WISDOM PANEL™

BCSYS: An Accurate and Scalable Local Ancestry Classifier

Daniel Garrigan, Jason Huff, and Rebecca Chodroff Foran

Abstract

Wisdom Health has updated our core ancestry algorithm with the goal of improving breed classification accuracy. The new algorithm, called BCSYS, is more computationally efficient than the legacy Wisdom Panel algorithm. This efficiency enables us to use a larger breed DNA reference panel. Large reference panels, in turn, allow for more breeds to be called and for increased accuracy, due to the inclusion of more reference samples per breed. Furthermore, the BCSYS algorithm was specifically tuned to improve accuracy for mixed breed samples. Finally, unlike our legacy algorithm, BCSYS is a *local ancestry classifier*, which means that in addition to calling the total proportion of breeds throughout an animal's genome, it also assigns ancestry labels to very specific small segments of chromosomes. One new feature is that we now use the local ancestry results to train a machine learning model that predicts the purebred status of an animal. However, the local ancestry classifier will also drive future product development, detailing how an animal's physical traits relate to their individual ancestry. In this document, we present the technical specifications of the BCSYS algorithm and compare its performance against that of industry-leading local ancestry classification algorithms. We find that BCSYS has both significantly higher accuracy for highly mixed samples and is orders of magnitude more computationally efficient than existing algorithms.

Introduction

Understanding how populations are structured through time and space remains an ongoing and important aspect of population genetics research. As population genomic data sets continue to accumulate, the problem of assigning new sequences to predefined population groups poses significant computational challenges. Initial efforts to assign population labels to a query sequence focused primarily on the total evidence from across the genome, particularly using single nucleotide polymorphism (SNP) data (Alexander *et al*. 2009; Falush *et al*. 2003; Patterson *et al*. 2006; Pritchard *et al*. 2000). These methods treat diploid SNP genotypes as independent realizations of a population structure model, thereby ignoring information from patterns of linkage disequilibrium and haplotype structure. For admixed query samples, the result is an estimate of *global ancestry*, or the proportions of the query genome assigned to different predefined reference populations.

As the methods for computational genotype phasing have improved, maternal and paternal haplotypes can now be reliably recovered from dense SNP data using only computational methods (Browning and Browning 2007; Loh *et al*. 2016). Subsequently, improvements in haplotype phasing has driven the development of methods for assigning *local ancestry*, which is the process of labelling specific segments of phased chromosomes to reference populations. Development of these new methods is motivated by the idea that local ancestry can more accurately control for the confounding effects of population structure in genome-wide association studies of admixed populations (Martin *et al*. 2017; Thornton and Bermejo 2014).

Several methods and software packages are currently available for performing local ancestry assignment: LAMP-LD (Baran *et al*. 2012), Globetrotter (Hellenthal *et al*. 2014), RFMix (Maples *et al*. 2013), MULTIMIX (Churchhouse and Marchini 2013), HapMix (Price *et al*. 2009), and MOSAIC (Salter-Townsend and Myers 2019). However, many of the above methods are computationally expensive and do not readily scale to handle large reference panels, nor do they efficiently process large numbers of query sequences. Here we introduce a novel local ancestry classifier that easily accommodates large reference panels, is computationally efficient, and produces results that are more accurate than the most widely used existing method. Finally, we are motivated to produce a local ancestry classifier that performs well with highly admixed genomes.

Methods

The Breed Classifier System (BCSYS) method assumes the existence of a curated reference panel comprising some number of haplotype sequences, each of which is labelled according to membership in some population group. The goal of the method is to classify an arbitrary query haplotype to one of the reference panel populations. The BCSYS workflow begins by using Beagle 5.1 (Browning *et al*. 2018) to phase both query and reference genotypes into maternal and paternal chromosomes. The phased data are then partitioned into non-overlapping 5 cM windows. The population assignment of each window is achieved by recovering all pairwise *set-maximal* matches (see definition in next subsection) between query and reference haplotypes using a positional Burrows-Wheeler transform algorithm (PBWT; Durbin 2014). The density of set-maximal matches between a given query and all reference haplotypes is calculated and the reference population with the highest relative density is selected as the “raw” assignment. Then a hidden Markov model (HMM) is run on the raw calls over windows grouped by chromosome to “smooth” the local ancestry assignments. Finally, the global ancestry proportions are aggregated from the local assignments and used in a global ancestry classifier to produce a population assignment for the entire diploid genome.

Recovering Short Matching DNA Segments

The set of algorithms inherent in the PBWT can efficiently recover matches between pairs of haplotype sequences in a collection (Durbin 2014). Several PBWT-based algorithms iterate through a collection of haplotype sequences and recover *set-maximal* matches, which are defined as the set of other sequences which show locally maximal, unbroken matches to the current sequence. In our case, the collection of sequences contains both query and reference haplotypes. The BCSYS method records all such matches to query sequences. Each set-maximal match is labeled by the reference population label of the matching haplotype (see **Figure 1**). To exclude small haplotype segments with elevated homozygosity across populations (and therefore are unlikely to arise due to recent common ancestry), only set-maximal matches longer than 0.5 centimorgans (cM) are considered in the analysis.

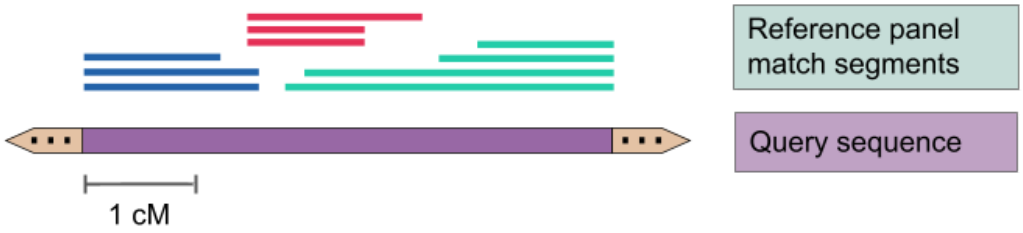

Figure 1. Illustration of marginal match length concept. The query sequence is depicted in purple at the bottom. Matches to reference panel sequences are shown above. Each match is colored according to its corresponding reference population label. The marginal match length sum per reference population is considered proportional to the likelihood of the query sequence originating from that reference population.

After running the PBWT-based algorithm described above, all set-maximal matches can be recovered between the query haplotype and the reference panel haplotypes and the marginal sum of match lengths from a particular population is considered proportional to the likelihood that the query sequence is sampled from that reference population. The final local assignment represents the reference population with the longest marginal match length. Marginal likelihoods can be computed as either the total length of reference panel matches or, can be corrected for reference panel sample size by taking the marginal average. The choice of correcting for reference panel sample size has the practical effect of upweighting small panels, while uncorrected likelihoods will favor large reference panel groups. We choose to use uncorrected likelihoods because we are motivated to develop a classifier that works well with admixed samples, for which admixture proportions are more likely to favor common reference populations that are well-represented in the panel.

In contrast to other local ancestry algorithms, the present method is a simple moment-based estimator, which minimizes reliance on complex underlying population genetics models that are often needed when a Dirichlet distribution is used as a conjugate prior for the categorical distribution (*e.g.*, Pritchard *et al*. 2000). Bayesian inference under a Dirichlet prior necessitates assumptions inherent in the simulation of highly stochastic population structure models with uncharacterized parameters, at the expense of scalability and increased computation time, often with an unknown improvement in accuracy. Rather than leveraging traditional simulation-based

Bayesian inference, we instead focus on improving assignment accuracy through application of machine learning models trained on reference panel samples.

Our method of using marginal reference population match lengths also lends increased robustness to haplotype phasing errors present in the reference panel haplotypes. The rationale for this assertion is that long matches broken by phase switches will still be recovered as separate matches by the algorithm and contribute equally to the marginal sum of match lengths. The only scenario for which this is not the case is when a phase switch breaks a long match and one (or both) of the resulting match fragments are too short (*i.e.*, < 0.5 cM) to be recorded by the method. In the worst case scenario, the estimate of the marginal population match length would be reduced by a maximum of 1 cM. One approach for addressing this case is to reduce the match length threshold.

Smoothing of Raw Local Ancestry

Hidden Markov models (HMM) are widely used in population genomics because they model the linear nature of features along chromosomes (*e.g.*, Li and Stephens 2003). In the present case, the ordered sequence of local ancestry labels is treated as the observed sequence in a HMM. In this framework, each reference population is considered a latent variable, or a "hidden state" of the query haplotype. The objective of employing the HMM in this manner is to eliminate spurious transitions between local ancestry assignments and to correct for common incorrect assignments. Again, because we are motivated to develop a classifier that performs well for admixed samples, we favor HMM parameters that facilitate mixing of the chain, such as adding pseudocounts to transition probabilities to ensure no probability is zero. Finally, the HMM is trained on the reference panel, for which the local ancestry assignments are assumed to be a source of truth.

The HMM emission probabilities are estimated by a leave-one-out procedure applied to all reference panel haplotypes. Each of the reference haplotypes is, in turn, used as a query sequence and assigned reference population labels from the estimated parameters of the categorical distribution. These estimates are aggregated over all query haplotype runs into a matrix binned by the "true" population label of the haplotype. The elements of the resulting population confusion matrix are used as the HMM emission probabilities. The transition matrix is also learned from the estimated sequence of population labels in the reference panel haplotypes. Finally, the vector of probabilities of starting in a given hidden state is estimated

from the global ancestry estimates resulting from the PBWT-based calls. A separate HMM is run for each chromosome using the backward-forward algorithm (Baum and Eagon 1967), and the most likely pathway through the hidden states is decoded using the Viterbi algorithm (Viterbi 1967).

Higher Order Global Ancestry Classification

The local ancestry assignments from the Viterbi path are aggregated over both maternal and paternal chromosome sets and used to calculate the global ancestry proportions for a given diploid sample. The global ancestry proportions are used as features to predict the single most likely population label for the entire diploid sample using a random forest of decision trees classifier (Breiman 1998), as implemented in the scikit-learn software package (https://scikit-learn.org). Prediction probabilities across populations (confidence scores) are recalibrated by logistic regression using the method of Platt (1999). The classifier can be trained on either the same reference panel as is used for the BCSYS algorithm or a different reference panel dedicated to global ancestry. In cases where a population of interest is fundamentally admixed from other populations, the global ancestry classifier training reference panel may include additional labels from such admixed populations. For populations that are extremely diverse or do not conform neatly to populations, such as street dogs and wild canids, the individuals can be omitted from the reference panel entirely.

Benchmarking the BCSYS Classifier

The data of Shannon *et al*. (2015) are used for benchmarking BCSYS with the widely used RFMix local ancestry classifier. The input data comprise genotypes from a semi-custom Illumina array with 173,662 SNPs from the CanineHD array (Vaysse *et al*. 2011), from which 84,414 autosomal variants are selected for analysis. Genomic coordinates map to release 3.1 of the canine reference genome (Lindblad-Toh *et al*. 2006). Diploid genotypes are cohort-phased into parental chromosomes using the Beagle 5.1 software (Browning *et al*. 2018).

The raw genotype data consist of 5,406 diploid samples, representing 234 dog breed groups. The data are filtered to exclude breed groups represented by fewer than 10 samples. Filtering retains 4,368 samples from 87 breed groups. From these 4,368 samples, 200 are randomly selected to be in the test set and the remaining 4,168 samples are used as the breed group reference panel. Only 53 breeds are represented in the test set. From this single-origin test set,

200 synthetic admixed individuals are generated, each from a randomly chosen (without replacement) eight individuals serving as the great grandparents of the synthetic sample. Recombination intervals are determined according to the reference recombination map.

Version 2 of the RFMix software (https://github.com/slowkoni/rfmix) is run on whole-chromosome input VCF files using 24 CPU threads and node size of 5. For the single-origin test set, all 5 cM intervals in the genome are assumed to have the same truth label. Error can then easily be measured as what proportion of the estimated assignments correctly assign the truth label. However, error measurement is more complicated for the synthetic-admixed data set because BCSYS and RFMix choose genomic windows in different ways. It is most straightforward to measure overall performance based on the resulting global ancestry proportions. The true global ancestry proportions are compiled from the generation of the synthetic-admixed individuals, resulting in a vector of real numbers that sum to unity. Two similarly structured vectors with estimated global ancestry proportions are obtained from the output of both BCSYS and RFMix. The accuracy of the two methods is calculated as the mean squared error (MSE) between the estimated and the true vectors.

Results

Comparison of Accuracy Using BCSYS and RFMix

Both BCSYS and RFMix are run on two query sets of haplotypes against a reference panel of 4138 diploid individuals from 87 breed groups. The first query set comprises 200 single-origin individuals from 53 different breed groups. The second query set comprises 200 simulated admixed individuals. The mean squared error (MSE) of the resulting global ancestry proportions are calculated for all four sets of results. **Table 1** shows the distribution of MSE for each of the runs. For single-origin query samples, both BCSYS and RFMix show similarly high levels of accuracy (**Figure 2**). A paired sample *t*-test indicates no significant difference between MSE for BCSYS compared to RFMix for single-origin samples (t = -1.075; P = 0.283). **Figure 3** provides a magnified view of the natural log of MSE to see the covariation more clearly (note: pseudocounts of one were added to all 200 observations to prevent undefined behavior of the natural log function).

Table 1. Descriptive statistics for mean squared error in local ancestry estimates using BCSYS and RFMix on 200 single-origin and 200 synthetic-admixed samples.

Software	Data set	Mean	StDev	Min	Max
BCSYS	single-origin	0.000113	0.000931	0	0.012872
	synthetic-admixed	0.000009	0.000022	2.435E-09	0.000135
RFMix	single-origin	0.000217	0.001009	0	0.011417
	synthetic-admixed	0.001647	0.000955	0.000223	0.005358

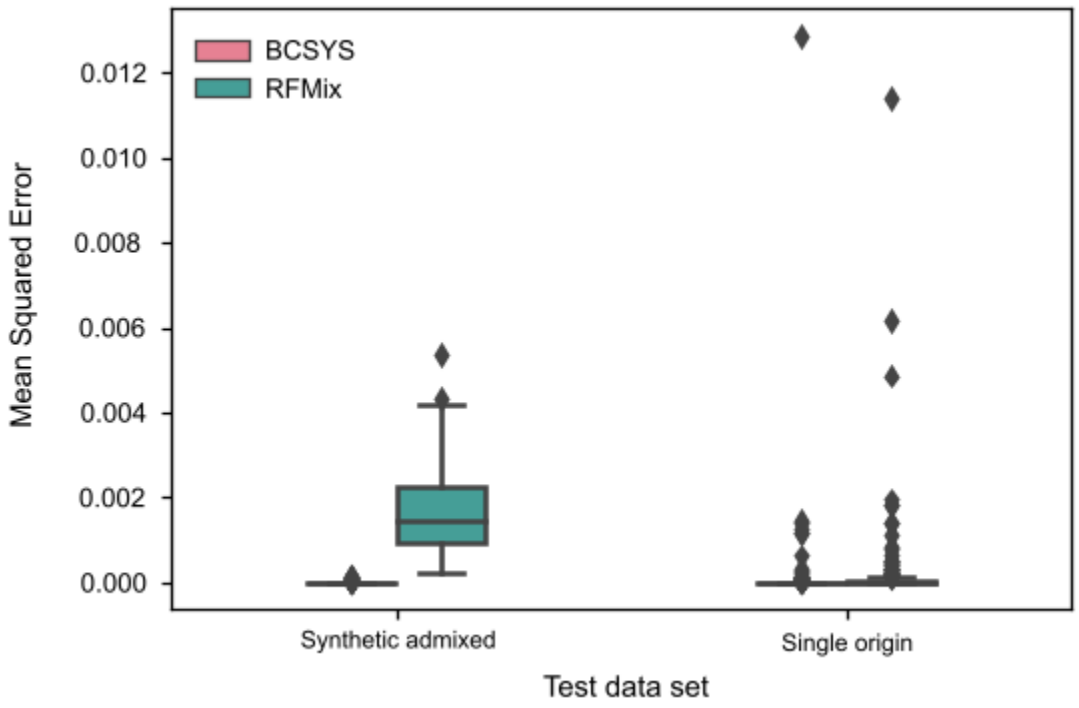

Figure 2. Distributions of mean squared error of local ancestry assignments in two data sets with 200 single-origin samples and 200 synthetic-admixed samples using both BCSYS and RFMix.

In contrast to the analysis of single-origin samples, we do observe a significant difference between BCSYS and RFMix for the synthetic-admixed samples (t = -24.246; P < 0.01). Intriguingly, BCSYS shows lower MSE for synthetic-admixed samples compared to single-origin samples, although this difference is not significant (t = -1.579; P = 0.116). Although the difference in BCSYS results between single-origin and synthetic-admixed samples is not significant, it suggests that the BCSYS HMM step shows a greater propensity for switching states than does RFMix. Conversely, the single-origin and synthetic-admixed sample MSE values are significantly different for the RFMix results (t = 14.127; P < 0.01).

Comparison of Computational Efficiency

We observe dramatic differences in the compute resources utilized by BCSYS and RFMix. To generate the results reported here, BCSYS requires a maximum of 2 Gb of RAM and the entire workflow takes an average of 6 minutes to complete for all chromosomes. However, RFMix requires a maximum of 60 Gb of RAM and takes an average of 3 hours to complete a single chromosome data set. To run both workflows in a commodity cloud environment (for example, Amazon Web Services), RFMix would require a r5a.4xlarge instance type currently priced at $0.904 per hour and an average runtime of 3 hours to run a single chromosome data set for 200 samples. These requirements translate to a cost of $0.515 per sample. The requirements for BCSYS are a m5.4xlarge instance type currently priced at $0.768 per hour, run for an average of 6 minutes for 200 samples, for all chromosomes means the price would be approximately $0.000384 per sample. Finally, in an analysis not detailed here, we found that runs of RFMix with larger reference panels (>16,000 samples, the approximate size of the proprietary reference panel routinely used by BCSYS) resulted in memory segmentation faults on machines with 144 Gb of RAM (results not shown). This final anecdote illustrates how difficult it can be for many local ancestry classifiers to scale to large numbers of reference samples.

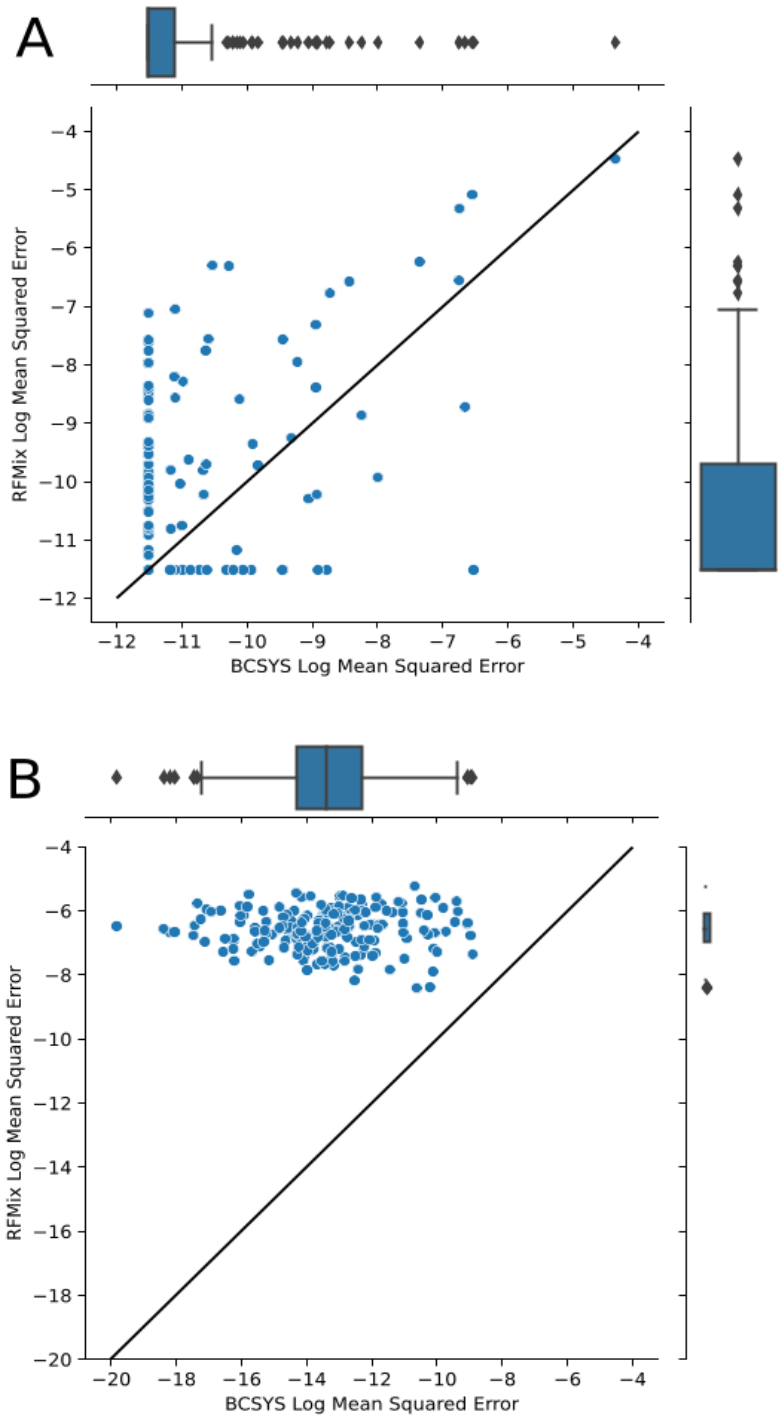

Figure 3. Detailed view of mean squared error comparison between BCSYS and RFMix. Panel (A) shows the covariation of log MSE for single-origin samples, while panel (B) shows covariation of log MSE for synthetic-admixed samples.

Accuracy of Global Ancestry Classifier

The global ancestry classifier takes the output of the local ancestry classifier and uses those labels as features to predict an aggregate label for the entire diploid sample. While this is currently used as a classifier for purebred animals, it has great potential to address the fact that not all breeds are genetically distinguishable. Some dog breeds may be phenotypically distinct, but are, in fact, recently derived from other, older breeds that are genetically distinguishable. A classifier that only accounts for the global ancestry proportions of animals can therefore apply a greater number of labels from a limited set of local ancestry labels. Hence, the global ancestry classifier can have a reference panel that is separate from the reference panel used by the BCSYS local ancestry classifier. However, in the present benchmark experiment, we use the same reference panel, meaning that 87 labels can be predicted from 87 features.

When a leave-one-out cross-validation is run on the BCSYS results described in the previous sections, the global ancestry classifier shows a recall of 0.993. Likewise, when the predictions are grouped by the 87 breed labels, the recall weighted by breed is 0.985. Finally, when the confidence scores for each prediction are used as a determination of a positive or negative prediction result, we can then generate a receiver operating characteristic (ROC) curve (**Figure 4**). The area under the curve (AUC) for the random forest implementation of the global ancestry classifier is 0.988.

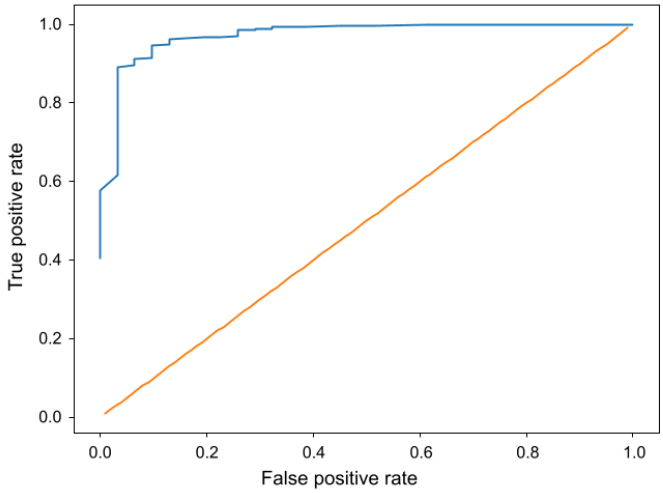

Figure 4. ROC curve for random forest implementation of the global ancestry classifier.

To assess purebred prediction accuracy across the live ancestry classifier, we started with the full reference panel covering 351 breeds and populations. We refined and balanced a subset as a global ancestry reference panel. This included removing some populations that likely have had little or no purebred dog ancestry, including street dogs and wild canids. In two-fold cross-validation using this reference panel, the global ancestry classifier shows a recall of 0.981.

Conclusions

We are motivated to develop a novel local ancestry classification method that 1) is more computationally efficient than existing classifiers, 2) does not compromise on the accuracy of local ancestry assignments, and 3) maintains accuracy even when highly admixed genomes are used as input. The BCSYS method presented here capitalizes on efficiencies inherent in the positional Burrows-Wheeler transform data structure and on sensitivity to genome-wide patterns learned by a hidden Markov model. We benchmark the accuracy of the BCSYS method to that of a widely-used local ancestry method implemented in the RFMix software. When single-origin samples (all segments should correctly be assigned to a single reference population) are used, BCSYS and RFMix do show equivalent levels of accuracy. However, when synthetically generated admixed samples are used as query input, BCSYS demonstrates significantly improved accuracy compared to RFMix.

One important distinction between BCSYS and other classifiers, including RFMix, is that BCSYS is more computationally scalable, in that it uses far less memory and runs much faster. This is important because the computational scalability of BCSYS allows for larger reference panels to be used for local ancestry assignment, including more breed groups and a greater number of individual samples per breed group. Larger numbers of samples per breed group is an important factor for improving the overall accuracy of the classifier going forward.

We also show that a novel global ancestry classifier can be used in conjunction with BCSYS to accurately predict a population label for an entire sample. This is important because a global ancestry classifier can be used in cases where the local ancestry reference panel is unable to distinguish populations based solely on the haplotype sequences within a given window. In addition, some recognized breed groups may not necessarily be genetically distinct, but instead are unique combinations of admixture proportions rather than a unique collection of haplotype sequences. We show that the proposed global ancestry classifier is highly accurate with an

observed recall greater than 98%. Each prediction produced by the global ancestry classifier has an associated confidence score that can be flexibly calibrated using a well-designed training set.

Finally, we note that accurate local ancestry prediction for a cohort enables a myriad of downstream applications, including improved accuracy of genome-wide association studies and the construction of robust machine learning models for the prediction of traits, particularly when used in conjunction with genotypes from loci of large effect. In this way, the next generation of Wisdom Health ancestry prediction will evolve to include discovery of novel genotype-phenotype associations in large cohorts of animals and increasingly accurate prediction of complex traits and health conditions.

Literature Cited

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664.

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., *et al*. (2012). Fast and accurate inference of local ancestry in Latino populations. Bioinformatics 28, 1359–1367.

Baum, L.E., and Eagon, J.A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bull. Amer. Math. Soc. 73, 360–364.

Breiman, L. (1998). Arcing classifier. Ann Stat 26, 801–849.

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.

Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. 103, 338–348.

Churchhouse, C., and Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. Genet. Epidemiol. 37, 1–12.

Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics 30, 1266–1272.

Falush, D., Stephens, M., and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587.

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science 343, 747–751.

Li, N., and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165, 2213–2233.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C., *et al.* (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438, 803–819.

Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., *et al.* (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. 48, 1443–1448.

Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. 93, 278–288.

Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. 100, 635–649.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. 2, e190.

Platt, J.C. (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in large margin classifiers. Cambridge: MIT Press.

Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5, e1000519.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Salter-Townshend, M., and Myers, S. (2019). Fine-scale inference of ancestry segments without prior knowledge of admixing groups. Genetics 212, 869–889.

Shannon, L.M., Boyko, R.H., Castelhano, M., Corey, E., Hayward, J.J., McLean, C., White, M.E., Abi Said, M., Anita, B.A., Bondjengo, N.I., *et al.* (2015). Genetic structure in village dogs reveals a Central Asian domestication origin. Proc Natl Acad Sci USA 112, 13639–13644.

Thornton, T.A., and Bermejo, J.L. (2014). Local and global ancestry inference and applications to genetic association analysis for admixed populations. Genet. Epidemiol. 38 Suppl 1, S5–S12.

Vaysse, A., Ratnakumar, A., Derrien, T., Axelsson, E., Rosengren Pielberg, G., Sigurdsson, S., Fall, T., Seppälä, E.H., Hansen, M.S.T., Lawley, C.T., *et al.* (2011). Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. PLoS Genet. 7, e1002316.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inform. Theory 13, 260–269.