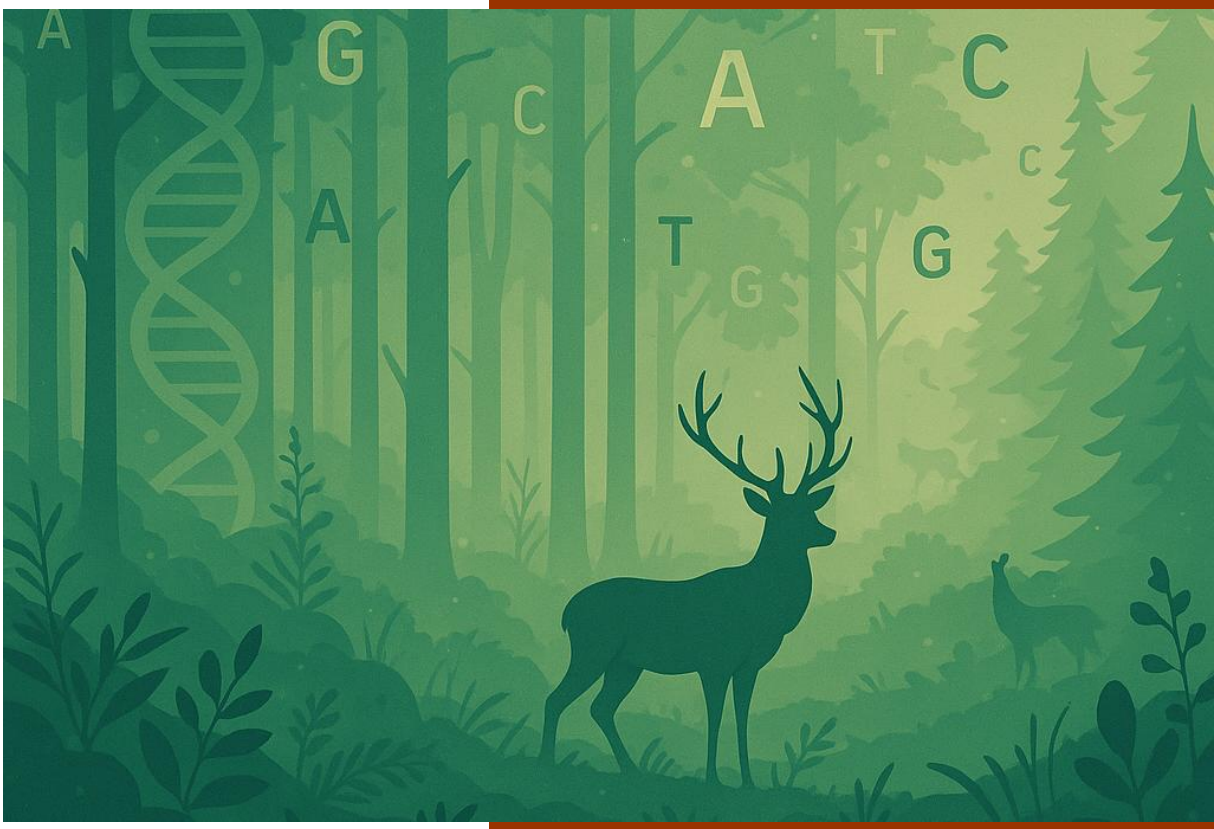
Clustering-Based Identification - Gene Variants in Deer Using Sequencing Data



Bioinformatics Course

Faculty of Electrical Engineering and Computing

Max Henrotin (max.henrotin@fer.hr)

Lynn Limbach (lynn.limbach@fer.hr)

Krešimir Križanović (kresimir.krizanovic@fer.hr)

8.6.2025

Inhalt

Introduction and Goal

In this project, clustering algorithms are applied to DNA sequencing reads from deer to identify different variants of a gene called the Major Histocompatibility Complex (MHC). The MHC plays a key role in the immune system by allowing the organism to recognize foreign molecules such as bacteria or viruses. Deer with a higher diversity of MHC gene variants are potentially able to recognize and respond to a broader range of pathogens.

To achieve this goal, sequencing data from 47 FASTQ files are provided. The file names include identifiers ranging from J1 to J41, suggesting that the dataset includes samples from 41 individual deer. However, files sharing the same numeric identifier but differing in other parts of their filenames (e.g., suffixes like *GK*, *S*, *L*) indicate that multiple technical or experimental variants may exist for the same deer. The exact nature of these differences is not specified.

Each file contains between 200 and 3,000 extractions of the MHC gene.

The objective of the program is to:

1. Eliminate corrupted or erroneous sequences.
2. Correct measurement errors in the gene extractions by averaging the ~1,000 sequences to identify their actual gene variants.
3. Detect and distinguish multiple gene variants present within a single deer, possibly due to inherited mutations preserved by natural selection.

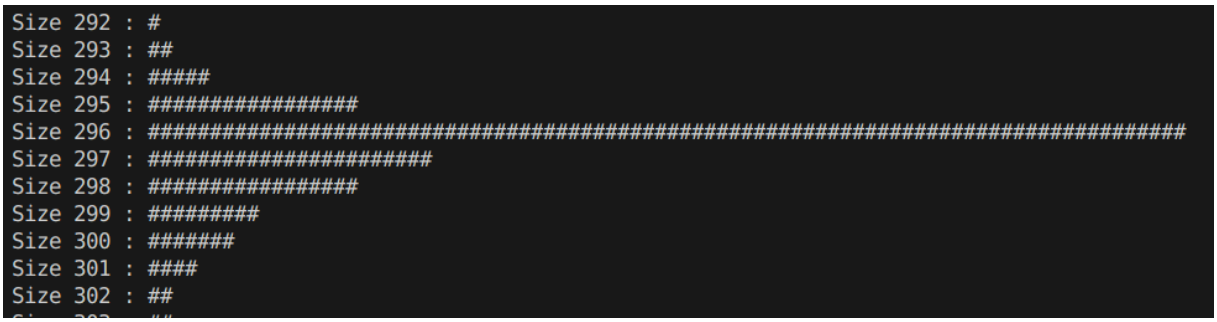
The program is designed to take a FASTQ file as input (corresponding to one deer) and output the *k* main variants of the gene for that deer. The value of *k* is provided by the user.

Design and Implementation

The first step was to isolate each sequence from the FASTQ file and store them in a vector:

vector<string> extract_sequences(const string& path)

To identify and eliminate corrupted sequences, we plotted a histogram showing the distribution of sequence lengths. This analysis revealed that most corrupted sequences are shorter than 200 base pairs, while the majority of valid sequences lands around 296 base pairs.



Based on this observation, we decided to filter and work exclusively with sequences of exact length 296 or, optionally, within a range of 290–305 base pairs. As later results showed, the exact length threshold had only a marginal effect on clustering outcomes.

From the beginning, we believed that the best approach would be to use the **k-means clustering** algorithm to group similar sequences into clusters, with each cluster representing a distinct variant of the gene. We implemented our main method that given a chosen k, could give the k most realistic gene variant it found from an array of sequences:

vector<string> k_centroid(vector<string> sequences, int k, int nbr_step_max)

A brief explanation of the k-means algorithm can be found here:

https://www.youtube.com/watch?v=R2e3Ls9H_fc

The key tasks to implement our k-mean clustering function were:

- **Finding a distance metric** to compare sequences and assign them to clusters.
- **Calculating centroids** to represent the consensus of each cluster.

Using SPOA for Alignment and Consensus

We initially explored several basic methods for generating centroids but found them ineffective. Based on project suggestions, we switched to using the **SPOA** library for multiple sequence alignment (MSA):

The documentation was limited, making it challenging to understand how to use all its features. However, two main SPOA functionalities turned out to be very helpful:

1. **MSA alignment**: By aligning all sequences, we could account for insertions and deletions, enabling us to:

a. Use distance metrics (like Hamming distance) that require equal-length sequences.

b. Obtain more accurate distance calculations by aligning homologous regions.

2. **Consensus generation**: We used SPOA's graph-based alignment to generate the centroid (consensus sequence) of a cluster:

   string consensus = graph.GenerateConsensus();

Distance Metrics

We initially hoped to use SPOA to directly compute distances between sequences using MSA alignment and precomputed sequences graph, but this feature was not documented or accessible. As a result, we implemented our own distance functions:

- **Hamming Distance**: Fast and simple, but only applicable to equal-length sequences.
- **Levenshtein Distance**: More flexible (works with unequal lengths), but significantly slower.

After testing, we observed no meaningful difference in clustering quality between the two methods. Since Hamming distance is 10–100 times faster, we chose to use it.

The idea is simple: increment the distance every time a base of the sequence is different. We can now clearly see why the SPOA MSA alignment is crucial:

```
for (size_t i = 0; i < s1.size(); ++i) {
if (s1[i] != s2[i]) {
++distance;
}
}
```

Cluster Evaluation Strategy

To evaluate the performance of our clustering approach, we compared all identified clusters from several test samples against all ground truth clusters from both J29B and J30B.

Our evaluation considered the following variables:

•Sequence count per file:
We included samples with varying numbers of 296-length sequences: approximately 200–350, ~500, 800–1200, and ~3500 sequences.

• Sequence length range:
We tested both strictly length-296 sequences and out of curiosity a broader range of 290–305 bp as well.

• Number of clusters (k):
We performed clustering with both 3 and 6 clusters to observe how granularity affects results.

• Sample file variation:
Compared similarly named FASTQ files (identified recurring patterns in prefixes: J1–J40; segments: S, L, B, GK; and suffixes: IonXpress_XXX) to check for clustering or content similarity.

The goal of this evaluation was to avoid both underfitting and overfitting by identifying a balanced configuration that reliably reveals the underlying gene variants. Additionally, we accounted for the biological possibility that different deer may possess different combinations of variants.

To better interpret the relevance of each cluster, we also printed the number of sequences assigned to each cluster. This allowed us to distinguish between biologically meaningful clusters and empty or low-relevance clusters.

Example Output

This is how an example output of three clusters would look like (without the histogram seen on the first picture in the documentation, because it is very big):

./code fastq/J14_L_CE_IonXpress_034.fastq

1306 sequences in total

305 sequences of size 296

567 sequences of a size from 290 to 305

3-Clustering over all sequences of lenght from 290 to 305 :

K-clustering STEP : 1/10

K-clustering STEP : 2/10

K-clustering has converged !!!

cluster no 1 : 109

cluster no 2 : 279

cluster no 3 : 179

GATCCTCTCTCTGCAGCACATTTCCTGGAGTATGCTAAGAGCGAGTGTCATTTCTCCAACGGG
ACGCAGCGGGTGCGGTTCCTGGACAGATACTTCTATAACCGGGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGTCCGCCAAGTA
CTGGAACAGCCAGAAGGATTTCATGGAGCAGAAGCGGGCCGAGGTGGACACGGTGTGCA
GACACAACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAA

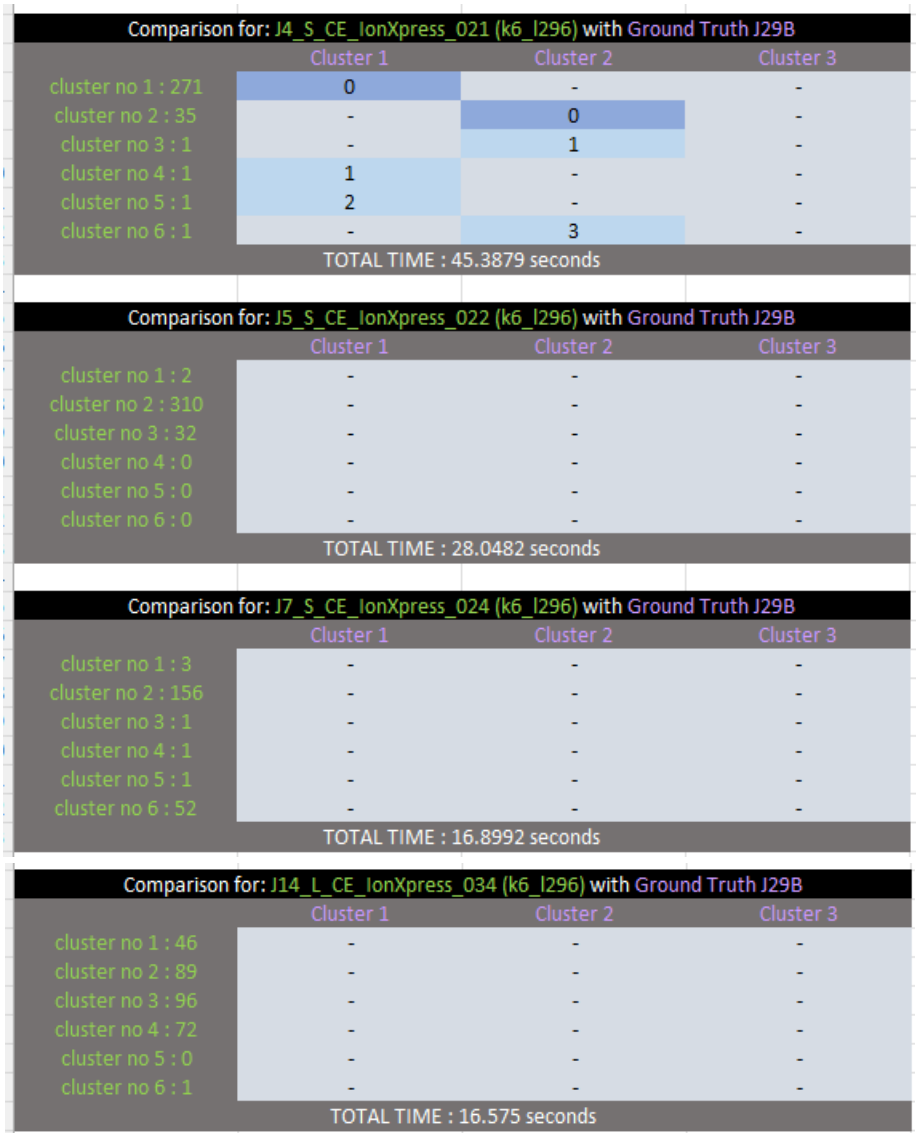
GATCCTCTCTCTGCAGCACATTTCCTGGAGCATCTTAAGGCCGAGTGTCATTTCTTCAACGGG
ACGGAGCGGATGCAGTTCCTGGCGAGATACTTCTATAACGGAGAAGAGTACGCGCGCTTCG
ACAGCGACGTGGGCGAGTTCCGGGCGGTGACCGAGCTGGGGCGGCCGGACGCCAAGTA
CTGGAACAGCCAGAAGGAGATCCTGGAGCAGCACCGGGCAGAGGTGGACAGGTACTGCA
GACACAACTACGGGGTCGGTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAAA

GATCCTCTCTCTGCAGCACATTTCCTGGAGTATCATAAGAGCGAGTGTCATTTCTCCAACGGG
ACGCAGCGGGTGGGGTACCTGGAGAGATACATCTATAACCGGGAAGAGTACGTGCGCTTCG
ACAGCGACTGGGGCGAGTACCGGGCGGTGACCGAGCTGGGGCGGCCGTCTGCCAAGTA
CATGAATAGCCAGAAGGAGCTCCTGGAGCGGAAGCGGGCCAATGTGGACACGTACTGCAG
ATACAACTACGGGGTTATTGAGAGTTTCACTGTGCAGCGGCGAGGTGACGCGAAA

Small Files (200–350 sequences of length 296)

Across all tested samples in this group, there was no significant difference between the 3-cluster and 6-cluster configurations. The last three clusters (clusters 4–6) in the 6-cluster setup were typically very small or even empty, containing at most one or two sequences. This suggests that k=3 is sufficient for smaller datasets. The three largest clusters in the 6-cluster runs corresponded roughly to the clusters found with k=3, both in size and composition.

Unfortunately, most files in this range showed no meaningful matches with either of the ground truth sets (J29B and J30B). In cases where a match could be identified, the largest ground truth cluster almost always corresponded to cluster 1, while the second-largest cluster aligned with cluster 2.



Comparison for: J4_S_CE_IonXpress_021 (k6_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 271	0	-	-
cluster no 2 : 35	-	0	-
cluster no 3 : 1	-	1	-
cluster no 4 : 1	1	-	-
cluster no 5 : 1	2	-	-
cluster no 6 : 1	-	3	-
TOTAL TIME : 45.3879 seconds			

Comparison for: J5_S_CE_IonXpress_022 (k6_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 2	-	-	-
cluster no 2 : 310	-	-	-
cluster no 3 : 32	-	-	-
cluster no 4 : 0	-	-	-
cluster no 5 : 0	-	-	-
cluster no 6 : 0	-	-	-
TOTAL TIME : 28.0482 seconds			

Comparison for: J7_S_CE_IonXpress_024 (k6_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 3	-	-	-
cluster no 2 : 156	-	-	-
cluster no 3 : 1	-	-	-
cluster no 4 : 1	-	-	-
cluster no 5 : 1	-	-	-
cluster no 6 : 52	-	-	-
TOTAL TIME : 16.8992 seconds			

Comparison for: J14_L_CE_IonXpress_034 (k6_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 46	-	-	-
cluster no 2 : 89	-	-	-
cluster no 3 : 96	-	-	-
cluster no 4 : 72	-	-	-
cluster no 5 : 0	-	-	-
cluster no 6 : 1	-	-	-
TOTAL TIME : 16.575 seconds			

Comparison for: J4_S_CE_IonXpress_021 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 271	-	-	-
cluster no 2 : 35	-	-	-
cluster no 3 : 1	-	-	-
cluster no 4 : 1	-	-	-
cluster no 5 : 1	-	-	-
cluster no 6 : 1	-	-	-
TOTAL TIME : 45.3879 seconds			

Comparison for: J5_S_CE_IonXpress_022 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 2	-	-	-
cluster no 2 : 310	-	-	-
cluster no 3 : 32	-	-	-
cluster no 4 : 0	-	-	-
cluster no 5 : 0	-	-	-
cluster no 6 : 0	-	-	-
TOTAL TIME : 28.0482 seconds			

Comparison for: J7_S_CE_IonXpress_024 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 3	-	-	-
cluster no 2 : 156	-	-	-
cluster no 3 : 1	-	-	-
cluster no 4 : 1	-	-	-
cluster no 5 : 1	-	-	-
cluster no 6 : 52	-	-	-
TOTAL TIME : 16.8992 seconds			

Comparison for: J14_L_CE_IonXpress_034 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 46	0	-	-
cluster no 2 : 89	-	0	-
cluster no 3 : 96	-	-	-
cluster no 4 : 72	-	-	-
cluster no 5 : 0	0	-	-
cluster no 6 : 1	-	-	-
TOTAL TIME : 16.575 seconds			

Comparison for: J14_L_CE_IonXpress_034 (k3_l290_305) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 109	0	-	-
cluster no 2 : 279	-	0	-
cluster no 3 : 179	-	-	-
TOTAL TIME : 22.7689 seconds			

Comparison for: J4_S_CE_IonXpress_021 (k3_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 36	-	0	-
cluster no 2 : 276	0	-	-
cluster no 3 : 1	-	1	-
TOTAL TIME : 22.5554 seconds			

Comparison for: J5_S_CE_IonXpress_022 (k3_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 2	-	-	-
cluster no 2 : 316	-	-	-
cluster no 3 : 32	-	-	-
TOTAL TIME : 14.2582 seconds			

Comparison for: J7_S_CE_IonXpress_024 (k3_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 56	-	-	-
cluster no 2 : 157	-	-	-
cluster no 3 : 1	-	-	-
TOTAL TIME : 8.72743 seconds			

Comparison for: J14_L_CE_IonXpress_034 (k3_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 48	-	-	-
cluster no 2 : 157	-	-	-
cluster no 3 : 100	-	-	-
TOTAL TIME : 12.8072 seconds			

Comparison for: J4_S_CE_IonXpress_021 (k3_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 36	-	-	-
cluster no 2 : 276	-	-	-
cluster no 3 : 1	-	-	-
TOTAL TIME : 22.5554 seconds			

Comparison for: J5_S_CE_IonXpress_022 (k3_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 2	-	-	-
cluster no 2 : 316	-	-	-
cluster no 3 : 32	-	-	-
TOTAL TIME : 14.2582 seconds			

Comparison for: J7_S_CE_IonXpress_024 (k3_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 56	-	-	-
cluster no 2 : 157	-	-	-
cluster no 3 : 1	-	-	-
TOTAL TIME : 8.72743 seconds			

Comparison for: J14_L_CE_IonXpress_034 (k3_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 48	0	-	-
cluster no 2 : 157	-	0	-
cluster no 3 : 100	-	-	-
TOTAL TIME : 12.8072 seconds			

Comparison for: J4_S_CE_IonXpress_021 (k3_l290_305) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 224	-	0	-
cluster no 2 : 568	0	-	-
cluster no 3 : 2	-	-	-
TOTAL TIME : 56.0646 seconds			

Mid-sized Files (800–1200 sequences of length 296)

This group gave clearly improved results. It appears that a higher number of sequences leads to more robust clustering. Nearly every file in this range contained at least one cluster matching one of the ground truth sequences, most often J30B.

Cluster 1 (ground truth) was consistently matched to the largest cluster in our test results, while Cluster 2 aligned with the second-largest. Interestingly, clusters 4–6 in the 6-cluster configuration occasionally contained a notable number of sequences and sometimes even showed perfect matches with ground truth clusters. This suggests they may capture real variants in some cases, but they were also usually redundant, as the same variants were already found in the 3-cluster configuration. Hence, while k=6 may uncover some additional structure, it is generally not required.

One possible explanation for matches occurring in smaller clusters could be random centroid assignment or the necessity of the algorithm to assign all sequences, resulting in a small group coincidentally clustering around a known variant. However, these small clusters often contained very few sequences and are thus likely not biologically meaningful.

In nearly all cases where the clustering matched a ground truth variant, closely related sequences were also found in the same cluster (e.g., differing by only 1–2 bases). This further validates the biological relevance of the identified variants.

Comparison for: J7_GK_CE_IonXpress_002 (k6_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 6	-	-	-
cluster no 2 : 1006	-	-	-
cluster no 3 : 6	-	-	-
cluster no 4 : 89	-	-	-
cluster no 5 : 1	-	-	-
cluster no 6 : 1	-	-	-
TOTAL TIME : 207.056 seconds			

Comparison for: J7_GK_CE_IonXpress_002 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 6	0	-	-
cluster no 2 : 1006	0	-	-
cluster no 3 : 6	1	-	-
cluster no 4 : 89	-	-	-
cluster no 5 : 1	-	-	-
cluster no 6 : 1	4	-	-
TOTAL TIME : 207.056 seconds			

Comparison for: J29_B_CE_IonXpress_005 (k6_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 1	-	2	-
cluster no 2 : 33	-	0	-
cluster no 3 : 8	-	2	-
cluster no 4 : 791	0	-	-
cluster no 5 : 4	1	-	-
cluster no 6 : 1	-	3	-
TOTAL TIME : 110.663 seconds			

Comparison for: J4_GK_CE_IonXpress_051 (k6_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 225	-	-	-
cluster no 2 : 445	-	-	-
cluster no 3 : 1	-	-	-
cluster no 4 : 1	-	-	-
cluster no 5 : 2	-	-	-
cluster no 6 : 296	-	-	-
TOTAL TIME : 170.284 seconds			

Comparison for: J5_GK_CE_IonXpress_052 (k6_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 196	-	-	-
cluster no 2 : 765	-	-	-
cluster no 3 : 2	-	-	-
cluster no 4 : 1	-	-	-
cluster no 5 : 9	-	-	-
cluster no 6 : 0	-	-	-
TOTAL TIME : 102.605 seconds			

Comparison for: J30_B_CE_IonXpress_006 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 3	1	-	-
cluster no 2 : 4	-	-	-
cluster no 3 : 493	-	1	-
cluster no 4 : 642	0	-	-
cluster no 5 : 1	4	-	-
cluster no 6 : 1	-	2	-
TOTAL TIME : 140.463 seconds			

Comparison for: J4_GK_CE_IonXpress_051 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 225	-	2	-
cluster no 2 : 445	1	-	-
cluster no 3 : 1	-	4	-
cluster no 4 : 1	4	-	-
cluster no 5 : 2	-	-	-
cluster no 6 : 296	-	-	-
TOTAL TIME : 170.284 seconds			

Comparison for: J5_GK_CE_IonXpress_052 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 196	-	1	-
cluster no 2 : 765	0	-	-
cluster no 3 : 2	3	-	-
cluster no 4 : 1	2	-	-
cluster no 5 : 9	0	-	-
cluster no 6 : 0	0	-	-
TOTAL TIME : 102.605 seconds			

Comparison for: J29_B_CE_IonXpress_005 (k3_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 40	-	0	-
cluster no 2 : 787	0	-	-
cluster no 3 : 11	1	-	-
TOTAL TIME : 61.7132 seconds			

Comparison for: J4_GK_CE_IonXpress_051 (k3_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 231	-	-	-
cluster no 2 : 739	-	-	-
cluster no 3 : 1	-	-	-
TOTAL TIME : 49.8881 seconds			

Comparison for: J5_GK_CE_IonXpress_052 (k3_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 196	-	-	-
cluster no 2 : 770	-	-	-
cluster no 3 : 9	-	-	-

Comparison for: J7_GK_CE_IonXpress_002 (k3_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 6	-	-	-
cluster no 2 : 1103	-	-	-
cluster no 3 : 6	-	-	-
TOTAL TIME : 63.7496 seconds			

Comparison for: J30_B_CE_IonXpress_006 (k3_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 646	0	-	-
cluster no 2 : 4	-	-	-
cluster no 3 : 494	-	1	-

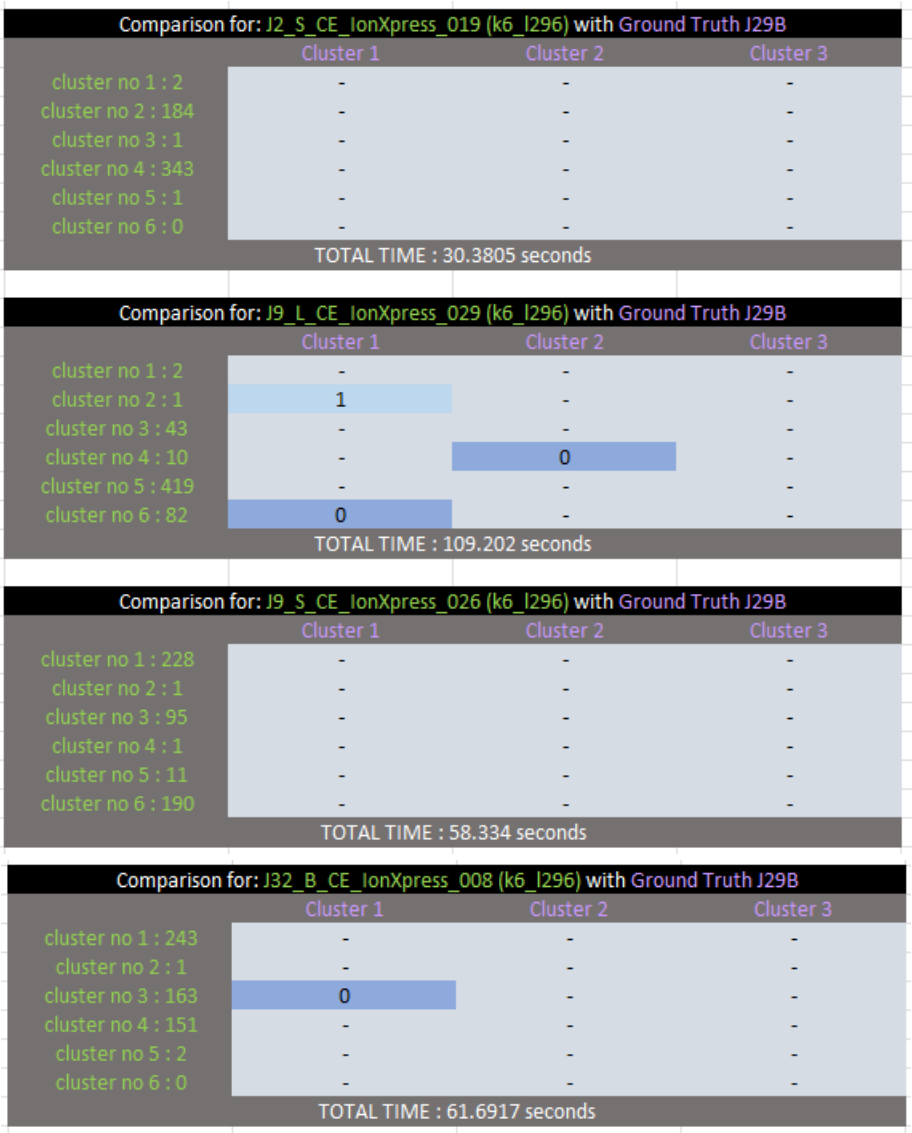
Comparison for: J4_GK_CE_IonXpress_051 (k3_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 231	-	2	-
cluster no 2 : 739	1	-	-
cluster no 3 : 1	-	4	-
TOTAL TIME : 49.8881 seconds			

Comparison for: J5_GK_CE_IonXpress_052 (k3_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 196	-	1	-
cluster no 2 : 770	0	-	-
cluster no 3 : 9	0	-	-

Comparison for: J7_GK_CE_IonXpress_002 (k3_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 6	0	-	-
cluster no 2 : 1103	0	-	-
cluster no 3 : 6	1	-	-
TOTAL TIME : 63.7496 seconds			

Intermediate Case (~500 sequences of length 296)

This group also showed more frequent matches with J30B, aligning with the trend observed in the 800–1200 range. Some cluster 4–6 configurations in k=6 contained more sequences than in the smaller files, but no dramatic difference was noted.



Comparison for: J2_S_CE_IonXpress_019 (k6_l296) with Ground Truth J29B

	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 2	-	-	-
cluster no 2 : 184	-	-	-
cluster no 3 : 1	-	-	-
cluster no 4 : 343	-	-	-
cluster no 5 : 1	-	-	-
cluster no 6 : 0	-	-	-

TOTAL TIME : 30.3805 seconds

Comparison for: J9_L_CE_IonXpress_029 (k6_l296) with Ground Truth J29B

	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 2	-	-	-
cluster no 2 : 1	1	-	-
cluster no 3 : 43	-	-	-
cluster no 4 : 10	-	0	-
cluster no 5 : 419	-	-	-
cluster no 6 : 82	0	-	-

TOTAL TIME : 109.202 seconds

Comparison for: J9_S_CE_IonXpress_026 (k6_l296) with Ground Truth J29B

	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 228	-	-	-
cluster no 2 : 1	-	-	-
cluster no 3 : 95	-	-	-
cluster no 4 : 1	-	-	-
cluster no 5 : 11	-	-	-
cluster no 6 : 190	-	-	-

TOTAL TIME : 58.334 seconds

Comparison for: J32_B_CE_IonXpress_008 (k6_l296) with Ground Truth J29B

	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 243	-	-	-
cluster no 2 : 1	-	-	-
cluster no 3 : 163	0	-	-
cluster no 4 : 151	-	-	-
cluster no 5 : 2	-	-	-
cluster no 6 : 0	-	-	-

TOTAL TIME : 61.6917 seconds

Comparison for: J2_S_CE_IonXpress_019 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 2	0	-	-
cluster no 2 : 184	-	0	-
cluster no 3 : 1	1	-	-
cluster no 4 : 343	0	-	-
cluster no 5 : 1	-	2	-
cluster no 6 : 0	0	-	-
TOTAL TIME : 30.3805 seconds			

Comparison for: J9_L_CE_IonXpress_029 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 2	-	-	-
cluster no 2 : 1	-	-	-
cluster no 3 : 43	-	-	-
cluster no 4 : 10	-	-	-
cluster no 5 : 419	-	-	-
cluster no 6 : 82	-	-	-
TOTAL TIME : 109.202 seconds			

Comparison for: J9_S_CE_IonXpress_026 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 228	0	-	-
cluster no 2 : 1	-	2	-
cluster no 3 : 95	-	-	-
cluster no 4 : 1	-	-	-
cluster no 5 : 11	-	-	-
cluster no 6 : 190	-	0	-
TOTAL TIME : 58.334 seconds			

Comparison for: J32_B_CE_IonXpress_008 (k6_l296) with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 243	0	-	-
cluster no 2 : 1	2	-	-
cluster no 3 : 163	-	-	-
cluster no 4 : 151	-	0	-
cluster no 5 : 2	-	0	-
cluster no 6 : 0	0	-	-
TOTAL TIME : 61.6917 seconds			

Large File (3500 sequences of length 296)

We tested one exceptionally large file in this group. No matches with ground truth could be found. This might be due to overcomplexity or noise introduced by too many sequences, which may degrade clustering performance. Alternatively, it could simply be due to random variation in the single file tested. More large files would need to be tested to confirm this trend.

Comparison for: J38_B_CE_IonXpress_014 (k6_l296) with Ground Truth J29B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 1	-	-	-
cluster no 2 : 10	-	-	-
cluster no 3 : 2270	-	-	-
cluster no 4 : 1156	-	-	-
cluster no 5 : 1	-	-	-
cluster no 6 : 10	-	-	-
TOTAL TIME : 975.179 seconds			

Comparison for: J38_B_CE_IonXpress_014 with Ground Truth J30B			
	Cluster 1	Cluster 2	Cluster 3
cluster no 1 : 1	-	-	-
cluster no 2 : 10	-	-	-
cluster no 3 : 2270	-	-	-
cluster no 4 : 1156	-	-	-
cluster no 5 : 1	-	-	-
cluster no 6 : 10	-	-	-
TOTAL TIME : 975.179 seconds			

Runtime Considerations

Unsurprisingly, the k=3 clustering consistently ran in roughly half the time compared to k=6 clustering. The effect was especially visible in large files, where runtime became more significant.

Sample File Variation

We also observed that the files with nearly identical filenames, especially numbers (e.g., differing only by a suffix or character), produced entirely different gene content and clustering outcomes. This suggests that file naming does not necessarily reflect genetic similarity, and samples must be treated independently.

Conclusion

Through systematic clustering and comparison of deer gene sequence data, we found that clustering quality and accuracy improve with a moderate number of input sequences—particularly around 800 to 1200 reads. Using k=3 clusters generally yielded stable results, with additional clusters (k=6) providing little extra benefit, except in rare cases. Importantly, strong matches with the J30B ground truth were consistently more common than with J29B, especially in larger and mid-sized datasets.

We also observed that file names do not reliably indicate genetic similarity, and each sample should be treated independently. Overall, our approach was able to reconstruct realistic gene variants from raw sequence data using clustering techniques, with Hamming distance and SPOA-based alignment forming a robust foundation for analysis. Further refinement could focus on dynamic k selection and better pre-filtering to handle extreme dataset sizes.

Sources

- lectures of Bioinformatics at FER
- recommendations and explanations from supervisor Mr. Krešimir Križanović
- EPFL course: https://edu.epfl.ch/coursebook/en/algorithms-i-CS-250
- GitHub SPOA library: https://github.com/rvaser/spoa