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. ***Excercise 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 $P < 1e\text{-}7$ 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.

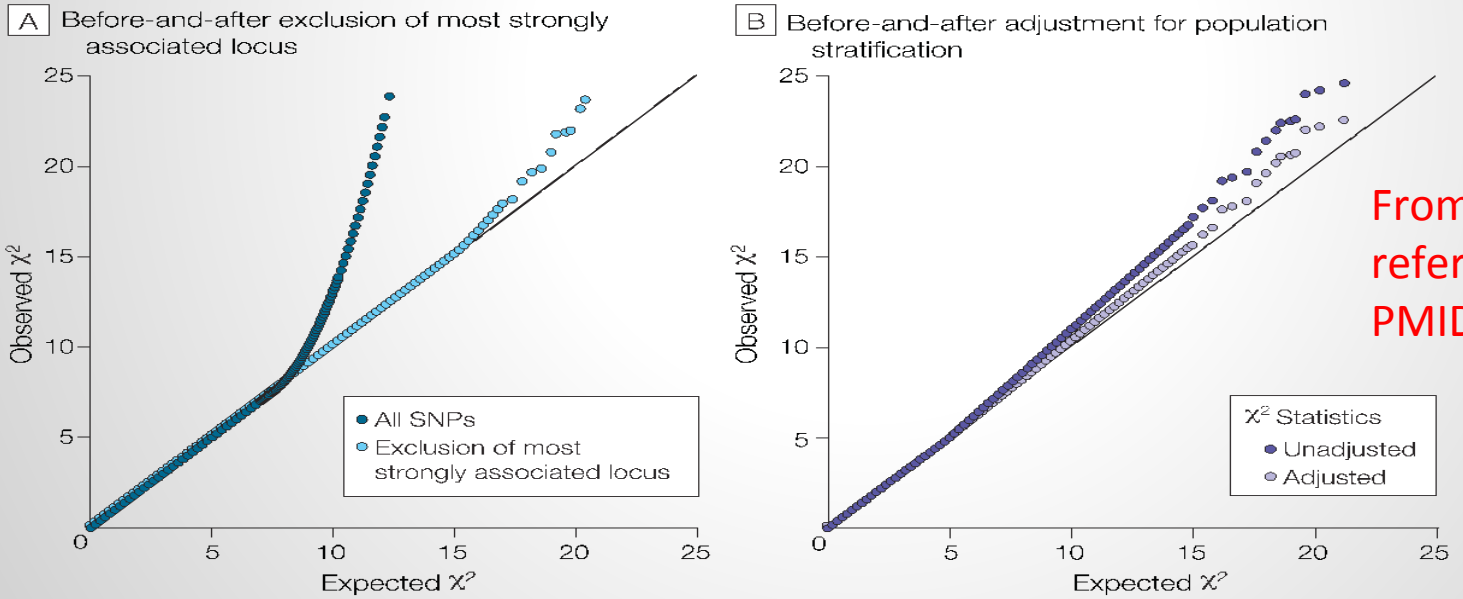
An introduction to genome-wide approaches in genetic epidemiology…

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

Figure 1. Hypothetical Quantile-Quantile Plots in Genome-wide Association Studies



From a great reference article PMID:18349094

Early GWAS of Parkinson's disease…

Parkinson disease

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

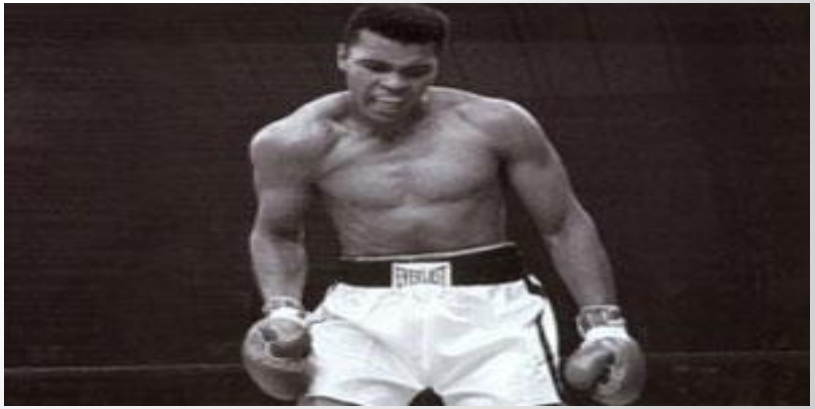
- Dementia in 30-50% of cases

Unknown environmental etiology

More common in males

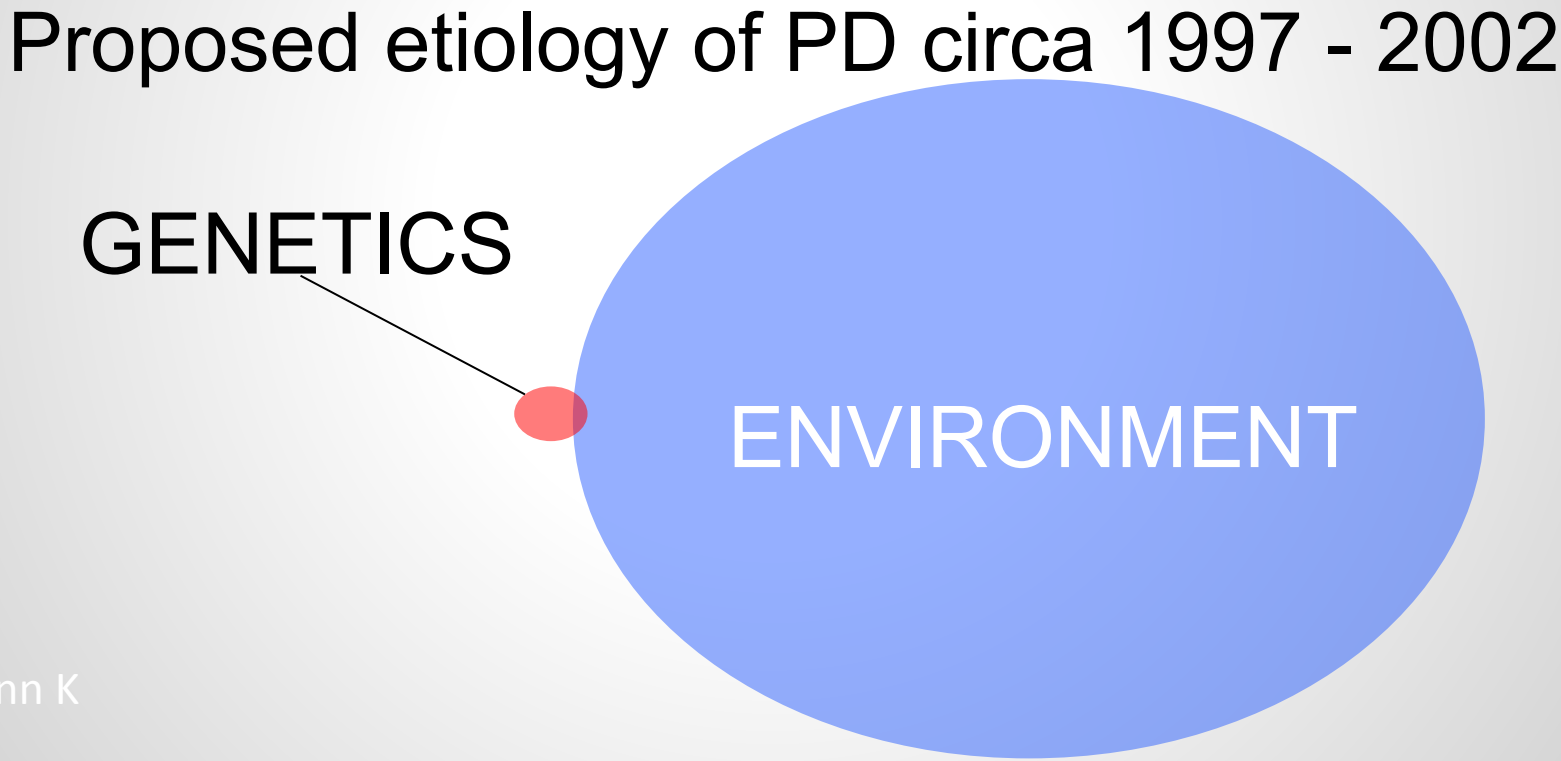
Less common in smokers and coffee drinkers

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.

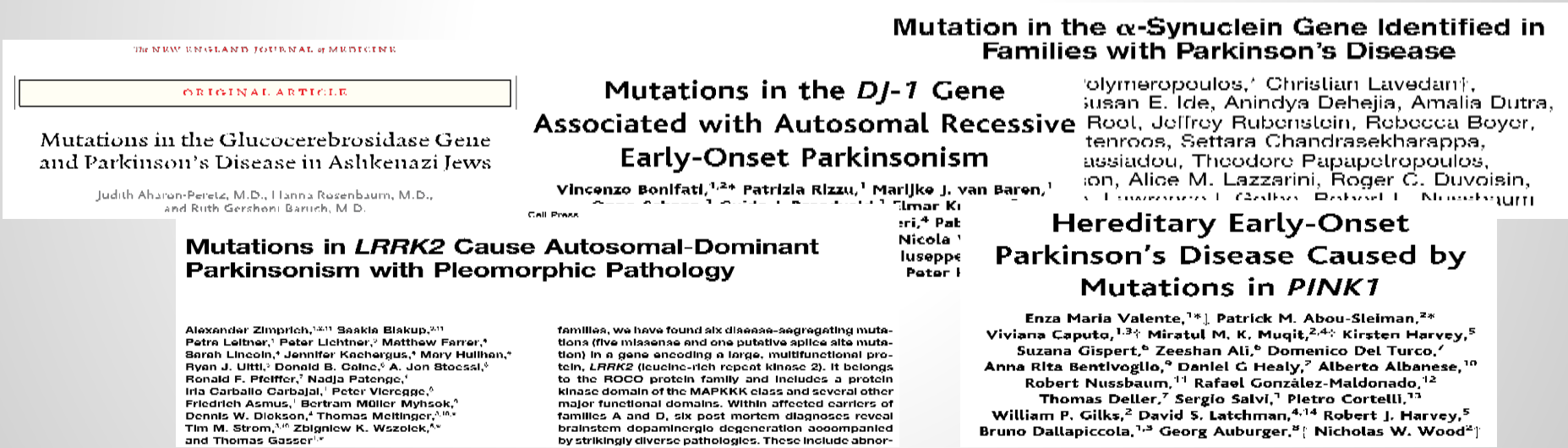


Early GWAS of Parkinson's disease…

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.



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Mutations in the Glucocerebrosidase Gene and Parkinson's Disease in Ashkenazi Jews

Judith Aharon-Peretz, M.D., Hanna Rosenbaum, M.D., and Ruth Gershoni-Baruch, M.D.

Mutation in the α-Synuclein Gene Identified in Families with Parkinson's Disease

olymeropoulos,* Christian Lavedan†, usan E. Ide, Anindya Dehejia, Amalia Dutra, Root, Jeffrey Rubenstein, Rebecca Boyer, tenroos, Settara Chandrasekharappa, assiadou, Theodore Papapetropoulos, on, Alice M. Lazzarini, Roger C. Duvoisin,

Mutations in the *DJ-1* Gene Associated with Autosomal Recessive Early-Onset Parkinsonism

Vincenzo Bonifati,[1,2]* Patrizia Rizzu,[1] Marijke J. van Baren,[1]

Cell Press

Mutations in *LRRK2* Cause Autosomal-Dominant Parkinsonism with Pleomorphic Pathology

Alexander Zimprich,[1,2,11] Saskia Biskup,[2,11] Petra Leitner,[1] Peter Lichtner,[2] Matthew Farrer,[3] Sarah Lincoln,[4] Jennifer Kachergus,[4] Mary Hulihan,[4] Ryan J. Uitti,[5] Donald B. Calne,[6] A. Jon Stoessl,[6] Ronald F. Pfeiffer,[7] Nadja Patenge,[1] Ira Carballo Carbajal,[1] Peter Vieregge,[8] Friedrich Asmus,[1] Bertram Müller-Myhsok,[9] Dennis W. Dickson,[4] Thomas Meitinger,[2,10,*] Tim M. Strom,[2,10] Zbigniew K. Wszolek,[4,*] and Thomas Gasser[1,*]

families, we have found six disease-segregating mutations (five missense and one putative splice site mutation) in a gene encoding a large, multifunctional protein, *LRRK2* (leucine-rich repeat kinase 2). It belongs to the ROCO protein family and includes a protein kinase domain of the MAPKKK class and several other major functional domains. Within affected carriers of families A and D, six post-mortem diagnoses revealed brainstem dopaminergic degeneration accompanied by strikingly diverse pathologies. These include abnor-

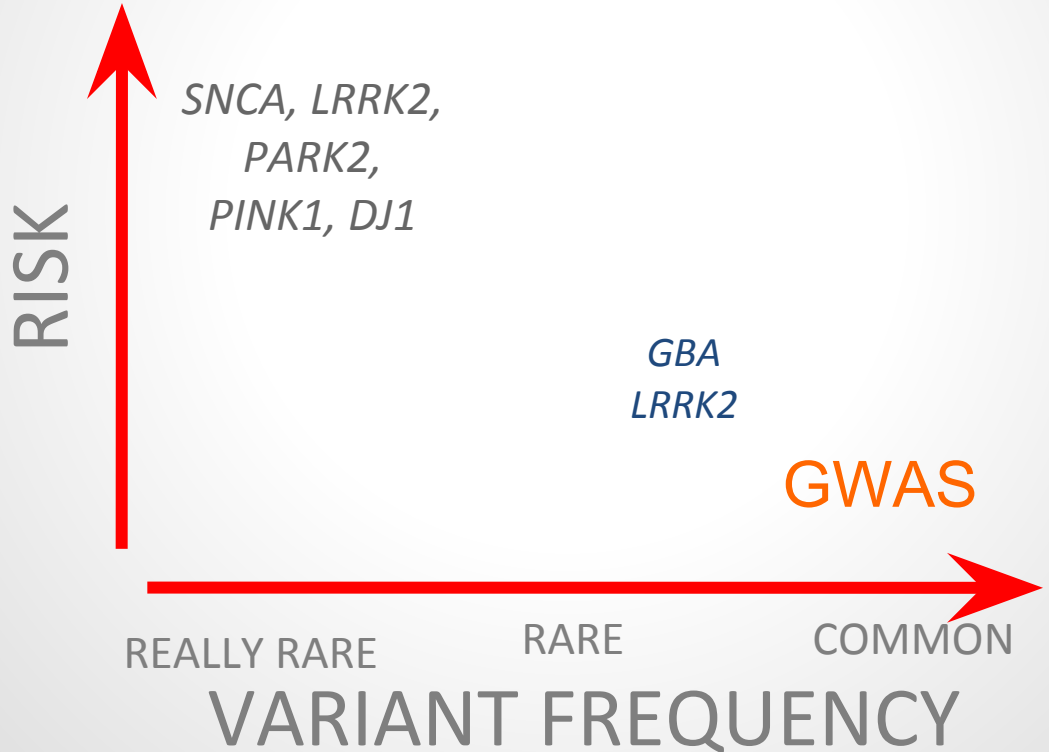
Hereditary Early-Onset Parkinson's Disease Caused by Mutations in *PINK1*

Enza Maria Valente,[1,*] Patrick M. Abou-Sleiman,[2,*] Viviana Caputo,[1,3] Miratul M. K. Muqit,[2,4] Kirsten Harvey,[5] Suzana Gispert,[6] Zeeshan Ali,[6] Domenico Del Turco,[7] Anna Rita Bentivoglio,[8] Daniel G. Healy,[7] Alberto Albanese,[10] Robert Nussbaum,[11] Rafael González-Maldonado,[12] Thomas Deller,[7] Sergio Salvi,[1] Pietro Cortelli,[13] William P. Gilks,[2] David S. Latchman,[4,14] Robert J. Harvey,[5] Bruno Dallapiccola,[1,3] Georg Auburger,[6] Nicholas W. Wood[2]

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



Early GWAS of Parkinson's disease…

2006

Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data

Hon-Chung Fung, Sonja Scholz, Mar Matarin, Javier Simón-Sánchez, Dena Hernandez, Angela Britton, J Raphael Gibbs, Carl Langefeld, Matt L Stiegert, Jennifer Schymick, Michael⁵ Okun, Ronald J Mandel, Hubert H Fernandez, Kelly D Foote, Ramón L Rodriguez, Elizabeth Peckham, Fidelma Wavrant De Vrieze, Katrina Gwinn-Hardy, John A Hardy, Andrew Singleton

2009

LETTERS

nature genetics

Genome-wide association study reveals genetic risk underlying Parkinson's disease

Javier Simón-Sánchez[1,2,22], Claudia Schulte[3,22], Jose M Bras[1,4,22], Manu Sharma[5,22], J Raphael Gibbs[1,3], Daniela Berg[2], Coro Paisan-Ruiz[3], Peter Lichtner[6], Sonja W Scholz[1,3], Dena G Hernandez[1,3], Rejko Krüger[3], Monica Federoff[1], Christine Klein[7], Alison Goate[8], Joel Perlmutter[9], Michael Bonin[3], Michael A Nalls[4], Thomas Illig[10], Christian Gieger[10], Henry Houlden[5], Michael Steffens[11], Michael S Okun[12], Brad A Racette[9], Mark R Cookson[1], Kelly D Foote[13], Hubert H Fernandez[13], Bryan J Traynor[1], Stefan Schreiber[14], Sampath Arepalli[1], Ryan Zonozi[1], Katrina Gwinn[14], Marcel van der Brug[1,15], Grisel Lopez[16], Stephen J Chanock[17], Arthur Schatzkin[17], Yikyung Park[17], Albert Hollenbeck[18], Jianjun Gao[19], Xuemei Huang[20], Nick W Wood[5], Delia Lorenz[21], Günther Deuschl[21], Honglei Chen[19], Olaf Riess[9], John A Hardy[5], Andrew B Singleton[4] & Thomas Gasser[3]

ights reserved.

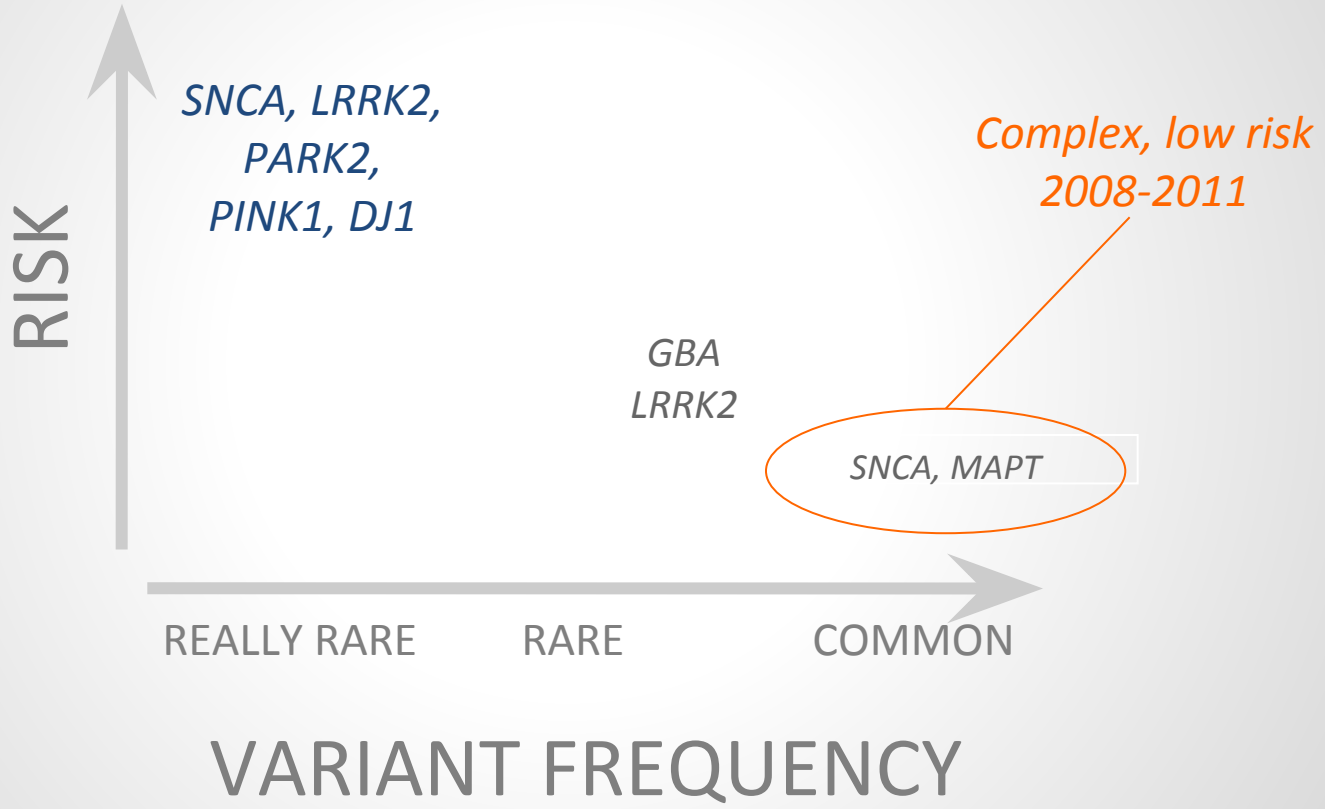
Our lab conducted the two first GWAS studies of Parkinson's disease.

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
Complex, low risk
2008-2011
GBA
LRRK2
SNCA, MAPT
RISK
REALLY RARE
RARE
COMMON
VARIANT FREQUENCY

Exercise 1 begins now!

We are going to impute some data.

First lets make a directory in your data directory called ‘biowulfClass’

Next lets copy the following files from the /data/classes/nalls directory to your new data directory…

- cohortA
- cohortB
- ChunkChromo

Exercise 1.

Then lets 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 damples, 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 imputaiton parameters

Exercise 1.

Lets take a look at the shell scripts…

- These are set up a 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, lets 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 lets 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.

Current methods: the meta-analysis of GWAS data

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 < 1E-4 in discovery phase (>200 loci)

• META2

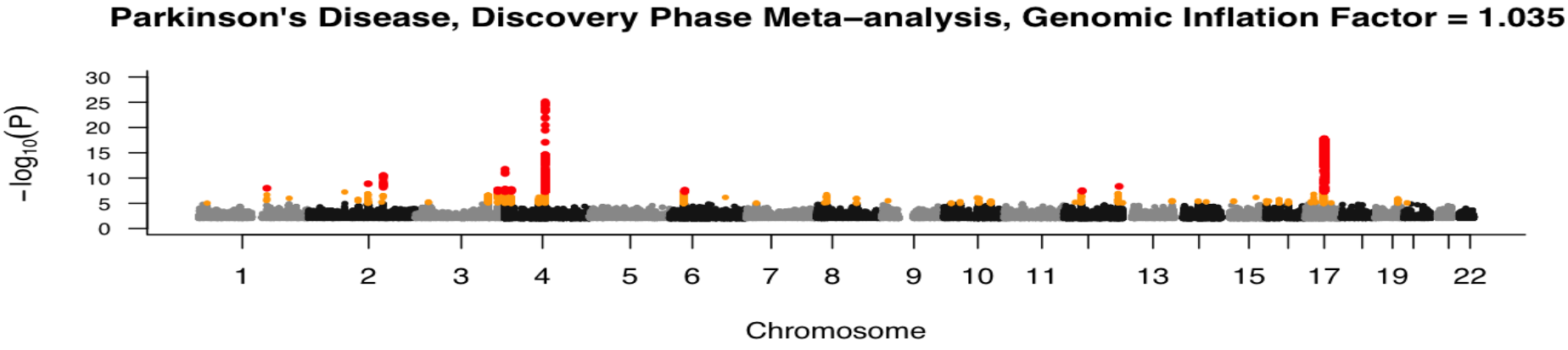
- Larger, more powerful than META1, but replication would be external (SLOW)
  - Combined meta-analysis of discovery and replication series from META1
  - Suggestive loci were built into the ImmunoChip for this reason
  - Validate new loci with external collaborators

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*



Current methods: the meta-analysis of GWAS data

Results from META1

	C	Position (bp)	MAF in discovery phase	Minor/ major alleles	Candidate gene	Discovery phase		Replication phase		Combined PAR estimate (95% CI)
						OR (SE) per minor allele dose	Fixed effects p value	OR (SE) per minor allele dose	Fixed effects p value	
chr1:154105678	1	154105678	0·02	T/C	*SYT11*	1·67 (0·09)	$1·02\times10^{-8}$	1·44 (0·08)	$1·18\times10^{-6}$	1·21% (0·34–1·47%)
rs6710823	2	135308851	0·19	A/G	*AMCSD*	1·38 (0·05)	$1·35\times10^{-9}$	1·07 (0·02)	0·003161	4·05% (1·66–6·82%)
rs2102808	2	168825271	0·13	T/G	*STK39*	1·28 (0·04)	$3·31\times10^{-11}$	1·12 (0·04)	0·001639	2·29% (1·11–2·98%)
rs11711441	3	184303969	0·14	A/G	*MCCC1/LAMP3*	0·82 (0·04)	$2·10\times10^{-8}$	0·87 (0·03)	$6·92\times10^{-5}$	13·71% (9·05–17·70%)
chr4:911311	4	911311	0·28	C/G	*GAK*	1·21 (0·03)	$1·80\times10^{-12}$	1·14 (0·02)	$7·46\times10^{-8}$	4·87% (2·68–6·38%)
rs11724635	4	15346199	0·45	C/A	*BST1*	0·87 (0·03)	$1·85\times10^{-8}$	0·87 (0·02)	$2·43\times10^{-9}$	7·82% (5·30–9·47%)
rs356219	4	90856624	0·39	G/A	*SNCA*	1·30 (0·03)	$7·90\times10^{-26}$	1·27 (0·02)	$4·23\times10^{-23}$	9·71% (6·68–10·27%)
chr6:32588205	6	32588205	0·15	G/A	*HLA-DRB5*	0·70 (0·06)	$2·58\times10^{-8}$	0·80 (0·04)	$9·30\times10^{-8}$	17·68% (11·04–23·00%)
rs1491942	12	38907075	0·21	G/C	*LRRK2*	1·19 (0·03)	$3·23\times10^{-8}$	1·30 (0·05)	$1·06\times10^{-6}$	2·09% (1·00–2·50%)
rs12817488	12	121862247	0·46	A/G	*CCDC62/HIP1R*	1·16 (0·03)	$4·43\times10^{-9}$	1·13 (0·03)	$9·06\times10^{-7}$	5·56% (3·20–7·37%)
rs2942168	17	41070633	0·22	A/G	*MAPT*	0·76 (0·03)	$1·62\times10^{-18}$	0·80 (0·03)	$1·37\times10^{-13}$	17·57% (12·92–20·78%)

Only loci with $p<5\times10^{-8}$ in the meta-analysis are shown. The SNP with the smallest p value per locus on the basis. 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^{-5} from tile upon request. C=chromosome. MAF=minor allele frequency. OR=odds ratio. PAR=population-attributable risk. I^2 index=I^2 index of heterogeneity. I^2 p value=hetero

- 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.

Current methods: the meta-analysis of GWAS data

Results from META2

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

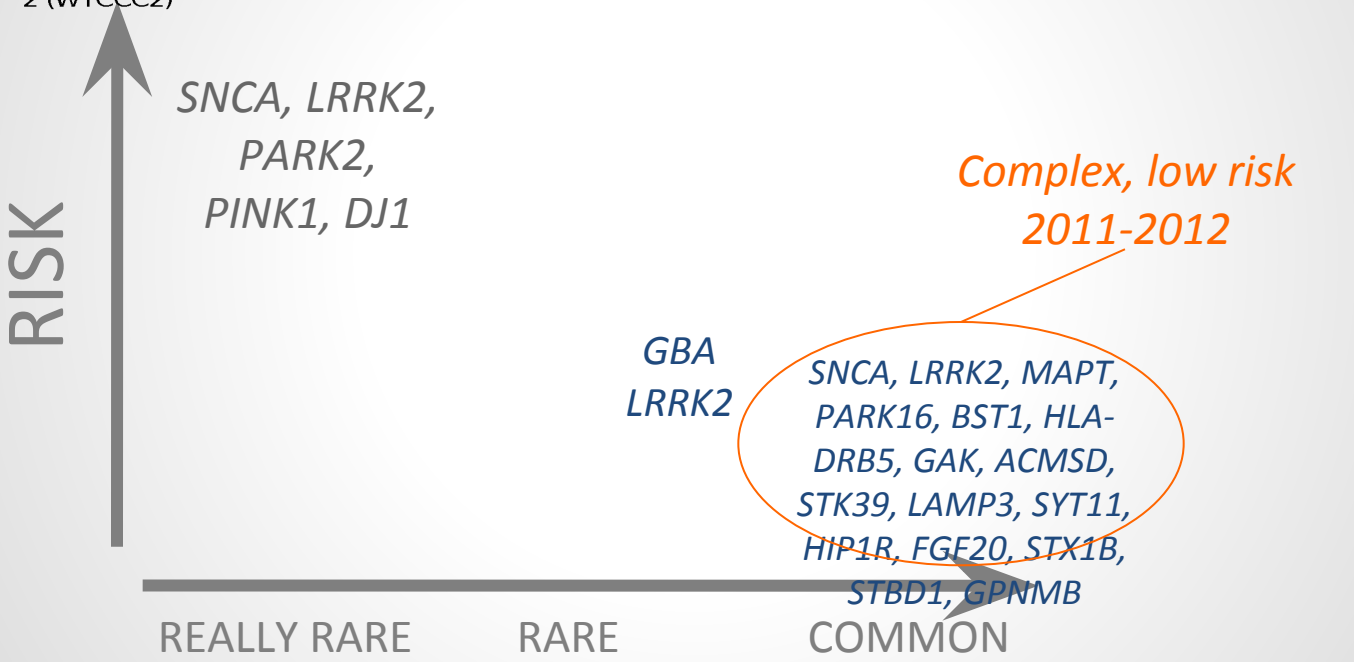
- 7 genome-wide significant loci, confirming 3 implicated loci (blue) and 4 novel (red)
- 5 independently replicated (MMP16 and NMD3 problematic)

Current methods: the meta-analysis of GWAS data

OPEN ACCESS Freely available online PLOS GENETICS

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)"



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.

Lets 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:

Alz ~ SNP[i..j] + C1 + 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 lets 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 lets 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.

cd ./imputedData/cohortA

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

cd ../

cd ./imputedData/cohortB

cat *.results | grep -w ‘Alz’ | grep ',' | grep -v ']' | grep -w -v 'NA' | sed -e 's/I,I/ I,I/g' -e 's/R,D/ R,D/g' -e 's/D,R/ D,R/g' -e 's/I,R/ I,R/g' -e 's/R,I/ R,I/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.

Lets 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 replciation 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:

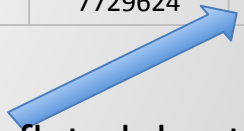
1. Fixed effects as per our previous work
   -Conservative compared to method below, it is the common standard of GWAS

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

Current mega-analyses

Sample Inventory for Mega-meta discovery

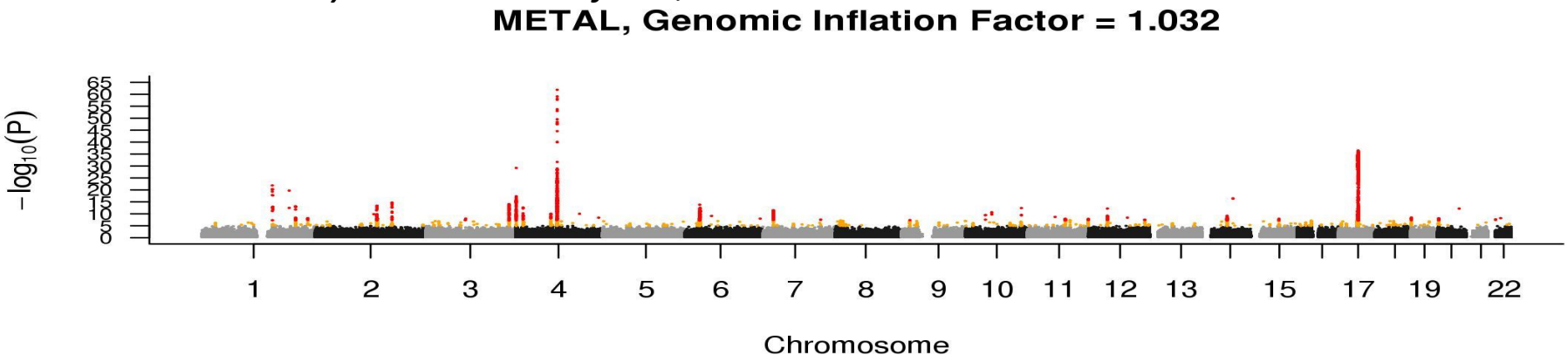
Study	Abbrev	N Cases	N Controls	Total N	mean AAO	%Male Cases	%Male Controls	Markers	Markers Passing QC	λRaw
Ashkenazi	AJ	268	178	446	TBD	TBD	TBD	11572500	7241832	1.006
deCODE	DC	604	4916	5520	TBD	TBD	TBD	11572501	6698963	1.061
Dutch	NL	744	2019	2763	*55.3*	*63.60%*	*43.82%*	11217965	7576956	1.061
France	FR	985	1984	2969	48.9	58.80%	67.00%	11572501	7641834	0.854
Germany	GER	667	937	1604	*56.0*	*60.20%*	*52.00%*	11210634	7486133	1.025
HIHG	HIHG	574	619	1193	57.2	63.07%	34.57%	11914767	7613933	0.998
NGRC	NGRC	1956	1982	3938	58.6	67.74%	38.70%	11914767	8163392	1.013
NIA	NIA	937	1896	2833	*55.9*	*39.50%*	*47.20%*	11247278	7620408	1.035
PROGENI/GenePD	PGPD	828	852	1680	62.1	59.90%	39.79%	11914767	7249203	1.009
UK	UK	1705	5200	6905	65.8	56.70%	50.50%	11272513	7686314	1.034
23andMe	TTM	4127	62037	66164	TBD	60.58%	59.48%	7840733	7729624	1.212

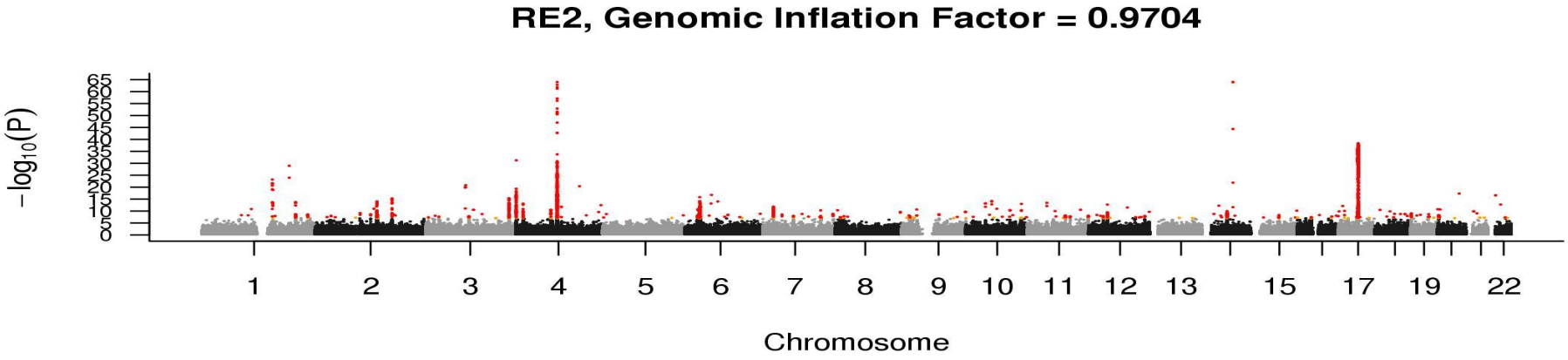


This high lambda for 23&Me was troubling, and 23&Me rationalized it as inflated due to their use of a 12:1 ratio of cases and controls, lambdas easily rescaled for case:control effect on inflation.

Initial results were very optimistic…

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





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)**!**

Accepted Manuscript

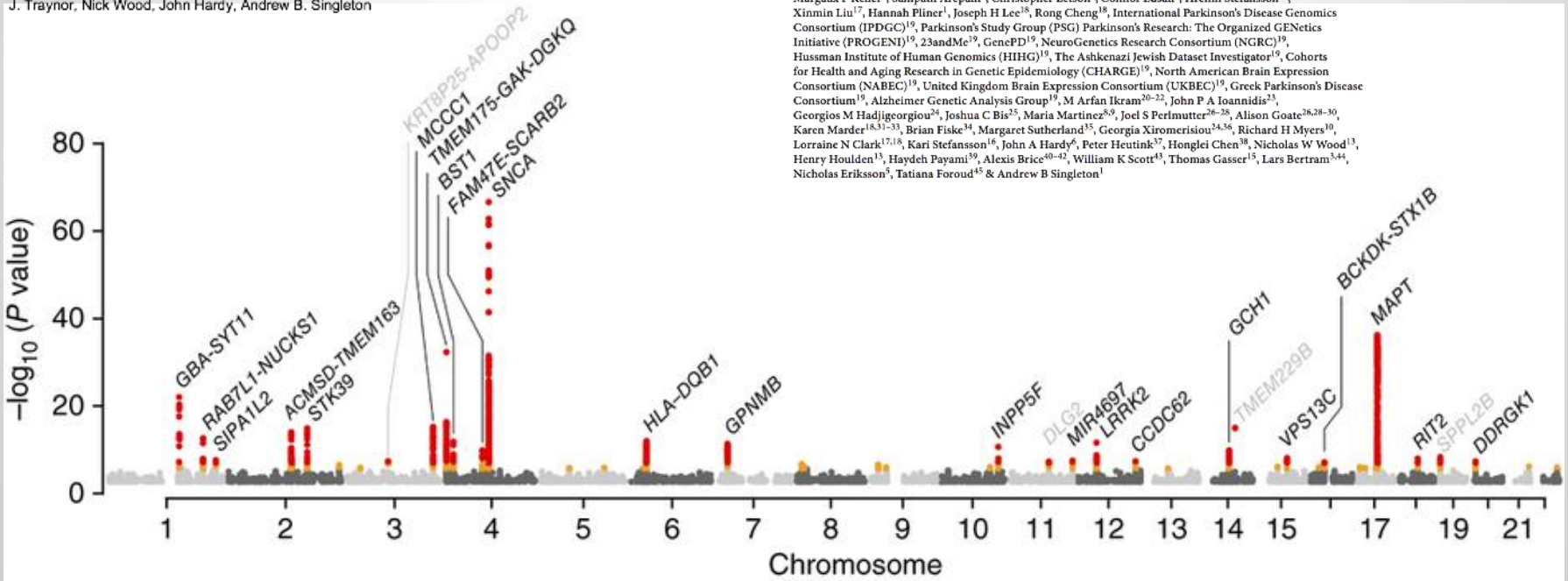
NeuroX, a Fast and Efficient Genotyping Platform for Investigation of Neurodegenerative Diseases

Mike A. Nalls, Jose Bras, Dena G. Hernandez, Margaux F. Keller, Elisa Majounie, Alan E. Renton, Mohamad Saad, Iris Jansen, Rita Guerreiro, Steven Lubbe, Vincent Plagnol, J. Raphael Gibbs, Claudia Schulte, Nathan Pankratz, Margaret Sutherland, Lars Bertram, Christina Lill, Anita L. DeStefano, Tatiana Faroud, Nicholas Eriksson, Joyce Y. Tung, Connor Edsall, Noah Nichols, Janet Brooks, Sampath Arepalli, Hannah Pliner, Chris Letson, Peter Heutink, Maria Martinez, Thomas Gasser, Bryan J. Traynor, Nick Wood, John Hardy, Andrew B. Singleton

LETTERS

Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease

Mike A Nalls[1,46], Nathan Pankratz[2,46], Christina M Lill[3,4], Chuong B Do[5], Dena G Hernandez[1,6], Mohamad Saad[7–9], Anita L DeStefano[10–12], Eleanna Kara[13], Jose Bras[13], Manu Sharma[14,15], Claudia Schulte[15], Margaux F Keller[1], Sampath Arepalli[1], Christopher Letson[1], Connor Edsall[1], Hreinn Stefansson[16], Xinmin Liu[17], Hannah Pliner[1], Joseph H Lee[18], Rong Cheng[18], International Parkinson's Disease Genomics Consortium (IPDGC)[19], Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI)[19], 23andMe[19], GenePD[19], NeuroGenetics Research Consortium (NGRC)[19], Hussman Institute of Human Genomics (HIHG)[19], The Ashkenazi Jewish Dataset Investigator[19], Cohorts for Health and Aging Research in Genetic Epidemiology (CHARGE)[19], North American Brain Expression Consortium (NABEC)[19], United Kingdom Brain Expression Consortium (UKBEC)[19], Greek Parkinson's Disease Consortium[19], Alzheimer Genetic Analysis Group[19], M Arfan Ikram[20–22], John P A Ioannidis[23], Georgios M Hadjigeorgiou[24], Joshua C Bis[25], Maria Martinez[8,9], Joel S Perlmutter[26–28], Alison Goate[26,28–30], Karen Marder[18,31–33], Brian Fiske[34], Margaret Sutherland[35], Georgia Xiromerisiou[24,36], Richard H Myers[10], Lorraine N Clark[17,18], Kari Stefansson[16], John A Hardy[6], Peter Heutink[37], Honglei Chen[38], Nicholas W Wood[13], Henry Houlden[13], Haydeh Payami[39], Alexis Brice[40–42], William K Scott[43], Thomas Gasser[15], Lars Bertram[3,44], Nicholas Eriksson[5], Tatiana Foroud[45] & Andrew B Singleton[1]

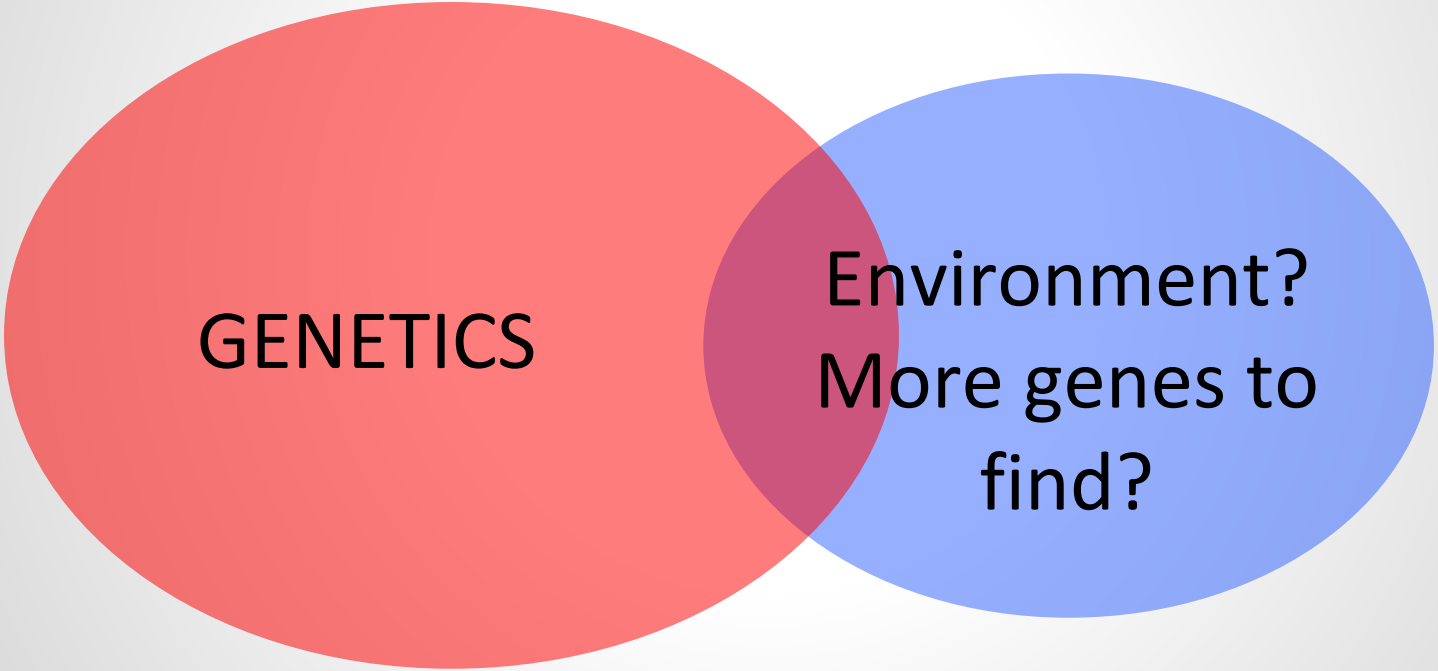


In total, we identified and replicated 6 new loci and confirmed an additional 22 suspected loci...

SNP Information							Discovery phase (13,728 cases and 95,282 controls)					Replication phase (5,353 cases and 5,551 controls)					Joint phase (19,081 cases and 100,833 controls)				
SNP	C	Position (bp)	Nearest gene(s)	Effect allele	Alternate allele	Effect allele frequency	I2	Beta	Odds ratio	Standard error	P	I2	Beta	Odds ratio	Standard error	P	I2	Beta	Odds ratio	Standard error	P
Genome Wide Significant, Discovery Phase																					
rs35749011*	1	155,135,036	GBA/SYT11	a	g	0.017	0	0.566	1.762	0.057	6.09x10-23	0	0.836	2.307	0.148	7.48x10-09	65.5	0.601	1.824	0.053	1.37x-29
rs823118	1	205,723,572	RAB7L1/NUCKS1	t	c	0.559	55.7	0.119	1.126	0.016	1.36x10-13	0	0.104	1.109	0.029	1.43x10-04	0	0.116	1.122	0.014	1.66x-16
rs10797576	1	232,664,611	SIPA1L2	t	c	0.14	0	0.13	1.139	0.023	1.19x10-08	26	0.104	1.11	0.039	3.38x10-03	0	0.123	1.131	0.020	4.87x-10
rs6430538	2	135,539,967	ACMSD/TMEM163	t	c	0.43	0	-0.136	0.873	0.017	5.56x10-15	47.9	-0.126	0.882	0.029	9.42x10-06	0	-0.133	0.875	0.015	9.13x-20
rs1474055*	2	169,110,394	STK39	t	c	0.128	9.3	0.193	1.213	0.024	7.12x10-16	54.4	0.198	1.218	0.042	1.07x10-06	0	0.194	1.214	0.021	1.15x-20
rs115185635*	3	87,520,857	KRT8P25/APOOP2	c	g	0.035	91	0.582	1.789	0.104	2.18x10-08	30.4	-0.071	0.931	0.07	0.846	96.3	0.133	1.142	0.058	0.022
rs12637471	3	182,762,437	MCCC1	a	g	0.193	26.6	-0.17	0.844	0.021	3.32x10-16	59	-0.179	0.836	0.036	3.72x10-07	0	-0.172	0.842	0.018	2.14x-21
rs34311866	4	951,947	TMEM175/GAK/DGKQ	t	c	0.809	52.4	-0.243	0.784	0.02	3.58x10-33	55.4	-0.234	0.791	0.035	6.29x10-12	0	-0.241	0.786	0.017	1.02x-43
rs11724635	4	15,737,101	BST1	a	c	0.553	14.8	0.116	1.122	0.016	8.07x10-13	20.6	0.129	1.138	0.028	2.73x10-06	0	0.119	1.126	0.014	9.44x-18
rs6812193	4	77,198,986	FAM47E/SCARB2	t	c	0.364	29.5	-0.108	0.897	0.017	7.17x10-11	10.7	-0.067	0.935	0.029	0.011	32.8	-0.098	0.907	0.015	2.95x-11
rs356182	4	90,626,111	SNCA	a	g	0.633	48.5	-0.306	0.737	0.018	3.23x10-67	34.4	-0.196	0.822	0.028	1.75x10-12	90.8	-0.274	0.760	0.015	4.16x-73
rs9275326*	6	32,666,660	HLA-DQB1	t	c	0.094	2.1	-0.227	0.797	0.032	5.82x10-13	0	-0.105	0.9	0.05	0.018	76.3	-0.192	0.826	0.027	1.19x-12
rs199347	7	23,293,746	GPNMB	a	g	0.59	10	0.116	1.123	0.017	2.37x10-12	26.6	0.07	1.072	0.029	7.66x10-03	46.6	0.104	1.110	0.015	1.18x-12
rs117896735*	10	121,536,327	INPP5F	a	g	0.014	15.2	0.569	1.767	0.084	1.21x10-11	0	0.339	1.404	0.111	1.10x10-03	63.4	0.485	1.624	0.067	4.34x-13
rs3793947*	11	83,544,472	DLG2	a	g	0.443	0	-0.092	0.912	0.017	2.59x10-08	0	-0.024	0.976	0.028	0.201	76.8	-0.074	0.929	0.015	3.96x-07
rs329648	11	133,765,367	MIR4697	t	c	0.354	0	0.095	1.1	0.017	1.65x10-08	48.5	0.114	1.121	0.029	4.38x10-05	0	0.100	1.105	0.015	9.83x-12
rs76904798	12	40,614,434	LRRK2	t	c	0.143	0	0.157	1.17	0.022	1.33x10-12	26.4	0.104	1.11	0.039	3.69x10-03	28.6	0.144	1.155	0.019	5.24x-14
rs11060180	12	123,303,586	CCDC62	a	g	0.558	42.8	0.097	1.101	0.017	2.14x10-08	0	0.108	1.114	0.028	7.26x10-05	0	0.100	1.105	0.015	6.02x-12
rs11158026	14	55,348,869	GCH1	t	c	0.335	32.1	-0.118	0.889	0.018	7.13x10-11	0	-0.054	0.948	0.03	0.039	70.1	-0.101	0.904	0.015	5.85x-11
rs1555399*	14	67,984,370	TMEM229B	a	t	0.468	97.2	-0.138	0.872	0.017	5.53x10-16	0	-0.03	0.971	0.028	0.144	90.8	-0.109	0.897	0.015	6.63x-14
rs2414739	15	61,994,134	VPS13C	a	g	0.734	29	0.108	1.114	0.018	4.13x10-09	1.1	0.104	1.109	0.033	7.96x10-04	0	0.107	1.113	0.016	1.23x-11
rs14235	16	31,121,793	BCKDK/ STX1B	a	g	0.381	25.5	0.09	1.094	0.016	3.89x10-08	27.5	0.125	1.133	0.029	7.72x10-06	10.4	0.098	1.103	0.014	2.43x-12
rs17649553	17	43,994,648	MAPT	t	c	0.226	0	-0.261	0.771	0.021	4.86x10-37	0	-0.269	0.764	0.035	7.03x10-15	0	-0.263	0.769	0.018	2.37x-48
rs12456492	18	40,673,380	RIT2	a	g	0.693	11.2	-0.1	0.905	0.017	5.12x10-09	49.1	-0.105	0.9	0.03	2.16x10-04	0	-0.101	0.904	0.015	7.74x-12
rs62120679*	19	2,363,319	SPPL2B	t	c	0.314	47.1	0.132	1.141	0.022	2.53x10-09	40.3	-0.002	0.999	0.034	0.518	90.9	0.093	1.097	0.019	5.57x-07
rs8118008*	20	3,168,166	DDRGK1	a	g	0.657	32.8	0.105	1.111	0.019	2.32x10-08	0	0.107	1.113	0.029	1.18x10-04	0	0.106	1.111	0.016	3.04x-11
Previously Reported as Significant in Genome Wide Studies																					
rs34016896	3	160,992,864	NMD3	t	c	0.319	37.8	0.077	1.08	0.017	7.68x10-06	0	0.028	1.028	0.03	0.174	50.5	0.065	1.067	0.015	1.08x-05
rs591323	8	16,697,091	FGF20	a	g	0.275	0	-0.083	0.921	0.019	1.30x10-05	1.7	-0.103	0.902	0.032	6.16x10-04	0	-0.088	0.916	0.016	6.68x-08
rs60298754	8	89,373,041	MMP16	t	c	0.024	53.6	0.075	1.078	0.056	0.181	-	-	-	-	-	0	0.075	1.078	0.056	0.181
rs7077361	10	15,561,543	ITGA8	t	c	0.874	0	0.104	1.11	0.025	3.24x10-05	34.4	0.043	1.044	0.042	0.154	35.8	0.088	1.092	0.022	4.16x-05
rs11868035	17	17,715,101	SREBF/RAI1	a	g	0.298	56.2	-0.065	0.937	0.018	2.17x10-04	0	-0.055	0.947	0.031	0.036	0	-0.063	0.939	0.016	5.98x-05
rs2823357	21	16,914,905	USP25	a	g	0.37	57	0.035	1.036	0.016	0.032	55.9	0.018	1.018	0.029	0.267	0	0.031	1.031	0.014	0.027

META4 is currently underway and we are busy looking for additional replication samples, but in summary we can conclude...

Etiology of PD, 2013-2014



Exercise 3.

Lets 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.
- Its also a bit faster than R.
- “Industry standard”

Exercise 3.

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

Lets 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-vlaue < 5E-8).

Lets 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, lets 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.

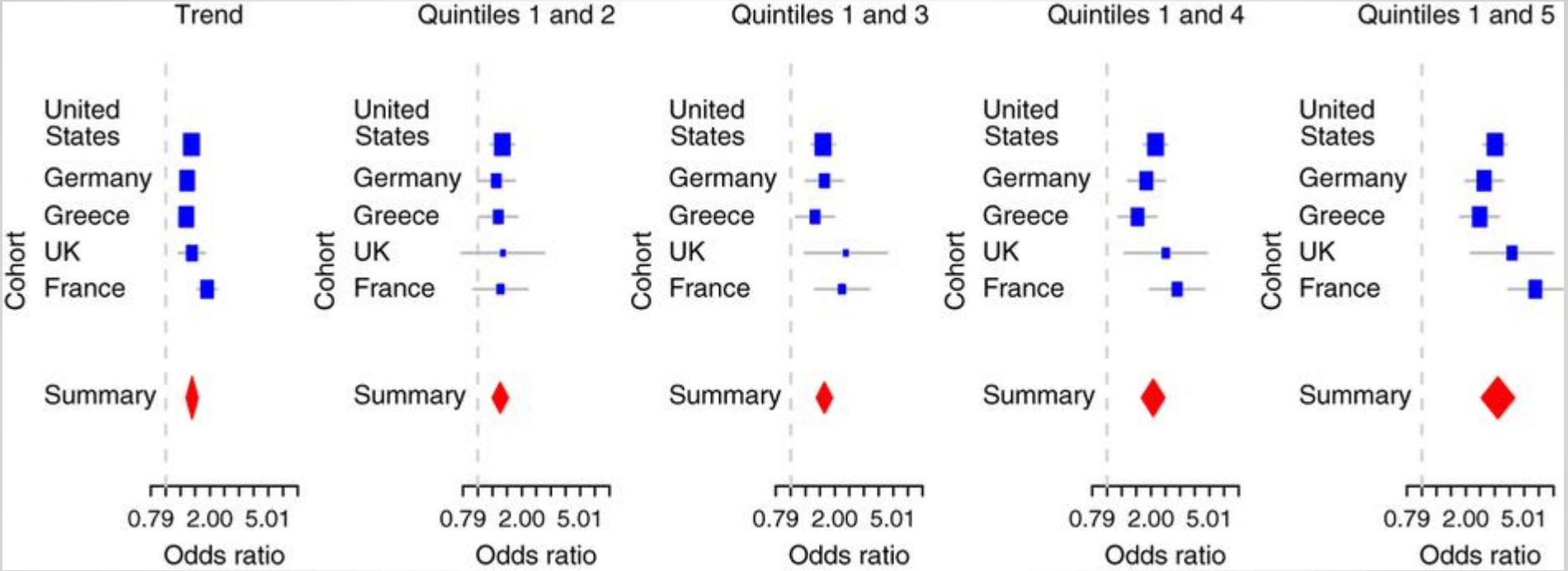
We employed risk profiling to quantify cumulative risk attributable to all of these variants of interest.

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



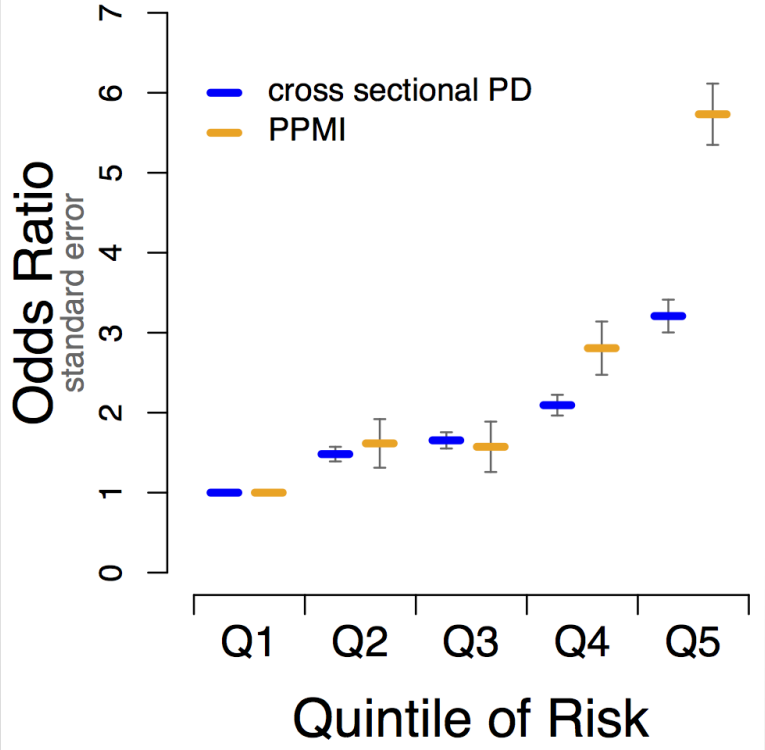
Trend → a genetic risk profile score greater than 1 s.d. from the population mean, indicative of a roughly 34% increase in genetic risk score above the mean for controls, had a significantly higher risk of Parkinson's disease (OR = 1.51, $P = 2 \times 10^{-16}$).

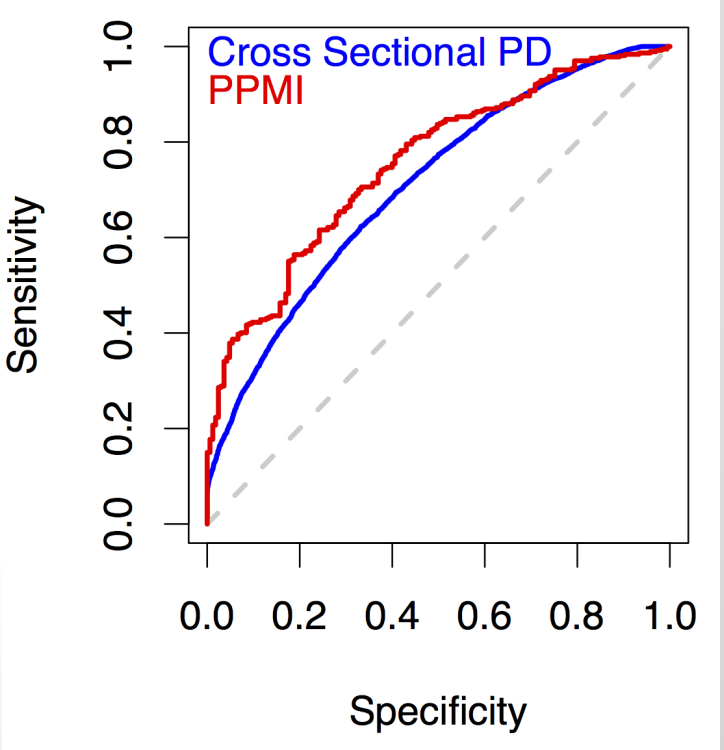
Outliers → fifth quintile of genetic risk scores to the first quintile of genetic risk as a reference; the OR was 3.31 ($P = 2 \times 10^{-16}$).

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%).



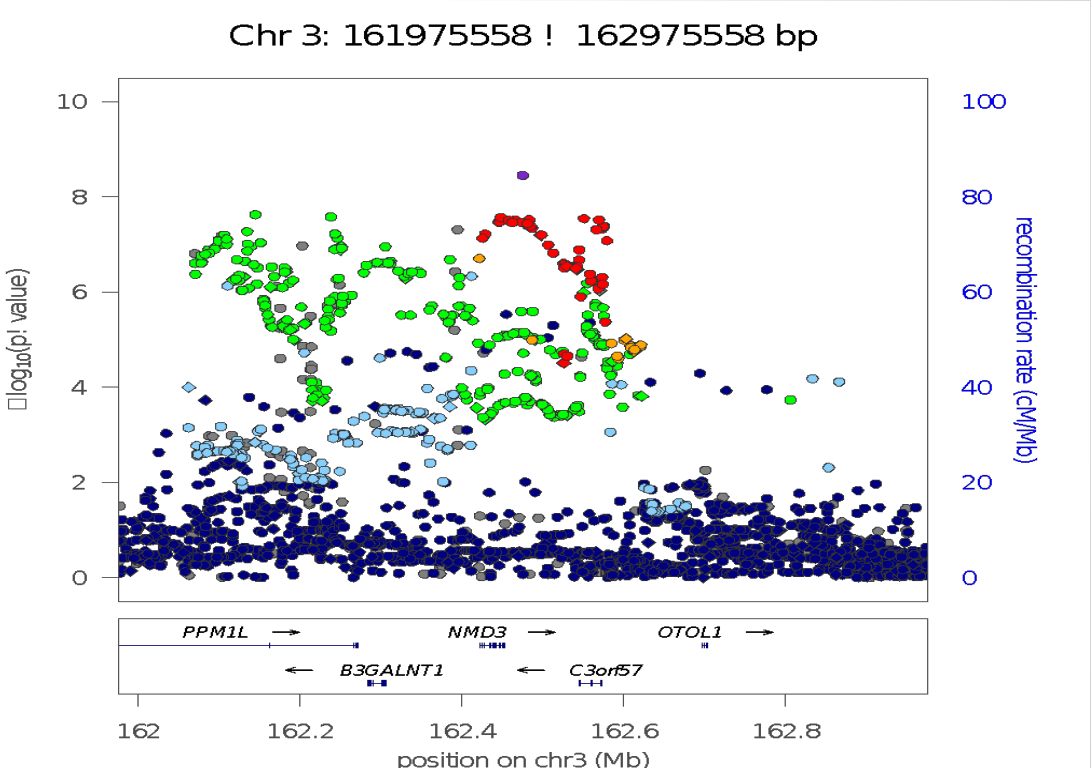


Functional inferences from methylation and expression data

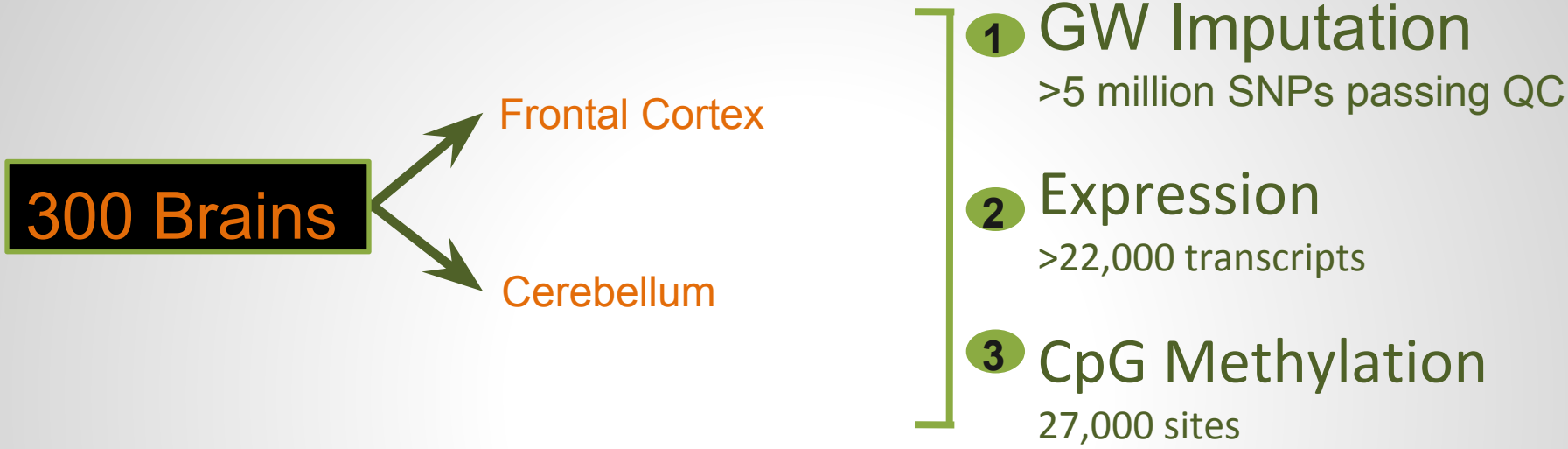
So at this point we have discovered a number of risk SNPs associated with PD
In general we discuss the most significant SNP in a region, we need to remember these are actually loci, anywhere from a handful to thousands of correlated proximal SNPs all associated with disease to some degree.

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



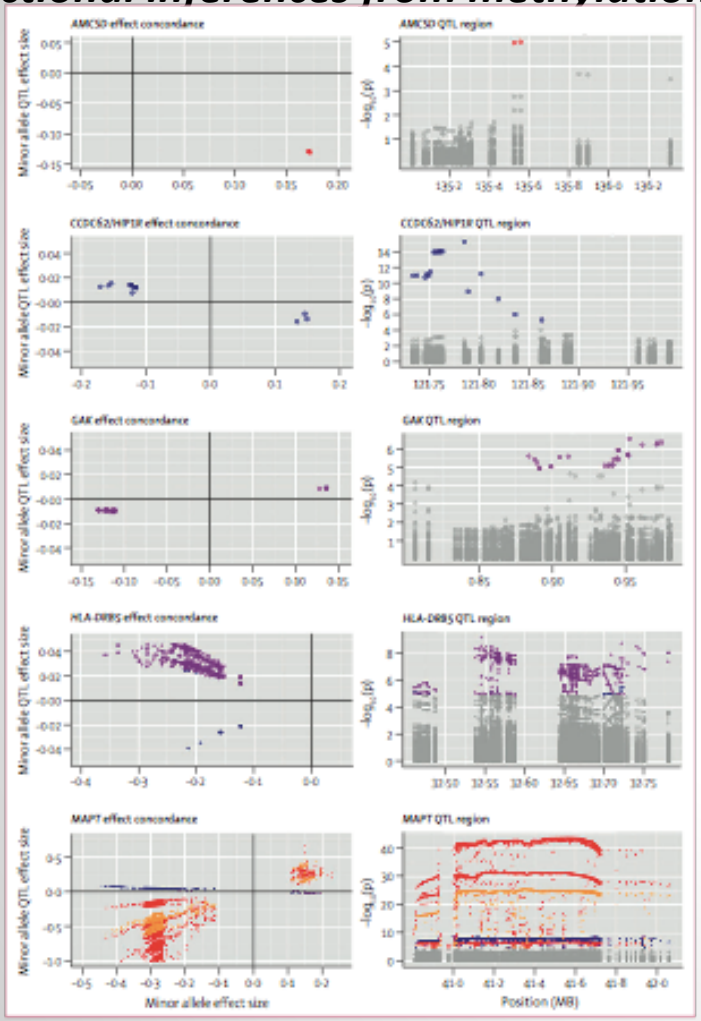
Allows for inference of biological processes at risk loci.

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.

Orange circles=expression assayed in the frontal.
Red circles=expression assayed in the cerebellum.
Purple circles=methylation assayed in the frontal.
Blue circles=methylation in the cerebellum.
Grey circles=not significant.

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.

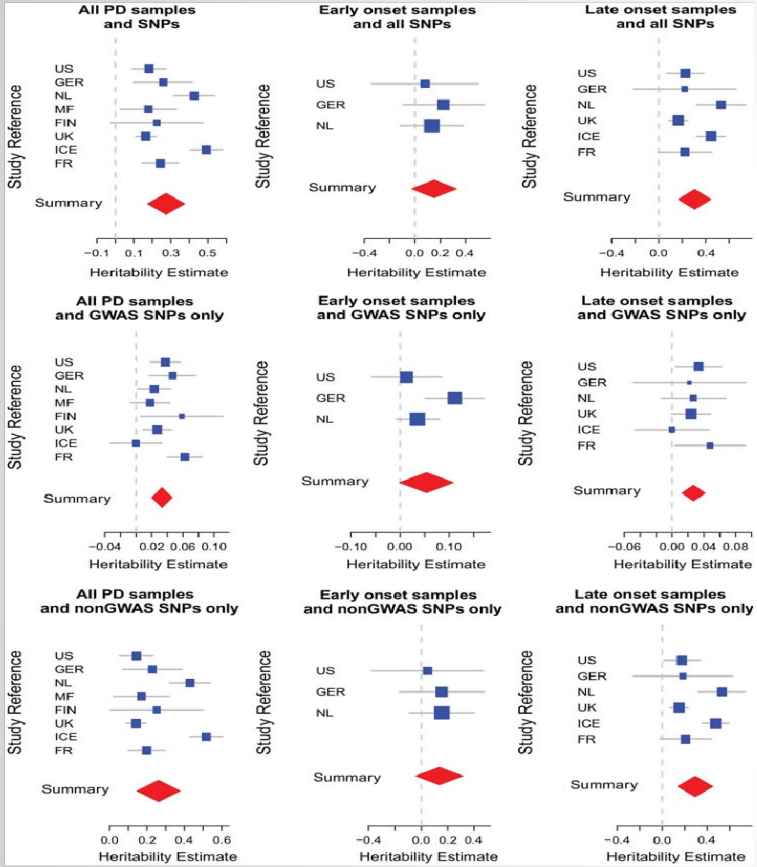
Recent methods have been developed to estimate heritability in ostensibly outbred populations based on low levels of background "relatedness" within the population.

- -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



Human Molecular Genetics, 2012, Vol. 21, No. 22 **4996–5009**
doi:10.1093/hmg/dds335
Advance Access published on August 13, 2012

Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease

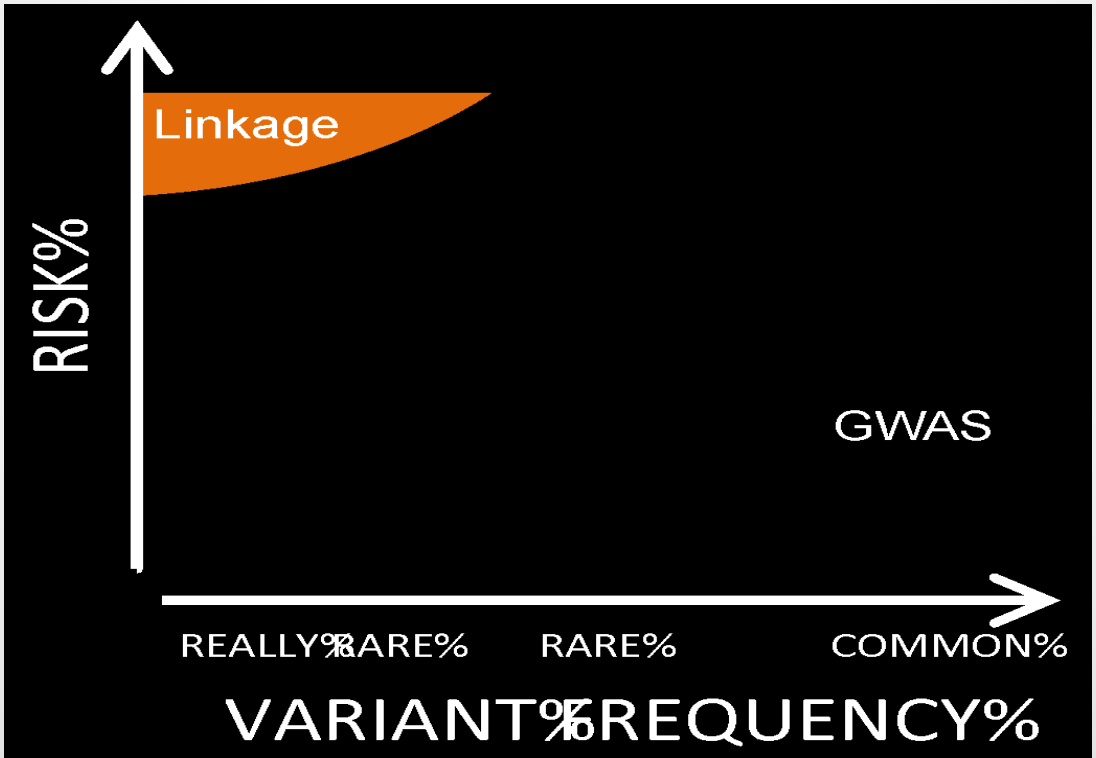
Margaux F. Keller[1,2], Mohamad Saad[3,4], Jose Bras[5], Francesco Bettella[7], Nayia Nicolaou[8], Javier Simón-Sánchez[8], Florian Mittag[3], Finja Büchel[3], Manu Sharma[9,10], J. Raphael Gibbs[1,5], Claudia Schulte[9,10], Valentina Moskvina[11,12], Alexandra Durr[13,14,15,16], Peter Holmans[11,12], Laura L. Kilarski[11,12], Rita Guerreiro[5], Dena G. Hernandez[1,5], Alexis Brice[13,14,15,16], Pauli Ylikotila[17], Hreinn Stefánsson[7], Kari Majamaa[18], Huw R. Morris[11,12], Nigel Williams[11,12], Thomas Gasser[9,10], Peter Heutink[7], Nicholas W. Wood[5,6], John Hardy[5], Maria Martinez[3,4], Andrew B. Singleton[1] and Michael A. Nalls[1,*] for the International Parkinson's Disease Genomics Consortium (IPDGC) and The Wellcome Trust Case Control Consortium 2 (WTCCC2)[†]

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.

PD type	SNPs included in analysis	Heritability estimate from random effects	Lower 95% confidence interval	Upper 95% confidence interval	*P*-value from random effects	Heterogeneity of variance from random effects (%)	Heterogeneity *P*-value
All	All SNPs	0.27	0.17	0.38	8.80E − 08	0.02	0.00E + 00
	GWAS SNPs in PD loci/regions	0.03	0.02	0.05	5.23E − 07	0.00	3.20E − 02

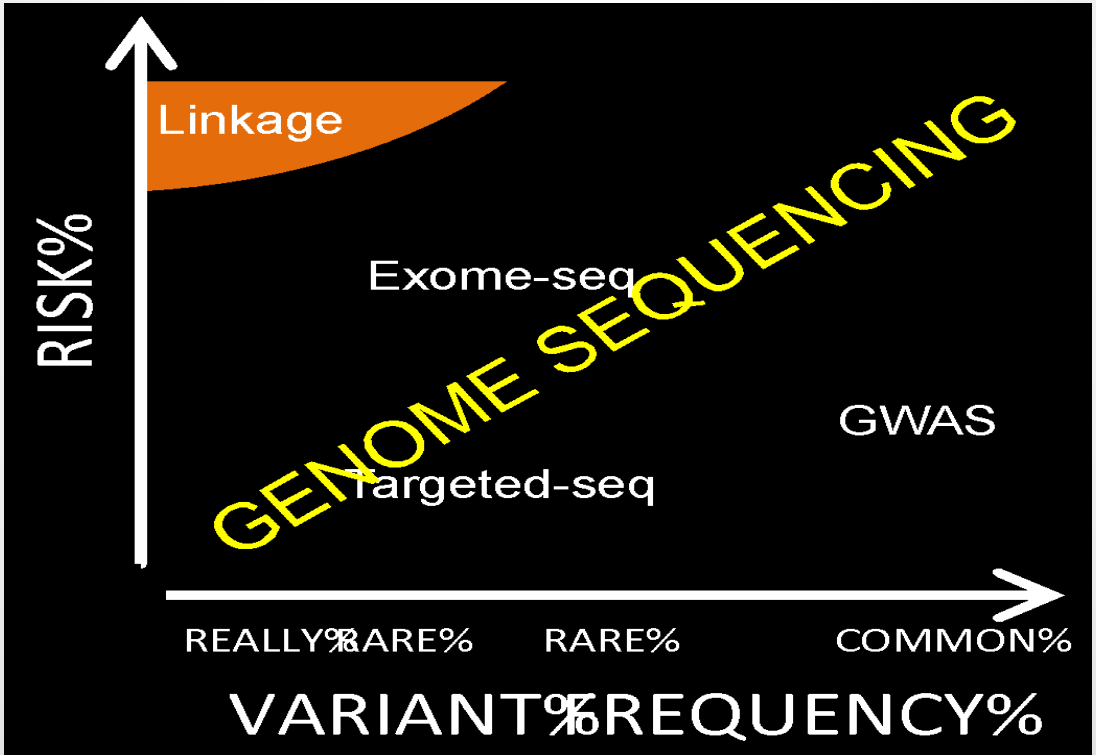
Where we stand now…

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 gnome 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.

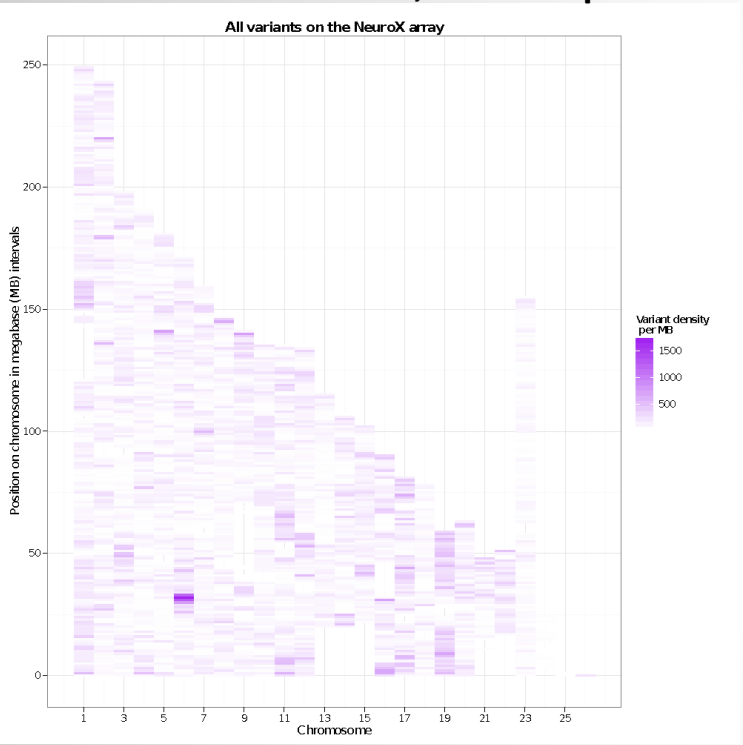
Full exome sequencing based variants standard to the Illumina exome array (242901 variants) and neurological and neurodegenerative disease focused content that may be added to other existing arrays (24706 variants).

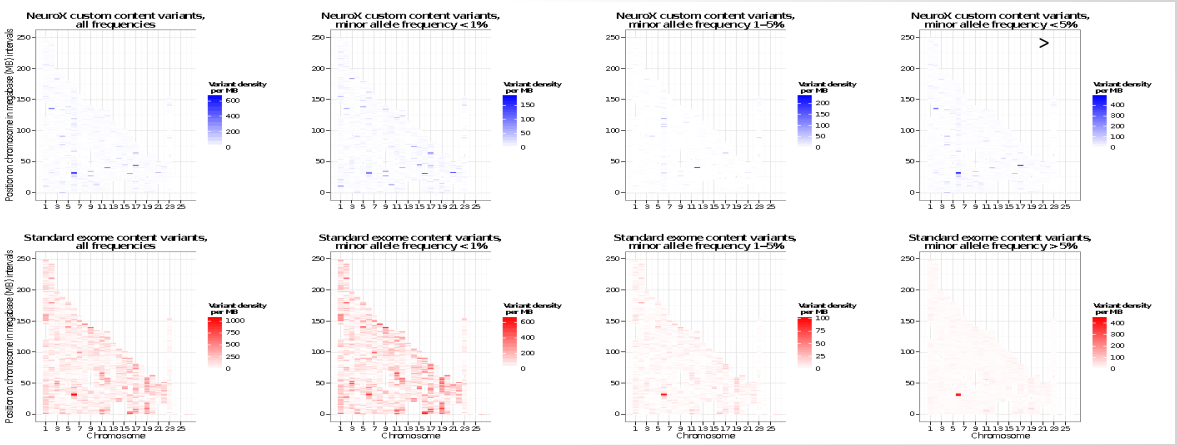
This array covers a majority of easily assay-able coding variation in the genome at a fraction of the price of sequence-based data with major attention to rare variants.

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…

The new england journal of medicine

original article

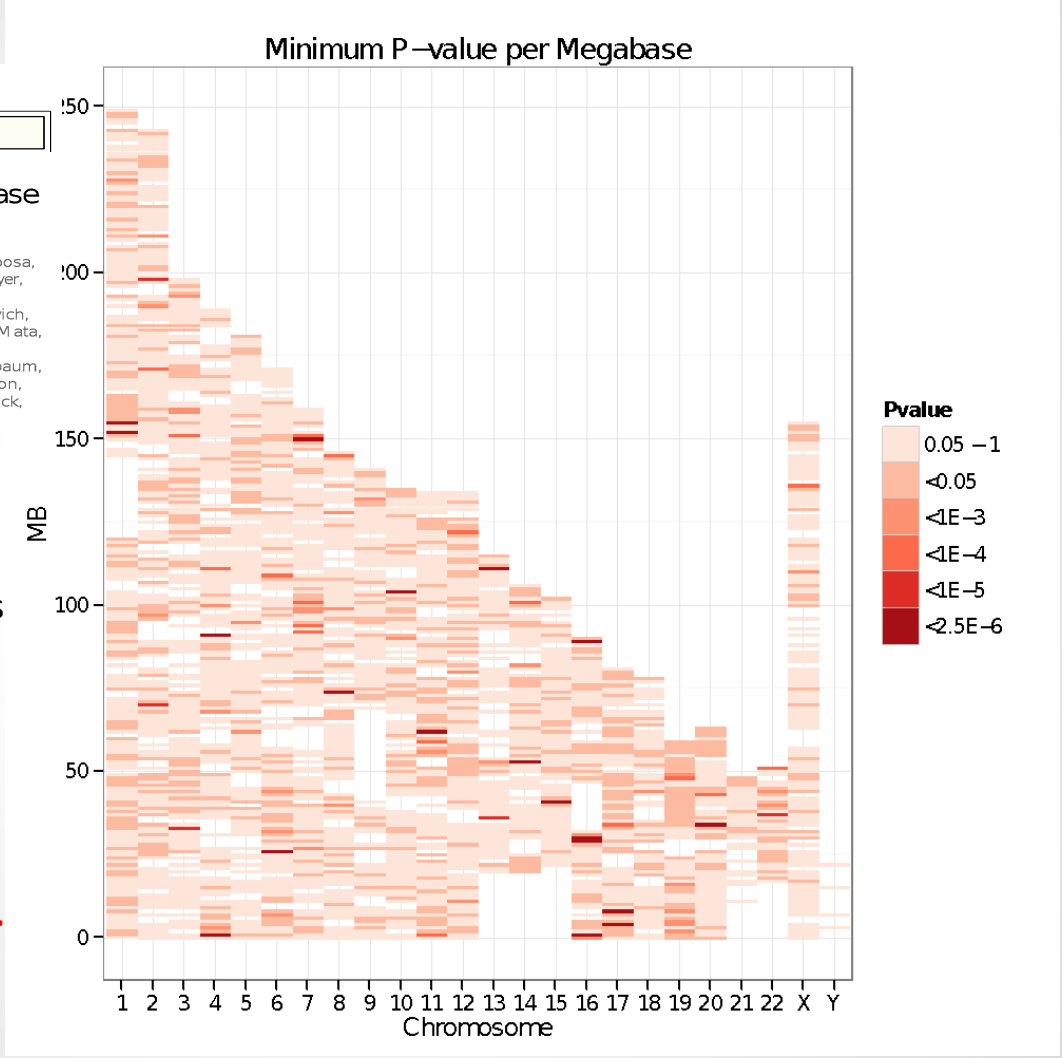
Multicenter Analysis of Glucocerebrosidase Mutations in Parkinson's Disease

E. Sidransky, M.A. Nalls, J.O. Aasly, J. Aharon-Peretz, G. Annesi, E.R. Barbosa, A. Bar-Shira, D. Berg, J. Bras, A. Brice, C.-M. Chen, L.N. Clark, C. Condroyer, E.V. De Marco, A. Dürr, M.J. Eblan, S. Fahn, M.J. Farrer, H.-C. Fung, Z. Gan-Or, T. Gasser, R. Gershoni-Baruch, N. Giladi, A. Griffith, T. Gurevich, C. Januario, P. Kropp, A.E. Lang, G.-J. Lee-Chen, S. Lesage, K. Marder, I.F. Mata, A. Mirelman, J. Mitsui, I. Mizuta, G. Nicoletti, C. Oliveira, R. Ottman, A. Orr-Urtreger, L.V. Pereira, A. Quattrone, E. Rogaeva, A. Rolfs, H. Rosenbaum, R. Rozenberg, A. Samii, T. Samaddar, C. Schulte, M. Sharma, A. Singleton, M. Spitz, E.-K. Tan, N. Tayebi, T. Toda, A.R. Troiano, S. Tsuji, M. Wittstock, T.G. Wolfsberg, Y.-R. Wu, C.P. Zabetian, Y. Zhao, and S.G. Ziegler

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.



Big thanks for big data (consortia members):

Margaux F. Keller (Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA), Michael A. Nalls (Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA), Vincent Plagnol (UCL Genetics Institute, London, UK), Dena G. Hernandez (Laboratory of Neurogenetics, National Institute on Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Manu Sharma (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tu¨bingen, and DZNE, German Center for Neurodegenerative Diseases, Tu¨bingen, Germany), Una-Marie Sheerin (Department of Molecular Neuroscience, UCL Institute of Neurology), Mohamad Saad (INSERM U563, CPTP, Toulouse, France; and Paul Sabatier University, Toulouse, France), Javier Simo´n-Sa´nchez (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre, Amsterdam, The Netherlands), 5006 Human Molecular Genetics, 2012, Vol. 21, No. 22 Downloaded from http://hmg.oxfordjournals.org/ at NIH Library on April 3, 2013 Claudia Schulte (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Suzanne Lesage [INSERM, UMR-S975 (formerly UMR-S679), Paris, France; Universite´ Pierre et Marie Curie-Paris, Centre de Recherche de l'Institut du Cerveau et de la Moelle e´pinie`re, Paris, France; and CNRS, Paris, France], Sigurlaug Sveinbjo¨rnsdo´ttir (Department of Neurology, Landspı´tali University Hospital, Reykjavı´k, Iceland; Department of Neurology, MEHT Broomfield Hospital, Chelmsford, Essex, UK; and Queen Mary College, University of London, London, UK), Sampath Arepalli (Laboratory of Neurogenetics, National Institute on Aging), Roger Barker (Department of Neurology, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK), Yoav Ben-Shlomo (School of Social and Community Medicine, University of Bristol), Henk W. Berendse (Department of Neurology and Alzheimer Center, VU University Medical Center), Daniela Berg (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Kailash Bhatia (Department of Motor Neuroscience, UCL Institute of Neurology), Rob M.A. de Bie (Department of Neurology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands), Alessandro Biffi (Center for Human Genetic Research and Department of Neurology, Massachusetts General Hospital, Boston, MA, USA; and Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA), Bas Bloem (Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands), Zoltan Bochdanovits (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Michael Bonin (Department of Medical Genetics, Institute of Human Genetics, University of Tu¨bingen, Tu¨bingen, Germany), Jose Bras (Department of Molecular Neuroscience, UCL Institute of Neurology), Kathrin Brockmann (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Janet Brooks (Laboratory of Neurogenetics, National Institute on Aging), David J. Burn (Newcastle University Clinical Ageing Research Unit, Campus for Ageing and Vitality, Newcastle upon Tyne, UK), Gavin Charlesworth (Department of Molecular Neuroscience, UCL Institute of Neurology), Honglei Chen (Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, NC, USA), Patrick F. Chinnery (Neurology M4104, The Medical School, Framlington Place, Newcastle upon Tyne, UK), Sean Chong (Laboratory of Neurogenetics, National Institute on Aging), Carl E. Clarke (School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, UK; and Department of Neurology, City Hospital, Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK), Mark R. Cookson (Laboratory of Neurogenetics, National Institute on Aging), J. Mark Cooper (Department of Clinical Neurosciences, UCL Institute of Neurology), Jean Christophe Corvol (INSERM, UMR_S975; Universite´ Pierre et Marie Curie-Paris; CNRS; and INSERM CIC-9503, Ho^pital Pitie´-Salpe^trie`re, Paris, France), Carl Counsell (University of Aberdeen, Division of Applied Health Sciences, Population Health Section, Aberdeen, UK), Philippe Damier (CHU Nantes, CIC0004, Service de Neurologie, Nantes, France), Jean-Franc¸ois Dartigues (INSERM U897, Universite´ Victor Segalen, Bordeaux, France), Panos Deloukas (Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK), Gu¨nther Deuschl (Klinik fu¨r Neurologie, Universita¨tsklinikum Schleswig-Holstein, Campus Kiel, Christian-Albrechts-Universita¨t Kiel, Kiel, Germany), David T. Dexter (Parkinson's Disease Research Group, Faculty of Medicine, Imperial College London, London, UK), Karin D. van Dijk (Department of Neurology and Alzheimer Center, VU University Medical Center), Allissa Dillman (Laboratory of Neurogenetics, National Institute on Aging), Frank Durif (Service de Neurologie, Ho^pital Gabriel Montpied, Clermont- Ferrand, France), Alexandra Du¨rr (INSERM, UMR-S975; Universite´ Pierre et Marie Curie-Paris; CNRS; and AP-HP, Pitie´-Salpe^trie`re Hospital), Sarah Edkins (Wellcome Trust Sanger Institute), Jonathan R. Evans (Cambridge Centre for Brain Repair, Cambridge, UK), Thomas Foltynie (UCL Institute of Neurology), Jianjun Gao (Epidemiology Branch, National Institute of Environmental Health Sciences), Michelle Gardner (Department of Molecular Neuroscience, UCL Institute of Neurology), J. Raphael Gibbs (Laboratory of Neurogenetics, National Institute on Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology), Alison Goate (Department of Psychiatry, Department of Neurology, Washington University School of Medicine, MI, USA), Emma Gray (Wellcome Trust Sanger Institute), Rita Guerreiro (Department of Molecular Neuroscience, UCL Institute of Neurology), O´mar Gu´stafsson (deCODE genetics and Department of Psychiatry, Oslo University Hospital, N-0407 Oslo, Norway), Clare Harris (University of Aberdeen), Jacobus J. van Hilten (Department of Neurology, Leiden University Medical Center, Leiden, The Netherlands), Albert Hofman (Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands), Albert Hollenbeck (AARP, Washington, DC, USA), Janice Holton (Queen Square Brain Bank for Neurological Disorders, UCL Institute of Neurology), Michele Hu (Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Xuemei Huang (Departments of Neurology, Radiology, Neurosurgery, Pharmacology, Kinesiology, and Bioengineering, Pennsylvania State University—Milton S. Hershey Medical Center, Hershey, PA, USA), Heiko Huber (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Gavin Hudson (Neurology M4104, The Medical School, Newcastle upon Tyne, UK), Sarah E. Hunt (Wellcome Trust Sanger Institute), Johanna Huttenlocher (deCODE genetics), Thomas Illig (Institute of Epidemiology, Helmholtz Zentrum Mu¨nchen, German Research Centre for Environmental Health, Neuherberg, Germany), Pa´lmi V. Jo´nsson (Department of Geriatrics, Landspı´tali University Hospital, Reykjavı´k, Iceland), Jean-Charles Lambert (INSERM U744, Lille, France; and Institut Pasteur de Lille, Universite´ de Lille Nord, Lille, France), Cordelia Langford (Cambridge Centre for Brain Repair), Andrew Lees (Queen Square Brain Bank for Neurological Disorders), Peter Lichtner (Institute of Human Genetics, Helmholtz Zentrum Mu¨nchen, German Research Centre for Environmental Health, Neuherberg, Germany), Patricia Limousin (Institute of Neurology, Sobell Department, Unit of Functional Neurosurgery, London, UK), Grisel Lopez (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI, National Institutes of Health), Delia Lorenz (Klinik fu¨r Neurologie, Universita¨tsklinikum Schleswig-Holstein), Alisdair McNeill (Department of Clinical Neurosciences, UCL Institute of Neurology), Catriona Moorby (School of Clinical and Experimental Medicine, University of Birmingham), Matthew Moore (Laboratory of Neurogenetics, National Institute on Aging), Huw R. Morris (MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff, UK), Karen E. Morrison (School of Clinical and Experimental Medicine, University of Birmingham; and Neurosciences Department, Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK), Ese Mudanohwo (Neurogenetics Unit, UCL Institute of Neurology and National Hospital for Neurology and Neurosurgery), Sean S. O'Sullivan (Queen Square Brain Human Molecular Genetics, 2012, Vol. 21, No. 22 5007 Downloaded from http://hmg.oxfordjournals.org/ at NIH Library on April 3, 2013 Bank for Neurological Disorders), Justin Pearson (MRC Centre for Neuropsychiatric Genetics and Genomics), Joel S. Perlmutter (Department of Neurology, Radiology, and Neurobiology at Washington University, St Louis, MO, USA), Hjo¨rvar Pe´tursson (deCODE genetics; and Department of Medical Genetics, Institute of Human Genetics, University of Tu¨bingen), Pierre Pollak (Service de Neurologie, CHU de Grenoble, Grenoble, France), Bart Post (Department of Neurology, Radboud University Nijmegen Medical Centre), Simon Potter (Wellcome Trust Sanger Institute), Bernard Ravina (Translational Neurology, Biogen Idec, MA, USA), Tamas Revesz (Queen Square Brain Bank for Neurological Disorders), Olaf Riess (Department of Medical Genetics, Institute of Human Genetics, University of Tu¨bingen), Fernando Rivadeneira (Departments of Epidemiology and Internal Medicine, Erasmus University Medical Center), Patrizia Rizzu (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Mina Ryten (Department of Molecular Neuroscience, UCL Institute of Neurology), Stephen Sawcer (University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's Hospital, Cambridge, UK), Anthony Schapira (Department of Clinical Neurosciences, UCL Institute of Neurology), Hans Scheffer (Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands), Karen Shaw (Queen Square Brain Bank for Neurological Disorders), Ira Shoulson (Department of Neurology, University of Rochester, Rochester, NY, USA), Ellen Sidransky (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI), Colin Smith (Department of Pathology, University of Edinburgh, Edinburgh, UK), Chris C.A. Spencer (Wellcome Trust Centre for Human Genetics, Oxford, UK), Hreinn Stefa´nsson (deCODE genetics), Stacy Steinberg (deCODE genetics), Joanna D. Stockton (School of Clinical and Experimental Medicine), Amy Strange (Wellcome Trust Centre for Human Genetics), Kevin Talbot (University of Oxford, Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Carlie M. Tanner (Clinical Research Department, The Parkinson's Institute and Clinical Center, Sunnyvale, CA, USA), Avazeh Tashakkori-Ghanbaria (Wellcome Trust Sanger Institute), Franc¸ois Tison (Service de Neurologie, Ho^pital Haut-Le´ve^que, Pessac, France), Daniah Trabzuni (Department of Molecular Neuroscience, UCL Institute of Neurology), Bryan J. Traynor (Laboratory of Neurogenetics, National Institute on Aging), Andre´ G. Uitterlinden (Departments of Epidemiology and Internal Medicine, Erasmus University Medical Center), Daan Velseboer (Department of Neurology, Academic Medical Center), Marie Vidailhet (INSERM, UMR-S975, Universite´ Pierre et Marie Curie-Paris, CNRS, UMR-7225), Robert Walker (Department of Pathology, University of Edinburgh), Bart van de Warrenburg (Department of Neurology, Radboud University Nijmegen Medical Centre), Mirdhu Wickremaratchi (Department of Neurology, Cardiff University, Cardiff, UK), Nigel Williams (MRC Centre for Neuropsychiatric Genetics and Genomics), Caroline H. Williams-Gray (Department of Neurology, Addenbrooke's Hospital), Sophie Winder-Rhodes (Department of Psychiatry and Medical Research Council and Wellcome Trust Behavioural and Clinical Neurosciences Institute, University of Cambridge), Ka´ri Stefa´nsson (deCODE genetics), Maria Martinez (INSERM U563; and Paul Sabatier University), John Hardy (Department of Molecular Neuroscience, UCL Institute of Neurology), Peter Heutink (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Alexis Brice (INSERM, UMR-S975, Universite´ Pierre et Marie Curie-Paris, CNRS, UMR-7225, AP-HP, Pitie´-Salpe^trie`re Hospital), Wellcome Trust Case Control Consortium 2 (webappendix, p. 13), Thomas Gasser (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, and DZNE, German Center for Neurodegenerative Diseases), Andrew B. Singleton (Laboratory of Neurogenetics, National Institute on Aging), Nicholas W. Wood (UCL Genetics Institute; and Department of Molecular Neuroscience, UCL Institute of Neurology).

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!