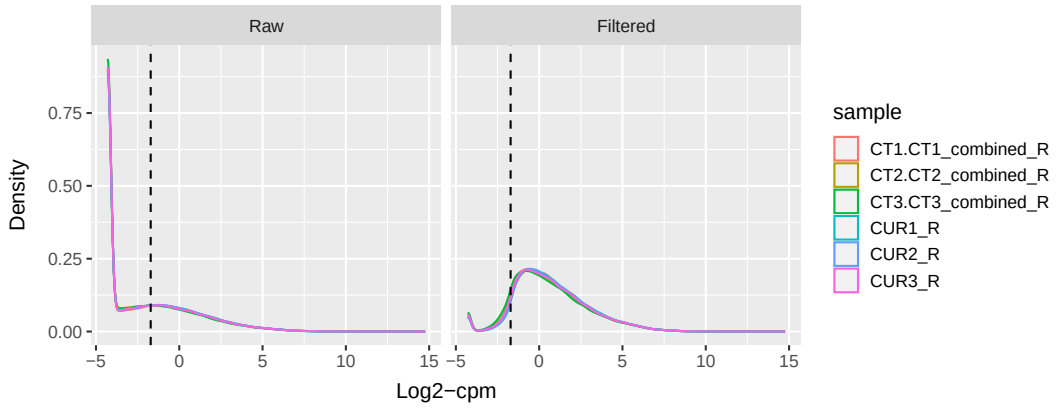


Figure 1: Filter low expression genes

Figure 2为图 nomalize genes expression 概览。

(对应文件为 `Figure+Table/nomalize-genes-expression.pdf`)

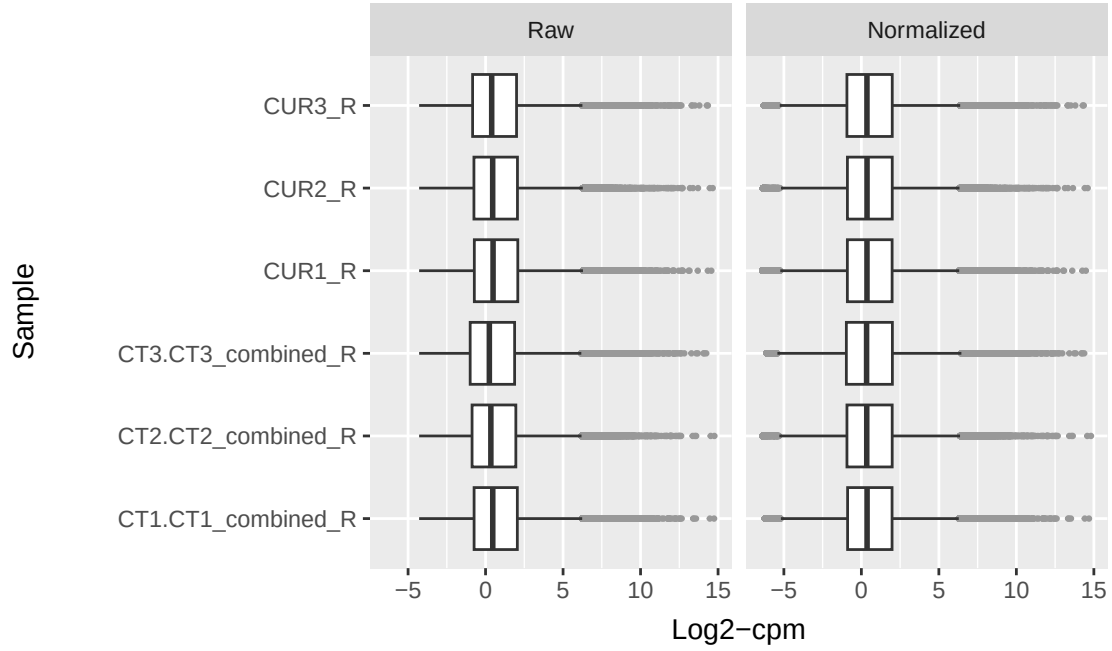


Figure 2: Nomalize genes expression

线形回归拟合 `model.matrix(~ 0 + group)`[2,8]，并统计检验。

Table 5为表格 linear regression and contrast fit results 概览。根据 P.Value (0.05) 和 |log2FC| (0.3) 过滤得到结果。

(对应文件为 `Figure+Table/linear-regression-and-contrast-fit-results.xlsx`)

> 注：表格共有 7204 行 14 列，以下预览的表格可能省略部分数据；表格含有 7204 个唯一 'ensembl_transcript_id'。

Table 5: Linear regression and contrast fit results

ensem……1	ensem……2	entre…	hgnc_…	chrom…	start…	end_p…	descr…	logFC	AveExpr	…
ENST0…	ENSG0…	23657	SLC7A11	4	13816…	13824…	solut…	1.096…	6.693…	…
ENST0…	ENSG0…	2316	FLNA	X	15434…	15437…	filam…	-0.60…	10.38…	…
ENST0…	ENSG0…	1728	NQO1	16	69706996	69726668	NAD(P…	1.540…	7.379…	…
ENST0…	ENSG0…	3486	IGFBP3	7	45912245	45921874	insul…	-1.77…	5.925…	…
ENST0…	ENSG0…	3880	KRT19	17	41523617	41528308	kerat…	-1.40…	5.649…	…
ENST0…	ENSG0…	NA		16	69709874	69710583	novel…	1.519…	4.998…	…
ENST0…	ENSG0…	682	BSG	19	571277	583494	basig…	-1.89…	6.649…	…
ENST0…	ENSG0…	3488	IGFBP5	2	21667…	21669…	insul…	-1.63…	6.279…	…
ENST0…	ENSG0…	128239	IQGAP3	HG251…	83962	131161	IQ mo…	-1.08…	5.548…	…
ENST0…	ENSG0…	128239	IQGAP3	1	15652…	15657…	IQ mo…	-1.08…	5.548…	…
ENST0…	ENSG0…	1728	NQO1	16	69706996	69726668	NAD(P…	1.374…	5.529…	…
ENST0…	ENSG0…	4176	MCM7	7	10009…	10010…	minic…	-1.04…	6.143…	…
ENST0…	ENSG0…	9537	TP53I11	11	44885903	44951306	tumor…	-1.01…	5.866…	…
ENST0…	ENSG0…	1728	NQO1	16	69706996	69726668	NAD(P…	1.757…	3.238…	…
ENST0…	ENSG0…	994	CDC25B	20	3786772	3806121	cell …	-0.80…	6.245…	…
…	…	…	…	…	…	…	…	…	…	…

Figure 3为图 volcano plot of differential expression genes 概览。

(对应文件为 `Figure+Table/volcano-plot-of-differential-expression-genes.pdf`)

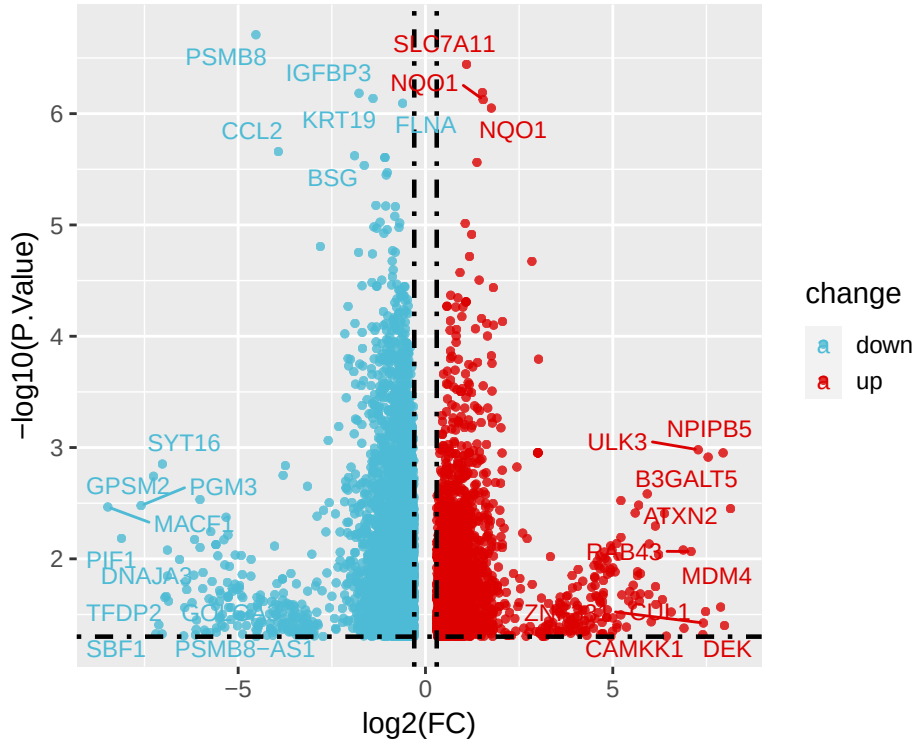


Figure 3: Volcano plot of differential expression genes

4.5 基因共表达分析

4.5.1 建立基因共表达模块

将上述（4.4.4, Fig. 2）标准化过的**差异表达**基因数据 (Tab. 5) 用于 WGCNA 分析[3]。

Figure 4为图 cluster sample 概览。

(对应文件为 `Figure+Table/cluster-sample.pdf`)

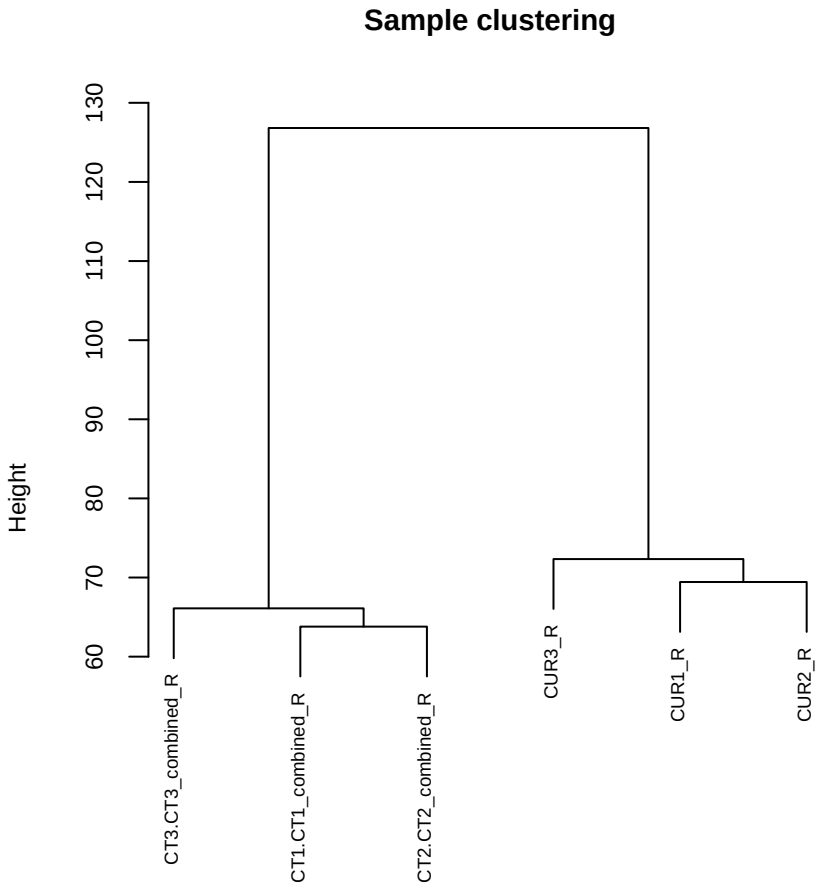


Figure 4: Cluster sample

由于样本数量较少，没有明显合适的 'soft threshold'。这里，选择 'soft threshold' 为 3。

Figure 5为图 pick soft thereshold 概览。

(对应文件为 `Figure+Table/pick-soft-thereshold.pdf`)

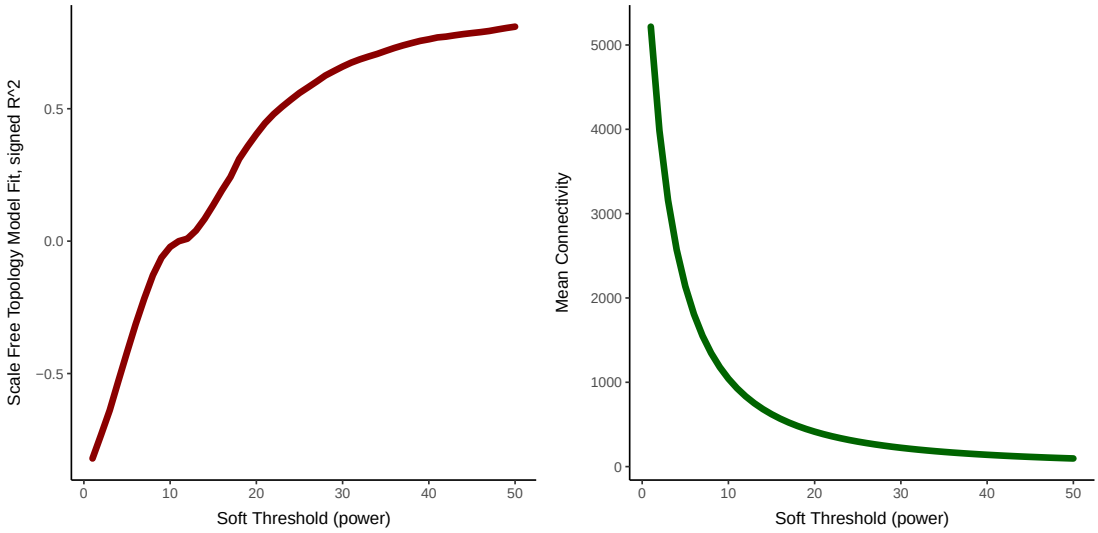


Figure 5: Pick soft thereshold

Figure 6为图 gene modules 概览。

(对应文件为 `Figure+Table/gene-modules.pdf`)

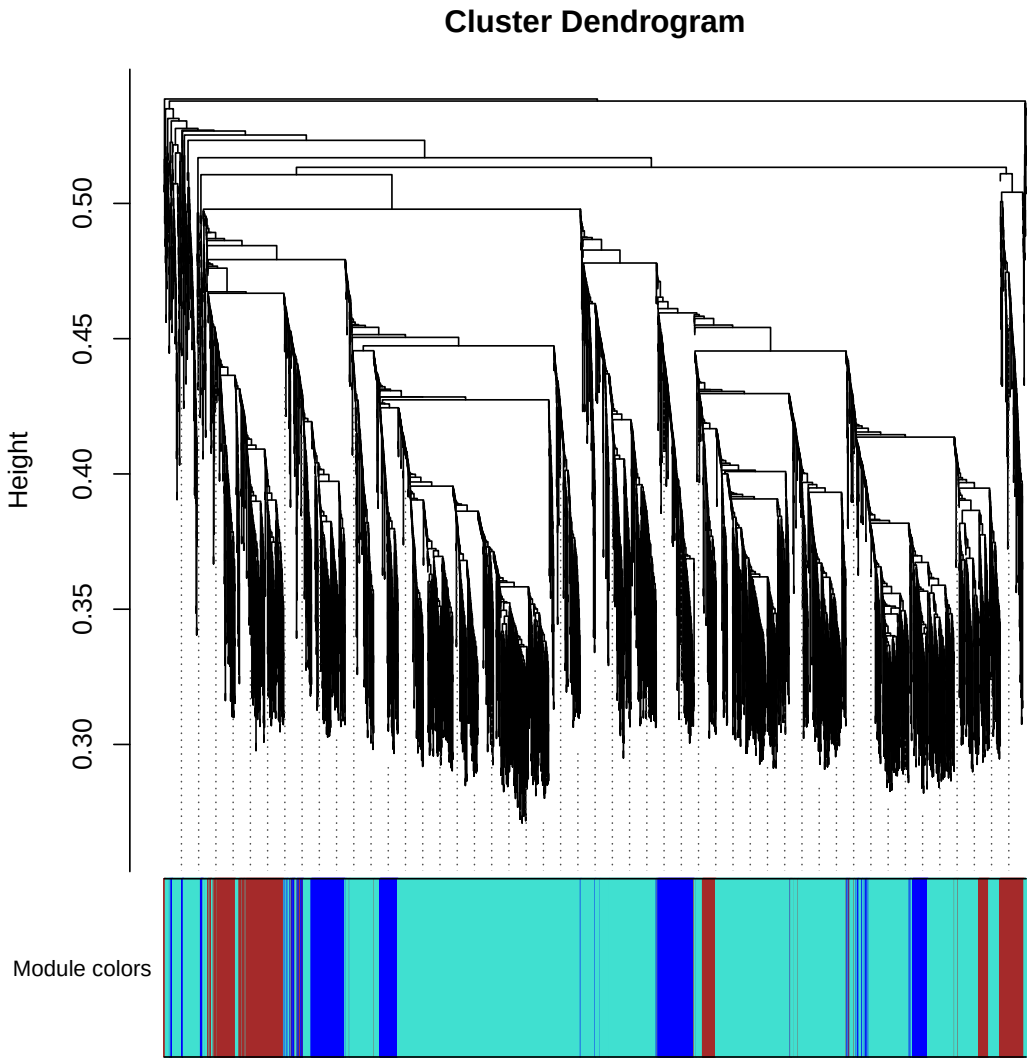


Figure 6: Gene modules

4.5.2 共表达模块和基因的关联性

计算 'gene module' 和 genes 之间的关联性（module membership）。

Table 6为表格 module membership 概览。

(对应文件为 `Figure+Table/module-membership.xlsx`)

> 注：表格共有 11853 行 15 列，以下预览的表格可能省略部分数据；表格含有 6009 个唯一 'gene'。

Table 6: Module membership

gene	module	cor	pvalue	log2...	signi...	sign	p.adjust	ensem...	entre...	hgnc_...	chrom...	start...	...
ENST0..	ME1	-0.84	0.0346	4.853...	< 0.05	*	0.078...	ENSG0..	2101	ESRRA	11	64305497.	
ENST0..	ME1	0.89	0.0166	5.912...	< 0.05	*	0.061...	ENSG0..	64847	SPATA20	17	50543058.	
ENST0..	ME2	-0.85	0.0319	4.970...	< 0.05	*	0.076...	ENSG0..	64847	SPATA20	17	50543058.	
ENST0..	ME1	-0.82	0.0471	4.408...	< 0.05	*	0.088...	ENSG0..	57414	RHBDD2	7	75842602.	
ENST0..	ME2	0.89	0.0171	5.869...	< 0.05	*	0.062...	ENSG0..	57414	RHBDD2	7	75842602.	
ENST0..	ME3	0.85	0.0305	5.035...	< 0.05	*	0.075...	ENSG0..	3675	ITGA3	17	50055968.	
ENST0..	ME1	0.96	0.002	8.965...	< 0.05	*	0.039...	ENSG0..	2067	ERCC1	19	45407334.	
ENST0..	ME2	-0.82	0.0435	4.522...	< 0.05	*	0.086...	ENSG0..	2067	ERCC1	19	45407334.	
ENST0..	ME3	-0.86	0.0287	5.122...	< 0.05	*	0.073...	ENSG0..	2067	ERCC1	19	45407334.	
ENST0..	ME2	0.86	0.0294	5.088...	< 0.05	*	0.074...	ENSG0..	8635	RNASET2	6	16692...	...
ENST0..	ME1	0.99	3e-04	11.70...	< 0.001	**	0.022...	ENSG0..	5010	CLDN15	3	17041...	...
ENST0..	ME2	-0.9	0.015	6.058...	< 0.05	*	0.059...	ENSG0..	5010	CLDN15	3	17041...	...
ENST0..	ME3	-0.95	0.0038	8.039...	< 0.05	*	0.046...	ENSG0..	5010	CLDN15	3	17041...	...
ENST0..	ME1	-0.9	0.014	6.158...	< 0.05	*	0.058...	ENSG0..	84957	RELT	11	73376399.	
ENST0..	ME2	0.85	0.0311	5.006...	< 0.05	*	0.075...	ENSG0..	84957	RELT	11	73376399.	
...	...	...	...	...	...	...	...	...	...	...	...	...	...

4.5.3 TCF4 所在的基因表达模块

确认 TCF4 或 TCF-AS1 所在基因模块。

Table 7为表格 TCF4 in modules memberships 概览。TCF4 所在基因模块为 'ME1' 和 'ME2'（TCF4 和 TCF-AS1 不存在共表达关系）。

（对应文件为 `Figure+Table/TCF4-in-modules-memberships.csv`）

注：表格共有 4 行 15 列，以下预览的表格可能省略部分数据；表格含有 3 个唯一 'gene'。

Table 7: TCF4 in modules memberships

gene	module	cor	pvalue	log2...	signi...	sign	p.adjust	ensem...	entre...	hgnc_...	chrom...	start...	...
ENST0..	ME1	0.85	0.033	4.921...	< 0.05	*	0.077...	ENSG0..	6925	TCF4	18	55222185.	
ENST0..	ME1	0.89	0.0171	5.869...	< 0.05	*	0.062...	ENSG0..	6925	TCF4	18	55222185.	
ENST0..	ME2	-0.86	0.0292	5.097...	< 0.05	*	0.074...	ENSG0..	6925	TCF4	18	55222185.	
ENST0..	ME2	-0.83	0.0412	4.601...	< 0.05	*	0.084...	ENSG0..	6925	TCF4	18	55222185.	

过滤 Tab. 6 数据，根据 p.adjust < 0.05, 以及 module 为 'ME1' 和 'ME2'。随后，使用 `biomaRt` 获取基因对应的蛋白质的序列，同时，获取 TCF4 和 TCF-AS1 的序列；将这些序列转化为 'fasta' 格式（数量大于 500 个的 'fasta' 文件被切分）。

4.5.4 使用 'catRAPID omics v2.1' 预测 RBPs

4.5.4.1 上传 catRAPID 服务器 catRAPID omics v2.1[4] 可同时计算多对 RNA 和蛋白质的结合（一次最多接受 500 个序列）。

结果可见于服务器：

- http://crg-webservice.s3.amazonaws.com/submissions/2023-08/729560/output/index.html?unlock=c9f3fccec3
- http://crg-webservice.s3.amazonaws.com/submissions/2023-08/729563/output/index.html?unlock=77c11a2b6a
- http://crg-webservice.s3.amazonaws.com/submissions/2023-08/729565/output/index.html?unlock=6449ff7496

4.5.4.2 结果整理 Table 8为表格 all results include positive or negtive 概览。

(对应文件为 `Figure+Table/all-results-include-positive-or-negtive.tsv`)

注：表格共有 162666 行 13 列，以下预览的表格可能省略部分数据；表格含有 1291 个唯一 'Protein_ID'。

Table 8: All results include positive or negtive

Prote…	RNA_ID	rnaFr……3	rnaFr……4	Annot…	Inter…	Z_score	RBP_P…	RNA_B…	numof……10	…
ERCC5	TCF4	6973	7306	-	119.58	1.47	0.43	PF007…	2	…
DLG3	TCF4	6973	7306	-	115.51	1.34	0.5	PF006…	7	…
NRDC	TCF4	6973	7306	-	114.44	1.3	0.51	PF161…	3	…
INCENP	TCF4	6973	7306	-	112.79	1.25	0.63	PF121…	2	…
ERC1	TCF4	6973	7306	-	107.5	1.08	0.29	PF101…	2	…
DLG3	TCF4.AS1	251	302	-	105.42	5.79	0.5	PF006…	7	…
KIF2C	TCF4	2201	2534	-	105.19	1.01	0.41	PF002…	2	…
ERCC5	TCF4.AS1	251	302	-	103.23	5.65	0.43	PF007…	2	…
GTSE1	TCF4	6973	7306	-	103.09	0.94	0.41	PF15259	1	…
LIG1	TCF4	6973	7306	-	102.39	0.92	0.43	PF010…	3	…
DLG3	TCF4.AS1	276	327	-	102.28	5.59	0.5	PF006…	7	…
FILIP1L	TCF4	6973	7306	-	101.81	0.9	0.29	PF09727	1	…
TPX2	TCF4	6973	7306	-	101.34	0.89	1	PF122…	3	…
KIF15	TCF4	6973	7306	-	101.24	0.88	0.37	PF002…	3	…
NUP107	TCF4	6973	7306	-	100.24	0.85	0.23	PF04121	1	…
…	…	…	…	…	…	…	…	…	…	…

关于结果表格和各类评分的解释可以参考：http://service.tartaglialab.com/static_files/shared/documentation_omics2.html。

接下来，按照不同条件筛选结果：

- `RBP_Propensity == 1`,
- `Interaction_Propensity > 0`,
- `numof.RNA.Binding_Domains_Instances > 0`,
- `numof.RNA_Binding_Motifs_Instances > 0`

Figure 7为图 intersects of sets of filtering conditions 概览。

(对应文件为 `Figure+Table/intersects-of-sets-of-filtering-conditions.pdf`)

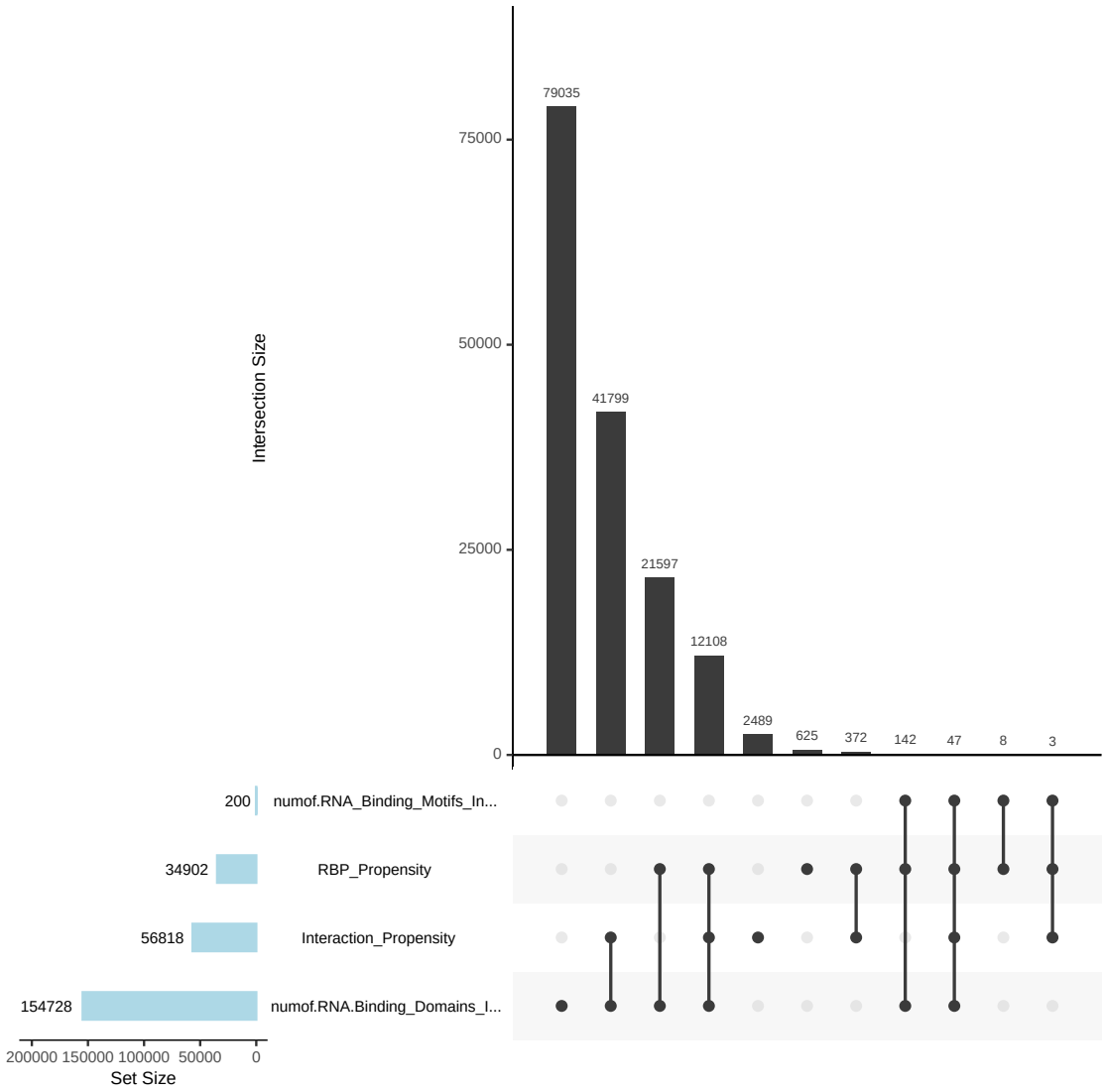


Figure 7: Intersects of sets of filtering conditions

可以发现，将四个数据集取交集，能得到包含少量数据的结果。

Table 9为表格 top candidates 概览。

(对应文件为 `Figure+Table/top-candidates.xlsx`)

> 注：表格共有 47 行 13 列，以下预览的表格可能省略部分数据；表格含有 10 个唯一 'Protein_ID'。

Table 9: Top candidates

Prote…	RNA_ID	rnaFr……3	rnaFr……4	Annot…	Inter…	Z_score	RBP_P…	RNA_B…	numof……10	…
PPIG	TCF4	6973	7306	-	70.11	-0.1	1	PF00160	1	…
LARP4	TCF4	2367	2700	-	46.53	-0.85	1	PF05383	1	…

Prote…	RNA_ID	rnaFr……3	rnaFr……4	Annot…	Inter…	Z_score	RBP_P…	RNA_B…	numof……10	…
PPIG	TCF4	5313	5646	-	33.58	-1.26	1	PF00160	1	…
LARP4	TCF4	2325	2658	-	26.88	-1.47	1	PF05383	1	…
LARP4	TCF4	6973	7306	-	26.8	-1.47	1	PF05383	1	…
CPEB2	TCF4	3363	3696	-	25.41	-1.52	1	PF163…	3	…
CPEB2	TCF4	167	500	-	23.38	-1.58	1	PF163…	3	…
CPEB2	TCF4	3031	3364	-	19.65	-1.7	1	PF163…	3	…
CPEB2	TCF4	209	542	-	16.25	-1.81	1	PF163…	3	…
PPIG	TCF4	5355	5688	-	8.39	-2.06	1	PF00160	1	…
CPEB2	TCF4	3321	3654	-	7.93	-2.07	1	PF163…	3	…
PPIG	TCF4	5479	5812	-	5.62	-2.15	1	PF00160	1	…
IGF2BP1	TCF4	2325	2658	-	42.85	-0.96	1	PF000…	2	…
IGF2BP1	TCF4	2159	2492	-	37.92	-1.12	1	PF000…	2	…
PCBP2	TCF4	2159	2492	-	34.91	-1.22	1	PF00013	1	…
…	…	…	…	…	…	…	…	…	…	…

Tab. 9 包含 RBPs 与 TCF4 结合或 TCF-AS1 结合的可能性，以下取它们的交集。

Figure 8为图 unique candidate of RBP binding with TCF4 and TCF AS1 概览。

(对应文件为 `Figure+Table/unique-candidate-of-RBP-binding-with-TCF4-and-TCF-AS1.pdf`)

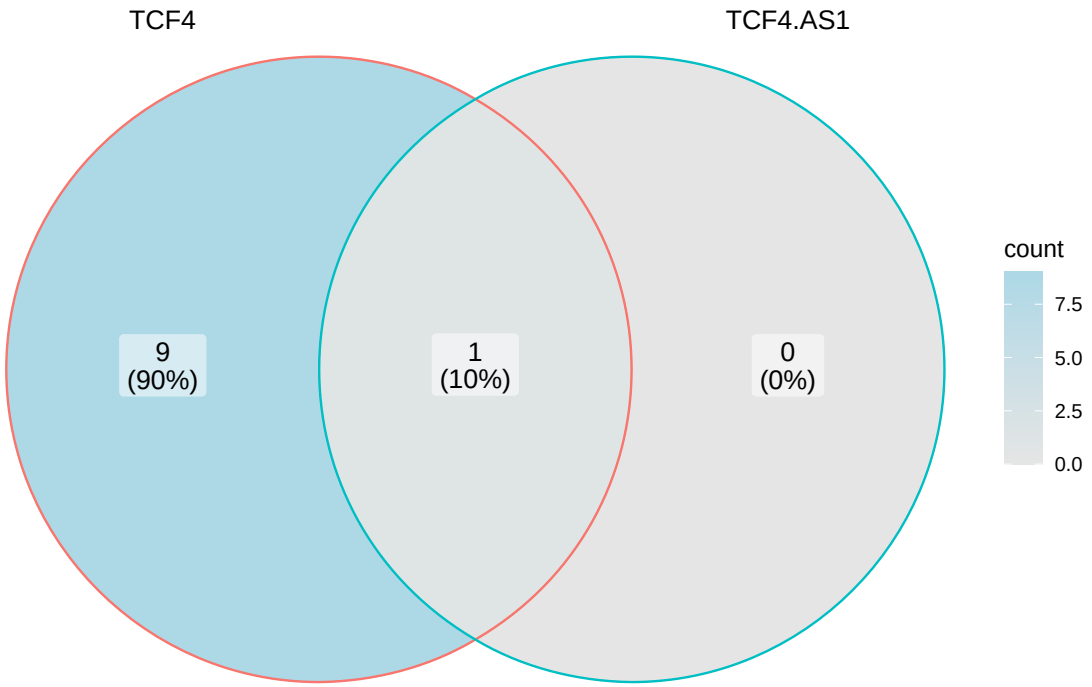


Figure 8: Unique candidate of RBP binding with TCF4 and TCF AS1

Fig. 8 中包含的蛋白的 'symbol' 为:

```
## $TCF4
##  [1] "PPIG"    "LARP4"   "CPEB2"   "IGF2BP1" "PCBP2"   "YTHDF3"  "HNRNPH1" "KHDRBS3" "MBNL1"   "DAZAP
##
## $TCF4.AS1
## [1] "HNRNPH1"
```

5 结论

将 RNA-seq 数据结合差异分析、基因共表达分析，并利用 catRAPID 工具预测 RBPs，成功筛选出一批 RBPs。随后，根据 RBP 倾向（`RBP_Propensity`）、结合倾向（`Interaction_Propensity`）等条件筛选，获得唯一 RBP：HNRNPH1

Reference

1. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic rna-seq quantification. *Nature Biotechnology* **34**, (2016).

2. Ritchie, M. E. *et al.* Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).

3. Langfelder, P. & Horvath, S. WGCNA: An r package for weighted correlation network analysis. *BMC Bioinformatics* **9**, (2008).

4. Armaos, A., Colantoni, A., Proietti, G., Rupert, J. & Tartaglia, G. G. catRAPIDomics v2.0: Going deeper and wider in the prediction of proteinRNA interactions. *Nucleic Acids Research* **49**, (2021).

5. Peng, X. *et al.* RBP-tstl is a two-stage transfer learning framework for genome-scale prediction of rna-binding proteins. *Briefings in Bioinformatics* **23**, (2022).

6. Su, Y., Luo, Y., Zhao, X., Liu, Y. & Peng, J. Integrating thermodynamic and sequence contexts improves protein-rna binding prediction. *PLOS Computational Biology* **15**, (2019).

7. Orenstein, Y., Wang, Y. & Berger, B. RCK: Accurate and efficient inference of sequence- and structure-based proteinRNA binding models from rnacompete data. *Bioinformatics* **32**, (2016).

8. Law, C. W. *et al.* A guide to creating design matrices for gene expression experiments. *F1000Research* **9**, 1444 (2020).