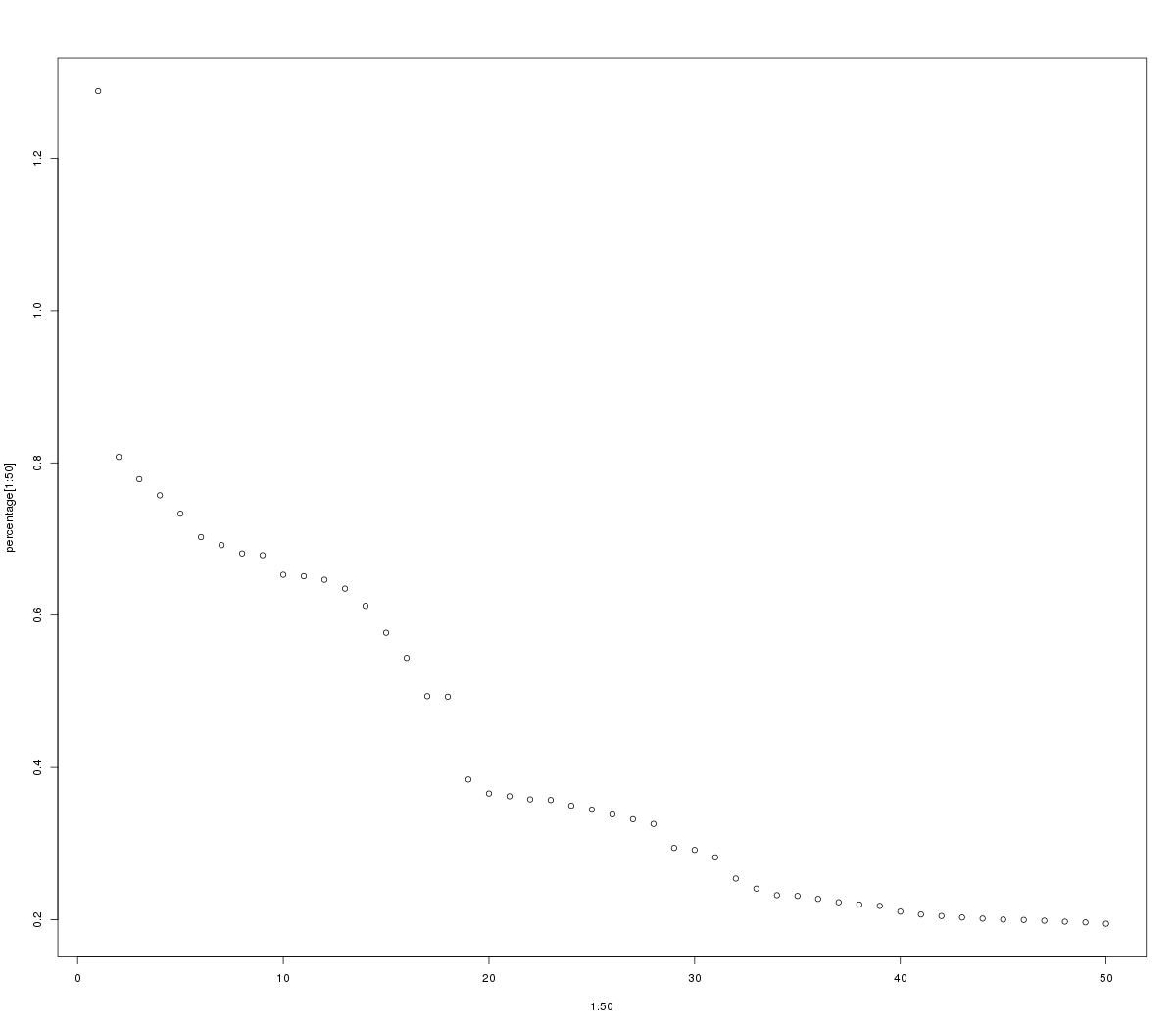
**Dimension Reduction**

Multiple Correspondence Analysis (MCA) [9] 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 [10] 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.

**Dimension Reduction – Result**

The percentages of variance of the first 50 dimensions in the new matrix (MCA processed matrix) are plotted below:



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:

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

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 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**

Unsupervised clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). [11] In our case, clustering the first choice to find 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.

* **Spectral Clustering**

Spectral Clustering does a low-dimension embedding of the affinity matrix between samples, followed by a KMeans in the low dimensional space. It works well for a small number of clusters but is not advised when using many clusters. [12] We use a python implementation in Scikit-learn package [13].

* **Birch**

The Birch builds a tree called the Characteristic Feature Tree (CFT) for the given data. The data is essentially lossy compressed to a set of Characteristic Feature nodes (CF Nodes). [12]

* **DBSCAN**

The [DBSCAN](http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html#sklearn.cluster.DBSCAN) algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k-means which assumes that clusters are convex shaped. [12]

* **Hierarchical Clustering**

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. [12]

* **Partitioning Around Medoids (PAM)**

PAM is a way of implementing k-medoids clustering, a more robust version of k-means clustering [2]. To get the best clustering scheme, we use silhouette score [3] 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.

* **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) [4]. 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 [5]. 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 [6]. It has been reported that the heterogeneity information with the data can be reflected by the network [7]. Based on the constructed bipartite network, we measure modularity using method developed by Newman [8]. This is an especially powerful method compared to the above mentioned ones, because the cluster specific SNPs are also highlighted.

**Unsupervised Clustering -- Result**

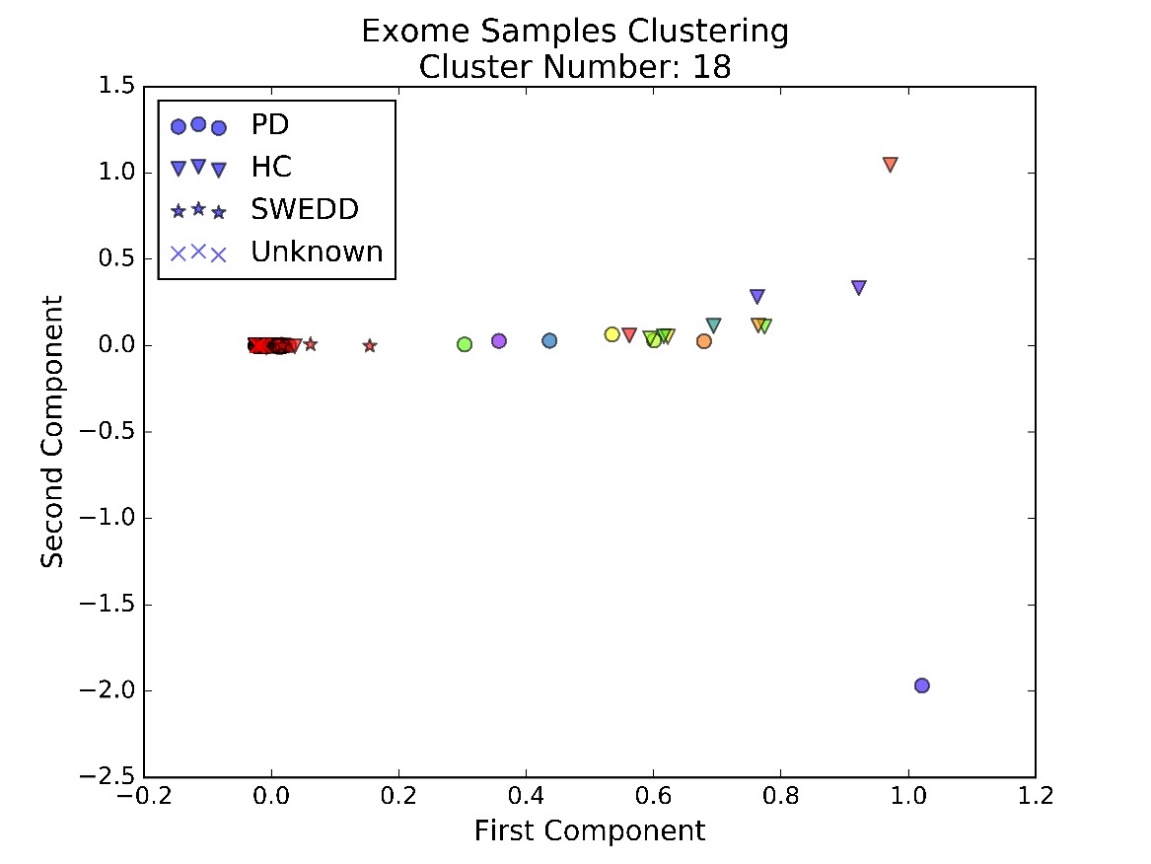
In practice, outliers always exist and will impede and distort the clustering result. It is also the case in our experiments. Exome-seq data are composed of 645 samples, while around 20 samples may be deemed as outliers, although they may not be and it is just due to limited samples. In our analysis,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

There are 4 kinds of cohorts composing all the 645 samples, i.e. Parkinson’s Disease (PD), Health Control (HC), 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 20 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.

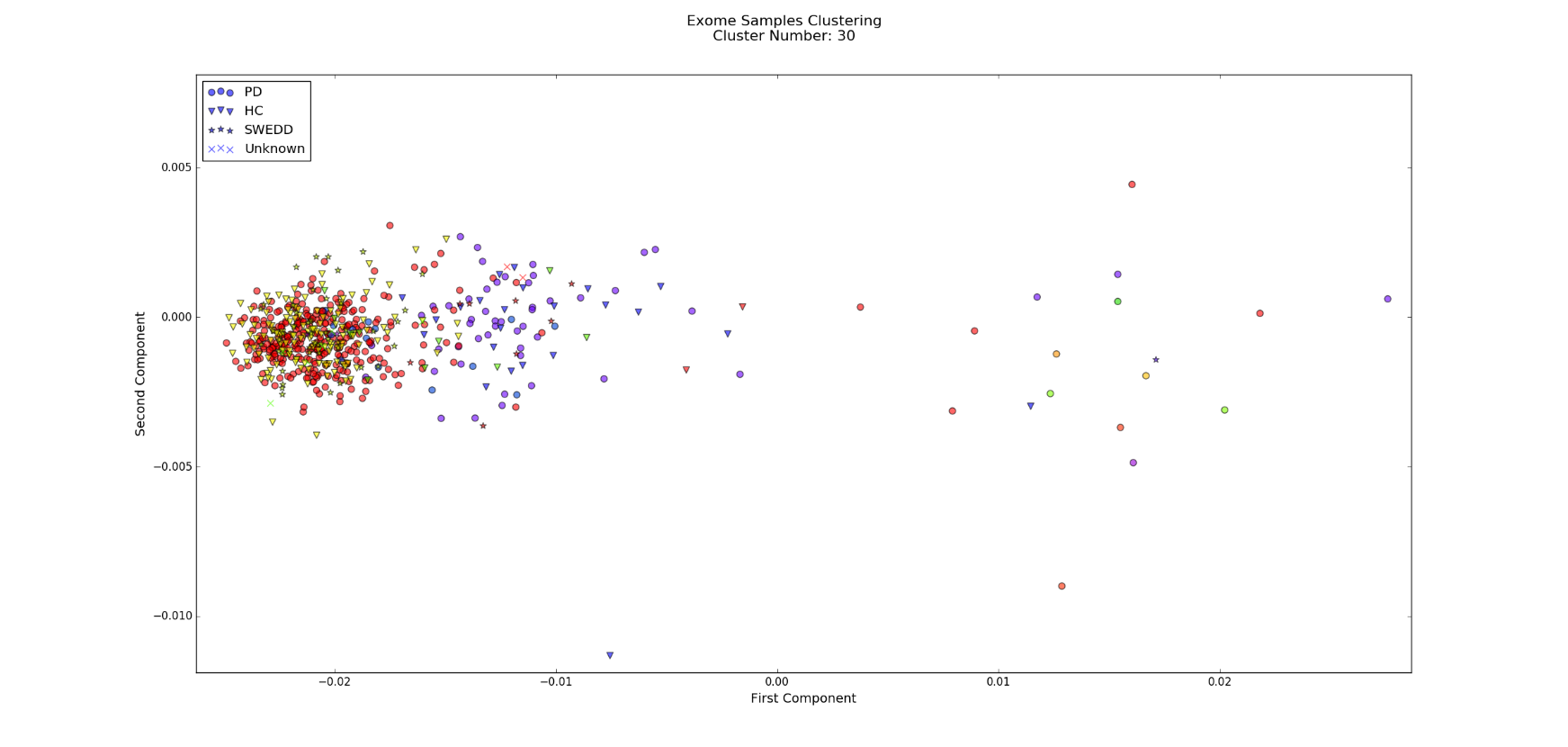
In the following figures, samples from same cohorts are shown with the same shape, while samples clustered within the same group are shown with the same color.

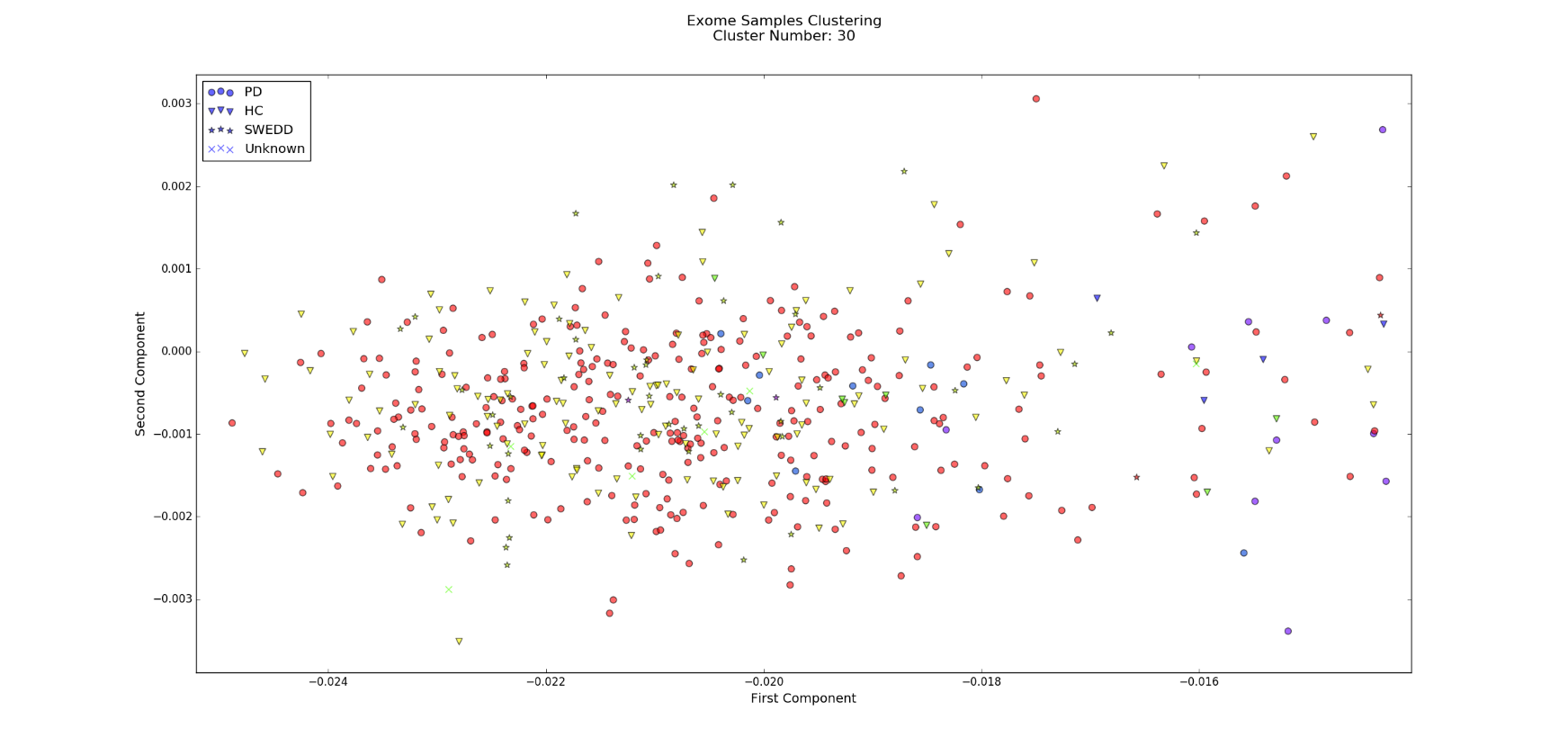
When cluster number equals 20:



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):





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 other side, although PD red cluster and yellow cluster are overlapped, they are still seperated 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 to 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.

**References**

1. <http://bioconductor.org/packages/release/bioc/manuals/VariantAnnotation/man/VariantAnnotation.pdf>
2. Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2), 3336-3341.
3. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.
4. Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning, 52(1-2), 91-118.
5. Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. science, 315(5814), 972-976.
6. Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nature genetics, 45(10), 1127-1133.
7. Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826.
8. Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. Physical review E, 69(6), 066133.
9. Le Roux; B. and H. Rouanet (2004). Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis. Dordrecht. Kluwer: p.180.
10. Lê, S., Josse, J. & Husson, F. (2008). [FactoMineR: An R Package for Multivariate Analysis](http://factominer.free.fr/docs/article_FactoMineR.pdf" \t "_blank).*Journal of Statistical Software*. **25(1)**. pp. 1-18.
11. <https://en.wikipedia.org/wiki/Cluster_analysis>
12. <http://scikit-learn.org/stable/modules/clustering.html#clustering>
13. [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.