生信文章修改甲基化测序

2024-05-31

LiChuang Huang

@ 立效研究院

Contents

List of Figures

List of Tables

1 摘要

原文中，对 DMR 的总体统计未修改，增补了一部分图片和 CpG Island 的统计。后续富集分析和 StringDB 的 PPI 网络等内容都重做了。详细见 4

重要说明：

目前，对于原数据表格中，'Model-vs-Model-Cure'，是按照 Model 比 Treatment 来认定的，而重新分析时，判定的是 Treatment vs Model，也就是，这里将原先的 Delta 值乘以 (-1) 实现转换。详情见 6.1.1

2 前言

3 材料和方法

3.1 材料

3.2 方法

Mainly used method:

- R package `biomaRt` used for gene annotation[1].
- The `biomart` was used for mapping genes between organism (e.g., mgi_symbol to hgnc_symbol)[1].
- R package `Gviz` were used for methylation data visualization[2].
- R package `ClusterProfiler` used for gene enrichment analysis[3].
- Databses of `DisGeNet`, `GeneCards`, `PharmGKB` used for collating disease related targets[4–6].
- R package `STEINGdb` used for PPI network construction[7,8].
- R package `rtracklayer` used for UCSC data query[9].
- The CpG islands data was downloaded from http://www.rafalab.org (generated by R package `makeCGI`)[10].
- R package `pathview` used for KEGG pathways visualization[11].
- The MCC score was calculated referring to algorithm of `CytoHubba`[8].
- R version 4.4.0 (2024-04-24); Other R packages (eg., `dplyr` and `ggplot2`) used for statistic analysis or data visualization.

4 分析结果

4.1 Methyl-seq DMR 分析

这部分的 DMR 数据和原先的内容是一样的，只是补充或替换了以下图：

- DMR 分布见 Fig. 1a，DMR 存在于基因的分布见 Fig. 1b。
- DMR 的筛选 with |delta| > 0.3, FDR < 0.05 (与原先相同) ，见 Fig. 1c。
- 补充了 DMR 存在于 CpG Island 的注释，在各个染色体的分布见 Fig. 1d，Fig. 1e。

Figure 1 (下方图) 为图 MAIN Fig 1 概览。

(对应文件为 `./Figure+Table/fig1.pdf`)

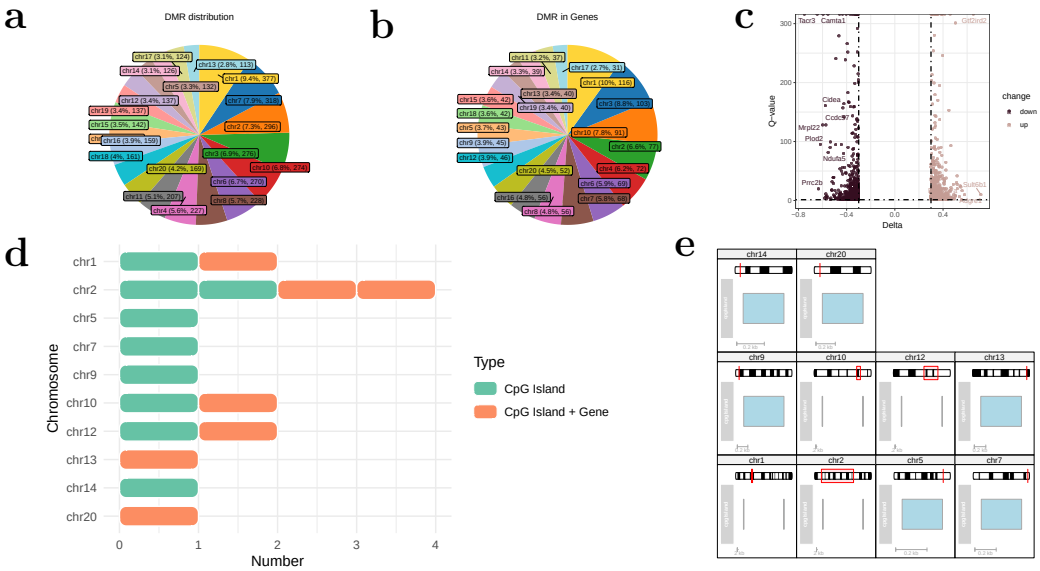

Figure 1: MAIN Fig 1

4.2 富集分析

对所有的 DMR 基因做了 KEGG 富集分析和 GO 富集分析，见 Fig. 2a，c。KEGG 富集分析发现 DMR 富集于 'Type II diabetes mellitus' (T2DM) 通路。见 Fig. 2b，其中，Ins2 基因甲基化程度升高，而 Pik3cb 甲基化程度下降。Pik3cb 在染色体 8 (chr8) 中，甲基化位置出于基因的中段 (Fig. 3a，b) 。Ins2 基因处于染色体 1 (chr1) (见 Fig. 3c，d) 。胰岛素信号通路 PI3K/Akt/mTOR 通路被认为与胰岛素抵抗 insulin resistance 相关密切[12]。DNA 甲基化改变影响 T2DM 发展中的胰岛素分泌和胰岛素抵抗[13]。Zuogui pill 给药后，改变了 Pik3cb (PI3K 的亚基) 的甲基化，可能进一步影响到了 PI3K 的活性，以及下游的信号通路，从而对胰岛素抵抗发挥调控作用。

Figure 2 (下方图) 为图 MAIN Fig 2 概览。

(对应文件为 `./Figure+Table/fig2.pdf`)

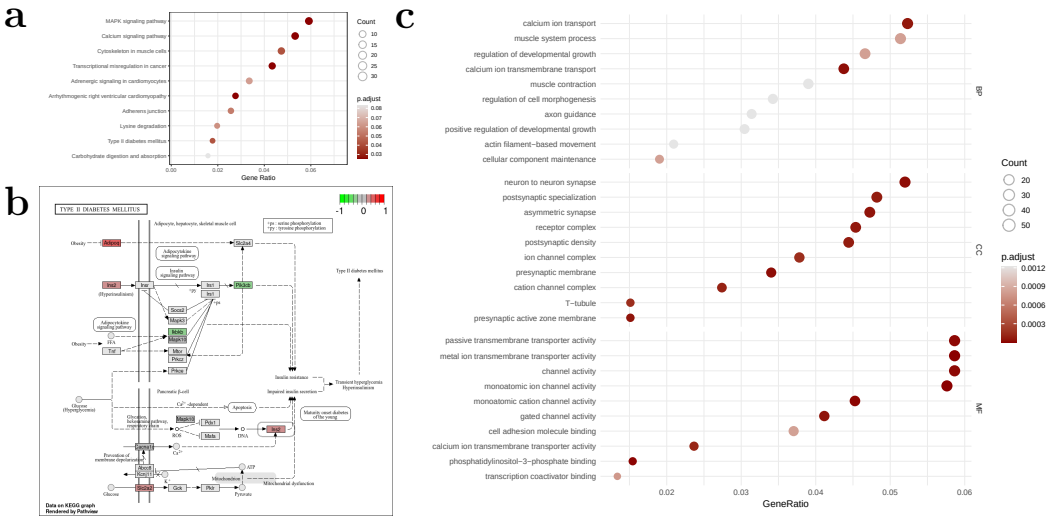

Figure 2: MAIN Fig 2

Figure 3 (下方图) 为图 MAIN Fig 3 概览。

(对应文件为 `./Figure+Table/fig3.pdf`)

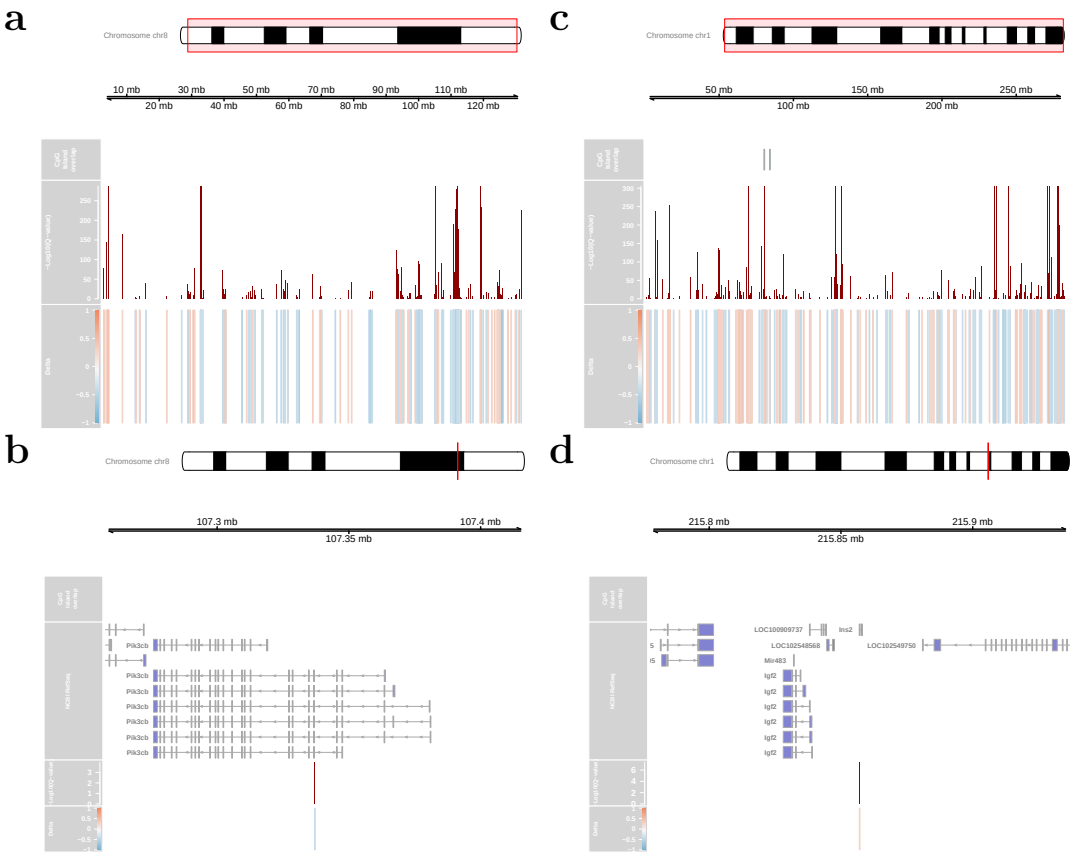

Figure 3: MAIN Fig 3

4.3 StringDB PPI

获取了 DM 相关的基因集，来源于 Fig. 4a 所示数据库 (这些数据库主要为人类的基因集，这里，使用 Biomart 将这些基因从 hgnc symbol 映射到 rgd symbol, 大鼠的基因)，取合集，与 DMR 取交集，发现有 201 个重叠基因，Fig. 4b。以重叠基因构建 PPI 网络，Fig. 4c。随后，筛选 TOP 30 的 Hub 基因，发现 Pik3cb、Ins2 在列。此外还有 Ikbkb。这些基因与 Fig. 4e 所示的其它基因存在互作关系，这可能涉及这些基因的上游或下游机制，与 T2DM 的发展机制以及甲基化在其中发挥的作用相关。

Figure 4 (下方图) 为图 MAIN Fig 4 概览。

(对应文件为 ./Figure+Table/fig4.pdf)

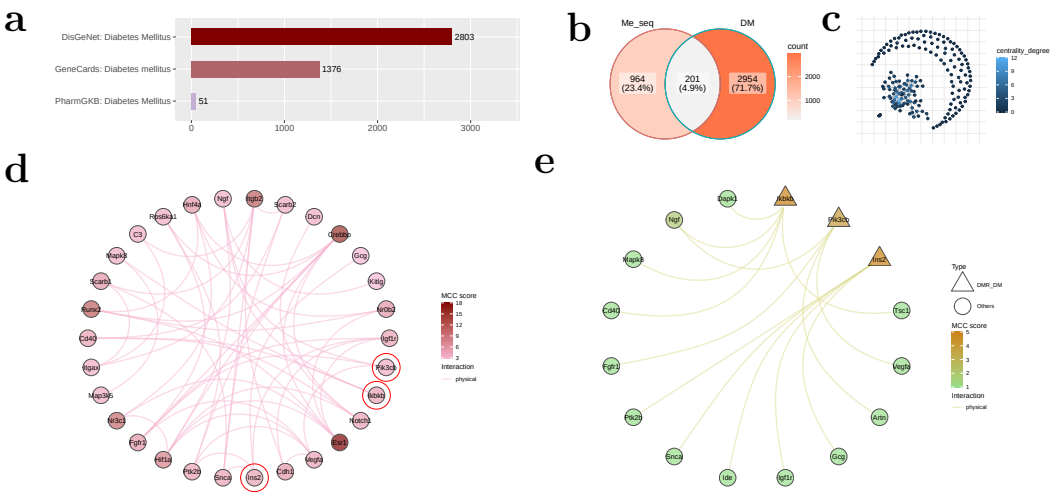

Figure 4: MAIN Fig 4

5 结论

Zuogui pill 给药涉及了分布于各染色体的 DMR，部分 DMR 处于 CpG Island。富集分析表明，总体 DMRs 与 T2DM 相关。对 Pik3cb 和 Ins2 基因的甲基化改变可能是 Zuogui pill 发挥药效的重要机制。

'Tiff figures' 数据已全部提供。

(对应文件为 ./Figure+Table/TIFF)

注：文件夹./Figure+Table/TIFF 共包含 4 个文件。

1. fig1.tiff
2. fig2.tiff
3. fig3.tiff
4. fig4.tiff

6 附：分析流程

- 许凯霞需求
- 浙江百越 4 例 WGBS 信息采集与分析

6.1 Methyl-seq

6.1.1 DMR data

- 数据来源：‘浙江百越 4 例 WGBS 信息采集与分析/结果/01_ 甲基化差异表达 DMR.csv’
- 注释来源 (基因)：‘浙江百越 4 例 WGBS 信息采集与分析/结果/01_ 甲基化差异表达基因.tsv’

重要说明:

目前，对于数据表格中，‘Model-vs-Model-Cure’，是按照 Model 比 Treatment 来认定的。而 Treatment vs Model, 则需要对原来的数据乘以 -1 转换。(如果测序公司或客户那边，实际上是相反的话，就要重新调整所有的分析了；不过，一般情况下应该是如此，只是在分析的过程中，发现结果与预期好像不是特别相符，因此这里有疑惑) 为了说明这一点, 这部分的数据处理提供了源代码:

R input

```
# ftibble <- function(x) tibble::as_tibble(data.table::fread(x))
t.genes <- ftibble("/media/echo/My Passport/浙江百越 4 例 WGBS 信息采集与分析/结果/01_ 甲基化差异表达基因
t.diff <- ftibble("/media/echo/My Passport/浙江百越 4 例 WGBS 信息采集与分析/结果/01_ 甲基化差异表达 DMR.
t.diff <- dplyr::select(t.diff, dmr_id, dplyr::ends_with("Model-vs-Model-Cure"))
t.diff <- dplyr::mutate(t.diff,
  chr = strx(dmr_id, "chr[0-9]+"),
  start = strx(dmr_id, "(?<=_)[0-9]+(?=_)"),
  end = strx(dmr_id, "[0-9]+$"),
  ## 以下为转换得到 Treatment vs Model:
  DMR_Treatment_vs_Model = -`dmr_diff_cg_Model-vs-Model-Cure`,
  DMR_Qvalue = `dmr_qvalue_cg_Model-vs-Model-Cure`
)
dmrDat <- dplyr::select(t.diff, chr, start, end, tidyselect::starts_with("DMR", F), dmr_id)
dmrDat <- dplyr::arrange(dmrDat, DMR_Qvalue)

dmrDat.genes <- map(dmrDat, "dmr_id", t.genes, "dmr_id", "gene", col = "symbol")
dmrDat.genes
# dplyr::filter(dmrDat.genes, symbol == "Ins2")
```

Table 1 (下方表格) 为表格 RAW DMR data 概览。

(对应文件为 `Figure+Table/RAW-DMR-data.csv`)

注：表格共有 4143 行 7 列，以下预览的表格可能省略部分数据；含有 21 个唯一 'chr；含有 1189 个唯一 'symbol'。

1. symbol: 基因或蛋白符号。
2. chr: chromosome (for the variant, same as gene_chr for cis-eQTLs)

Table 1: RAW DMR data

chr	start	end	DMR_Treatm...	DMR_Qvalue	dmr_id	symbol
chr10	27750001	27750200	-0.469714	0	chr10_2775...	NA
chr10	29567001	29567200	0.415147	0	chr10_2956...	NA
chr10	49174801	49175000	-0.335429	0	chr10_4917...	NA
chr10	62273401	62273600	0.33725	0	chr10_6227...	Wdr81
chr10	87251601	87251800	0.34196	0	chr10_8725...	NA
chr11	19652401	19652600	-0.632739	0	chr11_1965...	NA
chr11	25930601	25930800	-0.426453	0	chr11_2593...	NA
chr11	33534001	33534200	0.341885	0	chr11_3353...	NA
chr11	33678401	33678600	0.320595	0	chr11_3367...	NA
chr11	44229601	44229800	0.350355	0	chr11_4422...	St3gal6
chr11	59809001	59809200	-0.596561	0	chr11_5980...	NA
chr12	18027401	18027600	-0.347761	0	chr12_1802...	NA
chr12	25517401	25517600	0.651515	0	chr12_2551...	Gtf2ird2
chr12	25934201	25934400	0.340203	0	chr12_2593...	NA
chr12	39451801	39452000	-0.434265	0	chr12_3945...	Ift81
...	...	...	...	...	...	...

6.1.2 DMR distribution

Figure 5 (下方图) 为图 All DMR volcano plot 概览。

(对应文件为 `Figure+Table/All-DMR-volcano-plot.pdf`)

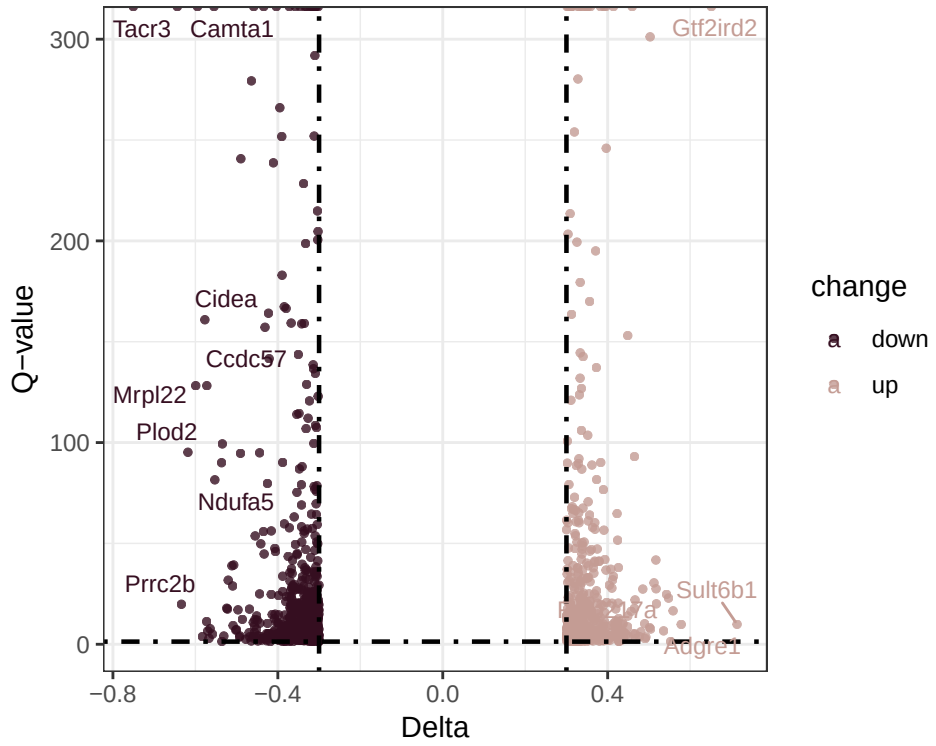

Figure 5: All DMR volcano plot

Figure 6 (下方图) 为图 DMR distribution 概览。

(对应文件为 `Figure+Table/DMR-distribution.pdf`)

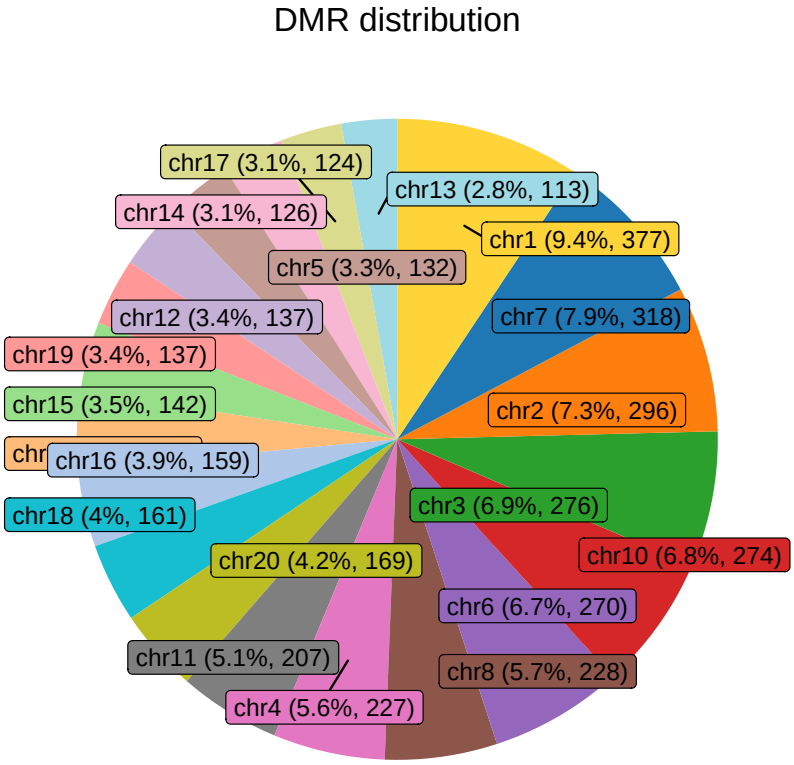

Figure 6: DMR distribution

Figure 7 (下方图) 为图 DMR in Genes 概览。

(对应文件为 `Figure+Table/DMR-in-Genes.pdf`)

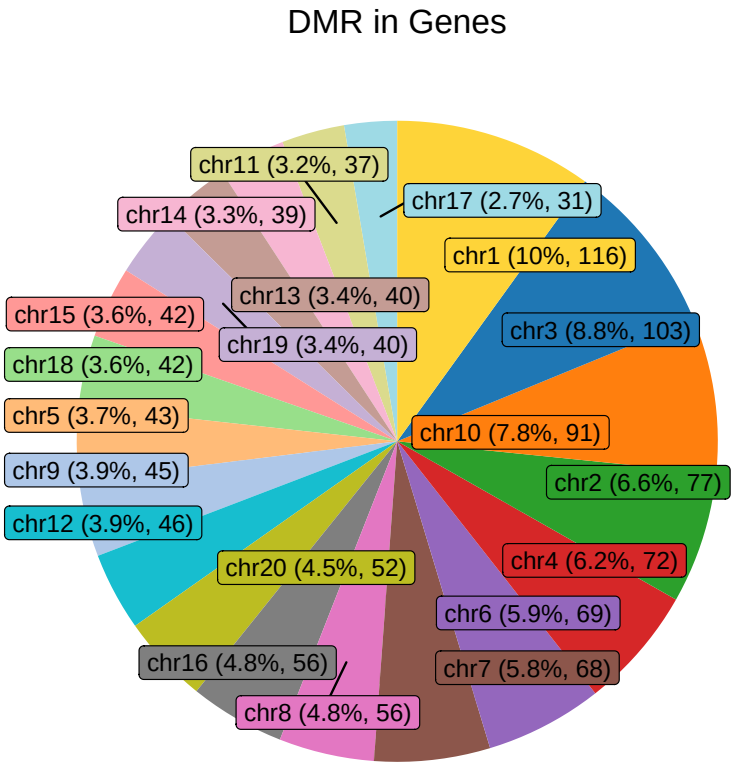

Figure 7: DMR in Genes

6.1.3 CpG Island

Figure 8 (下方图) 为图 Specific methylation 概览。

(对应文件为 `Figure+Table/Specific-methylation.pdf`)

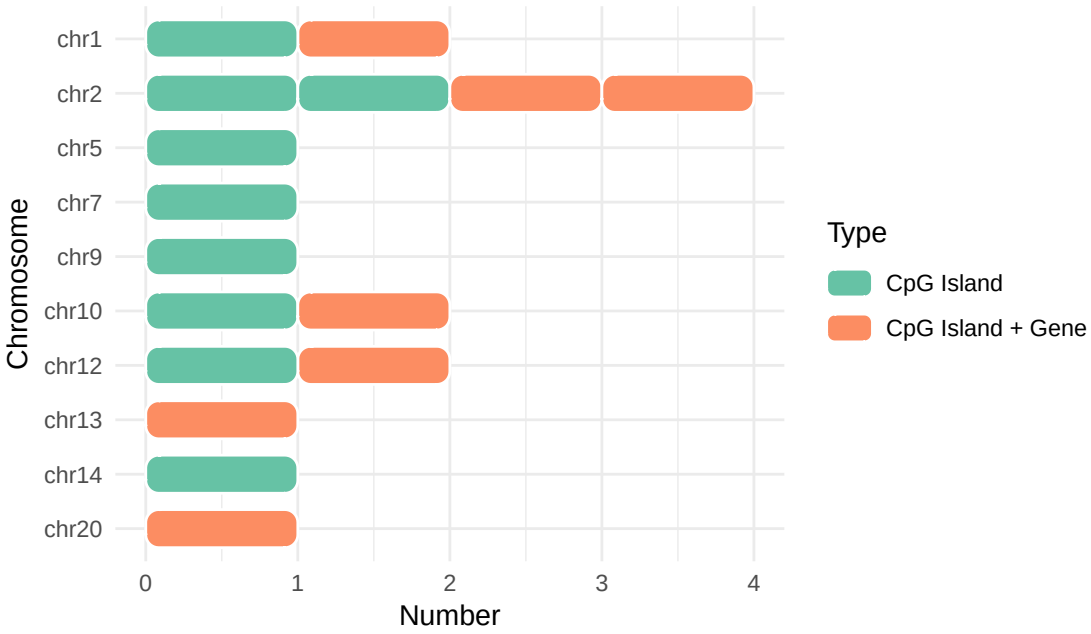

Figure 8: Specific methylation

Figure 9 (下方图) 为图 CpG Island methylation 概览。

(对应文件为 `Figure+Table/CpG-Island-methylation.pdf`)

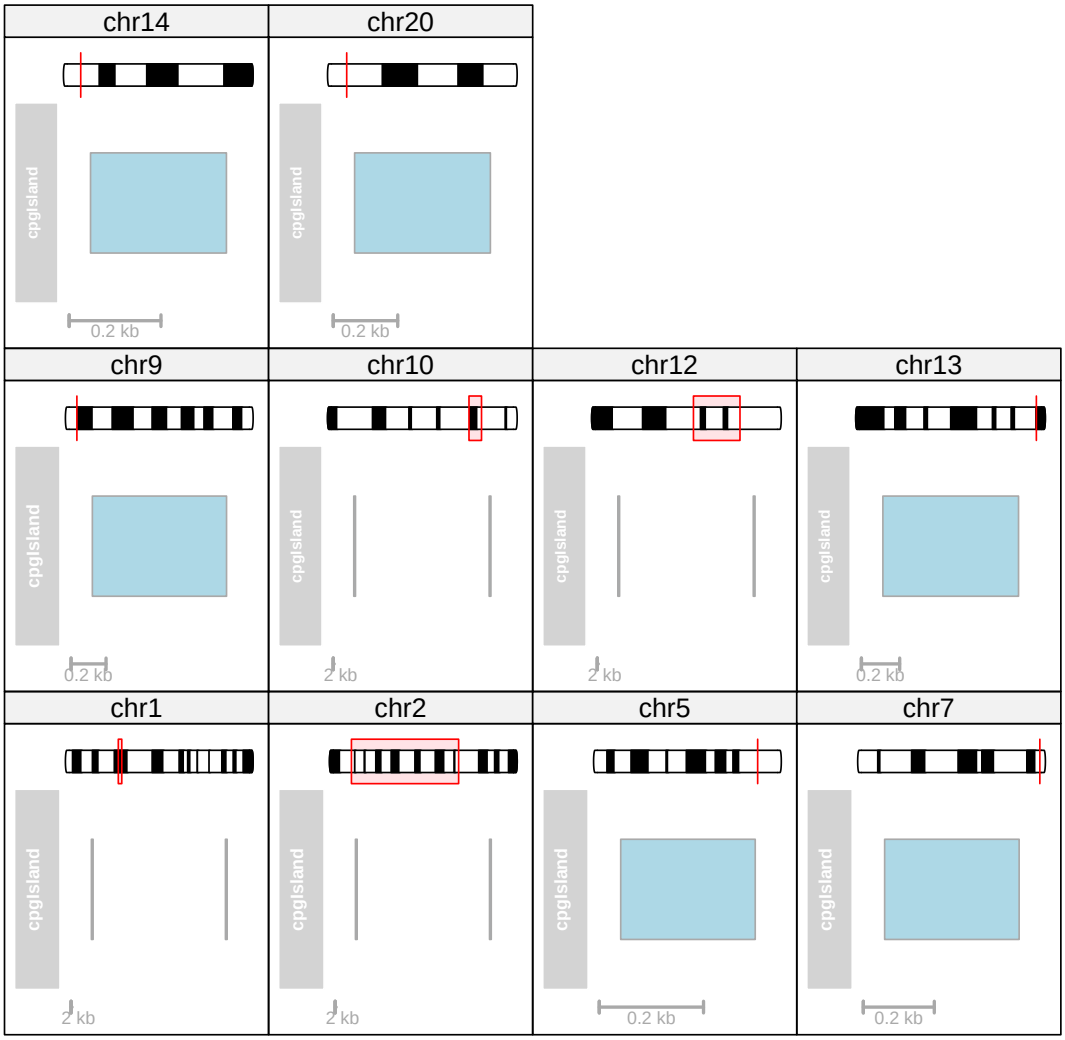

Figure 9: CpG Island methylation

6.1.4 DMR plot

6.1.5 富集分析

6.1.6 Enrichment

Figure 10 (下方图) 为图 DMR GO enrichment 概览。

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

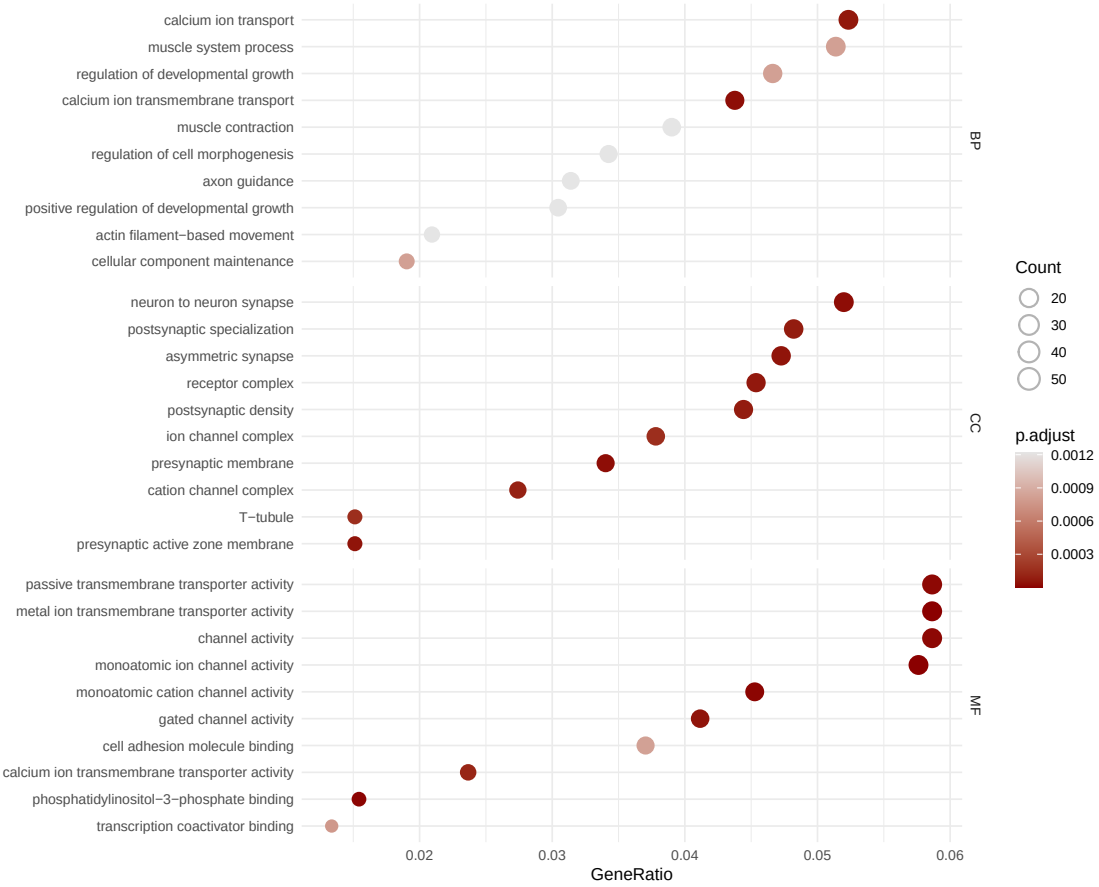

Figure 10: DMR GO enrichment

Figure 11 (下方图) 为图 DMR KEGG enrichment 概览。

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

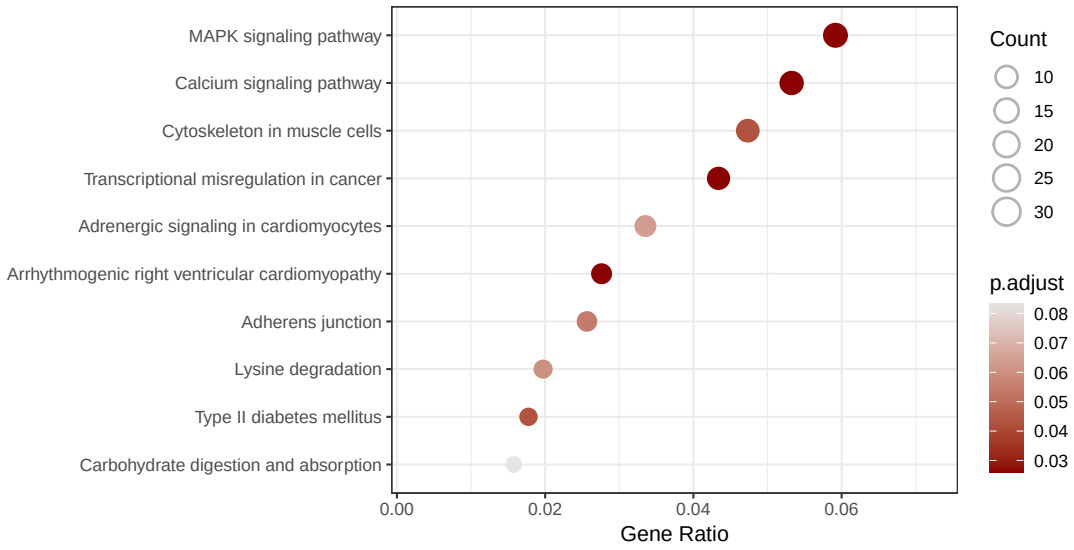

Figure 11: DMR KEGG enrichment

Figure 12 (下方图) 为图 DMR rno04930 visualization 概览。

(对应文件为 `Figure+Table/DMR-rno04930-visualization.png`)

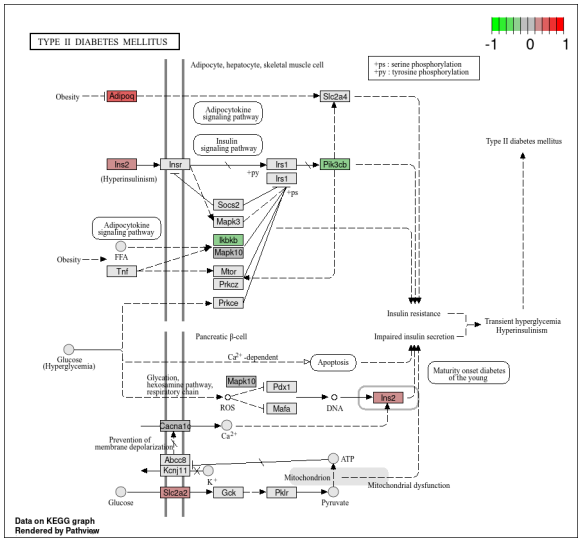

Figure 12: DMR rno04930 visualization

Interactive figure :

https://www.genome.jp/pathway/rno04930

Enriched genes :

Pik3cb, Ins2, Adipoq, Ikbkb, Mapk10, Cacna1c, Slc2a2

6.1.7 Ins2 和 Pik3cb

Figure 13 (下方图) 为图 Chr8 Pik3cb DMR annotation 概览。

(对应文件为 `Figure+Table/Chr8-Pik3cb-DMR-annotation.pdf`)

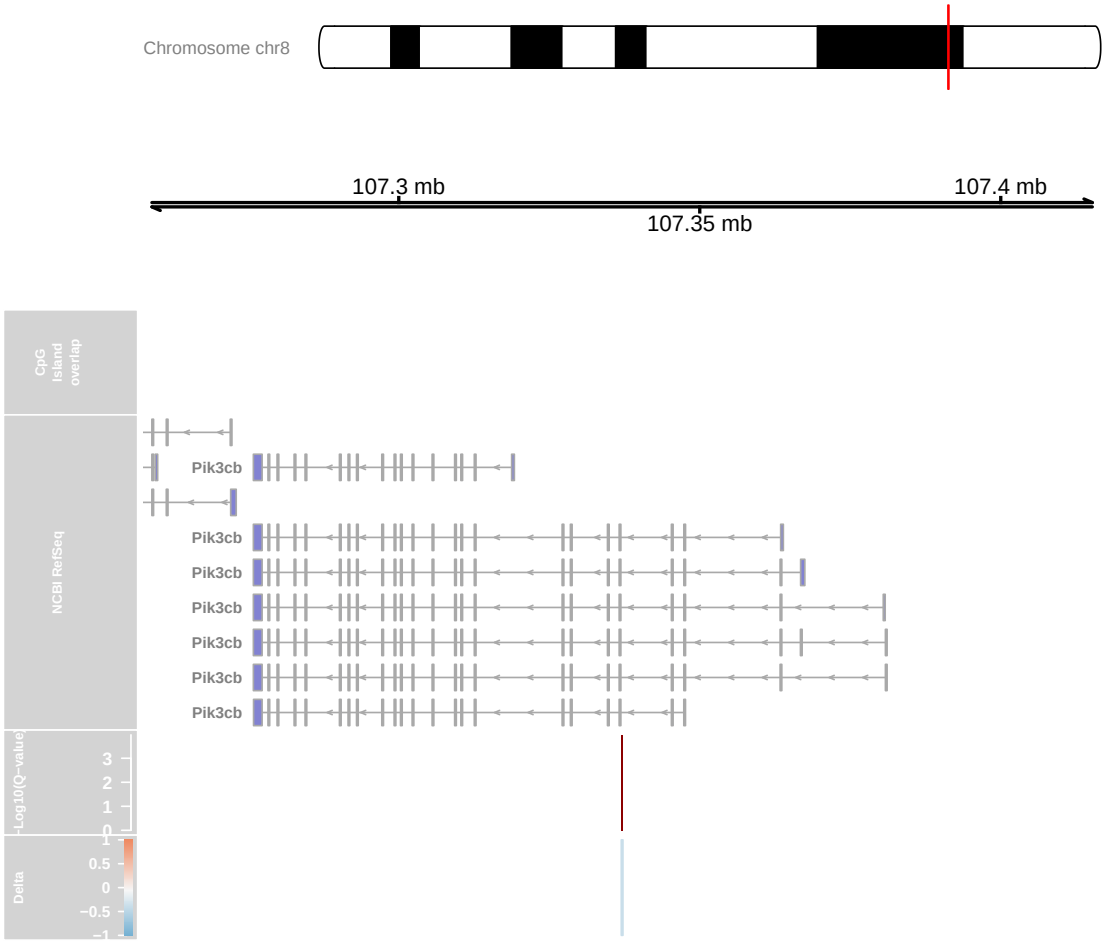

Figure 13: Chr8 Pik3cb DMR annotation

Figure 14 (下方图) 为图 Chr8 DMR annotation 概览。

(对应文件为 `Figure+Table/Chr8-DMR-annotation.pdf`)

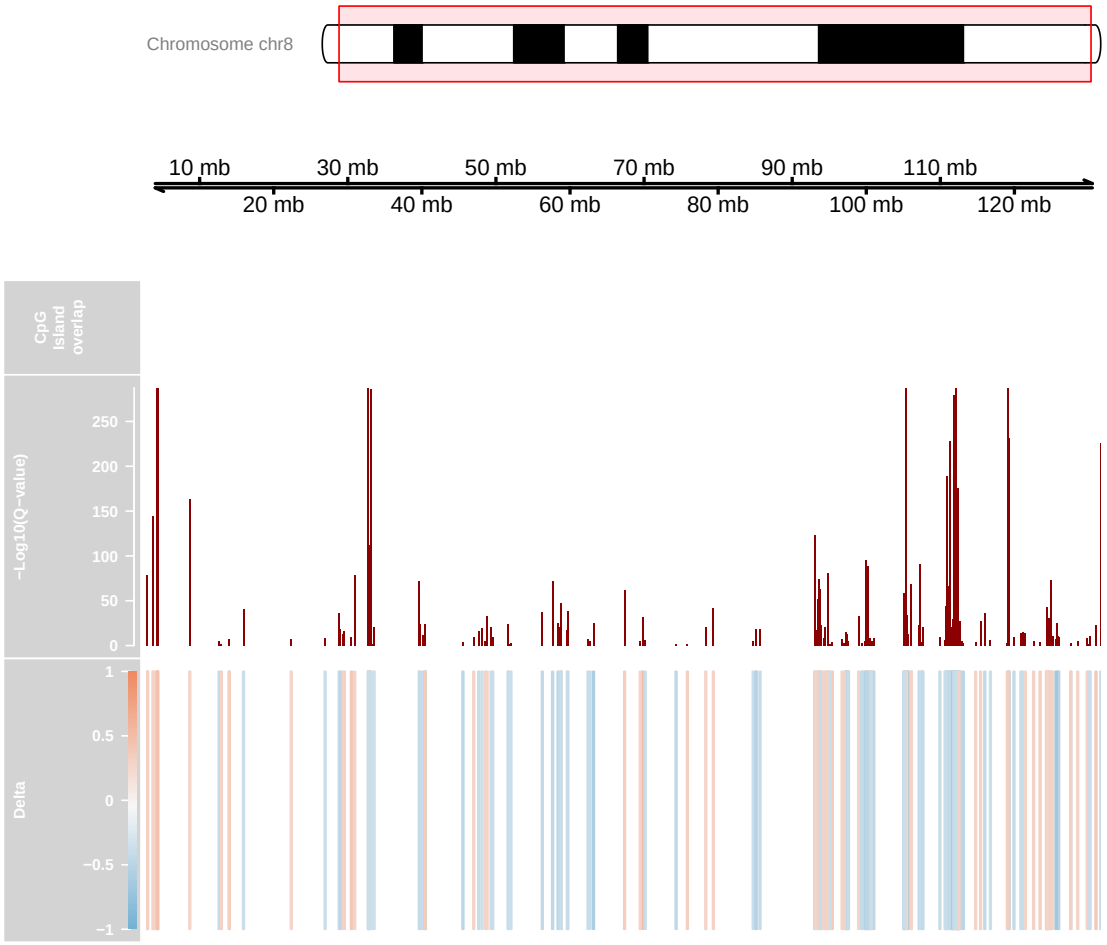

Figure 14: Chr8 DMR annotation

Figure 15 (下方图) 为图 Chr1 Ins2 DMR annotation 概览。

(对应文件为 `Figure+Table/Chr1-Ins2-DMR-annotation.pdf`)

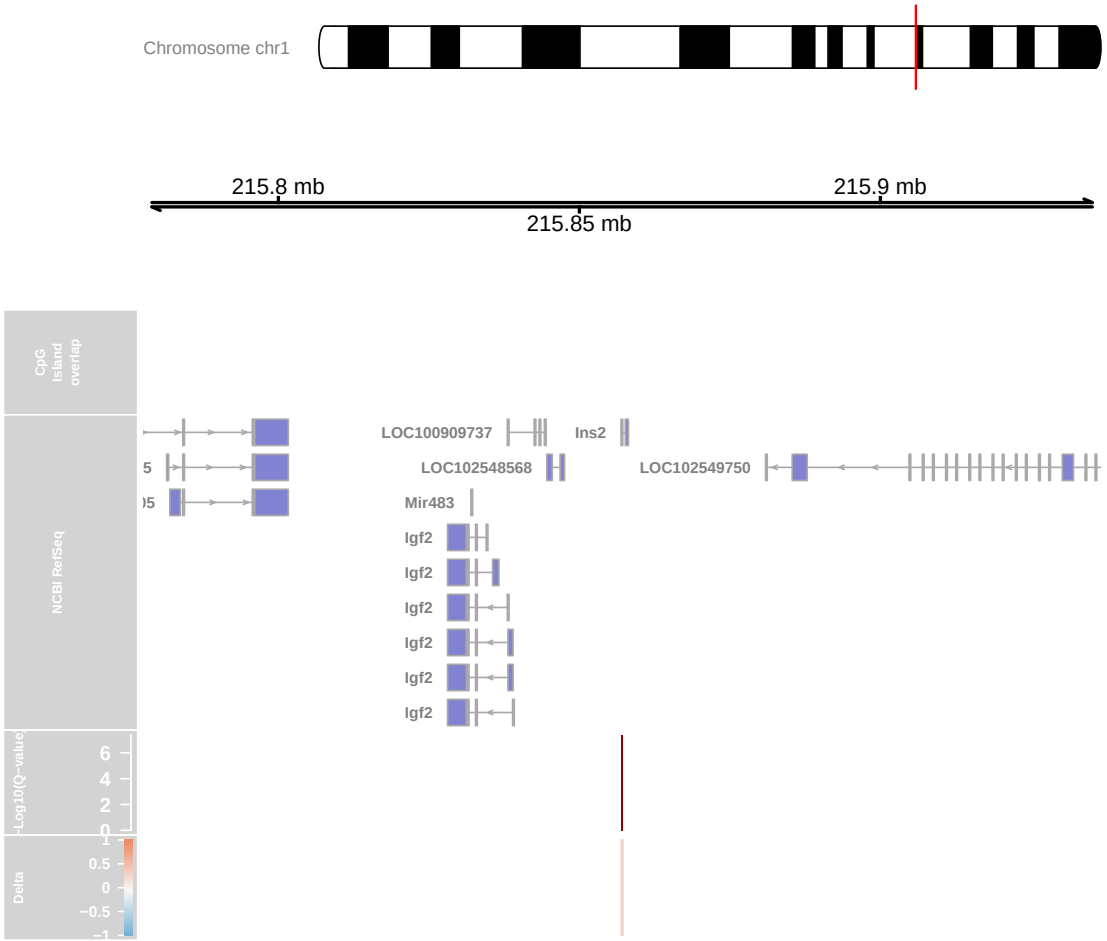

Figure 15: Chr1 Ins2 DMR annotation

Figure 16 (下方图) 为图 Chr1 DMR annotation 概览。

(对应文件为 `Figure+Table/Chr1-DMR-annotation.pdf`)

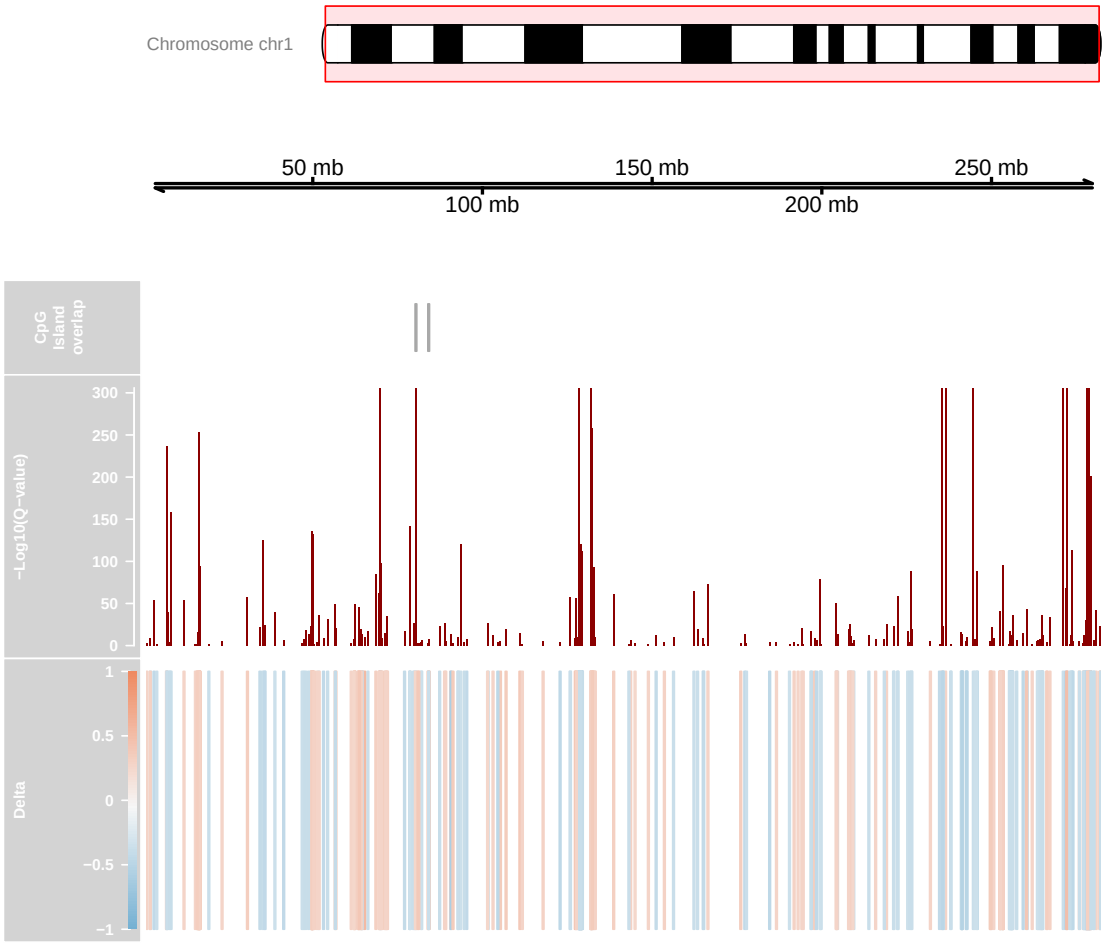

Figure 16: Chr1 DMR annotation

6.2 Diabetes mellitus

Figure 17 (下方图) 为图 Overall targets number of datasets 概览。

(对应文件为 `Figure+Table/Overall-targets-number-of-datasets.pdf`)

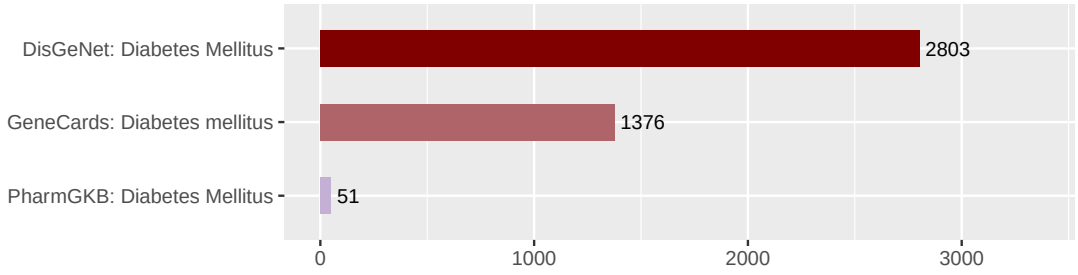

Figure 17: Overall targets number of datasets

Table 2 (下方表格) 为表格 mapped from human to rat 概览。

(对应文件为 `Figure+Table/mapped-from-human-to-rat.csv`)

注：表格共有 3245 行 2 列，以下预览的表格可能省略部分数据；含有 2789 个唯一 'hgnc_symbol'。

1. hgnc_symbol: 基因名 (Human)

Table 2: Mapped from human to rat

hgnc_symbol	rgd_symbol
RBM45	Rbm45
MFAP1	Mfap1a
THBS2	Thbs2
ACSS2	Acss2
NDUFV1	Ndufv1
NDUFAF5	Ndufaf5
CST3	Andpro
ARNTL	Arntl
POLD1	Pold1
KCNQ1	Kcnq1
PDE4D	Pde4d
NOX4	Nox4

hgnc_symbol	rgd_symbol
FOXO1	Foxo1
UMOD	Umod
AQP1	Aqp1
...	...

6.3 Methyl-seq 与 DM

6.3.1 Intersection

Figure 18 (下方图) 为图 Intersection of Me seq with DM 概览。

(对应文件为 `Figure+Table/Intersection-of-Me-seq-with-DM.pdf`)

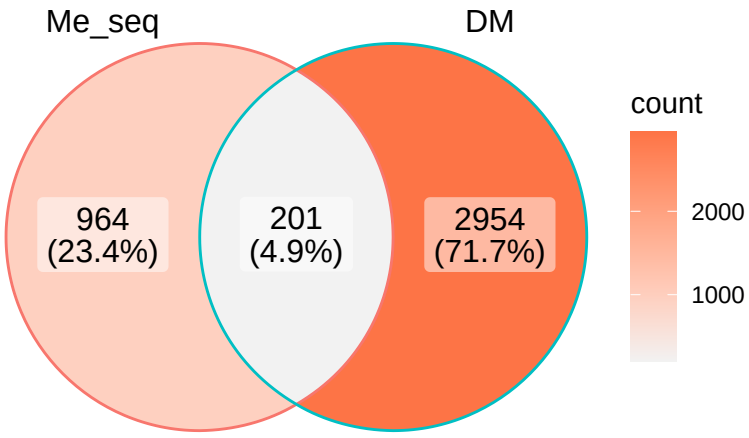

Figure 18: Intersection of Me seq with DM

All_intersection：

Gtf2ird2, Phlpp1, Ush2a, Fbxo25, Trpm7, Lpin1, Kitlg, Ahi1, Opa1, Cidea, P2rx7, Hmgcs2, P2rx4, Satb1, Agxt2, Sema3e, Nr0b2, Igf1r, Ppard, Atp2b2, Bcl2l11, Slpi, Ebf2, Itgax, Thrb, Gcg, Tg, Tcf4, Dcn, Alcam, Ece1, Tp63, Pex6, Fgfr1, Tshr, Atp2a2, Ephb1, Fat1, Ngf, Ube2q2, Spg7, Hemgn, Txn2, Nphp1,...

(上述信息框内容已保存至 `Figure+Table/Intersection-of-Me-seq-with-DM-content`)

6.3.2 StringDB

Figure 19 (下方图) 为图 Raw PPI network 概览。

(对应文件为 `Figure+Table/Raw-PPI-network.pdf`)

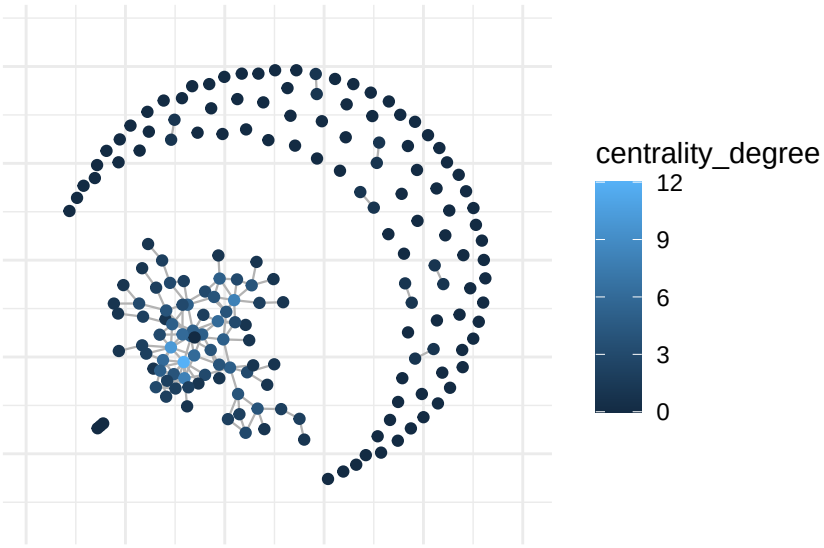

Figure 19: Raw PPI network

Figure 20 (下方图) 为图 Top30 MCC score 概览。

(对应文件为 `Figure+Table/Top30-MCC-score.pdf`)

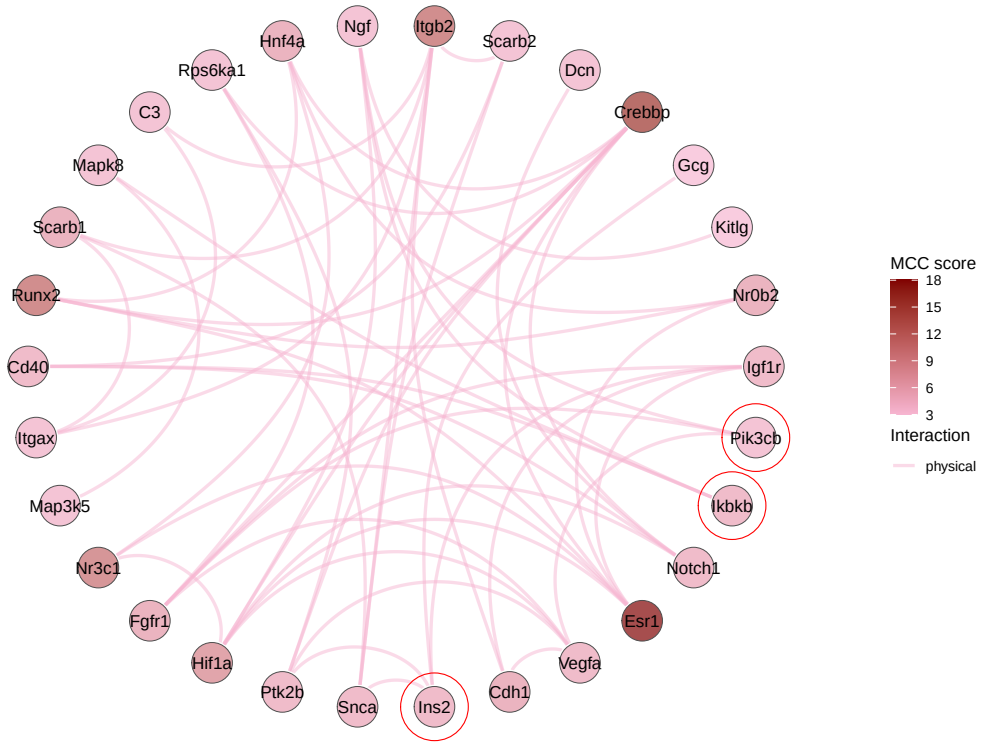

Figure 20: Top30 MCC score

Figure 21 (下方图) 为图 DME DM genes to other genes 概览。

(对应文件为 `Figure+Table/DME-DM-genes-to-other-genes.pdf`)

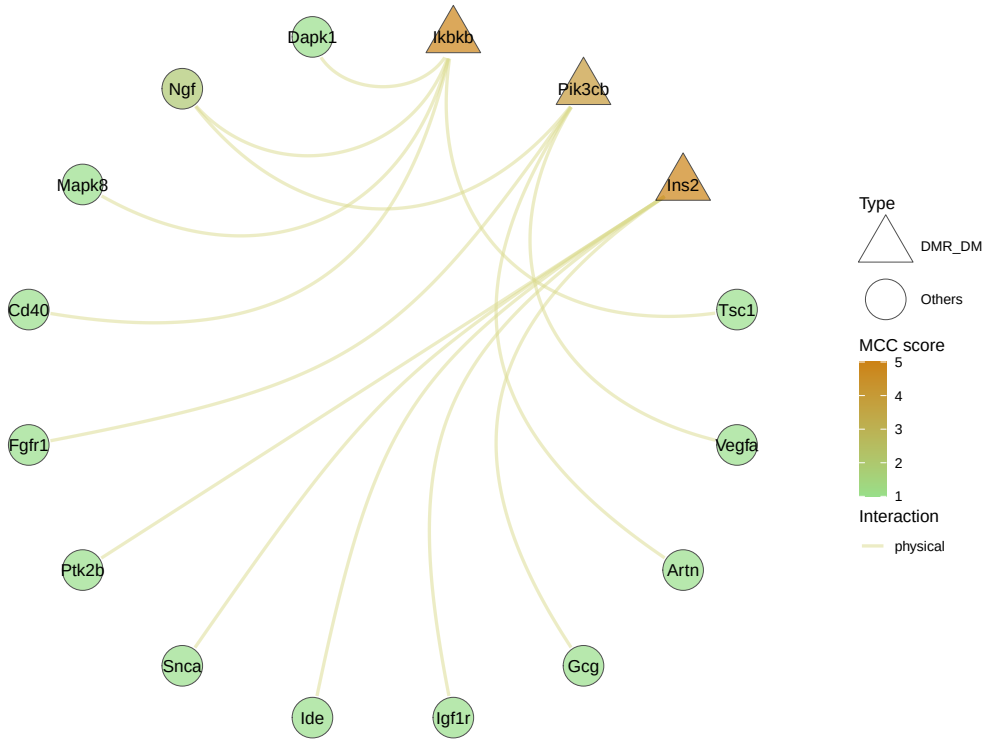

Figure 21: DME DM genes to other genes

Reference

1. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomaRt. *Nature protocols* **4**, 1184–1191 (2009).

2. Hahne, F. & Ivanek, R. Visualizing genomic data using gviz and bioconductor. in *Methods in Molecular Biology* 335–351 (Springer New York, 2016). doi:10.1007/978-1-4939-3578-9_16.

3. Wu, T. *et al.* ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, (2021).

4. Piñero, J. *et al.* The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* (2019) doi:10.1093/nar/gkz1021.

5. Stelzer, G. *et al.* The genecards suite: From gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics* **54**, 1.30.1–1.30.33 (2016).

6. Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley interdisciplinary reviews. Systems biology and medicine* **10**, (2018).

7. Szklarczyk, D. *et al.* The string database in 2021: Customizable proteinprotein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605–D612 (2021).

8. Chin, C.-H. *et al.* CytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology* **8**, S11 (2014).

9. Lawrence, M., Gentleman, R. & Carey, V. Rtracklayer: An r package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

10. Wu, H., Caffo, B., Jaffee, H. A., Irizarry, R. A. & Feinberg, A. P. Redefining cpg islands using hidden markov models. *Biostatistics (Oxford, England)* **11**, 499–514 (2010).

11. Luo, W. & Brouwer, C. Pathview: An r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics (Oxford, England)* **29**, 1830–1831 (2013).

12. Ramasubbu, K. & Devi Rajeswari, V. Impairment of insulin signaling pathway pi3k/akt/mTOR and insulin resistance induced ages on diabetes mellitus and neurodegenerative diseases: A perspective review. *Molecular and cellular biochemistry* **478**, 1307–1324 (2023).

13. Zhou, Z., Sun, B., Li, X. & Zhu, C. DNA methylation landscapes in the pathogenesis of type 2 diabetes mellitus. *Nutrition \& Metabolism* **15**, (2018).