2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A Hierarchical Learning Approach to Calibrate Allele Frequencies for SNP Based Genotyping of DNA Pools

Andrew D. Hellicar, Daniel Smith, Ashfaqur Rahman, Ulrich Engelke
Computational Informatics, CSIRO
Hobart, TAS, 7000, Australia

John Henshall
Division Animal, Food and Health Sciences, CSIRO
Armidale, NSW 2350, Australia

Abstract— **The combination of low density SNP arrays and DNA pooling is a fast and cost effective approach to genotyping that opens up basic genomics to a range of new applications and studies. However we have identified significant limitations in the existing approach to calculating allele frequencies with DNA pooling. These limitations include a reduced ability to deal with SNP to SNP variation via the standard interpolation method. Our contribution is a new hierarchical learning framework which resolves these drawbacks. The framework involves a hierarchy of two greedily trained layers of learners. The first layer learns the bias of each SNP then applies a calibration to reduce SNP bias by mapping into a common coordinate system across all SNPs. The second layer learns an allele frequency function exploiting the global SNP data. A range of algorithms have been applied including linear regression, neural network and support vector regression. The framework has been tested on pooled samples of Black Tiger prawns that have been genotyped with low density Sequenom iPLEX panels. Analysis of pooled samples and the corresponding individually genotyped SNP samples indicate the pooling approach introduces an allele frequency RMS error of 0.12. The existing calibration approach corrects ~14% of the error. Our hierarchical approach is 4.5 times as effective by correcting for ~64% of the introduced error. This is a significant reduction and has the potential to enable genetic studies previously not possible due to allele frequency error. Although testing so far is limited to low density SNP arrays the approach was developed to generalize to other SNP genotyping technologies.**

Keywords—Machine learning, DNA

I. INTRODUCTION

Singular Nucleotide Polymorphisms (SNPs) based genotyping is a fast and cost effective approach to identify functionally important polymorphisms of a species [1]. Such gene markers can be linked with disease, complex traits or be used to provide family information. Multiplex microarray systems have been developed for SNP genotyping to provide sufficiently dense coverage for genome wide association studies [2, 3]. For instance, Affymetrix array technology [2] can interrogate 906,600 SNPs, whilst Illumina [3] released the Human Omni5 Beadarray that can genotype 4.3 million SNPs. The costs associated with developing such high density technologies are still prohibitive for genotyping many species. In particular, species where research has not been conducted to identify polymorphic markers that provides coverage of the genome or genes of interest. Even for species where SNP microarrays have been developed, SNP based association studies still require a large number of samples to be genotyped. This is an expensive exercise given each microarray can only be used once. DNA pooling is an attempt to address this issue by combining multiple DNA samples prior to genotyping. Each pool is genotyped as a single sample; greatly reducing the number of microarrays, and hence, cost and time required to undertake a study [4-5].

For the typical case of genotyping individuals, SNP alleles are arbitrarily labeled as A or B and the SNP genotype is one of AA, AB, BB, due to the presence of two copies of the DNA. The raw output of a microarray is therefore quantized into one of three possible values. This quantization means the value can be retrieved despite the presences of genotyping noise. In contrast to individual SNPs, DNA pools require the raw array output to be used to directly compute a quantitative genotype of each SNP [4]. This is known as the allele frequency. Pooled samples are subject to greater genotyping inaccuracies than individual samples [4-6]. This is a consequence of the continuity of pooled allele frequency estimates, which are more susceptible to genotyping noise than the discrete alleles of bi-allelic SNP data.

Low density SNP array technologies have been utilized to genotype species where genomic research is limited and where it is economically infeasible to invest in expensive, higher density microarrays. The Sequenom MassARRAY iPLEX platform [7] is one such low cost technology that genotypes between tens to a few thousand SNPs. By combining low density SNP technology with DNA pooling, costs can be greatly reduced, opening up genomics to a range of low cost studies and applications. Aquaculture is one such application, and this paper forms part of an evaluation of pooled based genotyping of Black Tiger prawns (Penaeus monodon) for a selective breeding program. The study genotyped 22 pooled samples (each comprised of between 18 and 23 prawns) using a Sequenom iPLEX panel of 63 SNPs. The pooled assays are being considered to construct the pedigree of individual farmed prawns, which are not sufficiently valued to employ high density, individual genotyping.

There have been few feasibility studies using low density and low cost arrays with pooled samples. Pooling studies have generally focused upon higher density SNP based microarrays that provide genome wide coverage [8-11]. In this paper, we investigate the accuracy of using low density SNP arrays to

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

genotype pooled samples. To examine the effect of pooling upon these lower cost technologies, pooled allele frequencies are compared to "ground truth" allele frequencies computed from genotypes of the individuals belonging to the same pool.

The contribution of this paper is an allele frequency estimation method based upon supervised machine learning which we propose to correct for errors associated with pooled allele frequencies. The estimation method involves two stages; the first calibrates each SNP by correcting for the bias exhibited in the SNPs raw outputs. These biases exist for Sequenom [12] and Illumina systems and are caused by combinations of differential amplification and hybridization [4]. Bias creates errors in the pooled allele frequencies when estimating allele frequency directly. To correct for this bias, the pooled results are mapped onto a common domain across all of the SNP. The second stage then involves training a model which estimates allele frequency as a function on this common domain. Both stages used supervised training based on "ground truth" allele frequencies. The machine learning algorithms are implemented with SVM and radial basis neural networks. In addition, we examine the linearity of the error characteristics by comparing the accuracy of linear and non-linear methods.

The paper is structured is follows. Section II outlines previous related work that has considered the correction of allele frequency estimates from DNA pools and discusses how our machine learning based approach differs. In Section III, we investigate the accuracy of the pooled allele frequency estimates with the Sequenom iPLEX platform. This investigation is used to motivate our proposed calibration method for pooled allele estimation that is described in Section IV. In Section V, we present the results of our calibration methodology and draw our conclusions in Section VI.

II. BACKGROUND

Genotyping errors are recognized to have an impact on the conclusions drawn from a study; however, they are too often neglected. In [13] genotyping errors are defined as a discrepancy between the observed and true genotype of an individual. Four classes of errors are identified related to (1) variation in DNA sequence, (2) low quantity and quality of DNA, (3) biochemical artifacts, and (4) human factors. A protocol for estimating error rates within these classes is proposed.

In this work, we are particularly concerned with errors stemming from incorrect allele frequency estimates, and hence misleading biological conclusions, from pooled DNA samples as compared to individual DNA samples. In SNP genotyping systems, such as the ones by Sequenom and Illumina, these errors are a result of biochemical reactions during the genotyping process and hence fall mainly into category (3) of the above classification (not withstanding any other potential error sources). Whereas Illumina systems typically utilize a process based on differential hybridization and fluorescent detection, the Sequenom iPLEX platform is based on single nucleotide primer extension and mass spectrometry [14]. These errors may have a minor impact on genotype calling for individuals, they have a major impact on the estimation of allele frequencies in genotyping based on DNA pooled data though. Identifying and correcting for the bias and errors is therefore instrumental for the success of the estimation of allele frequencies in pooled DNA experiments.

Several research contributions aimed at solving this problem exist. In [15] the degree of bias is quantified using the coefficient of preferential amplification/hybridization (CPA), which is defined as the ratio of average peak intensities between two alleles. It was found that lognormal distributions adequately model bias introduced through preferential hybridization, resulting in reduced error of allele frequency estimation for the human genome. The authors in [16] propose a SNP genotyping method based on a general linear model that accounts for the nested structure of the data. The proposed method does not require the CPA to be known and hence, avoids the need for individual SNP genotyping to determine allelic ratio of hybridization, therefore scaling up to arrays with many thousands of SNPs. Finally, in [17] piecewise linear interpolation of pooled alleles was used to correct for the bias in pooled DNA data of the human genome.

Unlike any of the previous methods, we use a hierarchical machine learning approach to achieve both individual SNP calibration and bias corrected allele frequency estimation on DNA pooled data. We show that linear piecewise interpolation is not always optimal and that for some SNPs Lagrange or Hermite interpolation result in improved performance. To the best of our knowledge, this is the first line of research that utilizes these techniques to successfully correct for bias in DNA pooled data. Given the application to low-cost arrays, this framework is particularly beneficial in the agriculture and aquaculture domain and is in the following demonstrated on pooled Black Tiger prawns (Penaeus monodon) data.

III. ACCURACY OF ALLELE FREQUENCY ESTIMATES

We investigate the accuracy of the pooled allele frequency estimates with the Sequenom iPLEX platform.

The existing approach for calculating allele frequencies follows the process shown on left of Fig. 1. The following paragraphs describe each of the steps in the process.

A. Data sets

We have four data sets including three data sets from the Sequenom iPLEX platform. Two data sets (iE1, iE2 Fig. 1) contain SNP results for individuals, and a third (pE Fig. 1) contains SNP results for pools of individuals. A final pool membership database (pM) identifies which individuals were in each of the pooled samples.

Additional sub-sets of the available data were generated for analysis. 48 individuals had duplicate samples in iE2 allowing measurement repeatability to be assessed. 78 individuals in iE2 were also sequenced in iE1. Of the 63 SNPs called by the iPLEX platform 2 SNPs consistently failed and were initially

This work was supported in part by a grant from Tasmanian Government which is administered by the Tasmanian Department of Economic Development, Tourism and the Arts and in part by the CSIRO Food Futures Flagship.

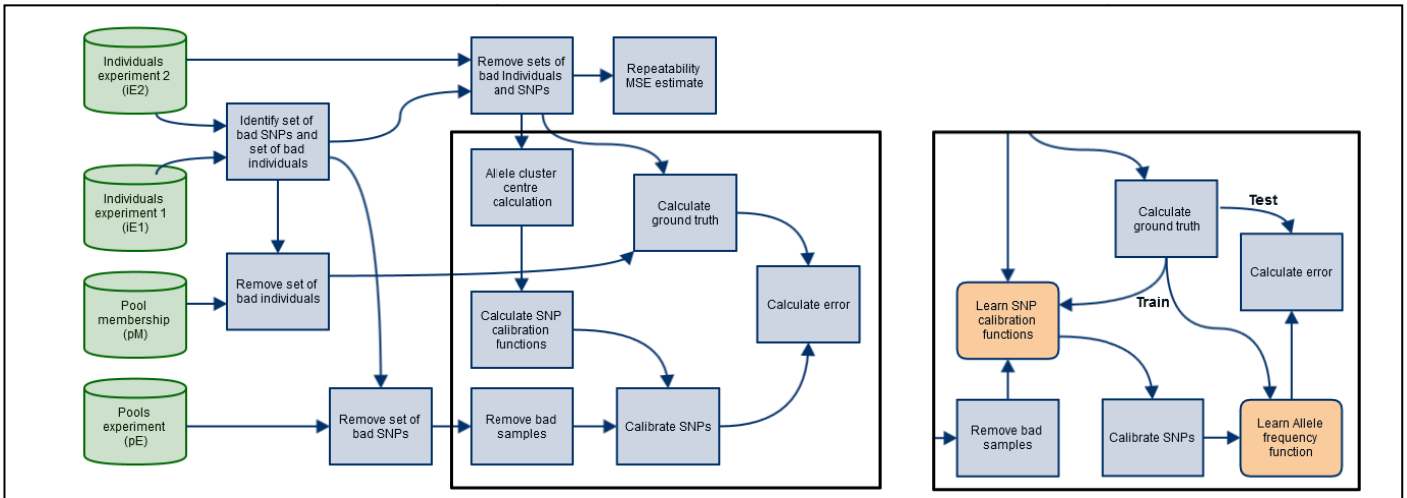

Fig. 1. Flowchart of existing approach. Proposed new approach, box in new approach replaces box in existing approach.

removed. A summary of the experimental result data sets is given in Table I.

The iPLEX platform generates (x, y) pairs where x is the platform response to the A allele and y the response to B allele. Based on the (x, y) values the platform calls the 61 SNP as being AA, AB or BB. iE2 and pE contain (x, y) values and allele calls, whereas iE1 only contained calls; however, the accuracy of experiment iE1 was more rigorous and this enabled assessment of the calling accuracy of the SNPs in iE2.

B. *Cleaning of data*

Initial cleaning of data involves culling a subset of SNPs from the data and a subset of individuals from the data.

1) *Bad SNP subset identification*

We use the iE1 measurements to determine iE2 SNP quality by comparing the SNP calls on the shared set of 78 individuals. Results are shown in Table II. SNPs were identified as bad if they failed to demonstrate 45 identical calls between the experiments. 45 calls was selected as a threshold as the number of errors rapidly increased below this threshold.

TABLE I. SUMMARY OF AVAILABLE EXPERIMENTAL DATA

Data set	Raw	Cleaned	units
iE1	111		individuals
iE2	1041	850	individuals
Intersection (iE1, iE2)	78		individuals
Duplicates (iE2)	48	41	individuals
SNPs	61	48	SNPs
Pooled data (pM)	22	19	pools
Pooled samples (pE)	1342	901	(x,y) pairs

2) *Bad Individual subset identification*

Individuals were placed in the bad subset if they failed to exhibit a minimum of 50 SNPs calls (~0.8 fraction of SNPs called). Figure 2 shows that approximately 20% of the individuals had less than 50 SNPs called. More precisely this corresponded to 191 bad individuals (out of 1041 in total) removed from iE1. 114 of the pool members were bad individuals and were removed from pool membership. The bad SNP subset was removed from both the individual results and pooled results leaving 48 SNPs from the original 61.

C. *Allele Clustering*

The (x, y) pairs from iE2 are plotted for four SNPs in Fig. 3. The AA, AB and BB alleles form three distinct clusters for each SNP, which is expected as in the ideal case x depends on quantity of A allele, and y of B allele; however, the centres and spatial distributions of the clusters vary from SNP to SNP. Because of this variation the standard approach to calling alleles on individual samples uses a clustering algorithm [12]. The AA, AB and BB cluster centres (x_{AAj}, y_{AAj}), (x_{ABj}, y_{ABj}) and (x_{BBj}, y_{BBj}) are then calculated for each SNP using the individual values (iE2) and converted to polar coordinates such that $\emptyset_{AAj}$, $\emptyset_{ABj}$ and $\emptyset_{BBj}$ are the angles of the cluster centres for SNP j.

TABLE II. ERRORS IN SNP CALLS

Error Type	% reads in error[a]
No error	61.4
No read -> read	27.8
Heterozygous/ Homozygous	8.8
Homozygous to other Homozygous	2.0

[a] Calculated by comparing iE2 results with iE1 for 78 shared individuals.

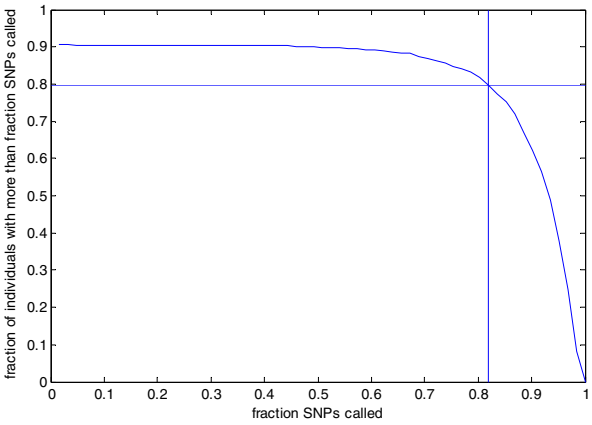

Fig. 2. Fraction of individuals with greater than fraction of SNPs called. ~80% of individuals have more than ~80% SNPs called.

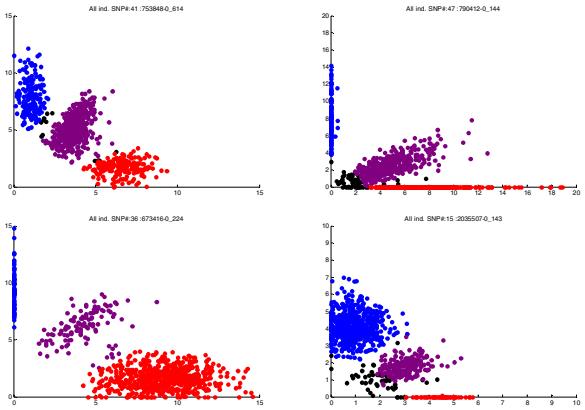

Fig. 3. (x, y) points for four example SNPs highlighting different cluster types.

D. SNP Calibration functions and SNP calibration

For pooled sample k the normalised angle $\hat{\emptyset}_{kj}$ can be used as an estimate of allele frequency of SNP j:

$$\hat{\emptyset}_{kj} = \frac{\tan^{-1}\left(\frac{x_{kj}}{y_{kj}}\right)}{\frac{\pi}{2}}$$

Improved accuracy results by accounting for the cluster centre variation. This is achieved by introducing a SNP calibration function which transforms $\hat{\emptyset}_{kj}$ into a more accurate pooled allele frequency. One common calibration approach involves defining a piece-wise linear calibration function which maps cluster centres AA, AB, BB to allele frequency values 1.0, 0.5, 0.0 and linearly interpolates values between. The calibrated allele frequency $\hat{f}_{kj}$ of pool k for SNP j is:

$$\hat{f}_{kj} = \begin{cases} 0.0, & \hat{\emptyset}_{kj} < \hat{\emptyset}_{\mathrm{AA}j} \\ 0.5\left(\frac{(\hat{\emptyset}_{kj} - \hat{\emptyset}_{\mathrm{AA}j})}{(\hat{\emptyset}_{\mathrm{AB}j} - \hat{\emptyset}_{\mathrm{AA}j})}\right), & \hat{\emptyset}_{\mathrm{AA}j} < \hat{\emptyset}_{kj} < \hat{\emptyset}_{\mathrm{AB}j} \\ 0.5 + 0.5\left(\frac{(p_{kj} - \hat{\emptyset}_{\mathrm{AB}j})}{(\hat{\emptyset}_{\mathrm{BB}j} - \hat{\emptyset}_{\mathrm{AB}j})}\right), & \hat{\emptyset}_{\mathrm{AB}j} < \hat{\emptyset}_{kj} < \hat{\emptyset}_{\mathrm{BB}j} \\ 1.0, & \hat{\emptyset}_{kj} > \hat{\emptyset}_{\mathrm{BB}j} \end{cases} \quad (1)$$

where $\hat{\emptyset}_{\mathrm{AA}j}$, $\hat{\emptyset}_{\mathrm{AB}j}$, and $\hat{\emptyset}_{\mathrm{BB}j}$ are the cluster centre normalised angles $\hat{\emptyset}_{\mathrm{AA}j} = \emptyset_{\mathrm{AA}j}/(\frac{\pi}{2})$.

This calibration function was applied to the raw pooling results to generate a calibrated continuous allele frequency.

E. Repeatability tests

The majority of individuals correspond to a single sample in the iE2 test; however, to test the experimental repeatability 48 individuals had duplicate samples. After bad individual removal 41 individuals were available. The RMS Error of the $\hat{\emptyset}$ values were calculated for these 41 individuals resulting in 0.065 RMS Error. This value indicates the inherent noise of the measurement system which cannot currently be eliminated and allows an estimation of the lower bound for the analysis accuracy.

F. Ground truth calculation

The ground truth continuous allele frequency for each pool was calculated from the individual results and knowledge of which individuals were in each pool. By calculating the contribution of the individuals to the pool the allele frequency can be calculated. Individual allele frequencies are in the set $\{1, 0.5, 0.0\}$, the pooled result is simply the average over the individuals in the pool.

G. Bad sample removal

Pooling samples were removed where the amplitude was less than a threshold ~1 which resulted in a further ~15% results being discarded corresponding to 3 pools entirely and 11 samples across the pool SNPs.

H. Calculate allele frequency function

The final allele residual is calculated by comparing the calibrated allele frequency with the ground truth. With no calibration the residual RMSE in $\hat{\emptyset}$ is 0.135, with calibration the RMSE in $\hat{f}$ falls to 0.120.

IV. Proposed Method for Pooled Allele Estimation

A. Limitations of existing approach

Observation of the clusters in Fig. 3 show significant variation from SNP to SNP of not only the cluster centre locations, but the spread and shape of the clusters. This motivates the idea that the calibration function should not only interpolate the cluster centres, but the form of the calibration function should vary from SNP to SNP. This proposition is validated by experimentation with different interpolating polynomials (Fig. 4). By comparing the calibrated result with the ground-truth allele frequency the form of the most accurate polynomial can be determined (Table III). Tested polynomials include piece-wise linear, second order Lagrange and piece-wise Hermite polynomials with zero derivatives enforced at end points. An equal domain Hermite variant enforces $\hat{\emptyset}$ domain to be identical for both segments. Because of the limited samples available per SNP the calibration function is kept as a function of angle only.

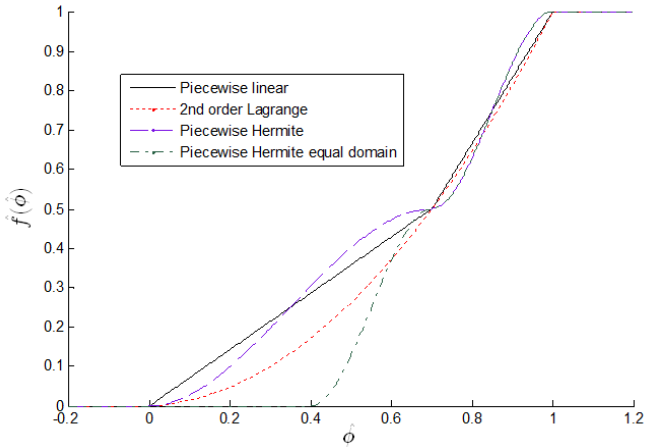

Fig. 4. Four example interpolation functions mapping cluster centres (x-axis) $\hat{\emptyset}_{AA} = 0$, $\hat{\emptyset}_{AB} = 0.7$ and $\hat{\emptyset}_{BB} = 1$ to calibrated allele frequencies (y-axis) $\hat{f}_{kj}(\hat{\emptyset}_{AA}) = 0$, $\hat{f}_{kj}(\hat{\emptyset}_{AB}) = 0.5$ and $\hat{f}_{kj}(\hat{\emptyset}_{BB}) = 1.0$. Hermite interpolation restricted to zero derivative at end points.

The results in Table III clearly demonstrate a limitation in the existing approach by using a single piece-wise linear interpolation. Improvement can be achieved by generalizing the calibration polynomial.

A second limitation of the existing approach occurs because the calibrated value $\hat{f}_{kj}$ is the computed allele frequency. However $\hat{f}_{kj}$ is not an ideal representation of allele frequency as it misses information common to all SNPs that is not calibrated by SNP calibration functions. This includes amplitude dependant distortion and measurement artifacts that become apparent with the larger dataset.

TABLE III. BEST INTERPOLATION POLYNOMIAL CALCULATED FOR EACH SNP

Interpolation polynomial	% of SNPs where polynomial is best
Piecewise linear	56%
2nd order Lagrange	33%
Piecewise Hermite	4%
Piecewise Hermite equal domain	6%

B. *Proposed learning framework structure*

We propose a hierarchical learning approach to simultaneously resolve the limitations in the previous Section. The framework (Fig. 5) improves the individual SNP calibration and allows universal distortions common to all SNPs introduced by the hardware to be corrected. The first learning algorithm finds a calibration function for each SNP based on the angle ($\hat{\emptyset}$) only. The cluster centres are not enforced explicitly as we train the learner on both pool values and the full available set of individual values. After training these 48 SNP learners, the outputs are fed to the higher layer. The higher layer learns an offset function dependant on both amplitude and angle. This offset is added to the output of the first layer. It is trained on the residual error at the output of the first layer. The inclusion of this hierarchical learner into the analysis process involves replacing the box in Fig. 1 left, with the box in Fig. 1 right.

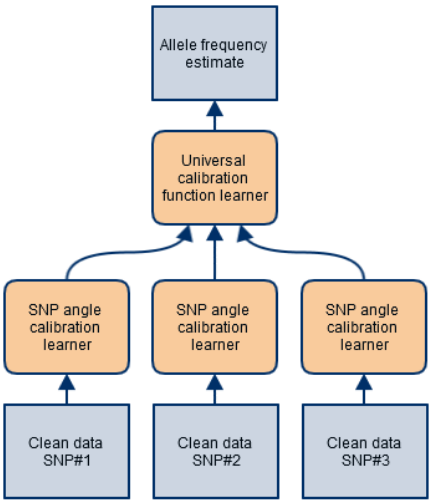

Fig. 5. Hierarchical machine learning approach

A standard 10-fold cross-validation was applied to the 901 cleaned pool SNP samples. 90 samples were withheld for testing and the remaining samples used for training. The 901 samples were ordered to ensure each SNP had a similar number of training samples available for calibration function generation. The training samples and corresponding ground-truth values were used to train the full hierarchical learning system. The 90 samples were then used to test the full system and estimate error in allele frequency. This process was repeated by the cycling the withheld testing samples 10 times through the data set such that the testing sets partition 900 samples in all. The final allele frequency MSE was averaged over 900 testing sets to calculate RMSE.

C. *Learning algorithms*

Learning algorithms implemented include a linear regression, a multi-layer perceptron neural network and lib-svm (mu-SVM). The WEKA package [18] implementations were used for these algorithms. Optimal hyper parameter tuning was achieved via parameter sweep on the 10-fold cross-validation with just 2 SNPs and selecting the best results for each learner.

V. RESULTS

All combinations of learners for calibration and allele frequency layers were tested. Baseline cases including no calibration or calibration via piecewise linear interpolation (PwL) were included for comparison. Results are presented in Table IV.

As expected inclusion of per SNP calibration improves the accuracy of allele frequency estimation. Results for the hierarchical learners further improve the accuracy with the

best value having an RMS error of 0.078. The error in allele frequency can be decomposed into the platform's measurement error (E_{RMS}^{M}) and the error introduced by the calibration approach (E_{RMS}^{C}) such that total mean square error

$$E_{RMS}{}^{2} = (E_{RMS}^{M})^{2} + (E_{RMS}^{C})^{2}.$$

Given the measurement process RMS error which was described in Section III.E (E_{RMS}^{M}~0.065) we can calculate the error introduced by the various calibration approaches (E_{RMS}^{C}) from the results in Table IV. These results (E_{RMS}^{C}) are shown in Table V after normalising to RMS error introduced with no calibration. The results show the standard approach reduces normalised RMS error to 0.86 whereas the hierarchical approach reduces it further to 0.36. The standard approach accounts for ~14% of the distortion whereas our approach accounts for ~64% of the distortion. Linear regression does a good job of accounting for the majority of the error when combined in the hierarchical model. LibSVM is able to achieve superior results when used in the initial calibration step.

TABLE IV. RMS ERROR OF ALLELE FREQUENCY ESTIMATE FOR ALL SINGLE SNP CALIBRATION TECHNIQUES AND HIGHER LAYER LEARNERS

Calibration learner	Universal Learner			
	None	***L.reg***	***MLP***	***libSVM***
None	0.135 (0.017)	0.110 (0.013)	0.113 (0.013)	0.110 (0.012)
PwL	0.120 (0.017)	0.094 (0.013)	0.102 (0.013)	0.094 (0.081)
L.reg.	0.087 (0.014)	0.081 (0.013)	0.084 (0.011)	0.082 (0.012)
MLP	0.092 (0.015)	0.089 (0.014)	0.090 (0.013)	0.088 (0.013)
libSVM	0.081 (0.015)	0.078 (0.014)	0.082 (0.013)	0.078 (0.013)

TABLE V. COMPARISON OF BEST LEARNING APPROACH TO EXISTING METHODS

Calibration approach	No universal learner	Best universal learner
None	1	0.75
Standard calibration	0.86	0.41
Best calibration learner	0.41	0.36

VI. CONCLUSION

The contribution of this paper is a new hierarchical learning framework. The framework solves the problem of SNP to SNP bias by applying learning algorithms to calibrate out this variation. The framework also takes advantage of the entire SNP dataset to learn a global allele frequency correction. As a result the framework achieves superior performance over the existing approach when estimating pooled allele frequency. The impact of this improved allele frequency estimation is to enable genetic studies using pooled samples which previously were not possible due to elevated allele frequency error.

Future work includes testing the framework on other SNP genotyping platforms such as Illumina or Affymetrix. Possible extensions include introducing amplitude dependant SNP calibration where enough ground truth is available per SNP, or alternatively clustering similar SNPs and calibrating on a SNP cluster basis where less ground truth data is available.

ACKNOWLEDGMENT

We would like to acknowledge Gold Coast Marine Aquaculture for their contribution towards the development of the Black Tiger Prawn SNP assay used in this study, and for the tissue samples used in evaluating the methods. We are grateful to Leanne Dierens and Melony Sellars who undertook sample collection and DNA extractions.

[1] G. Luikar, P. England, D. Tallmon, S. Jordan and P.Taberlet, "The power and promise of population genomics: from genotyping to genome typing," in *Nature Review Genetics,* vol. 4, 2003, pp. 981-994.

[2] K. Gunderson, F. Steemers, G. Lee, L. Mendoza and M. Chee, "A genome wide scalable SNP genotyping assay using microarray technology", *Nature Genetics,* vol. 37, 2005, 549-554.

[3] G. Kennedy, H. Matsuzaki, S. Dong, W. Liu, J. Huang, G. Liu, X. Su, M. Chu… "Large-scale genotyping of complex DNA", *Nature Biotechnology,* vol. 21, 2003, 233-237.

[4] P. Sham, J. Bader, I. Craig, M. O'Donavon and M. Owen, "DNA Pooling: a tool for large-scale association studies", *in Nature Review Genetics*, vol. 3, 2002, pp. 862-871.

[5] N. Norton, N. Williams, H. Williams , G. Spurlock, G. Kirov, D. Morris, B. Hoogendorn, M. Owen and M. O'Donovan, "Universal, robust, highly quantitative allele frequency measurement in DNA pools", *in Human Genetics*, vol. 110, 2002, pp. 471-478.

[6] A. Earp, M. Rahmani, K. Chew and A. Brooks-Wilson, "Estimates of array and pool-construction variance for planning efficient DNA-pooling genome wide association studies," *BMC Med Genomics,* vol: 4 (81), 2011.

[7] S. Gabriel, L. Ziuagra and D. Tabbaa, "SNP Genotyping using the Sequenom MassARRAY iPLEX Platform," in *Current Protocols in Human Genetics*, vol. 60(2), 2009, pp. 12.1–2.12.16

[8] G. Kirov, I. Nikolov, L. Georgrieva. V. Moskvina,. M. Owen and M. O'Donavon "Pooled DNA genotyping on Affymetrix SNP genotyping arrays," in *BMC Genomics*, vol. 7:27, 2006, pp. 15–64.

[9] S. MacGregor, Z.Z. Zhao, A. Henders, N. Martin, G. Montgomery and P. Visscher, "Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays," in *Nucleic Acids Research*, vol. 36 (6), 2008.

[10] R. Abraham**,** V. Moskvina**,** R. Sims**,** P. Hollingworth**,** A. Morgan**,** L. Georgieva**,** K. Dowzell**,** S. Cichon**,** A. Hillmer**,** M. O'Donovan**,** J. Williams**,** M. Owen and George Kirov**,** "A genome-wide association study for late-onset Alzheimer's disease using DNA pooling", in *BMC Medical Genomics*, vol 1(44)**,** 2008

[11] L. Butcher, O. Davis, I Craig and R. Plomin, "Genome-wide quantatitive trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays, in *Genes Brain Behaviour*, vol. 7, 2008, pp. 435-446.

[12] Sequenom, 'Model based clustering of genotyping samples using a "Mixture of Gaussians" approach, Sequenom Technical Note, 2007.

[13] F. Pompanon,, A. Bonin, E. Bellemain and P. Taberlet "Genotyping Errors: Causes, Consequences, and Solutions," in Nature Reviews Genetics, 2005, pp. 847–859.

[14] A.C. Syvänen, "Accessing Genetic Variation: Genotyping Single Nucleotide Polymorphisms," in Nature Reviews Genetics, 2001, pp. 930–942.

[15] H. Yang, Y. Liang, M. Huang, L Li, C. Lin, J. Wu, Y. Chen and C. Fann "A genome-wide study of preferential amplification/hybridization in microarray-based pooled DNA experiments," in Nucleic Acids Research vol. 34 (15), 2006.

[16] S. Macgregor, P. Visscher and G. Montgomery, "Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates," in Nucleic Acids Research vol. 34(7), 2006

[17] D. Peiffer, J. Le, F. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C. Shaw, J. Belmont, S. Cheung, R. Shen, D. Barker, K. Gunderson, "High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping," in Genome Research, 2006 pp. 1136–1148

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten "The WEKA Data Mining Software: An Update," in SIGKDD Explorations, Vol. 11(1), 2009.