Supplementary Material

Integrative Multi-Omics Framework for Causal Gene Discovery in Long COVID

Contents

1 Datasets

1.1 Genome-wide Association Studies (GWAS)

1.1.1 Description

Table 1 summarizes the Genome-wide Association Studies (GWAS) datasets available from Lammi et al., 2023 [1], which include genetic data for Long COVID cases and controls categorized into broad and strict definitions. Broad cases refer to Long COVID patients tested or untested for SARS-CoV-2, while strict cases are those with confirmed Long COVID based on SARS-CoV-2 test verification. Similarly, broad controls are from the general population, and strict controls are SARS-CoV-2-positive individuals who did not develop Long COVID.

In this study, we used only GWAS1 for the Mendelian Randomization (MR) analysis. GWAS1 pairs strictly verified Long COVID cases with broad controls, making it the most appropriate choice for this analysis. The strict case definition reduces misclassification bias, ensuring that the genetic associations identified are specific to Long COVID. Broad controls provide a larger sample size, increasing the analysis's statistical power.

The other datasets, GWAS2, GWAS3, and GWAS4, were not used as they either relaxed the definition of cases or limited the control group. For instance, GWAS2 includes broad cases, which may introduce noise into the analysis. GWAS3 and GWAS4 use strict controls, which, while specific, result in smaller sample sizes, reducing statistical power. GWAS1 was, therefore, the optimal choice for this study, as it strikes a balance between specificity in cases and a sufficient control group size to support reliable causal inference.

Table 1: **Genome-wide Association Studies (GWAS) datasets used in this study [1].** Release: 7, Ensembl: 109, Human Genome Build: GRCh38. **Broad Cases** refers to Long COVID cases that were tested and untested for SARS-CoV-2 infection. **Strict Cases** refers to Long COVID cases that were only test-verified for SARS-CoV-2 infection. **Broad Controls** are from the general population, while **Strict Controls** are SARS-CoV-2 cases that did not develop Long COVID.

Dataset	**Cases**	**Controls**	**SNPs**
GWAS 1	Strict: 3,018	Broad: 1,093,995	9,510,587
GWAS 2	Broad: 6,450	Broad: 1,093,995	9,722,678
GWAS 3	Strict: 3,018	Strict: 46,208	9,738,584
GWAS 4	Broad: 6,450	Strict: 46,208	9,753,825
TOTAL	Unique: 6,450	Unique: 1,093,995	9,722,678

Figure 1 illustrates the distribution of cases and controls across the four Long COVID GWAS datasets sourced from Lammi et al., 2023 [1]. The datasets distinguish between **Broad Cases**, which include Long COVID patients regardless of SARS-CoV-2 testing status, and **Strict Cases**, which consist only of test-verified Long COVID patients. Similarly, **Broad Controls** are drawn from the general population, while **Strict Controls** are SARS-CoV-2-positive individuals who did not develop Long COVID. For this study, GWAS1 was selected due to its use of strictly verified Long COVID cases combined with broad population controls, providing the necessary specificity and statistical power for robust analysis.

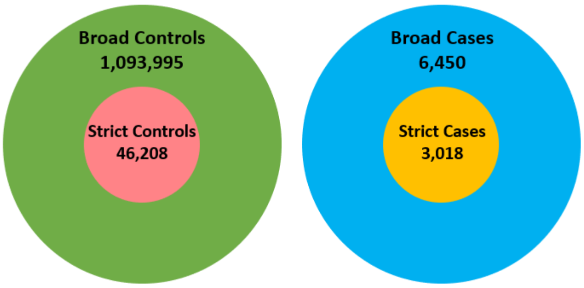


Fig. 1: **Cases and controls for the four Long COVID GWAS datasets used in the analysis and sourced from Lammi et al., 2023** [1]. **Broad Cases** refers to Long COVID cases that were both tested and untested for SARS-CoV-2 infection. **Strict Cases** refers to Long COVID cases that were only test-verified for SARS-CoV-2 infection. **Broad Controls** are from the general population, while **Strict Controls** are SARS-CoV-2 cases that did not develop Long COVID.

1.1.2 Example

Table 2 presents the top five rows from one of the original Long COVID GWAS datasets described by Lammi et al., 2023 [1]. Each row represents a genetic variant with its associated details, including the chromosome number (**Chr**), genomic position (**Position**), unique variant identifier (**Variant ID**), reference allele (**Ref All**), alternate allele (**Alt All**), log odds ratio (**logOR**), effect size estimate (**Beta**), standard error of the effect size (**SE**), and frequency of the alternate allele (**Freq**). These data highlight key attributes of the genetic variants that were used to identify potential associations with Long COVID phenotypes. The log odds ratio (**logOR**) and effect size estimate (**Beta**) provide insights into the direction and magnitude of the variant's impact. In contrast, the standard error (**SE**) reflects the variability in these estimates. The frequency of the alternate allele (**Freq**) aids in understanding the distribution of genetic variation within the population.

1.1.3 Symptoms and Ancestries

The symptoms and ancestries represented in the Long COVID GWAS1 dataset used in this study reflect the diverse clinical presentations and populations affected by Long COVID, as summarized in Table 1.1.3.

Table 2: **Top 5 rows from one of the original Long COVID GWAS datasets** [1]. It shows the chromosome number, the variant's genomic position, the genetic variant's unique identifier, the reference and alternate alleles, the log odds ratio, the effect size estimate, the standard error of the effect size, and the frequency of the alternate allele.

Chr	Position	Variant ID	Ref All	Alt All	logOR	Beta	SE	Freq
1	727242	rs61769339	G	A	0.660	−0.0891	0.0725	0.142
1	729886	rs539032812	T	C	0.204	−0.0856	0.175	0.0278
1	758351	rs12238997	A	G	0.642	−0.0838	0.0695	0.150
1	758443	rs61769351	G	C	0.462	−0.0672	0.0712	0.148
1	770988	rs12029736	A	G	0.332	0.0452	0.0619	0.491

The dataset includes a comprehensive list of symptoms commonly reported by Long COVID patients. These symptoms span various systems and manifestations, highlighting the condition's heterogeneity. Key symptoms include fatigue, shortness of breath, memory and concentration problems, anosmia, persistent cough, and insomnia. Symptoms affecting other systems, such as gastrointestinal issues (e.g., abdominal pain, nausea/vomiting, diarrhea) and musculoskeletal complaints (e.g., myalgia, arthralgia), are also represented.

Moreover, the GWAS1 dataset has individuals from six major ancestry groups: Mixed American, African, East Asian, European, Middle Eastern, and South Asian. This broad representation ensures that findings are inclusive and applicable across diverse populations. By considering multiple ancestries, the study minimizes the risk of population-specific bias and enhances the generalizability of the results. Furthermore, this diversity is crucial for understanding how genetic factors may influence Long COVID risk and symptoms differently across populations.

Symptoms:

- Abdominal pain
- Anosmia
- Arthralgia
- Chest pain
- Chills
- Confusion/Disorientation
- Depression
- Diarrhea
- Dysphagia
- Fatigue
- Fever
- Headache
- Hoarseness
- Insomnia
- Myalgia
- Nausea/Vomiting
- Numbness/Tingling
- Persistent cough
- Problems with memory/concentration
- Reduced appetite
- Rhinorrhea
- Shortness of breath
- Sore throat
- Weight loss

Ancestries:

- Admixed American
- African
- East Asian
- European
- Middle Eastern
- South Asian

1.2 Expression Quantitative Trait Loci (eQTL)

Table 3 summarizes the expression Quantitative Trait Loci (eQTL) datasets used in this study, obtained from the Genotype-Tissue Expression (GTEx) project (Version 8, Ensembl 99, GRCh38) [2]. These datasets provide a comprehensive resource for understanding the relationship between genetic variants and gene expression levels across 49 distinct human tissues. For each tissue, the table lists the number of samples, unique genes, and gene-SNP associations analyzed, reflecting the depth and breadth of the GTEx project.

The tissues include a wide range of systems, such as the nervous system (e.g., amygdala, cortex, hippocampus), cardiovascular system (e.g., aorta artery, coronary artery, left ventricle), digestive system (e.g., stomach, esophagus, colon), and others. The number of samples varies across tissues, with skeletal muscle having the largest sample size (706), while kidney cortex has the smallest (73). The diversity of tissues and sample sizes ensures robust tissue-specific gene expression regulation exploration.

These datasets are particularly valuable for identifying regulatory variants that affect gene expression tissue-dependently. This enables the integration of genetic and transcriptomic data, which is critical for finding mechanisms underlying complex traits and diseases, including Long COVID. For instance, lung, blood, and brain tissues are particularly interesting in this study due to their relevance to Long COVID symptoms, ranging from respiratory issues to neurological and systemic effects.

Table 3: **Summary of Expression Quantitative Trait Loci (eQTL) datasets**. It indicates the number of samples, unique genes, and gene-SNPs associations for 49 distinct tissues, obtained from the GTEx project (Version 8, Ensembl 99, GRCh38) [2].

Tissue	**Samples**	**Genes**	**SNPs**
Adrenal gland	233	23,820	23,264
Amygdala brain	129	24,069	23,609
Anterior cingulate cortex	147	24,342	23,843
Aorta artery	387	23,959	23,371
Atrial appendage	372	23,194	22,747
Breast mammary tissue	396	25,849	25,294
Caudate (basal ganglia)	194	24,718	24,323
Cerebellar hemisphere	175	25,144	24,404
Cerebellum	209	25,461	24,737
Coronary artery	213	24,529	24,095
Cortex	205	24,849	24,419
Cultured fibroblasts	483	22,050	21,416
EBV-transformed lymphocytes	147	22,759	22,199
Esophagus mucosa	497	23,949	23,340
Esophagus muscularis	465	23,871	23,288
Frontal cortex	175	24,676	24,265
Gastroesophageal junction	330	24,168	23,634
Hippocampus	165	24,420	24,087
Hypothalamus	170	25,096	24,649
Kidney cortex	73	24,807	24,395
Left ventricle	386	21,353	20,991
Liver	208	22,262	21,870
Lung	515	26,095	25,464
Minor salivary gland	144	25,579	25,020
Not sun-exposed skin (suprapubic)	517	25,279	24,676
Nucleus accumbens (basal ganglia)	202	24,890	24,463
Ovary	167	25,325	24,792
Pancreas	305	22,615	22,129
Pituitary	237	26,854	26,218
Prostate	221	26,529	25,969
Putamen (basal ganglia)	170	23,804	23,428
Sigmoid colon	318	24,483	23,951
Skeletal muscle	706	21,031	20,560
Small intestine terminal ileum	174	26,182	25,694
Spinal cord	126	24,669	24,167
Spleen	227	25,479	24,900
Stomach	324	24,290	23,862
Subcutaneous adipose	581	24,665	24,010
Substantia nigra	114	24,044	23,626
Sun-exposed skin (lower leg)	605	25,196	24,564
Testis	322	35,007	34,164
Thyroid	574	26,054	25,184
Tibial artery	584	23,304	22,652
Tibial nerve	532	25,873	25,092
Transverse colon	368	25,379	24,816
Uterus	129	25,188	24,637
Vagina	141	25,778	25,245
Visceral omentum adipose	469	24,724	24,167
Whole blood	670	20,315	19,701

1.3 Whole Genome Sequence (WGS) Data for Linkage Disequilibrium (LD) Analysis

Table 4 shows the top five rows from the Whole Genome Sequence (WGS) BIM file used to calculate the Linkage Disequilibrium (LD) matrix, sourced from the GTEx project (Ensembl 88, GRCh38) [3]. This file contains information about genetic variants, including the chromosome number (**Chr**), variant identifier (**Variant ID**), distance from the start of the chromosome (**Distance**), genomic position on the chromosome (**Position**), and the reference (**Ref all**) and alternate alleles (**Alt all**).

The provided rows demonstrate the structure and organization of the WGS BIM file. For example, the first row describes a variant located at position 13,526 on chromosome 1, with a reference allele of T and an alternate allele of C. This dataset forms the foundation for calculating LD matrices, which are crucial for understanding the correlation between genetic variants and their co-inheritance patterns.

Table 4: **Top 5 rows of the Whole Genome Sequence (WGS) BIM file.** This dataset was used for calculating the Linkage Disequilibrium (LD) matrix. The table provides details on the chromosome, variant ID, distance from the start of the chromosome, position on the chromosome, and reference and alternate allele. It was sourced from GTEx (Ensembl 88, GRCh38) [3].

Chr	Variant ID	Distance	Position	Ref all	Alt all
1	chr1_13526_C_T_b38	0	13526	T	C
1	chr1_13550_G_A_b38	0	13550	A	G
1	chr1_14451_CTCT_C_b38	0	14451	C	CTCT
1	chr1_14469_C_T_b38	0	14469	T	C
1	chr1_14470_G_A_b38	0	14470	A	G

Table 5 presents the top five rows from the GWS FAM file, containing metadata for the 836 European individuals used in calculating the LD matrix, sourced from GTEx (Ensembl 88, GRCh38) [3]. The table includes the family ID (**Family ID**), individual ID (**Individual ID**), paternal and maternal IDs (**Paternal ID** and **Maternal ID**), sex (**Sex**, where 1 represents male and 2 represents female), and phenotype status (**Phenotype**, with -9 indicating missing phenotype data).

This metadata ensures the accurate identification of individuals and their relationships, which is essential for LD matrix calculations. The uniform phenotype status (-9) reflects the absence of case/control definitions in this dataset, as it is primarily intended for population-level analyses.

Table 5: **Top 5 rows of the Whole Genome Sequence (GWS) FAM file**. This dataset has 836 European male and female individuals and it was used for calculating the Linkage Disequilibrium (LD) matrix, sourced from Genotype-Tissue Expression (GTEx) (Ensembl 88, GRCh38) [3]. The table provides details on the family ID, individual ID, paternal and maternal IDs, gender, and phenotype status.

Family ID	Individual ID	Paternal ID	Maternal ID	Sex	Phenotype
GTEX-1117F	GTEX-1117F	0	0	2	-9
GTEX-111CU	GTEX-111CU	0	0	1	-9
GTEX-111FC	GTEX-111FC	0	0	1	-9
GTEX-111VG	GTEX-111VG	0	0	1	-9
GTEX-111YS	GTEX-111YS	0	0	1	-9

1.4 RNA-seq Gene Expression and Clinical Data

The participants from the RNA-sequencing (RNA-seq) dataset used in this study [4], totaling 567 individuals (both males and females), have an age range from 0 to 90 years and are represented by a diverse racial background as follows :

- Black or African American
- Asian
- White
- American Indian/Alaska Native
- Native Hawaiian or Other Pacific Islander
- Individuals identifying with multiple races

Table 6 presents the top five rows and columns of the RNA-seq gene expression dataset used in this study, sourced from the Gene Expression Omnibus (GEO) - National Center for Biotechnology Information (NCBI) database (GSE215865, Ensembl GRCh37) [4]. This dataset includes gene expression measurements for 58,884 unique genes across a cohort of 567 participants, consisting of 495 acute and Long COVID patients and 72 controls. The participant's age range is between 0 to 90 years, and represents diverse racial backgrounds, ensuring the dataset reflects population heterogeneity.

The table highlights the Ensembl Gene IDs and their corresponding expression values for a subset of samples. Missing expression values (**NA**) are present for some genes, reflecting the sparsity often observed in RNA-seq data for lowly expressed or unexpressed genes in specific samples. For instance, *ENSG00000227232.5* has measurable expression across most samples, while other genes, such as *ENSG00000223972.5* and *ENSG00000243485.5*, show no detectable expression in the displayed rows.

To ensure the dataset was suitable for downstream analyses, rows and columns containing only **NA** values were deleted. For the remaining missing values, the mean of the respective gene or sample was used to impute missing expression levels, ensuring a complete dataset while minimizing potential biases.

This RNA-seq dataset forms the basis for transcriptomic analyses conducted in this study, helping identify key genes and pathways associated with Long COVID and acute COVID conditions.

Table 6: **Top 5 rows and columns of the RNA-sequencing (RNA-seq) gene expression dataset used in this paper.** This dataset consists of gene expression data from 495 acute and Long COVID patients and 72 controls, covers 58,884 unique genes, and includes data from a diverse cohort of 567 participants of varying ages (0 to 90 years) and multiple racial backgrounds. The table showcases the Ensembl Gene ID and expression values for a subset of the participants. It was sourced from the GEO - NCBI database (GSE215865, Ensembl GRCh37) [4].

Gene ID	**Subj1_Sample1**	**Subj1_Sample2**	**Subj2_Sample1**	**Subj2_Sample2**
ENSG00000223972.5	NA	NA	NA	NA
ENSG00000227232.5	NA	1.415	1.499	1.706
ENSG00000278267.1	NA	-0.669	-0.858	0.229
ENSG00000243485.5	NA	NA	NA	NA
ENSG00000284332.1	NA	NA	NA	NA

1.5 Protein-Protein Interaction (PPI)

Table 7 displays the top five rows from the Protein-Protein Interaction (PPI) dataset used to construct the Long COVID network, adapted from Vinayagam et al., 2011 [5]. This dataset provides insights into the functional relationships between genes based on their interactions within biological networks. The table includes the gene identifiers (**Gene ID**), gene names (**Gene Name**), and their classification into node subtypes (**Node Subtype**).

Node subtypes categorize genes based on their role and resilience within the network:

- **Type I genes:** Highly connected and robust against changes or disruptions, representing critical hubs within the network.
- **Type II genes:** Less connected and more susceptible to minor network disruptions, making them potential points of vulnerability in the system.

For example, *CREBBP*, classified as a Type I gene, is a key regulatory hub with significant connectivity, suggesting its role in maintaining network stability. In contrast, *HMOX2* and *GBP2*, both Type II genes, exhibit less resilience and may serve as potential targets for understanding network fragility in Long COVID pathophysiology.

This PPI dataset is integral to the construction of a robust Long COVID interaction network, providing a framework to study gene-level contributions to disease mechanisms and their potential as therapeutic targets.

Table 7: **Top 5 rows of the Protein-Protein Interaction (PPI) dataset used for building the Long COVID network.** The table showcases the genes and their associated subtypes. Type-I genes exhibit greater resilience and have more connections, being more robust against network alterations, while Type-II genes are more susceptible to even minor network disruptions. This dataset was sourced from Vinayagam et al., 2011 [5].

Gene ID	Gene Name	Node Subtype
3163	HMOX2	Type II
1387	CREBBP	Type I
2634	GBP2	Type II
5499	PPP1CA	Type I
6642	SNX1	Type II

2 Framework

2.1 Overview

Figure 2 illustrates the first part of the analytical framework employed in this study, outlining the relationship between genetic variants, gene expression, and Long COVID outcomes. The framework integrates three key components:

- **Instrumental Variables (IVs):** Genetic variants (SNPs) that serve as instruments to explore causal relationships.
- **Exposure:** Gene expression levels (eQTL) representing the intermediate step between genetic variants and the outcome.
- **Outcome:** Long COVID phenotypes derived from GWAS datasets.

The figure also highlights the presence of potential confounders that may simultaneously influence gene expression and Long COVID outcomes. This design adheres to the MR, which uses genetic variants as natural experiments to infer causal relationships, minimizing confounding and reverse causation.

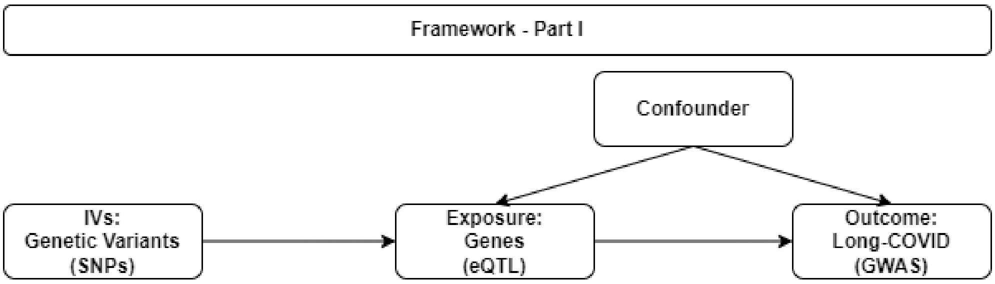
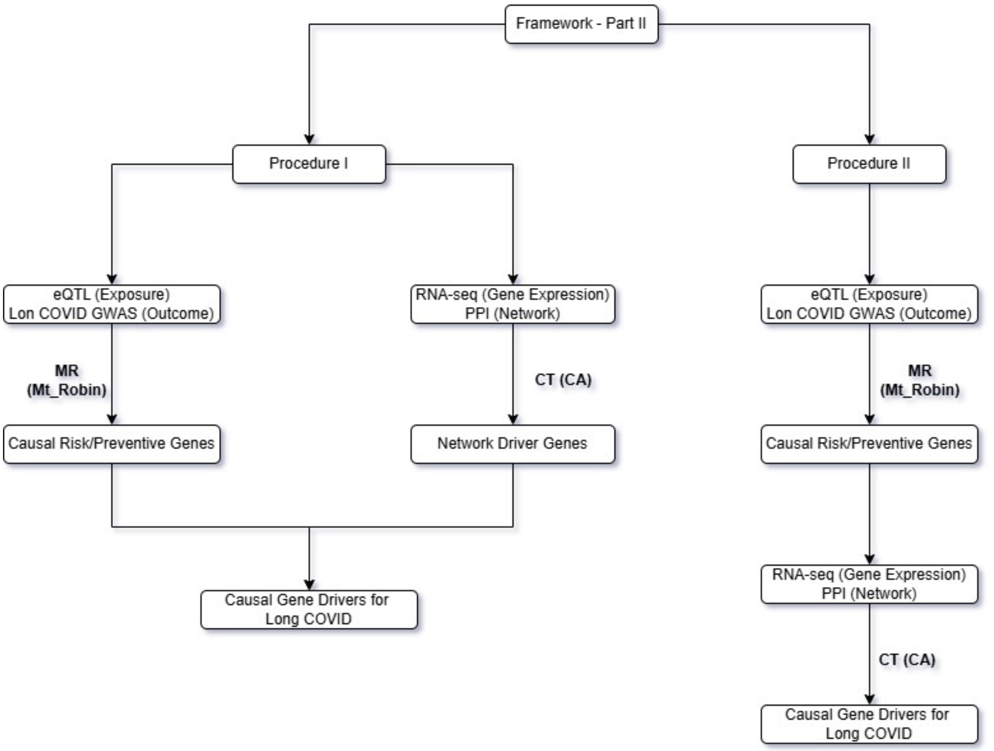


Fig. 2: **Mendelian Randomization (MR) Framework - Part I.** This framework applies MR to identify causal relationships between genetic variants (instrumental variables, IVs), gene expression (eQTL) as exposures, and Long COVID Genome-Wide Association Studies (GWAS) as outcomes. The model accounts for confounding factors to ensure robust causal inference, establishing a pathway from genetic variants through eQTL to Long COVID outcomes.

Figure 2 outlines the second part of the framework, which details two procedural approaches for identifying causal and critical genes in Long COVID.

- **Procedure I:**
  - Combines eQTL data and GWAS results to perform MR using the `Mt_Robin` tool, identifying causal genes associated with Long COVID.
  - RNA-seq gene expression data and PPI networks are analyzed using Control Theory (CT), specifically the Controllability Analysis (CA) method to identify critical genes within the Long COVID network.
  - The overlap between causal and critical genes provides a set of biologically significant targets for Long COVID.
- **Procedure II:**
  - Begins with MR to identify causal genes.
  - These genes are filtered and further analyzed using RNA-seq and PPI network data.
  - CT is applied to identify critical genes, producing a refined set of causal and critical genes for Long COVID.

This framework integrates MR and network controllability to identify and prioritize key genes comprehensively, bridging genetic causality with network-level importance in Long COVID research.



Integrative Multi-Omics Framework for Causal Gene Discovery in Long COVID - Part II. This framework combines two procedures to identify causal gene drivers for Long COVID by integrating genetic and transcriptomic data. Procedures I and II use expression Quantitative Trait Loci (eQTL) as exposures and Long COVID Genome-Wide Association Studies (GWAS) as outcomes, using Mendelian Randomization (MR) implemented via Mt-Robin to identify causal risk/preventive genes, and RNA sequencing (RNA-seq) data and Protein-Protein Interaction (PPI) networks, applying Controllability Analysis (CA), a kind of Control Theory (CT) method, to detect network driver genes. These steps collectively converge to uncover causal gene drivers for Long COVID.

2.2 Mendelian Randomization (MR): Mt-Robin

The Multi-tissue transcriptome-wide Mendelian Randomization method ROBust to INvalid instrumental variables (Mt-Robin) analysis for GWAS1 [1], and eQTL [2] represents a comprehensive approach to identifying causal relationships between genetic variants and Long COVID susceptibility. This analysis integrates GWAS data with eQTL information to detect potential causal genes.

2.2.1 Data Pre-processing

Overview

Data pre-processing for the Long COVID GWAS analysis involved multiple steps designed to ensure data quality, compatibility, and reliability. This process was essential for preparing the data for downstream MR analysis using the Mt-Robin method. The pre-processing workflow consisted of several key stages:

Data Cleaning

- Removal of missing or incomplete data entries
- Standardization of genetic variant identifiers
- Quality control of allele frequencies and effect sizes
- Verification of data consistency across files

Data Format Standardization

- Standardization of column names across datasets
- Conversion of data types to appropriate formats
- Organization of variant information in a consistent structure
- Implementation of uniform coordinate systems

Quality Control Measures

- Verification of variant ID consistency
- Cross-reference of allele frequencies
- Assessment of effect size distributions
- Evaluation of standard error measurements
- Check for duplicate entries

2.2.2 Forward Selection Analysis

Forward selection represents a stepwise approach to identifying instrumental variables while controlling for LD and expression patterns.

Instrumental Variable Selection

The selection process employs three key parameters:

$$\begin{aligned} \text{ld\_thresh} &= 0.5, \\ \text{pval\_thresh} &= 0.001, \\ \text{nTiss\_thresh} &= 1. \end{aligned} \tag{1}$$

where:

- ld_thresh: Maximum allowed linkage disequilibrium (LD) between SNPs to ensure independence.
- pval_thresh: Threshold for the p-value to select significant SNPs.
- nTiss_thresh: Minimum number of tissues where SNP expression is observed to retain significance.

These thresholds ensure a robust selection of independent genetic instruments while maintaining biological relevance.

Selection Process

The iterative selection process follows a structured approach:

1. **SNP Pool Creation:** The initial pool is formed from SNPs meeting the specified p-value threshold.
2. **LD Matrix Calculation:** Linkage Disequilibrium (LD) is computed using PLINK format files:

$$\text{LD} = \text{cor}(\text{df\_subset, use=”pairwise.complete”}) \tag{2}$$

The squared correlation coefficient is then calculated as:

$$\text{LD\_r2} = \text{LD}^2 \tag{3}$$

where:

- LD: Linkage Disequilibrium calculated as the correlation between genetic variants.
- df_subset: Subset of the data frame containing the genetic variant data.
- LD_r2: Squared correlation coefficient representing the strength of LD.

3. **Iterative Selection:** SNPs are iteratively selected based on:

- The minimum p-value in the remaining pool.
- LD threshold constraints to ensure independence.
- Tissue-specific expression patterns to prioritize biologically relevant SNPs.

This systematic approach ensures the selection of independent and biologically relevant genetic instruments.

2.2.3 Mt-Robin Setup

The Mt-Robin setup phase establishes the framework for analyzing causal relationships between genetic variants and disease outcomes.

Data Preparation

The setup process involves:

1. Loading eQTL and GWAS data
2. Matching SNPs across datasets
3. Extracting beta values and standard errors
4. Organizing tissue-specific expression data

This careful preparation ensures data compatibility and reliability for the subsequent analysis.

Mixed Model Implementation

The linear mixed model implements a sophisticated statistical framework:

$$\text{beta\_x} = \text{beta\_y} + (\text{beta\_y}|\text{snpID}) + \epsilon \tag{4}$$

where:

- beta_x: eQTL effect sizes.
- beta_y: GWAS effect sizes.
- snpID: Random effects grouping.
- ϵ: Error term.

This model structure captures both fixed and random effects while accounting for the hierarchical nature of genetic data.

2.2.4 Statistical Analysis

The statistical analysis phase ensures robust identification of significant associations while controlling for multiple testing.

P-value Calculation

P-values are computed through a rigorous process:

1. Bootstrap resampling (10,000 iterations)
2. Null distribution generation
3. Two-sided test comparison

This approach provides reliable statistical inference while accounting for the complex structure of genetic data.

Multiple Testing Correction

Multiple testing correction is implemented using the Benjamini-Hochberg procedure:

$$\text{FDR} = \text{p.adjust(p\_values, method = "BH")} \tag{5}$$

where:

- FDR: False Discovery Rate, representing the expected proportion of false positives among the declared significant results.
- p.adjust: Function used to adjust p-values for multiple testing.
- p_values: The original p-values from statistical tests.
- method = ”BH”: Benjamini-Hochberg method, a procedure for controlling the False Discovery Rate.

This correction controls the false discovery rate while maintaining reasonable statistical power.

2.2.5 Output Processing

The final results compilation includes comprehensive information for downstream analysis:

- Gene identifiers: Unique identifiers for each analyzed gene
- Effect sizes: Quantification of genetic effects
- Standard errors: Measures of estimation precision
- P-values and FDR: Statistical significance measures
- Number of instrumental variables: Count of independent genetic instruments
- Tissue-specific information: Expression patterns across tissues

This rich output facilitates detailed interpretation and follow-up analyses of identified associations.

2.2.6 Conclusion

The Mt-Robin analysis for GWAS1 and eQTL represents a comprehensive approach to identifying causal genetic factors in Long COVID susceptibility. Through careful data processing, robust statistical analysis, and stringent quality control, this analysis provides valuable insights into the genetic architecture of Long COVID while maintaining high statistical rigor and biological relevance. The complete code for this analysis is available in Notebook 1.

2.3 Control Theory (CT): Controllability Analysis (CA)

2.3.1 Clinical Data Processing

We generated four distinct subsets from the main clinical dataset [4]:

- PASC case subset
- Positive Acute COVID control subset
- Negative Acute COVID control subset
- Unknown Acute COVID control subset

2.3.2 Gene Expression Processing

Missing Value Treatment

We processed the gene expression matrix by:

- Removing columns containing only NA values
- Eliminating rows where all values except the first column were NA

Gene Identifier Processing

The gene identifiers were processed through the following sequential steps:

1. Removal of version numbers from Ensembl Gene IDs
2. Mapping to external gene names via biomaRt
3. Addition of gene type information
4. Filtering to retain only protein-coding genes

2.3.3 Control Theory Implementation

Procedure 1: Direct Control Theory Application

1. Applied Control Theory directly to PASC case data
2. Overlapped results with GWAS1 findings

Procedure 2: Overlap-First Approach

1. Overlapped PASC case data with GWAS1 results
2. Applied Control Theory to the overlapped dataset

2.3.4 Network Controllability Analysis

Type I Node Classification

Nodes were classified based on their impact on driver nodes (N_D):

$$\text{Node Type I} = \begin{cases} \text{Critical} & \text{if } N_D \text{ increases in absence} \\ \text{Redundant} & \text{if } N_D \text{ decreases in absence} \\ \text{Ordinary} & \text{if } N_D \text{ remains unchanged} \end{cases}$$

Type II Node Classification

Nodes were classified based on their presence in driver node sets:

$$\text{Node Type II} = \begin{cases} \text{Critical} & \text{if in all driver node sets} \\ \text{Redundant} & \text{if in no driver node sets} \\ \text{Ordinary} & \text{if in some driver node sets} \end{cases}$$

2.3.5 Network Metrics

Topological Measures

We calculated the following network metrics:

- Total number of nodes (N)
- Total number of edges (E)
- Average degree ($\langle k \rangle$)
- Number of driver nodes (N_d)

2.3.6 Mt-Robin Result Integration

Integration Analysis

We performed an analysis with the results from the Mt-Robin results to identify:

- Driver nodes overlapping
- Network properties implicated genes
- Critical control points

2.3.7 Conclusion

The control theory analysis framework presented here combines sophisticated network analysis with biological data integration to identify critical control points in Long COVID gene networks. Through systematic data processing, implementation of two complementary procedures, and comprehensive network metric calculations, this approach provides a robust methodology for understanding network controllability in the context of Long COVID pathogenesis. By integrating Mt-Robin results with network controllability analysis, we established a bridge between genetic association signals and their functional implications in network control. The complete analysis pipeline is available in Notebook 2, ensuring reproducibility and transparency of our methodological approach.

2.4 Integration

2.4.1 MR Score

MR analysis results were processed to include all significant causal genes with their respective beta values, p-values, and FDR scores. A normalized MR score (MR_Score_{norm}) was calculated as:

$$MR\_Score_{norm} = \frac{MR\_Score - \min(MR\_Score)}{\max(MR\_Score) - \min(MR\_Score)} \tag{6}$$

where:

- MR_Score: Raw Mendelian Randomization score.
- $\min(MR\_Score)$: Minimum value of the MR scores in the dataset.
- $\max(MR\_Score)$: Maximum value of the MR scores in the dataset.

Genes were classified into three categories based on their statistical significance and effect direction:

- **Risk genes:** Genes with a positive beta value, indicating an increased likelihood of the associated phenotype.
- **Preventive genes:** Genes with a negative beta value, suggesting a protective effect against the associated phenotype.
- **Non-significant genes:** Genes with a p-value greater than 0.05 or an FDR greater than 0.05, indicating no statistically significant association.

2.4.2 CT Score

CT analysis classified nodes based on two types:

- **Type I:** Critical (0), Redundant (1), or Ordinary (2)
- **Type II:** Critical (0), Redundant (1), or Ordinary (2)

Weight Calculation:

$$\text{Weight} = \begin{cases} 2 & \text{if TypeI} = 0 \text{ and TypeII} \neq 0, \\ 1 & \text{if TypeI} \neq 0 \text{ and TypeII} = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

where:

- TypeI: Classification of nodes based on their impact on driver nodes (e.g., critical, redundant, or ordinary).
- TypeII: Classification of nodes based on their presence in driver node sets (e.g., critical, redundant, or ordinary).

CT Score Computation:

$$CT\_Score = K \times \text{Weight} \tag{8}$$

where:

- CT_Score: Computed score representing the contribution of a node to network controllability.
- K: A constant or scaling factor determined by the analysis context.
- Weight: A classification-based multiplier determined by the node's type (e.g., critical, redundant, or ordinary).

CT Score Normalization:

$$CT\_Score_{norm} = \frac{CT\_Score - \min(CT\_Score)}{\max(CT\_Score) - \min(CT\_Score)} \tag{9}$$

where:

- CT_Score_norm: Normalized CT score, scaled to a range between 0 and 1.
- CT_Score: Raw CT score calculated for a node.
- min(CT_Score): Minimum CT score observed across all nodes in the analysis.
- max(CT_Score): Maximum CT score observed across all nodes in the analysis.

2.4.3 Final Score

The final score for each gene was computed as:

$$Final\_Score = \alpha \times MR\_Score_{norm} + (1 - \alpha) \times CT\_Score_{norm} \tag{10}$$

where:

- Final_Score: Combined score representing the integration of MR and CT analyses.
- α: Weighting parameter ($0 \leq \alpha \leq 1$) determining the relative contribution of MR and CT scores.

- MR_Score_norm: Normalized Mendelian Randomization score, scaled between 0 and 1.
- CT_Score_norm: Normalized Control Theory score, scaled between 0 and 1.

2.4.4 Alpha Value Exploration

Five different α values were evaluated:

- $\alpha = 1.00$: Pure MR score
- $\alpha = 0.75$: MR-weighted combination
- $\alpha = 0.50$: Balanced combination
- $\alpha = 0.25$: CT-weighted combination
- $\alpha = 0.00$: Pure CT score

2.4.5 Result Generation

Final output files included:

- Gene ranking based on Final Score
- Gene effect classification (Risk/Preventive/Non-significant)
- Critical gene classification (Type I/Type II/Not Critical)
- Normalized scores (MR and CT)

2.4.6 Conclusion

This scoring system provides a flexible framework for integrating MR and CT. The complete implementation code is available in Notebook 6.

2.5 Enrichment Analysis (EA)

2.5.1 Analysis Definition and Purpose

Enrichment Analysis (EA) was applied to determine whether gene sets are enriched for specific biological attributes compared to chance. The analysis follows a systematic process:

- Gene Set Selection: Disease-associated genes
- Background Set Definition: Complete genome reference
- Statistical Testing: Hypergeometric and Fisher's exact tests
- P-value Correction: Multiple testing adjustment
- Result Interpretation: Enriched annotation analysis

2.5.2 Analysis Implementation

We applied the following steps:

- Enrichment analysis
- Genome annotation
- Pathway analysis
- Visualization

2.5.3 Enrichment Categories

We ran an analysis across three ontologies:

- Biological Process (BP)
- Molecular Function (MF)
- Cellular Component (CC)

And three different pathways:

- KEGG pathways
- Reactome pathways
- WikiPathways

2.5.4 Core Gene Analysis

Processing of 21 core genes:

- GO term enrichment calculation
- KEGG pathway mapping
- Reactome pathway analysis

2.5.5 Extended Analysis

Analysis of 32 core genes:

- Comprehensive pathway mapping
- Cross-database enrichment
- Functional annotation clustering

2.5.6 Statistical Processing

Statistical settings include:

- p-value adjustment: Benjamini-Hochberg
- q-value threshold: 0.05
- Alpha threshold: 0.5 for sensitivity analysis

2.5.7 Conclusion

The EA provided comprehensive pathway insights for Long COVID genetic factors. Complete implementation code is available in Notebook 3.

2.6 Gene Expression Clustering

The gene expression data was processed through the following steps:

1. Subsetting data for causal genes
2. Removing metadata columns
3. Cleaning column names
4. Transposing data matrix
5. Aggregating by Subject-ID using median values

The analysis explored multiple parameter combinations:

- **Cluster Number (clusterNum $\in [2, 5]$):** This parameter defines the range of cluster numbers to evaluate during clustering analysis. The values between 2 and 5 specify the minimum and maximum number of clusters to test, helping identify the optimal number of clusters that best represent the data.
- **Maximum Number of Clusters (maxK $\in [5, 7]$):** The maximum number of clusters to consider when performing clustering algorithms. Values in this range allow flexibility while ensuring computational efficiency by not testing excessively high numbers of clusters.
- **Proportion of Items (pItem $\in [0.5, 0.9]$):** This parameter represents the proportion of items (samples) to be resampled during each iteration of the clustering process. Resampling between 50% and 90% of the data ensures robust cluster identification while maintaining computational efficiency.
- **Proportion of Features (pFeature $\in [0.5, 1.0]$):** Indicates the proportion of features (variables or genes) to include in each iteration of clustering. Testing values between 50% and 100% ensures that enough features are considered to identify meaningful patterns while avoiding overfitting.

Two clustering algorithms were implemented:

1. **Hierarchical clustering (hc):** This clustering method builds a hierarchy of clusters by iteratively merging or splitting them based on their similarity. In this study, hierarchical clustering helps visualize relationships between data points by creating a dendrogram, where closer branches represent higher similarity.
2. **Partitioning Around Medoids (pam):** PAM is a robust clustering method that partitions the data into a pre-specified number of clusters by selecting representative data points (medoids) as cluster centers. Unlike k-means, PAM minimizes the dissimilarity between points and their assigned medoids, making it less sensitive to noise and outliers.

The **ExecuteCC** function was implemented with error handling and data validation:

1. Data preprocessing and scaling
2. Consensus clustering execution
3. Distance matrix computation
4. Result extraction and validation

2.6.1 Clinical Data Analysis

Clinical data was processed to analyze symptom patterns:

1. Preprocess symptoms function implementation
2. Symptom frequency calculation
3. Statistical testing for significance

Statistical tests were performed:

- χ^2 **Test for Large Samples:** The chi-squared test is a statistical method used to assess whether there is a significant association between two categorical variables in large sample sizes. It compares the observed frequencies in a contingency table to the expected frequencies under the assumption of independence. This test is suitable when all expected cell frequencies are greater than or equal to 5, ensuring the validity of the test results.
- **Fisher's Exact Test for Small Samples ($n < 5$):** Fisher's Exact Test is a precise statistical method used for small sample sizes, specifically when the expected frequencies in a contingency table are less than 5. Unlike the chi-squared test, Fisher's Exact Test does not rely on large-sample approximations and calculates the exact probability of the observed data under the null hypothesis, making it ideal for sparse or small datasets.

Two types of heatmaps were generated:

1. Hierarchical clustering (hc)
2. Partitioning Around Medoids (pam)

2.6.2 Output Generation

Multiple output files were created:

1. Cluster assignments
2. Statistical test results
3. Frequency tables
4. Visualization outputs

2.6.3 Conclusion

The implemented analysis pipeline represents a comprehensive approach to analyzing Long COVID gene expression data. The methodology follows a systematic workflow:

1. Initial data preparation and quality control
2. Robust consensus clustering with parameter optimization
3. Clinical correlation analysis
4. Statistical validation
5. Visualization of results

Implementing both hierarchical and PAM clustering algorithms provided complementary approaches to pattern discovery. Integrating clinical data with gene expression profiles enabled a multi-dimensional understanding of Long COVID manifestations.

References

[1] Vilma, L. *et al.* Genome-wide Association Study of Long COVID. *medRxiv* (2023).

[2] GTEx portal - datasets (2023). URL https://gtexportal.org/home/datasets. Accessed 8 Sep 2023.

[3] GTEx portal - protected data access (2023). URL https://gtexportal.org/home/protectedDataAccess. Accessed: 09/08/2023.

[4] NCBI GEO - GSE215865 (2023). URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE215865. Accessed 11 Feb 2023.

[5] Vinayagam, A. *et al.* A directed protein interaction network for investigating intracellular signal transduction. *Sci Signal* **4** (2011).