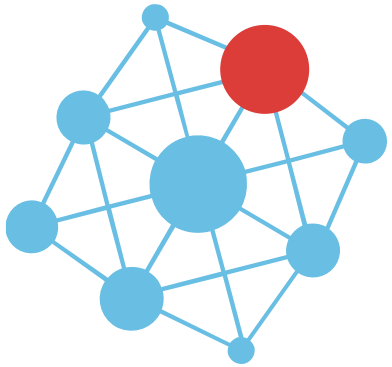
中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

BIO303 Final Year Project

Student id: 1931391

2024-05-22

Xi'an Jiaotong-Liverpool University
西交利物浦大学



Ⅰ. Introduction

- Concepts of batch effects
- Bulk RNA-seq work flow
- Quantification pipeline

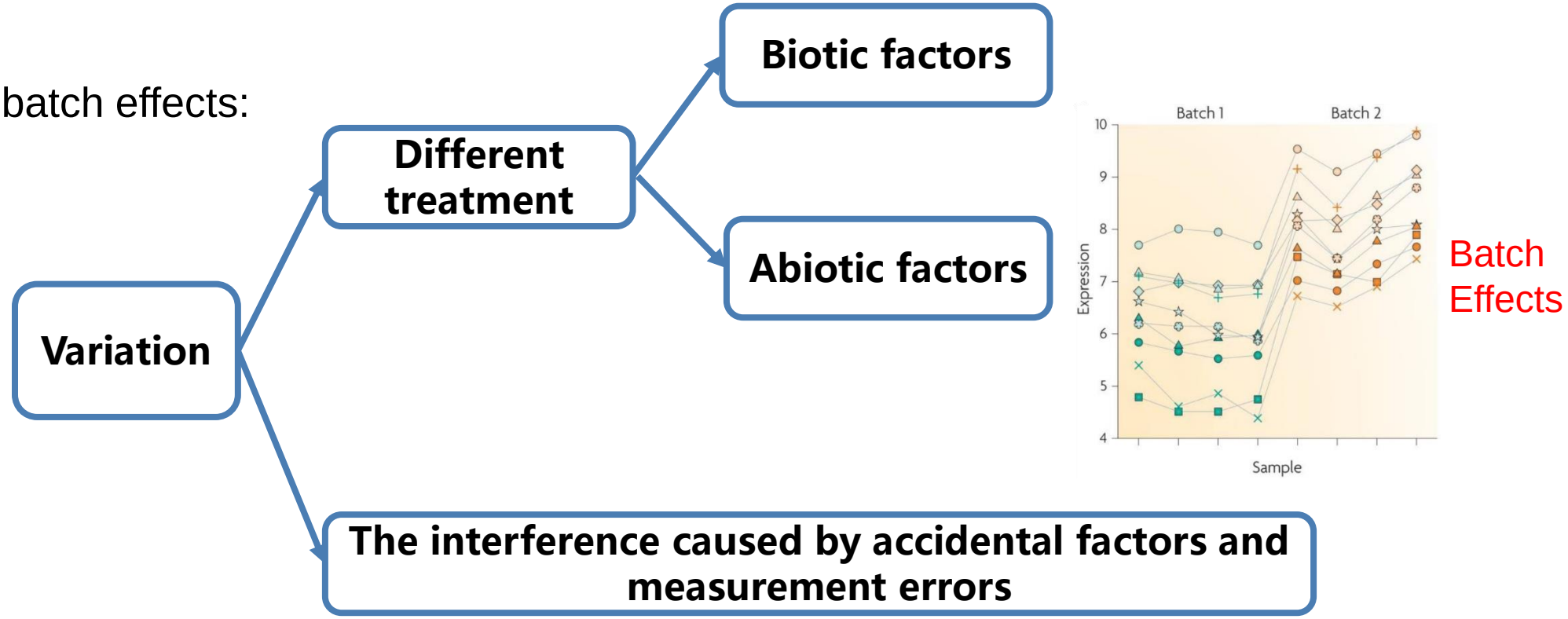
Ⅱ. Methods

Ⅲ. Results & Discussion

1 Concept of batch effects

Xi'an Jiaotong-Liverpool University 西交利物浦大学

- What is batch effects:



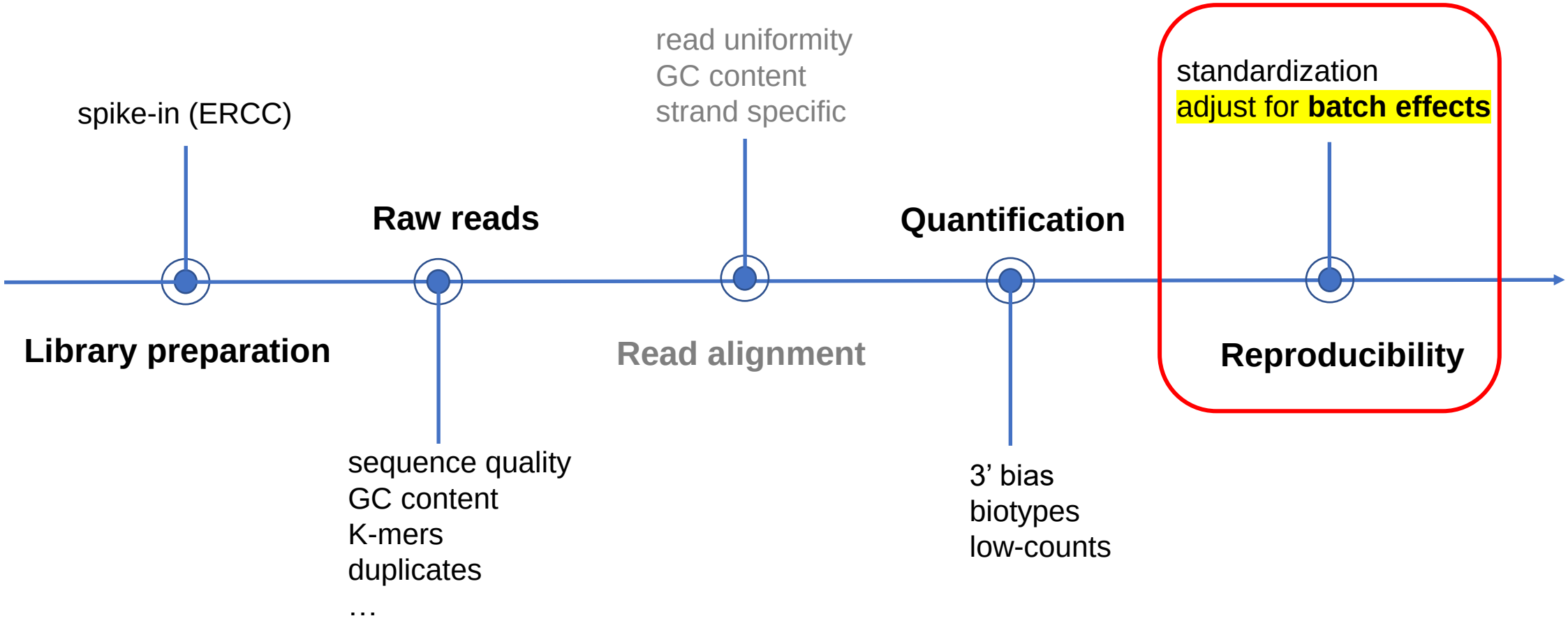
- Benefits of adjust for batch effects: 1. Increase repeatability of results; 2. Reduce false positive or false negative results.

Conesa et al., *Genome Biology*, 2016

1 RNA-seq quality control steps

Xi'an Jiaotong-Liverpool University 西交利物浦大学

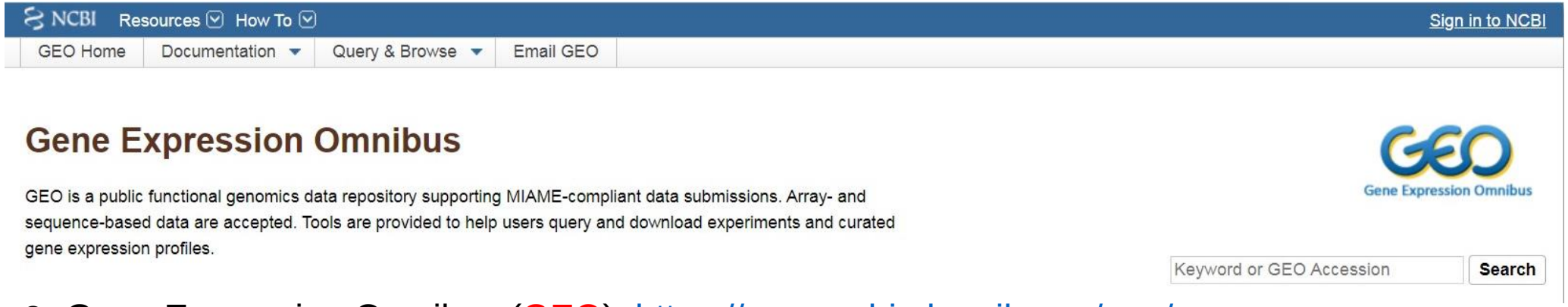
Data quality control is maintained throughout the RNA-seq analysis



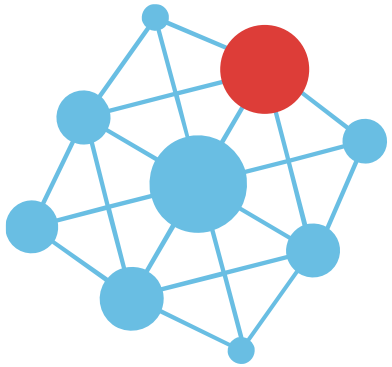
Conesa et al., *Genome Biology*, 2016

1 Quantification pipeline

- Gene Expression Nebulas (GEN): https://ngdc.cncb.ac.cn/gen/
- One standardized gene expression quantification pipeline to quantify all datasets



- Gene Expression Omnibus (GEO): https://www.ncbi.nlm.nih.gov/geo/
- Different researchers use different quantification methods to quantify their dataset



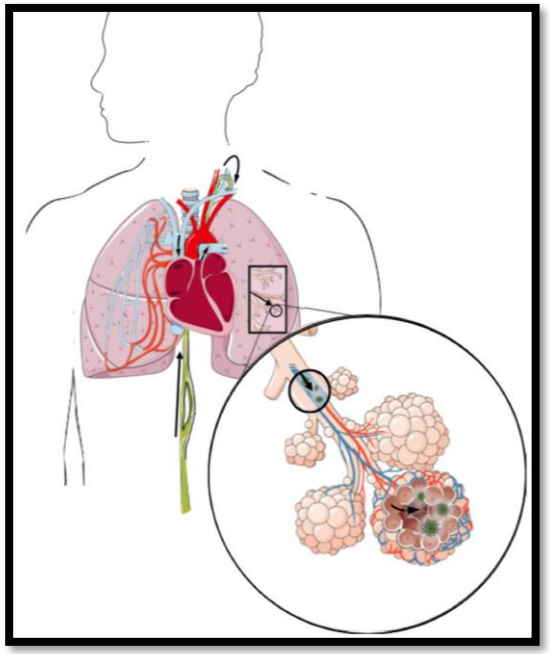
Ⅰ. Introduction

Ⅱ. Methods

- Materials
- Project design

Ⅲ. Results & Discussion

2 Materials

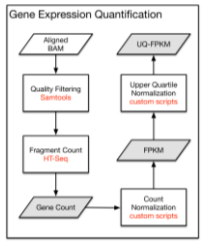


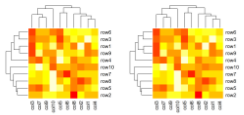
- **GSE147507**: Transcriptional response to SARS-CoV-2 infection (**3 Control** & **3 Treatment**).
- **GSE150962**: A safe inhalational treatment prevents SARS–CoV-2 viral replication in human airway epithelial cells (**3 Control** & **3 Treatment**).

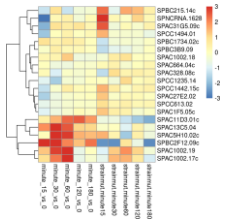
Both datasets are about SARS-COV-2 pulmonary airway epithelial cell infection, and their sequencing platform is GPL18573 Illumina NextSeq 500 (Homo sapiens). In these two experimental datasets, cells infected with SARS-CoV-2 from primary human bronchial epithelial cell were selected as the experimental group, while cells not infected with SARS-CoV-2 were selected as the control group.

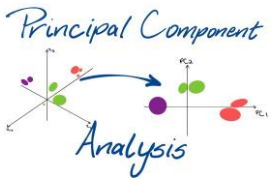
2 Project design

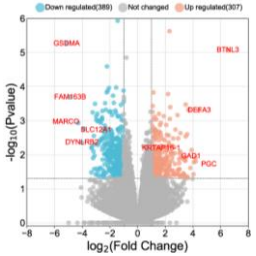
Xi'an Jiaotong-Liverpool University 西交利物浦大学











B1 fastq | **B2 fastq**

Fastp v0.20.0 ●
Hisat2 ●
RseQC v2.6.4 ●
STAR 2.7.1a ●
RSEM v1.3.1 ●

GEN
Standardized pipeline

● Fastp v0.20.0
● Hisat2
● RseQC v2.6.4
● STAR 2.7.1a
● RSEM v1.3.1

B1 count matrix (GEN) | **B2 count matrix (GEN)**

Combine

6 healthy control group (3C+3C)
6 SARS-CoV-2 infection group (3T+3T)

Combined raw count matrix (GEN)

Adjusted for batch effects (RUVs)

Adjusted count matrix (GEN)

Downstream analysis

...

B1 fastq | **B2 fastq**

SRA toolkit ●
Trimmomatic v0.38 ●
STAR ●
Sambamba ●
Subread package v1.61 ●

GEO
Discrete pipeline

● NextSeq Control Software 4.0.1.41
● RTA 2.11.3
● STAR v.2.5.2b
● subread package v.1.5.1

B1 count matrix (GEO) | **B2 count matrix (GEO)**

Combine

6 healthy control group (3C+3C)
6 SARS-CoV-2 infection group (3T+3T)

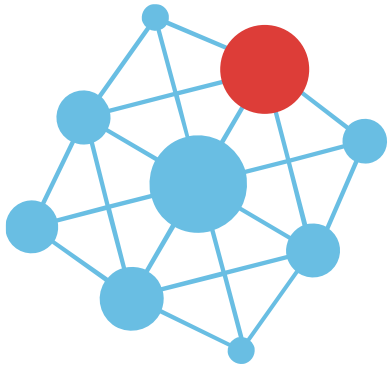
Combined raw count matrix (GEO)

Adjusted for batch effects (RUVs)

Adjusted count matrix (GEO)

Downstream analysis

...



Ⅰ. Introduction

Ⅱ. Methods

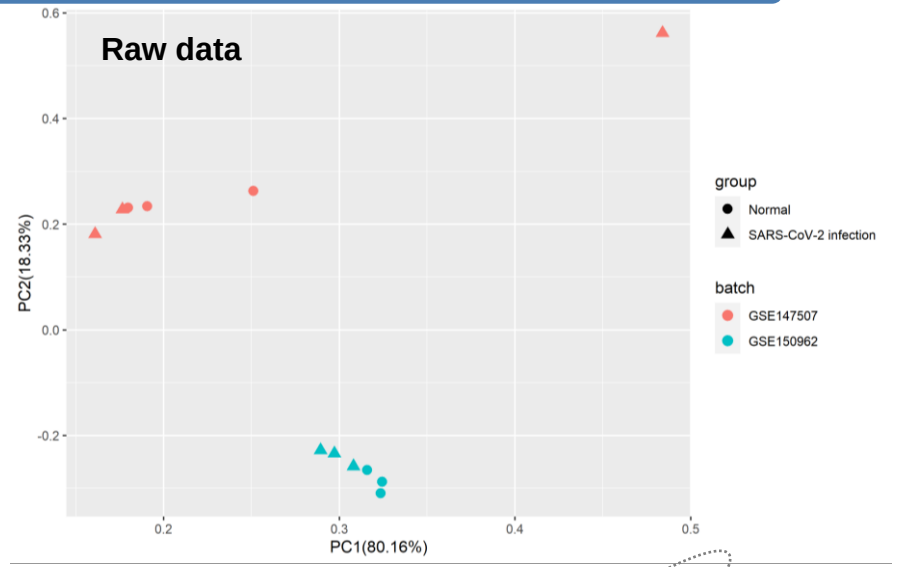
Ⅲ. Results & Discussion

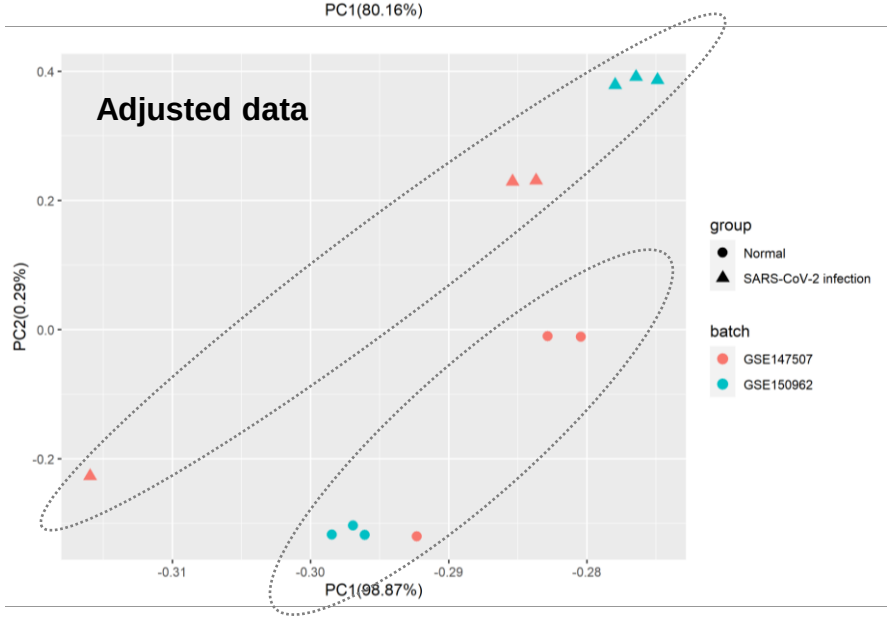
- Principal component analysis
- Heatmap
- Detect differential expression genes
- Go enrichment analysis

3 Principal Component Analysis

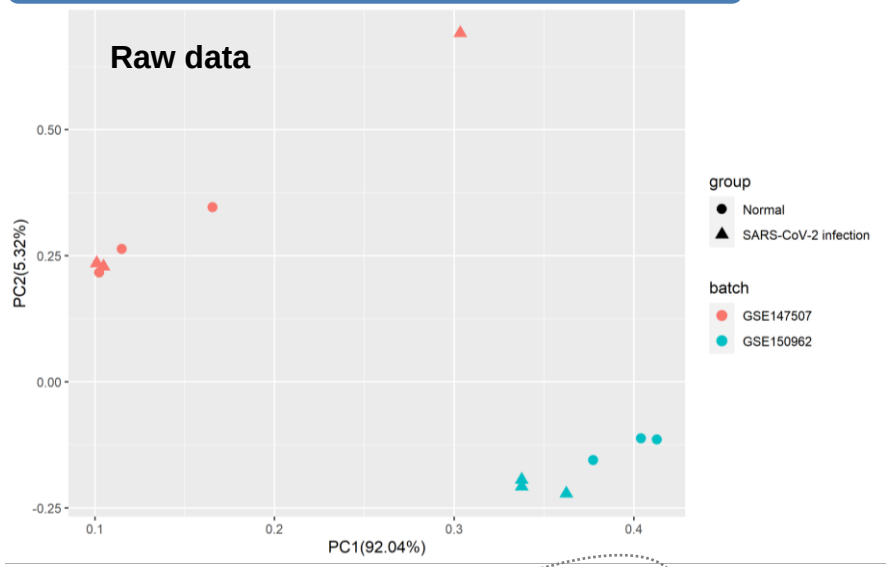
Xi'an Jiaotong-Liverpool University 西交利物浦大学

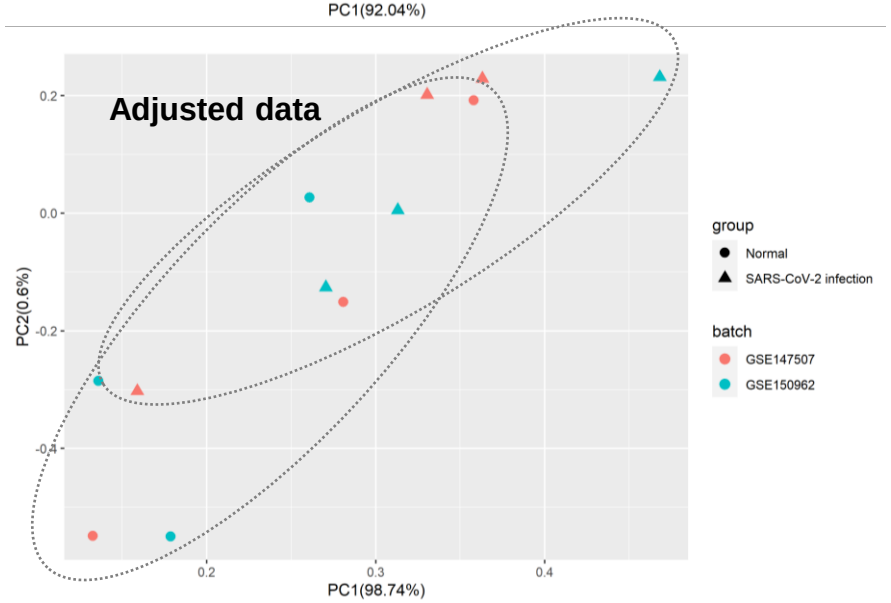
GEN standardized quantification pipeline





GEO discrete quantification pipeline

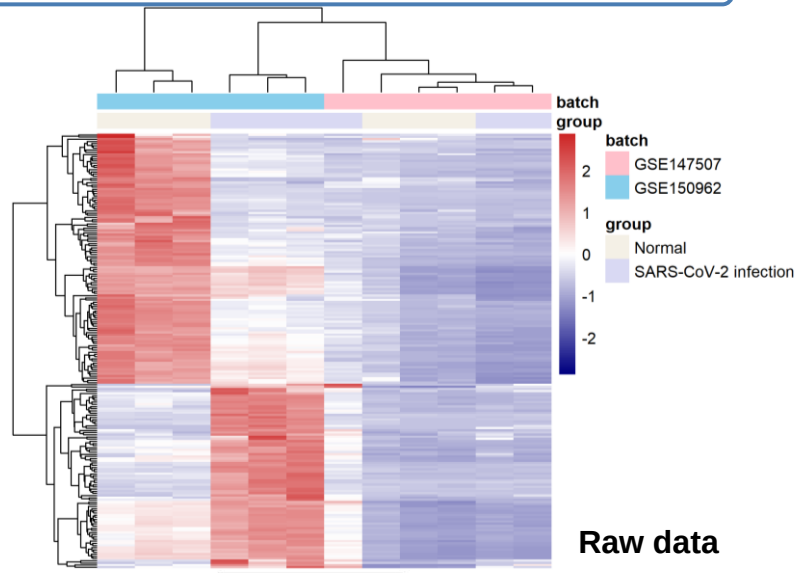




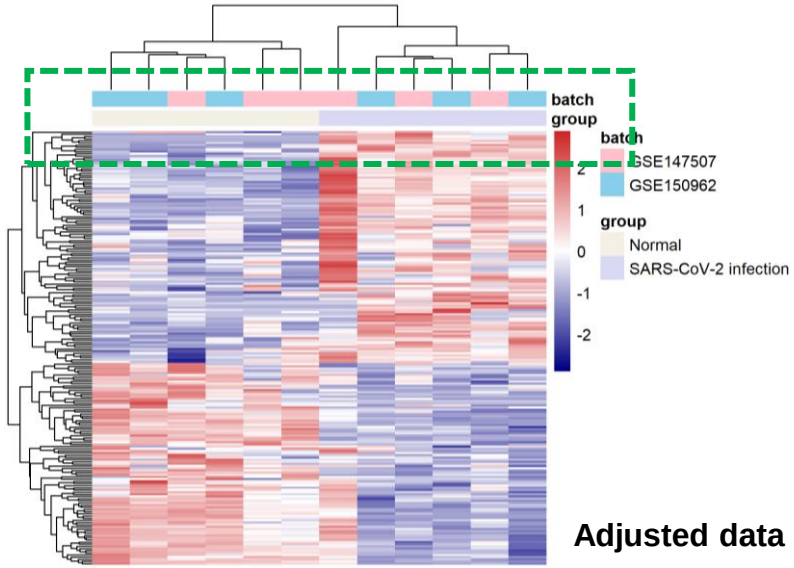
3 Heatmap

Xi'an Jiaotong-Liverpool University 西交利物浦大学

GEN standardized quantification pipeline

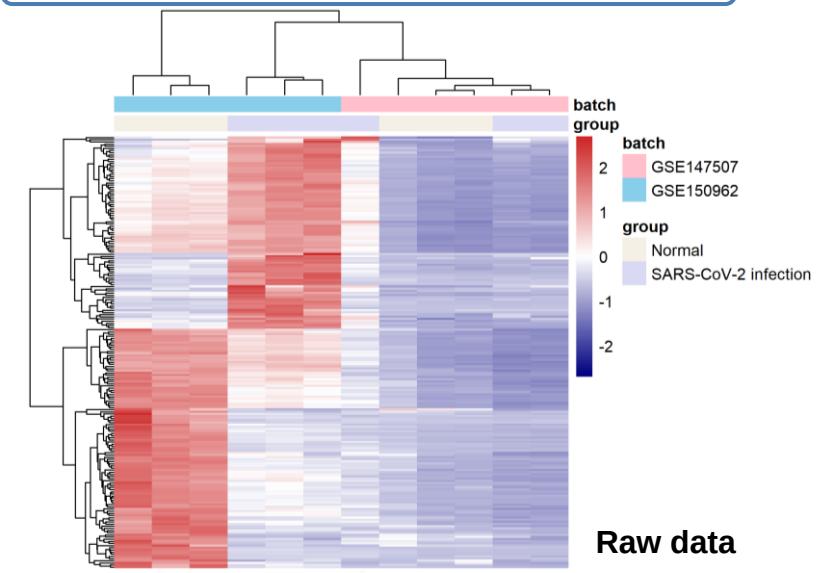


Raw data

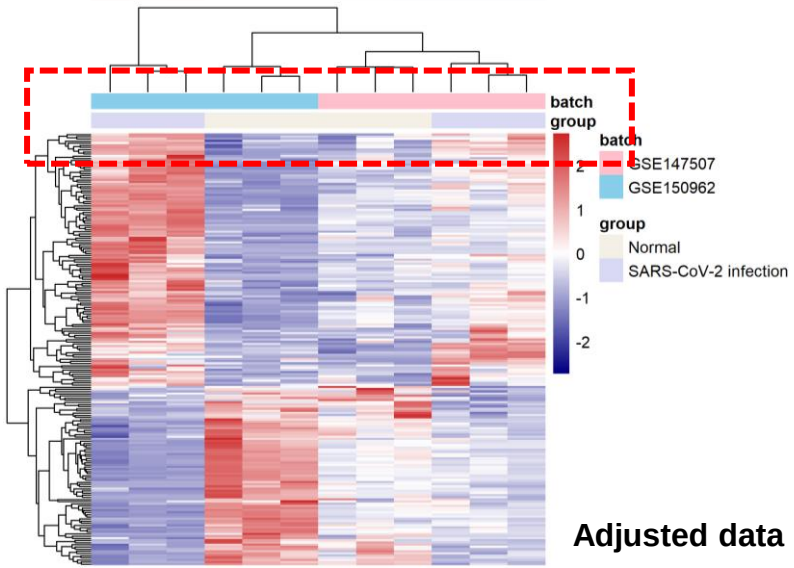


Adjusted data

GEO discrete quantification pipeline



Raw data

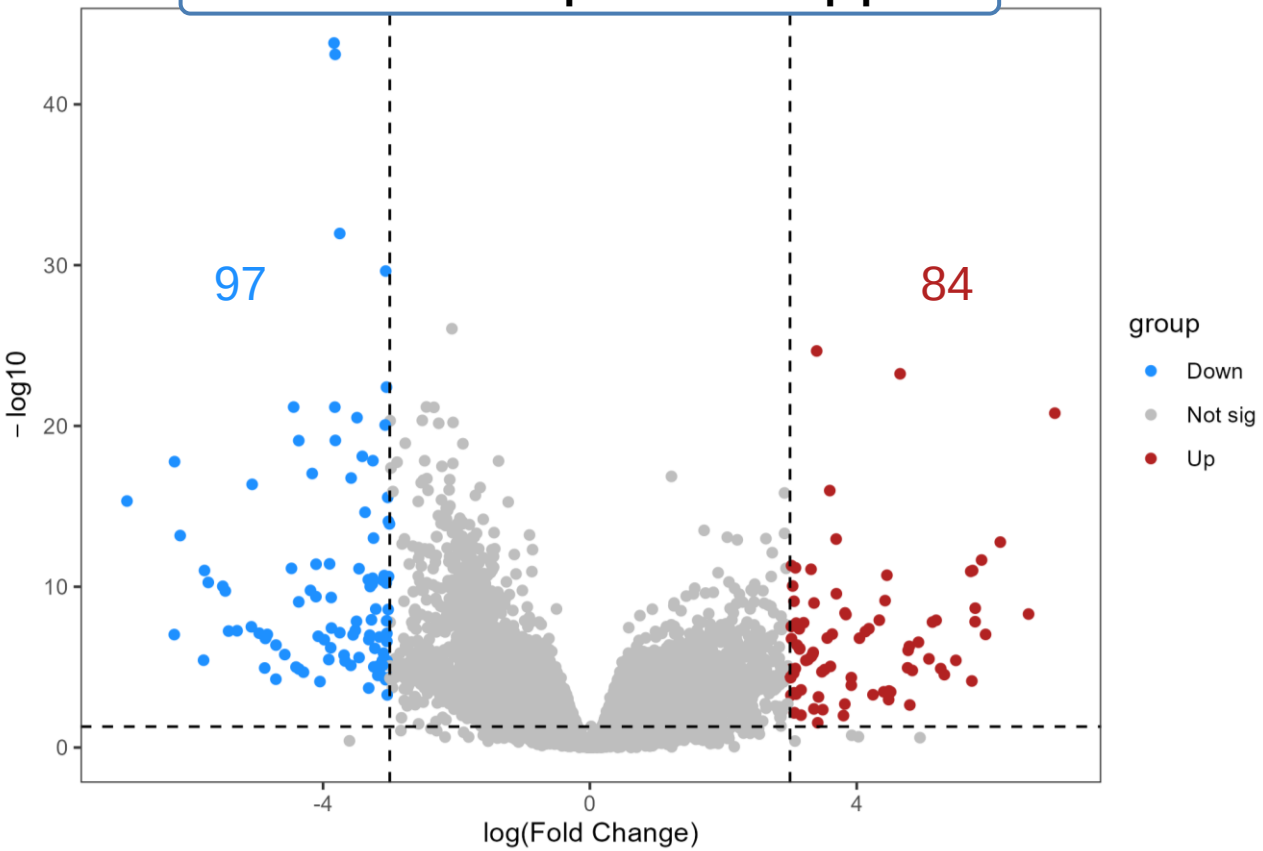


Adjusted data

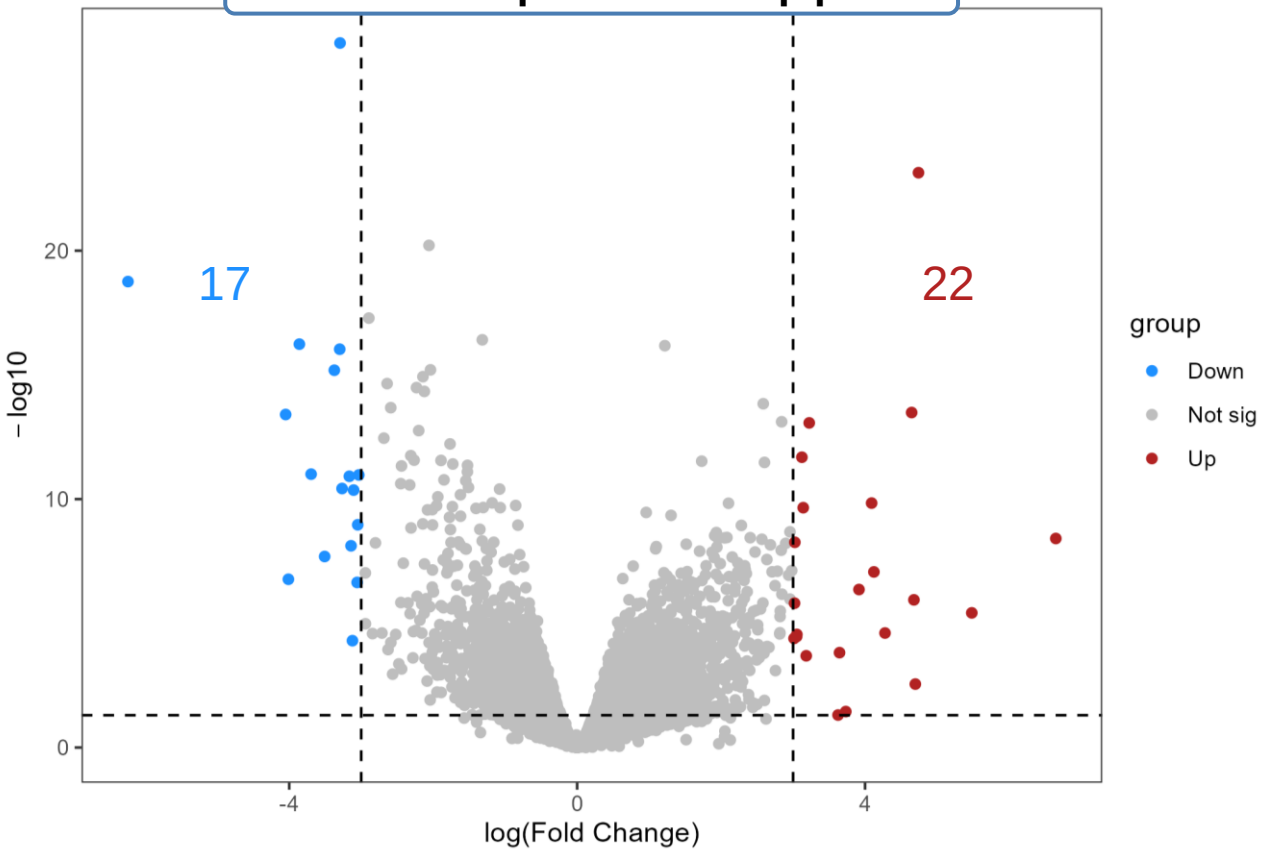
3 Volcano plot

Xi'an Jiaotong-Liverpool University 西交利物浦大学

GEN standardized quantification pipeline



GEO discrete quantification pipeline

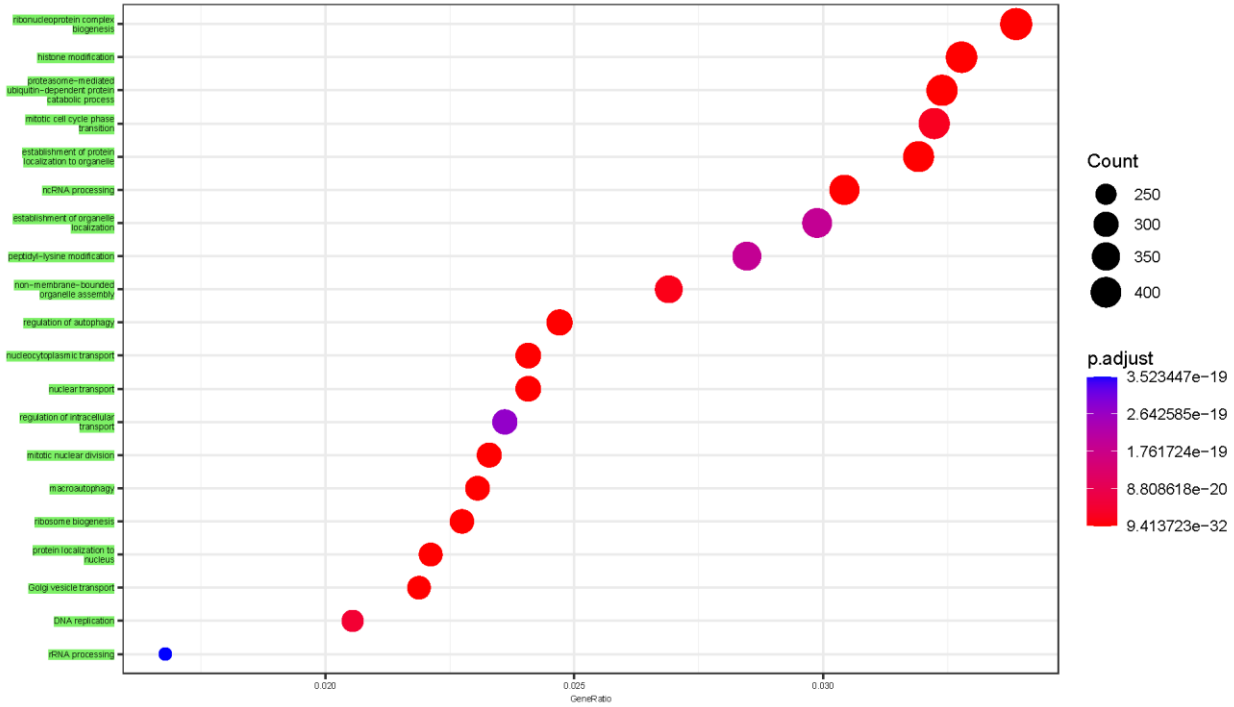


- Differential expression genes threshold: `P-value≤0.05` & `logFC≤-3 | logFC≥3`

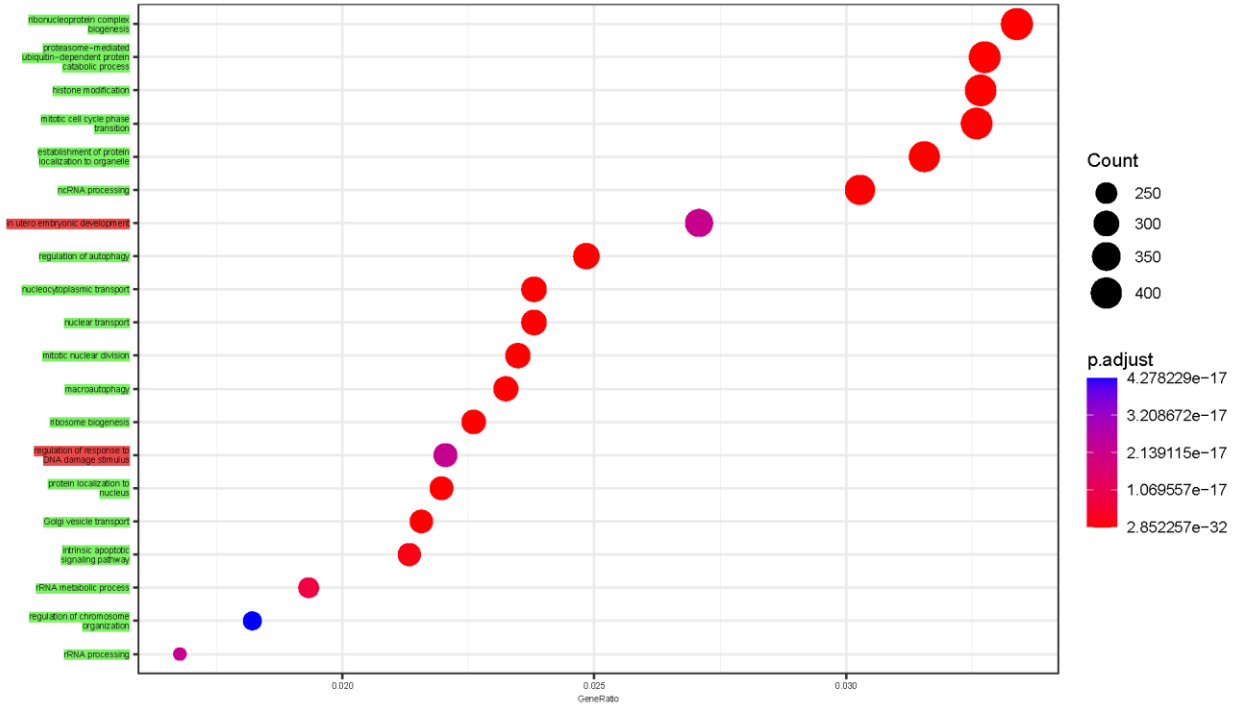
3 GO bubble plot

Xi'an Jiaotong-Liverpool University 西交利物浦大学

GEN standardized quantification pipeline



GEO discrete quantification pipeline



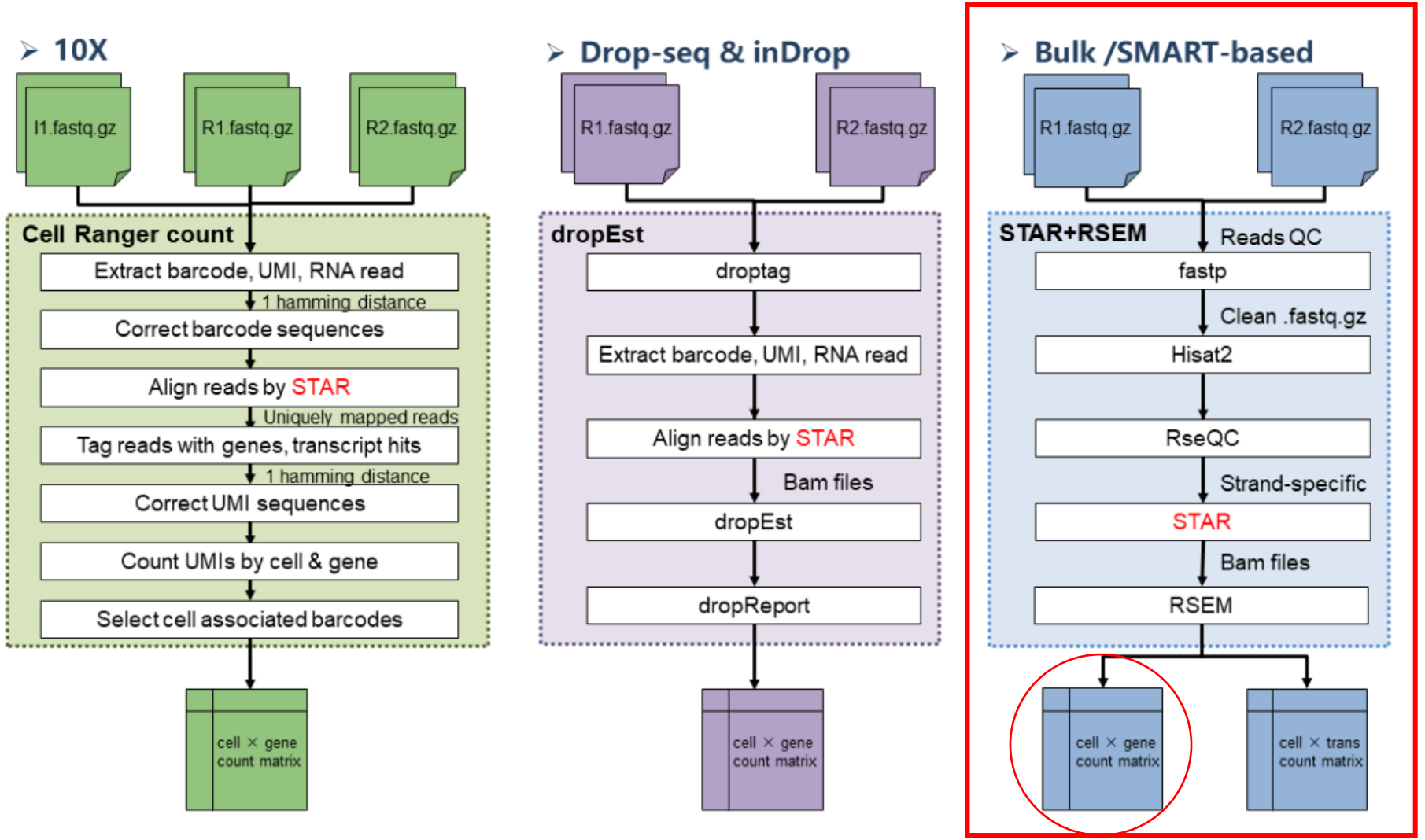
Use the corrected dataset from the discrete GEO quantification pipeline to do GO enrichment analysis will obtain more information that are unrelated to the experimental project compared with the standardized pipeline. For example, "in utero embryonic development" and "regulation of response to DNA damage stimulus" are indicative of a larger **false positive**.

Reference list

- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. **A survey of best practices for RNA-seq data analysis.** Genome Biol. 2016 Jan 26;17:13. doi: 10.1186/s13059-016-0881-8. Erratum in: Genome Biol. 2016;17(1):181. PMID: 26813401; PMCID: PMC4728800.
- Adiconis, X., Borges-Rivera, D., Satija, R. et al. **Comparative analysis of RNA sequencing methods for degraded or low-input samples.** Nat Methods 10, 623–629 (2013).
- Zhang Y, Zou D, Zhu T, Xu T, Chen M, Niu G, Zong W, Pan R, Jing W, Sang J, Liu C, Xiong Y, Sun Y, Zhai S, Chen H, Zhao W, Xiao J, Bao Y, Hao L, Zhang Z. **Gene Expression Nebulas (GEN): a comprehensive data portal integrating transcriptomic profiles across multiple species at both bulk and single-cell levels.** Nucleic Acids Res. 2022 Jan 7;50(D1):D1016-D1024. doi: 10.1093/nar/gkab878. PMID: 34591957; PMCID: PMC8728231.
- Risso D, Ngai J, Speed TP, Dudoit S. **Normalization of RNA-seq data using factor analysis of control genes or samples.** Nat Biotechnol. 2014 Sep;32(9):896-902. doi: 10.1038/nbt.2931. Epub 2014 Aug 24. PMID: 25150836; PMCID: PMC4404308.
- Fachrul M, Méric G, Inouye M, Pamp SJ, Salim A. **Assessing and removing the effect of unwanted technical variations in microbiome data.** Sci Rep. 2022 Dec 23;12(1):22236. doi: 10.1038/s41598-022-26141-x. PMID: 36564466; PMCID: PMC9789116.
- Daamen AR, Bachali P, Owen KA, Kingsmore KM, Hubbard EL, Labonte AC, Robl R, Shrotri S, Grammer AC, Lipsky PE. **Comprehensive transcriptomic analysis of COVID-19 blood, lung, and airway.** Sci Rep. 2021 Mar 29;11(1):7052. doi: 10.1038/s41598-021-86002-x. PMID: 33782412; PMCID: PMC8007747.
- CNCB-NGDC Members and Partners. **Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021.** Nucleic Acids Res. 2021 Jan 8;49(D1):D18-D28. doi: 10.1093/nar/gkaa1022. PMID: 33175170; PMCID: PMC7779035.

THANKS

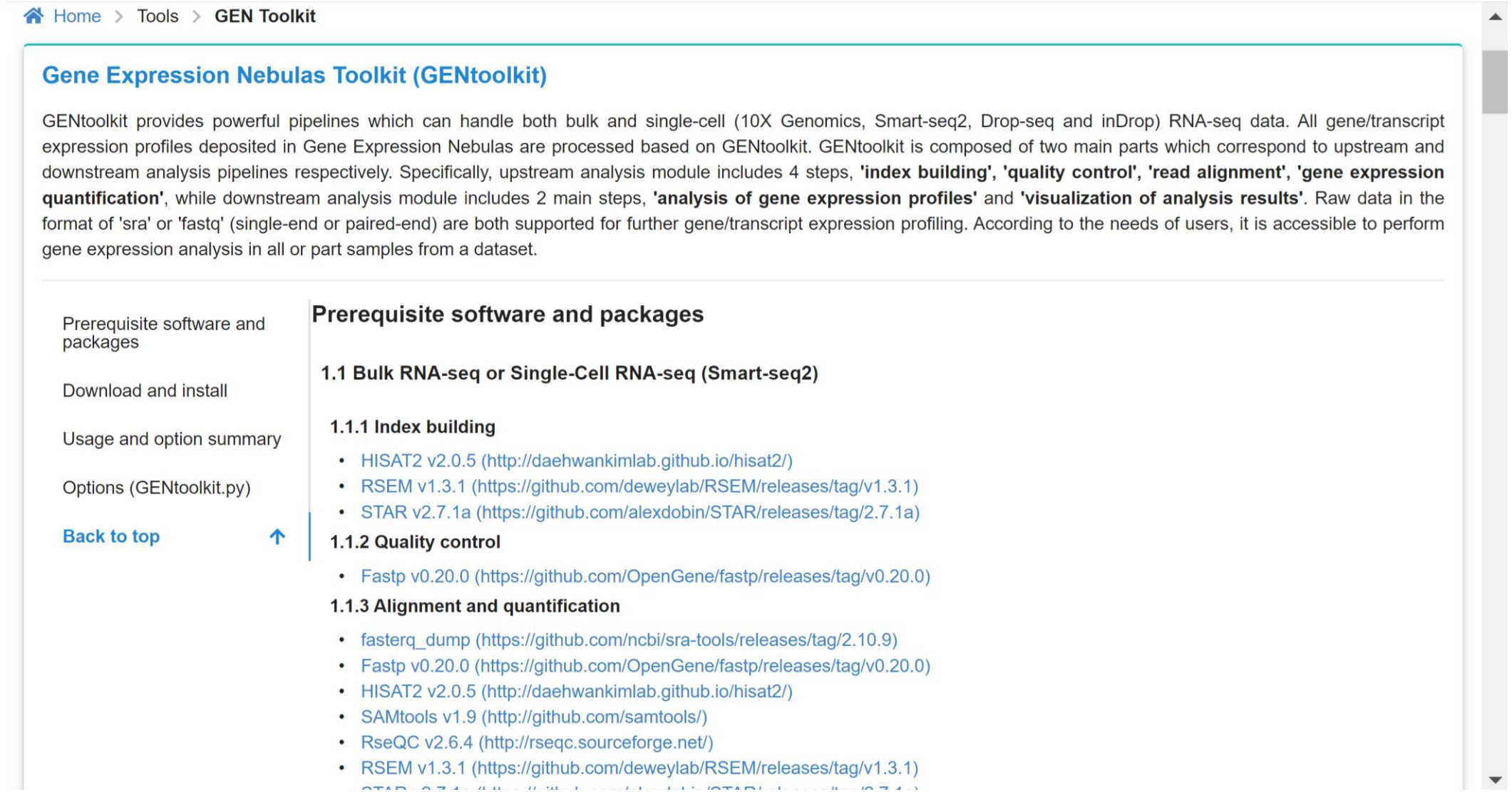
Appendix: GENToolkit Overview



- GENToolkit aim to analysis **Bulk** and **Single cell RNA-Seq** datasets

Appendix: GEN RNA-seq upstream analysis tools

Xi'an Jiaotong-Liverpool University 西交利物浦大学



Home > Tools > GEN Toolkit

Gene Expression Nebulas Toolkit (GENtoolkit)

GENtoolkit provides powerful pipelines which can handle both bulk and single-cell (10X Genomics, Smart-seq2, Drop-seq and inDrop) RNA-seq data. All gene/transcript expression profiles deposited in Gene Expression Nebulas are processed based on GENtoolkit. GENtoolkit is composed of two main parts which correspond to upstream and downstream analysis pipelines respectively. Specifically, upstream analysis module includes 4 steps, **'index building', 'quality control', 'read alignment', 'gene expression quantification'**, while downstream analysis module includes 2 main steps, **'analysis of gene expression profiles'** and **'visualization of analysis results'**. Raw data in the format of 'sra' or 'fastq' (single-end or paired-end) are both supported for further gene/transcript expression profiling. According to the needs of users, it is accessible to perform gene expression analysis in all or part samples from a dataset.

Prerequisite software and packages

Download and install

Usage and option summary

Options (GENtoolkit.py)

Back to top

Prerequisite software and packages

1.1 Bulk RNA-seq or Single-Cell RNA-seq (Smart-seq2)

1.1.1 Index building

- HISAT2 v2.0.5 (http://daehwankimlab.github.io/hisat2/)
- RSEM v1.3.1 (https://github.com/deweylab/RSEM/releases/tag/v1.3.1)
- STAR v2.7.1a (https://github.com/alexdobin/STAR/releases/tag/2.7.1a)

1.1.2 Quality control

- Fastp v0.20.0 (https://github.com/OpenGene/fastp/releases/tag/v0.20.0)

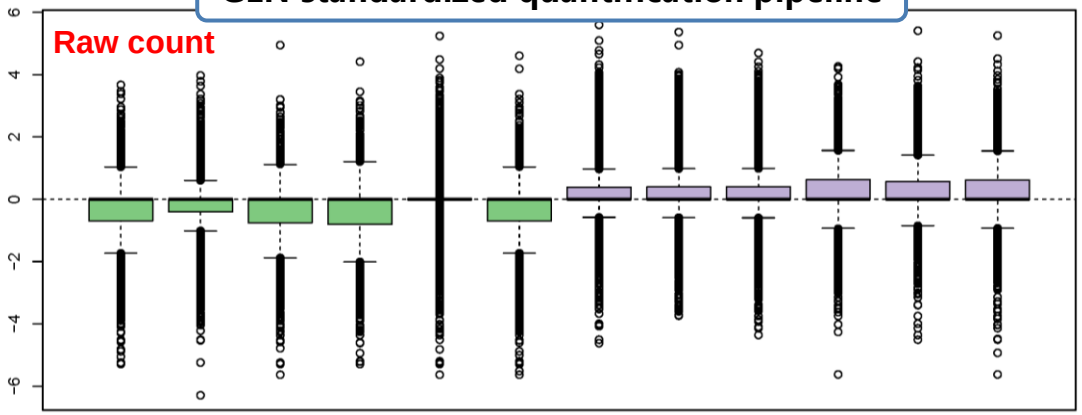
1.1.3 Alignment and quantification

- fasterq_dump (https://github.com/ncbi/sra-tools/releases/tag/2.10.9)
- Fastp v0.20.0 (https://github.com/OpenGene/fastp/releases/tag/v0.20.0)
- HISAT2 v2.0.5 (http://daehwankimlab.github.io/hisat2/)
- SAMtools v1.9 (http://github.com/samtools/)
- RseQC v2.6.4 (http://rseqc.sourceforge.net/)
- RSEM v1.3.1 (https://github.com/deweylab/RSEM/releases/tag/v1.3.1)

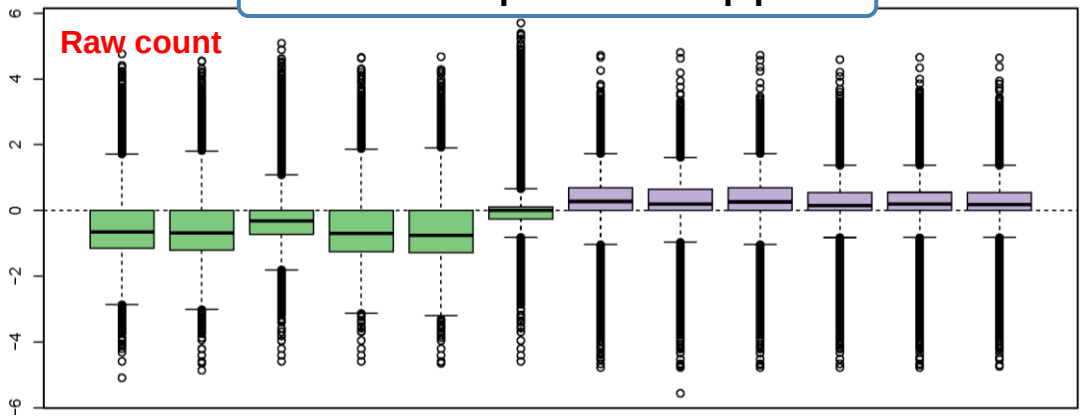
Appendix: Relative Log Expression

Xi'an Jiaotong-Liverpool University 西交利物浦大学

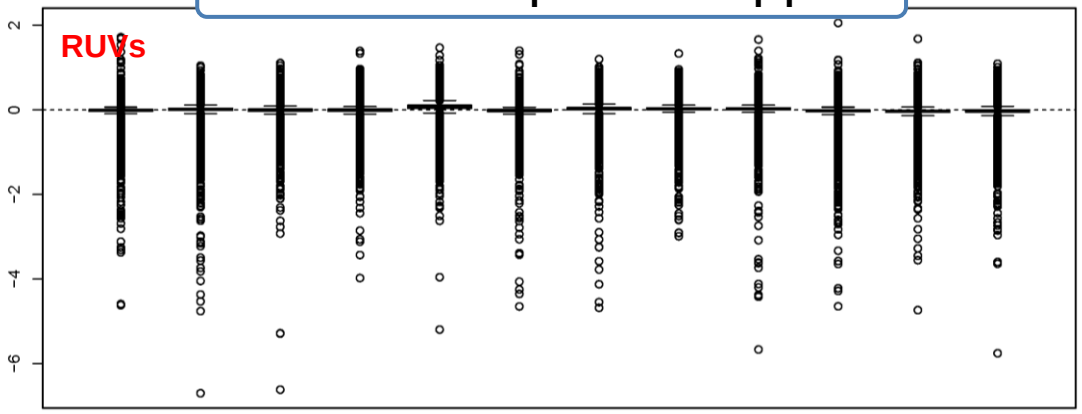
GEN standardized quantification pipeline



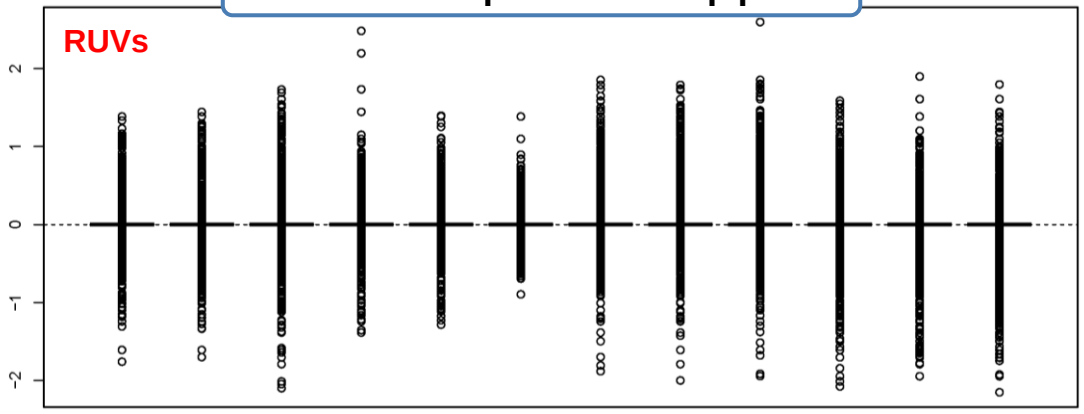
GEO discrete quantification pipeline



GEN standardized quantification pipeline



GEO discrete quantification pipeline



Appendix: DEGs table

Xi'an Jiaotong-Liverpool University 西交利物浦大学

Summary of DEGs, using the dataset corrected by RUVs from the standardized GEN quantification pipeline

	logFC	logCPM	LR	PValue	Sig
APOB	-3.26162	-1.9355	42.64943	6.55E-11	Down
CORO7-PAM16	-6.14183	-2.59922	56.19463	6.56E-14	Down
THEG	-3.04764	-3.23063	16.45275	4.99E-05	Down
HGD	-3.03061	-0.87847	66.96163	2.77E-16	Down
ATP6V1B1	-3.83562	1.328429	196.0392	1.53E-44	Down
… (97 DEGs in total)	…	…	…	…	…

	logFC	logCPM	LR	PValue	Sig
WNT16	3.359673	-2.14685	37.26895	1.03E-09	Up
NR1H4	3.415362	-3.20815	4.775572	0.028866	Up
CYP26A1	6.575954	-2.03149	34.18609	5.01E-09	Up
TNFSF18	3.081788	-1.47829	47.12096	6.67E-12	Up
TMEM35A	3.320433	-2.90279	22.43831	2.17E-06	Up
… (84 DEGs in total)	…	…	…	…	…

- *Detail information of Table 1.1: https://github.com/cupoftea0315/BIO303-project-appendices/blob/master/GEN%20database/GEN%20plots/Downstream%20analysis/GEN_down-regulate(RUVs).csv*
- *Detail information of Table 1.2: https://github.com/cupoftea0315/BIO303-project-appendices/blob/master/GEN%20database/GEN%20plots/Downstream%20analysis/GEN_up-regulate(RUVs).csv*

Summary of DEGs, using the dataset corrected by RUVs from the discrete GEO quantification pipeline

	logFC	logCPM	LR	PValue	Sig
ACAN	-3.69586	-1.98391	46.34798	9.90E-12	Down
APOB	-3.266	-1.6467	43.74034	3.75E-11	Down
AQP4	-3.37345	-1.06643	65.28147	6.49E-16	Down
CLDN19	-4.01153	-2.77436	27.36391	1.69E-07	Down
HMGCS2	-3.50794	-2.74435	31.45361	2.04E-08	Down
… (17 DEGs in total)	…	…	…	…	…

	logFC	logCPM	LR	PValue	Sig
ARC	4.647283	-0.72109	57.56098	3.28E-14	Up
CCDC178	3.224355	-1.27557	55.67377	8.56E-14	Up
CCDC65	4.698419	-2.97185	8.938764	0.002792	Up
CEND1	4.122975	-2.61784	28.68117	8.53E-08	Up
CYP26A1	6.65022	-1.71992	34.69834	3.85E-09	Up
… (22 DEGs in total)	…	…	…	…	…

- *Detail information of Table 1.3: https://github.com/cupoftea0315/BIO303-project-appendices/blob/master/GEO%20database/GEO%20plots/Downstream%20analysis/GEO_down-regulate(RUVs).csv*
- *Detail information of Table 1.4: https://github.com/cupoftea0315/BIO303-project-appendices/blob/master/GEO%20database/GEO%20plots/Downstream%20analysis/GEO_up-regulate(RUVs).csv*

Appendix: Software to correct batch effects

Xi'an Jiaotong-Liverpool University 西交利物浦大学

Name	Principle	Cite
ComBat-seq	negative binomial regression model	143
ComBat_Cor	classification，two-step	1
DESeq2	model parameter	35857
limma	model parameter	16604
RUVs	estimating the factors of unwanted variation using replicate samples	1192
RUVr	estimating the factors of unwanted variation using residuals	1192
RUVg	estimating the factors of unwanted variation using control genes	1192
svaseq	surrogate variable	230
psva	preserving biological heterogeneity with a permuted surrogate variable analysis	65
fsva (chip)	for prediction	52
GFS	fuzzy scoring	18
BatchI	predict the number of batches	15
BatchQC	ComBat，sva	40
DEBrowser	every step of differential analysis	112