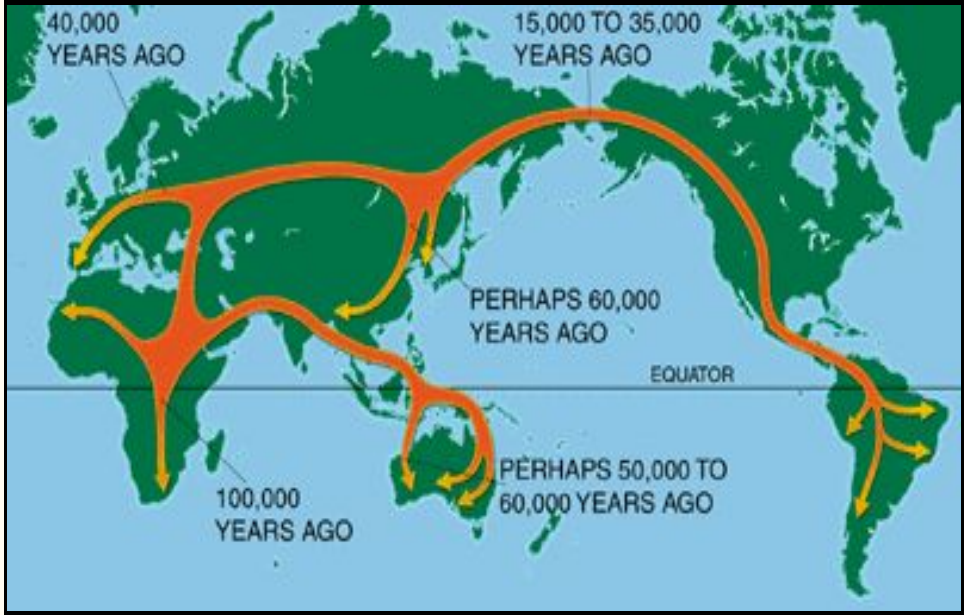
# Odd 1’s Out : How different are you ???

Group 17

# What is “Out of Africa” theory (OOA) ?

# “Out of Africa” theory

Also known as
**Single-origin hypothesis (RSOH),**
Proposes that modern humans evolved in East Africa, and then began to disperse throughout the world roughly 70,000 to 135,000 years ago.
**So basically we all are fractional African for sure !!!!**



# Did you know ?

Researchers compared 650,000 genetic markers in nearly a thousand individuals from 51 populations around the globe—an unprecedented level of detail for a human genetic study.

***"You get less and less variation the further you go from Africa,"***

- Marcus Feldman

(an evolutionary biologist at Stanford University)

# Then Why is it that we look so different ??

# Here’s Why ...

- As each small group of people broke away to found a new region, it took only a sample of the parent population's genetic diversity.

- People evolved into slightly different forms in the different environments as conditions in them were different and **chance mutations** proved to be beneficial.

For example, A darker skin protects you a little more from the damaging effects of sunlight.

# How is it related to Data Science !

- Doing Genetic Research on many level is difficult and very expensive process.
- Too vast data and too much of raw data at disposal.
- But definitely has a pattern to it !
- Different species could have same sequencing on genetic level(genotypes) for the same kind of traits (phenotypes).

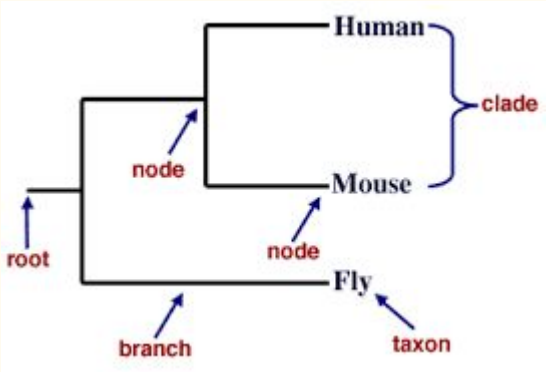
So Here  we are trying to **find the similarity between different individuals using their sequencing data and find the connection in their traits** through their genetics

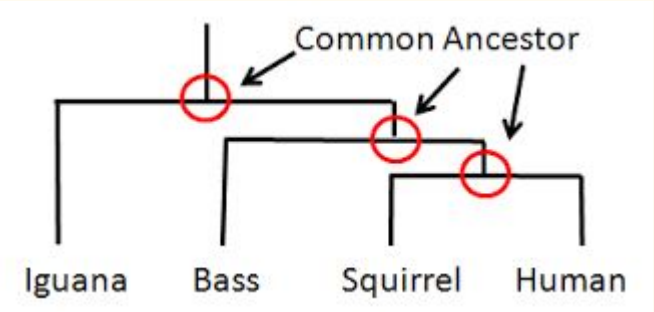
# Galaxy

- Galaxy is an open source, web-based platform for data intensive biomedical research.
- It enables user to perform computational analyses through the web.
- A user interacts with Galaxy through the web by uploading and analyzing the data.
- It provides many tool which are helpful for statistical analyses,mapping,sorting, Graphical analyses etc...

# Phylogenetic Tree

- A phylogenetic tree is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities,their phylogeny is based upon similarities and differences in their physical or genetic characteristics.
- Using the tool that generates phylogenetic tree we can determine a kind of "genetic distance" between each pair of individuals.



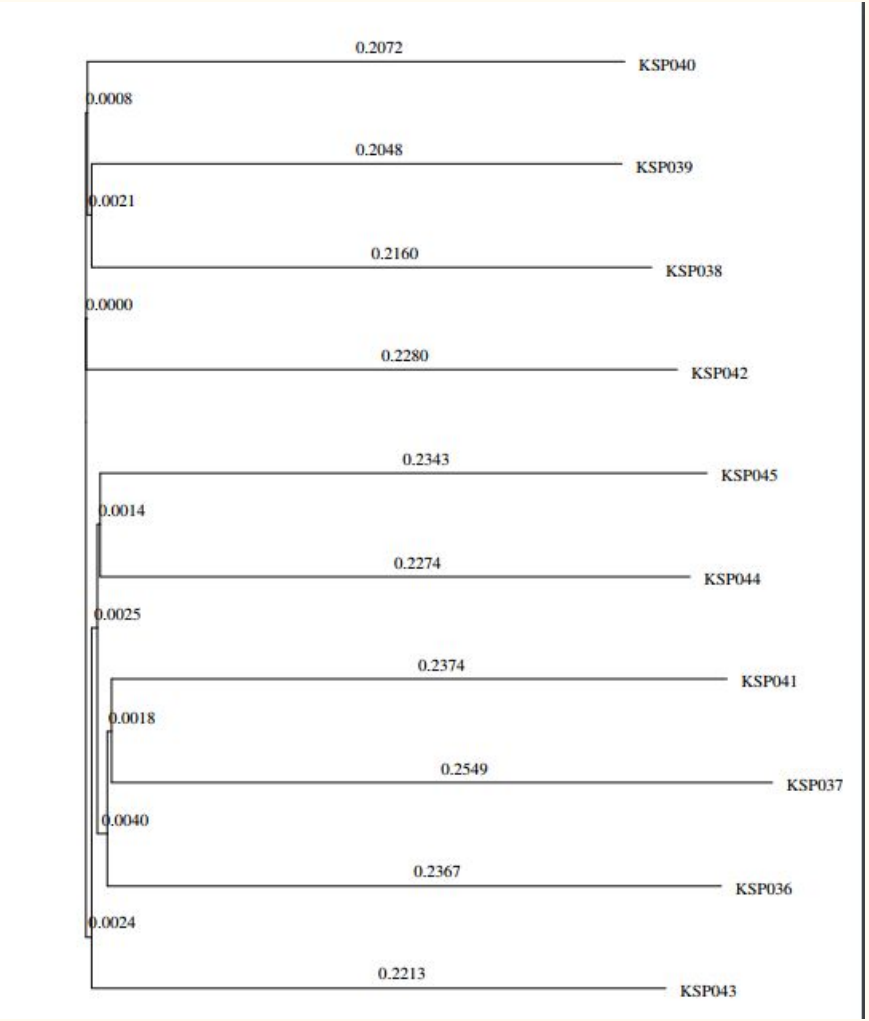



# Phylogenetic Tree (cont..)

**Input** : 485 individuals data is collect and converted to **gd_snp** format as required by the phylogenetic tree tool in galaxy.

**Output** : Visualization of phylogenetic tree for all the individuals,File containing the coordinates of individuals in the phylogenetic tree, Lower triangular distance matrix(mega matrix)  and a complete dissimilarity square matrix are obtained as an output of phylogenetic tree tool.

Phylogenetic tree for 10 individuals is shown in the following slide.


0.2072
KSP040
0.0008
0.2048
KSP039
0.0021
0.2160
KSP038
0.0000
0.2280
KSP042
0.2343
KSP045
0.0014
0.2274
KSP044
0.0025
0.2374
KSP041
0.0018
0.2549
KSP037
0.0040
0.2367
KSP036
0.0024
0.2213
KSP043

# Clustering Algorithms

Few clustering Algorithms that takes distance matrix as input parameter are considered, which includes

- Agglomerative clustering
- DBSCAN
- HDBSCAN
- MDS (Multidimensional Scaling)

# Silhouette Coefficient

- Silhouette refers to a method of interpretation and validation of consistency within clusters of data.
- The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
- The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

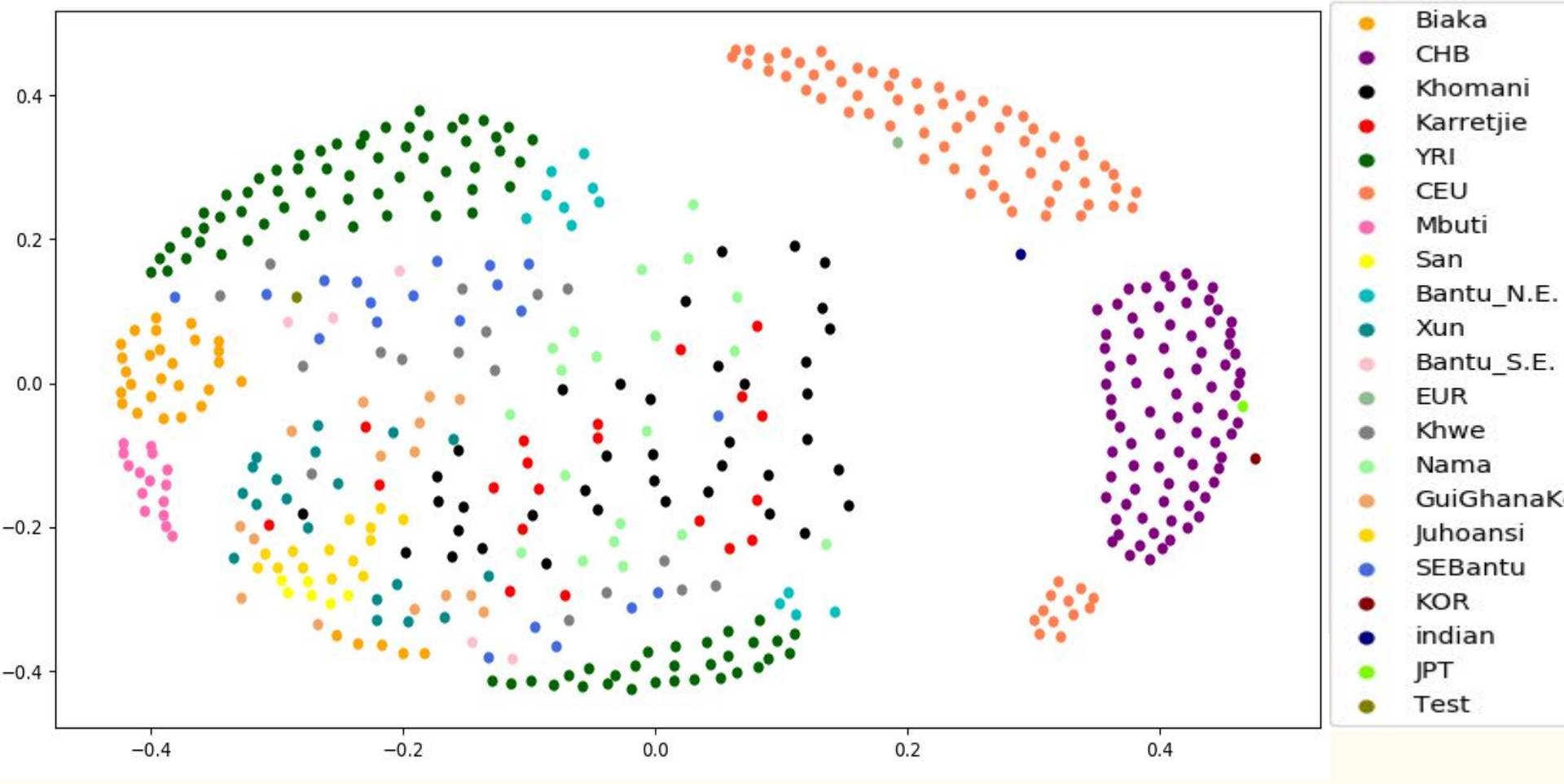
# Evaluation and best fit of clusters

| Method | N = 2 | N = 3 | N = 4 | N = 5 |
|---|---|---|---|---|
| Agglomerative | 0.1854427 | 0.180096 | 0.097132 | 0.0861307 |
| DBSCAN | | | 0.1222 | |
| HDBSCAN | | 0.180096 | | |

# MDS (Multidimensional Scaling)

- It is a means of visualising the similarity of individuals of a dataset.
- Aims to place each object in N- dimensional space such that between object distances are preserved.
- It is non- linear dimensionality reduction technique.

MDS is applied on our data and object locations are optimised to a 2-D scatterplot using the precomputed distance matrix. Following slide shows the plot of the transformed coordinates.


0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
Biaka
CHB
Khomani
Karretjie
YRI
CEU
Mbuti
San
Bantu_N.E.
Xun
Bantu_S.E.
EUR
Khwe
Nama
GuiGhanaK
Juhoansi
SEBantu
KOR
indian
JPT
Test

Utah (CEU)

Han Chinese

Biaki Tribe

Khomani / San

Yoruba

Karretjie

# Types of Individuals

- **CEU** - Utah residents (CEPH) with Northern and Western European ancestry
- **CHB** - Han Chinese in Beijing, China (East Asia)
- **YRI** - Yoruba in Ibadan, Nigeria (African)
- **JPT** - Japanese in Tokyo
- **Indian**

**San** - Southern Africa(Botswana,Namibia,Angola,Zambia,Zimbabwe,South Africa)
**Biaka** - Southern region of Central African Republic (Tropical rain forests)
**Mbuti** - Congo region of Africa
**Bantu_S.E.** - South Africa in sub-saharian Africa
**Bantu_N.E.** - Kenya in subsaharian Africa
**SEBantu** - Southeastern Africa Bantu speakers

# Types of Individuals

**Nama** - Ethnic group of South Africa, Botswana and Namibia
**Khomani, Xun and Khwe** are the indigenous groups of Southern Africa.
Xun and Khwe (Immigrants from Angolia via Namibia)
Khomani - Southern Kalahari desert
**Karretjie** - Found in The great Karoo in South Africa
**Guighanakgal** - Botswana
**Juhoansi** - Namibia

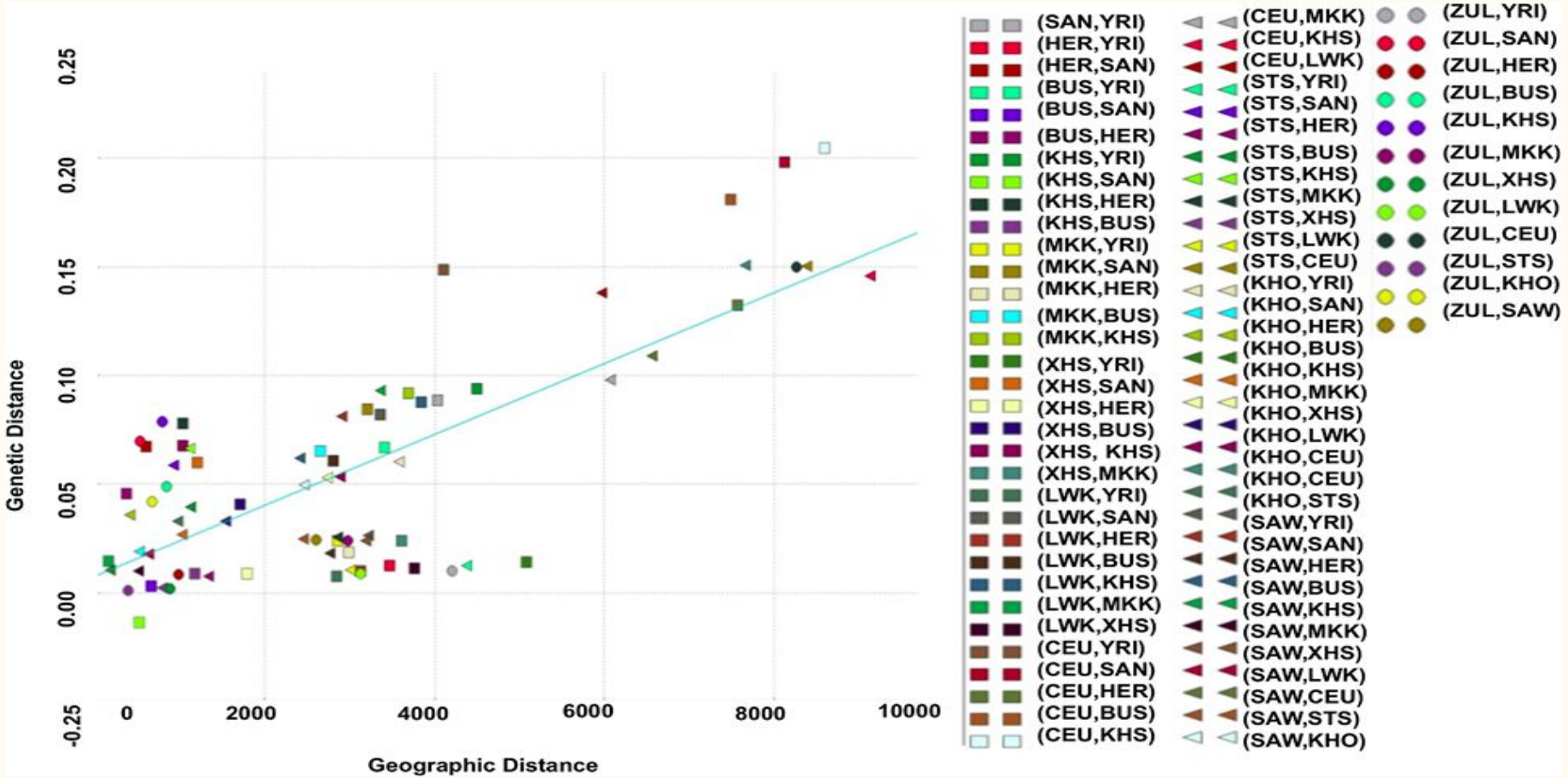
# Analysis of clustering results

- The dataset that we considered consists of individuals from the above mentioned regions
- In general the similarity of genes depends on the geometric distance of the regions in which people are living.
- The genes of two individuals vary to a large extent if (for example) one individual belongs to India and other individual belong to SouthAfrica (i.e) both of them must belong to two different clusters.
- The genes of two individuals are more similar if(for example) both the individuals belong to different regions among SouthAfrica(i.e) both of them must belong to the same cluster.

# Analysis of clustering results

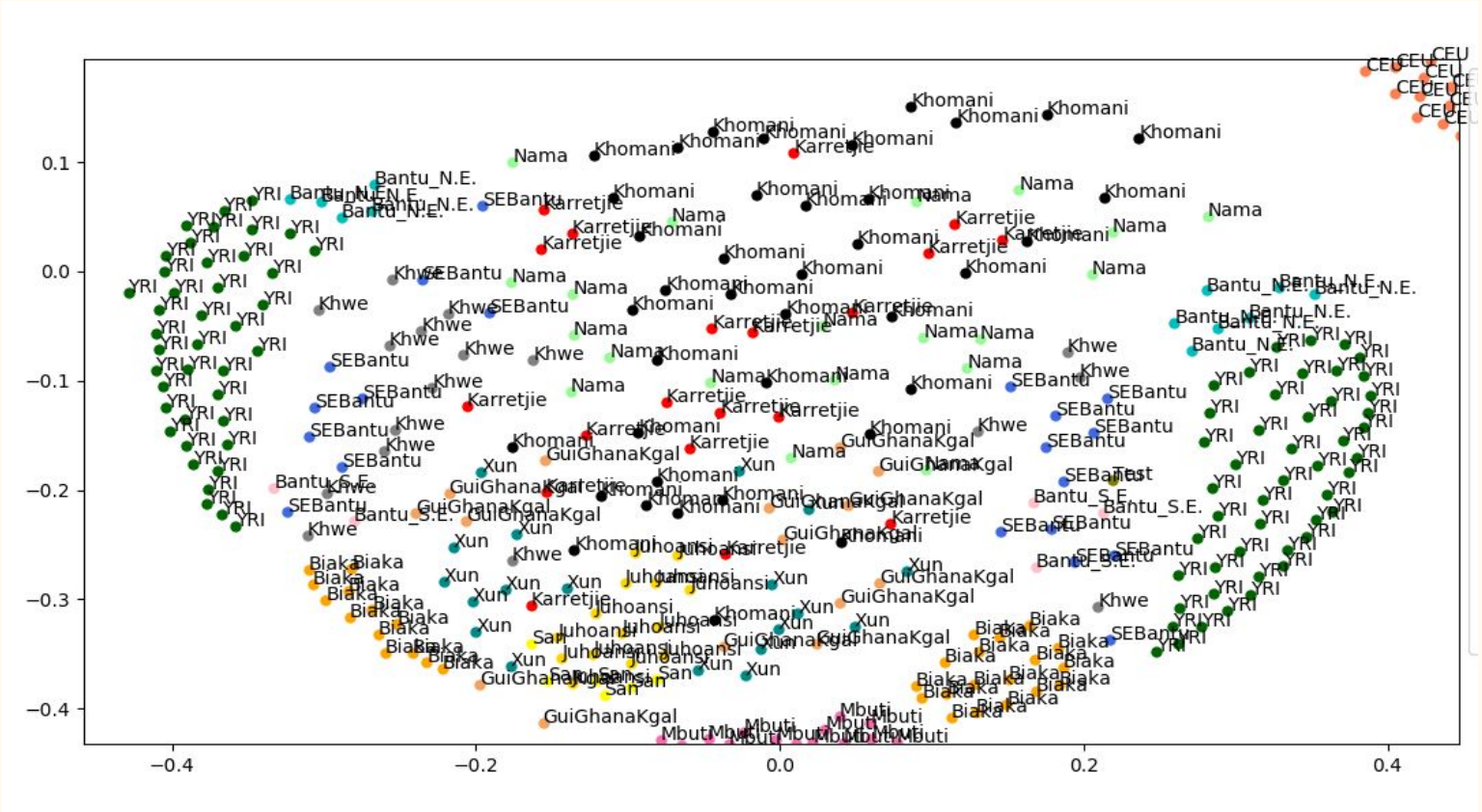
- Also the CHB and CEU cluster is so tightly packed showing the close resemblance among each other while in the African cluster it is very well spread.
- Could be Because that the individual picked from African cluster could be geographically spread out.
- In a study published in March 2011, Brenna Henn and colleagues found that the **Khomani San were the most genetically diverse of any living humans** studied. Due to which that Black point cluster is not tightly bounded.

# Variation of genetic Vs geometric distances



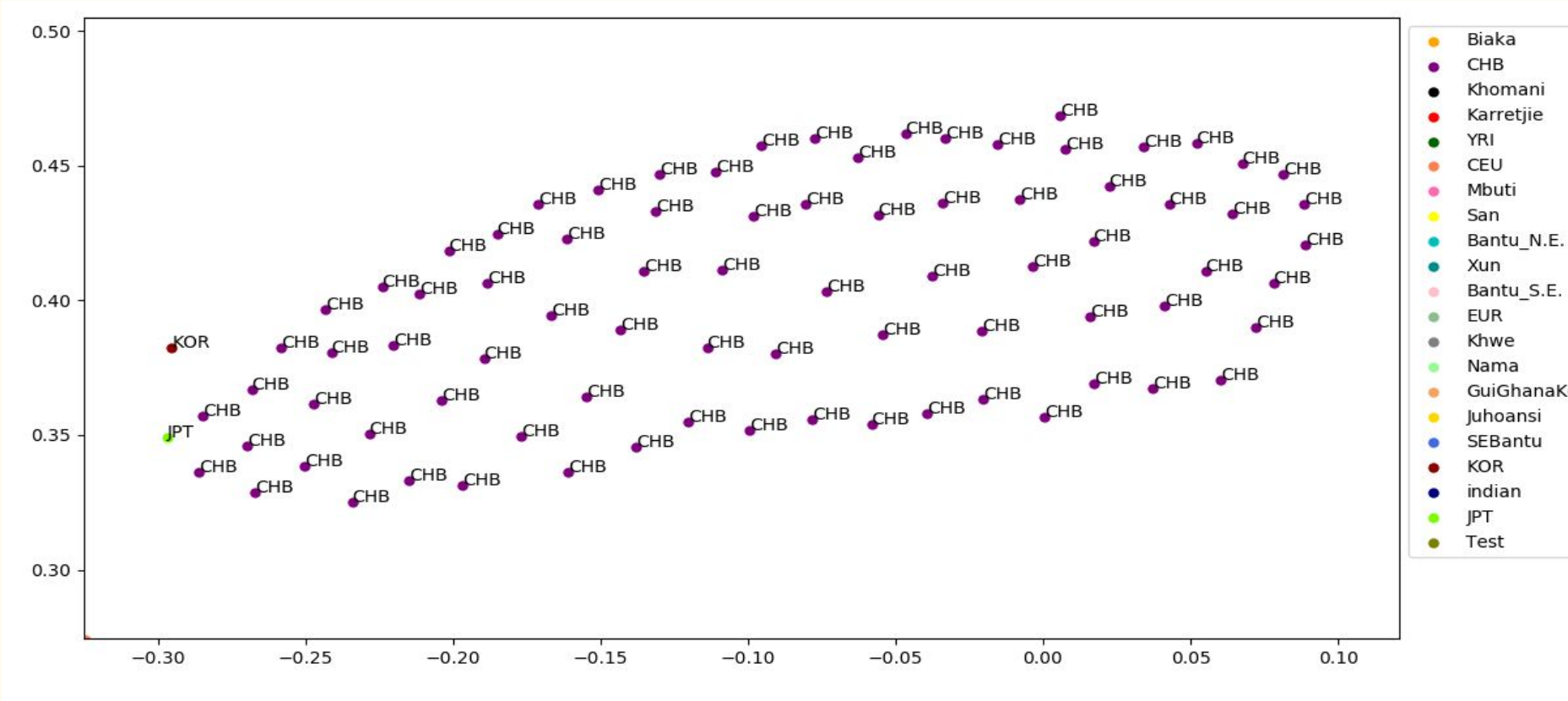
# Analysis of clustering results

Cluster - 1 consists of individuals from different parts of Africa



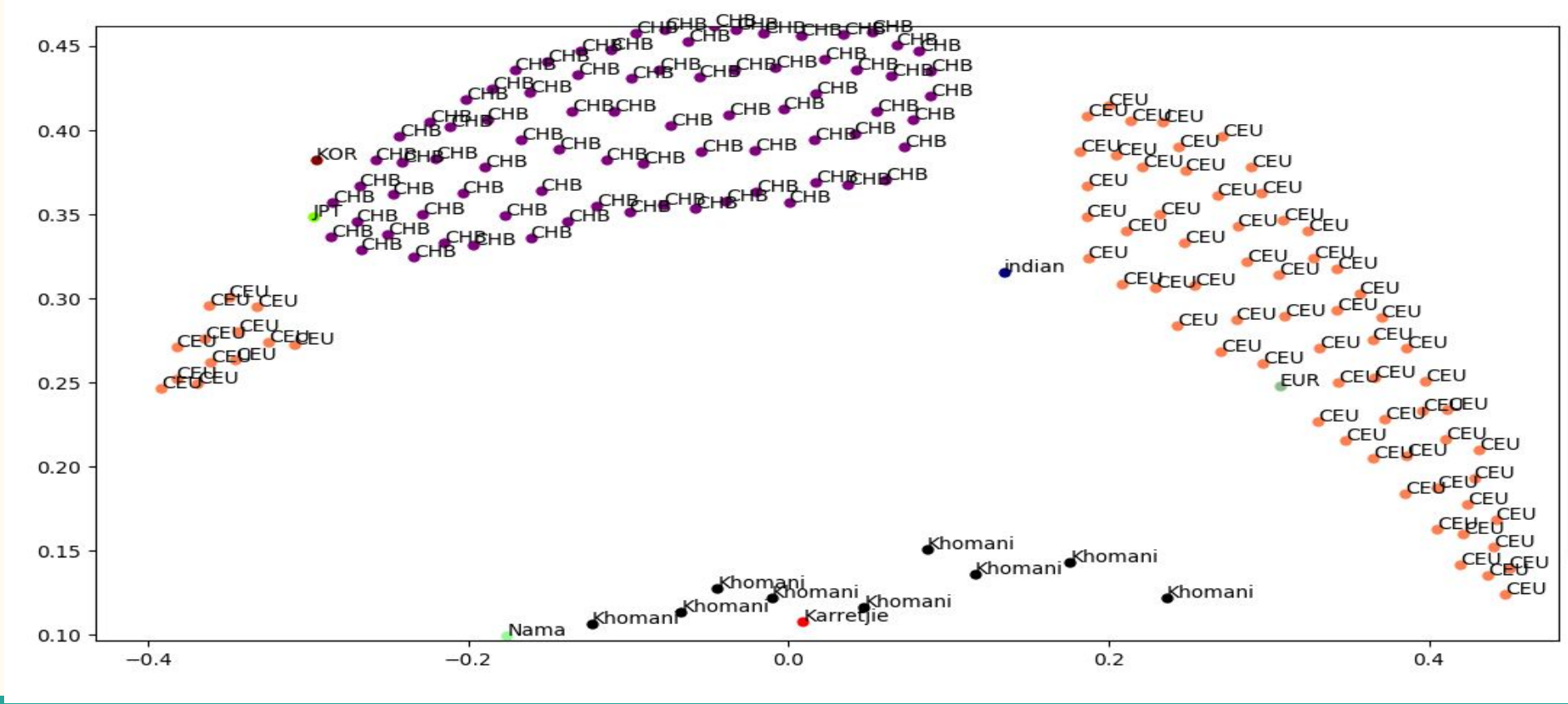
# Analysis of clustering results

Cluster - 2 consists of individuals from Han Chinese in Beijing (China).



# Analysis of clustering results

Cluster - 3 consists of individuals from CEU (Northern and European Ancestry)



# Outliers (Clustering based Approach)

- Clustering based approach is used in order to find the outliers among the given dataset of individual genomes.
  - **DBSCAN** algorithm is used and the cluster containing least number of data points is considered as an outlier cluster.

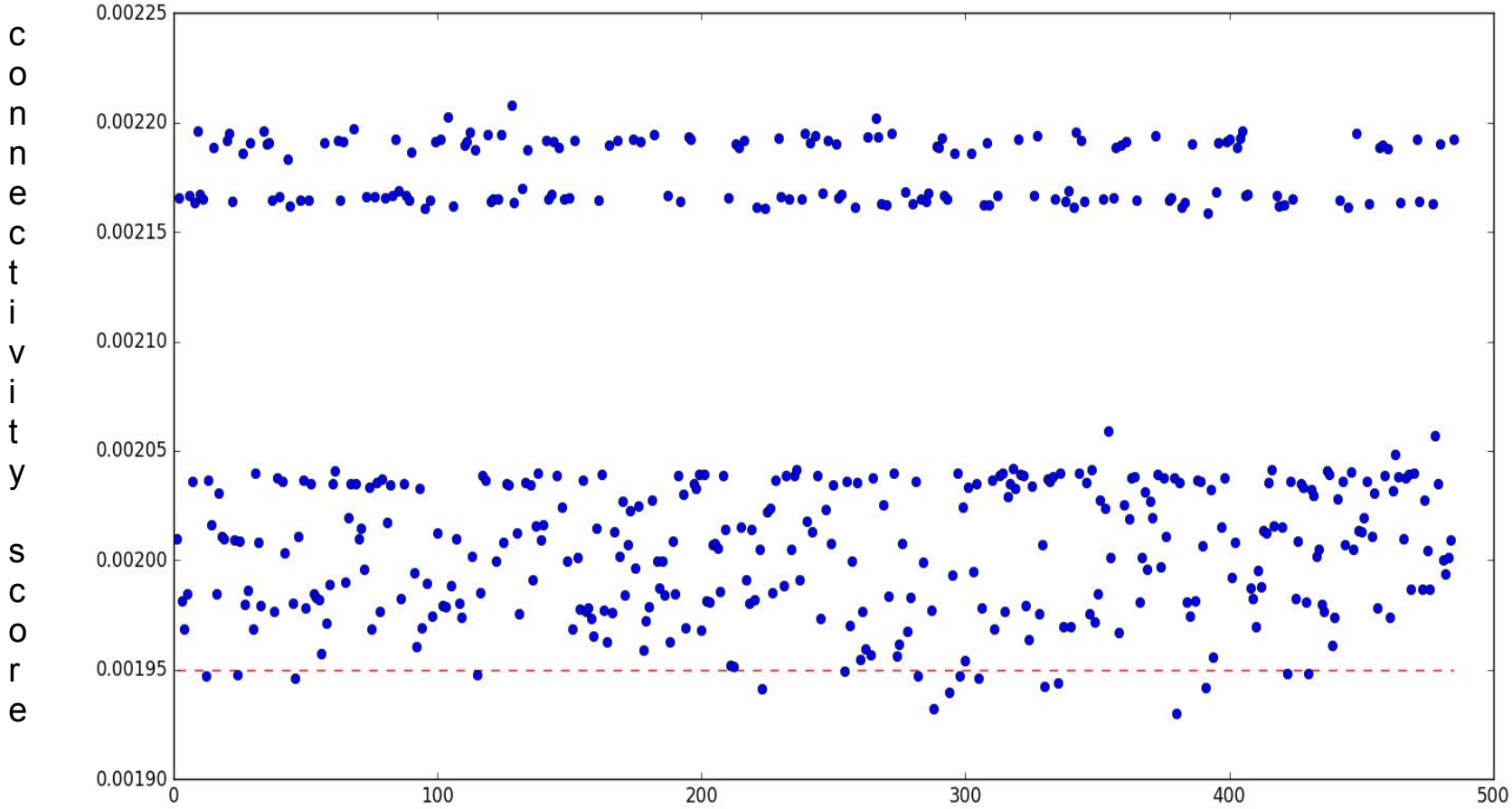
# Output analysis

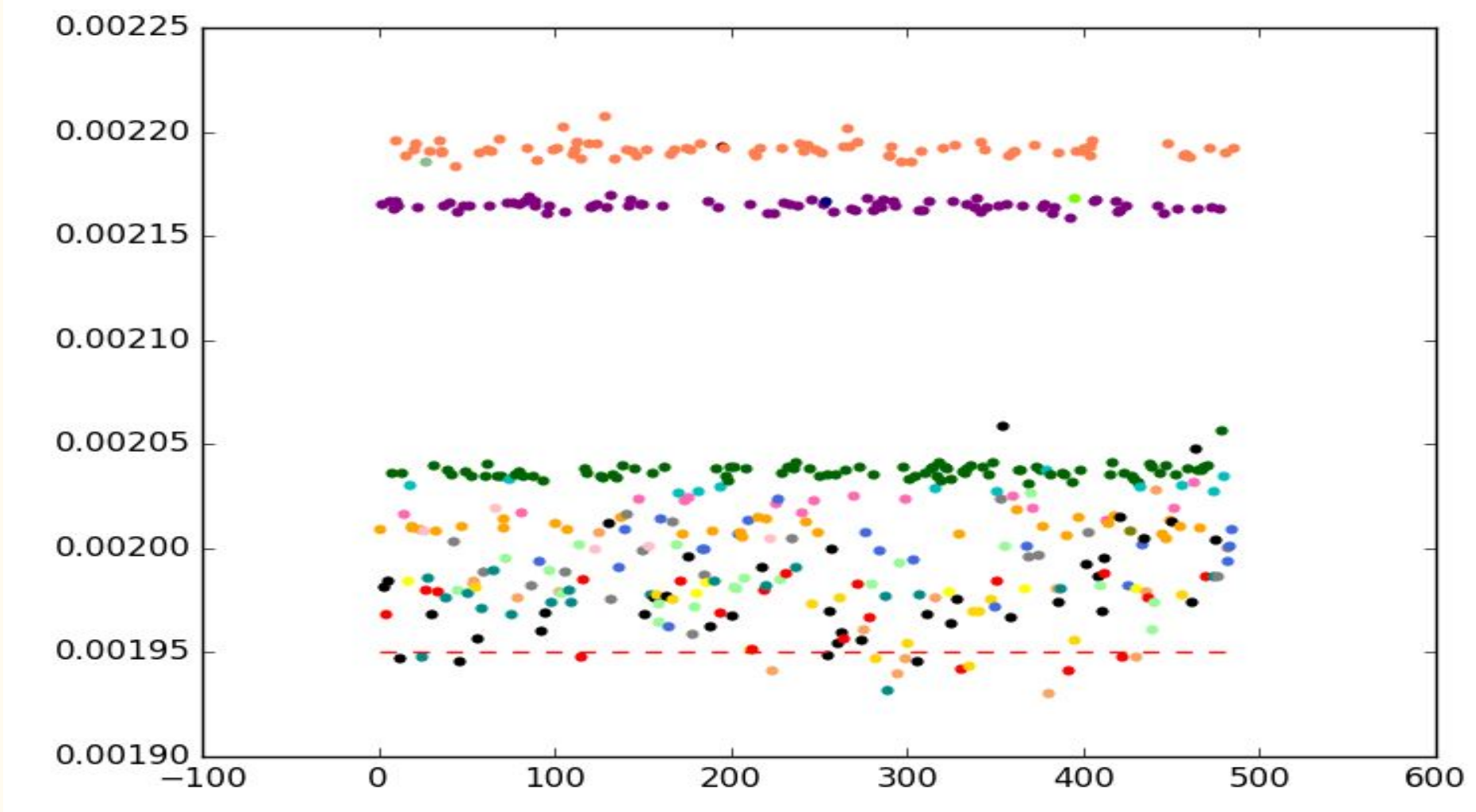
- DBSCAN results in four different clusters among which one cluster is having only 2 data points.
  - So both of these points can be treated as outliers.

# Outliers(Graph based approach)

- A graph based outlier detection framework using random walk is used for detection of outliers.
- Each datapoint is given a Connectivity score based on which the ranking is assigned to the data points.
- If the Connectivity score is high then ranking of that point is high.
- A threshold value is given so that the data points with the rank value less than the given threshold are considered as outliers.

Reference :
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.419.4844&rep=rep1&type=pdf


connectivity score
0.00225
0.00220
0.00215
0.00210
0.00205
0.00200
0.00195
0.00190
0
100
200
300
400
500


connectivity score
0.00225
0.00220
0.00215
0.00210
0.00205
0.00200
0.00195
0.00190
−100
0
100
200
300
400
500
600

# References

https://usegalaxy.org/

http://www.internationalgenome.org/data-portal/population

http://glaros.dtc.umn.edu/gkhome/views/cluto

http://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html

# THANK YOU

GROUP MEMBERS :
Y.ANUSHA ES14BTECH11024
B.SRAVANI ES14BTECH11019
B.SHRUTI CS14BTECH11007