Combining Genetic Similarities Among Known Relatives that Connect to an Unknown Relative

Stephen P Smith
email: hucklebird@aol.com

Cambrian Lopez

Nicole Lam
Kaiser Permanente Labor & Delivery
KSDHCPA (UNAC/UHCP) Hospital President
email: Nicolelam@att.net

May 2017

Abstract. Various DNA testing companies promise their customers a collection of genetic matches to facilitate finding family members. The matches are in centimorgans (cM), where the higher the cM value the closer the relationship to a customer (R). Unless the relationship is close, such as parent-offspring or among 1st cousins, a single cM value is not that informative if the goal is to locate family. This paper describes a statistical method that combines a collection cM values from a cluster of unknown relatives of R, but where the cluster members are known among themselves being for example 2rd and 3th cousins. A presumed envoy is attached to the cluster, where R is a descendant of the envoy, and the various cM values are combined to provide an overall cM value between R and the envoy. The envoy's cM comes with a statistical error to judge significance. Unlike a single cM value on a typical unknown relative, the envoy's cM can be quite large and indicative of a real genetic path to R that has previously been undiscovered. This paper describes the method for two sisters, where the path from the envoy led to their lost father, a father that was later discovered.

1. Introduction

When the computer age augmented itself and led to the internet, the study of genealogy benefitted tremendously among many other fields of study. Moreover, while these technological developments occurred, the fields of genetics, molecular biology and computational science produced many innovations that synergistically benefitted each other, as well as advancing genealogy research further. Today, the public can join various private companies, such as Ancestry, My Heritage, 23andMe, My Family Tree DNA, and have their DNA sampled and tested as an aid to genealogic research that is now conducted by the public at large.

Once DNA is sampled, it is put through laboratory analysis, then the results are put through computer algorithms to produce matches with others that are part of the respective databases kept by the various organizations. The costumers then receive a collection of DNA matches with others that indicate a possible relationship, like 2nd cousin, 2nd cousin 1x removed, 3rd cousin, distant cousin, etc. The strength of the match is evaluated in terms of pair-wise comparisons: the shared total centimorgans (cM)[1] found and the number of DNA segments involved where the sharing occurred.

A customer reviewing the reported matches can anticipate disappointment, however, because mere matches absent other supporting evidence provides little except in the extraordinary matches found among immediate family that had been lost. In a website footnote, Ancestry indicates that cM for any 3rd cousin can vary between 90 and 180, and for 4th cousins it can vary between 20 to 85. However, Bettinger (2016, page 106) reports wider variation where ranges overlap significantly. Typical variation can be defined precisely as the range provided by the 10th and 90th percentiles, however, based on 36 measurements[2] on known relatives the ranges were much lower than the expectation set by Ancestry: where the cM varied between 16 and 117 for 3rd cousins, and between 6 and 29 for 4th cousins. The proposed relationship that is supplied by the vendor can tend to be conservative for several reasons, e.g., penalizing the probability of a false positive much more than a false negative, or because the calculation of shared cM is directly impacted from efforts meant to err on the conservative side. In any regard, standards that characterize the random behavior of cM measurements are not well described, in part because testing companies use different methods to measure shared cM.

It is not untypical for new technology to disappoint because methodology is missing or is not perfected or because additional research is neglected. It is now advertised that the best use of DNA matches is to supplement existing genealogical findings. The findings that are most useful are those that are summarized as an existing family tree that includes branches for collateral relatives that may act to conduit possible genetic matches. Geneticists refer to this information as pedigree information. Presently, most customers that take DNA tests provide no pedigree information, and they will likely be disappointed with the matches they receive unless they can match with someone that can supply the missing pedigree information. Even if an existing pedigree finds a possible 3rd or 4th cousin, that comes with DNA confirmation, if the confirmation only involves one cM measurement the situation is less than perfect because of the innate cM variation hinted at above. For 5th cousins, or distant cousins, the cM's are expected to fall off dramatically, and even less utility is found with genetic matches that come as only one cM measurement. Unexplained large cM's can also be found, however,

---

[1]Wikipedia (see https://en.wikipedia.org/wiki/Centimorgan) provides an adequate introduction.

[2]An accounting made by the senior author.

presumably because of deep relationships that did not fall off as abruptly as anticipated, for whatever reason that the DNA is identical by state rather than descent.

In this paper, total shared cM is adopted as the measure of genetic similarity (inversely related to distance), and the following definitions are made. A known relative is someone that can be placed in a supplied pedigree. A collection of known relatives belong to the one pedigree. The unknown relative is someone that might connect to the pedigree through a hypothetically placed envoy. The unknown relative is a presumed descendant of the envoy. The envoy is the immediate offspring of the central ancestors (husband and wife) in the pedigree that have many living descendants identified. Each central ancestor has a mother and father, identified as the paternal common ancestors and the maternal common ancestors. The pedigree information that is to be used will represent all the descendants of the maternal and paternal common ancestors so defined, i.e., beyond the descendants of the central ancestors. Deeper genetic relationships are to be ignored, but in theory could be included. The living descendants of the paternal and maternal common ancestors are mapped out as known relatives in the provided pedigree. Some of the known relatives took DNA tests and matched with the unknown relative. This paper will describe a statistical method to combine all those cM values, into one estimate that can be assigned to the envoy that comes with a standard error. The overall fit can also be judged by a chi-square statistic that also informs on the innate variation of the cM values.

It will be demonstrated that combining cM values from known relatives can provide incontrovertible evidence of descent from the central ancestors, whereas no such proof comes from one cM unless it is between immediate family.

2. Data Requirements

Taking a DNA test that returns a set of unspecified matches is a necessary starting place, given that there can be no known relatives without first establishing an unknown relative that stands opposed to a collection of known relatives that belongs to one pedigree. It is necessary for the unknown relative to happen upon the set of relatives known to be placed in one pedigree, and initially this search has the accuracy of a scatter gun. Getting siblings, half-siblings and 1st cousins to take the same DNA test helps sharpen the focus of the search by eliminating possibilities and allowing more comparisons involving shared matches. Connecting with immediate relatives helps confirm which parts of a family tree are better known and come with well-researched genealogy, and which part of the tree remains open to discovery.

Downloading raw DNA results, and uploading the results to a service like My Family Tree DNA, can broaden the search because such searches are limited by the size of the respective databases.

Having happened upon a set of relatives that belong to one pedigree, what can be

observed is that the matches will cluster in groups as is apparent when viewing shared matches. However, this observation cannot be made by the unknown relative. It can only be made by the knowledgeable steward of the pedigree information, that also knows a-prior which known relative in the pedigree took the DNA test, and notes the fresh observation that the unknown relative is found matching with most of the known relatives that took the DNA test. In other words, the unknown relative must find and ask the knowledgeable steward of the outside pedigree for help.

Only now can the data requirements that meet statistical standards be specified, because those requirements describe the three clusters of information first observed by the knowledgeable steward: that the unknown relative is found matching strongly with descendants of a pair of central ancestors (Cluster 1); that the unknown relative is also found matching with descendants of siblings of the paternal central ancestor (Cluster 2); and that the unknown relative is found matching with descendants of siblings of the maternal central ancestor (Cluster 3). It must be possible to identify enough descendants belonging to the three clusters that took the DNA test, whether or not they match with the unknown relative. Those DNA tests must be in sufficient numbers and excluding no results (e.g., a non-match is a real data point), like 4 or 5 tests for each of the three clusters. Lastly, the match information as genetic similarity (cM) to the unknown relative is recorded for each of the known relatives that took the DNA test, ideally recording close to 15 observations and excluding no results.

There is a forth cluster of cMs that ads to the noise, those coming from the deeper common ancestors beyond the central ancestors, or their parents. It is theoretically possible to incorporate this information in a more sophisticated statistical analysis, but this is ambitious and well beyond the scope of the present paper.

Having serendipitously found possible common ancestors, the central ancestors in someone else's family tree, one has to consider the data requirements that have been adopted by the knowledgeable steward in building the outside family tree, otherwise the pedigree information is being taken for granted. The knowledgeable steward made available a rich family tree. The rich family tree has to go back one generation on the central ancestors to connect to their parents, and then forward as many generations as feasible while collecting information on all collateral relatives. The rich family tree may include all know ancestors (of the steward) going back into antiquity, but include enough collateral relatives to keep track of all the steward's 3rd cousins and possibly stopping with the generation that fought Word War II to avoid privacy concerns. Most people that take DNA tests do not provide a rich family tree. Therefore, it was necessary for the knowledgeable steward to build a rich family tree, to identify known relatives that took DNA tests even when most of the known relatives do not provide much of a family tree. A rich family tree is very untypical of what is actually provided by people that take DNA tests. Nevertheless, this requirement does not go away. If a rich family tree does not exists then it must be built, otherwise the three clusters will never be recognized and what is found only remains as a scatter shot of DNA matches.

3. Statistical Model

The remarkable observation is that the cM measurements follow a pattern of inheritance, a pattern that is different to that described by quantitative geneticists for additive genetic effects, but a pattern nevertheless between parent and offspring. Like the pattern found for additive genetic effects, that pattern found for the cM measurements allows the specification of a linear model that permits the best linear unbiased prediction (BLUP) of cM measurements for all the relatives in the pedigree by combining the observed cM measurements found on some of the living relatives. This prediction includes the envoy, and it comes with a standard error to judge significance. Just like dairy cattle are "blupped" to predict breeding values to aid selective breeding, dead people in the pedigree are blupped to predict the cM of the envoy thereby possibly proving that the unknown relative descended from the central ancestors. We are spared from having to exhume skeletons from graves long ago sealed, and performing DNA tests on the bones to prove paternity, even if its known what closet the skeletons are buried in.

The pattern of inheritance for the cM measurements fall into three categories, defined below.

A. From one parent to an offspring, with the parent not a common ancestor or central ancestor.

If u_P is the cM measurement between any parent, identified as P, and the unknown relative R, then let u_O be the cM measurement between that parent's offspring, identified as O, and relative R. Moreover, define Pr(P = R) as the probability[3] that a random gene taken from P (at a given loci) is identical by decent to a random gene taken from the relative R (at the same loci) that is now assumed to pass through the envoy by following a stipulated path. For relationships removed from immediate family, the expectation of u_P is approximately 6800×Pr(P = R) .

It is apparent with meiosis and crossovers (i.e., genetic recombination) that half of the parents genes will be passed on to the offspring, implying that

$$u_O = \tfrac{1}{2}u_P + \epsilon_O \quad (1)$$

where ϵ_O is a random residual, with a variance that must be approximated. If u_O is distributed as a Poisson distribution with mean parameter $\tfrac{1}{2}u_P$, then the variance is well approximated as 3400×Pr(P = R). This is a good variance to use as an approximation because it tends to be a small variance relative to what is typically observed, and this tends to create a sensitive goodness of fit test that points to a poor statistical fit with the

[3]This probability is an element of the numerator relationship matrix that can be computed by following known recursion formulae (Van Vleck, 1979, pg 35).

slightest departure from model expectations, pointing again at extra-Poisson variation that can be measured and then used to estimate statistical errors. The extra variation is merely tacked on at the end of computation.

More realistically, if u_O is distributed as an approximate binomial distribution of N_e effective DNA segments of equal length αcM, each segregating with the binomial probability ½, then u_O has mean $\frac{1}{2}\alpha N_e$ and variance $\frac{1}{4}\alpha^2 N_e$. Because α is defined such that $\alpha N_e = u_P$, we find that the variance is approximated as $\frac{1}{4}\alpha u_P$ or $\alpha 1700 \times Pr(P = R)$. It will be assumed that α is approximately constant for different variations of N_e and u_P., i.e., DNA fragments that segregate independently tend to be the same length, with some variation that can be ignored.

The means for Poisson and binomial distributions are identical. The difference in the variance for the Poisson and binomial distributions comes from a proportionality constant $\frac{1}{2}\alpha$ (a distinction that only repeats in the categories B and C that follow), and as such $\frac{1}{2}\alpha$ has no effect in the calculation of the linear predictions. The extra-Poisson variation can be estimated from the chi-square statistic that is calculated following linear prediction and tacked on at the end to compute standard errors[4], and so there is no need to further consider the binomial distribution as a special case.

However, rather that stopping with the variance approximation noted above for the Poisson distribution, it can be advantageous to seek a better estimate of variance once there has been the initial round of linear prediction. An additional improvement is found by using the fresh prediction of u_P, say $\hat{u}_P$, to approximate for the variance of ϵ_O with $\frac{1}{2}\hat{u}_P$, which is a variance conditional on $u_P = \hat{u}_P$ for a Poisson distribution. Similar improvements are found with categories B and C that follow. What is discovered is that the variances can be better approximated by plugging the linear predictions in as prescribed, and then continuing to a second round of prediction. If the chi-square goodness of fit statistics falls then this iteration is recommended. The procedure now is to continue iterating, at each round plugging the linear predictions in for variances. This is called re-weighted iteration, and it is continued while the chi-square statistic stabilizes. The calculation of the extra-Poisson variation is moved to the end of re-weighted iteration.

B. From common ancestors, or central ancestors, to an offspring, when the offspring is not one of the central ancestors.

There is a need to consider the case when one parent is known (as in category A above), and when two parents are known. However, in the case where one parent is not related to the unknown relative, it might as well be assumed that the second parent is unknown and restrict most of the statistical treatments to follow category A, thereby

[4]A similar adaptation was presented by Breslow (1984).

being more frugal with the numerical calculations. This shortcut[5] can be taken for most of the pedigree except for the case where the parents are central ancestors or the paternal and maternal common ancestors.

As long as the flow of genes that are common to the unknown relative flow down from the two parents (P1 and P2) to an offspring (O), we can apply model (1) to represent the uniting gametes from the two parents. Any allele that is common to the unknown relative can only occupy one loci across both parents, and hence the common genes are passed on independently with (1) applied twice to give model (2).

(2) $$u_O = \tfrac{1}{2} u_{P1} + \tfrac{1}{2} u_{P2} + \epsilon_O$$

The term ϵ_O is again the residual, but now with a variance that can be approximated by:

$\text{Variance}(\epsilon_O) \approx 3400 \times [\Pr(P1 = R) + \Pr(P2 = R)]$ for the first round,
or
$\text{Variance}(\epsilon_O) \approx \tfrac{1}{2} \times [\hat{u}_{P1} + \hat{u}_{P2}]$ during re-weighted iteration.

C. From a paternal (or maternal) central ancestor back against the flow of genes to the paternal (or maternal) common ancestors.

The paternal common ancestors are the *parents* for the male central ancestor, and the maternal common ancestors are the *parents* for the female central ancestor. Here gene flow is reversed to place the common genes, i.e., found identical in the unknown relative, that are also in the central ancestors, but now finding them into the noted parents. This makes two equations given by (3) for the parents rather than one for the offspring, and done for both the paternal and maternal sides of the central ancestors.

(3) $$u_{P1} = \tfrac{1}{2} u_O + \epsilon_{P1}$$
$$u_{P2} = \tfrac{1}{2} u_O + \epsilon_{P2}$$

Every common allele (at a particular loci) found in P1 is an allele missing in P2, and visa versa. Therefore ϵ_{P1} and ϵ_{P2} have a perfect negative correlation, and the associated 2×2 variance-covariance matrix is a rank-1 matrix, approximated by the following.

$$Var\begin{bmatrix} \varepsilon_{P1} \\ \varepsilon_{P2} \end{bmatrix} = \begin{bmatrix} v & -v \\ -v & v \end{bmatrix}$$

[5]Look ahead to Display 1 for example.

Where

$v = 3400 \times Pr(O=R)$ for the first round,
or
$v = ½ \hat{u}_O$ during re-weighted iteration.

There are now equations and residuals coming form (1), (2) and (3) for all individuals in the pedigree that is to be analyzed, excluding the central ancestors and the envoy, and the envoy's descendants leading to the unknown relative. This information can be collected and expressed in matrix notation as follows

$\mathbf{Pu} = \epsilon$

$Var(\epsilon) = \mathbf{R}$

Where **P** is a rectangular matrix with two more columns than rows, with each row representing an equation of the form (1) or (2) or where two rows are given by (3), where most elements in any row are set to zero expect for the numbers 1 or -½ that are found at the appropriate places. The column vectors **u** and ϵ represent the sets of shared cM values and residuals for the known relatives. The variance matrix **R** is almost completely diagonal, except for two 2×2 blocks that correspond to the paternal and maternal common ancestors. The fact that **R** is rank deficient is to be treated correctly with the matrix tools that are described below. Because the cM values for the common ancestors only impact the observed values as the sum of the paternal common ancestor cMs, or the sum of the maternal common ancestor cMs, there is no loss of degrees of freedom caused by the rank deficiency of **R**. The fact that there are no equations for the central ancestors has the desired effect of treating those two cM values as fixed effects, in much the same way fixed genetic groups (Westell, Quaas, Van Vleck 1988) or fixed animal effects (Graser, Smith and Tier 1987) can be introduced into linear mixed models that typify animal breeding studies. Introducing two fixed effects spends two degrees of freedom.

The data are the observed cM values found on living relatives that also belong to the pedigree. The linear model for the observations is the following.

$\mathbf{y}=\mathbf{Zu}$

Where **y** us a N×1 column vector containing the N shared cM values observed on some of the living relatives, and **Z** is an incidence matrix containing mostly zeros except for a single entry containing the number one in each row that picks out the appropriate element in **u** so that it is matched with the corresponding element in **y**. The linear model for the observations contains no additional error terms, meaning that the elements of **u** are treated as intrinsic measurements that won't vary if re-sampled. Rather than smoothing the estimates of **u**, not using an additional error term has the effect of

returning the cM values as estimates that now equal exactly the cM values that had been observed on some of the living relatives; the rest being best predictions.

No complication is found with **R** rank-deficient, or with the null matrix representing the variance matrix for observations that have no additional error vector, because the normal equations or Henderson's (1973) mixed model equations, won't be used. Rather, a method suitable to handle a singular variance matrix is to be used, described by Siegel (1965) and given by equation (3.9) of Goldberger (1962).

For a given linear model of the form **w**=**Xb**+**e**, with Var(**e**)=**V**, and **w** is observed and where **X** is known, Siegel recommended solving the following indefinite linear system of equations for estimating **b** by a generalized least-squares, as $\hat{\mathbf{b}}$.

$$\begin{bmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}^{\intercal} & \end{bmatrix} \begin{bmatrix} \mathbf{t} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{w} \\ \mathbf{0} \end{bmatrix}$$

It is convenient to augment the coefficient matrix with the right-hand side, producing the following square matrix **M** that is symmetric and indefinite.

(4) $$\mathbf{M} = \begin{bmatrix} \mathbf{V} & \mathbf{X} & \mathbf{w} \\ \mathbf{X}^{T} & & \\ \mathbf{w}^{T} & & \end{bmatrix}$$

The empty space in **M** is understood to be entries of the number zero. As Smith(2001a) demonstrated, the matrix **M** can be subjected to the Cholesky decomposition (generalized for indefinite matrices) or elementary row-operations to decompose **M** by the LU factorization, leading to maximum likelihood estimation of dispersion parameters, and linear estimation and prediction, which includes the calculation of the total sums of square minus the reduction of sums of square - the chi-square statistic. The beauty in this approach is that it works even for singular **V**, and all that is needed is to specify the linear model thereby building the matrix **M** directly using simple plug-ins, then the analysts turns to standardized computer algorithms to apply elementary row-operations, forward and backward substitution, and even backward differentiation, and gone is any reference to the mixed-model equations or the normal equations because those become a redundant by-product of a particular order of row-operations.

For the present example we need not employ the heavy equipment that involves backward differentiation of a likelihood function that is computed from a Cholesky decomposition, as the present application is limited to linear prediction with quasi-known dispersion parameters. Referring to the form on the linear model, substituting

values in for **V**, **X** and **w** in (4) to represent the present case, gives the following partition matrix.

$$\mathbf{M} = \begin{bmatrix} \mathbf{R} & & \mathbf{P} & \\ & & \mathbf{Z} & \mathbf{y} \\ \mathbf{P}^T & \mathbf{Z}^T & & \\ & \mathbf{y}^T & & \end{bmatrix}$$

The empty space in **M** is again understood to represent entries of zero. The computations follow in the outline below.

1. The matrix **M** is constructed as described above, using simple variances derived for the Poisson distribution and the probabilities of identity by descent.

2. A permutation matrix **Q** is found dynamically with the implementation of the LU factorization (see Smith 2001b), to compute the unit lower triangular matrix **L** and an upper triangular matrix **U** such that $\mathbf{LU}=\mathbf{QMQ}^T$, while restricting the permitted permutations to leave the last row and column of **M** fixed in the last position.

3. The chi-square statistic (χ^2) with N – 2 degrees of freedom is retrieve in the last diagonal element of **U** which is present as $-\chi^2$. The expectation is that this statistics will show significance because the Poisson distribution comes with small relative variances, and it is therefore easy to generate a poor fit. Significance implies the presence of extra-Poisson variation, with variance term σ^2 noted below.

$$\sigma^2 = \frac{\chi^2}{N-2}$$

4. To calculate the predictions of shared cM for all the known relatives (i.e., to calculate the prediction of the vector **u**), retrieve the last column **U** but excluding the last element where the chi-square statistic was found, and put it in the work vector **r**. Remove the last row and column of **U**, making a smaller upper triangular matrix $\bar{\mathbf{U}}$. The column vector **r** has already been subjected to implicit forward substitution with the LU factorization. Complete the process now by solving $\hat{\mathbf{s}}$ in $\bar{\mathbf{U}}\hat{\mathbf{s}}=\mathbf{r}$ by backward substitution. The prediction of **u**, now defined as $\hat{\mathbf{u}}$, is found scattered in $\hat{\mathbf{s}}$ depending on the permutations. However, because the permutations are done implicitly by software, $\hat{\mathbf{u}}$ is found in $\hat{\mathbf{s}}$ as if there had been no permutations.

5. With the chi-square statistic significant, the matrix **M** can be rebuilt for re-weighted iteration by using the current value of $\hat{\mathbf{u}}$. The calculation then returns to Step 2 above, and this iteration repeated as many of times as necessary until the chi-square statistic stabilizes. Ideally, the chi-square statistic should fall initially, if only a little, otherwise re-

weighted iteration is not recommended.[6] Once this is done, the last estimate of σ^2 found in Step 3 is taken as the extra-Poisson variation.

6. To predict the shared cM for the envoy add the cM predictions for the central ancestors together; i.e., add two elements of **û** together. Initialize the work vector **r** used in Step 4 to zero everywhere except for the two entries that correspond to the central ancestors that are set to the number one. With the permutations treated implicitly, use forward substitution to solve for the vector **s** in $\bar{U}^T\mathbf{s}=\mathbf{r}$. Calculate the negative weighted sum of squares,

$$\delta^2 = -\sum_i \bar{U}_i s_i^2$$

where $\bar{U}_i$ is the i-th diagonal of $\bar{U}$, and s_i is the i-th element of **s**. The standard error for the shared cM prediction for the envoy is $\sigma\delta=(\sigma^2\delta^2)^{½}$.

4. Numerical Example

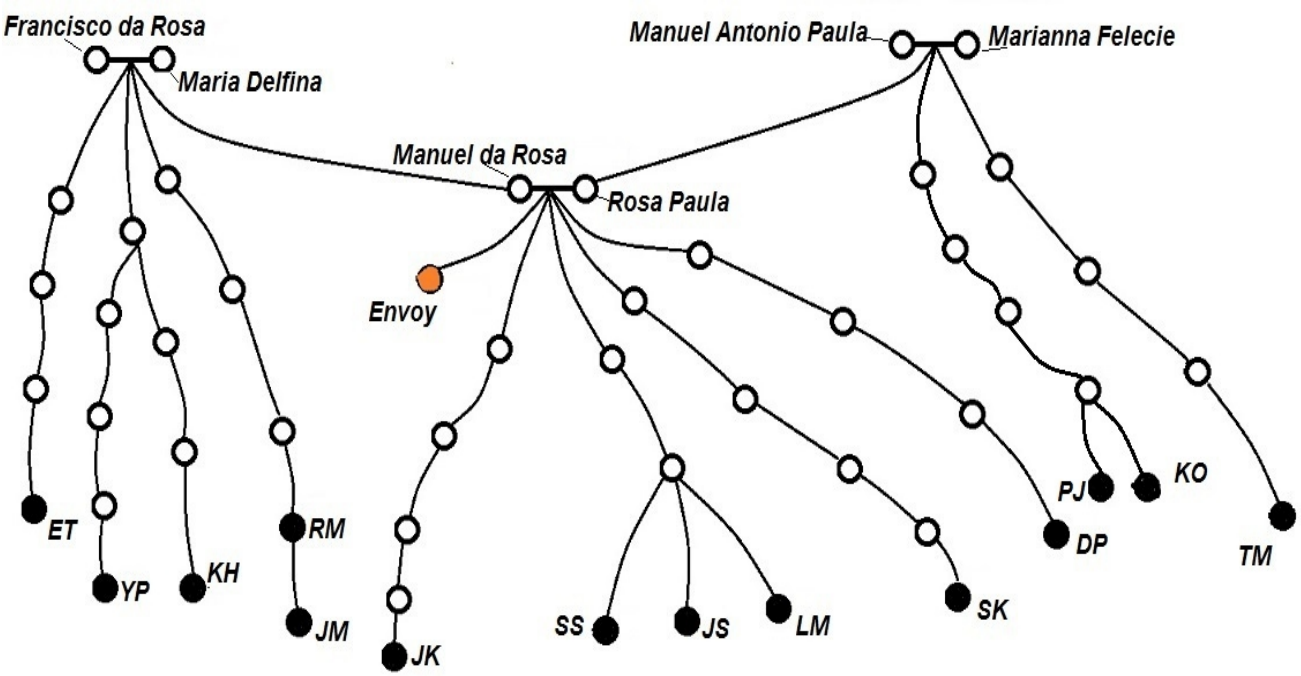
The pedigree information that is used to illustrate the method is presented in Display 1, showing the central ancestors (Manuel da Rosa and Rosa Paula), the paternal common ancestors (Francisco da Rosa and Maria Delfina) and the maternal common ancestors (Manuel Antonio Paula and Marianna Felecie). Fourteen living descendants were identified that took Ancestry's DNA test. The family tree is 4 to 5 generations deep showing relationships between 2nd and 3rd cousins.

Two sisters also took Ancestry's DNA test, and were previously not known related to the family tree shown in Display 1. However matches were found among the 14 individuals shown in Display 1, coming with various degrees of strength as measured in cM. Those cM values are listed in Table 1.

Table 1. Ancestry's cM values between the sisters and 14 individuals that belong to the Rosa and Paula families.

Individual	ET	YP	KH	RM	JM	JK	SS	JS	LM	SK	PJ	KO	DP	TM
Sister 1	36	11	31	28	21	33.2	72	182	108	26	<6	15.5	60	60
Sister 2	15.5	11.4	29.4	26	<6	36	47	135	24.3	<6	<6	<6	39	13.3

[6]Re-weighted iteration defeats any claim of having a best linear unbiased predictor.



Display 1. Family tree with central ancestors Manuel da Rose and Rosa Paula, with paternal common ancestors Francisco da Rosa and Maria Delfina, and with maternal common ancestors Manuel Antonio Paula and Marianna Felecie. Circles indicate individuals involved with gene flows that are common by decent between the envoy (red circle) and living descendants that took DNA tests (dark circles).

Following the method of Section 3, a pedigree of 56 individuals was built, including the envoy and 52 known relatives belonging to the Rosa and Paula families. The 14 cM values was evaluated for each sister in turn, predicting the cM values for all 38 relatives that did not come with a measured cM value.

Regarding the five cM values in Table 1 that correspond to non-matches because cM<6 were found, the corresponding cM values were set to 3 to permit the calculations. Some zero cM values are expected from chance alone even if the sisters are related to the 14 individuals as implied by Display 1. However, setting cM to zero can complicated re-weighted iteration where positive weights are required, and so setting the non-matches to cM=3 (the mid-point) rather than to cM=0 is preferred (not that it matters much).

After the first iteration for the Sister 1, the initial chi-square of 87.47 fell to 75.68 with subsequent re-weighted iteration, and so re-weighted iteration was performed prior to calculating the extra-Poisson variation and predicting the cM values for all the relatives. The cM prediction between Sister 1 and the envoy was calculated as 1004, coming with a standard error of 158. Using a normal approximation, this implies that the actual cM between Sister 1 and the envoy is greater than 745 with 95% probability.

No re-weighted iteration was performed for Sister 2 because the initial chi-square of 85.35 did not decline with re-weighted iteration. The extra-Poisson variation and the predicted cM values were calculated after the initial iteration. The cM prediction between Sister 2 and the envoy was calculated as 567, coming with a standard error of 154. Using a normal approximation, this implies that the actual cM between sister 2 and the envoy is greater than 314 with 95% probability.

The genetic signal between the sisters and the 14 individuals that belong to a known pedigree is stronger in Sister 1 than Sister 2. This difference is entirely expected from genetic recombination. Moreover, the pattern found is consistent with the possibility that the envoy is a great-grandparent[7] of the sisters when their results are taken together.

The evidence is compelling with the cM values observed on the 14 relatives. There remains a small tendency for a confirmation bias in seeing the envoy as a great-grandparent given that the model and its pedigree information is assumed correct. However, setting all the 14 cM values to 3 (i.e., to what is defined to be a non-match), only induces a predicted cM of 64.55 between a sister and the envoy. Therefore, any confirmation bias coming from the model is small. What actually was calculated for the cM between the envoy and a sister was much larger, and was closer to that expected for a great-grandparent. The models treats that the associated cM values for the central ancestors as "fixed effects"[8] that are un-impacted by prior information. These fixed effects are free to respond to the 14 measured cM values, even the non-matches.

The method is also robust to an unknown number of generations separating the sisters and the envoy. It is only necessary for the sisters to have descended from the envoy. To perform the calculation the envoy was assumed to be a great-grandparent of the sisters, but this only impacts the **R** matrix as a proportionality constant (during the first iteration) and has no impact with re-weighted iteration, leaving the linear predictions of the cM values unaffected.

5. Conclusion

The statistical model, and its calculation methods, were successful in combining the cM values of 14 relatives and concentrating those measurements into a single cM value between a hypothetical envoy and the previously unknown relative (actually a pair of full sib sisters). These results were from a combination of statistical linear prediction and genealogical research. However, the exercise also resembled a cluster analysis, where

---

[7]Bettinger (2016, pg 106) expects the cM values for great-grandparents to vary between 547 to 1110.

[8]Despite the unfortunate connotation of the word "fixed," the fixed effect originates from sampling theory and represents a standalone parameter that is free to be anywhere, without any bias from a prior distribution.

the 14 relatives were found clustered by being members in one pedigree that was the product of genealogical research. It is possible to utilize a more comprehensive cluster analysis of all the pair-wise cM values found in a large database, thereby finding many clusters of genetic relatives without the foundation provided by genealogy. There may be some utility in developing clustering tools that can be used to query the database beyond what is presently available[9]. For example, each sisters can only access the 14 cM values in the cluster define by the one pedigree, however, each of the 14 members have an additional 13 pair-wise cM measurements with other members of the cluster, and all such pair-wise cM values go into defining the cluster. There are many such clusters in the database, including clusters that are near-by and overlap, and these can all be identified in principle by a more thorough cluster analysis.

In is unlikely that a more ambitious cluster analysis will ever substitute completely for genealogy. Something must be known about a cluster before a linear model can be defined that connects the unknown relative to the cluster through the presumed envoy. That extra information comes from genealogy in the form of pedigree information.

Having found an envoy with a significantly large cM, as was done for the two sisters, further detective work had been required. The envoy is the child of Manuel da Rosa and Rose Paula, two Portugese emigrants that came to Northern California 150 years ago. That provides a valuable clue on how the envoy might relate to the family tree of the sisters, given that the envoy is a presumed great-grandparent of the sisters, and Manuel and Rosa had 10 children. What had been missing in the sister's family tree was one of several possibilities that was very unclear at first: a misidentified parent, grandparent or great-grandparent, and the sisters were initially thought to be half-sibs. In a remarkable set of discoveries that followed in the wake of the envoy's discovery by statistical analysis, what had gone missing was a lost father, a living grandson of the envoy. At the time of the writing of this paper, the lost father has agreed to take Ancestry's DNA test to confirm the discovery. It is usual for a parent-offspring discovery to come directly from a vary powerful DNA match between parent and offspring, but in the present case 14 relatives were first matched thereby creating a sharper focus in the search, and the discovery of the biological father followed.

References

Bettinger, B.T., 2016, *The Family Tree Guide to DNA Testing and Genetic Genealogy*, Family Tree Books, Cincinnati, Ohio.

Breslow, N.E., 1984, Extra-Poisson Variation in Log-Linear Models, *Applied Statistics*,

[9]My Family Tree DNA and Ancestry both permit shared comparisons, but these are not uniformly defined and are limited to the matches that are visible to the test taker, and a more comprehensive tool (or set of tools) can be very useful.

33 (1): 38-44.

Goldberger, A.S., 1962, Best Linear Unbiased Prediction in the Generalized Linear Regression Model, *Journal of the American Statistical Association*, 57 (298): 369-375.

Graser, H.-U., S.P. Smith and B. Tier, 1987, A Derivative Free Approach for Estimating Variance Components in Animal Models by REML, *Journal of Animal Science*, 64: 1362-1370.

Henderson, C.R., 1973, Sire Evaluation and Genetic Trends, In *Proceeding of the Animal Breeding and Genetics Symposium in Honor of Dr Jay L. Lush*, ASAS and ADSA, Champaign, Illinois, 10-41.

Siegel, I.H., 1965, Deferment of Computation in the Method of Least Squares, *Mathematics of Computation*, 19 (90): 329-331.

Smith, S.P., 2001a, Likelihood-Based Analysis of Linear State-Space Models Using the Cholesky Decomposition, *Journal of Computational and Graphical Statistics*, 10 (2): 350-369.

Smith, S.P., 2001b, Factorability of Symmetric Matrices, *Linear Algebra and Its Application*, 335: 63-80.

Van Vleck, D., 1979, *Notes on the Theory and Application of Selection Principles for the Genetic Improvement of Animals*, Cornell University, Ithaca, New York.

Westell, R.A., R.L. Quaas, and D. Van Vleck, 1988, Genetic Groups in an Animal Model, *Journal of Dairy Science*, 71 (5): 1310-1318.