Imputing big data from GWAS

Overview, exercises and case study
What we’ll be covering today...

...we will discuss genome-wide aspects of genetic epidemiological studies focusing on meta-analyses of imputed data.

Discuss basic theory behind GWAS, imputation and large-scale meta-analyses.

We will cite examples from recent work in Parkinson’s disease.

Share code, work on some examples.

If you have further questions after class feel free to email m.nalls.working@gmail.com.
Agenda:

1. Introduction to GWAS and basic history of GWAS
2. Exercise 1: data formatting and imputation
3. Meta-analysis and study design
4. Exercise 2: run analyses on study level
5. Current mega-analyses

!!!COFFEE BREAK / QUESTIONS!!!

6. Exercise 3: meta-analyses
7. Risk profiling
8. Functional inference
9. Heritability
10. Moving past basic disease GWAS

!!!MORE COFFEE /MORE QUESTIONS!!!
An introduction to genome-wide approaches in genetic epidemiology…

The field of genome-wide genetic epidemiology studies arose out of advancing technology.

The technology available at a low cost beginning in the mid 2000’s allowed for cost effective genotyping of hundreds of thousands of single nucleotide polymorphism (SNPs).

SNPs are simple common variants with generally 2 possible options at any location in your genome. Either AA, AB or BB at any particular location.

You have millions of SNPs in you genome, this is what makes us unique (unless you are a monozygotic twin).

These SNPs can have risk or protective effects, but for common SNPs that are variable across populations and individuals, the effects of these variants are small in complex traits or diseases for the most part.

Until companies like Illumina or Affymetrix developed genome-wide SNP assays, it took millions of dollars and considerable time/infrastructure to assay a few hundreds of thousands of variants in one sample.

Now you can genotype millions of SNPs for about $100 per sample.
An introduction to genome-wide approaches in genetic epidemiology…

Now we spend a little money and gain information on hundreds of thousands of SNPs for about $100/sample in the mid 2000’s

- The question arose of how could this wealth of data help with studies of disease genetics?
- Study design was a major issue as were statistical considerations and QC.
- Dealing with relatively common variants of small effect, not functional mutations that cause Mendelian diseases.

Therefore methods had to be developed to test these SNPs for associations with disease, treating each SNP as a separate exposure.

**SIMPLE MODELING:**
- Logistic regression
  - Disease ~ SNP + covariates
  - SNP parameter is the dosage of one allele (0, 1 or 2 copies)
  - Testing likelihood that particular allele at SNP has a significantly higher frequency in cases than controls, after using covariates to exclude effects of other factors from the calculations. *Essentially, we only want to see risk associated with the SNP!*
An introduction to genome-wide approaches in genetic epidemiology…

The premise is simple, hundreds of thousands of SNPs, one or a few of which may be associated with disease.

Here begins the a-hypothetical research paradigm:
- Test all available variants
- Look for small consistent effects, no-one expects huge disease risk effects for common SNPs
- Huge penalty for multiple testing
  To reduce false positives after hundreds of thousands of tests, significance is declared $P < 1e^{-7}$ instead of the 0.05 for a single statistical test
- Every SNP is tested under the assumption of a possible association with disease, not some prior biological or epidemiological knowledge of the SNP or the gene it resides in

So, the issue is statistical power. Small effects and the prohibitive nature of $$$ and assay time limited early GWAS from detecting or replicating many real association signals, primarily due to small sample sizes.
An introduction to genome-wide approaches in genetic epidemiology…

So what do you get?
Summary statistics for each tested SNP (per cohort analyzed).
These should be normally distributed, and the lambda value / qq plot is the most important metric of quality results

From a great reference article
PMID:18349094
Age related; ~1% of the population > age 50
Neurological disease
Cardinal Clinical Features:
  - Resting tremor
  - Bradykinesia
  - Rigidity
  - Postural instability
Pathology:
  - Loss of neurons including SNpc
  - α-synuclein-, Ub- positive inclusions
Progressive course
  - Dementia in 30-50% of cases
Unknown environmental etiology
More common in males
Less common in smokers and coffee drinkers
Head trauma likely associated
Early GWAS of Parkinson’s disease…

Our lab was one of the early adopters/testers of genome-wide technologies, with a focus on what was at the time, a bleak field of Parkinson’s disease (PD) genetics.

Proposed etiology of PD circa 1997 - 2002
Early GWAS of Parkinson’s disease…

At this point, we only had mono-genic or Mendelian genetic risk factors for Parkinson’s disease related to primarily familial cases.

These were identified through classical linkage analyses and family studies.

Before the initial GWAS were started, there was only a hunch that common SNPs were involved in PD risk.
Early GWAS of Parkinson’s disease…

GWAS was designed to target common variants for common diseases, outside of previous high and moderate risk studies focused on rare and relatively rare genetic variants.

At the time, not logistically feasible to conduct population studies of rare SNPs.
Our lab conducted the two first GWAS studies of Parkinson’s disease. The 2006 study was severely underpowered and found nothing. Data was expanded and combined with a series of German samples to increase the sample series to almost 2000 cases and 4000 controls. The 2009 study used more refined methods and found MAPT and SNCA SNPs associated with PD after multiple test correction. Imputation “guesses” genotypes based on more densely genotyped references, primary means of standardizing data genotyped on different arrays but also increases density and power for association studies.
SNCA, LRRK2, PARK2, PINK1, DJ1

SNCA, MAPT

GBA, LRRK2

Complex, low risk 2008-2011
Exercise 1 begins now!

We are going to impute some data.

First let's make a directory in your data directory called 'biowulfClass'.

Next let's copy the following files from the /data/classes/nalls directory to your new data directory:

- cohortA
- cohortB
- ChunkChromo
Exercise 1.

Then let's run the following code from the ChunkChromo directory...

● `tar -zxvf generic-ChunkChromosome-2011-08-05.tar.gz`
● `cd ./generic-ChunkChromosome`
● `make all`

This will compile the ChunkChromosome utility in your directory.

Also run the following…

● `module load nalls-class`

This loads some additional utilities and should be loaded every time you want to repeat this workflow we will be going over today.
Exercise 1.

Directories ‘cohortA’ and ‘cohortB’...

● 2 cohorts of simulated alzheimer’s disease data
● In interest of time, only chromosome 19 not the whole genome
● cohort A - 400 samples, 1:1 case to control ratio, ~9500 SNPs, PLINK format binary files
● cohort B - 500 samples, 1:1 case to control ratio, ~4700 SNPs, PLINK binary files
Exercise 1.

Both cohort A and cohort B also contain directories called ‘CustomImputationScripts’...

- The shell script ‘RunScripts.sh’ executes all 7 other scripts to go from formatting to imputed data in a few hours to a few days depending on the size of your dataset
- Steps 1-5 are formatting
- Step 6 calls the imputation and generates swarms
- impute-biowulf.pl contains imputation parameters
Exercise 1.

Let's take a look at the shell scripts...

- These are set up as shell scripts but could easily be run as swarms for steps 1-5 in sequence.
- `$USER` in the file paths specifies your data directory automatically.
- Steps 1-5 are just formatting, step 6 starts the real work.
- Once you finished finding and replacing, let's discuss `impute-biowulf.pl` (in 5 minutes)
Exercise 1.

impute-biowulf.pl ← an overview

- Imputes from 1000 genomes VCF files as reference
- Default settings of Mach1 and miniMac from the Abecasis Lab at University of Michigan
- 1 core for mach1, 4 cores for miniMac
- Sequential runs of mach1 → miniMac
- logs swarm files and intermediate files in realtime
Exercise 1.

impute-biowulf.pl ← an overview (cont’d)

- lines 11 and 12 should have same path to reference VCF file directory
- lines 13 and 14 allow you to set size of genotyped SNP chunks and their overlap … this really effects processing time and numbers of files generated.
- lines 15 and 16 set the memory … both mach1 and miniMac “spike” memory on biowulf causing the occasional segmentation faults with larger datasets.
- replace lines 28 with 29 and 47 with 48 to run a whole genome
- Set up for autosomes … non-autosomal is quite involved and can be covered via email later.
Exercise 1.

Now it seems you are familiar with what is going on in the scripts and have changed the file paths and options appropriately.

Feel free to run the following

```
cd /data/$USER/biowulfClass/cohortA/CustomImputationScripts/
sh RunScripts.sh > chunking.log

cd /data/$USER/biowulfClass/cohortB/CustomImputationScripts/
sh RunScripts.sh > chunking.log
```

Now let's take a 10 minute break for questions and help!
Current methods: the meta-analysis of GWAS data

By 2008-2010, methods for GWAS became more refined and false positives were less of a plague in the field, replication was easier.

Prices of GWAS arrays dropped slightly per sample, and the density of coverage increased

Principal components analyses were used to adjust for natural population substructure.

More rigorous QC of genotypes.

More rigorous QC of results (examining p-value distributions across genome via QQ plots etc.)

Slightly larger sample sizes

Although cost, time and sample availability were still prohibiting acquiring enough genotyped samples for any disease by any one group to make major gains in statistical power.

Difficult for individual sites/institutions to majorly move past early GWAS sample sizes and people were not finding many new results.
Current methods: the meta-analysis of GWAS data

The obvious solution to combine data across cohorts to increase power to detect new risk loci at minimal new cost (time and $$$).

Pooling data was not possible due to IRBs, sample privacy issues and the fact that different arrays genotyped slightly different SNPs.

This led to the necessity of genotype imputation!
- All common variation is generally correlated to nearby variation – linkage disequilibrium
- Dense genotyping in reference samples from HapMap (2.2 million SNPs) publicly available
- Use your genotyped SNPs to make the best guess at the genotype of nearby SNPs
- HapMap samples of similar continental ancestry as a reference to SNPs not in your study
- Once completed, all studies have larger standardized datasets for analysis

Imputation uses dosages, or non-integer genotypes that are weighted for uncertainty.

Allows for summary statistics from regression models to be combined across studies without sharing participant level data.

Standard meta-analytic techniques similar to clinical epidemiology.
To this end we formed a consortium of investigators with their own cohorts of PD cases and controls with GWAS data. These cohorts included our cohort at NIA, as well as German, French, British, and Dutch collaborators aka The International Parkinson’s Disease Genomics Consortium (IPDGC).

Logistically challenging and time consuming but worth the effort.

We were early adopters of the preliminary 1000 Genomes data (haplotypes) to use as our reference for imputation.

- Over 7 million SNPs from genome-wide sequencing in a number of European ancestry populations from around the world that is publicly available
- Massive standardized datasets
- Drafting of a standardized analysis plan for cohort implementation to ensure compliance
- Increased statistical power due to more samples and denser genotyping
- Fixed-effects meta-analysis of cohort level summary statistics from logistic regression
- Accounting for population substructure at the cohort and meta-analysis level to reduce likelihood of false positives

Also, no samples available for replication, so we needed more genotyping!!!
2 stages of analysis resulting in 2 papers (META1 and META2)

**META1**
- 2 stage design, built in replication (**FAST**)
  - US, UK, French, German, Dutch and Icelandic cohorts, aka the IPDGC
    - Imputed > 7 million SNPs
    - 5333 cases and 12019 controls
    - Meta-analysis
  - Replication via ImmunoChip
    - 7053 cases and 9007 controls
    - Targeted genotyping Loci at $P < 1 \times 10^{-4}$ in discovery phase (>200 loci)

**META2**
- Larger, more powerful than META1, but replication would be external (**SLOW**)
  - Combined meta-analysis of discovery and replication series from META1
  - Suggestive loci were built into the ImmunoChip for this reason
  - Validate new loci with external collaborators
Current methods: the meta-analysis of GWAS data

Initial META1 results from a genome-wide prospective

Imputation of sequence variants for identification of genetic risks for Parkinson’s disease: a meta-analysis of genome-wide association studies

International Parkinson Disease Genomics Consortium*

Parkinson's Disease, Discovery Phase Meta–analysis, Genomic Inflation Factor = 1.035
Current methods: the meta-analysis of GWAS data

Results from META1

<table>
<thead>
<tr>
<th>C</th>
<th>Position (bp)</th>
<th>MAF in discovery phase</th>
<th>Minor/major alleles</th>
<th>Candidate gene</th>
<th>Discovery phase</th>
<th>Replication phase</th>
<th>Combined PAR estimate (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OR (SE) per minor allele dose</td>
<td>Fixed effects p value</td>
<td>OR (SE) per minor allele dose</td>
</tr>
<tr>
<td>chr1:154105678</td>
<td>1 154105678</td>
<td>0.02</td>
<td>T/C</td>
<td>SYT1</td>
<td>1.67(0.09)</td>
<td>1.02×10^-8</td>
<td>1.44(0.08)</td>
</tr>
<tr>
<td>rs6710823</td>
<td>2 135308851</td>
<td>0.19</td>
<td>A/G</td>
<td>AMCSD</td>
<td>1.38(0.05)</td>
<td>1.35×10^-9</td>
<td>1.07(0.02)</td>
</tr>
<tr>
<td>rs2102808</td>
<td>2 168825271</td>
<td>0.13</td>
<td>T/G</td>
<td>STX39</td>
<td>1.28(0.04)</td>
<td>3.31×10^-11</td>
<td>1.12(0.04)</td>
</tr>
<tr>
<td>rs11711441</td>
<td>3 184303693</td>
<td>0.14</td>
<td>A/G</td>
<td>MCCC1/LAMP3</td>
<td>0.82(0.04)</td>
<td>2.10×10^-6</td>
<td>0.87(0.03)</td>
</tr>
<tr>
<td>chr4:913111</td>
<td>4 911311</td>
<td>0.28</td>
<td>C/G</td>
<td>GAK</td>
<td>1.21(0.03)</td>
<td>1.80×10^-12</td>
<td>1.14(0.02)</td>
</tr>
<tr>
<td>rs11724635</td>
<td>4 15346199</td>
<td>0.45</td>
<td>C/A</td>
<td>BST1</td>
<td>0.87(0.03)</td>
<td>1.85×10^-8</td>
<td>0.87(0.02)</td>
</tr>
<tr>
<td>rs356219</td>
<td>4 90856624</td>
<td>0.39</td>
<td>G/A</td>
<td>SNCA</td>
<td>1.30(0.03)</td>
<td>7.90×10^-10</td>
<td>1.27(0.02)</td>
</tr>
<tr>
<td>chr6:32588205</td>
<td>6 32588205</td>
<td>0.15</td>
<td>G/A</td>
<td>MHC-DRB5</td>
<td>0.70(0.06)</td>
<td>2.58×10^-8</td>
<td>0.80(0.04)</td>
</tr>
<tr>
<td>rs1491942</td>
<td>12 38907075</td>
<td>0.21</td>
<td>C/G</td>
<td>ZRRK2</td>
<td>1.19(0.03)</td>
<td>3.23×10^-9</td>
<td>1.30(0.05)</td>
</tr>
<tr>
<td>rs12817488</td>
<td>12 121862247</td>
<td>0.46</td>
<td>A/G</td>
<td>CCDC62/HIP1R</td>
<td>1.16(0.03)</td>
<td>4.43×10^-8</td>
<td>1.13(0.03)</td>
</tr>
<tr>
<td>rs2942168</td>
<td>17 41070633</td>
<td>0.22</td>
<td>A/G</td>
<td>MAPT</td>
<td>0.76(0.03)</td>
<td>1.62×10^-10</td>
<td>0.80(0.03)</td>
</tr>
</tbody>
</table>

Only loci with p<5×10^-8 in the meta-analysis are shown. The SNP with the smallest p value per locus is shown. Webappendix pp 15-31 provide additional details for the associated loci described above. An expanded version of this table that shows all p values less than 1×10^-8 from the main analysis is provided in the Supplementary file.

• 11 genome-wide significant loci, confirming 4 previously implicated loci (blue) and 5 novel (red)
• All replicated successfully
Current methods: the meta-analysis of GWAS data

META1 was successful in identifying multiple new loci and replicating these definitively.

As part of study design, suggestive but not significant loci from the discovery phase of META1 were built into the replication array since there was “room leftover” on the array.
   - These sub-significant loci are the orange regions on the “Manhattan plot”

Using identical meta-analysis techniques, replication and discovery phase samples were combined for overlapping SNPs.
   - 12,386 PD cases and 21,026 controls in total

Although we had officially burnt through all our available samples to replicate anything we found, so new collaborators must be sought out (Do et al., 23&Me)
   - Offered back to back publications in the same journal as a joint submission if both groups exchanged summary statistics for competing papers since they had no replication samples

Essentially META1 and META2 are 2 papers for the price of 1.
<table>
<thead>
<tr>
<th>SNP</th>
<th>Chrom</th>
<th>Gene(s)</th>
<th>Stage 1 OR (95%CI)</th>
<th>Stage 1 P</th>
<th>Stage 2 OR (95%CI)</th>
<th>Stage 2 P</th>
<th>Stage 1+2 OR (95%CI)</th>
<th>Stage 1+2 P</th>
<th>Do et al OR (95%CI)</th>
<th>Do et al P</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs708723</td>
<td>1q32</td>
<td>RAB7L1/PARK16</td>
<td>0.905 (0.862-0.95)</td>
<td>6.68x10^-5</td>
<td>0.863 (0.824-0.905)</td>
<td>9.47x10^-10</td>
<td>1.00x10^-12</td>
<td>0.758 (0.65-0.88)</td>
<td>2.12x10^-6</td>
<td></td>
</tr>
<tr>
<td>rs34016896</td>
<td>3q26</td>
<td>NMD3</td>
<td>1.14 (1.09-1.2)</td>
<td>3.00x10^-7</td>
<td>1.08 (1.02-1.14)</td>
<td>0.00399</td>
<td>1.81x10^-8</td>
<td>1.002 (0.95-1.06)</td>
<td>0.954</td>
<td></td>
</tr>
<tr>
<td>rs6812193</td>
<td>4q21</td>
<td>STBD1</td>
<td>0.886 (0.843-0.932)</td>
<td>2.52x10^-6</td>
<td>0.906 (0.864-0.95)</td>
<td>5.29x10^-5</td>
<td>7.46x10^-10</td>
<td>0.839 (0.79-0.89)</td>
<td>7.55x10^-10</td>
<td></td>
</tr>
<tr>
<td>rs156429</td>
<td>7p15</td>
<td>GPNMB</td>
<td>0.894 (0.849-0.942)</td>
<td>2.15x10^-5</td>
<td>0.893 (0.852-0.937)</td>
<td>3.86x10^-6</td>
<td>3.27x10^-10</td>
<td>0.901 (0.85-0.95)</td>
<td>0.000193</td>
<td></td>
</tr>
<tr>
<td>rs591323</td>
<td>8p22</td>
<td>FGF20</td>
<td>0.884 (0.836-0.935)</td>
<td>1.59x10^-5</td>
<td>0.875 (0.83-0.923)</td>
<td>8.49Ex10^-7</td>
<td>7.45x10^-11</td>
<td>0.932 (0.88-0.99)</td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td>chr8:89442157</td>
<td>8q21</td>
<td>MMP16</td>
<td>1.38 (1.21-1.57)</td>
<td>1.10x10^-6</td>
<td>1.29 (1.12-1.49)</td>
<td>0.000451</td>
<td>2.26x10^-9</td>
<td>0.969 (0.86-1.09)</td>
<td>0.589</td>
<td></td>
</tr>
<tr>
<td>rs4889603</td>
<td>16p11</td>
<td>STXB1</td>
<td>1.12 (1.06-1.18)</td>
<td>4.13x10^-5</td>
<td>1.15 (1.1-1.21)</td>
<td>8.21x10^-9</td>
<td>2.66x10^-12</td>
<td>1.070 (1.01-1.13)</td>
<td>0.014</td>
<td></td>
</tr>
</tbody>
</table>

- 7 genome-wide significant loci, confirming 3 implicated loci (blue) and 4 novel (red)
- 5 independently replicated (MMP16 and NMD3 problematic)
Current methods: the meta-analysis of GWAS data

A Two-Stage Meta-Analysis Identifies Several New Loci for Parkinson’s Disease

International Parkinson’s Disease Genomics Consortium (IPDGC), Wellcome Trust Case Control Consortium 2 (WTCCC2)

SNCA, LRRK2, PARK2, PINK1, DJ1

Complex, low risk 2011-2012

SNCA, LRRK2, MAPT, PARK16, BST1, HLA-DRB5, GAK, ACMSD, STK39, LAMP3, SYT11, HIP1R, FGF20, STX1B, STBD1, GPNMB
Exercise 2.

Your data may have finished imputing but likely it did not.

Anyways, it would be good to download the directory 'imputedData' into your working directory ‘biowulfClass’

ImputedData contains the following:
2 directories one for cohortA one for cohortB
Within each subdirectory is all you will need to run cohort level analyses. These include *.dose, *.info, *.ped and *.dat files
Exercise 2.

Within ./imputedData, we have the following for each cohort:

- *.dose and *.info files (3 for the more densely genotyped cohortA, only 1 for cohortB). These are the main output of the pipeline we ran in Exercise1.
- *.dose is the dosages of alleles per SNP estimated using the 1000 genomes reference data.
- *.info is the quality control output from miniMac detailing alleles, frequencies and imputation qualities.
- *.ped is the phenotype info for mach2dat, which we will use to run logistic regression models. This includes affection status and two PCA derived covariates. Note, these are genome-wide covariates, not specific to this chromosome.
- *.dat denotes contents of the *.ped file
Exercise 2.

Let's get started with the regression modeling…

```
cd to /data/$USER/biowulfClass/imputedData
swarm -f mach2dat.swarm -g 85 --module nalls-class --R gpfs
```

What this will do is run regressions for each chunk under the following model:

\[ \text{Alz} \sim \text{SNP}[i..j] + \text{C1} + \text{C2} \]

Where each SNP dosage is tested individually for an association with Alzheimer status while controlling for population structure (covariates C1 and C2).

This swarm should take 5-10 minutes. So let's handle some questions now.
Exercise 2.

Your swarm should have finished running by now. If not, please download the appropriate results files from the ‘regressionResults’ directory in the class’ folder, place this in a directory called `biowulfClass` on your local machine.

Everything from here on out will be able to be done on a local machine just as it would on a single cluster node.

Right now you should have 3 `*.results` files for cohort A and 1 for cohort B, each corresponding to the chunks.

You can use `head -100 *.results` to check logs, also use `tail * results` to check that everything completed running.
Exercise 2.

Now let’s begin extracting and filtering data. We will pull data based on the phenotype keyword ‘Alz’ and aggregate across all chunks.

This will be the first step from going from cohort level summary stats to meta-analysis results.
Exercise 2.

```bash
cd ./imputedData/cohortA

cat *.results | grep -w 'Alz' | grep ',' | grep -v ']' | grep -w -v 'NA' | sed -e 's/l/l/ I, l/g' -e 's/R,D/ R,D/g' -e 's/R,l/ R,l/g' -e 's/R/I/ R/I/g' -e 's/R,R/' R/R/g' -e 's/l,l/ l,l/g' > ../rawResultsCohortA.txt

cd ../
cd ./imputedData/cohortB

cat *.results | grep -w 'Alz' | grep ',' | grep -v ']' | grep -w -v 'NA' | sed -e 's/l/l/ I, l/g' -e 's/R,D/ R,D/g' -e 's/R,l/ R,l/g' -e 's/R,I/ R,I/g' -e 's/R,R/' R/R/g' -e 's/l,l/ l,l/g' > ../rawResultsCohortB.txt

This is contained in the shell script ExtractData.sh and can be run on genome-wide scale!

Run this from ./biowulfClass/imputedData on your local machine
```
Exercise 2.

What ExtractResults.sh does…

1. concatenates all chunked results per cohort
2. pulls results for trait of interest (you can analyze more than 1 trait at a time, longest ‘time thief’ is loading data)
3. removes monomorphic variants
4. removes poor model fit variants
5. reformats spacing to deal with some variant naming issues
6. splits alleles (column 3 is now column 3 and 4)

After this, results are ready for some quick QC before meta-analyses begin.
Excercise 2.

The R-script ‘formatResults.R’ is very basic and will do the following minimal QC for your cohorts.

1. attach headers to data
2. filter based on imputation quality
3. filter based on minor allele frequency
4. filter based on impossible effect estimates

There are any number of utilities/programs/packages out there that will do similar.

For this dataset, formatting the results should take only a minute or so.

Let's take a few minutes to open formatResults.R and go over the contents.
Excercise 2.

Now, run `R CMD BATCH formatResults.R` from your `/imputedData` directory and check the log.

At this point we have two results sets that have undergone basic formatting and QC. One for Cohort A and one for Cohort B.

In large consortia, usually these analysis, QC and formatting will be outlined for study level analysts by consortia guidelines and analysis plans.

Next step, meta-analysis using METAL - a meta-analysis package also from the Abecasis Lab at University of Michigan.

But first, lets go over MEGA-analyses.
Current mega-analyses

The current lack of funding for more samples but the desire to find more risk loci have spawned a trend towards “mega-analyses”.

In a mega-analyses, formerly competing groups combine samples using meta-analyses of summary statistics as before, but on an even larger scale.

Similar cross study harmonization must occur.
  - Uniform analysis plans
  - Identical statistical models
  - Compatible imputation procedures
  - Data transfer, storage and management issues

Currently, all NINDS funded groups interested in PD GWAS are conducting a mega-analysis for all samples with genome-wide data.

We have employed an identical strategy as set forth in META1 and META2.

We have designed a new replication array with sub-significant associations tagged on the replication array (NeuroX, more on that later) in addition to SNPs necessary to replicate the mega-analysis
  - This will be essentially META3 and META4
  - Strategies for replication of META4 will be “interesting”
Current mega-analyses

Basic premise
• Collaboration between competing groups to achieve largest possible meta-analysis of PD GWAS data.
• Total sample size > 13K cases and > 82K controls
• Based on 1K Genomes Project haplotypes to successfully impute ~11 million SNPs
• Standard methods used to generate and combine summary statistics from GWAS across studies in a more conservative fixed-effects model
• Also tested liberal method of meta-analysis (RE2) in addition to the commonly used fixed effects model

Two methods of meta-analyses were utilized:

1. Fixed effects as per our previous work
   - Conservative compared to method below, it is the common standard of GWAS

2. RE2 – Random effects modeling using the Han & Eskin method
   - More “liberal” than fixed effects
   - Good for large numbers of studies
   - Flexible, as it allows for a few strong p-values to drive a SNP to significance in spite of weak results from smaller studies
<table>
<thead>
<tr>
<th>Study</th>
<th>Abbrev</th>
<th>N Cases</th>
<th>N Controls</th>
<th>Total N</th>
<th>mean AAO</th>
<th>%Male Cases</th>
<th>%Male Controls</th>
<th>Markers</th>
<th>Markers Passing QC</th>
<th>λRaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ashkenazi</td>
<td>AJ</td>
<td>268</td>
<td>178</td>
<td>446</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>11572500</td>
<td>7241832</td>
<td>1.006</td>
</tr>
<tr>
<td>deCODE</td>
<td>DC</td>
<td>604</td>
<td>4916</td>
<td>5520</td>
<td>TBD</td>
<td>TBD</td>
<td>TBD</td>
<td>11572501</td>
<td>6698963</td>
<td>1.061</td>
</tr>
<tr>
<td>Dutch</td>
<td>NL</td>
<td>744</td>
<td>2019</td>
<td>2763</td>
<td>55.3</td>
<td>63.60%</td>
<td>43.82%</td>
<td>11217965</td>
<td>7576956</td>
<td>1.061</td>
</tr>
<tr>
<td>France</td>
<td>FR</td>
<td>985</td>
<td>1984</td>
<td>2969</td>
<td>48.9</td>
<td>58.80%</td>
<td>67.00%</td>
<td>11572501</td>
<td>7641834</td>
<td>0.854</td>
</tr>
<tr>
<td>Germany</td>
<td>GER</td>
<td>667</td>
<td>937</td>
<td>1604</td>
<td>56.0</td>
<td>60.20%</td>
<td>52.00%</td>
<td>11210634</td>
<td>7486133</td>
<td>1.025</td>
</tr>
<tr>
<td>HIHG</td>
<td>HIHG</td>
<td>574</td>
<td>619</td>
<td>1193</td>
<td>57.2</td>
<td>63.07%</td>
<td>34.57%</td>
<td>11914767</td>
<td>7613933</td>
<td>0.998</td>
</tr>
<tr>
<td>NGRC</td>
<td>NGRC</td>
<td>1956</td>
<td>1982</td>
<td>3938</td>
<td>58.6</td>
<td>67.74%</td>
<td>38.70%</td>
<td>11914767</td>
<td>8163392</td>
<td>1.013</td>
</tr>
<tr>
<td>NIA</td>
<td>NIA</td>
<td>937</td>
<td>1896</td>
<td>2833</td>
<td>55.9</td>
<td>39.50%</td>
<td>47.20%</td>
<td>11247278</td>
<td>7620408</td>
<td>1.035</td>
</tr>
<tr>
<td>PROGENI/GenePD</td>
<td>PGPD</td>
<td>828</td>
<td>852</td>
<td>1680</td>
<td>62.1</td>
<td>59.90%</td>
<td>39.79%</td>
<td>11914767</td>
<td>7249203</td>
<td>1.009</td>
</tr>
<tr>
<td>UK</td>
<td>UK</td>
<td>1705</td>
<td>5200</td>
<td>6905</td>
<td>65.8</td>
<td>56.70%</td>
<td>50.50%</td>
<td>11272513</td>
<td>7686314</td>
<td>1.034</td>
</tr>
<tr>
<td>23andMe</td>
<td>TTM</td>
<td>4127</td>
<td>62037</td>
<td>66164</td>
<td>TBD</td>
<td>60.58%</td>
<td>59.48%</td>
<td>7840733</td>
<td>7729624</td>
<td>1.212</td>
</tr>
</tbody>
</table>

This high lambda for 23&Me was troubling, and 23&Me rationalized it as inflated due to their use of a 12:1 ratio of cases and controls, lambdas easily rescaled for case:control effect on inflation.
Initial results were very optimistic…

What you see is lots of hits, 39 conservative loci and 143 liberal loci.

METAL, Genomic Inflation Factor = 1.032

RE2, Genomic Inflation Factor = 0.9704

The liberal model hits are a “super-set” of the conservative model, which is reassuring.
Current mega-analyses

Re-analysis of the discovery phase

- One of the cohorts was forced to be split into 2 cohorts
  - Different chips for genotyping
  - Sampling bias
  - Possible imputation batch artifacts
- Analyses were then carried out identically as before on subsetted data

Discovery phase results

- The number of loci identified by the discovery phase was reduced to 26
- Technically 25 loci as one locus was primarily driven by UK data (chr3:87520857)
- Loci were defined as regions of genome-wide significant hits within +/- 250kb of each other
  - Defining loci in this way sets the stage for conditional analyses to identify secondary and tertiary hits at each locus
- Significant loci identified by both conservative and liberal models were identical

Conditional analyses

- For each significant locus, the core 9 studies re-analyzed all SNPs within the 25 significant and high quality loci, adjusting for the top SNP per locus
- Identical statistical models except the additional adjustment for SNPs
- In the first round of conditional analyses, 8 secondary semi-independent loci existed that passed locus specific correction for multiple testing
Luckily all this analysis and re-analysis paid off and we finally published “META3” using replication from the NeuroX array we designed in house (grey loci failed replication)!
In total, we identified and replicated 6 new loci and confirmed an additional 22 suspected loci...

| rs37549011 | rs823118 | rs10797576 | rs6430558 | rs1447055 | rs115185653 | rs12637471 | rs343118664 | rs117348356 | rs8612193 | rs359182 | rs8275352* | rs199347 | rs117896735 | rs3793947 | rs329648 | rs76904798 | rs11068010 | rs11158026 | rs1555399* | rs2414739 | rs14235 | rs17645553 | rs12456492 | rs6210676* | rs8118008* |
|-----------|----------|------------|-----------|-----------|-------------|----------|------------|------------|----------|--------|-----------|---------|-------------|----------|--------|-------------|------------|----------|-----------|---------|----------|---------|
| GBA/SYT11  | RAB7L1/NUCKS1 | SIPA1L2  | ACMSD/TMEM163 | STX9 | KRT8P25/APOPO2 | MCCC1 | TEMEM175/GAVDGKQ | BST1 | DGK2 | MIR6497 | HLACQ1 | GPMB | INPF5F | DLG2 | MIR4459 | LLRK2 | CDDD2 | GCH1 | TMEM229B | VPS13C | BCKDK/STX1B | MAPT | DRRK1 |
| a          | g         | c          | t          | c          | g            | t        | t           | t          | t        | g       | t       | t       | t        | g        | t       | g           | t          | t       | t         | t        | t       | t        | t       |
| 0.017      | 0.559    | 0.14       | 0.43       | 0.128     | 0.035        | 0.193    | 0.809       | 0.563      | 0.384    | 0.633   | 0.094   | 0.59    | 0.014     | 0.545    | 0.354   | 0.059        | 0.558      | 0.335    | 0.468     | 0.734    | 0.325    | 0.138    | 0.319    
| 0.666      | 0.57     | 0.13       | 0.873      | 0.93       | 0.982       | 0.844    | 0.784       | 1.116      | 0.708    | 0.73    | 0.227   | 0.116   | 1.211     | 0.92    | 0.095   | 1.1          | 0.57      | 0.811    | 0.56      | 0.265    | 0.059    | 0.512    | 0.019    
| 1.762      | 1.126    | 0.139      | 0.873      | 1.214      | 1.768       | 0.842    | 0.784       | 1.122      | 0.708    | 0.73    | 0.227   | 0.116   | 1.211     | 0.92    | 0.095   | 1.1          | 0.57      | 0.811    | 0.56      | 0.265    | 0.059    | 0.512    | 0.019    
| 0.057      | 0.016    | 0.023      | 0.017      | 0.024      | 0.018       | 0.021    | 0.017       | 0.016      | 0.023    | 0.017   | 0.024   | 0.018   | 0.017     | 0.021    | 0.017   | 0.017        | 0.024      | 0.017   | 0.017     | 0.024   | 0.017   | 0.017   | 0.017    
| 6.069×10⁻⁶ | 1.36×10⁻⁶ | 1.19×10⁻⁸ | 5.56×10⁻¹⁵ | 7.12×10⁻⁸ | 3.16×10⁻⁷   | 2.58×10⁻⁹ | 3.37×10⁻⁸   | 8.07×10⁻⁷ | 7.17×10⁻⁷ | 3.23×10⁻⁷ | 5.82×10⁻⁶ | 1.05×10⁻⁹ | 1.72×10⁻⁷ | 2.58×10⁻⁹ | 3.37×10⁻⁸ | 1.15×10⁻⁷ | 7.17×10⁻⁷ | 3.23×10⁻⁷ | 5.82×10⁻⁶ | 1.05×10⁻⁹ | 1.72×10⁻⁷ | 2.58×10⁻⁹ | 3.37×10⁻⁸ | 1.15×10⁻⁷ |
META4 is currently underway and we are busy looking for additional replication samples, but in summary we can conclude...

Etiology of PD, 2013-2014

GENETICS

Environment? More genes to find?
Exercise 3.

Let's run METAL…

- Fixed-effects meta-analysis. The standard for GWAS discovery efforts.
- Always quantify heterogeneity across studies.
- Always meta-analyze effect estimates, not p-values.
- Use genomic control to adjust for inflation when running genome-wide but never on loci or single chromosomes in general.
- It's also a bit faster than R.
- “Industry standard”
Exercise 3.

First, make sure the script ‘RunMetal.txt’ is copied to your /imputedData directory.

Let's now take a minute to go over the contents of the script and answer any questions.

Also, please copy your metal executable to the /imputedData directory.

Then just run `metal RunMetal.txt > Metal.log` in the standard command line from the /imputedData directory.
Exercise 3.
Let's briefly take a minute to go over the log file and the additional descriptive file ‘METAANALYSIS1.TBL.info’.

The *.TBL.info file describes the contents of the columns in the actual meta-analysis file.

The log file shows that you have a “hit” that breaks genome-wide significance (usually regarded as a p-value < 5E-8).

Let's now investigate your actual meta-analysis results using the R-script ‘PostProcess.R’ that can also be found in the /imptuedData directory. You can run this using ‘R CMD BATCH PostProcess.R’.

When this is finished in 2 minutes, let's go over the output.
Exercise 3.

Now you have “rediscovered” the APOE locus that is the best known risk factor for Alzheimer’s…

● You will only see genome-wide significant hits in Cohort A, because Cohort B’s genotyping did not have sufficient density to tag enough of the risk haplotype at this locus. Although they do tag the very edges of the haplotype as evidenced in your list of “candidates”

● You have done what most large-scale GWAS consortia do in one morning, just on a limited scale and with less conference calls and paper work.

● You have learned some of the subtleties of this type of data analysis as well as received code that can be slightly modified prior to applying to your own data.

● You should be ready to begin your own analyses now building on this foundation!
Risk profiling

With META1 and META2 we were able to identify a total of 16 replicated loci in the span of 18 months. With META3 almost 2 years later we have increased to 28, and 30 if you include rare variants in LRRK2 and GBA (0.1-1% of cases).

Common SNPs and even some rare variants with low risk estimates don’t mean much by themselves to the average scientist.

We employed risk profiling to quantify cumulative risk attributable to all of these variants of interest.

Summing of the total number of known risk alleles per sample.

Scale risk allele counts by specific variant’s reported odds ratio
  - SNPs don’t all have the same effect
  - More realistic and specific model
  - More appropriate than population attributable risk (PAR) for SNPs of variable frequencies
  - Apply to cohort(s) not used in discovery
Risk profiling

Trend → a genetic risk profile score greater than 1 s.d. from the population mean, indicative of a roughly 34% increase in genetic risk score above the mean for controls, had a significantly higher risk of Parkinson's disease (OR = 1.51, \( P = 2 \times 10^{-16} \)).

Outliers → fifth quintile of genetic risk scores to the first quintile of genetic risk as a reference; the OR was 3.31 (\( P = 2 \times 10^{-16} \)).

These OR estimates are larger in comparison to those in earlier publications and might be due to the finer-scale imputation in META3, as well as to the inclusion of additional loci and, to some degree, differing distributions of cumulative genetic risk scores across populations in the analysis.
Risk profiling

Risk profiling, machine learning and similar risk prediction is of high interest and a hot topic for the foreseeable future.

We have fit these model parameters from our cross-sectional GWAS data to the Michael J Fox Foundation’s PPMI study (http://www.ppmi-info.org/), with an almost 10% increase in predictive power compared to previous modeling effort (from an AUC or ~63% to 72%).
Functional inferences from methylation and expression data

So at this point we have discovered a number of risk SNPs associated with PD. In general we discuss the most significant SNP in a region, we need to remember these are actually loci, anywhere from a handful to thousands of correlated proximal SNPs all associated with disease to some degree.

We should really think in terms of loci and not simply genes or SNPs!

And there is biology that we can try to understand within these loci!
Functional inferences from methylation and expression data

300 Brains

Frontal Cortex

Cerebellum

1. GW Imputation
   >5 million SNPs passing QC

2. Expression
   >22,000 transcripts

3. CpG Methylation
   27,000 sites

Allows for inference of biological processes at risk loci.

Expression relates to activity at the locus.

Methylation relates to estimates of “regulation”.

We are interested in concordance between risk and/or one of these two factors for the same allele at the same SNP.
Functional inferences from methylation and expression data

The left column shows the concordance between meta-analysis effect estimates and QTL effect estimates for SNPs at five loci with significant QTL associations.

The right column shows the position of significantly associated SNPs from the QTL analyses within every region of interest.

Heritability of risk

Now that we have identified all of these risk loci based on SNPs from GWAS studies, where does that leave us in terms of heritable risk.

Recent methods have been developed to estimate heritability in ostensibly outbred populations based on low levels of background “relatedness” within the population.

- mixed model
- maximum likelihood
- GCTA method (http://www.complextraitgenomics.com/software/gcta/)

We decided to compare genome-wide heritability versus heritability at GWAS identified loci across IPDGC cohorts

- estimate heritability
- meta-analyze estimates across cohorts (random-effects)
- identify heritability missed by GWAS (i.e. the difference between genome-wide and locus-specific estimates of variance explained by SNPs)
Heritability of risk

In the simplest terms, GWAS identified loci only account for ~3% of PD risk, but it is estimated that at least ~25% more of total PD risk is attributable to genetics.
Where we stand now…

At this point we know linkage and GWAS methods are missing something.
Where we stand now…

- Cost drops in sequencing have allowed us to begin investigating rare but larger effect variants using a variety of technologies to chase this “missing heritability” of disease.
- Exome arrays are a particularly cost effective tool for this, although analytic methods are “under construction” but improving rapidly.
The future direction…

Introducing NeuroX, exome sequencing / arrays and whole genome sequencing

The NeuroX array arose out of the need to custom genotype a multitude of markers for the replication of hits from the Mega-meta project.

It was almost as cheap to add custom content to an existing exome array as it was to build a custom array.

With the availability of cost effective custom content to supplement current exome arrays from Illumina, the idea basically presented itself as replication for META3 with free exome content.

1% of genome that is protein coding is focus!

We opted to utilize 30K bead types to add to the Illumina Exomev1.1 array to cover primarily Mega-meta replication (10K beadtypes) as well as other neurodegenerative diseases.

Full exome sequencing based variants standard to the Illumina exome array (242901 variants) and neurological and neurodegenerative disease focused content that may be added to other existing arrays (24706 variants).

This array covers a majority of easily assay-able coding variation in the genome at a fraction of the price of sequence-based data with major attention to rare variants.

This will be a supplement to current exome sequencing projects underway
The future direction, a focus on rare variants…

Coverage of genome by variants included on the NeuroX array. As a note, chromosome 23 is the X chromosome, 24 is the Y chromosome, 25 is the pseudoautosomal XY region and 26 represents mitochondrial DNA.
The future direction…

Currently analyses focus on gene burden tests
- genes enriched for more rare variants in cases compared to controls
- a “crutch” for the low statistical power related to testing rare variants by themselves
- lower penalty for multiple testing based on ~20K genes instead of 200K single variants

Similar a-hypothetical paradigm as GWAS

Test all genes to see if the cumulative burden of rare variants for a gene is enriched in cases
- all variants below a certain minor allele frequency
- only coding variants

Tests include
- T1, an enrichment of variants below 1% minor allele frequency enriched in cases
- T5, an enrichment of variants below 5% minor allele frequency enriched in cases
- SKAT, sequence kernel association test which scores variant loadings per gene based on variance-components and can be applied to a variety of frequencies (bidirectional test)
The future direction…

Burden and single variant testing is underway in ~6K cases and 6K controls assayed on NeuroX

Modeling parameters for different burden tests are being fine-tuned
  - weighting parameters based on annotation for predicted damaging effects and frequency
  - variant classes included in model (all variants, nonsynonymous coding changes, loss of function, etc)
  - frequency spectrum being analyzed

Analytic framework
  - single variant and burden tests are stratified by ancestry (i.e. USA, UK, Germany, Greece or France) then meta-analyzed
  - all analyses adjusted for principal components derived from SNPs outside of known PD risk loci to account for population substructure and reduce likelihood of false positive indicated by lambda inflation
  - pooled analyses when possible
  - meta-analyses by cohort to assess heterogeneity
The future direction...

Proof of concept:
From T5 test including only non-synonymous coding changes, GBA is one of the most significant associations as expected. Published associations from targeted sequencing studies have show the same results and similar effect estimates ... In addition to ~20 new candidate genes for further study. This will supplement exome sequence data being aggregated at the moment.
Congratulations!

Thanks for your participation. I’m happy to discuss this subject matter anytime, please feel free to email m.nalls.working@gmail.com.

Also, if you or someone you know is interested in a fellowship in analytics / biostats / genetic epidemiology, please let me know via email because I’m hiring!