Report of Analysis

Huang LiChuang of Wie-Biotech

Contents

1 摘要 **2**
1.1 fastp 质控 . . . 2
1.2 全外显子分析流程 . . . 2
1.3 结果可视化 . . . 2

2 研究设计流程图 **3**

3 材料和方法 **3**

4 分析结果 **4**
4.1 fastp 质控 . . . 4
4.2 WES 变异筛选 . . . 4
4.2.1 ANNOVAR 注释 . . . 5
4.2.2 `maftools` 可视化 . . . 5
4.3 下游分析 . . . 6
4.3.1 获取 Genecards 与胆汁相关疾病的基因 . . . 6
4.3.2 通路富集分析 . . . 7

5 结论 **11**

6 其它 **11**
6.1 新生儿心脏骤停 . . . 11
6.1.1 数据来源 . . . 11
6.2 新生儿胎粪性腹膜炎差异基因 . . . 12
6.3 胎儿宫内窘迫 . . . 12
6.4 死胎 . . . 12
6.5 新生儿呼吸窘迫综合征 . . . 13

Reference **13**

List of Figures

1 Summary of mutations in samples . . . 5

2 Proportion of SNPs mutation 6
3 Intersect of variants with Genecards prediction 7
4 KEGG enrichment 8
5 Intersection of filtered variants with KEGG pathway 9
6 Intersects of the pathways related variants in all samples 10

List of Tables

1 Genecards genes relative with bild acids 6
2 Bile acids related variants occurs in all ICP samples 10

1 摘要

根据客户提供的材料分析基因突变及信号通路，筛选出研究的对象基因。相关疾病是“高胆汁酸血症”或是“妊娠期肝内胆汁淤积症（Intrahepatic cholestasis of pregnancy，ICP）

1.1 fastp 质控

- 去低质量碱基
- 去接头
- 生成报告

1.2 全外显子分析流程

WES 一般分析流程为:

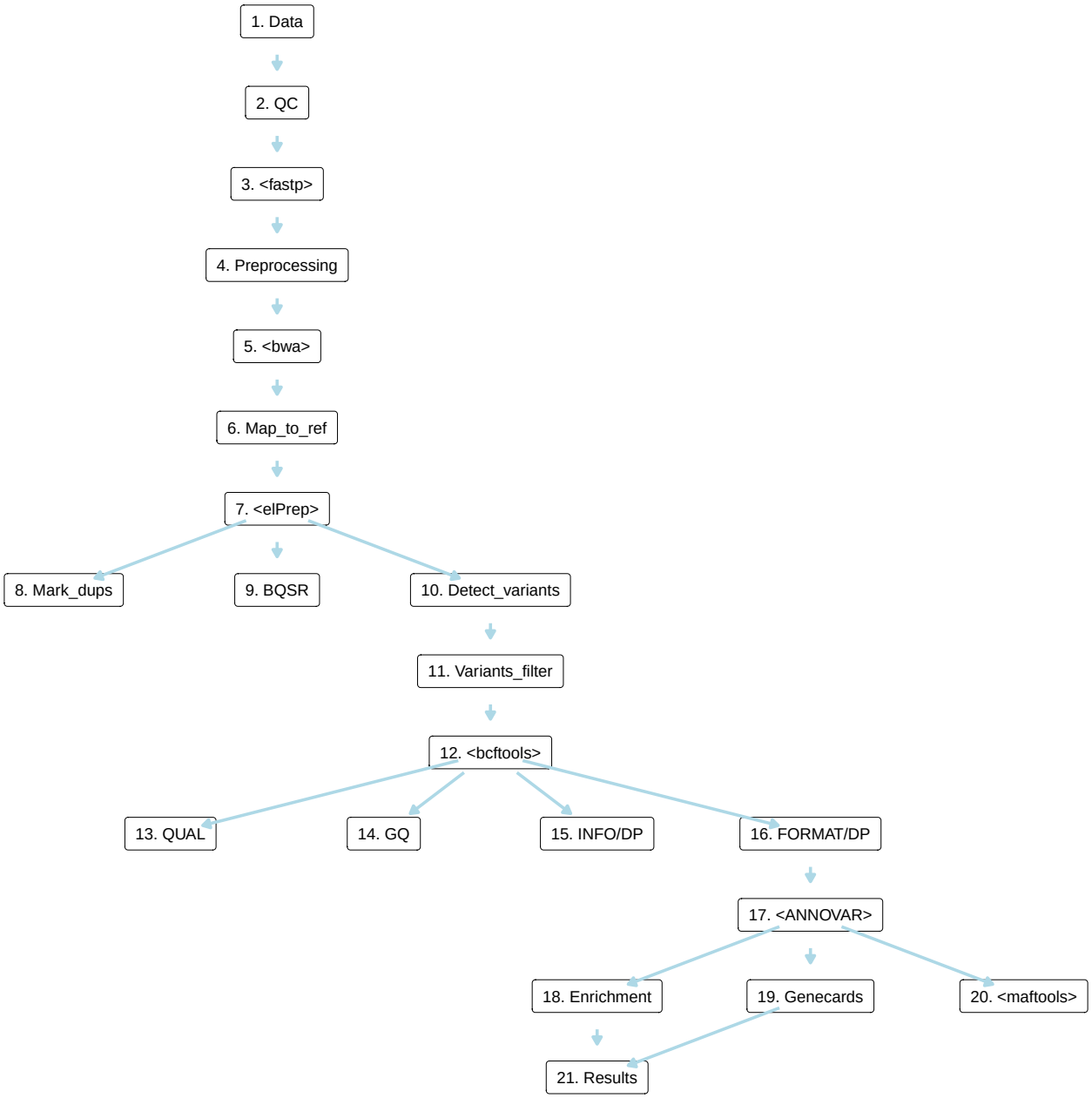
https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows:

- Preprocessing https://gatk.broadinstitute.org/hc/en-us/articles/360035535912-Data-pre-processingfor-variant-discovery
  - 比对到参考基因组
  - 标记重复
  - 基础校准（Base (Quality Score) Recalibration）
- Variant discovery https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-shortvariant-discovery-SNPs-Indels-
  - 获取变异注释文件
  - 变异检测
  - 变异质控和过滤
  - 变异注释

1.3 结果可视化

使用 maftools 对变异注释结果可视化。

2 研究设计流程图



3 材料和方法

- fastp (https://github.com/OpenGene/fastp)

以下可以通过 https://gatk.broadinstitute.org/hc/en-us/articles/360041320571--How-to-Install-all-software-packages-required-to-follow-the-GATK-Best-Practices 获取安装。

- bwa
- ...

使用 elPrep[1] 替代 GATK4 做 WES 分析（见 1.2）。

使用 bcftools[2] 过滤 vcf。

使用 ANNOVAR 变异注释。

使用 R maftools 可视化 ANNOVAR 注释结果。

使用 clusterProfiler 富集分析（KEGG）。

参考基因组：

- https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/latest/hg38.fa.gz

SNPs 和 Indels:

(https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0)

- `1000G_phase1.snps.high_confidence.hg38.vcf`
- `Mills_and_1000G_gold_standard.indels.hg38.vcf`

4 分析结果

4.1 fastp 质控

'Fastp report files' 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. V350065014_L01_93_.html
> 2. V350065014_L01_94_.html
> 3. V350065014_L02_94_.html
> 4. V350065026_L03_86_.html
> 5. V350065026_L04_85_.html
> 6. ...

注：客户提供 8 个病人的数据，每个病人的目录下有 2 个子文件，因此共 16 个样本数据。硬盘中有个别 fastq 文件有损坏。损坏的文件未纳入分析流程中（没有报告生成的为损坏的文件）。

4.2 WES 变异筛选

以下流程相较于 GATK4 Best Practice (https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows) 有所变化。

- 使用 elPrep 5 完成检测流程（流程类似于 GATK4，但速度更快）[1]。

得到变异信息文件（vcf）后，使用 bcftools 过滤（QUAL>10 && GQ>10 && FORMAT/DP>10 && INFO/DP>100）。

4.2.1 ANNOVAR 注释

使用 ANNOVAR (https://annovar.openbioinformatics.org/en/latest/) 注释后，滤除同义突变。

'Exonic annotation by ANNOVAR' 数据已全部提供。

(对应文件为 `exonic-annotation-by-ANNOVAR`)

> 注：文件夹 exonic-annotation-by-ANNOVAR 共包含 6 个文件。
>
> 1. 1_X220325_I26_V350065014_L1_22L01298712.93.csv
> 2. 10_X220325_M038_V350065181_L04_22L01298713.24.csv
> 3. 11_X220325_M038_V350065181_L04_22L01298716.25.csv
> 4. 12_X220325_M120_V350065026_L03_22L01298714.86.csv
> 5. 13_X220325_M120_V350065026_L04_22L01298711.85.csv
> 6. ...

4.2.2 `maftools` 可视化

参考 https://www.bioconductor.org/packages/release/bioc/vignettes/maftools/inst/doc/maftools.html#910_Mutational_Signatures

Figure 1为图 summary of mutations in samples 概览。

(对应文件为 `Figure+Table/summary-of-mutations-in-samples.pdf`)

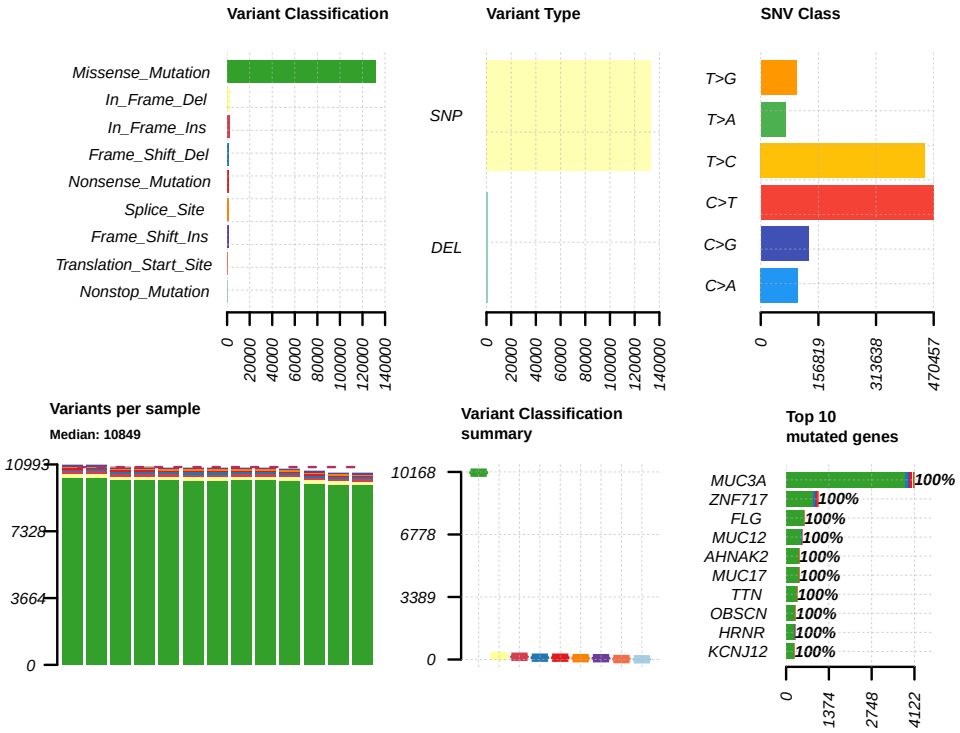


Figure 1: Summary of mutations in samples

Figure 2为图 proportion of SNPs mutation 概览。

(对应文件为 `Figure+Table/proportion-of-SNPs-mutation.pdf`)

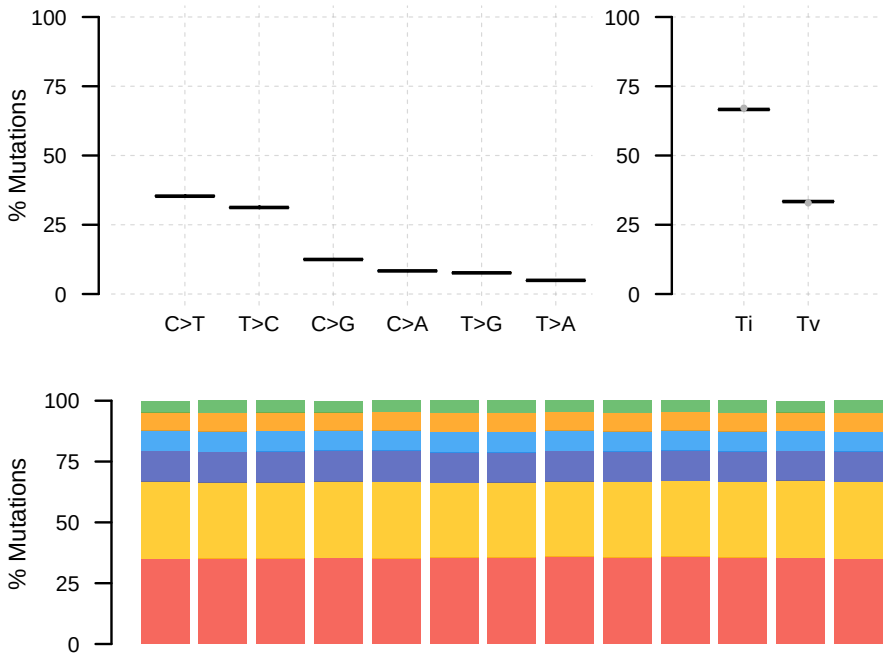


Figure 2: Proportion of SNPs mutation

4.3 下游分析

4.3.1 获取 Genecards 与胆汁相关疾病的基因

Table 1为表格 Genecards genes relative with bild acids 概览。

(对应文件为 `Figure+Table/Genecards-genes-relative-with-bild-acids.xlsx`)

> 注：表格共有 1494 行 7 列，以下预览的表格可能省略部分数据；表格含有 1494 个唯一 'Symbol'。

Table 1: Genecards genes relative with bild acids

Symbol	Descr...	Category	UniPr...	GIFtS	GC_id	Score
BAAT	Bile ...	Prote...	Q14032	47	GC09M...	93.50
AKR1D1	Aldo-...	Prote...	P51857	48	GC07P...	88.14
ABCB11	ATP B...	Prote...	O95342	50	GC02M...	85.41
SLC10A1	Solut...	Prote...	Q14973	47	GC14M...	83.40
NR1H4	Nucle...	Prote...	Q96RI1	52	GC12P...	83.18
AMACR	Alpha...	Prote...	Q9UHK6	49	GC05M...	76.57

Symbol	Descr...	Category	UniPr...	GIFtS	GC_id	Score
HSD3B7	Hydro...	Prote...	Q9H2F3	45	GC16P...	70.66
CYP7B1	Cytoc...	Prote...	O75881	50	GC08M...	65.39
GPBAR1	G Pro...	Prote...	Q8TDU6	44	GC02P...	63.83
CYP7A1	Cytoc...	Prote...	P22680	47	GC08M...	60.99
ACOX2	Acyl-...	Prote...	Q99424	47	GC03M...	59.23
SLC51B	Solut...	Prote...	Q86UW2	38	GC15P...	58.25
ABCB4	ATP B...	Prote...	P21439	51	GC07M...	56.92
SLC27A5	Solut...	Prote...	Q9Y2P5	45	GC19M...	55.16
ALB	Albumin	Prote...	P02768	53	GC04P...	51.11
...	...	...	...	...	...	...

4.3.2 通路富集分析

取 Tab. 1 的基因与 ?? 的所有基因的交集。

Figure 3为图 intersect of variants with Genecards prediction 概览。

(对应文件为 `Figure+Table/intersect-of-variants-with-Genecards-prediction.pdf`)

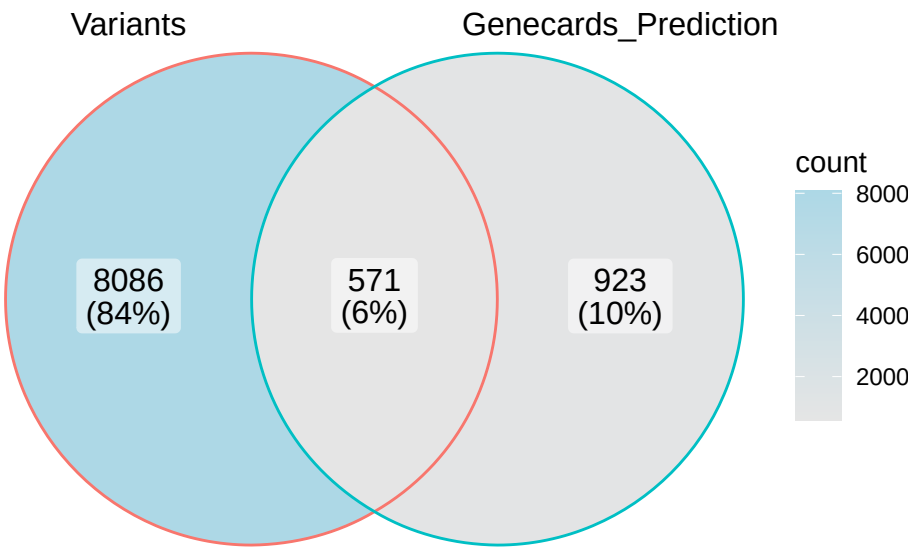


Figure 3: Intersect of variants with Genecards prediction

以交集基因通路富集。

Figure 4为图 KEGG enrichment 概览。

(对应文件为 `Figure+Table/KEGG-enrichment.pdf`)

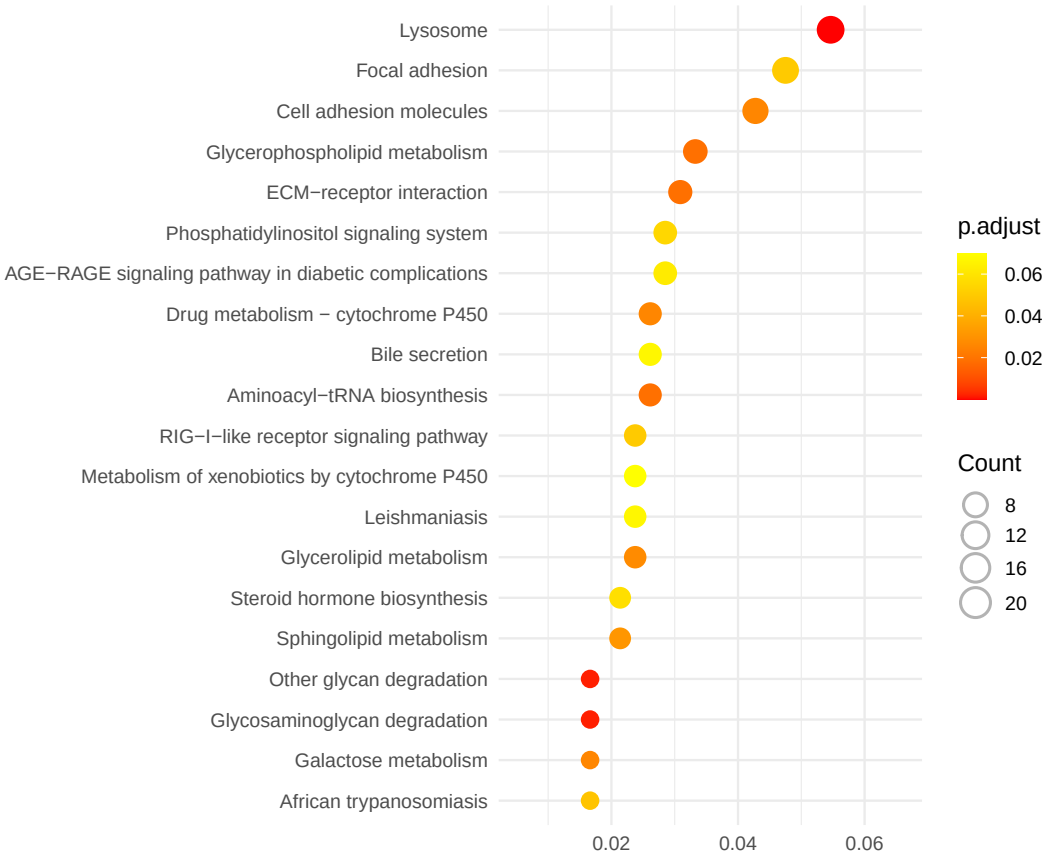


Figure 4: KEGG enrichment

取 'Bile secretion' 和 'Cholesterol metabolism' 相关的基因。

Figure 5为图 Intersection of filtered variants with KEGG pathway 概览。

(对应文件为 `Figure+Table/Intersection-of-filtered-variants-with-KEGG-pathway.pdf`)

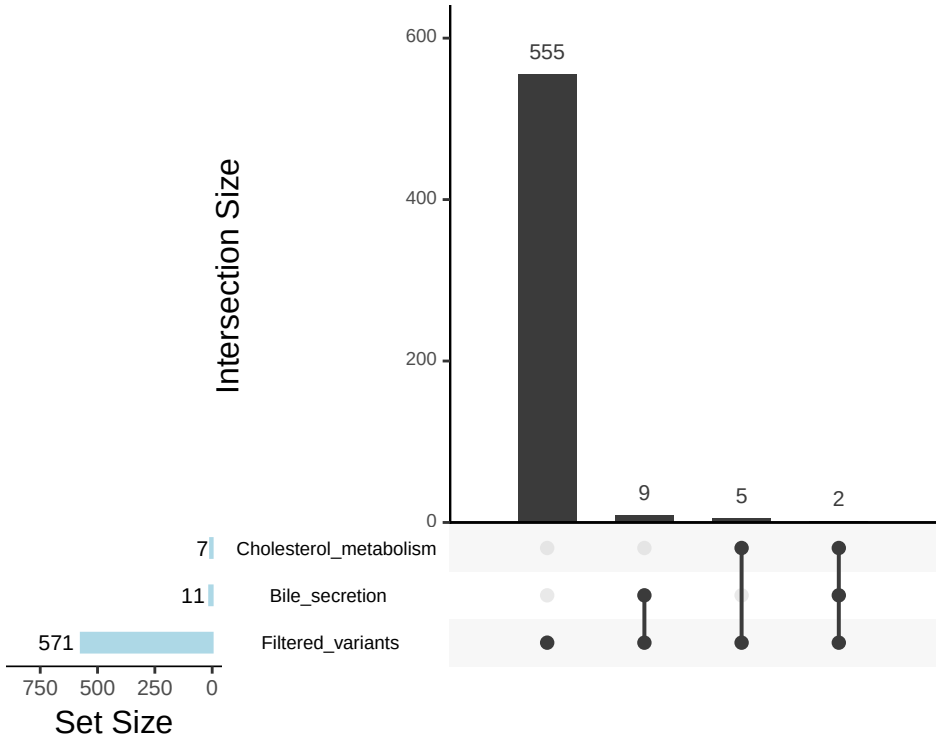


Figure 5: Intersection of filtered variants with KEGG pathway

取 Fig. 5 所示的两条通路的基因交集。

将这些交集基因回归到所有样本的变异数据中，取共同发生的突变结果。

Figure 6为图 intersects of the pathways related variants in all samples 概览。

(对应文件为 `Figure+Table/intersects-of-the-pathways-related-variants-in-all-samples.pdf`)

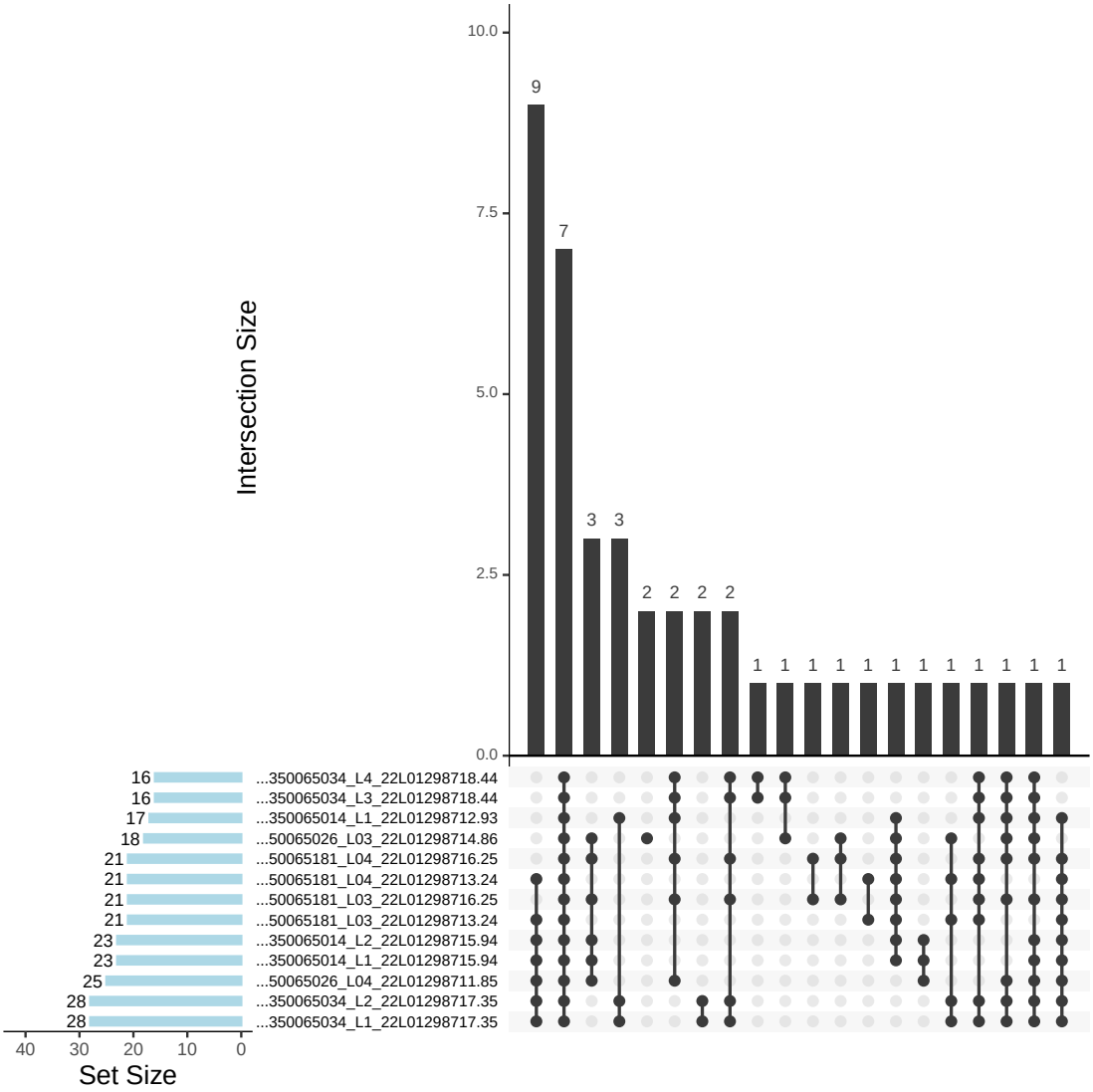


Figure 6: Intersects of the pathways related variants in all samples

有 7 个变异同时发生在所有样本中。

Table 2为表格 Bile acids related variants occurs in all ICP samples 概览。

(对应文件为 `Figure+Table/Bile-acids-related-variants-occurs-in-all-ICP-samples.xlsx`)

> 注：表格共有 7 行 14 列，以下预览的表格可能省略部分数据；表格含有 7 个唯一‘hgnc_symbol’。

Table 2: Bile acids related variants occurs in all ICP samples

hgnc_...	prote...	Chr	Start	End	Ref	Alt	Func....	Gene....	GeneD...	Exoni...	AACha...
LRP1	p.Q2900P	chr12	57196001	57196001	A	C	exonic	LRP1		nonsy...	LRP1:...
SLC10A1	p.S267F	chr14	69778476	69778476	G	A	exonic	SLC10A1		nonsy...	SLC10...

hgnc_...	prote...	Chr	Start	End	Ref	Alt	Func....	Gene....	GeneD...	Exoni...	AACha...
AQP9	p.T214A	chr15	58184082	58184082	A	G	exonic	AQP9		nonsy...	AQP9:...
APOH	p.V266L	chr17	66214639	66214639	C	A	exonic	APOH		nonsy...	APOH:...
ABCB11	p.V444A	chr2	16897...	16897...	A	G	exonic	ABCB11		nonsy...	ABCB1...
LRP2	p.A2872T	chr2	16919...	16919...	C	T	exonic	LRP2		nonsy...	LRP2:...
TSPO	p.T147A	chr22	43162920	43162920	A	G	exonic	TSPO		nonsy...	TSPO:...

5 结论

见 Tab. 2。

ICP 相关对象基因：

LRP1, SLC10A1, AQP9, APOH, ABCB11, LRP2, TSPO。

突变形式 (hgvs) 为：

p.Q2900P, p.S267F, p.T214A, p.V266L, p.V444A, p.A2872T, p.T147A。

6 其它

6.1 新生儿心脏骤停

6.1.1 数据来源

检索：neonatal cardiac arrest

数据来源于[3] (piglets mRNA-seq)：

- PMID: 31005300
- GSE120863

data_processing :

RNA sequencing reads were aligned to the Pig genome (Sscrofa11.1) using Star.

data_processing.1 :

featureCounts was used to map the reads to the exons of genes.

data_processing.2 :

DESeq2 was used to normalize the data using regularized-logarithm.

data_processing.3 :

Genome_build: Sscrofa11.1

data_processing.4 :

Supplementary_files_format_and_content: 0618_striatum.feature_counts.all_samples.txt includes the counts.

data_processing.5 :

Supplementary_files_format_and_content: sham_DHCA_striatum_mRNA_seq.txt includes the fold changes and statistics.

6.2 新生儿胎粪性腹膜炎差异基因

检索：neonatal meconium peritonitis

无相关数据。

6.3 胎儿宫内窘迫

检索：fetal distress

无相关数据。

6.4 死胎

检索：stillbirth

无相关数据。

6.5 新生儿呼吸窘迫综合征

neonatal respiratory distress syndrome

无相关数据。

Reference

1. Multithreaded variant calling in elPrep 5. *PLOS ONE* **16**, 1–13 (2021).

2. Danecek, P. *et al.* Twelve years of samtools and bcftools. *GigaScience* **10**, (2021).

3. Tu, L. N. *et al.* Transcriptome profiling reveals activation of inflammation and apoptosis in the neonatal striatum after deep hypothermic circulatory arrest. *The Journal of Thoracic and Cardiovascular Surgery* **158**, (2019).