IGC

Imaging Genetics Center

Searching for Common Variants

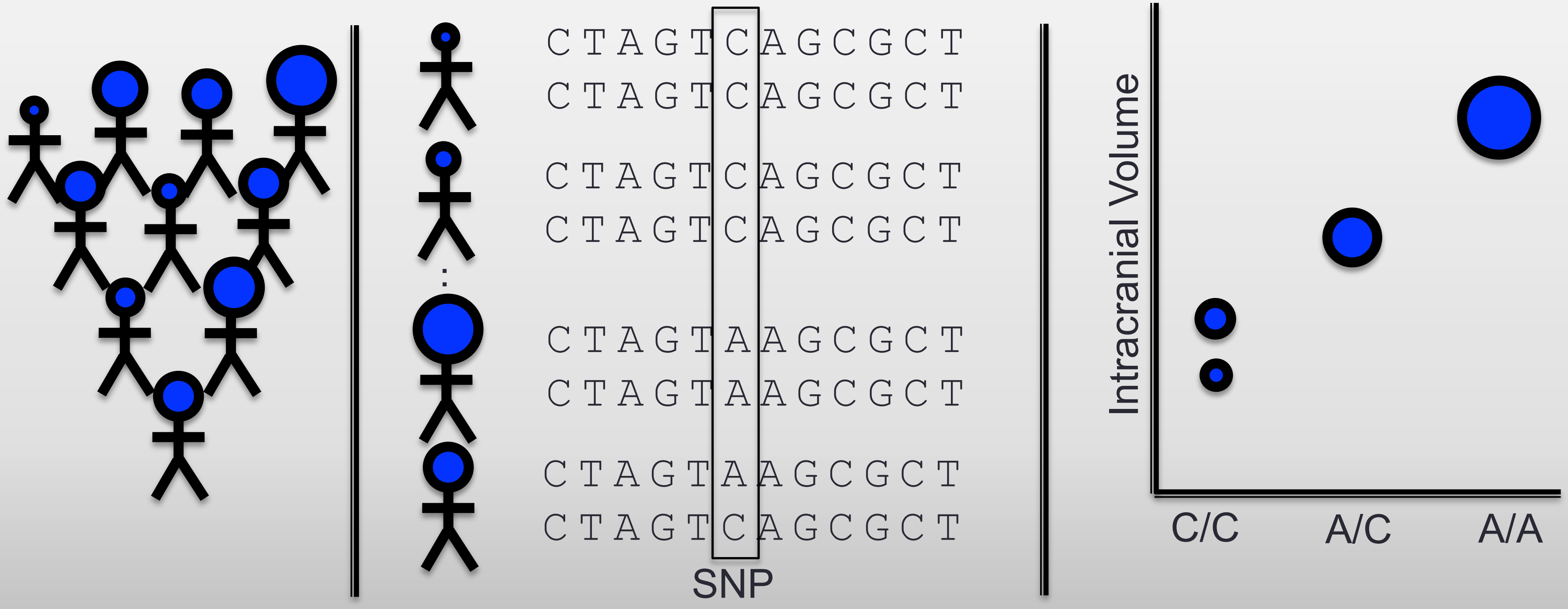
Derrek Paul Hibar
derrek.hibar@ini.usc.edu

USC Mark and Mary Stevens
Neuroimaging and Informatics Institute

Objectives

- Obtain the ADNI Genetic Data
- Quality Control Procedures
  - Missingness
  - Testing for relatedness
  - Minor allele frequency (MAF)
  - Hardy-Weinberg Equilibrium (HWE)
  - Testing for ancestry (MDS analysis)
- Image-wide genetic analysis!

Finding Common Genetic Variants Influencing Brain Structure



Phenotype

Genotype

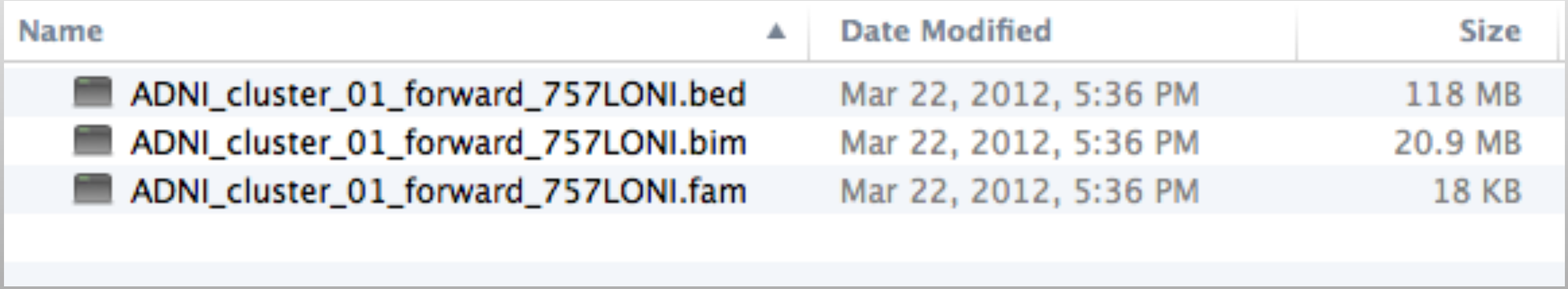
Association

Software and Scripts

- PLINK
  - http://pngu.mgh.harvard.edu/~purcell/plink/
- R
  - http://cran.us.r-project.org/
- PLINK 2 (beta)
  - https://www.cog-genomics.org/plink2
- Files for ADNI_Diagnosis and MDS Plots
  - https://github.com/dhibar/OHBMImagingGenetics2015

PLINK Formatted Genotypes

- Download/Unzip PLINK formatted ADNI1 data.
  - ida.loni.usc.edu/
- Data are in binary (compressed) PLINK format:
  - http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#bed



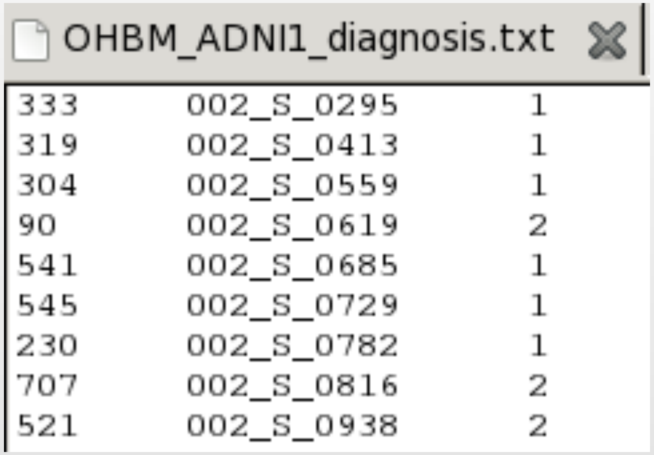
Name	Date Modified	Size
ADNI_cluster_01_forward_757LONI.bed	Mar 22, 2012, 5:36 PM	118 MB
ADNI_cluster_01_forward_757LONI.bim	Mar 22, 2012, 5:36 PM	20.9 MB
ADNI_cluster_01_forward_757LONI.fam	Mar 22, 2012, 5:36 PM	18 KB

PLINK Formatted Genotypes

- We need to update the files to include diagnostic status.
  - This is important for later steps (specifically for HWE testing).
- Download Diagnosis Information:
  - From the LONI IDA
  - Patients = 2; Controls = 1

PLINK Formatted Genotypes

Example of OHBM_ADNI1_diagnosis.txt:



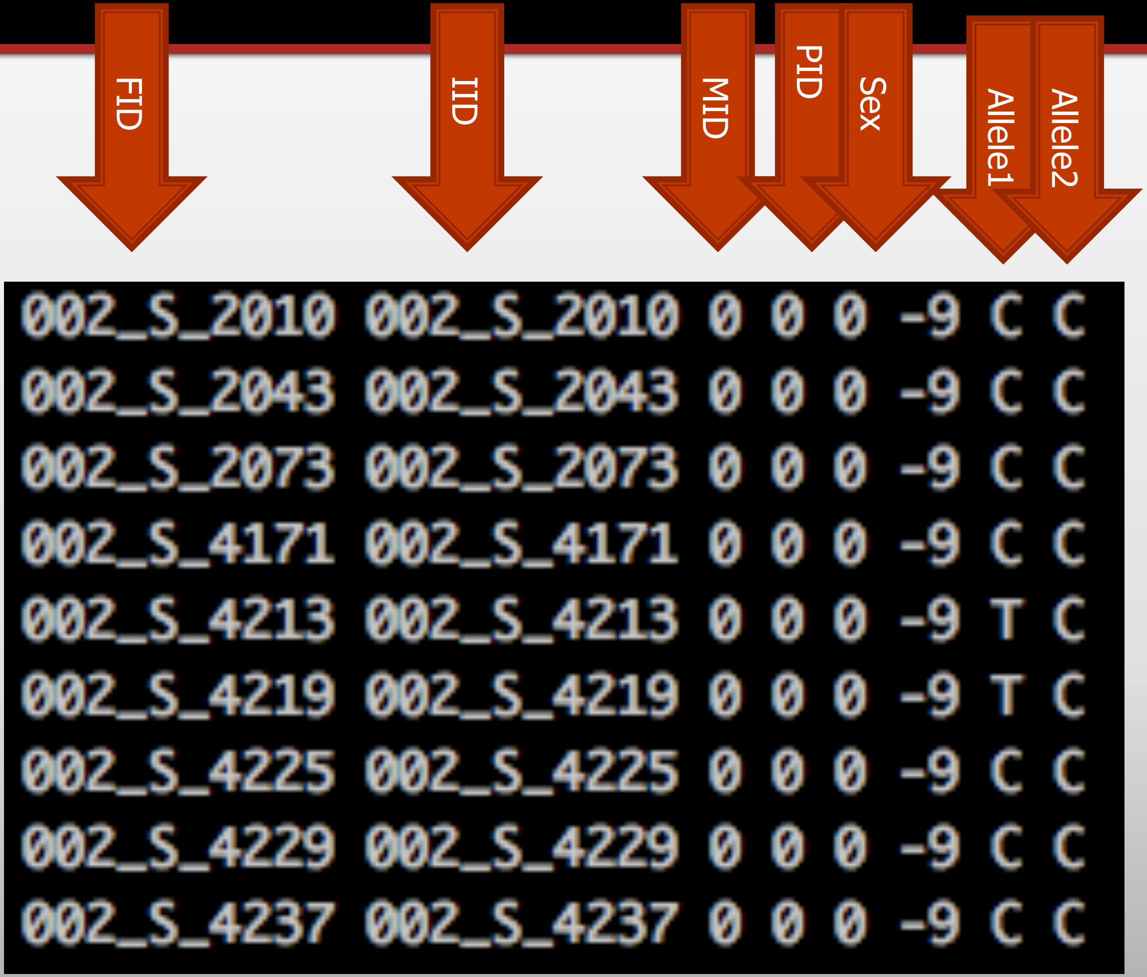
```
333         002_S_0295        1
319         002_S_0413        1
304         002_S_0559        1
90          002_S_0619        2
541         002_S_0685        1
545         002_S_0729        1
230         002_S_0782        1
707         002_S_0816        2
521         002_S_0938        2
```

```
plink
--bfile ADNI_cluster_01_forward_757LONI
--pheno OHBM_ADNI1_diagnosis.txt
--noweb
--make-bed
--out ADNI1_Genotypes_Unfilt
```

PLINK Formatted Genotypes

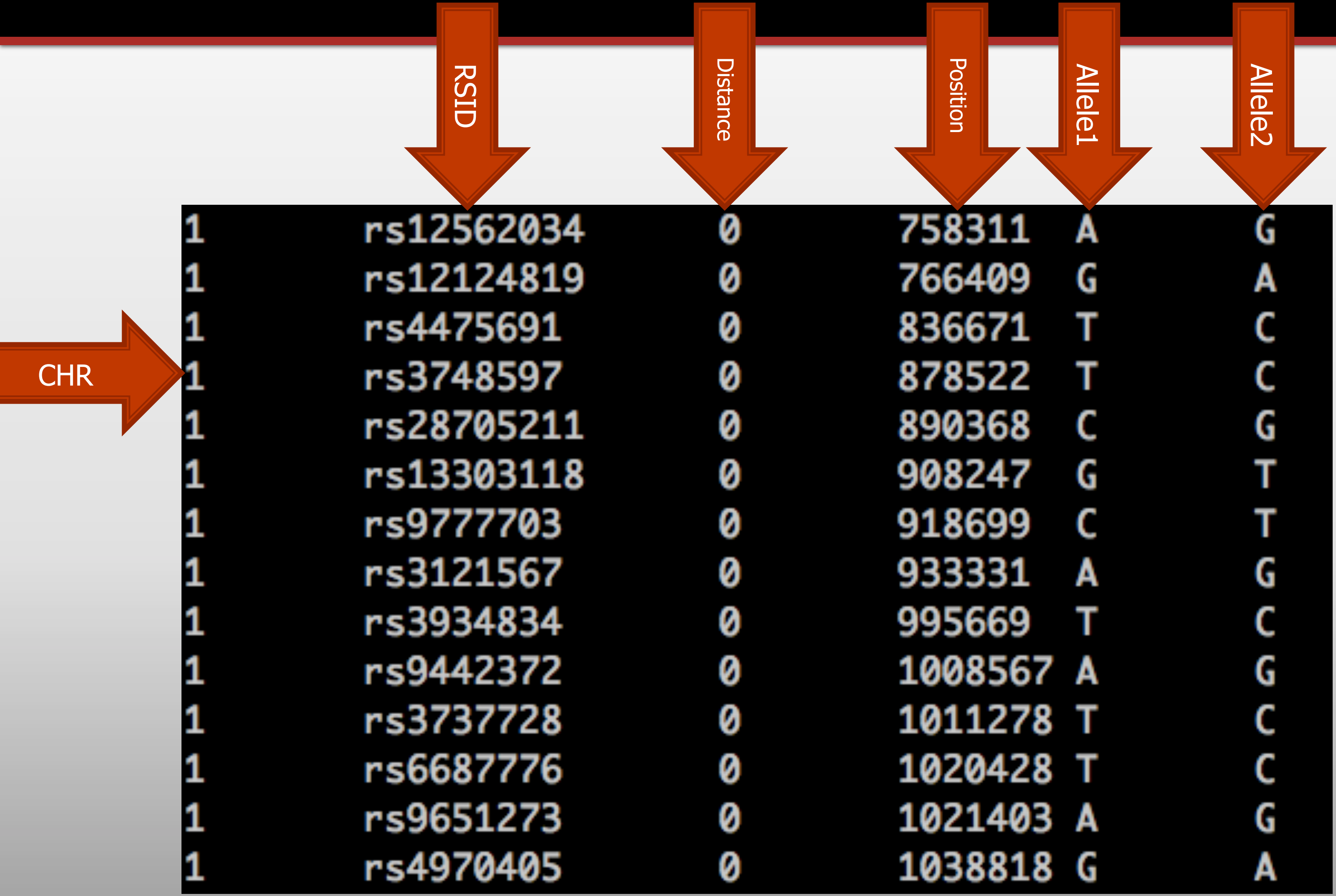
- We name our PLINK formatted ADNI1 genotype data:
  - ADNI1_Genotypes_Unfilt.bed
  - ADNI1_Genotypes_Unfilt.bim
  - ADNI1_Genotypes_Unfilt.fam
- The files contain 757 subjects
  - 449 Males and 308 Females
  - 177 AD, 366 MCI, and 214 CTLs

ADNI1_Genotypes_Unfilt.bed



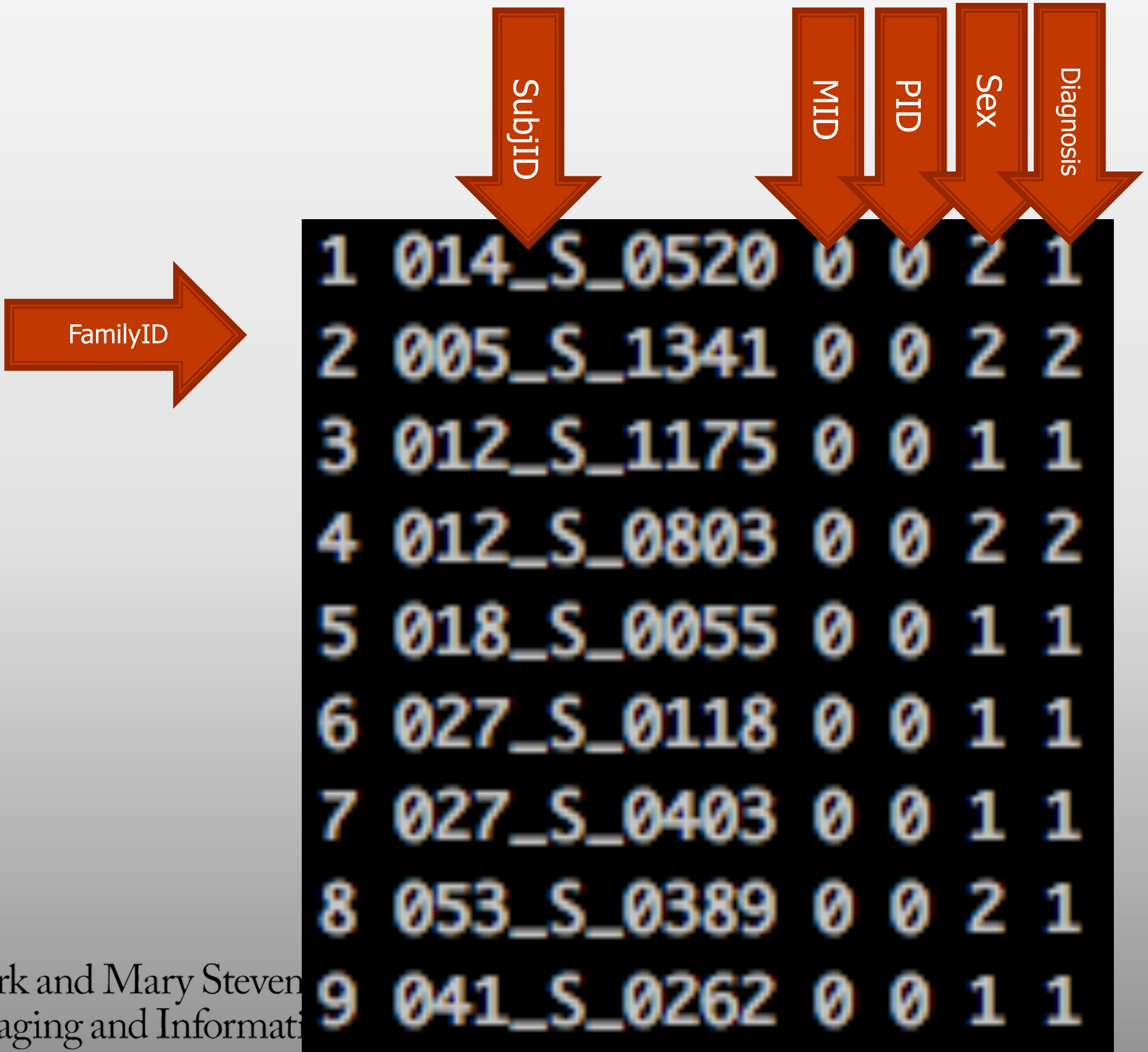
```
002_S_2010 002_S_2010 0 0 0 -9 C C
002_S_2043 002_S_2043 0 0 0 -9 C C
002_S_2073 002_S_2073 0 0 0 -9 C C
002_S_4171 002_S_4171 0 0 0 -9 C C
002_S_4213 002_S_4213 0 0 0 -9 T C
002_S_4219 002_S_4219 0 0 0 -9 T C
002_S_4225 002_S_4225 0 0 0 -9 C C
002_S_4229 002_S_4229 0 0 0 -9 C C
002_S_4237 002_S_4237 0 0 0 -9 C C
```

ADNI1_Genotypes_Unfilt.bim



```
1       rs12562034   0      758311  A      G
1       rs12124819   0      766409  G      A
1       rs4475691    0      836671  T      C
1       rs3748597    0      878522  T      C
1       rs28705211   0      890368  C      G
1       rs13303118   0      908247  G      T
1       rs9777703    0      918699  C      T
1       rs3121567    0      933331  A      G
1       rs3934834    0      995669  T      C
1       rs9442372    0      1008567 A      G
1       rs3737728    0      1011278 T      C
1       rs6687776    0      1020428 T      C
1       rs9651273    0      1021403 A      G
1       rs4970405    0      1038818 G      A
```

ADNI1_Genotypes_Unfilt.fam



```
1 014_S_0520 0 0 2 1
2 005_S_1341 0 0 2 2
3 012_S_1175 0 0 1 1
4 012_S_0803 0 0 2 2
5 018_S_0055 0 0 1 1
6 027_S_0118 0 0 1 1
7 027_S_0403 0 0 1 1
8 053_S_0389 0 0 2 1
9 041_S_0262 0 0 1 1
```

#1 - Quality Control Procedures

- #1 - Check for Discordant Sex Information
  - Use genotype data from X chromosomes to determine sex (females have two copies, males have only one).
  - Compare the genotyped sex to the sex reported in the study. % heterozygosity on the X chromosome is used to determine genotypic sex.
  - Consider removing subjects with discordant sex information in PLINK using the --remove command
    - http://pngu.mgh.harvard.edu/~purcell/plink/dataman.shtml#remove

#1 - Quality Control Procedures

- To check sex with PLINK:

  ```
  plink
  --bfile ADNI1_Genotypes_Unfilt
  --check-sex
  --out ADNI1_sex
  ```

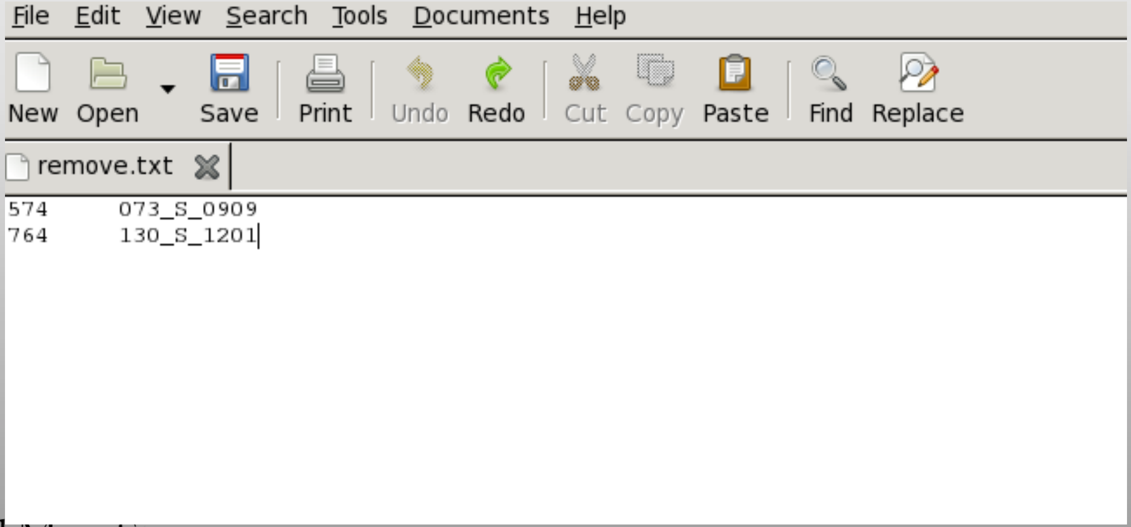
- Print out any discordant subjects:

  ```
  grep "PROBLEM" ADNI1_sex.sexcheck
  ```

#1 - Quality Control Procedures

FID	IID	PEDSEX	SNPSEX	STATUS	F
574	073_S_0909	2	0	PROBLEM	0.2268
764	130_S_1201	2	0	PROBLEM	0.2273

- Remove these subjects from the dataset:
  - Store the FID and IID in a text file called remove.txt


File Edit View Search Tools Documents Help
New Open Save Print Undo Redo Cut Copy Paste Find Replace
remove.txt

```
574     073_S_0909
764     130_S_1201
```

#2 - Quality Control Procedures

- #2 Test for Missingness
  - We excluded genotypes with GC Scores < 0.15 and marked them as missing. If >10% of the total set of SNPs genotyped are missing it might indicate a poorly genotyped subject.
- Using PLINK:

  -
    ```
    plink --bfile ADNI1_Genotypes_Unfilt
    --remove remove.txt --noweb --missing
    --out missingness
    ```
- Print the subjects with >10% missingness:

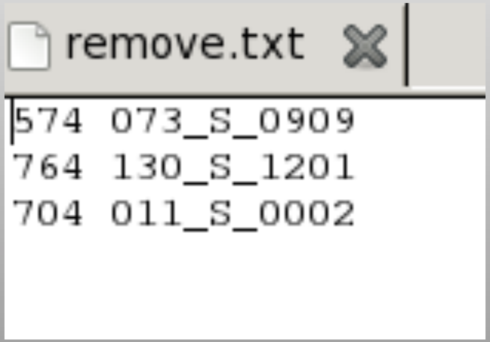
  -
    ```
    awk '{if($6 > 0.1) print $0}'
    missingness.imiss
    ```

#2 - Quality Control Procedures

- In our data it looks like one subject might have excessive missingness:

FID	IID	MISS_PHENO	N_MISS	N_GENO	F_MISS
011_S_0002	011_S_0002	N	63407	620901	0.1021

Update the remove.txt file to exclude this subject as well:



```
574 073_S_0909
764 130_S_1201
704 011_S_0002
```

#3 - Quality Control Procedures

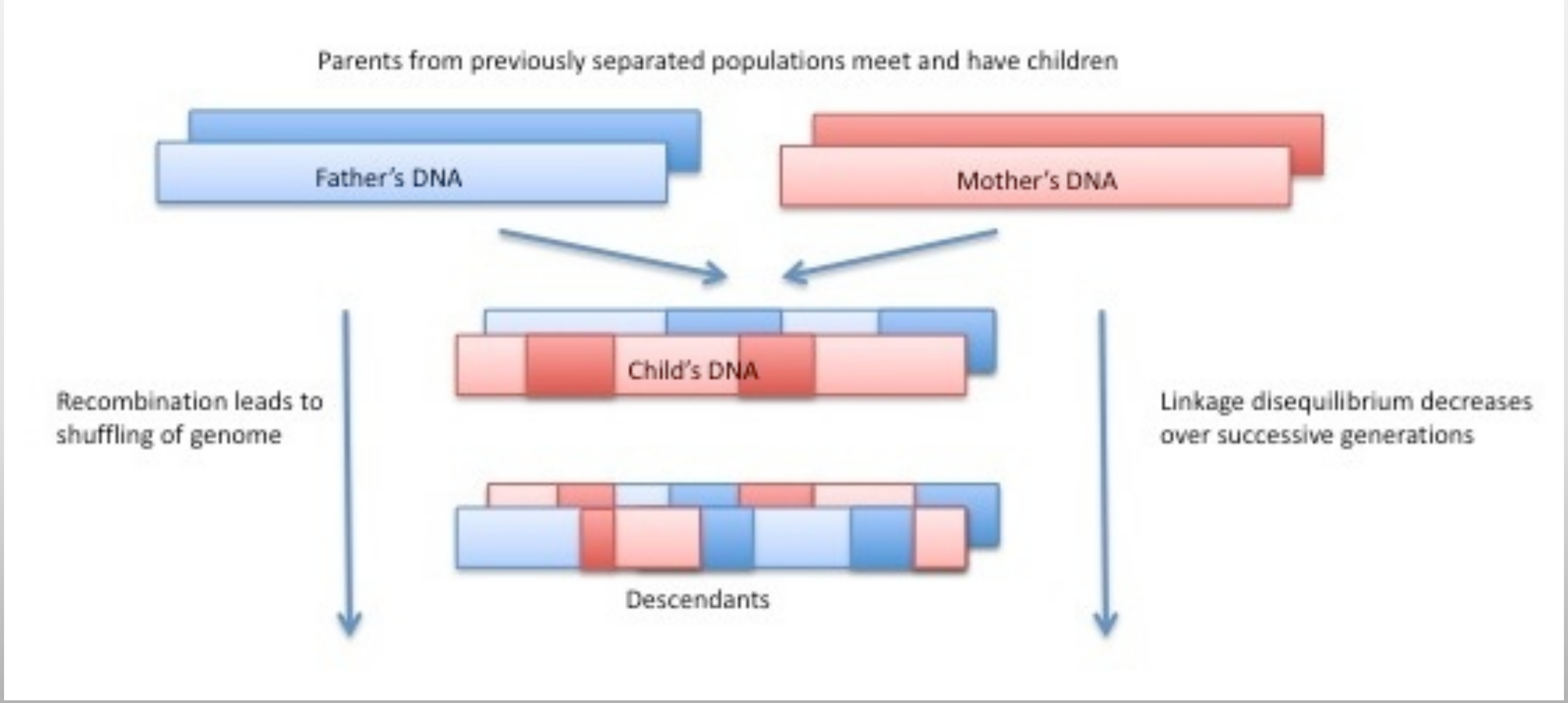
- #3 - Identifying Related Subjects
- Prune down high-LD regions:

```
plink --bfile ADNI1_Genotypes_Unfilt
--indep-pairwise 50 5 0.2
--remove remove.txt
--out relatedness
--noweb
```

- Generate an IBS Matrix:

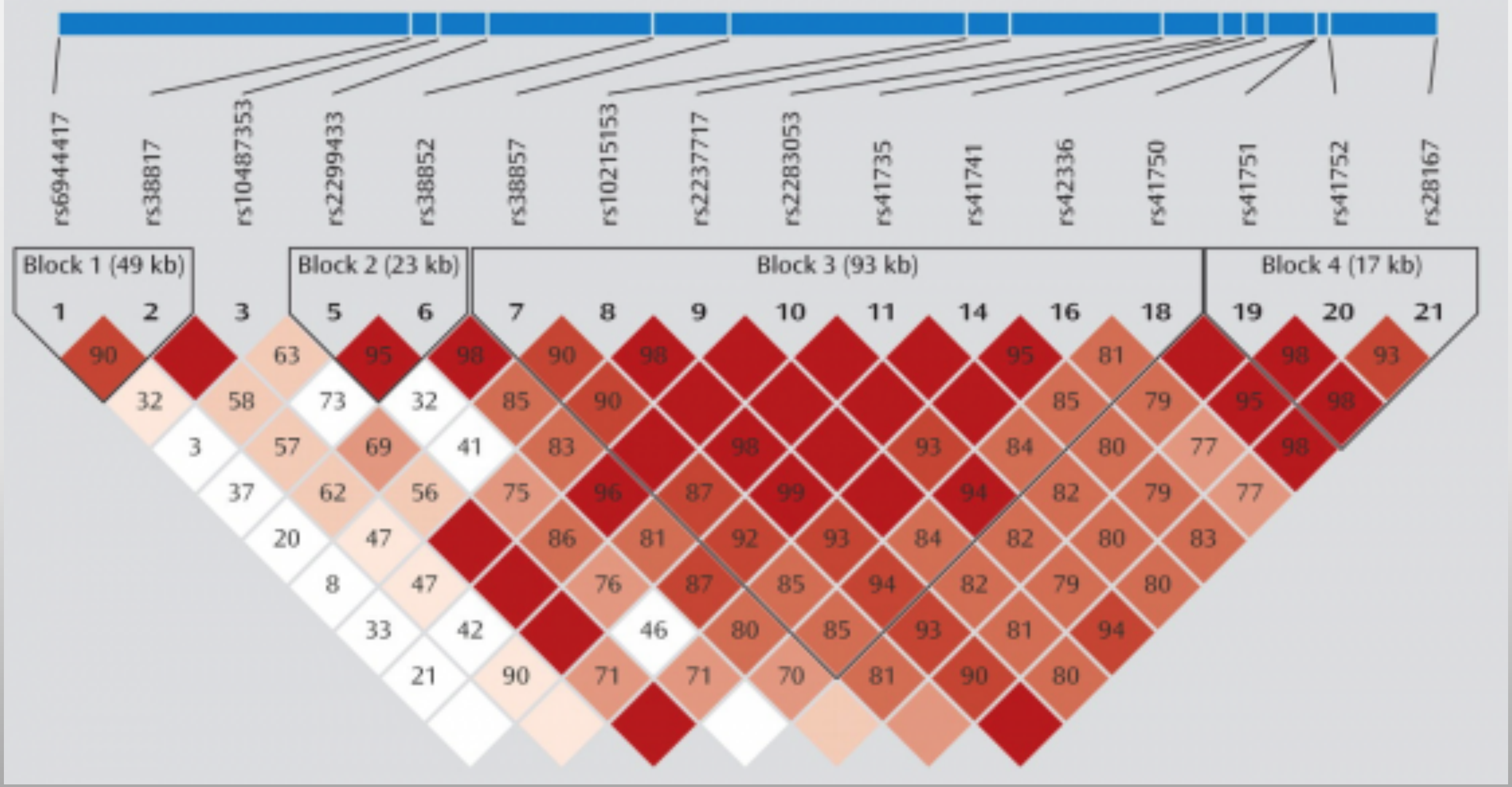
```
plink --bfile ADNI1_Genotypes_Unfilt
--extract relatedness.prune.in
--genome
--out relatedness
--noweb
```

What is Linkage Disequilibrium?


Parents from previously separated populations meet and have children
Father's DNA
Mother's DNA
Child's DNA
Recombination leads to shuffling of genome
Linkage disequilibrium decreases over successive generations
Descendants

http://shared.web.emory.edu/whsc/news/img/whsc/linkage_disequilibrium.jpg

What is Linkage Disequilibrium?



USC Mark and Mary Stevens
Neuroimaging and Informatics Institute

https://estrip.org/articles/read/tinypliny/44920/Linkage_Disequilibrium_Blocks_Triangles.html

#3 - Quality Control Procedures

- Identify related subjects, with an IBS > 0.2

  -
    ```
    plink --bfile ADNI1_Genotypes_Unfilt
    --extract relatedness.prune.in
    --min 0.2
    --genome
    --genome-full
    --out relatedness
    --noweb
    ```
- Remove one subject from each related pair, keep the one with the highest genotyping rate:
  - `grep "057_S_0643" missingness.imiss`
  - `grep "057_S_0934" missingness.imiss`
    - Add 057_S_0934 to remove.txt (has high missingness)

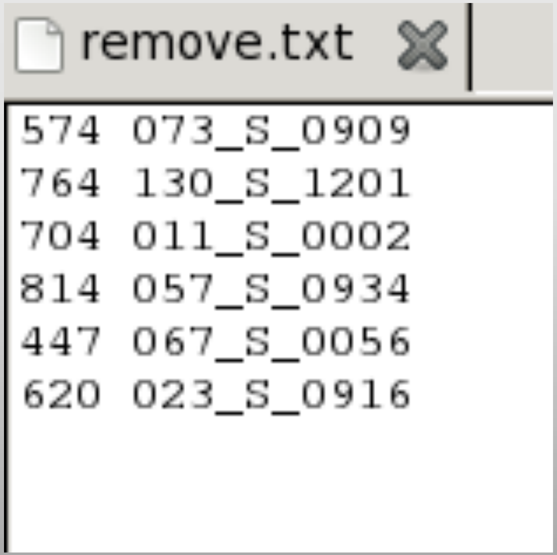
#3 - Quality Control Procedures

- We found that 6 subjects were related:

FID1	IID1	FID2	IID2	PI_HAT	PHE	DST	PPC	RATIO
56	057_S_0643	814	057_S_0934	0.524	-1	0.877099	1	11.868
359	067_S_0059	447	067_S_0056	0.4746	-1	0.865581	1	9.8591
591	023_S_0058	620	023_S_0916	0.5266	0	0.87768	1	11.0342

#3 - Quality Control Procedures

- Add one subject from each pair to the remove.txt list (the ones with the highest missingness):



```
574 073_S_0909
764 130_S_1201
704 011_S_0002
814 057_S_0934
447 067_S_0056
620 023_S_0916
```

A Pre-Cleaned Dataset

- Create a new PLINK file that removes each of the subjects in the remove list that can then be carried forward for additional QC:

```
plink --bfile ADNI1_Genotypes_Unfilt
--remove remove.txt
--make-bed
--out ADNI1_Genotypes_Unfilt_preclean
--noweb
```

#4 - Quality Control Procedures

- Now that we have carefully looked at our dataset and removed bad samples we can filter the dataset:
  - plink

```
--bfile ADNI1_Genotypes_Unfilt_preclean
--maf 0.01
--geno 0.05
--hwe 5e-7
--make-bed
--out ADNI1_Genotypes_Filt
--noweb
```

Filtering Criteria

- --maf 0.01
  - Removes “rare” SNPs, if the minor allele occurs fewer than 1% of the total alleles.
- --geno 0.05
  - Removes SNPs that have >5% of alleles missing. This is related to the subject-wide missingness (--mind)
- --hwe 5e-7
  - Removes SNPs that significantly deviate from Hardy-Weinberg Equilibrium. The option 5e-7 we give here is the p-value threshold from the HWE test we use to exclude tests.
  - Note: HWE can detect deviations in allele frequency that might be due to poor genotyping. However, if you are looking at a case-control cohort alleles may deviate from HWE just because they are overrepresented in your patient population. So it is good practice to only run the HWE tests in controls (this is the default behavior PLINK, but you have to first include diagnosis information).

#5 - Quality Control Procedures

- Before we can use our cleaned files for genetic association testing we need to examine the ethnicities of our samples.
- For genetic tests we can only compare samples of the same ancestry, or else we risk discovering spurious results due to Population Stratification.
  - Li, C. C. "Population subdivision with respect to multiple alleles." Annals of human genetics 33.1 (1969): 23-29.

Eric Lander's example

- Say you want to study the "trait" of ability to eat with chopsticks
- Decide to look at the HLA-A1 allele in San Francisco
- We know that the HLA-A1 allele is more common among Asians than Caucasians
- So when looking for an association we would conclude that Asian ethnicity is associated with the phenotype of interest
  - But obviously we know that immune response does not play a role in your ability to use chopsticks.

Quality Control Procedures

- Using Multi-Dimensional Scaling (MDS) Analysis we can estimate the ancestry of each sample in our study by comparing their genetic footprint with other subjects of known ancestry.

Quality Control Procedures

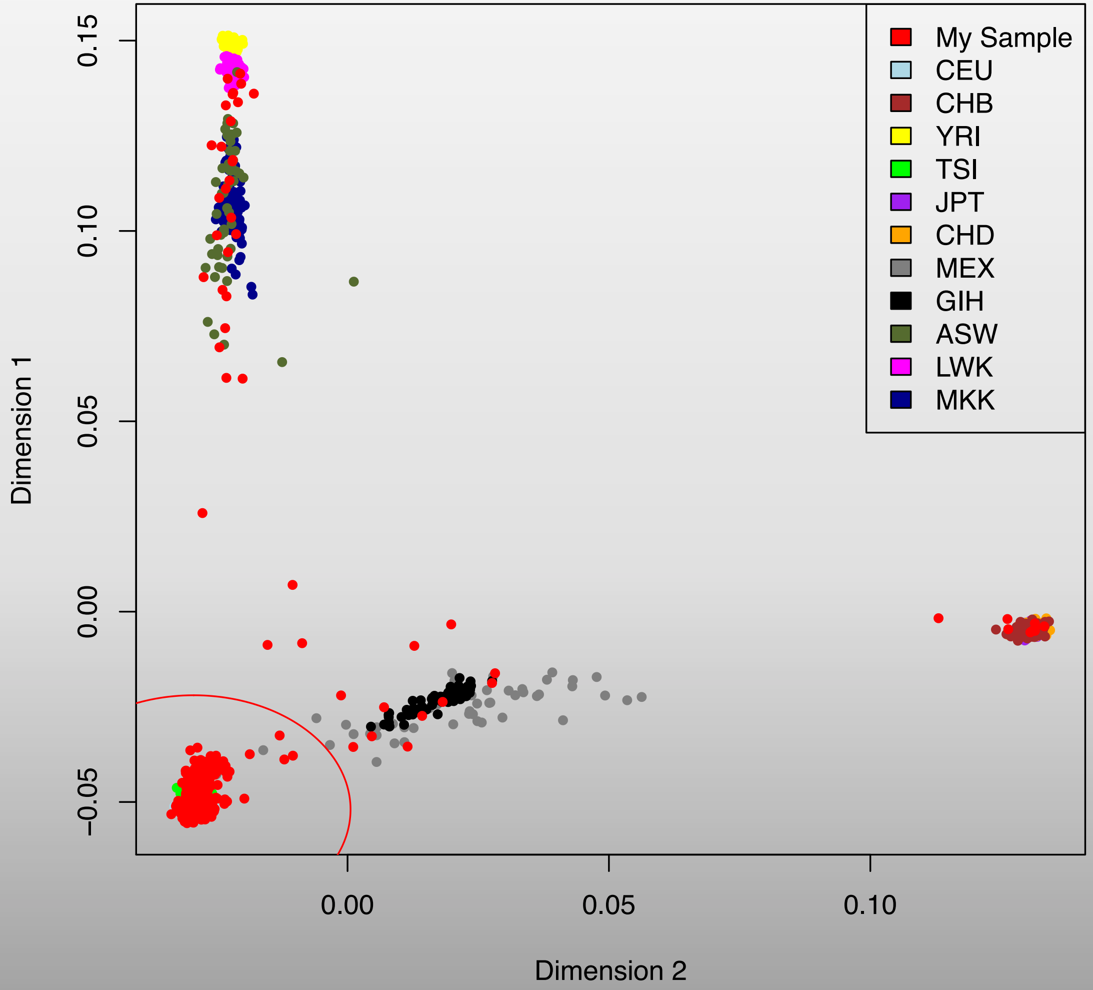
- Performing an MDS Analysis

  ```
  awk 'BEGIN{OFS=","};{print $1, $2, $3, $4, $5, $6, $7}' >> HM3mds2R.mds.csv HM3mds.mds
  ```

Quality Control Procedures

- Visualize results in R
  - R
  - `source(mdsplot.R)`
  - #Read our MDS analysis output into R
  - `mds.cluster = read.csv(as.character("HM3mds2R.mds.csv"), header=T)`
  - #Plot our data
  - `mdsplot(mds.cluster,pop.interest="CEU",pruningf=0.03,plotfinal=FALSE,flip.x=FALSE,flip.y=FALSE)`

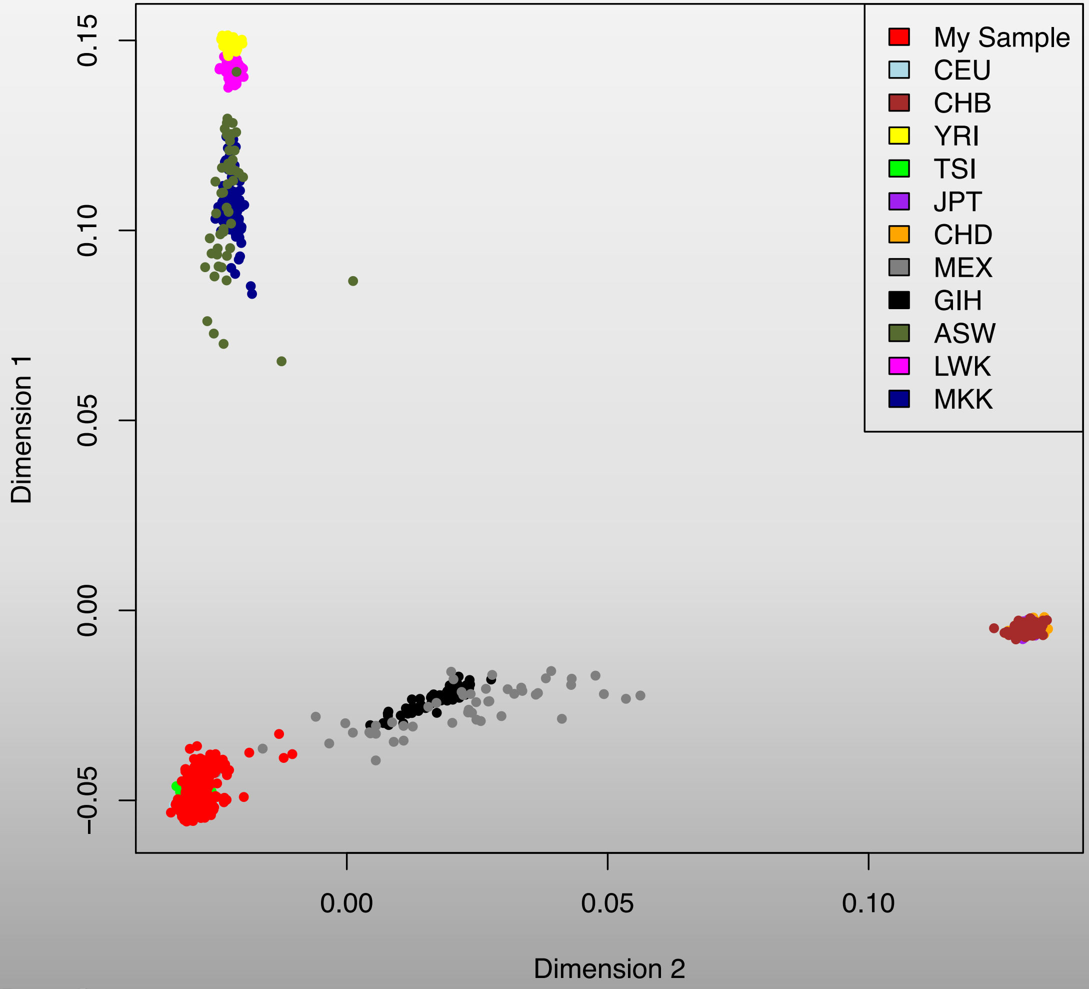
Quality Control Procedures



Quality Control Procedures

- Plot data with outliers removed
  - `mdsplot(mds.cluster,pop.interest="CEU",pruningf=0.03,plotfinal=TRUE,flip.x=FALSE,flip.y=FALSE)`
- A final plot will be outputted as well as a file called HM3mds_Pruned_0.03_CEU.txt which will contain the list of subjects to keep in the analysis.

Quality Control Procedures



Quality Control Procedures

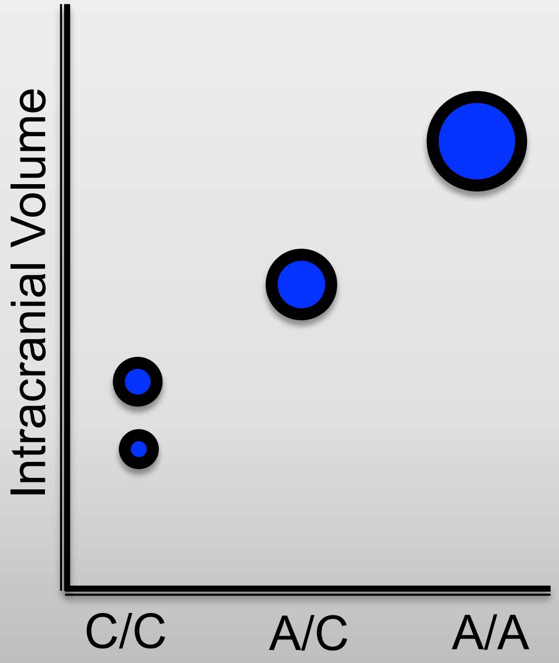
- We still need to drop the ancestry outliers from our dataset.

  -
    ```
    awk 'NR > 1{print $1, $2}' HM3mds_Pruned_0.03_CEU.txt > Subjects.list
    ```

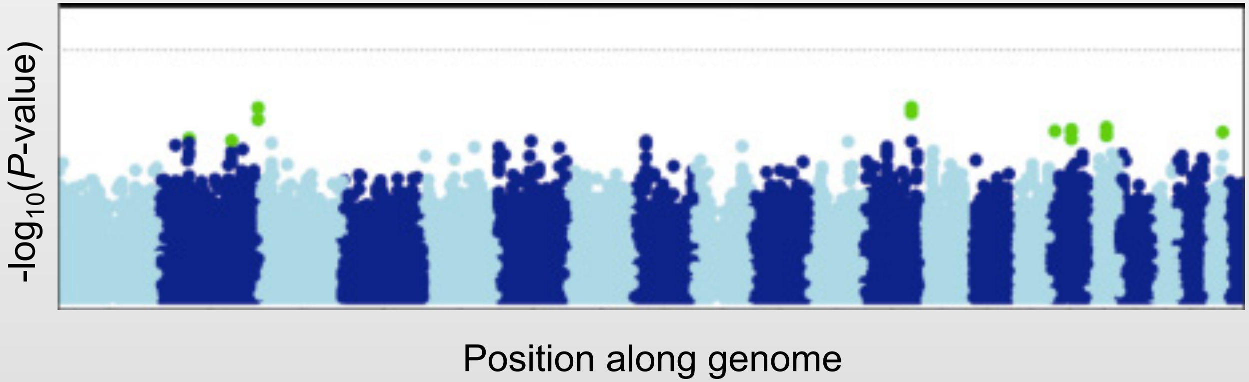
  -
    ```
    plink --bfile ADNI1_Genotypes_Filt --noweb --keep Subjects.list --make-bed --out ADNI1_Genotypes_Filt_CEU
    ```

Genome-wide Association Study

One SNP



Millions of SNPs



An unbiased search to find where in the genome a common variant is associated with a trait.

Genome-wide Association Study

- You can run a GWAS very easily in PLINK.
- First, you need to create a text file for the phenotype (imaging trait) that you want to test.
- The --pheno file is just a text file organized in excel and saved as a tab delineated text file. The first column is the subject FamilyID, second column is the SubjectID, and the third column is the value/ phenotype you are testing. No column headers.
  - Lets run a GWAS on temporal lobe volume, saving the phenotype info in a file called temporal.txt

1	014_S_0520	3090
2	005_S_1341	4039
3	012_S_1175	3847
4	012_S_0803	5983
5	018_S_0055	2999
6	027_S_0118	3485

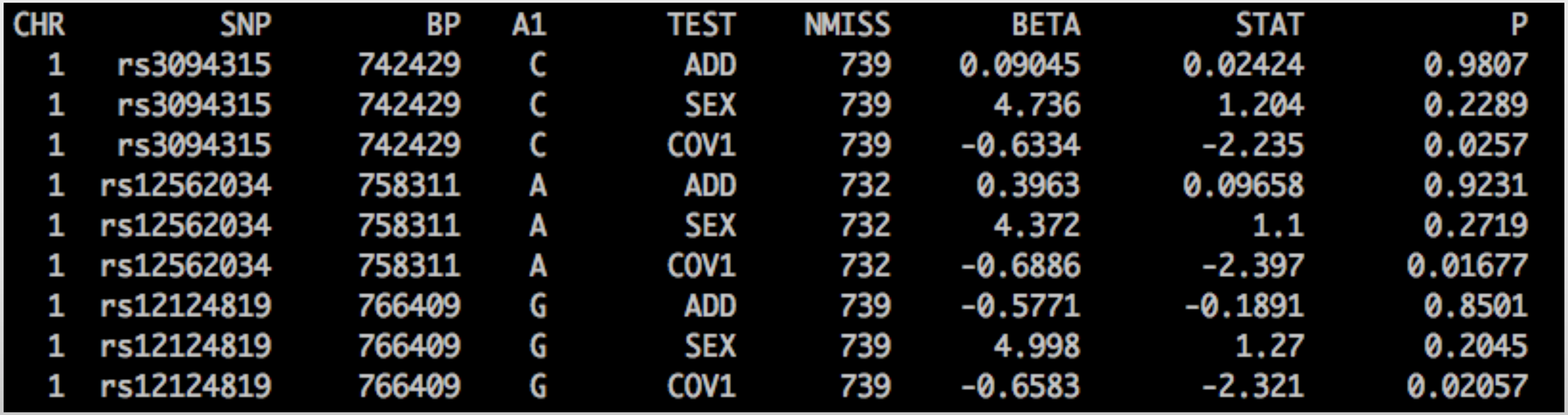
Genome-wide Association Study

- What about covariates? You can include covariates like age, sex, intracranial volume, etc. by creating a text file just like the phenotype file. The first two columns are the FamilyID and IndividualID and each column after that is a covariate.
  - NOTE: if sex is already included in your PLINK file then you do not have to add it to your covariates file, you can include it as a covariate by adding the --sex option to your PLINK GWAS command.
- For this analysis we just control for age and sex, in a file called covars.txt

Genome-wide Association Study

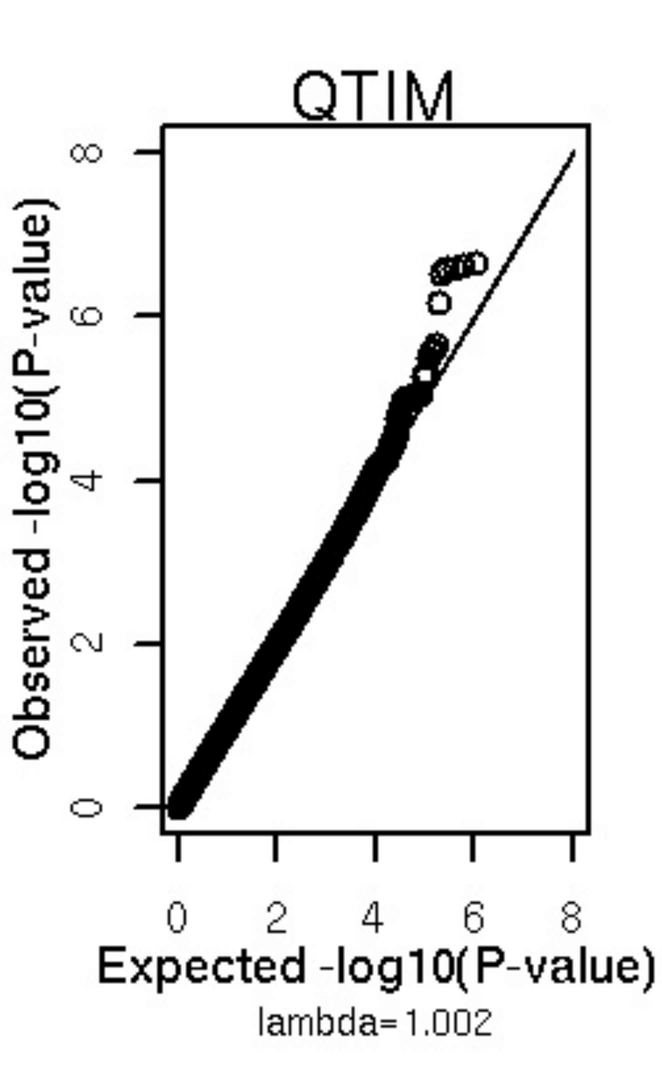
- We’re ready to run a GWAS!
-
  ```
  plink --bfile ADNI1_Genotypes_Filt_CEU --noweb --linear --covar covars.txt --pheno temporal.txt --out temporal_lobe_gwas
  ```
- This will output a file called temporal_lobe_gwas.assoc.linear which is described on the PLINK site:
  - http://pngu.mgh.harvard.edu/~purcell/plink/anal.shtml#glm

Genome-wide Association Study

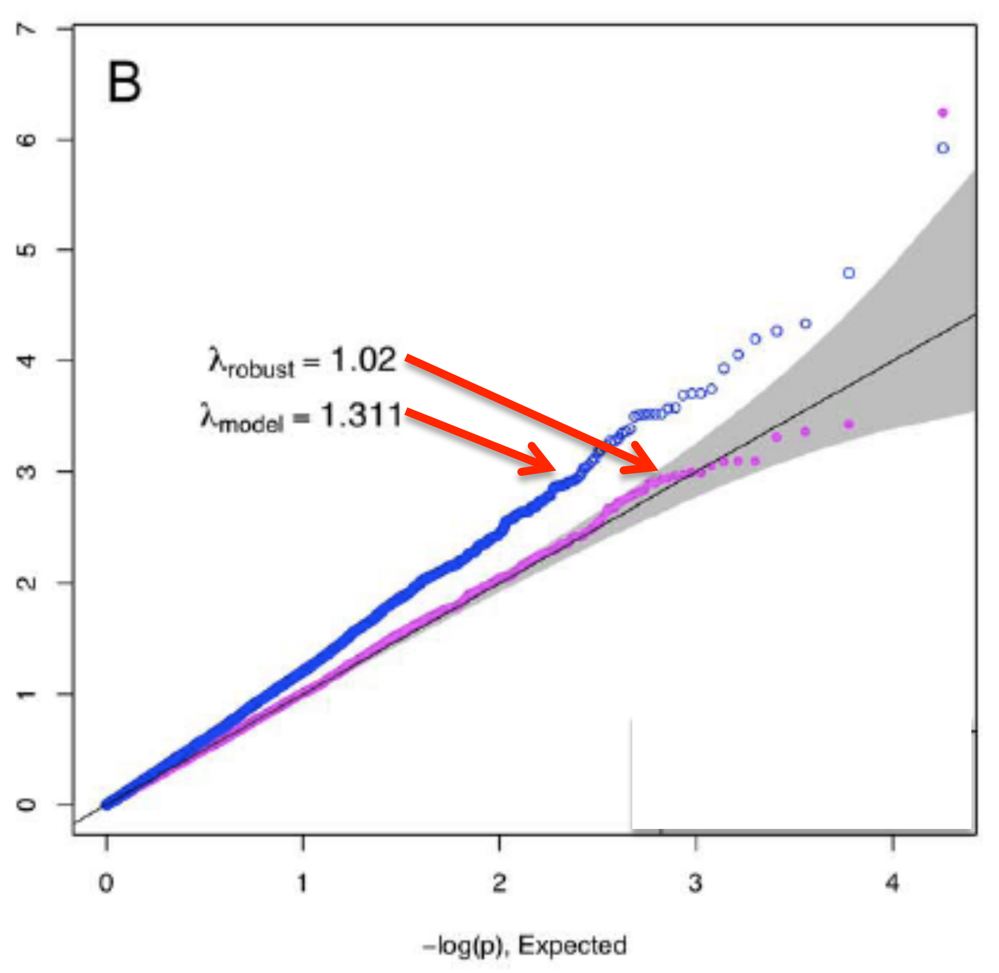


```
CHR         SNP        BP   A1       TEST    NMISS       BETA       STAT           P
  1   rs3094315    742429    C        ADD      739    0.09045    0.02424      0.9807
  1   rs3094315    742429    C        SEX      739      4.736      1.204      0.2289
  1   rs3094315    742429    C       COV1      739    -0.6334     -2.235      0.0257
  1  rs12562034    758311    A        ADD      732     0.3963    0.09658      0.9231
  1  rs12562034    758311    A        SEX      732      4.372        1.1      0.2719
  1  rs12562034    758311    A       COV1      732    -0.6886     -2.397     0.01677
  1  rs12124819    766409    G        ADD      739    -0.5771    -0.1891      0.8501
  1  rs12124819    766409    G        SEX      739      4.998       1.27      0.2045
  1  rs12124819    766409    G       COV1      739    -0.6583     -2.321     0.02057
```

Individual Site QQ Plots



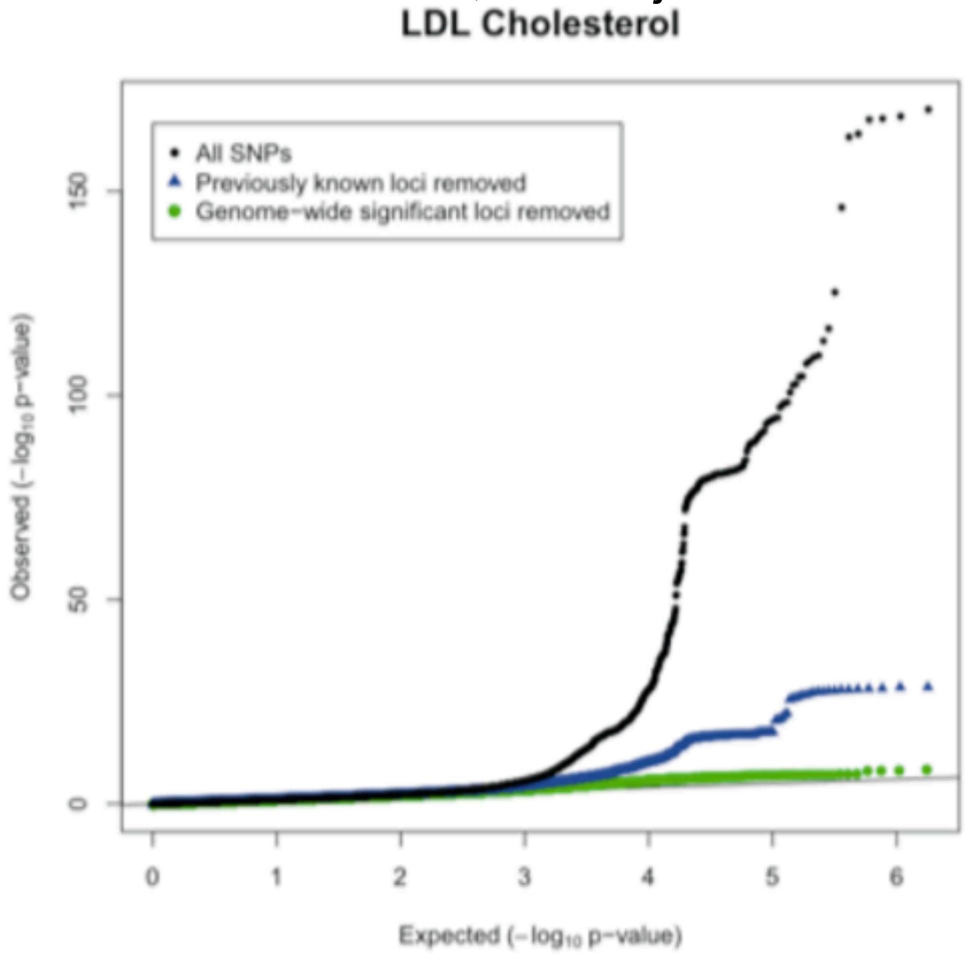
When QQ Plots Go Wrong



Can show evidence of unaccounted for population stratification, cryptic relatedness, or just that your data does not follow expected distributions

When QQ Plots Go Really Well

N>100,000 subjects



(Teslovich et al., 2010)

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.

Selecting A Single SNP for Further Analysis

- You can output a single SNP from your PLINK formatted dataset to be used in other forms (e.g. testing the effects of the SNP at each voxel in the brain).
- To output an additive coded SNP from your dataset use the --recodeA option:

  -
    ```
    plink --bfile ADNI1_Genotypes_Filt_CEU
    --noweb --snp rs6265 --recodeA --out
    bdnf
    ```
- This will output a text file called bdnf.raw. The 7th column gives the total number of minor alleles each subject has (each subject is a row).

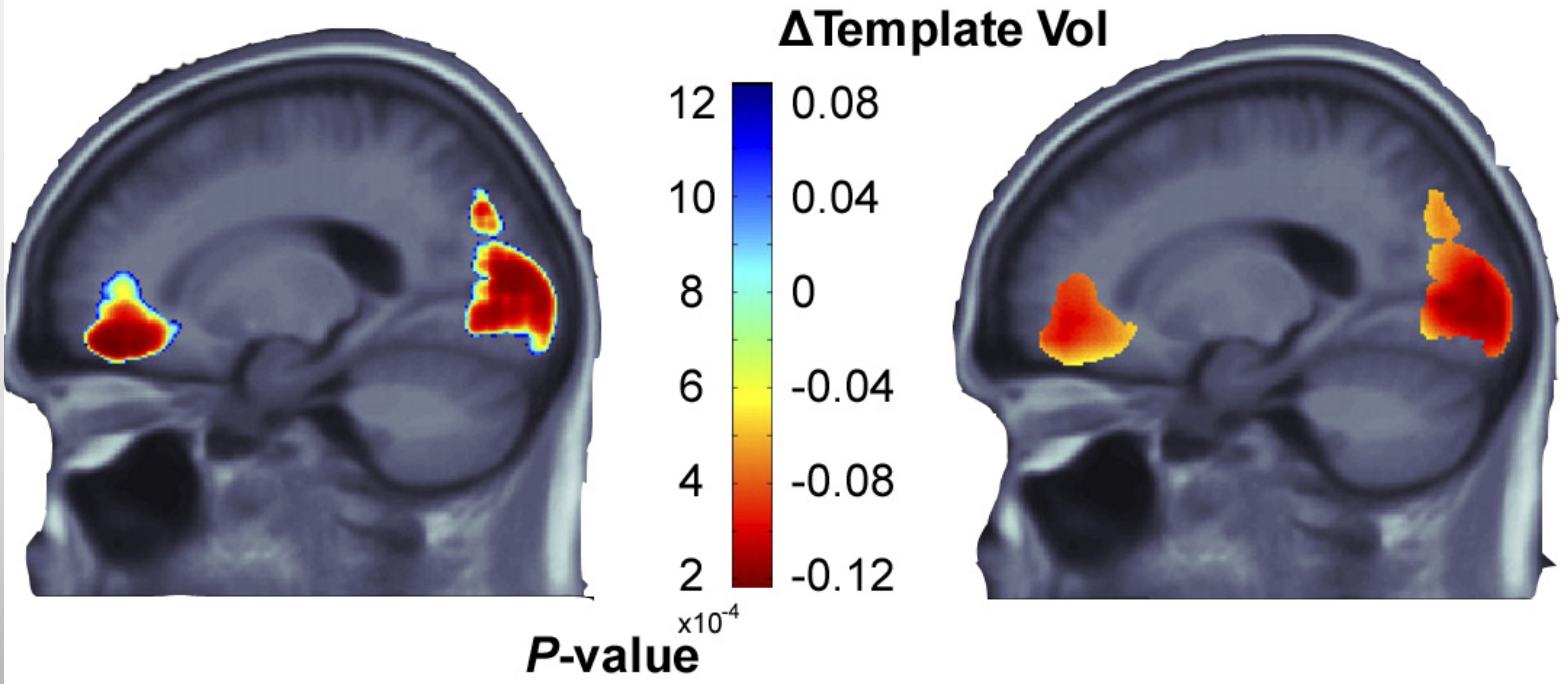
Image-wide genetic analysis

- You can use this extracted SNP for further analyses. One interesting analysis is to look at a SNP's effects in the full brain.
- You can get directions and code for testing a SNP for association at each point in the brain here:
  - https://github.com/dhibar/VoxelwiseRegression
  - All you need are images and a mask file.

Image-wide genetic analysis

- To download the files, go to https://ida.loni.usc.edu -> Project ADNI -> Search. In your search panel, please click <post-processed> under Image Types and enter <TBM*> under Image Description. The full description is TBM Jacobian Maps [MDT - Screening] . You should find N=817 files and then Select All -> Add to a Collection. The Jacobian maps were created by nonlinearly warping the screening scan to the average group template or MDT, thus the Jacobian values indicate regional volume differences between the screening scan and the MDT. You can download a copy of the MDT here:
  - http://users.loni.usc.edu/~thompson/XUE/MDT/ADNI_ICBM9P_MDT.nii

Pretty Pictures



Useful web resources

UCSC genome browser: http://genome.ucsc.edu/cgi-bin/hgGateway
Genome visualization magic.

Hapmap: http://hapmap.ncbi.nlm.nih.gov/
Allele frequencies in multiple populations.

Allen Brain Atlas: http://www.brain-map.org/
See where a gene is expressed.

Entrez Gene: http://www.ncbi.nlm.nih.gov/gene/
See the gene ontology (what it does).

dbSNP: http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp
The database of every documented genetic variation.

Plink: http://pngu.mgh.harvard.edu/~purcell/plink/
Incredibly useful tool for genome-wide analysis, organization, etc. Excellent documentation.

dbGaP: http://www.ncbi.nlm.nih.gov/gap/
Database of genotypes and phenotypes.

Acknowledgements

IGC (USC)
Paul Thompson (Advisor)
Jason L Stein
Neda Jahanshad

QTIM (Australia)
Sarah Medland
Margie Wright
Katie McMahon
Nick Martin
Greig de Zubicaray

IGC
Imaging Genetics Center

USC Mark and Mary Stevens
Neuroimaging and Informatics Institute