On Assigning Individuals from Cryptic Population Structures to Optimal Predicted Subpopulations: An Empirical Evaluation of Non-parametric Population Structure Analysis Techniques

Pornchalearm Deejai[1], Anunchai Assawamakin[2], Pongsakorn Wangkumhang[2], Kanokwan Poomputsa[3], and Sissades Tongsima[2,*]

[1] Bioinformatics and Systems Biology Program, King Mongkut University of Technology
Thonburi, Bangkok 10140, Thailand
[2] Biostatistics and Informatics Laboratory, Genome Institute,
National Center for Genetic Engineering and Biotechnology, Thailand Science Park,
Paholyothin Road, Pathumthani 12120, Thailand
[3] School of Bioresources and Technology, King Mongkut University of Technology
Thonburi, Bangkok 10140, Thailand
sissades@biotec.or.th

Abstract. Many algorithms have been proposed to analyze population structures from the single nucleotide polymorphism (SNP) genotyping data of some number of individuals and try to assign individuals to genetically similar groups. These algorithms can be categorized into two computational paradigms: parametric and non-parametric approaches. Although the parametric-based approach is a gold standard for population structure analysis, the computational burden incurred by running these algorithms is unacceptable for large complex dataset. As genotyping platforms incorporating more SNPs, analyzing ever larger and more complex datasets are becoming a standard practice. Hence, the computationally efficient non-parametric methods for analysis of genotypic datasets are needed to reveal the population structure. In this study, we evaluated two leading non-parametric population structure analysis techniques, namely ipPCA and AWclust, on their abilities to characterize the genetic diversity and population structure of two complex SNP genotype datasets (as many as 243855 SNPs). The head-to-head comparisons were conducted on two major aspects: ability to infer the number of genetically related subpopulations (K) and ability to correctly assign individuals to these subpopulations. The experimental results suggested that AWclust could be more suitable when applying to a small and less complex dataset. However, with a large and more complex dataset, ipPCA is a much better choice yielding higher accuracy on assigning genetically similar individuals to the inferred groups.

Keywords: Population genetic, Population genetic structure, parametric-based method, non-parametric-based method.

[*] Corresponding author.

J.H. Chan, Y.-S. Ong, and S.-B. Cho (Eds.): CSBio 2010, CCIS 115, pp. 58–70, 2010.

1 Introduction

Population genetics is concerned with the structure of different populations, which can be observed by frequency differences among the populations. Population structure analysis is important to genetic association studies [1-4] and evolutionary investigations [5-7]. Since most studies of human variation focus on sampling from predefined "populations" using culture and/or their geographical origins, these populations may not reflect the underlying genetic relationships [8-9]. Many algorithms have been proposed to analyze population structures from the single nucleotide polymorphism (SNP) genotyping data of some number of individuals and try to assign individuals to genetically similar groups. These algorithms can be categorized into two major computational paradigms: parametric and non-parametric approaches.

Parametric approaches require assumption of genetic model to assign individuals with similar genetic background to a predefined number of subpopulations (K). Such an assignment is carried out based on statistical likelihood using assumptions such as Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE) among loci for each population [10,11]. The parametric approach, e.g., STRUCTURE, has been used as standard practice on population structure analyses. Nonetheless, the computational burden incurred by running these algorithms is unacceptable for solving large complex dataset. Furthermore, the statistical estimators of HWE and LE may not hold by any statistical estimators due to randomness in sampling. For this scenario, the non-parametric methods are more appropriate for analyzing population structure than parametric methods. These non-parametric approaches use standard statistical techniques to search for relatedness of genetic signal among data instead of finding the -best-fit likelihood of presumed genetic model. Two most recent reports of the algorithms in this class include ipPCA [12] and AWclust [13].

As SNP genotyping data becomes ever larger, it is increasingly difficult to efficiently analyze population structure by means of parametric statistical techniques due to their computational intensive requirements. The non-parametric approaches are becoming viable tools for researchers to understand population diversity and structure. Both ipPCA and AWclust tools have both advantages and disadvantages, but it is still not clear how these methods differ in their power to analyze population structure and suitability for analyzing large and complex SNP genotype data in terms of individual assignment and estimation of the optimal number of subpopulations (K). Hence, this paper aims to empirically evaluate these two aspects (inferring K and individual assignment) of these non-parametric algorithms. This evaluation was conducted on large complex datasets, 1) worldwide human dataset from Xing et al [14] containing 586 individuals from 28 populations 243855 SNPs and 2) BovineHapMap dataset containing 497 samples from 19 predefined breeds 27203 SNPs. 3) Simulated dataset from program GENOME [15] containing 20 subpopulations 10000 SNPs. These comparison results from of all datasets can suggest researchers to select non-parametric tools to analyze their datasets.

2 Material and Methods

2.1 Dataset

There are one simulated and two real datasets used in this study. For the simulated dataset, we create a population model shown in **Fig. 1** and use the program GENOME [15] to generate the genotypic data under the Wright-Fisher neutral coalescent model (backward in time) [16]. The simulated model contains 20 subpopulations derived from three ancestral populations. The simulated data contains 400 individuals with 10000 SNPs. This simulated model was used to test both AWclust and ipPCA by simulating 30 datasets from this model as the inputs to these algorithms. These simulated datasets with only 10000 SNPs are much less complex when comparing with the real datasets. The first real dataset represents a complex dataset with a large number of subpopulations but with a smaller number of SNP markers. The second group of real dataset represents a very complex dataset both in number of subpopulations and the number of SNP markers. The first real dataset is the SNP genotype of 497 cattle from BovineHapMap Project obtained from 19 different biologically diverse breeds. Due to computational limitation of Gap Statistics, AWclust demarcates the number of maximum inferred subpopulations to 16. In order to perform the experiment, the dataset was reduced to 15 breeds, containing 368 individuals with 27203 SNPs, by dropping individuals labeled as HOL, HFD, JER, and GNS from the original dataset. We also use a complex dataset from [14] to test ipPCA algorithm. This dataset represent a large population from 27 worldwide populations from Africa, Asia, and Europe in which we added our 32 samples from Thai population making up 586 individuals with 243855 SNPs. Tables 1 and 2 present the detail information of these two complex datasets.

Table 1. The number of individuals of each cattle breed in BovineHapMap. The total number of subpopulations is 19 using the three letter labeling as follows: ANG, Angus; BMA, Beefmaster; BRM, Brahman; BSW, Brown Swiss; CHL, Charolais; GIR, Gir; LMS, Limousin; NDA, N'Dama; NEL, Nelore; NRC, Norwegian Red; PMT, Piedmontese; Gertrudis; SHK, Sheko.RGU, Red Angus; RMG, Romagnola SGT, Santa; HFD, Hereford; GNS, Guernsey; HOL, Holstein; JER, Jersey. The asterisk (*) symbol indicates the breeds that were removed from the experiment so as to meet the K=16 limitation imposed by AWclust.

Breed	Count	Breed	Count
CHL	24	RMG	24
GIR	24	SHK	20
HFD *	27	BSW	24
ANG	27	NDA	25
BRM	25	NEL	24
HOL *	53	BMA	24
JER *	28	GNS *	21
LMS	42	NRC	25
PMT	24	SGT	24
RGU	12		
		Total	**497**

Table 2. The number of individuals from 27 subpopulations reported in [14]. The 32 Thai samples (unpublished data) were added to this dataset. The total number of individuals is 586 samples with 243855 SNPs.

Ethnic Group	No. of individuals	Ethnic Group	No. of individuals
Alur	10	CHB	45
Hema	15	JPT	45
Pygmy	25	Luhya	24
Brahmin	25	Tuscan	25
Utah	25	Kung	13
Khmer	5	Pedi	10
Chinese	7	Sotho_Tswana	8
Dalit	13	Stalskoe	5
Irula	24	Iban	25
Japanese	13	Chinese_TW	3
Madiga	10	Tamil	14
Mala	11	Urkarah	18
CEU	60	Veitnam	7
YRI	60	Nguni	9
		Thai	32
		Total	**586**

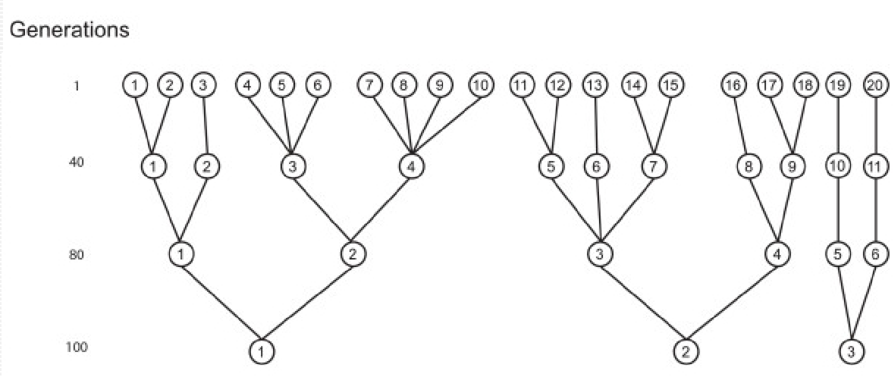


Fig. 1. Population history trees for generating simulated datasets. The GENOME tool [15] was used to generate the simulated datasets.

2.2 Comparison of the Two Non-parametric Algorithms

In this paper, the two non-parametric population structure analysis algorithms were studied in terms of their performance. Both algorithms claimed that they could efficiently operate on dataset on which the parametric STRUCTURE algorithm would be infeasible to operate. The following paragraphs describe fundamental non-parametric techniques deployed in each algorithm.

The ipPCA makes use of an exploratory data analysis technique, called principal component analysis (PCA), to observe common pattern from given genetic data by means of covariance analysis. The algorithm markedly improves resolution of population substructure by an iterative pruning process. It first performs PCA on the dataset and uses fuzzy c-mean algorithm [17] to cluster the PCA result to split the transformed data into two prominent. The process is repeated on each of the split group. The terminating condition is verified, for every run of PCA, using the TW test statistic described in EIGENSTRAT/SmartPCA [18,19]. The default TW p-value threshold used for detecting structure is conservative ($p < 10^{-12}$). This software is publicly available from http://www4a.biotec.or.th/GI/tools/ippca.

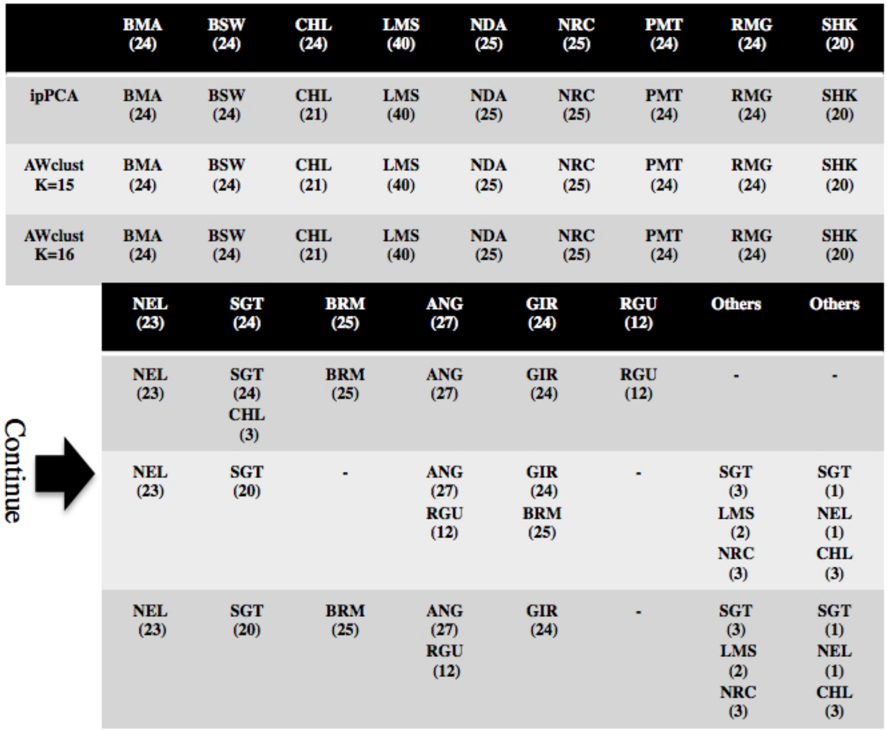
The AWclust software calculates the allele sharing distance (ASD) matrix, which represents the underlying genetic distance between every pair of individuals. It performs non-parametric exploration with the SNP data set by generating multidimensional scaling (MDS) 2D/3D plots to get a general idea of how the data clusters and to detect any outliers in the dataset. The MDS plot helps reveal outliers in the dataset and identify clusters and general relationships among individuals. AWclust calculates the Gap statistic for estimating the optimal number of groups based on the sample genetic relatedness. The Gap statistics compares the pooled within-cluster sum of squares with its expectation from a null reference distribution. Hence, the precision of this method requires multiple simulations from the null reference distribution. This process, however, is computationally intensive. The data points are then plotted ranging by cluster sizes and the optimal size maximizes the distance between the observed and expected pooled within-cluster sum of squares. The resulting hierarchical plots may also help interpret Gap statistic plots [13]. This software is publicly available from http://awclust.sourceforge.net/

3 Results

In this section, we present the results obtained by running ipPCA and AWclust algorithms to analyze BovineHapMap SNP dataset. Individual assignment tables were created to report results obtained from the two algorithms. Each column represents the genetically related group inferred by each algorithm. To make the tables more readable, we labeled each group to match the breed name originally given when the samples were first collect.

3.1 ipPCA Analysis

The ipPCA program was used to analyze the dataset with 27203 SNPs of BovineHapMap hosting 15 breeds. By observing the terminal nodes produced by ipPCA, this dataset can be re-organized into K=15 genetically related clusters. This is in concordance with the breed labels previously assigned at the sample collection time. **Fig. 2** presents the individual assignment to the terminal nodes of ipPCA. Most of the assignments agree with the breed labels previously specified, except the 3 samples from the CHL breed. We also experimented on the whole data set of BovineHapMap (with 19 breeds). The ipPCA algorithm was able to predict 19 genetically similar groups (K=19), which is the same as the breeds. The assignment was preformed yielding the assignment accuracy as high as 99.2 percent as shown in **Fig. 3**.



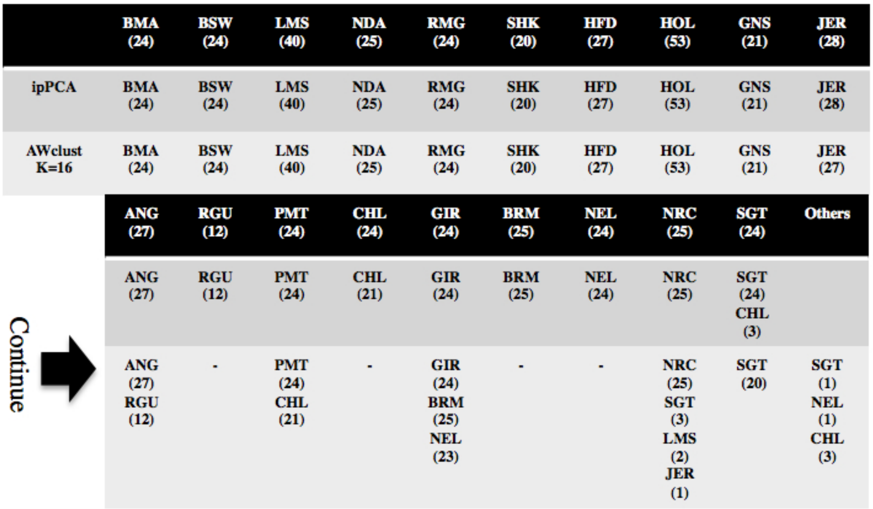
	BMA (24)	BSW (24)	CHL (24)	LMS (40)	NDA (25)	NRC (25)	PMT (24)	RMG (24)	SHK (20)	NEL (23)	SGT (24)	BRM (25)	ANG (27)	GIR (24)	RGU (12)	Others	Others
ipPCA	BMA (24)	BSW (24)	CHL (21)	LMS (40)	NDA (25)	NRC (25)	PMT (24)	RMG (24)	SHK (20)	NEL (23)	SGT (24) CHL (3)	BRM (25)	ANG (27)	GIR (24)	RGU (12)	-	-
AWclust K=15	BMA (24)	BSW (24)	CHL (21)	LMS (40)	NDA (25)	NRC (25)	PMT (24)	RMG (24)	SHK (20)	NEL (23)	SGT (20)	-	ANG (27) RGU (12)	GIR (24) BRM (25)	-	SGT (3) LMS (2) NRC (3)	SGT (1) NEL (1) CHL (3)
AWclust K=16	BMA (24)	BSW (24)	CHL (21)	LMS (40)	NDA (25)	NRC (25)	PMT (24)	RMG (24)	SHK (20)	NEL (23)	SGT (20)	BRM (25)	ANG (27) RGU (12)	GIR (24)	-	SGT (3) LMS (2) NRC (3)	SGT (1) NEL (1) CHL (3)

Continue

Fig. 2. Analysis results of reduced BovineHapMap dataset (15 breeds). This figure presents the individual assignment ipPCA (observed at the terminal nodes generated by ipPCA tree) and the individual assignment results obtained from the cut tree of AWclust. Each column represents a genetically similar group, which both algorithms assigned the samples to. The columns labeled "others" represent the extra groups suggested by Gap statistics. The number of samples is shown in parentheses. For demonstration purpose, we tried to put the same assigned breed name in the same column. The "-" symbol indicates no such group name, implying that the samples might be in the same group with other samples. The first row indicates the assignment results done by ipPCA. These results demonstrate that most of the assignments agree with the breed labels previously specified, except the 3 samples from the CHL breed mixing with the SGT one. The second and third rows of the table reports the assignment results of AWclust when setting K=15 and 16 respectively.

We also applied ipPCA to analyze the large and complex dataset [14] combining with our unpublished SNP genotype data of 32 Thai individuals. This combined dataset forms 28 different ethnics groups, geographically distributed around the world. ipPCA was able to infer 15 genetically similar groups (K=15). The individual assignment results observed at each terminal nodes of the ipPCA bifurcation tree are shown in **Fig 5**.

Moreover, we also applied ipPCA framework to analyze the simulated dataset, which was generated from program GENOME [15] and derived from three ancestral populations. To be able to compare with AWclust, the simulated data was reduced to



	BMA (24)	BSW (24)	LMS (40)	NDA (25)	RMG (24)	SHK (20)	HFD (27)	HOL (53)	GNS (21)	JER (28)
ipPCA	BMA (24)	BSW (24)	LMS (40)	NDA (25)	RMG (24)	SHK (20)	HFD (27)	HOL (53)	GNS (21)	JER (28)
AWclust K=16	BMA (24)	BSW (24)	LMS (40)	NDA (25)	RMG (24)	SHK (20)	HFD (27)	HOL (53)	GNS (21)	JER (27)

Continue	ANG (27)	RGU (12)	PMT (24)	CHL (24)	GIR (24)	BRM (25)	NEL (24)	NRC (25)	SGT (24)	Others
	ANG (27)	RGU (12)	PMT (24)	CHL (21)	GIR (24)	BRM (25)	NEL (24)	NRC (25)	SGT (24) CHL (3)	
	ANG (27) RGU (12)	-	PMT (24) CHL (21)	-	GIR (24) BRM (25) NEL (23)	-	-	NRC (25) SGT (3) LMS (2) JER (1)	SGT (20)	SGT (1) NEL (1) CHL (3)

Fig. 3. ipPCA and AWclust results of BovineHapMap dataset with 19 subpopulations. The ipPCA algorithm was able to predict 19 genetically similar groups (K=19), which is the same as the breeds. Using K=16, AWclust was not able to correctly assign the individuals to the pre-allocated groups.

15 subpopulations with 300 individuals, we found that ipPCA was able to infer the correct K with results swinging between K=14 and K=15 (**Fig 6**.). When we test ipPCA against the full dataset (20 pops, 400 individuals), the classification accuracy was much improved. The individual assignment results observed at each terminal nodes of the ipPCA bifurcation tree are shown in **Fig 7**.

3.2 AWclust Analysis

AWclust calculated the allele sharing distance (ASD) matrix from the raw data and calculate the Gap statistics for estimating the number of K. To predict the optimal K, the module Gap statistics must be performed incrementally; the highest Gap statistics value (y-axis) indicates the most probable K (x-axis). We present the Gap statistics values ranging from K= 1 to 16 in **Fig 4**.

AWclust was applied to analyze 27,203 SNPs of the reduced BovineHapMap dataset. The Gap statistics suggested the optimal K to be either 15 or more than 16 (the maximum inferred K is 16 for AWclust). **Fig 4** presents the Gap statistic results suggesting the value of K to be used in the individual assignment step. Next, AWclust used this K number to create a cut on dendogram plot in order to inform us which individuals belong to what groups (see AWclust user manual for more information on the hierarchical clustering and its dendogram plot). Since it is not certain if inferred K should be 15 or more groups, we tried to create different cuts based on K=15 and 16. The assignment results are tabulated in **Fig 2**. In this figure, the assignment results of AWclust when using K=16 are worse than those results when using K=15. Extra groups were created with mixed individuals from different breeds assigned to the group (see the columns “others” in **Fig 2**).

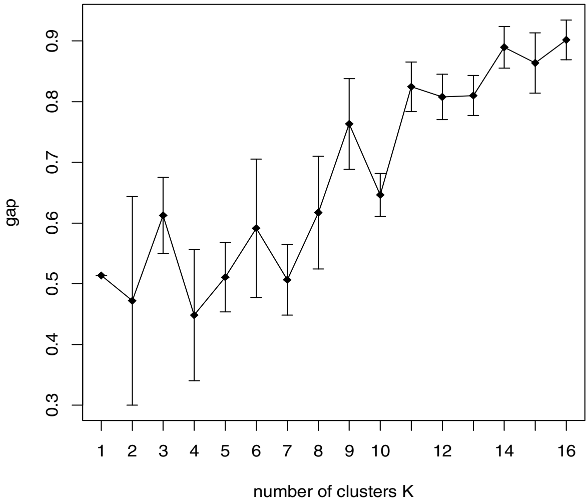
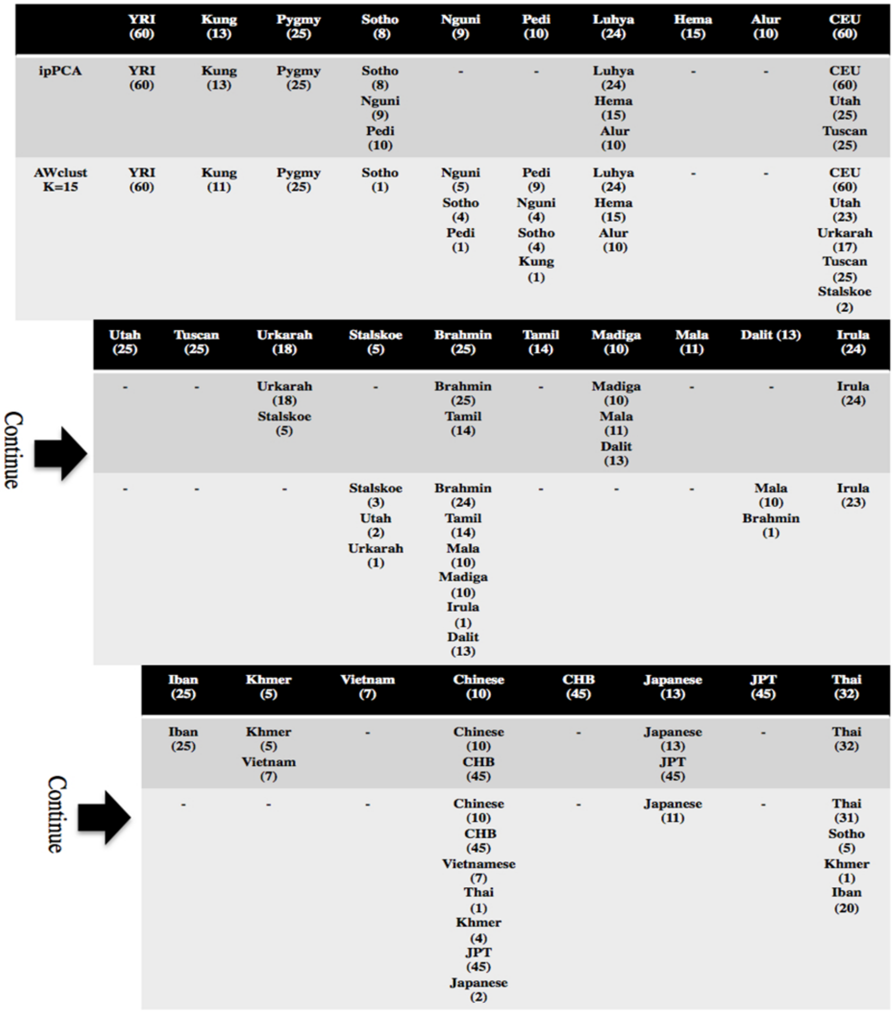


Fig. 4. Gap statistic result from reduced BovineHapMap dataset. The numbers of inferred Ks ranging from 1 to 16 are shown in the graph. The x-axis represents different possible Ks and the y-axis represents the gap value (higher is better).

Since we cannot verify if the Gap statistics can accurately predicting the correct K, for the full set of BovineHapmap data containing 19 breeds, we omitted the Gap statistics step and applied different larger Ks (K=15 to K=18) to test the individual assignment function of AWclust. Similar to the case K=16, larger K values resulting in incorrect cuts on the hierarchical clustering dendogram. **Fig. 3** presents the assignment results on the full BovineHapmap dataset. However, too many mis-assignments, i.e., having a group of combined breeds or the same breed get split to two or more different groups, are observed. Due to the page limitation of this conference, the assignment data when K=17 and 18 are not shown in this paper.

AWclust was deployed to analyze a very large and complex dataset, the 28 worldwide population dataset. Since the number of geographic subpopulations is far greater than the limit which was set for Gap statistic, the experiments on this dataset will only test the individual assignment accuracy. Similar to the BovineHapmap situation, we assume that Gap statistic could infer K to be any value larger than 16. We then used these numbers to create cut trees, which in turn gave us the individual assignment results. **Fig 5** shows the assignment given by AWclust assuming K=15. Unlike, the results shown in ipPCA row, the AWclust assignment tends to group unrelated individuals together. These groups are in the form of mixed populations in which some populations were split and assigned to several other groups. The AWclust assignment results got worst when increasing the number of K (data not shown).

AWclust was tested against the reduced simulated dataset (15 pops with 300 individuals having 10000 SNPs each). The experiment was repeatedly performed for 30 times on both reduced and full simulated datasets. We found that AWclust was able to infer the correct K with 100% accuracy for individual assignment of 15 subpopulations (see **Fig 6**.). On the other hand, when the dataset become more complex, we found AWclust failed to correctly infer K (see **Fig 7.).**

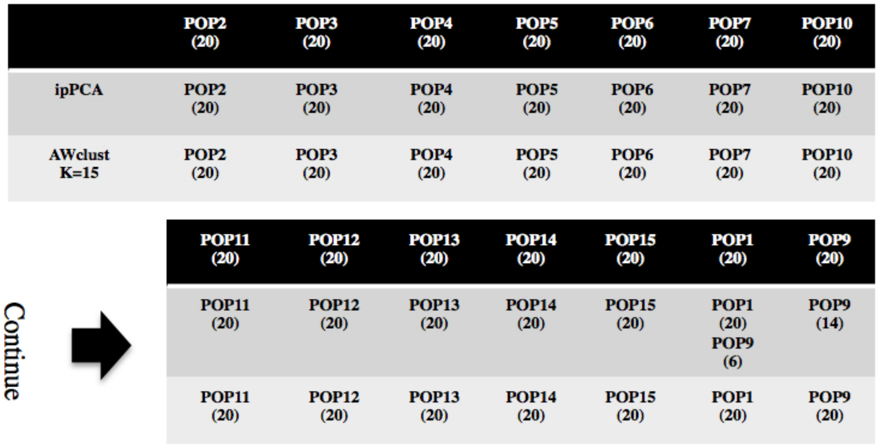

Continue
Continue

	YRI (60)	Kung (13)	Pygmy (25)	Sotho (8)	Nguni (9)	Pedi (10)	Luhya (24)	Hema (15)	Alur (10)	CEU (60)
ipPCA	YRI (60)	Kung (13)	Pygmy (25)	Sotho (8) Nguni (9) Pedi (10)	-	-	Luhya (24) Hema (15) Alur (10)	-	-	CEU (60) Utah (25) Tuscan (25)
AWclust K=15	YRI (60)	Kung (11)	Pygmy (25)	Sotho (1)	Nguni (5) Sotho (4) Pedi (1)	Pedi (9) Nguni (4) Sotho (4) Kung (1)	Luhya (24) Hema (15) Alur (10)	-	-	CEU (60) Utah (23) Urkarah (17) Tuscan (25) Stalskoe (2)

Utah (25)	Tuscan (25)	Urkarah (18)	Stalskoe (5)	Brahmin (25)	Tamil (14)	Madiga (10)	Mala (11)	Dalit (13)	Irula (24)
-	-	Urkarah (18) Stalskoe (5)	-	Brahmin (25) Tamil (14)	-	Madiga (10) Mala (11) Dalit (13)	-	-	Irula (24)
-	-	-	Stalskoe (3) Utah (2) Urkarah (1)	Brahmin (24) Tamil (14) Mala (10) Madiga (10) Irula (1) Dalit (13)	-	-	-	Mala (10) Brahmin (1)	Irula (23)

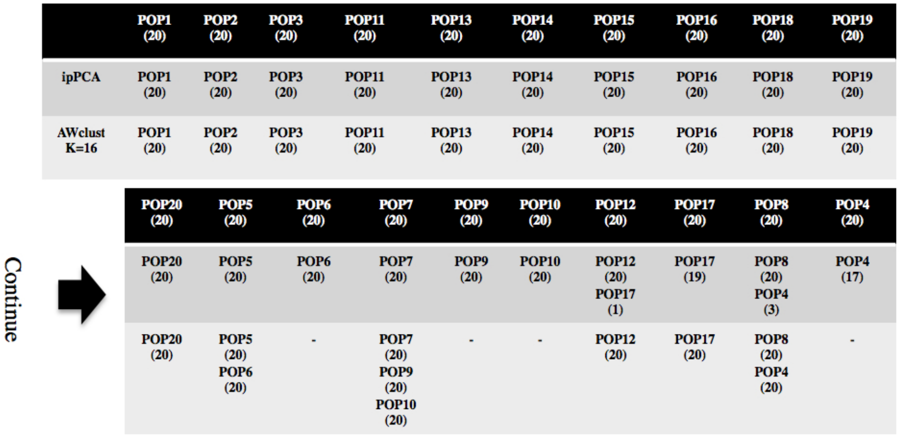
Iban (25)	Khmer (5)	Vietnam (7)	Chinese (10)	CHB (45)	Japanese (13)	JPT (45)	Thai (32)
Iban (25)	Khmer (5) Vietnam (7)	-	Chinese (10) CHB (45)	-	Japanese (13) JPT (45)	-	Thai (32)
-	-	-	Chinese (10) CHB (45) Vietnamese (7) Thai (1) Khmer (4) JPT (45) Japanese (2)	-	Japanese (11)	-	Thai (31) Sotho (5) Khmer (1) Iban (20)

Fig. 5. Result of analysis with dataset from [14] combining with 32 Thai samples. ipPCA was able to infer 15 genetically similar groups (K=15) and assigned most related individuals to these predicted groups (see [14] for detail discussion on these populations]. AWclust, however, was not able to predict the K due to the Gap statistic limitation. To make it comparable with ipPCA, we set K=15 for AWclust. The AWclust assignment is shown in the row under that of ipPCA. Since there are 28 populations being observed while only 15 groups to assign individuals to, mixed populations of different combinations can be expected. For demonstration purpose, the table was split into three parts to accommodate different mixed-pop combinations.



	POP2 (20)	POP3 (20)	POP4 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP10 (20)	POP11 (20)	POP12 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP1 (20)	POP9 (20)
ipPCA	POP2 (20)	POP3 (20)	POP4 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP10 (20)	POP11 (20)	POP12 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP1 (20) POP9 (6)	POP9 (14)
AWclust K=15	POP2 (20)	POP3 (20)	POP4 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP10 (20)	POP11 (20)	POP12 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP1 (20)	POP9 (20)

Fig. 6. Analysis results of reduced simulated dataset (15 subpopulations). This figure presents the individual assignment ipPCA (observed at the terminal nodes generated by ipPCA tree) and the individual assignment results obtained from the cut tree of AWclust. Each column represents a genetically similar group, which both algorithms assigned the samples to. The number of samples is shown in parentheses. For demonstration purpose, we tried to put the same assigned subpopulation name in the same column. The first row indicates the assignment results done by ipPCA. These results demonstrate that most of the assignments agree with the subpopulation labels previously specified, except the 6 samples from the population 9 mixing with the population 1. The second row of the table reports the assignment results of AWclust when setting K=15, we found the accuracy of individual assignment has 100%.



	POP1 (20)	POP2 (20)	POP3 (20)	POP11 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP16 (20)	POP18 (20)	POP19 (20)	POP20 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP9 (20)	POP10 (20)	POP12 (20)	POP17 (20)	POP8 (20)	POP4 (20)
ipPCA	POP1 (20)	POP2 (20)	POP3 (20)	POP11 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP16 (20)	POP18 (20)	POP19 (20)	POP20 (20)	POP5 (20)	POP6 (20)	POP7 (20)	POP9 (20)	POP10 (20)	POP12 (20) POP17 (1)	POP17 (19)	POP8 (20) POP4 (3)	POP4 (17)
AWclust K=16	POP1 (20)	POP2 (20)	POP3 (20)	POP11 (20)	POP13 (20)	POP14 (20)	POP15 (20)	POP16 (20)	POP18 (20)	POP19 (20)	POP20 (20)	POP5 (20) POP6 (20)	-	POP7 (20) POP9 (20) POP10 (20)	-	-	POP12 (20)	POP17 (20)	POP8 (20) POP4 (20)	-

Fig. 7. ipPCA and AWclust results of simulated dataset with 20 subpopulations. The ipPCA algorithm was able to predict 20 genetically similar groups (K=20), which is the same as the population label. Using K=16, AWclust was not able to correctly assign the individuals to the pre-allocated groups.

4 Discussion

Labeling the breed or subpopulation according to their pedigree or ethnics could be inaccurate. Genetic profile of each individual is more appropriate to distinguish subpopulations. Both non-parametric algorithms strive to infer the optimal number of genetically related groups, which may differ from the number of original labels. The number of inferred groups (K) heavily influences the accuracy in assigning individuals to the groups. The following discussion points out advantages and disadvantages on using ipPCA versus AWclust.

Practicality of ipPCA and Comparison with AWclust

In view of practical use, both ipPCA and AWclust tools are convenient to use because both of them provide graphical user interface, which is required by many life science scientists. Since both programs make use of different algorithms, the running time of these tools are also different. AWclust utilizes Gap statistics, which is computational intensive due to iterative statistical inference process. Hence, AWclust demands more computational resource than ipPCA, which makes use of PCA technique. Moreover, the larger number of SNPs dramatically slow down the execution of AWclust while this does not happen to ipPCA since it makes use of singular value decomposition (SVD), which reduces the size of correlation matrix down to the matrix rank (set by the number of individuals). Due to the slowness of Gap statistics, AWclust set a hard limit of the maximum number of inferred K to be 16; this value already doubled the upper limit set in the previous version of AWclust. For this aspect, AWclust is clearly not suitable to perform large-scale population genetic analysis of current genome wide SNP array platform.

Assignment of Individual Samples to Inferred Group K in Real Datasets

AWclust tends to perform better than ipPCA when the number of SNP markers is small and the data contain less variety of individuals (smaller number of inferred K). The simulated results of 15 subpopulations with 300 individuals and 10000 SNPs demonstrate that AWclust consistently yielded correct inferred K and able to re-assign these individuals to their original subpopulations. This experiment was repeatedly performed for 30 times on the same simulated data in order to test the robustness of these two algorithms. However, ipPCA was able to infer the correct K for merely half of all the experiments (swinging between K=14 and K=15). For the real datasets, which contain more SNPs and samples, ipPCA outperformed AWclust on both inferring K and assigning individuals to correct subpopulations. This performance discrepancy stems from the different core algorithms used in these two programs. In other words, the exploratory data analysis in PCA offers better results only when the number of informative attributes, SNPs, is large enough so that the eigenanalysis of PCA can thoroughly explore the variance profile among the input samples. AWclust core algorithms, however, rely heavily on Gap statistics to correctly predict K, which is later used to create a cut point on the hierarchical clustering dendogram. For a not so complex dataset, Gap statistics can correctly predict the optimal K. Furthermore, the hierarchical clustering in AWclust can also produce a decent dendogram based on allele sharing distance (ASD) matrix. This simple two-step process is nearly deterministic rendering AWclust to outperform ipPCA for a small simulated dataset. On the other hand, for small dataset with not many SNPs, the clustering step, during each iteration of ipPCA, tends to perform inconsistently. It is also worth noting that both non-parametric approaches can discover the genetical differences within populations that the parametric STRUCTURE approach could not see. In particular, the paper [18] used STRUCTURE to analyze BovineHapmap data and it failed to differentiate the three admixed breed, namely NEL, BRM and GIR [20]. These breeds appeared as one group in STRUCTURE view. However, the prior information suggests us that these three breeds are distinct by their nature. Both AWclust and ipPCA were able to put them in three genetically different groups. Although the discussion on STRUCTURE is beyond the scope of this work, we suspected that the genetic model used by STRUCTURE may not be able to work well on this BovineHapMap dataset.

5 Conclusion

This study empirically demonstrated the performance of non-parametric-based population structure analysis methods in the aspect of individual assignment and prediction of the number of genetically similar subpopulations when applying the algorithms to the large complex datasets. The results showed that ipPCA is more suitable when applying to large dataset. This conclusion was derived from the ability to assign individuals to K subpopulations. Furthermore, we observed that the predicted K played a significant role in the overall prediction accuracy performance. AWclust used Gap statistics, which was claimed to be optimal by the authors, can accurately infer the optimal K with small datasets. For the complex dataset with large number of SNP markers, AWclust was not able to predict the correct number of subpopulations. Furthermore, if the number of correct subpopulations was given, AWclust cannot assign individuals to the correct group. Thus, from these real experimental results ipPCA is a better choice for handling complex dataset with large number of SNPs with large variety of subpopulations. However, from the result tested on the less complex simulated dataset, PCA-based technique was not able to accurately observe the overall trend from less number of SNPs for which AWclust outperformed ipPCA. Consequently, researchers can choose the tools to analyze the data based on the number of SNP markers and the number of subpopulations.

Acknowledgments. This work was supported by the National Center for Genetic Engineering and Biotechnology of Thailand, School of Information Technology and School of Bioresources and Technology King Mongkut's University of Technology Thonburi. Anunchai Assawamakin was supported by BIOTEC postdoctoral grant.

References

1. Lander, E.S., Schork, N.J.: Genetic Dissection of Complex Traits. Science 265(5181), 2037–2048 (1994)
2. Risch, N.J.: Searching for Genetic Determinants in the New Millennium. Nature 405, 847–856 (2000)

3. Marchini, J., Cardon, L.R., Phillips, M.S., Donnelly, P.: The Effects of Human Population Structure on Large Genetic Association Studies. Nat. Genet. 36(5), 512–517 (2004)
4. Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., Pato, M.T., Petryshen, T.L., Kolonel, L.N., Lander, E.S., Sklar, P., Henderson, B., Hirschhorn, J.N., Altshuler, D.: Assessing the Impact of Population Stratification on Genetic Association Studies. Nat. Genet. 36, 388–393 (2004)
5. Cavalli-Sforza, L.L., Menozzi, P., Piazza, A.: The History and Geography of Human Genes. Princeton University Press, Princeton (1994)
6. Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J., Cavalli-Sforza, L.L.: High Resolution of Human Evolutionary Trees with Polymorphic Microsatellites. Nature 368, 455–457 (1994)
7. Mountain, J.L., Cavalli-Sforza, L.L.: Multilocus Genotypes, a Tree of Individuals, and Human Evolutionary History. Am. J. Hum. Genet. 61, 705–718 (1997)
8. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W.: Genetic Structure of Human Populations. Science 298, 2381–2384 (2002)
9. Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M., Jones, K.W.: The Genomic Distribution of Population Substructure in Four Populations Using 8,525 Autosomal SNPs. Hum. Genomics 1, 274–276 (2004)
10. Pritchard, J.K., Stephens, M., Donelly, P.: Inference of Population Structure Using Multilocus Genotype Data. Am. J. Hum. Genet. 67, 945–959 (2000)
11. Purcell, S., Sham, P.: Properties of Structured Association Approaches to Detecting Population Stratification. Hum. Hered. 58, 93–107 (2004)
12. Intarapanich, A., Shaw, P.J., Assawamakin, A., Wangkumhang, P., Ngamphiw, C., Chaichoompu, K., Piriyapongsa, J., Tongsima, S.: Iterative Pruning PCA Improves Resolution of Highly Structured Populations. BMC Bioinf. 10(382) (2009)
13. Gao, X., Starmer, J.D.: AWclust: Point-and-Click Software for Non-parametric Population Structure Analysis. BMC Bioinf. 9(77) (2008)
14. Xing, J., Watkins, W.S., Witherspoon, D.J., Zhang, Y., Guthery, S.L., Thara, R., Mowry, B.J., Bulayeva, K., Weiss, R.B., Jorde, L.B.: Fine-Scaled Human Genetic Structure Revealed by SNP Microarrays. Genome Res. 19, 815–825 (2009)
15. Liang, L., Zollner, S., Abecasis, G.R.: GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics (Oxford, England) 23(12), 1565–1567 (2007)
16. Ewens, W.J.: Mathematical Population Genetics. Springer, Berlin (1979)
17. Bezdec, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
18. Parsons, L., Haque, E., Liu, H.: Subspace Clustering for High Dimensional Data: a Review. ACM SIGKDD Explor. Newslett. 6(1), 15 (2004)
19. Patterson, N., Price, A.L., Reich, D.: Population Structure and Eigenanalysis. PLoS genet. 2(12), e190 (2006)
20. Gibbs, R.A., Tassell, C.V., Weinstock, G., Green, R., Hamernik, D., Kappes, S., Liu, G., Matukumalli, L., Matukumali, A., Sonstegard, T., Silva, M.: Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 24, 528–532 (2009)