ENIGMA & Large Scale Imaging Association

Jason L. Stein
Geschwind Laboratory /
Imaging Genetics Center, Laboratory of Neuro Imaging
University of California, Los Angeles
steinja@ucla.edu

June 10, 2012
Organization for Human Brain Mapping
Introduction to Imaging Genetics Educational Course
Beijing, China
Imaging Genomics
Meta-Analysis

- The need for large scale meta-analysis
- Step 1: Set up the Consortium
- Step 2: Receiving the data
- Step 3: Filtering data
- Step 4: Processing Data
- Step 5: Quality Checking
- Step 6: Meta-analysis
- Conclusions
Imaging Genomics
Meta-Analysis

• The need for large scale meta-analysis
• Step 1: Set up the Consortium
• Step 2: Receiving the data
• Step 3: Filtering data
• Step 4: Processing Data
• Step 5: Quality Checking
• Step 6: Meta-analysis
• Conclusions
Finding Genetic Variants Influencing a Quantitative Trait

Is there a relationship between genotype and phenotype?
Genome-wide association study

An unbiased search to find where in the genome a common variant is associated with a trait.
The Powerlessness of the Individual Researcher

When looking at genetic variants, there are (unfortunately!) small effects and many tests conducted. This means you need huge sample sizes to significantly detect variants.

However, both imaging and genotyping are extraordinarily expensive (compared to disease diagnosis and family history), so a consortium is needed.

E. Nigma
Imaging Genomics
Meta-Analysis

- The need for large scale meta-analysis
- Step 1: Set up the Consortium
- Step 2: Receiving the data
- Step 3: Filtering data
- Step 4: Processing Data
- Step 6: Quality Checking
- Step 7: Meta-analysis
- Conclusions
<table>
<thead>
<tr>
<th>Options for how to run a consortium</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mega-analysis</strong></td>
</tr>
<tr>
<td>-Raw phenotype and genotype data are uploaded for each subject to a central server</td>
</tr>
<tr>
<td><strong>Meta-analysis</strong></td>
</tr>
<tr>
<td>- Collaboration is necessary for significant results, but sharing data is difficult because of IRB and ownership issues.</td>
</tr>
<tr>
<td>- ENIGMA is a data free results filled network, we get uploads of statistical results – no raw data</td>
</tr>
</tbody>
</table>
Advantages of each

Mega-analysis

- Ability to check every part of the analysis
- Moves as fast as you make it (no waiting)
- Greater analysis possibilities (polygenic score, structural equation modeling)

Meta-analysis

- (We think) more groups are willing to share results than data
- Distributed approach
Replication through collaboration

http://enigma.loni.ucla.edu

Enhancing Neuro Imaging Genetics through Meta-Analysis

The ENIGMA Network brings together researchers in imaging genomics, to understand brain structure and function, based on MRI, DTI, fMRI and genomewide association scan (GWAS) data. The ENIGMA Network has several goals:

- To create a network of like-minded individuals, interested in pushing forward the field of imaging genetics.
- To ensure promising findings are replicated via member collaborations, in order to satisfy the mandates of most journals.
- To share ideas, algorithms, data, and information on research studies and methods.
- To facilitate training, including workshops and conferences on key methods and emerging directions in imaging genetics.

"Data sharing with other members of the ENIGMA Network is optional and by no means a requirement of joining the network."

Paul Thompson
Laboratory of Neuro Imaging, UCLA

Nick Martin
Genetic Epidemiology Unit, Queensland Institute of Medical Research

> 200 scientist members from 12 countries all over the world!

Imaging Segmentation, Imputation, and Association protocols provided at our website.
Project Description

**Imaging Protocols**
- T₁-weighted structural
- Hippocampal segmentation
- Tissue segmentation to calculate brain volume (white matter + grey matter)
- Registration to template to calculate ICV

**Genetics Protocols**
- Genome-wide genotypes imputed to 1,387,466 autosomal single nucleotide polymorphisms (SNPs) based on HapMap III reference panels

**Analytical Steps**
- Genome-wide association to imaging phenotypes using dosage data (accounting for kinship in related samples)

Completed at the level of the individual site.
Project Description

Complete

Uploaded to ENIGMA server for analysis at central site

17 sites uploaded (N=7795)

Quality Checking and Filtering (MAF < 0.01, R² < 0.3)

Fixed effects meta-analysis
Random effects meta-analysis

Phenotypes
- Hippocampal Volume
- Brain Volume
- ICV

Covariates
- Brain Volume
- ICV
- Age
- Sex
- Age²
- Sex*Age
- Sex*Age²
- 4 MDS components
- Dummy covariates for acquisitions

Genome-wide association to imaging phenotypes using dosage data (accounting for kinship in related samples)

Healthy Only

17 sites uploaded (N=5776)

Quality Checking and Filtering (MAF < 0.01, R² < 0.3)

Fixed effects meta-analysis
Random effects meta-analysis
Imaging Genomics
Meta-Analysis

• The need for large scale meta-analysis
• Step 1: Set up the Consortium
• Step 2: Receiving the data
• Step 3: Filtering data
• Step 4: Processing Data
• Step 6: Quality Checking
• Step 7: Meta-analysis
• Conclusions
Goal is to get as much information at the beginning as you possibly can...
ENIGMA Automated Submission System

<table>
<thead>
<tr>
<th>HapMap Count Post Filter</th>
<th>210,737</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Uploaded SNPs</td>
<td>1,387,488</td>
</tr>
<tr>
<td>RSQR Failed</td>
<td>170,058</td>
</tr>
<tr>
<td>P Val Failed</td>
<td>0</td>
</tr>
<tr>
<td>Freq Failed</td>
<td>40,701</td>
</tr>
<tr>
<td>Total SNPs For Analysis</td>
<td>1,176,729</td>
</tr>
</tbody>
</table>

### Stage 1

**File Sample:**

| 1 | rs3131967 | T | C | 0.110 | HippoVol | -123.792 | 164.682 | 2.509 | 0.123 | 0.4522 | 0.1099 |
| 1 | rs1048488 | C | T | 0.160 | HippoVol | -118.594 | 138.708 | 3.160 | 0.159 | 0.3926 | 0.1598 |
| 1 | rs12562034 | G | A | 0.898 | HippoVol | -13.048 | 36.165 | 0.026 | 0.028 | 0.7183 | 0.8979 |
| 1 | rs12124819 | A | G | 0.788 | HippoVol | 5.057 | 99.895 | 0.007 | 0.001 | 0.9596 | 0.7885 |
| 1 | rs404617 | G | A | 0.114 | HippoVol | -122.766 | 158.387 | 2.538 | 0.130 | 0.4383 | 0.1135 |
| 1 | rs2505036 | T | Z | 1.000 | HippoVol | -0.000 | 100000.000 | 0.000 | 0.000 | 1.0000 |

### Stage 2

**File Sample:**

| rs11778460 | C | 39.631 | G | 82.726 | 0.912 | 0.03685 |
| rs333277 | A | 22.477 | G | -13.554 | 0.629 | 0.5465 |
| rs10806671 | T | 22.523 | C | 22.571 | 0.410 | 0.3163 |
| rs579035 | T | 23.847 | G | 4.090 | 0.650 | 0.8838 |
| rs7173425 | C | 40.506 | T | 18.975 | 0.917 | 0.6395 |
| rs2239669 | G | 25.125 | A | 11.094 | 0.587 | 0.6588 |
| rs734883 | G | 21.425 | T | 3.442 | 0.483 | 0.6563 |
| rs10459518 | G | 21.100 | A | 0.834 | 0.8337 | 0.103 | 0.016 | 0.4018 | 0.3980 | 0.2867 | 0.6735 | 0.9986 |
| rs5980169 | T | 3.641 | 0.409 | 0.4980 | 0.005 | 0.005 | 0.8716 | 0.9984 | 2.531 | 0.9984 |

### Stage 3

**File Sample:**

SNP EFFECT, ALLELE, NON_EFFECT, ALLELE, BETA, FREQ, P_VAL
rs11778460 | C | 39.631 | G | 82.726 | 0.912 | 0.03685 |
rs333277 | A | 22.477 | G | -13.554 | 0.629 | 0.5465 |
rs10806671 | T | 22.523 | C | 22.571 | 0.410 | 0.3163 |
rs579035 | T | 23.847 | G | 4.090 | 0.650 | 0.8838 |
rs7173425 | C | 40.506 | T | 18.975 | 0.917 | 0.6395 |
rs2239669 | G | 25.125 | A | 11.094 | 0.587 | 0.6588 |
rs734883 | G | 21.425 | T | 3.442 | 0.483 | 0.6563 |
rs10459518 | G | 21.100 | A | 0.834 | 0.8337 | 0.103 | 0.016 | 0.4018 | 0.3980 | 0.2867 | 0.6735 | 0.9986 |
rs5980169 | T | 3.641 | 0.409 | 0.4980 | 0.005 | 0.005 | 0.8716 | 0.9984 | 2.531 | 0.9984 |
Imaging Genomics
Meta-Analysis

- The need for large scale meta-analysis
- Step 1: Set up the Consortium
- Step 2: Receiving the data
- Step 3: Filtering data
- Step 4: Processing Data
- Step 5: Quality Checking
- Step 6: Meta-analysis
- Conclusions
Filtering by Quality of Imputation

- **RSQR_HAT**: Estimates the squared correlation between imputed and true genotypes.
- Typically, a cut-off of 0.30 will flag most of the poorly imputed SNPs but only a small number (<1%) of well imputed SNPs

*The true genotype is generally unknown so how do you get this?*
Take the ratio of the empirically observed variance of the allele dosage (from imputed results) to the expected binomial variance of HWE calculated by allele frequency from HapMap

(Li et al., 2009)
How to filter based on imputation quality using mach2qtl output

```
% awk '{if (NR>32 && $5>=0.3) print $1,$2,$3,$4,$5,$6,$7,$8,$9,$10}' chr3-mach2qtl.out > chr3-rsqrfilt.out
```
Filter by low Minor Allele Frequency (MAF)

Associations with low MAF can be driven by outliers in the data. It is common to exclude SNPs with MAF < 0.01.
How to filter based on frequency using mach2qtl output

```awk
awk '{if (NR>32 && ($4 >= 0.01 && $4 <= 0.99) ) print $1,$2,$3,$4,$5,$6,$7,$8,$9,$10}' chr3-mach2qtl.out > chr3-freqfilt.out
```
Filtering P-values

P-value = NA
P-value = -1

Incorrect: Perfect collinearity with intercept
Correct: Remove one dummy covariate
Imaging Genomics Meta-Analysis

• The need for large scale meta-analysis
• Step 1: Set up the Consortium
• Step 2: Receiving the data
• Step 3: Filtering data
• Step 4: Processing Data
• Step 5: Quality Checking
• Step 6: Meta-analysis
• Conclusions
The effect allele is crucial in meta-analysis. Depending on how the association is coded, you can get the same P-value, but different direction of effect in meta-analysis.
## Assignment of Effect Allele

<table>
<thead>
<tr>
<th>Program</th>
<th>Effect Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>mach2qtl</td>
<td>A2</td>
</tr>
<tr>
<td>Plink</td>
<td>A1</td>
</tr>
<tr>
<td>Merlin-Offline</td>
<td>A1</td>
</tr>
<tr>
<td>ProbABEL</td>
<td>A1</td>
</tr>
<tr>
<td>snptest</td>
<td>allele_B</td>
</tr>
<tr>
<td>Quicktest</td>
<td>alleleB</td>
</tr>
<tr>
<td>Solar</td>
<td>A1</td>
</tr>
</tbody>
</table>

Effect allele varies by association and imputation program. It is not always an easy or readily apparent thing to figure out either so you have to delve into the manuals!
### Frequency Allele

Effect allele is not the same as the frequency allele in mach2qt!

A simple $1 - \text{FREQ1}$ gives you the frequency of the effect allele
## Assignment of Frequency Allele

<table>
<thead>
<tr>
<th>Program</th>
<th>Effect Allele</th>
<th>Frequency Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>mach2qtl</td>
<td>A2</td>
<td>A1</td>
</tr>
<tr>
<td>Plink</td>
<td>A1</td>
<td>A1</td>
</tr>
<tr>
<td>Merlin-Offline</td>
<td>A1</td>
<td>A1</td>
</tr>
<tr>
<td>ProbABEL</td>
<td>A1</td>
<td>A1</td>
</tr>
<tr>
<td>snptest</td>
<td>allele_B</td>
<td>Genotype counts given</td>
</tr>
<tr>
<td>Quicktest</td>
<td>alleleB</td>
<td>Genotype counts given</td>
</tr>
<tr>
<td>Solar</td>
<td>A1</td>
<td>A1</td>
</tr>
</tbody>
</table>

Frequency allele varies by association and imputation program. It is not always an easy or readily apparent thing to figure out either so you have to delve into the manuals!
Correct scale for beta / SE

Units will affect beta values and standard errors in regression so need to be very careful to know what units analyses were run in. You can use simple scaling factors if you know the conversion:

\[
\beta_{\text{mm}} = \beta_{\text{cm}} \times 1000
\]

\[
\text{SE}_{\text{mm}} = \text{SE}_{\text{cm}} \times 1000
\]
When Scaling Goes Wrong

- Forest Plots can show something is off in units relative to other groups.

- Assuming similar effect sizes and standard errors between groups which should be a safe assumption.
A side note on family based samples

Calculation of association statistics in family based samples is generally done through a minimization algorithm (see SOLAR or merlin-offline) which generally fails with large phenotype values.

You can use this Beta/SE trick in your favor in this case!

Divide the phenotype values by a constant prior to association, then multiply the Beta/SE values by the constant after association so meta-analysis is in the same scale.
HapMap3 Specific Troubles

SNPs with prefix of AFFX have no white space between alleles and AFFX in mach2qtl (messes up parsing)

```
sed 's/A,/A,/g; s/C,/C,/g; s/T,/T,/g; s/G,/G,/g' chr3-mach2qtl.out > chr3-AFFXfix.out
```
Imaging Genomics
Meta-Analysis

- The need for large scale meta-analysis
- Step 1: Set up the Consortium
- Step 2: Receiving the data
- Step 3: Filtering data
- Step 4: Processing Data
- Step 5: Quality Checking
- Step 6: Meta-analysis
- Conclusions
Frequency Differences Histograms Relative to CEU

Histograms centered around zero with little variance imply that imputation and genotyping were done well (not different from the reference HapMap sample). Also will tell you if frequency is assigned to the correct allele!

Here you can see that all samples are centered around zero which is evidence of good imputation and genotyping. Samples of different ethnicity (GOBS) have greater variance.
When Frequency Difference Plots go Wrong

Generally indicative of poor imputation or poor genotyping.
Right Hippocampus Phenotype

Histograms are a great way to visualize your data and a way to identify poorly segmented outliers.
Neurologically normal adults (Weis et al., 1989; Watson et al., 1992; Jack et al., 2000) and children (Giedd et al., 1996; Pfluger et al., 1999; Utsunomiya et al., 1999) have significantly larger right hippocampi. Therefore the mean L/R ratio should generally be less than 1.
Slice by Slice QC of Hippocampal Segmentations in Sagittal View

Good Segmentation
Common Errors: Missing Large Lateral Sections of Hippocampus
Common Errors: Missing Anterior Portions of Hippocampus
Common Errors: Extension into amygdala
Less Common Error: Segmentation Completely Off
Subjects with Well Delineated but Small Hippocampi

Right Hippocampus

Hippocampal Volume
Mean: 3602.24173628975
Subjects with Good but Big Hippocampi
Subject with good segmentations
large L/R Ratio

L/R Ratio = 1.525996
Poor registration gives poor estimates of the scaling used to determine estimated total intracranial volume.
Individual Site QQ Plots

Expected -log10(P-value)

Observed -log10(P-value)

lambda=1.002
When QQ Plots Go Wrong

Can show evidence of unaccounted for population stratification, cryptic relatedness, or just that your data does not follow expected distributions
The observed distribution only deviates from the expected at low P-values. Would not expect something like this without huge effect sizes or huge sample sizes.
Individual Site Manhattans
When Manhattan Plots Go Wrong

Can be evidence that imputation failed on one chromosome or that somebody just typed something incorrectly when running association
Population Stratification

Instead of a few hits, most everything is $P < 1e-5$

Lambda should be near 1
Other Interesting Data

Something was run incorrectly

Expected $-\log_{10}(P\text{-value})$
$\lambda = 0.07397$
MDS Plots

Can show you which HapMap population best describes your sample

Good for finding and removing ancestry outliers
Imaging Genomics
Meta-Analysis

- The need for large scale meta-analysis
- Step 1: Set up the Consortium
- Step 2: Receiving the data
- Step 3: Filtering data
- Step 4: Processing Data
- Step 5: Quality Checking
- Step 6: Meta-analysis
- Conclusions
Fixed Effects MA Description

Fixed effects assume that the genetic effects are the same across the combined investigations and all differences are due to chance.

\[ z_{\text{meta}} = \sum_i z_i \times w_i \]

\[ w_i = \sqrt{\frac{N_i}{N_{\text{total}}}} \]

P-value based meta-analysis
(Weighted sum of Z-scores)

the units of the beta coefficients and standard errors need not be the same across studies
Fixed Effects MA Description

\[
\langle \beta \rangle = \frac{\sum_i [\beta_i/(SE_i)^2]}{\sum_i [1/(SE_i)^2]},
\]

\[
\langle SE \rangle = \sqrt{\frac{1}{\sum_i [1/(SE_i)^2]}}
\]

\[
z_{\text{meta}} = \frac{\langle \beta \rangle}{\langle SE \rangle}
\]

pooled inverse variance weighted

the units of the beta coefficients and standard errors **must be** the same across studies
How to run METAL: Stage 1

http://genome.sph.umich.edu/wiki/METAL_Documentation

<table>
<thead>
<tr>
<th>SNP</th>
<th>NON_EFFECT_ALLELE</th>
<th>EFFECT_ALLELE</th>
<th>BETA</th>
<th>FREQ</th>
<th>N</th>
<th>P_VAL</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs9586302</td>
<td>C</td>
<td>A</td>
<td>24.297</td>
<td>0.0501</td>
<td>550</td>
<td>0.5288</td>
<td>38.575</td>
</tr>
<tr>
<td>rs354417</td>
<td>T</td>
<td>G</td>
<td>8.233</td>
<td>0.815</td>
<td>550</td>
<td>0.7102</td>
<td>22.154</td>
</tr>
<tr>
<td>rs1927207</td>
<td>T</td>
<td>C</td>
<td>10.124</td>
<td>0.0827</td>
<td>550</td>
<td>0.7388</td>
<td>30.361</td>
</tr>
<tr>
<td>rs9582391</td>
<td>A</td>
<td>C</td>
<td>-22.281</td>
<td>0.2143</td>
<td>550</td>
<td>0.2723</td>
<td>20.296</td>
</tr>
<tr>
<td>rs507529</td>
<td>A</td>
<td>G</td>
<td>-26.772</td>
<td>0.4256</td>
<td>550</td>
<td>0.1227</td>
<td>17.343</td>
</tr>
<tr>
<td>rs7490444</td>
<td>T</td>
<td>C</td>
<td>90.795</td>
<td>0.1035</td>
<td>550</td>
<td>0.005564</td>
<td>32.75</td>
</tr>
<tr>
<td>rs12584241</td>
<td>T</td>
<td>C</td>
<td>-54.848</td>
<td>0.0876</td>
<td>550</td>
<td>0.06505</td>
<td>29.729</td>
</tr>
</tbody>
</table>

Make sure all data are in the same format as white space delimited text files. Pay special attention to making sure effect_allele is correct!
How to run METAL: Stage 2

Minimum sample size needed to calculate meta-analysis

/home/enigma/allsites.metal
How to run METAL: Stage 2

Can specify P-value based meta-analysis or inverse standard error weighted meta-analysis here

```
MARKERLABEL  SNP
EFFECTLABEL  BETA
WEIGHTLABEL  N
PVALUELABEL  P_VAL
MINWEIGHT    1000
SCHEME       SAMPLESIZE
AVERAGEFREQ  ON
MINMAXFREQ   ON
FREQLABEL    FREQ
GENOMICCONTROL ON
OUTFILE      MetaOutput .tbl

ALLELELABELS  EFFECT_ALLELE NON_EFFECT_ALLELE

PROCESS /home/enigma/site1.tbl  #Site 1
PROCESS /home/enigma/site2.tbl  #Site 2
PROCESS /home/enigma/site3.tbl  #Site 3
PROCESS /home/enigma/site4.tbl  #Site 4

ANALYZE       HETEROGENEITY
QUIT

/home/enigma/allsites.metal
```
How to run METAL: Stage 2

Adjust the test statistics at the individual site level using the lambda factor

```bash
MARKERLABEL  SNP
EFFECTLABEL  BETA
WEIGHTLABEL  N
PVALUENAME   P.VAL
MINWEIGHT    1000
SCHEME       SAMPLESIZE
AVERAGEFREQ  ON
MINMAXFREQ   ON
FREQLABEL    FREQ
GENOMICCONTROL  ON
OUTFILE      MetaOutput .tbl

ALLELELABELS  EFFECT_ALLELE NON_EFFECT_ALLELE

PROCESS /home/enigma/site1.tbl  #Site 1
PROCESS /home/enigma/site2.tbl  #Site 2
PROCESS /home/enigma/site3.tbl  #Site 3
PROCESS /home/enigma/site4.tbl  #Site 4

ANALYZE       HETEROGENEITY
QUIT

/home/enigma/allsites.metal
```
How to run METAL: Stage 2

```
MARKERLABEL   SNP
EFFECTLABEL   BETA
WEIGHTLABEL   N
PVALUELABEL   P_VAL
MINWEIGHT     1000
SCHEME        SAMPLESIZE
AVERAGEFREQ   ON
MINMAXFREQ    ON
FREQLABEL     FREQ
GENOMICCONTROL ON
OUTFILE       MetaOutput .tbl

ALLELELABELS  EFFECT_ALLELE NON_EFFECT_ALLELE

PROCESS /home/enigma/site1.tbl  #Site 1
PROCESS /home/enigma/site2.tbl  #Site 2
PROCESS /home/enigma/site3.tbl  #Site 3
PROCESS /home/enigma/site4.tbl  #Site 4

ANALYZE       HETEROGENEITY
QUIT
```

/home/enigma/allsites.metal
metal /home/enigma/allsites.metal > /home/enigma/allsites.metal.log

Specify the files to process
The Final Product

Each SNP P-value is the combined evidence from all contributing studies.
Visualization of Results

Manhattan plot showing all SNPs with p<0.001

http://enigma.loni.ucla.edu/enigma-vis/
Practical aspects of imputation-driven meta-analysis of genome-wide association studies

Paul I.W. de Bakker\textsuperscript{1,2,*}, Manuel A.R. Ferreira\textsuperscript{2,3,†}, Xiaoming Jia\textsuperscript{4}, Benjamin M. Neale\textsuperscript{2,3}, Soumya Raychaudhuri\textsuperscript{2,3,5} and Benjamin F. Voight\textsuperscript{2,3}

Identification of common variants associated with human hippocampal and intracranial volumes
Conclusions

• Consortia are important because **effect sizes are small**
• However, the more data you combine with different groups the more **quality checking and filtering** matters
• Clear demonstration of **Murphy’s Law**
• Did I mention quality checking?
Acknowledgements


207 co-authors

Paul Thompson
UCLA

Rudy Senstad
UCLA

Derrek Hibar
UCLA

Sarah Medland
Queensland Institute of Medical Research, Australia

Alejandro Arias Vasquez
Radboud University Nijmegen, Nijmegen, The Netherlands