Clustering of the human genome

Menghor THUO, Matthieu DESPREZ, Skander GORGI

Description of the data

The objective of this project is to apply clustering algorithms to genotype data derived from a large-scale genomic study, specifically focusing on a dataset from the 1000 Genomes Project. By clustering genotypes, we aim to:

1. Identify subpopulations within the dataset.
2. Examine the relationship between genetic variation and phenotypic traits.

The genotype array has the following characteristics:

- **Shape**: (1103547, 2504, 2)
    - **1103547**: The number of variants (rows), which corresponds to the number of different genetic variants the VCF file.
    - **2504**: The number of samples (columns), indicating how many individual samples are represented in the dataset.
    - **2**: The number of alleles (for diploid organisms, each individual has two alleles at each variant site).

There is a list of fields to be extracted from a VCF (Variant Call Format) file when using the allel library in Python. Each field corresponds to a specific type of information stored in the VCF file format. Here's a breakdown of each field in the list:

1 **variants/CHROM**:

- **Description**: This field contains the chromosome identifier where the variant is located.
- **Example**: It has values like chr1, chr2, chrX, etc. This indicates which chromosome the variant appears on.

2 **variants/POS**:

- **Description**: This field represents the position of the variant on the specified chromosome.
- **Example**: It is a numeric value (e.g., 123456) that indicates the exact base pair position on the chromosome where the variant occurs.

3 **variants/ID**:

- **Description**: This field contains an identifier for the variant, if one exists.
- **Example**: Commonly, this will be an rsID from dbSNP (e.g., rs123456). If there is no identifier for a particular variant, this field might be represented as a dot (.).

4 **variants/REF**:

- **Description**: This field shows the reference allele at the specified position.
- **Example**: It will contain the nucleotide(s) that are present in the reference genome (e.g., A, C, G, T).

5 **variants/ALT**:

- **Description**: This field lists the alternate alleles present at the position, which represent the variant(s) that differ from the reference.
- **Example**: This could be a single nucleotide like G (for a SNP) or multiple alleles in case of multi-allelic variants (e.g., G, T for a variant that can be either G or T).

6 **calldata/GT**:

- **Description**: This field provides genotype information for each sample at the variant position. It indicates the alleles present in the samples.
- **Example**: Genotype representations include:
    - 0/0: Homozygous reference (both alleles are the reference allele)
    - 0/1: Heterozygous (one reference and one alternate allele)
    - 1/1: Homozygous alternate (both alleles are the alternate allele)

The dataset was found on http://www.internationalgenome.org/ .