ARTICLE

# PADRE: Pedigree-Aware Distant-Relationship Estimation

Jeffrey Staples,[1] David J. Witherspoon,[2] Lynn B. Jorde,[2] Deborah A. Nickerson,[1] the University of Washington Center for Mendelian Genomics,[1] Jennifer E. Below,[3,5,*] and Chad D. Huff[4,5,*]

Accurate estimation of shared ancestry is an important component of many genetic studies; current prediction tools accurately estimate pairwise genetic relationships up to the ninth degree. Pedigree-aware distant-relationship estimation (PADRE) combines relationship likelihoods generated by estimation of recent shared ancestry (ERSA) with likelihoods from family networks reconstructed by pedigree reconstruction and identification of a maximum unrelated set (PRIMUS), improving the power to detect distant relationships between pedigrees. Using PADRE, we estimated relationships from simulated pedigrees and three extended pedigrees, correctly predicting 20% more fourth- through ninth-degree simulated relationships than when using ERSA alone. By leveraging pedigree information, PADRE can even identify genealogical relationships between individuals who are genetically unrelated. For example, although 95% of 13th-degree relatives are genetically unrelated, in simulations, PADRE correctly predicted 50% of 13th-degree relationships to within one degree of relatedness. The improvement in prediction accuracy was consistent between simulated and actual pedigrees. We also applied PADRE to the HapMap3 CEU samples and report new cryptic relationships and validation of previously described relationships between families. PADRE greatly expands the range of relationships that can be estimated by using genetic data in pedigrees.

## Introduction

Accurate prediction and verification of relationships among individuals is essential in a variety of genetic studies. Errors in pedigrees are common[1–3] and have adverse consequences, including biased phasing and family-based imputation results, inaccurate identification of Mendelian errors, and reduction of power to detect linkage[4] or family-based associations. Therefore, ensuring that the genetic relationships among the DNA samples match the reported pedigree structure is critical for accurate family-based genetic analysis.[5] Detecting cryptic relationships can be important as well.[6] Genetic relationships identified in population studies can be leveraged for improved haplotype phase inference, detection of population structure, genotype imputation, and study designs such as identical-by-descent (IBD) mapping and tests to detect multiple rare and common variants that contribute to disease.[7–13] The identification of relatives also plays an important role in forensics in criminal investigations,[14] identification of victims of mass disaster,[15] and discovery of family history.

With close relationships (first through third degree), pedigree reconstruction can provide the kinship structure of the individuals in a genetic dataset.[5] However, genetic datasets often contain relationships that are more distant than third degree, resulting in sparsely connected pedigrees that are unsuitable for reconstruction. Algorithms that consider IBD segment data, such as ERSA (estimation of recent shared ancestry),[16,17] can accurately predict pairwise relationships up to ninth-degree relatives (e.g., fourth cousins), but do not reconstruct pedigrees, nor can they utilize information from known or observed pedigree structures in the data. Here, we introduce PADRE (pedigree-aware distant-relationship estimation), which leverages the pedigree reconstruction of known or cryptic first- to third-degree relatives by PRIMUS (pedigree reconstruction and identification of a maximum unrelated set)[5] along with the accurate distant relationship predictions by ERSA.[17] PADRE, which has been implemented as an extension of PRIMUS and ERSA, uses ERSA-generated relationship likelihoods to identify the highest composite likelihood connection between family networks reconstructed by PRIMUS (Figure S1), significantly improving the accuracy of the predictions and expanding the range of relationships that can be predicted.

## Subjects and Methods

### PADRE Algorithm

PADRE combines reconstructed pedigree information with distant pairwise relationship predictions to identify distant relationships between pedigrees and requires results from PRIMUS (v.1.8.0) and ERSA (v.2.1) as input (Figure S1). PRIMUS identifies family-based networks of individuals within a dataset, where each family network consists of the set of individuals with a detected first- through third-degree relationship to at least one other individual in the network. When PRIMUS reconstructs a dataset into family networks (i.e., $Net_1$, $Net_2$, … $Net_n$), each network can be represented by one or more possible pedigrees that fit the genetic data, annotated here with subscripts (i.e., $Net_{11}$, $Net_{12}$, … and $Net_{1j}$). PADRE tests for significant relationships between each

[1] Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA; [2] Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA; [3] Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [4] Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

[5] These authors contributed equally to this work

*Correspondence: jennifer.e.below@uth.tmc.edu (J.E.B.), chad@hufflab.org (C.D.H.)

http://dx.doi.org/10.1016/j.ajhg.2016.05.020.

pair of networks by using the likelihood ratio test in ERSA. This test compares the null model that two founders in two different networks, $x$ and $y$, are unrelated to the alternative model that the two individuals are $N^{\text{th}}$-degree relatives. If a significant relationship is detected between any two founders, PADRE will then identify the best fitting relationship between the two networks by using a composite likelihood framework. The significance threshold is 0.05 by default but can be adjusted for the number of founder-to-founder relationships tested by application of a Bonferroni correction with the command-line argument adding "–PADRE_multiple_test_correct." For a significant relationship detected between founders $x$ and $y$, PADRE calculates the maximum composite likelihood for each possible $N^{\text{th}}$-degree relationship between $x$ and $y$ by multiplying the cross-network pairwise relationship likelihoods in ERSA for each pair of pedigrees $Net_{1i}$ and $Net_{2j}$:

$$\widehat{L}_{1i2j}(x,y\,|\,N) = L_{Net_{1i}}L_{Net_{2j}} \prod_{\substack{\forall a\in Net1_i \\ \forall b\in Net2_j}} \widehat{L}_{ab}(s_{ab}\,|\,D_{ab}), \qquad \text{(Equation 1)}$$

where $a$ and $b$ are individuals in the pedigrees $Net_{1i}$ and $Net_{2j}$, respectively, $D_{ab}$ is the degree of relatedness between $a$ and $b$ given the two pedigrees and that founders $x$ and $y$ are $N^{\text{th}}$-degree relatives, and $s_{ab}$ is a set containing the length of each detected IBD segment between $a$ and $b$. $\widehat{L}_{ab}(s_{ab}|D_{ab})$ is the maximum likelihood of the observed IBD segments shared by $a$ and $b$ conditioned on the degree of relationship distance $D_{ab}$ specified by $N$, $Net_{1i}$ and $Net_{2j}$. $L_{Net1i}$ and $L_{Net2j}$ are the composite pedigree likelihoods consisting of the product of the PRIMUS likelihoods for each pairwise relationship specified by $Net_{1i}$ and $Net_{2j}$, respectively. When $D_{ab}$ is less than 10, $\widehat{L}_{ab}$ includes two additional estimated parameters compared to the model with no relationship (relationship distance and number of shared segments conditioned on relationship distance); these likelihoods are calculated by ERSA.[16]

Because many 10th- and most 11th-degree human relatives share no autosomal IBD segments from their most recent common ancestor,[16] models involving relationships more distant than ninth degree require special consideration. Although such models also include two additional parameters, the maximum likelihood estimate for the number of shared genetic segments is typically 0, resulting in a compressed free parameter space. Maximizing the likelihood of such models without accounting for the reduced free parameter space over-penalizes such distant relationships. We address this problem with the following approximation. Given that two individuals, $a$ and $b$, are genetic ninth-degree relatives, the unconditional maximum likelihood of a 10th-degree relationship for individual $a$ and the offspring of individual $b$ is as follows: with 50% probability, the shared segment is inherited by the offspring of individual $b$, and the likelihood is equal to the ninth-degree relationship likelihood. Otherwise, the likelihood is equal to the unrelated likelihood. This approximation holds for relationship distances detectable by PADRE and beyond (see Figure S2) and leads to the following formula to approximate the likelihood of 10th-degree and more-distant relationships in PADRE:

$$\begin{aligned}\widehat{L}_{ab}(s_{ab}\,|\,D_{ab} > 9) = {} & (0.5)^{D_{ab}-9}\widehat{L}_{ab}(s_{ab}\,|\,D'_{ab}=9) \\ & + \left(1-(0.5)^{D_{ab}-9}\right)\widehat{L}_{ab}(s_{ab}\,|\,\text{unrelated}),\end{aligned} \qquad \text{(Equation 2)}$$

where the effective degrees of freedom is given by:

$$g(D_{ab}) = \begin{bmatrix} 2, & D_{ab}\le 9 \\ (0.5)^{D_{ab}-9}, & D_{ab} > 9 \end{bmatrix}. \qquad \text{(Equation 3)}$$

To identify the best fitting model, PADRE calculates the composite likelihood Akaike information criterion (CL-AIC) for each possible fourth- through ninth-degree relationship between the founders of the two family networks via Equation 1.[18] Because each network could have more than one possible pedigree, we evaluate all pairs of possible pedigrees identified by PRIMUS for each network and identify the pair of pedigrees that minimizes the CL-AIC of the two networks. For a given pair of pedigrees $Net_{1i}$ and $Net_{2j}$, the CL-AIC is calculated according to Equation 4:

$$AIC_{1i2j}(x,y\,|\,N) = 2k_{1i2j}(x,y,N) - 2ln\widehat{L}_{1i2j}(x,y\,|\,N) - lnL_{Net_{1i}} - lnL_{Net_{2j}}, \qquad \text{(Equation 4)}$$

where $k$ is equal to the effective number of parameters in $\widehat{L}_{1i2j}(x,y|N)$. The value for $k$ is given by Equation 5:

$$k_{1i2j}(x,y,N) = \sum_{\substack{\forall a\in Net1_i \\ \forall b\in Net2_j}} g(D_{ab}). \qquad \text{(Equation 5)}$$

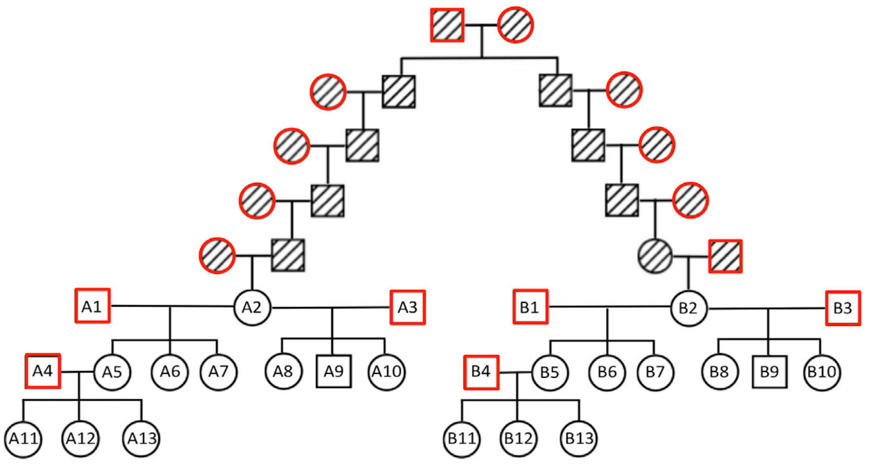

**Figure 1. Pedigree Structure Used to Simulate Ninth-Degree Pedigrees**
100 ninth-degree pedigrees, each with different genotypes, were generated with A2 and B2 related as ninth-degree relatives. The same pedigree structures for samples A1–A13 and B1–B13 were also used to generate 100 pedigrees, each with different genotypes, where A2 and B2 were fourth-, fifth-, sixth-, seventh-, eighth-, and ninth-degree relatives. The number of ancestral relatives was adjusted to account for the different degree of relatedness.

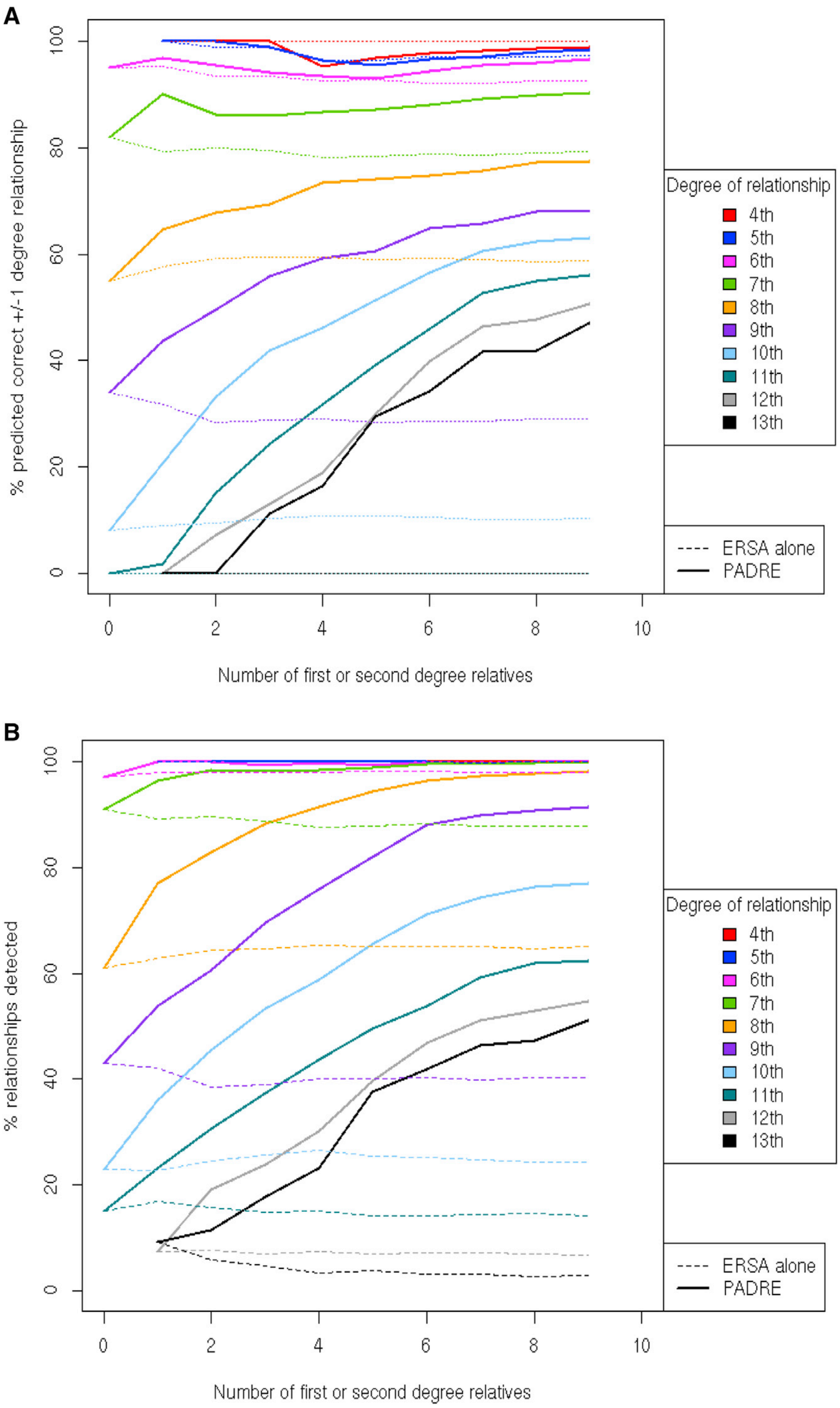

**Figure 2. Comparison of PADRE versus ERSA Accuracy in Simulated Data by Degree of Relationships**

(A) The observed accuracy in the pedigree predictions increases as additional first- and second-degree relatives are added.

(B) Power of PADRE and ERSA alone to detect simulated relationships as additional first- and second-degree relatives were added to the pedigree. The ERSA results fluctuate slightly because additional pairwise estimates are added as more individuals are included, as described with the generation of the simulations.

Finally, PADRE evaluates all pairs of possible pedigrees identified by PRIMUS for each network to identify the model that minimizes the CL-AIC of the two networks:

$$AIC_{\min}(Net1, Net2) = \min_{\substack{x \in Net1 \\ y \in Net2 \\ 4 \le N \le 9 \\ 1 \le i \le Net1_n \\ 1 \le j \le Net2_n}} AIC_{1i2j}(x, y \mid N). \qquad \text{(Equation 6)}$$

For each pair of family networks, PADRE reports the pair of founders, their degree of relatedness, and the two pedigrees from the model specified by the $AIC_{\min}$. In a separate output file, PADRE provides the degree of relatedness between each pair of samples in the model specified by the $AIC_{\min}$.

PADRE takes, as a command line option, the maximum degree of relatedness PRIMUS uses to reconstruct and then adjusts the range of ERSA predictions to test all relationships greater than the maximum degree of relatedness in PRIMUS. By default, the maximum degree of relatedness is three, and PADRE thus considers all fourth- through ninth-degree relationships in ERSA.

## Simulations

We simulated pedigrees to evaluate the accuracy and relative benefit of using PADRE to detect distant relationships. We used two identical 13-person, three-generation pedigree structures and connected a founder in each pedigree by varying the number of generations to their recent common ancestor. Figure 1 illustrates a simulated pedigree in which founders A2 and B2 are ninth-degree relatives. To test the full range of predictions beyond the third degree, we generated versions of the pedigree in which individuals A2 and B2 are fourth- through ninth-degree relatives. For each of these versions of simulated pedigree structures, we created 100 different sets of genotypes by using the method described in Morrison.[19] We randomly selected haplotypes with ~1 M SNPs from among the unrelated HapMap3[20] CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) samples and assigned them to the all founders (individuals with red symbols in Figure 1). The unrelated set of CEU samples was determined by running ERSA (v.2.1) on all the HapMap3 CEU samples and then running the IMUS algorithm within PRIMUS[21] to identify the maximum unrelated set of individuals. We then used Morrison's recombination simulation software to propagate the founder genotypes through the pedigree. This method simulates recombination events as a homogeneous Poisson process by using the genetic map provided with the HapMap3 data, disregarding the centromere. Genotypes were removed for all individuals not included in either of the 13-person pedigrees. IBD estimates were calculated with PLINK v.1.9[22]

```
plink --file [data_file_root_name] --genome --maf 0.05
        --geno 0.1 --out[data_file_root_name],
```

and all simulated pedigrees were reconstructed with PRIMUS (v.1.8):

```
run_PRIMUS.pl --p [data_file_root_name].genome.
```

We obtained ERSA (v.2.1) results for each simulation as described below.

To test improvements in relationship predictions by PADRE as the size and density of genotyped individuals increased, we first used PRIMUS, ERSA, and PADRE to analyze individuals A6 and B6 in each simulated pedigree (Figure 1). We repeated the analyses iteratively, including genotypes of an additional randomly selected first- or second-degree relative of A6 and B6 in each iteration. We then performed a final analysis using all 13 individuals in each pedigree (see Figure 1).

### Runtime

We evaluated PADRE runtime on a machine with Intel Xeon CPU E5-2670 v.2 at 2.50 GHz with 14 GB of memory, subtracting the time needed to load the ERSA likelihood files. The number of comparisons is the number of pairwise likelihoods that were looked up during the PADRE analysis and is the single best estimate of runtime. Each comparison is conducted at the lowest level of five nested for loops: (1) for each pair of networks, (2) for each pair of possible pedigrees within the networks, (3) for each pair of founders between each of the pedigrees in different networks, (4) for degrees of relatedness between the fourth and ninth degrees, and (5) for each pair of non-missing individuals between the two pedigrees.

The variability in the comparisons per second is due to variability in the other PADRE calculations. PRIMUS reconstruction was unable to complete for all family networks when it was run on the European ancestry dataset using third-degree relationships as a cutoff because some family networks resulted in too many possible pedigree structures consistent with the genetic dataset. The results of these runtime comparisons are shown in Table S1.

### Extended Pedigree Samples

We analyzed Affymetrix 6.0 SNP microarray data on 169 individuals from three previously described extended pedigrees with predominantly northern European ancestry.[16] The three pedigrees were validated as described in Huff et al.,[16] are composed of 24, 30, and 115 genotyped individuals, and included a total of 7,266 previously described relationships between pairs of individuals.

### HapMap3 CEU Samples

Using 165 CEU individuals from HapMap3 release 2[20] obtained from the HapMap website (see Web Resources), we reconstructed pedigree structures in this dataset with PRIMUS as described below by using the default settings. We applied a Bonferroni correction when detecting initial relationships between family networks identified in PRIMUS of $p = 5.5 \times 10^{-6}$ (0.05/9,074 founder-to-founder relationships).

### Pedigree Reconstruction with PRIMUS

PRIMUS uses genome-wide IBD estimates to identify families and reconstruct all possible pedigrees that fit the genetic data by using relationships as distant as third-degree relatives. We used the pre-PRIMUS IBD pipeline[5] to generate genome-wide average IBD estimates between all samples in each pedigree and used PRIMUS (v.1.8) to reconstruct pedigrees. The command line options used were "--file [data_file_root_name] and --genome." Due to the sparse number of individuals genotyped in the three European ancestry pedigrees and in many of the simulations which lead to long runtimes in PRIMUS, we applied a relatedness threshold of second degree in PRIMUS to both datasets by adding the command line option "--degree_rel_cutoff 2." We used the default relatedness cutoff of third-degree relatives for the HapMap3 CEU dataset.[20]

### Distant Relationships Prediction with ERSA

We applied the IBD detection pipeline described by Glusman et al.[17] by first phasing all genetic data with Beagle (v.3.3.2)[11] by using the phasing pipeline provided on the GERMLINE website (see Web Resources). We analyzed the phased data in GERMLINE (v.1.4.0)[23] for each chromosome with the following command:

```
germline –homoz –err_het 1 –err_hom 2 –map
[data_root_name_chrN].map –min_m 2.5  <
[data_root_name_chrN_options].txt
```

We analyzed the GERMLINE output files with ERSA (v.2.1) to calculate the likelihood of each possible pairwise relationship (from the first through 39th degrees) among all samples in the dataset. We controlled for potential false-positive IBD segments by masking genomic regions from the 1000 Genomes Project[24] CEU samples with greater than a 4-fold excess of pairwise IBD (mask_region_threshold = 4) as previously described:[17]

```
ersa --segment_files = [sample_data_germline.match_files] --
model_output_file model_likelihoods.txt --output_file =
ersa_results --
confidence_level 0.999 --mask_common_shared_regions true --
control_files = [CEU_germline.match_files]
```

## Results

To evaluate the improvements in relationship prediction, we ran PADRE on 600 simulated pedigrees, each with ten different patterns of genotyped individuals, and compared the accuracy of the resulting pairwise relationship predictions (see Subjects and Methods). Figure 2A shows that PADRE and ERSA alone exhibited the same accuracy when the individuals had no other first- or second-degree relatives in the pedigree. However, as additional genotyped individuals were included in the pedigrees, PADRE accurately predicted up to 56% more of the simulated relationships. In addition to higher relationship prediction accuracy, Figure 2A demonstrates that PADRE predicted relationships that are undetectable by methods that consider only pairwise genetic data. For example, PADRE

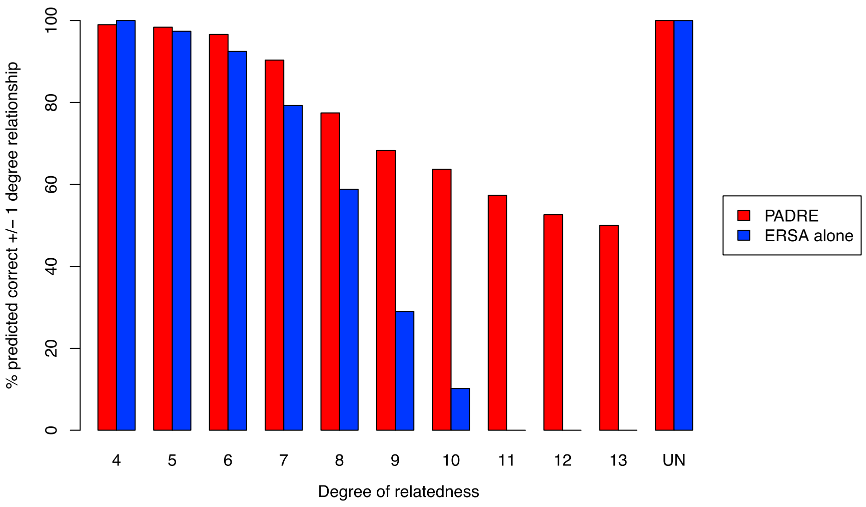

**Figure 3. PADRE and ERSA Prediction Accuracy on Simulated Pedigrees Where All Individuals Have Been Genotyped**
PADRE more accurately predicts fifth- through tenth-degree relationships relative to ERSA and frequently identifies 11th- through 13th-degree relatives who were undetectable in ERSA.

detected over 50% of 13th-degree relationships, although 95% of 13th-degree relatives share no genetic material through their most recent common ancestors (in humans). PADRE provided a substantial increase in power by correctly detecting up to 83% of seventh- through 13th-degree relationships in our simulations (Figure 2B).

Figure 2 displays the ERSA and PADRE results for simulated pedigrees as large as 20 individuals, and Figure 3 summarizes the results for the simulated pedigrees with all 26 individuals. PADRE predicted the exact degree of relationship for 20% additional fourth- through ninth-degree relationships, relative to ERSA alone. For 10th- through 13th -degree relationships, ERSA accurately predicted only 4% of relationships to within one degree. In comparison, PADRE accurately predicted 59% of the simulated 10th- through 13th-degree relationships to within one degree, even though approximately 71% of such relatives share no DNA segments that are IBD (additional comparative data are shown in Figure S3). This can be accomplished because genetic relationships across pedigree founders propagate through pedigrees, allowing for multiple pairwise comparisons, which improves accuracy and results in accurate estimates of distant genealogical relationships even in pairs of descendants who inherited no genomic segments in common. Thus, by utilizing the pairwise sharing across all members of both pedigrees, PADRE is frequently able to predict very distant genealogical relationships between deeply genotyped pedigrees that are undetectable from single pairwise genetic comparisons.

We also analyzed Affymetrix 6.0 microarray data from 169 individuals in three previously described extended pedigrees with predominantly northern European ancestry.[16] The three pedigrees were composed of 24, 30, and 115 genotyped individuals and included a total of 7,266 pairs of related individuals. As expected, ERSA and PADRE attained the same accuracy for pairs of individuals with no genotyped first- or second-degree relatives. However, when we considered pairs of individuals who had

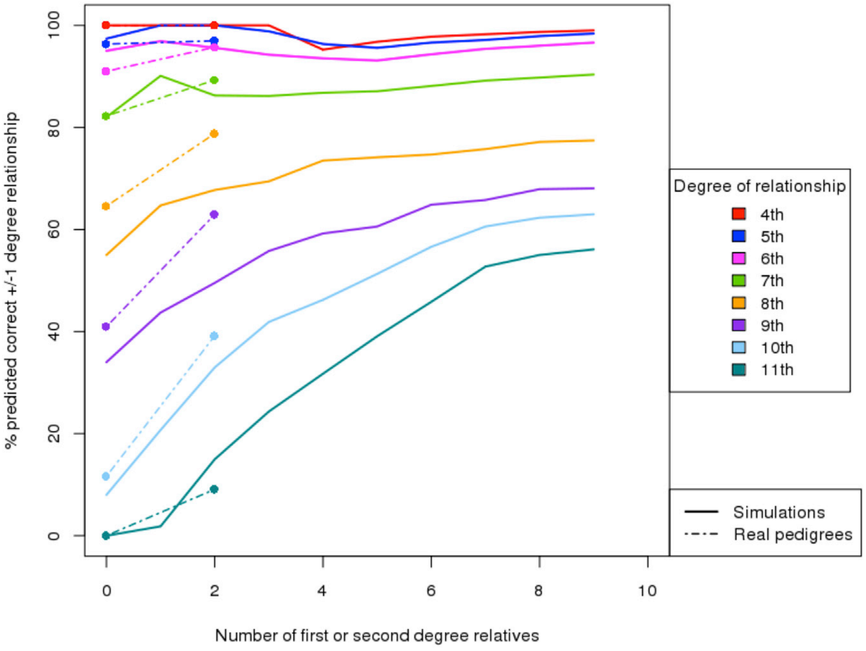

**Figure 4. Percentage of Relationships Correctly Predicted by PADRE to Within ± One Degree in Real Pedigrees of European Ancestry and Simulated Pedigrees**
Relationship detection accuracy was broadly consistent between the real and simulated pedigrees. Because the real pedigrees included two or fewer first- or second-degree relatives, PADRE's estimated relationship detection accuracy for pedigrees with three or more sampled relatives is based solely on simulated data.

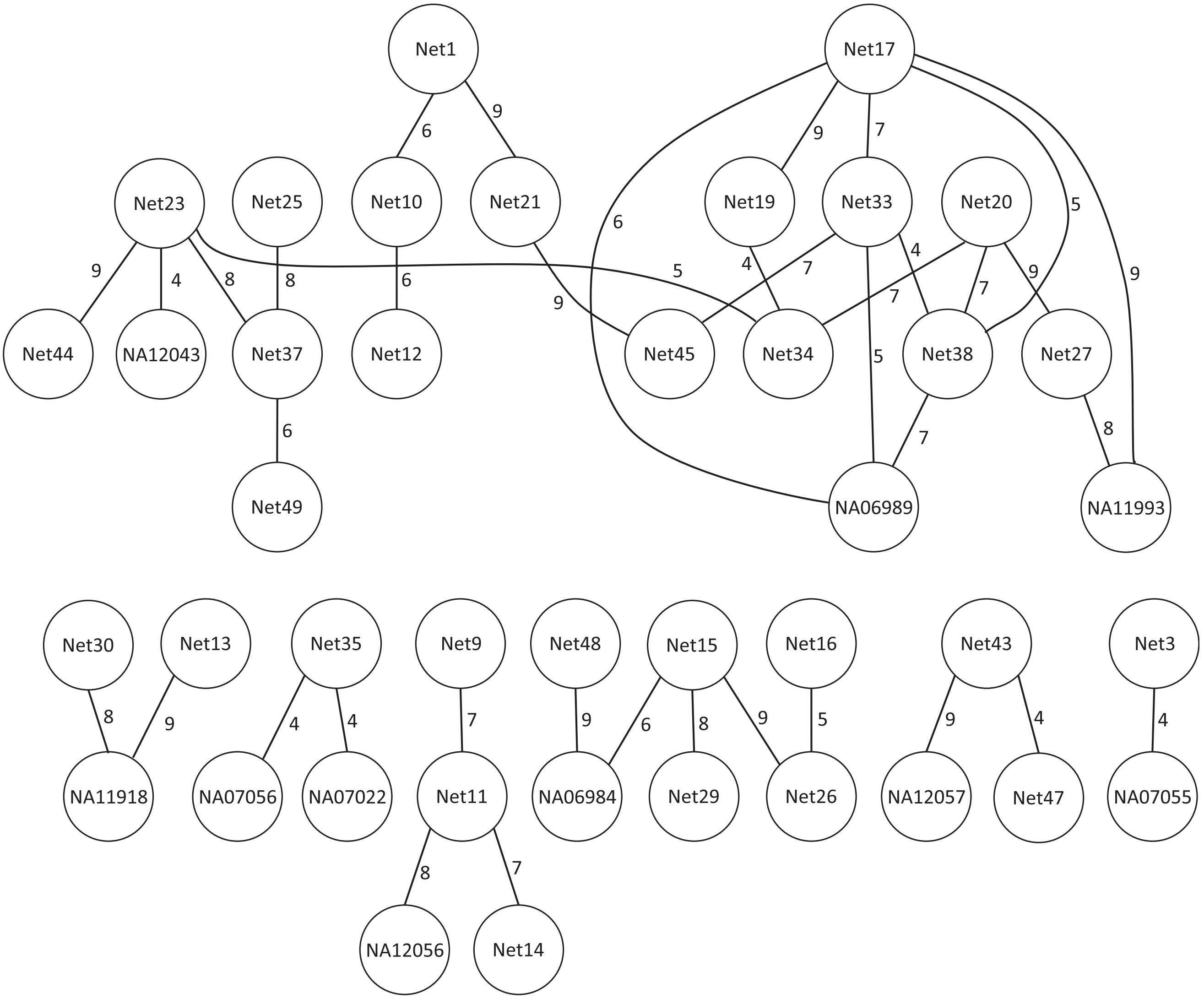

**Figure 5. A Graph of PADRE-Estimated Relationships among the CEU Samples with a Bonferroni-Adjusted Threshold of $\alpha = 0.05/9{,}090 = 5.5 \times 10^{-6}$**

Each node corresponds to a PRIMUS reconstructed network number, and an edge between nodes indicates a significant relationship predicted by PADRE using pairwise relationship likelihoods obtained by ERSA. The number next to each edge indicates the degree of relationship connecting a founder in the reconstructed pedigree of each network. This type of network graph is the standard output of PADRE.

two first- or second-degree relatives, we observed a substantial improvement in accuracy with PADRE (Figure 4), whereas ERSA's accuracy rate was unchanged. PADRE correctly predicted 39% (95% confidence interval: 38% to 40%) of the 10th-degree relationships within one degree of relatedness when the individuals had two first- or second-degree relatives in the pedigree, in comparison to 23% (95% confidence interval: 22% to 24%) for ERSA alone. In addition, PADRE was able to detect 9% of the 11th-degree relationships, whereas ERSA did not detect any. The relationship prediction accuracy in this dataset increased as the number of first- and second-degree relatives in the pedigree increased, broadly matching the improvement we observed in our simulations (Figure 4). Effects of Bonferroni correction on relationship estimation accuracy in these data are shown in Figure S4.

We previously reconstructed 51 separate pedigrees within the HapMap3 CEU dataset.[20] These pedigrees contain between two and six individuals. PADRE identified relationships between 40 pairs of pedigrees consisting of 594 pairs of individuals via previously unknown fourth- through ninth-degree relationships (Figure 5). Figure 6 illustrates one example in which PADRE predicts relationships connecting founders from four previously described CEU pedigrees.

We have demonstrated through simulated and actual data that PADRE can leverage pedigree reconstruction results from PRIMUS and distant pairwise relationship predictions from ERSA to improve both the sensitivity and accuracy of distant relationship estimation. The power to detect relationships more distant than ninth-degree relatives was dependent on the number of generations in the

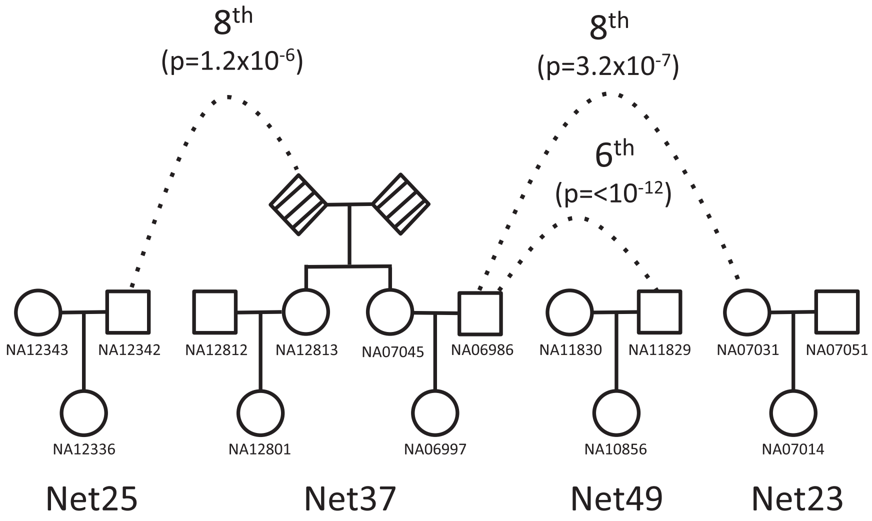

**Figure 6. An Example of Four Distantly Related HapMap3 CEU Pedigrees with Relationships Predicted by PADRE**
Although the trios and the full-sibling relationship between NA12813 and NA07045 have been previously reported, PADRE is able to identify statistically significant relationships connecting these distantly related pedigrees. The related pairs of founders are marked with the dotted lines, and the degree of relationship is labeled next to the line.

pedigrees with genotype data. For instance, PADRE detected up to 13th-degree relationships in the simulated pedigrees with three generations of genotype data and the founders of the pedigrees (A2 and B2, Figure 1). As the depth of the pedigrees connected by PADRE increases, so will the distance of relationships that PADRE will be able to predict. Relationship estimation accuracy in PADRE improved as the number of genotyped individuals within each pedigree increased (Figures 2 and 4) and was most accurate when all individuals within a pedigree were genotyped (Figure 3).

We note that PADRE assumes absence of consanguinity and thus does not look for distant relationships within the reconstructed pedigree structures identified by PRIMUS. However, these types of relationships can be detected in other ways, for example, by using ISCA[25] and ERSA (v.2) to evaluate regions of the genome that are shared IBD on both chromosomes (IBD2) between founders within a pedigree. Although PADRE can connect a single pedigree, and even a single founder, to multiple other pedigrees, the algorithm is currently limited to establishing a maximum of one distant relationship between founders of any given pair of pedigrees. Allowing for multiple relationships between founders of a pair of pedigrees will require modeling of independently inherited shared segments to prevent confounding and is a direction of work for future releases of PADRE.

## Discussion

PADRE has several important and immediate applications in human genetic analysis, especially in large case-control studies. PADRE can detect cryptic fourth- through 13th-degree relationships, even in small datasets, as shown in our analysis of the CEU data (Figures 5 and 6). By identifying and appropriately modeling these relationships, studies can avoid findings biased by relatedness[26] and in some cases might be able to leverage familial relationships to improve power.[27] This is particularly important for detecting relatively high-penetrance risk alleles segregating in distantly related pedigrees.

Existing prediction algorithms for detecting distant pairwise relationships use the number and size of shared IBD segments between two individuals to estimate their degree of relatedness.[16,28] However, as the degree of relatedness increases, the number of shared segments drops to zero. Most 11th-degree human relatives share no segments of their autosomal DNA IBD;[16] therefore, their degree of relatedness cannot be estimated by existing pairwise comparison programs. In some scenarios, PADRE can leverage reconstructed pedigrees to identify genealogical relationships between individuals who are genetically unrelated, i.e., share no portion of their genome IBD through their most recent common ancestors.

PADRE runtime increases combinatorially depending on the number of family networks, the number of possible pedigrees within each family network, the number of founders in each of the pedigrees, and the number of non-missing individuals in each pedigree structure in the PRIMUS results. These numbers are difficult to predict prior to running PRIMUS and depend heavily on how densely the pedigrees have been sampled (Figure S5). For some datasets, it will be necessary to use a closer relatedness cutoff for the PRIMUS reconstruction in order to limit the number of possible pedigrees generated. This adjustment will in turn improve the runtime of PADRE. We have employed this technique with the European ancestry pedigrees due to the sparse sampling of individuals. Table S1 and accompanying text provides additional information on runtimes and computational limitations of PADRE.

There has been a resurgence of interest in large and deeply genotyped pedigrees in the search for genetic heritability of complex disease traits. Pedigrees have become especially relevant in the detection of rare variant effects on diseases because pedigrees are well-suited for the study of rare variation.[9] Under the hypothesis that multiple rare and common variants contribute to complex disease, projects such as the Alzheimer's Disease Sequencing Project, the San Antonio Mexican American Family Studies, and the Jackson Heart Study have all undertaken deep whole-genome sequencing of members of clinically ascertained pedigrees. Projects such as these could particularly benefit directly from verification and detection of distant relatedness in PADRE.

PADRE leverages genome-wide average IBD sharing, as well as the size and distribution of shared IBD segments, to achieve a substantial improvement in accuracy over existing methods. PADRE has immediate relevance to a host of applications within genetics, allowing investigators to more accurately estimate cryptic relatedness, verify very distant relationships, and maximize power in analytic design. PADRE is freely available for academic use (see Web Resources).

## Data Access

Access to PADRE input data for the extended pedigrees has been made publicly available. The ERSA-derived shared segments (as described in Huff et al. 2011[16]) as well as the PRIMUS-derived pedigree likelihoods for the extended pedigree samples are available on the PADRE website (see Web Resources).

## Supplemental Data

Supplemental Data include five figures and one table and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.05.020.

## Acknowledgments

We thank Lauren Petty for her helpful comments on the manuscript. J.S. was supported by the National Science Foundation Graduate Research Fellowship under grant DGE-0718124. D.A.N. was supported by the University of Washington Center for Mendelian Genomics (UW-CMG), funded by the National Human Genome Research Institute and the National Heart, Lung and Blood Institute, grant U54HG006493. C.D.H. was supported by R01 GM104390.

Received: January 21, 2016
Accepted: May 16, 2016
Published: June 30, 2016

## Web Resources

GERMLINE, http://www.cs.columbia.edu/~gusev/germline/phasing_pipeline.tar.gz
HapMap, http://hapmap.ncbi.nlm.nih.gov
PADRE, http://www.hufflab.org/software/padre

## References

1. Bellis, M.A., Hughes, K., Hughes, S., and Ashton, J.R. (2005). Measuring paternal discrepancy and its public health consequences. J. Epidemiol. Community Health *59*, 749–754.
2. Kerr, S.M., Campbell, A., Murphy, L., Hayward, C., Jackson, C., Wain, L.V., Tobin, M.D., Dominiczak, A., Morris, A., Smith, B.H., and Porteous, D.J. (2013). Pedigree and genotyping quality analyses of over 10,000 DNA samples from the Generation Scotland: Scottish Family Health Study. BMC Med. Genet. *14*, 38.
3. Wolf, M., Musch, J., Enczmann, J., and Fischer, J. (2012). Estimating the prevalence of nonpaternity in Germany. Hum. Nat. *23*, 208–217.
4. Boehnke, M., and Cox, N.J. (1997). Accurate inference of relationships in sib-pair linkage studies. Am. J. Hum. Genet. *61*, 423–429.
5. Staples, J., Qiao, D., Cho, M.H., Silverman, E.K., Nickerson, D.A., and Below, J.E.; University of Washington Center for Mendelian Genomics (2014). PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am. J. Hum. Genet. *95*, 553–564.
6. Palamara, P.F., Francioli, L.C., Wilton, P.R., Genovese, G., Gusev, A., Finucane, H.K., Sankararaman, S., Sunyaev, S.R., de Bakker, P.I., Wakeley, J., et al.; Genome of the Netherlands Consortium (2015). Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. Am. J. Hum. Genet. *97*, 775–789.
7. Saad, M., and Wijsman, E.M. (2014). Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees. Genet. Epidemiol. *38*, 579–590.
8. Saad, M., and Wijsman, E.M. (2014). Power of family-based association designs to detect rare variants in large pedigrees using imputed genotypes. Genet. Epidemiol. *38*, 1–9.
9. Wijsman, E.M. (2012). The role of large pedigrees in an era of high-throughput sequencing. Hum. Genet. *131*, 1555–1563.
10. Browning, S.R., and Browning, B.L. (2012). Identity by descent between distant relatives: detection and applications. Annu. Rev. Genet. *46*, 617–633.
11. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.
12. Browning, S.R., and Thompson, E.A. (2012). Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics *190*, 1521–1531.
13. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. *10*, e1004234.
14. Alvarez-Cubero, M.J., Saiz, M., Martinez-Gonzalez, L.J., Alvarez, J.C., Eisenberg, A.J., Budowle, B., and Lorente, J.A. (2012). Genetic identification of missing persons: DNA analysis of human remains and compromised samples. Pathobiology *79*, 228–238.
15. Lin, T.H., Myers, E.W., and Xing, E.P. (2006). Interpreting anonymous DNA samples from mass disasters–probabilistic forensic inference using genetic markers. Bioinformatics *22*, e298–e306.
16. Huff, C.D., Witherspoon, D.J., Simonson, T.S., Xing, J., Watkins, W.S., Zhang, Y., Tuohy, T.M., Neklason, D.W., Burt, R.W., Guthery, S.L., et al. (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res. *21*, 768–774.
17. Li, H., Glusman, G., Hu, H., Shankaracharya, Caballero, J., Hubley, R., Witherspoon, D., Guthery, S.L., Mauldin, D.E., Jorde, L.B., et al. (2014). Relationship estimation from whole-genome sequence data. PLoS Genet. *10*, e1004144.
18. Ng, C.T., and Joe, H. (2014). Model comparison with composite likelihood information criteria. Bernoulli *20*, 1738–1764.

19. Morrison, J. (2013). Characterization and correction of error in genome-wide IBD estimation for samples with population structure. Genet. Epidemiol. *37*, 635–641.
20. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
21. Staples, J., Nickerson, D.A., and Below, J.E. (2013). Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. Genet. Epidemiol. *37*, 136–141.
22. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
23. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326.
24. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A.; The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.
25. Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science *328*, 636–639.
26. Voight, B.F., and Pritchard, J.K. (2005). Confounding from cryptic relatedness in case-control association studies. PLoS Genet. *1*, e32.
27. Hu, H., Roach, J.C., Coon, H., Guthery, S.L., Voelkerding, K.V., Margraf, R.L., Durtschi, J.D., Tavtigian, S.V., Shankaracharya, Wu, W., et al. (2014). A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat. Biotechnol. *32*, 663–669.
28. Browning, B.L., and Browning, S.R. (2011). A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. *88*, 173–182.