METK Barley SNP-Chip: Linking Genetics to Protein Content

Agnes Kivistik, Liselle Velner, Alexandra Voit

Introduction

Our project, utilizing exclusive 2024 data from METK, explores the genetic factors influencing barley's protein content—a key attribute for both nutritional and agricultural purposes. By analyzing the correlation between barley's genetic makeup and its protein levels, we aim to provide insights that could guide future crop breeding efforts and support the growing demand for plant-based protein sources.

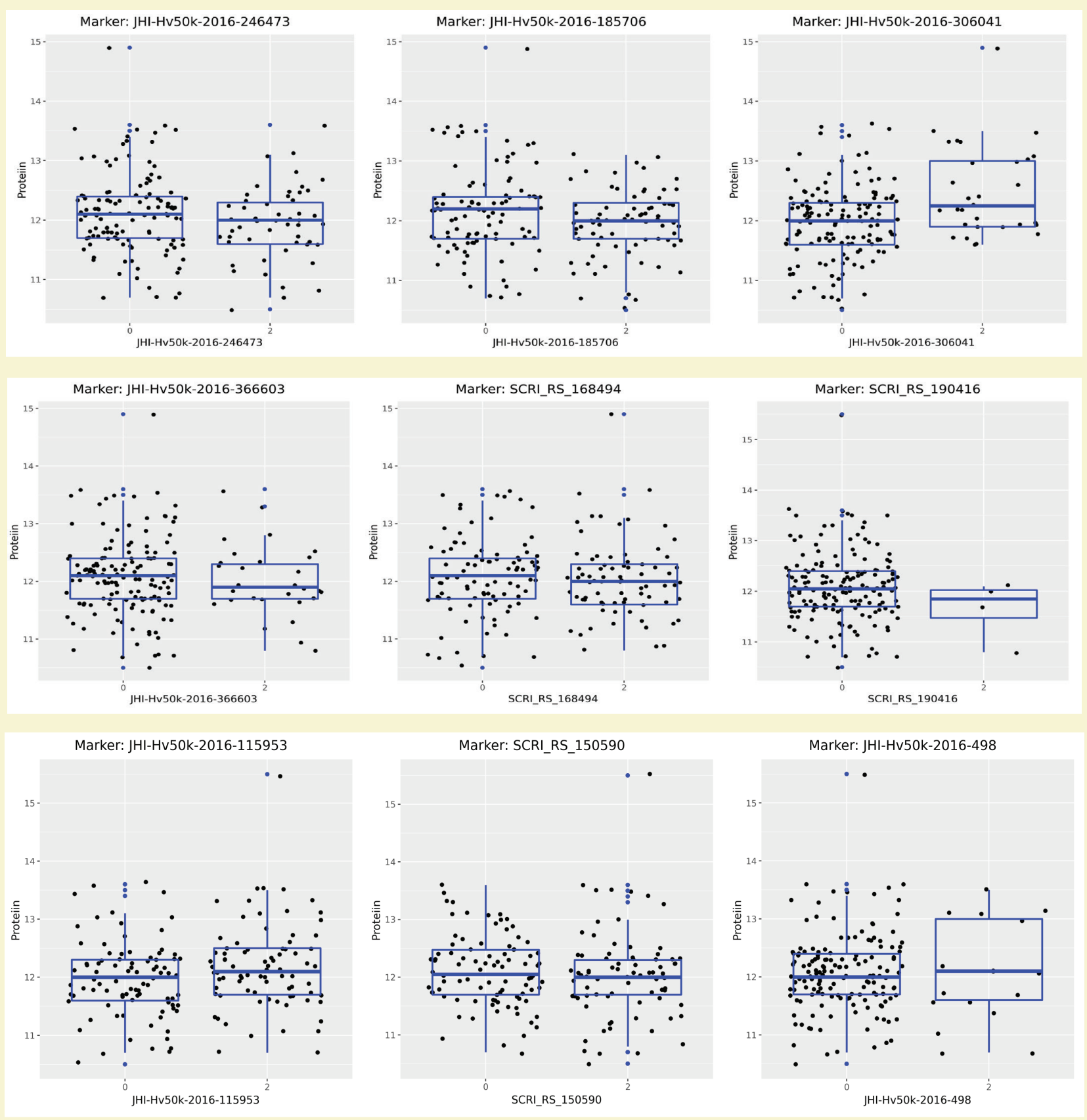

Figure 1: The scatter and box plots provided a clear visual representation of how protein content varies across different genotypes for each marker. These visualizations helped identify SNP markers that are associated with larger differences in protein content.

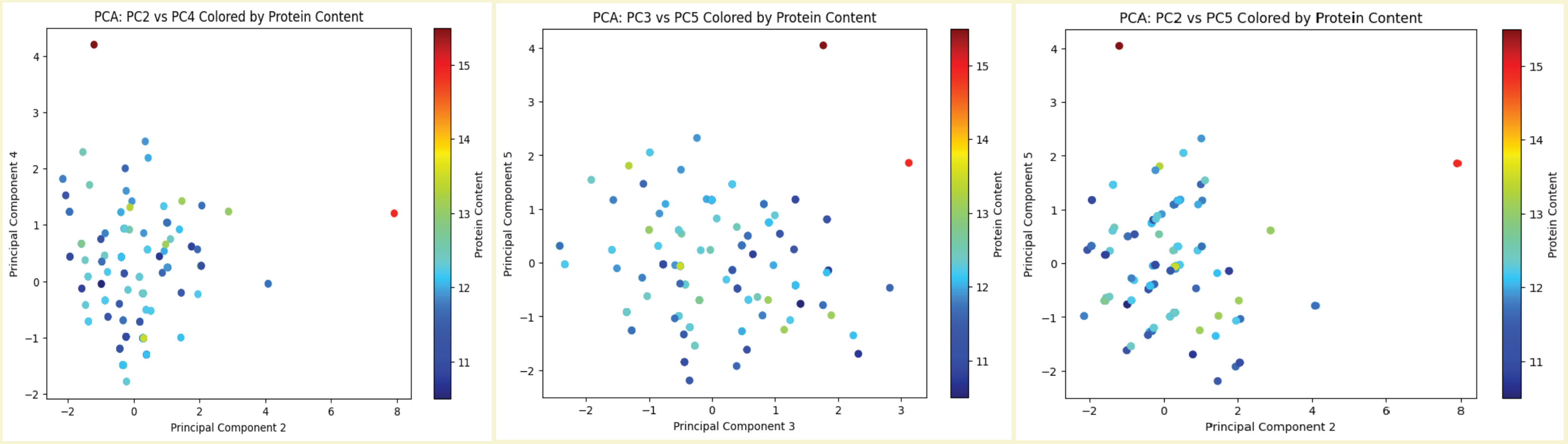

Figure 2: A scatter plot was created to visualize the PCA results, with points colored according to protein content. This helped assess how protein content was distributed across clusters and PCA components.

Data analysis

Data Cleaning: By filtering out SNP markers with insufficient variation (where one genotype or NaN values dominated), we focused only on markers with useful genetic diversity.

Protein Content Variation: The analysis showed which SNP markers had significant variation in protein content between different genotypes. Some SNP markers exhibited a large range of protein content differences, indicating that the genotype at these markers could have a notable effect on protein expression. (Figure 1)

PCA: We performed a series of steps to analyze SNP (Single Nucleotide Polymorphism) data in relation to protein content, with the goal of identifying any potential patterns or clusters in the data. (Figure 2)

Dendrogram: We used Hamming distance to quantify the similarity between SNP samples, which is appropriate for binary or categorical data. Missing data was handled using mean imputation, ensuring that incomplete rows wouldn't disrupt the analysis. Applied hierarchical clustering to group similar samples based on their SNP profiles, helping to identify patterns or clusters of genetic similarity. (Figure 3)

Prediction Models: For protein content prediction different models were used: RandomForestRegressor, DecisionTreeRegressor, Simple Linear, Lasso, and Ridge Regression. Ridge Regression had the best R2 score (0.089) and MSE (0.535) out of the selection. Unfortunately, these models are probably not sophisticated enough and a more specialised software such as MEGA should be used.

Results

2 markers found that have a stronger correlation to protein: JHI-Hv50k-2016-306041 and SCRI_RS_190416. Both are located on chromosome 5H. 2 varieties (Kornelija, Ilma) are genetically very different from the rest, having more unique DNA as well as a higher protein content. Although current findings do not point towards any significant correlation between protein content and SNP markers, this is a solid base for future research.

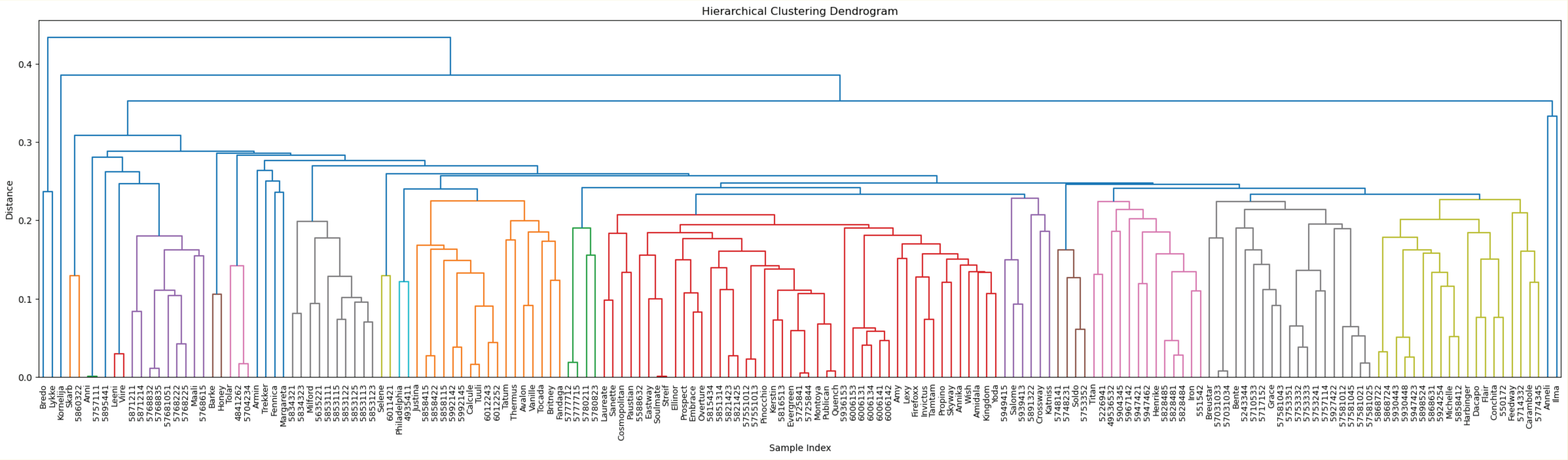

Figure 3: Dendrogram highlighted sample relationships and distinct clusters.

METK Maaelu Teadmuskeskus