Relatedness inference using RNA-seq data from Glossophaga soricina bats

Bat-pop-pangenomic team

July 25, 2023

1 Introduction

Due to the relatively low proportion of the genome covered by RNA-seq, coupled with skewed coverage of different genes and the presence of allele-specific expression, the precision of kinship inference can be limited. Despite these challenges, the potential of RNA-seq data for detecting kinship between individuals through pairwise identity by descent (IBD) estimates remains possible (Blay, NAR, 2019).

In this study, we assess the potential of using RNA-seq data to detect kinship in the Neotropical nectarivorous bat (Glossophaga soricina). These bats, which are among the smallest mammals, exhibit specialised exploitation of flowers and consume large amounts of dilute nectar each night. In terms of their reproductive habits, breed once or twice a year depending on the geographical location, with a gestation period of about 3-4 months. These bats are known to form maternity colonies, where multiple adult females congregate to give birth and raise their offspring. The social and reproductive structures of these colonies, coupled with the bats' polygynous mating system, can have significant impacts on their genetic structure and the pattern of kinship relationships, providing a compelling backdrop for our analysis. Though most accounts indicate that G. soricina either breeds continuously throughout the year or is bimodally polyestrous. Normally only single offspring, but twins have been reported. In general, female bats are sexually mature near the end of their first year. Male bats take slightly longer and will reach sexual maturity after their first year (Lee, JMAL, 2002).

The genome of the Neotropical nectarivorous bat, Glossophaga soricina, has been sequenced and found to consist of 15 pairs of autosomes and one pair of sex chromosomes. Like many other mammals, these bats have an XY sex determination system, with males being XY and females XX.

2 Materials and Methods

Data for this study were derived from Glossophaga soricina bats, with RNA-seq data specifically taken from the liver. The selection of variants for analysis was conducted selecting only those variants with a Qual score greater than 500, a Hardy-Weinberg p-value higher than 0.001, a Minor Allele Frequency (MAF) of at least 0.10 across the entire population, and a MAF of at least 0.05 within each subpopulation (referred to as LC and GC).

This process resulted in a total of 52,422 high-quality Single Nucleotide Polymorphisms (SNPs) being identified for further investigation. Tailored scripts were then employed to convert this data into the .map and .ped file formats required by the PLINK2 software.

Subsequent kinship analysis utilised the PLINK2 software package, specifically the –make-king option, which provides accurate results even when good population allele frequency estimates are unavailable. This approach is based on the methodology described by Manichaikul et al. (2010) in their study on robust relationship inference in Genome-Wide Association Studies (GWAS). The algorithm they developed was presented for enabling the precise inference of relationships between pairs of individuals through robust estimation of their kinship coefficient, irrespective of sample composition or population structure.

In this study, a total of 5,460 pairwise comparisons were made (105 individuals), with only those featuring more than 50,000 shared SNPs being considered for further analysis. The output data includes the ID of the two individuals compared, the number of shared SNPs, the proportion of shared heterozygous SNPs, the proportion of SNPs where the individuals share no alleles (IBS0), and the estimated kinship coefficient.

Additional metadata were provided for each bat and were used for grouping and analysing the data based on individual characteristics. This metadata included each bat's ID, the subpopulation (LC or GC) it belongs to, its sex, and a minimum age estimate.

Based on the kinship coefficient values, the following categories were established: a coefficient greater than 0.354 corresponding to duplicate/Monozygotic (MZ) twin relationships, a coefficient within the range 0.177 to 0.354 representing 1st-degree relationships, a coefficient within the range 0.0884 to 0.177 indicating 2nd-degree relationships, and a coefficient within the range 0.0442 to 0.0884 corresponding to 3rd-degree relationships. Any relationships with a coefficient less than 0.0442 were deemed to be unrelated. This values were chosen based on King software documentation (https://www.kingrelatedness.com/manual.shtml).

The remaining analyses and graphics were generated using various libraries in Python.

3 Results

The dataset comprises 105 Glossophaga soricina bats, split into two subpopulations: 66 individuals belong to the LC colony and 39 to the GC colony.

In terms of sex, the dataset includes 72 females and 33 males. The sex distribution within the colonies is as follows: within the GC colony, there are 20 males and 19 females, while within the LC colony, there are 53 females and 13 males.

Age data is available for 66 individuals, all of whom belong to the LC colony. The median of the estimated minimum age for these individuals is 5.41 years.

Figure 1 presents the distribution of kinship coefficients resulting from the pairwise kinship analysis. The kinship coefficient is a measure of genetic relatedness, with higher values indicating a closer genetic relationship. The histogram provides an overview of the range and frequency of these kinship coefficients among pairs of bats in the dataset, highlighting the diversity of genetic relationships within the study population. Furthermore, supplementary Figure S1 presents a heatmap of the kinship coefficients between pairs of individuals. The heatmap visualises the degree of genetic relatedness between pairs of bats, with warmer colours indicating closer genetic relationships. The visualisation reveals clusters of closely related individuals. Notably, certain individuals, such as GSO-81-e, appear to be unrelated to any others, indicating the presence of significant genetic diversity within the population.

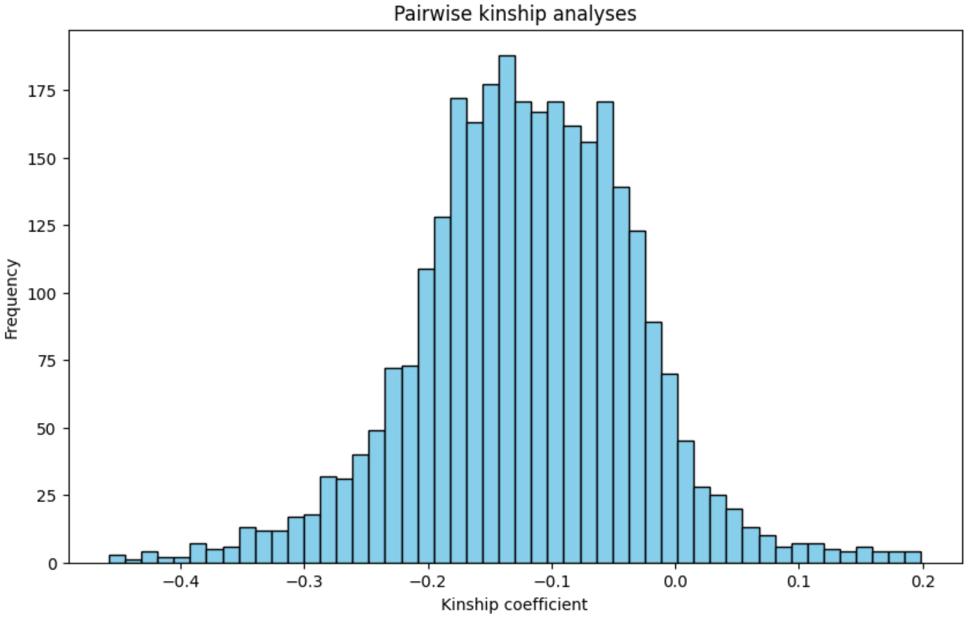

Figure 1: The histogram depicts the distribution of kinship coefficients obtained from pairwise kinship analysis. Each bin represents a range of kinship coefficients, with the height of the bin indicating the frequency of pairs with a kinship coefficient within that range.

The analysis presented in Figure 2 shows the closest relationship found in the bat population finding 1st and 2nd degree relatives. No relationships were identified as duplicates or monozygotic twins. Relationships were further classified by whether pairs of individuals come from the same or different colonies. This analysis reveals that the higher kinship coefficients are primarily observed among individuals from the same colony. This finding suggests that closer genetic relationships are more prevalent within colonies than between them, highlighting the influence of colony structure on genetic relatedness among the bats.

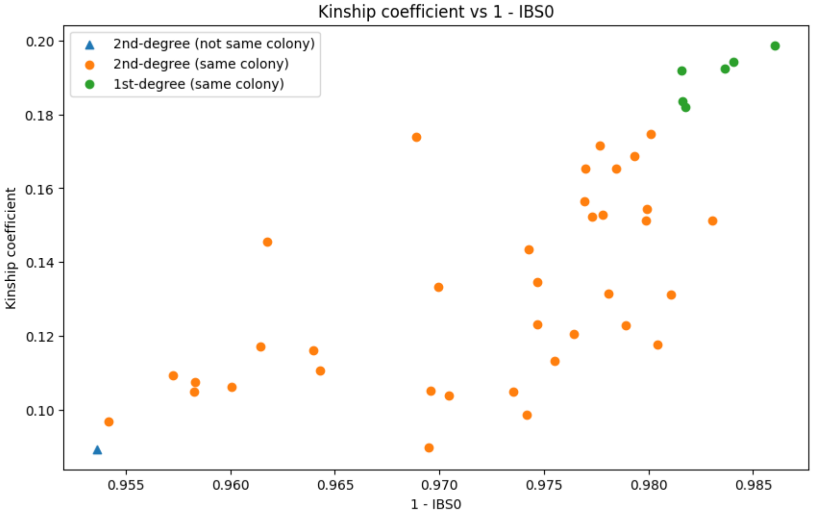

Figure 2: The scatter plot illustrates the relationship between the kinship coefficient and $1-IBSO$ for pairs of individuals. Each point represents a pair of individuals, with the kinship coefficient on the y-axis and $1 - IBSO$ on the x-axis. Points are colour-coded according to the inferred degree of relatedness, with separate colours for 1st degree, 2nd degree, and 3rd degree relatives. The plot shows a clear positive relationship, with higher kinship coefficients associated with higher values of $1 - IBS0$. This suggests that pairs of individuals with closer genetic relationships tend to have a greater proportion of SNPs where they share at least one allele.

To further explore the relationship between the obtained kinship coefficients and the bats' colony membership, we analysed the kinship coefficients between individuals from different colonies, as well as the coefficients obtained for individuals from the same colony. As shown in Figure 3, higher kinship coefficients are primarily observed among individuals from the same colony, suggesting a strong influence of colony membership on genetic relatedness.

Given our findings that higher kinship coefficients are primarily observed among individuals from the same colony and the low or negligible reproduction between individuals from different colonies, we proceeded to analyse the LC and GC colonies separately. Figure 4 shows scatter plots of the kinship coefficient against 1 $1 - IBS0$ for pairs of individuals from the LC and GC colonies. The analyses reveal that the LC colony exhibits a higher number of close relationships between individuals compared to the GC colony, where relationships are fewer and tend to be more distant.

To explore the kinship relationships within the LC colony in greater detail, a network graph was constructed using the 1st and 2nd-degree relationships identified in the kinship analysis (Figure 5). Each node in the graph represents an individual bat, with males and females distinguished by different shapes (square male, circle female). The size of each node

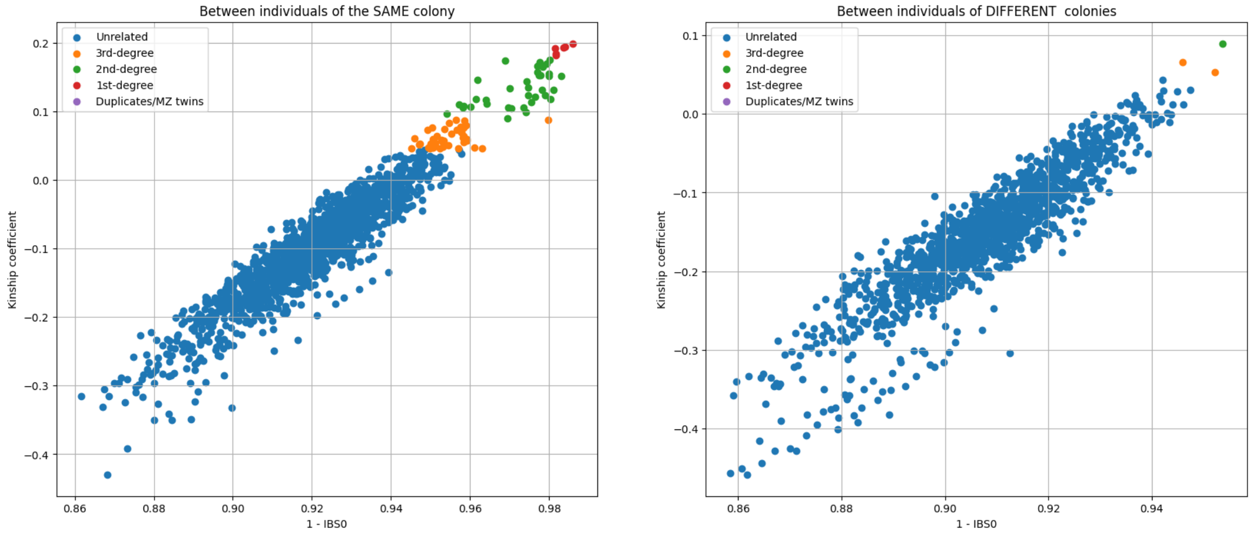

Figure 3: This figure consists of two scatter plots that illustrate the relationship between the kinship coefficient and $1 - IBS0$ for pairs of individuals from the same colony (left) and pairs from different colonies (right). Each point represents a pair of individuals, with the kinship coefficient on the y-axis and $1 - IBS0$ on the x-axis. Points are colour-coded according to the inferred degree of relatedness.

corresponds to the age of the individual bat. The relationships between individuals are shown as edges, with 1st-degree relationships shown in red and 2nd-degree relationships in blue.

Notably, certain individuals appear to have multiple close relationships within the colony. For instance, individual GSO-129-b shows 1st-degree relationships with both GSO-84-m and GSO-75-d, who in turn have a 2nd-degree relationship with each other. This pattern could be consistent with a pedigree in which GSO-129-b is the parent of both GSO-84-m and GSO-75-d, who are half-siblings. However, to confirm this interpretation, additional types of markers such as mitochondrial DNA and X and Y chromosomes would need to be considered.

4 Conclusion

The analysis of RNA-seq data from 105 bats from the Glossophaga soricina species has revealed patterns of kinship within bat colonies. Notably, we observed a clear differentiation between the LC and GC colonies in terms of the degree of relatedness among individuals. The LC colony exhibited a greater number of close relationships (1st and 2nd degree). In contrast, the GC colony displayed fewer and more distant relationships, suggesting a less related group of bats.

The network graph of relationships within the LC colony has provided further insights into the potential family structures within this group. However, while this analysis has suggested

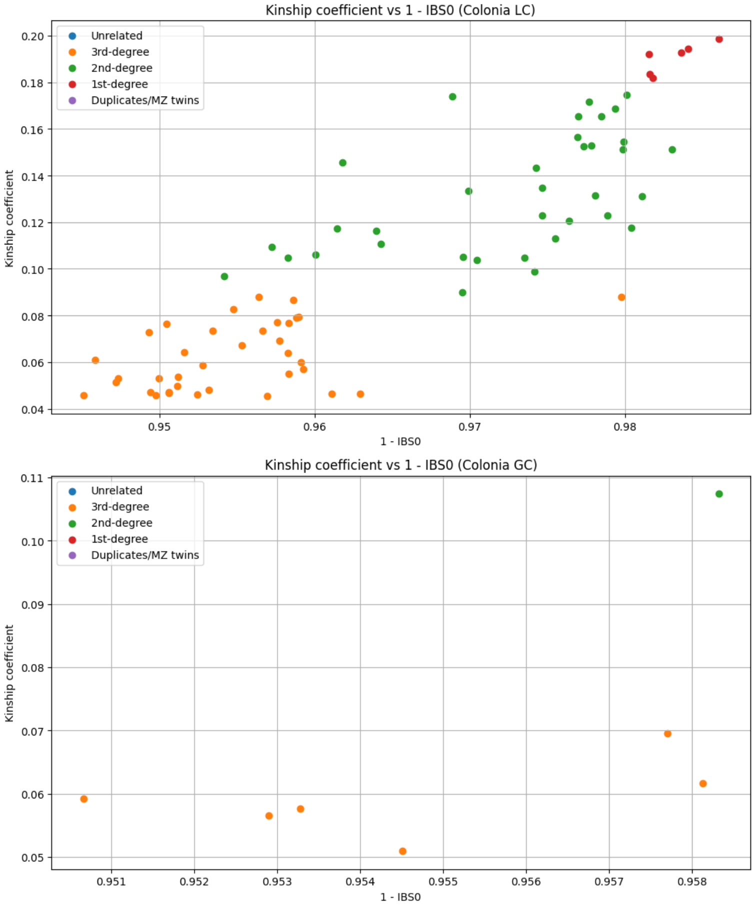

Figure 4: This figure consists of two scatter plots that illustrate the relationship between the kinship coefficient and $1 - IBS0$for pairs of individuals from the same colony (left) and pairs from different colonies (right). Each point represents a pair of individuals, with the kinship coefficient on the y-axis and $1 - IBS0$ on the x-axis. Points are colour-coded according to the inferred degree of relatedness.

the presence of parent-offspring, half-sibling relationships and other relationships, the precise pedigree structure remains to be resolved. More comprehensive data, incorporating additional types of genetic markers such as mitochondrial DNA and X and Y chromosomes, would provide a more detailed picture of the familial relationships within this colony.

Nevertheless, our analysis suggest the potential of RNA-seq data to elucidate kinship relationships within bat colonies. Despite the inherent limitations of this type of data for pedigree analysis, we were able to detect evidence of relatedness among bats, providing a starting point for further investigations.

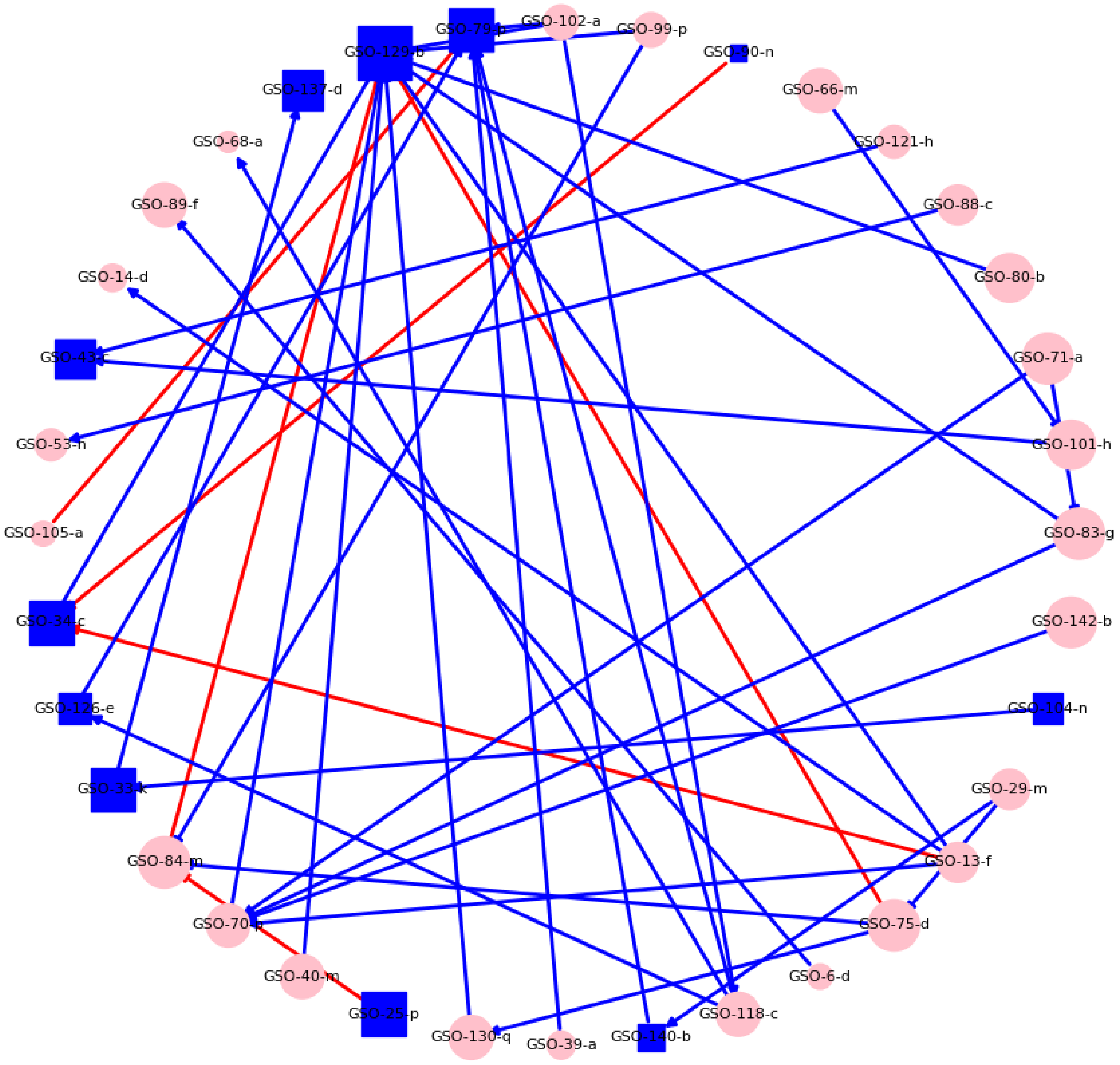

Figure 5: This figure consist in a network, where nodes represent bat individuals and edges relationships between them. Red edges indicates 1st degree relationship and blue edges 2nd degree. Circules correspond to females and squares to males.

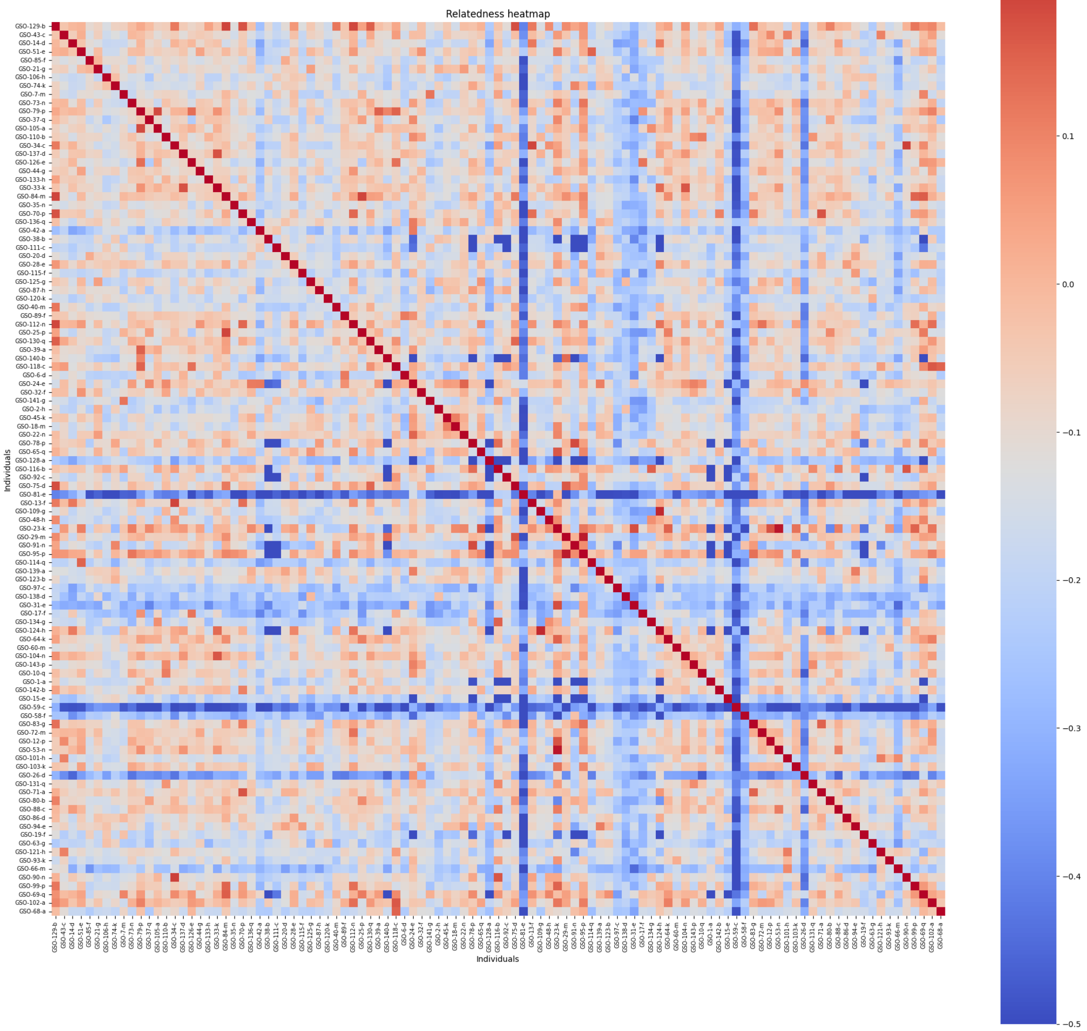

Fig S1. 6: The heatmap visualises the kinship coefficients between pairs of individuals, as calculated from the pairwise kinship analysis. Each cell in the heatmap represents a pair of individuals, with the colour of the cell indicating the kinship coefficient. A colour scale ranging from cool (lower kinship) to warm (higher kinship) is used. The heatmap reveals clusters of individuals with higher relatedness, as well as certain individuals, such as GSO-81-e, who appear to be unrelated to any others.