The Genetic Chain Rule for Probabilistic Kinship Estimation[1]

Brian S. Helfer[a], Philip Fremont-Smith[a], Darrell O. Ricke[a]

[a] *Bioengineering Systems and Technologies, MIT Lincoln Laboratory, Lexington, MA 02421 USA*

Abstract

Accurate kinship predictions using DNA forensic samples is limited to first degree relatives. High throughput sequencing of single nucleotide polymorphisms and short tandem repeats (STRs) can be used to expand DNA forensics kinship prediction capabilities. Current kinship identification models incorporate STR size profiles to statistical models that do not adequately depict genetic inheritance beyond the first degree, or machine learning algorithms that are prone to over optimization and requiring similar training data. This work presents an alternative approach using a computational framework that incorporates the inheritance of single nucleotide polymorphisms (SNPs) between specific relationships(patent pending)[1]. The impact of SNP panel size on predictions is visualized in terms of the distribution of allelic differences between individuals. The confidence of predictions is made by calculating log likelihood ratios. With a panel of 39108 SNPs evaluated on an in silico dataset, this method can resolve parents from siblings and distinguish 1st, 2nd, 3rd, and 4th degree relatives from each other and unrelated individuals.

Keywords: Single nucleotide polymorphism (SNP), Kinship Prediction, probabilistic model, familial search, high throughput sequencing

1. Introduction

High throughput sequencing (HTS) is revolutionizing capabilities in the fields of forensics, biology, and medicine. DNA forensics is evolving from sizing short tandem repeats (STRs) to sequencing STRs and single nucleotide polymorphisms (SNPs)[2, 3]. Currently, DNA forensics uses STRs sized by capillary electrophoresis to perform both individual identification and familial searching. Familial searching is employed when the database being searched does not contain an exact match to a query STR profile. The use of identity by state(IBS), and/or likelihood ratio based searches enable query STR profiles to match potentially related individuals[4]. In familial searching, likelihood ratio searches are often referred to as a kinship index(KI). Lineage testing is then performed using mitochondrial DNA or Y chromosome STRs to confirm paternal relationships. Familial searches are limited to first degree relatives due to the small number of STRs used (20 loci for US Combined DNA Index System - CODIS)[4], and the high probability of false positive matches when familial searching is expanded beyond first degree relations[5].

Ancestry prediction companies[6, 7, 8] use DNA SNP microarrays and the aforementioned methods to predict close and distant relatives. These DNA SNP microarrays require a lot more DNA than is typically available for forensic samples. Illumina SNP microarrays require 200 ng of input DNA[9] while as little as 1 ng of input DNA is recommended using Ion S5 HTS technology[10]. Small input DNA requirements enable the use of HTS technology when only trace DNA quantities exist at a crime scene.

Machine learning and forensic HTS SNP panels have been used to predict familial relationships across a set of three families[11]. This work trained a support vector machine based on features including the KING coefficient, IBS, and IBD[12]. This model was able to accurately predict all first degree and three quarters of second degree relationships. While machine learning models have the potential for accurate performance, they are highly dependent on the consistency of the training data, and are prone to over-optimization.

Enhanced kinship prediction capabilities can be obtained by incorporating genetic inheritance into a statistical framework applied to HTS data. This paper formalizes the expected relationship between any two individuals using the aforementioned approach with applications to HTS forensic SNP panels. The Genetic Chain Rule for Probabilistic Kinship Estimation provides a mathematical model that can predict likely relationship between two individuals. This model does not require training data, thereby increasing generalizability. Furthermore, this work reflects the biological underpinnings of inheritance allowing for improved kinship predictions.

[1]This material is based upon work supported by the Office of the Secretary of Defense under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of the Secretary of Defense.

2. Methods

2.1. Input Data

All results were tested on an in silico dataset that simulated millions of individuals from African American, Estonian, Korean, and Palestinian ethnic groups[13]. Minor allele frequencies for 39,108 SNPs well characterized across all four groups were taken from the Allele Frequency Database (ALFRED) [14]. The data were simulated across 9 generations with inter and within group marriage rates set to reflect public census data. The last four generations of data were used to simulate individuals with relationships spanning first through fifth degree and strangers.

2.2. Data Representation

All data are represented as a series of SNPs, with each locus having a minor allele, coupled with a minor allele frequency (mAF). The probability of the major allele occurring across a population is represented as p, while the probability of the minor allele occurring across a population is represented as q. As the SNPs analyzed have one major, and one minor allele, p, and q are set such that:

$$p + q = 1 \tag{1}$$

This follows Hardy-Weinberg equilibrium and leads to the number of people with a homozygous major genotype to occur with a frequency of p^2, the number of individuals with a heterozygous genotype to occur with a frequency of $2pq$, and the number of homozygous recessive genotypes to occur with a frequency of q^2. By satisfying Eq. 1, it is ensured that:

$$p^2 + 2pq + q^2 = 1 \tag{2}$$

Using this information, the conditional probability of any genotype occurring, given another individual of a known relationship having a certain genotype is derived.

2.3. Parent Child Relationships

2.3.1. Child Given Known Parent

The first relationship calculated is the probability of a child having a particular genotype G_c, given that their parent has a genotype G_{p1}. Given that one parent's genotype is known, the possible alleles that they could pass on to their child is also known. The probability of inheriting each allele from G_{p1} for the child is $\frac{1}{2}$. The probability for a G_c allele not in G_{p1} is zero. The genotype of the second parent is represented as G_{p2}. While the alleles for the second parent are not known, Pr($G_c$allele=A) $= p_2$ and Pr($G_c$allele=a) $= q_2$ for the child for alleles inherited from G_{p2}. Where A is a major allele, and a is a minor allele.

$$\begin{aligned} Pr(G_c|G_{p1}) = Pr(G_c allele1|G_{p1}) * Pr(G_c allele2|G_{p2}) \\ + Pr(G_c allele1|G_{p2}) * Pr(G_c allele2|G_{p1}) \end{aligned} \tag{3}$$

2.3.2. Parent Given Known Child

Leveraging the information presented in Table 2.3.1, it is possible to calculate the probability of a parent having a particular genotype G_p, given that the child has a known genotype G_c. This is formulated through an application of Bayes' rule.

$$Pr(G_c|G_{p1}) = \frac{Pr(G_c|G_p) * Pr(G_p)}{\sum_{G_i} Pr(G_c|G_p = G_i) Pr(G_p = G_i)} \tag{4}$$

In the above and subsequent equations, G_i, is used to represent all possible allele combinations. As a result, G_i can be expressed as:

$$G_i \in (AA, Aa, aa) \tag{5}$$

2.4. Sibling Relationships

It is possible to use this information to further compute the probability that a child will have a genotype G_{c1} given that a sibling of theirs has an observed genotype G_{c2}. In order to properly compute the probability of genotype G_{c1} occurring, it is essential to factor in genotypes of the two parents G_{p1}, and G_{p2}. Using this information, the desired sibling-sibling conditional probability is computed as the probability of Child 1 having a genotype G_{c1} given the possible genotypes that their parents could have, multiplied by the probability of the two parents having genotypes G_{p1} and G_{p2} given the known genotype of Child 2.

$$\begin{aligned} Pr(G_{c1}|G_{c2}) = \sum_{G_{i1}} \sum_{G_{i2}} Pr(G_{c1}|G_{p1} = G_{i1}, G_{p2} = G_{i2}) \\ * Pr(G_{p2} = G_{i1}, G_{p2} = G_{i2}|G_{c2}) \end{aligned} \tag{6}$$

Given an assumption that the two parents are not close relatives, it is possible to further rewrite Eq. 6 as:

$$\begin{aligned} Pr(G_{c1}|G_{c2}) = \sum_{G_{i1}} \sum_{G_{i2}} Pr(G_{c1}|G_{p1} = G_{i1}, G_{p2} = G_{i2}) \\ * Pr(G_{p1} = G_{i1}|G_{p2} = G_{i2}, G_{c2}) \\ * Pr(G_{p2} = G_{i2}|G_{c2}) \end{aligned} \tag{7}$$

As shown in Eq. 7, it becomes possible to compute the probability of a child having a particular genotype given that their sibling has a known genotype. This is done by computing the product of the probability of a child having genotype G_{c1} normalized across the genotypes for both parents G_{p1}, and G_{p2}, and the probability of each parent having a genotype given the knowledge of the second childs genotype. This equation is the product of three terms, where the first term represents the probability of a child having a genotype given specified genotypes for their parents, the second term represents the probability of the first parent having a specified genotype given the

Ind_α	Ind_β	$Pr(Ind_\beta = Child \mid Ind_\alpha = Parent_1)$ $\text{Parent}_1 : p_1, q_1 Parent_2 : p_2, q_2$	$Pr(Ind_\beta = Parent_1 \mid Ind_\alpha = Child)$ $\text{Parent}_1 : p_1, q_1 Parent_2 : p_2, q_2$
AA	AA	p_2	$\frac{p_2 * p_1^2}{p_2 * p_1} = p_1$
AA	Aa	q_2	$\frac{0.5p_2 * 2p_1q_1}{p_2 * p_1} = q_1$
AA	aa	0	0
Aa	AA	$0.5p_2$	$\frac{q_2 * p_1^2}{p_1q_2 + q_1p_2}$
Aa	Aa	$0.5p_2 + 0.5q_2 = 0.5$	$\frac{(0.5p_2 + 0.5q_2) * 2p_1q_1}{p_1q_2 + q_1p_2} = \frac{p_1q_1}{p_1q_2 + q_1p_2}$
Aa	aa	$0.5q_2$	$\frac{p_2 * q_1^2}{p_1q_2 + q_1p_2}$
aa	AA	0	0
aa	Aa	p_2	$\frac{0.5q_2 * 2p_1q_1}{q_1 * q_2} = p_1$
aa	aa	q_2	$\frac{q_2 * q_1^2}{q_1 * q_2} = q_1$
Ind_α	Ind_β	$Pr(Ind_\beta = child_1 \mid Ind_\alpha = child_2)$ $\text{Parent}_1 : p_1, q_1, Parent_2 : p_2, q_2$	$Pr(Ind_\beta = Grandchild \mid Ind_\alpha = Grandparent_1)$ $\text{Parent}_1 : p_1, q_1, Parent_2 : p_2, q_2,$ $\text{Grandparent}_1 : p_3, q_3, Grandparent_2 : p_4, q_4$
AA	AA	$p_1p_2 + 0.5p_1q_2 + 0.5q_1p_2 + 0.25q_1q_2$	$p_2p_4 + 0.5p_2q_4$
AA	Aa	$0.5(p_1q_2 + q_1p_2 + q_1q_2)$	$q_2p_4 + 0.5q_4$
AA	aa	$0.25q_1q_2$	$0.5q_2q_4$
Aa	AA	$0.5(p_1p_2 + \frac{p_1q_1p_2q_2}{p_1q_2 + q_1p_2})$	$0.5p_2p_4 + 0.25p_2$
Aa	Aa	$0.5(p_1p_2 + q_1q_2) + \frac{p_1^2q_2^2 + p_1q_1p_2q_2 + q_1^2p_2^2}{p_1q_2 + q_1p_2}$	$0.5p2q_4 + 0.25 + 0.5q_2p_4$
Aa	aa	$0.5(q_1q_2 + \frac{p_1q_1p_2q_2}{p_1q_2 + q_1p_2})$	$0.25q_2 + 0.5q_2q_4$
aa	AA	$0.25p_1p_2$	$0.5p_2p_4$
aa	Aa	$0.5(p_1q_2 + q_1p_2 + p_1p_2)$	$0.5p_4 + p_2q_4$
aa	aa	$q_1q_2 + 0.5p_1q_2 + 0.5q_1p_2 + 0.25p_1p_2$	$0.5p_2q_4 + q_2q_4$

Table 1: Probability of the event (α) an individual with a given genotype (Ind_α), conditioned (β) on another individual (Ind_β) having a given genotype. The genotype letter (**A**) represents the major allele with population frequency p_i for individual i, while genotype letter (**a**) represents the minor allele with population frequency q_i; this allows individuals to have different ethnicities.

child's sibling and their other parent's genotype, and the third term represents the probability of the second parent having a particular genotype given the child's sibling.

2.5. Bayesian Chain Rule of Kinship

The above framework can be generalized to compute the probability of a particular genotype given any relationship between two people. This formulation is defined as the Bayesian Chain Rule of Kinship. The Bayesian Chain Rule of Kinship expresses any relationship between individuals as the product of a series of relationships. For instance, if one wished to compute the cousin relationship between M_{31} and F_{33} as shown in Fig. 1, one would represent this as the relationship between child and parent, parent and sibling, and the parent's sibling and their child. As can be noted, all components of the chain rule take the form of child given parent to move up the tree, sibling given sibling to move across the tree, and parent given child to move down the tree. These operations allow for complete navigation between any two individuals. Expressed another way:

$$Pr(P_A|P_B) = Pr(P_A|x_1)Pr(X_n|P_B)\prod_{x=2}^{n} Pr(x-1|x) \quad (8)$$

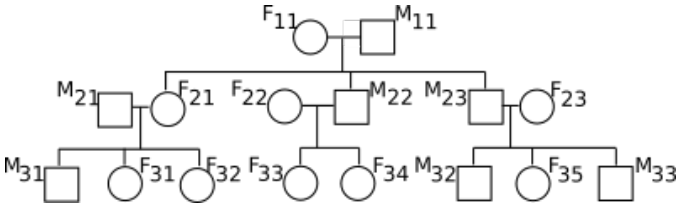

Figure 1: Reference Family Tree

This takes the product of all people between two individuals, and uses the Bayesian Chain Rule of Kinship to compute a probability of a genotype given a particular relationship.

2.6. Extended Relationships

Extended relationships can be computed using the previously defined Bayesian Chain Rule of Kinship.

Figure 1 shows a family tree where each individual is identified as male (M) or female (F), and with two indices identifying their generation, along with a unique identifier for that individual within the generation. For instance, F_{23} represents the third unique woman appearing in the second generation.

2.6.1. Grandchild Given Grandparent

The probability of a child (M_{31}) having a given genotype G_c, given their grandparent (M_{11}) has a known genotype G_g can be computed using the Markov and chain rule assumptions to model the child (M_{31}) as dependent on their parent (F_{21}), and the parent (F_{21}) to be dependent on the grandparent (M_{11}).

$$Pr(G_c|G_g) = \sum_{G_i} Pr(G_c|G_p = G_i) * Pr(G_p = G_i|G_g) \tag{9}$$

Since child(M_{31}) can only inherit DNA from parent(F_{21}), and parent(F_{21}) can only inherit DNA from grandparent(M_{11}), the probability equation decomposes into the child being directly dependent on their parent, and the parent being directly dependent on the grandparent. It is unnecessary to condition the child's genotype on the grandparent's genotype, as that is already factored into the parent's genotype. Given that the parent has an unknown genotype, G_i is used to marginalize over all possible genotypes for that parent.

2.6.2. Child Given Aunt/Uncle

The same principles apply to identify the likelihood that a child will have a genotype given that their aunt/uncle have a known genotype G_{au}. In this case, the probability of the child's genotype is decomposed into the relationship between child and parent, and parent and sibling.

$$Pr(G_c|G_{au}) = \sum_{G_{i1}} \sum_{G_{i2}} Pr(G_{c1}|G_{p1} = G_{i1}, G_{p2} = G_{i2}) * Pr(G_{p1} = G_{i1}|G_{c2}) * Pr(G_{p2} = G_{i2}|G_{c2}) \tag{10}$$

2.7. Log Likelihood Calculation

The above formulations can be further used to calculate the log likelihood of two individuals having a particular relationship given the observed data. The log likelihood is defined as the probability of data (D) given a hypothesis (H).

$$L = log(Pr(D|H)) \tag{11}$$

In the case of familial identification, this is computed by taking the product of all conditional probabilities across SNPs. This allows for the computation of the likelihood of any relationship given the observed genotypes of two individuals.

2.8. Current Limitations

The current calculations rely on the independence of inheritance of all alleles. This simplifies the calculation; however, it does not account for haploblocks, or sex chromosomes. As a result, it is not possible to distinguish between different relationships that are two or more generations apart, or the direction of the relationship (e.g. Parent given child, versus child given parent). However, this framework is generalizable, and fully capable of incorporating haploblocks and sex chromosome SNPs.

3. Results

The previously defined mathematical relationships were validated using the the silico database of four ethnicities and the lower four of the nine generations[13]. The data was further subdivided into four different ethnic groups which have separate mAF values across the 39,108 sampled SNPs.

3.1. Data Relationship Separability

The relationship separability was examined as a function of the number of differences across SNPs. A difference was defined as the number of discordant alleles at each locus with a value between zero and two. The number of discrepancies was summed across all SNPs for a single pairwise relationship. This was then done for one thousand examples of each relationship. A kernel density estimate was fitted to this distribution and then shown in the figures below.

The number of differences across degree were plotted while varying the number of SNPs used in the comparison. Fig 2 plots differences across distributions using the full panel of 39k SNP loci. The level of separation is then plotted for half this panel, utilizing 20k SNPs as shown in

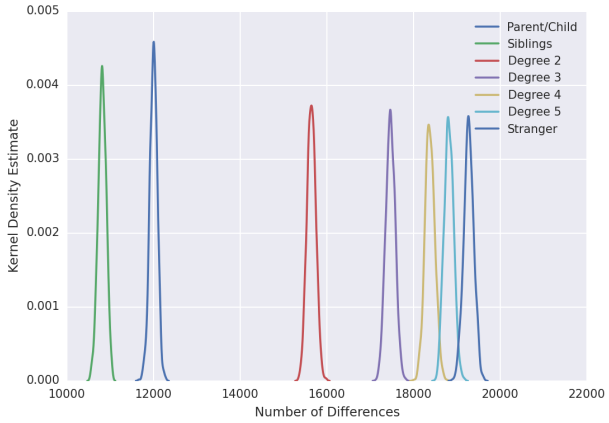

Figure 2: Differences separated by degree of relationship for 39k SNP Panel

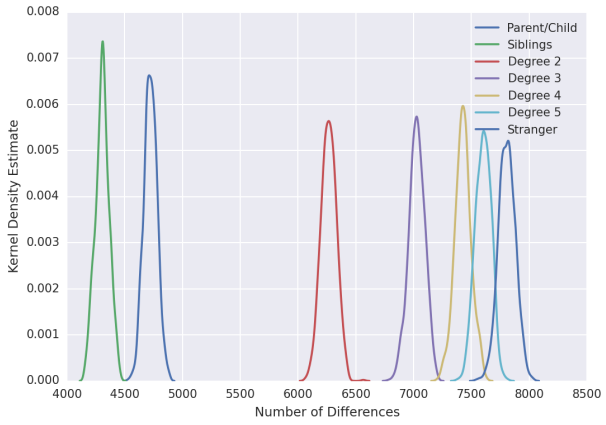

Figure 3: Differences separated by degree of relationship for 20k SNP Panel

Fig. 3, and finally the number of differences is examined with the panel reduced to only 2k SNPs, as shown in Fig. 4.

3.2. Log Likelihood Prediction

After examining differences between individuals, the log likelihood was then used to predict the degree of relation across pairs of individuals. At each pair, the algorithm identifies if it is a parent-child relationship, a sibling-sibling relationship, 2^{nd} to 5^{th} degree relationship, or two unrelated individuals. The performance for this assessment is shown in the Table 2.

4. Discussion

In this study, in silico data were used to identify the degree of relatedness between individuals spanning four generations. Figure 2 demonstrated a clear separability between parent-child relationships and siblings, as well as

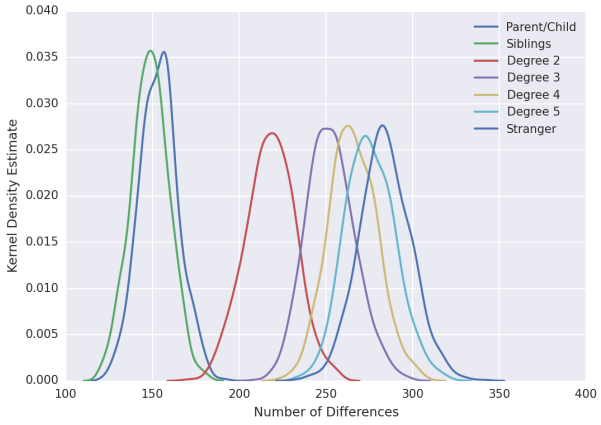

Figure 4: Differences separated by degree of relationship for 2k SNP Panel

Table 2: Confusion matrix for degree prediction

	Parent-Child	Sibling	Degree 2	Degree 3	Degree 4	Degree 5	Unrelated
Parent-Child	2000	0	0	0	0	0	0
Sibling	0	1000	0	0	0	0	0
Degree 2	0	0	4000	0	0	0	0
Degree 3	0	0	0	4997	3	0	0
Degree 4	0	0	0	1	949	50	0
Degree 5	0	0	0	0	53	903	44
Stranger	0	0	0	0	0	46	954

between individuals with second, third, and unrelated levels of relationship. This was reflected by the use of log likelihood values to fully and correctly identify the difference between individuals of these different degrees. As the degree increases, the curves become closer together. The upper tail of the 4th degree relatives is near the lower tail of unrelated individuals. For the 39k SNP panel, the distribution for 5th degree relatives overlap the distributions for 4th degree relatives and unrelated individuals. Larger SNP panels are required to separate 5th degree relatives from 4th degree relatives and unrelated individuals. The confusion matrix shown in Table 2 for this method illustrates the high accuracy on these in silico pedigrees. The impact of reducing the SNP panel size is illustrated in Figure 3 for 20k SNPs and Figure 4 for 2k SNPs. For the 2k SNPs panel, the different relationships become much less separable. The curves also become wider as a function of racial heterogeneity/admixture. As the amount of mixed ancestry increased the standard deviation of the distributions also increases. This also increases the difficulty of distinguishing levels of relationship in less related individuals.

5. Conclusion

In this work, we present a Bayesian framework for identifying the level of relation between different individuals.

This framework builds on the biology of inheritance, along with Bayesian statistics to predict degree of relation without requiring a training database or parameter optimization. This allows for further improvement by incorporating more biological properties into the model.

References

[1] D. O. Ricke, DNA mixtures from one or more sources and methods of building individual profiles therefrom, US patent pending 62/534,590.

[2] B. Sobrino, M. Brión, A. Carracedo, SNPs in forensic genetics: a review on SNP typing methodologies, Forensic science international 154 (2) (2005) 181–194.

[3] A. J. Pakstis, W. C. Speed, J. R. Kidd, K. K. Kidd, Candidate SNPs for a universal individual identification panel, Human Genetics 121 (3-4) (2007-05-01) 305(13).

[4] J. Ge, R. Chakraborty, A. Eisenberg, B. Budowle, Comparisons of familial DNA database searching strategies, Journal of Forensic Sciences 56 (6) (2011) 1448–1456. doi:10.1111/j.1556-4029.2011.01867.x.
URL http://dx.doi.org/10.1111/j.1556-4029.2011.01867.x

[5] E. Niedzwiecki, S. Debus-Sherrill, M. B. Field, Understanding familial DNA searching: Coming to a consensus on terminology.

[6] Ancestry, https://www.ancestry.com/.

[7] 23andme, https://www.23andme.com/.

[8] Parabon snapshot, https://snapshot.parabon-nanolabs.com/.

[9] Illumina, https://www.illumina.com/techniques/popular-applications/genotyping/targeted-genotyping.html.

[10] Ion torrent s5, https://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-s5-ngs-targeted-sequencing.html.

[11] A. Shcherbina, D. O. Ricke, E. Schwoebel, T. Boettcher, C. Zook, J. Bobrow, M. Petrovick, E. Wack, Kinlinks: Software toolkit for kinship analysis and pedigree generation from HTS datasets, in: Technologies for Homeland Security (HST), 2016 IEEE Symposium on, IEEE, 2016, pp. 1–6.

[12] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W.-M. Chen, Robust relationship inference in genome-wide association studies, Bioinformatics 26 (22) (2010) 2867–2873.

[13] B. Helfer, D. O. Ricke, SNP identity searching across ethnicities, kinship, and admixture, in preparation.

[14] K. Cheung, M. Osier, J. Kidd, A. Pakstis, P. Miller, K. Kidd, ALFRED: an allele frequency database for diverse populations and DNA polymorphisms, Nucleic Acids Res. 28 (1) (2000) 361–363.