Using Dimension Reduction Techniques to Model Genetic Relationships for Association Studies

Thesis Proposal for

Andrew Crossett

Carnegie Mellon University
Department of Statistics

Committee: Kathryn Roeder[1], Ann Lee[1], Alessandro Rinaldo[1], and Bernie Devlin[2]

[1] Department of Statistics
Carnegie Mellon University

[2] Department of Psychiatry
University of Pittsburgh School of Medicine

Abstract

Beyond a few degrees of relationship pedigrees are rarely known with absolute certainty. This uncertainty is often elevated in population isolates, in which all extant individuals trace their ancestry to a limited number of founders. Cryptic relatedness can have a detrimental impact on nominal false positive rates for genetic association tests. An algorithm overcoming this problem is as follows: first estimate the relatedness of all pairs of individuals assessed for association; then adjust the test for association on the basis of relatedness. Methods exist by which relatedness can be estimated using genotypes obtained as part of a genome wide association study (GWA). It is important to recognize that using genotype information to estimate relationships between pairs of individuals can be very noisy.

Treelets are an adaptive approach to dealing with noisy, high-dimensional and unordered data. Treelets simultaneously construct a hierarchical tree and an orthonormal basis that represent the internal structure of the data. We propose to use treelets on estimated relationship data by examining each individuals relationship to everyone else. Noise is removed by identifying the most important features of the basis and then reconstructing the data. We apply these techniques to data from Palau, an Oceanic nation of relatively recent origin in human history. These data are part of an ongoing project to understand the genetic basis of schizophrenia.

Introduction

Genome wide association studies (GWAS) are an extremely useful tool for finding genetic variants associated with complex diseases. It has been well documented that spurious associations between these diseases and gene locations may exist in the presence of cryptic relatedness [1]. Cryptic relatedness can arise in a number of ways. One such way occurs when ancestry is not properly accounted for in a case-control sampling design. Due to demographic, biological and random forces, genetic variants differ in allele frequency in populations around the world, creating population structure or stratification reflected by ancestry [2]. These inherent ancestral differences may lead to spurious associations. One way to alleviate this problem is to use family-based designs [3; 4; 5; 6]. These designs model the transmission of alleles, conditional on the family structure, and so they eliminate the problem of population stratification. Unfortunately, family-based designs can be difficult to collect. It is quite common that a researcher would have both types of data (case-control and family-based) when performing a GWAS.

We propose a method to analyze family-based samples together with unrelated cases and controls. The method builds on the idea of matched case-control analysis using conditional logistic regression. For each trio within the family, a case (the proband) and matched pseudo-controls are constructed, based upon the transmitted and untransmitted alleles. Unrelated controls, matched by genetic ancestry, supplement the sample of pseudo-controls; likewise unrelated cases are also paired with genetically matched controls. Within each matched stratum, the case genotype is contrasted with control/pseudo-control genotypes via conditional logistic regression, using a method we call *matched*-conditional logistic regression (mCLR).

Spectral graph analysis of numerous genotypes provides a tool for mapping genetic ancestry [7]. The result of such an analysis can be thought of as a multidimensional map, or eigenmap, in which the relative genetic similarities and differences amongst individuals is encoded in the map. Once constructed, new individuals can be projected onto the ancestry map based on their genotypes. Successful differentiation of individuals of distinct ancestry depends upon having a diverse, yet representative sample from which to construct the ancestry map. Once samples are well-matched, mCLR yields comparable power to competing methods but ensures excellent control over Type I error. This work has previously been completed [8].

Another source of cryptic relatedness may be due to a researchers desire to sample from individuals of some isolated population. There is no reason to estimate each individual's ancestry as it is known that they are all related. For instance, Palau, an Oceanic island nation of relatively recent origin in human history has been shown to have an unusually high rate of schizophrenia [9]. Unfortunately, standard GWAS can not be directly applied due to the cryptic relatedness of this isolated island's residents. Many algorithms have been developed to account for this relatedness [10; 11]. Each of these methods estimates how related individuals are and then uses this relatedness to adjust their respective test statistics. These estimates of relatedness can be very noisy.

The most common way of dealing with relatedness is to examine the kinship coefficient (Φ) between pairs of individuals [12]. Loosely defined, the kinship coefficient is the probability that any pair of randomly chosen genes between two individuals are identical. This property is known as Identity by Descent (IBD). Kinship coefficients can be calculated in one of two ways: (i) using pedigree or family tree information and (ii) using genotype information.

The first of which involves counting the number of paths between pairs of individuals until a common ancestor. Some common pedigree-based kinship coefficients can be found in Table 1[13]. The second method uses a likelihood based approach. Given only genotype information, it is very difficult to determine if two genes are IBD. Typically, we do not know if the alleles are inherited from a common ancestor, but we do know if the alleles are of the same form. This is known as Identity by State (IBS). Therefore, we can construct a locus specific likelihood based on the (known) possible IBS modes, conditional on the probabilities of the (unknown) number of IBD alleles. Maximum-likelihood techniques can then be applied [14; 15; 16; 17]. For the rest of the paper, we will predominately be dealing with (ii), as errors in this method are more statistical in nature.

Ultimately, we want to use all pairwise relationships in the sample to improve our estimates of each pairwise relationship. In other words, if two individuals are related to each other then they are likely to have similar kinship coefficients with others. Therefore, we can essentially extract information we need about either from some single transformation of the two individuals, or variables in general.

Two of the most common approaches to the transformation problem above is to use principal component analysis (PCA) and wavelets. PCA is a simple technique but fails to capture localized structure among the variables due to its global preserving properties. Wavelets are more sensitive to localized structure but are ill equipped to handle unordered data. One way around this problem is to order the data in some meaningful way and *then* use wavelets. Therefore, we propose to use Treelets [18], an adaptive technique that simultaneously orders the data by building an agglomerative hierarchical dendrogram and builds an orthogonal basis that reflects the internal structure of the data. This is done through a series of Jacobi rotations between pairs of variables [19]. This method is similar to work done by Murtagh [20] and Singh, et. al [21]. Murtagh uses a balanced Haar transform at every level of an already established hierarchical tree. It does not simultaneously build the tree as well. Also, balanced Haar wavelets on a dendrogram do not produce an orthonormal basis. The second method supplements building a hierarchical tree with using an unbalanced Haar basis representation. The main difference between it and Treelets is in the goal of each method. Singh's method actively attempts to find the most sparse transform in general, where as Treelets do not. Also, Singh's method is not interested in determining relationships among individuals, or in their case, nodes of a network. They are only interested in the coordinates that correspond to non-zero means of signal strength.

Once we have determined a basis that captures the internal structure of the data, we can use the corresponding expansion coefficients to reconstruct our original matrix. Noise is removed by thresholding the expansion coefficients and *then* reconstructing the matrix. We show that Treelets do indeed capture closely related individuals and that thresholding the expansion coefficients does remove noise from our matrix. This process is examined through simulations.

Data

We have both pedigree-based and genotype-based kinship coefficient values ($2 * \Phi$) for 556 individuals from Palau. Here, we will take $2 * \Phi_{ij}$ to be twice the kinship coefficient between individual i and individual j. Because Palau is a young, isolated population, most individuals are at least somewhat related. Nevertheless, many of these relationships are quite distant. Consequently, most kinship coefficients are close to zero. Therefore, we transform the data to the **Coefficient of Relationship (R)** [13]:

$$\begin{aligned} \rho &= 2^{-R} = 2\Phi \\ &\Rightarrow R = \frac{-log(2\Phi)}{log(2)}. \end{aligned}$$

This transformation is not arbitrary, as it corresponds nicely to a natural scale of relationship. For instance, $R_{ij} = 0$ when i and j are MZ twins or when $i = j$, $R_{ij} = 1$ if i and j are parent-offspring or full siblings and $R_{ij} = 2$ when i and j are aunt/uncle-niece/nephew or grandparent-grandchild. Because this is a log-transformation, I added 5e-07 whenever $\Phi_{ij} = 0$. Therefore, we are dealing with a maximum value of about 21. A plot of both 2Φ and R for the genotype-based Palau dataset can be found in Figure 1

Simulations

To thoroughly evaluate our approach we simulate some (realistic) data to verify that using treelets is reasonable. One can also evaluate our methods using pedigree-based degree of relationship values; however, entries beyond 5 are not reliable. A simple simulation algorithm for constructing pedigree-based R values is as follows:

1. Start with 80 founders or completely unrelated individuals (40-40 males/females) and marry them

2. Choose $X \sim Bin(4, .5)$ children for each couple from (1)

3. Marry children from (2) using the following rules:

   - Can't marry siblings
   - Randomly choose a female for every male (if enough). If chosen female is a sister, then neither get married.
   - If there are more males than females, randomly choose group of males the same size as group of females

4. Choose $X \sim Bin(4, .5)$ children for every couple from (3)

5. Repeat previous steps for 7 generations (not including founding generation)

6. Calculate pedigree-based 2Φ for all individuals [22]

   - Number the people in the pedigree in such a way that every parent precedes his or her children
   - Φ is built recursively by considering $i = 1, \ldots, n$:
   - If the current individual i is a founder then:
     - Set $\Phi_{ii} = 1/2$.
     - For each previously considered person j (i.e. $j < i$), set $\Phi_{ij} = \Phi_{ji} = 0$.
   - If i is not a founder then:
     - Let i have parents k and l.
     - Set $\Phi_{ii} = 1/2 + 1/2\Phi_{kl}$. This is due to the fact that we are equally likely to choose either the same gene twice or both maternally and paternally derived genes once.
     - Set $\Phi_{ij} = \Phi_{ji} = 1/2\Phi_{jk} + 1/2\Phi_{jl}$. This is due to the fact that in this case, we are equally likely to compare either the maternal gene of i or the paternal gene of i to a randomly drawn gene from j.

7. Transform to R

8. Remove founding and first generation from analysis

9. Add $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$ to every point. Variance is dependent on the variance of similar values from Palau dataset. See Figure 2

   - Binned genotype-based R values from Palau dataset
   - Calculated median absolute deviation (MAD) of pedigree-based R values for first six bins from above
   - Ran OLS simple linear regression through the origin on MAD values versus bin number ($\hat{\beta} = .1892$).
   - Extrapolated to get σ_{ij}. Did not add any noise to R values at artificial bound

Ended up with 621 individuals in simulated dataset. Simulated data and noisy data can be found in Figure 3.

It is important to note that our simulated degree of relationship values are pedigree-based. I have previously mentioned that we would only consider using genotype-based values because the error structure is more statistical in nature and because these are what are more commonly used in practice. We are currently working on developing simulations using a technique known as gene dropping [23]. This method constructs marker information by simulating gene flow: genes are "dropped" down a pedigree according to Mendel's laws and realistic levels of recombination. As was mentioned previously, genotype-based degree of relationship values can then be estimated from this simulated marker information using likelihood based approaches.

Methods

Treelets

Suppose we have n individuals whose degree of relationship (R) has been calculated for all pairs. Then, we have a symmetric n x n matrix that we will build our treelets on. In short, the treelet algorithm builds a hierarchical tree by starting at $l = 0$ (bottom of the tree) and combining the two most similar columns (or rows, since it is symmetric) of R into a coarse "sum variable" and a residual "difference variable" by local PCA. Put another way, local PCA finds an angle, θ, such that the covariance between the two chosen variables is 0. The angle chosen, along with the pair of indices form the basis. Here, the sum variable is the projection of the two variables in the main principal direction and the difference variable is the projection of the two in the orthogonal direction. It should be noted that using a fixed $\theta = \frac{\pi}{4}$ is the same as the balanced Haar transform. Only sum variables are considered further as we increase $l = 1, \ldots, n-1$. At every level of the hierarchical tree, nodes are associated with the sum variables. More specifically:

- At level $l = 0$ (the bottom of the tree), let $\mathbf{x}^{(0)} = [r_{0,1}, \ldots, r_{0,n}]$, where $x_k^{(0)} = r_{0,k}$ is the k-th column of the degree of relationship matrix, R. Define the basis at this level, B_0, to be the n x n Identity matrix. Compute the covariance and similarity matrices $\hat{\Sigma}^{(0)}$ and $\hat{M}^{(0)}$. Initialize the set of "sum variables," $S = \{1, 2, \ldots, n\}$.

- Repeat for $l = 1, \ldots, L$:

  - Find α and β that satisfy:

    $$(\alpha, \beta) = \text{argmax}_{i,j \in S} \hat{M}_{ij}^{(l-1)}$$

  - Perform local PCA on α and β. Find the orthonormal rotation of the two dimensional plane spanned by $x_\alpha^{(l-1)}$ and $x_\beta^{(l-1)}$. This is done using a Givens rotation matrix [24],

    $$G_l = I + \Theta(\alpha_l, \beta_l, \theta_l)$$

    where I is the identity matrix and $\Theta(\alpha_l, \beta_l, \theta_l)$ is defined as:

    $$\Theta_{ij} = \begin{cases} cos(\theta_l) - 1 & \text{if } i = j = \alpha_l \text{ or } i = j = \beta_l \\ sin(\theta_l) & \text{if } i = \alpha_l \text{ and } j = \beta_l \\ -sin(\theta_l) & \text{if } i = \beta_l \text{ and } j = \alpha_l \\ 0 & \text{otherwise.} \end{cases}$$

    This transformation corresponds to a change of basis $B_l = B_{l-1} G_l$ and new coordinates $\mathbf{x}^{(l)} = G_l^t \mathbf{x}^{(l-1)}$. Calculate $\hat{M}^{(l)}$ and $\hat{\Sigma}^{(l)}$ accordingly.

  - For ease of notation, assume that $\hat{\Sigma}_{\alpha\alpha}^{(l)} \geq \hat{\Sigma}_{\beta\beta}^{(l)}$. Then, α and β correspond to the first and second principal components, respectively. Set the sum and difference variables as $s_l = x_\alpha^{(l)}$ and $d_l = x_\beta^{(l)}$. Define the scaling and detail functions ϕ_l and ψ_l as columns α and β of B_l, respectively. As is the case with standard wavelet analysis, only sum variables are considered further. In other words, set $S = S \backslash \{\beta\}$.

Once we have constructed an orthonormal basis $B_L = (b_1, \ldots, b_n)$, every row of R, r, can be completely reconstructed as:

$$r = \sum_{i=1}^{n} c_i b_i$$

where $c_i = \langle r, b_i \rangle$. In order to actually smooth the vector, we employ hard thresholding using a universal threshold on the c's (components). In other words,

$$\delta_{\lambda}^{H}(c_i) = \begin{cases} c_i, & \text{if } |c_i| > \lambda, \\ 0, & \text{otherwise}, \end{cases}$$

where $i = 1, \ldots, n$ and λ is some threshold value chosen a priori.

Finally, because relationships are symmetric ($R_{ij} = R_{ji}$) it stands to reason that our smoothed matrix $\tilde{R}$ should be symmetric. Given that we are treating each row individually, there is no guarantee that this will be the case. Therefore, we simply symmetrize it to get $\hat{R} = (\tilde{R} + \tilde{R}^t)/2$.

Similarity Score

In the above treelet algorithm, at each level of the tree building process, we are required to find the two "most similar" variables to combine. A natural choice could simply be the correlation coefficient, ρ_{ij}, between variables x_i and x_j. Unfortunately, this metric is not completely suitable for our purposes due to natural "outliers" in our data. Essentially, by the way we have constructed the data, there will be numerous outliers and influential points between any two given variables. These can be seen by examining Figure 4.

This figure is of two randomly chosen individuals from the simulated dataset. It turns out that $R_{ij} \approx 3$. This means that individual i and individual j are probably first cousins. Each point on the plot represents the degree of relationship of a third individual, k, to both i and j. A few items of note about this plot are the "boundaries" along the right side and top. These points correspond to individuals who are essentially unrelated to individual i (or j) but are related to individual j (or i). This may seem counter intuitive at first, but it makes sense if you imagine family members being related on one side of the family (i.e. in-laws). For example, Figure 5 is a simple pedigree. Without any other knowledge of possible relatives, we can see that individual 1 is related to everyone else, except 6. Similarly, individual 2 is related to everyone else, except 3. These are both because individuals 3 and 6 "married" into this family. Another important subset of points lie on the top right diagonal. These are individuals who are basically unrelated to both i and j. It is clear that most informative points of this plot lie in the mass of individuals who are related to both chosen individuals. The correlation coefficient of the points will not capture this information. For instance, the correlation using all the points is about .469, the correlation among points only among the "good" points is .414 and the correlation after removing the influential points in the top right is only .143. Therefore, it is clear that the correlation coefficient is not a sufficient way of choosing the most similar variables. We will have the same problem when choosing the optimal angle, θ, between any two variables. This is because θ is based on $\hat{\Sigma}_{ij}$, the sample covariance of x_i and x_j.

There have been numerous attempts to try to ameliorate the problem of PCA's sensitivity to outliers. Previous work has centered around using robust estimators of the covariance matrix [25; 26]. Instead of decomposing the original covariance matrix, Σ, they perform an eigenanalysis on a more robust M-estimator, $\tilde{\Sigma}$. Define

$$\tilde{\Sigma} = \frac{\sum_{i=1}^{n} w_i^2(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t}{(\sum_{i=1}^{n} w_i^2 - 1)},$$

and

$$\bar{\mathbf{x}} = \sum_{i=1}^{n} w_i \mathbf{x_i} / \sum_{i=1}^{n} w_i$$

where

$$w_i = \omega(d_i)/d_i$$

and

$$d_i = \{(\mathbf{x}_i - \bar{\mathbf{x}})^t \tilde{\Sigma}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})\}^{\frac{1}{2}}$$

Here, ω is known as a bounded influence function [27] and is defined by the authors to be:

$$\begin{aligned} \omega(d) &= d \text{ if } d \leq d_0 \\ &= d_0 \exp\{-1/2(d - d_0)^2/b_2^2\} \text{ if } d > d_0 \\ \text{where } d_0 &= \sqrt{p} + b_1/\sqrt{2} \end{aligned}$$

The authors provide an iterative procedure for calculating $\bar{\mathbf{x}}$ and $\tilde{\Sigma}$. They also provide practical values for b_1 and b_2. Unfortunately, this method cannot resist many outliers.

Croux and Haesbroeck [28] examined minimum covariance determinant (MCD) methods [29] and other S-estimators [30; 31] to solve this problem. MCD methods calculate the determinant of the sample covariance matrix of every possible subset of size h from the original observations. A robust measure of the MCD covariance is simply the sample covariance of the h-subset that minimizes those determinants. This method can be very computationally intensive in high dimensions. There is also the limitation that the MCD estimator can only be computed when $h > p$. Otherwise, the covariance matrix of any h-subset will be singular. This maximum percentage of points that can be considered outliers and still yield a bounded estimate is known as the breakdown value.

Another approach is to use projection pursuit (PP) ideas when looking for robust PCA [32; 33; 34]. This technique searches for the direction in which the projected observations have the largest robust scale (i.e. eigenvalues). Then, every subsequent step produces directions constrained to be orthogonal to all previous directions. One such algorithm is as follows:

1. Center each data point by its L^1-median, $\hat{\mu}^R$. Here,

   $$\hat{\mu}^R = \text{argmin}_\theta \sum_{i=1}^{n} \|\mathbf{x}_i - \theta\|.$$

   Call these new data points $\mathbf{x}_i^{(1)}$ and this new matrix $\tilde{X}$.

2. Calculate the first eigenvector by:

   $$\mathbf{v_1} = \text{argmax}_{\mathbf{a} \in A_1} Q_n(\mathbf{a}^t \mathbf{x}_1^{(1)}, \ldots, \mathbf{a}^t \mathbf{x}_n^{(1)})$$

   where $A_1 = \{\mathbf{x}_i^{(1)} / \|\mathbf{x}_i^{(1)}\|; i = 1, \ldots, n\}$ and
   $Q_n(x_1, \ldots, x_n) = 2.2219 c_n \{|x_i - x_j|; i < j\}_{(k)}$. Here,
   c_n is some constant and $k = \begin{pmatrix} n \\ 2 \end{pmatrix}$.

3. Calculate $\mathbf{x}_2^{(2)} = (I_{p,p} - \mathbf{v}_1 \mathbf{v}_1^t)\mathbf{x}_i^{(1)}$. This step ensures orthogonality.

4. Repeat steps 1 and 2 until $k \leq r = \text{rank}(\tilde{X})$ eigenvectors are found.

5. Suppose that $\mathbf{v}_1, \ldots, \mathbf{v}_{l-1}$ eigenvectors have been constructed then:

$$\mathbf{v}_l = \text{argmax}_{\mathbf{a}\in A_l} Q_n(\mathbf{a}^t\mathbf{x_1}^{(l)}, \ldots, \mathbf{a}^t\mathbf{x}_n^{(l)})$$

where $A_l = \{\mathbf{x}_i^{(l)}/\|\mathbf{x}_i^{(l)}\|; i = 1, \ldots, n\}$. Also, $\mathbf{x}_i^{l+1} = \mathbf{x}_i^{(l)} - \mathbf{v}_l\mathbf{v}_l^t\mathbf{x}_i^{(1)}$.

6. For all $l = 1, \ldots, r$, the robust scale s_l^R is defined as $s_l^R = Q_n(\mathbf{v}_l^t\mathbf{x}_1^{(l)}, \ldots, \mathbf{v}_l^t\mathbf{x}_n^{(l)})$.

Although this method has been shown to have a higher breakdown value than previous methods and allows for $p > n$, there are still some deficiencies with this approach. Most recently, Hubert et al. [35] developed a technique that utilizes both of the previous ideas. PP is used as an initial dimension reduction technique. MCD is then applied to this lower-dimensional space.

Each of the previous algorithms first attempt to locate outliers that could prove detrimental in PCA. They do this by assigning an "outlyingness" value to every point and then correcting for it. In our case, we are not interested in locating outliers, only how to deal with them. Because we are only dealing with two dimensions at a time it is very easy to see what points are considered outliers. Our main goal is to determine how influential the already chosen outliers are and then figure out how to deal with them in the Treelet framework.

Results of Simulated Data

Results focus on two types of analysis: keeping all the points (scenario A) and not including points where two individuals in question are both unrelated to a third individual (i.e. influential points) in the calculation of pairwise correlations (scenario B). In each case, I examined the overall mean square error (MSE) between the smoothed matrix, $\hat{R}$, and the true R values at varying threshold values, λ. Here, $MSE = 1/(n^2)\sum_{ij}(\hat{R}_{ij} - R_{ij})^2$. I also looked at the MSE in just the "good" points, along with the MSE of points at the artificial boundary. Good points refer to any entries in the matrix where $R < 21$. Boundary points are any entries such that $R = 21$. For comparison, overall MSE of the noisy matrix (before smoothing) is .3729. The MSE of the good points is .5997. Obviously, the MSE at the boundary is 0, since I did not add any error to those terms.

One can see from Table 2 that Treelets do, in fact, remove noise from our relationship matrix. It is also important to point out that scenario B did almost uniformly better than scenario A across both λ and types of matrix entries examined. This is especially true in the boundary points. In other words, by simply removing these extremely influential points from the calculation of pairwise correlations, we were able maintain low overall and good point MSE values while simultaneously lowering the MSE at the boundary points. This is probably due to the algorithm merging variables that are more closely related at every level of the tree. Using all points in the calculation may artificially inflate the correlation coefficient if the two individuals in question are unrelated to nearly everyone else. This is counter intuitive to the natural process of related individuals between related to the same third parties and shows precisely why robust PCA techniques must be incorporated in this instance. Finally, one can see that the MSE is very dependent on the choice of λ. Currently, I have chosen λ in a very ad hoc manner. I would like to have λ chosen automatically and make it row (or column) of R dependent. This is because I am treating every row of R as its own signal.

Another important feature of Treelets is building the hierarchical tree from the variables. In order for the algorithm to work properly it should build the tree in a similar fashion to a pedigree. In other words, closely related individuals should be clustered together and visa versa. We start by examining the highest energy Treelet basis vectors across the (ordered) individuals. Suppose we have n individuals, and thus, have $r^1, \ldots, r^n$ rows of our degree of relationship matrix. Here, $r^i \in \mathbf{R}^n$. The Treelet algorithm will produce an orthonormal basis $B = (b_1, \ldots, b_n)$. We can then assign a normalized energy score, ϵ to each vector b_i by:

$$\epsilon(b_i) = \frac{\sum_{j=1}^{n} |b_i \cdot r^j|^2}{\sum_{j=1}^{n} \|r^j\|^2}$$

Sort the vectors according to decreasing energy. Figure 6 shows the 5 highest energy Treelets. We can see that the second, third and fourth capture the localized nature of closely related family members. The highest energy Treelet corresponds to the scaling function and thus, should not be considered further as we are primarily interested in the detail functions. In fact, we can see exactly how the individuals within the Treelet "peaks" merge both in the algorithm and in the actual pedigree. Figures 8-11 display how the individuals from the maximal energy Treelet "peaks" relate within the overall (known) pedigree. Filled in cells are the individuals clustered by Treelets. To add to the overall picture, Figure 7 displays the dendrograms of the chosen individuals within the hierarchical tree. We can see that for the most part, we are doing an excellent job grouping siblings and parents together and a pretty good job clustering the more extended family structure.

Proposed Work

Thresholding

I need to determine a more appropriate way to threshold expansion coefficients.

- λ should be row-dependent on R
- FDR approach[36]: Authors propose to exploit sparsity by controlling the false discovery rate of possible non-zero expansion coefficients.
- Bootstrap approach[37]: Authors propose bootstrapping original matrix R and examining distribution of expansion coefficients when there is no structure
- Best basis selection[38; 39; 18]: Instead of running Treelet algorithm on all $L = n - 1$ levels and dealing with basis B_L, want to choose a basis, B_k where $k < n - 1$ is chosen to maximize (or minimize) some energy or entropy criterion. May want to consider only scale functions left at that level.

Using Robust PCA Techniques in Treelet Algorithm

We would like our algorithm to be more robust to outlying points (as in Figure 4). This is especially true when calculating the similarity matrix, $\hat{\Sigma}$, to be used for merging individuals. One possibility is to use weighted correlation and/or covariance. In other words, measure the angular distance between points on the boundary and the line that best rotates the "good points". Let θ_i be the angle created by point i and let $\hat{\theta}$ be the angle of the line that best rotates these good points. Set $w_i = \frac{|\theta_i - \hat{\theta}|}{c}$. Here, $\theta < \frac{\pi}{4}$ and c is some normalizing constant. Then, one similarity measure could be:

$$M_{XY} = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x^*})(y_i - \bar{y^*})}{s_X^* s_Y^* \sum_{i=1}^{n} w_i}.$$

Here, $\bar{x}^*$ and $\bar{y}^*$ are the weighted means. Similarly, s_X^* and s_Y^* are the weighted standard deviations. Points not on the boundary are assigned a weight of 1 and points that are unrelated to both individuals would have weights close to 0.

References

[1] Li CC. Population subdivision with respect to multiple alleles. *Annals of Human Genetics* 1972; **33**:23–29.

[2] Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994; **265(5181)**:2037–2048.

[3] Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics* 1987; **57**:455–464.

[4] Schaid DJ, Sommer SS. Genotype relative risks: Methods for design and analysis of candidate-gene association studies. *American Journal of Human Genetics* 1993; **53**:1114–1126.

[5] Schaid DJ, Sommer SS. Comparison of statistics for candidate-gene association studies using cases and parents. *American Journal of Human Genetics* 1994; **55**:402–409.

[6] Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American Journal of Human Genetics* 1993; **52**:506–516.

[7] Lee AB, Luca D, Klei L, Devlin B, Roeder K. Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology* 2010; **34**:51–59.

[8] Crossett A, Kent BP, Klei L, Ringquist S, Trucco M, Roeder K, Devlin B. Using ancestry matching to combine family-based and unrelated samples for genome-wide association studies. *Statistics in Medicine* In Press; .

[9] Devlin B, Klei L, Worsley MM, Tiobech J, Otto J, Byerley W, Roeder K. Genetic liability to schizophrenia in oceanic palau: a search in the affected and maternal generation. *Human Genetics* 2002; **121**:675–684.

[10] Choi Y, Wijsman EM, Weir BS. Case-control association testing in the presence of unknown relationships. *Genetic Epidemiology* 2009; **33**:668–678.

[11] Thornton T, McPeek MS. Roadtrips: Case-control association testing with partially or completely unknown population and pedigree structure. *American Journal of Human Genetics* 2010; **86**:172–184.

[12] Malecot G. *Les Mathematiques de l'Heredite*. Paris: Masson, 1948.

[13] Thomas DC. *Statistical methods in genetic epidemiology*. Oxford University Press, 2004.

[14] Thompson EA. The estimation of pairwise relationships. *Annals of Human Genetics* 1975; **39**:173–188.

[15] McPeek MS, Sun L. Statistical tests for detection of misspecified relationships by use of genome-screen data. *American Journal of Human Genetics* 2000; **66**:1076–1094.

[16] Millilgan BG. Maximum-likelihood estimation of relatedness. *Genetics* 2003; **163**:1153–1167.

[17] Lynch M, Ritland K. Estimation of pairwise relatedness with molecular markers. *Genetics* 1999; **152**:1753–1766.

[18] Lee AB, Nadler B, Wasserman L. Treelets - an adaptive multi-scale basis for sparse unordered data. *Annals of Applied Statistics* 2008; **2**:435–471.

[19] Golub G, van Loan CF. *Matrix Computations*. 3 edn., Johns Hopkins University Press, 1996.

[20] Murtagh F. The haar wavelet transform of a dendrogram. *Journal of Classification* 2007; **24**:3–32.

[21] Singh A, Nowak R, Calderbank R. Detecting weak but hierarchically-structured patterns in networks. *Artificial Intelligence and Statistics* In Press; .

[22] Lange K. *Mathematical and statistical methods for genetic analysis*. Spring: Statistics for Biology and Health, 1997.

[23] Terwilliger JD, Speer M, Ott J. Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genetic Epidemiology* 1993; **10**:217–224.

[24] Givens W. Computation of plane unitary rotations transforming a general matrix to triangular form. *Journal of the Society for Industrial and Applied Mathematics* 1958; **6(1)**:26–50.

[25] Maronna RA. Robust m-estimators of multivariate location and scatter. *The Annals of Statistics* 1976; **4**:51–67.

[26] Campbell NA. Robust procedures in multivariate analysis i: robust covariance estimation. *Applied Statistics* 1980; **29**:231–237.

[27] Hampel FR. The influence curve and its role in robust estimation. *Journal of American Statistics* 1974; **69**:383–393.

[28] Croux C, Haesbroeck G. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika* 2000; **87**:603–618.

[29] Rousseeuw PJ. Least median of squares regression. *Journal of the American Statistical Association* 1984; **79**:871–880.

[30] Davies L. Asymptotic behavior of s-estimators of multivariate location and dispersion matrices. *The Annals of Statistics* 1987; **15**:1269–1292.

[31] Rousseeuw PJ, Leroy A. *Robust regression and outlier detection.* New York: Wiley, 1987.

[32] Li G, Chen Z. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. *Journal of the American Statistical Association* 1985; **80**:759–766.

[33] Croux C, Ruiz-Gazen A. A fast algorithm for robust principal components based on projection pursuit. *Proceedings in Computational Statistics* 1996; :211–217.

[34] Hubert M, Rousseeux PJ, Verboven S. A fast method for robust principal components with application to chemometrics. *Chemometrics and Intelligent Laboratory Systems* 2002; **60**:101–111.

[35] Hubert M, Rousseeuw PJ, Vanden Branden K. Robpca: a new approach to robust principal component analysis. *Technometrics* 2005; **45**:301–320.

[36] Abramovich F, Benjamini Y, Donoho DL, Johnstone IM. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics* 2006; **34(2)**:584–653.

[37] Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genetics* 2007; **3(9)**:1724–1735.

[38] Coifman R, Wickerhauser M. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory* 1992; **38(2)**:713–718.

[39] Saito N, Coifman R. On local orthonormal bases for classification and regression. *IEEE Signal Processing Society* 1995; **35**:2841–2852.

Relationship	Φ
Ind.-Themself	1/2
MZ. Twins	1/2
Parent-Off.	1/4
Full Sibs.	1/4
Half Sibs.	1/8
First Cousins	1/16
Uncle-Nephew	1/8
Unrelated	0

Table 1: Pedigree-based kinship coefficients

	MSE					
	Overall		Good pts.		Boundary	
λ	A	B	A	B	A	B
0	.3729	.3729	.5997	.5997	0	0
.5	.3600	.3593	.5746	.5743	.0073	.0058
1	.3084	.3026	.4807	.4802	.0249	.0106
1.5	.2645	.2492	.3900	.3884	.0581	.0205
2	.2649	.2450	.3656	.3687	.0995	.0416
2.5	.3215	.2957	.4082	.4228	.1788	.0868

Table 2: MSE value of all points of R (Overall), only points where $R < 21$ (Good pts.) and points where $R = 21$ (Boundary). A corresponds to running Treelet algorithm on all data points and B corresponds to running Treelet algorithm after throwing away points where two individuals in question are both unrelated to third individual

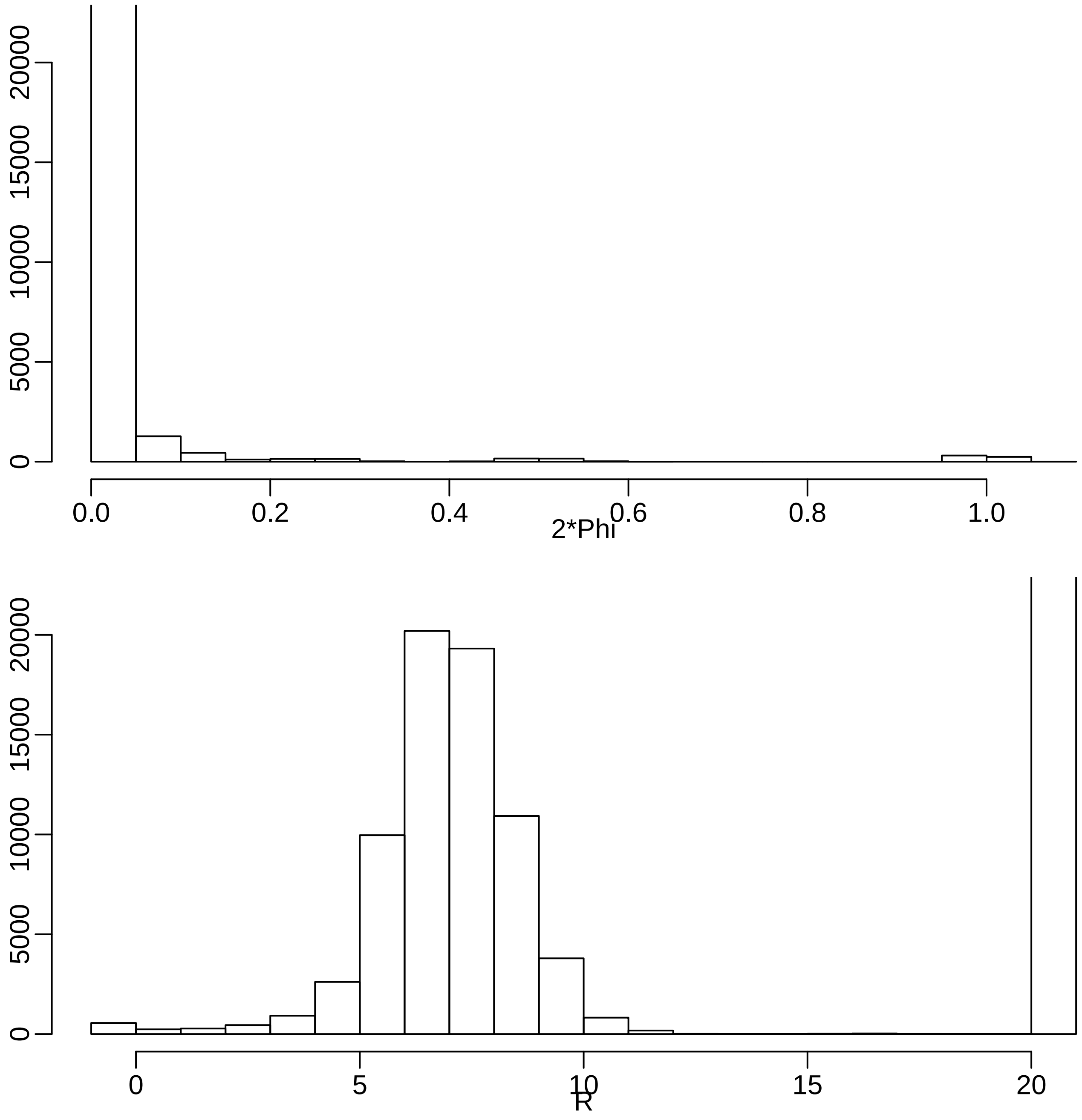

Figure 1: Top: Upper triangle of genotype-based 2Φ. Bottom: Upper triangle of R. The truncated cell, capped at 25000 observations, actually includes...

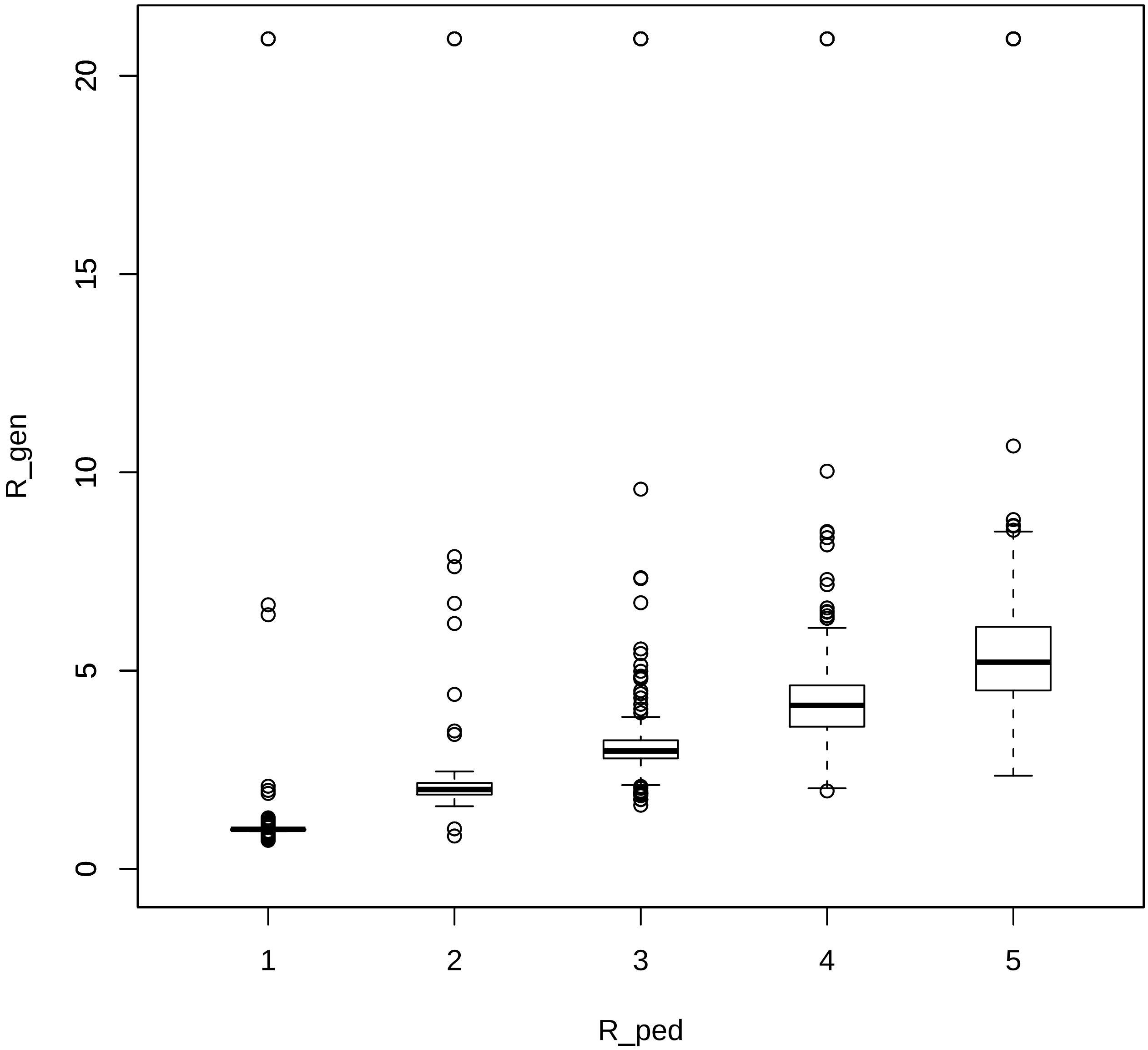

Figure 2: Pedigree-based R values versus binned genotype-based R values from Palau dataset

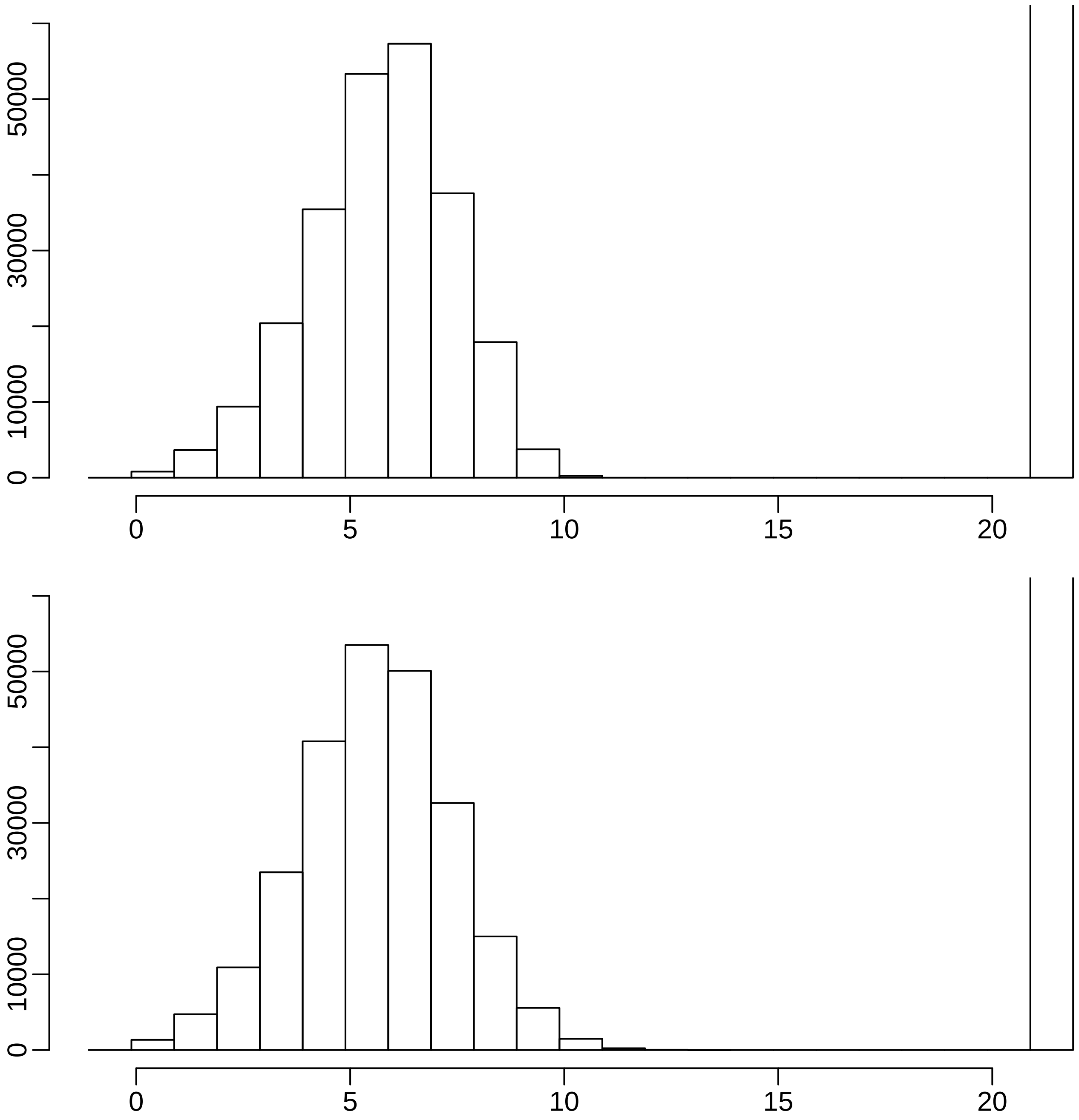

Figure 3: Top: R values of 621 simulated individuals. Bottom: Top, but with added noise. Noise based on (9) of Simulations.

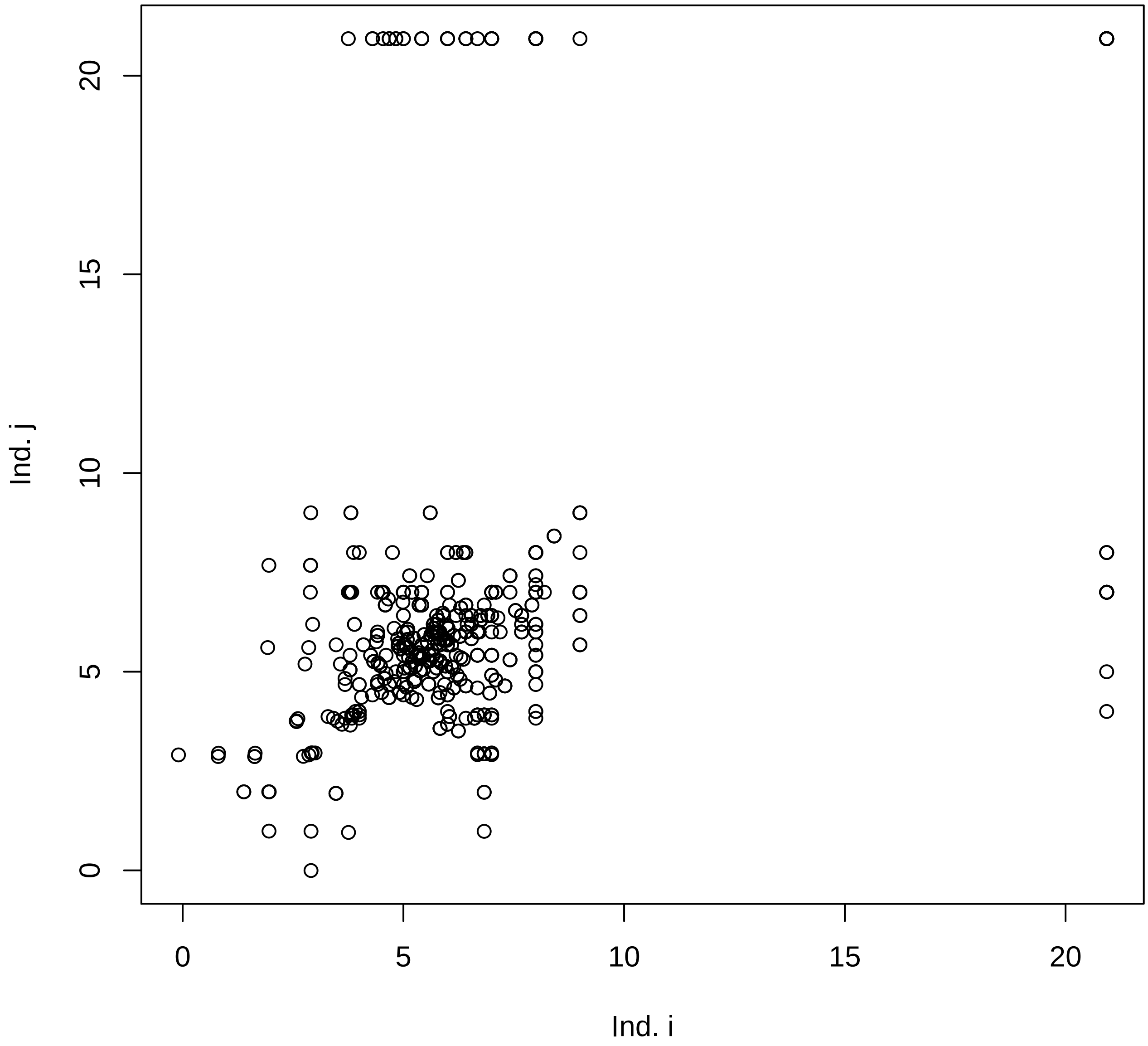

Figure 4: Plot of two randomly chosen columns from simulated R.

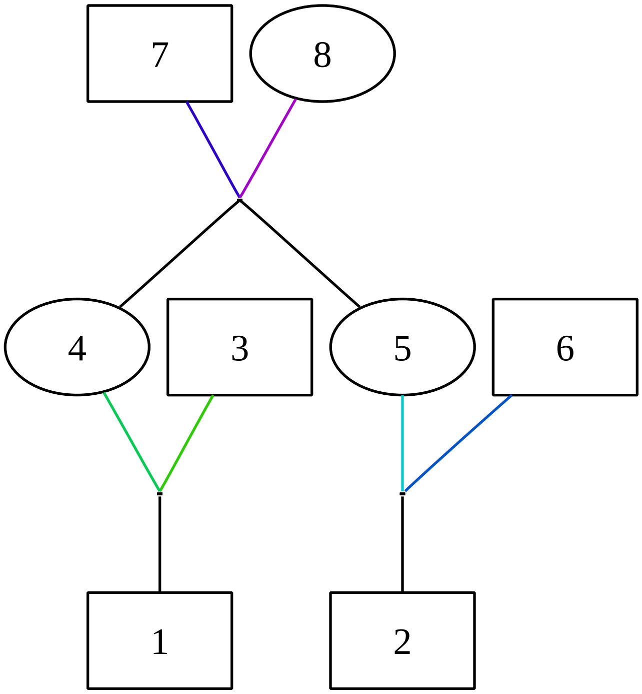

Figure 5: Simple pedigree of a family of 8 individuals.

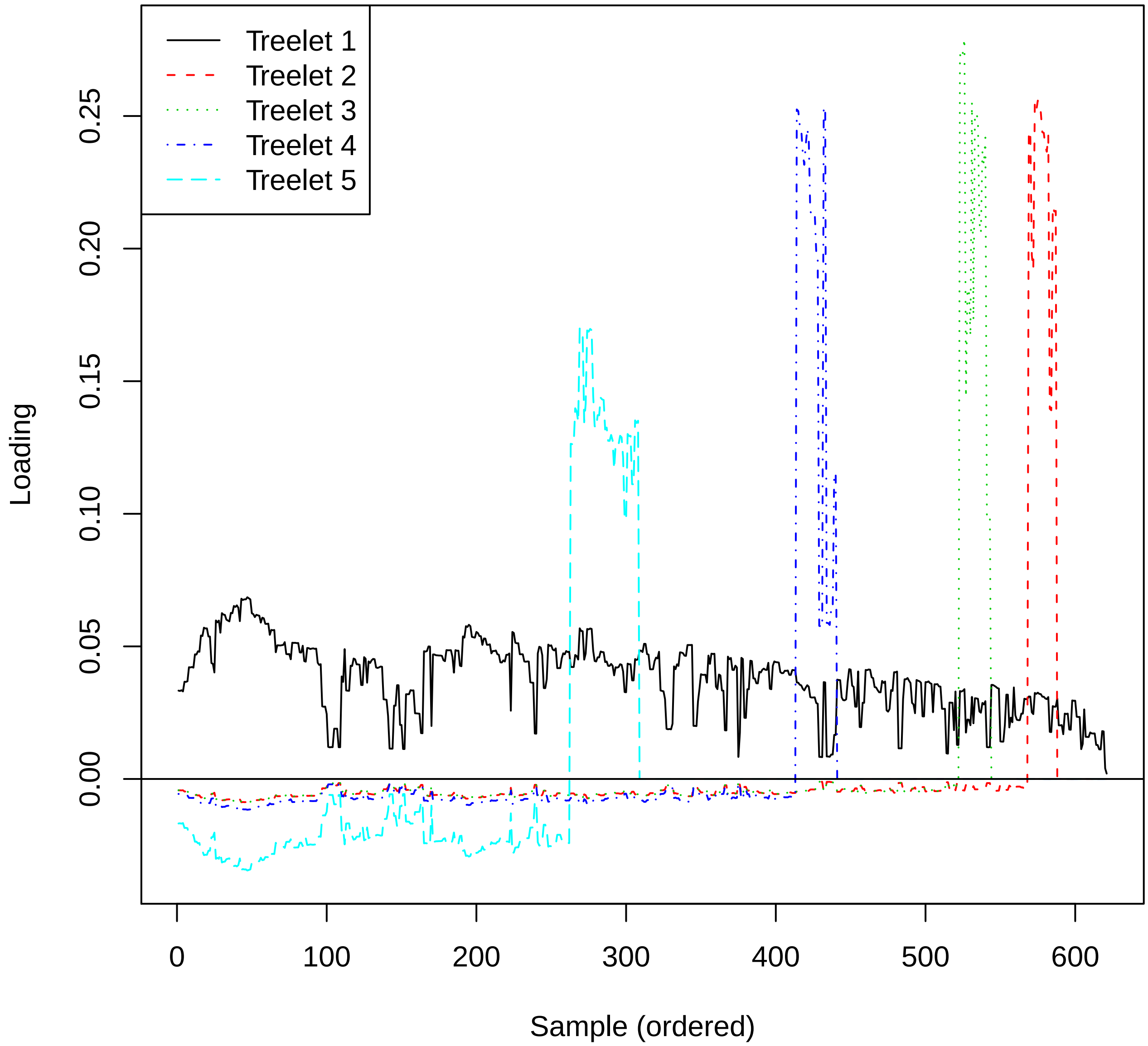

Figure 6: Plot of maximal energy treelets. Y-axis are the normalized energy scores and the X-axis are the ordered individuals (based on how hierarchical tree was built)

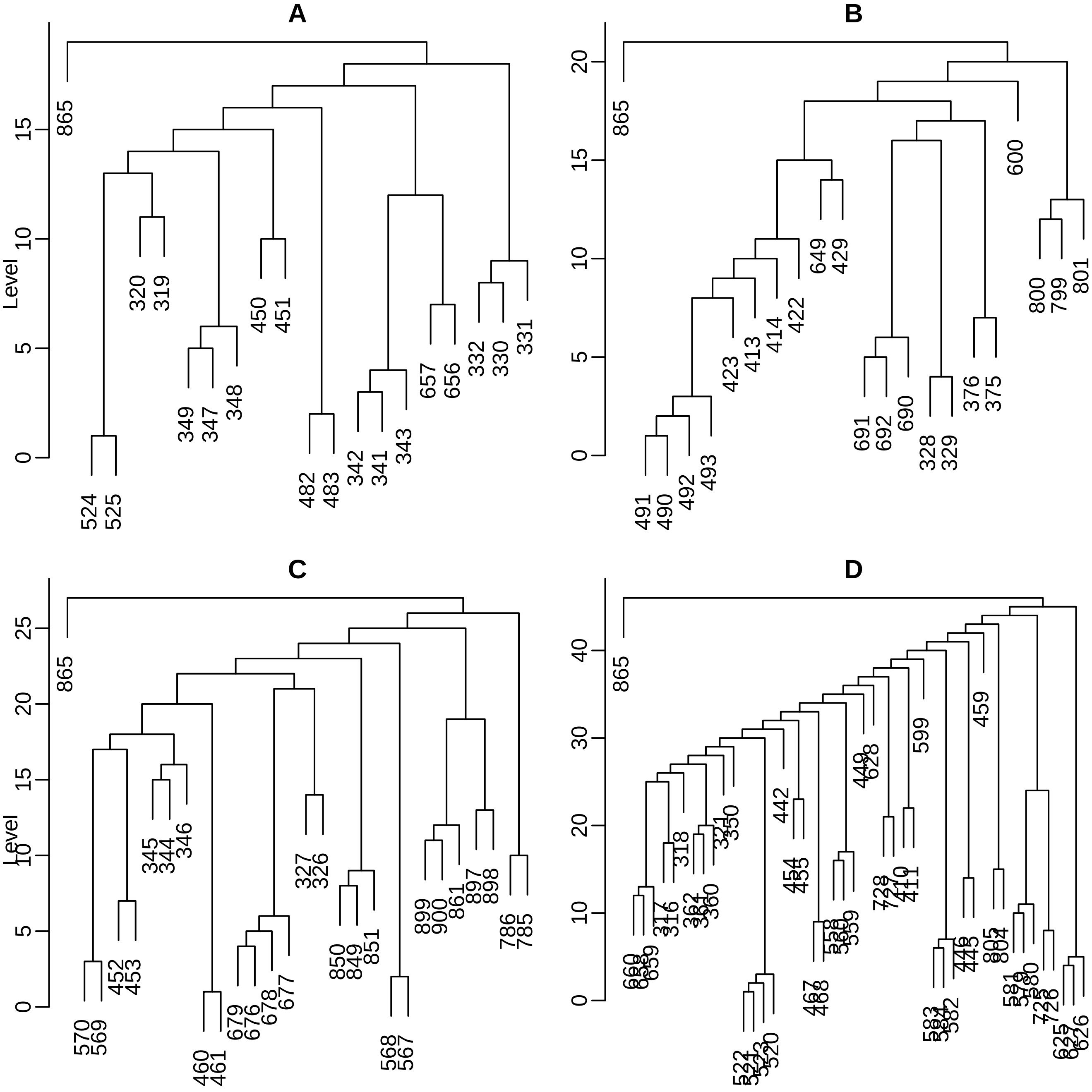

Figure 7: Snapshots of dendrogram produced by treelet algorithm. A is of individuals chosen by second highest energy treelet. B is of third highest. C is of fourth highest. D is of fifth highest. Y-axis is the relative level that individuals were merged.

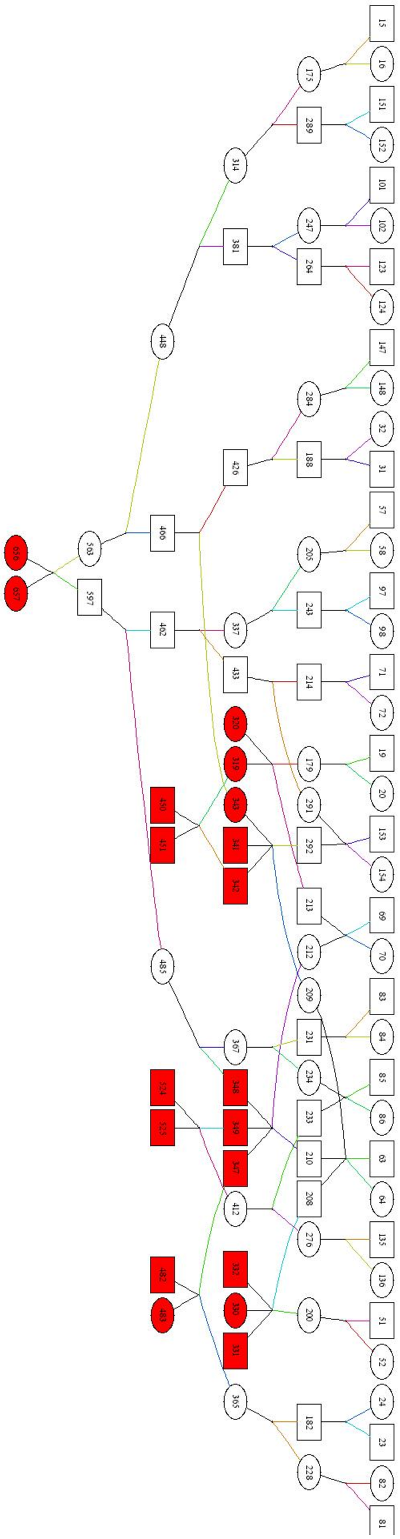

Figure 8: Snapshot of actual pedigree relating to individuals chosen from second highest energy treelet. Filled in cells are the individuals within that specific cluster.

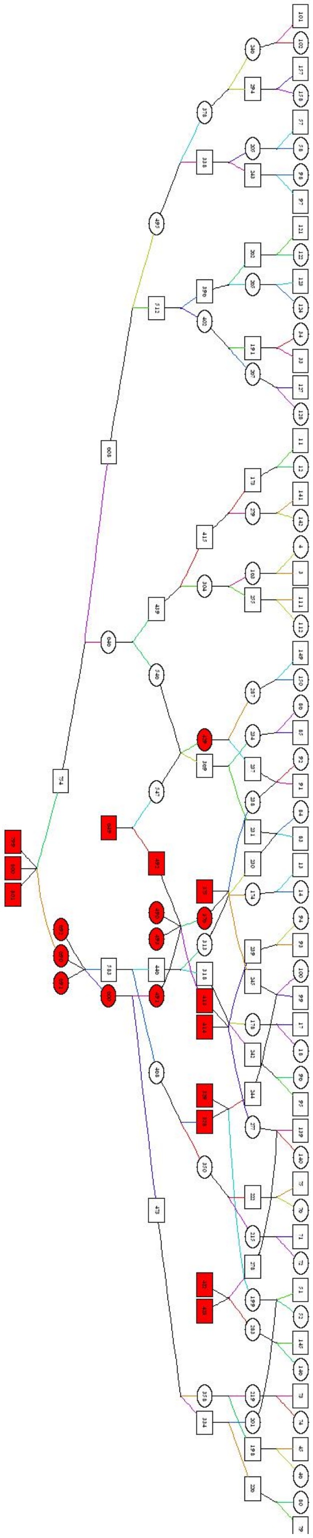

Figure 9: Snapshot of actual pedigree relating to individuals chosen from third highest energy treelet. Filled in cells are the individuals within that specific cluster.

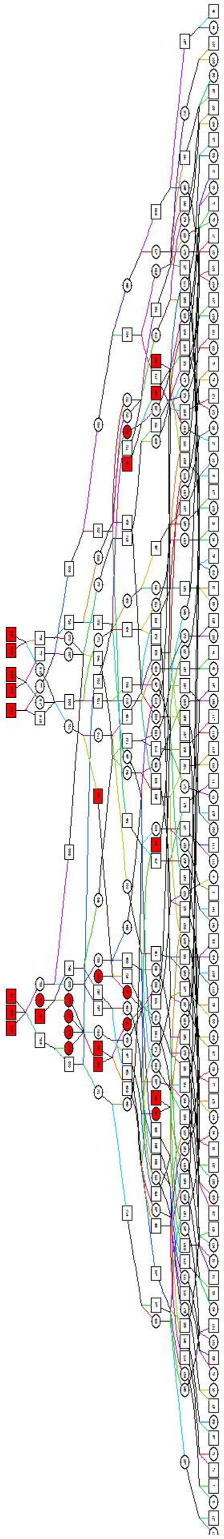

Figure 10: Snapshot of actual pedigree relating to individuals chosen from fourth highest energy treelet. Filled in cells are the individuals within that specific cluster.

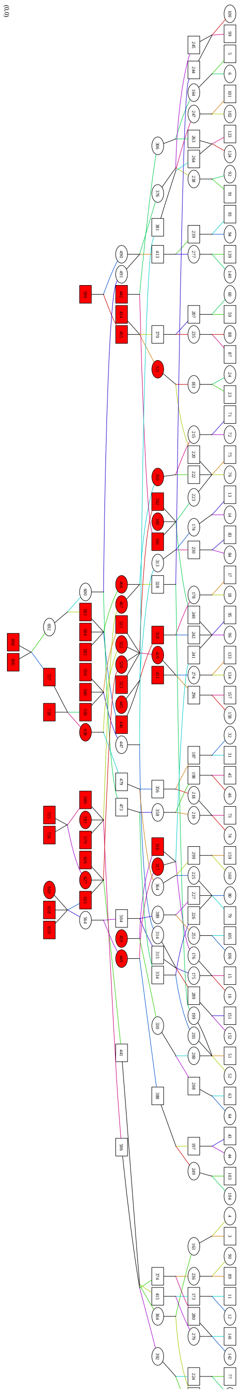

Figure 11: Snapshot of actual pedigree relating to individuals chosen from fifth highest energy treelet. Filled in cells are the individuals within that specific cluster.