OPEN ACCESS Freely available online

PLoS one

Genetic Classification of Populations Using Supervised Learning

Michael Bridges[1], Elizabeth A. Heron[2], Colm O'Dushlaine[2], Ricardo Segurado[2], The International Schizophrenia Consortium (ISC)[¶], Derek Morris[2], Aiden Corvin[2], Michael Gill[2], Carlos Pinto[2]*

1 Astrophysics Group, Cavendish Laboratory, Cambridge, United Kingdom, **2** Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College, Dublin, Ireland

Abstract

There are many instances in genetics in which we wish to determine whether two candidate populations are distinguishable on the basis of their genetic structure. Examples include populations which are geographically separated, case–control studies and quality control (when participants in a study have been genotyped at different laboratories). This latter application is of particular importance in the era of large scale genome wide association studies, when collections of individuals genotyped at different locations are being merged to provide increased power. The traditional method for detecting structure within a population is some form of exploratory technique such as principal components analysis. Such methods, which do not utilise our prior knowledge of the membership of the candidate populations. are termed *unsupervised*. Supervised methods, on the other hand are able to utilise this prior knowledge when it is available. In this paper we demonstrate that in such cases modern supervised approaches are a more appropriate tool for detecting genetic differences between populations. We apply two such methods, (neural networks and support vector machines) to the classification of three populations (two from Scotland and one from Bulgaria). The sensitivity exhibited by both these methods is considerably higher than that attained by principal components analysis and in fact comfortably exceeds a recently conjectured theoretical limit on the sensitivity of unsupervised methods. In particular, our methods can distinguish between the two Scottish populations, where principal components analysis cannot. We suggest, on the basis of our results that a supervised learning approach should be the method of choice when classifying individuals into pre-defined populations, particularly in quality control for large scale genome wide association studies.

Citation: Bridges M, Heron EA, O'Dushlaine C, Segurado R, The International Schizophrenia Consortium (ISC), et al. (2011) Genetic Classification of Populations Using Supervised Learning. PLoS ONE 6(5): e14802. doi:10.1371/journal.pone.0014802

Editor: Daniel J. Kliebenstein, University of California, United States of America

Received April 13, 2010; **Accepted** December 1, 2010; **Published** May 12, 2011

Funding: This project has not been directly funded by any agency. The authors employed on research contracts are supported by the Wellcome Trust (http://www.wellcome.ac.uk), Science Foundation Ireland (http://www.sfi.ie), and the UK Science and Technology Research Council (http://www.stfc.ac.uk). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: capinto@tcd.ie

¶ Membership of the International Schizophrenia Consortium is provided in the Acknowledgments.

Introduction

The advent of the new large-scale genotyping and sequencing technologies has resulted in unprecedented quantities of data becoming available to the genetics community. Geneticists are now confronted with new and challenging problems in data analysis and interpretation, and novel approaches and techniques will be required to fully exploit these new resources. In view of the fact that other scientific fields have already gone through a similar process of development, it is likely that cross-disciplinary collaborations in data analysis will yield fruitful results in genetics. This paper represents such a collaboration.

We apply machine learning techniques previously used in cosmology to the problem of genetic classification. Such techniques involve the use of automated algorithms to mimic the learning capabilities of animal brains. They have proved extremely useful in the analysis of complex data in many scientific disciplines. There are two basic approaches – *supervised* learning, where the data is pre-classified according to some hypothesis and *unsupervised* learning where the data is unclassified (usually, but not always, because the potential classes are *a priori* unknown). Genetics has, to date, relied mainly on unsupervised methods, such as principal components analysis (PCA), to classify individuals on the basis of their genetic data.

PCA is a standard tool in population genetics, and has been used, for example in a study of 23 European populations [1] and more recently of 25 Indian populations [2]. It is also commonly used in quality control in genetic studies. For example, a dataset destined for a disease association study may be pre-screened using PCA in order to detect and remove population structure so as to minimise noise in the final study. In many of the large scale collaborations now being undertaken it is of interest to determine whether genetic differences exist between groups of controls ascertained from different geographic locations, or genotyped at different laboratories. If the differences are sufficiently small, these groups can be merged to achieve greater power. The aim of this work is to demonstrate and quanmtify the superiority of supervised learning techniques when applied to this problem.

We have adapted two supervised learning algorithms, artificial neural networks (ANN) and support vector machines (SVM) for

this purpose. We use sets of control samples genotyped by the International Schizophrenia Consortium (ISC) [3] as our test data. For comparison we also conduct a conventional PCA analysis.

The paper is organised as follows. In the Methods section we briefly discuss the PCA methodology that we use and give a short introduction to ANNs and SVMs. We also include a description of the data used for the analysis. The first part of the Results section presents the PCA analysis and results. The second and third sections describe the ANN and SVM analyses respectively. Finally, the Discussion section contains our interpretation of the analyses and some suggestions for potential applications of the methods.

Methods

We examine three approaches to the problem of genetic classification, given pre–existing candidate populations. More precisely, we wish to determine the confidence with which the individuals in these populations can be distinguished on the basis of their genetic structure. We first consider PCA, the most commonly used unsupervised method. Next, we investigate a sophisticated non–linear supervised classifier, a probabilistic ANN. Lastly we consider a simpler but more limited linear supervised classifier, an SVM.

We would expect the supervised methods to perform better than PCA, since they utilise more information. The aim is to quantify this difference. We therefore adopt a sliding window approach, using genetic windows of different sizes in order to to assess the perfomance of the classifiers given different amounts of genetic data.

According to a recent hypothesis, discussed below, unsupervised methods cannot distinguish between two populations if the amount of data available falls below a certain threshold value. It is therefore of interest to determine whether supervised methods can classify below this limit, and we investigate this question also.

Principal Components Analysis

The PCA technique is well known and commonly used in genetics and we do not describe it in detail here. Briefly, the aim is to determine the direction of maximum variance in the space of data points. The first principal component points in the direction of maximum variance, the second component maximises the remaining variance and so on. Any systematic difference between groups of individuals will manifest itself as a differential clustering when the data points are projected on to these principal components.

We use the smartpca component of the eigensoft (v3.0) software package [4] for our analysis. In addition to the principal components, smartpca produces a biased but asymptotically consistent estimate of Wright's F_{ST} parameter [5]. We use this estimator as our measure of effect size.

The authors of SMARTPCA use a result obtained by [6] and [7], to conjecture the existence of a phase transition (the Baik, Ben Arous, Péché or BBP transition) below which population structure will be undetectable by PCA [4]. They further conjecture that this threshold represents an absolute limit for *any* (presumably unsupervised) classification method. For two populations of equal size, the critical F_{ST} threshold is given by:

$$F_{ST}(crit) = \frac{1}{\sqrt{N_{SNP}S}}$$

where N_{SNP} is the number of single nucleotide polymorphisms (SNPS) and S is the total number of individuals in the dataset.

A measure of statistical significance between any pair of populations is also produced by SMARTPCA. This is obtained by computing the ANOVA F-statistics for the difference in mean values along each principal component. A global statistic is calculated by summing over all components; this statistic follows a χ^2 distribution. We use the associated p-value as our measure of statistical significance.

It is important to point out that we are using the p-value as a quantitative measure. This quantity is more usually used in a hypothesis testing framework, where the decision to accept or reject is made on the basis of some pre-determined threshold. We do not set such a threshold; rather, we use the p-value to detect the onset of the BBP phase transition, when its value drops by many orders of magnitude.

We determine the effectiveness or otherwise of PCA by comparing the estimated value of F_{ST} with the critical value in a sliding window across the chromosome.

Artificial Neural Networks

ANNs are relatively uncommon in genetics and may be unfamiliar to many geneticists. Furthermore the network we employ possesses some novel features particularly relevant to genetic analysis. We therefore give a somewhat more detailed overview in this section.

ANNs are a methodology for computing, based on massive parallelism and redundancy, features also found in animal brains. They consist of a number of interconnected processors each of which processes information and passes it to other processors in the network. Well-designed networks are able to 'learn' from a set of training data and to make predictions when presented with new, possibly incomplete, data. For an introduction to the science of neural networks the reader is directed to [8].

The basic building block of an ANN is the *neuron*. Information is passed as inputs to the neuron, which processes them and produces an output. The output is typically a simple mathematical function of the inputs. The power of the ANN comes from assembling many neurons into a network. The network is able to model very complex behaviour from input to output. We use a three-layer network consisting of a layer of input neurons, a layer of "hidden" neurons and a layer of output neurons. In such an arrangement each neuron is referred to as a node. Figure 1 shows a schematic design for this network with 7 input nodes, 3 hidden nodes and 5 output nodes.

The outputs of the hidden layer and the output layer are related to their inputs as follows:

$$\text{hidden layer}: \quad h_j = g^{(1)}\left(f_j^{(1)}\right); \; f_j^{(1)} = \sum_l w_{jl}^{(1)} x_l + b_j^{(1)}, \quad (1)$$

$$\text{output layer}: \quad y_i = g^{(2)}\left(f_i^{(2)}\right); \; f_i^{(2)} = \sum_j w_{ij}^{(2)} h_j + b_i^{(2)}, \quad (2)$$

where the output of the hidden layer h and output layer y are given for each hidden node j and each output node i. The index l runs over all input nodes. The functions $g^{(1)}$ and $g^{(2)}$ are called activation functions. The non-linear nature of $g^{(1)}$ is a key ingredient in constructing a viable and practically useful network. This non-linear function must be bounded, smooth and monotonic; we use $g^{(1)} = \tanh x$. For $g^{(2)}$ we simply use $g^{(2)}(x) = x$. The

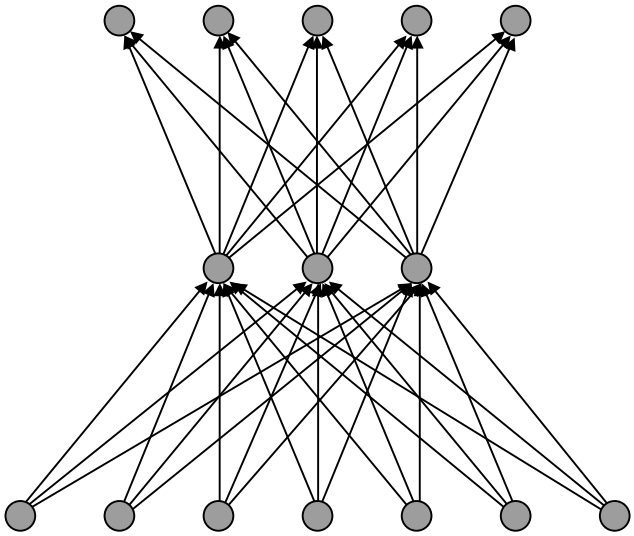

Figure 1. An example of a 3-layer neural network with 7 input nodes, 3 nodes in the hidden layer and 5 output nodes. Each line represents one weight.
doi:10.1371/journal.pone.0014802.g001

layout and number of nodes are collectively termed the *architecture* of the network.

The weights **w** and biases **b** effectively define the network and are the quantities we wish to determine by some *training* algorithm. We denote **w** and **b** collectively by **a**. As these parameters vary during training, a very wide range of non-linear mappings between inputs and outputs is possible. In fact, according to a 'universal approximation theorem' [9], a standard three-layer feed-forward network can approximate any continuous function to *any* degree of accuracy with appropriately chosen activation functions. However a network with a more complex architecture could well train more efficiently.

The use of ANNs in genetics to date has been limited. A comprehensive review is given in [10]. Previous work has focused mainly on investigating the optimum network architecture for specific applications, using a small number of genetic markers. A case-control scenario was considered in [11]. Their networks typically consisted of four input nodes, representing four markers, with two hidden layers incorporating up to three hidden nodes each. The output was the case or control status of the individual. The authors explored a variety of different architectures and assessed the performance of each. In common with other authors such as [12], they noted that the performance of the network was strongly dependent on the choice of architecture. Nevertheless, many authors such as [13] and [14] have successfully used ANNs with pragmatic choice of architecture based on trial and error searching.

A more serious problem is the size of networks that it is possible to train when using traditional back-propagation or quasi-newtonian gradient descent methods. Most such methods are very inefficient in navigating the weight space of a network and can therefore handle only relatively small genetic datasets.

Both these problems are addressed in the MemSys package [15] which we use to perform the network training. This package uses a non–deterministic algorithm which allows us to make *statistical* decisions on the appropriate classification. This makes possible the fast efficient training of relatively large network structures on large data sets. Moreover the MemSys package computes a statistic termed the Bayesian evidence (see for example [16] for a review). The evidence provides a mechanism for selecting the optimum number of nodes in the hidden layer of our three–layer network.

We apply this ANN to our genetic classification problem by associating each input node with the value of a genetic marker from an individual and the output nodes with the probabilities of the individual's membership of each class. As in the case of the PCA analysis we perform the classification in a sliding window across the chromosome.

Support Vector Machines

The ANN described in the previous section is a sophisticated classifier, able to amplify weak signals and to detect non–linear relationships in the data. This feature is potentially of great significance in genetic analysis, since non–linearity is likely to arise due to long-range interactions between genes at different physical locations. It is also of interest to investigate the performance of a more conventional linear supervised classifier on the genetic classification problem. We therefore conduct a parallel analysis with an SVM.

The principle of an SVM is intuitively very simple. The space of data points is partitioned by finding a hyperplane that places as many of the points as possible into their pre-defined class. The SVM algorithm iterates through trial planes, computing the shortest combined distance from the plane to the closest of the data points in each class while simultaneously ensuring all data points of each class remain in the same partition. An example of a two-dimensional feature space partitioned in three different ways is shown in Figure 2.

In the example pictured the plane p3 does not partition the space correctly. The plane p2 produces an adequate classification with all of the data points appropriately divided. However two data points lie very close to the plane and leave little margin for future generalisation to unseen examples. The plane p1 is an

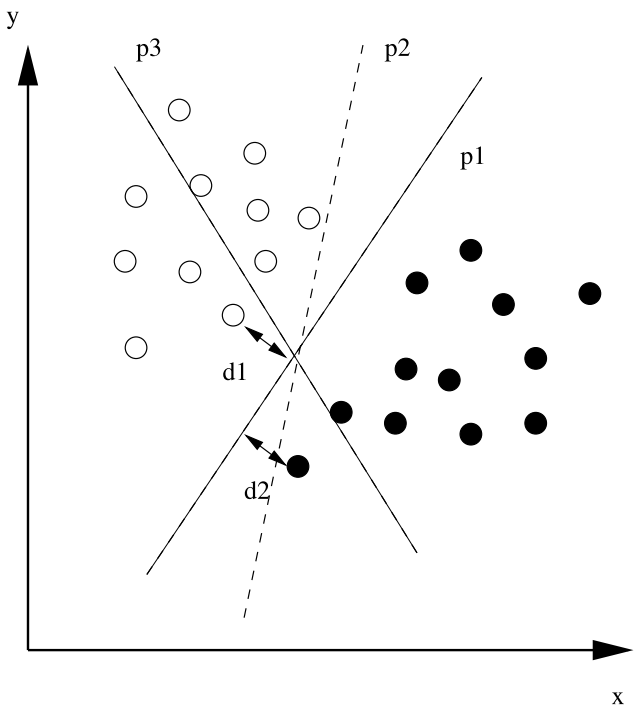

Figure 2. An example of a two-dimensional feature space $x-y$ for data of known class divided by three hyperplanes p1, p2 and p3. Clearly p1 divides most efficiently.
doi:10.1371/journal.pone.0014802.g002

optimum partitioning, maximising the combined distance d1 + d2. The function of an SVM is to attempt to identify this optimum partition. In this work we make use of the LIBSVM library of SVM routines [17].

The SVM has the advantage of being simpler to use in practice, but has certain limitations compared with our ANN. Firstly it is a linear classifier and cannot allow for non–linear relationships in the data. Secondly it is deterministic, providing a unique solution for each problem. It is therefore impossible to develop an estimate of the accuracy of the solution–that is, to place confidence limits on the classification. Our ANN, on the other hand, is probabilistic, producing a slightly different solution on each iteration. This allows us to assess the stability of the solution. Thirdly, the classification is binary–an individual either does, or does not, belong to a particular class. The ANN, in contrast, provides probabilities of class membership for each class.

Data

Our test populations are a subset of the data obtained by the International Schizophrenia Consortium (ISC). The consortium collected genome-wide case–control data from seven sample collection sites across Europe. The final post quality controlled (QC) dataset contained 3322 cases and 3587 controls. The controls from three sites were used for the purposes of this study:

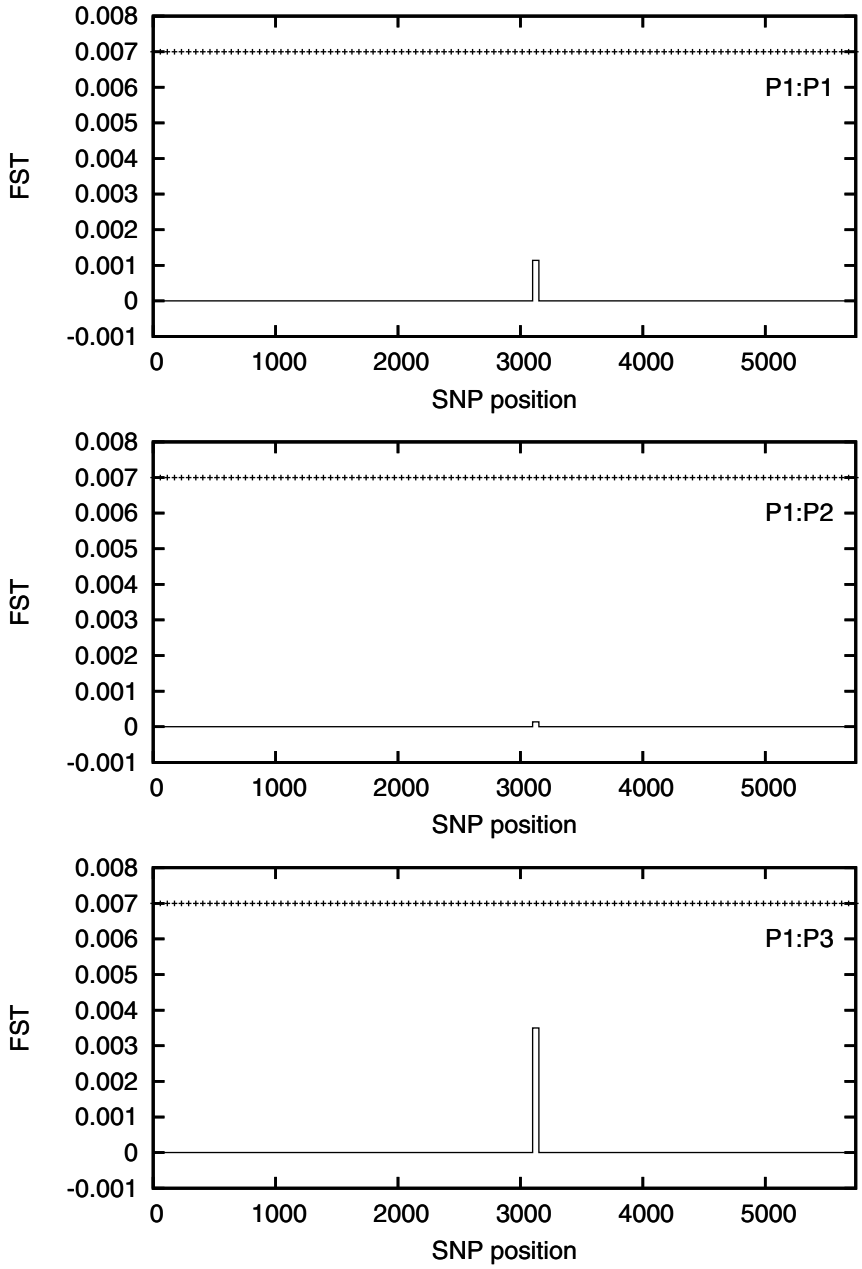

Figure 3. Estimated F_{ST} values for a 50 SNP sliding window for P1:P1 (top), P1:P2 (middle), P1:P3 (bottom). The F_{ST} is essentially zero everywhere except for a small region approximately halfway along the chromosome. The horizontal dotted line is the value of $F_{ST}(crit)$.
doi:10.1371/journal.pone.0014802.g003

- **Aberdeen Site (P1)** A set of 702 controls, consisting of volunteers recruited from general practices in Scotland. These were genotyped on an Affymetrix 5.0 genotyping array.
- **Edinburgh Site (P2)** A set of 287 controls recruited through the South of Scotland Blood Transfusion Service, typed on an Affymetrix 6.0 array.
- **Cardiff Site (P3)** A set of 611 controls recruited from several sources in the two largest cities in Bulgaria, typed on an Affymetrix 6.0 array.

Quality control was performed by the ISC [18]. In addition to the usual genotype and sample QC procedures, attempts were made to resolve technical differences arising from the different genotyping arrays used by the various ISC sites. A multi-dimensional scaling analysis was also performed to detect population stratification and remove outliers from each population.

We start with the cleaned ISC data comprising 739,995 SNPs, all samples having a call rate >0.95 and all SNPs having minor allele frequencies >0.01, with population outlier identifiers removed [18]. For the purposes of this study we examine a linkage-disequilibrium (LD) pruned set of 5739 SNPs ($r^2<0.2$) on chromosome 1, selecting only those that were common to both the Affy 5.0 and Affy 6.0 platforms. PLINK v1.06 [19] software was used for this data reduction. The parameters of the three test populations are given in Table S1.

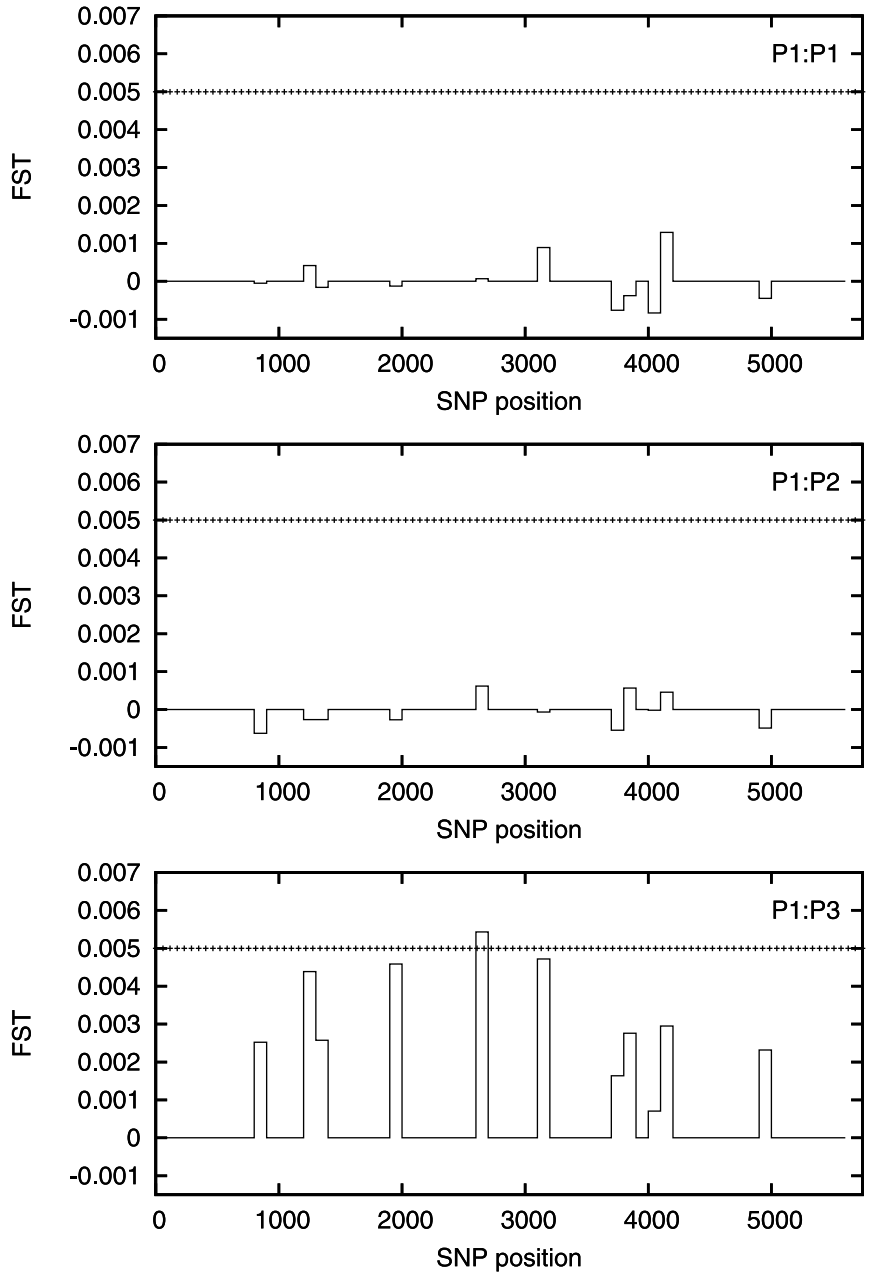

Figure 4. Estimated F_{ST} values for a 100 SNP sliding window for P1:P1 (top), P1:P2 (middle), P1:P3 (bottom). The horizontal dotted line is the value of $F_{ST}(crit)$. Note that although F_{ST} is always non-negative, the estimator may become negative for small values of F_{ST}.
doi:10.1371/journal.pone.0014802.g004

Results

We first perform a principal components analysis (PCA) on the three populations to determine whether the populations can be distinguished using an unsupervised learning approach. We then carry out both ANN and SVM supervised learning classifications on the same three populations.

PCA Classification

We first test for structure *within* each of our three populations. In each case the population is divided into two disjoint subsets. For P1 and P3 each subset consists of 200 samples. In the case of P2, only 287 samples are available in total, so we divide these into two subsets of 140 samples each. We do not remove any residual (post QC) outliers, in order to maximise any signal.

In all three cases we find that the estimated F_{ST} values are vanishingly small, less than 0.0001 even when all 5739 SNPs are used. In no case do the estimated levels of F_{ST} exceed $F_{ST}(crit)$. By comparison a recent study [20] found values ranging as high as 0.023 across Europe. The ANOVA p-values for the three populations P1, P2, and P3 are 0.050, 0.559 and 0.022 respectively. Although two of these p-values fall at or below the conventional threshold of 0.05 this does not in itself imply the ability to detect structure in the absence of a reasonable effect size. The PCA plot for the most significant case ($p = 0.022$) shows that the populations do not separate (Figure S1). We conclude that

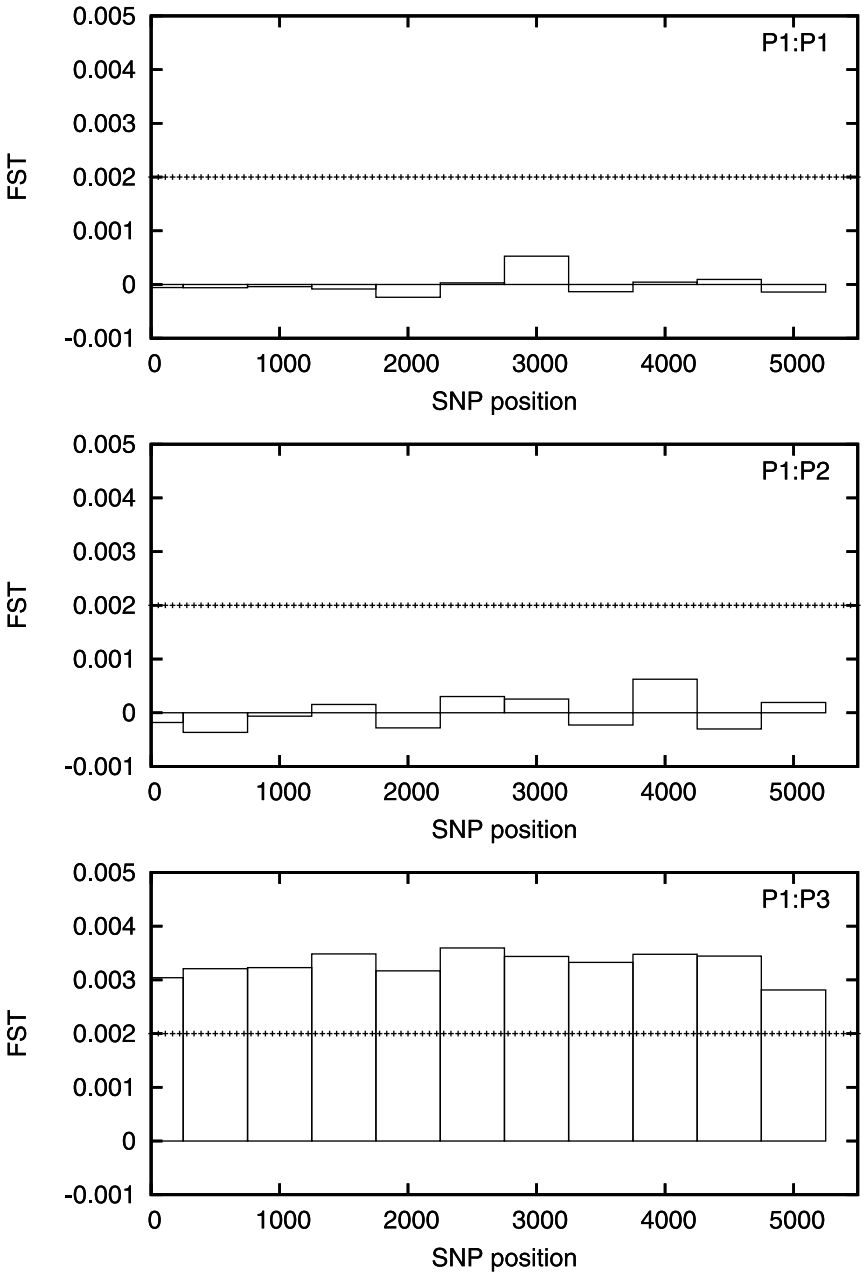

Figure 5. Estimated F_{ST} values for a 500 SNP sliding window for P1:P1 (top), P1:P2 (middle) and P1:P3 (bottom). The horizontal dotted line is the value of $F_{ST}(crit)$.
doi:10.1371/journal.pone.0014802.g005

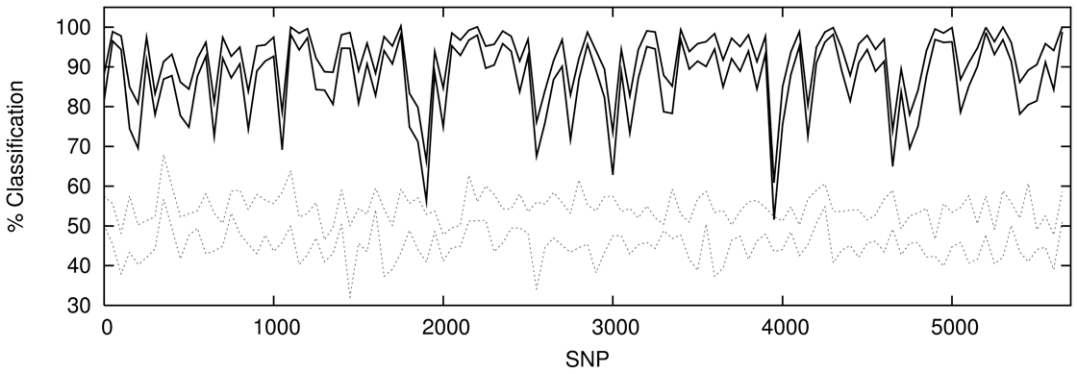

Figure 6. Classification with windows of 50 contiguous, non-overlapping SNPs for P1 against P2 (solid lines) with classification results for a sample of P1 against P1 (dotted lines) shown for comparison. The regions enclosed between the lines illustrate 1σ confidence intervals.
doi:10.1371/journal.pone.0014802.g006

PCA fails to detect structure between the subsets tested in each of our three populations; that is, each population is essentially homogeneous.

We next test for differences between our three populations. We perform a sliding window PCA analysis with non–overlapping windows of length 50, 100 and 500 SNPs. The estimated F_{ST} values are plotted in Figures 3, 4 and 5 with the corresponding critical value shown for comparison.

The estimated F_{ST} is negligible at the 50 SNP level, except for one window about halfway along the chromosome, and even here it does not approach $F_{ST}(crit)$. Some signals are visible for the P1**:**P3 comparison at the 100 SNP level, but $F_{ST}(crit)$ is exceeded in only one window. At the 500 SNP level the PCA analysis can distinguish between the P1 and P3 populations, with the estimated F_{ST} exceeding $F_{ST}(crit)$ everywhere along the chromosome but the P1**:**P2 comparison still shows negligible signal. The full results from this analysis are given in Table S2. Sample PCA plots showing the BBP transition given in Figures S2–S4 and S5–S7.

We may summarise the results of our PCA analysis as follows. As expected, no internal structure is detectable within any of the three populations. Moreover, PCA is unable to distinguish the two Scottish populations even when using the full input set of 5739 SNPs. The two Scottish populations can, however, be distinguished from the Bulgarian population, given an input data set of around 500 SNPs, anywhere along the chromosome.

ANN Classification

We next attempt to classify the same data using the ANN. The pre-classified data available is divided into a *training set* used to train the network and a *hold-out set* used to assess the accuracy of the network after training. Since we merely wish to determine whether the ANN is able to classify or not, it is desirable to to maximise the size of the training set while retaining a large enough testing set to ensure statistically meaningful results. In practice we find that a ratio of 80% : 20% to be satisfactory and all the results presented here use this ratio.

As with the PCA analysis we use samples of 200 from each population, except in the P2:P2 case, where we use 140 for each sub-population. We perform multiple repetitions of the network training, drawing a different random starting point (of the weights and biases) on each occasion. In this way we are able to obtain an ensemble of trained classifiers from which we can draw a standard 1σ error on the network classification. For all of the results below we use >20 repetitions. We present all of our results in terms of % accuracy of classification on the hold-out set, where 100% defines a perfect classifier and 50% is no better than random.

To explore the variation of classification across the chromosome we use an input set of non-overlapping windows each containing 50 SNPs. Figures 6–8 show the classification rate along the chromosome for each population combination. In addition each figure illustrates a reference null classification of two sub-samples

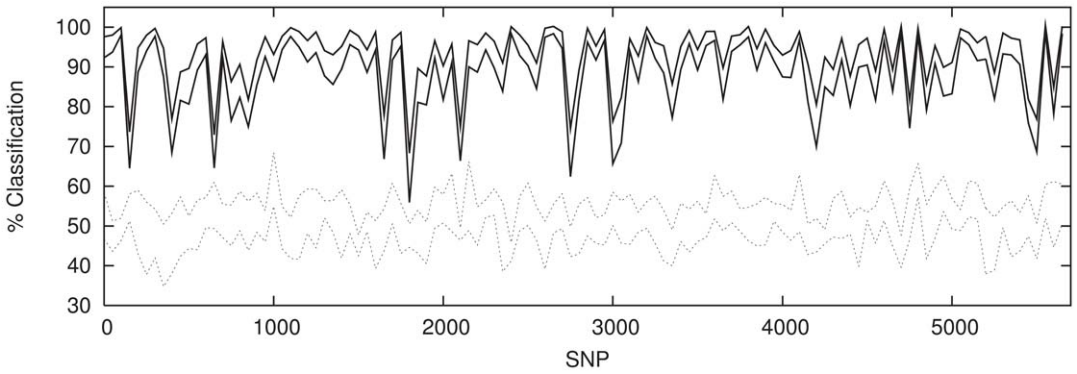

Figure 7. Top panel shows classification with windows of 50 contiguous, non-overlapping SNPs for P1 against P3 (solid lines) with classification results for a sample of P3 against P3 (dotted lines) shown for comparison. The regions enclosed between the lines illustrate 1σ confidence intervals.
doi:10.1371/journal.pone.0014802.g007

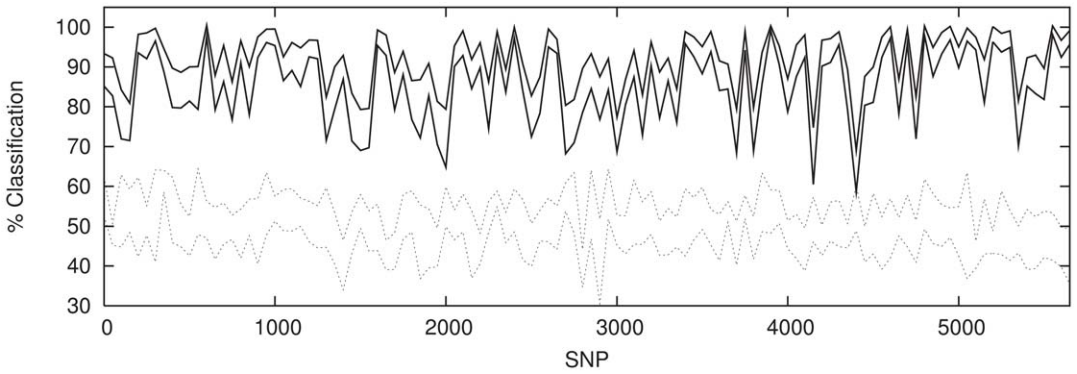

Figure 8. Top panel shows classification with windows of 50 contiguous, non-overlapping SNPs for P2 against P3 (solid lines) with classification results for a sample of P2 against P2 (dotted lines) shown for comparison. The regions enclosed between the lines illustrate 1σ confidence intervals.
doi:10.1371/journal.pone.0014802.g008

from each of the three populations to demonstrate the internal homogeneity of each population.

It is notable that a classification rate of $>80\%$ is achieved across the majority of the chromosome for *both* populations P1:P2 *and* P1:P3. This demonstrates that the network can successfully amplify a much weaker, intra-Scottish population signal to roughly the same level as that obtained for the Scotland-Bulgaria comparison.

We next investigate the variation in performance as the window size is varied. Figure 9 shows results for the classification of P1:P2 with window sizes of 20, 50 and 100 SNPs.

For a window size of 20, one sees considerable structure along the chromosome, with some regions classifying well, and others poorly. As the window size increases, with each window now containing both "good" and "bad" regions, we find that the classification rate converges to the best, rather than the worst rate. This shows that even when the network is presented with a large window that contains a small proportion of informative SNPs it can successfully filter out the extraneous inputs and produce a classifier with the same level of accuracy as would have been obtained with a reduced set of informative inputs. This feature has many important implications within genetics where data is often noisy or incomplete.

It is common in signal processing to represent the efficiency of a classifier graphically, using a receiver operating characteristic (ROC) curve which plots the true positive rate (TPR) versus the false positive rate (FPR) for increments of the classifier's discrimination threshold. The default threshold is normally 0.5, but variation of this criterion allows classifiers to be tuned to minimise the FPR while simultaneously maximising the TPR. An ideal classifier has a ROC curve that resembles a step-function with a TPR of 1.0 for all values of the threshold, while the ROC curve for a random classifier is a line with slope of unity from a TPR of 0 to 1. Figures 10 and 11 illustrate the ROC curves for the network classifier in two different regimes along the chromosome spectrum. Figure 10 shows the ROC curve of the classifier trained

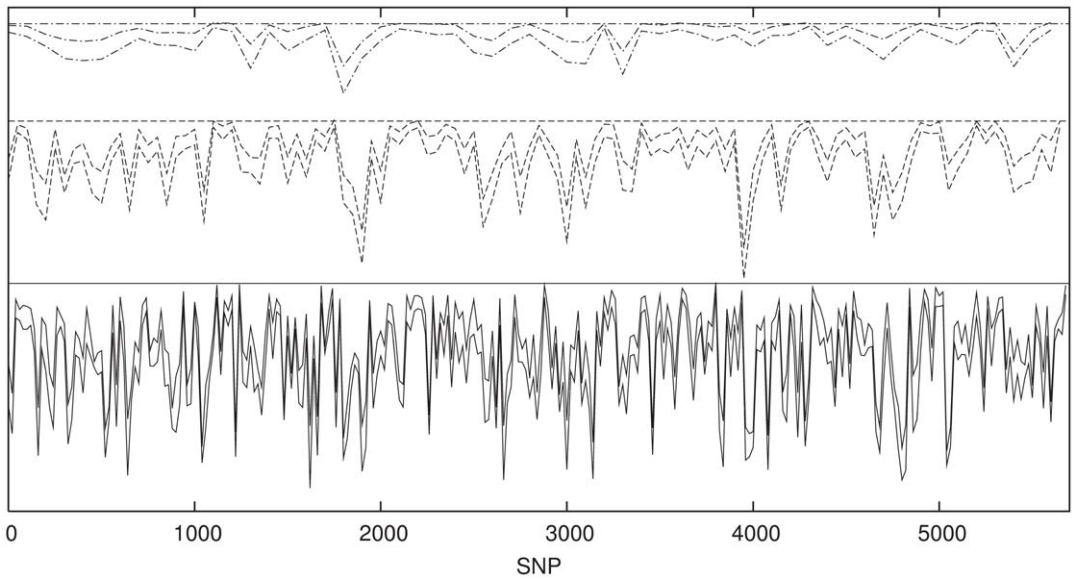

Figure 9. Classification with windows of 100 (dot-dashed), 50 (dashed) and 20 (solid) contiguous, non-overlapping SNPs for P1 against P2. Note that as the window size increases, the accuracy converges to the *most* accurate classification, indicating that the ANN is successfully discarding irrelevant information. For clarity we have added an offset to each spectrum and omitted the ordinate axis, the horizontal lines represent 100% classification in each case.
doi:10.1371/journal.pone.0014802.g009

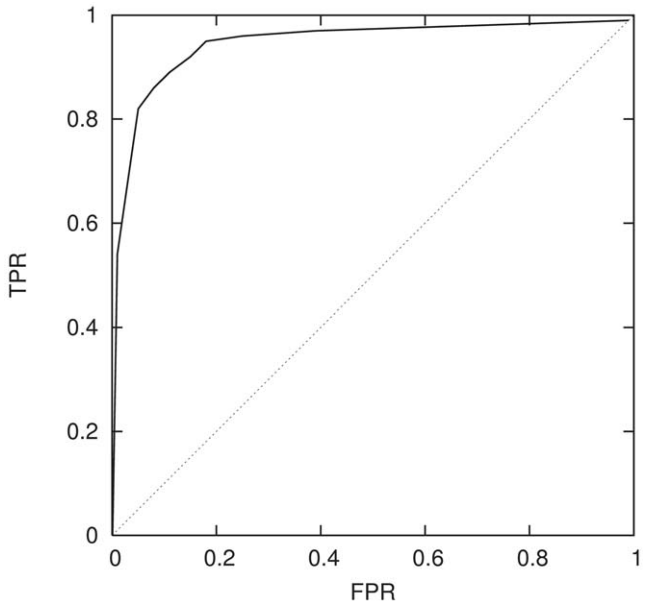

Figure 10. Receiver Operating Characteristic (ROC) curve, that is a plot of true positive rate (TPR) against false positive rate (FPR) of the neural network classifier trained using the first 50 SNPs using P1:P2 (solid curve). A random classifier (dotted curve) is shown for comparison.
doi:10.1371/journal.pone.0014802.g010

using the first 50 SNPs. As is evident from Figures 6–8 this region produces a classifier that is capable of distinguishing the two population groups at the 90% level. The quality of this classifier is then clearly discernible by a ROC curve that approaches a step-function. For comparison we performed the same test on a part of the chromosome spectrum where the classifier was relatively poor, at a SNP window of $1950-2000$. This ROC curve, shown in Figure 11 appears very close to the random classifier line, as would be expected. Along with multiple network realisations computed for each classifier these tests provide a useful way to confirm the stability of the classifiers.

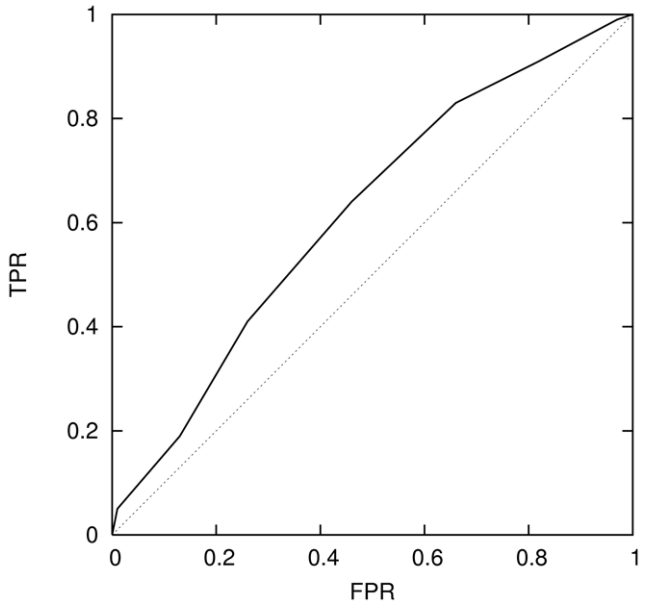

Figure 11. Receiver Operating Characteristic (ROC) curve of the neural network classifier trained using 50 SNPs form 1950 to 2000 also for P1:P2 (solid curve). A random classifier (dotted curve) is shown for comparison.
doi:10.1371/journal.pone.0014802.g011

The architecture of our three layer network is determined entirely by the number of nodes in the hidden layer. This number in turn can be estimated from the Bayesian evidence. We find that our results are insensitive to the number of hidden nodes. In fact, reducing the number of hidden nodes from 20 to zero results in negligible degradation in performance, indicating that the signal we detect is essentially linear. It is of course possible to identify such a linear signal using PCA for example, given a signal of sufficient strength, as was demonstrated in the earlier part of this paper. The reason for the increased sensitivity of our ANN here is its utilisation of our prior knowledge of class membership and its efficiency in exploring the space of all *possible* linear (and non-linear) mappings and identifying the choice that maximises the classifier's sensitivity automatically.

In summary, we find that the ANN exhibits considerably greater sensitivity than PCA. In particular, while PCA cannot distinguish between the two Scottish populations, the ANN can do so given fewer than 100 SNPs. Moreover, the ANN can classify on a dataset well below the BBP limit. Furthermore, as we have seen, the ANN can also efficiently eliminate noise. Our results indicate that the signal the ANN is identifying is linear, but nevertheless too weak for PCA to detect.

SVM Classification

In view of the fact that the dominating signal in the data is linear, we would expect the SVM to perform equivalently. We do not repeat the entire analysis here, but simply show the sliding window analysis for the population combination P1 and P2 in Figure 12 (with the equivalent ANN results for comparison in Figure 13). Since the SVM for a given dataset is entirely deterministic it is not possible to generate multiple realisations of the classifier and thus build up 1σ confidence intervals. However it is clear that SVM performs comparably with the ANN on this dataset, locating strikingly similar features in the classification spectrum across the chromosome. It is also of interest to compare the speed of each method. The SVM takes roughly 10 seconds to build a classifier on a 50 SNP window, using a currently standard desktop computer. A single iteration of the ANN takes a roughly equal amount of time, with 1σ limits being generated in a $n_{\text{iterations}}$ multiple of this time.

Discussion

We demonstrate in this paper that supervised learning classification is to be preferred to unsupervised learning in genetics, when we have an *a priori* definition of class membership from some non-genetic source. The classification then serves to determine whether or not the pre-defined populations are *genetically* distinguishable.

Both the techniques investigated in this paper (SVMs and ANNs) significantly outperform PCA on the data presented here. It is noteworthy that the sensitivity of these methods exceeds the conjectured BBP limit on the sensitivity of supervised approaches.

Although ANNs have been previously discussed in the context of genetics, they have yet to come into common use in this field. This is probably due, in part, to the limited number of input nodes that it was possible to handle, and in part to the difficulty of determining the optimal network architecture. Our ANN allows us to handle very large numbers of inputs, an essential feature in many applications in genetics. The problem of deciding on the

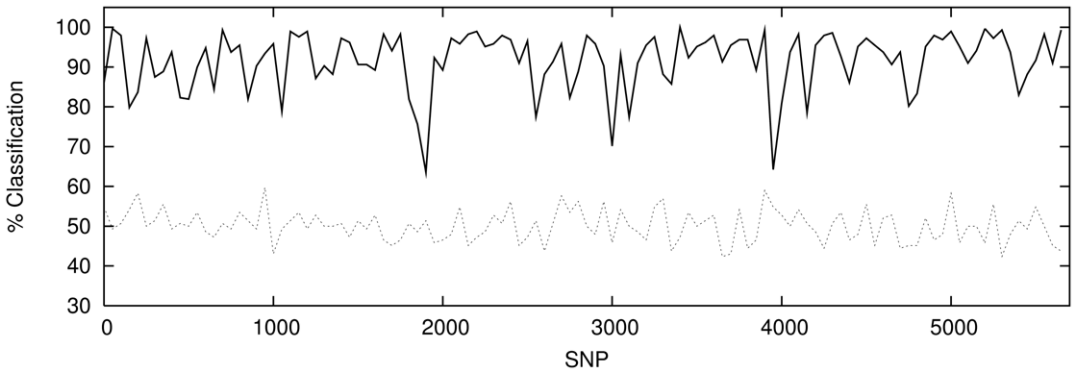

Figure 12. SVM classification with windows of 50 contiguous, non-overlapping SNPs for P1 against P2 (solid lines) with classification results for a sample of P1 against P1 (dotted lines) shown for comparison.
doi:10.1371/journal.pone.0014802.g012

optimal network architecture, much discussed by previous authors, reduces, in the case of a 3-layer network, to deciding on the number of hidden nodes; the MemSys package provides a rigorous method of determining this number.

In the event, we observe a predominantly *linear* signal on this dataset, easily detectable by both SVM and ANN but too weak to be detected by PCA. In a sense, this is be expected, since the SVM and ANN utilise our prior knowledge of class membership to find the optimal linear mapping for classifying the data. In the absence of such prior information, PCA finds the linear mapping that maximises the variance; this is not necessarily the optimal mapping. However the sensitivity of the supervised methods and the small number of SNPs that they need in order to classify efficiently is noteworthy. A further important consequence of this fact is that the SVM and ANN can *localise* the sources of genetic difference along the chromosome and indeed the results of both methods are consistent with each other in this respect.

The linearity of the signal means that the SVM and ANN perform comparably. (The main novelty here is the large number of inputs that our ANN can accept). This linearity is not altogether surprising, since non–linear effects would arise as a result of long–range correlation between loci. The relatively small size of our SNP windows greatly reduces the probability of seeing such correlations. (Short range correlations, which arise from linkage disequilibrium, carry no useful information and were eliminated by LD pruning our data).

When a linear signal is present, both the ANN and the SVM can classify with equal efficiency and we recommend that both be considered for use in genetic classification. The ANN, however, possesses three advantages over the SVM. Firstly the stochastic nature of the classification means that we can place confidence limits on our results. Secondly, the ANN supplies explicit probabilities for the classification of each individual. This provides the potential to "clean" our datasets by removing those individuals who classify with very high (or very low) probability. Thirdly, the ANN is capable of being applied to more general datasets where non–linear signals are significant.

It is noteworthy that the supervised learning methods are able to classify individuals from two populations within Scotland. One would expect sufficient gene flow to occur within this region to homogenise the populations. The differences detected are not necessarily due to ancestry, but may be a consequence of the fact that the two population samples were drawn from different datasets, genotyped on different platforms, at different sites. These differences, whatever their origin, are nevertheless too small to detect using PCA, but in many applications the presence of such differences may be of critical importance.

The behaviour of our ANN in the presence of significant non–linear effects remains to be investigated; one possible target is the common disease common variant (CDCV) model of complex diseases. These are associated with many common genetic variants, each of individually small effect. Interactions between

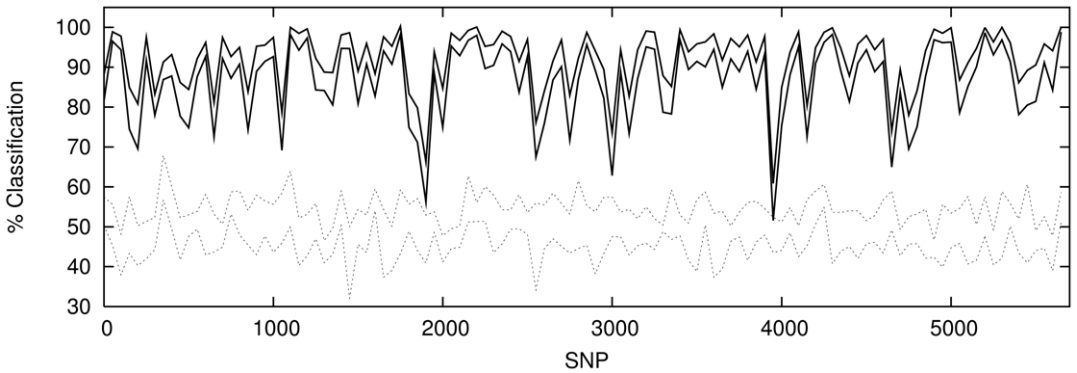

Figure 13. ANN classification with windows of 50 contiguous, non-overlapping SNPs for P1 against P2 (solid lines) with classification results for a sample of P1 against P1 (dotted lines) shown for comparison.
doi:10.1371/journal.pone.0014802.g013

these variants are likely to result in non–linear effects suitable for study with ANNs.

We suggest, on the basis of the evidence presented in this paper, that supervised learning methods have a useful role to play in genetic applications where we are interested in differences between pre–defined groups of individuals. Possible applications include population genetics, case–control studies and quality control for genetic data gathered at different sites or on different platforms.

Additional Information

Software. The LIBSVM library of SVM routines is publicly available [17]. The MemSys algorithms can be made available for academic use. We have developed an interface to both the MemSys and LIBSVM packages for our specific genetic application and are currently developing it for more general applications. We would be happy to collaborate with interested parties to facilitate this development process.

Supporting Information

Figure S1. Intra- population projection of the P3 population (5739 SNPs, $p = 0.022$), along the two most significant axes. It is clear that despite the nominally significant p-value, the two sub-populations fail to separate along these axes.
Found at: doi:10.1371/journal.pone.0014802.s001 (0.02 MB EPS)

Figure S2. Inter-population projection of the P1 and P2 population along the first most significant set of axes for each value of N. F_{ST}(crit) is never exceeded and the populations do not separate.
Found at: doi:10.1371/journal.pone.0014802.s002 (0.02 MB EPS)

Figure S3. Inter-population projection of the P1 and P2 population along the second most significant set of axes for each value of N. F_{ST}(crit) is never exceeded and the populations do not separate.
Found at: doi:10.1371/journal.pone.0014802.s003 (0.02 MB EPS)

Figure S4. Inter-population projection of the P1 and P2 population along the third most significant set of axes for each value of N. F_{ST}(crit) is never exceeded and the populations do not separate.
Found at: doi:10.1371/journal.pone.0014802.s004 (0.02 MB EPS)

Figure S5. Inter-population projection of the P1 and P3 population along the first most significant set of axes for each value of N. The populations separate as F_{ST}(crit) is exceeded.
Found at: doi:10.1371/journal.pone.0014802.s005 (0.02 MB EPS)

Figure S6. Inter-population projection of the P1 and P3 population along the second most significant set of axes for each value of N. The populations separate as F_{ST}(crit) is exceeded.
Found at: doi:10.1371/journal.pone.0014802.s006 (0.02 MB EPS)

Figure S7. Inter-population projection of the P1 and P3 population along the third most significant set of axes for each value of N. The populations separate as F_{ST}(crit) is exceeded.
Found at: doi:10.1371/journal.pone.0014802.s007 (0.02 MB EPS)

Table S1. Parameters of the reduced dataset used for analysis.
Found at: doi:10.1371/journal.pone.0014802.s008 (0.00 MB TXT)

Table S2. PCA results for inter-population tests. P_R and P_C are the reference and comparison datasets, M_R and M_C the respective sample sizes and N the number of SNPs used. F_{ST}(crit) is the value of F_{ST} at which the phase transition is expected. \hat{F}$_{ST}$ is the estimate of the F_{ST} and SE is its standard error. Pval is the ANOVA p-value. The 50 SNP and 500 SNP sets were a contiguous set starting from the 1000th data point along the chromosome. Note the sharp drop in p-value at the BBP transition when \hat{F}$_{ST}$ exceeds F_{ST}(crit).
Found at: doi:10.1371/journal.pone.0014802.s009 (0.00 MB TXT)

Acknowledgments

We thank the individuals and families who contributed data to the International Schizophrenia Consortium. We are grateful to the reviewers for their constructive comments. We thank the members of the Statistical Genetics Unit in the Neuropsychiatric Genetics Group for helpful comments and advice at all stages of this work. We acknowledge Anthony Ryan for reviewing and commenting on this manuscript. The authors would like to acknowledge support from the Cambridge Centre for High Performance Computing where this work was carried out, and also to Stuart Rankin for computational assistance. Additionally we acknowledge Steve Gull for useful discussions and for the use of MemSys in this application.

Members of the International Schizophrenia Consortium:

Trinity College Dublin Derek W. Morris, Colm O'Dushlaine, Elaine Kenny, Emma M. Quinn, Michael Gill, Aiden Corvin;

Cardiff University Michael C.O'Donovan, George K. Kirov, Nick J. Craddock, Peter A. Holmans, Nigel M.Williams, Lucy Georgieva, Ivan Nikolov, N. Norton, H. Williams, Draga Toncheva, Vihra Milanova, Michael J. Owen;

Karolinska Institutet/University of North Carolina at Chapel Hill Christina M. Hultman, Paul Lichtenstein, Emma F.Thelander, Patrick Sullivan;

University College London Andrew McQuillin, Khalid Choudhury, Susmita Datta, Jonathan Pimm, Srinivasa Thirumalai, Vinay Puri, Robert Krasucki, Jacob Lawrence, Digby Quested, Nicholas Bass, Hugh Gurling;

University of Aberdeen Caroline Crombie, Gillian Fraser, Soh Leh Kuan, Nicholas Walker, David St Clair;

University of Edinburgh Douglas H. R. Blackwood, Walter J. Muir, Kevin A. McGhee, Ben Pickard, Pat Malloy, Alan W. Maclean, Margaret Van Beck;

Queensland Institute of Medical Research Naomi R. Wray, Peter M. Visscher, Stuart Macgregor;

University of Southern California Michele T. Pato, Helena Medeiros, Frank Middleton, Celia Carvalho, Christopher Morley, AymanFanous, David Conti, James A. Knowles, Carlos Paz Ferreira, AntonioMacedo, M. Helena Azevedo, Carlos N. Pato;

Massachusetts General Hospital Jennifer L. Stone, Douglas M. Ruderfer, Manuel A. R. Ferreira,

Stanley Center for Psychiatric Research and Broad Institute of MIT and Harvard Shaun M. Purcell, Jennifer L. Stone, Kimberly Chambert, Douglas M. Ruderfer, Finny Kuruvilla, Stacey B. Gabriel, Kristin Ardlie, Mark J. Daly, Edward M. Scolnick, Pamela Sklar.

Author Contributions

Conceived and designed the experiments: MB CP. Performed the experiments: MB CP. Analyzed the data: MB EH CTO RS CP. Contributed reagents/materials/analysis tools: MB TISC AC DWM MG. Wrote the paper: MB CP.

References

1. Lao O, Lu T, NothNagel M, Junge O, Freitag-Wolf S, et al. (2008) Correlation Between Genetic and Geographic Structure in Europe. Curr Biol 18: 1241–1248.
2. Reich D, Thangaraj K, Patterson N, Price A, Singh L (2009) Reconstructing Indian Population History. Nature 461: 489–494.
3. International Schizophrenia Consortium website. Available: http://pngu.mgh.harvard.edu/isc. Accessed 2011.
4. Patterson N, Price A, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genetics 2: 2074–2093.

5. Reich D, Kumarasamy T, Patterson N, Price AL, Singh L (2009) Reconstructing Indian Population History. Nature 461: 489–494.
6. Baik J, Ben Arous G, Péché S (2005) Phase Transition of the Largest Eigenvalue for Nonnull Complex Sample Covariance Matrices. Ann Probability 33: 1643–1697.
7. Baik J, Silverstein JW (2006) Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models. J Multivariate Anal 97: 1382–1408.
8. Bailer-Jones C (2001) Automated Data Analysis in Astronomy. New Delhi: Narosa Publishing House. 363 p.
9. Leshno M, Ya Lin V, Pinkus A, Schocken S (1993) Multilayer Feedforward Networks with a Nonpolynomial Activation Function can Approximate Any Function. Neural Networks 6: 861–867.
10. Motsinger-Reif A, Ritchie M (2008) Neural Networks in Genetic Epidemiology; Past, Present and Future. BioData Min 1: 3.
11. Curtis D (2007) Comparison of Artificial Neural Network Analysis with Other Multimarker Methods for Detecting Genetic Association BMC Genet 8: 49.
12. North B, Curtis D, Cassell P, Hitman G, Sham P (2003) Assessing Optimal Neural Network Architecture for Identifying Disease Associated Multi-Marker Genotypes Using a Permutation Test, and Application to Calpain 10 Polymorphisms Associated with Diabetes. Ann Hum Genet 67: 348–356.
13. Seretti A, Smeraldi E (2004) Neural Network Analysis in Pharmacogenetics of Mood Disorders. BMC Med Genet 5: 27.
14. Penco S, Grossi E, Cheng S, Intraligi M, Maurelli G, et al. (2004) Assessment of the Role of Genetic Polymorphism in Venous Thrombosis Through Artificial Neural Networks. Ann Hum Genet 69: 693–706.
15. Gull S, Skilling J (1999) Quantified maximum entropy: MemSys 5 Users' Manual. Royston: Maximum Entropy Data Consultants Ltd.
16. Jaynes E (2003) Probability Theory: The Logic of Science. Cambridge: Cambridge University Press. 727 p.
17. Chang C, Lin C (2001) LIBSVM: A Library for Support Vector Machines LIBSVM. Available: http://www/csie.ntu.edu.tw/cjlin/libsvm. Accessed 2011.
18. The International Schizophrenia Consortium (2009) Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder. Nature 6: 748–52.
19. Purcell S, Neal B, Todd-Brown K, Thomas L, Ferreira Mal (2007) PLINK: A Toolset for Whole-Genome Association and Population-Based Linkage Analysis. Am J Hum Genet 81: 559–575.
20. Nelis M, Esko T, Magi R, Zimprich F, Zimprich A et al (2009) Genetic Structure of Europeans: A View from the North-East. PLoS ONE 4(5): e4572.