# Discovery of Novel Subtypes in Parkinson's Disease Using Germline Mutation Landscape

Xiang GU[1,2], Hongxu Ding[1,3], Jinghao Sun[1,4]

1. contribute equally, names not listed in order

2. University of Texas Health Science Center San Antonio

3. Columbia University

4. Tsinghua University

## Abstract

As a long-term neurodegenerative disease, Parkinson's Disease (PD) is less well understood and with no available therapeutics. This study utilized and mined the germline mutation landscapes of 645 individual samples consist of healthy controls, PD patients and patients without Dopaminergic Deficit. Using multiple unsupervised clustering algorithm and social networks models, we identified three subtypes of PD, which are distinct from other healthy controls and SWEDD. This is the first study using germline mutation landscapes to subtype the PD, providing significant insight into understanding PD.

## Introduction

Parkinson's disease (PD), a long-term neurodegenerative disease, mainly affects the motor neural system. Patients of PD bear with disabilities in obvious shaking, rigidity and slow movement. Unfortunately, there is no available therapeutics against the underlying neurodegenerative process currently. Therefore, it is of great importance to understand how the disease develops, how many subtypes within the disease and how to perform health care for different patients. This study aims to classify the patients based on their genetic data to identify novel subtyping mechanism for PD.

Genetic information has been successfully applied in subtyping and classifying other diseases including complex disease as cancer. Genetic mutations in the protein coding region imply potential alteration of protein structure and function which contribute to the disease development and progression. Therefore, patients can be classified the mutations they harbor which is in correlation with the cause of their disease. This theory has been proved successful in many cancer studies, including breast cancer[1] and glioblastoma[2]. In this study, we analyzed the mutation information of 645 patients obtained via Exome-seq which sequenced only the protein coding regions in current cohort, to identify novel subtyping scheme for PD.

Subtyping and identifying potential subtypes of a disease falls in the regime of unsupervised clustering problem. Numerous methods for unsupervised clustering have been developed with each's own advantages and disadvantages depending on the format and features of unlabeled data. These methods have been widely used in other fields to understand both real-life problems such as pattern mining in customers and biomedical problems such as subtyping of breast cancer[1]. Under the concept of social networks, analysis on bipartite graph formulated by the patients and their mutations proposes a novel dimension in studying subtyping of a disease[3, 4]. Therefore, in this study, we apply both unsupervised machine learning algorithms and social networks to identify the subtypes of PD.

# Method

## Preprocessing of Exome-seq Data

Exome-seq of 645 patients was downloaded from Parkinson's Progression Markers Initiative (http://www.ppmi-info.org/). Variants were annotated using hg19 with ANNOVAR and Bioconductor package VariantAnnotation[5] in R. Only the unique single nucleotide polymorphism (SNP) passing quality control filter, with amino acid sequence changed were selected. The 179,754 overlapping filtered SNPs from both packages were selected as final candidate features used for subtyping.

## Dimension Reduction

Multiple Correspondence Analysis (MCA)[6] method, which is a generalization of Principal Component Analysis (PCA), has been applied to reduce dimensions before clustering. We use an R implementation of MCA in the package: FactoMineR[7] to do our experiments. After preprocessing steps, we get an indicator matrix with each row representing a patient and each column as a SNP. Every element in the indicator matrix is either ‘y’ or ‘n’ standing for having this SNP and not having it representatively. Applying MCA to the indicator matrix with a given reduction dimension, e.g. 100 in our experiment, will gives us a new matrix with only 100 columns and every element in the new matrix has been converted to continuous value instead of categorical value. In the new matrix, every column as a dimension can be interpreted as a combination of original columns in the indicator matrix, i.e. SNPs and we can get the variance of each dimension and which SNPs contribute most to this dimension after computation. Thus, by checking the result of MCA, we may tell which groups of SNPs play significant roles in the determination of different subtypes of PD.

## Partitioning Around Medoids (PAM)

PAM is a way of implementing k-medoids clustering, a more robust version of k-means clustering[8]. To get the best clustering scheme, we use silhouette score[9] to evaluate the clustering performance under each k value (from 2 to 20). Result shows silhouette score increase as k goes up, reaches maximum value (0.22) when k ranges from 8 to 11and decreases as k is greater than 12, indicating the optimal classification scheme lies within k=8, 9, 10 and 11.

## Unsupervised Clustering

Unsupervised clustering was performed using dimensionally reduced feature matrix. Clustering is the first choice to identify subtypes of PD. There are many popular clustering techniques nowadays with their own advantages and disadvantages. We utilize a few distinct clustering methods to our reduced matrix from MCA, aiming to take advantage of their strengths and avoid their drawbacks.

In our experiments, robustness to outliers of every clustering method affect the clustering result a lot. For most clustering methods, we only use the first 10 dimensions of reduced matrix from MCA, considering that they account for a large part of variance and by doing this, we hope to eliminate much noise and avoid curse of dimensionality in some certain methods.

In this study, we applied python implementations in Scikit-learn package[10] of Spectral Clustering, Birch, DBSCAN and Hierarchical Clustering.

## Consensus Clustering

Consensus clustering can 1) determine the number of clusters and 2) assess the stability of the discovered clusters by evaluating the consensus across multiple runs of a clustering algorithm (in our case PAM clustering)[11]. Result shows a general trend that as k increases, the PAC (proportion of ambiguous clustering) score decreases. Also, no significant decrease of PAC is observed if k goes beyond 7, indicating the optimal classification scheme lies within k greater than 7.

## Affinity Propagation

Affinity propagation determines heterogeneities within data by exchanging messages between data points. Such process is repeated until a high-quality set of exemplars and corresponding clusters gradually emerges[12]. Affinity propagation gives clusters with few patients, and we consider those as non-representative. After removing clusters with less than 10 patients, we have 12 representative clusters (negative distance as pairwise similarity, clustering scheme slightly variates when different pairwise similarity measurements methods are used).

## Bipartite Network Modularity

The relationship between SNPs and patients can be modeled with a bipartite network[3]. It has been reported that the heterogeneity information with the data can be reflected by the network[4]. Based on the constructed bipartite network, we measure modularity using method developed by Newman[13]. This is an especially powerful method compared to the above mentioned ones, because the cluster specific SNPs are also highlighted.

# Result

## Dimension Reduction

The percentages of variance of the first 50 dimensions in the new matrix (MCA processed matrix) are plotted below (Figure 1). We can find that the first dimension accounts for the most variance of all, which is nearly as twice as the second one.

After calculating the correlation coefficients between every SNP and every dimension in the MCA processed matrix, here we list most influential SNPs on the first dimension along with their R-squared statistics and p-values:

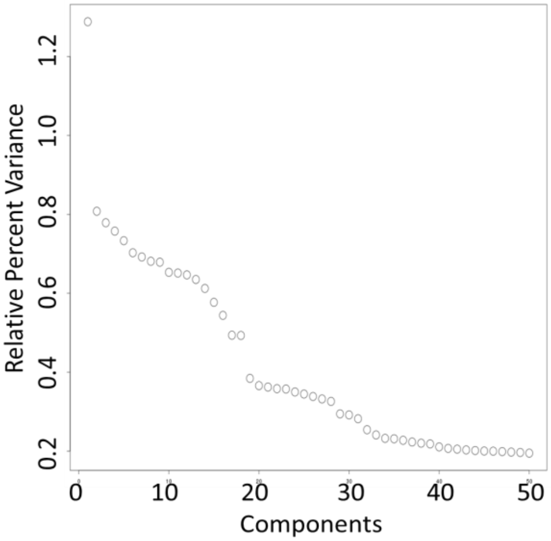

**Figure 1. Minor Components and their relative contribution in MCA.**

By looking over the most influential genes for the first several dimensions, we find that most genes are related to cell cytoskeleton and cell motion, as well as visual and olfactory functions. Also, some are associated with cell surface receptors. Although genes of great contributions to each of the first several dimensions seem not to have direct relations and interactions and even the functions of some gene remain unknown, we believe that, within each group of genes in one dimension, the mutations of genes will cooperate with others to advance Parkinson's Disease (PD) in a certain direction, which may be very different with that in other

**Table 1. Most Influential SNPs on ohe First Dimension**

| Rank | R2 | p-value | SNP |
|---|---|---|---|
| **1** | 0.784723 | 1.29E-216 | OR1L1:NM_001005236:exon1:c.A283G:p.S95G |
| **2** | 0.776604 | 1.90E-211 | RP1L1:NM_178857:exon2:c.C335G:p.T112S |
| **3** | 0.754697 | 2.23E-198 | SIGLEC9:NM_001198558:exon1:c.A391C:p.K131QSIGLEC9:NM_014441:exon1:c.A391C:p.K131Q |
| **4** | 0.754697 | 2.23E-198 | SIGLEC9:NM_001198558:exon5:c.T1046C:p.V349ASIGLEC9:NM_014441:exon5:c.T1046C:p.V349A |
| **5** | 0.712313 | 4.11E-176 | NRK:NM_198465:exon19:c.C2978A:p.A993E |
| **6** | 0.712268 | 4.32E-176 | HSD3B1:NM_000862:exon3:c.A235G:p.I79V |
| **7** | 0.705472 | 7.89E-173 | DUOX1:NM_175940:exon22:c.T2885C:p.I962TDUOX1:NM_017434:exon23:c.T2885C:p.I962T |
| **8** | 0.698271 | 1.87E-169 | SLC39A4:NM_130849:exon6:c.C1114G:p.L372VSLC39A4:NM_017767:exon5:c.C1039G:p.L347V |
| **9** | 0.695923 | 2.27E-168 | SIGLEC9:NM_001198558:exon4:c.C947A:p.A316DSIGLEC9:NM_014441:exon4:c.C947A:p.A316D |
| **10** | 0.658847 | 2.70E-152 | EPPK1:NM_031308:exon2:c.G1151A:p.G384E |

dimensions. Moreover, this result may give us some hints about the genes whose functions are still unclear and we may consider the genes in the list as biomarkers for PD diagnosis and detection and even prognosis in the near future.

After MCA, we are able to embed every sample in a 2-D plane constructed by the first and second dimensions to visualize the distance between every sample. We will show this in the clustering section.

## Unsupervised Clustering

In practice, outliers always exist and will impede and distort the clustering result. It is also the case in our experiments. Exome-seq data are composed of 645 samples, while around 20 samples may be deemed as outliers, although they may not be and it is just due to limited samples. In our analysis, we choose to regard them as outliers to prevent their influence on other clusters which are with greater confidence.

There are 4 kinds of cohorts composing all the 645 samples, i.e. Parkinson's Disease (PD), Health Control (HC), Scans without Evidence for Dopaminergic Deficit (SWEDD), Unknown (PPMI doesn't provide their information). After comparison with various result in our experiments, we think that there are five parts of samples: one part for HC, three main subtypes of PD and one part for outliers, as illustrated below. We will use Birch method to show the result.

We first set the cluster number to around 18 to tell which samples are outliers. In this case, all the other four parts are grouped together as one or two clusters and outliers are assigned to different clusters, usually one alone as one cluster. Then, we set cluster number to 30. Now, the leftmost part is divided into several clusters, each with ample samples.

In the following figures, samples from same cohorts are shown with the same shape, while samples clustered within the same group are shown with the same color.

When cluster number equals 18:

We can see that over 600 samples in the blue rectangle are clustered together.

Next, when cluster number equals 30 (show the blue rectangle area):

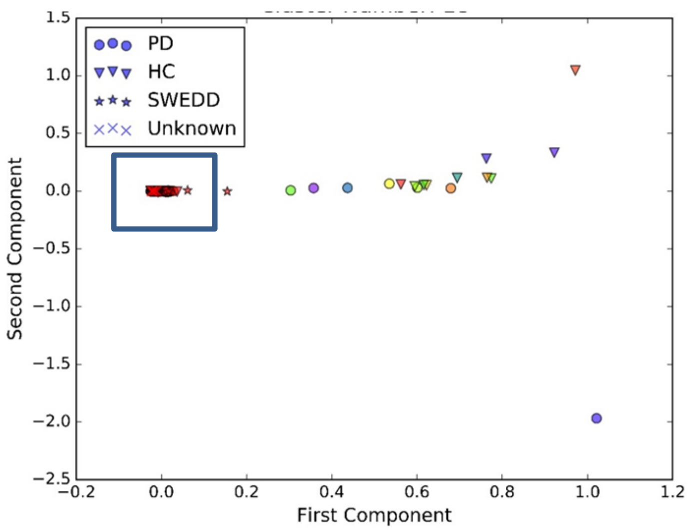

**Figure 2. Exome Samples Clustering in 18 Clusters.**

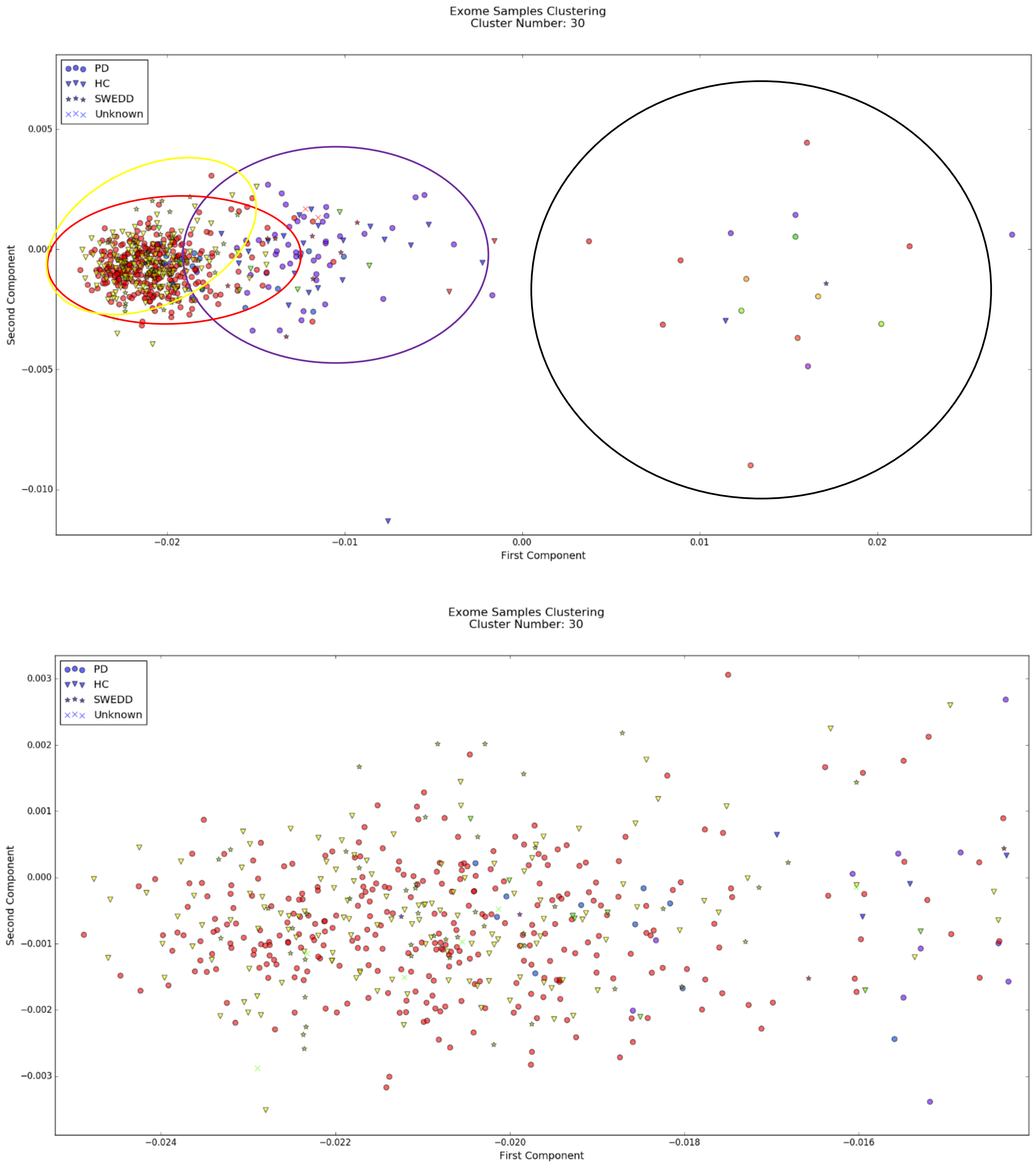

**Figure 3. Exome Samples Clustering in 30 Clusters.** Figure above shows a zoomed in view on samples in blue rectangle in Figure 2. Figure below shows a zoomed view on samples in red clusters.

We can find that most HC and SWEDD are clustered in yellow in the left most. This on one side provides a novel and very accurate way to distinguish PD patients with SWEDD patients, and on the other side, although PD red cluster and yellow cluster are overlapped, they are still separated clearly, which proves that our clustering method are very reliable!

Besides, if we observe three PD subtype clusters carefully, we may discover that these three subtypes may take Gaussian distribution with different mean and variance in the embedded space. Thus Gaussian Mixture Model method may also be applied to this problem.

Given the differences between subtypes, we are able to predict whether a new patient has PD and which subtype he belongs to if his exome-seq data or even just risk gene mutations profile is available and thus, may be able to provide him with more precise prognosis and personalized treatments and medications.

We also set cluster number as 50 to repeat the Birch. The result shows that the red subtype cluster above is divided into two subtypes. It shows us the hierarchical structure nature of subtypes of PD. But since 645 samples are in fact limited, this result may take the risk of over-clustering. So we will stop at 30 and discard this 50-cluster result until enough samples are acquired and enough proofs are got.

## Conclusion

Germline mutation landscape provides valuable insights into subtyping PD patients. With an increasing size of cohort and incorporation of other diagnostic measurements, careful examination into each subtype will facilitate the research of PD and health care of PD patients.

## References

1. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
2. Verhaak, R.G., et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.
3. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers.* Nat Genet, 2013. **45**(10): p. 1127-1133.
4. Girvan, M. and M.E. Newman, *Community structure in social and biological networks.* Proc Natl Acad Sci U S A, 2002. **99**(12): p. 7821-6.
5. Obenchain, V., et al., *VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants.* Bioinformatics, 2014. **30**(14): p. 2076-8.
6. *B. Le Roux and H. Rouanet, Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis, Dordrecht, Kluwer, 2004, pp. xi + 475.* Journal of Classification, 2008. **25**(1): p. 137-141.
7. Lê, S., J. Josse, and F. Husson, *FactoMineR: An R Package for Multivariate Analysis.* 2008, 2008. **25**(1): p. 18.
8. Park, H.-S. and C.-H. Jun, *A simple and fast algorithm for K-medoids clustering.* Expert Systems with Applications, 2009. **36**(2, Part 2): p. 3336-3341.
9. Rousseeuw, P.J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 1987. **20**: p. 53-65.

10. Pedregosa, F.a.V., Ga{\"e}l and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and Vanderplas, Jake and Passos, Alexandre and Cournapeau, David and Brucher, Matthieu and Perrot, Matthieu and Duchesnay, {\'E}douard, *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011.

11. Monti, S., et al., *Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data.* Machine Learning, 2003. **52**(1): p. 91-118.

12. Frey, B.J. and D. Dueck, *Clustering by Passing Messages Between Data Points.* Science, 2007. **315**(5814): p. 972-976.

13. Newman, M.E., *Fast algorithm for detecting community structure in networks.* Phys Rev E Stat Nonlin Soft Matter Phys, 2004. **69**(6 Pt 2): p. 066133.