FAMILY FINDER MATCHING 5.0

Matching Algorithm and Relationship Estimation

White Paper 2021-08-18

Rui Hu • Paul Maier • Göran Runfeldt • Roberta Estes • Efrain Rocha
Andrew Walker • Patrick Baur • Linh Ty

Contents

Introduction

Genetic genealogy is the practice of combining genetic profiles with traditional genealogical methods to infer pedigree relationships. All humans share one large and tangled pedigree.

Largely thanks to the recent explosion in the sizes of human genomic databases, it has become clear that all humans share ancestors on surprisingly recent time scales. For example, if the pedigrees of any two Europeans are compared back in time 1,000 years (approximately 33 generations), they share approximately the same set of ancestors [1]. Each person's set of 2^{33} ancestors would include more than 100× the size of the European population in approximately the year 1000 A.D.

Due to the randomness of genetic recombination, we do not have the same set of genetic ancestors. A genetic ancestor is any ancestor who is a genetic contributor to you. Although there is nearly a 100% probability that any one of your 8 great grandparents (GGPs) contributed DNA to you, that probability drops off for earlier ancestors: 5th GGPs (97%), 7th GGPs (67%), 9th GGPs (29%), 11th GGPs (10%), etc. Any two Europeans are estimated to share one genetic ancestor between 1,000 and 2,000 years ago and 10 genetic ancestors looking 3,000 years into the past [1]. Of course, this is only an average, and many more genetic ancestors are shared within the same population, particularly within endogamous populations. (For more information on this subject, please see our myOrigins v3 White Paper.)

Humans therefore share an increasingly large amount of DNA across historic, prehistoric, and ancient time scales.

At the heart of genetic genealogy is "DNA matching": the comparison of DNA segments between individuals for the purpose of identifying genetic relatives within recent times. Therefore, it is imperative to distinguish between DNA segments that are identical-by-descent (IBD) from one most recent common ancestor (MRCA) and DNA that is merely identical-by-state (IBS). Segments of DNA that are IBS may be identical but ubiquitous in the population, as people inherited them from many different ancestors (and ultimately from one very ancient ancestor). Whereas the size distribution of IBD segments can be used to infer recent relationships, IBS segments are likely to mislead such inference if they are erroneously treated as IBD. Thus, the challenge in genetic genealogy is to leverage IBD segments to accurately predict relationships while accounting for the presence of IBS segments.

The Family Finder autosomal DNA test utilizes Illumina SNP array results that include genotypes for approximately 700,000 autosomal and X-chromosome SNP loci distributed across the genome. SNPs are the ~1% of DNA sites that are known to be variable within the human population. Thus, they can resolve whether and where two DNA segments are identical or different. Both your mother and father produce recombined versions of their DNA during a process called meiosis and pass along their copies to you. In other words, you inherit from your mother a copy of her chromosomes that are a patchwork of your maternal grandparents' DNA. Our algorithm will find that you are half-identical to your parents across every chromosome, and therefore parent-child segments span the entire chromosome (except for the X in fathers/sons). However, your siblings will inherit DNA that underwent a separate meiosis. Since you and your

siblings are separated by two meioses, you will share shorter segments, fractions of each chromosome. First, second, and third cousins share segments of decreasing size and number because they are separated by four, six, and eight meioses, respectively. Thus, shorter and fewer IBD segments are the result of more random recombination and generally indicate more distant relationships [2,3].

More distant relationships are expected to have fewer and shorter IBD segments; however, genealogically uninformative IBS segments also tend to be short. Most IBS segments have a much older origin and experienced many meioses that reduced their size before they spread across the population. Thus, cumulative IBS segments can potentially make unrelated people appear to be distant relatives and make distant relatives appear to be closer. To further complicate matters, identical-by-chance (IBC) segments can inflate relationship estimates even more. IBC is the phenomenon whereby you separately inherit an IBS segment from each parent, but by chance, they are adjacent to one another. Their apparently longer combined length can make IBC segments difficult to distinguish from true IBD segments.

The pattern of shared IBD and IBS/IBC segments between genealogical relatives of a certain degree can vary greatly across different parts of the world, depending on from which ancestral population their shared ancestry is derived. Many populations have a long history of intermarriage within the cultural group that over time can result in lower genetic diversity, which is referred to as endogamy. Two matches with shared ancestry from the same endogamous population will share more IBS and IBC segments than matches with ancestry from non-endogamous (or different endogamous) populations. For this reason, many matching systems make special adjustments for matches between Ashkenazi Jewish, Native American, and Polynesian individuals, among others.

Our new matching system introduces a novel genealogy relationship prediction system that utilizes machine learning. The system is based on shared IBD segments and incorporates an endogamy score based on IBS segments. The endogamy adjustment reduces false-positive matches and provides more accurate relationship predictions for connections from endogamous populations including Ashkenazi, Native American, Polynesian, Finnish, and others.

In brief, there are five modules to Family Finder matching:

1. A SNP-by-SNP comparison for each pair of customers to find shared segments and generate match statistics.
2. A support vector machine (SVM) classifier that determines the extent of DNA matching due to a shared endogamous population based on the pattern of IBS micro-segments.
3. An SVM classifier that filters out false-positive matches using the pattern of shared centimorgans (cM); both total cM and longest cM are used.
4. Naïve Bayes classifiers to predict the range of relationship based on input from previous steps #2 and #3.
5. The match list, relationship estimates, and their ranges are displayed.

Our new algorithm has resulted in changed match thresholds based in part on the endogamy adjustment, as well as improved relationship estimates. It is important to note: the new match thresholds may not always be immediately intuitive. IBS micro-segments are false-positive (i.e., genealogically irrelevant) IBD segments; therefore, it is possible to share less total DNA with a true genealogical match than a false one. For example, a true match may have one 7 cM segment that is IBD, whereas a false match may have one 7 cM segment and several 2 cM segments.

These improvements have changed the customer experience in the following ways:

- All segments below 6 cM have been removed from the reported total cM.
- The lowest possible match threshold requires at minimum one 7 cM segment, but the actual match threshold, especially for individuals with endogamy in their pedigree, may vary based on their endogamy score.
- More accurate IBD matching with fewer false IBS/IBC matches.
- More accurate relationship predictions especially for endogamous populations, such as Ashkenazi, Native American, and Polynesian populations.

We are very excited to introduce these improvements as they will significantly help your journey to find your ancestors!

Matching Algorithm

The purpose of the newly improved matching algorithm is to detect identical-by-descent (IBD) segments shared by a pair of DNA samples. Prior to matching, DNA test results are imputed to a union SNP set for the supported chip types and versions, using industry-standard imputation software. Low-quality imputed genotypes are marked as no-calls and information about whether the SNP call was genotyped or imputed is preserved. The imputed data are then fed into the matching workflow (illustrated in Figure 1), which consists of the following steps:

(1) Perform SNP by SNP comparison for the whole genome including autosomal chromosomes 1–22 and chromosome X. This generates an array of 0s and 1s, where 0 and 1 represent mismatched and matched SNPs, respectively. A matched SNP is defined as two genotypes sharing at least one common allele, e.g., A/A vs. A/A, or A/A vs. A/C; whereas a mismatched SNP is defined as two homozygous genotypes with different alleles, e.g., A/A vs. C/C. A no-call in either result is not counted as a matched SNP.

(2) Consolidate consecutive matched SNPs into matched segments and scan all segments to find qualified seed segments. Seed segments can be extended by joining adjacent micro-segments. For increased accuracy, we require a seed segment to have at least 900 matched SNPs without any mismatches.

(3) Extend the seed segment in both forward and reverse directions to merge nearby matched segments using a predefined error rate. The genotypes in the input file could have errors from either genotyping or imputation. Therefore, mismatches below the error rate threshold are allowed in a match segment in order to tolerate such errors.

(4) Calculate the centimorgan (cM) value based on the centimorgan map (build 37) from 1000 Genomes Project, and SNP density for each segment, then prune off any unqualified segments. The criteria for a qualified (micro or macro) segment are at least 480 SNPs, 2 cM, and 105 SNPs/cM.

(5) Generate match statistics. We use a pre-defined centimorgan threshold to categorize segments into micro or macro-segments. We then calculate the total centimorgans and segment counts for both segment size classes, which are used for downstream relationship estimations (micro-segments for endogamy score and macro-segments for IBD relationship prediction).

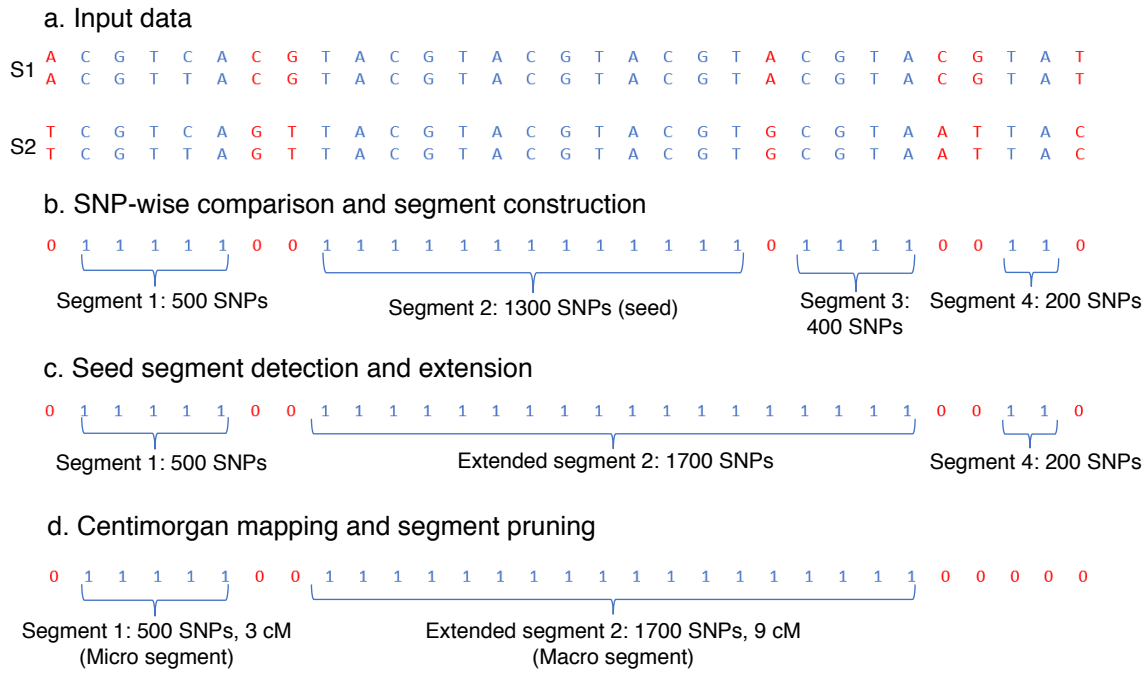

Figure 1. Family Finder matching workflow. (a) S1 and S2 are two sample files to compare. (b) Results of SNP-wise comparison and segment construction. Matched SNPs are denoted by '1' whereas mismatched SNPs are denoted by '0.' Each number represents 100 SNPs for illustration purposes. Four segments are found with the lengths of 500, 1300, 400, and 200 SNPs, respectively. (c) All segments are scanned, and those with >900 SNPs are marked as seed segments (segment 2 in this figure). Seed segments are extended in forward and reverse directions to merge with nearby segments. Here, segment 3 is merged with segment 2 to form a longer segment due to only one mismatch between them. (d) The cM length is calculated for each segment, and tiny segments are pruned out based on the number of SNPs, cM, and SNP density. Here, segment 4 is removed due to the low SNP count. Finally, segments are categorized into micro- and macro-segments. Micro-segments (segment 1), which are more likely to be IBS segments, are used to predict endogamy score. Macro-segments (segment 2), which are more likely to be IBD segments, are used to predict the relationship.

Relationship Estimation

Background

The goal of relationship estimation is to predict the relationship level between two samples based on the amount of shared DNA from one or more recent common ancestors. This is a challenging task due to the following facts:

(1) IBD segments must have been inherited by descendants of one single ancestor and no one else. In contrast, false-positive identical-by-state (IBS) segments are inherited from a much older or ancient ancestor, and as a result, many very distantly related people in the population possess them. They are difficult to differentiate from true IBD segments using any matching algorithm.

(2) Inheritance patterns are different for populations with different levels of endogamy. The diversity within endogamous populations is smaller, leading to longer runs of homozygosity (ROH) and more population-based IBS segments.

(3) The variation in the amount of DNA sharing among matches is large due to the randomness of inheritance. For example, a 5^{th} cousin sometimes can share more DNA than a 4^{th} cousin [4,5].

(4) The access to a large amount of accurate training data is limited.

To overcome these challenges, we invented a novel relationship estimation method, as shown in the following diagram (Figure 2). The input is the list of match statistics generated from the matching algorithm. This is fed into an endogamy classification module to assess the endogamy pattern, followed by a match classification step to filter out false matches from the list of candidate matches. Finally, the match statistics and endogamy coefficient for qualified matches are used to estimate the relationships and their associated probabilities. In our new methodology, we utilize the distribution of match segment size (including both IBD and IBS segments) to detect the level of endogamy, and ensembled models based on both total centimorgans and the longest segment to generate more accurate predictions for both close and distant matches. Each module in the pipeline is described in the following sections.

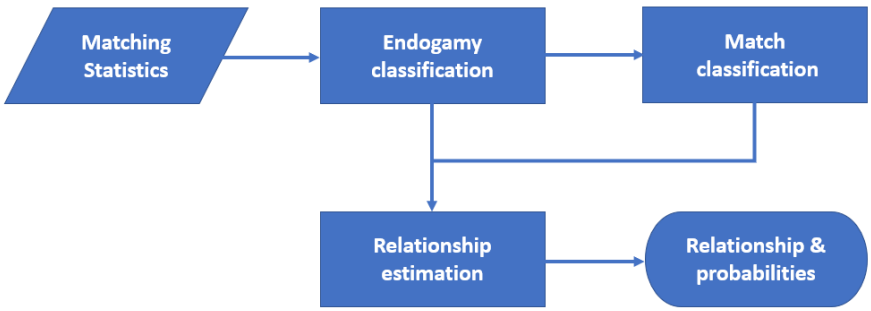

Figure 2. Diagram of relationship estimation.

Endogamy Classification

One of the challenges in relationship estimation is the different matching patterns in populations with different levels of endogamy. Figure 3 shows the histogram of total centimorgans and the longest centimorgan for the non-Jewish European matches and Ashkenazi Jewish matches from a European Jewish admixed individual.

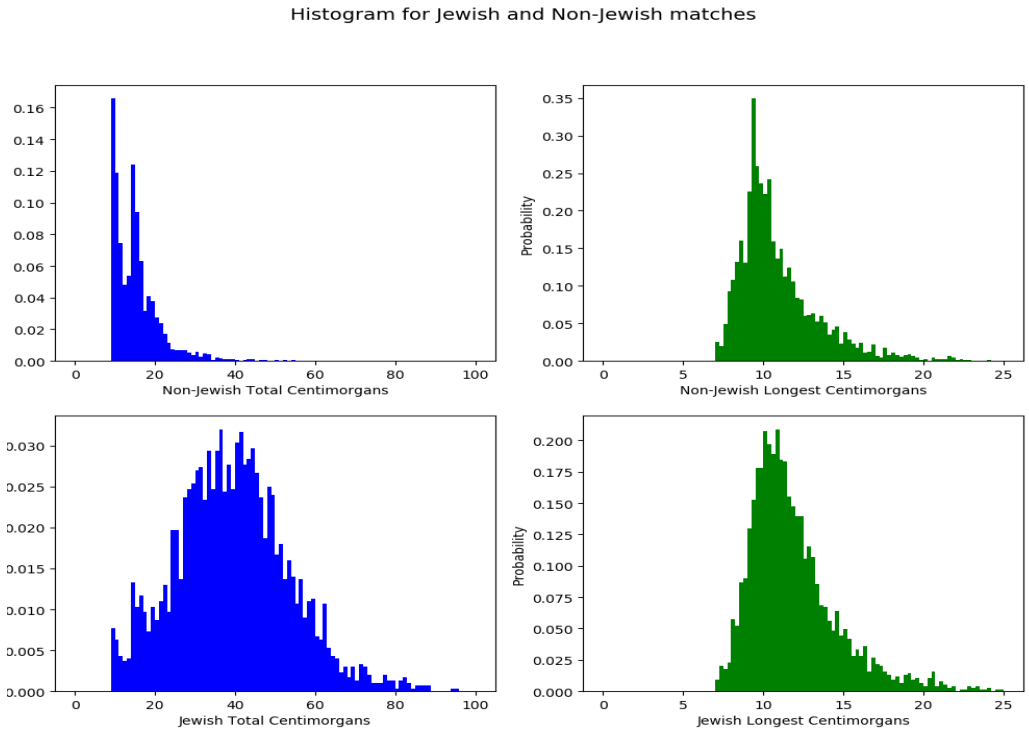

Figure 3. Histogram of total centimorgans and longest centimorgan for non-Jewish and Jewish matches, for an individual with both Jewish and non-Jewish ancestry.

Here we can see the overall total centimorgans for Jewish matches is higher than non-Jewish matches due to the high level of shared homozygosity (IBS segments) among Ashkenazim. To further investigate the matching pattern, a principal components analysis (PCA) was conducted for three different populations: British, Ashkenazi, and Native American, using the matching statistics as features (Figure 4). We can see that the endogamous populations show a larger number of micro-segments. Therefore, we built an SVM classifier to detect the probability of endogamy for a given pair of matches [6]. The predicted endogamous probability can be used as a weighted coefficient to combine endogamous and non-endogamous models in the downstream steps for relationship estimation.

Match Classification

The existence of IBS segments can cause false positives when detecting DNA matches. Usually, false-positive matches will contain one or more short DNA segments. These micro-segments are a useful signature of false IBD matches, so certain criteria can be established to reduce or

eliminate them. One potential solution is to use a pre-defined flat threshold of total centimorgans or the longest centimorgan. However, a single threshold of total centimorgans or the longest centimorgan does not effectively capture complex differences between true and false matches at different relationship levels. Therefore, we built an SVM classifier utilizing both total centimorgans and the longest centimorgan to automatically classify true or false matches.

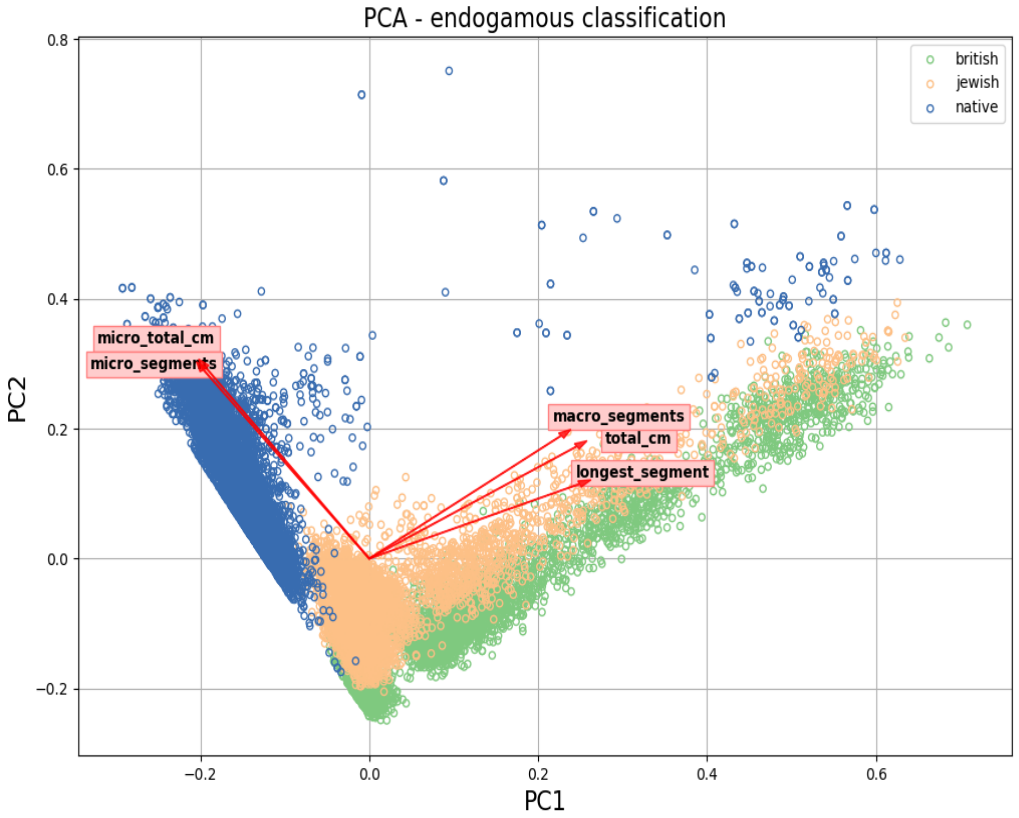

Figure 4. PCA plot for three populations: British (green), Ashkenazi (tan), and Native American (blue). The latter two populations are endogamous.

Given that there is a paucity of training samples with known status (related vs. unrelated), especially for distant relatives, we created a simulated dataset with customized pedigrees to train our model for match classification. We randomly selected 4,000 unrelated samples as founders from Western European and Ashkenazi Jewish populations to represent non-endogamous and endogamous populations, respectively. The simulated relationship ranged from parent/child and full siblings up to 6th cousins as shown in the following pedigree (Figure 5).

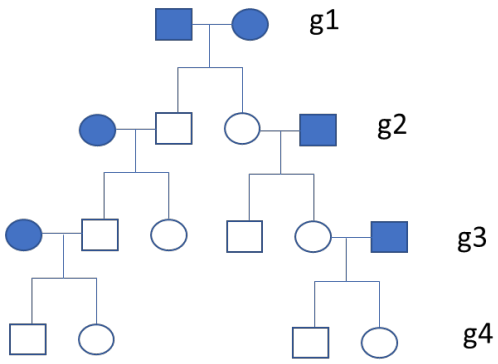

Figure 5. Simulated family trees. Blue symbols indicate founders. We simulated multiple trees up to 8th generation (6th cousin level). For simplicity, only one tree with 4 generations is shown.

We ran the matching algorithm on the simulated dataset, and both the total centimorgans and the longest centimorgan were selected as features to train the match classifier. Figure 6 shows the decision boundaries of the trained classifiers. As shown in Figure 6, the true matches have higher total centimorgans and higher longest centimorgan in general. When the longest centimorgan segment is relatively short, a relatively large number of total centimorgans is required to qualify a match as being true. But when the ratio of the longest centimorgan to the total centimorgans decreases, the chance of qualifying a true match also decreases, as it is more likely to be multiple short IBS segments aggregating into a high total centimorgan value. We trained a non-endogamous model and an endogamous model based on their match patterns. To combine the prediction results of endogamous and non-endogamous models, the endogamous probability generated in the previous step was used as a coefficient of a weighted sum of the match probabilities using the following equation:

$$P(m) = P(e_{ne})P(m_{ne}) + \big(1 - P(e_{ne})\big)P(m_{en}) \quad (1)$$

where $P(m)$ is the final match probability used to determine if a match candidate qualifies a true match, $P(e_{ne})$ is the score of being non-endogamous, and $P(m_{ne})$ and $P(m_{en})$ are the match probabilities from non-endogamous and endogamous models, respectively. For example, if a pair of matches are 20% from a non-endogamous line, and the probabilities of being a true match by the non-endogamous model and endogamous model are 90% and 30%, respectively, then the final match probability is $20\% \times 90\% + (1 - 20\%) \times 30\% = 42\%$.

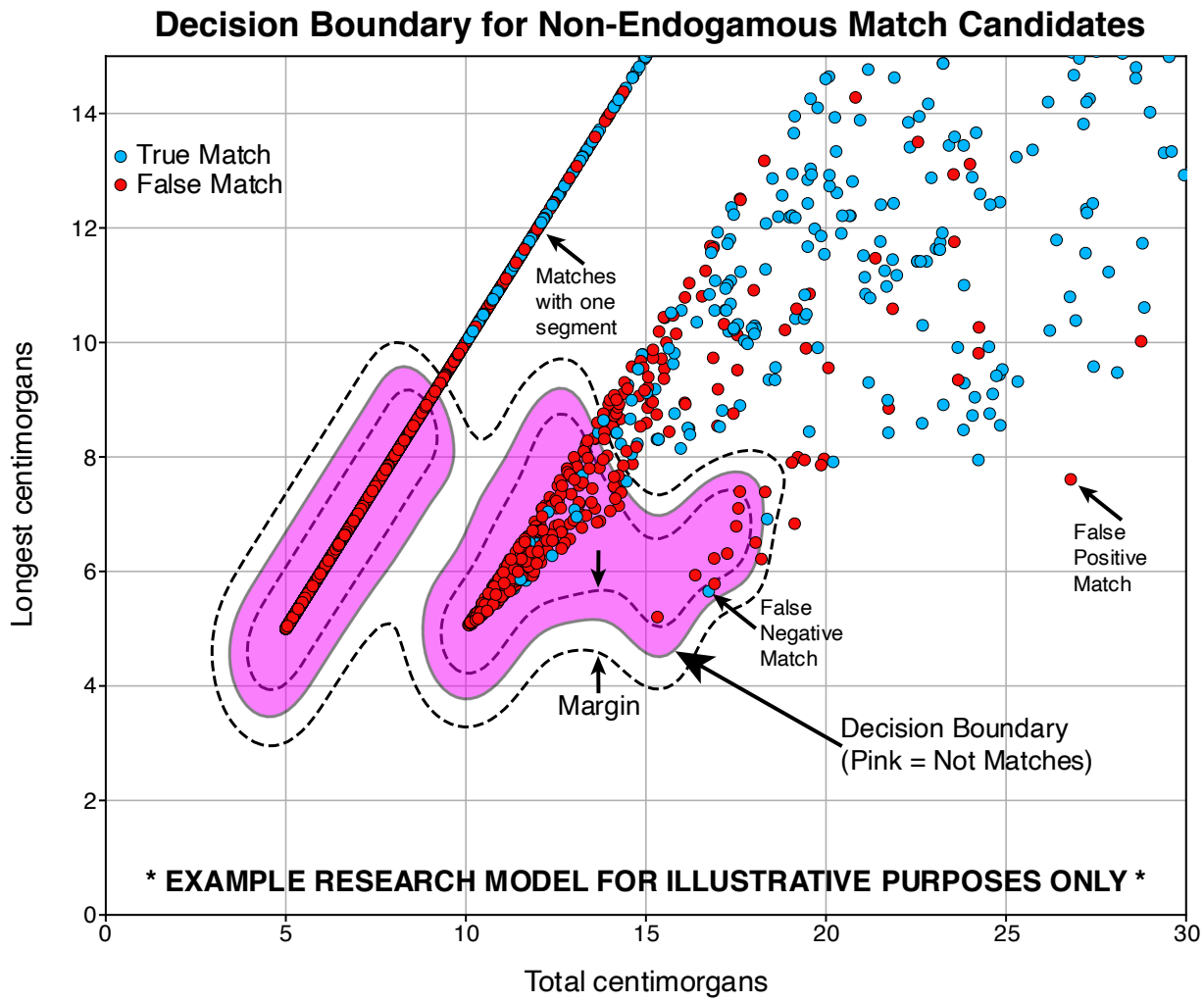

Figure 6. Example decision boundary for the non-endogamous match classifier. Red and blue dots represent false and true matches, respectively. Pink area indicates prediction for false matches. If a match with the total centimorgans (horizontal axis) and the longest centimorgan (vertical axis) falls into the pink area, then it will be determined as a nonmatch. Blue dots inside the pink area are false negative matches, while red dots outside the pink area are false positive matches.

Relationship Estimation

Once a list of true matches is generated for a customer, the goal is to estimate the relationship level between each pair of samples based on the match statistics generated by the matching algorithms. The possible degrees of relatedness are summarized below (Table 1). Degree of relatedness is the distance between two nodes in the family tree. For example, parent and child are directly connected with a distance of one in the family tree, so the degree is one. Full siblings are connected with one step up and one step down, so the total distance is two. First cousins have two steps up to the grandparents and two steps down with a total distance of four. For each degree between two people, one meiosis has reduced the amount of their shared DNA.

To accomplish the goal of estimating a relationship level, first it is important to investigate the distribution of IBD segments for each relationship level. Here, we use the total centimorgan count of the macro-segments to eliminate the effect of IBS segments, since longer segments have a much higher chance of being IBD segments. We initially found 116,114 pairs of samples from the family trees of FamilyTreeDNA customers, which had already been annotated by linked relationships. Given that family trees were self-reported by customers and occasionally had incorrect relationships, we removed outliers for each relationship group. This resulted in a final set of 111,429 pairs of samples in the training set.

Table 1. Description of degree of relatedness for relationship estimation.

Degree	Relationship(s)
1	Parent/child
2	Full sibling
3	Half sibling, uncle/aunt/niece/nephew, grandparent/grandchild
4	1C, great/half uncle/aunt/niece/nephew
5	1C1R, half 1C
6	2C, half 1C1R, 1C2R
7	2C1R, half 2C, 1C3R, half 1C2R
8	3C, 2C2R, half 2C1R, half 1C3R
9	3C1R, half 3C, 2C3R, half 2C2R
10	4C, 3C2R, half 3C1R, half 2C3R
11	4C1R, half 4C, 3C3R, half 3C2R
12	5C, 4C2R, half 4C1R, half 3C3R

Figure 6 shows the conditional probability of total centimorgans given the relationship level, which can be denoted as $P(m|d)$, where m is the total centimorgans, and d is the degree of relationship. According to Bayes' theorem, we can construct a Naïve Bayes classifier [7] by:

$$P(d|m) = \frac{P(m|d) \times P(d)}{P(m)} \tag{2}$$

where $P(d|m)$ is the target probability of relationship given the matching statistics, $P(d)$ is the prior probability, and $P(m)$ is a normalized factor that is unrelated to d.

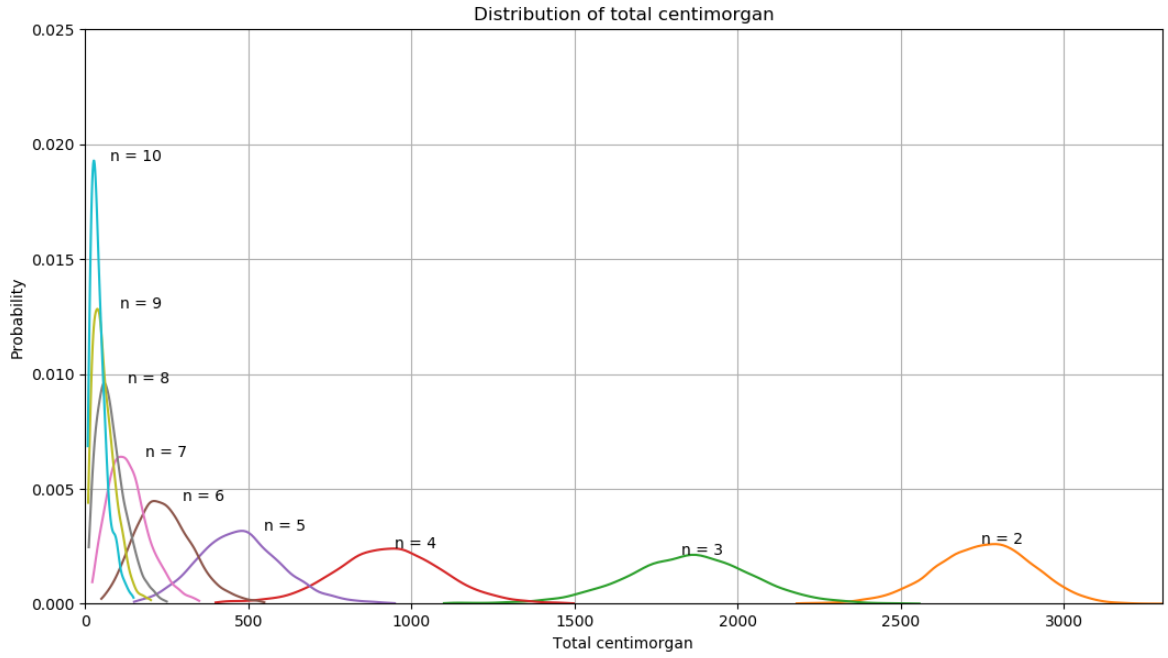

Figure 7. Distribution of total centimorgans for different relationship levels.

In addition, Figure 3 shows an offset in the distribution of total centimorgans between endogamous matches and non-endogamous matches. Therefore, we built distributions of total centimorgans for endogamous populations in a similar way and the predicted probability for endogamous matches can be calculated by Equation 2. Then, the predicted probability was combined by both endogamous and non-endogamous probabilities using the following equation:

$$P(d_{total}|m) = P(e_{ne})P(d_{ne}|m) + \big(1 - P(e_{ne})\big)P(d_{en}|m) \quad (3)$$

where $P(e_{ne})$ is the probability of non-endogamous populations predicted by SVM classifier, and $P(d_{ne}|m)$ and $P(d_{en}|m)$ are the estimated probabilities using non-endogamous and endogamous models, respectively. The idea is similar to what is conducted in the match classification step. We generated the predicted probabilities by the non-endogamous and endogamous models, then used the endogamous score to assign different weights to each model. The endogamous prediction is weighted more for endogamous matches and vice versa.

We discovered that the centimorgan value for the longest matched segment carries more representative information for relationship estimation in distant relatives and endogamous populations. For example, distant matches in Native American populations sometimes can share total centimorgans over 200 cM, among which the longest segment can be less than 10 cM. In this case, relationship estimates based solely on the total centimorgans would cause bias in the prediction. Since the distribution of the longest centimorgan is very similar in endogamous and non-endogamous populations alike (Figure 3), we modeled the conditional probability of the longest centimorgan given the relationship level in one single model (Figure 8).

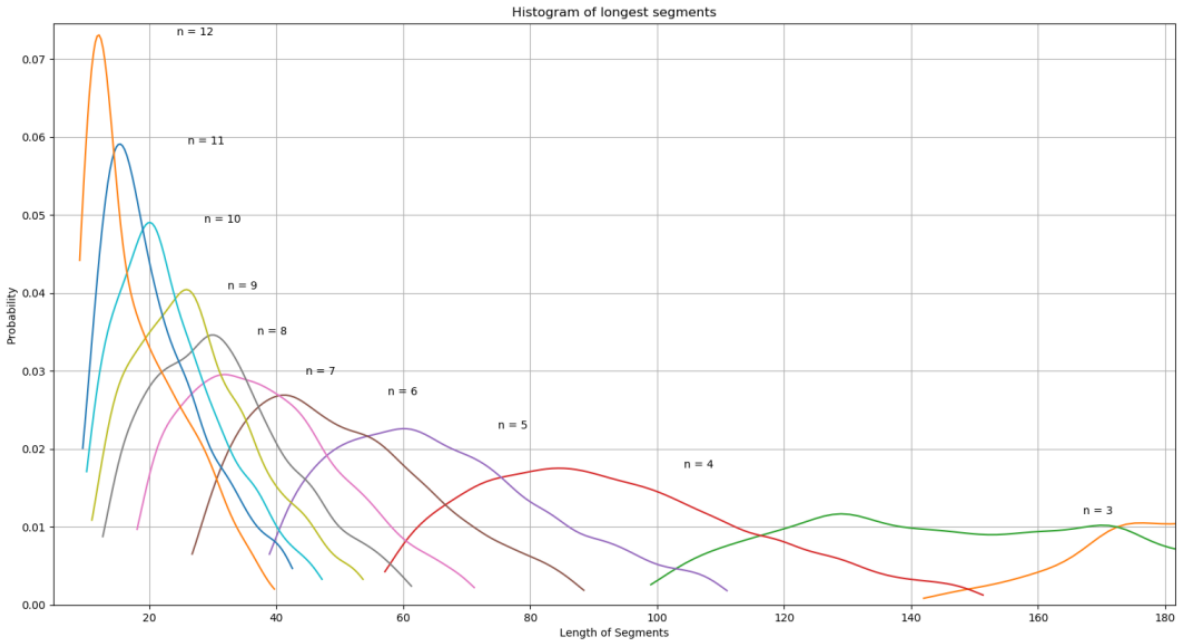

Figure 8. Distribution of the longest centimorgan for different relationship levels.

We constructed a Naïve Bayes classifier in the same way as in Equation 2 to estimate the probability of relationship level given the longest centimorgan value. Finally, we ensembled the total centimorgan models and the longest centimorgan models by the following equation:

$$P(d_{final}|m) = c \times P(d_{total}|m) + (1 - c) \times P(d_{longest}|m) \quad (4)$$

where c is the coefficient to combine both models by favoring the total centimorgan model for closer matches and the longest centimorgan model for more distant matches. As the overlap between the distributions of distant relationships is wide, due to large variation in shared DNA, we generated relationship ranges as an approximate 90% confidence interval by removing all relationships with <5% probability.

Validation and Results

Matching Segments

In relationship estimation, it is essential to use IBD segments rather than IBS segments to predict the relatedness. Therefore, it is important that the matching algorithm can detect IBD segments accurately so that downstream analyses will not be misled. To evaluate the accuracy of the matching segments, we ran our matching algorithms on the simulated dataset described above. In the simulated dataset, we know which segment is inherited from which ancestor, so those

segments can be used as ground truth data to compare with the detected segments for validation purposes. Results are summarized in Table 2. The true segments are from the simulated dataset.

Table 2. False positive rate according to different lengths of match segments.

cM	True IBD segments	False IBS or IBC segments	False positive rate
1	295	798731	99.96%
2	2064	574497	99.64%
3	4225	135698	96.98%
4	5538	24377	81.49%
5	5926	5434	47.83%
6	5745	1413	19.74%
7	5496	592	9.72%
8	5278	364	6.45%
9	5404	186	3.33%
10	5273	109	2.03%
11	5019	69	1.36%
12	4674	26	0.55%
13	4400	27	0.61%
14	4429	13	0.29%

Table 2 shows that with increasing segment lengths, the false-positive rate decreases. For micro-segments that are merely 1–2 cM, the false positive rate is even higher than 99%! This means that nearly every one of these tiny segments is not genealogically helpful and was not inherited from a recent common ancestor. When the segment length is above 6 cM, the false-positive rate reduces to <20%. Therefore, we choose 6 cM as a reasonable threshold to report IBD segments.

Endogamy Classification

Our classifier tested the endogamy level on a set of 3,176 pairs of matches with ancestry from British (proxy for non-endogamous), Ashkenazi Jewish, and Native American populations. We also selected testing matches evenly from different levels of relationship groups to avoid the bias caused by different relatedness. The results are summarized in a confusion matrix (Table 3). Each row of the matrix represents the true endogamy class, and each column represents the predicted class. For examples, out of 988 pairs of matches from the British population, 942 were predicted correctly, whereas 46 were predicted incorrectly as being Ashkenazi Jewish. The balanced accuracy, which is the average accuracy of all classes, was 95.7%.

The advantage of using an endogamous classifier on the match level is that each pair of matches can be evaluated independently. We do not need to wait until population ancestry estimates (i.e., myOrigins) are completed. Also, for samples that are admixed between endogamous and non-endogamous populations, each match can be evaluated based on the match pattern to assign an endogamous score individually. In addition, endogamous populations other than Ashkenazi

Jewish population that are less represented in our database, such as Native American and Polynesian, can also get improved relationship predictions.

Table 3. Confusion matrix for endogamous classification. Values on the diagonal are correctly classified.

Actual \ Predicted	Non-endogamous	Endogamous (Ashkenazi Jewish)	Endogamous (Native American)
Non-endogamous	942	46	0
Endogamous (Ashkenazi Jewish)	78	970	0
Endogamous (Native American)	1	8	1131

Match Classification

We tested the accuracy of the match classification using the same simulated dataset described above. The data was split into a training and a testing set. The training dataset was used to train the match classifier, then the classifier was validated using the testing dataset. The confusion matrix of the match classifier is shown in Table 4, and the balanced accuracy was 96.2%.

Table 4. Confusion matrix for match classification.

Actual \ Predicted	True Match	False Match
True Match	2440	52
False Match	98	1702

We also tested the match classifier on real data. We used data from customers whose parents are both tested and compared the child's match list with the parents' match list. If a person on the child's match list does not match either parent, then that person is very likely a false-positive match for the child. In this way, we devised a rough estimate of the false-positive match rate. We calculated the percentage of matches not in common with parents for the current pipeline and our previous version of the pipeline (Table 5).

Table 5. False positive match rate for current and previous version.

Method	Mean of false positive rate	Standard deviation of false positive rate
Previous	21.6%	7.7%
Current	18.2%	6.0%

Table 5 shows that the false positive rate using the current method is lower than using our previous version. Additionally, we compared the average number of matches between the two versions. On average, our current method produces 10% additional matches (with 14% standard deviation) compared to the previous version. This indicates that the proposed method is able to detect more matches with a lower false-positive rate.

Relationship Estimation

To test the accuracy of our relationship estimation method, we validated our results on a set of 13,803 pairs of linked matches from customer self-created family trees. These matches were randomly sampled from our customer database and not used for training the models. Among all the testing matches, 7,336 pairs got correct point estimation, and 12,868 pairs (93%) got correct range estimation. More details for decompositions for each relationship level are as follows:

Table 6. True and false positive predictions for each relationship category.

Degree	Relationship(s)	False	True	Total
1	Parent/child	0	500	500
2	Full sibling	0	945	945
3	Grand parent/child, uncle/aunt/nephew/niece	2	1,016	1,018
4	First cousin	0	1,003	1,003
5	First cousin once removed	63	1,437	1,500
6	Second cousin	12	1,488	1,500
7	Second cousin once removed	174	1,326	1,500
8	Third cousin	93	1,407	1,500
9	Third cousin once removed	196	1,304	1,500
10	Fourth cousin	4	1,516	1,520
11	Fourth cousin once removed	254	553	807
12	Fifth cousin	137	373	510
Grand Total		**935**	**12,868**	**13,803**

Table 6 shows that prediction accuracy is lower for more distant relationships. This is an effect of shared IBD variation between distant matches due to the randomness of Mendelian inheritance. We can see that two pairs of matches in the grandparent/grandchild level were predicted incorrectly. Those are in the lower bound of the centimorgan range, and the prediction probability is too low to display in a 90% confidence interval. Additionally, the error rate for an odd-numbered relationship degree is higher. An example of an odd-numbered degree is any half

relationship such as half 2nd cousin, with degree of 7 (see Table 1). Half relationships have higher error rates because we predict full relationship ranges such as 1st to 3rd cousin. Therefore, any odd degree that falls between the boundary of two nearby ranges is more likely to be outside of the predicted range.

Fourth cousin predictions have particularly high accuracy (Table 6) for the following reason. The most distant relationship ranges we predict are 2nd–4th cousin, 3rd–4th cousin, 3rd–5th cousin, and 4th–remote cousin. All of them include "4th cousin." Therefore, the chance of a 4th cousin being predicted correctly is higher than for other relationship levels.

To further investigate how different the predicted range is from the actual relationship degree, we used the actual relationship degree and its closest predicted relationship to generate a confusion matrix, shown in Figure 9. Most of the incorrect predictions are just one or two degrees off from the true label. For example, all incorrect predictions for a 5th degree relationship (1st cousin once removed) are predicted as within the 2nd cousin range.

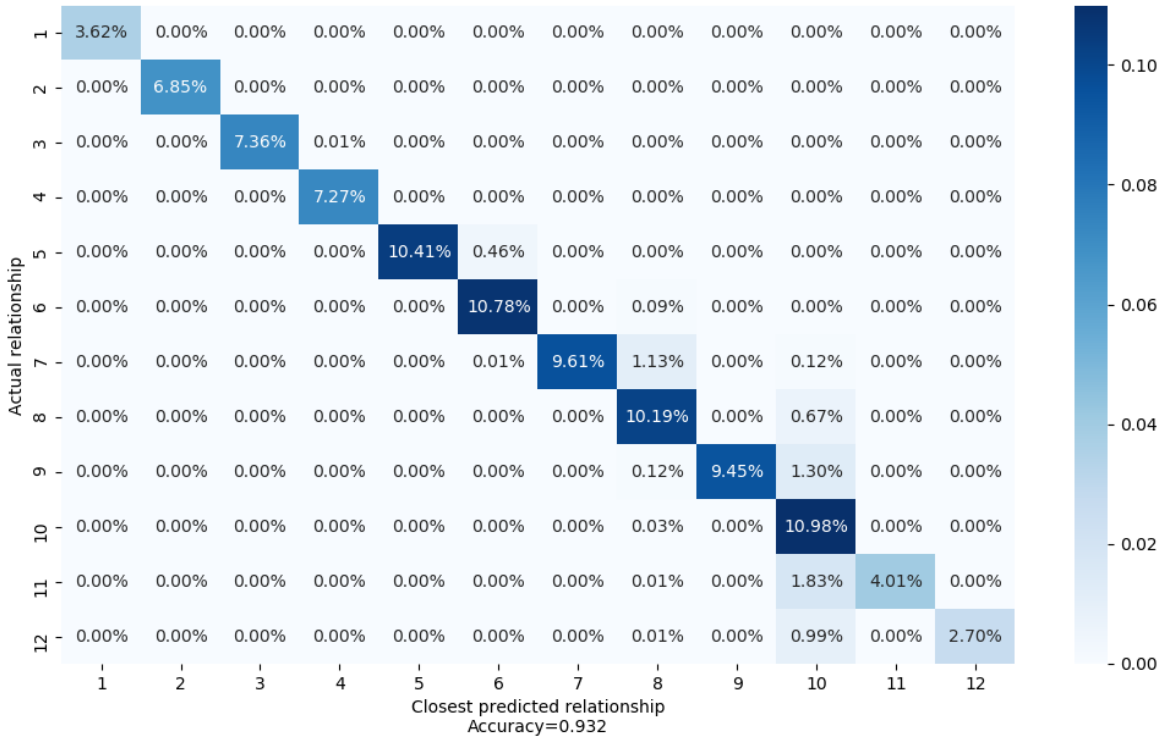

Figure 9. Confusion matrix heatmap for relationship prediction.

Summary and Limitations

In this paper, we have described the pipeline for the new Family Finder matching algorithm and relationship estimation.

(1) For each individual pair of matches, our pipeline has implemented an endogamous classifier to generate an endogamous score based on the matching pattern, including IBD and IBS segments.

(2) The match classifier, which is based on total centimorgans and the longest centimorgan value, captures complex pattern of true matches so that it can detect more matches while keeping the false positive rate low.

(3) In addition, the relationship estimation is not solely based on total centimorgans. The longest centimorgan value is also utilized to estimate the relationship, which provides additional information and more accurate relationship range estimations.

(4) The new method provides significant improvements, especially for matches from endogamous populations, such as Polynesian and Native American populations.

However, there are also some limitations in our methods.

(1) The endogamous classification is sometimes less effective when classifying matches that are partially mixed between endogamous and non-endogamous populations, because the IBS pattern is not sufficiently distinct given the small portion of endogamous admixture.

(2) The match classifier was trained on a simulated dataset and may not cover all possible patterns of real-world data, especially for endogamous populations. A growing dataset with known pedigree information will help train a more accurate classifier.

(3) The accuracy of point prediction is lower for distant matches. Distant relationships are difficult to predict accurately due to the large variation in DNA sharing by distant relatives.

As our database grows and more data with family tree information are accessible, we will continue improving our methods to provide more accurate results to our customers.

References

1 Ralph, P. and Coop, G. (2013) The geography of recent genetic ancestry across Europe. *PLoS Biol.* 11, e1001555

2 Li, H., Glusman, G., Hu, H., Shankaracharya, Caballero, J., et al. (2014) Relationship estimation from whole genome sequence data. *PLoS Genet.* 10, e1004144

3 Huff, C.D., Whitespoon, D.J., Simonson, T.S., Xing, J., Watkins, W.S., et al. (2011) Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res.* 21, 768–774

4 Hill, W.G., Weir, B.S. (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93, 47–64

5 Bettinger, B.T. (2016) The shared cM project: a demonstration of the power of citizen science. *J. Genet. Geneal.* 8, 38–42

6 Chang, C.C. and Lin, C.J. (2011) LIBSVM: a library for support vector machines. *ACM T. Syst. Techn.* 2, 1–27

7 Zhang, H. (2004) The optimality of Naive Bayes. FLAIRS conference