

Maladie de Parkinson

**Une entité cliniquement définie, une
seule entité étiologique ?**

OUTLINE

- Parkinson's Disease
- Rational of Genome-Wide Association Study
- Association testing for a Disease
- Specific issues of GWAS
- GWAS of Parkinson
- Missing Heritability & the question of a single disease entity
- MeMoDeeP project

I.

What is Parkinson's Disease?

Described by James Parkinson in 1817

Maladie **neuro-dégénérative** (perte progressive d'une population spécifique de neurones : les neurones à dopamine de la substance noire du cerveau)

Ces neurones sont impliqués dans le contrôle des **mouvements**

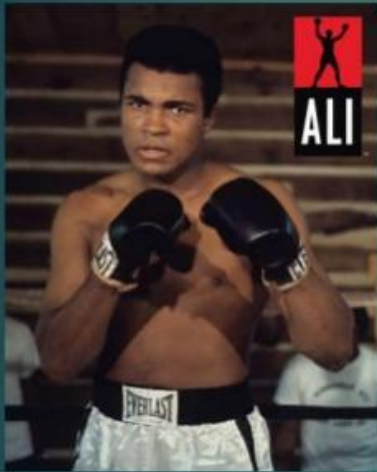
Neurodegeneration: A major health care burden

In a pop of 50 M about 1.5 M will be directly affected by neurodegenerative disease and, as the population ages, this number will increase

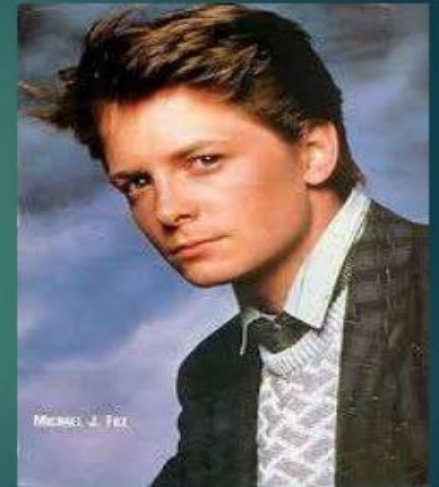
- **Alzheimer's** afflicts 1% at age 60; 20% at age 80y
- **Parkinson's** afflicts 1% over age 60; 10% over 80y
 - Slight Male predominance
 - Typical onset between 50 and 75 years
 - Average age at onset : 62 years
 - 10% have onset before 40 years
 - May be less prevalent in China and other Asian countries and in Afro-Americans

Famous personalities

3

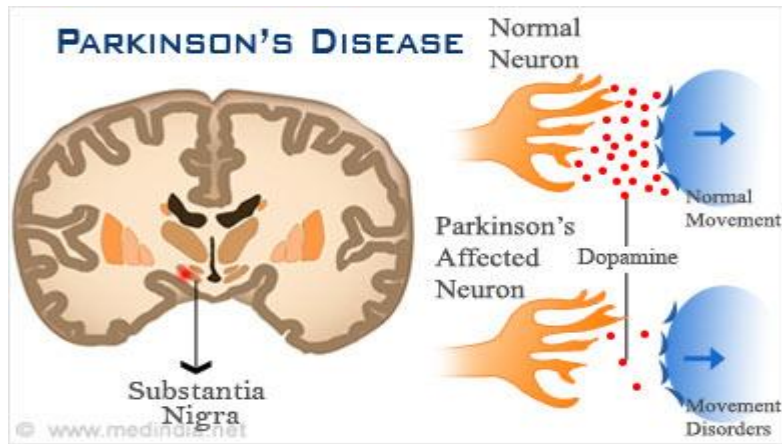


4



Caused by the loss of dopamine brain cells (neurons)

PD affects the region of the brain that controls movement



Dopamine cell death

-> lesser dopamine

-> impairment of movement
in body parts

Symptoms begin when ~ 50%-80% of dopamine neurons have died -> Movement disorder & worsens over time

• **No permanent cure:** Medication (**L-Dopa**) & surgery manage its symptoms

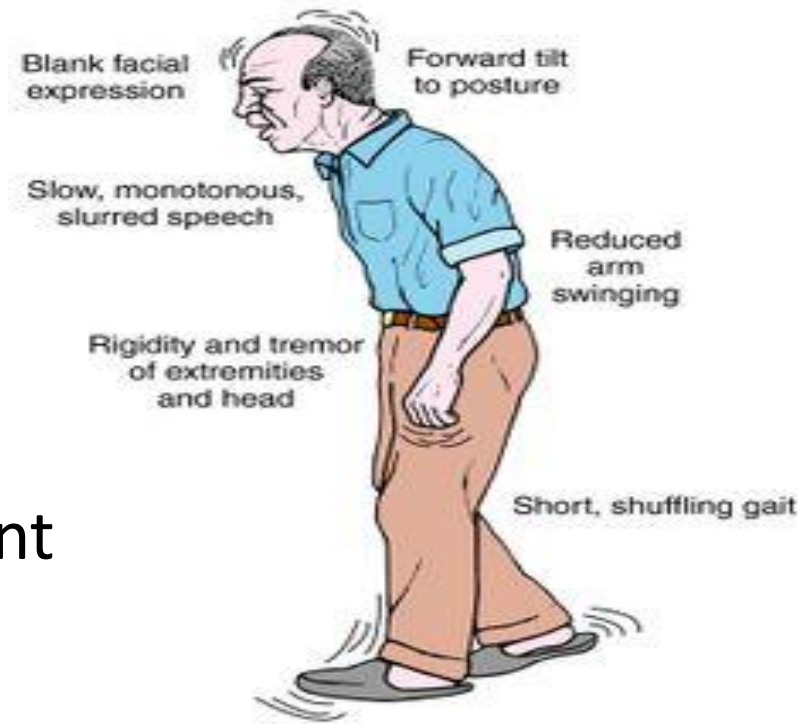
Major Clinical features

- Tremor
- Rigidity
- Bradykinesia: slowed movement
- Dyskinesia : involuntary movement (head, neck or upper extremities)
- Postural instability

Major symptoms: Motor

Heterogeneous (Dominant-tremor & non-dominant- tremor patients)

& Non-Motor symptoms



PD: Heterogeneous clinical symptoms

Major symptoms: Motor

Heterogeneous (Dominant-tremor & non-dominant- tremor patients)

& Non-Motor symptoms

Olfactory dysfunction, Cognitive impairment, Depression, sleep disorders, Constipation,...

Objective biomarkers are lacking & access to DNA/RNA profiles from damaged CNS-specific tissues in a large number of patients is, up to now, relatively limited

Unraveling the genotype-phenotype relationship in PD: Important challenge towards the dissection of its complex etiology

Limitation of GWAS: Build on large but retro-prospective samples of PD patients with often typical and sparse clinical measures

➔ Beyond empirical stratification of patients

Mathematical and statistical models based on prospective of recently diagnosed patients (longitudinal data)

High-dimensionality & mixture of data & genomic

-> **MeMoDeeP project**

PD : Clinical Diagnosis

No standard diagnostic test - > **Diagnosis is made clinically**

- Examination & neurological test
- Exclusion of other disorders ,...
- Imaging of the head (CT, MRI) may help

Etiology of PD

1. Environmental factors (caffeine, toxins, pesticides,..)?

- Caffeine, Tobacco : decreased risk of PD
- Pesticide & increased risk of PD : Rural > Urban ?

Smoking & Decreased PD Risk

- A meta-analysis of 44 studies (6,814 cases, 11,791 controls) has revealed a highly consistent reduction.
- Current smokers are 60% less likely to develop PD than non-smokers.
- Ex-smokers are 40% less likely to develop PD than non-smokers
- Potential mechanisms
 - third variable?
 - ↑ dopamine
 - Inhibition of MAO_B

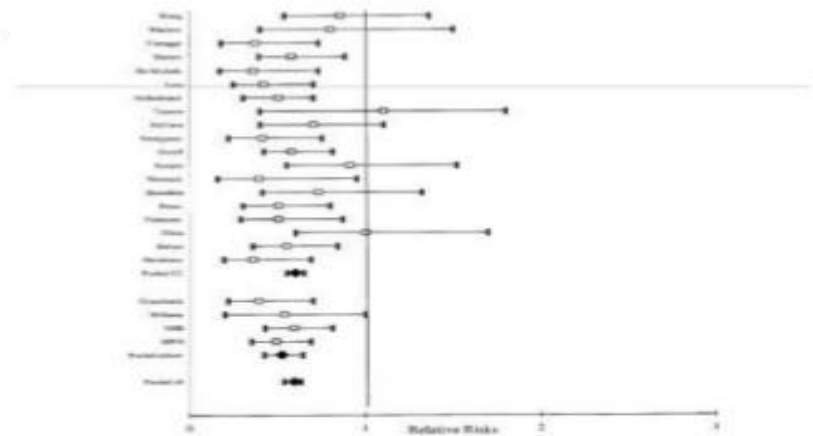


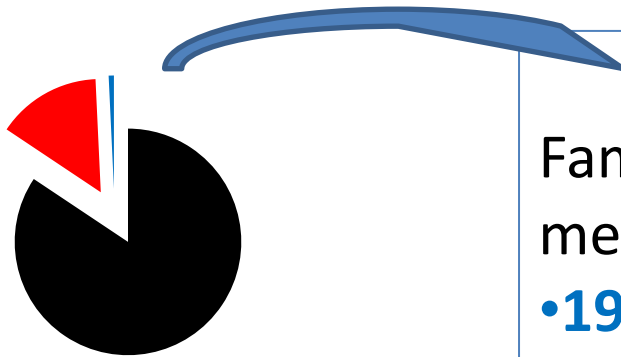
Fig 3. Study-specific and pooled relative risks from case-control and cohort studies on cigarette smoking and Parkinson's disease. CC = case-control study; NHS = Nurses' Health Study; HPS = Health Professionals' Follow-up Study.

$$\text{Relative Risk} = \frac{\text{Probability Exposed}}{\text{Probability Unexposed}}$$

Etiology of PD: Genetic factors ?

Traditionally, PD has been considered a non-genetic disorder

- Heritability ~30%
- Sibling Relative Risk ~5; Parent RR ~2
- Most PD cases (85%) : no affected relatives (censoring?)



- Isolated
- Familial-1st-degree relative affected
- Familial-Multiple Affected

Rare PD forms (1%)

Families with multiple affected members in several generations

•1997-2004

Linkage mapping -> Identified

Dominant mutations: SNCA, LRRK2, ..

Recessive mutations : Parkin, PINK1,..

Overall > **12 genes**; some Pop-specific freq.; allelic heterogeneity

Simple Diseases

- Generally rare
- Simple pattern of inheritance (D/R)
- Monogenic disorder (single/few genes; i.e., genetic heterogeneity is low)
- Environment: Low influence

Examples: Cystic Fibrosis, Huntington

Complex Diseases

- Common (1-5%)
- Cluster in families but no simple pattern of inheritance
 - Adult onset of the disease (Age is a strong risk factor: censored family data)
- Multiple risk factors (Environment, Genetic)

Examples: diabetes, asthma, cardiovascular disease, cancers, Alzheimer's disease, Parkinson's disease, and many more..

Complex Diseases

2 major theories: very controversial !

- **Common Disease – Common Variant (CD/CV)**
 - Alleles that existed prior to the global dispersal of humans or those subject to positive selection represent a significant proportion of the susceptibility alleles for common disease
- **Common Disease – Rare Variant (CD/RV)**
 - Most mutations underlying common disease have occurred after the divergence of populations
 - Expect heterogeneity in genes of common diseases

II.

Rational of Genome Wide Association Study of Complex Traits

CD/CV hypothesis

**The Future of Genetic Studies of
Complex Human Diseases**

Neil Risch and Kathleen Merikangas

Science, 1996

**On the allelic
spectrum of human
disease**

David E. Reich and Eric S. Lander

Trends in Genetics, 2001

Complex traits under CD/CV hypothesis

Susceptibility alleles confer *moderate risk* and occur at relatively *high rates* in the population (Minor Allele Frequency >1%)

Because they are *frequent* in the *population* the magnitude of their *attributable risk* (% of people affected due to them) may be *large*
-> making them of **public health importance**

Reich & Lander, 2001

Complex traits under CD/CV hypothesis

- Tests of linkage for genes of modest effects are of low power

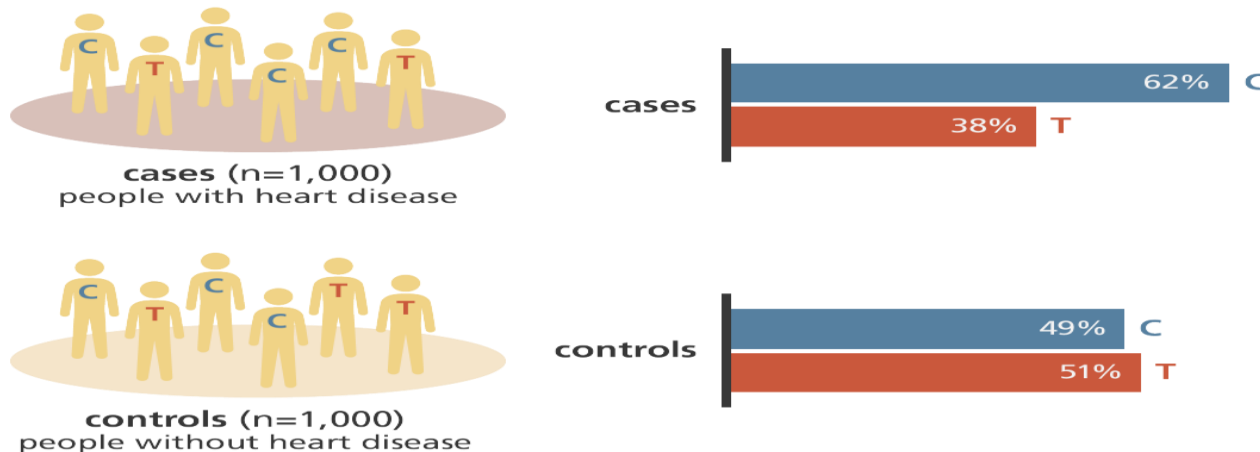
➔ The identification of the genetic basis of complex human diseases « can best be accomplished by *combining power* of the *human genome project* with **association studies**»

SNPs (Single Nucleotide Polymorphism) have facilitated this type of study: - easy to measure, stable in population

Association Study of Disease

Case-Control Design

Single-locus test



Question: are the alleles or genotypes at a genetic marker associated with disease status?

Case and control Selection

Case and control samples may be :

- Matched for known risk factors (age, gender, ..)
- Chosen to increase magnitude of contrast

Case samples may be selected to be enriched for predisposing variants(s)

- Family history
- Early age of onset
- Increased severity of disease

Control samples may be selected to be “very healthy” or “super controls”

- Individuals who have normal glucose at age 70
- Control selection just as important (and tricky) as for any case-control study

Measures of association

Absolute (Population Attributable Risk) or ***Relative*** (Odds Ratio) **differences** between groups being compared

Measure:

Usual application:

PAR

- Primary prevention impact

% of affected subjects due to the risk factor

OR

- Search for causes

$$\frac{\% \text{ affected among subjects exposed to the risk factor}}{\% \text{ affected among subjects not exposed to the risk factor}}$$

Single-locus Testing for Association with Disease

- Usual statistical machinery get **estimates** of *measures of association* and to test for association for each of the SNPs

One typical approach: additive genetic model, logistic regression

$$\text{logit} (E(Y_i)) = \log\left(\frac{P_{\text{disease}}}{1 - P_{\text{disease}}}\right) = \alpha + \beta S_i + \gamma_1 X_{1i} + \dots$$

- P_{disease} = Rate of cases in the case-control sample
 - α = baseline
 - $S_1 = 0, 1, 2$: num of minor alleles
 - X_j = Covariates
 - β, γ = regression coeff regression of genetic marker, covariates
- $\exp(\hat{\beta})$ **estimate of SNP OR** • Test of association: $\beta = 0$ (1 df)

Types of association

- Positive/Negative

Spurious association: due to chance, bias or confounding

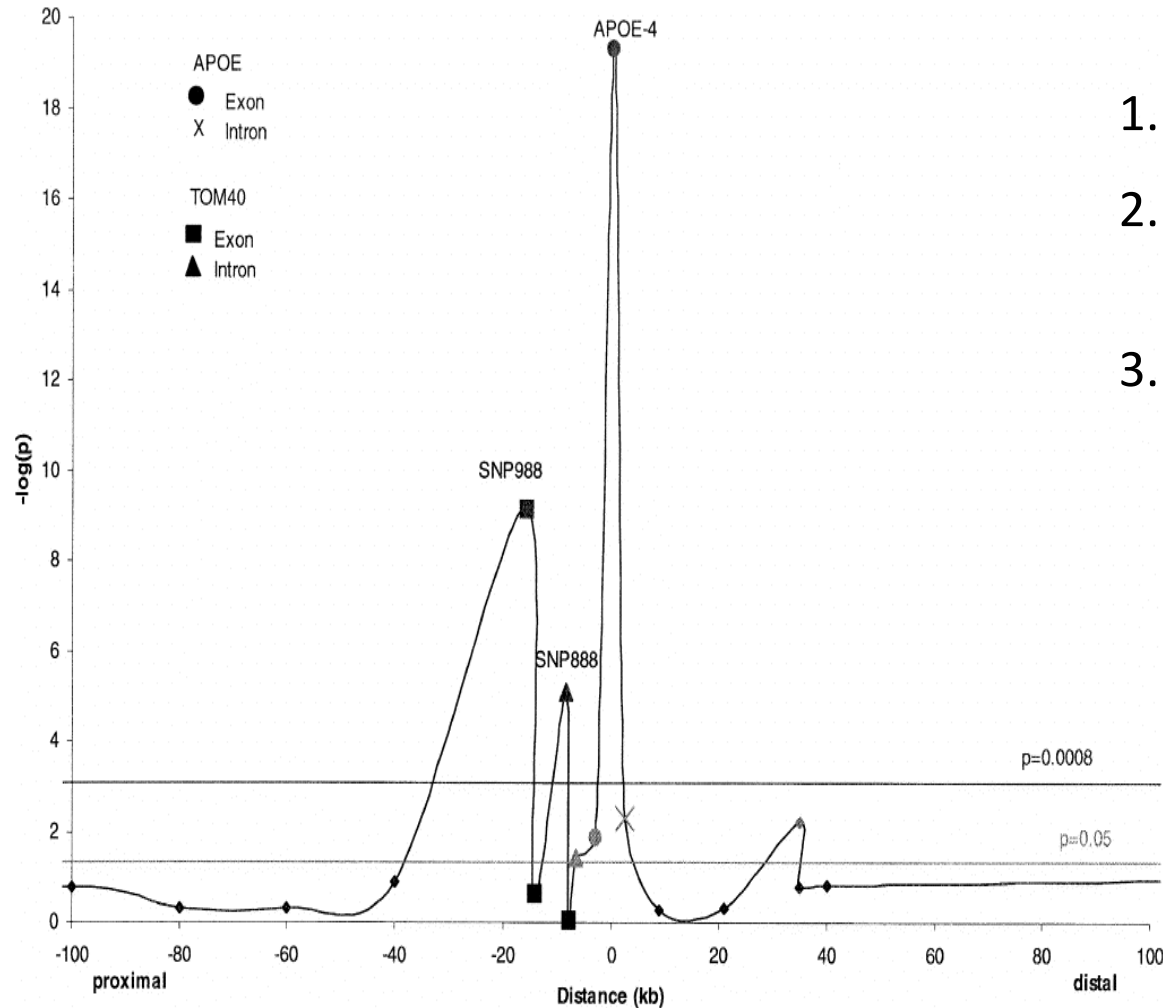
- Direct/Indirect:

Linkage Disequilibrium (allelic correlations) genetic marker and disease-risk variant

- Causal/Non causal ?

Ex: Association directe vs indirecte: Maladie d'Alzheimer & APOE

Signification statistique de l'association de différents SNPs autour de APOE (<100kb) [Martin et al., AJHG, 2000]



Puissance:

1. Max quand M=APOE
2. Diminue avec la distance M-APOE (<50kb)
3. Variation non-linéaire avec la distance : dépend de la force du DL entre M & APOE

Human Genomics brings breakthroughs in common diseases

1) Human genome sequence



2001-04

2) Haplotypic maps



2005

3) High density DNA variant arrays



2006

4) Genome Wide Association Studies

Vol. 445 | 22 February 2007 | doi:10.1038/nature05616

nature

12th of February 2007

ARTICLES

A genome-wide association study identifies novel risk loci for type 2 diabetes

Robert Sladek^{1,2,4}, Ghislain Rocheleau^{1*}, Johan Rung^{2*}, Christian Dina^{2*}, Lishuang Shen¹, David Serre¹, Philippe Boutin³, Daniel Vincent⁴, Alexandre Belisle⁴, Samy Hadjadj⁵, Beverley Balkau⁷, Barbara Heude², Guillaume Charpentier⁶, Thomas J. Hudson^{1,8}, Alexandre Montpetit⁴, Alexey V. Pshezhetsky¹⁰, Marc Prentke^{10,11}, Barry I. Posner^{2,12}, David J. Balding^{1,3}, David Meyre³, Constantin Polychronakos^{1,3} & Philippe Froguel^{5,14}

... >800 loci found through GWAS of common diseases



2007

Human Genome Project – a 13 year effort (1)

Human genome 3 billion bases

- Only < 2% of the human genome encodes proteins
 - ~20,000 protein-coding genes
 - Size of genes : ~3,000 bases on average
High variability (up to 2.5 million bases; largest= dystrophin)
- Other than protein coding genes
 - Noncoding RNAs (rRNA, tRNA, miRNAs,..)
 - Structural sequences

HGP – a 13 year effort (2)

- Humans share 99.9% of sequence identity
- The other 0.1% are mostly SNPs
- 10 million SNPs
 - SNPs occur every ~1,000 bases (1kb)
 - SNPs can cause silent, harmless, harmful or latent changes
 - Most SNPs not in coding regions (99% not in genes)

<http://www.genome.gov/10001665>

Build Map of Haplotype Blocks



- ➔ Regions of high LD interspersed by regions of low LD
- The high LD regions can form haplotype blocks

Haplotype block partition results for the three populations

Population	Blocks	Average size, kb	Requires SNPs*
African-American	235,663	8.8	570,886
European-American	109,913	20.7	275,960
Han Chinese	89,994	25.2	220,809

*Minimum number of SNPs required to distinguish common haplotype patterns with frequencies (MAF) $\geq 5\%$

Adapted from Hinds et al., Science 2005

Has had major impact on:

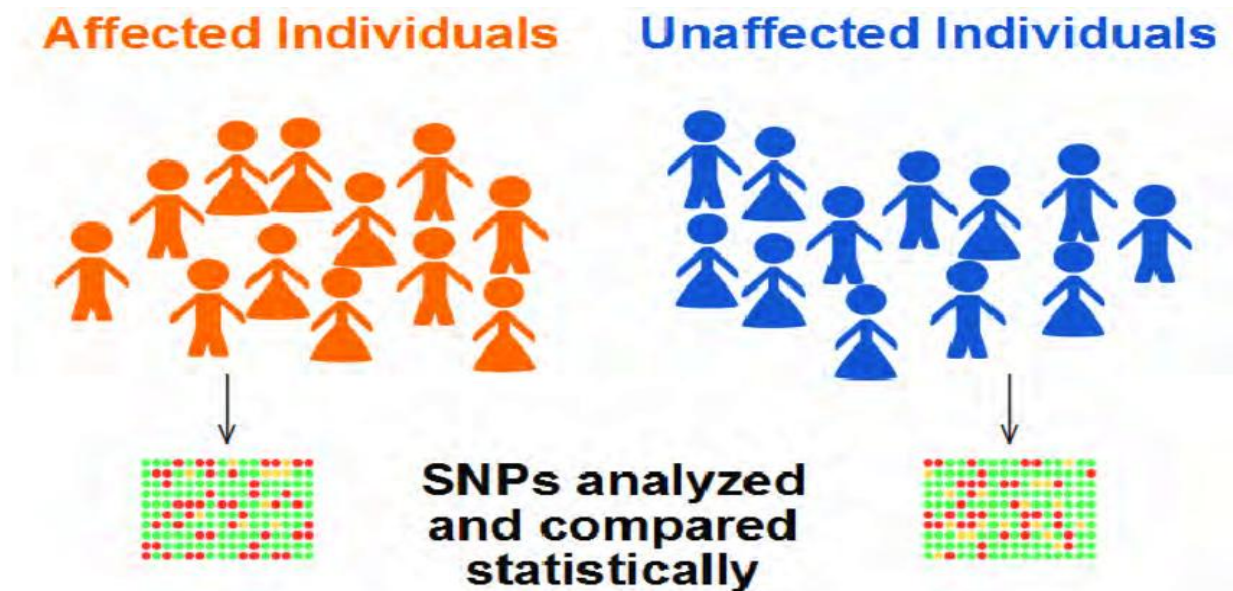


- Understanding of human pop history as reflected in genetic diversity and similarity
 - Design and analysis of genetic association studies
 - 10 M common SNPs ($MAF > 1\%$)
 - tag SNPs allow for identification of a person's haplotype
 - Estimated 300,000-600,000 tag SNPs in genome
 - High-throughput genotyping of tag SNPs (Affymetrix, Illumina chips)
- ➔ **Whole-Genome Association testing of tag SNPs**

GWAS : Design and Analysis Strategies

additional challenges

Agnostic screen
-> $M \gg \gg N$



More DATA \neq (?) More INFORMATION

For a given Complex trait

- Most genotypes are NOISE
- Many potential sources of systematic errors that might lead to false positive results
- More Tests require greater Significance at any one
Trade-off type I error vs Type II error (1-power)
- Large Ns required
Increase sensitivity to confounding & biases
- High-throughput genotyping quality control issues particularly important

Interpreting Genetic Association



Statistical
false positive?



The variant is in
linkage disequilibrium
with a disease-causing
variant (or variants)



Confounding
by population
stratification?

- Family Wise Rate Bonferroni « genome-wide » significant $P = 5 \times 10^{-8}$
- Replication study design
- Meta-analysis

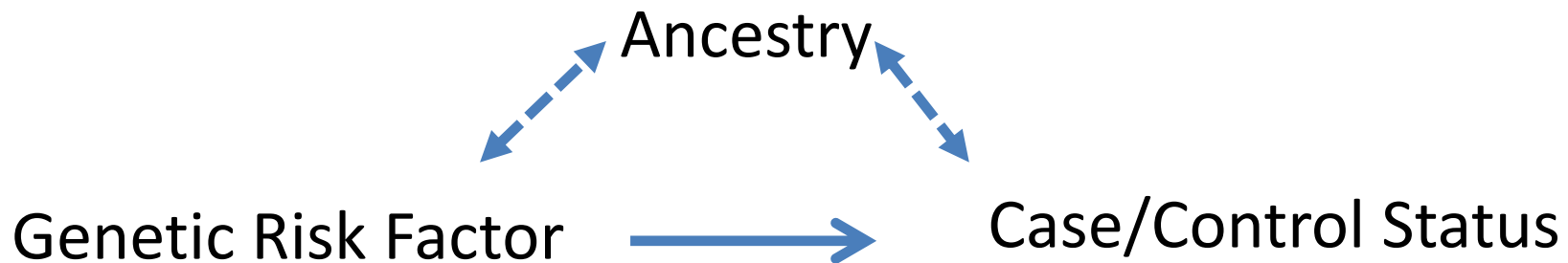
- Combine data with HapMap
- PC Analyses:
 - Identify & exclude « outliers »
 - Adjust for pop stratification

Fine-Mapping: Step-wise regression
-> Independent assoc signals

Confounding by Ancestry (Population Stratification)

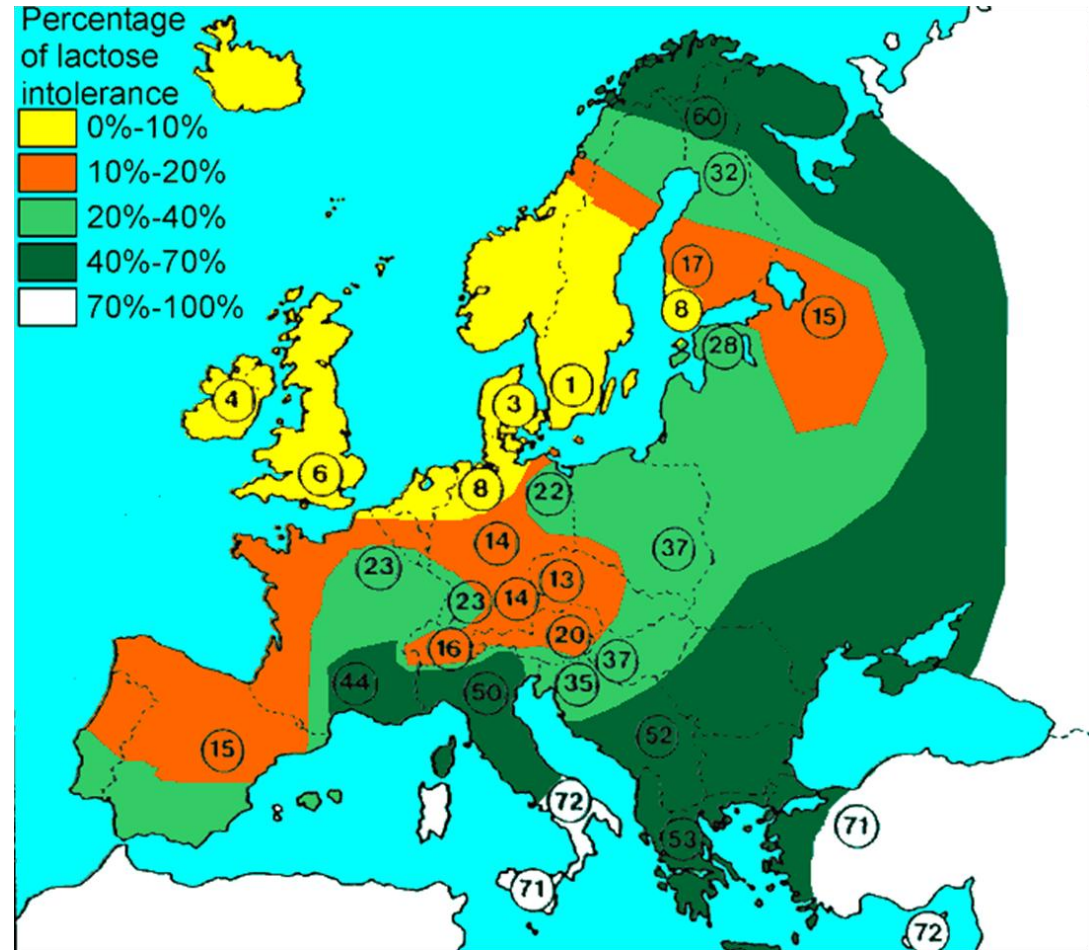
Cases / Controls selection is critical, as always

- Confounding by ancestry: Distortion of the relationship between the genetic risk factor & the outcome of interest ***due to ancestry*** that is related to both the frequency of the genetic risk factor and phenotype of the subject



Ex: Distribution of the allele frequency of lactase gene in Europe

Allele frequencies vary widely across Europe



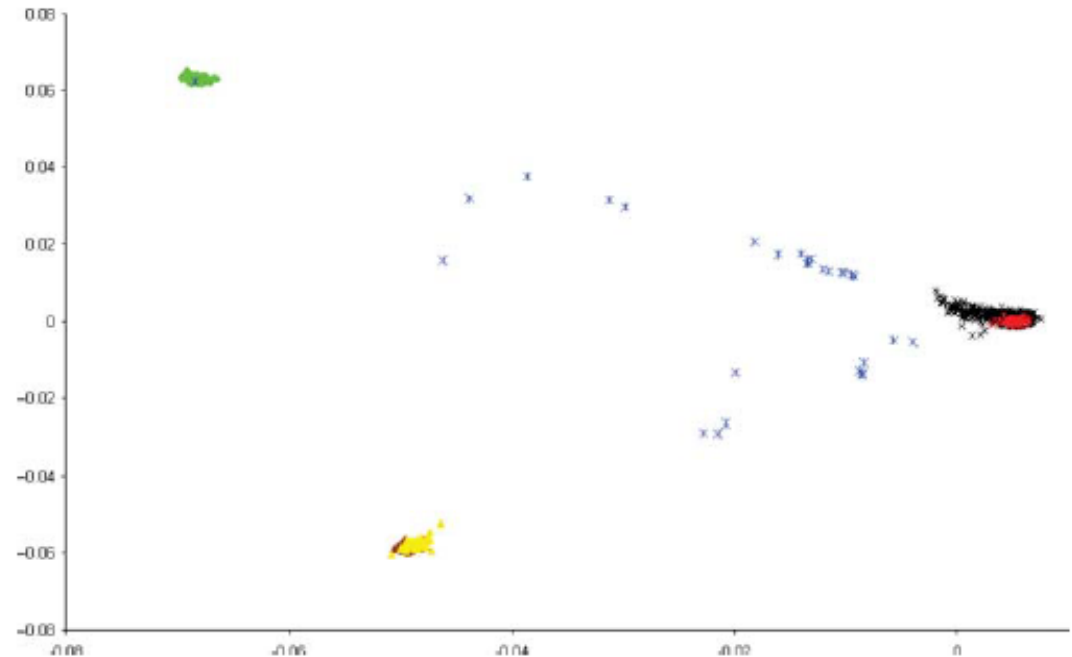
Correct/Adjust for Population stratification

- Use PCA analysis in WGA data combined with HapMap (EU, AS, AF)
 - > identify outliers and exclude them from downstream analyses
- Check empirical distribution of single-locus association test :
$$\lambda = \text{emp median of } \chi^2 / (\text{theor median value, } \chi^2 \text{ 1df})$$
if skewed, i.e., $\lambda \gg 1$
 - Genomic control: alter each single-locus test , χ^2 / λ
 - Adjust for hidden pop stratification: Main PCs included as covariates in the logistic regression

Plot of first 2 principal components from FR-GWAS data combined with HapMap data.

Human Molecular Genetics, 2011, Vol. 20, No. 3

Ethnicity of HapMap:
Africa (green),
Japan (brown),
Chinese (yellow)
Europe (red)



Study samples identified to be non-European or not clustering with European samples (**outliers**) are colored in **blue** and the remaining samples assumed to be of European origin are colored in black.

Sample size N required to achieve 80% power at a significance level $P < 10^{-6}$

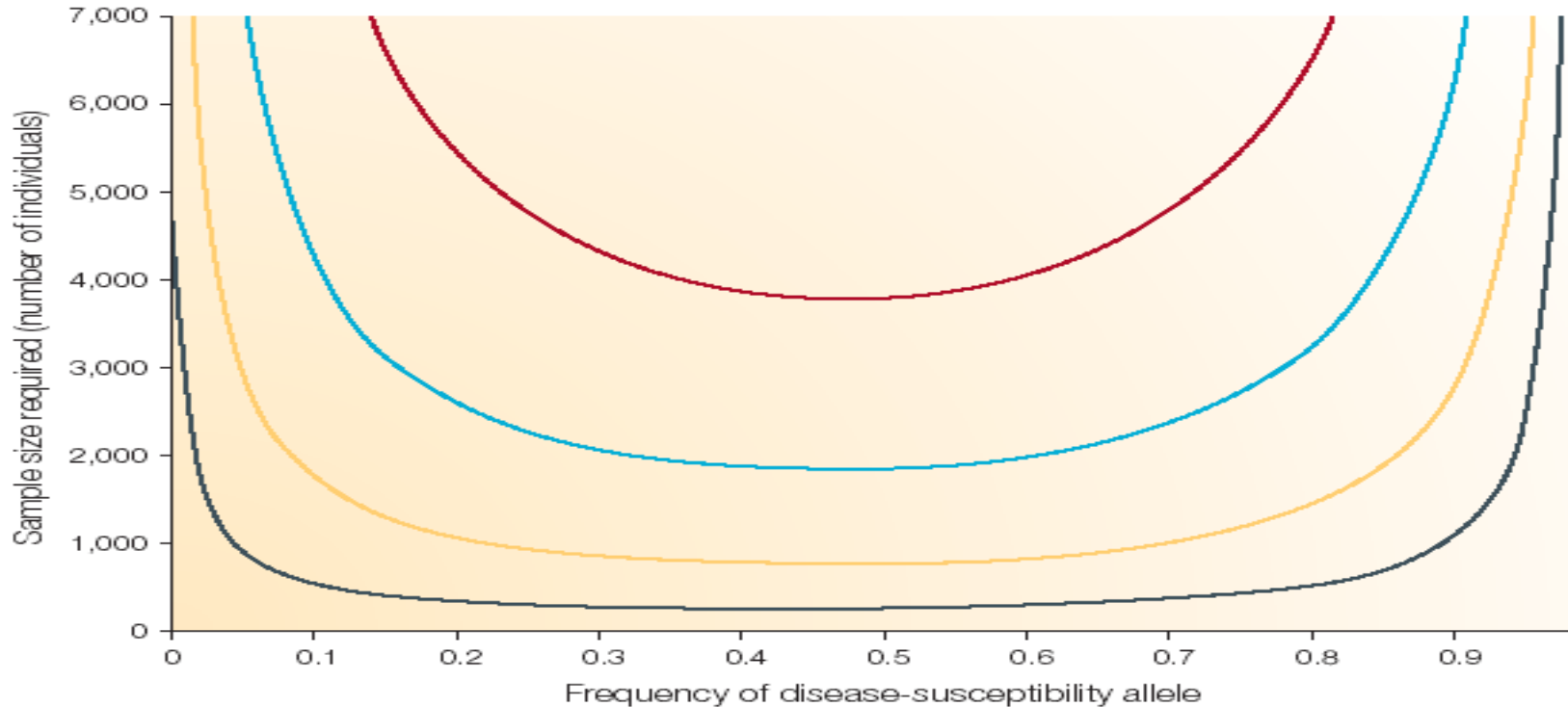


Figure 1 | **Effects of allele frequency on sample-size requirements.** The numbers of cases and controls that are required in an association study to detect disease variants with allelic odds ratios of 1.2 (red), 1.3 (blue), 1.5 (yellow) and 2 (black) are shown. Numbers shown are for a statistical power of 80% at a significance level of $P < 10^{-6}$, assuming a multiplicative model for the effects of alleles and perfect correlative linkage disequilibrium between alleles of test markers and disease variants.

Allelic OR= 1.2; 1.3; 1.5 and 2
Complete LD ($r^2=1$)

Multiple independent Samples

- « **Winner's Curse** »: the odds ratios (genetic effects) of the most significant SNPs in a GWAS are biased : higher than the true odds ratios
- **REPLICATION: → Unbiased estimates of ORs**
 - Requires independent samples from the same population as the original study (***Discovery sample***)
 - Replication should be in the same direction and be consistent with the same genetic model

Replication limited to the top K most associated SNPs (identified in the discovery stage) -> preserves power

Significance in replication studies : $P < 5 \times 10^{-8}$

Meta-Analysis

- Meta-analysis involves combining the results of several studies to obtain an overall conclusion
- Meta-analysis of GWAS can discover more associated SNPs than the individual studies
 - Greater sample size -> **more power**
- Results of meta-analysis also need replication
 - Meta-analysis retains any biases present in the individual studies
- 2 main methods: Fixed effects & Random effects meta-analysis (assumption on homogeneity of ORs)

Multilocus GWAS Models

Risk prediction

- **Polygenic Risk Scores**

k indpt SNPs associated can be combined

$$\text{PRS} = \beta_1 S_1 + \beta_2 S_2 + \dots + \beta_k S_k$$

β_j estimated from previous independent sample

- **Mixed linear models**

Limited to the **k loci** ($h^2_{\text{GWA-loci}}$) or not ($h^2_{\text{Whole-Genome}}$)

III.

GWAS of Parkinson's Disease

Individual GWAS of PD: discovery samples & results

2010, Hamza
2000 PD / 1986 Cont
US

SNCA; HLA

2011, Do (23andMe)
3426 PD / 29624 Cont
US

**SNCA; MAPT; LRRK2;
HLA; GBA + 2 new loci**

2009, Simon-Sanchez
1713 PD / 3978 Cont
US/GE

**SNCA
MAPT
LRRK2**

2009, Satake
988 PD / 2521 Cont
Jap

**SNCA
+ 2 new
PARK16
BST1**

2011, Saad
1039 PD / 1984 Cont
FR

**SNCA
MAPT
BST1**

2011, WTTC-PD
1705 PD / 5175 Cont
UK

**SNCA
MAPT**

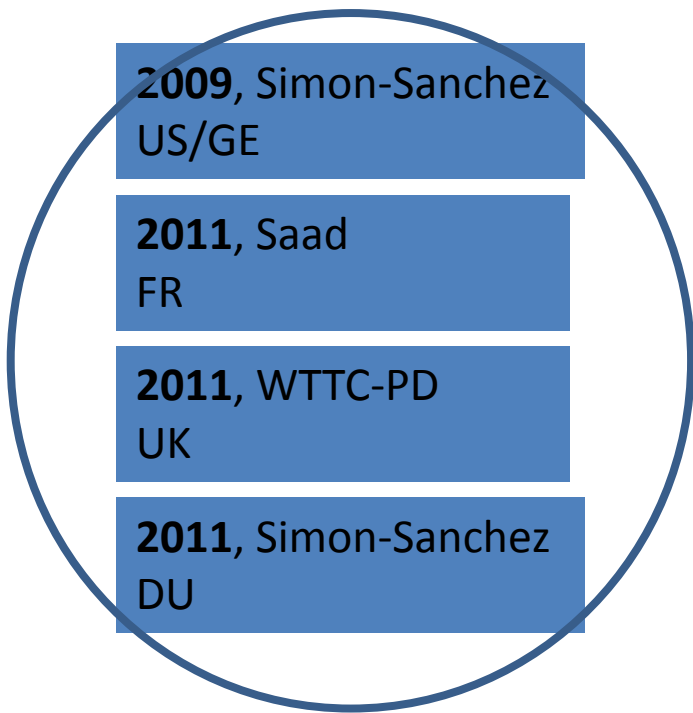
2011, Simon-Sanchez
772 PD / 2024 Cont
DU

**SNCA
MAPT
BST1**

- ➔ **Shared genetic factors between multifactorial forms of PD and**
- Monogenic forms of PD (SNCA, LRRK2) but heterogeneity in risk variants
 - Other neurological disorders (taupathies : MAPT; Gaucher: GBA)
- ➔ **Few “knew” risk variants detected: Individual GWAS lack power**

Transition to Meta-Analyses:

International Parkinson Disease Genetic Genomics Consortium



2009, Simon-Sanchez
US/GE

2011, Saad
FR

2011, WTTC-PD
UK

2011, Simon-Sanchez
DU

-Combine statistical data (OR, SD, direction of effects)

-Discovery Phase: 5,333 PD/12,019 Controls

-Replication phase: 7,053 PD/9,007 Controls
Complemented with imputations

+ Follow-up study using 23andMe data

- **17 loci genome-wide significant** (Lancet, 2012)

- 6 previously reported (SNCA; MAPT; HLA-DRB5, BST1, GAK, LRRK2) and PARK16 (Japanese pop)

- **new loci**

- **PAR (Attributable Risk) = 60%**

REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

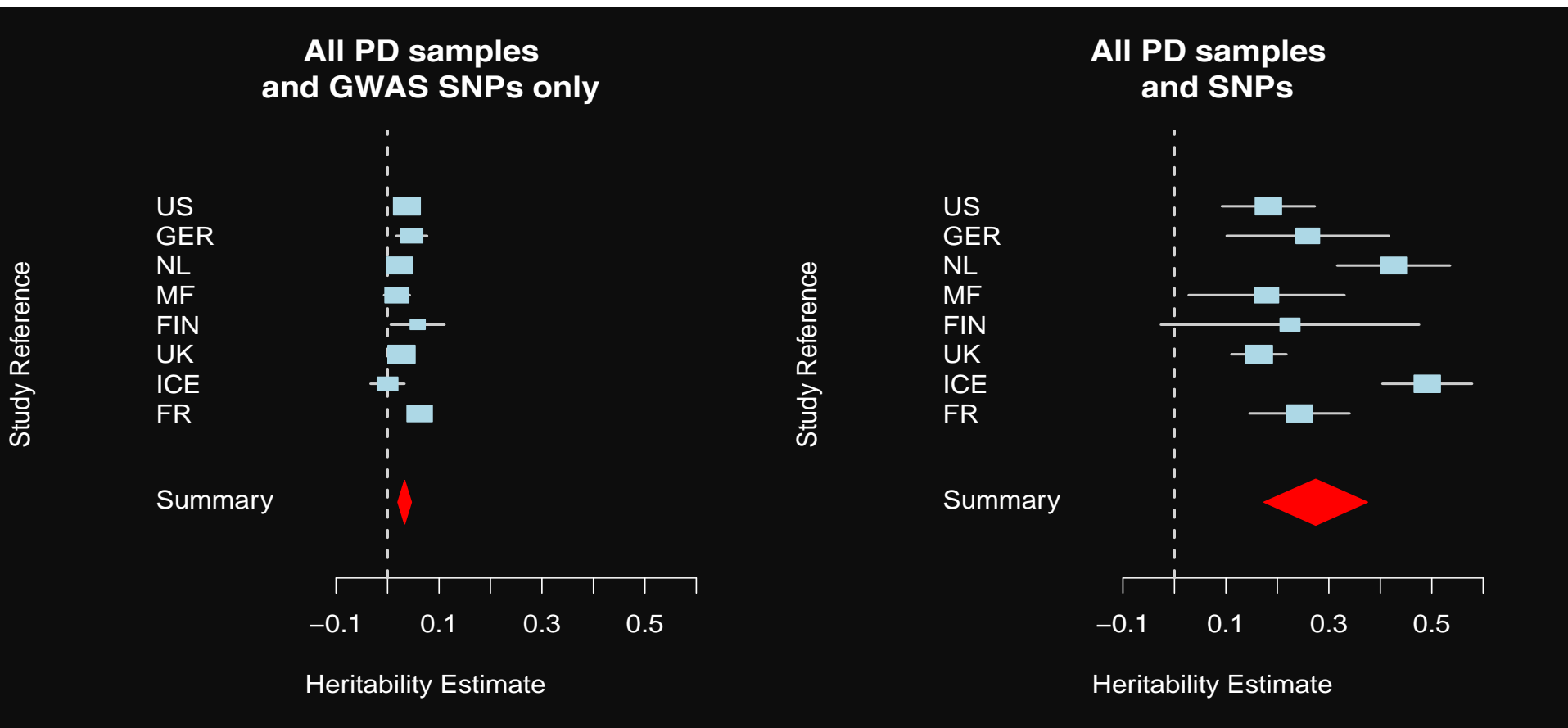
Jian Yang,^{1,*} S. Hong Lee,¹ Michael E. Goddard,^{2,3} and Peter M. Visscher¹

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

Margaux F. Keller^{1,2}, Mohamad Saad^{3,4}, Jose Bras⁵, Francesco Bettella⁷, Nayia Nicolaou⁸, Javier Simón-Sánchez⁸, Florian Mittag³, Finja Büchel³, 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 Guerreiro⁵, Dena G. Hernandez^{1,5}, Alexis Brice^{13,14,15,16}, Pauli Ylikotila¹⁷, Hreinn Stefánsson⁷, Kari Majamaa¹⁸, Huw R. Morris^{11,12}, Nigel Williams^{11,12}, Thomas Gasser^{9,10}, Peter Heutink⁷, Nicholas W. Wood^{5,6}, John Hardy⁵, Maria Martinez^{3,4}, Andrew B. Singleton¹ and Michael A. Nalls^{1,*} for the International Parkinson’s Disease Genomics Consortium (IPDGC) and The Wellcome Trust Case Control Consortium 2 (WTCCC2)[†]

GCTA analysis – Heritability estimates

The 17 GWA loci account for ~10% of the detectable heritability -> more to be found



Further Extending PD data: PD Mega-Meta-GWA -- Discovery Phase

Study	N Cases	N Controls	Total N	%Male Cases	%Male Controls	Markers	Markers Passing QC	λ_{Raw}	λ_{1000}
Ash Jewish	268	178	446	TBD	TBD	11572500	7241832	1.006	1.028
IPDGC-DC	604	4916	5520	TBD	TBD	11572501	6698963	1.061	1.057
IPDGC-FR	985	1984	2969	58.80%	67.00%	11572501	7641834	0.854	0.889
IPDGC-GE	667	937	1604	60.20%	52.00%	11210634	7486133	1.025	1.032
IPDGC-NE	744	2019	2763	63.60%	43.82%	11217965	7576956	1.061	1.056
IPDGC-NIA	937	1896	2833	39.50%	47.20%	11247278	7620408	1.035	1.028
IPDGC-UK	1705	5200	6905	56.70%	50.50%	11272513	7686314	1.034	1.013
HIHG	574	619	1193	63.07%	34.57%	11914767	7613933	0.998	0.997
NGRC	1956	1982	3938	67.74%	38.70%	11914767	8163392	1.013	1.007
PGPD	828	852	1680	59.90%	39.79%	11914767	7249203	1.009	1.011
23&Me	4127	62037	66164	60.58%	59.48%	7840733	7729624	1.212	1.027
	13395	82620	96015						

Nalls et al., Nat Genet, 2014

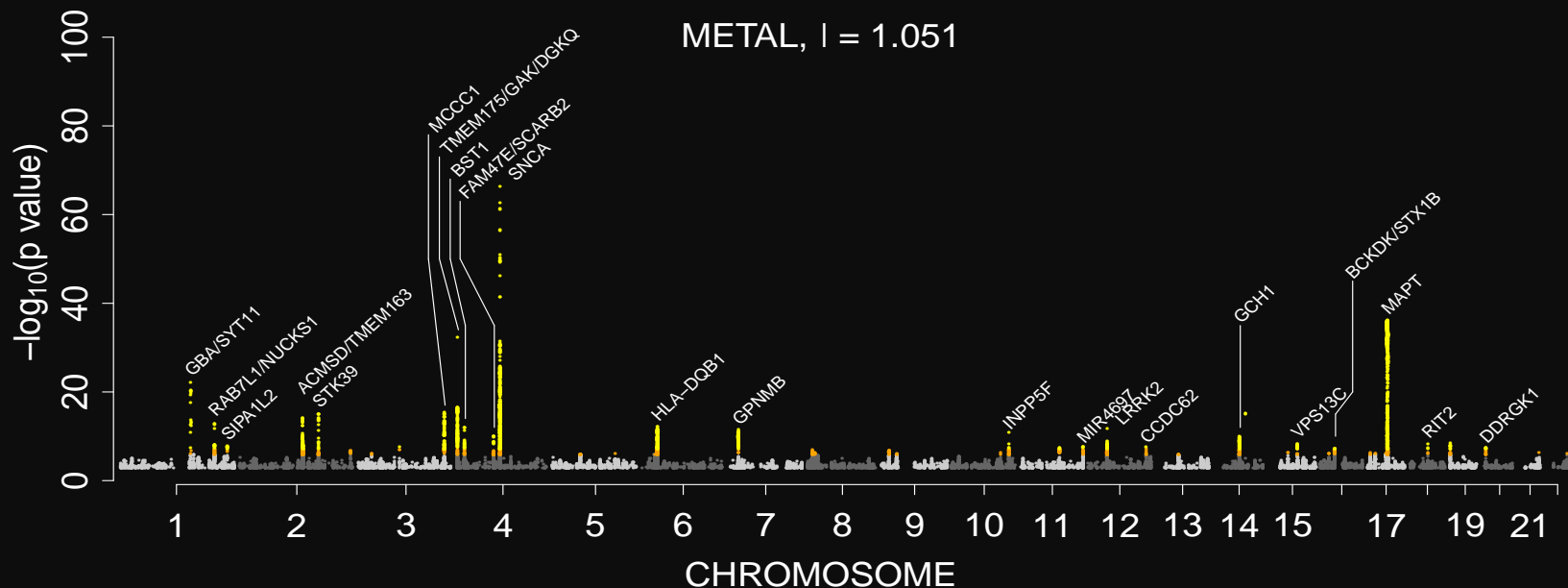
Mega-Meta-Analysis: RESULTS

1. Discovery phase : 30 genome-wide significant risk loci identified (12 novel)

2. Replication Phase (7000 PD/7000 cont): → 28 independent risk loci

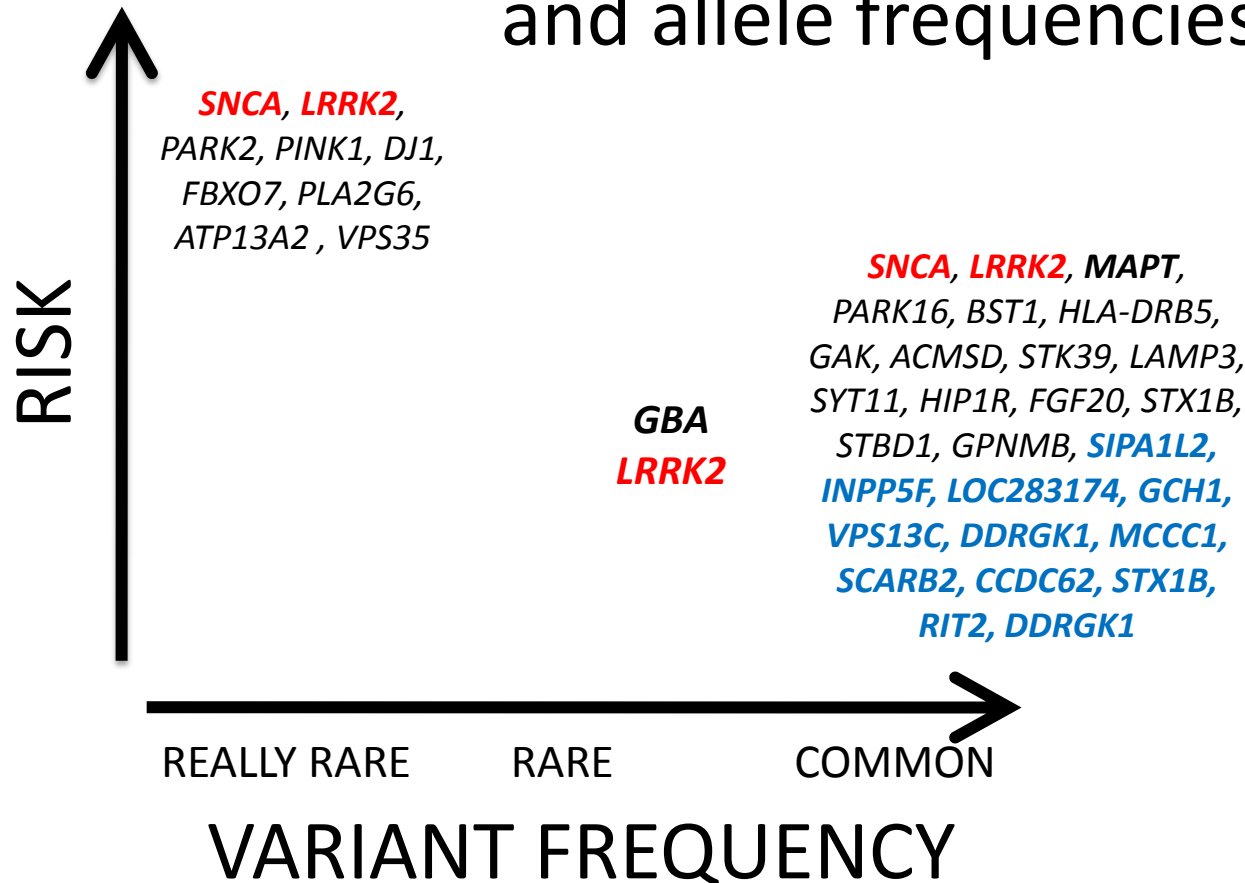
GBA, *GAK/DGKQ*, *SNCA* & *HLA* each contain > 1 independent risk allele

Manhattan plot:



Genetic Architecture of PD

Continuum of variants of different effects strengths and allele frequencies



Red: Genes explaining Monogenic forms of PD; Bold: Genes involved in other neurological disorders; Blue: Novel genes identified in the Mega-Meta-Analyses

Insights from GWAS of PD

- Mendelian vs Complex disease:
Continuum with substantial within-gene allelic heterogeneity
- Identification of **novel** candidate loci & new biological knowledge about genes and PD that was otherwise absent a decade ago
- However, enthusiasm should be tempered
How do the associated variants influence disease risk?

How do the associated variants influence PD risk?

- Most of these SNPs (GWAS-SNPs) have weak effect; the likelihood to develop PD only increased by factor 1.2–2.0
- The association signal can span multiple genes.
- Indirect association = Functional variant not identified

-> GWAS-SNPs explain little of genomic heritability

-> Not useful to predict an individual risk to develop PD

IV.

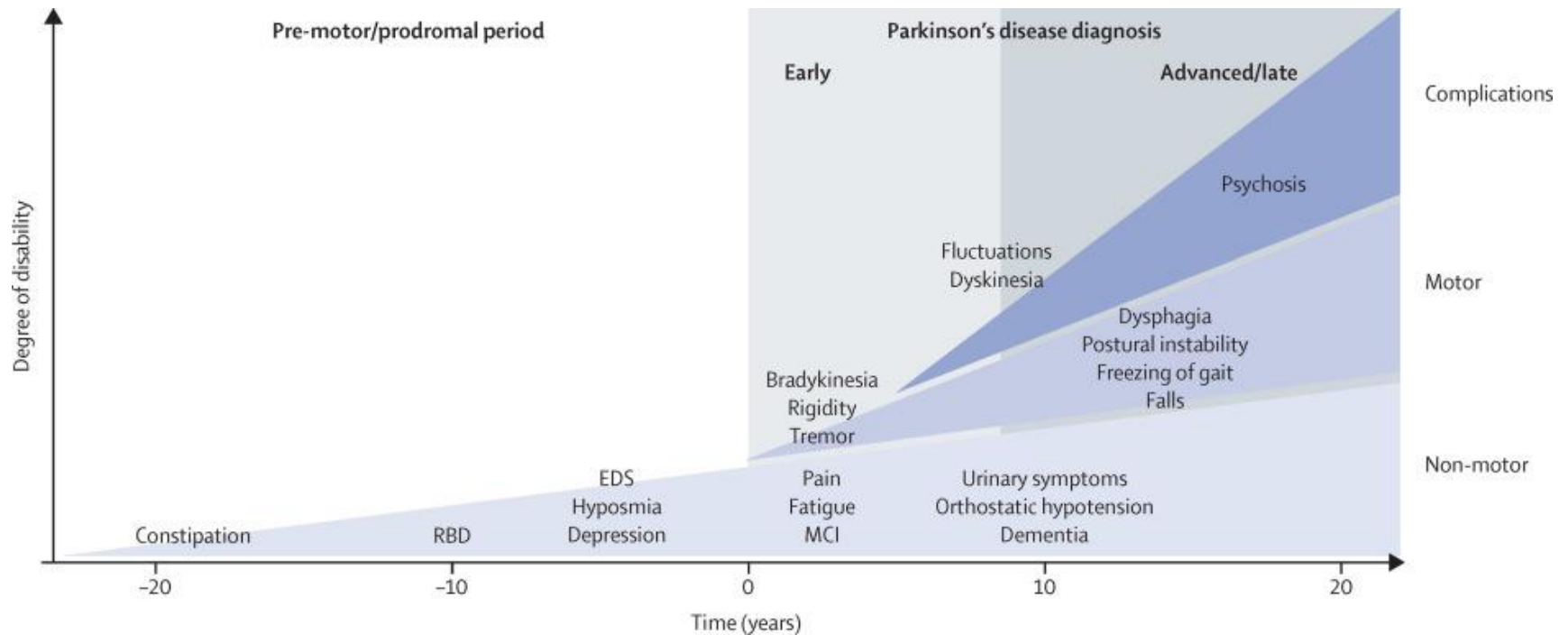
Missing Heritability & assumption of a Single Etiological Disease entity in GWAS

Differentiating specific subphenotypes can help to define more accurately the PD spectrum and the prediction of disease risk/progression

PD is associated with **non-motor symptoms**

-Olfactory dysfunction, Cognitive impairment, Depression, sleep disorders, Constipation,..

Some non-motor symptoms may precede the motor dysfunction by a decade



Unraveling the genotype-phenotype relationship in PD:

Important challenge towards the dissection of its complex etiology

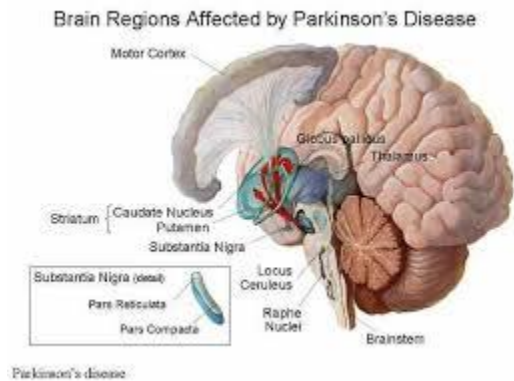
Limitation of GWAS: Build on *large but retro-prospective* samples of PD patients with often typical and sparse clinical measures

➔ Beyond empirical stratification of patients

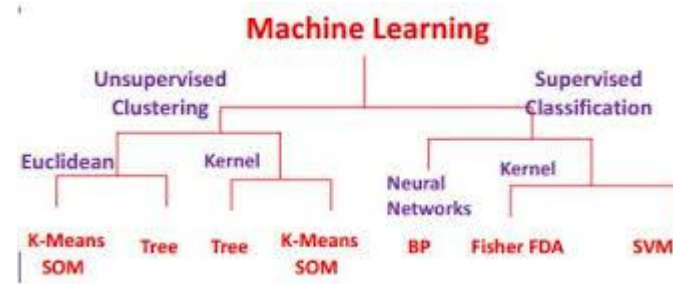
Mathematical & statistical models based on prospective of recently diagnosed patients (*longitudinal data*)

High-dimensionality & mixture of data & genomic

-> MeMoDeeP project



Methods & Models for Deep screening of subphenotypes in Parkinson's Disease



MeMoDeeP Project

Mathematical & statistical models on longitudinal data

1- Unsupervised clustering methods

Disentangle the different underlying basic-components (i.e., clusters, subphenotypes) for the joint analysis of a mixture (binary, quantitative, categorical and longitudinal) of outcomes.

2. Linear mixed models (MLM)

Assess the ***meaning of the identified subsets***

An inherent limitation of unsupervised learning procedures is that they may identify disease subtypes that may not involve specific biological processes or genetic architectures.

1- Unsupervised clustering methods

Challenging the issue of integrating large-scaled data types into a single framework and within the context of

- (1) small sample size compared to the large number of measured features;
- (2) noise in each data source;
- (3) redundancy of information provided by different data types/sources. Concatenating different data sources may result in a loss of information. Also, data-dimension reduction, as pre-filtering features (e.g. select the genes “most” differentially expressed) can certainly alleviate the large-scaled problem; but it can also lead to biased analyses.

Overall, ***current multilevel integration data approaches have yet to address these challenges and need to be evaluated.***

1- Unsupervised clustering methods

To our knowledge, no study has yet investigated such current unsupervised clustering approaches in the context of integrating a large-scaled level source of data as whole-genome DNA (several millions of SNPs) and with the objective of identifying etiologically heterogeneous components of a disease spectrum.

MeMoDeeP : We aim to evaluate the feasibility and properties of multiple kernel-based clustering methods

2- Mixed Linear Models

Univariate & Bivariate MLM analyses:

- Estimates of genetic variance & covariance to estimates
- To infer if the identified subphenotypes are genetically the same or not.

Traits that are seemingly correlated at the observational (phenotypic) level might not be so at the genetic level and vice-versa, because of confounding environmental (but not genetic) effects that push phenotypes in the same or in opposite directions.

Understanding the genetic relationship between traits allows conceiving strategies for the analysis; for instance two genetically “identical” traits can be joined together, whereas joining two distinct traits may result in spurious noise.

Resources: DIG-PD prospective cohort of PD

500 consecutive PD patients recruited at 4 sites in France (Paris, Toulouse, Nantes and Clermont-Ferrand) & followed annually for up to 6 years.

- recruitment : March 2009 - March 2016

Inclusion criteria: aged of ≥ 18 years, with a diagnosis of PD since 5 years or less

By now: 416 patients; 981 variables collected /patient at different visits; 1698 visits

Resources: DIG-PD prospective cohort of PD

At baseline:

Demographic characteristics of the patients were recorded (age, sex, ethnicity, sociocultural level, age at PD diagnosis, familial history of PD or other neurodegenerative diseases, medical and treatment history)

Questionnaire : to assess patient exposure to environmental factors (pesticides, tobacco, caffeine, alcohol,..)

At baseline and at each year of follow-up:

Clinical examination (height, weight, and blood pressure) & clinical evaluation (variety of questionnaires to assess PD severity and progression of motor and non-motor symptoms, drug adverse events, treatment history and change over time)

1- Unsupervised clustering methods

Joint analyses of longitudinal variables

Issues: number of longitudinal variables; number of genetic variants; accounting for subject's effect

Model developed by Marie Courbariaux (PostDoc, Christophe Amboise, CNRS, Evry)

Limited to <10,000 SNPs

So far, applied to the joint clustering of 4 longitudinal data

Quadratic relation with time (age / visit)

Vraisemblance du modèle : 1 variable longitudinale

Modèle de mélange à poids|logistiques

$$(Y_{v,i,j}|Z_i = k) = \alpha_{v,0,k} + \alpha_{v,1,k}t_{i,j} + \alpha_{v,2,k}t_{i,j}^2 + \sigma_{v,k}\varepsilon_{v,i,j}, \quad \varepsilon_{v,i,j} \underset{iid}{\sim} \mathcal{N}(0, 1),$$

$$\mathbb{P}(Z_i = k) = \frac{e^{\omega_{0,k} + \omega_k^T \mathbf{G}_i}}{\sum_{k'=1}^K e^{\omega_{0,k'} + \omega_{k'}^T \mathbf{G}_i}},$$

- v : variable clinique, i : patient, j : numéro de visite
- $t_{i,j}$: temps depuis le diagnostic
- \mathbf{G}_i : données génétiques
- Z_i : classe du patient i
- α , σ et ω : paramètres (à estimer).

Références : Samé et al. (2011); Montuelle et al. (2014); Schulam and Saria (2015); Courbariaux et al. (2017);...

Hypothèses du modèle : Régression $Y_v \sim \text{temps}$

$$(Y_{v,i,j} | Z_i = k) = \alpha_{v,0,k} + \alpha_{v,1,k} t_{i,j} + \alpha_{v,2,k} t_{i,j}^2 + \sigma_{v,k} \varepsilon_{v,i,j}, \quad \varepsilon_{v,i,j} \underset{iid}{\sim} \mathcal{N}(0, 1)$$

❑ Effet sujet ignoré

Ecart à l'hypothèse d'indépendance des obs

- > Paramètres de la régression: estimations biaisées
- > Inférence sur relation $Y \sim X$: biaisée (\uparrow Taux Faux positifs)
- > Biais dans l'inférence des clusters?

1.B Unsupervised clustering methods for *longitudinal variables* accounting for subject's effect:?

Latent class mixed models for longitudinal data (Proust-Lima & Jacqmin-Gadda, INSERM Bordeaux)

Issues: number of longitudinal variables; number of genetic variants; accounting for subject's effect