Jan 03, 2021

Protocols for study of "Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease"

Nana Liu[1], Jiayuan Xu[1], Huaigui Liu[1], Shijie Zhang[2], Miaoxin Li[3], Yao Zhou[2], Wen Qin[1], Mulin Jun Li[2], Chunshui Yu[1]

[1]Department of Radiology and Tianjin Key Laboratory of Functional Imaging, Tianjin Medical University General Hospital, Tianjin 300052, P.R. China;

[2]The Province and Ministry Co-sponsored Collaborative Innovation Center for Medical Epigenetics, Tianjin Key Laboratory of Medical Epigenetics, Department of Pharmacology, Tianjin Medical University, Tianjin 300070, P.R. China;

[3]Department of Medical Genetics, Center for Genome Research, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, 510080, P.R. China

1 Works for me   dx.doi.org/10.17504/protocols.io.bp4amqse

Nana Liu

DOI

dx.doi.org/10.17504/protocols.io.bp4amqse

PROTOCOL CITATION

Nana Liu, Jiayuan Xu, Huaigui Liu, Shijie Zhang, Miaoxin Li, Yao Zhou, Wen Qin, Mulin Jun Li, Chunshui Yu 2021. Protocols for study of "Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease". **protocols.io** https://dx.doi.org/10.17504/protocols.io.bp4amqse

CREATED

Nov 25, 2020

LAST MODIFIED

Jan 03, 2021

PROTOCOL INTEGER ID

44898

1 Step1. Quality control (QC) and imputation for genotype data from GTEx version7

1.1 Pre-imputation QC

Tools: PLINK (https://www.cog-genomics.org/plink/) [1]

The sample-level QC:

a. Genotyping call rate per individual (> 98%)

```
plink --bfile ${genotype_data} --missing --out ${missingness}
awk '{if($6 > 0.02)print$0}' ${missingness.imiss} >> ${remove}
```

b. Sex concordance check

```
plink --bfile ${genotype_data} --check-sex --out ${sexcheck}
```

Citation: Nana Liu, Jiayuan Xu, Huaigui Liu, Shijie Zhang, Miaoxin Li, Yao Zhou, Wen Qin, Mulin Jun Li, Chunshui Yu (01/03/2021). Protocols for study of "Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease". https://dx.doi.org/10.17504/protocols.io.bp4amqse

c. Identity check

```
plink --bfile ${genotype_data} --indep-pairwise 50 5 0.2 --out ${relatedness}
plink --bfile ${genotype_data} --extract ${relatedness.prune.in} --min 0.2 --
genome --genome-full --out ${relatedness}
```

The SNP-level QC:

a. SNP call rate (> 85%)

b. Hardy-Weinberg Equilibrium (HWE) ($p > 1 \times 10^{-6}$)

c. Minor allele frequency (MAF) (> 1%)

```
plink --bfile ${genotype_data} --hwe 1e-6 --geno 0.15 --maf 0.01 --make-bed --
out ${output}
```

d. Remove the ambiguous strand SNPs (no A/T or C/G SNPs)

```
awk '{ if (($5=="T" && $6=="A")||($5=="A" && $6=="T ")||($5=="C" && $6=="G")||
($5=="G"&& $6=="C")) print $2, "ambig" ;else print $2;}' ${data.bim} | grep -v
ambig > ${remove_ambig.txt}
plink --bfile ${genotype_data} --extract ${remove_ambig.txt} --make-bed --out
${output}
```

1.2 Imputation

Tools: SHAPEIT2 (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html) [2]; IMPUTE2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) [3]
Reference panel: 1000 Genomes Phase 3
Shapeit

```
shapeit --input-bed chr${chromosome}.bed chr${chromosome}.bim
chr${chromosome}.fam --input-ref ${hapFile} ${legendFile} ${sampleFile} --
exclude-snp ${excludeFile} --input-map ${mapFile} -O chr${chromosome}.phased --
thread 4 --force
```

Imputation

```
impute2 -use_prephased_g -known_haps_g chr${chromosome}.phased.haps -m
${mapFile} -h ${hapFile} -l ${legendFile} -int $chunkStart $chunkEnd -Ne 20000
-o chr${chromosome}-${chunkStart}-${chunkEnd}.imputed
```

1.3 After-imputation QC

Tools: PLINK; VCFtools (https://vcftools.github.io/index.html) [4]; BCFtools (https://samtools.github.io/bcftools/) [5]
Convert to VCF format

```
plink --gen ${data.imputed} --oxford-single-chr ${chr} --sample
${phased.sample} --recode-vcf --out ${output}
bcftools concat ${data_chr*.vcf} -o ${data.vcf.gz} -O z
```

Biallelic and single-character allele codes only
Remove the ambiguous strand SNPs (no A/T or C/G SNPs)
SNP call rate = 100%
HWE $p > 1 \times 10^{-6}$
MAF > 0.01
IMPUTE info quality score > 0.8

```
vcftools --gzvcf ${data.vcf.gz} --snps ${imputed_info0.8_snp} --remove-indels -
-maf 0.01 --hwe 1e-6 --max-missing 1 --recode --out ${output}
```

2 Step2. Training prediction models by genotype and RNA-seq data

Citation: Nana Liu, Jiayuan Xu, Huaigui Liu, Shijie Zhang, Miaoxin Li, Yao Zhou, Wen Qin, Mulin Jun Li, Chunshui Yu (01/03/2021). Protocols for study of "Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease". https://dx.doi.org/10.17504/protocols.io.bp4amqse

2.1 Conditional analysis of *cis*-expression quantitative trait loci (*cis*-eQTLs)

Tools: QTLtools (https://qtltools.github.io/qtltools/) [6]

Independent *cis*-eQTLs were identified in a *cis*-window of ± 1Mb from the transcription start site (TSS) and with a moderate threshold ($p < 0.01$).

```
QTLtools cis --vcf ${genotype_data} --bed ${expression_data} --cov
${covariates_data} --mapping ${thresholds_file} --chunk n1 n2 --out ${output}
```

2.2 Training gene expression prediction models

Tools: the nested cross validated elastic-net procedure following the GTEx V7 pipeline (https://github.com/hakyimlab/PredictDB_Pipeline_GTEx_v7) [7]

1. Samples were split into 5 folds.
2. One fold was removed at a time, the remaining samples (four folds) were used to train the prediction models by elastic net with 10-fold cross-validation to tune the lambda parameter.
3. The prediction models were applied to the samples of the removed fold to evaluate the correlations between the predicted and measured expression levels and get test statistics.
4. Assessing the performance of each prediction model by average Pearson correlation coefficient of the 5 times 10-fold nested cross validation tests.
5. Training a new elastic-net model using 10-fold cross validation based on all the samples to calculate weights.

3 Step3. Transcriptome-wide association study (TWAS)

Tools: S-PrediXcan (https://github.com/hakyimlab/MetaXcan) [8]

SNP-Alzheimer's disease (AD) associations were derived from genome-wide association study (GWAS) summery data, SNP-expression associations were assessed by the weighted value for each SNP's relative contribution to the gene's expression level of the prediction models, and linkage disequilibrium (LD) reference set was created by the prediction models.

```
python MetaXcan.py --model_db_path ${db_file} --covariance ${LD_ref} --gwas_folder
${gwas_folder} --gwas_file_pattern ${gwas_file} --snp_column ${SNP_name} --
effect_allele_column ${effect_allele} --non_effect_allele_column ${non_effect_allele} --
beta_column ${beta_name} --se_column ${se_name} --output_file ${output}
```

4 Step4. Gene-level fine-mapping

Tools: FOCUS (fine-mapping of causal gene sets) (https://github.com/bogdanlab/focus) [9]

Weighting table of hippocampal tissue we trained and the suggested multiple tissue, multiple eQTL reference panel weight database (https://github.com/bogdanlab/focus/wiki) were used to perform FOCUS.

4.1 Cleaning GWAS summary data

```
python focus munge ${GWAS_summary_data} --output ${GWAS_summary_data_cleaned}
```

4.2 Importing prediction modules

```
Python focus ${db_file} predixcan --tissue Hippocampus --name GTEx --assay
rnaseq --output ${output}
```

4.3 FOCUS

```
python focus finemap ${GWAS_summary_data} ${LD_ref_data} ${db_file} --chr
```

Citation: Nana Liu, Jiayuan Xu, Huaigui Liu, Shijie Zhang, Miaoxin Li, Yao Zhou, Wen Qin, Mulin Jun Li, Chunshui Yu (01/03/2021). Protocols for study of "Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease". https://dx.doi.org/10.17504/protocols.io.bp4amqse

```
${chr} --out ${output}
```

5 **Step5. Network topology-based analysis**

Tools: WEB-based GEne SeT AnaLysis Toolkit (Webgestalt, https://www.webgestalt.org) [10]

The reference database was the human PPI of the Biological General Repository for Interaction Datasets (BIOGRID) (Build 3.5.167) [11].

6 **Step6. Statistical over-representation test**

Tools: PANTHER classification system (v.14.0) (http://pantherdb.org/) [12]

The reference pathway was gene ontology (GO) biological process, Fisher's exact test was used to calculate *p*-value and BH-FDR correction was used for multiple testing ($q_c < 0.05$).

7 **Step7. Hippocampal tissue functional modules detection**

Tools: the HumanBase online tool (https://hb.flatironinstitute.org/)[13]

Functional modules were built in the context of hippocampal tissue networks. Functional enrichment was performed based on GO terms and the statistical significance of each GO term was tested by one-sided Fisher's exact test and multiple testing was corrected by BH-FDR ($q_c < 0.05$).

8 **Step8. Quality control (QC) and imputation for genotype data from Alzheimer's Disease Neuroimaging Initiative (ADNI)**

8.1 **Pre-imputation QC**

The sample-level QC:

a. Genotyping call rate per individual (> 90%)

```
plink --bfile ${genotype_data} --missing --out ${missingness}
awk '{if($6 > 0.1)print$0}' ${missingness.imiss} >> ${remove}
```

b. Sex concordance check

```
plink --bfile ${genotype_data} --check-sex --out ${sexcheck}
```

c. Identity check

```
plink --bfile ${genotype_data} --indep-pairwise 50 5 0.2 --out ${relatedness}
plink --bfile ${genotype_data} --extract ${relatedness.prune.in} --min 0.2 --
genome --genome-full --out ${relatedness}
```

d. Exclude unqualified subjects

e. Multidimensional scaling (MDS) analysis with HapMap phase III data (build 37 version)

```
plink --bfile ${genotype_and_HM3_merge} --cluster --mind 0.05 --mds-plot 4 --
extract ${snplist.txt} --out ${mds}
```

The SNP-level QC:

a. SNP call rate (> 85%)

b. HWE ($p > 1 \times 10^{-6}$)

c. MAF (> 1%)

```
plink --bfile ${genotype_data} --hwe 1e-6 --geno 0.15 --maf 0.01 --make-bed --
out ${output}
```

d. Remove the ambiguous strand SNPs (no A/T or C/G SNPs)

```
awk '{ if (($5=="T" && $6=="A")||($5=="A" && $6=="T ")||($5=="C" && $6=="G")||
($5=="G"&& $6=="C")) print $2, "ambig" ;else print $2;}' ${data.bim} | grep -v
ambig > ${remove_ambig.txt}
plink --bfile ${genotype_data} --extract ${remove_ambig.txt} --make-bed --out
${output}
```

Citation: Nana Liu, Jiayuan Xu, Huaigui Liu, Shijie Zhang, Miaoxin Li, Yao Zhou, Wen Qin, Mulin Jun Li, Chunshui Yu (01/03/2021). Protocols for study of "Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease". https://dx.doi.org/10.17504/protocols.io.bp4amqse

8.2 Imputation (same as step 1.2)

8.3 After-imputation QC

Convert to PLINK format

```
plink --gen ${data.imputed} --oxford-single-chr ${chr} --sample
${phased.sample} --make-bed --out ${output}
```

IMPUTE info quality score > 0.8

```
plink --bfile ${genotype_data} --extract ${imputed_info0.8_snp} --make-bed --
out ${output}
```

Merge genotype data

```
plink --bfile ${genotype_data} --bmerge ${genotype_data2.bed}
${genotype_data2.bim} ${genotype_data2.fam} --make-bed --out ${output}
```

SNP call rate > 85%

HWE $p > 1 \times 10^{-6}$
MAF > 0.01

```
plink --bfile ${genotype_data} --hwe 1e-6 --geno 0.15 --maf 0.01 --make-bed --
out ${output}
```

9 Step9. Predicting gene expression in ADNI

Tools:Predixcan [7]
Integrating genotype data and weighted value for each SNP's relative contribution to the gene's expression level to predict gene expression in brain tissues.

9.1 Make dosage file

```
python convert_plink_to_dosage.py -p plink -b ${genotype_data} -o ${chr}
```

9.2 Predicting expression

```
python PrediXcan.py --predict --weights ${db_file} --dosages ${dosage_folder} -
-dosages_prefix ${chr} --samples ${genotype_data.fam} --output_prefix ${output}
```

10 Step10. Validating AD-related genes in ADNI data

1. The binary logistic regression was used to compare the difference in the gene expression in hippocampal tissue between AD and cognitively normal (CN) groups as well as between mild cognitive impairment-conversion (MCI-C) and mild cognitive impairment-stable (MCI-S) groups.
2. Linear regression was performed to explore the correlation between hippocampal gene expression and hippocampal volume.
3. Multiple linear regression was used to identify the total effect of multiple genes on hippocampal volume.

11 Step11. Mediation analysis

Tools: The PROCESS macro for SPSS (v3.4) [14]
The hippocampal *cis*-genetically regulated expression for each gene was defined as an independent variable, the mean hippocampal volume as a mediator variable, the disease states (AD versus CN) as a binary dependent variable, hippocampal volume was adjusted by a linear regression with MR field strength, the covariates included age, gender, and education.

Citation: Nana Liu, Jiayuan Xu, Huaigui Liu, Shijie Zhang, Miaoxin Li, Yao Zhou, Wen Qin, Mulin Jun Li, Chunshui Yu (01/03/2021). Protocols for study of "Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease". https://dx.doi.org/10.17504/protocols.io.bp4amqse

12 **Reference**

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81(3):559-75. doi: 10.1086/519795.
2. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013;10(1):5-6. doi: 10.1038/nmeth.2307. PubMed PMID: 23269371.
3. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529. doi: 10.1371/journal.pgen.1000529. PubMed PMID: 19543373; PubMed Central PMCID: PMCPMC2689936.
4. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8. doi: 10.1093/bioinformatics/btr330.
5. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93. Epub 2011/09/10. doi: 10.1093/bioinformatics/btr509. PubMed PMID: 21903627; PubMed Central PMCID: PMCPmc3198575.
6. Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. A complete tool set for molecular QTL discovery and analysis. Nat Commun. 2017;8:15452. doi: 10.1038/ncomms15452. PubMed PMID: 28516912; PubMed Central PMCID: PMCPMC5454369.
7. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091-8. doi: 10.1038/ng.3367. PubMed PMID: 26258848; PubMed Central PMCID: PMCPMC4552594.
8. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9(1):1825. doi: 10.1038/s41467-018-03621-1. PubMed PMID: 29739930; PubMed Central PMCID: PMCPMC5940825.
9. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, et al. Probabilistic fine-mapping of transcriptome-wide association studies. Nat Genet. 2019;51(4):675-82. doi: 10.1038/s41588-019-0367-1. PubMed PMID: 30926970; PubMed Central PMCID: PMCPMC6619422.
10. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic acids research. 2019;47(W1):W199-W205.
11. Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Research. 2019;47(D1):D529-D41. doi: 10.1093/nar/gky1079.
12. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v. 14.0). Nature protocols. 2019;14(3):703-21.
13. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015;47(6):569-76. doi: 10.1038/ng.3259. PubMed PMID: 25915600; PubMed Central PMCID: PMCPMC4828725.
14.Hayes AF. Introduction to mediation, moderation, and conditional process analysis: A regression-based approach: Guilford publications; 2017.

Citation: Nana Liu, Jiayuan Xu, Huaigui Liu, Shijie Zhang, Miaoxin Li, Yao Zhou, Wen Qin, Mulin Jun Li, Chunshui Yu (01/03/2021). Protocols for study of "Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease". https://dx.doi.org/10.17504/protocols.io.bp4amqse