Genetic association studies, GWAS, polygenic risk scores and Mendelian randomisation

Dr Alastair Noyce
Reader in Neurology and Neuro-epidemiology
Preventive Neurology Unit, Wolfson Institute of Preventive Medicine
Consultant Neurologist, Barts Health NHS Trust
Objectives

• Discuss genetic association studies and genome wide association studies of PD risk and progression.

• Discuss the creation, uses and limitations of polygenic risk scores.

• Explain the rationale for Mendelian randomization, show some applications and discuss limitations.
Is it genetic, doctor???
Genetic association studies

- Collection of methods to identify genetic risk factors for complex diseases.
- Correlates the presence of genetic variation with case status.
- Genetic variation identified using single nucleotide polymorphisms (SNPs), or other markers.

A single nucleotide polymorphism
Genetic association studies

• Significant’ associations for a genotyped SNP can be:
  – Directly associated – SNP is the causal variant conferring susceptibility.
  – Indirectly associated – SNP is in linkage disequilibrium with a causal variant.
  – False positive – arising due to bias (such as population stratification).
Genome wide association studies (GWAS)

- Common method in complex trait genetics.
- Extends the principles of genetic association to a genome-wide approach.
- Evaluates genetic associations with apparently sporadic disease.
- Has the potential to:
  - identify disease mechanisms and pathways.
  - yield actionable targets.

https://www.ebi.ac.uk/training-beta/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/
GWAS

• Requires:
  – Large number of unrelated cases & controls.
  – Practically, this means multi-site or global collaboration.
  – Bioinformatics expertise
    • Next Generation Sequencing
    • Quality control
    • Imputation
- 37.7K cases, 18.6K ‘proxy-cases’
- 1.4M controls
- 90 independent GWAS hits
- 11-15% of PD risk heritability
  (out of total heritability 20-30%)
1 SD deviation in the PRS associated with lower AAO by 0.8 years

- ~28.8K cases
- 2 GWAS hits
  - SNCA and TMEM 175
  - Both PD risk hits
- 11% of PD AAO heritability
Genomewide Association Study of Parkinson’s Disease Clinical Biomarkers in 12 Longitudinal Patients’ Cohorts

Hirotsuka Iwaki, MD, PhD,1,2 Corneli Blauwendaal, PhD,3,4 Hampton L. Leonard, MS,1,2 Jonggeol J. Kim, BA,1 Ganyqiang Liu, PhD,1,2,4,5 Jodi Maple-Grader, PhD,5,6 Jean-Christophe Convol, MD, PhD,4 Lasse Pihlstrom, MD, PhD,9 Marlies van Nimwegen, PhD,10 Samantha J. Hutsen, PhD,11 Khanh-Dung H. Nguyen, PhD,12 Jacqueline Rick, PhD,13 Shirley Eberly, MS,14 Faraz Faghri, MS,11,15 Peggy Aurier, MS,7,16 Kristen M. Scott, MRCGP, MPH,17 Ruwani Wijeyesekon, MRCGP, PhD,17 Viviana M. Van Deelen, MD, PhD,18 Dena G. Hernandez, PhD,1 J. Raphael Gibbs, PhD,1 International Parkinson's Disease Genomics Consortium, Kumaraswamy Naidu Chitrale, PhD,19 Aaron G. Day-Williams, PhD,20,21 Alexis Bruce, MD,21,22,23 Guido Alves, MD, PhD,21,22,23,24 Alastair J. Noyce, MRCGP, PhD,25,26 Oie-Bjorn Tyros, MD, PhD,26,27 Jonathan R. Evans, MRCGP, PhD,20,28 David P. Breen, MRCGP, PhD,29,30 Karol Estrada, MD, PhD,29 Claire E. Wegel, MPH,14 Fabrice Danjou, MD, PhD,22 David K. Simon, MD, PhD,23,30 Oie Andreassen, MD, PhD,30,37,38 Bernard Ravina, MD, PhD,29,40 Mathias Tolf, MD, PhD,9,41 Peter Heutink, PhD,42,43 Bastiaan R. Bloem, MD, PhD,10 Daniel Weiner, PhD, MD,42,44 Roger A. Barker, MRCGP, PhD,16 Caroline H. Williams-Grey, MRCGP, PhD,46 Bart P. van de Warrenburg, MD, PhD,10 Jacobs J. Van Hulsen, MD, PhD,42 Clemens R. Scherzer, MD,4,5 Andrew B. Singleton, PhD,1 and Mike A. Nalls, PhD,1,2,4

- ~4000 cases (22k observations)
- Average follow up 3.8 years
- 25 phenotypes investigated
- 1 hit for HY3 rate, 1 for insomnia
- 9 risk variants associated
- 2 GBA SNPs assoc. with motor/cognitive
- 1 APOE SNPs assoc. with cognitive

Cohort level (12 cohorts/13 datasets*)

Phenotypes (Maximum) | Genetics data | Analysis
--- | --- | ---
- Constipation | SNP-level filtering | Binomial trait
- Cognitive impairment | - Ror < 0.3 (if imputed) | - OR at the baseline
- Depression | - Call rate < 0.95 | - HR in the follow-ups
- Daytime sleepiness | - MAF < 0.01 | - Continuous trait
- Dyskinesia | - HWE < 1E-4 | - Mean difference over time
- HY3 | | Covariates fixed
- Hyposmia | - high-missingness (< 0.95) | - Age at diagnosis (AAD)
- Insomnia | - Sex discordance | - Year from diagnosis (YTD)
- Motor fluctuation | - Extreme heterozygosity (F > 0.15) | - Sex
- REM Sleep Behavior Disorder | - Non-European ancestry | - PC1, PC2, PC3
- Restless legs syndrome | | Covariates maybe selected
- SEADL <= 70 | | - Quadratic AAD
- HY scale | | - Quadratic YTD
- MMSE | | - Years of education
- MoCA | | - HY score 2 or more at the baseline
- SEADL | | - Medication status
- UPDRS subscores/total |

Meta-analysis (for each phenotype)

<table>
<thead>
<tr>
<th>Input</th>
<th>Output filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed effect model</td>
<td>- MAF &gt; 0.05</td>
</tr>
<tr>
<td>Inverse variance weighting</td>
<td>- Max_MAF - Min_MAF &gt; 0.15</td>
</tr>
<tr>
<td>Genomic correction</td>
<td>- Total N of Participants &gt; 1000</td>
</tr>
<tr>
<td>No heterogeneity (Q-test &gt; 0.05)</td>
<td></td>
</tr>
</tbody>
</table>

Targeted assessment

Conduct analyses-wide tests for specific variants. Significant variants in above step / Previously identified PD risk variants/ coding variants of GBA/APOE associated variant

Summary statistics available online
The genetic architecture of Parkinson’s disease

Cornelis Blauwendraat, Mike A Nalls, Andrew B Singleton

Lancet Neurol 2020; 19: 170-78

Figure 1: Timeline of genetic discoveries from GWASs for Parkinson’s disease
GWAS=genome-wide association study.
After GWAS

Not specific follow-up of GWAS hits using fine mapping and deep resequencing etc.

Using GWAS summary data:

- Polygenic scores
- [Linkage Disequilibrium Score Regression (LDSC)]
- Mendelian randomization
Polygenic scores

- GWAS hits are independent & generally have small effect sizes.
- Effects can be combined to produce a weighted score according to the number of risk alleles in an individual.
- Polygenic scores may relate to a binary (e.g. risk) or continuous (AAO) outcome.

*Misconception – polygenic scores include only GWAS significant hits
Can in fact create polygenic scores according to a broad range of parameters*
Polygenic scores

• Calculations
  – At each risk locus
    • 0 – no risk alleles, 1 – single risk allele, 2 – two risk alleles
  – Weight scores by effect size per risk allele.
  – Z score transformation to normalize scores (mean 0, SD 1).
  – Binary outcome (e.g. risk) use logistic regression &
    continuous outcome (e.g. AAO) use linear regression.
Polygenic scores

• Uses
  – Investigating shared genetic architecture.
  – Investigate G*G and G*E interactions.
  – Personalized medicine – stratification & sub-phenotyping.
  – Mendelian randomization.

• Limitations
  – [Prediction and Diagnosis]
    • Risk distributions for cases & controls overlap significantly.
  – Focus on European ancestry (like GWAS in general)

Nalls et al, Lancet Neurol 2019;18(12):1091-1102
Undertaken using PPMI data
PRS alone
Integrated model AUC 0.92

Integrated model included:
• PRS
• Family history
• Age
• Gender
Parkinson’s disease determinants, prediction and gene-environment interactions in the UK Biobank

Benjamin M. Jacobs, Daniel Belete, Jonathan P Bestwick, Cornelis Blauwendraat, Sara Bandres-Ciga, Karl Heibron, Ruth Dobson, Mike A. Nalls, Andrew B. Singleton, John Hardy, Gavin Giovannoni, Andrew J. Lees, Annette Schrag, Alastair J Noyce, for The International Parkinson’s Disease Genomics Consortium (IPDGC)

doi: https://doi.org/10.1101/2020.02.15.950733

This article is a preprint and has not been certified by peer review [what does this mean?].

Risk factors

Risk algorithm – AUC 0.75
PRS + risk algorithm – AUC 0.76

Nagelkerke pseudo-R2 improved model fit compared to null model
P = 2.11x10^{-9}

PRS *
environmental/comorbid risk factors
Mendelian randomization - origins

**THE LANCET, MARCH 1, 1986**

**APOLIPOPROTEIN E ISOFORMS, SERUM CHOLESTEROL, AND CANCER**

Six... It is unclear whether the relation between low serum cholesterol levels and cancer is causal. In many studies occult tumour may have depressed cholesterol levels though in others the relation was found when serum cholesterol had been measured many years before the cancer was diagnosed. The relation is probably not explained by diet, because in the Seven Countries Study cohorts with widely different diets and corresponding differences in mean cholesterol levels experienced similar mean cancer rates. On the other hand, within each region cancer incidence was higher in men with a serum cholesterol in the lowest part of the cholesterol distribution for that country. Thus, naturally low cholesterol levels are sometimes associated with increased cancer risk.

Differences in the amino acid sequence of apolipoprotein E (apo E) are major determinants of differences in plasma cholesterol levels within a population. Apo E has a key role in the clearance of cholesterol from plasma. The synthesis of apo E is under the control of three independent alleles, located at a single gene locus, coding for the major isoforms E-2, E-3, and E-4 with respective population frequencies of about 8%, 77%, and 15%. The homozygous E-3/E-3 is the most common phenotype encountered and E-2/E-2 is the least common. From apo E-2 to apo E-3, one cysteine residue is replaced by arginine, and from apo E-3 to apo E-4 another cysteine residue is replaced. As a result the avidity of apo E containing lipoproteins for lipoprotein receptors increases from apo E-2 to apo E-3 to apo E-4. In several populations, including the Finns and the Japanese (Dr G. Utermann, personal communication), the gradient in serum cholesterol levels in the population is associated with a gradient in apo E phenotype, E-2 being associated with lower serum low-density lipoprotein and total cholesterol levels than E-3 and E-4. Thus, if a naturally low cholesterol favours tumour growth, then subjects with the E-2/E-2 or E-2/E-3 phenotype should have an increased risk of cancer.

Unlike most other indices of lipid metabolism, apolipoprotein aminocacid sequences are not disturbed by disease, and the apo E phenotype found in a patient will have been present since birth. A comparison of apo E phenotypes in cancer patients with those in matched controls might thus shed light on the relation between low cholesterol and cancer. If it is causal then the E-2 allele should be more common among patients and E-3 and E-4 more common among controls. On the other hand, equal distribution of apo E phenotypes among cases and controls would suggest that the association between low cholesterol and cancer is spurious.

Measurement of apo E phenotype by isolectric focusing of plasma is a routine determination in lipid laboratories; epidemiologists interested in cholesterol and cancer should include it in their studies.

Department of Human Nutrition, Agricultural University, 6703 BC Wageningen, Netherlands

**MARTIJN B. KATAN**

Nature’s RCT
Many observational studies evaluate associations between risk factors and disease

Risk factor → Disease

$p<0.05$
Sometimes associations between risk factors and disease arise from reverse causation rather than causation.

Risk factor $\leftarrow$ Disease

$p < 0.05$
Another explanation is confounding, that is another factor explains an apparent association between a risk factor and disease.
Associations between behavioural, socioeconomic and physiological factors assumed to be independent occur more frequently than expected by chance.
Z is an instrumental variable if:
1. It is robustly associated with X
2. Independent of C
3. Given X and C, independent of Y (exclusion restriction criterion)
MR assumptions

Z is an instrumental variable if:
1. It is robustly associated with X
2. Independent of C
3. Given X and C, independent of Y (exclusion restriction criterion)
MR assumptions

Z used for causal inference about effect of X on Y

1. Use the association between Z and Y & Z and X (ratio) to determine magnitude of effect of X on Y

2. In MR, Z is a SNP (or many SNPs) associated with a given exposure/risk factor
Mendelian Randomization—
A Journey From Obscurity to Center Stage
With a Few Potholes Along the Way
Multiple SNPs comprise $Z$ and capture maximum variance in BMI ($X$)
MR - instruments

Sample 1
BMI GIANT consortium

Sample 2
PD IPDGC consortium

Wald ratio

\[ Z \xrightarrow{\beta} X \]

\[ Z \xrightarrow{\text{Log odds ratio}} Y \]

\[ X_Z \xrightarrow{\text{Log odds ratio}} Y \]

\[ X_{Z1} X_{Z2} X_{Zn} \]
MR - instruments

Advantages of multiple variants

• Maximum variation in exposure trait capture which in turn increases statistical power
  • Similar analogy to multiple RCTs being conducted
  • Can pool effects using standard meta-analysis methods
• Can explore effects/influence of horizontal pleiotropy
MR - instruments

If MR assumptions are upheld, each SNP represents an independent experiment.

Effect estimates can be pooled together to ascertain the overall causal effect.

Use standard meta-analysis methods weighted by inverse variance.
MR – handling pleiotropy

The problem is, these assumptions are rarely upheld.
MR – handling pleiotropy

Horizontal pleiotropy can be identified using methods for heterogeneity
Cochran’s Q - used in meta-analysis to assess heterogeneity between studies
I2 statistic and p-value

‘Significant’ heterogeneity in Wald ratios could indicate a variety of problems

• One (or several or all) of the SNPs is exhibiting horizontal pleiotropy
• Non-collapsibility of binary trait, different covariate distribution, different causal relationship
After clumping, there were 78 independent SNPs:

- Associated with BMI ($p<5\times10^{-8}$)
- Together these explained 2.2% of the variance in BMI ($R^2=0.022$)
Causal estimate of the effect that 5 kg/m² higher BMI has on PD

Suggests 18% lowering of PD risk
R² 7%
80% power OR
<0.9 or >1.1

Results suggest modulation of urate should not be prioritized for neuroprotection in PD

Kia et al, Ann Neurology 2018;84(2):191-199
Williams et al, Ann Neurology 2020 [early view]
Researchers recently announced that SURE-PD3, a Phase III clinical trial evaluating the potential of inosine to slow Parkinson’s progression, will end earlier than planned. At a regularly scheduled meeting, the study’s Data and Safety Monitoring Board (DSMB), a group of independent experts, reviewed trial progress and available data. Based on the study’s primary measure and timeline, the board determined that SURE-PD3 would be unable to show that inosine slows Parkinson’s progression. While disappointing, this illustrates the importance of solid trial design and measurement tools and the value of research participation. Though the outcome is not what participants and researchers hoped for, there is still much to gain.

Parkinson’s drug trial gives definitive answer on possible statin treatment

Initial results show simvastatin does not slow progression of Parkinson’s disease and should not be investigated further as therapy for the condition.

A major clinical trial of a potential new treatment for Parkinson’s disease (PD) has found that the statin under investigation holds no promise as a protective therapy in PD.

PD-STAT, which has been running since 2016, and at its start was the largest academic study in the UK investigating neuroprotective drugs in PD.

It examined whether simvastatin, a widely-used cholesterol-lowering drug, had the potential to reduce the rate of neurodegenerative decline in patients with PD of moderate severity.

The study has now provided robust evidence that simvastatin, in comparison with a placebo, was ‘futile’ in slowing the rate of progression of Parkinson’s disease and that a phase III trial should not be recommended. A futile study is designed to test a new treatment over a relatively short period to determine whether it is worthy of larger and longer-term studies or should be abandoned.

PD-STAT was led by Dr Carmel Carroll, Associate Professor in the University of Plymouth’s Faculty of Health and Honorary Consultant Neurologist at University Hospitals Plymouth NHS Trust (UHP). It was managed by the Peninsula Clinical Trials Unit (PCTU), part of the University’s Faculty of Health, and sponsored by UHP.
MR pitfalls

Good practice

- Strong instrument (F stat)
- Sufficient sample sizes
- Clumping thresholds
- Multi-variant
- Sensitivity analyses
- Steiger filtering

What/Why?

- Weak instrument bias
- Power calcs, low power nulls
- Prevents double counting
- Single variant pitfalls
- Consistency across these
- Avoids reverse causation
Is it genetic, doctor???
Objectives

• Discuss genetic association studies and genome wide association studies of PD risk and progression.
• Discuss the creation, uses and limitations of polygenic risk scores.
• Explain the rationale for Mendelian randomization, show some applications and discuss limitations.
Extra viewing/reading

• parkinsonsroadmap.org/gp2/
  – Training and Development page
    • Introduction to complex trait genetics
    • GWAS and secondary analysis
    • Beginner Bioinformatics for Parkinson’s disease Genetics
    • Genetics for non-geneticists [coming soon]
Acknowledgements

- Include IPDGC
- GP2
- 23andMe
- QMUL
- UCL
Thanks for listening
a.noyce@qmul.ac.uk