PPMI — Genetics Update

Christian Bracher

August 20, 2014

Overview

- Genetics Markers in the PPMI Biomarkers Database
- The Role of Apolipoprotein E (ApoE)
- Single Nucleotid Polymorphisms (SNPs) in Parkinson's-Related Genes

Genetics Markers in PPMI

PPMI provides different types of genetics markers:

- Apolipoprotein E (*ApoE*) genetic allele data
  - Data for three common forms of cholesterol transporter: ε2, ε3, ε4
- Single-nucleotid polymorphisms (SNPs)
  - Genetic data for 33 point mutations in genes coding for PD-related proteins
- SNCA multiplication data
  - All but one subject had normal number of *SNCA* copies; therefore ignored
- Illumina NeuroX Array and ImmunoChip Array data
  - Data for ca. 500,000 SNPs (!), compressed in two PLINK files
  - No attempt to decompress/analyze data yet

Why Apolipoprotein E?

ApoE has been implicated in neurodegenerative diseases:

- Individuals homozygous in the ε4 allele (ε4/ε4 carriers) have a massively enhanced risk to develop Alzheimer's Disease

  See e.g.: E. Corder *et al.*, Gene dose of ApoE type 4 allele and the risk of Alzheimer's disease in late onset families, Science **261** (1993), 921–923.

- Conflicting studies about role of *ApoE* alleles in PD:
  - Earlier studies have implicated the *ApoE* ε2 allele as a PD risk factor

    See e.g.: X. Huang *et al.*, APOE-ε2 allele associated with higher prevalence of sporadic PD, Neurology **62** (2004), 2198–2202.
  - Later studies found no connection between *ApoE* genetic status and PD

    See e.g.: M. Federoff *et al.*,. A large study reveals no association between APOE and PD, Neurobiol. Dis. **46** (2012), 389–392.

- What insights does PPMI provide?

ApoE and Parkinson's Disease

- *ApoE* genetic status determined for 558 subjects:
  - 155 healthy controls (HC)
  - 351 Parkinson's patients, confirmed by DaTSCAN (PC)
  - 52 Parkinson's patients with normal DaTSCAN (SWEDD)
- Average number of alleles in subjects:

	ApoE Allele e4 [SC]	ApoE Allele e2 [SC]	ApoE Allele e3 [SC]
global mean	0.293907	0.168459	1.537634
std dev	0.511623	0.406776	0.620866
HC mean	0.290323	0.167742	1.541935
PD mean	0.276353	0.159544	1.564103
SWEDD mean	0.423077	0.230769	1.346154

ApoE: The SWEDD Connection

- Result:
  - HC, PD subjects share very similar allele frequencies
  - In SWEDD patients, ε2 and ε4 are more common, while ε3 is suppressed
- Suspicion: *ApoE* genetic status is related to SWEDD, not PD!
- Dig a bit deeper – compare to population statistics:

	Number in PPMI	Percentage in PPMI	Caucasian Population
ApoE e2	94	0.084229	0.084
ApoE e3	858	0.768817	0.779
ApoE e4	164	0.146953	0.137

Source: L. Farrer *et al.*, Effects of Age, Sex, and Ethnicity on the Association Between ApoE and Alzheimer Disease, JAMA **278** (1997), 1349–1356.

ApoE ε4 slightly enhanced because of presence of small SWEDD cohort?

ApoE: The SWEDD Connection

- What is the distribution of the alleles within the PPMI cohorts?

	PPMI HC	Ratio HC	PPMI PD	Ratio PD	PPMI SWEDD	Ratio SWEDD
ApoE e2	26	0.083871	56	0.079772	12	0.115385
ApoE e3	239	0.770968	549	0.782051	70	0.673077
ApoE e4	45	0.145161	97	0.138177	22	0.211538

- Distribution for HC, PD subjects is very close to population average
- In SWEDD patients, both ε2 and ε4 are enhanced by close to 50%

A SWEDD – Alzheimer Relation?

- Summarize:
  - No apparent connection between *ApoE* status and PD
  - Clear correlation between *ApoE* genotypes ε2, ε4, and SWEDD
- Is there a possible link of SWEDD to Alzheimer's Disease?
  - Presence of the ε4 allele strongly increases Alzheimer risk
  - ε2 and ε4 are significantly more prevalent in SWEDD patients
  - SWEDD (but not PD) cohort has increased levels of Amyloid-β_{42} in CSF – the protein forming the hallmark Alzheimer 'plaques'
  - But: The ε2 allele has a protective effect against Alzheimer!

Source: L. Farrer *et al.*, Effects of Age, Sex, and Ethnicity on the Association Between ApoE and Alzheimer Disease, JAMA **278** (1997), 1349–1356.

Single-Nucleotid Polymorphisms in PPMI

- Genotyping results for 33 PD-related loci have been entered in the PPMI Biomarkers database
- Examine: What relationship is there between PD, SWEDD, and variants of these genes?
- Method: "Numerify" entries for sample statistics
  - Use numerical weights: 0 for mixed genotype, ±1 for homozygotes (lexicographical order: A < C < G < T)
  - Ignore three rare SNPs that are homogeneous in the PPMI population:
    - rs34995376_LRRK2_p.R1441H
    - rs35801418_LRRK2_p.Y1699C
    - rs35870237_LRRK2_p.I2020T

PPMI SNP Statistics

Methods:

- Use (3×3) contingency table for log-likelihood significance testing (three cohorts, three genotypes each)
- Consider p-values ≤ 0.05 significant, $p \leq 0.01$ very significant
  - Caveat: Rather random cut-off
  - Statistics suffers from low number of healthy controls, SWEDD subjects in study; analysis should be repeated with general population data
- Raw data for 30 SNPs in PPMI biomarker database

PPMI SNP Statistics – Results

	SNP rs10797576 -C+T [SC]	SNP rs11060180 -A+G [SC]	SNP rs11158026 -C+T [SC]	SNP rs114138760 -C+G [SC]	SNP rs115462410 -C+T [SC]
global mean	-0.685714	-0.121429	-0.319643	0.973214	-0.821429
std dev	0.512257	0.724899	0.665780	0.161601	0.410382
HC mean	-0.768750	-0.125000	-0.256250	0.981250	-0.768750
PD mean	-0.667622	-0.117479	-0.343840	0.968481	-0.862464
SWEDD mean	-0.549020	-0.137255	-0.352941	0.980392	-0.705882
p_PD	0.046825	0.565880	0.388488	0.591485	0.040399
p_SWEDD	0.000419	0.833413	0.088746	0.555024	0.707678
p_both	0.002186	0.878554	0.131778	0.658028	0.033078

	SNP rs11724635 -A+C [SC]	SNP rs118117788 -C+T [SC]	SNP rs11868035 -A+G [SC]	SNP rs12456492 -A+G [SC]	SNP rs12637471 -A+G [SC]
global mean	-0.082143	-0.958929	0.376786	-0.346429	0.608929
std dev	0.713054	0.198633	0.683956	0.669810	0.550419
HC mean	0.037500	-0.962500	0.418750	-0.368750	0.587500
PD mean	-0.157593	-0.951289	0.369628	-0.340974	0.618911
SWEDD mean	0.058824	-1.000000	0.294118	-0.313725	0.607843
p_PD	0.013057	0.734918	0.451204	0.209535	0.700560
p_SWEDD	0.981878	0.306556	0.241583	0.864458	0.902203
p_both	0.026880	0.089812	0.518683	0.420839	0.885036

PPMI SNP Statistics – Results

	SNP rs14235 -A+G [SC]	SNP rs17649553 -C+T [SC]	SNP rs1955337 -G+T [SC]	SNP rs199347 -C+T [SC]	SNP rs2414739 -A+G [SC]
global mean	0.230357	-0.582143	-0.703571	0.183929	-0.491071
std dev	0.681055	0.580268	0.498285	0.687292	0.630125
HC mean	0.275000	-0.600000	-0.706250	0.106250	-0.425000
PD mean	0.214900	-0.570201	-0.707736	0.226361	-0.510029
SWEDD mean	0.196078	-0.607843	-0.666667	0.137255	-0.568627
p_PD	0.614452	0.769203	0.233678	0.176273	0.171744
p_SWEDD	0.767963	0.988153	0.898362	0.959429	0.069569
p_both	0.891678	0.958135	0.433128	0.442113	0.150195

	SNP rs329648 -C+T [SC]	SNP rs34311866 -A+G [SC]	SNP rs34637584_LRRK2_p.G2019S -A+G [SC]	SNP rs34884217 -G+T [SC]	SNP rs356181 -C+T [SC]
global mean	-0.248214	-0.564286	0.991071	0.810714	-0.014286
std dev	0.705838	0.576346	0.094152	0.431198	0.715139
HC mean	-0.250000	-0.650000	1.000000	0.775000	0.093750
PD mean	-0.234957	-0.550143	0.985673	0.816619	-0.088825
SWEDD mean	-0.333333	-0.392157	1.000000	0.882353	0.156863
p_PD	0.510076	0.106839	1.000000	0.545141	0.011642
p_SWEDD	0.720284	0.016191	1.000000	0.199196	0.779030
p_both	0.695934	0.039363	1.000000	0.441908	0.016132

PPMI SNP Statistics – Results

	SNP rs3910105 -C+T [SC]	SNP rs55785911 -A+G [SC]	SNP rs591323 -A+G [SC]	SNP rs6430538 -C+T [SC]	SNP rs6812193 -C+T [SC]
global mean	0.123214	0.285714	0.460714	-0.064286	-0.267857
std dev	0.703299	0.677105	0.631799	0.753864	0.676492
HC mean	0.056250	0.343750	0.400000	0.037500	-0.281250
PD mean	0.171920	0.263610	0.495702	-0.083095	-0.257880
SWEDD mean	0.000000	0.254902	0.411765	-0.254902	-0.294118
p_PD	0.105898	0.291823	0.146248	0.178694	0.122539
p_SWEDD	0.597025	0.719203	0.985865	0.053984	0.415375
p_both	0.170665	0.567059	0.369086	0.098140	0.334784

	SNP rs71628662 -C+T [SC]	SNP rs76763715_GBA_p.N370S -C+T [SC]	SNP rs76904798 -C+T [SC]	SNP rs8192591 -C+T [SC]	SNP rs823118 -C+T [SC]
global mean	0.967857	0.983929	-0.717857	-0.933929	0.137500
std dev	0.176537	0.125863	0.473674	0.248629	0.705730
HC mean	0.981250	0.993750	-0.800000	-0.956250	0.112500
PD mean	0.957020	0.979943	-0.687679	-0.931232	0.128940
SWEDD mean	1.000000	0.980392	-0.666667	-0.882353	0.274510
p_PD	0.244406	0.414451	0.034928	0.359419	0.944664
p_SWEDD	0.750138	0.978122	0.194640	0.136882	0.266291
p_both	0.060561	0.442090	0.115243	0.191335	0.592092

PPMI SNP Statistics – Interpretation

Results:

- Seven SNPs show significant differences between HC and patients
- Among these, only one (rs10797576) is unspecific:
  - rs10797576 is strongly significant for SWEDD ($p < 0.005$), less significant for PD ($p \approx 0.05$)
  - Affected gene: *SIPA1L2* complex locus on chromosome 1 (function unknown)
  - *SIPA1L2* is tentatively associated with amyotrophic lateral sclerosis (ALS)

PPMI SNP Statistics – Interpretation

Two SNPs are specific for SWEDD only:

- rs34311866 has fairly strong significance ($p \approx 0.015$):
  - Affected gene: TMEM 175 (transmembrane protein 175) on chromosome 4
- rs6430538 is less significant ($p \approx 0.05$):
  - Mutation affects an intergenic region on chromosome 2

PPMI SNP Statistics – Interpretation

Four SNPs are specific for PD only:

- rs115462410 (a/k/a rs9275326) has limited significance ($p \approx 0.05$):
  - Affected gene: HLA-DQB 1 (human leukocyte antigen) on chromosome 6
- rs11724635 is strongly significant ($p \approx 0.01$):
  - Affected gene: BST1 (bone marrow stromal cell antigen) on chromosome 4
- rs356181 is strongly significant ($p \approx 0.01$):
  - Affected gene: SNCA (α-synuclein) on chromosome 4
- rs76904798 is less significant ($p \approx 0.03$):
  - Affected gene: LRRK2 (leucine-rich repeat kinase 2) on chromosome 12

Cohort Identification by Genetic Status

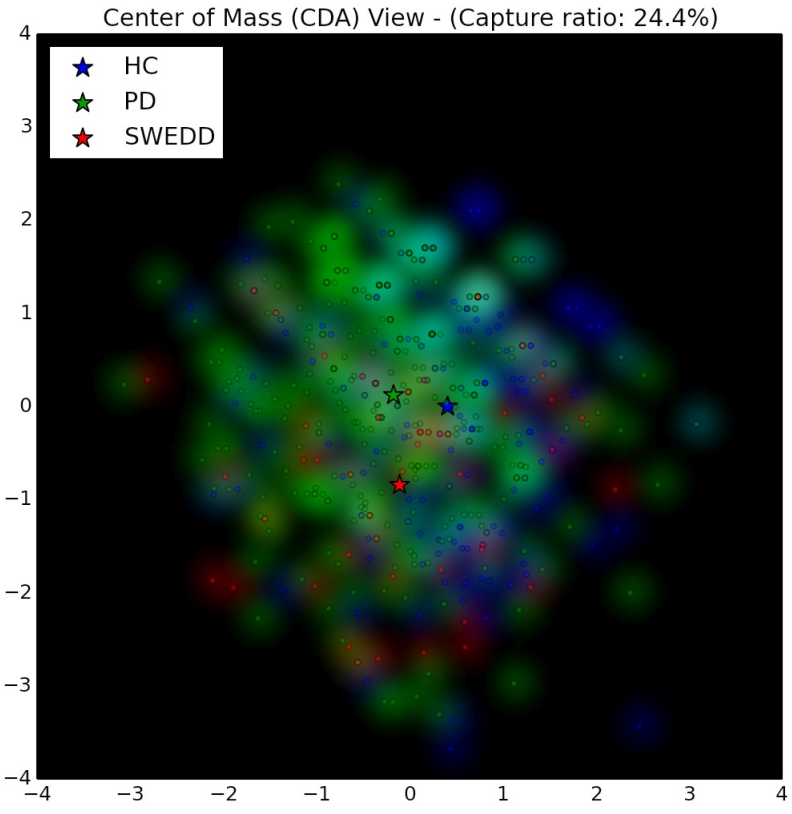
Genotype carries information about cohort membership

- SNPs represent a 30-dimensional feature space; one may add the three-dimensional space of *ApoE* genetic allele data

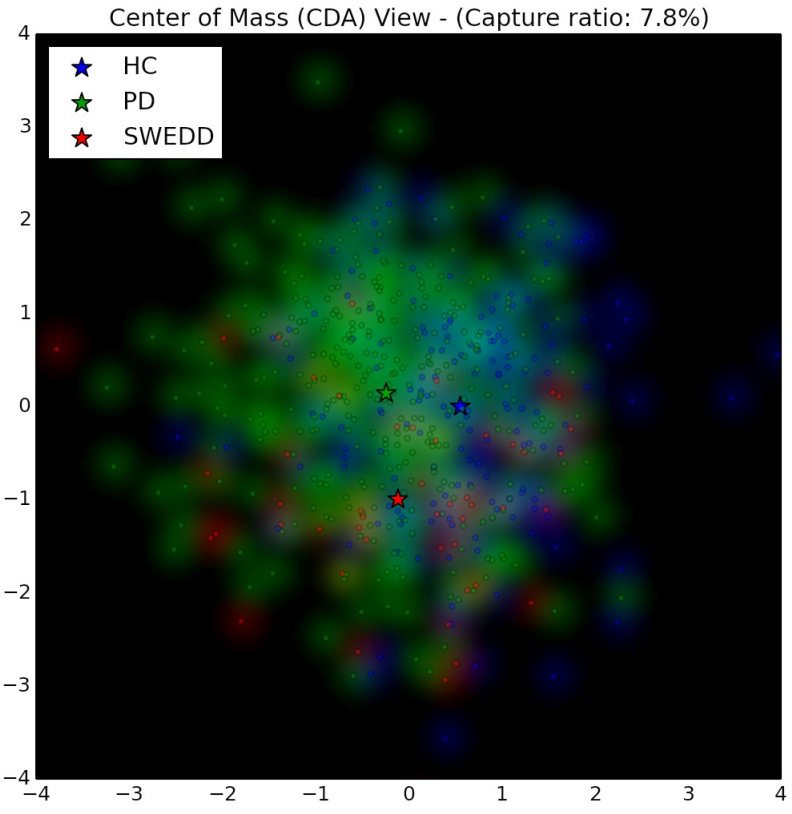
Problem: Visualize high-dimensional data structure

- Method #1: Projection – Canonical Discriminant Analysis (CDA)
  Projection on plane that maximizes distance between cohort averages
- Method #2: Clustering – Stochastic Neighborhood Embedding
  Nonlinear mapping that tries to preserve 'clusters' of data points while reducing the dimensionality of the representation

PPMI Data in SNP Space



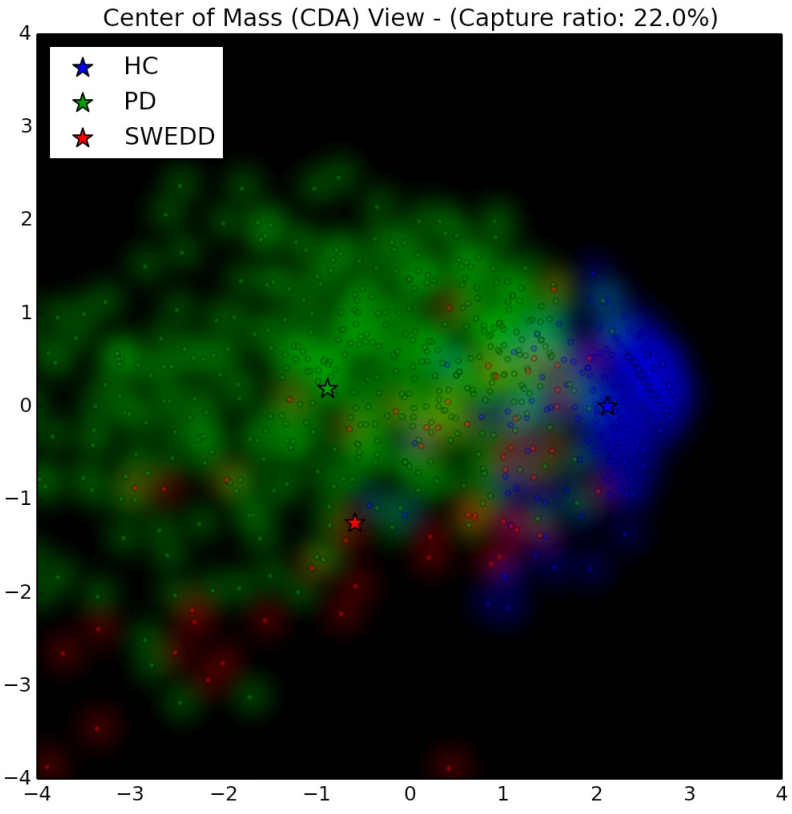
CDA – ApoE and selected SNPs



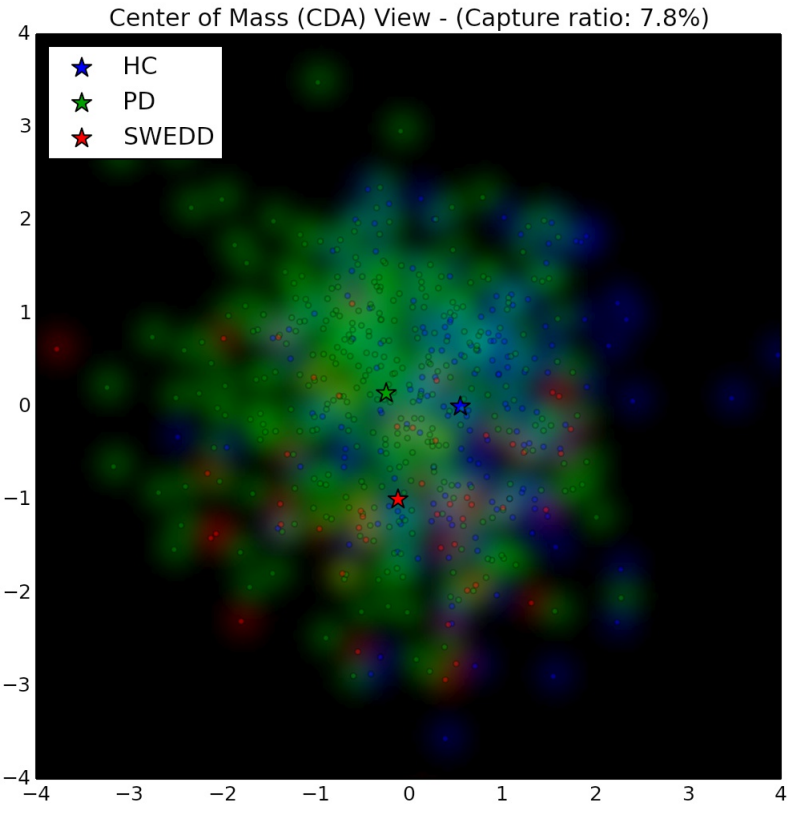
CDA – ApoE and all 30 SNPs

Separation between cohorts increases with number of genetic features included

REM Sleep-Motor-Smell vs. Genetic Data



CDA – REM Sleep-Motor-Smell features

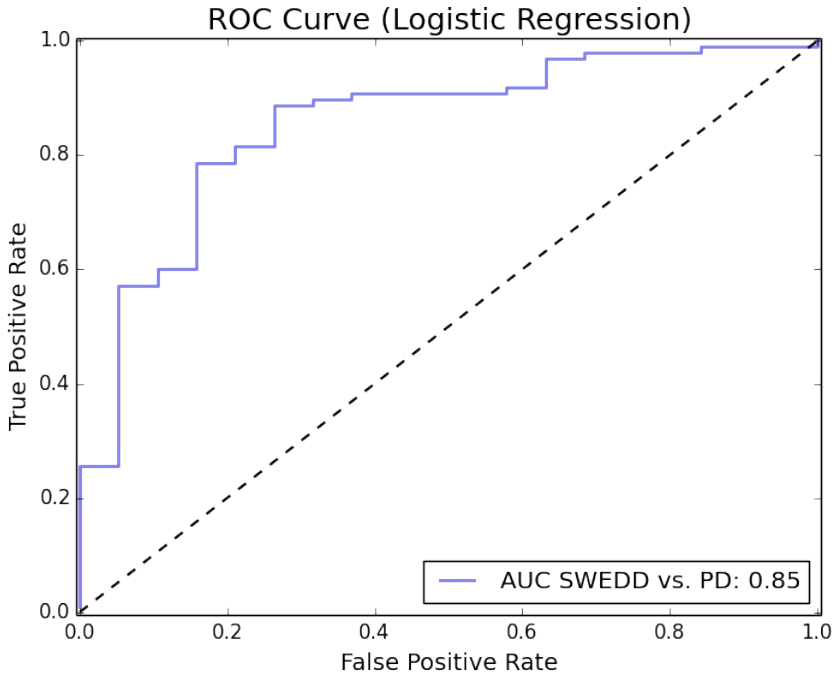


CDA – ApoE/SNP genetic features

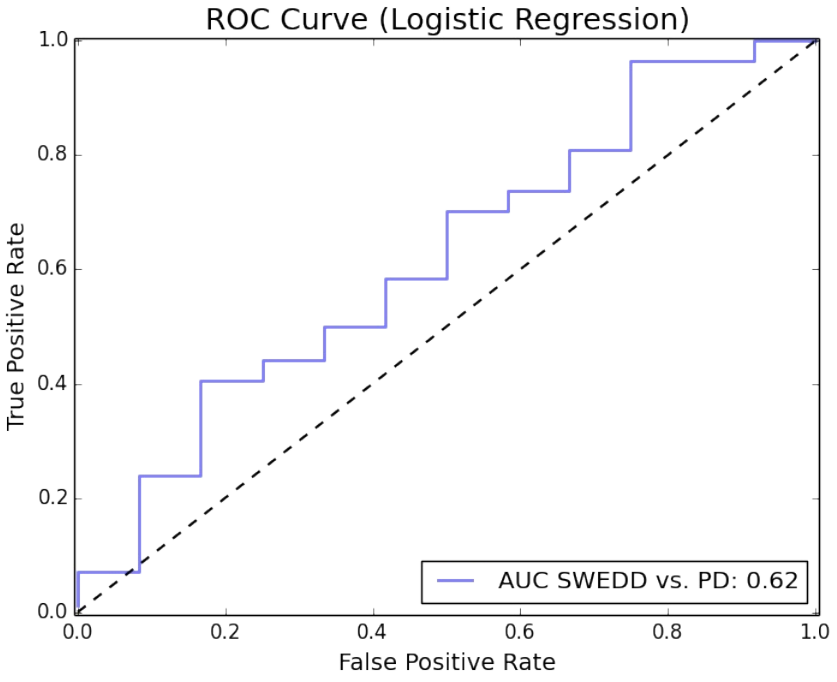
REM Sleep-Motor-Smell yields cleaner separation between PD and SWEDD cohorts

Comparison of Predictive Power

Supervised learning success for REM Sleep-Motor-Smell vs. ApoE-SNPs feature sets:



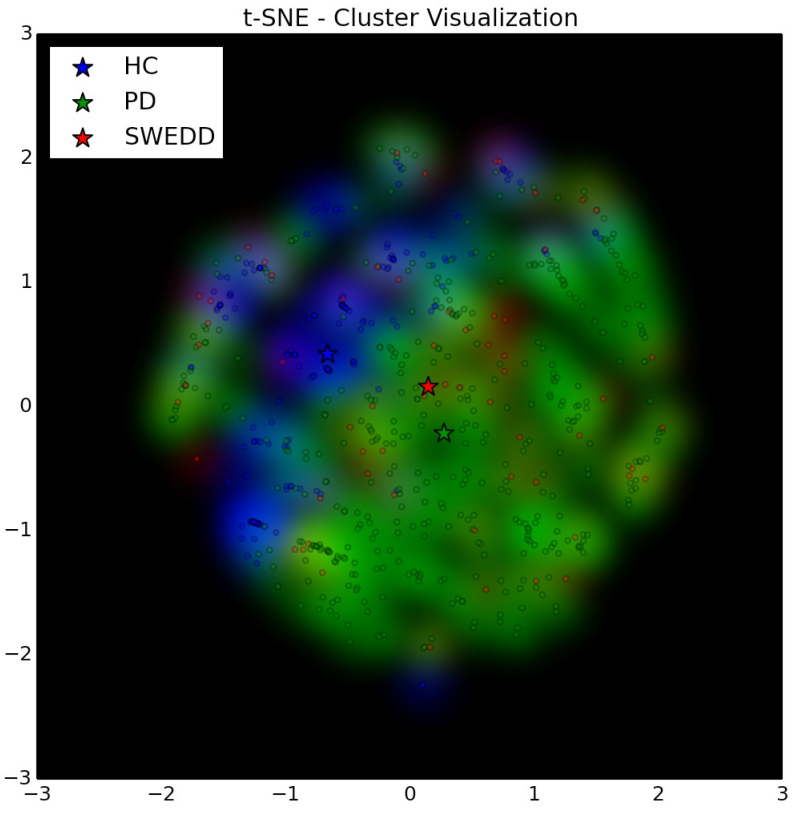
REM Sleep-Motor-Smell feature model



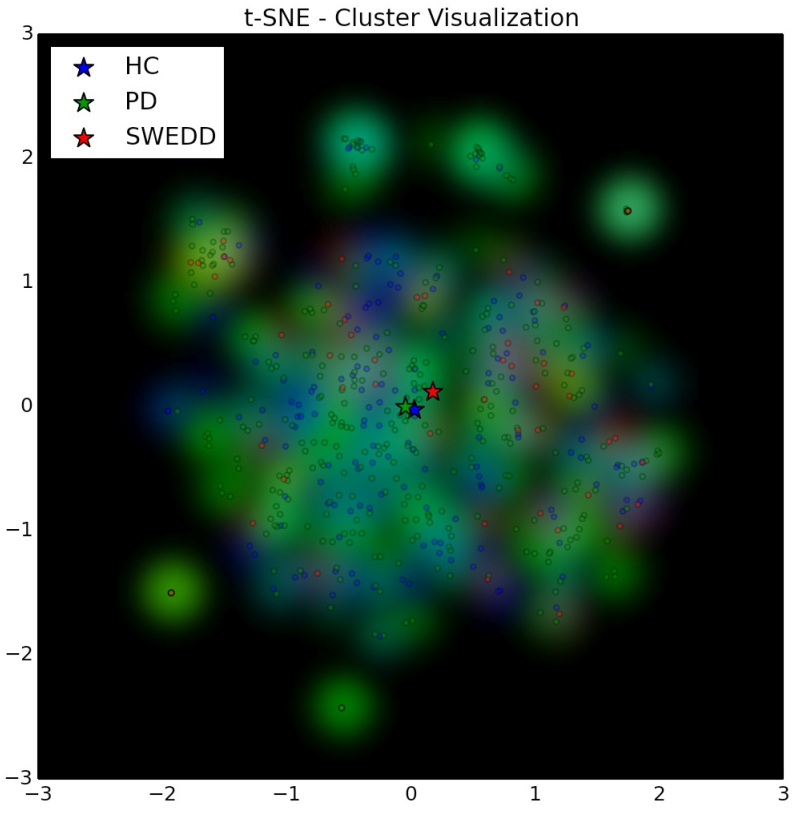
Genetic data feature model

Better separation of data yields higher success rate!

First Attempts at Clustering



t-SNE – REM Sleep-Motor-Smell



t-SNE – ApoE/SNP genetic features

Clustering algorithm (t-SNE) generates distinct “satellites” in the genetic data set

Summary: A First Look at PPMI Genetic Data

- I investigated the genetic data immediately available in the PPMI Biomarkers database: *ApoE* genotype and status for 33 SNPs.
- The ε2 and ε4 alleles of *ApoE* seem to carry an elevated risk for SWEDD. No correlation between *ApoE* genotype and PD was found.
- The PPMI population was heterogeneous for 30 of the 33 SNPs. The small size of the SWEDD cohort (about 50) and healthy controls limited the statistical power of the data.
- I found seven SNPs that carried significant risk for PD (four loci) or SWEDD (two loci). One SNP was associated with risk for both.
- The genetic features have less predictive power than the REM Sleep-Motor-Smell triad, but may be more suited to clustering analysis.

Further Investigations – Genetic Data

- Unpack the vast amount of genetic information stored in the NeuroX and Immunochip Array data files.
- Improve statistics of disease relevance of genetic features:
  - Compare to allele frequencies in general population
  - Add genetic data for SWEDD cohort to alleviate imbalance
- Examine interactions between genetic loci:
  - Study collective effects of genetic markers (correlations)
  - Identify subject clusters, find common features (Ayasdi)
  - Search for links between RNA transcription rates, protein levels, genotype

Contact Information

- Christian Bracher

cbracher69@gmail.com

http://www.christianbracher.net/

http://www.linkedin.com/in/christianbracher/

https://github.com/cbracher69/PD-Learn/