**Instructions**

**ENIGMA-CNV working group, v2.1**

**Table of Contents**

Joining ENIGMA-CNV with data – short overview 2

A. Administrative - sign up for ENIGMA-CNV 3

B. Structural imaging data 3

C. Diffusion Tensor Imaging data 3

D. CNVs - calling and visualization 4

1. Organize your data and identify appropriate files for CNV calling 4

a. Genetic information files 4

b. Cohort-generated files 6

2. Download necessary software, files and scripts 7

a. Software (independent of dataset) 7

b. Scripts for running analysis 8

c. Genotyping-chip-dependent files 8

3. Call and visualize CNVs 10

a. Adjust the script 10

b. Run the analysis 10

4. Return data to ENIGMA-CNV 11

E. Covariate files 11

General covariates 11

MDS-covariates 12

F. Send data 13

Notes for cohorts that ran the ‘old’ protocol and should visualize more CNVs 14

**Purpose:**

This document describes the different steps for joining ENIGMA-CNV including data analysis and submission.

Please address any questions to: [enigmacnvhelpdesk@gmail.com](mailto:enigmacnvhelpdesk@gmail.com)

# Joining ENIGMA-CNV with data – short overview

Welcome to the ENIGMA-CNV working group! We are pleased to have you onboard.

**Steps involved in joining:**

A. Administrative – sign up for ENIGMA-CNV to the working group chairs.

Analysis and data submission:

B. Structural imaging data

-minimal protocol

-extended protocol

C. Diffusion Tensor Imaging data

D. CNV data - call and visualize

E. Send data or request a secure transfer-link from [enigmacnvhelpdesk@gmail.com](mailto:enigmacnvhelpdesk@gmail.com) (preferred) or [i.e.sonderby@medisin.uio.no](mailto:i.e.sonderby@medisin.uio.no)

Note: All files including this instruction and helper-files are available at: [https://github.com/ENIGMA-git/ENIGMA-CNV](https://github.com/ENIGMA-git/ENIGMA-CNV" \t "_blank)

Note: The ENIGMA-CNV working group is happy to help both with CNV calling and visualization as well as imaging processing if provided with raw data. Please contact us.



# A. Administrative - sign up for ENIGMA-CNV

**Please sign up for the ENIGMA-CNV e-mailing list to:**

[enigmacnvhelpdesk@gmail.com](mailto:enigmacnvhelpdesk@gmail.com) and [i.e.sonderby@medisin.uio.no](mailto:i.e.sonderby@medisin.uio.no).

**Indicate: (**google forms: https://forms.gle/dbSEAg9mSWzEar3V7**)**

-cohort name

-names of individuals involved, e-mail address, role [PI, main contact point, analysis of either imaging/CNV-data].

# B. Structural imaging data



Time estimate:

Please follow the protocol “ENIGMA-CNV, sMRI protocol, v2.0”.

# C. Diffusion Tensor Imaging data

Time estimate:

Work in progress

# D. CNVs - calling and visualization

*Overview:*



This part allows you to call raw CNVs and filter them with the commonly used CNV caller, PennCNV. Likewise, all CNVs of Interest as well as CNVs >50 kb are visualized.

We advice that you skim through the entire protocol before starting to get an overview/not do repeated work.

Please address any questions to: [enigmacnvhelpdesk@gmail.com](mailto:enigmacnvhelpdesk@gmail.com)

## 1. Organize your data and identify appropriate files for CNV calling

Create a folder called “ENIGMA-CNV\_Analysis/” where you wish your output and software for the ENIGMA-CNV protocol to be. Henceforward referred to as “the Analysis-folder”.

Please place all generated files in the Analysis-folder.

### a. Genetic information files

CNVs are called based on data from genotyping chips:

Illumina: The Illumina Final Report or LRR/BAF-files

Affymetrix: CEL-files or LRR/BAF-files

#### Illumina - *IlluminaFinalReport.txt/LRR-BAF-files*

-the Illumina Final report file of your Dataset/samples with, as a minimum:

SNP Name Sample ID B Allele Freq Log R Ratio.

Note: Allele 1 - Forward (or Allele 1 - Top), Allele 2 - Forward (or Allele 2 - Top), X & Y are requested by e.g. the Psychiatric Genetics consortium for use in the iPattern CNV caller. They are not used for this protocol but may be relevant for other CNV endeavors. Likewise, idat files (for loading in Genomestudio) may become handy if wanting to call CNVs with a different algorithm than PennCNV.

The Illumina report file can be generated directly from the BeadStudio project files, go to GenomeStudio, click the "Analysis" menu, select "Report", then select "Final Report", then make sure to drag the "Log R Ratio" and "B Allele Freq" field from the "Available Fields" to "Displayed Fields" so that these two signal intensity measures are exported to the final report file. Removing all other fields like GType, GC score, X Raw etc from the "Displayed Fields" can speed up the process and in addition decreases file size.

*Format of Illumina final report file:*

[Header]

BSGT Version 3.2.23

Processing Date 10/31/2008 11:42 AM

Content sample.bpm

Num SNPs 45707

Total SNPs 45707

Num Samples 48

Total Samples 192

[Data]

**SNP Name Sample ID B Allele Freq Log R Ratio**

Snp10000 KS2231000715 1.0000 -0.0558

Snp20000 KS2231000715 1.0000 -0.0422

Etc…

OBS – file may not contain the first 8 lines

Note – a previously extracted Illumina Final Report with (as minimum) the mentioned columns is just as good.

Please place the IlluminaFinalReport.txt-file in the Analysis-folder.

Likewise, LRR-BAF files produced with the PennCNV command ‘split\_illumina\_report.pl’ can be used directly. This saves time and computation (see the script and adapt). Please note that if these files are present in several subfolders, the script might run into issues – please contact the helpdesk for help if this is the case.

*Format of (tab-delimited) LRR-BAF file [Name: SampleID]:*

**Name NORMENT1.Log R Ratio NORMENT1.B Allele Freq**

SNP1 -0.0038 0.0161

SNP2 0.0073 0.9943

rs10000023 -0.0307 0.0026

etc…

#### Genetic information files – Affymetrix CEL files

Affymetrix CEL files acquires an additional step for conversion to LRR/BAF-files.

Already generated LRR/BAF-files can be used directly.

If you need to convert Affymetrix CEL files to LRR-BAF-files, perform step 1 of the protocol at: <http://penncnv.openbioinformatics.org/en/latest/user-guide/affy/>. ENIGMA-CNV does provide a ‘helper script’ (*ENIGMA-CNV\_AffyPrep\_Protocol\_v2.sh*) for performing this step – please contact the helpdesk.

#### SNP Position File

For generation of the PFB-file (population frequency file, see later), you need to provide a SNP-position file.

The SNP-position-file is a tab-delimited file with the positions of the SNPs on your chip, containing at least these columns:

**Name Chromosome Position**

rs1000000 12 126890980

rs1000002 3 183635768

rs10000023 4 95733906

These are, for instance, present in the SNP-map-file (.map) generated together with your IlluminaReport or the Illumina manifest file [.bpm]. Note that the header needs to have the exact headlines above to be used.

The PFB-file is specific to genome-version. First wave of ENIGMA-CNV CNVs was based on the hg18 genome build. Most genotyping chips are now released with hg19 or hg38 coordinates. In order of preference, we now aim for: hg38, hg19 and hg18.

You need to take note of the genome version of this file as this info is needed for CNV calling.

Please place the SNPPosition-file in the Analysis-folder.

### b. Cohort-generated files

#### Sex - *SexFile.txt*

-please produce a tab-delimited file specifying SampleID (as in the lluminaFinalReport-file) and sex in the format:

**SampleID sex**

SampleID1 female

SampleID2 male

etc…

[please note that sex MUST be in the format "female" or "male"; if missing sex-info, just leave the individual out of the list].

If sex is missing for some individuals, please note the number down in the CNV calling script [see later].

Place the file in the Analysis-folder.

#### Individuals to remove - *RemoveFile.txt*

Please produce a text-file with the individuals you want to remove (if any) in the analysis.

Each row has the SampleID [and ONLY this ID] specified as in the IlluminaReportFile

**SampleID**

SampleID1

SampleID2

The RemoveFile.txt should contain individuals that you know are sample mixups or subjects that for instance withdrew from the study so that these are kept out of the analysis. Please note that these are the ONLY initial exclusion measures we wish you to apply as some CNVs are so low in frequency that we want to be as inclusive as possible.

We request that you keep both duplicates and related individuals (and indicate these in the duplication/relatedness file below) since these will be helpful in QC as well as in analysis.

Place the file in the Analysis-folder

#### Individuals to keep - *KeepFile.txt*

Needed if you have files without imaging in your IlluminaFinalReport-file (or for others reasons just want to keep some individuals in the analysis)

-produce a text-file with the individuals you want to keep in the analysis

Each row has the SampleID [and ONLY this ID] specified as in the IlluminaReportFile:

**SampleID**

SampleID1

SampleID2

Place the file in the Analysis-folder

#### Duplication/relatedness *DupsRelatives.txt*

Please produce a tab-delimited list of duplicate individuals and related individuals:

**SampleID1 SampleID2 PI\_HAT Relation**

SampleID10 SampleID24 0.5 Mother\_child

etc…

-where *SampleID1* and *SampleID2* are the related individuals, *PI\_HAT* the proportion of the genomic variation shared IBD and *Relation* (if known or suspected - e.g. duplicate, sibling, parent/child, uncle/nephew, grandparent/grandchild, halfsibling, cousin if known). If Pihat is unknown, keep Pihat as “\_“, but instead indicate Relation. Please only include PI\_HAT>0.125.

Pihat can be calculated with the plink –genome command (<https://zzz.bwh.harvard.edu/plink/ibdibs.shtml>).

Place the file in the Analysis-folder.

## 2. Download necessary software, files and scripts

Please place all scripts and files downloaded in the Analysis-folder.

### a. Software (independent of dataset)

#### a. Singularity software

To minimize the impact of differences in computer hardware and software between sites, we developed a containerized pipeline - all required software is included in this container and its runs independent of hardware.

You first need to install Singularity (sylabs.io)[[1]](#footnote-1) along with some required dependencies (if already installed on your system (v3.0 or later), proceed to b.).

Please follow the instructions on the page here:

<https://sylabs.io/guides/3.0/user-guide/installation.html>

to install Singularity version 3.0 or later (latest stable release at time of writing is v3.8.5.). If you need admin rights to install, please consult your IT representative.

Note that a Linux based operating system is required. For users without a Linux-system, there are a few alternatives (not recommended, see footnote[[2]](#footnote-2)). Likewise, if unable to install 3.0 due to an old machine, see footnote[[3]](#footnote-3).

Note, storage required to install:

Singularity: ~140MiB disk space (once compiled and installed)

Container: 1.3 GB

#### b. Singularity container (*enigma-cnv.sif*) – download and install

The software in the *enigma.si f[[4]](#footnote-4)* container is:

* PennCNV v1.0.5, used for CNV calling
* R 3.3.1 + relevant R packages including iPsychCNV for visualization

*Download with singularity already installed*

singularity build enigma-cnv.sif  docker://bayramalex/enigma-cnv:latest

*enigma-cnv.sif* (the name of the container) will be downloaded to your $PWD.

*Download without singularity installed*

(e.g. if you need to import into a system without internet-access and need to download first):

<https://github.com/comorment/containers/blob/main/singularity/enigma-cnv.sif>

Place the singularity container in the Analysis-folder.

### b. Scripts for running analysis

Please download scripts here:

Github…

ENIGMA\_CNV\_CNVProtocol\_v2.sh

Also the Rscript (ENIGMA-CNV\_visualize.R)…

compile\_pfb\_new.pl

### c. Genotyping-chip-dependent files

As mentioned previously, for genetic data overall, files are specific to genome-version. First wave of ENIGMA-CNV was based on the hg18 genome build. Most genotyping chips are now released with hg38 coordinates. In order of preference, we now aim for: hg38, hg19 and hg18.

In addition to the correct genome version, appropriate files according to your genotyping chip must be used:

-*The PFB-file* (population frequency of B-allele file) - supplies the PFB information for each marker, and can give the chromosome coordinate information to PennCNV for CNV calling.

-*The GCMODEL file* - specifies the GC content of the 1Mb genomic region surrounding each marker (500kb each side) (calculated using the UCSC GC annotation file).

*-the HMM-file* - tells the program what would be the expected signal intensity values for different copy number state, and what is the expected transition probability for different copy number states (For more information, see here: http://penncnv.openbioinformatics.org/en/latest/user-guide/input/)

#### Generating your own PFB-file

For cohorts with more than 300 individuals with good quality genotyping data (for a suggestion, see footnote[[5]](#footnote-5)), we advise that you create your own PFB-file and GC-model files – this procedure is part of the script.

If you have less than 300 individuals in your cohort, you need to use a generic PFB- and GCmodel file. Please confer with the ENIGMA-CNV working group with [enigmacnvhelpdesk@gmail.com](mailto:enigmacnvhelpdesk@gmail.com) for advice.

#### The HMM-file

There are several HMM-files:

*hhall.hmm* – can be used for all Illumina arrays (path: /opt/PennCNV-1.0.5/lib/) except:

*exome.hmm[[6]](#footnote-6)* – more appropriate for Illumina Infinium HumanExome BeadChip and Infinium HumanCoreExome-24

*affygw6.hmm* - Affymetrix Genome-Wide Human SNP Array 6.0 and SV (path: (/opt/PennCNV-1.0.5/affy/libgw6/)

*exomechip.hmm* needs to be downloaded from the ENIGMA-CNV github repository – the others are present in the enigma-cnv.sif-container.

#### CNV of interest file based on genotyping chip

We visualize a selected set of CNVs - both with set and non-set boundaries.

For this we need one of the following files:

*CNVsofInterest\_ENIGMA-CNV\_hg18.csv*

*CNVsofInterest\_ENIGMA-CNV\_hg19.csv*

*CNVsofInterest\_ENIGMA-CNV\_hg38.csv*

This list of CNVs is compiled from the paper from UK biobank: Kendall et al., 2017 (doi: 10.1016/j.biopsych.2016.08.014)[[7]](#footnote-7) with a few additions.

Download according to your genome version and put in your Analysis-folder.

Filtering genomic regions files

Several genomic regions are known to harbor spurious CNV calls that might represent cell-line artifacts.

Therefore, we remove CNVs in certain genomic regions overlapping >50%:

-centromeric regions

-telomeric regions

-segmental duplication regions

-immunoglobulin regions

# -repeat masked regions

You can find the corresponding files here:

*telo\_hg38.txt*

*immune\_hg38.txt*

*segmentaldups\_hg38.txt*

*centro\_hg38.txt*

Download according to your genome version and put in your Analysis-folder.

## 3. Call and visualize CNVs

Memory of ≥???GB is highly recommended.

OBS – note that if you have data from more than one type of chip, please run the protocol separately for each different chip set

Time estimate: ???

### a. Adjust the script

a. Please rename the script *ENIGMA\_CNV\_CNVProtocol\_v2.sh* to *${Dataset}\_ENIGMA\_CNV\_CNVProtocol\_v2.sh* [i.e. the name of your file with your dataset name [as stated below] as prefix]

b. edit the script by changing to your dataset-specific files as directed in section 0a [labeled “USER-INPUT needed] in the script.

d. Regarding deidentification – the protocol has an option to add an additional step of deidentification to the IDs – this leaves a key in the hands of the cohort submitting data. Adding this can give additional complications (if trying to track back to data). Please only tick this if required by your IRB-approvals.

### b. Run the analysis

Run the ${Dataset}\_ENIGMA\_CNV\_CNVProtocol\_v2.sh script in the terminal:

bash ./${Dataset}\_ENIGMA\_CNV\_CNVProtocol\_v2.sh

After the run is done, please check the checklist found in ${Dataset}\_checklist.txt (visualization folder) to see if things seem to make sense.

## 4. Return data to ENIGMA-CNV

Please send the compressed folder ${Dataset}\_visualize.tar.gz to [enigmacnvhelpdesk@gmail.com](mailto:enigmacnvhelpdesk@gmail.com) or request a transfer-link.

# E. Covariate files

Time estimate: variable

The covariate files are produced by manually constructing a covariate-file and obtaining ancestry covariates following the standard protocols of ENIGMA genetics.

## General covariates

Please create a csv-file called *${Dataset]\_Covar\_ENIGMACNV.csv*

The table will look like this:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SubjID** | **GeneticID** | **PID** | **MID** | **DiseaseType** | **AffectionStatus** | **Affectionstatus2** | **Age** | **Sex** | **ScannerSite** |
| Subj1 | A2035 | A1003 | A6008 | Population | 0 | Healthy\_bipolar | 30.0 | 1 | Oslo2 |
| Subj2 | D1000 | NA | NA | Population | 1 | Bipolar | 35.5 | 2 | Oslo2 |

(use excel or your favorite spreadsheet program saving as a csv-file)

**General NOTES on the columns**

* Missing values should be coded as NA.
* Please include individuals, even if they are missing one or two covariates (say AffectionStatus) – the data might still be valuable. Just fill out the missing values as NA.

**Specific NOTES on the columns**

* **SubjID**: Must match the SubjID in the imaging files.
* **GeneticID:** The ID used in the ENIGMA-CNV-calling protocol (“ENIGMA-CNV\_CNVCalling\_Protocol\_final.sh”) prior to de-identification. If this is the same as the SubjID, fill in the column accordingly.
* **Sex**: Must be coded as follows: Male=1, Female=2
* **DiseaseType**: Type of cohort. ENIGMA-CNV receives both population cohorts (or volunteer-cohorts) and disease-cohorts (e.g. epilepsy, psychiatric disease, dementia).
* For *population-studies,* please note the best fitting term (copy to all fields):
  + Twinstudy
  + Population
  + Volunteers
* For *case-controls-studies,* please note the best fitting term(copy to all fields):
  + E.g bipolar, schizophrenia, ADHD, autism, dementia, stroke, ???.
* **AffectionStatus**: a binary indicator variable where Patients = 1 & Controls = 0. If your cohort does not have patients, code = 0. If this value is missing, the individual will be removed from downstream analysis (based on the Rscript). Thus, make sure you fill this in!
* **AffectionStatus2**:
* For *population-studies*: please code as NA in all columns
* For *case-controls-studies*:
  + Please code patients with
    - a standardized version of the diagnosis per individual such as DSM or ICD (We are particularly interested in ‘Mental, behavioural and neurodevelopmental disorders’, corresponding to ICD-10 F01-99 and ‘Diseases of the nervous system, corresponding to ICD-10 G00-G99.
    - If lacking such detailed information, SCZ (schizophrenia), BD (bipolar disorder), ADHD, dementia or the likes will suffice.
  + Please code controls as:
    - **Healthy\_psych, Healthy\_epilepsy**, **Healthy\_dementia** (controls have been ascertained/screened for lack of e.g. psychiatric diseases, epilepsy or dementia)
* **ScannerSite:** The name of your scanner site.

In parallel, please fill out *ScannerInfoSheet.xlxs***.**

Information required each scannersite from your cohort (IF the same site used different field strengths, these must be named as two different sites (e.g. Oslo\_T1.5, Oslo\_T3)):

ScannerSite, Sequence, FieldStrength, AcquisitionDirection, NumberOfSlices, SliceGap, VoxelSize\_m3, TI, TE, TR, FlipAngle, Citation, Segmentation

OBS: Make sure names correlate in *ScannerInfoSheet.xlxs* and *Covar\_ENIGMACNV.csv.*

If you have previously run the ENIGMA genetics association protocol, you can find the info on sex and age in: *SubCortCovs\_nopatients.csv* [methodA], *SubCortCovs\_havepatients.csv* [method B], *SubCortCovs\_related\_nopatients.csv* [method C]

## MDS-covariates

MDS-covariate-file: *HM3\_b37mds2R.mds.csv* (version July 13, 2017), alternatively *HM3mds2R.mds.csv (*version July 27, 2012).

This file contains covariates for the ancestry of each subject in your cohort.

If you previously participated in ENIGMA-genetics (and followed the protocols of ENIGMA), you have this file already.

If not, go to the ENIGMA-imaging-protocol-homepage:

[**http://enigma.ini.usc.edu/protocols/genetics-protocols/**](http://enigma.ini.usc.edu/protocols/genetics-protocols/):

**Text

Description automatically generated**

Please download:

“Imputation Protocol (PDF) – Updated (July 13, 2017)”.

Or use this link:

<https://enigma.ini.usc.edu/wp-content/uploads/2020/02/ENIGMA-1KGP_p3v5-Cookbook_20170713.pdf>

OBS: To obtain MDS-covariates, you only need to run the “Multi-dimensional Scaling (MDS) Protocol” part of the code in the ENIGMA imputation protocol, NOT the imputation part.

The output file, *HM3mds2R.mds.csv*, is a spreadsheet containing the following columns:

Family ID (FID), individual ID (IID), 4 MDS components (C1, C2, C3 and C4), and PLINK’s assigned solution code (SOL).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FID | IID | SOL | C1 | C2 | C3 | C4 |
| Fam1 | Subj1 | ... | ... | ... | ... | ... |
| Fam2 | Subj2 | ... | ... | ... | ... | ... |

OBS: Please confirm that either the FID OR the IID in MDS file match the SubjID in the imaging files.

# F. Send data

Files to transfer:

*${Dataset}\_visualize.tar.gz* [CNV-calling data]

*${Dataset}\_sMRI.tar.gz* [sMRI calling data]

*HM3\_b37mds2R.mds.csv* [MDS-data]

*${Dataset]\_Covar\_ENIGMACNV.csv* [covariates]

Script for joining imaging and covariate files:

-the script will print the dimensions of your input-files and the final output-file.

Please send the data to [enigmacnvhelpdesk@gmail.com](mailto:enigmacnvhelpdesk@gmail.com) or request a secure transfer-link.

The data will be stored at the secure server TSD (<https://www.uio.no/english/services/it/research/sensitive-data/>) in Oslo, Norway. Only registered users have access.

# Notes for cohorts that ran the ‘old’ protocol and should visualize more CNVs

**PennCNV, ipsych and dockers**

Files:

-instruction files

-script files

-HMM/PFB/GCMODEL-files

-example files of how illumina report/final file looks..

-video tutorial

-choice of HMM/PFB/etc

-hg18/hg19/hg38 files

The pre-compiled container: <https://github.com/comorment/containers/blob/main/singularity/enigma-cnv.sif>

1. Singularity is an open-source program that allows the ‘containerization’ of software packages, their dependences and libraries to build entire pipelines that are cross-platform compatible. [↑](#footnote-ref-1)
2. A less (non-)supported MacOS/Windows solution does exist here: <https://sylabs.io/guides/3.7/admin-guide/installation.html#installation-on-windows-or-mac>. Please note that we do not recommend the latter - we did not test this and it may not be suitable to run all analysis defined. Alternatively, we recommend installing Linux via dual-boot or as a virtual machine. [↑](#footnote-ref-2)
3. On some older systems and kernels, Singularity v3.x may not work. If that is the case you can install Singularity v2 – however, please note that the container has not been tested with V2 although it may work. [↑](#footnote-ref-3)
4. If you wish to modify/expand the container, the docker can be found here: <https://github.com/comorment/gwas/blob/main/containers/enigma-cnv/Dockerfile> [↑](#footnote-ref-4)
5. We suggest to follow the guidelines of the ENIGMA genetics imputation protocol, i.e filter out individuals genotype call rate <0.95 (and SNPs with Minor Allele Frequency < 0.01 and Hardy-­‐Weinberg Equilibrium < 1x10-­‐6).

   This can be done with plink, e.g.: plink -­‐-­‐bfile $datafileraw -­‐-­‐hwe 1e-­‐6 -­‐-­‐geno 0.05 -­‐-­‐maf 0.01 -­‐-­‐noweb -­‐-­‐make-­‐bed -­‐-­‐out ${datafileraw}\_filtered). For reference, please see: the ENIGMA genetics imputation protocol, <http://enigma.ini.usc.edu/protocols/genetics-protocols/>. [↑](#footnote-ref-5)
6. Derived from this article: Szatkiewicz et al (2013): Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample. Mol Psychiatry, 18(11):1178-84] [↑](#footnote-ref-6)
7. Originally in hg19 – lifted over to hg18 and hg38 with UCSC liftover tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). [↑](#footnote-ref-7)