Big Data and Artificial Intelligence Modeling for Drug Discovery

2024 International Academic Exchange Conference

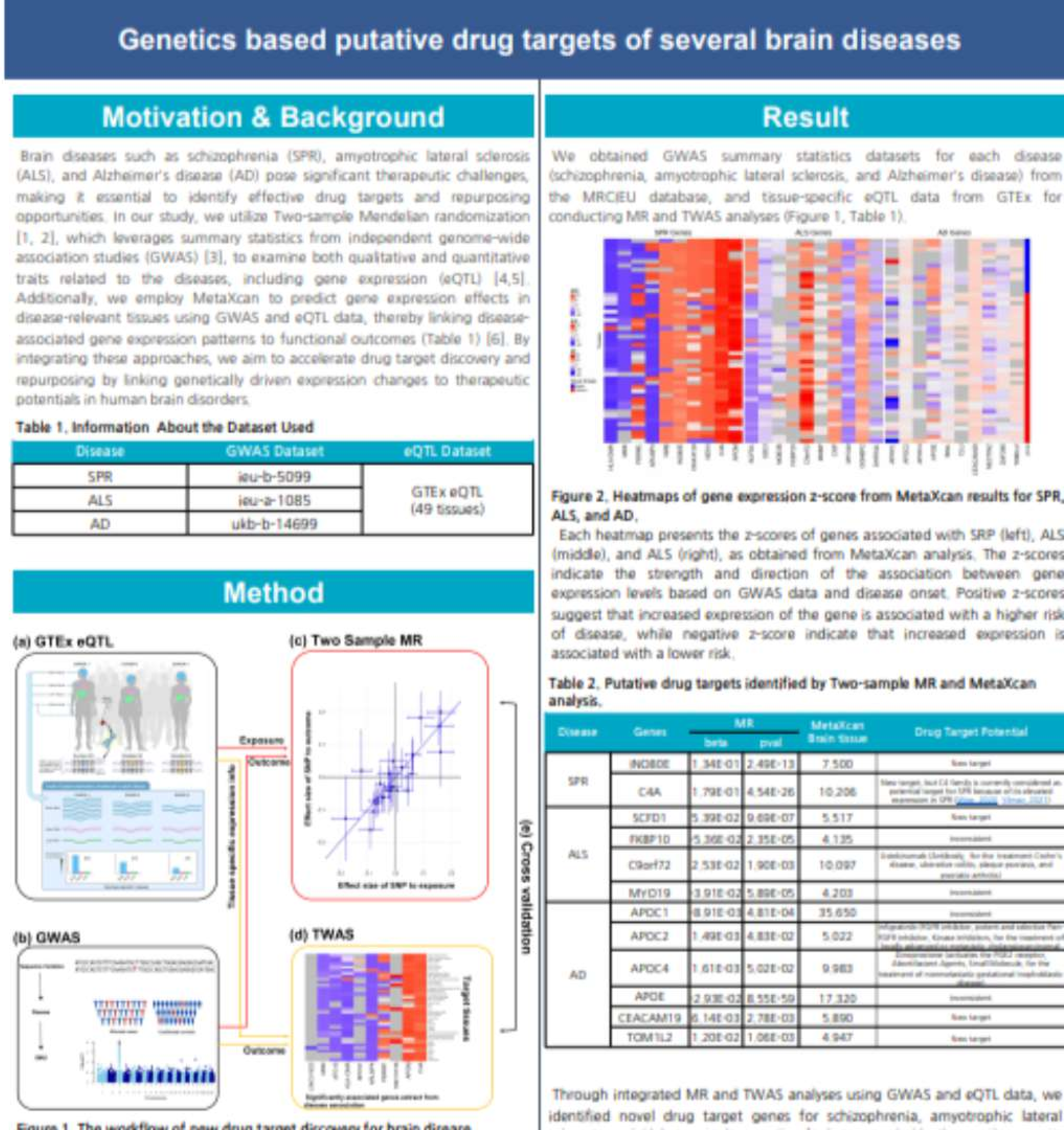
KOREA UNIVERSITY SEJONG CAMPUS   TOKAI UNIVERSITY

2024.11.16

Kim SangJin
Department of Big Data Science, Korea University Graduate School

Index

1. Project
2. Method 1: MR
   Virtual Environment Set up
   Files and Data
3. Method 2: MetaXcan
   AWS EC2
4. Results



Genetics based putative drug targets of several brain diseases

Motivation & Background

Brain diseases such as schizophrenia (SPR), amyotrophic lateral sclerosis (ALS), and Alzheimer's disease (AD) pose significant therapeutic challenges, making it essential to identify effective drug targets and repurposing opportunities. In our study, we utilize Two-sample Mendelian randomization [1, 2], which leverages summary statistics from independent genome-wide association studies (GWAS) [3], to examine both qualitative and quantitative traits related to the diseases, including gene expression (eQTL) [4,5]. Additionally, we employ MetaXcan to predict gene expression effects in disease-relevant tissues using GWAS and eQTL data, thereby linking disease-associated gene expression patterns to functional outcomes (Table 1) [6]. By integrating these approaches, we aim to accelerate drug target discovery and repurposing by linking genetically driven expression changes to therapeutic potentials in human brain disorders.

Table 1. Information About the Dataset Used

Disease	GWAS Dataset	eQTL Dataset
SPR	ieu-b-5099	GTEx eQTL (49 tissues)
ALS	ieu-a-1085	
AD	ukb-b-14699	

Method

(a) GTEx eQTL
(b) GWAS
(c) Two Sample MR
(d) TWAS
(e) Cross validation
Exposure
Outcome
Outcome
Tissue specific expression info
Effect size of SNP to outcome
Effect size of SNP to exposure
Target tissues
Significantly associated genes extract from disease association

Figure 1. The workflow of new drug target discovery for brain disease

Result

We obtained GWAS summary statistics datasets for each disease (schizophrenia, amyotrophic lateral sclerosis, and Alzheimer's disease) from the MRCIEU database, and tissue-specific eQTL data from GTEx for conducting MR and TWAS analyses (Figure 1, Table 1).

Figure 2. Heatmaps of gene expression z-score from MetaXcan results for SPR, ALS, and AD.

Each heatmap presents the z-scores of genes associated with SRP (left), ALS (middle), and ALS (right), as obtained from MetaXcan analysis. The z-scores indicate the strength and direction of the association between gene expression levels based on GWAS data and disease onset. Positive z-scores suggest that increased expression of the gene is associated with a higher risk of disease, while negative z-score indicate that increased expression is associated with a lower risk.

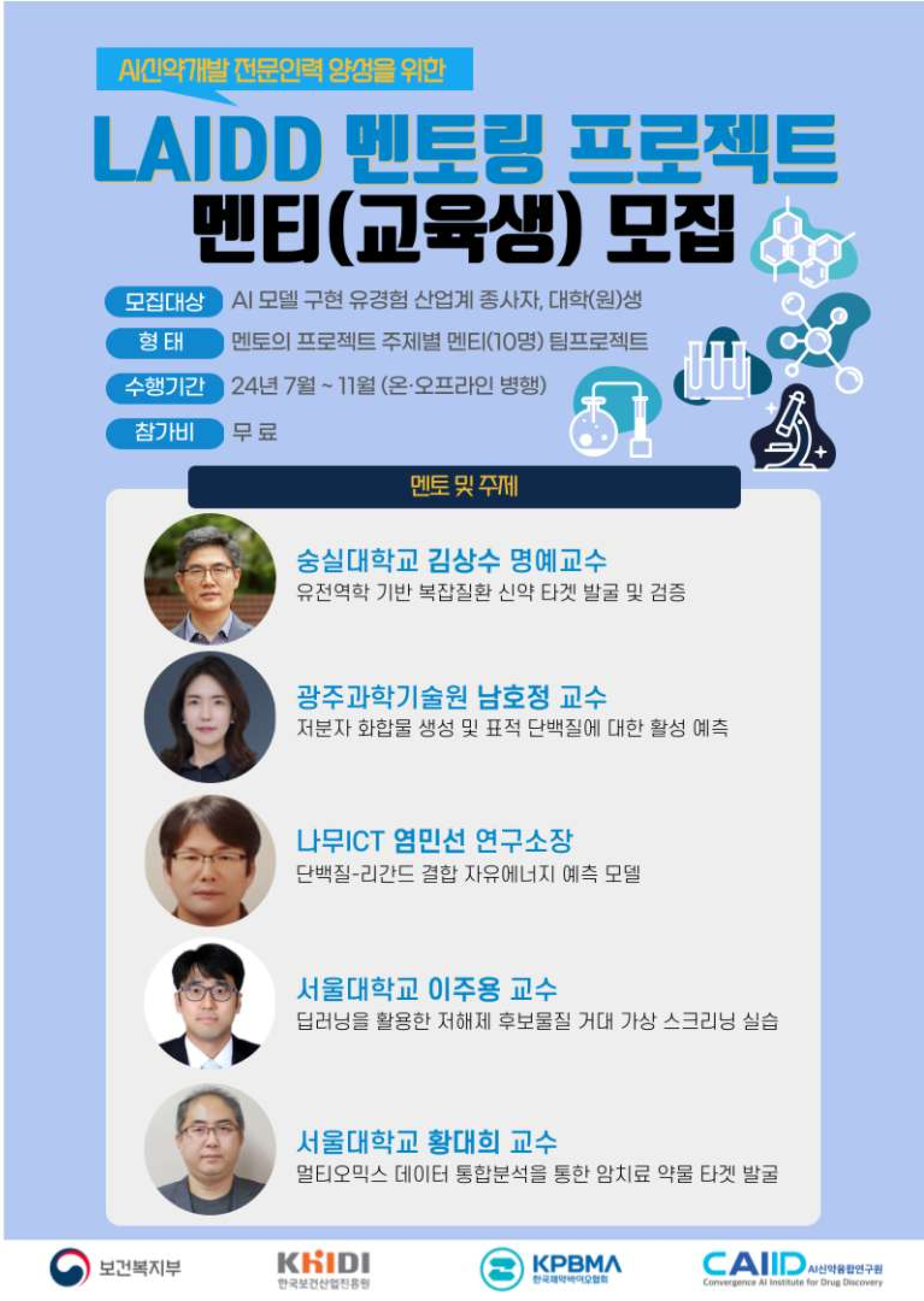
Table 2. Putative drug targets identified by Two-sample MR and MetaXcan analysis.

Disease	Genes	MR		MetaXcan Brain tissue	Drug Target Potential
		beta	pval		
SPR	INO80E	1.34E-01	2.49E-13	7.500	New target
	C4A	1.79E-01	4.54E-26	10.206	New target, but C4 family is currently considered as potential target for SPR because of its elevated expression in SPR ([illegible] [illegible])
ALS	SCFD1	5.39E-02	9.69E-07	5.517	New target
	FKBP10	5.36E-02	2.35E-05	4.135	Inconsistent
	C9orf72	2.53E-02	1.90E-03	10.097	Ustekinumab (Antibody, for the treatment Crohn's disease, ulcerative colitis, plaque psoriasis, and psoriatic arthritis)
	MYD19	3.91E-02	5.89E-05	4.203	Inconsistent
AD	APOC1	8.91E-03	4.81E-04	35.650	Inconsistent
	APOC2	1.49E-03	4.83E-02	5.022	Infigratinib (FGFR inhibitor, potent and selective Pan-FGFR inhibitor, Kinase Inhibitors, for the treatment of locally advanced or metastatic [illegible])
	APOC4	1.61E-03	5.02E-02	9.983	Dinoprostone (activates the PGE2 receptor, Abortifacient Agents, SmallMolecule, for the treatment of nonmetastatic gestational trophoblastic disease)
	APOE	2.93E-02	8.55E-59	17.320	Inconsistent
	CEACAM19	6.14E-03	2.78E-03	5.890	New target
	TOM1L2	1.20E-02	1.06E-03	4.947	New target

Through integrated MR and TWAS analyses using GWAS and eQTL data, we identified novel drug target genes for schizophrenia, amyotrophic lateral

1. Project

Discovery and Validation of Drug Targets for Complex Diseases Based on Genetic Epidemiology



AI신약개발 전문인력 양성을 위한

LAIDD 멘토링 프로젝트
멘티(교육생) 모집

모집대상	AI 모델 구현 유경험 산업계 종사자, 대학(원)생
형 태	멘토의 프로젝트 주제별 멘티(10명) 팀프로젝트
수행기간	24년 7월 ~ 11월 (온·오프라인 병행)
참가비	무 료

멘토 및 주제

숭실대학교 김상수 명예교수
유전역학 기반 복잡질환 신약 타겟 발굴 및 검증

광주과학기술원 남호정 교수
저분자 화합물 생성 및 표적 단백질에 대한 활성 예측

나무ICT 염민선 연구소장
단백질-리간드 결합 자유에너지 예측 모델

서울대학교 이주용 교수
딥러닝을 활용한 저해제 후보물질 거대 가상 스크리닝 실습

서울대학교 황대희 교수
멀티오믹스 데이터 통합분석을 통한 암치료 약물 타겟 발굴

보건복지부 KHIDI 한국보건산업진흥원 KPBMA 한국제약바이오협회 CAIID AI신약융합연구원 Convergence AI Institute for Drug Discovery

1. Project Framework

Data Collection

Exposure

Index of /downloads/eqtlgen/cis-eqtl/

```
../
[illegible]
```

Outcome

GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

ieu open gwas project

Data preprocessing

Input data preparing

R

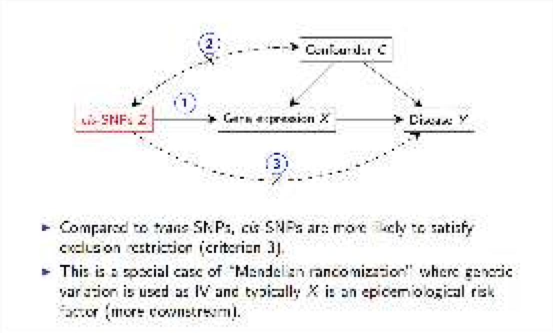
GitHub

Split the gene into 50 chunk

BASH
THE BOURNE-AGAIN SHELL

Running Methods

MR

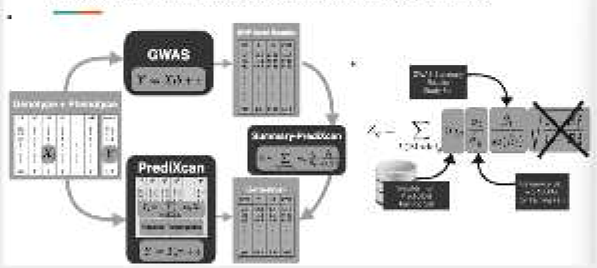


- Compared to trans-SNPs, cis-SNPs are more likely to satisfy exclusion restriction (criterion 3).
- This is a special case of "Mendelian randomization" where genetic variation is used as IV and typically X is an epidemiological risk factor (more downstream).

TWAS

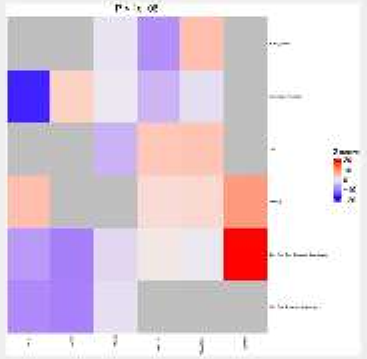
Barbeira, A.N., Dickinson, S.P., Bonazzola, R. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 9, 1825 (2018).

Transcriptome-wide Association Study (TWAS)



Results validation

Visualization

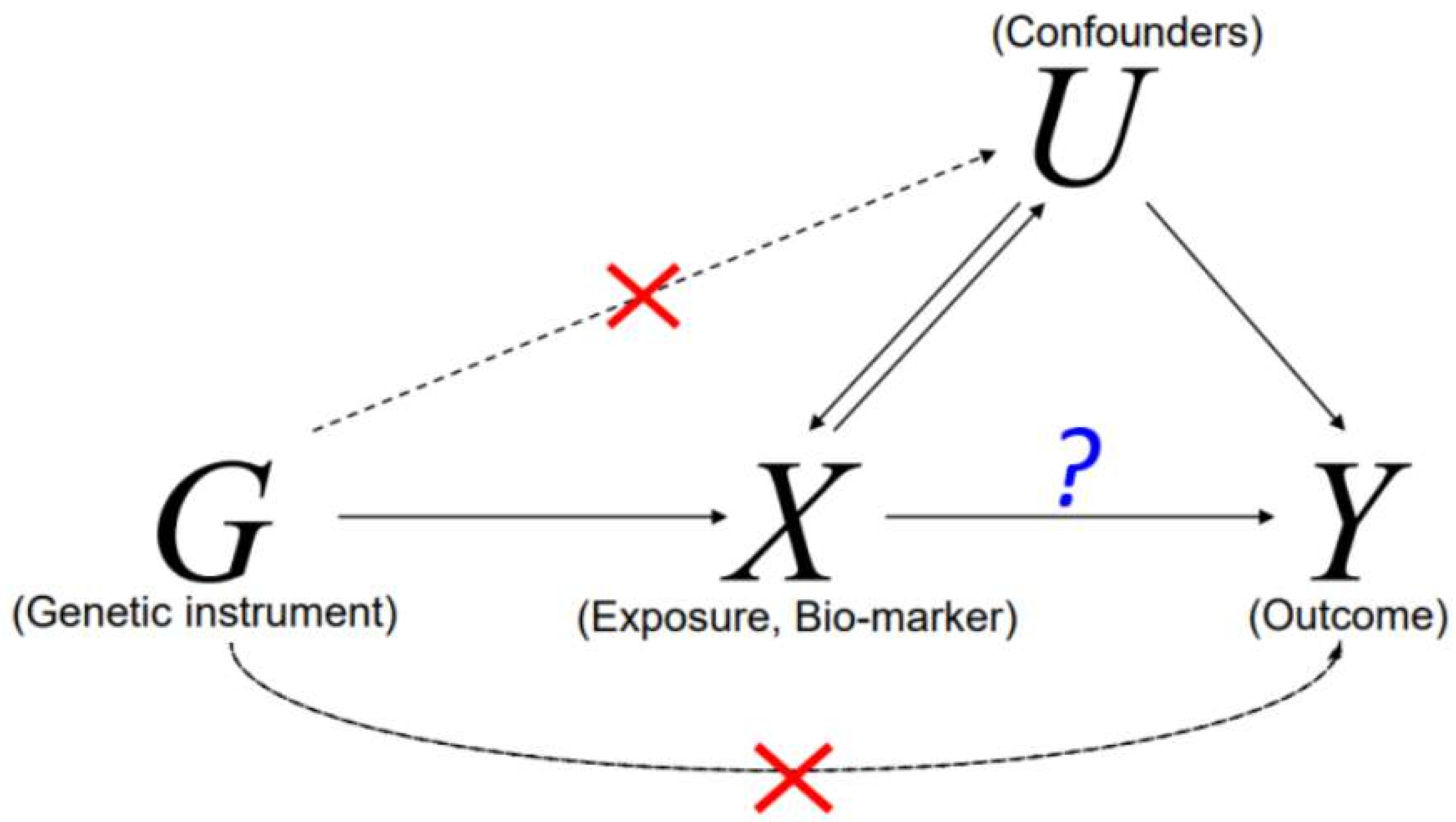


Searching DB

GeneCards®
THE HUMAN GENE DATABASE

DRUGBANK Online

2. Method: Mendelian Randomization



Mendelian Randomization(MR) is an epidemiological research method that uses genetic variants to explain how modifiable exposures causally affect various outcomes. It is based on Mendel's laws of inheritance and utilizes instrumental variable (IV) estimation methods to reduce confounding and infer causal effects.

Virtual Environment Set up

System Langauge: Linux

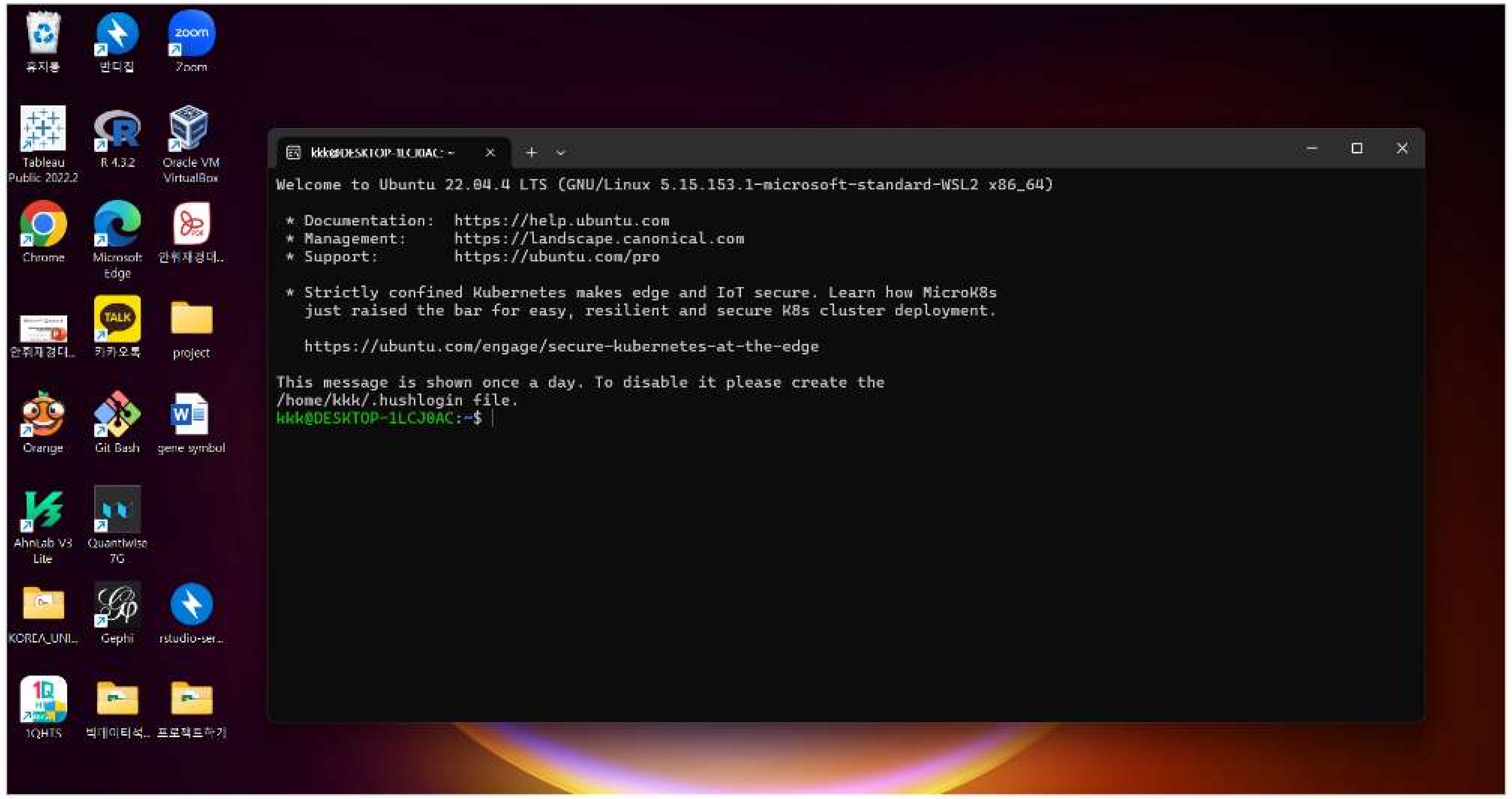
Many development tools and packages (e.g., Git, Python, Node.js) are better supported or easier to install on Linux. Using WSL allows for easy access to these tools.

We can use Linux through WSL (Windows Subsystem for Linux).

Virtual Environment Set up

Using WSL (Windows Subsystem for Linux), you can install several Linux distributions at MS store, one of which is Ubuntu. Ubuntu is one of the most popular Linux distributions, known for being user-friendly and easy to install and use.

Virtual Environment Set up



```
Welcome to Ubuntu 22.04.4 LTS (GNU/Linux 5.15.153.1-microsoft-standard-WSL2 x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/pro

 * Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
   just raised the bar for easy, resilient and secure K8s cluster deployment.

   https://ubuntu.com/engage/secure-kubernetes-at-the-edge

This message is shown once a day. To disable it please create the
/home/kkk/.hushlogin file.
kkk@DESKTOP-1LCJ0AC:~$
```

Ubuntu terminal and shell

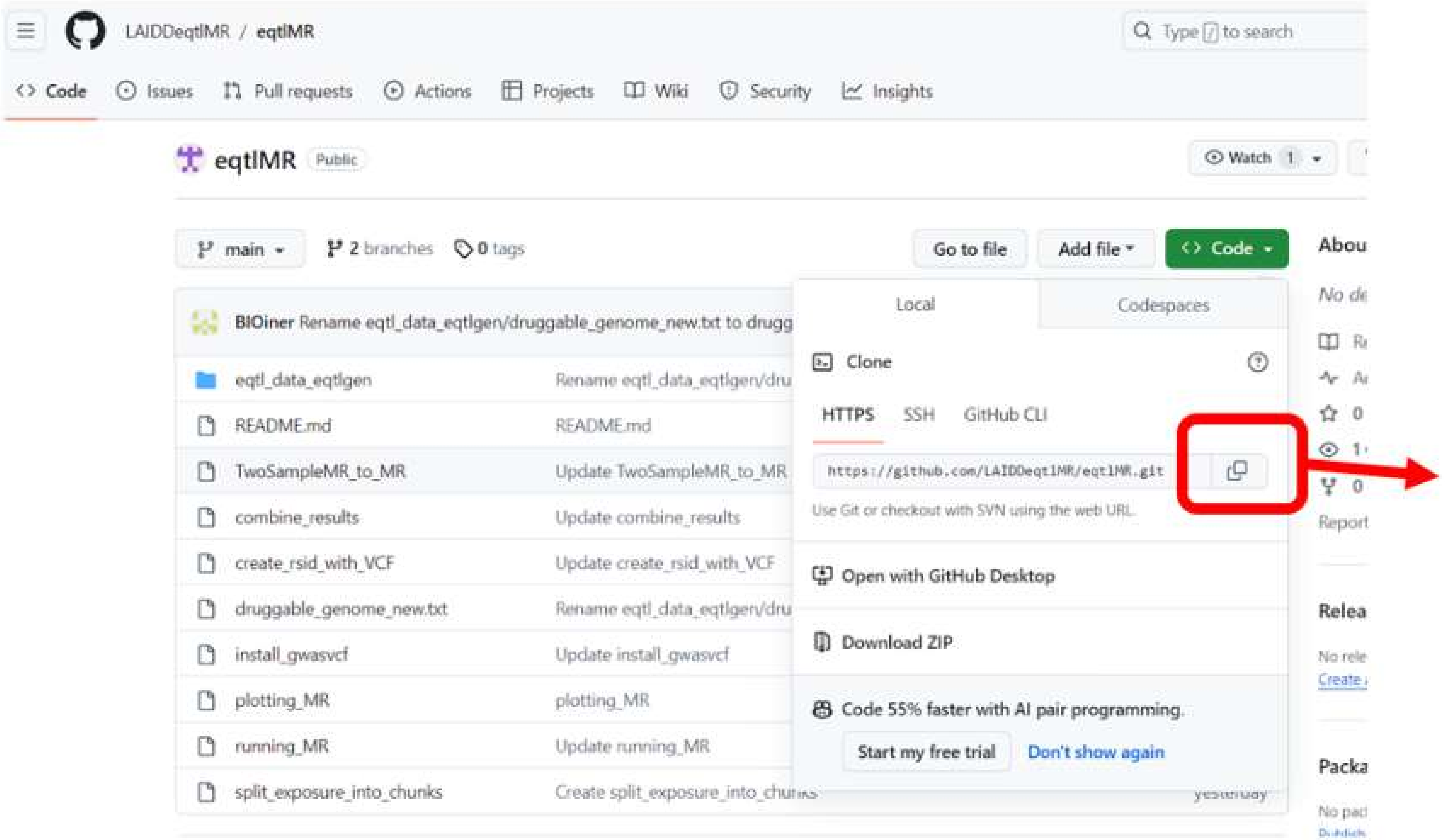
When a command is entered through the CLI(command line interface) in the terminal, the shell interprets it and passes it to the operating system for execution. The most widely used shell scripting language is Bash.

Virtual Environment Set up

Programming Language: R

Install R Studio on Ubuntu and install the necessary libraries for the project.

Virtual Environment Set up



BIOiner Rename eqtl_data_eqtlgen/druggable_genome_new.txt to drugg

eqtl_data_eqtlgen	Rename eqtl_data_eqtlgen/dru
README.md	README.md
TwoSampleMR_to_MR	Update TwoSampleMR_to_MR
combine_results	Update combine_results
create_rsid_with_VCF	Update create_rsid_with_VCF
druggable_genome_new.txt	Rename eqtl_data_eqtlgen/dru
install_gwasvcf	Update install_gwasvcf
plotting_MR	plotting_MR
running_MR	Update running_MR
split_exposure_into_chunks	Create split_exposure_into_chunks

```
https://github.com/LAIDDeqtlMR/eqtlMR.git
```

Use Git or checkout with SVN using the web URL.

Code 55% faster with AI pair programming.

yesterday

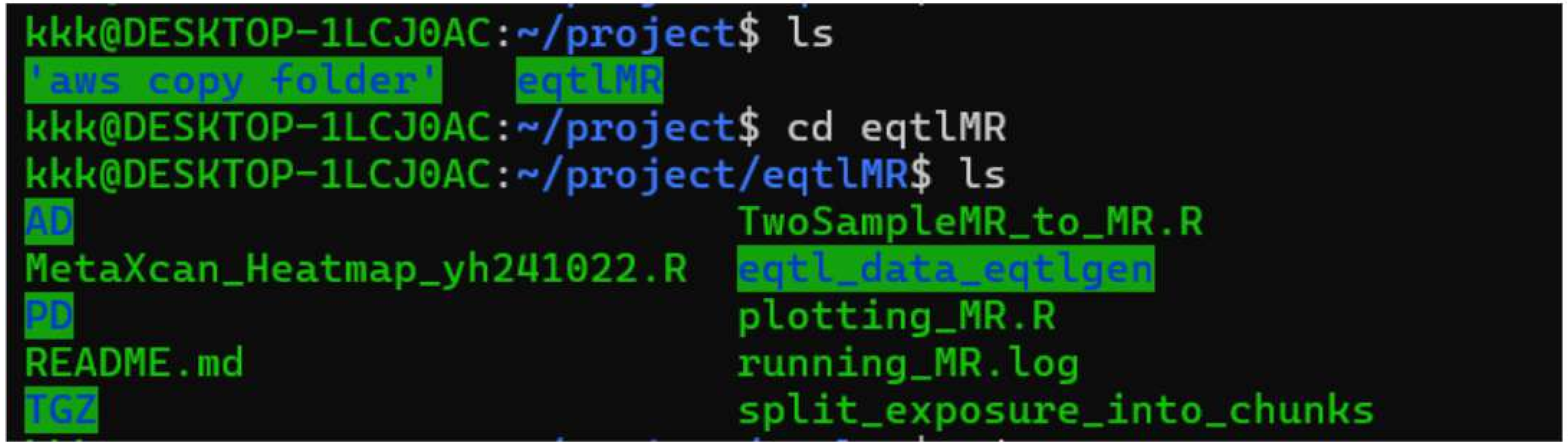
Copy link
https://github.com/LAIDDeqtlMR/eqtlMR.gi
t

https://github.com/LAIDDeqtlMR/eqtlMR

File Scripts from GitHub

Download the file for preprocessing and analysis from GitHub and modify it later as needed.

Files and Data



```
kkk@DESKTOP-1LCJ0AC:~/project$ ls
'aws copy folder'   eqtlMR
kkk@DESKTOP-1LCJ0AC:~/project$ cd eqtlMR
kkk@DESKTOP-1LCJ0AC:~/project/eqtlMR$ ls
AD                             TwoSampleMR_to_MR.R
MetaXcan_Heatmap_yh241022.R    eqtl_data_eqtlgen
PD                             plotting_MR.R
README.md                      running_MR.log
TGZ                            split_exposure_into_chunks
```

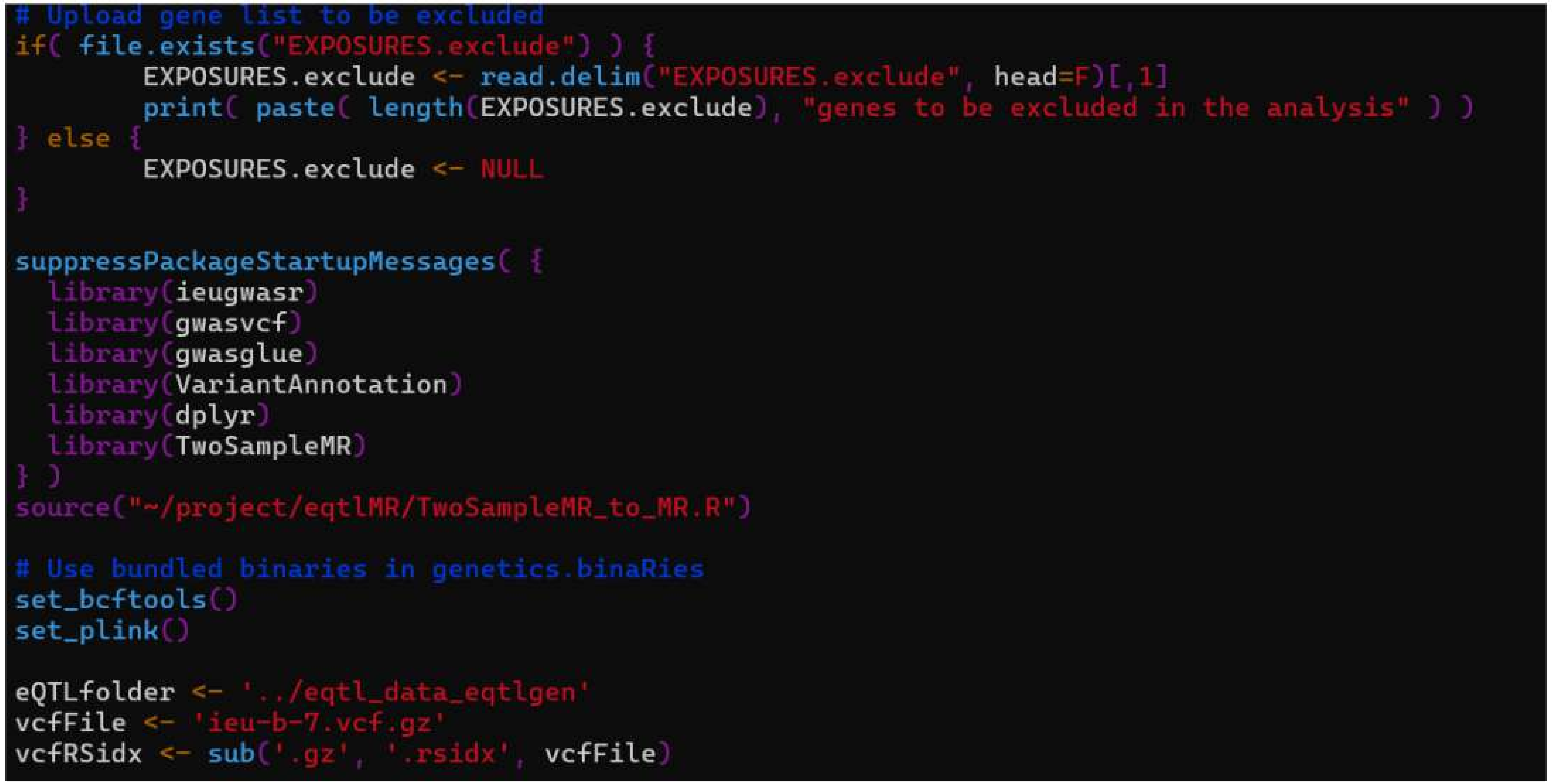
Download the entire script files from GitHub

```
~/project$ git clone https://github.com/LAIDDeqtlMR/eqtlMR.git
```

Download the data files from Web

```
~/project/eqtlMR/eqtl_data_eqtlgen$ wget https://molgenis26.gcc.rug.nl/downloads/eqtlgen/cis-eqtl/2019-12-11-cis-eQTLsFDR0.05-ProbeLevel-CohortInfoRemoved-BonferroniAdded.txt.gz
```

Files and Data

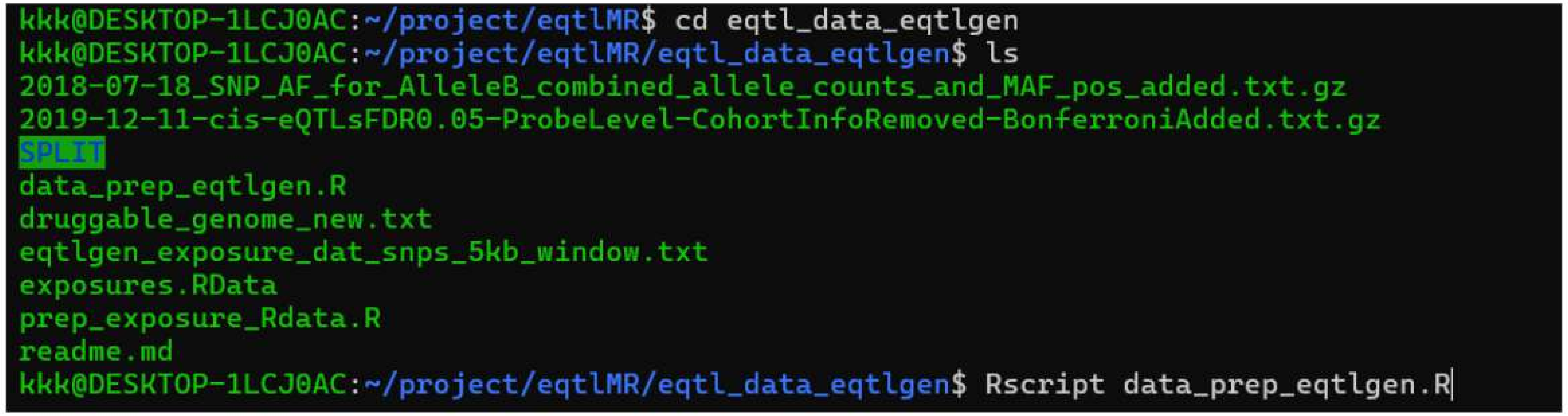


```
# Upload gene list to be excluded
if( file.exists("EXPOSURES.exclude") ) {
        EXPOSURES.exclude <- read.delim("EXPOSURES.exclude", head=F)[,1]
        print( paste( length(EXPOSURES.exclude), "genes to be excluded in the analysis" ) )
} else {
        EXPOSURES.exclude <- NULL
}

suppressPackageStartupMessages( {
  library(ieugwasr)
  library(gwasvcf)
  library(gwasglue)
  library(VariantAnnotation)
  library(dplyr)
  library(TwoSampleMR)
} )
source("~/project/eqtlMR/TwoSampleMR_to_MR.R")

# Use bundled binaries in genetics.binaRies
set_bcftools()
set_plink()

eQTLfolder <- '../eqtl_data_eqtlgen'
vcfFile <- 'ieu-b-7.vcf.gz'
vcfRSidx <- sub('.gz', '.rsidx', vcfFile)
```

Modify the file according to the research objectives.

~/project/eqtlMR/PD$ vi running_MR.R

Files and Data

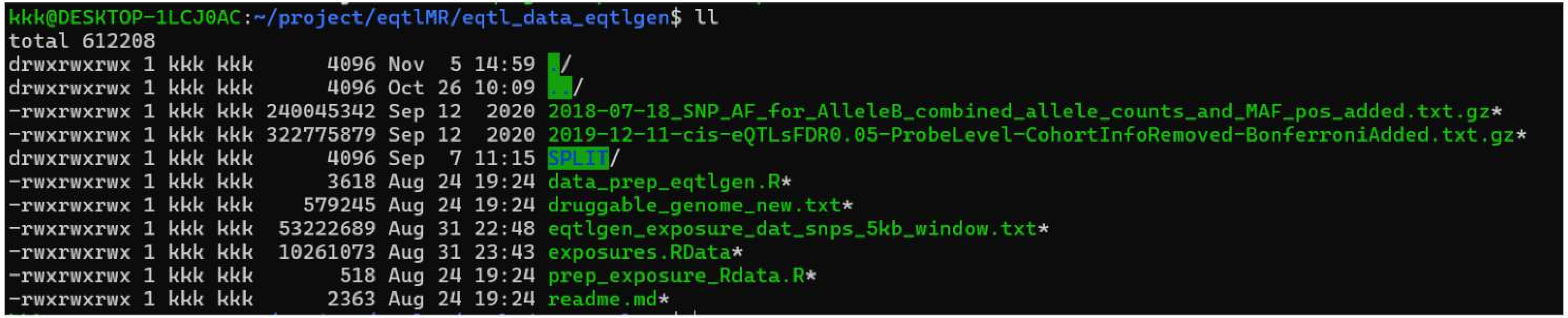


```
kkk@DESKTOP-1LCJ0AC:~/project/eqtlMR$ cd eqtl_data_eqtlgen
kkk@DESKTOP-1LCJ0AC:~/project/eqtlMR/eqtl_data_eqtlgen$ ls
2018-07-18_SNP_AF_for_AlleleB_combined_allele_counts_and_MAF_pos_added.txt.gz
2019-12-11-cis-eQTLsFDR0.05-ProbeLevel-CohortInfoRemoved-BonferroniAdded.txt.gz
SPLIT
data_prep_eqtlgen.R
druggable_genome_new.txt
eqtlgen_exposure_dat_snps_5kb_window.txt
exposures.RData
prep_exposure_Rdata.R
readme.md
kkk@DESKTOP-1LCJ0AC:~/project/eqtlMR/eqtl_data_eqtlgen$ Rscript data_prep_eqtlgen.R
```

Execute a R file in a Linux environment

~/project/eqtlMR/eqtl_data_eqtlgen$ Rscript data_prep_eqtlgen.R
~/project/eqtlMR/eqtl_data_eqtlgen$ Rscript Prep_exposure_Rdata.R

Files and Data

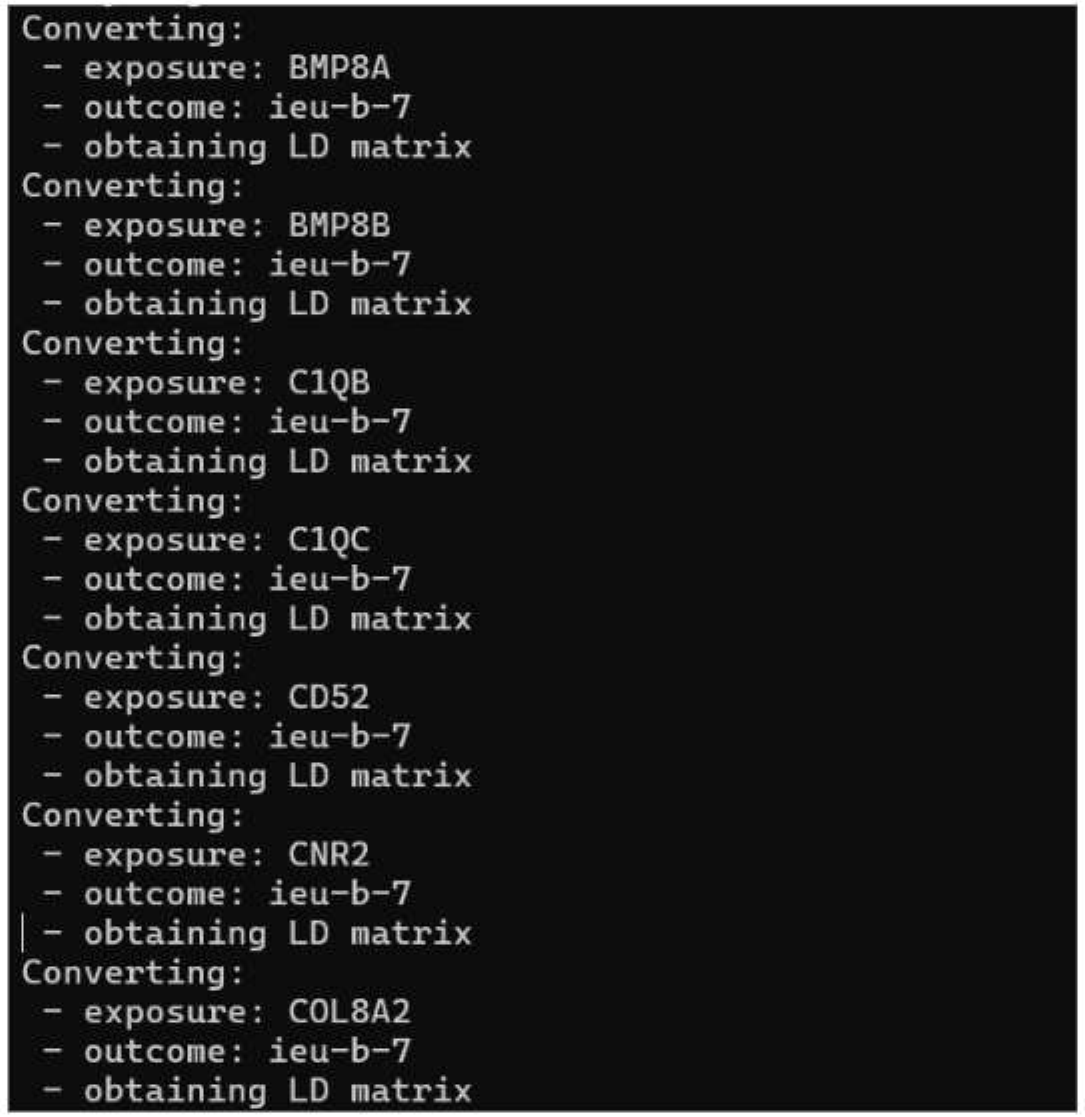


```
kkk@DESKTOP-1LCJ0AC:~/project/eqtlMR/eqtl_data_eqtlgen$ ll
total 612208
drwxrwxrwx 1 kkk kkk      4096 Nov  5 14:59 ./
drwxrwxrwx 1 kkk kkk      4096 Oct 26 10:09 ../
-rwxrwxrwx 1 kkk kkk 240045342 Sep 12  2020 2018-07-18_SNP_AF_for_AlleleB_combined_allele_counts_and_MAF_pos_added.txt.gz*
-rwxrwxrwx 1 kkk kkk 322775879 Sep 12  2020 2019-12-11-cis-eQTLsFDR0.05-ProbeLevel-CohortInfoRemoved-BonferroniAdded.txt.gz*
drwxrwxrwx 1 kkk kkk      4096 Sep  7 11:15 SPLIT/
-rwxrwxrwx 1 kkk kkk      3618 Aug 24 19:24 data_prep_eqtlgen.R*
-rwxrwxrwx 1 kkk kkk    579245 Aug 24 19:24 druggable_genome_new.txt*
-rwxrwxrwx 1 kkk kkk  53222689 Aug 31 22:48 eqtlgen_exposure_dat_snps_5kb_window.txt*
-rwxrwxrwx 1 kkk kkk  10261073 Aug 31 23:43 exposures.RData*
-rwxrwxrwx 1 kkk kkk       518 Aug 24 19:24 prep_exposure_Rdata.R*
-rwxrwxrwx 1 kkk kkk      2363 Aug 24 19:24 readme.md*
```

Split Large File to several chunks

In Ubuntu, we analyzed the data file "eqtlgen_exposure_dat_snps_5kb_window.txt" in R.
However, since the file size is large, it may take more than a day to run on a personal laptop.
To avoid potential errors during this long process, we splited the file into several smaller files.

Files and Data



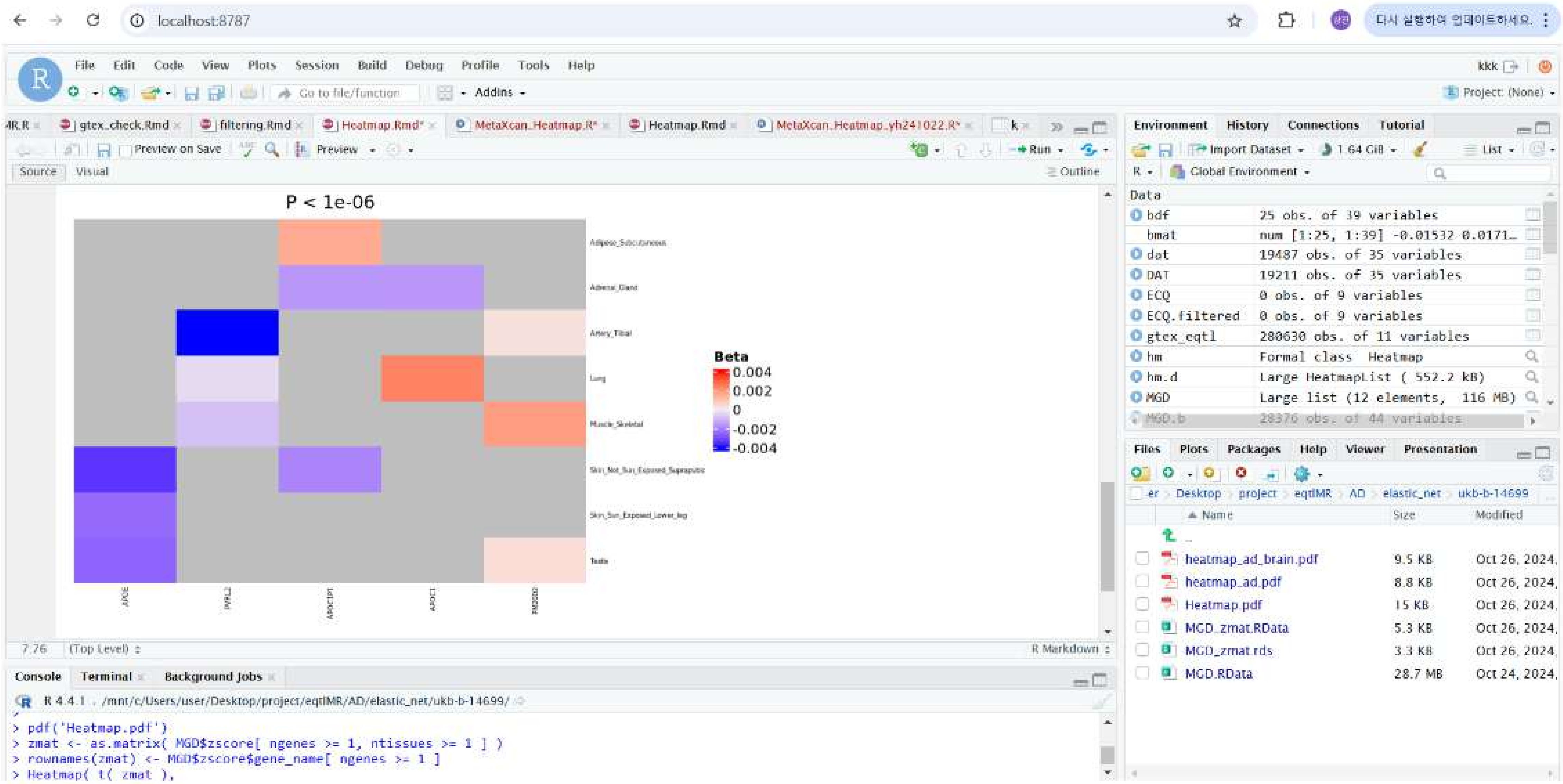
```
Converting:
 - exposure: BMP8A
 - outcome: ieu-b-7
 - obtaining LD matrix
Converting:
 - exposure: BMP8B
 - outcome: ieu-b-7
 - obtaining LD matrix
Converting:
 - exposure: C1QB
 - outcome: ieu-b-7
 - obtaining LD matrix
Converting:
 - exposure: C1QC
 - outcome: ieu-b-7
 - obtaining LD matrix
Converting:
 - exposure: CD52
 - outcome: ieu-b-7
 - obtaining LD matrix
Converting:
 - exposure: CNR2
 - outcome: ieu-b-7
 - obtaining LD matrix
Converting:
 - exposure: COL8A2
 - outcome: ieu-b-7
 - obtaining LD matrix
```

Recording specific outputs to a log file

By logging error outputs to the file, it becomes possible to trace and analyze problems that occur during exection. This is very useful in the debugging process.

```
~/project/eqtlMR/PD$ Rscript running_MR.R 100 151 2> running_MR.log
```

Visualizaton



```
R 4.4.1 . /mnt/c/Users/user/Desktop/project/eqtlMR/AD/elastic_net/ukb-b-14699/
> pdf('Heatmap.pdf')
> zmat <- as.matrix( MGD$zscore[ ngenes >= 1, ntissues >= 1 ] )
> rownames(zmat) <- MGD$zscore$gene_name[ ngenes >= 1 ]
> Heatmap( t( zmat ),
```

```
Data
bdf            25 obs. of 39 variables
bmat           num [1:25, 1:39] -0.01532 0.0171…
dat            19487 obs. of 35 variables
DAT            19211 obs. of 35 variables
ECQ            0 obs. of 9 variables
ECQ.filtered   0 obs. of 9 variables
gtex_eqtl      280630 obs. of 11 variables
hm             Formal class  Heatmap
hm.d           Large HeatmapList ( 552.2 kB)
MGD            Large list (12 elements,  116 MB)
MGD.b          28376 obs. of 44 variables
```

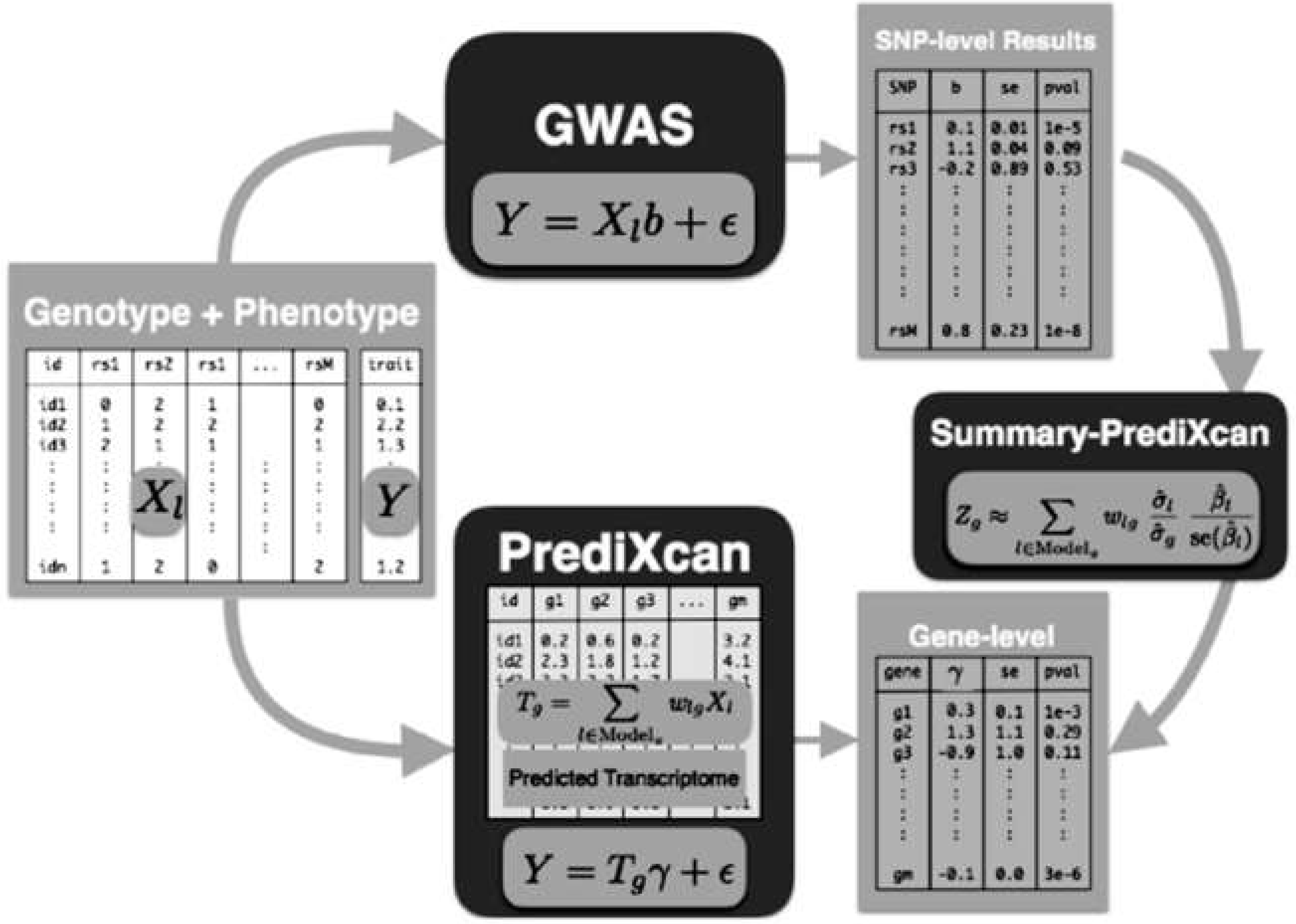
er › Desktop › project › eqtlMR › AD › elastic_net › ukb-b-14699

Name	Size	Modified
..		
heatmap_ad_brain.pdf	9.5 KB	Oct 26, 2024,
heatmap_ad.pdf	8.8 KB	Oct 26, 2024,
Heatmap.pdf	15 KB	Oct 26, 2024,
MGD_zmat.RData	5.3 KB	Oct 26, 2024,
MGD_zmat.rds	3.3 KB	Oct 26, 2024,
MGD.RData	28.7 MB	Oct 24, 2024,

To perform data visualization using RStudio

check the location of input data and the current working directory, and then execute the file

3. Method: MetaXcan using Elastic-net


GWAS
$Y = X_l b + \epsilon$
Genotype + Phenotype
X_l
Y
SNP-level Results
Summary-PrediXcan
$Z_g \approx \sum_{l \in \text{Model}_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)}$
PrediXcan
$T_g = \sum_{l \in \text{Model}_g} w_{lg} X_l$
Predicted Transcriptome
$Y = T_g \gamma + \epsilon$
Gene-level

id	rs1	rs2	rs1	...	rsM	trait
id1	0	2	1		0	0.1
id2	1	2	2		2	2.2
id3	2	1	1		1	1.3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
idn	1	2	0		2	1.2

SNP	b	se	pval
rs1	0.1	0.01	1e-5
rs2	1.1	0.04	0.09
rs3	-0.2	0.89	0.53
⋮	⋮	⋮	⋮
rsM	0.8	0.23	1e-8

id	g1	g2	g3	...	gm
id1	0.2	0.6	0.2		3.2
id2	2.3	1.8	1.2		4.1

gene	γ	se	pval
g1	0.3	0.1	1e-3
g2	1.3	1.1	0.29
g3	-0.9	1.0	0.11
⋮	⋮	⋮	⋮
gm	-0.1	0.0	3e-6

While Mendelian Randomization (MR) analyzes causal relationships to assess the impact of specific exposures (e.g., genetic variants) on outcomes (e.g., diseases), MetaXcan focuses on estimating the relationship between gene expression and phenotypes, primarily predicting phenotypes through changes in gene expression. MetaXcan uses Elastic Net regression to enable more accurate predictions through variable selection and model generalization.

AWS EC2

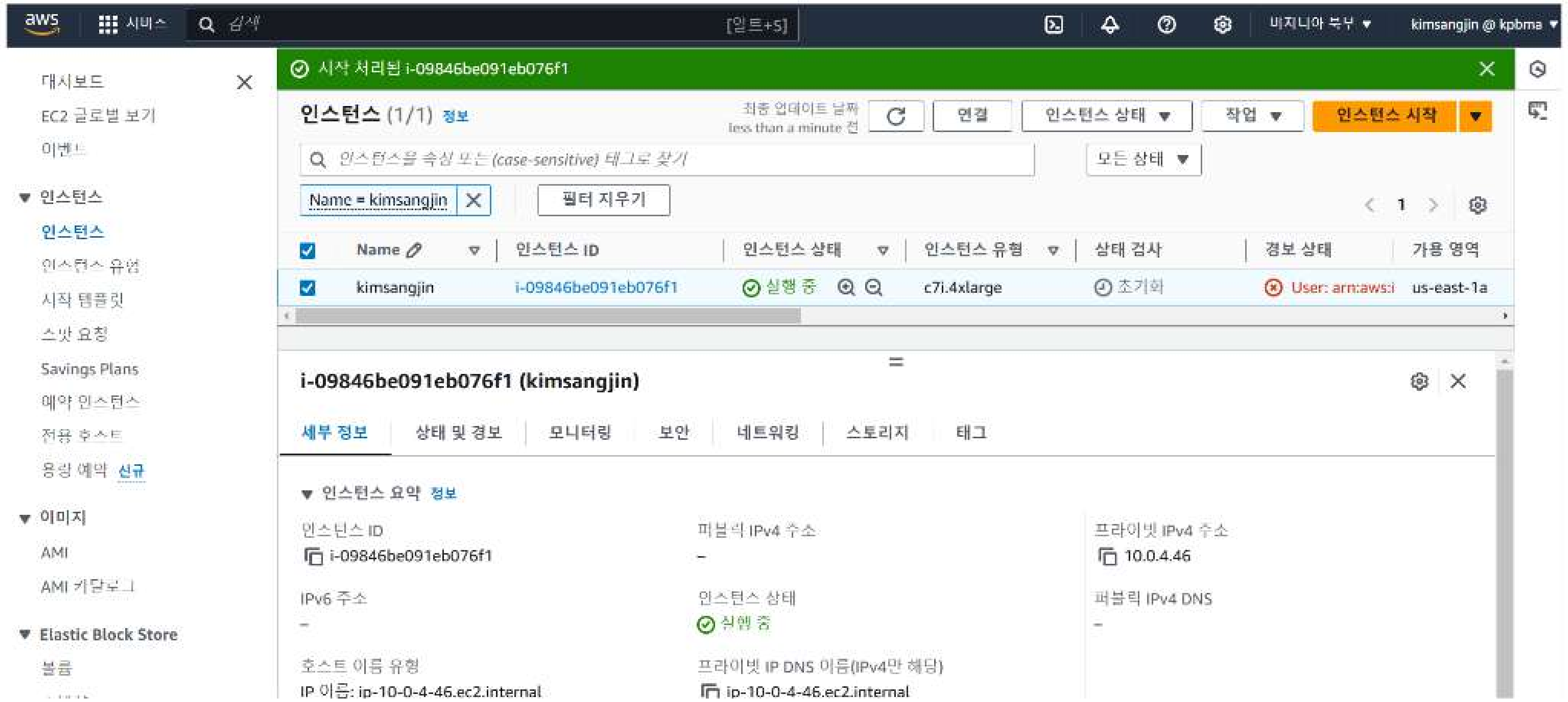
Amazon EC2

Amazon S3

AWS EC2 (Amazon Elastic Computer Cloud)

Amazon EC2 was to shared data and files and to facilitate efficient work processes. In AWS EC2, instances can be configured according to specific needs by selecting CPU, memory, storage capacity, and more. In a team project, team members are assigned different instance IDs, and they can use Amazon S3, an object storage service accessible from multiple instances, to store and share data.

AWS EC2



시작 처리됨 i-09846be091eb076f1

인스턴스 (1/1) 정보

최종 업데이트 날짜
less than a minute 전

Name = kimsangjin

☑	Name	인스턴스 ID	인스턴스 상태	인스턴스 유형	상태 검사	경보 상태	가용 영역
☑	kimsangjin	i-09846be091eb076f1	실행 중	c7i.4xlarge	초기화	User: arn:aws:i	us-east-1a

i-09846be091eb076f1 (kimsangjin)

▼ 인스턴스 요약 정보

인스턴스 ID: i-09846be091eb076f1

퍼블릭 IPv4 주소: –

프라이빗 IPv4 주소: 10.0.4.46

IPv6 주소: –

인스턴스 상태: 실행 중

퍼블릭 IPv4 DNS: –

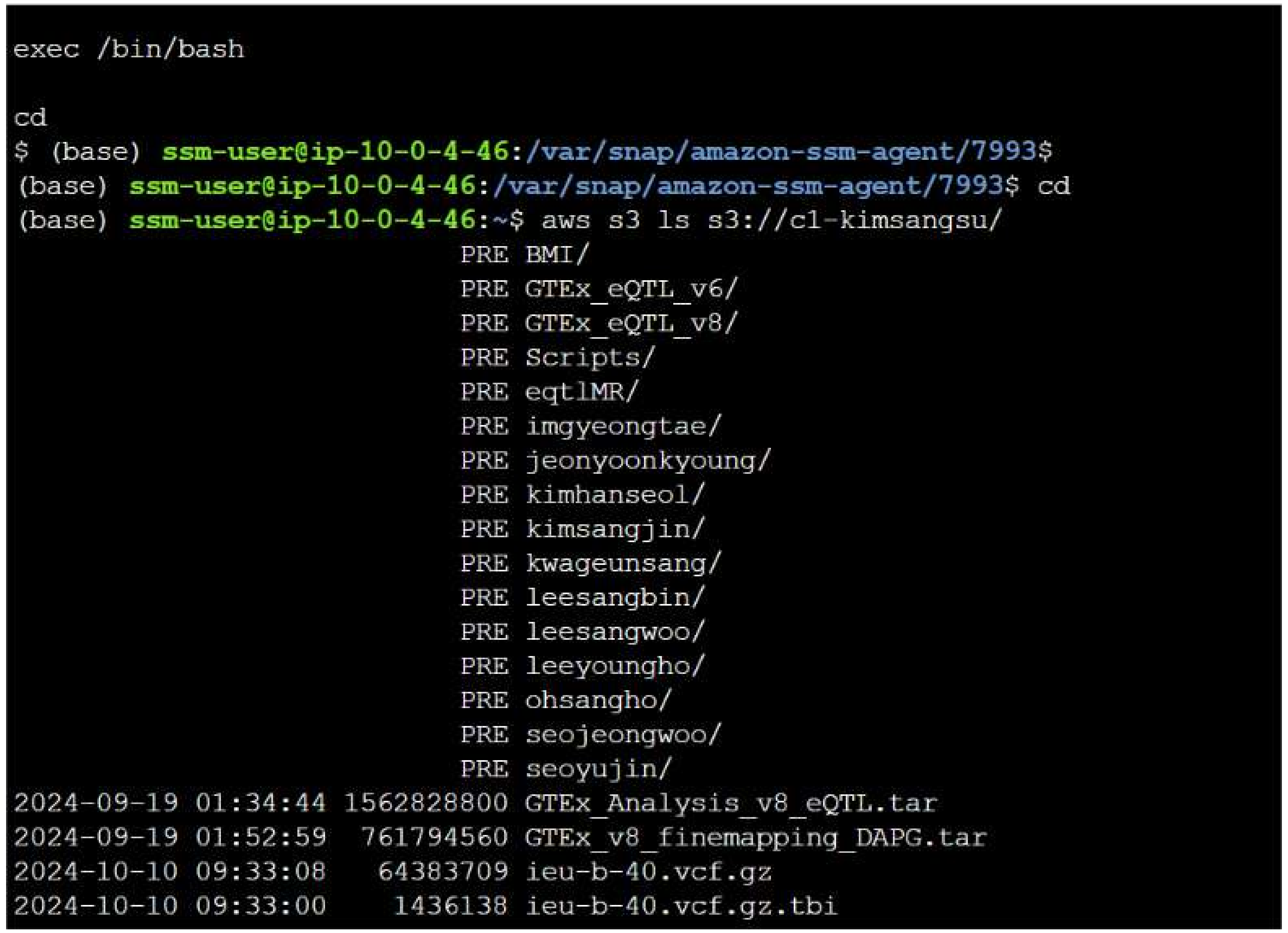
호스트 이름 유형: IP 이름: ip-10-0-4-46.ec2.internal

프라이빗 IP DNS 이름(IPv4만 해당): ip-10-0-4-46.ec2.internal

AWS EC2 (Amazon Elastic Computer Cloud)

Amazon EC2 was to shared data and files and to facilitate efficient work processes.In AWS EC2, instances can be configured according to specific needs by selecting CPU, memory, storage capacity, and more. In a team project, team members are assigned different instance IDs, and they can use Amazon S3, an object storage service accessible from multiple instances, to store and share data.

AWS EC2

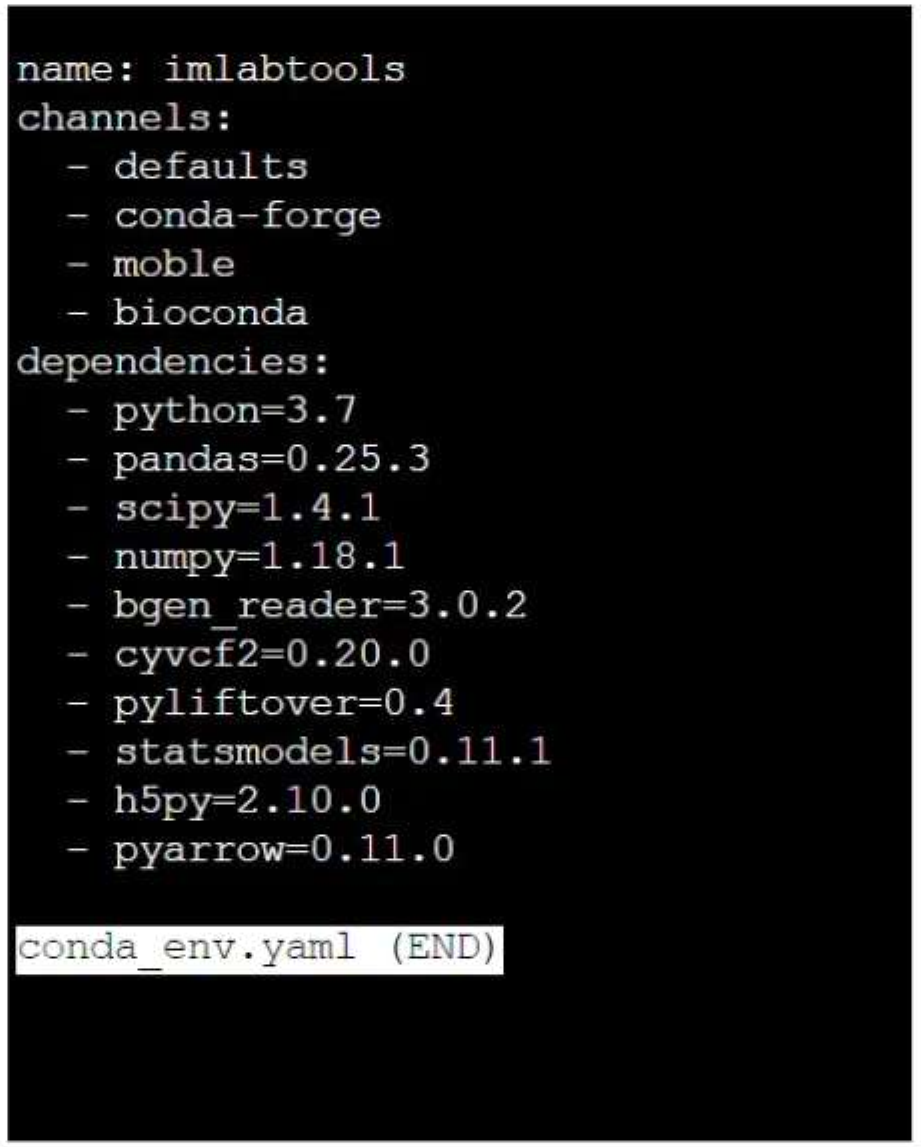


```
exec /bin/bash

cd
$ (base) ssm-user@ip-10-0-4-46:/var/snap/amazon-ssm-agent/7993$
 (base) ssm-user@ip-10-0-4-46:/var/snap/amazon-ssm-agent/7993$ cd
 (base) ssm-user@ip-10-0-4-46:~$ aws s3 ls s3://c1-kimsangsu/
                           PRE BMI/
                           PRE GTEx_eQTL_v6/
                           PRE GTEx_eQTL_v8/
                           PRE Scripts/
                           PRE eqtlMR/
                           PRE imgyeongtae/
                           PRE jeonyoonkyoung/
                           PRE kimhanseol/
                           PRE kimsangjin/
                           PRE kwageunsang/
                           PRE leesangbin/
                           PRE leesangwoo/
                           PRE leeyoungho/
                           PRE ohsangho/
                           PRE seojeongwoo/
                           PRE seoyujin/
2024-09-19 01:34:44 1562828800 GTEx_Analysis_v8_eQTL.tar
2024-09-19 01:52:59  761794560 GTEx_v8_finemapping_DAPG.tar
2024-10-10 09:33:08   64383709 ieu-b-40.vcf.gz
2024-10-10 09:33:00    1436138 ieu-b-40.vcf.gz.tbi
```

AWS EC2 (Amazon Elastic Computer Cloud)

Amazon EC2 was to shared data and files and to facilitate efficient work processes.In AWS EC2, instances can be configured according to specific needs by selecting CPU, memory, storage capacity, and more. In a team project, team members are assigned different instance IDs, and they can use Amazon S3, an object storage service accessible from multiple instances, to store and share data.

AWS EC2



```
name: imlabtools
channels:
  - defaults
  - conda-forge
  - moble
  - bioconda
dependencies:
  - python=3.7
  - pandas=0.25.3
  - scipy=1.4.1
  - numpy=1.18.1
  - bgen_reader=3.0.2
  - cyvcf2=0.20.0
  - pyliftover=0.4
  - statsmodels=0.11.1
  - h5py=2.10.0
  - pyarrow=0.11.0

conda_env.yaml (END)
```

Virtual environment set up

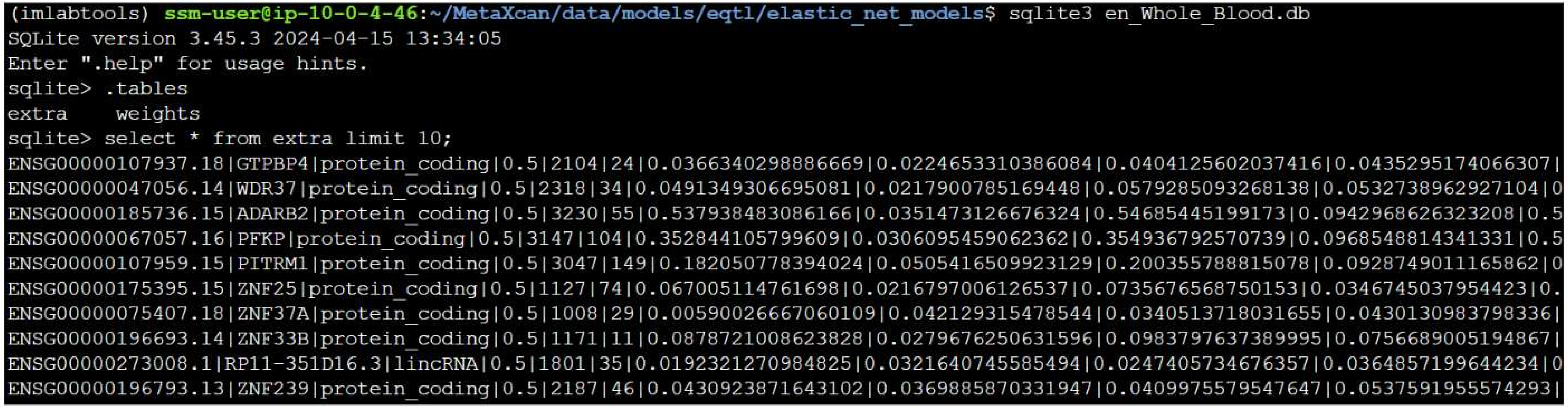
In AWS EC2, you can create a virtual environment for analysis, download the necessary packages, and activate that environment.

AWS EC2

ANACONDA®

```
# Activate virtual environment: base -> imlabtools

~$ conda activate imlabtools
```

AWS EC2

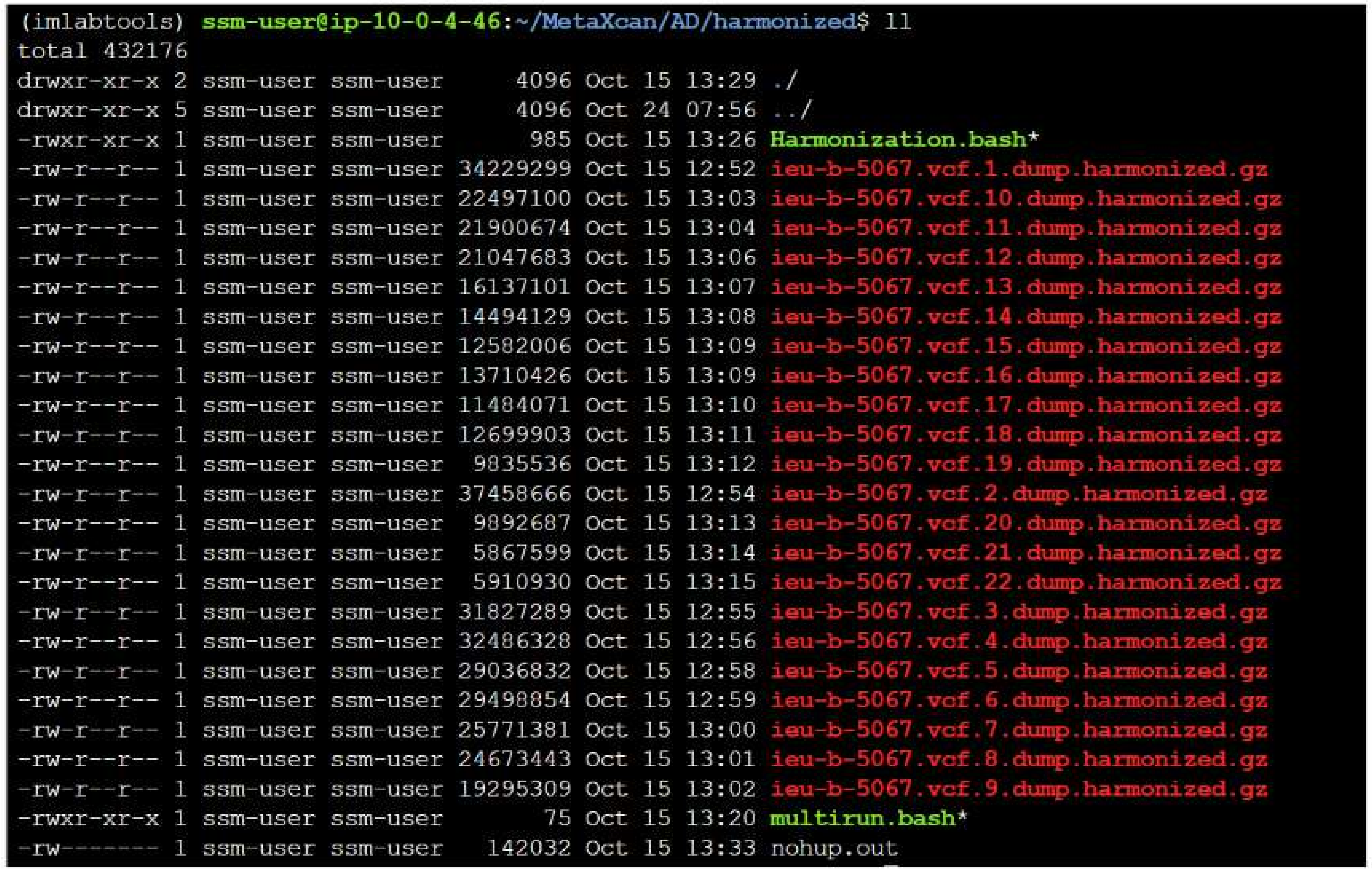


```
(imlabtools) ssm-user@ip-10-0-4-46:~/MetaXcan/data/models/eqtl/elastic_net_models$ sqlite3 en_Whole_Blood.db
SQLite version 3.45.3 2024-04-15 13:34:05
Enter ".help" for usage hints.
sqlite> .tables
extra    weights
sqlite> select * from extra limit 10;
ENSG00000107937.18|GTPBP4|protein_coding|0.5|2104|24|0.0366340298886669|0.0224653310386084|0.0404125602037416|0.0435295174066307|
ENSG00000047056.14|WDR37|protein_coding|0.5|2318|34|0.0491349306695081|0.0217900785169448|0.0579285093268138|0.0532738962927104|0
ENSG00000185736.15|ADARB2|protein_coding|0.5|3230|55|0.537938483086166|0.0351473126676324|0.54685445199173|0.0942968626323208|0.5
ENSG00000067057.16|PFKP|protein_coding|0.5|3147|104|0.352844105799609|0.0306095459062362|0.354936792570739|0.0968548814341331|0.5
ENSG00000107959.15|PITRM1|protein_coding|0.5|3047|149|0.182050778394024|0.0505416509923129|0.200355788815078|0.0928749011165862|0
ENSG00000175395.15|ZNF25|protein_coding|0.5|1127|74|0.067005114761698|0.0216797006126537|0.0735676568750153|0.0346745037954423|0.
ENSG00000075407.18|ZNF37A|protein_coding|0.5|1008|29|0.00590026667060109|0.042129315478544|0.0340513718031655|0.0430130983798336|
ENSG00000196693.14|ZNF33B|protein_coding|0.5|1171|11|0.0878721008623828|0.0279676250631596|0.0983797637389995|0.0756689005194867|
ENSG00000273008.1|RP11-351D16.3|lincRNA|0.5|1801|35|0.0192321270984825|0.0321640745585494|0.0247405734676357|0.0364857199644234|0
ENSG00000196793.13|ZNF239|protein_coding|0.5|2187|46|0.0430923871643102|0.0369885870331947|0.0409975579547647|0.0537591955574293|
```

Using SQL to manage .db files

It is stored as a .db file because it is advantageous for managing complex data structures and allows for easy access to required data through queries. Files that end with .db typically represent database files, and in many cases, they are SQLite database files. These files are used to store and manage structured data.

AWS EC2

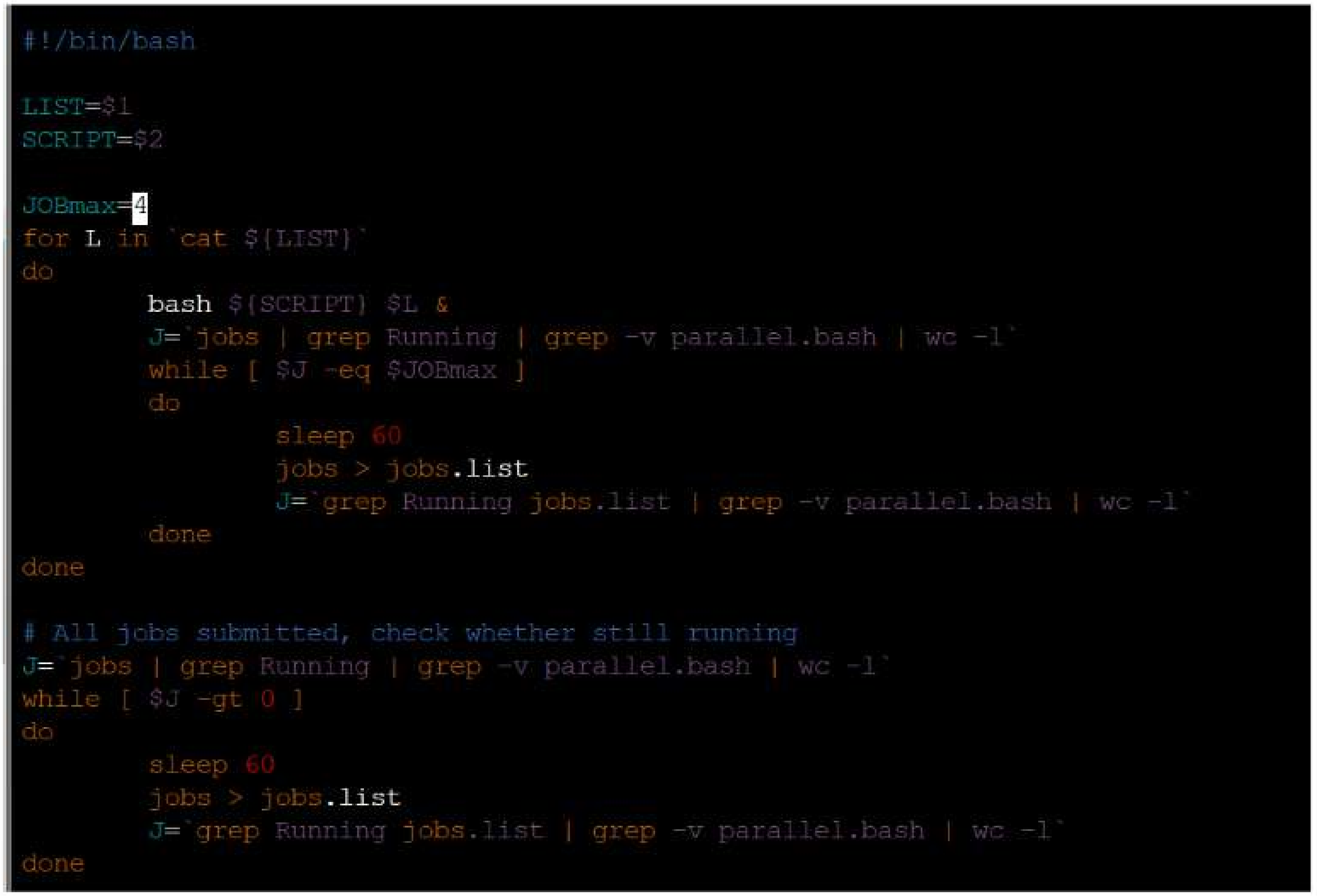


```
(imlabtools) ssm-user@ip-10-0-4-46:~/MetaXcan/AD/harmonized$ ll
total 432176
drwxr-xr-x 2 ssm-user ssm-user     4096 Oct 15 13:29 ./
drwxr-xr-x 5 ssm-user ssm-user     4096 Oct 24 07:56 ../
-rwxr-xr-x 1 ssm-user ssm-user      985 Oct 15 13:26 Harmonization.bash*
-rw-r--r-- 1 ssm-user ssm-user 34229299 Oct 15 12:52 ieu-b-5067.vcf.1.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 22497100 Oct 15 13:03 ieu-b-5067.vcf.10.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 21900674 Oct 15 13:04 ieu-b-5067.vcf.11.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 21047683 Oct 15 13:06 ieu-b-5067.vcf.12.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 16137101 Oct 15 13:07 ieu-b-5067.vcf.13.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 14494129 Oct 15 13:08 ieu-b-5067.vcf.14.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 12582006 Oct 15 13:09 ieu-b-5067.vcf.15.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 13710426 Oct 15 13:09 ieu-b-5067.vcf.16.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 11484071 Oct 15 13:10 ieu-b-5067.vcf.17.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 12699903 Oct 15 13:11 ieu-b-5067.vcf.18.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user  9835536 Oct 15 13:12 ieu-b-5067.vcf.19.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 37458666 Oct 15 12:54 ieu-b-5067.vcf.2.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user  9892687 Oct 15 13:13 ieu-b-5067.vcf.20.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user  5867599 Oct 15 13:14 ieu-b-5067.vcf.21.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user  5910930 Oct 15 13:15 ieu-b-5067.vcf.22.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 31827289 Oct 15 12:55 ieu-b-5067.vcf.3.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 32486328 Oct 15 12:56 ieu-b-5067.vcf.4.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 29036832 Oct 15 12:58 ieu-b-5067.vcf.5.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 29498854 Oct 15 12:59 ieu-b-5067.vcf.6.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 25771381 Oct 15 13:00 ieu-b-5067.vcf.7.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 24673443 Oct 15 13:01 ieu-b-5067.vcf.8.dump.harmonized.gz
-rw-r--r-- 1 ssm-user ssm-user 19295309 Oct 15 13:02 ieu-b-5067.vcf.9.dump.harmonized.gz
-rwxr-xr-x 1 ssm-user ssm-user       75 Oct 15 13:20 multirun.bash*
-rw------- 1 ssm-user ssm-user   142032 Oct 15 13:33 nohup.out
```

Large Chromosome data

You need to process the files containing chromosome information for analysis. The total size of all the files is 426 gigabytes, and each chromosome is approximately 20 gigabytes, with the largest being 34 gigabytes. Since processing each chromosome individually is too large, you plan to divide each chromosome into 10 batches, creating batches from 0 to 219. Once the processing is complete, you will regroup them back into sets of 10.

AWS EC2

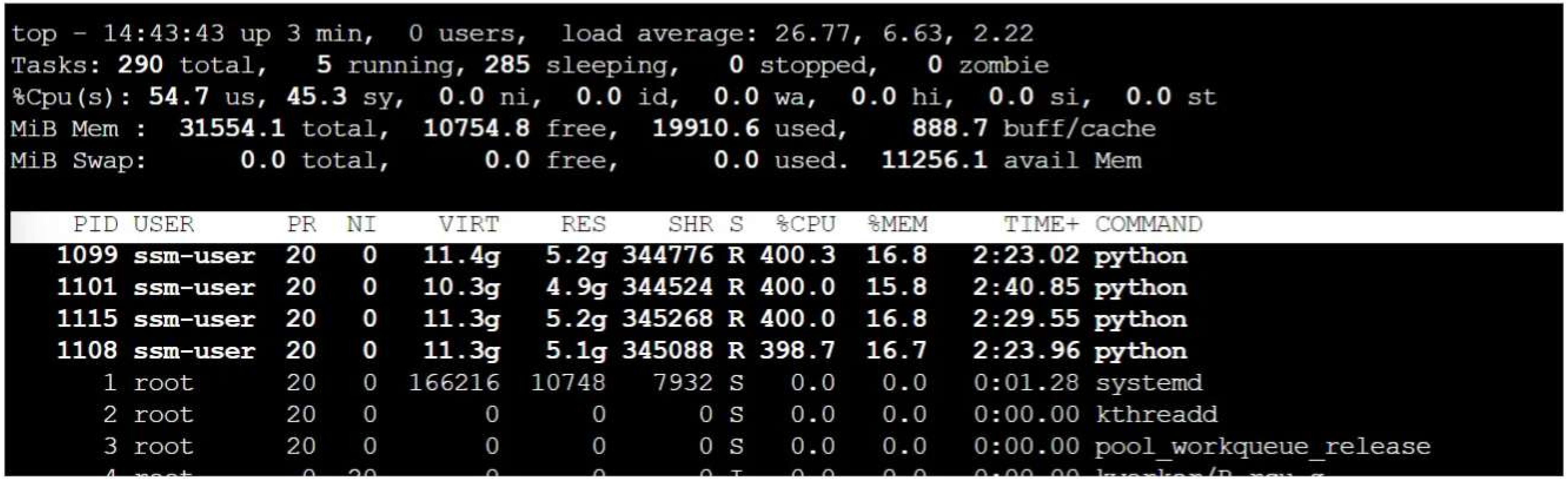


```
#!/bin/bash

LIST=$1
SCRIPT=$2

JOBmax=4
for L in `cat ${LIST}`
do
        bash ${SCRIPT} $L &
        J=`jobs | grep Running | grep -v parallel.bash | wc -l`
        while [ $J -eq $JOBmax ]
        do
                sleep 60
                jobs > jobs.list
                J=`grep Running jobs.list | grep -v parallel.bash | wc -l`
        done
done

# All jobs submitted, check whether still running
J=`jobs | grep Running | grep -v parallel.bash | wc -l`
while [ $J -gt 0 ]
do
        sleep 60
        jobs > jobs.list
        J=`grep Running jobs.list | grep -v parallel.bash | wc -l`
done
```

Parallel Computing

The JOBmax variable is used to limit the maximum number of jobs that can run simultaneously to n.
This value should be adjusted based on the computer's performance.

AWS EC2



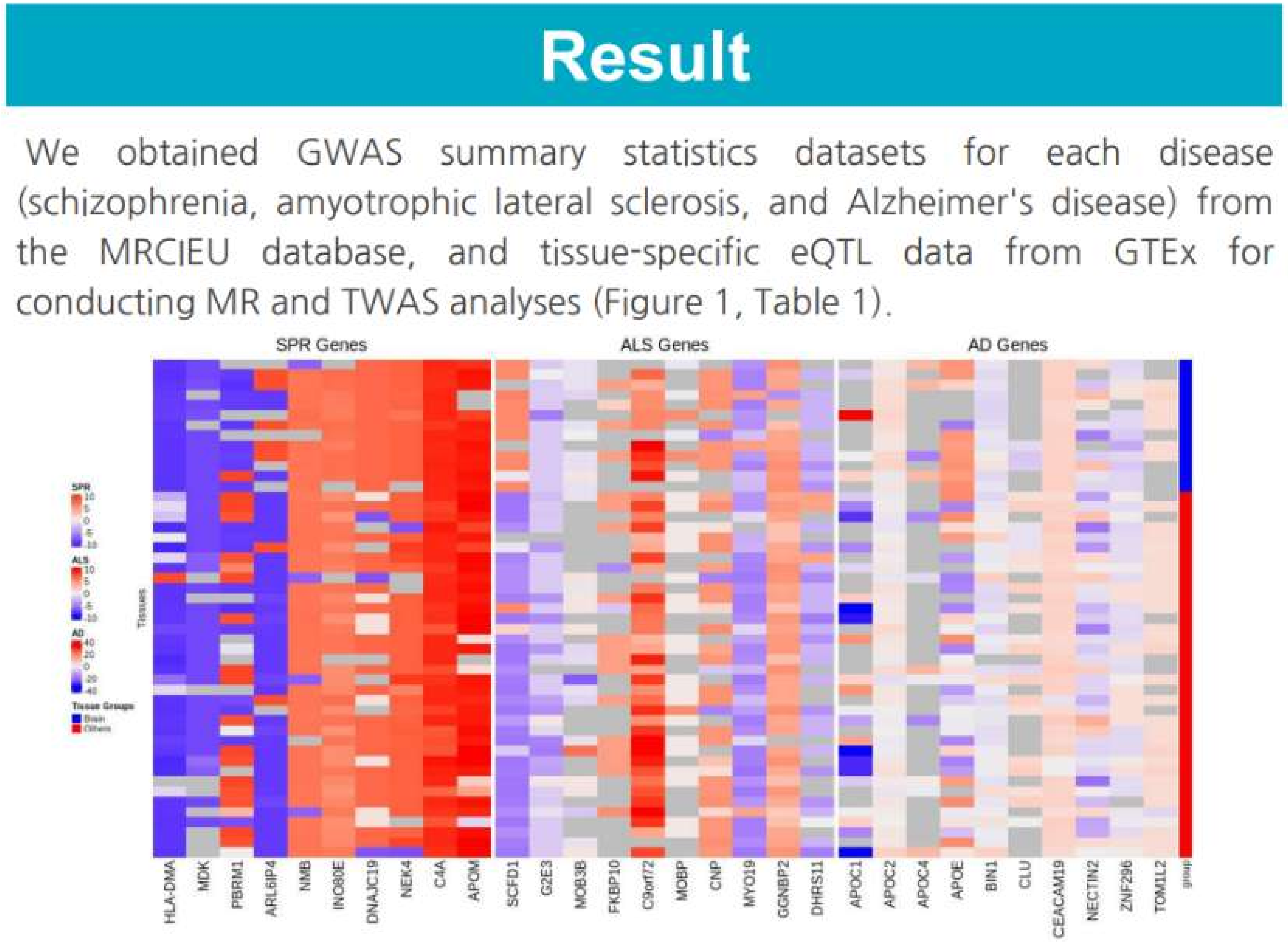
```
top - 14:43:43 up 3 min,  0 users,  load average: 26.77, 6.63, 2.22
Tasks: 290 total,   5 running, 285 sleeping,   0 stopped,   0 zombie
%Cpu(s): 54.7 us, 45.3 sy,  0.0 ni,  0.0 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
MiB Mem :  31554.1 total,  10754.8 free,  19910.6 used,    888.7 buff/cache
MiB Swap:      0.0 total,      0.0 free,      0.0 used.  11256.1 avail Mem

    PID USER      PR  NI    VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND
   1099 ssm-user  20   0   11.4g   5.2g 344776 R 400.3  16.8   2:23.02 python
   1101 ssm-user  20   0   10.3g   4.9g 344524 R 400.0  15.8   2:40.85 python
   1115 ssm-user  20   0   11.3g   5.2g 345268 R 400.0  16.8   2:29.55 python
   1108 ssm-user  20   0   11.3g   5.1g 345088 R 398.7  16.7   2:23.96 python
      1 root      20   0  166216  10748   7932 S   0.0   0.0   0:01.28 systemd
      2 root      20   0       0      0      0 S   0.0   0.0   0:00.00 kthreadd
      3 root      20   0       0      0      0 S   0.0   0.0   0:00.00 pool_workqueue_release
```

top

To monitor the CPU and memory usage of your currently running code using the top command.

4. Results

Visualization: Heatmap



Result

We obtained GWAS summary statistics datasets for each disease (schizophrenia, amyotrophic lateral sclerosis, and Alzheimer's disease) from the MRCIEU database, and tissue-specific eQTL data from GTEx for conducting MR and TWAS analyses (Figure 1, Table 1).

Figure 2. Heatmaps of gene expression z-score from MetaXcan results for SPR, ALS, and AD.

Each heatmap presents the z-scores of genes associated with SRP (left), ALS (middle), and ALS (right), as obtained from MetaXcan analysis. The z-scores indicate the strength and direction of the association between gene expression levels based on GWAS data and disease onset. Positive z-scores suggest that increased expression of the gene is associated with a higher risk of disease, while negative z-score indicate that increased expression is associated with a lower risk.

4. Results

Cross Validation with DataBase

GeneCards®
THE HUMAN GENE DATABASE

TTD

DRUGBANK Online

To compensate for the limitations of in-silico methods, validation using public databases or reference literature is necessary.

Linux™

R Studio®

python
Programming

SQL

GitHub

aws

Project Details

You can all see more details about the AIDD project through the QR code above, which links to my Notion page.

More Interest:AlphaFold

If you're curious about simple code for protein structure prediction using AlphaFold and other methods, try using this QR code to access Google Drive.

2024 International Academic Exchange Conference

KOREA UNIVERSITY
SEJONG CAMPUS

TOKAI UNIVERSITY

Thank You

Kim SangJin
Department of Big Data Science, Korea University Graduate School