Correcting For Unequal Sampling in Principal Component Analysis of Genetic Data

W. Burgos-Paz[1,2], **S. E. Ramos-Onsins**[1], **M. Pérez-Enciso**[1,2,3], **L. Ferretti**[4]

[1] Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain, [2] Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193, Bellaterra, Spain, [3] Institut Català de Recerca i Estudis Avançats (ICREA), Carrer de Lluís Companys 23, Barcelona, 08010, Spain, [4] UMR 7138, UPMC and CIRB, College de France, Paris, France

ABSTRACT: Principal component analysis (PCA) is one of the most widely used tools to explore variability of high dimensional data. PCA is used for population and quantitative genetics. Its popularity has recently increased due to the huge amount of molecular markers available in datasets worldwide. In genetics, a common issue due to external constraints is uneven sampling of populations, limiting the usefulness of PCA because of well-known sample size sensitivity and two-dimensional projection bias. Here we evaluated the use of weighted PCA (wPCA) in genetic data in order to correct uneven sampling bias. Simulations suggest that wPCA improves the two-dimensional projections of PCA data and, in some cases, recovers population relationships patterns, even when sample size is as low as n=1. We used this correction in pig data from populations with uneven sampling, recovering a more realistic structure than inferred with only PCA.

Keywords: SNP; population structure; phylogeography

Introduction

The Principal Component Analysis (PCA) is a widely used tool for visualization of the structure of a highly multidimensional data set. It has applications in many fields, including genetics (Novembre and Stephens (2008)). The idea behind this technique is to reduce the complexity of the data by retaining only the linear combinations of variables explaining the maximum variance in the data set. Individual data is then projected onto the space of these combinations and visualized as a low-dimensional plot. In population genetics, PCA is frequently applied to visualize the genetic structure of populations, and its popularity has increased in the last years because of the high throughput genotyping technologies (e.g., SNP arrays), where it is necessary to extract the underlying structure of the dataset (dimensionality reduction) in a computationally efficient manner (Paschou et al., (2007)).

Despite several advantages of PCA in the population genetics field, this technique is very sensitive to the choice of the dataset and the distortion of the PCA plots due to a biased or unequal sampling is a known problem (McVean (2009)). In genetic studies, it is often not possible to sample individuals according to some criteria chosen a priori. For instance, the choice of sampling could be biased by incomplete knowledge of the population structure, conservation or there could also be ethical issues. Further, many factors, like budget constraints or accessibility of geographical regions, availability of samples and technical problems in their conservation and sequencing limit sampling in the field.

McVean (2009) suggested to correct for unequal sampling by downsampling the different populations, but this may lead to a loss in power or in a less accurate picture of genetic structure. A natural framework for this is to correct for this issue by means of a weighted PCA (wPCA), where the variables measured for each sample can be assigned a different weight w_i (Kriegel et al., (2008)). Kriegel et al. proposed this correction for synthetic data to evaluate the influence of outliers on PCA but, to our knowledge, its properties in genetic data are unknown. Here, we evaluated the usefulness of this correction with data from SNPs polymorphism under different population relationships, and illustrate the method in pig populations where differences in sample size could under-over estimate population relationships.

Materials and Methods

Weighted PCA. Given a set of genetic data of n sequences/genotypes and s biallelic SNPs for each sequence, the first step of the PCA is to compute the covariance:

$$\boldsymbol{Cov}\left(\boldsymbol{x_s}, \boldsymbol{x_{s'}}\right) = \boldsymbol{E}\left(\boldsymbol{x_s}, \boldsymbol{x_{s'}}\right) - \boldsymbol{E}\left(\boldsymbol{x_s}\right)\boldsymbol{E}\left(\boldsymbol{x_{s'}}\right)$$

$$= \frac{1}{n}\sum_{ij} x_s^i x_{s'}^j - \left(\frac{1}{n}\sum_i x_s^i\right)\left(\frac{1}{n}\sum_j x_{s'}^j\right)$$

A common assumption is that all individuals should be equally important in the PCA (that is, the implicit weight of each individual is $1/n$). This is usually correct, since there is often no prior information about the origin and grouping of individuals. However, this is not true for samples taken from different breeds or structured natural populations living in different habitats or geographical locations, since the degree of relationship will differ between and within breeds.

In the wPCA paradigm, each sample's genotype can be assigned a different weight w_i; then, the covariance can be rewritten naturally as:

$$\boldsymbol{Cov}\left(\boldsymbol{x_s}, \boldsymbol{x_{s'}}\right) = \sum_i \boldsymbol{w_i} x_s^i x_{s'}^i - \left(\sum_i \boldsymbol{w_i} x_s^i\right)\left(\sum_j \boldsymbol{w_j} x_{s'}^j\right),$$

where the weights are numbers between $0 < \boldsymbol{w_i} < 1$ with $\sum_i \boldsymbol{w_i} = \boldsymbol{1}$. This reduces to the usual PCA when $w_i = 1/n$. We denote the number of populations by $npop$ and the number of sequences in population A by nA. Since the weight of population A is the sum of the weights of all the sequences belonging to A, a natural weight for each sequence of A is:

$$w_i = \frac{1}{n_{pop}} \cdot \frac{1}{n_A}, \qquad i \in A$$

Simulations. To evaluate the usefulness of wPCA in genetic data, we simulated six populations considering two scenarios where population relationships are mainly shaped by the individuals exchange (e.g., migrations or introgression). In the first scenario (M1), we considered that all populations are equally related in terms of exchange of individuals. This scenario may occur in cooperative breeding schemes where sires or dams are regularly interchanged between herds. In the second scenario (M2), a subpopulation structure was considered where some populations have high individual exchange rates among them but low with the rest. This will typically occur, e.g., when one nucleus provides genetic material to other herds that are otherwise not connected.

For each population, five hundred individuals with thousand independent SNPs each were simulated by coalescence using MLCOASIM.v2 (available at http://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html). Effective population size (Ne=1000) was similar for all populations. Further, each population was sampled retaining one hundred individuals each for downstream analyses. To simulate unbalanced sampling three and one populations for M1 and M2 respectively, were subject to reduction of sample size ranged from n=1 to n=15. Sampling and multivariate analyses were performed in R (R Development Core Team (2011)).

Pig data. Genotype data from Cuban pig populations (Burgos-Paz et al., (2013)), is an interesting case to validate this methodology. Briefly, these samples represent well-structured populations with unequal sample sizes 1, 5 and 12 for Central Cuba (CUCE), Eastern Cuba (CUEA) and Western Cuba (CUWE), respectively. Previous analyses in Burgos-Paz et al. revealed different admixture proportions with Iberian, Asian and commercial breeds in these Cuban populations, and we suspected that unequal sample size, especially of CUCE population, affected the two-dimensional PCA projection.

We evaluated the credibility of PCA projections with real data using coalescence simulations. To do this, we simulated three populations with 20 individuals each. For the demography model, we considered an initial bottleneck due to arrival of Iberian pigs to Cuba in the 16th century (Crossby (2003)), and a recent introgression (20th century) of Commercial breeds, which in turn derived from a strong introgression of Asian populations in 18th century to Europe (Giuffra et al., (2000)). A generation interval of 3 years was assumed. We studied two datasets, one with equal sampling size (n=20) and an unbalanced dataset of n = 1, 5, and 12, as in the real sampling scheme.

Results and Discussion

Simulated data. Each model leaves a distinct characteristic PCA projection if sample sizes are balanced across populations. A large and balanced simulated data PCA (Figure 1, left-column) represents the expected 'true' pattern of PCA projection. Middle and right columns show the PCA projection with an unbalanced finite dataset using the usual (middle) and weighted (right) approach. An unweighted PCA result in highly distorted PC projections, in agreement with McVean, (2009). The PCA distortion was especially noticeable in M1, where individuals from reduced populations were located in the middle of the populations with larger sample size. In M2, distortion may cause misinterpretation of subpopulation structure (i.e blue dot individuals).

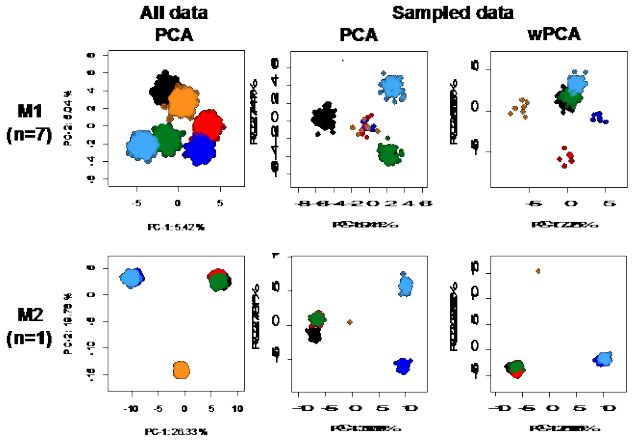

Figure 1. PCA and wPCA projections of simulated models. Left column shows PCA of all simulated data, the middle and the right columns shows PCA and wPCA of sampled data, respectively.

Considering the sample size of each population as weighted vector, application of wPCA recovered the expected population structure observed using balanced data (Figure 1, right-column). The proposed correction greatly improves the projections even if lower sample size in the population is n=1, as in M2. For M1, the minimum sample size required to improve the projections was higher (n=7) because of the low differentiation among populations. In this model, wPCA tends to overcorrect the projection with too small sample sizes. In hierarchical population structures (M2), wPCA recover the original structure with very low sample sizes and over correction is minimal.

wPCA in Cuban pig data. The previous results suggest that wPCA correction improves the PCA projection for structured populations, even in sample size n=1. Considering this, we used wPCA correction to explore the graphical representation of Cuban pig populations (Burgos-Paz et al., (2013)). Because samples were collected from three areas and no phenotypic criterion available, we used the size according to geographic origin as weight vector. Unweighted PCA projections showed the expected pattern where the population with fewest samples (CUCE) is shrunk towards the largest populations (Figure 2, left).

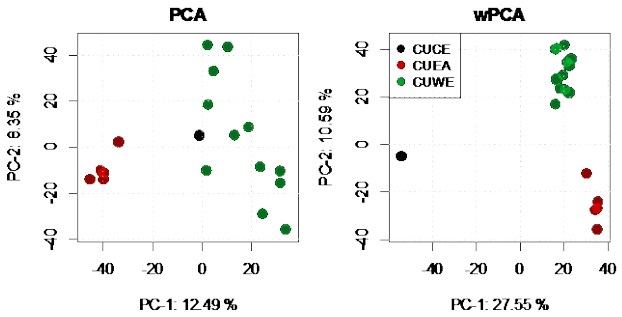

Figure 2. PCA and wPCA projections for observed data of Cuban pig populations.

Using wPCA (Figure 2, right), population with sample size n=1 (CUCE) was separated from the population with largest sample size (CUWE) and projection agrees with differentiation estimated either from F_{ST} or admixture analyses (Burgos-Paz et al., (2013)). To test whether the wPCA represents the most realistic true structure, we performed a simulation study with subsequent sample size reduction up to one individual in CUCE. The PCA projection of simulated data showed a triangle-like population arrangement mainly caused by differentiation of populations (Figure 3).

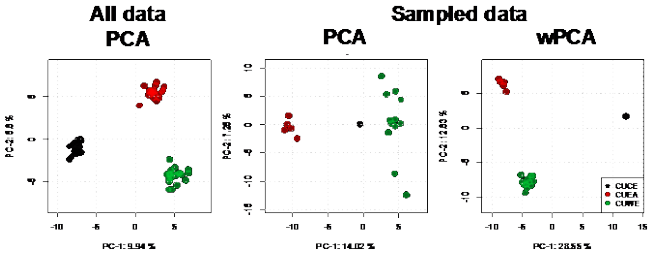

Figure 3. PCA and wPCA projection for the simulated model for Cuban pig populations

With unequal sample size and uncorrected PCA, we obtained a projection for PCA very similar to the real data (Figure 2, left), where the population with the lowest sample size (CUCE) clusters with the largest sampled population (CUWE). Additionally, similar values of the variance explained by the two first PC's were found. When wPCA was used, we recovered instead the expected triangle-like relationships as in PCA of complete data. The simple correction proposed here leads to a realistic interpretation of a likely demography model.

Conclusions

Application of weighted covariance in the estimation of PCA (wPCA) is a simple yet effective strategy to correct the effect of uneven sample size in genetic data, showing that population structure could be recovered in populations with low and unequal sample sizes. The results suggested also that, even in the presence of unbalanced migration and admixture, the application of weights in PCA could capture additional information of dataset, improving the two-dimensional projections of data.

Acknowledgments

WBP is funded by COLCIENCIAS (*Francisco José de Caldas* fellowship 497/2009, Colombia). Work funded by AGL2010-14822 grant (Spain) to MPE, CGL2009-09346 grant (Spain) to SERO.

Literature Cited

Burgos-Paz, W., Souza, C. A., Megens, H. J. et al. (2013). Heredity, 110: 321–330

Crossby, A. (2003). The Columbian Exchange, Praeger Publishers., Connecticut.

Giuffra, E., Kijas, J. M., Amarger, V. et al. (2000). Genetics, 154: 1785-1791

Kriegel, H., Kröger, P., Schubert, E. et al. (2008). In Scientific and Statistical Database Management. Springer., pp. 418–435.

McVean, G. (2009). PLoS Genetics, 5: e1000686.

Novembre, J., and Stephens, M. (2008). Nat. Genet., 40: 646–649.

Paschou, P., Ziv, E., Burchard, E. et al. (2007). PLoS Genet., 3: e160.

R Development Core Team. (2011). http://www.R-project.org/